EXPENDITURE MODELING WITH A MIXTURE OF LOGNORMAL DISTRIBUTIONS

Stanislav Kolenikov[†],
University of North Carolina at Chapel Hill
and Centre for Economic and Financial Research, Moscow,
and Serguei A. Aivazian,
Central Economics and Mathematics Institute, Russian Academy of Sciences

Keywords: Finite Mixture, Income Distribution, Missing Data, Maximum Likelihood, Parametric Bootstrap, RLMS.

AMS subject classification: primary 62P20.
JEL classification numbers: I32, C13, C15, D31, P29.

Motivation

The adequate evaluation of success of market reforms in transition economies necessarily includes the assessment of the reform social cost, including welfare redistribution. The main source of information on the distribution of income, expenditures and wealth are population surveys [1,2]. Various distortions and deficiencies of the available survey micro data complicate this assessment. Because of wage arrears, as well as high shares of informal economic activities, including home production, the welfare of a household is better represented by (per capita) expenditures than by the officially reported income. Besides, survey participation rates tend to differ in different welfare groups. One of the manifestations of those deficiencies is a huge discrepancy between the mean income as found from the macroeconomic statistics, and one found from survey data. For the time period analyzed in this paper, the macroeconomic mean income for the Q4 1998 as reported in [3] is 1211 rub., while the sample mean from the raw data [2] is 913 rub.

The distributional model currently used by the Russian statistical authority, Goskomstat (The State Committee on Statistics) is the lognormal distribution [4], for which the location parameter (mean or mode) is estimated from macroeconomic trade statistics, and the variance parameter is estimated from sample income data.

We propose several refinements to this model. The first one is to use expenditure information that seems to represent the household financial situation better than income. The second is to approximate the shape of expenditure distribution by a univariate mixture of lognormal components. Such a model can be estimated by the maximum likelihood method from survey data, with special attention paid to the choice of the appropriate number of the mixture components. Third, we introduce weights to account for propensity to avoid disclosing income information. Finally, having estimated the above model, we use a parametric bootstrap to reconstruct the observations from the range of very high expenditures not touched upon by the sample. The estimates of the expenditure distribution thus obtained are used to construct popular inequality and poverty indices. The results suggest that unadjusted estimates of income inequality and poverty (including the officially reported poverty rates and the values of Gini index) might be seriously biased downwards.

Assumptions and Hypotheses

The following assumptions and hypotheses are used throughout the analysis.

Hypothesis H_1 states that the per capita expenditures distribution of Russian households can be adequately described by a mixture of lognormal laws. This hypothesis can be verified by fit criteria such as the Pearson χ^2 or Kolmogorov-Smirnov test (the latter is known to have low power when there are parameters to be estimated though). A justification for such discrete mixture is that the contemporary Russian society is believed to be stratified into several income groups, including “old economy” workers, “new economy” workers, and entrepreneurs, with incomes varying about an order of magnitude between groups. Assuming that the distribution within each group is lognormal, the discrete mixture is a reasonable approximation once the groups are well separated. The hypothesis H_1 provides a flexible not-so-parametric approach to density estimation.

Hypothesis H_2 states that the probability of household refusal to participate in the official budget survey is a function of its social, economic, and geographical characteristics. This hypothesis was suggested by E. B. Frolova, the Head of Living Standards Department of Goskomstat, and was taken from field experience. As long as we use panel data (see description below in the Data section), we can find both fi-

[†] Corresponding author. Mail address: 117 New West, Cameron Ave, UNC CB#3260, Chapel Hill, NC 27599-3260, US. E-mail: skolenik@unc.edu

nancial and demographic characteristics for some of the households that declined to participate in the survey. We then fit a logistic regression to those available data, and significance of this regression indicates support for hypothesis H_2.

Assumption H_3 states that in the lognormal mixture model, the coefficient of variation is constant across components, or equivalently that the variance of logarithms is constant. This assumption simplifies estimation, since if we assume that the variance can vary across the mixture components, then the estimation procedure may fail to converge [5].

Finally, the assumption H_4 states that there is a latent range of expenditure unobserved in surveys at the upper of end of expenditure distribution, and the population expenditure distribution in this range is again lognormal with a shift parameter $x_{(n)} = \max\limits_{1 \le i \le n}\{x_i\}$ where x_i is the per capita expenditure of *i*-th household. Assumption H_3 of constant variance serves as an identifying one: having estimated the variance in the observed data range, the *same* variance can be used for the latent component, also.

The first two hypotheses H_1 and H_2 are verifiable, while the latter two H_3 and H_4 are identification conditions.

Thus, the density of the expenditure distribution in our model can be written as follows:

$$f(x\,|\,\Theta) = \sum_{j=1}^{k} q_j \frac{1}{\sqrt{2\pi}\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} +$$

$$q_{k+1} \frac{1}{\sqrt{2\pi}\sigma_j \cdot (x - x_0)} e^{-\frac{(\ln(x - x_{(n)}) - a_{k+1})^2}{2\sigma_j^2}} \qquad (1)$$

where the components of the parameter vector $\Theta = (k;\, q_1,\ldots,q_{k+1};\; a_1,\ldots,a_{k+1};\; x_0;\; \sigma_1^2,\ldots,\sigma_{k+1}^2)$ are: q_j, j=1,…,k+1, the component weights in the mixture, a_j are the component means, σ_j, the standard deviation of logs in the j-th component (we assume $\sigma_1 = \sigma_2 = \ldots = \sigma$ according to H_3); and $x_0 = x_{(n)}$ is the largest observed expenditure. The number of observed components is k (unknown), and the k+1-st component is unobserved. The parameters of the observed components can be estimated by maximum likelihood procedures, while those of the unobserved one, from additional macroeconomic data. We shall also need to assume that the population share of this latent component is much smaller than any of the ones estimated.

Data

The only publicly available household data for Russia is the Russian Longitudinal Monitoring Survey (RLMS) run by the Carolina Population Center at University of North Carolina at Chapel Hill[1]. The project started in 1991, but the sample quality was found to be poor, so a new sample was created in 1994. The new sample used a multistage clustered design with 38 strata, of which 3 are self-representative metropolitan areas, and 1 PSU per stratum. The original sample consists of 4718 households, and there are 3600–3800 households (~10 ths. individuals) actually participating in each round. The data is collected in the fourth quarter of the year (Round V, 1994; Round VI, 1995; Round VII, 1996; Round VIII, 1998; Round IX, 2000). This research used Round VIII data.

The RLMS questionnaire contains expenditures for a large number of goods and services. This data can be aggregated to broad categories of goods and services, and to total expenditures. The raw data include a wide range of items, though the time span in each category might be different. The expenditures for food (~60 items) are based on weekly reports; fuel, services (with a breakdown to about 10 items), rent, club payments, insurance premia, savings and credits have a one month window; non-food consumer goods and durables expenditures are measured on the quarterly basis. RLMS also traces annual home production, as well as intermediate expenditures for subsistence plots. All those data are rescaled on monthly basis and published separately from the raw data. We used this “cleaned” data in our analysis.

Auxiliary data on refusals were used in deriving household sample weights that included codes of the survey results, i.e. whether the survey was conducted, and if not, why, with a breakdown into about three dozen main reasons.

Estimation

The first stage of the estimation is procedure is the estimation of the sample weights. A logistic regression model was formulated that used the non-response variable as the dependent variable, and mean over all available years household expenditures (deflated with the standard RET deflator[2]), household head education level, and rural / urban / metropolitan areas dummies as regressors. There were 29 households (out of total 4239 households where the survey was conducted at least once in four rounds V-VIII) that stated they do not want to be surveyed be-

[1] http://www.cpc.unc.edu/rlms/

[2] Russian Economic Trends, publication of Russian-European Centre for Economic Policy, Moscow, Russia. See http://www.recep.org.

cause they did not want to disclose their income information. The pooled number of refusals across all refusal codes is 795. Thus, there are several possibilities for the dependent variable: a dummy for refusal because of the financial considerations, dummy for all refusals, and the proportion of time the household refused to participate in the survey. The first regression with the “don’t want to tell our income” dummy turned out to produce insignificant results, most likely because of a very low proportion of such responses. The one with the “ever refused” dummy, however, did produce sensible results summarized in Table 1.

Table 1. Non-response probabilities.

Prob[refrain from survey at least once]=logit(•)

log(mean expenditure)	Urban, metro	Education
0.399 (0.079)**	++	--

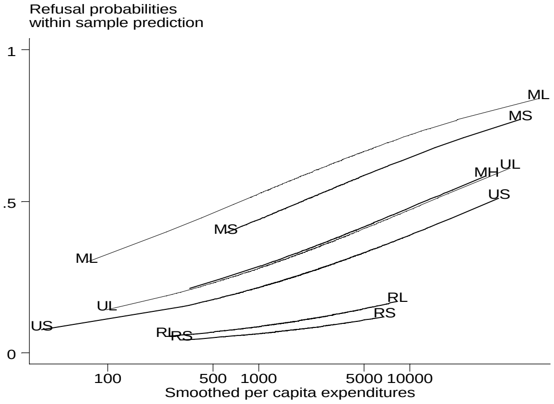

Figure 1. Response probabilities for population groups. Legend: R, rural; U, urban; M, metropolitan; H: higher education; S: secondary education; L: low education, less than 8 years of schooling.

We found that *ceteris paribus*, the effect of household expenditures is positive and significant. Households in the urban areas, and especially in the metropolitan areas (self representing strata of Moscow, St. Petersburg, and Moscow suburbs), are more likely to drop out of the sample, as opposed to the rural households, even controlling for incomes that are higher in the met areas. Households with better-educated heads refuse less often to answer, on average.

The weights obtained from the above logistic model were combined with the post-stratification weights provided along with the raw data. The effect of the weighting was not very substantial – see Table 3 below.

The mixture model (1) (or rather the observed part of it) can be estimated from the available micro data by using maximum likelihood procedures [5,6] or the EM-algorithm [7,8,9]. The weights obtained earlier can be used in the maximization procedure, as well.

One of the most important questions in this maximization problem is the estimation of the number of components k. A natural suggestion would be to consider a sequence of nested hypotheses on the number of components H_u: $k=u$, $u=1,2,\ldots$, and test them one by one with a likelihood ratio test, H_1 against H_2, H_2 against H_3, etc., until one fails to reject the increase in the number of components. It turns out however [10,11] that the likelihood ratio statistic does not have a convenient χ^2 distribution, but rather a mixture typical for estimation on the boundary.

To avoid the complications, information criteria or goodness of fit tests can be used to choose the appropriate k. We used the standard AIC and SBIC criteria, as well as another one of the family, ICOMP [12]. We also used Pearson χ^2 test to

Table 2. The mixture estimation results (9176 observations).

# of mixture comp-s / times converged	Log likelihood at maximum	Akaike criterion (AIC)	SBIC	ICOMP	Goodness of fit test $\chi^2(d.f.)$	p-value	$\hat{\sigma}^2$	Component parameters $\{a_i,q_i\}$
1 / 20	-11684.61	23373.21	23387.46	.	152.54(11)	10^{-27}	0.865	6.343
2 / 9	-11618.34	23244.67	23273.17	23244.78	96.39 (9)	10^{-16}	0.826	6.370: 98.9% 3.914: 1.1%
2 / 10	-11633.65	23275.29	23303.79	23278.39	98.24 (9)	10^{-17}	0.838	6.326: 99.36% 8.968: 0.64%
3 / 18	**-11532.21**	**23076.42**	**23119.17**	**23078.59**	**58.30 (7)**	$\mathbf{3\cdot10^{-10}}$	**0.756**	**6.340: 95.8%** **8.282: 2.3%** **4.159: 1.8%**
4 / 11	-11520.49	23056.98	23113.97	23058.96	58.40 (5)	$3\cdot10^{-11}$	0.716	6.297: 90.96% 7.618: 6.54% 4.235: 2.29% 9.790: 0.21%
5 / 8	-11515.97	23051.95	23123.19	23049.66	52.27 (3)	$3\cdot10^{-11}$	0.684	6.294: 87.62% 7.562: 8.40% 4.652: 3.54% 9.766: 0.23% 3.022: 0.22%

Source: authors’ calculations based on the RLMS data, with post-stratification and non-response adjustment weights.

compare the actual distribution with the hypothetical mixture with estimated parameters.

Stata 6 statistical software [13,14] was used throughout the analysis. The particular advantage of this software for our purposes is its open-end likelihood maximization procedure [15] that allows maximizing likelihood functions specified by the user. A Stata module `denormix` was developed that performs the numeric maximization using the ML procedure as well as a number of diagnostic tests that help determining the optimal number of components. The software is available from the corresponding author's webpage[3].

The results for 20 runs of the maximization procedure are reported in Table 2. We decided that the model with k=3 (and thus six parameters to be estimated) provides the best results, even though the p-value of the goodness of fit test is still very low. Models with a lower number of components do not fit data well (and there are multiple local maxima for the case of two components), while in those with a larger number of components, the maximization procedure failed to converge. It either was stuck in a flat region, or converged to a model with a smaller number of components (i.e. one of the estimated components coincided with another). The model seems to be misspecified even with fairly large number of components, as the p-values of the goodness of fit test suggest. We attribute this to the "large sample curse". Our experience with other unrelated samples of several hundreds observations shows that the estimation of mixture model may result in p-values of about 10%.

Economic Interpretations

We see from Table 2 that all estimated mixture models have a dominant modal component, and a number of components with rather small population shares at the tails. It thus seems that model misspecification is related to failure to fit the tails. The graphical analysis of the CDF curves (not reported here) shows that this is indeed the case, with the fit lacking mainly at the upper part of distribution. For some applications, lack of fit in tails can be tolerable, but it is crucial for poverty and inequality indices that are related to the tails of the expenditure distribution.

We are interested in estimating the following popular quantities: the Gini coefficient [16,17] and Foster-Greer-Thorbecke indices [18] with exponents 0 (head count ratio) and 2 (poverty depth). While the latter indices can be computed directly from mixture parameter estimates, using the expression for the incomplete moments of the lognormal distribution [19]:

$$\int_0^z x^m dF = \mu_k \Phi(\frac{\ln z - a - m\sigma^2}{\sigma}) \qquad (2),$$

$$\mathrm{E}x^m = \mu_m = \exp[ma + m^2\sigma^2/2] \qquad (3),$$

and are insensitive to the particular choice of the latent component parameters, the Gini coefficient cannot be expressed as a function of distribution moments. In its sample form, it is a linear combination of the order statistics. Thus to proceed to the estimation of inequality indices, we need to reconstruct the latent component by using the hypothesis H_4.

From (1)–(3), we can establish that the mean of the distribution (1) is

$$\mu = \sum_{j=1}^{\hat{k}} \hat{q}_j \, e^{\frac{1}{2}\hat{\sigma}_j^2 + \hat{a}_j} + q_{\hat{k}+1}\left(x_0 + e^{\frac{1}{2}\sigma_{\hat{k}+1}^2 + a_{\hat{k}+1}} \right) \qquad (4)$$

The first term in the sum, up to a straightforward scaling of the weights, is the mean of the estimated distribution in its observed range (reported in the second column of the Table 3 below). The mean expenditure μ is equated to that derived from macroeconomic statistics (column 3 of Table 3). After applying the identifying assumption H_3, the problem is reduced to the choice of two parameters such that

$$\mu(q_{\hat{k}+1}, a_{\hat{k}+1}) = \mu^{\mathrm{macro}} \qquad (5)$$

After choosing a particular point on this curve, we use parametric bootstrap from the estimated distribution to get the Gini coefficient. It turned out that the choice of the latent component parameters did not have much effect on the resulting number, so we used $a_{\mathrm{k+1}}$=13 with corresponding share $q_{\mathrm{k+1}}=4.3\cdot 10^{-4}$.

The results are reported in Tables 3 and 4, and the graphical representation of inequality comparisons by Lorenz curves[4] is given at Fig. 2. We can see that the lognormal model predicts the lowest inequality (the upper curve), with the model without the latent component and raw (unweighted) data Lorenz curves somewhat below it, and a drastically different curve evidencing much higher inequality, for the model with the latent component included. The value of the Gini coefficient of about 0.6 suggests that Russia is a country with very high inequality (like that in Brazil or South Africa). The values in high 30s – 40s are more typical for countries like the US,

[3] http://www.komkon.org/~tacik/stata/. The program is also available in the Statistical Software Components archive at Boston College; http://ideas.uqam.ca.

[4] Lorenz curve $L(p)$ of an income distribution is a proportion of income that p percent of population with the lowest incomes receive. The Gini coefficient is twice the area between the Lorenz curve and the 45° line.

Table 3. The results of the distribution calibration and inequality comparisons for Russia, Q4 1998.

Mean expenditure, ths. rbs.			Gini index	
Raw	Cali-brated	With the latent compo-nent	Raw data	With the latent compo-nent
0.913	0.952 (+2%)	1.211 (+29%)	0.478, 0.380 [3]	0.610

Source: [3]; authors' calculations based on the RLMS data. The second column is the mean with the weights accounting for post-stratification as well as non-response probability.

Table 4. Poverty indices for Russia, Q4 1998.

	Official figure [3]	Lognor-mal model	Mixture model	Sample
Poverty rate	28,4	52,5	52,8	53,9
Poverty depth (FGT(2))	N/A	0,139	0,130	0,137

Source: [3]; authors' calculations based on the RLMS data. Poverty rate is 636 rubles [3].

and values in high 20s–low 30s, for Nordic countries.

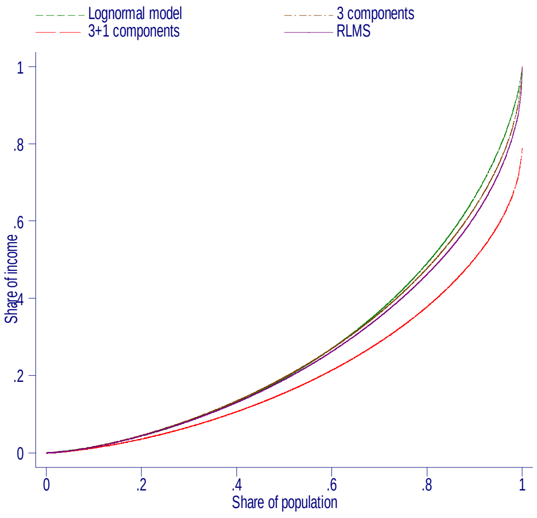

Figure 2. Lorenz curves for different models.

Finally, the analysis of sensitivity of the results with respect to some of the model assumptions was performed. The assumption H_4 of the existence of the latent component can be reformulated as follows: all of the discrepant expenditures are due to this unobserved component. An alternative assumption can be that all households can be observed, and the discrepancy is due to misreporting (i.e. households fail to report their true expenditure). There is a continuum between the two, and in the simplest form, we can assume that each household underreports its expenditures by a fraction of λ. So in terms of λ, the previous analysis assumed λ=0. The sensitivity analysis included recalculation of (4) corrected for λ, and drawing 20 parametric bootstrap samples of size 400000 from the estimated distribution. The sample size was chosen so that the smallest component will still be represented by at least a hundred observations.

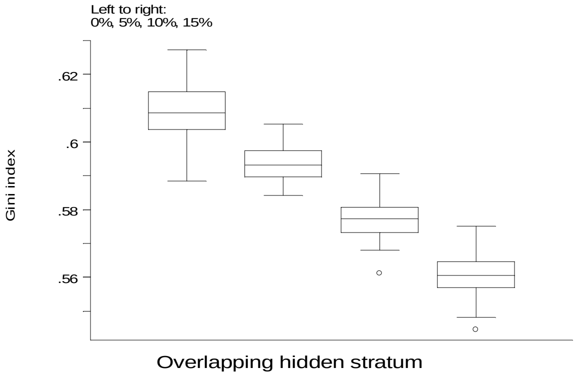

Figure 3. The Gini coefficient sensitivity to misreporting.

The results of the sensitivity analysis are presented on Fig. 3. Despite rather large sample size, the variability in the value of Gini coefficient across simulations is quite pronounced. Still, the downward trend is also notable, as we move to higher λ's, which corresponds to the gradual transition between the "3 components" and "3+1 components" curves on Fig. 2. Probably the assumption of some 5 to 10 per cent reporting error is more reasonable than one of no error at all, so a more realistic value of the Gini coefficient is 0.56–0.59.

Remarks

There are a number of nuances in the budget data analysis that were not pursued here. In particular, one of the options was to employ equivalence scales to account for the difference in needs of households of different size and composition. It can be argued that economies of scale are not substantial in Russia, as the rent and other fixed expenses of a household are pretty low. This and many other points are discussed in [20] that goes into deeper discussion of both statistical and economic issues raised by this research.

Conclusion

This paper proposed to estimate expenditure distribution by the finite mixture of lognormal components. The maximum likelihood is the most appropriate way to estimate the parameters of such mixture. By using Russian data for Q4 1998, we showed that the mixture model, with additional weights accounting for non-response related to particular factors, performed better than a standard lognormal one. The mixture with three components was found the most appropriate. A parametric bootstrap was suggested to recover the observations in the unobserved range of expenditures. The results differ substantially from the figures reported by Russian statistical authorities. In particular, the estimate of the Gini coefficient obtained from our model is 0.56-0.59, which is much greater than the official one of 0.38.

Acknowledgements

The authors are grateful to Michael Beenstock, Constantin Colonescu, John Earle, David Guilkey, Tom Mroz, Klara Sabirianova, Anthony Shorrocks, and Michel Sollogoub for substantial comments and references; to all participants of EERC workshops for useful discussions; to Elena Frolova for discussions and help with data collection; to Carolina Population Center and Polina Kozyreva for access to the RLMS data. The research was supported by Economics Education and Research Consortium (Grant No. 99-113), the Russian Humanitarian Scientific Foundation (Grant No 99-02-00270) and by the Moscow Public Science Foundation from the funds provided by the U.S. Agency for International Development-USAID (Grant No 020/1-01-O). The presentation at the JSM 2001 was sponsored by the ASA Sections of Government Statistics, Social Science Statistics, and Survey Research Methods. Views reflected in this paper are those of the authors only.

References

[1] Chernysheva, T.M. (2000). *Theoretical and practical sampling methods for household budget surveys.* Goskomstat, Moscow (in Russian. Russian title: Teoreticheskie i prakticheskie osnovy vyborki dlya obsledovaniya budgetov domashnikh khozyastv).

[2] Mroz, T., L. Henderson, and B.M. Popkin (2001). *Monitoring Economic Conditions in the Russian Federation: The Russian Longitudinal Monitoring Survey 1992-2000.* Report submitted to the U.S. Agency for International Development. Carolina Population Center, University of North Carolina at Chapel Hill.

[3] Goskomstat (1998). *Social and economic state of Russia, January--December*. Moscow (in Russian. Russian title: Sotsial'no-ekonomicheskoe polozhenie Rossii).

[4] Goskomstat (1996). *Methodological regulations on statistics*. Pt. 1. Goskomstat, Moscow (in Russian. Russian title: Metodologicheskie polozheniya po statistike).

[5] Day N.E. Estimating the Components of a Mixture of Normal Distributions. — *Biometrika*, **56**, pp. 463—474 (1969).

[6] Basford, K. E., and G. J. McLachlan (1985). Likelihood Estimation with Normal Mixture Models. *Appl. Stat.*, **34**, 282–289.

[7] McLachlan, G.J., and D. Peel (2000). *Finite mixture models*. New York, Wiley.

[8] Little R.J.A., and D.B.Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

[9] Rudzkis R., and M. Radavicius (1995) Statistical Estimation of a Mixture of Gaussian Distributions. *Acta Applicandae Mathematical*, **38**.

[10] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat*., **36**, 318–324.

[11] Feng, Z. D., and C. E. McCulloch (1996). Using bootstrap likelihood ratios in finite mixture models. *J. of the Royal Stat. Soc.*, B, **58**, 609–617.

[12] Bearse P., H. Bozdogan, and A. Schlottman (1997). Empirical Econometric Modelling of Food Consumption Using A New Informational Complexity Approach. *J. of Applied Econometrics*, **12**, 563--592.

[13] StataCorp. (1999). Stata Statistical Software: Release 6. College Station, TX. Stata Corporation.

[14] Kolenikov, S. (2001). Review of Stata 7. *J. of Applied Econometrics*, **16** (5), 637–646.

[15] Gould W., and Sribney W. *Maximum Likelihood Estimation with Stata*. Stata Corp. (1999).

[16] Atkinson A.B. On the Measurement of Inequality. *J. of Economic Theory*, **2**, 244-263 (1970)

[17] Lambert, P. (1989). *The Distribution and Redistribution of Income.* Blackwell.

[18] Foster J., Greer J., Thorbecke E. A Class of Decomposable Poverty Measures. *Econometrica*, **52** (3), 761—766 (1984).

[19] Aitchison J., Brown J. A. C. *The Lognormal Distribution*. Cambridge Univ. Press (1963).

[20] Aivazian S.A., and S. Kolenikov (2001). Poverty and Expenditure Differentiation of the Russian Population. *EERC Working Paper*, No. 01/01. EERC, Moscow. Also available online through the RePEc system, http://repec.org, http://ideas.uqam.ca.