# Wage distribution models

*Lubos Marek*
Dept. Statistics and Probability
University of Economics, Prague
Prague, Czech Republic
marek@vse.cz

*Michal Vrabec*
Dept. Statistics and Probability
University of Economics, Prague
Prague, Czech Republic
vrabec@vse.cz

*Abstract.* **In our article we try to contribute to the discussion of the possibility to predict the trend of the wage distribution. For this purpose we use data from Czech Republic. But our model is useable for all similar data types. 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 are divided into groups with respect to sex, age and regions.**

***Keywords* — wage distribution, probability models, mixture normal density functions**

## I. 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.

## II. WAGE DISTRIBUTION

### A. 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 represents the density of a two-parameter lognormal distribution with parameters $\mu$ and $\sigma^2$.

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}, \qquad x > 0,$$
$$= 0, \qquad x \le 0.$$

First figure below shows that the wage distribution could be modelled by lognormal distribution in the first 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.

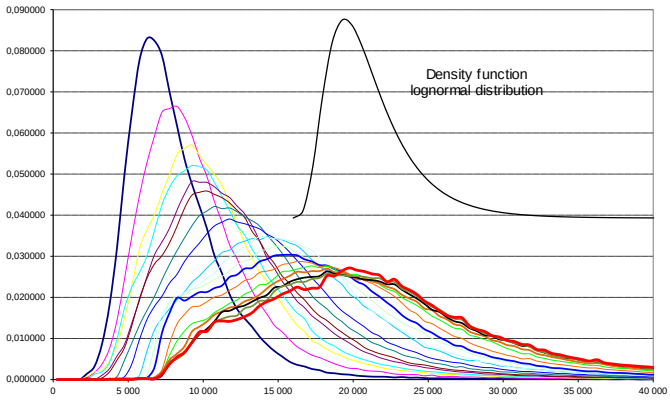


Fig. 1. *Empirical* wage *distribution*

On the other hand, the multimodal character might be well explained if the population is suitable subdivided. The next figures show a division by sex. 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.

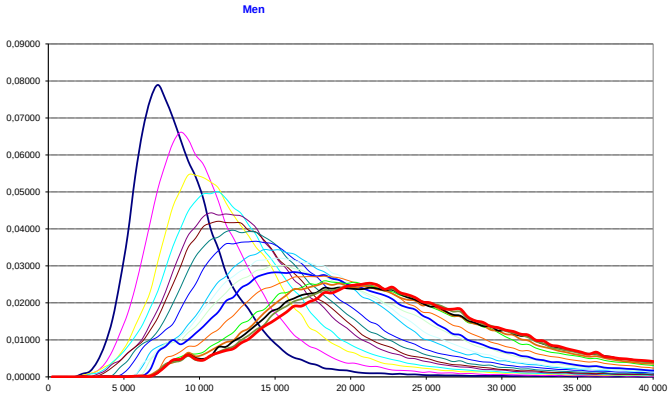


Fig. 2. *Empirical wage distribution - men*

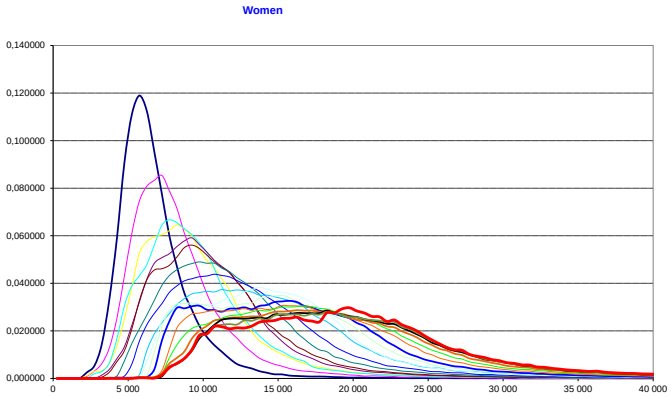


Fig. 3. *Empirical wage distribution - women*

## B. *Description the frequency distribution*

The probability density for a general model of a normal mixture can be written as follows

$$PDF\left(\acute{}NORMALMIX\acute{},x,n,p,\mu,\sigma\right)=\sum_{i=1}^{n}p_i\cdot PDF\left(\acute{}NORMAL\acute{},x,\mu_i,\sigma_i\right)$$

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

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

$\mu$ and $\sigma$ 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

$$PDF\left(\acute{}NORMAL\acute{},x,\mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

$$\mu=\mu_i,\ \sigma=\sigma_i,\ \forall i$$

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, 3n + 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.

## C. *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 sex, to men and women) we get n=2, are the expected 2013 wage values for men and women (respectively), and 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.

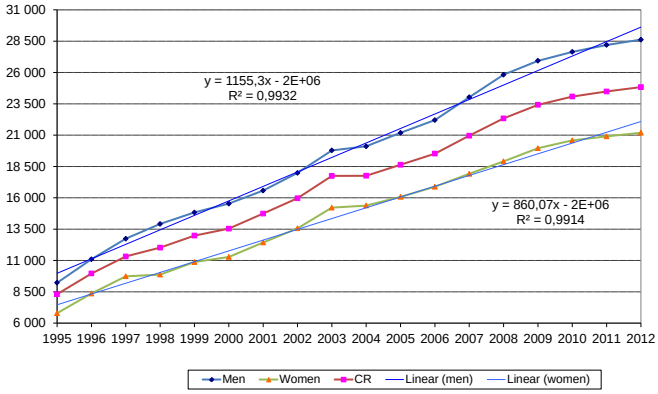


Fig. 4. *Average wage – men and women*

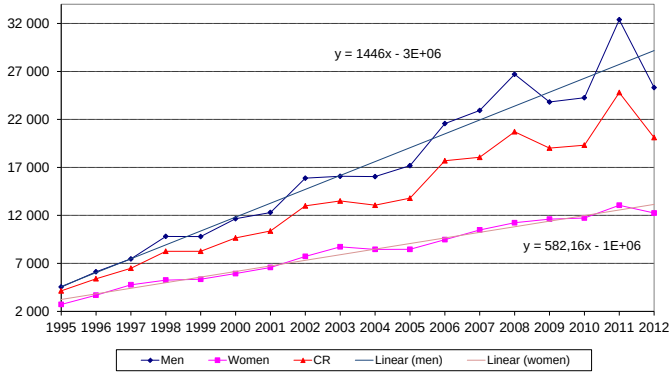


Fig. 5. *Standard deviation of wage – men and women*

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

TABLE I. EMPIRICAL PARAMETERS –GROUPS BY SEX

| | Men | | | Women | | |
|---|---|---|---|---|---|---|
| Year | Weight | Average | StdDev | Weight | Average | StdDev |
| 1995 | 0,646 | 9 221 | 4 538 | 0,354 | 6 794 | 2 720 |
| 1996 | 0,599 | 11 100 | 6 118 | 0,401 | 8 363 | 3 683 |
| 1997 | 0,532 | 12 737 | 7 462 | 0,468 | 9 740 | 4 766 |
| 1998 | 0,538 | 13 914 | 9 808 | 0,462 | 9 872 | 5 255 |
| 1999 | 0,535 | 14 835 | 9 790 | 0,465 | 10 878 | 5 345 |
| 2000 | 0,531 | 15 537 | 11 654 | 0,469 | 11 281 | 5 936 |
| 2001 | 0,557 | 16 580 | 12 299 | 0,443 | 12 435 | 6 569 |
| 2002 | 0,542 | 17 987 | 15 876 | 0,458 | 13 565 | 7 722 |
| 2003 | 0,554 | 19 784 | 16 078 | 0,446 | 15 217 | 8 726 |
| 2004 | 0,503 | 20 109 | 16 042 | 0,497 | 15 380 | 8 459 |
| 2005 | 0,502 | 21 188 | 17 183 | 0,498 | 16 076 | 8 463 |
| 2006 | 0,497 | 22 203 | 21 565 | 0,503 | 16 882 | 9 472 |
| 2007 | 0,497 | 24 026 | 22 933 | 0,503 | 17 916 | 10 480 |
| 2008 | 0,496 | 25 821 | 26 701 | 0,504 | 18 912 | 11 233 |
| 2009 | 0,496 | 26 929 | 23 814 | 0,504 | 19 957 | 11 605 |
| 2010 | 0,495 | 27 644 | 24 261 | 0,505 | 20 585 | 11 726 |
| 2011 | 0,491 | 28 196 | 32 390 | 0,509 | 20 903 | 13 056 |
| 2012 | 0,490 | 28 617 | 25 318 | 0,510 | 21 189 | 12 245 |

The resulting mixture (its parameters) is given by this formula (the last row in table):

$$PDF\left(\acute{}NORMALMIX\acute{}, x, 2, (0,49; 0,51), (28617; 21189), (25318; 12245)\right)$$

There is the corresponding estimated empirical density of the wage. The following Figure illustrates the estimated wage distribution in the Czech Republic for 2013 for model 1 - mixture 2 sex groups (men, women)

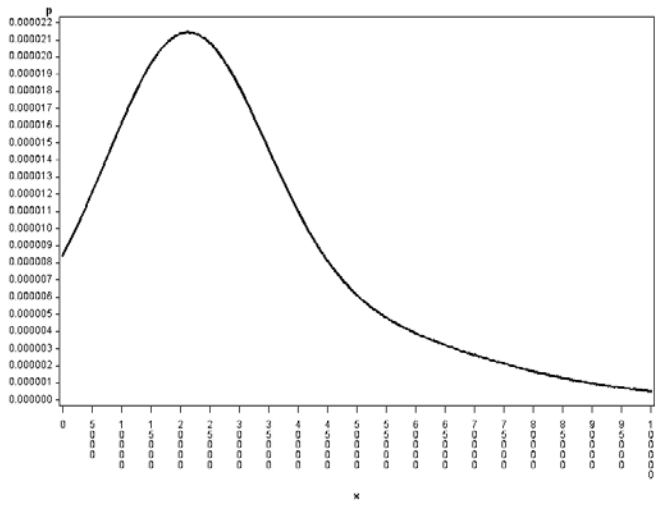

Fig. 6. *Wage distribution – model 1*

Parameters for the remaining subdivisions (groups by age and by regions) were estimated in a similar way.

The next Figure illustrates the estimated wage distribution in the Czech Republic for 2013 for model 2 - mixture 3 age groups (till 30, 30-50, over 50).

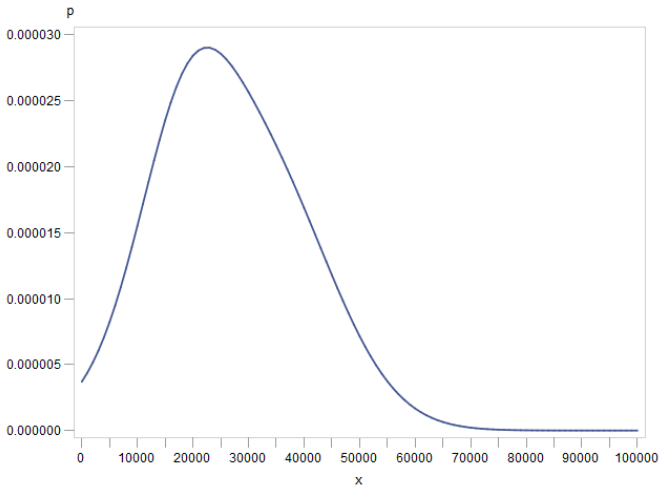


Fig. 7. *Wage distribution – model 2*

The last Figure illustrates the estimated wage distribution in the Czech Republic for 2013 for model 3 - mixture 14 region groups.

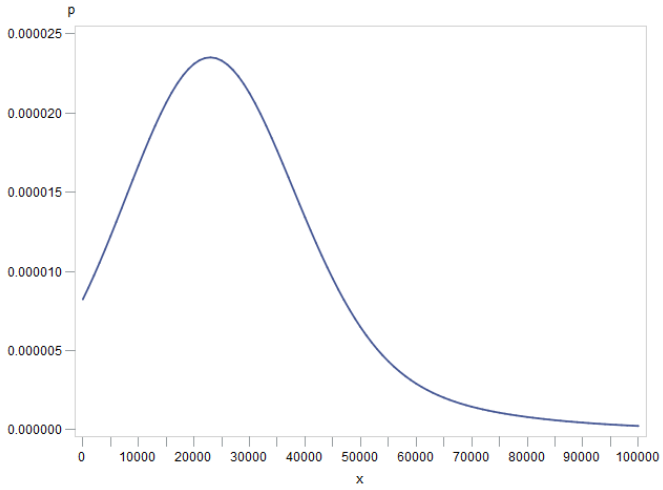


Fig. 8. *Wage distribution – model 3*

## D. *Conclusions*

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 × 3 × 14) components or with 420 components (by including five education categories).

## 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, 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.