Income Inequality in Europe and Russia: Cross-Countries Analysis

Irina Degtiareva

Advisor:

Surinov Alexander, Professor

Department of Statistics and Data Analysis

National Research University – Higher School of Economics

1 Introduction

Against the backdrop of technological advances and globalization, inequality has become a major obstacle for the global development. Reducing inequality within and between countries became one of the Sustainable Development Goals in 2015. Inequality becomes an essential and complicated issue, especially in the context of on-going economic integration processes between the countries in Europe, which seek to converge their economic, social and political systems. Thus, we need to measure how unevenly income is distributed among a certain group or the whole population. In the era of globalization, we need to better understand the inequality beyond the confines of the nation-state and analyze the cross-country variation of income and consumption inequalities.

1.1 Problem Statement

The household income provides information about the current standard of living and actual economic activity of the household members.

We aim to conduct a comparative analysis of the income inequality indicators of the Russian population and the comparable European countries based on micro-level data from the household surveys.

In the theoretical review we will discuss the approaches to the measurement of the inequality and income distribution modeling.

The study will be based on the Luxemburg Income Study Database (LIS Data Center), that stores harmonized data on income, wealth and other indicators for many European and non-European countries, that enables cross-national comparisons.

The main objectives of this research are to:

1. Define the group of comparable countries which will have similar demographic, economic and possibly cultural particular qualities.
2. Describe and calculate different inequality indicators and compare approaches to determining income.
3. Estimate the income distribution density function in Russia.

2 Literature Review

This section begins with a short review of the literature regarding the methods of measurements of income inequality and income distribution modeling using parametric and non-parametric methods.

According to Milanovic, there are three main approaches to inequality measurements: inequality among countries, inequality among individuals within the nation, and global inequality among all individuals without any limitation (Milanovic, 2016). Author claims that inequality gaps between countries are still dominant, even considering increase in mean incomes and reduction in inequality within countries (Milanovic, 2016).

The following problem will be addressed in the proposed research as we are going to conduct a cross-country analysis and compare in-state inequality level in Russia to the composition of inequality in developed European countries.

Frank Cowell suggests taking into account the household wealth as it provides additional information reflecting past financial well-being, including but not limited to savings that represent the excess of income over expenditure (Cowell, 2012). Piketty claims that wealth itself may become a source of income. Thus, accumulated wealth and capital (i.e., saving, debt, inheritance and etc.) reinforces inequality among both countries and individuals (Piketty, 2013).

However, we do not aim to provide comprehensive information on the origins of inequality in Russia. On the contrary, we would like to focus on more volatile variable of household income that reflects the flow of money household receives and current standard of living. This makes the analysis less complicated and allows us to define the middle-income households for a closer examination, based on the income distribution.

2.1 Defining Income

Using the analogy with poverty studies, we can conditionally divide the approaches to determining inequality into monetary and non-monetary (Atkinson, et al., 2015). Monetary approaches are subject of economic and statistical analysis, while non-monetary approaches are most commonly used in social sciences.

On the other hand, monetary estimates are represented by indicators of household income and consumption spending. Income and consumption are streamed and measured per unit of time (year, month, etc.). Income includes any cash and material assets received by a household over a certain period of time. It also includes compensation for labor (wages, bonuses and allowances), social transfers (pensions, scholarships and other social benefits), as well as capital income (rent, interest and dividends) (Gimpelson et al., 2014).

However, different articles use different approaches to determining income, which in turn changes estimates of inequality. Thus, it is necessary to understand whether the total gross income of the household or the disposable income (after taxes) is used. Besides, some researchers propose the use of a comprehensive definition of income, which assumes full accounting for changes in the value of assets, as well as public services expressed in monetary terms (medical programs, education, etc.) (Bernstein, et al., 2008).

In addition to the described approaches to determining income, the OECD methodology also proposes to take into account the size and composition of the household using a "reduction factor" known as an equivalence scale. In this case, household expenditures for the first adult household member are multiplied by 1, for subsequent adults by 0.5, and for children and dependents by 0.3 (Atkinson, 2015)

2.2 Inequality indicators

The most common indicator of inequality is the widely used Gini index. Its peculiarity is that different methods are used in its calculation, and, as a consequence, these estimates are incomparable (Lazaryan et al., 2017). For example, to calculate the Gini, researchers use disposable income, disposable income adjusted for taxes paid and transfers received, and disposable income adjusted for household size and composition. Thus, all these estimates are different.

$$G = \frac{1}{2n^2\overline{y}} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_j - y_i| \tag{1}$$

Where n is the population size, y_i and y_j are the incomes of the $i-th$ and $j-th$ households, respectively, and $\overline{y}$ is the average value of the population's income.

It is important to note that the Gini index shows how far the distribution is from the line of absolute equality, but does not take into account the nature and form of the distribution of the population's income. So, with the same values of the Gini coefficient, the shape of the income distribution density can vary greatly in countries, while the shapes of the Lorenz curves can have different shapes or intersect (Salmina, 2019).

Atkinson proposed a new indicator to measure the economic inequality - the Atkinson index. Its key difference from the Gini index is that it is able to take into account society's perceptions of the marginal level of inequality (Atkinson, 1970). The parameter ε regulates the weight of a certain part of the distribution, that is, with an increase in the parameter ε, an increasing proportion of the population is in favor of taking redistribution measures.

$$I_A^{\varepsilon}(F) = 1 - \frac{1}{\mu(F)} \left[\int x^{1-\varepsilon} dF(x) \right]^{\frac{1}{1-\varepsilon}} \tag{2}$$

Where $\mu(F)$ is the expected value (mean for the sample), the ε parameter is a measure of sensitivity to inequality. The parameter of sensitivity to inequality is defined in the range from 0 to 2. The closer the parameter value is to 2, the more it reacts to changes in incomes on the left side of the distribution (lowest incomes).

Along with these indices, decile ratios are widely used in the literature. The decile income ratio is the ratio of the average income of the 10% of the population with the highest income to the 10% of the population with the lowest income. The same way we can use a quintile coefficient, in which the sample will be divided into 5 groups in the same way, and the ratio will accordingly be considered by 20%

of the population with the highest incomes and 20% of the population with the lowest incomes (Lebedev et al., 2019).

2.3 Approaches to the estimation of the theoretical income distribution function

When using non-parametric methods, the researcher is not initially questioned about the presence of some basic distribution. Histograms are often used to get an initial idea of the data. However, they are highly dependent on the choice of the number and length of the interval in which the frequencies will be calculated (Scott, 2004). The main disadvantage of a histogram is that it is a set of intervals, that is, a discontinuous one. This does not allow differentiating and performing many other mathematical operations.

Nevertheless, today the development of software allows the use of kernel density estimates, which largely solve the problems associated with the use of a histogram. The principle of calculating the density at a point is to count the closest observations to this point, and so on for each subsequent point, the weight of which depends on the distance to the starting point (Rosenblatt, 1956). In this case, in order to ensure smooth differentiability, various functions of the kernel are used, most often it is the Gaussian kernel.

The kernel density function thus has the form:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{hn}\sum_{i=1}^{n}\left(\frac{x - x_i}{h}\right) \tag{3}$$

Where K is the kernel function, and $h > 0$ is a parameter that determines the interval on which the nearest points are searched. In this case, h is chosen as small as possible so that the function is smoother.

The Gaussian kernel is:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{4}$$

The Epanechnikov (parabolic) kernel has the form:

$$K(u) = \frac{3}{4}(1 - u^2), where \mid u \mid < 1 \tag{5}$$

Since the parabolic kernel, although having great efficiency, is defined only in the interval from -1 to 1, the Gaussian kernel is most often used. Further in the paper, a kernel estimate of the density function will be given using this approach.

When using parametric methods, the researcher must have a basic idea about the shape of the distribution: is it symmetrical, is there a right-sided or left-sided asymmetry, and so on. Since the distribution of income most often has a right-sided asymmetry, researchers have a question about using the log-normal, Pareto, exponential or Weibull distribution.

Pareto distribution is considered to be the first approximating model of income distribution (Arnold, 2008). However, further research has shown that the Pareto distribution is suitable for approximating high-income groups for distributions with a long “heavy” tail (Butaeva, 2016).

One-peak distributions with positive skewness, such as lognormal, are best suited to approximate low and middle incomes. Most often, in studies of population incomes, an approximation using a lognormal distribution gives best fit.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{6}$$

However, for a number of countries, it has been shown that the lognormal distribution does not approximate high incomes well. For a lognormal distribution, the “tail” decreases much faster than in empirical data for some countries (Butaeva, 2016).

Likewise, Russian researchers use mixture distribution models. S. Ayvazyan in his 1977 study used a five-component model of a mixture of lognormal distributions (Ayvazyan, 1997). In this case, the general density function is specified as the sum of the distribution densities of all groups.

$$f(x, \theta) = \sum_{j=1}^{k} \pi_j f_j(x, \theta_j) \tag{7}$$

Where $f_j(x, \theta_j)$ is the parametric density function, θ_j is the vector of the distribution parameters, π_j is the proportion of the sample belonging to the j-th group.

However, it is important to note that the use of complex models and functions with three, four or more parameters significantly complicates the analysis and, as a result, leads to misinterpretation of statistical results. While conservative tests may produce better model fitting results, the statistical interpretation and meaning of the model may be significantly worse (Cowell, 1988).

This paper will give the main characteristics of the distribution of income in Russia, carry out the selection and assessment of the theoretical distribution and check if the empirical distribution corresponds to the theoretical one.

3 Empirical Findings

Further research involves a comparative analysis of economic inequalities in Europe and Russia. Based on the World Bank data, we have selected several parameters by which we will select the countries with the closest welfare indicators. The variables used to classify countries are described in the Table 1.

Based on hierarchical classification and k-means clustering methods, we have selected countries for subsequent comparison with Russia[1].

[1] In this paper we will not describe the classification process. Details of comparable countries selection and R code for clustering are available at GitHub project

Variable	Specification
GDP per capita, PPP, current international dollars	The market value of all goods and services (adjusted for PPP) produced in the country per year on average per person
Unemployment rate, %	Reflects the proportion of the economically active population that wants to find a job, but objectively cannot
Gini index according to the World Bank, %	Inequality measure
Life expectancy at birth, years	Life expectancy at birth, provided that mortality rates at all ages remain the same as in the current year

Table 1: Description of the variables used for classification.

Thus, we see it sensible to compare Russia with the countries of Eastern Europe. For this clustering criterion, Poland was selected in both cases. In addition, Russia should be compared with the more developed countries of Western Europe so that the differences in welfare are not strong. This paper will analyze inequality indicators calculated from micro-level data from household surveys in Russia, Poland and Spain.

3.1 Household income variables

Since the purpose of our study is to analyze the general population, we need to provide weighted estimates of indicators. We will use the following variables: household disposable income (*dhi*), household weight (*hpopwgt*), number of household members (*nhhmem*), information on gross or net income (*grossnet*), factor income

($hifactor$), work-related insurance payments ($hpub_i$), universal benefits ($hpub_u$), social assistance benefits ($hpub_a$), private transfers ($hiprivate$), and social security taxes and contributions ($hxitsc$).

Many inequality measures are sensitive to the values at the bottom and/or top of the income distribution, and some are not defined for non-positive values of income (e.g., any measure that calculates a logarithm). Applying top and bottom-codes (often referred to as 'winsorising') will avoid this problem, as well as ensuring comparability between datasets that may have originally had different top- and bottom-coding.

Next, we calculate per capita income and the equivalised income - adjusted for household composition.

Equivalence scales are used to compare the incomes of different types of households (in terms of size and composition), since they cannot be directly compared. Imagine that we have a household of two adult earning individuals, we cannot compare it directly with a household of one adult, several children, and the elderly, since this comparison would be unequal. As discussed earlier, there are various approaches to using equivalence scales. We will use the scale suggested by the Luxembourg Income Survey: the square root of the number of household members.

As you can see from the Table 2, the equivalent income for Russia is between the

	Household income (no top or bottom codes)	Household income	Per capita income	Equivalised income
Mean	758 346	757 197.1	297 802.7	492 329.6
Median	614 181	614 181	240 419.7	414 419.8
Minimum	1 200	1 200	400	692.8
Maximum	11 048 410	6 141 810	6 141 810	6 141 810

Table 2: Approaches to determining income for Russia (rubles, per year)

disposable household income and the per capita income. This allows, among other things, economies of scale to be taken into account in the context of household analysis.

You can notice that when the top and bottom code is applied, the bottom border has not changed, but the top border has decreased. This is important to notice, since the income distribution function for Russia has a long and heavy “tail”, which will be shown below.

	Household income (no top or bottom codes)	Household income	Per capita income	Equivalised income
Mean	49 766.68	50 048.54	17 873.21	31 469.66
Median	42 638.52	42 638.52	15 388.38	27 708.63
Minimum	-1 433 087	0	0	0
Maximum	9 816 924	426 385.2	382 560	382 560

Table 3: Approaches to determining income for Poland (zlotys, per year)

The sample of Poland and Spain has an interesting feature - disposable income can be negative. This occurs in cases where the loss as such has been retained

	Household income (no top or bottom codes)	Household income	Per capita income	Equivalised income
Mean	27 894.62	27 879.12	11 202.9	18 376.79
Median	23 103	23 103	9 525.12	16 064.05
Minimum	-27 380	0	0	0
Maximum	347 583	231 030	161 724.9	161 724.9

Table 4: Approaches to determining income for Spain (euros, per year)

at the data source rather than using bottom-coding techniques. This usually happens with self-employment income or capital income; in rare cases, it happens that taxes exceed gross income due to different accounting periods of income or incorrect calculation of taxes.

At the same time, for the data for Poland and Spain, we see an urgent need to use the top and bottom code, since, the minimum values of disposable income for these countries are negative, which does not allow calculating indices and measures of inequality.

3.2 Comparison of the inequality indicators

Next, we will calculate the Gini index using different approaches to determining income, which were previously calculated for each country. Here is a summary table of the results for all countries:

Note that for all countries the Gini index calculated for household disposable income is higher than when using per capita income and equivalent income. When

	Household income	Per capita income	Equivalised income
Russia	0.366	0.357	0.331
Poland	0.349	0.313	0.289
Spain	0.378	0.359	0.340

Table 5: Gini index calculated for various types of income

using equivalence scales, the level of inequality in Russia, Poland and Spain is noticeably reduced, since economies of scale are assumed – individuals from large households can save on buying large goods (TV, car), using one for all.

The Gini index for per capita income is lower than for households as a whole. We can assume that low-income households are smaller than high-income households. Also, when we calculate per capita income, we assume that income is equally distributed among household members.

Next, consider alternative measures of inequality, such as the Atkinson index, as well as the ratios 90/10, 90/50, and 80/50. The latter measures take into account the distribution of income by quantiles.

To calculate the Atkinson index, we will use formula (2) given in the theoretical part of the work. We will use the parameters $\varepsilon = 0.5$ and $\varepsilon = 1$, implying low and medium perception of inequality, respectively. In all cases, the equivalent income calculated according to the methodology of the Luxenburg Income Survey (income corrected by the root of the number of household members) will be used as the base.

Country	Atkinson index $\varepsilon = 0.5$	Atkinson index $\varepsilon = 1$	Ratio 90/10	Ratio 90/50	Ratio 80/50
Russia	0.089	0.17	4.62	2.118	2.714
Poland	0.072	0.149	3.492	1.85	2.198
Spain	0.098	0.204	5.159	2.017	2.868

Table 6: Alternative measures of inequality

Note that Spain again becomes the first in almost all inequality indicators (Table 6). However, it is important to emphasize that using the ratios allows us to notice an interesting detail about the structure of inequality in Spain and Russia. With almost equal Gini indexes, we observe a difference in the ratio of 90/50 (incomes of the richest 10% to the median level). This figure in Spain is lower than in Russia. Thus, it can be stated that in Russia the median income in relation to the income of the richest households is lower than in Spain.

Thereby, in Russia inequality is largely explained by the high incomes of the rich 10% and low incomes of the population to the median. Which, in turn, brings us once again to the concept of a long and "heavy" tail of income distribution in Russia and low incomes in the zone before the median. In Spain, however, inequality occurs, among other things, due to inequality among the population with low incomes. Poland, against the background of Russia and Spain, shows low estimates of inequality, as is the case with the Gini index.

3.3 Plotting the Lorentz curve

Furthermore, consider the Lorenz curve for Russia, Poland and Spain. Figure 1 shows the curve in green for Poland, in red for Spain, and in blue for Russia. We note right away that the smallest deviation from the general equality curve (black line) is observed in Poland. Also, the distribution of income for Poland is the most even.

The following are Spain and Russia, for which the Gini coefficient is almost equal. Nevertheless, it can be seen that the Lorentz curve for Russia is noticeably biased towards higher-income groups (percentile 75-100). Thus, the greatest concentration of income in Russia is observed precisely among them.

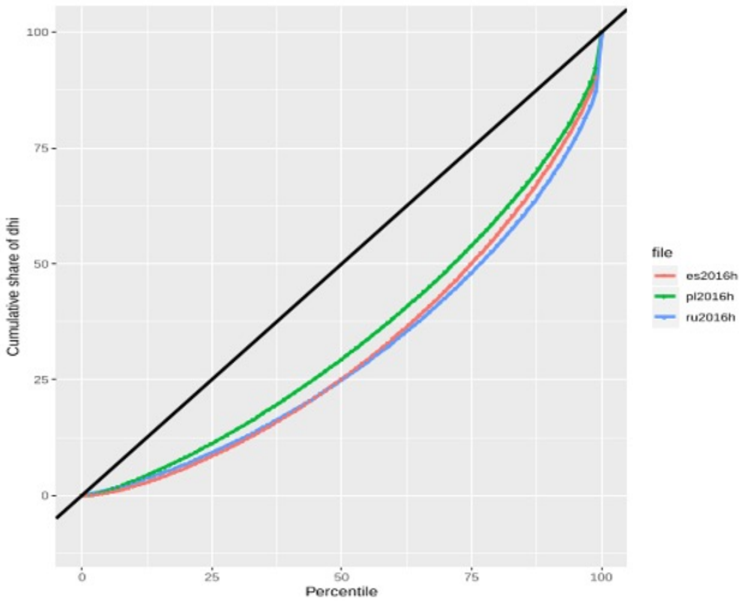

Figure 1: The Lorentz Curve for Russia, Poland and Spain

In Spain, on the other hand, there is a clear bias towards low-income groups.

It can be seen that the Lorentz curves for Russia and Spain intersect in the fifth quantile having almost equal Gini coefficients.

4 Estimation of the probability density function

This chapter will consider the probability density function of the disposable household income in Russia (variable *dhi*). We will analyze histograms, kernel density estimates for different intervals of the distribution. Next, we select the theoretical density function and give the calculation of the distribution parameters estimate. At the next stage, we will compare the empirical density estimate (kernel estimate) with the theoretical distribution and draw conclusions about the applicability of the used model for approximating the disposable household income in Russia.

4.1 Non-parametric methods

Earlier, we have already approximated the Lorentz curve and calculated the fund ratios, which also confirms our hypothesis about the long tail of the income distribution in Russia. It can be seen that the median is less than the mean, which indicates a right-sided asymmetry of the distribution. Also, the maximum value in the sample is almost 20 times higher than the income of the third quantile, which may indicate a long tail of the distribution.

The red dashed line marks the median of the distribution. Note that, to the

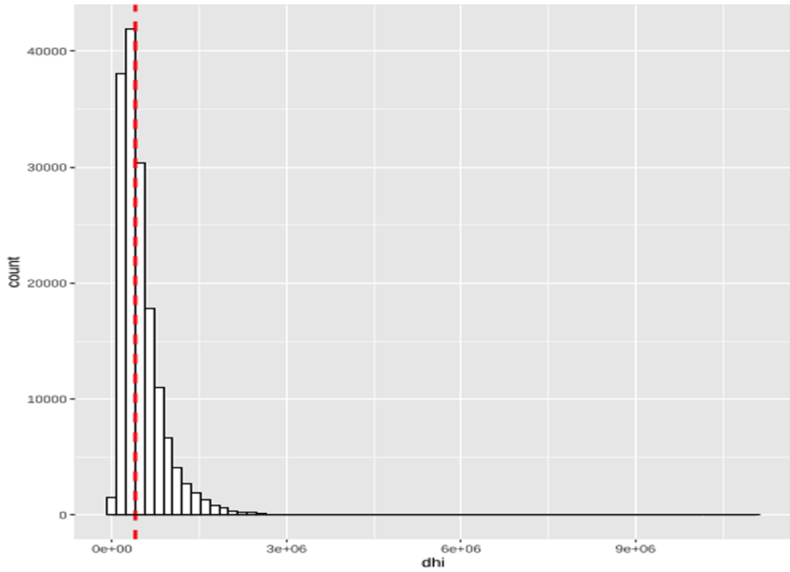

Figure 2: Histogram of the distribution of disposable income in Russia, 2016

median, density rises sharply in low- and middle-income groups. Further, there is a systematic decrease, forming a long tail of high-income groups of the population.

Once again, it can be suggested that the distribution function should have a right-sided skewness and that the tail of the income distribution should be well approximated.

Functions assumed at this stage:

1. Lognormal distribution with the density function (7) described in the theoretical part of the work.

2. Pareto distribution, described in the theoretical part of the work.

3. Gamma distribution.

The Weibull distribution will not account for the rapid growth in the low-income and middle-income zone.

Next, we will consider the kernel density estimate probability density function of income at two intervals - from 0 to 3 million rubles per year and more than 3 million rubles per year, since this will allow us to consider the distribution density function and its features in more detail. We will estimate the density function using the statistical programming environment R. The Gaussian kernel is used as the kernel function.

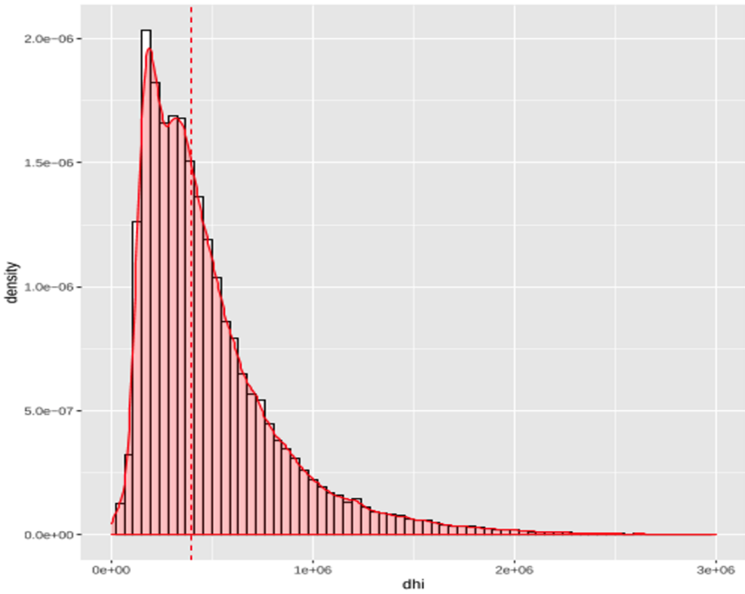

Figure 3: Kernel estimate of the density function of the disposable income in Russia in the range from 0 to 3,000,000 rubles. per year, 2016

Figure 3 shows the kernel estimate of the density function (red) as well as the median line (red dashed line). In this figure, two peaks of the distribution are clearly visible: the first and the highest in the region of low-income groups to the left of the median, the second - in the region of the median, slightly to the left of the median value. This type of the distribution density function differs from the adopted averaged income distribution density function. In developed countries, the peak falls on the zone closer to the median, the distance from the mode to the median, thus, decreases and a large share of the population falls into the middle class.

Consider the kernel estimate for the density function of the tail of the distribution:

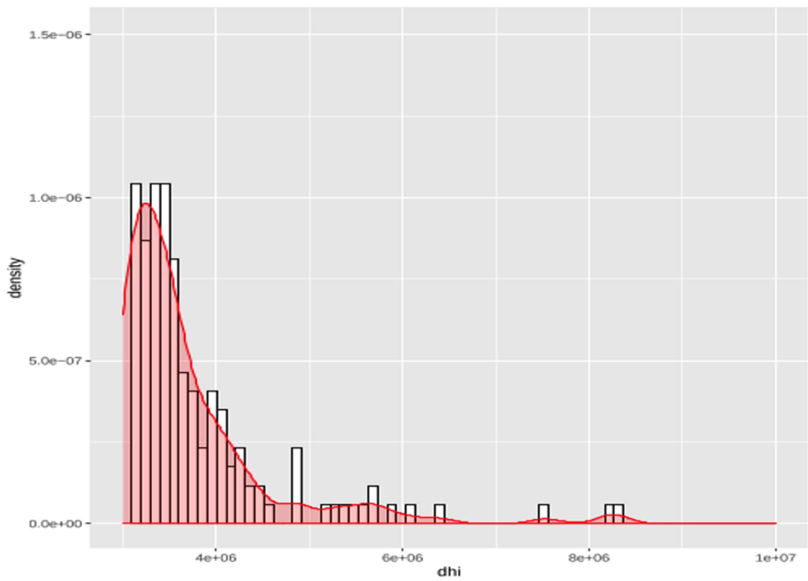

Figure 4: Kernel estimate of the density function of the disposable income in Russia in the range from 3,000,000 to 12,000,000 rubles per year, 2016

Note that the tail of the distribution is indeed quite long and the density of the distribution in the tail is high, relative to other income groups. This confirms the conclusion from the second chapter of the work that a large share of income is concentrated precisely in the tail of the distribution.

4.2 Parameter estimation

Having considered in detail the distribution function at the previous stages using nonparametric methods, we proceed to parametric estimation of the density function. We assume that the best approximation of the income distribution function is a lognormal distribution. We will estimate its parameters using the maximum likelihood method.

Maximum Likelihood is a popular estimation technique for many distributions because it picks the values of the distribution's parameters that make the data "more likely" than any other values of the parameters would make them. This is accomplished by maximizing the likelihood function of the parameters given the data.

To compute the Maximum Likelihood estimators, we start with the likelihood function.

$$\begin{aligned} L\left(\mu, \sigma^2 \middle| X\right) &= \prod_{i=1}^{n} \left[f(X_i|\mu, \sigma^2) \right] = \\ &= \left(2\pi\sigma^2\right)^{-n/2} \prod_{i=1}^{n} X_i^{-1} exp \left[\sum_{i=1}^{n} \frac{-\left(\ln\left(X_i\right) - \mu\right)^2}{2\sigma^2} \right] \end{aligned} \tag{8}$$

The log-likelihood function of the lognormal for the series of $X_i (i = 1, 2, ...n)$

is then derived by taking the natural log of the likelihood function:

$$\begin{aligned}\mathcal{L}\left(\mu,\sigma^2|X\right) &= \ln\left[\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\prod_{i=1}^{n}X_i^{-1}exp\left[\sum_{i=1}^{n}\frac{-\left(\ln\left(X_i\right)-\mu\right)^2}{2\sigma^2}\right]\right] = \\ = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) &- \sum_{i=1}^{n}\ln\left(X_i\right) - \frac{\sum_{i=1}^{n}\ln\left(X_i\right)^2}{2\sigma^2} + \frac{\sum_{i=1}^{n}\ln\left(X_i\right)\mu}{\sigma^2} - \frac{n\mu^2}{2}\end{aligned} \tag{9}$$

We now find $\hat{\mu}$ and $\hat{\sigma}$, which maximize $L(\mu,\sigma^2|X)$. To do this, we take the gradient of L with respect to μ and σ^2 and set it equal to 0: with respect to μ:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mu} = \frac{\sum_{i=1}^{n}\ln\left(X_i\right)}{\hat{\sigma}^2} - \frac{n\hat{\mu}}{2\hat{\sigma}^2} = 0 \\ \implies \hat{\mu} = \frac{\sum_{i=1}^{n}\ln\left(X_i\right)}{n}\end{aligned} \tag{10}$$

With respect to σ^2:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\sigma^2} = -\frac{n}{2\hat{\sigma}^2} - \frac{\sum_{i=1}^{n}\left(\ln\left(X_i\right)-\mu\right)^2}{2}\left(-\hat{\sigma}^2\right)^{-2} = 0 \\ \implies \hat{\sigma} = \frac{\sum_{i=1}^{n}\left[\ln(X_i) - \frac{\sum_{i=1}^{n}\ln(X_i)}{n}\right]^2}{n}\end{aligned} \tag{11}$$

Using R again to calculate the numerical parameters of the lognormal distribution for $n = 160,008$, we find that $\ln(\widehat{\sigma^2}) = 0.695$, that $\ln\left(\hat{\mu}\right) = 12.88$.

Let's plot both distributions (empirical and theoretical) in the same coordinate system to compare the simulation results. We will consider again two intervals: the main distribution up to 3,000,000 rubles per year and the tail of the distribution.

Figure 5 shows two estimates of the density functions. The empirical non-

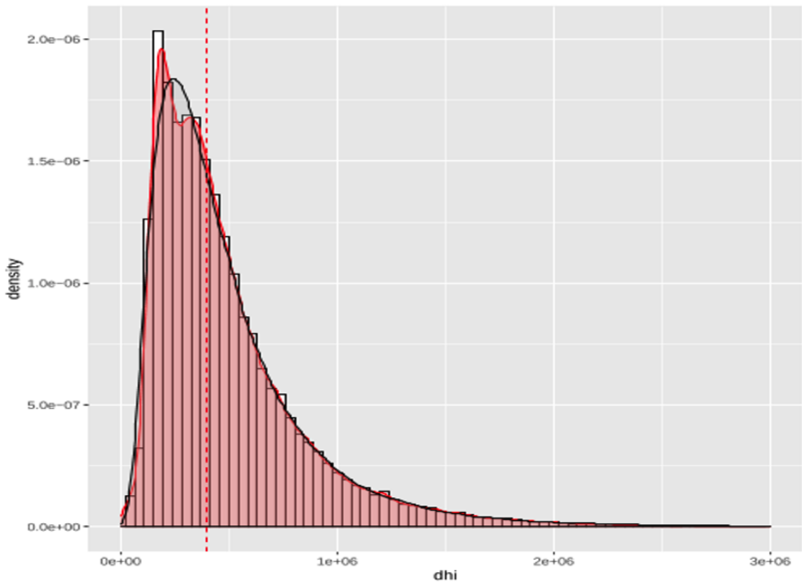

Figure 5: Parametric (lognormal) and non-parametric (kernel) estimates of the density function of the disposable income in Russia in the range from 0 to 3,000,000 rubles per year, 2016

parametric density function is shown in red. Black shows the estimate of the density function of the lognormal distribution for the parameters obtained above. Note that the resulting parametric density function underestimates the number of low-income groups. At the same time, there is an excessive density in the area of the median distribution – the middle class.

Thus, a lognormal estimate of the density function overestimates the size of the middle class in Russia and underestimates the size of low-income groups, as seen in Figure 5.

As shown in the theoretical part of the paper, the lognormal distribution density function incorrectly estimates the tail of the distribution. Figure 6 shows that

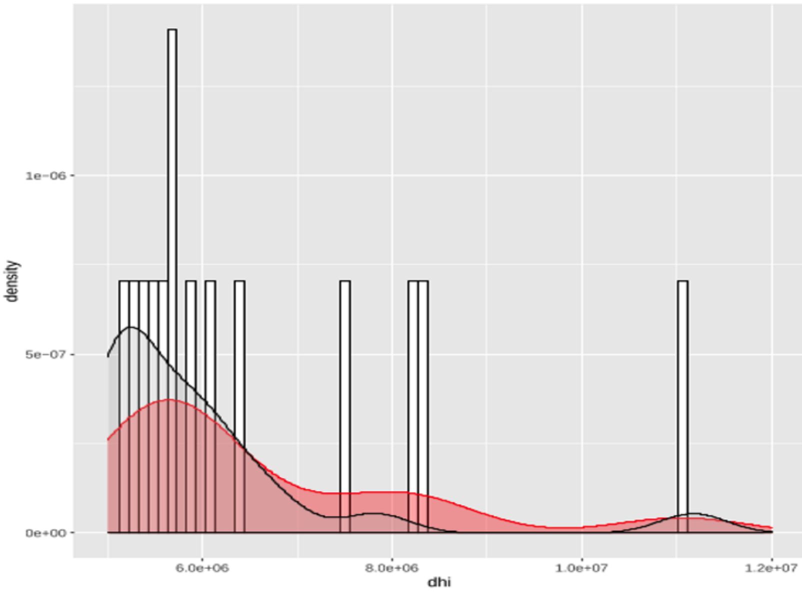

Figure 6: Parametric (lognormal) and non-parametric (kernel) estimates of the density function of the disposable income in Russia in the range from 5,000,000 to 12,000,000 rubles per year, 2016

in the tail of the distribution, the empirical density (red) significantly exceeds the density of the estimated lognormal distribution (black). Nevertheless, statistical tests, for example, the Kolmogorov-Smirnov test, do not reject the hypothesis that the theoretical and empirical distributions coincide in our sample. Indeed, if we take the natural logarithm of the vector of the initial data on the disposable income of the population, we will again see that the Kolmogorov-Smirnov test confirms the correspondence of the distribution (the logarithm of the initial data) to the normal one. However, if we consider other visual criteria, it can be seen that the logarithm of income does not follow the normal distribution:

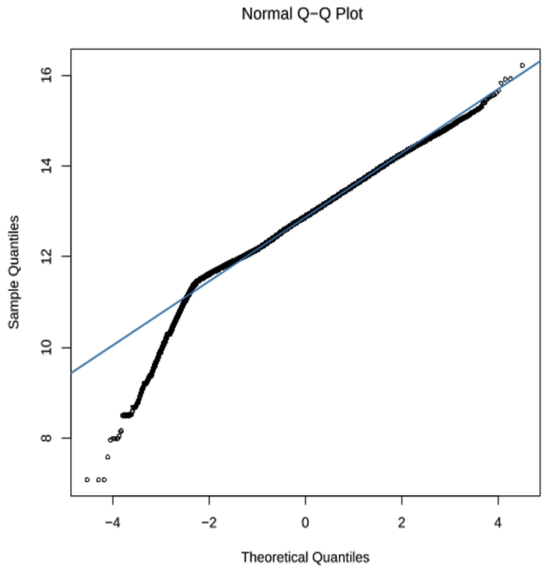

Figure 7: Normal Q-Q Plot of the logarithm of the initial data on disposable income

5 Conclusion

Thus, the assessment of the lognormal distribution as a model for the distribution of disposable incomes showed that the application of this distribution to the data for Russia gives a number of inaccuracies:

1. Lognormal distribution underestimates the size of the population in low-income groups. For Russia, as can be seen from the distribution density graph (non-parametric estimate), this group is the most numerous.

2. Lognormal distribution overestimates the size of the middle class around the median, which in turn can lead to misinterpretation of inequality coefficients.

3. Lognormal distribution underestimates the "heavy" tail of income distribu-

tion in Russia.

Nevertheless, the lognormal distribution is one of the basic estimates of the probability density function of the income. This comparison can shed light on many features of income distribution in a particular country. In our case, fitting the initial income distribution to the lognormal one also allowed us to identify a number of features of income distribution in Russia, described above:

1. The small number of the middle class in Russia.

2. Large share of the population is low-income.

3. Concentration of income in high income groups (tail of distribution).

These features, in turn, lead to the results we obtained for evaluating inequality in the second chapter of the paper.

6 References

1. Arnold Barry. Pareto and Generalized Pareto Distributions: Modeling Income Distributions and Lorenz Curves // Economic Studies in Equality, Social Exclusion and Well-Being. - New York : Springer, 2008 . - vol. 5.

2. Atkinson Anthony and Bourguignon François. Income Distribution Today [Article] // Handbook of Income Distribution. - 2015. - Vol. 2. - pp. xvii-xiv.

3. Atkinson Anthony. Inequality: What Can Be Done? . -: Harvard University Press, 2015.

4. Atkinson Anthony. On the measurement of inequality [Journal] // Journal of Economic Theory. - 1970. - Vol. 2 (3). - pp. 244–263.

5. Bernstein Bernstein, McNichol Elizabeth and Nicholas Andrew. Pulling Apart: A State-by-State Analysis of Income Trends [Book]. - Washington, DC : Center on Budget and Policy Priorities, 2008.

6. Cowell Frank and Flachaireb Emmanuel. Income distribution and inequality measurement: The problem of extreme values [Journal] // Journal of Econometrics. - 2007.

7. Cowell Frank. Measurament of Inequality // Handbook of Income Inequality. - : STICERD, London School of Economics and Political Science, 1988..

8. Cowell Frank. Wealth, Top Incomes and Inequality // National wealth: what is missing, why it matters. - 2017.. - pp. 175-204.

9. Davies James. Wealth and Economic Inequality [Journal] // The Oxford Handbook of Economic Inequality. - 2011.

10. European Social Policy Network. Revision of personal income tax in Poland: increase in the tax-free allowance for the lowest earners // ESPN Flash Report. - 2017. - vol. 13.

11. Milanovic Branko. Global inequality: A New Approach for the Age of Globalization [Book]. - [s.l.] : The Belknap Press of Harvard University Press, 2018.

12. Piketty Thomas. Capital in the Twenty-First Century [Book]. - [s.l.] : Harvard University Press , 2014.

13. Rosenblatt Murray. Remarks on Some Nonparametric Estimates of a Density Function. // The Annals of Mathematical Statistics. - 1956.. - 3 : vol. 27.

14. Scott David. Multivariate Density Estimation and Visualization // Humboldt-Universität zu Berlin. - Berlin : Center for Applied Statistics and Economics (CASE), 2004.

15. Solt Frederick. Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database [Journal] // Social Science Quarterly 101(3). - October 2020. - pp. 1183-1199.

16. Zucman Gabriel and Piketty Thomas. From Soviets to Oligarchs: Inequality and Property in Russia 1905-2016 [Article] // NBER Working Paper No. 23712. - August 2017

17. Luxembourg Income Study (LIS) Database, http://www.lisdatacenter.org (multiple countries; ru16h, pl16h, es16h). Luxembourg: LIS..