AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 215–221

Logarithmic-Normal Model of Income Distribution in the Czech Republic

Jitka Bartošová
University of Economics, Praque, Czech Republic

Abstract: While making a statistical model the main task is both, to find out a theoretical distribution function that would characterize empirical frequency distribution and to choose suitable methods to calculate parameters of the model. As a theoretical model of income distribution, logarithmic-normal distribution has been used so far. Recently used type of theoretical model was derived from the character of a particular feature and from a longtime experience with its behavior before revolution. One of the important tasks of the present time is to make statistical analysis of impact of economic changes on distribution of annual financial income per household, or per person after velvet revolution. This paper concentrates on verification of validity of the statistical model of income distribution in the Czech Republic that has been used so far. It also covers a suitable estimate of model's parameters.

Keywords: Parameter, Point Estimate.

1 Introduction

Economists' interest in income of the population in developed countries arises out of efforts to solve matters concerning the level of living standard of the population in an objective manner. Income models may be easily used to directly evaluate the level of standard or to compare the level of standard in different regions or nations. While making a statistical analysis of the level of living standard, we focus only on measurable elements of the level of living standard. In order to correctly quantify the element of the level of standard that directly depends on incomes, we need to characterize the level and structure of population income in their complexity, i.e. to find out suitable statistical models of income distributions both in different social classes and in the whole population, without regarding social classes. Knowledge of the current statistical model of income distribution, which is a simple approximation of sample distribution and knowledge of tendency of its parameters development may be used to predict behavior of the particular variable in the following period of time.

While making a statistical model it is important both, to find out a theoretical distribution function that would characterize empirical frequency distribution and to choose suitable methods to calculate parameters of the model. As a theoretical model of income distribution, logarithm-normal distribution has been used so far. Recently used type of theoretical model was derived from the character of a particular feature and from a longtime experience with its behavior before revolution. Especially the three-parameter logarithm-normal distribution has represented a good approximation of income distribution for most of social classes.

Before revolution, planned economy in the Czech Republic experienced high homogeneity of income in the population in all social classes. In presence, the transformation to market economic system has caused a significant change in income distribution. The variety of income sources and present process of differentiation of wages brings about discrepancies between empirical income distribution and the theoretical model. There are values of income that may be concerned as outliers and that causes contamination of the model (see Antoch and Vorlíčková, 1992; Jurečková, 2001; Bartošová, 2004d). Distributions in some social classes correspond to commixture of several theoretical curves etc. These differences are still deepening and in a different rate and inertia they are progressively reflected in both income distribution of the whole population without regarding social classes and income distribution of some social classes (see Bartošová, 2003, 2004a, 2004b, 2004c). There is some inertia in income distribution, so its changes would noticeably display in a term of several years after revolution. The first sample survey focusing on distribution of income of population, Mikrocensus, in which we may anticipate significant changes in its results was carried out by Czech Statistical Office in 1996. Therefore, one of the important tasks of the present time is to make statistical analysis of impact of economic changes on distribution of annual financial income in 1996. This contribution deals with the construction of the logarithm-normal model of households' incomes distribution in the Czech Republic in 1996 and its agreement with the empirical distribution.

2 Logarithm-Normal Model of Income Distribution

To characterize empirical households' income distribution, it's often convenient to use a parametric model. One of the most used parametric models of households' income distribution is the logarithm-normal distribution. This model is also used in economics to illustrate distribution of wages and job standardization. The logarithm-normal distribution competes with Weibull's distribution in the field of stipulation of the time to failure of the device. Besides, it's also frequently used for quality control and in the theory of reliability.

Initially, development of the logarithm-normal distribution was studied by Aitchison and Brown (1957). They were interested mainly in its application in astronomy, biology, sociology, economics or simulation of physical processes. These authors also formulated special reasons concerning the problem of the logarithm-normal distribution and income models. Wise (1966) described application of the logarithm-normal model in relation to the concentration of indicators that are time functions.

The biggest competitor of the logarithm-normal distribution in income modelling is Pater's distribution. The logarithm-normal curve corresponds to the empirical income distribution in a large central area, while in extremes it significantly diverges. On the contrary, Paret's curve is a suitable model of income distribution in extreme values (see Johnson et al., 1994). It means that it's suitable to use the logarithm-normal model for representation of income distribution of the majority of households of the central part of range and Paret's distribution for extremes.

Since the logarithm-normal distribution with parameters μ, σ^2 and γ, where γ is the theoretical minimum, represented a good approximation of income distribution in the Czech Republic before revolution in 1989, I used it also in income model after revolution.

Concordance rate of the empirical household's income distribution with the logarithm-normal model was quantified by the likelihood ration, i.e. statistic

$$LR(\mu, \sigma^2, \gamma|n) = 2[\ell(\vec{p}|n) - \ell(\vec{\pi}(\mu, \sigma^2, \gamma)|n)]\,, \tag{1}$$

where $\vec{p}$ is the vector of income empirical probability, $\vec{\pi}(\mu, \sigma^2, \gamma)$ is the vector of probabilities of occupation of particular classes and $\ell(\vec{p}|n)$, $\ell(\vec{\pi}(\mu, \sigma^2, \gamma)|n)$ are corresponding multinomial likelihood functions. Statistic (1) is asymptotically χ^2 with the number of degrees equal to the number of classes minus one, or minus four if the parameters μ, σ, γ are replaced by their estimates with required properties (see Anděl, 1993, 2002). Statistic LR was chosen from the wide range of concordance rate, because the method of maximal likelihood and its variation were used to estimate the parameters of the model.

The results are also influenced by the number of classes, where data are gathered during calculation. The problem of optimal number of classes is a matter of many titles. In this case I chose

$$k = 15 \cdot \sqrt[5]{\left(\frac{n}{100}\right)^2}\,, \tag{2}$$

which is suitable for sufficiently large sample, i.e. for $n > 80$ (see Williams, 2001).

I concentrated on the construction of the model of empirical distribution of two statistical features – income of the whole household and income per one member of the household. This way I got complex information about households' income in the Czech Republic in 1996 and its and its logarithm-normal model. Households with larger number of members usually had lower income per one member, so that these distributions must differ. To make the logarithm-normal model I used net year incomes of households in current prices. This is the only way to get results that reflect real state of income distribution in 1996, i.e. six years after Velvet Revolution.

2.1 Data Set

In 1996 sample survey Mikrocensus (including households' income data) was made in about 1% of households, which was about 28000 flats that time. While making statistical analysis of income distribution in 1996, I came out from complete data set that I got from Czech Statistical Office. Complete non-aggregated sample set enabled me to gain quality estimates of parameters of the models of the distribution. For purposes of research following data were chosen:

- Social class of the head of household
- Number of members in household
- Net income of household (CZK per year)

In connection with proceeding economical transformation new sources of incomes appeared and so social structure of sample sets was changing. Before Velvet Revolution households were divided into classes of workers, cooperative farmers, employees and retired. Social structure of sample set of households' incomes from.

1996 is in Table 1 (it reflects changes in sources of incomes as well as percentage representation of classes). As you can see in Table 1, three biggest social classes, i.e. households of workers, employees and retired without economically active members, formed

Table 1: Structure of set of households' incomes after Velvet Revolution (in 1996)

Social class	Size	%
Worker	8856	31.5
Self-employed (except agriculture)	1748	6.2
Employees	6915	24.6
Self-employed farmers	131	0.5
Farmers - member of cooperative	195	0.7
Retired with economically active members	1156	4.1
Retired without economically active members	8651	30.7
Unemployed	260	0.9
Others	236	0.8

86.8% of all households in the Czech Republic in 1996. New social classes that were born due to after revolution transformation, i.e. households of self-employed (except agriculture), self-employed farmers, unemployed and other households, formed about only 8.4% altogether. That's the reason why they can't significantly influence the income distribution of all households. Thus the character of the distribution will be determined mainly by the type of distribution of these three major classes.

2.2 Point Estimate of Parameters of the Model

An important part of the construction of the model is a choice of a suitable method of estimate of its parameters (see Anděl, 1993, 2002). As to assumed changes in income distribution, such as high variability, contamination etc., that came after Velvet Revolution, it's necessary to choose such method of parameters estimate giving good results also under these new conditions.

Therefore, I chose the method of maximal likelihood to estimate parameters μ, σ^2, and γ of the logarithm-normal model of households' income distribution in 1996. Because it's not possible to determine maximal likelihood estimate of the theoretical minimum directly and it's needy to use numerical maximization, I estimated this parameters by some more methods. To determine the minimum following methods were successively used

- Null (two-parametric model)
- Sample minimum
- Cohen's method, where the sample minimum is used as $100/(n+1)\%$ quintile of the theoretical distribution
- Method of maximal likelihood
- Method of likelihood ratio minimization

The first three methods are simple but at the same time they are totally insensitive or only a little sensitive to character of empirical income distribution. Thus it's interesting to compare their results with results of other two iteration methods.

The maximal likelihood estimates of parameter γ of three parameter logarithmic-normal distribution are numerically calculated as the maximum of the modified logarithmic likelihood function. In the case of sample size n, we have

$$\ell(\gamma) = -n\left[\hat{\mu}(\gamma) + \frac{1}{2}\log\hat{\sigma}^2(\gamma)\right], \tag{3}$$

where $\hat{\mu}(\gamma)$ and $\hat{\sigma}^2(\gamma)$ are maximal likelihood estimates of parameters and γ is a chosen value of theoretical minimum of the model.

Considering the character of the particular feature, I obtained the estimate by searching the maximum value of the function $\ell(\gamma)$ in the interval $(-x_{\max}, x_{\min})$, where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of incomes. The task was solved by iteration method – on a lattice, which density was increased in each iteration step.

The same iteration process was used to estimate parameter γ by way of numerical likelihood ratio minimization, i.e. by way of maximization of statistic $-LR(\gamma)$, given

$$-LR(\gamma) = 2\{\ell[\vec{\pi}(\hat{\mu}(\gamma), \hat{\sigma}^2(\gamma)] - \ell(\vec{p})\}\,. \tag{4}$$

3 Results

Quality of the logarithm-normal distribution of income distribution in 1996 is determined by quality of estimate of its parameters. Essentially the best estimate is the maximal likelihood estimate of all three parameters of the model. All the other models, which resulted from combining maximal likelihood estimates of μ and σ^2 with different estimate of γ, are compared by way of likelihood ratio in Tables 2 and 3. Besides information about likelihood ratio, there are values of corresponding 95% quintiles of χ^2 with the number of degrees equal to $k-4$ (k is defined by (2)).

As we can see in Tables 2 and 3, in almost all social classes $LR \approx \chi^2$ works, so that our logarithm-normal curves are mostly good approximations of empirical distribution of household' incomes in 1996. Only in the case of retired without EA members and the case of all households (without regarding social classes) the logarithm-normal model is totally unsuitable as model for empirical distribution. In both cases above statistics LR was more then ten times higher than corresponding quintile. Moreover graphs of kernel estimates of income distribution density per the whole household are two-peaked in both cases and it clearly signalizes commixture of two one-peaked curves. So that it's not possible to find such one-peaked parametric model that would well approximate this empirical distribution. On the other hand both corresponding income distributions per a member of household are one-peaked, therefore one should search for other, better parametric models.

It also follows from Tables 2 and 3 that besides logarithm-normal models with maximal likelihood estimates of all three parameters, as the most appropriate models can be considered models combining maximal likelihood estimates of μ and σ^2 with estimate of γ using likelihood ratio minimization. As the worst seems to be models with estimate of parameter by way of sample minimum. From the Tables above you can also see that even approximation of empirical income distribution by simple two-parametric variant of the logarithm-normal model is relatively good.

Table 2: Comparison of agreement of empirical distribution with logarithm-normal models by way of likelihood ratio (incomes of whole households)

Social class	$\gamma = 0$	$\gamma = x_{\text{min}}$	γ-Cohen	γ-LR	χ^2
Worker	475.9	907.3	300.8	232.8	108.6
Self-employed	77.4	108.6	81.3	77.1	59.3
Employees	145.2	264.6	114.1	109.4	99.6
Self-employed farmers	31.3	34.3	32.0	31.0	22.4
Farmers-member of cooperative	14.8	28.8	15.0	14.8	26.3
Retired with EA members	32.6	38.8	29.6	21.4	51.0
Retired without EA members	3874.5	3690.6	4027.7	3698.5	108.0
Unemployed	29.3	51.7	26.7	23.0	28.9
Others	26.9	21.8	33.3	25.3	27.6
All	3239.8	3233.1	3311.7	3224.4	167.5

Table 3: Comparison of agreement of empirical distribution with logarithm-normal models by way of likelihood ratio (incomes per a member of household)

Social class	$\gamma = 0$	$\gamma = x_{\text{min}}$	γ-Cohen	γ-LR	χ^2
Worker	170.6	279.7	174.7	163.1	108.6
Self-employed	94.6	55.3	97.2	55.8	59.3
Employees	167.4	189.1	104.5	104.2	99.6
Self-employed farmers	29.0	23.0	23.8	22.7	22.4
Farmers-member of cooperative	11.5	41.8	12.4	11.0	26.3
Retired with EA members	86.0	120.3	101.7	84.8	51.0
Retired without EA members	1520.6	1672.2	2675.8	1519.3	108.0
Unemployed	17.8	25.1	18.0	17.8	28.9
Others	76.9	38.3	58.4	44.4	27.6
All	2702.5	2533.3	3253.9	2534.2	167.5

4 Conclusions

Economical transformation in the Czech Republic from planned economy to market economic system (which started more then ten years ago) has brought some changes in level and structure of incomes of population. There's been a movement in level of incomes, but mainly a significant differentiation of incomes. In some social classes, there have arisen single households or groups of households with markedly higher or lower incomes. This way they can contaminate the chosen theoretical model. These facts cause rise of discrepancy of empirical distribution and the model used so far. Here comes the question: Isn't it necessary to look for another parametric model of households' income distribution?

From the obtained results it follows that in majority of social classes the logarithm-normal distribution can be considered as suitable model of household income distribution. Only in the case of retired without EA members, there's significant discrepancy between the empirical distribution and this model. The same situation appears in the case of all households – it's much influenced by the previous case.

Results of comparison of quality of models with other estimates of the theoretical minimum show (as one could expect) that the best agreement was achieved using the method of likelihood ratio minimization. The reasons of the slight difference from this result are probably restrictive conditions of iteration procedure. From the values of likelihood ratio it also follows that the differences among mentioned methods are not substantial.

References

Aitchison, J., and Brown, J. (1957). *The Lognormal Distribution with Special Reference to its Uses in Economics*. Cambridge: Cambridge University Press.

Anděl, J. (1993). *Statistical Methods (in Czech)*. Praha: MatfyzPress.

Anděl, J. (2002). *Basic Parts of Mathematical Statistics (in Czech)*. Praha: Charles University in Prague.

Antoch, J., and Vorlíčková, D. (1992). *Selected Methods of Statistical Data Analysis (in Czech)*. Praha: Academia.

Bartošová, J. (2003). Income models (in Czech). In J. Chajdiak, J. Luha, and S. Koróny (Eds.), *Proc. of the 12th International Seminar on Computational Statistics* (p. 7-11). Bratislava: SŠDS.

Bartošová, J. (2004a). Statistic model of households' annual income distribution in the Czech Republic. In J. Antoch (Ed.), *Compstat 2004, August 23-17, 2004, Praque, [CD-ROM]* (p. 1-8). Praha: Czech Statistical Society, for IASC.

Bartošová, J. (2004b). Contribution to analysis of distribution of household incomes (in Czech). In J. Antoch and G. Dohnal (Eds.), *Proc. of Robust, June 7-11, 2004, Třešť* (p. 451-458). Praha: JČMF.

Bartošová, J. (2004c). Statistical model of income distributions (in Czech). *Acta Universitatis Bohemiae Meridionales*, *7(2)*, 39-46.

Bartošová, J. (2004d). Contamination level estimate of household income distribution by distant observations in the (Czech Republic). *Forum Metricum Slovakum*, *8*, 74-78.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Vol. 1* (2nd ed.). New York: J. Wiley.

Jurečková, J. (2001). *Robust Statistical Methods (in Czech)*. Praha: Karolinum.

Williams, D. (2001). *Weighing the Odds: A Course in Probability and Statistics*. Cambridge: Cambridge Univ. Press.

Wise, M. (1966). Tracer dilution curves in cardiology and random walk and lognormal distributions. *Acta Physiologica Pharmacologica Neerlandica*, *14*, 175-204.

Author's address:

Jitka Bartošová
Department of Information Management
University of Economics in Prague
Jarošovská 1117/II
CZ-37701 Jindřichův Hradec, Czech Republic

Tel. +420 384 417 221
Fax +420 384 417
E-mail: bartosov@fm.vse.cz
http://www.fm.vse.cz