DOI: 10.7763/IPEDR. 2013. V59. 15

Mixture Normal Denssity Functions as a Model Wage Distribution

Lubos Marek[1+] and Michal Vrabec[1]

[1] Informatics and Statistics University of Economics, Prague

Abstract. In our article we try to contribute to the discussion of the possibility to predict the trend of the wage distribution in the Czech Republic. Classical models use the probability distribution such as lognormal, Pareto, etc., but their results are not very good. We suggest using a mixture of normal probability distribution (normal mixture) in our model. We focus mainly on the possibility of constructing a mixture of normal distributions based on parameter estimation. We estimate these parameters on the basis of their evolution in time. We work with data collected in the last 15 years. The data is divided into groups with respect to gender, age, and regions.

Keywords: Income Distribution, Probability Distribution, Mixture Normal Probability Distribution.

1. Introduction

We want to contribute to the discussion on suitability of the arithmetic mean as a characteristic for the wage level in the Czech Republic. There is recurring expression of surprise with the fact that „… the income of more than fifty percent of the population is lower than the average wage“. If the intended effect is to have "more" wage recipients above the officially announced level, a simple solution would be to use different characteristics of this level. For example, the median (50% quantile) is defined by the condition that exactly 50% wage recipients are below this value, while the remaining 50% are above it. Choosing a suitable quantile, we can always get the required percentage of wage recipients above the quantile level. E.g., 60% of wage recipients are above, and 40% below, the second pentile. Whichever characteristic is chosen, we have to keep in mind that it is a simplification. Another possible approach comprises monitoring a higher number of characteristics (of not only the location). In addition to location, we can also pay attention to variability, skewness, kurtosis, etc.

Another approach is to describe the frequency distribution of individual income groups. Apart from other advantages, this approach enables us to derive any of the above-mentioned characteristics at the required level of accuracy. We can also predict the future distribution on the basis of the time evolution of the parameters in the model.

2. Description the Wage Distribution

2.1. Description the Frequency Distribution

If the wage distribution is more or less "smooth", it can be adequately modelled with the aid of a suitable theoretic (continuous) distribution, such as a lognormal one [1 & 2]. The following formula (compliant with the notation used in SAS software [3]) represents the density of a two-parameter lognormal distribution.

$$PDF\left('LOGN',x,\theta,\lambda\right)=\begin{cases}0 & x\le 0\\ \dfrac{1}{x\lambda\sqrt{2\pi}}\exp\left(-\dfrac{\left(\log\left(x\right)-\theta\right)^2}{2\lambda^2}\right) & x\ge 0\end{cases}$$

Figure 1 below shows that the wage distribution could be modelled by lognormal distribution in the beginning years. It also indicates, however, that the wage distribution has been becoming multimodal in the recent years and the use of the lognormal model is thus problematic.

[+] Corresponding author. Tel.: +(420)224095481; fax: + (420)224095431.
E-mail address: marek@vse.cz.

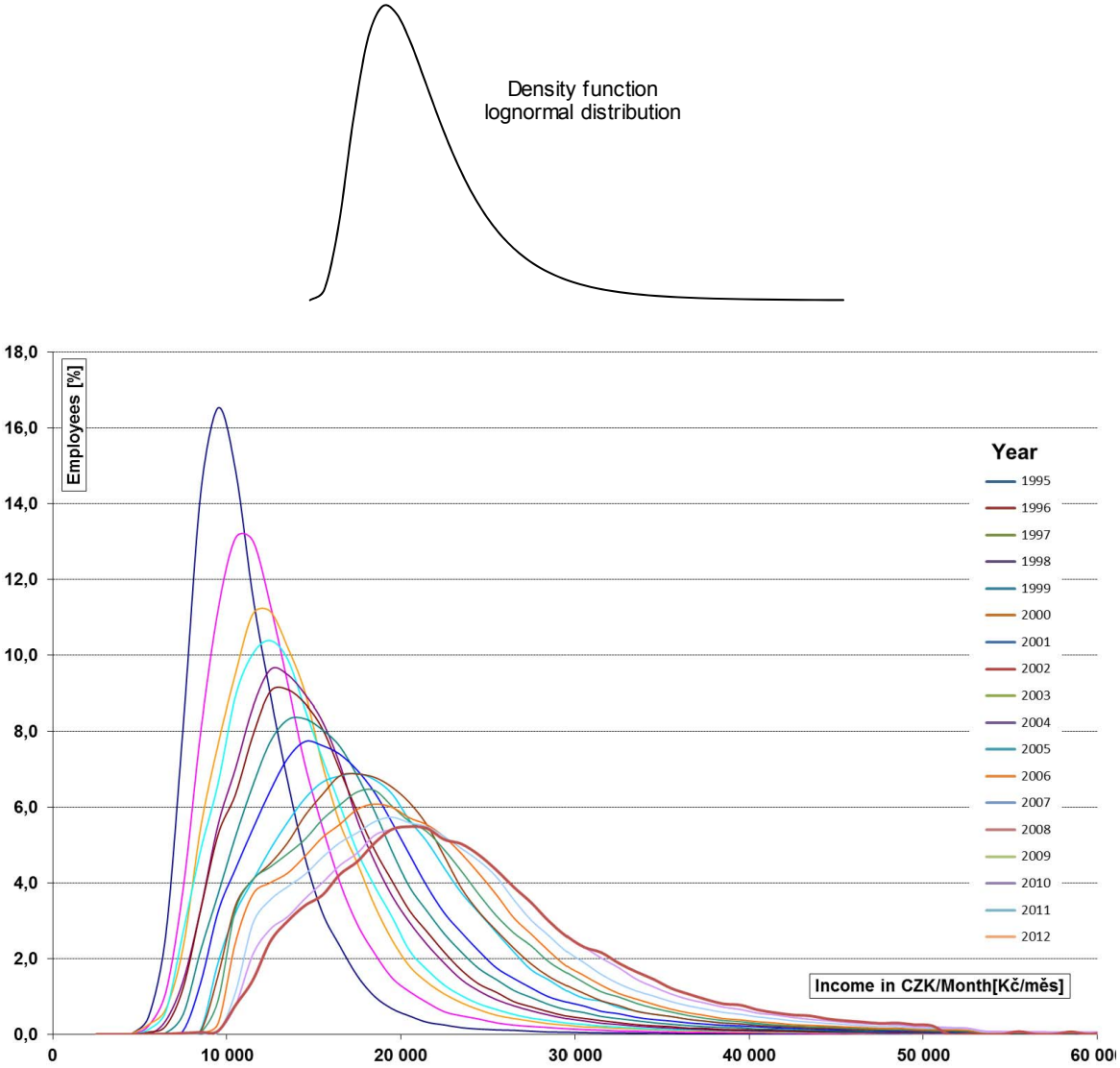

Fig. 1: Empirical distribution of income in the years 1995-2012.

On the other hand, the multimodal character might be well explained if the population is suitable subdivided. Fig. 2 shows a division by gender. A secondary effect of a subdivision is that skewness values of the component distributions are smaller. All these reasons led us to modelling the wage distribution with the aid of a mixture of normal distributions.

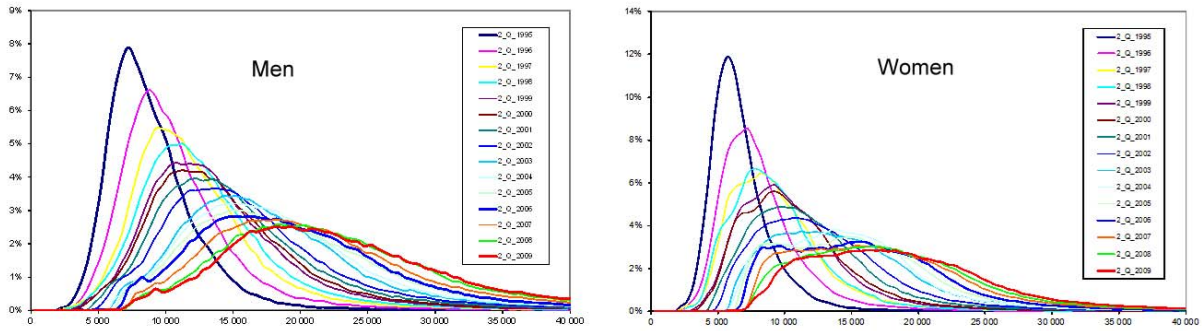

Fig. 2: Empirical distribution of income men - women.

2.2. Normal Mixture Distribution

The probability density for a general model of a normal mixture can be written as follows (where the standard SAS notation is used):

$$PDF\left('NORMALMIX', x, n, p, m, s\right) = \sum_{i=1}^{n} p_i \cdot PDF\left('NORMAL', x, m_i, s_i\right)$$

Here PDF stands for a probability density of a mixture of normal distributions (´NORMALMIX´) or a normal distribution as such (´NORMAL´) x for the argument, n for the number of components in the mixture, and p is the vector of weights, for which holds.

$$0 < p_i < 1, \forall i, \sum_{i=1}^{n} p_i = 1,$$

m and *s* are vectors of mean values and standard deviations of individual components in this mixture.

The density of normal distribution (of individual components in this mixture) is expressed by the following formula (again, in compliance with SAS):

$$PDF\left('NORMAL', x, \theta, \lambda\right) = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{\left(x-\theta\right)^2}{2\lambda^2}\right) \theta = m_i, \lambda = s_i, \forall i.$$

The standard approach (parameter estimation on the basis of selected optimisation criteria) is rather good for describing the history (even though interpretation is not easy) but it cannot be used for useful prediction of the future development. Several methods for estimating such parameters have been described in the literature (Expectation Maximisation (EM), Markov Chain Monte Carlo, Moment Matching, EF3M algorithm, etc.). The EM algorithm is most frequently used for practical applications – it is an iterative method for establishing the estimate with the aid of the Maximum Likelihood or MAP - Maximum Aposteriori Probability [4]. This algorithm is included in SAS [5]. In the general case, 3•n + 1 parameters have to be estimated (among them n itself). See [2] for details. Hence we decided for another method, namely, that of factual determination of parameters and a construction of the mixture on the basis of standard prediction of parameters within the mixture.

2.3. Factual Determination of Parameters

This approach brings about considerable advantages. The first such advantage is the factual interpretation. E.g., the simplest model (division of the population by gender, to men and women) we get n=2, m_1, m_2 are the expected 2013 wage values for men and women (respectively), and s_1, s_2 are the corresponding standard deviation values. Another advantage is a simple construction of the prediction for the future period (2013). The Figures below illustrate the linear evolution of these parameters in time.

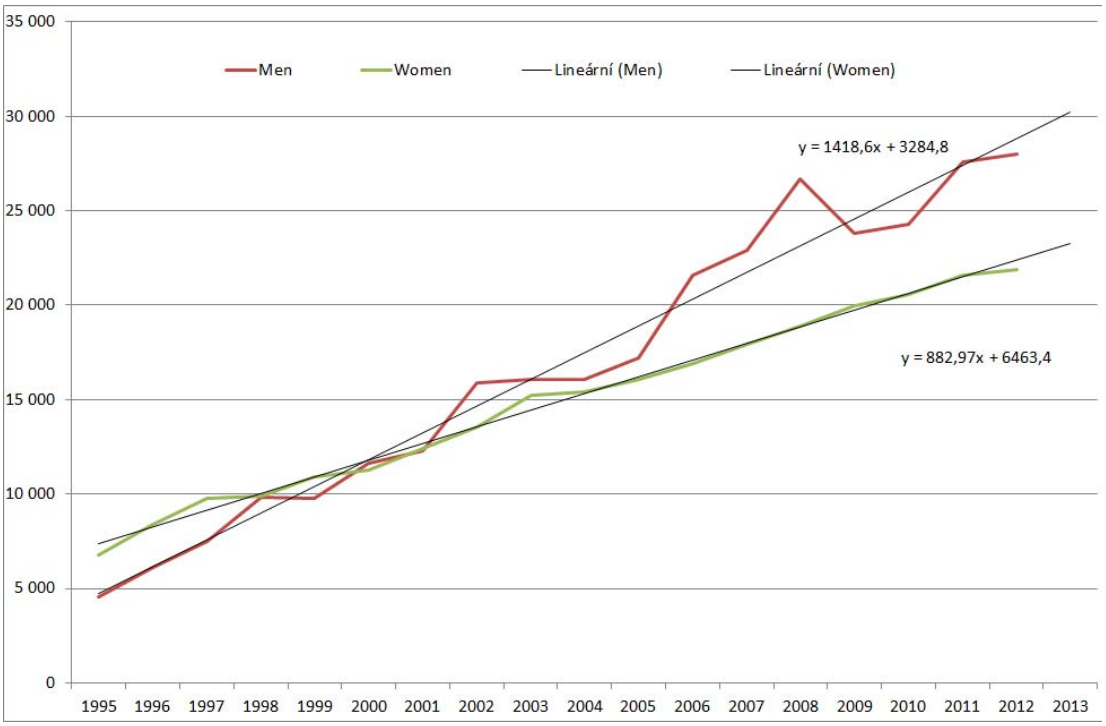

Fig. 3: Average incomes - men, women, CR.

Fig. 4: Variability of incomes - men, women, CR.

Hence we can estimate the mixture parameters for 2013 by a linear trend (cf. the Table below).

Table 1: Development parameters

	Men			Women		
Year	**Count**	**Average**	**StdDev**	**Count**	**Average**	**StdDev**
1995	65%	9 221	4 538	35%	6 794	2 720
1996	58%	11 100	6 118	42%	8 363	3 683
1997	52%	12 737	7 462	48%	9 740	4 766
1998	53%	13 914	9 808	47%	9 872	5 255
1999	54%	14 835	9 790	46%	10 878	5 345
2000	53%	15 537	11 654	47%	11 281	5 936
2001	56%	16 580	12 299	44%	12 435	6 569
2002	54%	17 987	15 876	46%	13 565	7 722
2003	55%	19 784	16 078	45%	15 217	8 726
2004	50%	20 109	16 042	50%	15 380	8 459
2005	50%	21 188	17 183	50%	16 076	8 463
2006	50%	22 203	21 565	50%	16 882	9 472
2007	50%	24 026	22 933	50%	17 916	10 480
2008	50%	25 821	26 701	50%	18 912	11 233
2009	50%	26 929	23 814	50%	19 957	11 605
2010	49%	27 644	24 261	51%	20 585	11 726
2011	48%	28 359	27 597	52%	21 583	12 690
2012	49%	29 374	27 985	51%	21 894	12 907
2013	**49%**	**30 124**	28 038	**51%**	**22 034**	**13 056**

The resulting mixture (its parameters) is given by this formula:

$$PDF\left('NORMALMIX', x, 2, (0,48; 0,52), (30124; 22034), (28038; 13056)\right)$$

There is the corresponding estimated empirical density of the wage distribution (only the values from 10 thousand to 52 thousand are tabulated, with a two-thousand step).

Table 2: Empirical density function

X	10	12	14	16	18	20	22	24	26	28	30
p(x)	0,015	0,029	0,049	0,072	0,093	0,105	0,104	0,093	0,077	0,061	0,051
X	32	34	36	38	40	42	44	46	48	50	52
p(x)	0,044	0,038	0,034	0,029	0,024	0,019	0,015	0,011	0,008	0,005	0,002

The following Figure illustrates the estimated wage distribution in the Czech Republic for 2013.

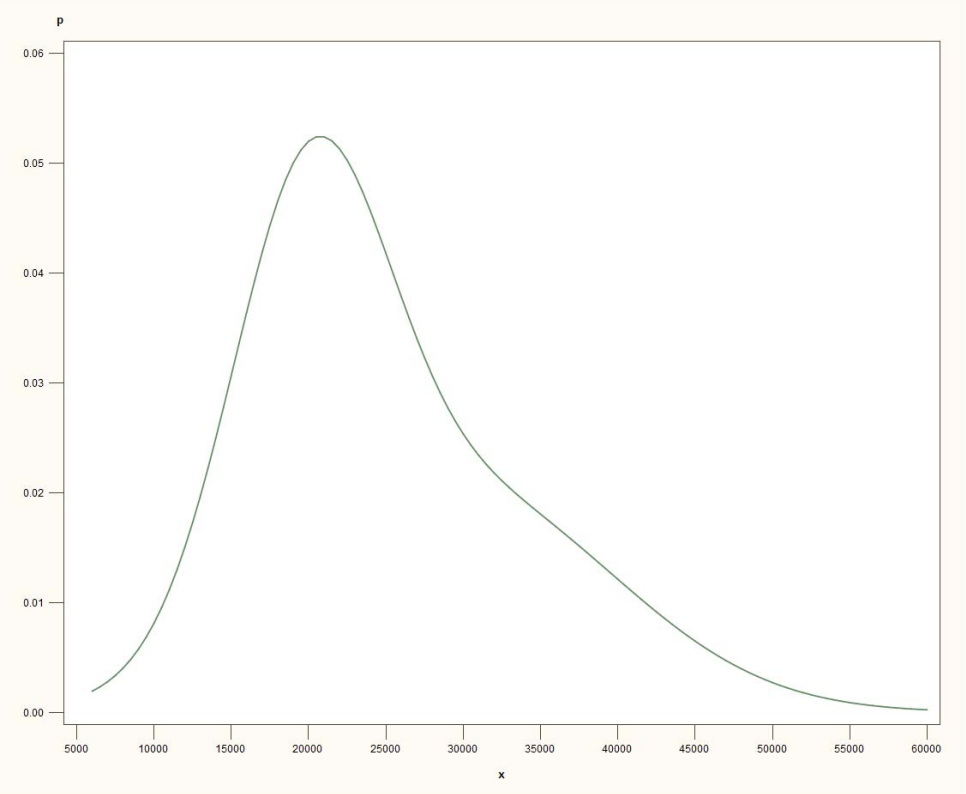

Fig. 5: Empirical density function.

Parameters for the remaining subdivisions were estimated in a similar way. These estimates are summarised in the Table below.

Table 3: Parameters for region

Region				
Praha	14	0,18	33 842	35 774
Středočeský		0,10	24 995	17 913
Jihočeský		0,06	22 625	19 735
Plzeňský		0,05	23 501	13 948
Karlovarský		0,02	22 330	16 023
Ústecký		0,07	23 080	15 617
Liberecký		0,04	23 432	14 573
Královehradecký		0,05	22 752	13 140
Pardubický		0,05	22 220	16 574
Vysočina		0,05	22 941	14 147
Jihomoravský		0,11	24 040	17 177
Olomoucký		0,06	22 341	14 966
Zlínský		0,05	21 842	14 222
Moravskoslezský		0,12	23 448	17 226

Table 3: Continuation parameters for gender and age

Sex	n	Count	Average	StdDev
Men	2	0,48	28 842	21 583
Women		0,52	27 597	12 690
Age				
under 30	3	0,16	21 356	10 216
30-50		0,50	26 406	24 433
over 50		0,34	24 984	21 378

3. Conclusion

Neither tables nor estimate charts for empiric densities are shown for these models. Differences in frequencies implied by individual subdivisions of the basic population are not very large. We can provide these results, together with the SAS code, to interested parties. If we are able to get structured data, we will try and formulate a comprehensive model with 84 ($2 \times 3 \times 14$) components or with 420 components (by including five education categories).

We take the approach outlined above for a possible method of how to make estimates of the wage distribution more accurate, facilitating subsequent analyses (such as the Income Tax yield).

4. References

[1] L. Marek, M. Vrabec. K moznostem modelovani mzdovych rozdeleni. Praha 13.12.2010 – 14.12.2010. In: *Reprodukce lidskeho kapitalu – Vzajemne vazby a souvislosti*. [online] Praha : KDEM VSE, 2010, s. 1–9. ISBN 978-80-245-1697-4. URL: http://kdem.vse.cz/resources/relik10/Index.htm.

[2] L. Marek, M. Vrabec. Forecast of the Income Distribution in the Czech Republic in 2011. Ras Al Khaimah 29.11.2010 – 03.12.2010. In: ICABR 2010 – VI. *International Conference on Applied Business Research*. Brno: Mendel University in Brno, 2010, s. 142. ISBN 978-80-7375-462-4.

[3] SAS Institute. *Base SAS(R) 9.2 Procedures Guide: Statistical Procedures*, Third Edition.

[4] A.P. DEMPSTER, N.M. LAIRD; D.B. RUBIN, (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1): 1–38. JSTOR 2984875. MR0501537.

[5] Edward P. Hughes and Trevor D. Kearney. *Optimization with the SAS® System*. SAS Institute Inc. 2012.