VOLUME 6, 2012

ISSN 2045-3345

ISSN 2045-3345

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

MODELLING OF CONDITIONAL DISTRIBUTIONS OF WAGES IN THE CZECH REPUBLIC

Ivana Malá[I]
University of Economics in Prague, Czech Republic

ABSTRACT

In the text conditional distributions of positive value random variable are studied. The conditions are described by information that the random variable is less or equal to a given value, it is included in a given (finite) interval or greater than a given value. Conditional distribution functions are derived and characteristics of the level and variability are evaluated with the use of numeric methods. The results are applied to monthly gross wages (in CZK) in the Czech Republic in 2008. In the text two models for the probability distribution of wages are used; three parametric lognormal distribution and a mixture of three parametric lognormal densities for men and women. All unknown parameters were estimated with the use of moment method of estimation. Conditional moment and quantile characteristics of location (median, expected value) and variability (quartile deviation, standard deviation, coefficient of variation) are evaluated and given in the tables in order to obtain estimated characteristics and to compare results from different models fitted to the same dataset.

JEL CLASSIFICATION & KEYWORDS

■ C51 ■ C13 ■ D31 ■ FINITE MIXTURE OF DISTRIBUTION ■ LOGNORMAL DISTRIBUTION ■ DISTRIBUTION OF WAGES ■ INCOME DISTRIBUTIONS

INTRODUCTION

Conditional distributions of random variable X (with continuous probability distribution) given conditions of the type $X \geq z$, $z_1 < X \leq z_2$ and $z < X$ for given finite quantities z, z_1 and z_2 are studied in the text. The results are used for the modelling of the probability distributions of wages in the Czech Republic in 2008.

When analysing income or wages data we have frequently information whether an income is less or equal to given amount, it is included in a given (finite) interval or it is greater than a given amount. For example various questionnaires incorporate questions of this type. The distribution of wages in the Czech Republic is given by intervals in the website of the Czech Statistical office (CZSO). We will be interested in possibilities how to operate with such information from the point of view of probability theory.

For the modelling of distributions of incomes and wages various probability distributions are used in huge number of applications. Lognormal distribution for incomes and wages in the Czech Republic is used in Bartošová (2006, 2009), Bílková (2008, 2011 and 2012) and Bílková, Malá (2012). In the literature another distributions are used as generalized lambda distribution (for wages in the Slovak Republic in Pacáková, Sipková (2007)), Dagum distribution, generalized gamma or beta distributions or Pareto distribution for high incomes. Incomes or wages are usually very non-homogenous and a treated population consists of different subpopulations. For this reason the mixtures of probability densities can be used for the modelling of these distributions. An overview of probability distributions that are frequently used for the modelling of wages and incomes with comments and recommendations is given in Kleiber, Kotz (2003).

In this text the values of estimated parameters of three parametric lognormal models for wages in the Czech Republic were taken from Bílková (2011); these values are shown in the Table 1. The author used lognormal distribution to fit wages for the whole population and then separately for men and women. Estimates of unknown parameters were performed with the use of moment method instead of usually used maximum likelihood estimation. In this text estimated distributions for 2008 are used to model the distribution of wages; three parametric lognormal distribution and then three parametric lognormal distributions are fitted to wages for men and women and mixed into one mixture distribution for the whole population. Conditional distributions over chosen intervals (Tables 2 and 3) are compared with the use of estimated characteristics of the level (expected value and median) and variability (standard deviation, quartile deviation and coefficient of variation).

Methodology

Suppose X to be a positive value random variable with continuous probability distribution. Suppose that we have information that value of this variable X is included in a finite interval (z_1, z_2) and nothing more about the distribution of X is known. In this case different approaches are possible to the conditional distribution over the given interval. If we have no preference in the interval, uniform distribution over the interval (z_1, z_2) can be used for the modelling of conditional distribution of X. In this case the (conditional) mean and median are $(z_1 + z_2)/2$ standard deviation equals and $\frac{z_2 - z_1}{\sqrt{12}}$ quartile deviation is equal to $(z_2 - z_1)/2$ (it follows from the properties of the uniform distribution). The finite length of the interval is crucial, and this distribution is not applicable for infinite interval $X > z$.

Let the random variable has probability density f and distribution function F. Then the conditional probability distributions over intervals $X \leq z$, $z_1 < X \leq z_2$ and $z < X$ (for given values z, z_1, z_2) may be derived with the use of straightforward probability computations. In this paragraph conditional distribution functions are given; conditional densities could be find applying derivative to formulas (1)-(3).

Suppose X to be included in an (finite) interval (z_1, z_2) it means information that $z_1 < X \leq z_2$ for given $0 < z_1 < z_2 < \infty$. In this case a conditional distribution function $F(x \mid z_1 < X \leq z_2)$ (given $z_1 < X \leq z_2$) can be evaluated from the definition of conditional probability as

$$F\left(X \leq x \middle| z_1 < X \leq z_2\right) = \begin{cases} 0 & x \leq z_1, \\ \dfrac{P\left(z_1 < X \leq x\right)}{P\left(z_1 < X \leq z_2\right)} = \dfrac{F(x) - F(z_1)}{F(z_2) - F(z_1)}, & z_1 < x < z_2, \\ 1 & x \geq z_2. \end{cases} \quad (1)$$

[I] malai@vse.cz

If the condition is given as $X \le z$ the conditional distribution function $F(x \mid X \le z)$ can be expressed (we put $z_1=0$ to (1)) for $x<z$ as

$$F\left(x|X \le z\right) = \frac{P\left(X \le x\right)}{P\left(X \le z\right)} = \frac{F(x)}{F(z)}, \qquad x < z, \qquad (2)$$

and 1 for $x \ge z$. For the condition $X>z$ the conditional distribution function $F(x \mid X>z)$ is equal to 0 for $x \le z$ and it can be written as (substituting $z_2=\infty$ to (1))

$$F\left(x|X > z\right) = \frac{P\left(z < X \le x\right)}{P\left(z < X\right)} = \frac{F(x) - F(z)}{1 - F(z)}, \qquad z < x. \quad (3)$$

The distribution function F of the three parametric lognormal distribution $LN(\theta, \mu, \sigma^2)$ can be expressed as $F(x) = \Phi\left(\frac{\ln(x-\theta)-\mu}{\sigma}\right)$, where Φ is the distribution function of standard normal distribution. In case of a mixture of two lognormal distributions $LN(\theta_1, \mu_1, \sigma_1^2)$ and $LN(\theta_2, \mu_2, \sigma_2^2)$ with a mixing proportion $0<\alpha<1$ we suppose the distribution function F to be

$$F(x) = \alpha\Phi\left(\frac{\ln(x-\theta_1)-\mu_1}{\sigma_1}\right) + (1-\alpha)\Phi\left(\frac{\ln(x-\theta_2)-\mu_2}{\sigma_2}\right). \quad (4)$$

Suppose that the unconditional distribution is unimodal (as lognormal distribution is). In this case it is easy to describe the shape of conditional density. In the case of the condition $X \le z$ the conditional distribution is unimodal if z is greater than unconditional mode (its value is included in the interval $(0, z)$. The conditional density is increasing function if x is less than the (unconditional) mode and decreasing function for x greater than the (unconditional) mode. The conditional and unconditional modes have the same value (with different value of the density). If the condition z is less than the unconditional mode, the conditional density is increasing function over the interval $(0, z)$.

For the condition $X>z$ the conditional distribution is unimodal (with the same mode as that of unconditional distribution) if the unconditional mode is included in the interval (z, ∞), and the density is decreasing otherwise. For the condition $z_1<X \le z_2$ the conditional distribution is unimodal if the unconditional mode is included in the interval (z_1, z_2), it is increasing function if unconditional mode is less than z_1 and it is decreasing otherwise. These shapes of conditional distributions are well visible in the Figures 1 and 2.

Explicit formulas are known for the characteristics of location and variability for the three parametric lognormal distribution (Bílková (2011), Kleiber, Kotz (2003)). In case of a mixture the expected value is evaluated as a weighted average of expected values of components (with weights equal to mixing proportions). Standard deviation of the mixture may be evaluated from the component expected values and component variances (McLachlan, Peel (2000)). There are not explicit formulas for the mode (maximum of the density) and quantiles. These values were evaluated with the use of numeric procedures as well as characteristics of conditional distributions. Numeric integration (for the conditional moment characteristics) and solving equation (for the conditional quantile characteristics) were used to evaluate all characteristics. All calculations were performed in MS Office 2010 Excel and R v.3.11 (RPROGRAM).

Data and results

In this part conditional distributions that were discussed in the previous section are applied to the distribution of wages in the Czech Republic. Suppose X in the previous section to be monthly gross wages (in CZK) in the Czech Republic in 2008. In 2008 the mean annual exchange rate was 24.942 CZK/EUR (CNB). Bílková (in Bílková (2011)) applied three parametric lognormal distribution as a model. Estimated values of unknown parameters for lognormal distribution of the whole population and for men and women separately are shown in the Table 1.

Densities for the subgroups of men and women from the Table 1 are mixed into a mixture of two lognormal distributions (4) with the mixing proportion α=0.57 for men and 1-α=0.43 for women (CZSO). The mixture of two lognormal distributions is not lognormal distribution; it can be bimodal but in the analysed case the mixture is unimodal (Figure 1) and all findings about the shape of conditional distributions from the previous part can be applied. Estimated expected value (Table 1) are in euro for the whole population 925 EUR (928 EUR form the mixture model resp.), 1 032 EUR for men and 790 EUR (only 77 % of the value for man) for women. If the modes are estimated, these values are 656 (655 resp.) for the whole population and 737 EUR for man and 565 EUR (77 % of 737) for women.

In the text intervals (described by amounts z_i) that were chosen for the analysis, given in the Tables 2 - 4 are chosen. These intervals are used to tabulate the distribution of incomes by the website of the Czech Statistical Office (CZSO). With this choice the first conditional distribution is given by the condition *X≤10 000 CZK* then nine intervals with conditional distributions are given by conditions $z_1<X \le z_2$ the first one *10 000<X≤12 000* and the last one *40 000<X≤50 000.* Last rows in the tables give the only infinite interval with conditional density given by the condition *X>50 000 CZK.* The estimated proportions of these intervals in the population are given in the Table 2. The modal interval is 12 000–16 000 CZK for both models of the whole population and for men; for women the modal interval is previous one 12 000 – 16 000 CZK. Very similar values were obtained by both models of the whole population; the distribution of wages for men is shifted to the higher values and the distribution for women to lower values. This contemplation once again illustrates the gap between wages of men and women. If we compare proportions estimated by the Czech Statistical Office the coincidence is good but not excellent. Usually the fit of lognormal distribution to income or wages describe well location but underestimates variability. As a consequence we obtain the differences in proportions in Table 2.

From the estimated parameters given in the Table 1 values of characteristics of location and variability for all estimated models can be evaluated. The mode was found 16 360 CZK in case of lognormal distribution and 16 336 CZK for the mixture (both values included in the interval 12 000–16 000 CZK). The mode is equal to 18 378 CZK for men and it is also included in the interval 16 000–20 000 CZK. For women the mode 14 085 CZK is included in the interval 12 000–16 000 CZK. It corresponds with modal intervals discussed above (Table 2). From the previous explanations, the conditional density with an extreme (unimodal density with the maximum) will be in the interval 16 000–20 000 CZK for overall density obtained with lognormal distribution, the mixture and for the subpopulation of men. The unimodal conditional density for wages of women occurs in the interval 12 000–16 000 CZK. All other densities are monotone functions of wages.

Medians (not given in the Table 1) are 22 232 CZK evaluated from lognormal distribution and 21 714 CZK for the mixture. For men median wage is 24 634 CZK and for women it equals 18 578 CZK. It can be seen, from previous figures (and Table 1), that median and mode for the mixture are smaller than for lognormal distribution and expected value

Table 1: Estimates of three unknown parameters of the lognormal distribution for monthly gross wages and salaries (in CZK) in the Czech Republic in 2008; estimated characteristics of location and variability (CZK)

	Estimated parameters				Estimated characteristics (CZK)		
	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\theta}$	$\hat{\alpha}$	expected value	mode	standard deviation
Czech Republic	9.851	0.345	2 920		23 061	16 360	20 141
Men	10.008	0.311	2 431	0.57	25 735	18 378	23 304
Women	9.520	0.400	4 949	0.43	19 714	14 085	14 765
Mixture					23 146	16 336	20 302

Source: own calculations, Bílková (2011), CZSO

Table 2: Estimated proportions of analysed intervals

Interval (CZK)		Model				
z_1	z_2	observed	lognormal	men	women	mixture
0	10 000	0.032	0.047	0.027	0.058	0.040
10 000	12 000	0.047	0.058	0.039	0.090	0.061
12 000	16 000	0.137	0.158	0.123	0.221	0.165
16 000	20 000	0.183	0.166	0.149	0.192	0.167
20 000	24 000	0.179	0.142	0.142	0.139	0.141
24 000	28 000	0.139	0.112	0.121	0.095	0.110
28 000	32 000	0.090	0.084	0.096	0.064	0.082
32 000	36 000	0.056	0.062	0.074	0.043	0.061
36 000	40 000	0.035	0.045	0.056	0.029	0.045
40 000	50 000	0.045	0.066	0.087	0.038	0.066
50 000	∞	0.057	0.061	0.086	0.029	0.062

Source: own calculations, CZSO

Table 3: Estimated quantile (median, CZK) and moment (expected value, CZK) characteristics of location for conditional distributions on analysed intervals

Interval (CZK)			Median (CZK)		Expected value (CZK)	
z_1	z_2	centre	lognormal	mixture	lognormal	mixture
to	10 000		8 817	8 981	8 593	8 783
10 000	12 000	11 000	11 090	11 093	11 061	11 063
12 000	16 000	14 000	14 103	14 078	14 070	14 053
16 000	20 000	18 000	17 958	17 939	17 972	17 959
20 000	24 000	22 000	21 895	21 886	21 930	21 924
24 000	28 000	26 000	25 865	25 864	25 910	25 909
28 000	32 000	30 000	29 851	29 853	29 900	29 902
32 000	36 000	34 000	33 844	33 849	33 896	33 899
36 000	40 000	38 000	37 842	37 848	37 895	37 898
40 000	50 000	45 000	44 041	44 080	44 352	44 379
50 000	or more		59 969	60 463	65 322	66 084

Source: own calculations

Table 4: Estimated quantile (quartile deviation) and moment (standard deviation, coefficient of variation) characteristics of variability for conditional distributions on analysed intervals

Interval (CZK)		Quartile deviation (CZK)		Standard deviation (CZK)		Coefficient of variation (%)	
z_1	z_2	lognormal	mixture	lognormal	mixture	lognormal	mixture
0	10 000	769	662	1 067	882	12.41	17.11
10 000	12 000	492	490	573	572	5.18	5.19
12 000	16 000	983	984	1 145	1 145	8.14	8.17
16 000	20 000	994	994	1 151	1 150	6.40	6.40
20 000	24 000	995	995	1 152	1 153	5.25	5.24
24 000	28 000	994	994	1 154	1 155	4.45	4.44
28 000	32 000	994	994	1 153	1 152	3.85	3.85
32 000	36 000	994	994	1 154	1 154	3.40	3.39
36 000	40 000	994	994	1 152	1 152	3.4	3.4
40 000	50 000	2 409	2 416	2 844	2 848	6.41	6.40
50 000	∞	8 248	8 660	20 407	17 787	31.24	26.95

Source: own calculations

is greater. Common inequalities that the smaller characteristic of location is mode, than median and the greater value is expected value due to the rare but high wages are true for both models. It follows from the Table 3, that in first three intervals conditional medians are greater than conditional expected values.

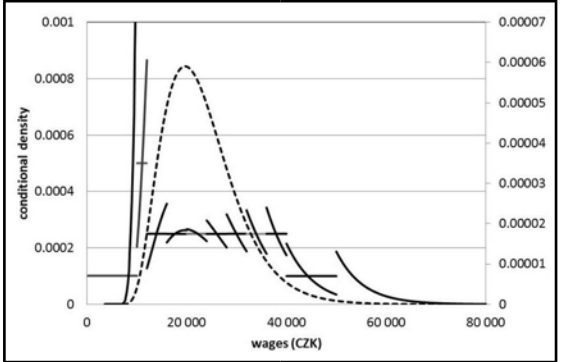
In the Tables 3 and 4 the estimated characteristics of the conditional distributions for all treated intervals are shown. In the Table 3 characteristics of location of all conditional distributions (median, expected value) are given. If a uniform distribution over the interval (for finite intervals) is used, all characteristics in the table are equal to the centre of the interval. If the lognormal or mixture distributions for wages are used, these values are included in the interval but are generally different from the centre. In the intervals to 16 000 CZK both values are greater than for uniform distribution, for intervals with greater values these values are less than for uniform distribution. No uniform distribution can be used for the last infinite interval (wages over 50 000 CZK).

In the Table 4 characteristics of variability of the conditional distributions (quartile deviation, standard deviation and coefficient of variation) are shown. As in case of characteristics of location the similar values were obtain by both models. As expected, values of standard deviations are higher than quartile deviations and absolute variability increases in all the models. Coefficients of variation are small, but most of intervals have the lengths of 4 000 CZK, so it is not surprising. Note that despite the fact that all values in Table 4 are in Czech koruna, the characteristic of relative variability coefficient of variation doesn't depend on the scale (currency).

In the Figures 1 (mixture model) and 2 (lognormal model) estimated conditional densities are given (evaluated as the derivative of conditional distribution functions (1) – (3)). Horizontal lines are uniform distribution densities over intervals. The dotted line is an estimated (unconditional) probability density of the model. Both Figures are in the same scale, because of really different values of densities, the unconditional density is plotted against the right axis (the values of unconditional density are smaller than values of conditional densities). We can see monotone densities with exception of the interval 16 000–20 000 CZK. Both modes are very close to the left limit (16 360 and 16 336 CZK) and the extremes are not well visible.

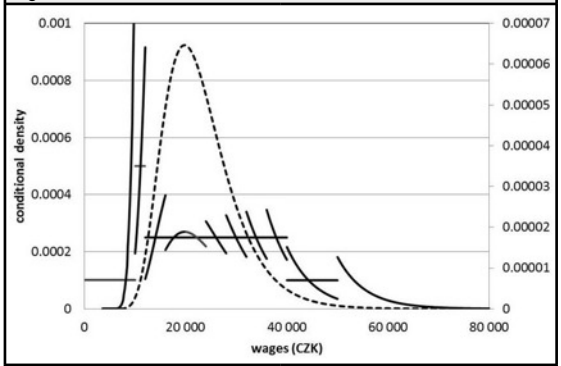
Both figures are very similar and at the first sight undistinguishable. Lognormal densities are greater than mixture densities (uniform densities are the same). The

Figure 1: Estimated density (dotted line, right axis) and estimated conditional densities (solid line, left axis), uniform distribution on analysed intervals (solid line, parallel with horizontal axis) - mixture model



Source: own calculations

Figure 2: Estimated density (dotted line, right axis) and estimated conditional densities (solid line, left axis), uniform distribution on analysed intervals (solid line, parallel with horizontal axis) - lognormal model



Source: own calculations

mixture has greater standard deviation (as it was mentioned lognormal distribution usually underestimates variability) and this fit is better from the point of view of chi-square goodness-of-fit. Due to the large samples all models are significally different ($P<0.01$) from the empirical distribution.

Conclusion

In the text an application of conditional probability distributions is given for wages in the Czech Republic in 2008. Easy probability calculations give interesting results connected with probability distributions of wages under information about the interval that includes its value. Lognormal distribution is used but this method is general enough to be applicable for different probability distributions frequently used for the modelling of wages or incomes. The proposed method needs only the choice of an overall probability distribution, but this choice enables more complex models than lognormal distribution or a mixture of two lognormal distributions. All computations are done with the use of numeric methods, so no close formulas are needed (and in fact it is not possible to obtain explicit formulas generally).

The proper choice of the distribution provides more precise conclusions about conditional distributions that can be used for uniform distribution. In case of inappropriate model distribution all results can be misleading and the use of uniform distribution is safer in this case.

It is shown (in the Tables 3 and 4 and in the Figures 1 and 2) that for relatively narrow intervals the values of estimated characteristics are similar (Tables 3 and for) even for diverse distributions (given by densities in Figures 1 and 2).

References

Bartošová, J. Logarithmic-Normal Model of Income Distribution in the Czech Republic. Austrian Journal of Statistics, Vol. 35, No. 23, s. 215 – 222. 2006.

Bartošová, J., Bína, V. Modelling of income distribution of Czech households in years 1996-2005. Acta Oeconomica Pragensia, roč. 17, č. 4, s. 3–18. 2009.

Bílková, D. Application of Lognormal Curves in Modeling of Wage Distributions. Journal of Applied Mathematics, Vol. 1, No. 2, pp. 341 – 352. 2008.

Bílková, D. Analysis of the Development in Wage Distributions of Men and Women in the Czech Republic in Recent Years. Statistika, 2011, roč. 48, č. 1, s. 40–57. 2011.

Bílková, D. Malá, I. Modelling the Income Distributions in the Czech Republic since 1992. Austrian Journal of Statistics, 41 (2), 133–152. 2012.

Bílková, D. Recent Development of the Wage and Income Distribution in the Czech Republic. Prague Economic Papers: 21 (2), 233–250. 2012.

Cohen A.C., Whitten J.B. Estimation in the three-parameter lognormal distribution, Journal of American Statistical Association, 75,399-404. 1980.

Kleiber, C., Kotz, S. Statistical size distributions in economics and actuarial sciences. Hoboken : John Wiley and Sons. 2003.

Malá. I. The use of finite mixtures of lognormal distributions in the modeling of incomes in the Czech Republic. Research Journal of Economics, Business and ICT, 2011, Vol. 4, 41–46. 2011.

McLachlan, G. J., Peel, D. Finite Mixture Models. Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section, New York. 2000.

Pacáková,V., Sipková,L. Generalized Lambda Distributions of Household\s Incomes. E + M Ekonomie a Management, Vol.10, Iss.1, s. 98 – 107. 2007.

CZSO. Czech statistical office. www.czso.cz. [2012-10-15]

CNB. Czech National Bank. www.cnb.cz. [2012-10-10]

RPROGRAM. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. http://www.R-project.org/. [2012-10-20]