THE USE OF THE LOGNORMAL DISTRIBUTION IN ANALYZING INCOMES

Jakub Nedvěd

Abstract

Object of this paper is to examine the possibility of using the lognormal distribution as a model of incomes distribution. The paper is focused on three-parametric lognormal distribution, because it is most common in analyzing incomes. Using data from the Informational system of average income this paper compares quality of models. The Informational system of average income gathers data about person's hourly wages from two sectors – business and non-business. Data set is divided by gender, education, type of job, age and so on. The Informational system of average income publishes on its web sites quartiles and other characteristics for each mentioned groups. Models are made separately for business and non-business sector. For both sectors are constructed 5 models which uses different method of estimating parameters of the lognormal distribution and this thesis describes quality and differences between these models. The thesis also shows some problems appearing when we use lognormal distribution in analyzing incomes. There are described options of usage the lognormal distribution in analyzing incomes. The thesis demonstrates the fact that the curve of lognormal distribution density function is applicable model which is reliable especially in the wide central part of incomes distribution.

Key words: lognormal distribution, Informational system of average income, distribution of incomes

JEL Code: C13, C16

Introduction

Research of the incomes is important in economics and politics. In economics we can compare socioeconomics groups like men and women or groups divided by education. For politics it is important that the incomes aren't uniformly distributed. We can quantify this non-uniformity by detailed analysis of income distribution. The distribution of income is very complicated, so in practice we use probability models. The model can be created as the curve of probability distribution or the curve of Pearson's or Johnson's system.

This paper is focused on three-parametric lognormal distribution that is frequently used for the modelling of income distribution. We use data about wage per hour from The Informational system of average income which is held by company TREXIMA. TREXIMA has long-term experience with collecting these data and we can rely on their high quality. Data from Informational system of average income are divided by business and non-business sector.

At the end of this paper is suggested an analysis of income distribution of business sector between the years 2000 and 2010 is presented.

1 Lognormal distribution

The lognormal distribution is the most frequently used distribution for the modelling of incomes and wages. For the purpose of analyzing incomes and another statistics attributes which are correlated with income is mainly used three-parametric lognormal distribution. Two-parametric and four-parametric lognormal distributions are in practice also used but not as frequent as three-parametric distribution.

The option of usage the analysis of income distribution is to investigate the effect of different taxation on the structure of income. The analyzing of incomes is important and useful because the level of income is connected with quality of life and it gives objective view and enables quantification. It is used to compare international or intrastate regions and we can predict future progress.

There are two required factors when we make a model of distribution of income. The first is to get the most exact similarity of model and reality. This factor means increasing the number of parameters of the model. The second factor is to get easy economic interpretation. It means reducing the number of parameters. Therefore we use the most frequently three-parametric lognormal distribution.

Lognormal distribution is one of many distributions used in analyzing incomes. "Two-parametric lognormal distribution fits well over a large part of middle income range, but gives a poor fit at the tails. However, in the middle income range it exaggerates skewness. Pareto distribution provides an excellent fit to the upper tail of the income distribution, but the fit over the whole range of income is poor. Gamma distribution provides a better fit than lognormal at the tails. In the middle range, both lognormal and gamma exaggerate skewness, but the tendency is more marked in case of lognormal. Dagum distribution performs better than lognormal and gamma distributions" (Chakravarty and Majumder, 1990). Another probability model of income distribution is generalized lambda distribution which is defined by its quantile function.

1.1 Characteristics of lognormal distribution

If a random variable $Z = \ln(X - \theta)$ has normal distribution with expected value μ and variance σ^2, variable *X* has three-parametric lognormal distribution with parameters θ, μ, σ^2 and its probability density function is given by formula

$$f(x) = \left[(x-\theta)\sqrt{2\pi} * \sigma\right]^{-1} * \exp\left[-\frac{\frac{1}{2}\{\ln(x-\theta)-\mu\}^2}{\sigma^2}\right], (x \geq \theta). \quad (1)$$

Parameter θ means a theoretical minimum of variable *X*. If parameter θ equals zero than variable *X* has two-parametric lognormal distribution. Distribution function of variable *X* is given by formula

$$F(x) = \Phi\left(\frac{\ln(x-\theta)-\mu}{\sigma}\right) \quad , \quad (2)$$

where Φ is a distribution function of standard normal distribution.

The income distributions have positive skewness and lognormal distribution meets this property. As the income distribution is usually highly skewed, the mean loses its ability to objective describe the distribution (mean is strongly affected by rare but very high values) and then the median seems to be better characteristic of the level of income.

1.2 The estimation of parameters

The most frequently used methods of estimating the parameters are moment method, quantiles method and method of maximum likelihood. We need three equations to find out the estimated values of three unknown parameters of three-parametric lognormal distribution and each method forms these equations from random sample differently. (See Johnson at all., 1994 for more.)

The maximum likelihood estimates have optimal asymptotic properties. For the lognormal distribution logarithm of likelihood function is maximized with respect to estimates $\hat{\mu}, \hat{\sigma}^2$ of parameters μ, σ^2 are calculated by equations (Johnson at all., 1994)

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i - \theta) \quad , \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left[\ln\left(X_j - \theta\right) - \hat{\mu}\right]^2 \quad . \tag{4}$$

The difficult problem is to estimate the value of parameter θ. For the large sample the minimum can be taken. Cohen's method combines method of maximum likelihood with the quantile method and parameter θ is estimated as $100*(n+1)^{-1}\%$ sample quantile.

The method of moments is relatively simple but it could be quite inaccurate. This method equates sample moments with the theoretic moments. The point estimates $\theta^+, \mu^+, \sigma^{2^+}$ of three unknown parameters are given by formulas

$$\theta^+ = \bar{x} - \mathrm{e}^{\mu^+ + \frac{\sigma^{2^+}}{2}} \quad , \tag{5}$$

$$\mu^+ = \frac{1}{2}\ln\frac{m_2}{e^{\sigma^{2^+}}\left(e^{\sigma^{2^+}} - 1\right)} \quad , \tag{6}$$

$$\sigma^{2^+} = \ln\left[\sqrt[3]{1+b_2+\sqrt{(1+b_2)^2-1}} + \sqrt[3]{1+b_2-\sqrt{(1+b_2)^2-1}} - 1\right] \tag{7}$$

,

where $b_2 = \frac{1}{2}\left(e^{\sigma^{2^+}} - 1\right)\left(e^{\sigma^{2^+}} + 2\right)^2$ and $m_2 = \bar{y} - \bar{x}^2, y = x^2$.

The method of quantiles is also relatively simply. This method uses three quantiles (to estimate three parameters) – the most frequently median, $100*\alpha\%$ and $100*(1-\alpha)\%$ quantiles. These values are published by The Informational system of average income on its web-site (http://www.ispv.cz/cz/Vysledky-setreni/Archiv.aspx). Equations (8), (9) and (10) give the point estimations where x^S are sample quantiles.

$$\sigma^{2^*} = \left[\frac{\ln\dfrac{x_\alpha^S - x_{0.5}^S}{x_{0.5}^S - x_{1-\alpha}^S}}{u_\alpha}\right]^2 \quad , \tag{8}$$

$$\mu^* = \ln\frac{x_\alpha^S - x_{1-\alpha}^S}{e^{\sigma^* u_\alpha} - e^{-\sigma^* u_\alpha}} \quad , \tag{9}$$

$$\theta^{*} = x_{0.5}^{S} - e^{\mu^{*}} \quad . \tag{10}$$

2 The Informational system of average income

The Informational system of average income (ISAI) is made primarily for the Ministry of labour and social affairs of the Czech Republic. It provides to inform about the income structure in the Czech Republic. The survey is for example used to assess the wages of civil servants.

ISAI was created on the beginning of ninetieth as the complement of statistics of wages. The basic principles were established by The Ministry of labour and social affairs and the Czech Statistical Office (http://czso.cz/eng/redakce.nsf/i/home). ISAI is the only source of information about the income standard based on the income of individuals in the Czech Republic. Processing of data is performed by company TREXIMA. ISAI monitors the wage per hour, gross monthly wage, worked and not worked time of individuals and organizations too. Data are classified by category of works KZAM-R[1]. Data set can be divided by gender, age, education etc.

The web-portal of ISAI was set off in the 2009. There are published results of surveys and selected statistical characteristics from the year 2000 to present on this portal. Descriptive statistical characteristics published by ISAI are median, mean, the first and the third quartiles, the first and the ninth deciles of hourly wage.

Results are divided into two sections – business and non-business sector. Non-business sector is researched every half-year and it is comprehensive survey. It covers about 14 550 economic subjects with approximately 660 thousand employees. Business sector is investigated selectively every quarter. The sample covers about 3500 economic subjects with more than 1.3 million employees.

Business and non-business sector's data sets have different characteristic. The statistical characteristics of data from non-business sector are affected only by non-sampling errors because it is comprehensive survey. However, the statistical characteristics of data set from business sector are affected by both sampling and non-sampling errors.

This paper uses data from ISAI to compare quality of different lognormal models. Data were randomly selected from ISAI of the second quarter 2009. First data set contains

[1] Since 2011 is used classification CZ-ISCO.

10 000 entries of income per hour in CZK from the business sector and the second data set contains 10 000 entries of income per hour from the non-business sector.

3 Modelling of income distribution

We use interval of the length 5 CZK and we compare theoretical and empirical frequency of intervals. The empirical frequency is given by number of entries in the corresponding interval in the data set from the business or non-business sector. The theoretical frequency of interval *i* is given by formula

$$n * \pi_i = n * \left[\Phi\left(\frac{\ln\left(x_i + \frac{v}{2} - \theta \right) - \mu}{\sigma} \right) - \Phi\left(\frac{\ln\left(x_i - \frac{v}{2} - \theta \right) - \mu}{\sigma} \right) \right], \quad (11)$$

where x_i is the middle value of *i*-th interval, *v* is the interval length (*v* = 5 CZK) and *n* is the sample size (*n* = 10 000). But we must extend the first and the last interval to the rest of the area under the model curve[2]. Quality of various models is compared by *S* statistic which is evaluated as a sum of absolute deviations of theoretical and empirical frequencies[3].

$$S = \sum_{i=1}^{k} \left| n_i - n * \pi_i \right| . \quad (12)$$

We made 5 models for business and non-business sector too. The first model (2_ml) is two-parametric lognormal distribution which uses method of maximum likelihood to estimate its parameters. Other models are three-parametric lognormal distributions. The second model (3_m) uses moment method to estimate parameters, the third (3_qI) uses quantiles method with $\alpha = 0.1$ in formulas (7), (8), (9) and the fourth model (3_qII) uses also quantiles method but $\alpha = 0.25.$ The fifth model (3_ml) uses method of maximum likelihood and parameter θ is evaluated by numeric minimizing the *S* statistic using MS Excel.

3.1 Non-business sector

Table 1 presents the estimates of the parameters, the basic characteristics and the *S* statistic for five models mentioned above for non-business sector together with descriptive characteristics of data set. Three-parametric lognormal model using method of maximum likelihood (3_ml) to estimate parameters has the lowest value of *S* statistic, but all models

[2] Then the sum of the theoretical frequency equals 1.

[3] We don’t use the chi-square statistic because we operate with large sample. In this case we get almost always the rejection of χ^2 at the conventional level of significance.

have quite similar quality. Interesting is the fact that all models undervalue the basic characteristics – modus, expected value (mean) etc..

Tab. 1: Estimates of the parameters and the basic characteristics and *S* statistic for the models of income distribution in non-business sector for the second quarter of 2009 and sample characteristics of data set

model		2_ml	3_m	3_qI	3_qII	3_ml	data set
method		maximum likelihood	moment	quantiles	quantiles	maximum likelihood	/
comment		/	9980 values[4]	$\alpha = 0.1$	$\alpha = 0.25$	θ by MS Excel	/
parameters	θ	/	4.2690	-76.3609	-181.0605	-25.4797	/
	μ	4.8500	4.8111	5.3286	5.7393	5.0422	
	σ^2	0.1484	0.1520	0.0582	0.0218	0.1020	
$E(X)$		137.5799	136.8305	135.8663	133.1808	137.4271	137.7376
$D(X)$		3,028.2919	2,884.1260	2,698.9003	2,172.5249	2,849.9914	3,342.7408
$\sigma(X)$		55.0299	53.7041	51.9509	46.6104	53.3853	57.8164
modus	$\hat{x}$	110.1221	109.8089	118.1260	123.0885	114.3142	125
S		1,344	1,359	1,320	1,277	1,255	/

Source: own computations

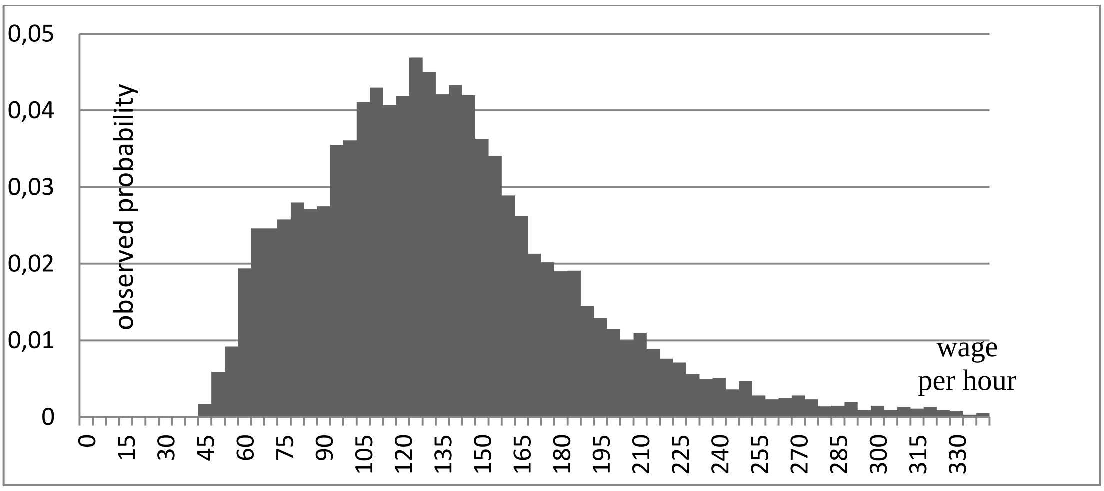
Tab. 2: Estimates of the parameters and the basic characteristics and *S* statistic for the models of income distribution in business sector for the second quarter of 2009 and sample characteristics of data set

model		2_ml	3_m	3_qI	3_qII	3_ml	data set
method		maximum likelihood	moment	quantiles	quantiles	maximum likelihood	/
comment		/	/	$\alpha = 0.1$	$\alpha = 0.25$	θ by MS Excel	/
parameters	θ	/	27.1020	27.7359	23.9340	21.6933	/
	μ	4.8726	4.5942	4.5935	4.6313	4.6653	
	σ^2	0.2073	0.3708	0.3314	0.2595	0.3061	
$E(X)$		144.9342	146.1583	144.3955	140.8038	145.4583	146.1583
$D(X)$		4,839.8554	6,362.6788	5,348.0305	4,047.6491	5,486.4353	6,362.6788
$\sigma(X)$		69.5691	79.7664	73.1302	63.6211	74.0705	79.7664
modus	$\hat{x}$	106.1938	95.3675	98.6955	103.1151	101.7142	115
S		1,475	1,605	1,280	1,086	1,303	/

Source: own computations

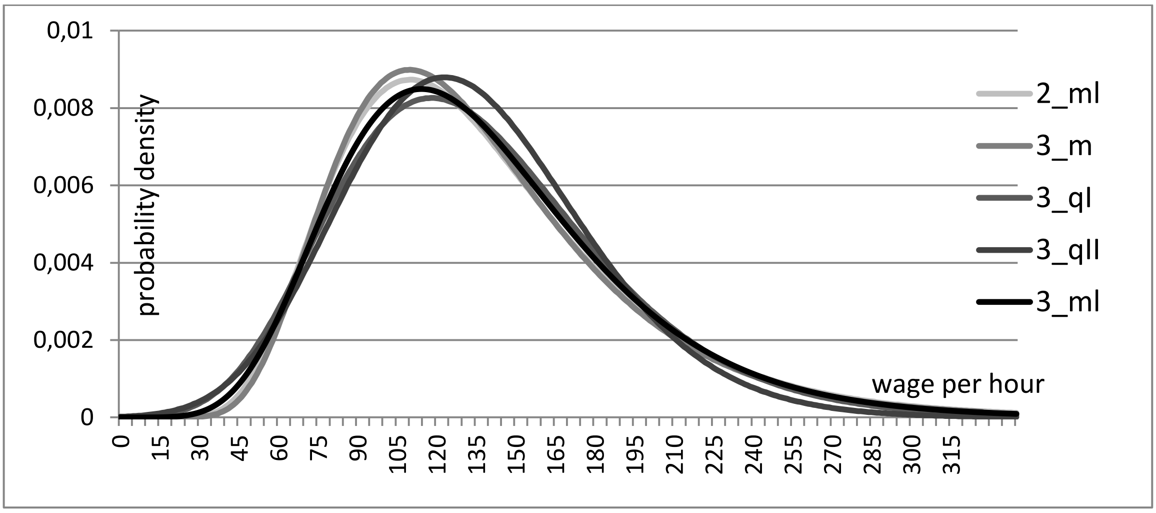
[4] The highest 20 values weren't used to estimate the parameters of model 3_m. It is said that approximately 1-2 percent of the highest values make the model worse. This paper doesn't find out the most quality model but compares the quality of different lognormal models so we don't need to leave out these values. But by estimating the parameters with moment method was the *S* statistic so bad that these 20 values weren't used to get the value of *S* statistic of model 3_m closer to the other *S* statistics.

Fig. 1: Observed frequencies of wage per hour (CZK) in non-business sector in the 2nd quarter of 2009



Source: own calculations

Fig. 2: Models of income distribution of non-business sector for the 2nd quarter of 2009



Source: own calculations

The worst model is model 3_m which uses moment method of estimation of the parameters although its quality was improved by not using the highest 20 values. Model 3_qII provides relatively good fit. Model 3_qII uses quartiles to estimate the parameters and these values are published by ISAI[5] so we can make a model for any year.

The observed frequencies of income per hour are shown in Figure 1 and fitted probability densities are shown in Figure 2. We can see in Figure 1 local extreme in the left

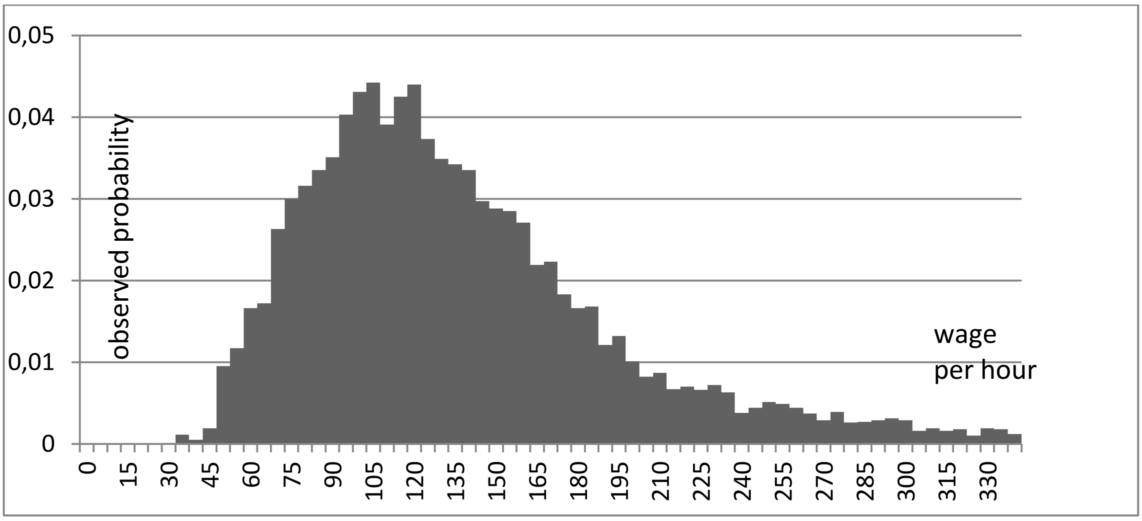
[5] The quartiles and the 1st and the 9th deciles for the years 2000 – 2010 are published on http://www.ispv.cz/cz/Vysledky-setreni/Archiv.aspx

part of the histogram. This paper models the income distribution by one curve of the lognormal distribution, more detailed analysis could use two or more curves in a mixture. In this approach a data set is divided into two or more groups and every single group is modeled by one probability density. The final model is generated as a weighted average of all curves. Lower values of income per hour (the left part of the histogram) in non-business sector should be modeled separately and the final model would be probably better. Nevertheless all models seem to be suitable for the modelling of the income distribution of non-business sector.

3.2 Business-sector

Table 2 presents the estimates of the parameters, the basic (estimated theoretical) characteristics and values of *S* statistic for the five models for business sector and sample characteristics of data set. The best fit gives the model 3_qII which uses quantile method of estimation of the parameters and chooses sample quartiles for the estimation. The second model which uses the quantile method of estimation of the parameters (3_qI) is also better than the model 3_ml. This is quite interesting because we expected the better model with the maximum likelihood estimates of the parameters. The theoretical estimated characteristics of mean, variance and modus are again undervalued compared to sample.

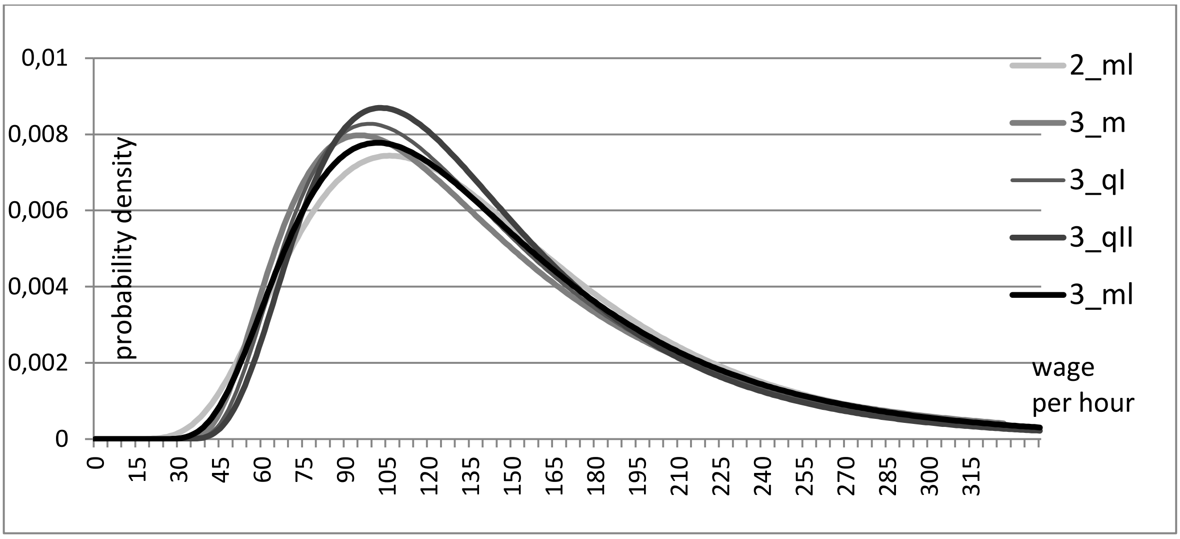
Fig. 3: Observed probabilities of wage per hour in business sector in the 2nd quarter of 2009



Source: own calculations

The observed frequencies of income per hour are shown in Figure 3 and fitted probability densities are shown in Figure 4. The estimated densities are very similar and all of them seem to be suitable for the modelling of the income distribution of business sector especially in central part of distribution.

Fig. 4: Models of income distribution of business sector for the 2nd quarter of 2009



Source: own calculations

4 The use of the lognormal distribution in analyzing incomes

The three-parametric lognormal distribution seems to fit well the data of business and non-business sector. Characteristics of this distribution can be used to compare different socioeconomic groups such as men and women or categories of education. Moreover, income model can be used to evaluate the standard of living. The model can show the influence of various tax rates on income distribution. The prognosis of future development of income distribution can be transferred to estimation of future development of the parameters of model based on three-parametric lognormal distribution.

The opportunity to analyse development of wage distribution is offered by The Informational system of average income. ISAI has published on its web sites[6] selected quantiles of wage per hour in business and non-business sector since 2000. We can use quantile method of estimation of the parameters to build a model and to compare it with models made for another year.

[6] http://www.ispv.cz/cz/Vysledky-setreni/Archiv.aspx

4.1 Analysis of income distribution of business sector between 2000 and 2010

The three-parametric lognormal model which uses quantile method of estimating the parameters gives acceptable model of income distribution of business sector. Values in Table 3 are taken from ISAI web sites and they are used to fit models of 2000 and 2010.

Table 4 presents the estimates of the parameters and the basic characteristics of the models of income distributions for 2000 and 2010. Comparing Table 4 with Table 3 we can see that both models undervalue the mean. This fact was mentioned in previous chapter so the modus in the real distribution of wages will be higher, too. Expected value of hourly wage has increased but it was caused mainly by increasing of price level. The interesting are values of other characteristics of distribution. They can be interpreted as the measure of non-uniformity. Variance has nearly quadrupled and it can be interpreted as relatively large increase of non-uniformity of income distribution between 2000 and 2010. It is also shown in Figure 5 where the model for 2000 seems to be more concentrated than the model for the year 2010 which shows increasing the number of people with higher wage per hour.

Tab. 3: Quantiles of income distribution (in CZK) of business sector

	quantiles			mean
period	1st decile	median	9th decile	
4th quarter 2000	45.10	76.94	134.08	88.29
4th quarter 2010	70.11	124.46	240.72	152.41

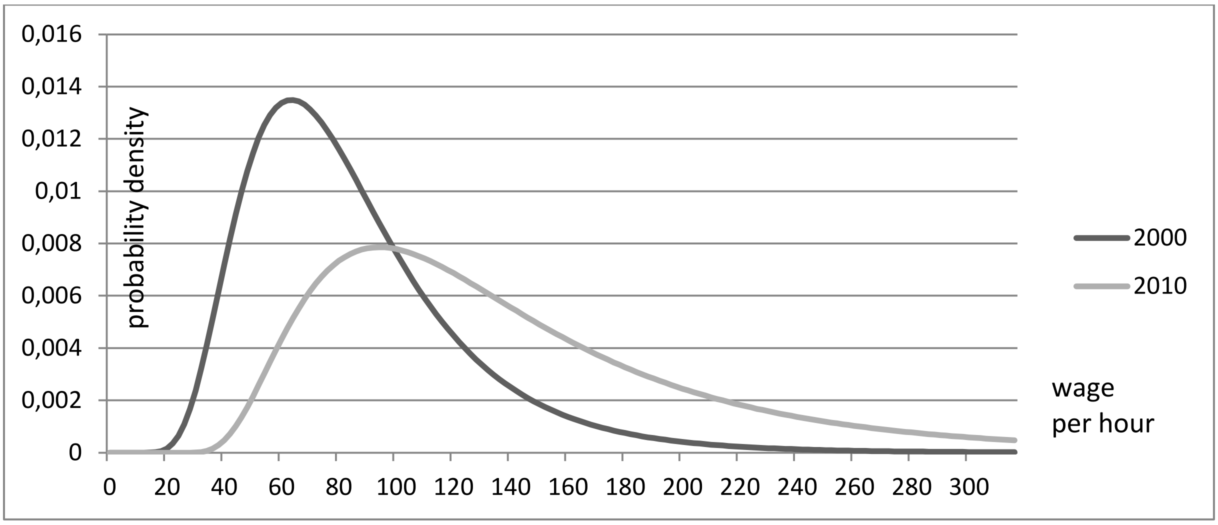
Source: TREXIMA

Tab. 4: Estimates of the parameters and the estimated characteristics of the distribution for the models of income distribution in business sector for 2000 and 2010

		2000	2010
parameters	σ^2	0.2082	0.3520
	μ	4.2754	4.6256
	θ	5.0294	22.3968
characteristics	$\hat{x}$	63.42	94.17
	$E(X)$	84.83	144.10
	$D(X)$	1,474.09	6,250.40
	$\sigma(X)$	38.39	79.06
	$\mu_3(U)$	1.55	2.22
	$\mu_4(U)$	7.58	12.90

Source: own computations

Fig. 5: Models of income distribution of business sector for 2000 and 2010



Source: own calculations

Conclusion

The probability density of three-parametric lognormal distribution can be used as a model of income distribution for business and non-business sector in the Czech Republic with relatively good confidence in central part of the distribution. The quality of the model depends on the used method of estimating of the parameters. For non-business sector the best fit is model that uses the method of maximum likelihood. For business sector seems to be the best model that uses quantiles method of estimating the parameters. Quality of models was described on data about hourly wage from The Informational system of average income from 2009.

Using the three-parametric lognormal distribution, two models of income distribution are made for business sector for the years 2000 and 2010. The models use the quantile method of estimating the parameters and quantiles are taken from the Informational system of average income. The analysis shows the increasing of non-uniformity of income distribution in business sector between the years 2000 and 2010.

References

1. Chakravarty, S. R., Majumder, A. "Distribution of Personal Income: Development of a new Model and its Application to U.S. Income Data" *Journal of Applied Econometrics* 5/2 Apr. – Jun., 1990: 189-196.
2. Johnson, N. L., and N. Balakrishnan, and S. Kotz. *Continuous univariate distributions*. Vol. 1. New York: John Wiley & Sons, 1994.

Contact

Jakub Nedvěd

The University of Economics in Prague, Faculty of Informatics and Statistics

nám. W. Churchilla 4, Praha, Czech Republic

xnedj06@isis.vse.cz