Lognormal distribution and using L-moment method for estimating its parameters

Diana Bílková

Abstract—L-moments are based on the linear combinations of order statistics. The question of L-moments presents a general theory covering the summarization and description of sample data sets, the summarization and description of theoretical distributions, but also the estimation of parameters of probability distributions and hypothesis testing for parameters of probability distributions. L-moments can be defined for any random variable in the case that its mean exists. Within the scope of modeling income or wage distribution we currently use the method of conventional moments, the quantile method or the maximum likelihood method. The theory of L-moments parallels to the other theories and the main advantage of the method of L-moments over these methods is that L-moments suffer less from impact of sampling variability. L-moments are more robust and they provide more secure results mainly in the case of small samples.

Common statistical methodology for description of the statistical samples is based on using conventional moments or cumulants. An alternative approach is based on using different characteristics which are called the L-moments. The L-moments are an analogy to the conventional moments, but they are based on linear combinations of the rank statistics, i.e. the L-statistics. Using the L-moments is theoretically more appropriate than the conventional moments because the L-moments characterize wider range of the distribution. When estimating from a sample, L-moments are more robust to the existence of the outliers in the data. The experience shows that in comparison with the conventional moments the L-moments are more difficult to distort and in finite samples they converge faster to the asymptotical normal distribution. Parameter estimations using the L-moments are especially in the case of small samples often more precise than estimates calculated using the maximum likelihood method.

This text concerns with the application of the L-moments in the case of larger samples and with the comparison of the precision of the method of L-moments with the precision of other methods (moment, quantile and maximum likelihood method) of parameter estimation in the case of larger samples. Three-parametric lognormal distribution is the basis of these analyses.

Keywords—Income distribution, L-moments, lognormal distribution, wage distribution.

Manuscript received October xx, 2011: Revised version received March xx, 2011. This work was supported by grant project IGS 24/2010 called "Analysis of the Development of Income Distribution in the Czech Republic since 1990 to the Financial Crisis and Comparison of This Development with the Development of the Income Distribution in Times of Financial Crisis − According to Sociological Groups, Gender, Age, Education, Profession Field and Region" from the University of Economics in Prague.

D. Bílková is with the University of Economics in Prague, Faculty of Informatics and Statistics, Department of Statistics and Probability, Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic (corresponding author to provide phone: +420 224 095 484; e-mail: bilkova@vse.cz).

Fig. 1 Basic information about the Czech Republic

I. Introduction

THE question of income and wage models is extensively covered in the statistical literature, see form example [8] − [9]. Data base for these calculations is composed of two parts: firstly, the individual data of a net annual household income per capita in the Czech Republic (in CZK), secondly, interval frequency distribution of gross monthly wage in the Czech Republic (in CZK). The aim of this work is to compare the accuracy of using the L-moment method of parameter estimation to the individual data with the accuracy of using this method to the data ordered to the form of interval frequency distribution. Another aim of this paper is to compare the accuracy of different methods of parameter estimation with the accuracy of the method of L-moments. Three-parametric lognormal distribution was a fundamental theoretical distribution for these calculations. Individual data on net annual household income per capita come from the statistical survey Microcensus (years 1992, 1996, 2002) and from the statistical survey EU-SILC − European Union Statistics on Income and Living Conditions (years 2005, 2006, 2007, 2008) organized by the Czech Statistical Office. The data in the form of interval frequency distribution come from the website of the Czech Statistical Office. Fig. 1 presents current basic information about the location of the Czech Republic in Europe and about the Czech Republic itself.

II. Methods

A. Three-Parametric Lognormal Distribution

The essence of lognormal distribution is treated in detail for example in [2]. Use of lognormal distribution in connection with income or wage distributions is described in [1] or [2].

Random variable X has three-parametric lognormal distribution LN(μ,σ^2,θ) with parameters μ, σ^2 and θ, where $-\infty < \mu < \infty$, $\sigma^2 > 0$ and $-\infty < \theta < \infty$, if its probability density function $f(x;\, \mu,\sigma^2,\theta)$ has the form

$$f(x;\mu,\sigma^2,\theta) = \frac{1}{\sigma\,(x-\theta)\sqrt{2\pi}}\, e^{-\frac{[\ln(x-\theta)-\mu]^2}{2\sigma^2}}, \quad x>\theta, \qquad = 0, \quad \text{else.} \tag{1}$$

Random variable

$$Y = \ln(X - \theta) \tag{2}$$

has a normal distribution N(μ,σ^2) and random variable

$$U = \frac{\ln(X-\theta)-\mu}{\sigma} \tag{3}$$

has a standardized normal distribution N (0, 1). The parameter μ is the expected value of random variable (2) and the parameter σ^2 is the variance of this random variable. Parameter θ is the theoretical minimum of random variable X. Figs. 2 and 3 represent the probability density functions of three-parametric lognormal curves depending on the values of their parameters.

The expected value (4) is the basic moment location characteristic of a random variable X having three-parametric lognormal distribution

$$E(X) = \theta + e^{\mu + \frac{\sigma^2}{2}}. \tag{4}$$

100 P% quantile is the basic quantile location characteristic of a random variable X

$$x_P = \theta + e^{\mu + \sigma u_P}, \tag{5}$$

where $0 < P < 1$ and u_P is 100·P% quantile of the standardized normal distribution. Substituting into the relation (5) $P = 0.5$, we get 50% quantile of three-parametric lognormal distribution, which is called median

$$\tilde{x} = \theta + e^{\mu}. \tag{6}$$

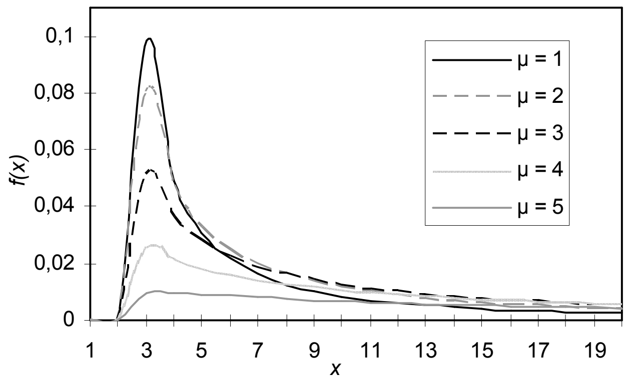

Fig. 2 Probability density function for the values of parameters $\sigma = 2$, $\theta = -2$

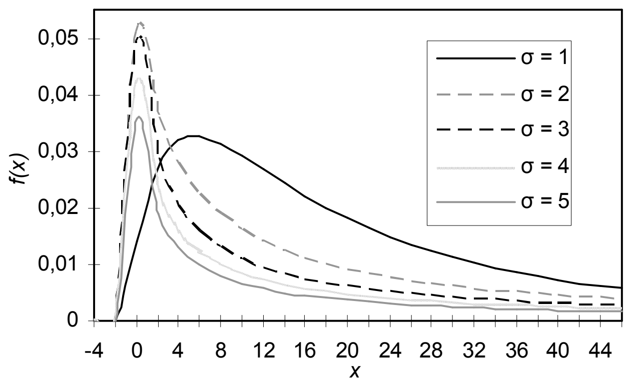

Fig. 3 Probability density function for the values of parameters $\mu = 3$, $\theta = -2$

The Median (6) divides the range of values of random variable X on the two equally likely parts. The mode (7) of random variable X is another often used location characteristic of three-parametric lognormal distribution

$$\hat{x} = \theta + e^{\mu - \sigma^2}. \tag{7}$$

The variance (8) of random variable X is a basic variability characteristic of three-parametric lognormal distribution

$$D(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1). \tag{8}$$

Standard deviation (9) is the square root of the variance and it represents another moment variability characteristic of the considered theoretical distribution

$$\sqrt{D(X)} = e^{\mu+\frac{\sigma^2}{2}}\sqrt{e^{\sigma^2} - 1}. \tag{9}$$

The coefficient of variation (10) is a characteristic of relative variability of this distribution and we get it by dividing the standard deviation to the expected value of the distribution

$$V(X) = \frac{e^{\mu+\frac{\sigma^2}{2}}\sqrt{e^{\sigma^2} - 1}}{\theta + e^{\mu+\frac{\sigma^2}{2}}}. \tag{10}$$

It is a dimensionless characteristic of variability.

The coefficient of skewness (11) and the coefficient of kurtosis (12) belong to basic moment shape characteristic of the distribution

$$\beta_1(X) = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}, \tag{11}$$

$$\beta_2(X) = e^{4\sigma^2} + 2\,e^{3\sigma^2} + 3\,e^{2\sigma^2} - 3. \tag{12}$$

B. Methods of Point Parameter Estimation

Question of parameter estimation of three-parametric lognormal distribution is already well developed in statistical literature, see for example [3]. We can use various methods to estimate the parameters of three-parametric lognormal distribution. We give as an example: moment method, quantile method, maximum likelihood method, method of L-moments, Kemsley's method, Cohen's method or graphical method.

Moment method

The essence of moment method of parameter estimation lies in the fact that we put the sample moments and the corresponding theoretical moments into equation. We can combine the general and the central moments. This method of estimating parameters is indeed very easy to use, but it is very inaccurate. In particular, the estimate of theoretical variance by its sample counterpart is very inaccurate. However, in the case of income and wage distribution we work with large sample sizes, and therefore the use of moment method of parameter estimation may not be a hindrance in terms of efficiency of estimators.

In the case of moment method of parameter estimation of three-parametric lognormal distribution we put the sample arithmetic mean $\bar{x}$ equal to the expected value of random variable X and we put the sample second central moment equal to the variance of random variable X. Furthermore, we put equal the sample third central moment m_3 with a theoretical third central moment of random variable X and we get the third equation. We obtain a system of moment equations

$$\bar{x} = \tilde{\theta} + e^{\tilde{\mu}+\frac{\tilde{\sigma}^2}{2}}, \tag{13}$$

$$m_2 = e^{2\tilde{\mu}+\tilde{\sigma}^2}(e^{\tilde{\sigma}^2} - 1), \tag{14}$$

$$m_3 = e^{3\tilde{\mu}+\frac{3}{2}\tilde{\sigma}^2}(e^{\tilde{\sigma}^2} - 1)^2(e^{\tilde{\sigma}^2} + 2). \tag{15}$$

We obtain from equations (14) and (15)

$$b_1^2 = m_3^2 \cdot m_2^{-3} = (e^{\tilde{\sigma}^2} - 1)(e^{\tilde{\sigma}^2} + 2)^2, \tag{16}$$

and therefore we also gain the moment parameter estimations of three-parametric lognormal distribution from the system of equations (13) to (15)

$$\tilde{\sigma}^2 = \ln\left[\sqrt[3]{1 + \frac{1}{2}b_1^2 + \sqrt{\left(1 + \frac{1}{2}b_1^2\right)^2 - 1}} + \sqrt[3]{1 + \frac{1}{2}b_1^2 - \sqrt{\left(1 + \frac{1}{2}b_1^2\right)^2 - 1}} - 1\right], \tag{17}$$

$$\tilde{\mu} = \frac{1}{2}\ln\frac{m_2}{e^{\tilde{\sigma}^2}(e^{\tilde{\sigma}^2} - 1)}, \tag{18}$$

$$\tilde{\theta} = \bar{x} - e^{\tilde{\mu}+\frac{\tilde{\sigma}^2}{2}}. \tag{19}$$

Quantile method and Kemsley's method

Quantile method of parameter estimation of three-parametric lognormal distribution is based on the use of three sample quantiles, namely there are $100{\cdot}P_1$% quantile, $100{\cdot}P_2$% quantile and $100{\cdot}P_3$% quantile, where $P_2 = 0{,}5$ and $P_3 = 1 - P_1$, and thus

$$u_{P_2} = 0 \quad \text{a} \quad u_{P_3} = -u_{P_1}.$$

We create a system of quantile equations by substituting to (5)

$$x_{P_1}^V = \theta^* + e^{\mu^* + \sigma^* u_{P_1}}, \tag{20}$$

$$x_{0,5}^V = \theta^* + e^{\mu^*}, \tag{21}$$

$$x_{(1-P_1)}^V = \theta^* + e^{\mu^* - \sigma^* u_{P_1}}, \tag{22}$$

where

$$x_{P_1}^V,\, x_{0,5}^V \quad \text{a} \quad x_{(1-P_1)}^V$$

are the corresponding sample quantiles. We obtain quantile parameter estimations of three-parametric lognormal distribution from the system of quantile equations (20) to (22)

$$\sigma^{2*} = \left[\frac{\ln \dfrac{x_{P_1}^V - x_{0,5}^V}{x_{0,5}^V - x_{(1-P_1)}^V}}{u_{P_1}} \right]^2 , \tag{23}$$

$$\mu^* = \ln \frac{x_{P_1}^V - x_{(1-P_1)}^V}{e^{\sigma^* u_{P_1}} - e^{-\sigma^* u_{P_1}}}, \tag{24}$$

$$\theta^* = x_{0,5}^V - e^{\mu^*}. \tag{25}$$

The sample median can be replaced by the sample arithmetic mean. Then we solve a similar system of equations as in the case of quantile method. This method is called Kemsley's method.

Maximum likelihood method and Cohen's method

If the value of the parameter θ in known, the likelihood function is maximized when the likelihood parameter estimations of three-parametric lognormal distribution have the form

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln(x_i - \theta)}{n}, \tag{26}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} [\ln(x_i - \theta) - \hat{\mu}]^2}{n}. \tag{27}$$

If the value of parameter θ is not known, this problem is considerably more complicated. If the parameter θ is estimated based on its sample minimum

$$\hat{\theta} = x_{\min}^V, \tag{28}$$

the likelihood function is unlimited. Maximum likelihood method is therefore sometimes combined with the Cohen's method. In this procedure, we put the smallest sample value to the equality with $100 \cdot (n+1)^{-1}$% quantile

$$x_{\min}^V = \hat{\theta} + e^{\hat{\mu} + \hat{\sigma}\, u(n+1)^{-1}}. \tag{29}$$

Equation (29) is then combined with a system of equations (26) and (27).

L-moment method

Question of L-moment is described in detail for example in [10]. We will assume that X is a real random variable with the distribution function $F(x)$ and quantile function $x(F)$ and $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$ are the rank statistics of the random sample of the size n selected from the distribution X. Then the r-th L-moment of the random variable X is defined as

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \quad r = 1, 2, 3, \ldots. \tag{30}$$

The letter 'L' in the name 'L-moments' is to stress the fact that r-th L-moment λ_r is a linear function of the expected rank statistics. Natural estimate of the L-moment λ_r based on the observed sample is furthermore a linear combination of the ordered values, i.e. the so called L-statistics. The expected value of the rank statistic is of the form

$$EX_{j:r} = \frac{r!}{(j-1)! \cdot (r-j)!} \cdot \int x\, [F(x)]^{j-1} \cdot [1 - F(x)]^{r-j}\, \mathrm{d}F(x). \tag{31}$$

If we plough the equation (31) in the equation (30), we get after some operations

$$\lambda_r = \int_0^1 x(F)\, P_{r-1}^*(F)\, \mathrm{d}F, \quad r = 1, 2, 3, \ldots, \tag{32}$$

where

$$P_r^*(F) = \sum_{k=0}^{r} p_{r,k}^*\, F^k \tag{33}$$

and

$$p_{r,k}^* = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k}, \tag{34}$$

where $P_r^*(F)$ represents the r-th shifted Legender's polynom which is related to the usual Legender's polynoms. Shifted Legender's polynoms are orthogonal on the interval (0,1) with a constant weight function. The first four L-moments are of the form

$$\lambda_1 = EX = \int_0^1 x(F)\, \mathrm{d}F, \tag{35}$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}) = \int_0^1 x(F) \cdot (2F - 1)\, \mathrm{d}F, \tag{36}$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2X_{2:3} + X_{1:3}) = \int_0^1 x(F) \cdot (6F^2 - 6F + 1)\, \mathrm{d}F, \tag{37}$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) = \int_0^1 x(F) \cdot (20F^3 - 30F^2 + 12F - 1)\, \mathrm{d}F. \tag{38}$$

Details about the L-moments can be found in [4] or [5]. The coefficients of the L-moments are defined as

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r = 3, 4, 5, ... \tag{39}$$

L-moments λ_1, λ_2, λ_3, …, λ_r and coefficients of L-moments τ_1, τ_2, τ_3, …, τ_r can be used as the characteristics of the distribution. L-moments are in a way similar to the conventional central moments and coefficients of L-moments are similar to the moment ratios. Especially L-moments λ_1 and λ_2 and coefficients of the L-moments τ_3 and τ_4 are considered to be characteristics of the location, variability and skewness.

Using the equations (35) to (37) and the equation (39), we get the first three L-moments of the three-parametric lognormal distribution LN(μ, σ^2, ξ), which is described e.g. in [5]. The following relations are valid for these L-moments

$$\lambda_1 = \xi + \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{40}$$

$$\lambda_2 = \exp\left(\mu + \frac{\sigma^2}{2}\right)\cdot \operatorname{erf}\left(\frac{\sigma}{2}\right), \tag{41}$$

$$\tau_3 = \frac{6\pi^{-1/2}}{\operatorname{erf}\left(\frac{\sigma}{2}\right)} \cdot \int_0^{\sigma/2} \operatorname{erf}\left(\frac{x}{\sqrt{3}}\right)\cdot \exp(-x^2)\,\mathrm{d}x, \tag{42}$$

where erf(z) is the so called error function defined as

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\cdot\int_0^z e^{-t^2}\,\mathrm{d}t. \tag{43}$$

Now we will assume that $x_1, x_2, \ldots, x_n$ is a random sample and $x_{1:n} \le x_{2:n} \le \ldots \le x_{n:n}$ is the ordered sample. The r-th sample L-moment is defined as

$$l_r = \binom{n}{r}^{-1} \cdot \sum_{1 \le i_1 \le i_2 \le \ldots \le i_r \le n} \sum \cdots \sum r^{-1} \cdot \sum_{k=0}^{r-1} (-1)^k \cdot \binom{r-1}{k} \cdot x_{i_{r-k}:n}, \quad r = 1, 2, \ldots, n. \tag{44}$$

We can write specifically for the first four sample L-moments

$$l_1 = n^{-1} \cdot \sum_i x_i, \tag{45}$$

$$l_2 = \frac{1}{2}\cdot\binom{n}{2}^{-1} \cdot \sum_{i>j}\sum (x_{i:n} - x_{j:n}), \tag{46}$$

$$l_3 = \frac{1}{3}\cdot\binom{n}{3}^{-1} \cdot \sum_{i>j>k}\sum\sum (x_{i:n} - 2x_{j:n} + x_{k:n}), \tag{47}$$

$$l_4 = \frac{1}{4}\cdot\binom{n}{4}^{-1} \cdot \sum_{i>j>k>l}\sum\sum\sum (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}). \tag{48}$$

Sample L-moments can be used similarly as the conventional sample moments because they characterize basic properties of the sample distribution and estimate the corresponding properties of the distribution from which were the data sampled. They might be also used to estimate the parameters of this distribution. In these cases L-moments are of then used instead of the conventional moments because as linear functions of the data they are less sensitive on the sample variability or on the error size in the case of the presence of the extreme values in the data than the conventional moments. Therefore it is assumed that the L-moments provide more precise and robust estimates of the characteristics of parameters of the population probability distribution.

Let us denote the distribution function of the standard normal distribution as Φ, then Φ^{-1} represents the quantile function of the standard normal distribution. The following relation holds for the distribution function of the three-parametric lognormal distribution LN(μ, σ^2, ξ)

$$F = \Phi\left[\frac{\ln(x-\xi) - \mu}{\sigma}\right]. \tag{49}$$

The coefficients of L-moments (39) are then commonly estimated using the following estimates

$$t_r = \frac{l_r}{l_2}, \quad r = 3, 4, 5, ... \tag{50}$$

The estimates of the three-parametric lognormal distribution can then be calculated as

$$z = \sqrt{\frac{8}{3}}\cdot\Phi^{-1}\left(\frac{1+t_3}{2}\right), \tag{51}$$

$$\hat{\sigma} \approx 0{,}999\,281 z - 0{,}006\,118 z^3 + 0{,}000\,127 z^5, \tag{52}$$

$$\hat{\mu} = \ln\left[\frac{l_2}{\operatorname{erf}\left(\frac{\hat{\sigma}}{2}\right)}\right] - \frac{\hat{\sigma}^2}{2}, \tag{53}$$

$$\hat{\theta} = l_1 - \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right). \tag{54}$$

More on L-moments is for example in [6], [11] or [12].

C. Appropriateness of the Model

It is also necessary to assess the suitability of the constructed model or choose a model from several alternatives, which is made by some criterion, which can be a sum of absolute deviations of the observed and theoretical frequencies for all intervals

$$S = \sum_{i=1}^{k} \left| n_i - n\pi_i \right| \tag{55}$$

or known criterion χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}, \tag{56}$$

where n_i are observed frequencies in individual intervals, π_i are theoretical probabilities of membership of statistical unit into the i-th interval, n is the total sample size of corresponding statistical file, $n \cdot \pi_i$ are the theoretical frequencies in individual intervals, $i = 1, 2, ..., k$, and k is the number of intervals.

The question of the appropriateness of the given curve for model of the distribution of income and wage is not entirely conventional mathematical-statistical problem in which we test the null hypothesis "H_0: The sample comes from the supposed theoretical distribution" against the alternative hypothesis "H_1: non H_0 ",because in goodness of fit tests in the case of income and wage distribution we meet frequently with the fact that we work with large sample sizes and therefore the test would almost always lead to the rejection of the null hypothesis. This results not only from the fact that with such large sample sizes the power of the test is so high at the chosen significance level that the test uncovers all the slightest deviations of the actual income or wage distribution and model, but it also results from the principle of the construction of test. But, practically we are not interested in such small deviations, so only gross agreement of the model with reality is sufficient and we so called "borrow" the model (curve). Test criterion χ^2 can be used in that direction only tentatively. When evaluating the suitability of the model we proceed to a large extent subjective and we rely on the experience and logical analysis. More is for example in [2].

D. Another Characteristics of Differentiation

There are various characteristics of variability of incomes and wages (or differentiation of incomes and wages) – variance, standard deviation, coefficient of variation or Gini index. In this article, only variance, standard deviation and a coefficient of variation are used. As L-moments are of interests, we give a few comments on the relation between the two- and three-parametric lognormal distribution and characteristics of differentiation.

If we substitute $\theta = 0$ into the formulas of three-parametric lognormal distribution, we obtain two-parametric lognormal distribution. It follows from the formula (10) that the coefficient of variation depends only on one parameter σ^2 in the case of two-parametric lognormal distribution

$$V(X) = \sqrt{e^{\sigma^2} - 1}\,. \tag{57}$$

Formulas for Gini coefficient can be found in the form

$$G = \operatorname{erf}\left(\frac{\sigma}{2}\right) \tag{58}$$

or equivalently in the form

$$G = 2\,\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1. \tag{59}$$

Unfortunately, in the case of the three-parametric lognormal distribution it is not true and both characteristics depend on all three parameters, see (10) for the case of coefficient of variation. We substitute $r = 2$ into the formula (30) and we obtain

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{2:1}) = \frac{1}{2} E\left| X_1 - X_2 \right| \tag{60}$$

and we conclude that Gini mean difference equals $2\lambda_2$ (see [5]). Gini coefficient can be evaluated as $\frac{\lambda_2}{\lambda_1}$. We obtain for the Gini coefficient G of the three-parametric lognormal distribution a formula

$$G = \frac{e^{\mu + \frac{\sigma^2}{2}} \cdot \operatorname{erf}\left(\frac{\sigma}{2}\right)}{\theta + e^{\mu + \frac{\sigma^2}{2}}}. \tag{61}$$

In this text, Gini coefficients are not included but from previous considerations the usefulness of L-moments in evaluating these characteristics is clear.

E. Four-Parametric Lognormal Distribution

Random variable X has four-parametric lognormal distribution $LN(\mu,\sigma^2,\theta,\tau)$ with parameters μ, σ^2, θ and τ, where $-\infty < \mu < \infty$, $\sigma^2 > 0$ and $-\infty < \theta < \tau < \infty$, if its probability density function $f(x;\, \mu,\sigma^2,\theta,)$ has the form

$$\begin{aligned} f(x;\mu,\sigma^2,\theta,\tau) &= \frac{(\tau-\theta)}{\sigma(x-\theta)(\tau-x)\sqrt{2\pi}}\, e^{-\frac{\left(\ln\frac{x-\theta}{\tau-x} - \mu\right)^2}{2\sigma^2}}, \quad \theta < x < \tau, \\ &= 0, \quad \text{else.} \end{aligned} \tag{62}$$

Random variable

$$Y = \ln \frac{X - \theta}{\tau - X} \qquad (63)$$

has a normal distribution N(μ,σ^2) and random variable

$$U = \frac{\ln \frac{X - \theta}{\tau - X} - \mu}{\sigma} \qquad (64)$$

has a standardized normal distribution N (0, 1). The parameter μ is the expected value of random variable (63) and the parameter σ^2 is the variance of this random variable. Parameter θ is the theoretical minimum of random variable X and parameter τ is the theoretical maximum of this variable. Figs. 4 and 5 represent the probability density functions of four-parametric lognormal curves depending on the values of their parameters.

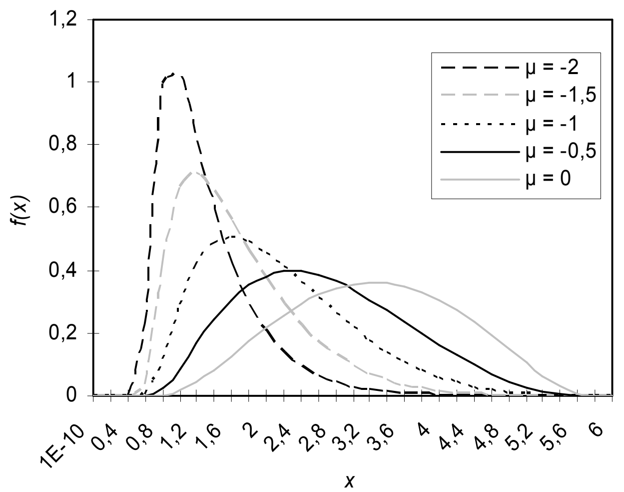

Fig. 4 Probability density function for the values of parameters σ = 0,8, θ = 0,5, τ = 6

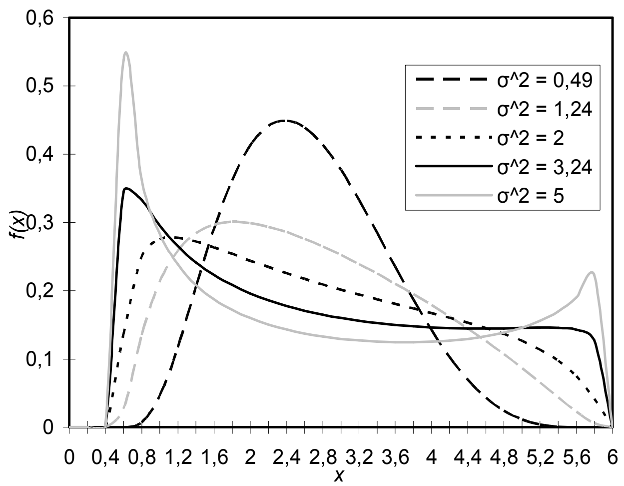

Fig. 5 Probability density function for the values of parameters μ = 0,5, θ = 0,5, τ = 6

III. Analysis and Results

Tabs. 1 to 14 present the estimated parameters of three-parametric lognormal curves using various methods of point parameter estimation (method of L-moments, moment method, quantile method and maximum likelihood method) and the sample characteristics on the basis of these the parameters were estimated. We can see from Tables 7, 11 and 13 that the value of the parameter θ (theoretical beginning of the distribution) is negative in many cases. This means that

Tab. 1 Sample L-moments − Income

	Sample L-moments		
Year	l_1	l_2	l_3
1992	35,246.51	7,874.26	2,622.14
1996	66,121.92	16,237.54	5,685.46
2002	105,029.89	27,978.40	10,229.62
2005	111,023.71	28,340.18	9,113.57
2006	114,945.08	28,800.68	9,286.18
2007	123,806.49	30,126.11	9,530.57
2008	132,877.19	31,078.96	9,702.45

Tab. 2 Parameter estimations of three-parametric lognormal distribution obtained using the L-moment method − Income

	Parameter estimation		
Year	μ	σ^2	θ
1992	9.696	0.490	14,491.687
1996	10.343	0.545	25,362.753
2002	10.819	0.598	37,685.637
2005	11.028	0.455	33,738.911
2006	11.040	0.458	36,606.903
2007	11.112	0.440	40,327.610
2008	11.163	0.428	45,634.578

Tab. 3 Sample characteristics (arithmetic mean $\bar{x}$, standard deviation s and coefficient of skewness b_1) − Income

	Sample characteristics		
Year	$\bar{x}$	s	b_1
1992	35,247	19,364	7.815
1996	68,286	51,102	17.606
2002	105,030	83,598	17.142
2005	111,024	77,676	14.907
2006	114,945	74,503	10.395
2007	123,806	74,578	7.727
2008	132,877	73,982	6.979

Tab. 4 Parameter estimations of three-parametric lognormal distribution obtained using the moment method − Income

	Parameter estimation		
Year	μ	σ^2	θ
1992	8.883	1.173	22,284.335
1996	9.154	1.780	45,269.967
2002	9.668	1.760	66,925.879
2005	9.710	1.656	73,299.950
2006	9.976	1.386	71,936.249
2007	10.242	1.165	73,575.417
2008	10.328	1.089	80,180.795

Tab. 5 Sample quartiles − Income

Year	Sample quartiles		
	$\tilde{x}_{0,25}$	$\tilde{x}_{0,50}$	$\tilde{x}_{0,75}$
1992	25,900	31,000	39,298
1996	47,550	57,700	76,550
2002	73,464	89,204	115,966
2005	79,600	97,050	124,068
2006	82,998	100,640	128,000
2007	90,000	108,744	138,000
2008	97,160	117,497	148,937

Tab. 6 Parameter estimations of three-parametric lognormal distribution obtained using the quantile method − Income

Year	Parameter estimation		
	μ	σ^2	θ
1992	9.490	0.521	17,766.792
1996	9.998	0.842	35,708.333
2002	10.551	0.619	50,986.446
2005	10.805	0.420	47,774.906
2006	10.813	0.423	50,970.817
2007	10.862	0.436	56,577.479
2008	10.961	0.417	59,909.386

Tab. 7 Parameter estimations of three-parametric lognormal distribution obtained using the maximum likelihood method − Income

Year	Parameter estimation		
	μ	σ^2	θ
1992	10.384	0.152	-0.342
1996	10.995	0.180	52.236
2002	11.438	0.211	73.525
2005	11.503	0.206	-2.050
2006	11.542	0.199	-8.805
2007	11.623	0.190	-42.288
2008	11.703	0.177	-171.167

Tab. 8 Sample L-moments − Wage

Year	Sample quartiles		
	l_1	l_2	l_3
2002	17,437.49	4,251.48	1,267.44
2003	18,663.18	4,524.95	1,251.90
2004	19,697.57	5,001.34	1,586.09
2005	20,738.14	5,262.93	1,636.67
2006	21,803.28	5,454.74	1,738.23
2007	23,882.83	6,577.65	2,627.93
2008	25,477.59	6,993.72	2,737.94

Tab. 9 Parameter estimations of three-parametric lognormal distribution obtained using the L-moment method − Wage

Year	Parameter estimation		
	μ	σ^2	θ
2002	9.238	0.388	4,952.259
2003	9.402	0.332	4,364.869
2004	9.313	0.442	5,872.138
2005	9.392	0.424	5,908.390
2006	9.393	0.447	6,795.207
2007	9.222	0.724	9,349.280
2008	9.319	0.693	9,719.297

Tab. 10 Sample characteristics (arithmetic mean $\bar{x}$, standard deviation s and coefficient of skewness b_1) − Wage

Year	Sample characteristics		
	$\bar{x}$	s	b_1
2002	17,437	8,321	1.817
2003	18,663	8,657	1.354
2004	19,698	9,804	1.614
2005	20,738	10,180	1.481
2006	21,803	10,477	1.419
2007	23,883	13,776	2.338
2008	25,478	14,485	2.191

Tab. 11 Parameter estimations of three-parametric lognormal distribution obtained using the moment method − Wage

Year	Parameter estimation		
	μ	σ^2	θ
2002	9.492	0.264	2,311.688
2003	9.837	0.166	-1,681.293
2004	9.779	0.221	-25.695
2005	9.906	0.193	-1,339.601
2006	9.979	0.180	-1,805.527
2007	9.734	0.377	3,509.924
2008	9.851	0.345	2,920.381

Tab. 12 Sample quartiles − Wage

Year	Sample quartiles		
	$\tilde{x}_{0,25}$	$\tilde{x}_{0,50}$	$\tilde{x}_{0,75}$
2002	11,944	15,545	20,215
2003	12,728	16,735	22,224
2004	13,416	17,709	23,077
2005	14,063	18,597	24,470
2006	14,717	19,514	25,675
2007	15,769	20,910	27,545
2008	16,853	22,225	29,404

Tab. 13 Parameter estimations of three-parametric lognormal distribution obtained using the quantile method − Wage

Year	Parameter estimation		
	μ	σ^2	θ
2002	9.663	0.149	-185.316
2003	9.605	0.218	1,899.151
2004	9.974	0.110	-3,742.702
2005	9.897	0.147	-1,283.306
2006	9.983	0.138	-2,144.719
2007	10.036	0.143	-1,919.373
2008	9.968	0.185	887.792

Tab. 14 Parameter estimations of three-parametric lognormal distribution obtained using the maximum likelihood method − Wage

Year	Parameter estimation		
	μ	σ^2	θ
2002	8.977	0.828	6,364.635
2003	9.024	0.615	6,679.910
2004	9.363	0.306	3,090.038
2005	9.400	0.329	4,134.624
2006	9.159	0.742	8,070.167
2007	9.487	0.369	2,586.616
2008	9.593	0.341	3,324.455

lognormal curve gets into negative territory at the beginning of its course.

Because of a very tight contact of the lower tail of the lognormal curve with the horizontal axes, this fact does not have to be a problem for a good fit of the model. The advantage of the lognormal models is that the parameters have an easy interpretation. Also some parametric functions of these models have straight interpretation. In the case that the estimated value of parameter θ is negative, we can not really interpret this value.

Figs. 6 to 13 show the probability density functions of three-parametric lognormal curves, whose parameters were estimated using different methods of parameter estimation. We can also see from these figures the development of theoretical income distribution in the years in 1992, 1996, 2002, 2005 to 2008 (Figs. 6 to 9) and the development of theoretical wage distribution in the years 2002 to 2008 (Figs. 10 to 13). Although the shapes of probability density function of three-parametric lognormal curves differ considerably between the used methods of point parameter estimation, we can observe certain trends in their development. We can see form Figs. 6 to 13 that as in the case of income, so in the case of wage distribution, characteristics of the level of these distributions increase gradually and characteristics of income and wage differentiation increase gradually, too. Therefore, data can not be considered homoskedastic in terms of the same variability in the same distributions as the characteristics of absolute variability grow in time. We see also from Figs. 6 to 13 the gradual decline of characteristics of shape of the distribution (skewness and kurtosis).

Figs. 14 to 20 represent the histograms of observed interval frequency distribution of net annual household income per capita in 1992, 1996, 2002, 2005 to 2008. Histograms of observed interval frequency distribution of gross monthly wage in 2002 to 2008 could not be constructed due to non-uniform width of the individual intervals. The interval frequency distributions with unequal wide of intervals were taken from the official website of the Czech Statistical Office and the frequency distribution histogram would lose any visual informative about the shape of the frequency distribution in this case. Figs. 21 to 24 also provide approximate information about the accuracy of the used methods of parameter estimation. Figs. 21 and 23 represent the development of the sample arithmetic mean and the development of theoretical expected values of three-parametric lognormal distribution with parameters estimated using different methods of parameter estimation. Figs. 22 and 24 represent the development of the sample median and the development of theoretical medians of three-parametric lognormal distribution with parameters estimated using different methods of parameter estimation. It is important to note, however, that Figs. 21 and 23 give nothing about the accuracy of moment method of parameter estimation, because equality of the sample arithmetic mean and theoretical expected value represents one of three moment equations. In this case, the course of development of sample arithmetic mean coincides with the course of development of theoretical expected value of three-parametric lognormal distribution with

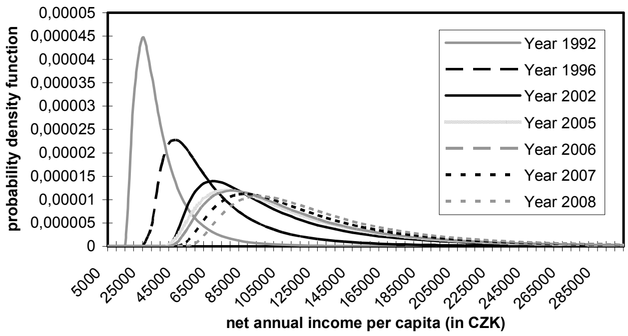

Fig. 6 Probability density function of net annual household income per capita − L-moment method

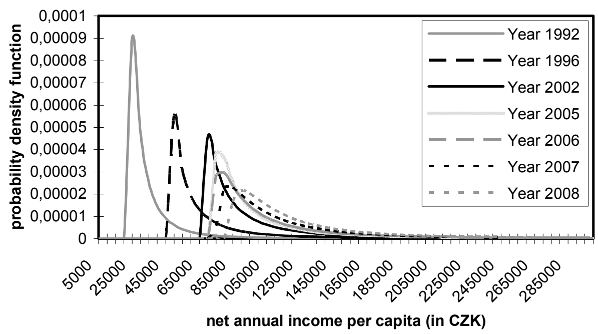

Fig. 7: Probability density function of net annual household income per capita − Moment method

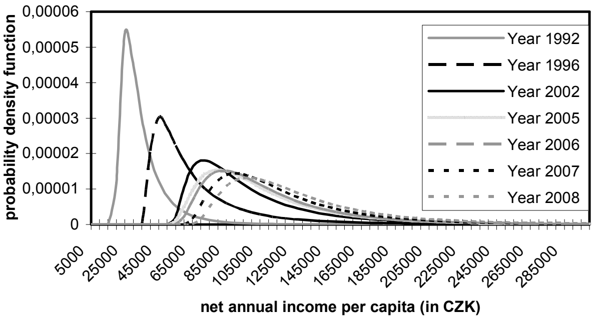

Fig. 8 Probability density function of net annual household income per capita − Quantile method

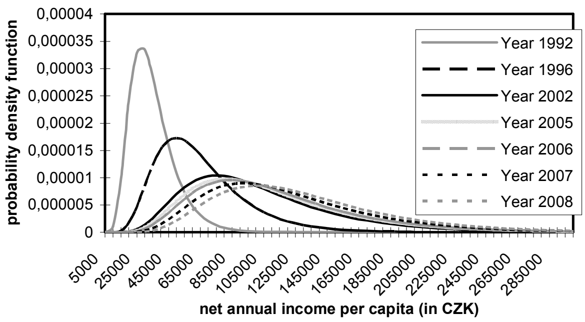

Fig. 9 Probability density function of net annual household income per capita − Maximum likelihood method

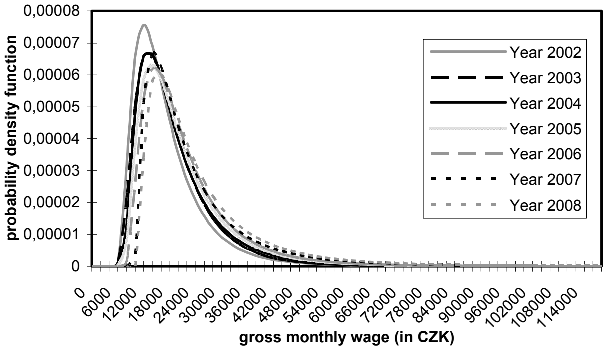

Fig. 10: Probability density function of gross monthly wage − L-moment method

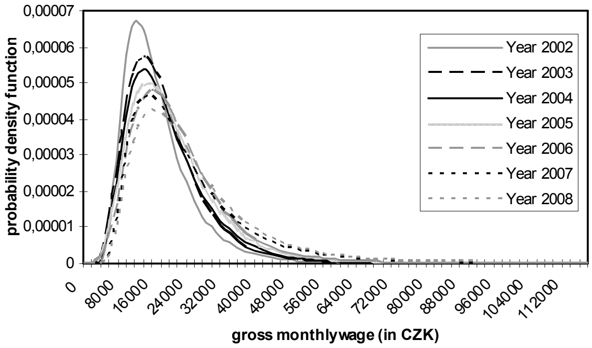

Fig. 11 Probability density function of gross monthly wage − Moment method

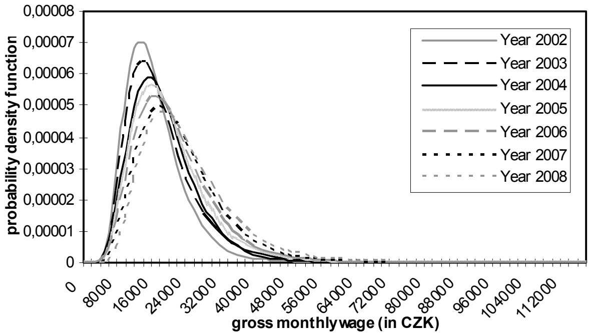

Fig. 12 Probability density function of gross monthly wage − Quantile method

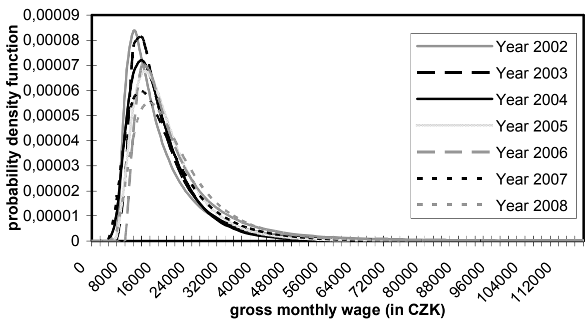

Fig. 13 Probability density function of gross monthly wage − Maximum likelihood method

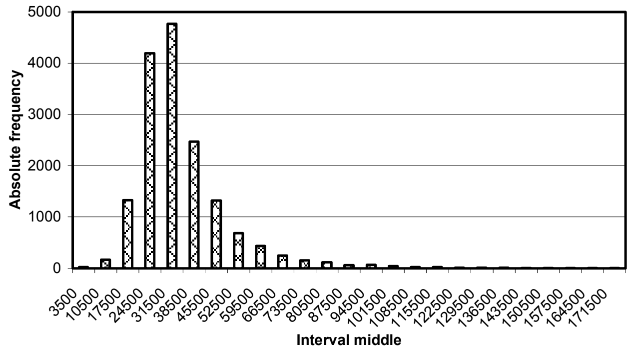

Fig. 14 Interval frequency distribution of net annual household income per capita in 1992

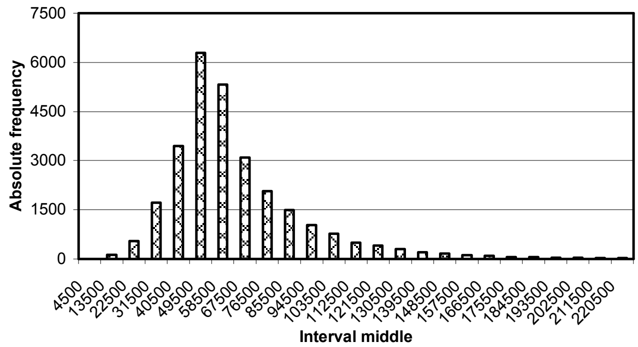

Fig. 15 Interval frequency distribution of net annual household income per capita in 1996

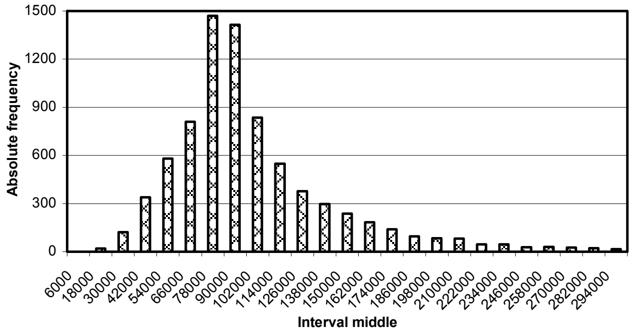

Fig. 16 Interval frequency distribution of net annual household income per capita in 2002

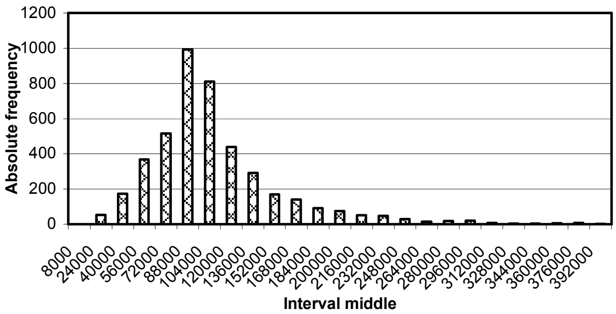

Fig. 17 Interval frequency distribution of net annual household income per capita in 2005

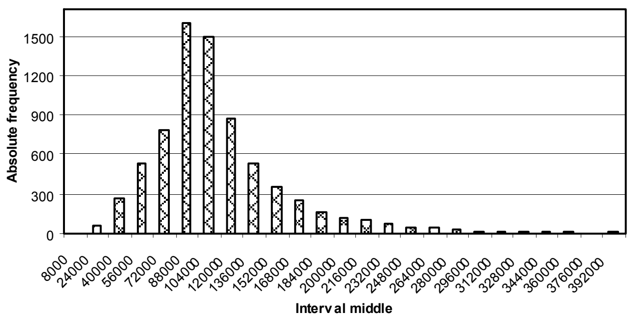

Fig. 18 Interval frequency distribution of net annual household income per capita in 2006

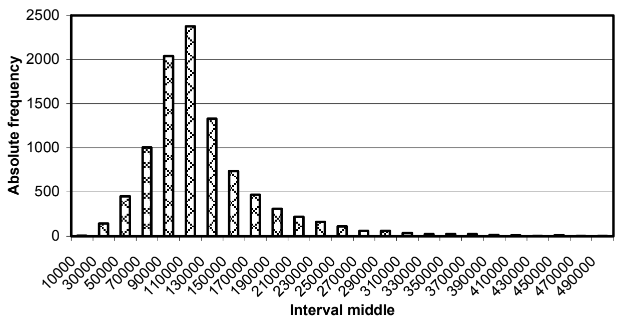

Fig. 19 Interval frequency distribution of net annual household income per capita in 2007

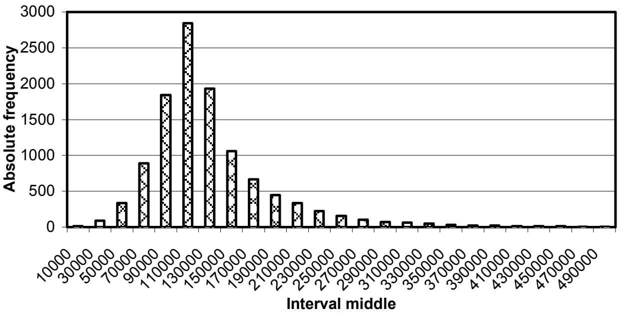

Fig. 20 Interval frequency distribution of net annual household income per capita in 2008

parameters estimated using the moment method of parameter estimation. Similarly situation is for Figs. 22 and 24 in the case of quantile method of parameter estimation, where equality of sample and theoretical median is one of three quantile equations and so the course of the development of sample median coincides with the course of development of theoretical median of three-parametric lognormal distribution with the parameters estimated using the quantile method of parameter estimation. Figs. 21 to 24 show a high accuracy of all four methods used to estimate parameters on these data.

Using moment parameter estimation has some unpleasant specifics in the case of the distribution of income and wage. The moments of higher order including the moment characteristic of skewness are very sensitive to inaccuracies on

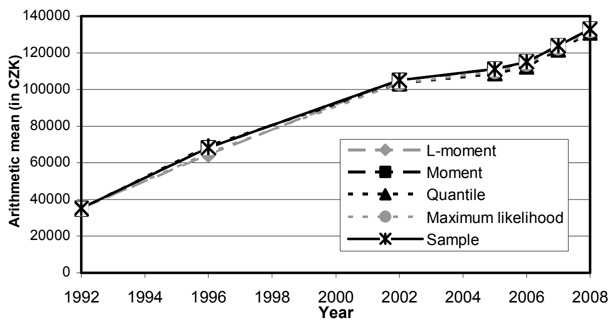

Fig. 21 Development of sample average net annual income per capita and the theoretical expected value (in CZK)

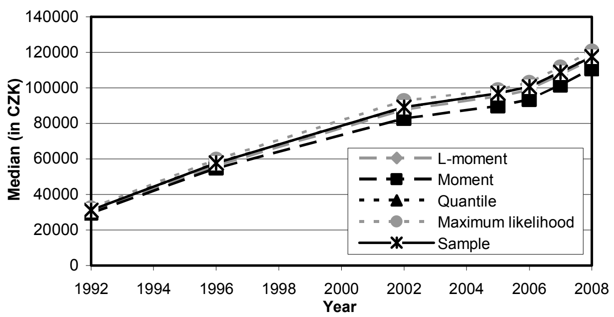

Fig. 22 Development of sample median of net annual income per capita and the theoretical median (in CZK)

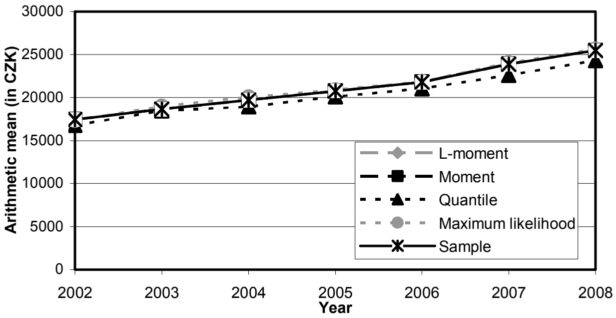

Fig. 23 Development of sample average gross monthly wage and the theoretical expected value (in CZK)

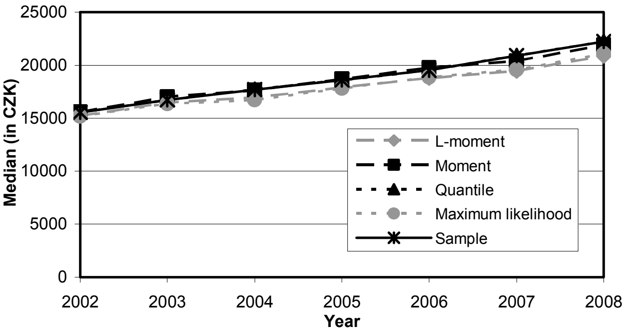

Fig. 24 Development of sample median of gross monthly wage and the theoretical median (in (CZK)

Tab. 15 Sum of absolute deviations of the observed and theoretical frequencies for all intervals − net annual household income per capita

Year	Method			
	L-moment	Moment	Quantile	Maximum likelihood
1992	2,661.636	5,256.970	3,880.846	2,933.275
1996	5,996.435	15,673.846	9,677.446	7,181.322
2002	2,181.635	3,888.523	3,206.585	2,236.348
2005	1,158.556	2,261.200	1,331.944	1,237.170
2006	2,197.016	3,375.662	2,984.503	2,217.975
2007	2,359.258	3,654.637	2,995.680	2,585.448
2008	2,251.531	4,282.314	3,277.620	2,889.890

Tab. 16 Sum of absolute deviations of the observed and theoretical frequencies for all intervals − gross monthly wage

Year	Method			
	L-moment	Moment	Quantile	Maximum likelihood
1992	134,846.633	314,497.134	292,479.483	289,279.267
1996	135,772.928	356,423.157	303,335.493	283,469.483
2002	252,042.801	357,087.483	335,019.202	295,900.939
2005	260,527.847	426,062.444	345,954.758	306,785.789
2006	277,661.535	448,632.374	372,420.681	357,828.202
2007	229,525.420	432,745.341	338,552.122	250,114.480
2008	255,510.389	441,371.539	372,924.579	289,621.287

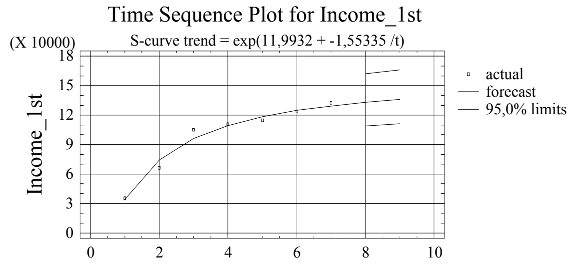

Fig. 25 The trend function in the development of the first sample L-moment of net annual household income per capita (forecasts: 133,122.0; 136,026.0)

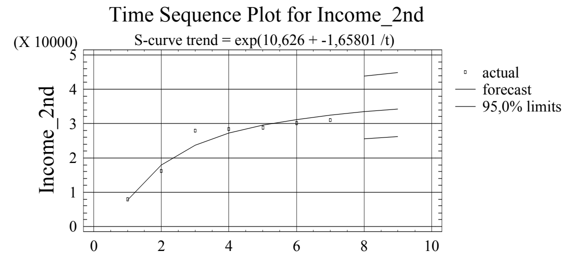

Fig. 26 The trend function in the development of the second sample L-moment of net annual household income per capita (forecasts: 33,482.6; 34,262.6)

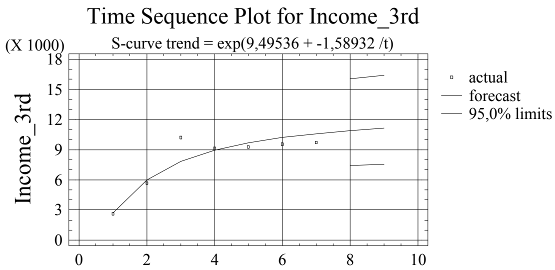

Fig. 27 The trend function in the development of the third sample L-moment of net annual household income per capita (forecasts: 10,901.9; 11,145.3)

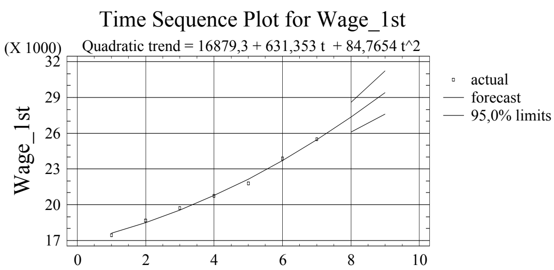

Fig. 28 The trend function in the development of the first sample L-moment of gross monthly wage (forecasts: 27,355.1; 29,427.5)

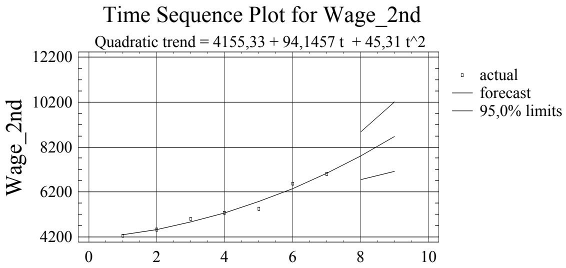

Fig. 29 The trend function in the development of the second sample L-moment of gross monthly wage (forecasts: 7,808.34; 8,672.75)

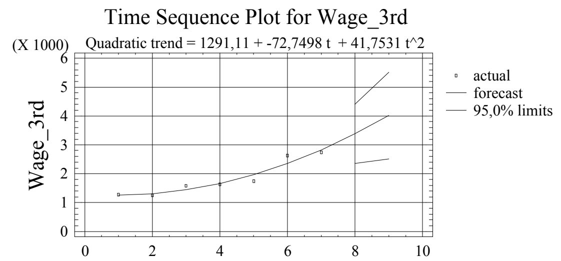

Fig. 30 The trend function in the development of the third sample L-moment of gross monthly wage (forecasts: 3,381.31; 4,018.36)

Tab. 17 Extrapolations of sample L-moments

Set	Year	Sample L-moments		
		l_1	l_2	l_3
Income	2009	133,122	33,483	10,902
	2010	136,026	34,263	11,145
Wage	2009	27,355	7,808	3,381
	2010	29,428	8,673	4,018

Tab. 18 Extrapolations of parameter estimations of three-parametric lognormal distribution obtained using the L-moment method

Set	Year	Parameter estimation		
		μ	σ^2	θ
Income	2009	11.176	0.467	42,913.996
	2010	11.201	0.466	43,631.177
Wage	2009	9.247	0.864	11,384.492
	2010	9.217	1.004	12,794.380

the both ends of the distribution. Registration errors, from which these inaccuracies arise, are just typical for the survey of income and wage. Moment method of parameter estimation does not guarantee maximum efficiency of the estimation, nevertheless it may not be a hindrance when working with the income and wage distributions due to a usually high sample size.

Tabs. 15 and 16 provide more accurate information about the used methods of parameter estimation. These tables contain the sum of absolute deviations of the observed and

Tab. 19 Extrapolations of the interval distribution of relative frequencies (in %) of net annual household income per capita for 2009 and 2010

Interval	Year	
	2009	2010
0 − 20,000	0.00	0.00
20,001 − 40,000	0.00	0.00
40,001 − 60,000	1.82	1.42
60,001 − 80,000	15.07	13.88
80,001 − 100,000	20.27	19.82
100,001 − 120,000	17.29	17.37
120,001 − 140,000	12.89	13.16
140,001 − 160,000	9.19	9.49
160,001 − 180,000	6.47	6.75
180,001 − 200,000	4.56	4.79
200,001 − 220,000	3.24	3.42
220,001 − 240,000	2.32	2.47
240,001 − 260,000	1.68	1.80
260,001 − 280,000	1.23	1.32
280,001 − 300,000	0.91	0.98
300,001 − 320,000	0.68	0.74
320,001 − 340,000	0.51	0.56
340,001 − 360,000	0.39	0.42
360,001 − 380,000	0.30	0.33
380,001 − 400,000	0.25	0.26
400,001 − ∞	0.93	1.02
Total	100.00	100.00

Tab. 20 Extrapolations of the interval distribution of relative frequencies (in %) of gross monthly wage for 2009 and 2010

Interval	Year	
	2009	2010
0 − 5,000	0.00	0.00
5,001 − 10,000	0.00	0.00
10,001 − 15,000	12.84	6.49
15,001 − 20,000	29.25	30.44
20,001 − 25,000	19.43	20.69
25,001 − 30,000	12.03	12.74
30,001 − 35,000	7.66	8.14
35,001 − 40,000	5.06	5.44
40,001 − 45,000	3.45	3.76
45,001 − 50,000	2.43	2.69
50,001 − 55,000	1.75	1.97
55,001 − 60,000	1.29	1.48
60,001 − 65,000	0.97	1.13
65,001 − 70,000	0.74	0.88
70,001 − 75,000	0.57	0.69
75,001 − 80,000	0.45	0.55
80,001 − 85,000	0.35	0.44
85,001 − 90,000	0.28	0.36
90,001 − 95,000	0.23	0.30
95,001 − 100,000	0.17	0.25
100,001 − ∞	1.05	1.56
Total	100.000	100.000

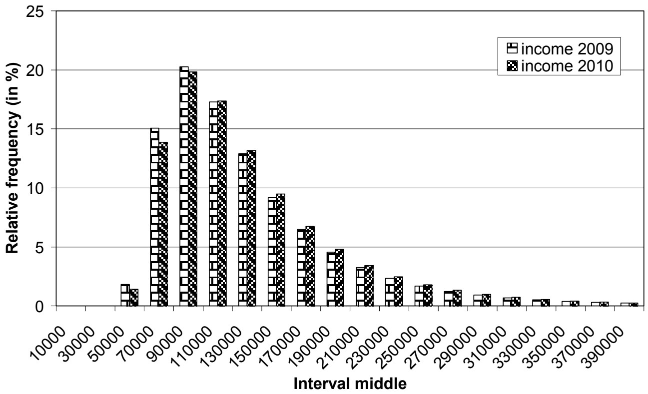

Fig. 31 Extrapolations of the interval distribution of relative frequencies (in %) of net annual household income per capita for 2009 and 2010

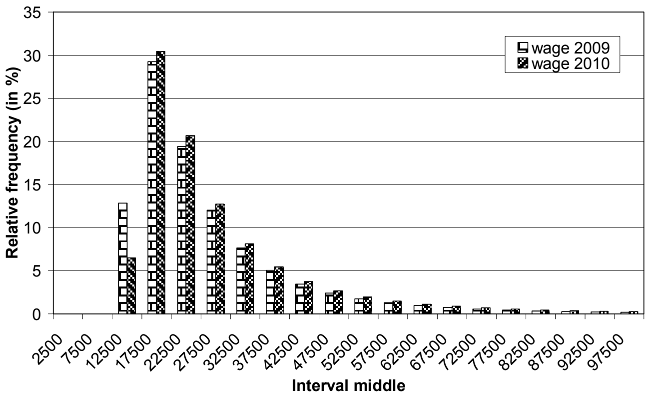

Fig. 32 Extrapolations of the interval distribution of relative frequencies (in %) of gross monthly wage for 2009 and 2010

theoretical frequencies for all intervals and therefore they serve as an objective criterion for evaluating the accuracy of used methods of parameter estimation. It should be noted here that in the case of income distribution on the one hand, and in the case of wage distribution on the other hand, we used the same number of intervals, whose width is expanded in time due to the rising level of the distributions. We can see from Tabs. 15 and 16 that the method of L-moments provides the most accurate results, which are even more accurate than results obtained using the maximum likelihood method. Already mentioned maximum likelihood method ended in the terms of accuracy of the estimations as the second best. Quantile method of parameter estimation follows as the third best (second worst). As expected, moment method of parameter estimation provides the least accurate results.

Values of test criterion (56) were also calculated for each income distribution or for each wage distribution. As it was mentioned, the tested hypothesis on the expected shape of the distribution is rejected even at 1% significance level in the case of each income or wage distribution. This situation is caused by large sample sizes, with whom we work in the case of income and wage distribution. Values of test criterion χ^2 are not therefore listed.

Interestingly in addition, Figs. 25 to 30 represent the trend functions for the development of sample L-moments in corresponding monitored periods, including their forecasts for the years 2009 and 2010 in parentheses. Tab. 17 represents the extrapolations of sample L-moments created on the basis of the trend functions from Figs. 25 − 30. Table 18 shows the extrapolations of parameter estimations of three-parametric lognormal curves obtained using the L-moment method based on the values from Table 17. Tabs. 19 and 20 and Figs. 31 and 32 show the extrapolations of income and wage distribution for the years 2009 and 2010 based on the parameter values from Table 18.

IV. Conclusion

Importance of lognormal curve as a model for the empirical distribution is indisputable, and it has found application in many areas from the sociology to astronomy. Characteristic features of the process described by this model are: successive appearances of interdependent factors; tendency to develop in a geometric sequence; overgrowth of random variability to the systematic variability − differentiation. Incomes and wages are among the many economic phenomena that lognormal model allows to interpret, which is confirmed by numerous practical experiences.

Three-parametric lognormal distribution (Johnson's curve of the type S_L) was used in the modelling of incomes and wages in this study. Various methods of parameter estimation were used in estimating the parameters of this distribution − moment method, quantile method, maximum likelihood method and finally the method of L-moments. In the case of small sample size, L-moment method usually provides markedly more accurate results than other methods of parameter estimation, including the maximum likelihood method, see for example [5]. However, it appears that even in the case of large samples tahat the L-moment method gives more accurate results than the other methods of parameter estimation (and again, including the maximum likelihood method). When calculating the sum of the absolute deviations of the observed and theoretical frequencies and also in calculating the value of test criterion χ^2, it showed that inaccuracies arise especially at the both ends of the distribution in the case of method of L-moment. If we abstracted from inaccuracies on both ends of the distribution, the results based on L-moment method would be much more accurate compared to other methods of parameter estimation in the case of large samples, too.

In addressing the question which method of parameter estimation of three-parametric lognormal distribution is most suitable, it was the high dependency of the value of χ^2 criterion due to the sample size. As it is usual with such a large sample size, all tests led to the rejection of the null hypothesis on the expected distribution. From the results it is clear that all four used methods of parameter estimation yielded relatively accurate results at such large samples, which were used in this research and which are typical of the income and wage distribution. Despite some differences in the accuracy of parameter estimation methods used were discovered. As it is evident from the outputs, the L-moment method gives again the most accurate results of parameter estimation. The method of maximum likelihood follows as the second most accurate. Quantile method of parameter estimation follows and method of moments has brought at least accurate results of parameter estimation, as expected. Notwithstanding the foregoing, the differences in accuracy between parameter estimation methods used are not relatively too high in the case of such large sample sizes, see outputs above.

REFERENCES

[1] J. Bartošová, "Logarithmic-Normal Model of Income Distribution in the Czech Republic", *Austrian Journal of Statistics*, vol. 35, no. 23, 2006, pp. 215 − 222.
[2] D. Bílková, "Application of Lognormal Curves in Modelling of Wage Distributions", *Journal of Applied Mathematics*, vol. 1, no. 2, 2008, pp. 341 − 352.
[3] A.C. Cohen and J.B. Whitten, "Estimation in the Three-parameter Lognormal Distribution", *Journal of American Statistical Association*, vol. 75, 1980, pp. 399 − 404.
[4] N.B. Guttman, "The Use of L-moments in the Determination of Regional Precipitation Climates", *Journal of Climate*, vol. 6, 1993, pp. 2309 − 2325.
[5] J.R.M. Hosking, "L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics", *Journal of the Royal Statistical Society (Series B)*, vol. 52, no. 1, 1990, pp. 105 – 124.
[6] J. Kyselý and J. Picek, "Regional Growth Curves and Improved design Value Estimates of Extreme Precipitation Events in the Czech Republic", *Climate Research*, vol. 33, 2007, pp. 243 – 255.
[7] I. Malá, "Conditional Distributions of Incomes and their Characteristics", in *ISI 2011*, Dublin, Ireland, 2011, pp. 1− 6.
[8] I. Malá, "Distribution of Incomes Per Capita of the Czech Households from 2005 to 2008", in *Aplimat 2011 [CD-ROM]*, Bratislava, Slovakia, 2011, pp. 1583 − 1588.
[9] I. Malá, "Generalized Linear Model and Finite Mixture Distributions", in *AMSE 2010 [CD-ROM]*, Demänovská Dolina, Slovakia, 2010, pp. 225 − 234.
[10] I. Malá, "L-momenty a jejich použití pro rozdělení příjmů domácností", in *MSED 2010 [CD-ROM]*, Prague, Czech Republic, 2010, pp. 1 − 9.
[11] J.C. Smithers and R.E. Schulze, "A Methodology for the Estimation of Short Duration Design Storms in South Africa Using a Regional Approach Based on L-moments", *J. Hydrol.*, vol. 241, 2001, pp. 42 – 52.
[12] T.J. Ulrych, D.R. Velis, A.D. Woodbury and M.D. Sacchi, "L-moments and C-moments", *Stoch. Environ. Res. Risk Asses*, vol. 14, 2000, pp. 50 − 68.
[13] http://www.czso.cz/
[14] http://puzzle.heureka.cz
[15] http://search.seznam.cz/