ECONOMICS EDUCATION AND RESEARCH CONSORTIUM RUSSIA

РОССИЙСКАЯ ПРОГРАММА ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

S. A. AIVAZIAN, S. O. KOLENIKOV

POVERTY AND EXPENDITURE DIFFERENTIATION OF RUSSIAN POPULATION

Interim report, November 1999

1 CONTENTS

2 ABSTRACT

The problem of poverty and inequality measurement in contemporary Russian society is considered in the framework of the general problem of social tension reduction via rational social assistance system. We argue that features specific to Russian transition stipulate poverty indicators (e.g. Foster-Greer-Thorbecke family) to be calculated on the basis of *expenditures* rather than incomes as it is usually done. These features are also accustomed for in the proposed econometric model of per capita expenditure distribution. The model includes special methods to calibrate, or to adjust, the distributions obtained from the official budget surveys' statistics. The preliminary results of the empirical approbation of the technique are reported which use the RLMS (Rounds 5–8) statistical data as well as budget surveys of Komi Republic, Volgograd and Omsk oblasts.

3 THE STATEMENT OF THE PROBLEM

Various measures of poverty and expenditure inequality serve as the key indicators of the quality of social policy and are used, in particular, to target social assistance to the aimed at reduction of the social tension in the society.

The indicators and estimation procedures used nowadays by Russian statistical authorities ([1]–[3]), as well as those proposed by other researchers ([4]–[6]), are based on the household budget survey data and suffer from certain drawbacks, even after correction for the macroeconomic balances of income and consumption[1)] and/or equivalence scales.

We see the following reasons to explain those distortions:

(i) The two-parameter lognormal income distribution model used by the statistical authorities (Goskomstat) for modeling regional and Russian income distribution is not valid. The main distortions are located on the tails of the distributions, while, evidently, the main contribution to inequality and poverty indicators are due to the distribution tails.

[1)] Some estimations (e.g., [3], [4], [6]) show that the ratio of the average income in the top quintile to that in the bottom one is biased downwards by the factor of at least 2, while the proportion of households with per capita income below subsistence level (i.e., poverty rate), by the factor of 1.5–2 for Russia as a whole and 2–5, for some regions.

(ii) The model has limited applicability even when it is adjusted to comply with both known social and demographic household composition and the level of average household per capita income obtained from macroeconomic income and expenditure balances [3]. The shape (lognormal) and parameters of the distribution (mode) are assumed to be constant.

(iii) Distribution approximation and weighting (calibration) techniques proposed by other researchers (e.g. [4], [5]) also tend to lead to substantial distortions. They do not allow for estimation of neither the share in the population nor the structure of the *unobserved* range of "rich" and "ultra rich" households as weighting only gives weights to the observed households, but *does not generate observations from the latent part* of distribution.

(iv) Head-count ratio, or the proportion of households with per capita expenditure below subsistence level, is usually analyzed as a proper poverty measure despite the goal of analysis ([7]–[9]). However, the choice of poverty indicator (or criteria to classify a household as poor) is to be determined by the final applied goal of economic analysis. There must be a set of criteria specific for the problem of targeted social assistance system development (e.g. [9]), and another set, to measure welfare, or social tension (e.g. [10]–[12]).

(v) The specific features of Russian transition economy prompt that **expenditure** rather than *income* is to be used for the purposes of poverty and inequality indicators calculation as well as dichotomy into poor or non-poor status of a household. If expenditures are used,

a) the problem of wage arrears in a household is solved;
b) intentionally or non-intentionally hidden income, including income from shadow economy, is accounted for;
c) the concept of household welfare is appropriately generalized to include land (subsidiary plot) and property (real estate, private transportation means, jewelry, etc.).

(vi) The problem of the optimal, in terms of a specific poverty indicator (see (iv) above), allocation of resources addressed to targeted assistance has never been stated, let alone solved, in Russian economic theory and policy.

3.1. ***The aim and the main tasks of the project***

The goals of the project are determined by the desire of the project participants to overcome the aforementioned drawbacks (i)–(vi). In particular, we are aiming at: development of the methodology of econometric analysis of per capita expenditure distribution based on Russian budget survey data; construction of the main characteristics of poverty and welfare inequality of Russian population and their statistical assessment; and formulation and theoretical solution to the problem of optimal allocation of the limited amount of resource dedicated to targeted assistance for the poor.

In general, problem statements are necessitated by the above goals. In their aggregated formulation, the two main problems are as follows.

Task 1 is to infer from theory and approve empirically an interpretable econometric model of per capita expenditure distribution within a region or a country. This task would also include the development of identification methodology based on the sample budget surveys and macroeconomic balances of income and expenditure.

The solution to this task is to be constructed conditionally on the contemporary Russian economy status, which would require the following hypotheses to be proposed on the theoretical grounds and, if possible, approved for statistically:

- The first hypothesis H_1 concerns the shape of the distribution function;
- The second hypothesis H_2 concerns the probability of refusal of a household to participate in the budget survey conditional on its income;
- The third hypothesis H_3 states that the coefficient of variation of per capita expenditures is constant across all strata;
- The fourth hypothesis H_4 deals with the shape of the distribution of household per capita expenditures within the *unobserved* range of expenditures (right distribution tail, the richest population strata).

The detailed description and foundation for all these hypotheses will be given in the main part of the report.

Task 2 is to consider a broad class of poverty indices based on the per capita expenditure distribution, and formulate the problem of an optimal allocation of a limited resource S devoted to targeted social assistance to the poor with the objective function from this class.

The following family of poverty indices would be considered:

$$I\big(w(x),\, f(x)\big) = \int_0^{z_0} w(x)\ f(x)\, dx\ , \tag{1}$$

where $f(x)$ is the pre capita expenditures density function, z_0, poverty line, and weighting function $w(x)$ is supposed to be differentiable, decreasing and convex at $[0, z_0)$ (the latter property is due to transfer principle). Apparently, the family (1) includes, with the appropriate weighting $w(x)$, such popular measures as Foster-Greer-Thorbecke family of indices, Dalton class indicators, and poverty-line-discontinuous measures [13]–[15].

Let S is the amount given for targeted assistance less than the poverty gap. Denote $\varphi(x|S)$, the rule of allocation of this resource among population with per capita expenditures $x < z_0$ (e.g. distribution density), and $\tilde{f}(x|\varphi, S)$, the population per capita expenditure distribution density observed *after* realization of social assistance according to $\varphi(x|S)$. The ex post indicator value would thus be:

$$I\big(w(x),\, \tilde{f}(x|\varphi;\, S)\big) = \int_0^{z_0} w(x)\ \tilde{f}(x|\varphi;\, S)\, dx\, . \tag{1'}$$

Task 2 is then reduced to the identification of the $\varphi_0(x|S)$ such that (1') achieves its minimum, given $w(x)$ and S:

$$\varphi_0(x|S) = \arg\min_{\varphi} \int_0^{z_0} w(x)\ \tilde{f}(x|\varphi;\, S)\, dx\, . \tag{2}$$

It needs to be emphasized that this problem is considered here within the framework of the **concrete** project of the **longer-term** poverty combat (see [8] and [9] from the References of this interim report). Two propositions are worth mentioning in this context. First, the thesis of relatively high income mobility is not applicable for the category of the permanently poor (cf. Bogomolova T. *et. al.*, EERC, new project, December 1999): if they were mobile, they would not be *permanently* poor. Second, the main instruments of (longer-term and permanent) poverty alleviation are *direct transfers to the poor* rather than creation of incentive structure, though the latter are certainly more efficient for temporarily poor (e.g. the potential middle class with high educational standard).

4 LITERATURE OVERVIEW

Let us first discuss the sources where problems closed to our Task 1 (see above) were posed. The model of per capita expenditure distribution developed in this project is supposed to develop and modify the basic model of population per capita *income* distribution pioneered in [16]. The *modification* includes an introduction and statistical estimation of budget survey refusal probability (see H_2 above); the replacement of income by *expenditures* in the lognormal mixture model; and calibration of the existent observations and Monte Carlo generation of some additional data points unobserved in the sample on the basis of known macroeconomic balance of household expenditures supplemented by hypotheses H_2 – H_4.

The sources [4]–[6], [16] contain arguments which prove the validity of our critique (i)~(iv) in the first part of the proposal. Velikanova et al. in [2] describe an approach which is also based on the mixture of lognormal distribution, but this source neither provide econometric tools to analyze this mixture nor proposes any ways to reconstruct the unobserved data. The approach by Ershov and Mayer in [5] is based on polynomial density approximation and seems to be too formal. It does not allow for establishment of an interpretable model of the phenomenon studied and does not account for the latent expenditure range.

The main drawback of the approach by Suvorov and Ulyanova in [6] is inadequacy of the basic assumption on lognormality of income distribution though the authors do study a three-parameter model, as opposed to two-parameter Goskomstat model. Nevertheless, the authors a) analyze income, but not expenditures; b) do not provide any convincing arguments in favor of the basic assumption on the correct estimation of the *modal* income out of the Goskomstat budget survey sample (which is considered substantially biased even by Goskomstat specialists, let alone independent experts); c) propose a formal approximation technique of unknown parameters fitting. While the *economic analysis* of stylized facts of income redistribution in Russia of the 1990s does clarify the mechanism of formation of the right distribution tail unobserved in the Goskomstat budget surveys and is thus really serious, the above cited drawbacks of the approach can be quite heavily criticized.

Special attention needs to be paid to the work of Shevyakov and Kiruta [4] as well as to the differences of the approach of theirs from the one proposed in our project. Their work is currently the most serious attempt to describe *realistically* the regional per capita income distribution with the information contained in the Goskomstat budget survey data and macroeconomic "Population Income and Expenditure Balances" (a special balance routinely calculated by State Committee in Statistics, Goskomstat). The attempt is based on the non-parametric approach to density estimation and includes, in particular, a technique to eliminate the Goskomstat sample bias, as well as description of the procedure to aggregate the regional data corrected for regional deflators and equivalence scales. To our view, the main drawbacks of Shevyakov & Kiruta approach are as follows:

a) The weighting (calibration) technique proposed in [4], actually, ignores population beyond the maximum income observed, i. e., the right tail of the distribution. In our model, the tail is recovered by using the set of hypotheses H_1–H_4.

b) The immediate consequence of the previous critique point is a principally erroneous inference that "the excessive economic inequality is in whole caused by the excessive poverty". Given that the authors ignore the right tail, there cannot be any other result.

c) A seemingly attractive "non-parametricity" of the approach has, in fact, two serious drawbacks. First, the estimate of the per capita income distribution thus obtained is a *purely formal approximation* of the unknown distribution analyzed and *cannot be interpreted in understandable terms*. Second, the model is not at all suitable for prediction purposes.

d) To estimate poverty rate, wealth inequality and other welfare indicators, expenditure side is more appealing in Russian situation than income, as long as it removes inconsistencies related to wage arrears, hidden income, etc.

Let us know focus on the works related to the Task 2. First of all, worth mentioning are the World Bank project [8] and pilot programs [9]. They do accomplish a rightful attempt to assess poverty according to re-estimation of realistic household per capita income (termed 'potential consumption expenditures' in [8]). Both approaches, however, still suffer from significant drawbacks analyzed by Aivazian in [17]. Besides, the only poverty index used is head-count ratio (i.e. (1) with $w(x) \equiv 1$), and the problem of optimal allocation of social assistance is not stated (i.e. problem (2) is not solved for).

A comprehensive overview of poverty indicators is given in [18]. This work discusses, in particular, a special case of criterion (1), i.e., Foster-Greer-Thorbecke set of indices, and reports the sample statistics of quarterly budget surveys as of 1996. Still, the index calculation relies on income distribution and is not related to targeted assistance optimization.

Thus, neither *Russian* economic theory nor practice states and solves Task 2. Nevertheless, various aspects of this problem are addressed in the Western literature though most authors still rely on income rather than expenditure distributions ([15], [21]–[24]). In particular, [23] proves that with FGT indices featuring

$$w(x) = \left(\frac{z_0 - x}{z_0} \right)^{\alpha}, \qquad 0 \le x < z_0, \quad \alpha > 1 \tag{3}$$

the optimal solution to (2) is the pure strategy of giving the poorest people enough income to raise their income to the threshold $\bar{z}_0 < z_0$, where $\bar{z}_0$ is found from

$$N \left(\int_0^{\bar{z}_0} f(x)\, dx \right) \left(\int_0^{\bar{z}_0} (\bar{z}_0 - x)\ f(x)\, dx \right) = S \tag{4}$$

where N is total population. This strategy is referred to as “allocation of p-type” in [15] and [23] and implies that each person with income below $x < \bar{z}_0$ is to receive a subsidy $\bar{z}_0 - x$. An alternative option is the allocation of mixed-type when a part of S is used to raise the incomes of the poorest up to $\bar{z}_0$, and this part $S_1 < S$ enters the RHS of (4), and the rest of S, to raise the incomes of the richest among poor to z_0. It is proved in [23] that mixed strategy can be optimal only if $w(z_0) = \delta > 0$, i.e. that the underlying poverty index is discontinuous. These type of indices are referred to as ‘poverty-line-discontinuous, or PLD, measures’ in [15].

5 MODEL SPECIFICATION AND PRELIMINARY ESTIMATION RESULTS

5.1. Verification of the basic working research hypotheses

The solution to the above stated Task 1 is based on the theoretical inference and/or empirical testing of a number of working hypotheses.

- **Hypothesis H_1** states that the distribution of Russian households by per capita expenditures can be adequately described by a *mixture of lognormal distributions.* This hypothesis can be verified by a fit criteria. An example for 1996 data is [16].

  Theoretical reasoning for this hypothesis is as follows.

  a) Per capita expenditure ξ distribution within a homogeneous strata follows lognormal distribution with parameters $a = \mathbf{E}(\ln \xi(a))$ and $\sigma^2(a) = \mathbf{D}(\ln \xi(a))$. Here, homogeneity refers to similar income sources, geographical, social, demographic, and professional characteristics of its representatives.

  b) If society as a whole can be represented by a continuous (in terms of the average log expenditures a) spectrum of such strata, then under a certain though natural shape of the mixing function $q(a)$, the population distribution by per capita expenditures is reproduced to be lognormal.

  c) If continuity of the spectrum is violated (i.e. some strata are eliminated, or crowded out), or $q(a)$ is not monotonically decreasing as its argument a increases from the global average a_0, then the population lognormality holds no longer, and the distribution is transformed into a discrete-type mixture.

  Let us now discuss each of the postulates.

The first statement is quite widespread in income distribution studies and results from multiplicative shocks to expenditures (incomes, wages) within the strata. The data generating mechanism is described in [25] as applied to wages.

The second postulate follows from the fact that if the within-strata-average log expenditures $a = \mathbf{E}(\ln \xi)$ are distributed normally with parameters $(a_0; \Delta^2)$ (i.e. if $q(a)$ is normal), then the resulting distribution of expenditure logarithms $(\zeta = \ln \xi)$

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma(a)} e^{-\frac{(z-a)^2}{2\sigma^2(a)}} q(a)\, da$$

is a composition of normal distributions and thus normal itself. If $\sigma^2(a) = \sigma^2 = const$, then the parameters of the resulting distribution are $a_0 = \mathbf{E}(\ln\xi)$ and $\sigma_0^2 = \sigma^2 + \Delta^2$. This fact is mentioned and proved in [25].

The third statement is apparent in a degenerate situation when the number of points where the mixing function $q(a)$ is different from zero is finite: $a_1, a_2, \ldots, a_k$. The realistic distribution of expenditures in Russian economy is, of course, more complicated. But it nevertheless is characterized by a significant transformation of the mixing function $q(a)$. The transition period do not abolish the a) and b) postulates though affected the shape of $q(a)$.

- **Hypothesis H$_2$** states that the probability of the household to refuse to participate in the official budget survey is an increasing function of its per capita expenditures. This hypothesis can also be verified against the data such as RLMS ([26]) and some additional information from Goskomstat. This hypothesis was prompted by the Head of Living Standards Department of Goskomstat E. B. Frolova and was apparently implied by the field experience.
- **Hypothesis H$_3$** states that the *coefficient of variation* of the household per capita expenditures is constant across the social strata, i.e., is independent of the strata number; this hypothesis can also be verified by criteria of variance homogeneity ([16]). As long as incomes and expenditures $\xi(j)$ of population of *j*-th homogeneous strata are distributed lognormally with the parameters $a(j) = \mathbf{E}(\ln\xi(j))$ and $\sigma^2(j) = \mathbf{D}(\ln\xi(j))$ (e.g. [27]), the hypothesis H$_3$ is equivalent to:

$$H_3': \mathrm{Var}[\ln \xi(j)] = \sigma^2 = \mathrm{const}$$

The equivalence of H$_3$ and H$_3$' follows from the relation between the moments of the lognormal distribution:

$$\frac{[Var(\xi(j))]^{\frac{1}{2}}}{\mathbf{E}\xi(j)} = \left(e^{\sigma^2} - 1\right)^{\frac{1}{2}}$$

- Hypothesis H4 states that the population per capita expenditures x in the latent range of $x > \max_{1 \le i \le n}\{x_i\}$, where xi is per capita expenditures in the i-th household surveyed, and n, total number of households, can be approximated by three parameter lognormal distribution with a shift parameter $x_{(n)} = \max_{1 \le i \le n}\{x_i\}$ and logarithm variance $Var(\xi(k) = \sigma^2$ where σ2 is independent of strata and estimated from the observed strata (see hypothesis H3 above).

This hypothesis cannot be directly verified from the data available with any statistical criteria as long as the data from the required expenditure ranges cannot be observed. Hence, hypothesis H4 is not a *hypothesis in statistical terms*, but rather a model assumption based on Russian stylized, as well as statistically established, facts. It can be established ex ante by some theoretical considerations, and ex post, by matching the levels of observed characteristics with the model output. The following evidence from the researchers and specialists supports the relevance of this model assumption.

One of the real consequences of the USSR and it economic system disintegration is the formation of "new Russian" group from the communist, state bureaucracy and managerial elites. By benefiting from the privatization campaign, they managed to get access to the rent flows that can be considered a part of the national wealth which could (and was) sold on the domestic and world markets. Some calculations (e.g. [6]) show that market intervention of 0.2–0.3% of Russian national wealth per annum is equivalent to the increase of gross population income in Russia by 20–30%. Evidently, the major part of this income is distributed into this novo riche group of population, which can be classified as a separate stratum as long as its representatives are homogeneous by their social status and power. It is this stratum which is referred to in the hypothesis H4.

Typically, the right tail of income / expenditure distribution above some x0 beyond the distribution mode is approximated by Pareto distribution. This assumption, however, is only valid if the density function decreases monotonically for all x>x0 (as it is the case in a well-functioning economy). In our case, we cannot rule out multimodal shape of the distribution density function, in particular, a local maximum in the unobserved richest strata to the right of x0, as long as we consider super-rich as a separate stratum with its specific unimodal log-normal distribution.

The exclusion of the super-rich stratum from the econometric analysis would not affect, due to their focus properties, poverty indicators used in targeted social assistance design as the objective functions of the corresponding transfer allocation optimization problem. In calculating poverty indicators, only the left tail of the analyzed distribution is used, while H4 is used to assess the shape of the distribution density at its right tail.

If this super-rich stratum is taken into account, however, the inequality indicators (Gini coefficient, decile ratio, etc.) tend to increase significantly[1]. In turn, these inequality indicators measure social tension (see, for instance, [35], where inequality is assumed to be the main factor to explain crime rates). That is why expenditure inequality (and, consequently, the super-rich population) is paid so much attention in this project.

5.2. The main variables to be used

1) Gross per capita expenditures ξ of a randomly sampled (i.e., surveyed) household x_i.

Following [7], we shall define (with the time quantum of a quarter) gross pecuniary expenditures of a household as the sum of:

- $\xi^{(1)}$ — *quarterly consumption expenditures*, which is the sum of food products expenditures, alcohol, non-food private consumption goods and private services;
- $\xi^{(2)}$ — *interim consumption expenditures* (household expenditures for subsidiary land plot);
- $\xi^{(3)}$ — *the quarterly average of the net household capital accumulation* (acquisition of land and property, jewelry, construction and dwelling maintenance expenditures);
- $\xi^{(4)}$ — *the quarterly total of taxes paid and other obligatory payments* (including alimony, debt, club and public payments);
- $\xi^{(5)}$ — *cash in hands and net savings increase* (including currency and stock accumulation, bank deposits);
- $\xi^{(6)}$ — *estimate of monetary equivalent of the household produced products*.

All in all,

$$\xi = \sum_{l=1}^{6} \xi^{(l)} ,$$

where $\xi^{(l)}$, l=1,2,…,6 are as defined above.

[1] Estimations in [16] based on 1995—96 data show that Gini coefficient for Russia rises from 0.376 to 0.531, and the decile ratio, from 12.9 to 22.8, due to adjustment for the households refused to participate in surveys. It was assumed that “richer” population evades the survey partially, and “super-rich” avoids survey participation altogether.

The observed values of x_i, $x_i^{(1)}$, $x_i^{(2)}$, …, $x_i^{(6)}$ of the random variables $\xi, \xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(6)}$ are the results of the (Goskomstat [7], RLMS [26] or any other) survey of the i-th household.

2) Regional/national average per capita expenditures μ_{macro} defined from macroeconomic characteristics, namely, quarterly Goskomstat "The Population Income and Expenditure Balances" ([28]). μ_{macro} has the same structure as ξ but is defined from regional trade, tax, bank and security market statistics rather than surveys.

3) The proportion of households $p(x)$ with per capita expenditure level x who refused to participate in the survey in the given period. The sources of information are supposed to be Goskomstat and RLMS.

4) Social and demographic composition of the region (regional averages on household size, number/proportion of children, retired, etc.).

Let us now describe in some detail the RLMS and Goskomstat budget survey data that comprise the information base of our research.

(1) Russian Longitudinal Monitoring Survey data, Rounds V–VIII (see [26]). This is a stratified sample survey covering 3.5+ ths. households and 10+ ths. individuals in all areas of Russia. It is believed to be representative on the national level. The survey is regularly conducted in the late fall (October–December, with November as the reference month). The first four rounds of RLMS (1992–1993) are not considered representative of Russian population, and that is why the sample was re-designed in 1994. The unit of the survey is household; the questionnaires used are the household, the individual, and the children ones. The RLMS household questionnaire contain expenditures on 56 food products, about 10 aggregate durables, fuel, about 10 types of services, housing, other household expenditures, and savings. This data can be aggregated to large groups of goods and services, and to total expenditures, respecting the time periods to which the expenditures relate. Currently, we have used variables **`totexpr*`** from the data files labeled **`heexpd*`**. These data have been verified by RLMS staff and include the appropriately scaled data on

weekly expenditures on food, as well as expenditures on other items with time horizons from month to year.

(2) Household budget survey data as of Q3 1998 on three regions of Russia, namely, Komi Republic, Volgograd and Omsk oblasts, with a supplementary questionnaire ([11], the questionnaire is attached). According to Goskomstat methodology, the sample is constructed to be representative of the household types, except the collective households, on the basis of 1994 microcensus. During the quarterly budget survey, a household fills in twice in the quarter a two-week daily log of expenditures, two bi-weekly logs, and is exposed to a intermediate monthly survey. From this primary data, Goskomstat infers the following aggregate indicators: pecuniary expenditures ("**`denras`**" variable in the Goskomstat survey datasets; the sum of actual expenditures made by household members in the period of account; includes consumption and non-consumption expenditures); consumption expenditures ("**`potras`**" variable; the proportion of pecuniary expenditures directed to acquisition of consumption goods and services); final household consumption expenditures ("**`konpot`**" variable; consumption expenditures sans food products transferred outside the household, plus the in-kind household income, i.e. the sum of non-cash and natural intakes of food products and subsidies); household disposable resources ("**`rasres`**" variable; the sum of pecuniary resources, "**`denres`**" variable, i.e., pecuniary expenditures and nominal savings by the end of the period, and natural intakes, "**`natdox`**" variable). The budget surveys referred to were supplemented with the questionnaire on quality of life (see [11], the questionnaire is attached).

5.3. *Model description and parameter interpretation*

Denote ξ (ths. rub.) the yearly average expenditure of a randomly selected representative of Russian population, and ξ_j (ths. rub.), per capita expenditures of the representative of j th homogeneous stratum. According to hypotheses H_1 and H_4, the distribution density of the random variable ξ is described by the model of lognormal mixture:

$$f(x|\Theta) = \sum_{j=1}^{k} q_j \frac{1}{\sqrt{2\pi}\,\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} + q_{k+1} \frac{1}{\sqrt{2\pi}\,\sigma_{k+1}\cdot(x-x_0)} e^{-\frac{(\ln(x-x_0)-a_{k+1})^2}{2\sigma_{k+1}{}^2}}, \quad (5)$$

where $\Theta = (k; q_1,\dots,q_{k+1};\ a_1,\dots,a_{k+1};\ x_0;\ \sigma_1^2,\dots,\sigma_{k+1}^2)$ are the model parameters interpreted as follows:

$k+1$ is the number of mixture components, or homogeneous stratum;

$q_j\ \ (j = 1,2,\dots,k+1)$ is the *ex ante* probability of the *j*-th mixture component, or the share of the respective stratum in the population;

x_0 is the threshold expenditure separating observed expenditures $(x \le x_0)$ from the unobserved ones $(x > x_0)$;

$a_j = \mathbf{E}(\ln \xi_j)\ \ (j = 1,2,\dots,k+1)$ are the model averages of logarithms within *j*-th stratum;

$\sigma_j^2 = \mathbf{D}(\ln \xi_j)\ \ (j = 1,2,\dots,k+1)$ are the respective expenditure logarithms variance.

We assume that per capita expenditures of the population of the richest $k+1$-th stratum exceed the threshold x_0, and that they always refuse to participate in surveys. The rest households are available to statistical investigation, thought also can escape from the survey with probability $p(x)$ monotonically increasing with x (see hypothesis H_2 above).

Econometric analysis of the model (5) implies estimation of the parameter vector Θ by survey data, as well as some social and demographic population characteristics necessary to derive individual distribution from the household one.

5.4. Econometric analysis methodology

5.4.1. *Estimation of the dependence p(x) of refusal probability on per capita expenditures.*

The specification of the dependence is a logit model:

$$p\,(x_i) = P\{\xi_i = 0 \mid x_i\} = \frac{1}{1 + e^{a + b\ln x_i}}, \tag{6}$$

where

$$\xi_i = \begin{cases} 0, & i\text{ - th household refused from participation in the survey;} \\ 1, & i\text{ - th household participated in the survey,} \end{cases}$$

and x_i is i -th household expenditure.

Appendix 1 cites the results of econometric estimation of the logit model parameters in (6) by RLMS data, Rounds V–VIII. The analysis was conducted in Stata 6 package. We have confirmed the statistically significant monotonic dependence of $p(x)$ upon x.

5.4.2. Calibration (weighting) of the existing observations

Let $f(x)$ be the density function of the per capita expenditure distribution of population of a Russian region. If n is the total size of the survey sample and x_i is a certain value of per capita expenditures, then the number $\nu(x_i)$ of observations in the Δ-neighborhood of the point x_i on the condition that no one escapes from the survey, is given by

$$\nu(x_i) \approx nf(x_i)\Delta \ . \tag{7}$$

The actual number of observations, however, would be adjusted for the probability of refusal $p(x)$:

$$\tilde{\nu}(x_i) \approx nf(x_i)\big(1-p(x_i)\big)\Delta \ . \tag{8}$$

(7) and (8) imply that

$$\nu(x_i) = \tilde{\nu}(x_i)\cdot\frac{1}{1-p(x_i)} \ . \tag{9}$$

In particular, by choosing x_i as the actually observed data on per capita expenditures and taking small enough Δ, we would have:

$$\tilde{\nu}(x_i) = 1$$

$$\nu(x_i) = \frac{1}{1-p(x_i)}.$$

It means that if we want to estimate the underlying density $f(x)$ from the existing sample

Observed x	x_1	x_2	$\dots$	x_n
Observation weights	$\frac{1}{n}$	$\frac{1}{n}$		$\frac{1}{n}$

(10)

then we should recalibrate, or re-weight, the sample in the following way:

Observed x	x_1	x_2	$\cdots$	x_n
Observation weights	ω_1	ω_2	$\cdots$	ω_n

(11)

where ω_i are found from

$$\omega_i = \frac{\dfrac{1}{1-p(x_i)}}{\sum_{j=1}^{n}\left(\dfrac{1}{1-p(x_j)}\right)}.$$

It is worth noting that ω_i increase with the refusal probability $p(x_i)$, and $\sum_{i=1}^{n}\omega_i = 1$

5.4.3. Estimation of the observed mixture components parameters

At this stage we solve the problem of estimation from the sample (11) of the mixture parameters $k, \tilde{q}, \ldots, \tilde{q}_k,\ a_1, \ldots, a_k,\ \sigma_1^2, \ldots, \sigma_k^2$ driving the distribution shape:

$$\tilde{f}(x) = \sum_{j=1}^{k} \tilde{q}_j \frac{1}{\sqrt{2\pi}\,\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} \qquad (12)$$

The problem is in fact reduced to that of parameter estimation of the mixture of *normal* distributions:

$$\tilde{\varphi}(y) = \sum_{j=1}^{k} \tilde{q}_j \frac{1}{\sqrt{2\pi}\,\sigma_j} e^{-\frac{(y-a)^2}{2\sigma_j^2}} \qquad (13)$$

by the sample

Observed y	y_1	y_2	$\cdots$	y_n
Observation weights	ω_1	ω_2	$\cdots$	ω_n

, $(8')$

with $y_i = \ln x_i \quad (i = 1,2,\ldots,n)$.

The problem is solved by EM algorithm (a version of maximum likelihood method) described in [29] and [30] and implemented in CLASSMASTER software developed in CEMI.

5.4.4. Test for the hypothesis H3 on constant coefficient of variance in the population per capita expenditure distribution.

As it was mentioned in 4.1, testing this hypothesis in the model (5) is equivalent to testing hypothesis H_3' of equivalent variances:

$$H_3^{'}:\ \mathbf{D}\left(\ln\xi(1)\right)=\mathbf{D}\left(\ln\xi(2)\right)=\ldots=\mathbf{D}\left(\ln\xi(k)\right)=\sigma^2 .$$

There would be two stages of the hypothesis testing.

1 stage. Classification the existing observations $y_1=\ln x_1,\ldots\ \ldots,y_n=\ln x_n$ into $\hat{k}$ by normal discriminant analysis techniques. The output of this stage is the $\hat{k}$ portions of data, $n_1,n_2,\ldots,n_k$ $\left(\sum_{j=1}^{\hat{k}} n_j = n)\right)$.

2 stage. Recalculation of the estimates $\hat{\sigma}_j^2=\frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_{j.})^2$ and homogeneity testing with Bartlett criterion.

5.4.5. Estimation of the unobserved mixture component and distribution as a whole

Let the (relative) weight of the unobserved $\hat{k}+1$-th mixture component is $q_{\hat{k}+1}$, and the mean logarithm of per capita expenditures is $a_{\hat{k}+1}$. Then the regional average μ from the model (5) based on the parameter estimates $\hat{k};\ \hat{\hat{q}}_1,\ldots,\hat{\hat{q}}_{\hat{k}};\ \hat{a}_1,\ldots,\hat{a}_{\hat{k}};\ \hat{\sigma}_1^2,\ldots,\hat{\sigma}_{\hat{k}}^2$ obtained earlier is given by

$$\mu=\int_0^\infty x\left(\sum_{j=1}^{\hat{k}}\hat{q}_j\frac{1}{\sqrt{2\pi}\ \hat{\sigma}_j x}e^{-\frac{(\ln x-\hat{a}_j)^2}{2\hat{\sigma}_j^2}}+q_{\hat{k}+1}\frac{1}{\sqrt{2\pi}\ \sigma_{\hat{k}+1}(x-x_0)}e^{-\frac{(\ln(x-x_0)-a_{\hat{k}+1})^2}{2\sigma_{\hat{k}+1}^2}}\right)dx, \quad (14)$$

where

$$\hat{q}_j=\hat{\hat{q}}_j(1-q_{\hat{k}+1}),\quad j=1,2,\ldots,\hat{k}\,. \quad (15)$$

Given the properties of lognormal distribution,

$$\mu = \sum_{j=1}^{\hat{k}} \hat{q}_j \; e^{\frac{1}{2}\hat{\sigma}_j^2 + \hat{a}_j} + q_{\hat{k}+1}\left(x_0 + e^{\frac{1}{2}\sigma_{\hat{k}+1}^2 + a_{\hat{k}+1}} \right). \tag{14'}$$

The value of μ from (14') depends on the unknown $q_{\hat{k}+1}, a_{\hat{k}+1}$, as well as on x_0 and $\sigma_{\hat{k}+1}^2$. By construction, x_0 is taken to be the maximum of the observed expenditures:

$$x_0 = \max_{1 \le i \le n} \{x_i\} \tag{16}$$

If we do not reject H_3', then the overall estimate $\hat{\sigma}^2$ of the variance of logarithms is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{\hat{k}} n_j \hat{\sigma}_j^2 \,, \tag{17}$$

and $\sigma_{\hat{k}+1}^2$ is taken to be equal $\hat{\sigma}^2$.

We can then graph the level line in the plane ($q_{\hat{k}+1}, a_{\hat{k}+1}$):

$$\mu(q_{\hat{k}+1}, a_{\hat{k}+1}) = \mu^{\text{macro}} \,, \tag{18}$$

where the model value $\mu(q_{\hat{k}+1}, a_{\hat{k}+1})$ is calculated by (14') with $x_0 = \max_{1 \le i \le n} \{x_i\}$ and $\sigma_{\hat{k}+1}^2 = \hat{\sigma}^2$, while μ^{macro} is obtained from the macroeconomic Balance of Population Incomes and Expenditures for the relevant region and time point.

The final selection of the point $(\hat{q}_{\hat{k}+1}, \hat{a}_{\hat{k}+1})$ on the line (18) requires some additional conditions, assumptions, or expert information.

When constructing the line (18), it is worth considering that:

(i) Apparently,

$$q_{\hat{k}+1} << \min_{1 \le j \le \hat{k}} \{q_j\}$$

where the sign << means “much less”, i.e. that q_{k+1} in at least an order of magnitude less than $\min_{1 \le j \le \hat{k}} \{q_j\}$.

(ii) The level line (18) may be represented by a table with the values of $q_{\hat{k}+1}$ as input and $a_{\hat{k}+1}$ from (14')–(18) as output. A possible range of values $q_{\hat{k}+1}$ could be chosen as follows (with $\min_{1\le j\le \hat{k}}\{q_j\} = m\cdot 10^{-2},\ \ 1\le m\le 9$, i.e. if the least of the stratum shares is at the level of several per cent):

$$q_{\hat{k}+1} = \begin{cases} \nu\cdot 10^{-2}, & \nu = m-1,\, m-2,\ldots,1; \\ \nu\cdot 10^{-3}, & \nu = 9,8,\ldots,1; \\ \nu\cdot 10^{-4}, & \nu = 9,8,\ldots,1. \end{cases}$$

(iii) The results of earlier studies (e.g. [16]) suggest that the solution is to be expected to lie in the neighborhood of the point $(q^{(0)}_{\hat{k}+1}, a^{(0)}_{\hat{k}+1})$ where

$$q^{(0)}_{\hat{k}+1} \le \nu\cdot 10^{-3};$$
$$a^{(0)}_{\hat{k}+1} \ge 11.$$

5.4.6. Targeted assistance to the poor

If we restrict the class of weighting functions $w(x)$ in (1) to the functions like (3), then we can use results of [23] on the optimal allocation of the financial aid to the poor. By combining those with the estimates of the per capita expenditure density function $f(x)$ (of the form (5)), we can formulate the following rule of targeted assistance:

(i) For given inputs of the model (such as the population N, poverty line z_0, total resource S for targeted assistance, density function $f(x)$ describing the population of the region per capita expenditures, and Foster-Greer-Thorbecke index parameter $\alpha > 1$), the threshold value $\bar{z}_0$ can be found from

$$N\cdot I_0^{(\bar{z}_0)}(f)\cdot \bar{z}_0 \cdot I_1^{(\bar{z}_0)}(f) = S; \qquad (4')$$

(ii) Each inhabitant of the region whose per capita expenditures x are below the threshold, $x < \bar{z}_0$, is then eligible to the lump sum transfer $\bar{z}_0 - x$.

Apparently, if the weighting function $w(x)$ is changed, the optimal allocation rule may need to be reformulated.

5.5. *The preliminary results of econometric estimation*

By the time this interim report had been compiled (October 1999), we did not have some of the data at our disposal necessary to accomplish the tasks outlined in items 4.3–4.4 above. In particular, macroeconomic data from Balance of Population Income and Expenditures for the three regions of Komi Republic, Omsk and Volgograd oblasts as of Q3 1998, as well as for Russia as a whole for 1998 was not available. This prevented us from completing the second stage of calibration of the distributions analyzed (see 4.4 and (14)–(18)).

The report, however, does include the preliminary results of statistical analysis of per capita expenditure distribution, including the first stage of calibration, for Goskomstat data on the three regions of Russia. Also, the statistical investigation of the refusal probability $p(x)$ as a function of per capita expenditures x is conducted by using RLMS data on Rounds V–VIII coupled with data on refuses of households to participate in RLMS and the reasons of those refusals which were kindly given to us by P. M. Kozyreva and E. Artamonova from RAS Institute of Sociology. Finally, the report summarizes the earlier conducted research on *income* distribution of Russian population so that the dynamics of the distribution and reproducibility of the results can be traced.

Let us shortly cite the results obtained so far in our study. The technical details like graphs and tables see in Appendix 1.

5.5.1. Russian per capita income distribution in 1996

The distribution was analyzed within the framework of a similar model of the Gaussian distributions mixture of log per capita incomes with unknown number of components. RLMS Round VI data were used for the analysis. The sample design was as follows (see [31]). Russian administrative division as of 1989 was used for identifying primary sampling design. Because of low population density or difficult availability, a number of regions were excluded partially or altogether, namely, Tyva Republic, Sakha-Yakutia Republic, Chechnya, Ingush Republic, Krasnoyarsk kray, Taymyr, Evenkia and Yamal-Nenets okrugs (national autonomous units within larger krays), Kamchatka, Sakhalin, Tumen and Kaliningrad oblasts. This reduced the population subject to investigation by 4.4%, or 6.5 mln. From the rest of the country, at-once sample of 4718 households was created by a multistage stratification using geographical and urbanization characteristics. Of those, 3781 households did actually participate in the study, including 237 from Moscow, 192, from Moscow region, 111, from Saint Petersburg, 107, from Vladivostok, 100, from Saratov, etc. A number of separate questionnaires were used, including household, individual, and children questionnaires. Besides, community data such as infrastructure and local prices was also collected.

In the study of year 1996 incomes, RLMS data were accompanied by the data of "Obshchestvennoye Mnenie" (Public Opinion) Foundation of a study covering 12,000 household members as of early 1997, with an appropriate deflating.

To assess the richest strata characteristics such as its size (q_5) and average income $(m_5 = \mathbf{E}\,\xi_5)$, the sources cited in [31] as well as macroeconomic estimates of the richest population income were used. The estimation of the weight of the 'super-rich' strata (q_5) was implemented in the framework defined above. Ambiguity in q_5 was eliminated via identification of the 'contents' of this stratum. The results are given in Appendix 1 (in particular, Table A1.1 and Figures A1.1 and A1.2). The composition of the mixture and its components interpretable in the form of the corresponding population groups is given below (see item 3 of Conclusion).

5.5.2. Probability of refusal to participate in RLMS as a function of per capita expenditures

This dependence was studied in the form of logistic regression by using RLMS data, Rounds V-VIII, together with refusal data from RAS Institute of Sociology (see detailed description in Appendix 1):

$$p(x) = \frac{1}{1 + e^{a + b \ln x}},$$

where $p(x)$ is the probability that a household with per capita expenditures x (rubles) would refuse to participate in a survey. Statistically significant though not very much expressed monotone dependence was found between $p(x)$ and x (details see in Table A1.2 and Fig. A1.3 in Appendix 1).

5.5.3. Analysis and calibration of per capita expenditures of Komi Republic, Omsk and Volgograd oblasts population, Q3 1998

The results of this analysis are presented in Fig. A1.4–A1.9 and Tables A1.3–A1.11 of Appendix 1. They evidence that:

(i) The per capita expenditures distribution has log normal shape;

(ii) The algorithm of automatic number k of mixture components identification implemented in CLASSMASTER statistical package, as a rule, gives an estimate $\hat{k} = 4$ which signals that the population distribution in the observed ranges can be represented as mixture of four social and economic strata, though this does not necessarily imply that four local density maxima exist.

(iii) As compared to 1996 picture (see Fig. A1.1 and A1.2), population expenditure stratification in 1998 is much vaguer. This result confirms the earlier statement about long run tendency to restore overall lognormality by the end of transition.

6 CONCLUSIONS

The specific features of Russian transition did not cancel the lognormal model of income/expenditure distribution, though it did affect the mixing function $q(a)$. The phenomenon of the *discrete* lognormal mixture (instead of, typical for stable economies, *continuous* mixture of special form which in turn reproduces the lognormal distribution) is explained by the structural labor, human capital and skills demand shifts during the transition. These changes have crowded out "soviet middle class", i.e. relatively qualified workers, who has to seek other, as a rule, less profitable, income sources. This search has been adversely affected by low labor mobility typical for Russia. At the same time, rich rent flows have been acquired by new "extra rich" population groups. Thus, a well-defined pattern of groups of income earners has developed which has led to the discrete character of distribution mixture, the distribution being lognormal within each group. Hence, it is natural to try to model the underlying distribution by a discrete lognormal mixture. It is worth noting that as transition draws to a close, i.e. the Russian economy evolves towards its steady state, the shape of the mixing function $q(a)$ (and, consequently, of the whole expenditure distribution) would tend to resemble a usual two parameter lognormal distribution. Preliminary estimation based on 1998 data and its comparison to the 1996 results confirms this tendency.

The econometric analysis of the proposed model implies: **a)** per capita expenditures denisty idenitification via lognormal finite mixture parameters $\Theta = (k; a_1,\ldots,a_k;\ \sigma_1^2,\ldots,\sigma_k^2)$ estimation by the appropriate statistical procedures (see [29], [30]); **b)** re-weighting of the distribution accounting for the probability of refusal to participate in the survey as a function of per capita expenditures; **c)** reconstruction of the unobserved $(k+1)$-th stratum with the second re-calibration of the model based on partially verifiable working hypotheses and macroeconomic income and expenditure balances.

The preliminary results of the model estimation are given in [16] and in Appendix 1. In particular, the following interpretable strata are identified:

- people with marginal status, unemployed with unstable income, non-working pensioners and stipend receivers, low-paid workers, families with many children (39%);
- workers financed from budget, artists, "white collars", unsuccessful industry (e.g. military equipment) workers (40%);
- workers of relatively successful (energy sector) industries, arts and intellectual labor specialties in demand (17%);
- owners, principal shareholders, key workers of successful enterprises and industries, middle shadow economy personnel (4%);
- top state bureaucracy and Mafia elite (0.03-0.05%).

Thus, identification of the proposed model of population per capita expenditures distribution and, in particular, identification of each stratum analyzed, would allow to bring into light the main social and economic population groups as well as to analyze the main exogenous strata determinants.

The proposed distribution model structure including two-stage calibration procedure accounting for both survey evasion and macroeconomic balances provides better estimation of poverty and inequality characteristics than those currently used in Russian practice ([1]–[3], [7], [28]) as well as by other researchers ([4]–[6]). In particular, it can be used in calculations related to organization of targeted assistance to the poor.

The techniques based on this model are to be implemented on the *regional level*. Generalization to Russia as a whole can only be made if the regional data are framed to be cross-comparable by using appropriate deflators and coefficients accounting for differences in purchasing power, the minimum consumption basket composition, subsistence level, etc.

7 REFERENCES

[1] К методике оценки распределения генеральной совокупности населения по величине среднедушевого дохода на основании эмпирических данных. — Материалы Госкомстата РФ (тезисы к докладу на Ученом совете в ЦУЖе), 1999.

[2] *Великанова Т., Колмаков И., Фролова Е.* Совершенствование методики и моделей распределения населения по среднедушевому доходу. — «*Вопросы статистики*», 1996, № 5.

[3] *Великанова Т.Б., Фролова Е.Б.* Методология прямой оценки величины уровня бедности, основанная на распространении результатов выборочного обследования на генеральную совокупность. — Материалы Управления статистики уровня жизни населения Госкомстата РФ, 1999.

[4] *Шевяков А.Ю., Кирута А.Я.* Экономическое неравенство, уровень жизни и бедность населения России и ее регионов в процессе реформ: методы измерения и анализ причинных зависимостей. — Финальный отчет по проекту EERC, июль 1999.

[5] *Ершов Э.Б., Майер В.Ф.* Методологические и методические проблемы определения уровня, объема и дифференциации доходов населения. — Материалы к заседанию ВЦУЖ 28 декабря 1998 г.

[6] *Суворов А.В., Ульянова Е.А.* Денежные доходы населения России: 1992-1996 гг. — «Проблемы прогнозирования», 1997 г.

[7] Определение основных показателей агрегирования данных обследования бюджетов домашних хозяйств. — Матриалы Госкомстата РФ, 1999.

[8] *Брейтуэйт Дж.* Адресность и относительно длительная бедность в России. Февраль 1999. Доклад на семинаре Всемирного банка 19 апреля 1999 г.

[9] Пилотные программы по введению адресной социальной поддержки малоимущих семей в Республике Коми, Воронежской и Волгоградской областях. Предварительные итоги. Министерство труда и социального развития МФ. — М.: 1999. — 104 с.

[10] *Айвазян С.А., Колеников С.О. и др.* Методологические рекомендации по информационному обеспечению базового уровня системы показателей качества жизни населения и по расчету интегрального показателя качества жизни в условиях имеющейся статистической базы для федерального и регионального уровней. Итоговый научный отчет по договору № 6/98 с Госкомстатом РФ. М: ЦЭМИ РАН, декабрь 1998.

[11] *Айвазян С.А., Герасимова И.А.* Социальная структура и социальное расслоение населения Российской Федерации (по материалам выборочного обследования населения трех регионов РФ). — М.: ЦЭМИ РАН, 1998.

[12] *Колеников С.О.* Методы анализа качества жизни. Серия РЭШ «Лучшие студенческие работы», 1999 (в печати).

[13] *Hagenaars A.* A Class of Poverty Indices. — «Interational Economic Review», 28, 1987, pp. 583—607.

[14] *Foster J., Greer J., Thorbeck E.* A Class of Decomposable Poverty Measures. — «Econometrica», 52(3), 1984, pp. 761—766.

[15] *Bourguignon F., Fields G.* Discontinuous loss from poverty, generalized P_α measures, and optimal transfers to the poor. — XI-th World Congress of the International Economic Association. Tunis, December 1995.

[16] *Айвазян С.А.* Модель формирования распределения населения России по величине среднедушевого дохода. — «*Экономика и математические методы*», том 33 (1997), № 4, с. 74-86.

[17] Национальная оценка и распространение информации. Предварительный отчет в рамках проекта Всемирного банка по теме «Структурная перестройка системы социальной защиты населения» №SPIL-2.2.2/11. ГУ-ВШЭ, Москва, 1999.

[18] *Корчагина И., Овчарова Л., Турунцев Е.* Система индикаторов уровня бедности в переходный период в России. — Научный доклад №98/04 по секции «Микроэкономика-2 (поведение домохозяйств)» Российской программы Консорциума экономических исследований. РПЭИ/Фонд Евразия, апрель 1999.

[19] *Sen A.K.* A Sociological Approach to Measurement of Poverty. — Oxford Economic Papers, 37, pp. 669—667.

[20] *Atkinson A.B.* On the Measurement of Poverty. — Econometrica, vol.55 (1987), N4, pp.749-764.

[21] *Kanbur S.M.R.* Measurment and Alleviation of Poverty. — «IMF Staff Papers», 34, 1987, pp. 60—85.

[22] *Foster J.E., Shorroks A.F.* Poverty Orderings. — «Econometrica», 56(1), 1988, pp. 173—177.

[23] *Bourguignon F., Fields G.S.* Poverty Measures and Anti-Poverty Policy. — «Recherches Economiques de Louvain», 56(3-4), 1990, pp. 409—427.

[24] *Ravallion M.* Poverty Comparisons. — Chur, Switzerland, Harwood Academic Publishers, 1994.

[25] *Айвазян С.А., Рабкина Н.Е., Римашевская Н.М.* Методика расчета ожидаемого распределения рабочих и служащих по размерам заработной платы. —М.: НИИ труда ГКСМСССР по вопросам труда и заработной платы, 1967.

[26] *Mroz T., Popkin B., Mancini D., Glinskaya T., Lokshin V.* Monitoring Economic Conditions in the Russian Federation: The Russia Longuitudinal Monitoring Sur-

vey 1992-1996. — Report submitted to the U.S. Agency for International Development. Carolina population Center. University of North Caroline at Chapel Hill. February, 1997.

[27] *Aivazian S.A.* Probabilistic-Statistical Modelling of the Distributary Relations in Society. — In: «Private and Enlarged Consumption», North-Holland Publ. Comp., 1976.

[28] Методологические положения по статистике. Выпуск 1. — М., Госкомстат РФ, 1996.

[29] *Aivazian S.A.* Mixture-Model Cluster Analysis Using the Projection Pursuit Method. — In: «Computational Learning and Probabilistic Reasoning», John Wiley and Sons Ltd, 1996, pp. 278-286.

[30] *Rudzkis R., Radavicius M.* Statistical Estimation of a Mixture of Gaussian Distributions. — In: «Acta applicandae Mathematical», vol. 38, N1, 1995.

[31] The Russia Longitudinal Monitoring Survey: «Family questionnaire» and «Sample of Russian Federation». Rounds V and VI. Technical Report. August-October 1996.

[32] Социально-экономическое положение России. Январь-сентябрь 1996 г., №9. Госкомстат РФ по статистике, 1996.

[33] Bogomolova T.Y., Tapilina V.S., Rostovtsev P. The impact of income mobility on the change of income inequality. EERC, new project, December 1999.

[34] Stata Reference Manual. Release 6. Volume 2, H–O. Stata Press, 1999.

[35] Andrienko Y. What determine crimes in Russian regions? EERC, new project, December 1999.

APPENDICES

APPENDIX 1. PRELIMINARY RESULTS OF ECONOMETRIC ANALYSIS OF THE MODEL

Per capita income distribution in Russia as of 1996

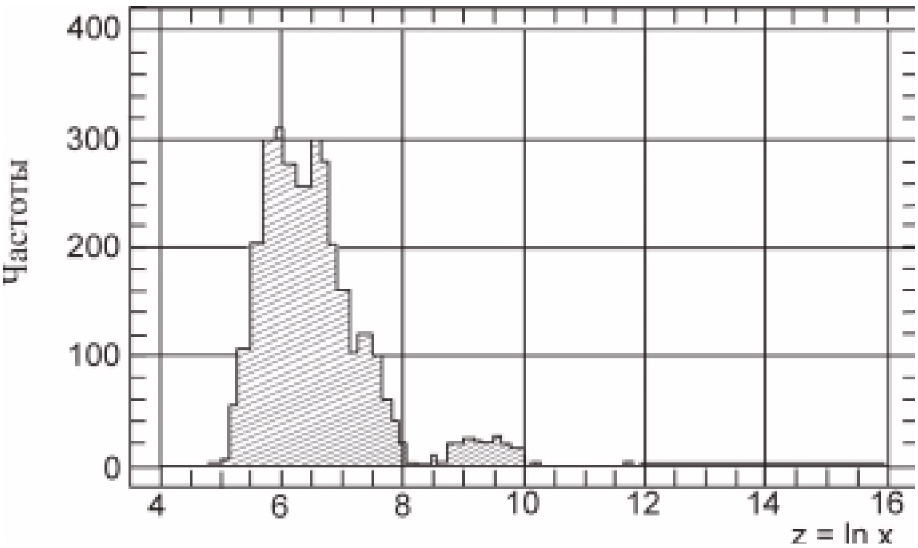

Fig A1.1. Histogram for distribution of Russian population by logarithm of per capita income. Sample size (for smoothed bootstrap) $n = 33330$. Left axis: frequency.

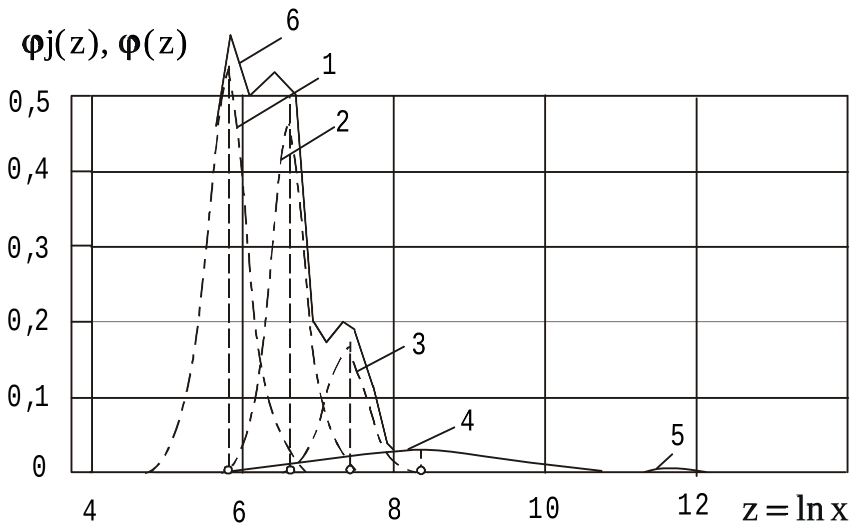

Fig. A1.2. Distribution density in corresponding strata; 6 – total density.

Table A1.1

The main characteristics of mixture components in the distribution of Russian households by monthly per capita income (September 1996).

Stratum (homogeneous group) No. j	Share in the population q_j	Mean per capita monthly income, th. rbs. m_j	Standard deviation of income, th. rbs. Δ_j	Ratio of the total income of the group to total Russian population $q_j m_j$	The share of the group in the total income $\gamma_j = \frac{q_j m_j}{\sum_{j=1}^{5} q_j m_j}$	Parameters μ_j and σ_j^2 for underlying normal distributions $\varphi(z \mid \mu_j; \sigma_j^2)$
1	0,3900	350	101	136,5	0,113	$\mu_1 = 5{,}818$ $\sigma_1^2 = 0{,}0802$ $\sigma_1 = 0{,}2832$
2	0,4000	700	210	280,0	0,232	$\mu_2 = 6{,}508$ $\sigma_2^2 = 0{,}0862$ $\sigma_2 = 0{,}2936$
3	0,1700	1640	525	278,8	0,231	$\mu_3 = 7{,}344$ $\sigma_3^2 = 0{,}0975$ $\sigma_3 = 0{,}3122$
4	0,0397	12000	5100	476,4	0,394	$\mu_4 = 9{,}310$ $\sigma_4^2 = 0{,}1660$ $\sigma_4 = 0{,}4075$
5	0,0003	120000	45000	36,0	0,030	$\mu_5 = 11{,}629$ $\sigma_5^2 = 0{,}1316$ $\sigma_5 = 0{,}3627$
total $\sum$	1,0000	-	-	1207,7[*)]	1,0000	-

*) 54% higher than the official Goskomstat figure for September, 1996 (see [32], p.176). According to Goskomstat, total money income of population in September 1996 were 116 bln rbs., i.e. 784 th rbs. per capita. The discrepancy is due to hidden incomes of the unobserved population, mainly of that from strata 4 and 5.

Probability of household refusal to participate in RLMS as a function of per capita expenditures

RLMS panel data were used to study the probability of a household to refuse to participate in a sociological survey.

For each of the 4718 households in the second wave RLMS sample, interviewers wrote down whether the household participated in the survey, and, if not, why. The codes registered (i.e., most typical responses) are reproduced in the Table A 1.2.

Table A1.2.

Visit result codes

01	Survey conducted
	Objective failure reasons
02	Uninhabited premises
03	No one lives in the house (apartment) at the moment
04	Apartment cannot be reached
05	Apartment is rented be foreigners
06	No one is at home
07	They neither open the door nor communicate
08	Survey impossible because of illness
09	Survey impossible because of handicap
10	No adults at home
11	Person opened the door is drunk
14	Family is absent during the whole period of the survey
15	Family is present only late in the evenings
16	Family actually lives in another location
18	Other
	Refusals
30	Refused to participate
	Communication circumstances
21	Refusal with the door closed
22	Refusal of the person opened the door
23	Refusal of the respondent
24	Refusal of another family member
25	Refusal when being interviewed
26	Refused by deceiving
27	Action against interviewer
28	Other
	Refusal reasons
41	Unmotivated refusal
42	“Too busy”
43	“Have no time”
44	“I never open the door”
45	“These surveys change nothing”
46	“Don’t want to tell about my life to anyone”
47	“I have a right not to answer”
48	“I want to have rest”
49	“I do not want to be in a computer”
50	“Participated in a sociological survey recently”
51	“We are temporarily here”
52	Family reasons
53	Not interested in the survey topic
54	Bored with politics
55	Refusal out of protest
56	Anxious of releasing information on political views
57	*Anxious of releasing information on family welfare level*
58	Do not trust the interviewer
59	Other

Table A1.3 reports the refusal rates in Rounds V–VIII.

Table A.1.3.

The RLMS attrition and refuses

	Round 5	Round 6	Round 7	Round 8
Survey not conducted	743	963	1118	1254
Refuses	410	539	489	701
Refuses because of unwillingness to provide information about household welfare			17	19
Survey conducted	3973	3781	3750	3831

The final goal of the analysis is the answer to the question: “Does the probability to refuse to participate in a sociological survey depend on the welfare of the household?” By using the above data on refusals combined with appropriate household data on income and/or expenditure level in the main RLMS data files, one can formulate binary dependent variable econometric model with per capita expenditure (or another of household welfare) as an explanatory variable and participation as a dependent one. Apparently, if the household had refused to participate in the survey in a given round, the data on its expenditures in this period cannot be obtained. But as long as RLMS provides *panel* data, i.e. the data on the same households in different periods[1], then, assuming that household welfare is approximately constant across time, one can impute this constant welfare from information in other rounds. The assumption of constant welfare may be subject to critique (see, e.g.,[33]). However, we propose average over several periods of per capita expenditure as a measure of household welfare. By smoothing out the period specific shocks, we obtain the indicator consistent, for instance, with Friedman life-cycle consumption hypothesis.

The software used (Stata 6) easily allows for such imputation via linear regression models. For each pattern of the missing data (in fact, for each observation where at least one of the variables to be imputed is missing), the regression model is constructed which includes all non-missing variables. This model is estimated by using all observations of the same missing value pattern, and then prediction of the values of the missing variable of interest is obtained ([34]).

[1] The information about sample attrition can be obtained from the Table A.1.3. No new households appear in the sample (at least, intentionally; some new households may actually appear if the previously registered ones move into another location). This was controlled for in selecting the sample for the analysis of the probability to avoid survey participation.

When the imputed values are used as regressors, the estimates of the respective coefficients are likely to be biased (due to measurement error effect) towards zero.

The basic RLMS variables used for the analysis of the refusal probability were per capita expenditures deflated to the same period (1992 prices; deflators from Russian Economic Trends publications), namely, **`totexpr*`** . The measure of household welfare is the average logarithm of deflated expenditures over Rounds V–VIII (years 1994–1996 and 1998). The variance of imputed log expenditures varies between 0.018 (i.e. the deviations of expenditures from its mean are less than 2% in these years) and 1.32 (i.e. the expenditures vary by a factor of 3.7) with median of 0.21. The dependent variable is an indicator whether the household refused to participate in RLMS at least once (in four periods).

The statistical model was that of logistic type regression: $p(x) = (1 + e^{a + b\ln x})^{-1}$, where $p\,(x)$ is the probability to refuse as a function of per capita expenditure x (ths. rbs). The ML estimates were as follows:

$\hat{\boldsymbol{a}} = \mathbf{1{,}613}$ (MSE =0,351);

$\hat{\boldsymbol{b}} = \mathbf{-0{,}250}$ (MSE =0,075);

likelihood ratio **LR (1)=18,47** (distributed as chi-squared with 1 d.f.; significant at $1.7 \cdot 10^{-5}$ level).

Fig A1.3 reports the graphical representation of the estimated refusal probability.

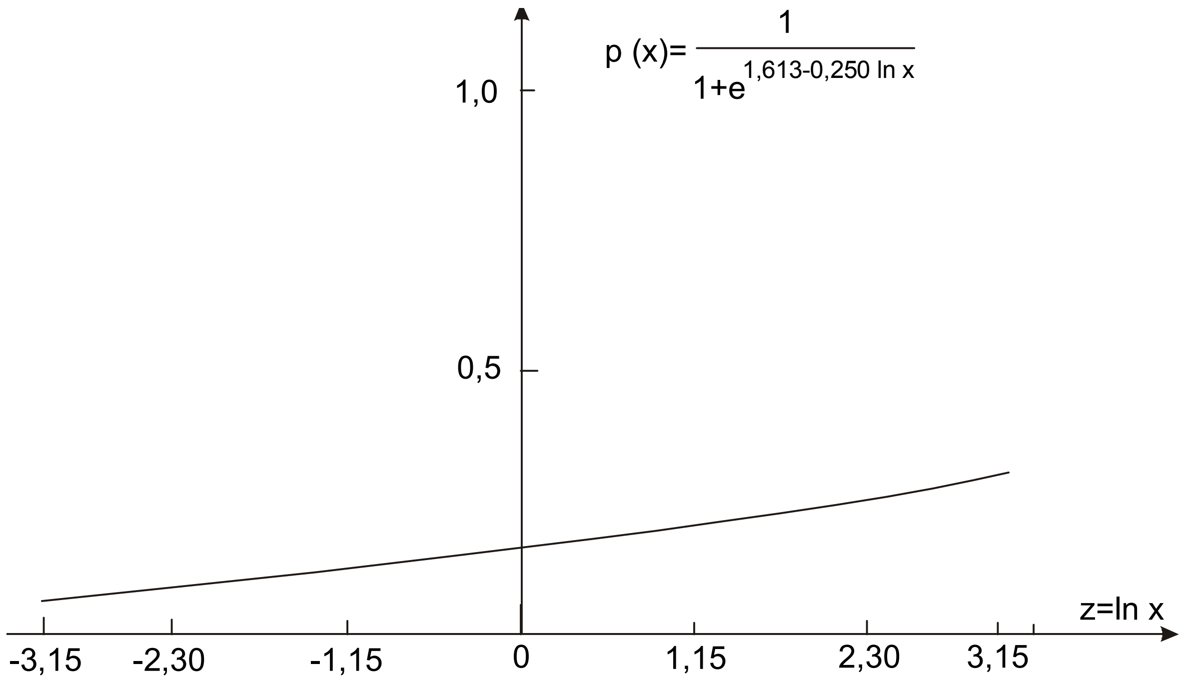

Fig. A1.3. Refusal probability $p(x) = \left(1 + e^{a + b \ln x}\right)^{-1}$ as a function of log per capita expenditures, ths. rbs. $(a = 1{,}613;\; b = -0{,}250)$.

The obtained results are interesting per se. We use them, however, only to adjust sample weights of the households in the sample. To improve the goodness of fit, which is the final goal of this logistic regression exercise, multivariate logistic regression model can be used with such regressors as household size, rural/urban dummies, etc. Besides, a more detailed measure of refusal probability as a dependent variable can be constructed as the ratio of the number of times the household did participate in the survey (which can vary between 0 and 4) to the total number of rounds (here, 4). Of course, if the household has never participated in the RLMS (i.e. has refused four times to fill in the questionnaire), no information can be obtained about welfare level of the household, so that the household is essentially lost.

The authors express gratitude to Kozyreva P.M. and Artamonova E. from RAS Institute of Sociology for the RLMS Rounds V–VIII refusal data.

<u>Estimation of the observable mixture components by the EM-algorithm and its modifications</u>

In this section, we describe the procedure to estimate the vector of parameters

$$\Theta(k) = \left(\tilde{q}_1,\ldots,\tilde{q}_k;\; a_1,\ldots,a_k;\; \sigma_1^2,\ldots,\sigma_k^2\right) \tag{A.1}$$

of the density function

$$\tilde{\varphi}_k\left(z|\Theta\right) = \sum_{j=1}^{k} \tilde{q}_j \varphi\left(z|a_j;\; \sigma_j^2\right) \tag{A.2}$$

from the information contained in the random sample (8') by the maximum likelihood method with *fixed* number of mixture components k. In (A.2), $\varphi(z|a_j;\; \sigma_j^2)$ is the density function of the normal distribution with the mean a_j and variance σ_j^2.

The problem is formualted as to find the parameter vector

$$\hat{\Theta}(k) = \left(\hat{\tilde{q}}_1,\ldots,\hat{\tilde{q}}_k;\; \hat{a}_1,\ldots,\hat{a}_k;\; \hat{\sigma}_1^2,\ldots,\hat{\sigma}_k^2\right), \tag{A.3}$$

so as to maximize the log likelihood function

$$l_k(\Theta(k)) = \sum_{i=1}^{n} \omega_i \left[\ln \sum_{j=1}^{k} \tilde{q}_j \varphi\,(z_i | a_j; \sigma_j^2) \right] \tag{A.4}$$

i.e.

$$\hat{\Theta}(k) = \arg\max_{\Theta(k)} l_k(\Theta(k)) \tag{A.5}$$

In (A.4), z_i are the observations in the sample (11), ω_i, their weights defined by (11'), and n, sample size.

Iterative EM-algorithm (Expectation–Maximization) solves (A.5) as follows ([29], [30]):

(i) log likelihood function (A.4) is decomposed as

$$l_k(\Theta(k)) = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij} \ln \tilde{q}_j + \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij} \ln \varphi(z_i | a_j; \sigma_j^2) - \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}\,, \tag{A.6}$$

where

$$g_{ij} = \frac{\tilde{q}_j \varphi(z_i | a_j; \sigma_j^2)}{\tilde{\varphi}_k(z_i | \Theta(k))} \tag{A.7}$$

together with the conditional probability formulae for $\varphi(z | a_j;\ \sigma_j^2)$ determine the probability to observe class j conditional on the observed value of z_i (i.e., *a posteriori* probability to observe the j-th class for observation i). This expression can also be used in the optimal Bayesian classification rule with uniform penalty for misclassification. Under this rule, the observation is classified into the class with the greatest *a posteriori* probability given by (A.7).

(ii) Expectation stage: let

$$\hat{\Theta}^{(t)}(k) = \left(\hat{\tilde{q}}_1^{(t)}, \ldots, \hat{\tilde{q}}_k^{(t)};\ \hat{a}_1^{(t)}, \ldots, \hat{a}_k^{(t)};\ \left(\hat{\sigma}_1^2\right)^{(t)}, \ldots, \left(\hat{\sigma}_k^2\right)^{(t)} \right) \tag{A.8}$$

be the estimate of the parameter $\Theta(k)$ obtained on the t-th step of the iterative procedure. By substituting this estimate into (A.7), one gets the vector of *a posteriori* probabilities $g_{ij}^{(t)}$.

(iii) Maximization stage: in turn, the probabilities $g_{ij}^{(t)}$ are put into the RHS of (A.6) instead of g_{ij}, and the next EM-algorithm iteration maximizes

$$l_k\left(\hat{\Theta}^{(t)}(k)\right) = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}^{(t)} \ln \hat{\tilde{q}}_j^{(t)} + \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}^{(t)} \ln \varphi\left(z_i | \hat{a}_j^{(t)}; (\hat{\sigma}_j^2)^{(t)}\right) - \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}^{(t)} \tag{A.9}$$

by $\hat{\Theta}^{(t)}(k)$ with *fixed* $g_{ij}^{(t)}$. The values in the point of the optimum

$$\hat{\hat{q}}_j^{(t+1)} = \sum_{i=1}^{n} \omega_i g_{ij}^{(t)},$$

$$\hat{a}_j^{(t+1)} = \frac{1}{\hat{\hat{q}}_j^{(t+1)}} \sum_{i=1}^{n} \omega_i g_{ij}^{(t)} z_i,$$

$$\left(\hat{\sigma}_j^2\right)^{(t+1)} = \frac{1}{\hat{\hat{q}}_j^{(t+1)}} \sum_{i=1}^{n} \omega_i g_{ij}^{(t)} \left(z_i - \hat{a}_j^{(t+1)}\right)^2,$$

$$j = 1,2,\dots,k.$$

are then launched into (A.7) to recalculate $g_{ij}^{(t+1)}$ $(j = 1,2,\dots,k)$ for the next iteration. [30] and other (later) works[2)] prove that under some rather general assumptions (the most strict among them being the boundness of the log likelihood), EM-algorithms have nice properties. In particular, the EM-estimates converge in probability to the solution of (A.5).

In our work, the EM-algorithm is modified in some technical points to adjust for our purposes. These points include the use of weights ω_i of observations z_i, and the use of background cluster supposed to follow *uniform* distribution (unlike the rest components with *normal* distribution) over the whole range of observed values. The detailed description of the EM-algorithm implementation in CLASSMASTER software can be found in [33].

The above framework of the ML $\hat{\Theta}(k)$ estimation is applicable when the number of components k is *known* in advance. Now, we need to estimate **the very number of components**, i.e. the components that can be revealed by statistical methods within the observed per capita expenditure range.

The estimation of k is conducted via consequent hypothesis testing with simple hypotheses

$H_0\colon\ k = j$

under the alternative

$H_1\colon\ k = j + 1 \quad , \ - \ j = 1,2,\dots,$ —

---

2) In fact, the procedure that was later named an EM-algorithm was first proposed in Шлезингер М.И. О самопроизвольном различении образов. — «Читающие автоматы», Киев, Наукова думка, 1965, с. 38—45 (Shlesinger O.M., On inadvertent image recognition. — Reading automates, Kiev, Naukova Dumka Publ., 1965, pp. 38—45). The main properties of these algorithms were also studied in this work. This publication is, however, hardly accessible and not known in the West.

by using the LR statistics

$$\gamma(j) = -2\ln\frac{l_j(\hat{\Theta}(j))}{l_{j+1}(\hat{\Theta}(j+1))}.$$

The earliest value $j = \hat{k}$ such that H_0 is not rejected is taken to be the estimate of the number of the components in (A.2). The procedure is also complemented by the technique of number of clusters estimation with projection pursuit approach descrobed in [31], as well as the analysis of the contents and composition of the classes obtained by EM-algorithm.

<u>Statistical analysis and calibration of per capita expenditure distribution of Komi Republic, Volgograd and Omsk oblasts, Q3 1998.</u> [2]

A. Komi Republic

Households studied: 330

Population studied: 1089

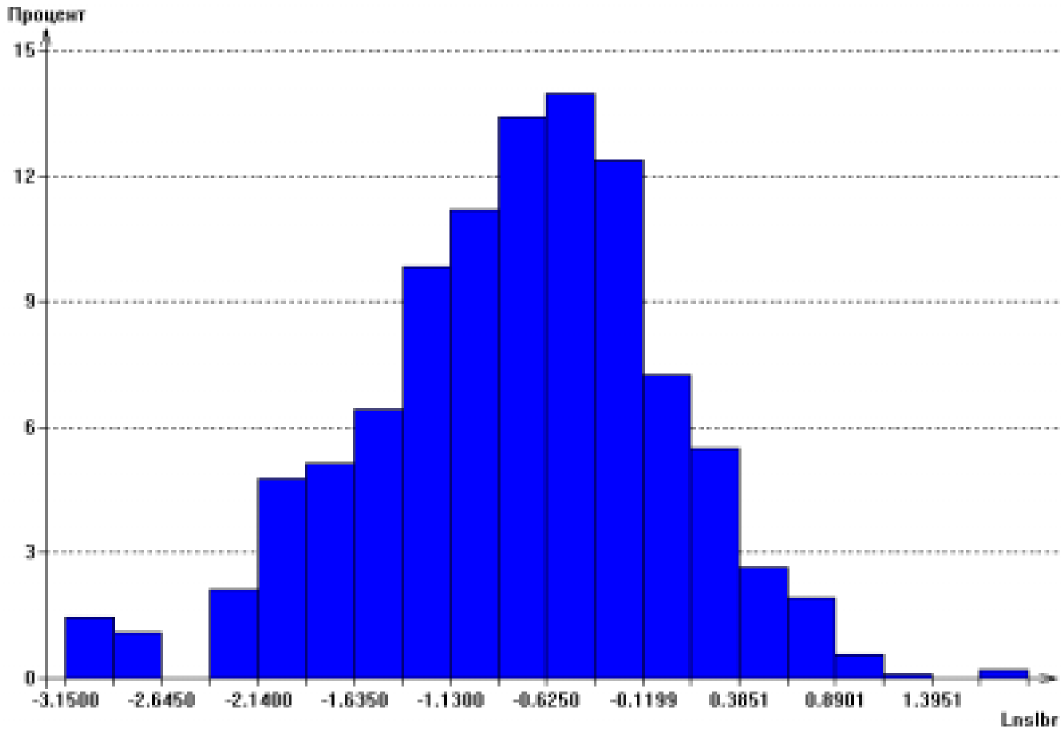

Fig. A1.4. Histogram of log per capita expenditure distribution of Komi Republic population, Q3 1998.

---

[2] Calculations are conducted by the senior researcher of RAS CEMI N. I. Makarchuk, candidate of science, by using CLASSMASTER statistical software.

Table A1.4

Distribution by bins

Bin number	Boundary points	Percentage of population within the bin
1.	-3.1500 — -2.8975	1.47
2.	-2.8975 — -2.6450	1.10
3.	-2.6450 — -2.3925	0
4.	-2.3925 — -2.1400	2.11
5.	-2.1400 — -1.8875	4.78
6.	-1.8875 — -1.6350	5.14
7.	-1.6350 — -1.3825	6.43
8.	-1.3825 — -1.1300	9.83
9.	-1.1300 — -0.8775	11.20
10.	-0.8775 — -0.6250	13.41
11.	-0.6250 — -0.3724	13.96
12.	-0.3724 — -0.1199	12.40
13.	-0.1199 — 0.1326	7.25
14.	0.1326 — 0.3851	5.51
15.	0.3851 — 0.6376	2.66
16.	0.6376 — 0.8901	1.93
17.	0.8901 — 1.1426	0.55
18.	1.1426 — 1.3951	0.09
19.	1.3951 — 1.6476	0
20.	1.6476 — 1.9001	0.18

Table A1.5

Summary statistics

Statistics	Observed value
Number of observations	1089
Mean (α)	-0.79075
Variance (σ^2)	0.61548
MSE (σ)	0.78453
Coefficient of variation	0.99
Min	-3.150
Max	1.900
Median	-0.700
20% quantile	-1.4340
80% quantile	-0.1600
Normality test	
Degrees of freedom	17
Chi-squared statistics	122.52
p-value	0

Table A1.6.

Mixture decomposition results

Number of classes (mixture components) = 4

Class No. (j)	Share of the class in population (q_j)	Mean (α_j)	MSE (σ_j)	Observations classified into the class (n_j)	Sample variance (s_j^2)
1.	0.12	-0.26	0.06	129	0.004059
2.	0.13	-0.62	0.05	143	0.002467
3.	0.03	-2.91	0.17	28	0.029063
4.	0.72	-0.83	0.80	789	0.634757

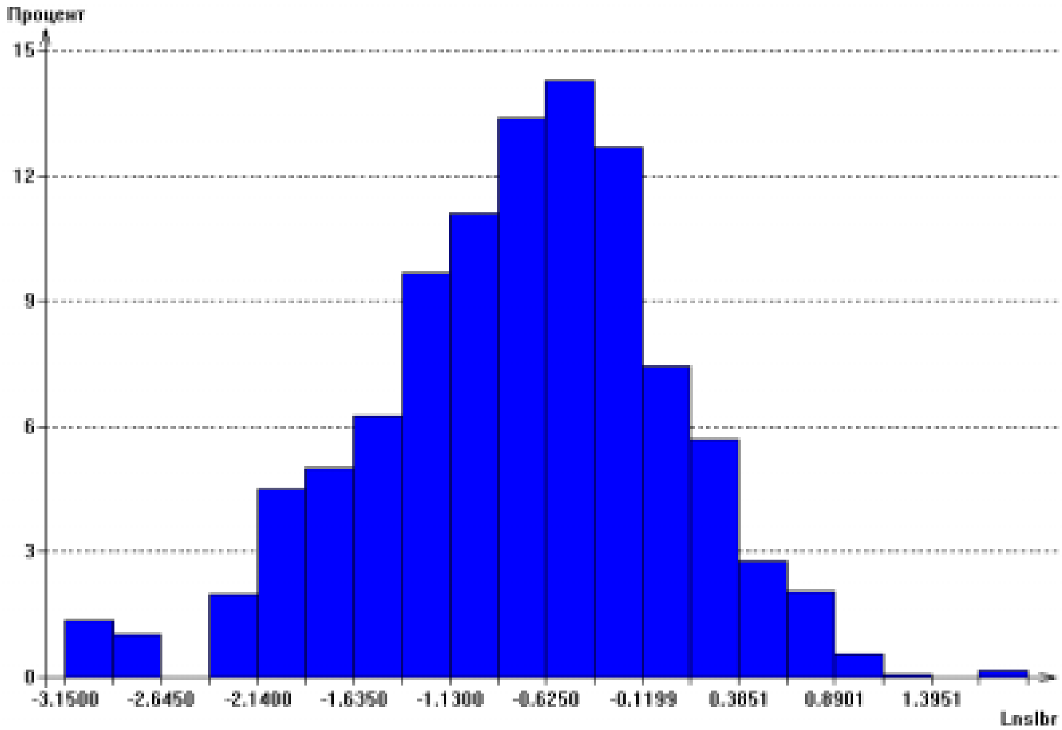

Fig. A1.4’. Histogram of log per capita expenditure distribution of Komi Republic population adjusted for the probability of refusal (total calibrated population n = 1262).

Table A1.4’

Distribution by bins

Bin number	Boundary points	Percentage of population within the bin
1.	-3.1500 — -2.8975	1.35
2.	-2.8975 — -2.6450	1.03
3.	-2.6450 — -2.3925	0
4.	-2.3925 — -2.1400	1.98
5.	-2.1400 — -1.8875	4.52
6.	-1.8875 — -1.6350	4.99
7.	-1.6350 — -1.3825	6.26
8.	-1.3825 — -1.1300	9.67
9.	-1.1300 — -0.8775	11.09
10.	-0.8775 — -0.6250	13.39
11.	-0.6250 — -0.3724	14.26

12.	-0.3724 — -0.1199	12.68
13.	-0.1199 — 0.1326	7.45
14.	0.1326 — 0.3851	5.71
15.	0.3851 — 0.6376	2.77
16.	0.6376 — 0.8901	2.06
17.	0.8901 — 1.1426	0.55
18.	1.1426 — 1.3951	0.08
19.	1.3951 — 1.6476	0
20.	1.6476 — 1.9001	0.16

Summary statistics

Number of observations	1262
Mean (α)	-0.76585
Variance (σ^2)	0.61299
MSE (σ)	0.782941
Coefficient of variation	1.022317

B. Volgograd oblast

Households studied: 400

Population studied: 1263

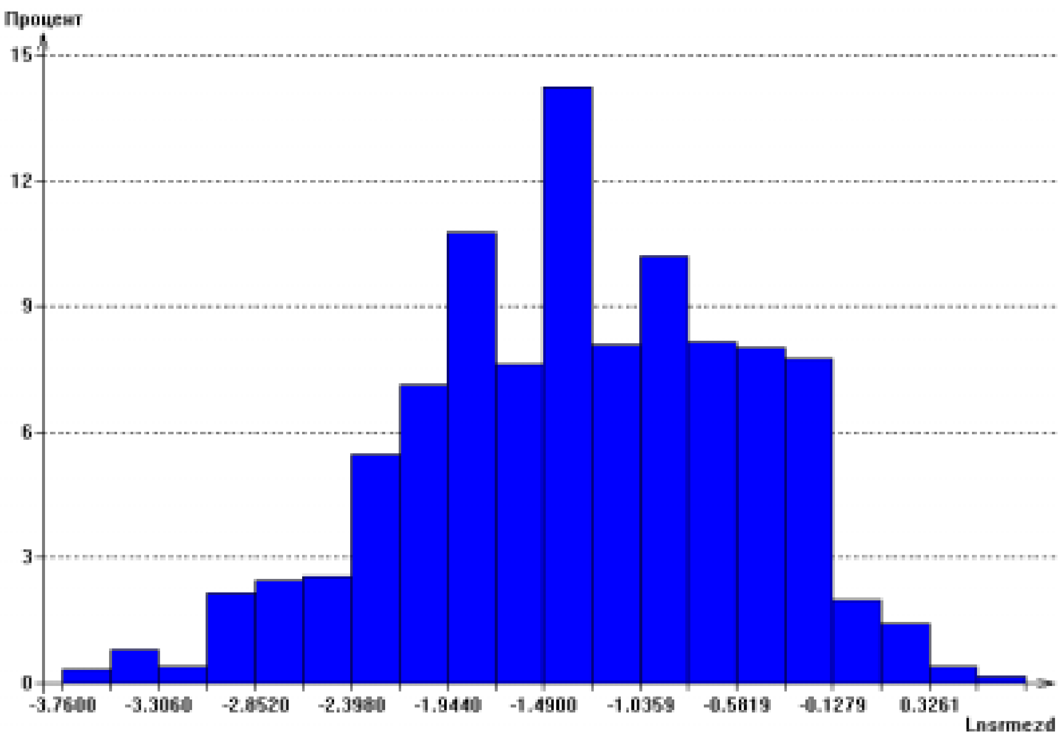

Fig. A1.5. Histogram of log per capita expenditure distribution of Volgograd oblast population Q3 1998.

Table A1.7

Distribution by bins

Bin number	Boundary points	Percentage of population within the bin
1.	-3.7600 — -3.5330	0.32
2.	-3.5330 — -3.3060	0.79
3.	-3.3060 — -3.0790	0.40
4.	-3.0790 — -2.8520	2.14
5.	-2.8520 — -2.6250	2.45
6.	-2.6250 — -2.3980	2.53
7.	-2.3980 — -2.1710	5.46
8.	-2.1710 — -1.9440	7.13
9.	-1.9440 — -1.7170	10.77
10.	-1.7170 — -1.4900	7.60
11.	-1.4900 — -1.2629	14.25
12.	-1.2629 — -1.0359	8.08
13.	-1.0359 — -0.8089	10.21
14.	-0.8089 — -0.5819	8.16
15.	-0.5819 — -0.3549	8.00
16.	-0.3549 — -0.1279	7.76
17.	-0.1279 — 0.0991	1.98
18.	0.0991 — 0.3261	1.43
19.	0.3261 — 0.5531	0.40
20.	0.5531 — 0.7801	0.16

Table A1.8

Summary statistics

Statistics	Observed value
Number of observations	1263
Mean (α)	-1.32176
Variance (σ^2)	0.61294
MSE (σ)	0.78290
Coefficient of variation	0.59
Min	-3.7600
Max	0.7800
Median	-1.3300
20% quantile	-1.9900
80% quantile	-0.5900
Normality test	
Degrees of freedom	17
Chi-squared statistics	104.36
p-value	0

Table A1.9

Mixture decomposition results

Number of classes (mixture components) = 4

Class No. (j)	Share of the class in population (q_j)	Mean (α_j)	MSE (σ_j)	Observations classified into the class (n_j)	Sample variance (s_j^2)
1.	0.04	-1.00	0.01	51	0.000165
2.	0.82	-1.33	0.86	1032	0.742936
3.	0.08	-1.42	0.03	104	0.000898
4.	0.06	-1.31	0.03	76	0.000857

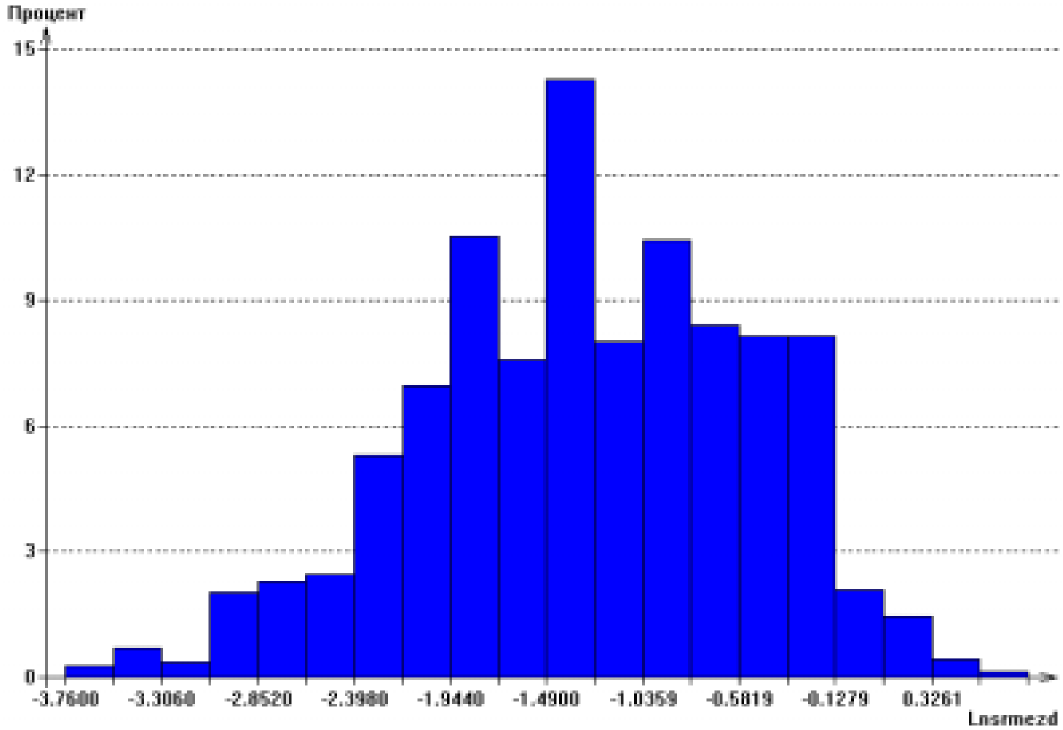

Fig. A1.5’. Histogram of log per capita expenditure distribution of Volgograd oblast population adjusted for the probability of refusal (total calibrated population n = 1436).

Table A1.7′

Distribution by bins

Bin number	Boundary points	Percentage of population within the bin
1.	-3.7600 — -3.5330	0.28
2.	-3.5330 — -3.3060	0.70
3.	-3.3060 — -3.0790	0.35
4.	-3.0790 — -2.8520	2.02
5.	-2.8520 — -2.6250	2.09
6.	-2.6250 — -2.3980	2.65

7.	-2.3980 — -2.1710	5.29
8.	-2.1710 — -1.9440	6.96
9.	-1.9440 — -1.7170	9.75
10.	-1.7170 — -1.4900	8.15
11.	-1.4900 — -1.2629	14.5
12.	-1.2629 — -1.0359	7.31
13.	-1.0359 — -0.8089	10.79
14.	-0.8089 — -0.5819	8.77
15.	-0.5819 — -0.3549	7.45
16.	-0.3549 — -0.1279	8.70
17.	-0.1279 — 0.0991	2.23
18.	0.0991 — 0.3261	1.46
19.	0.3261 — 0.5531	0.42
20.	0.5531 — 0.7801	0.14

Summary statistics

Number of observations	1436
Mean (α)	-1.30179
Variance (σ^2)	0.60615
MSE (σ)	0.7785573
Coefficient of variation	0.598066

C. Omsk oblast

Households studied: 365

Population studied: 1244

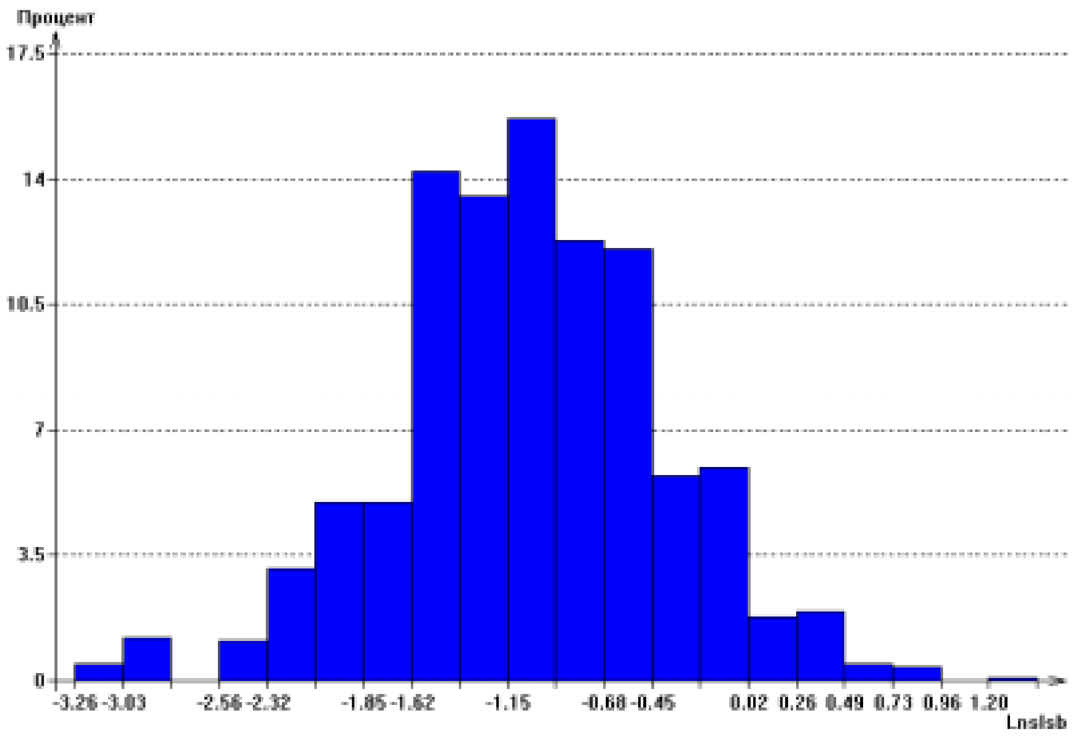

Fig. A1.6. Histogram of log per capita expenditure distribution of Omsk oblast population Q3 1998.

Table A1.10

Distribution by bins

Bin number	Boundary points	Percentage of population within the bin
1.	-3.26 — -3.03	0.48
2.	-3.03 — -2.79	1.21
3.	-2.79 — -2.56	0.0
4.	-2.56 — -2.32	1.13
5.	-2.32 — -2.09	3.14
6.	-2.09 — -1.85	4.98
7.	-1.85 — -1.62	4.98
8.	-1.62 — -1.38	14.23
9.	-1.38 — -1.15	13.59
10.	-1.15 — -0.92	15.68
11.	-0.92 — -0.68	12.78
12.	-0.68 — -0.45	11.82
13.	-0.45 — -0.21	5.71
14.	-0.21 — 0.02	5.87
15.	0.02 — 0.26	1.77
16.	0.26 — 0.49	1.69
17.	0.49 — 0.73	0.48
18.	0.73 — 0.96	0.40
19.	0.96 — 1.2	0
20.	1.20 — 1.43	0.08

Table A1.11

Summary statistics

Statistics	Observed value
Number of observations	1244
Mean (α)	-1.054
Variance (σ^2)	0.439
MSE (σ)	0.663
Coefficient of variation	0.63
Min	-3.26
Max	1.42
Median	-1.05
20% quantile	-1.53
80% quantile	-0.53
Normality test	
Degrees of freedom	17
Chi-squared	113.44
p-value	0

Table A1.12.

Mixture decomposition results

Number of classes (mixture components) = 4

Class No. (j)	Share of the class in population (q_j)	Mean (α_j)	MSE (σ_j)	Observations classified into the class (n_j)	Sample variance (s_j^2)
1.	0.57	-0.947	0.86	705	0.73295
2.	0.10	-1.452	0.035	126	0.00127
3.	0.13	-1.258	0.050	160	0.00253
4.	0.20	-1.003	0.086	253	0.00730

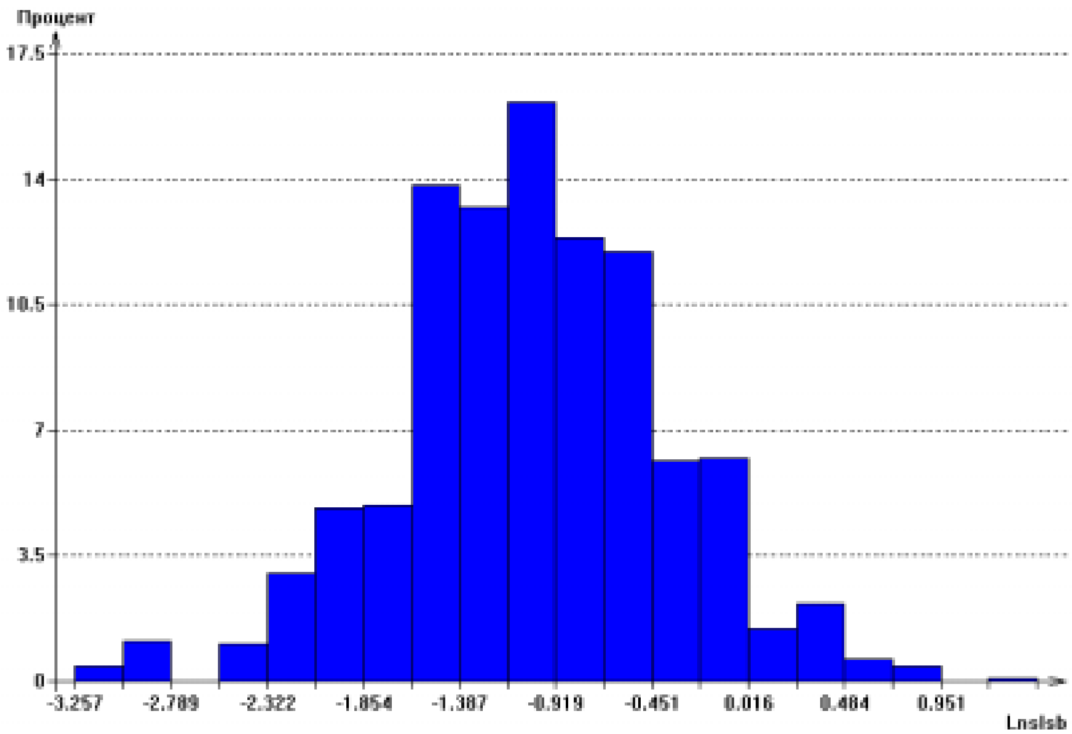

Fig. A1.6’. Histogram of log per capita expenditure distribution of Omsk oblast population adjusted for the probability of refusal (total calibrated population n = 1430).

Table A1.10′

Distribution by bins

Bin number	Boundary points	Percentage of population within the bin
1.	-3.26 — -3.03	0.42
2.	-3.03 — -2.79	1.12
3.	-2.79 — -2.56	0.0
4.	-2.56 — -2.32	1.05
5.	-2.32 — -2.09	3.01
6.	-2.09 — -1.85	4.83
7.	-1.85 — -1.62	4.90
8.	-1.62 — -1.38	14.06
9.	-1.38 — -1.15	13.43
10.	-1.15 — -0.92	15.73
11.	-0.92 — -0.68	12.38
12.	-0.68 — -0.45	12.24
13.	-0.45 — -0.21	5.90
14.	-0.21 — 0.02	6.15
15.	0.02 — 0.26	1.82
16.	0.26 — 0.49	2.03
17.	0.49 — 0.73	0.49
18.	0.73 — 0.96	0.42
19.	0.96 — 1.2	0
20.	1.2 — 1.43	0.07

Summary statistics

Number of observations	1430
Mean (α)	-1.04266
Variance (σ^2)	0.428138
MSE (σ)	0.654322
Coefficient of variation	0.627552

APPENDIX 2. DRAFT OF THE FINAL REPORT.

The final report will follow the outline of this interim report. The contents will be expanded with the following points.

1) Item 4.3, 'Model description and parameter interpretation' will be supplemented by the theoretical results on the optimal strategies of targeted social assistance under quite general assumptions about the weighting function $w(x)$ entering equation (1) for poverty indices.

2) Item 4.5 will be named 'Results of the model econometric analysis'. Its contents as well as that of Appendix 1 will be enlarged by

   (i) implementation of the 2nd stage of the distribution calibration adjusting for macroeconomic figures of per capita expenditures and simultaneous identification and estimation of the parameters of the richest strata by Q3 1998 three regions Goskomstat data and Round VIII RLMS data;

   (ii) working hypotheses H_1–H_3 testing (see p. 5 of this interim report) with the same data;

   (iii) estimation of the main poverty indices (FGT family, including head-count ratio) and main inequality characteristics (decile income ratio, Gini index, etc.) for the three regions of Russia and, if the necessary data are available, for Moscow;

   (iv) analysis of the strata identified upon the density $f(x)$ estimation and social, demographic and economic factors underlying these strata, in all three regions.