Stochastic Models of Wage Distributions: Empirical Comparison

Michal Vrabec
Dept. of Statistics
University of Economics, Prague
e-mail: vrabec@vse.cz

Petr Berka
Dept. of Information and Knowledge Engineering
University of Economics, Prague
and
Dept. of Informatics and Mathematics
University of Finance and Administration, Prague
e-mail: berka@vse.cz

Abstract

A number of stochastic models for modeling time series data can be found in the literature. Among them models based on Log-normal distribution are more traditional, while models using Johnson SB or Johnson SU distributions were introduced recently. We present basic properties of the above-mentioned distributions and discuss their usability to model economic data. Data concerning the wages of more than two million Czech employees collected for more than twenty years are used for the comparison.

1 Introduction

Statistical analysis of the development of the wage and income distribution is a crucial precondition for economic modeling of the labour market processes. One of the most discussed characteristics of the wage distribution is the average wage. There is an ongoing debate about the suitability of the average as a measure of the wage level. There are proposals to replace the average by median, and/or to consider additional characteristics like variability or percentiles. In our opinion, it is necessary to work with the entire wage distribution.

If the wage distribution is more or less "smooth", it can be adequately modeled with the aid of a suitable theoretic (continuous) distribution, such as a log-normal one ([2]). But as shown e.g. in ([3]), as far as wages are concerned, the log-normal

distribution is not the best-fitting one and this distribution is most often used mainly because of its convenient theoretical qualities. Following this argument, we present an empirical comparison of the log-normal distribution and log-logistic distribution with Johnson SB and Johnson SU distributions.

2 Used distributions

2.1 Log-normal Distribution

Log-normal distribution (sometimes also called Galton distribution) is a continuous probability distribution of a random variable whose logarithm is normally distributed. The formula (1) represents the density of a three-parameter log-normal distribution: here μ is the location parameter and σ is the scale parameter ($\sigma > 0$) for the normally distributed logarithm $\ln(X)$.

$$p(x) = \frac{1}{(x-\gamma)\sigma\sqrt{2\pi}} \times \exp\left[-\frac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}\right], \ x \geq 0 \tag{1}$$

If $\gamma = 0$ then the three-parameter log-normal distribution changes into two-parameter one, as shown in formula (2).

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \times \exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right], \ x \geq 0 \tag{2}$$

2.2 Log-logistic Distribution

Log-logistic distribution is the probability distribution of a random variable whose logarithm has a logistic distribution. The formula (3) represents the density of this distribution: here γ is the location parameter ($\gamma = 0$ yields a two-parameter distribution) and α is shape parameter ($\alpha > 0$) and β is scale parameter ($\beta > 0$).

$$p(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2} \tag{3}$$

2.3 Johnson Distribution

Johnson distributions [1] are based on a transformation of the standard normal variable. Given a continuous random variable X whose distribution is unknown and is to be approximated, Johnson proposed three normalizing transformations having the general form:

$$Z = \gamma + \delta f\left(\frac{X-\mu}{\sigma}\right) \tag{4}$$

where $f(.)$ denotes the transformation function, Z is a standard normal random variable, γ and δ are shape parameters, σ is a scale parameter and μ is a location parameter. Without loss of generality, it is assumed that $\delta > 0$ and $\sigma > 0$. Johnson distributions include three forms: log-normal, bounded and unbounded.

The most simple transformation defines the log-normal system of distributions denoted by S_L

$$Z = \gamma + \delta \ln\left(\frac{X-\mu}{\sigma}\right), \ X > \Theta \tag{5}$$

The bounded system of distributions S_B is defined by

$$Z = \gamma + \delta \ln\left(\frac{X-\mu}{\mu+\sigma-X}\right), \ \mu < X < \mu + \sigma \tag{6}$$

S_B curves cover bounded distributions. The distributions can be bounded on the lower end, the upper end or both ends. This family covers Gamma distributions, Beta distributions and many others.

The unbounded system of distributions S_U is defined by

$$Z = \gamma + \delta \ln\left\{\left(\frac{X-\mu}{\sigma}\right) + \left[\left(\frac{X-\mu}{\sigma}\right)^2 + 1\right]^{1/2}\right\}, \ -\infty < X < \infty \tag{7}$$

The S_U curves are unbounded and cover the t and normal distributions, among others.

Using the fact that, after the transformation in (4), Z follows standard normal distribution, the probability density function $p(y)$ of each of the family in the Johnson system can be derived. If X follows the Johnson distribution and $Y = \frac{X-\mu}{\sigma}$ then, for S_L family,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{y} \times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln\left(y\right)\right)^2\right] \ \mu < x < \infty \tag{8}$$

similarly, for the S_B family,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1-y}{y} \times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln\left(\frac{y}{1-y}\right)\right)^2\right] \ \mu < x < \mu + \sigma \tag{9}$$

and for the S_U family,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{\sqrt{y^2+1}} \times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln\left(y + \sqrt{y^2+1}\right)\right)\right] -\infty < x < \infty \tag{10}$$

In general, the the probability density function of X is given by

$$p(x) = \frac{\delta}{\sqrt{2\pi}} \times f'\left(\frac{x-\mu}{\sigma}\right) \times \exp\left[-\frac{1}{2}\left(\gamma + \delta f\left(\frac{x-\mu}{\sigma}\right)\right)^2\right] \quad x \in G \tag{11}$$

where

$$f'(y) = \begin{cases} \frac{1}{y}, & \text{for the } S_L \text{ family} \\ \frac{1}{y(1-y)}, & \text{for the } S_B \text{ family} \\ \frac{1}{\sqrt{y^2+1}}, & \text{for the } S_U \text{ family} \end{cases} \tag{12}$$

and

$$f(y) = \begin{cases} \ln(y), & \text{for the } S_L \text{ family} \\ \ln\left(\frac{y}{1-y}\right), & \text{for the } S_B \text{ family} \\ \ln\left(y + \sqrt{y^2+1}\right), & \text{for the } S_U \text{ family} \end{cases} \tag{13}$$

The support G of the distribution is:

$$G = \begin{cases} \langle \mu; \infty) , & \text{for the } S_L \text{ family} \\ \langle \mu; \mu + \sigma \rangle , & \text{for the } S_B \text{ family} \\ (-\infty; \infty), & \text{for the } S_U \text{ family} \end{cases} \tag{14}$$

3 Modeling wage distributions

3.1 Used data

We work with time series of wages in Czech Republic over the years 1995 - 2017. The annual data are reported in quarterly units; our study observes the average wages in the second quarter of each year. The scope of the data set on which the analyses were carried out was gradually increased from more than 300,000 observations in 1995 to more than two million in 2017. This data is structured in a very detailed way. The wage values are divided into intervals with widths of 500 CZK. Such a detailed structure enables us to achieve quite accurate results. We have basic characteristics of wages in the entire period at our disposal. The analysis was aimed at creating a model for probability distribution of wages (estimating the parameters of the probability density). We used only the data for the years 2015-2017 in the experiments. Table 1 shows basic characteristics of the data, all numbers except the sample size are in Czech crowns (CZK).

3.2 Parameter estimation

We used the SAS system and EasyFit program for computations. Figures (Fig. 1), (Fig. 2), (Fig. 3) and (Fig. 4) show the fit of the distributions on data

characteristics	year 2015	year 2016	year 2017
sample size	2 098 854	2 119 396	2 185 573
average wage	26 369	27 668	29 166
standard deviation	19 903	20 478	20 749
10th percentile	12 978	13 944	14 982
lower quartile	17 290	18 391	19 547
median	22 658	23 757	25 135
upper quartile	29 566	30 963	32 610
90th percentile	40 162	42 026	44 334
modus	8 635	9 275	10 296

Table 1: Basic characteristics of the used data

from the year 2015. We also performed the Kolmogorov-Smirnov test to assess the quality of the model. We tested the null hypothesis "H0: the data follow the specified distribution" against the alternative hypothesis "H1: the data do not follow the specified distribution" Table 2 gives the results of this test in terms of the Kolmogorov-Smirnov statistics and the rank of the model (in both cases, the lower is the value the better is the model).

distribution	year 2015		year 2016		year 2017	
	statistics	rank	statistics	rank	statistics	rank
2 par. log-normal	0,03808	3	0,03877	3	0,03949	3
3 par. log-normal	0,03739	2	0,03809	2	0,03886	2
log-logistic	0,01839	1	0,01667	1	0,01732	1
Johnson SB	0,06982	4	0,06605	4	0,0621	4

Table 2: Results of the Kolmogorov-Smirnov test

4 Conclusions

The aim of the analysis was to compare several models of probability distribution of wags in Czech Republic. The experiments show that the best model is the log-logistic distribution with three parameters. This confirms the previously achieved results ([4]). Anyway, as the wage variability grows over the years and empirical density's curves became less smooth, mixture models have potential to provide better models of wage distributions in the future. And good models that are able to make good predictions of the future wage distributions are necessary for various socio-economic considerations.

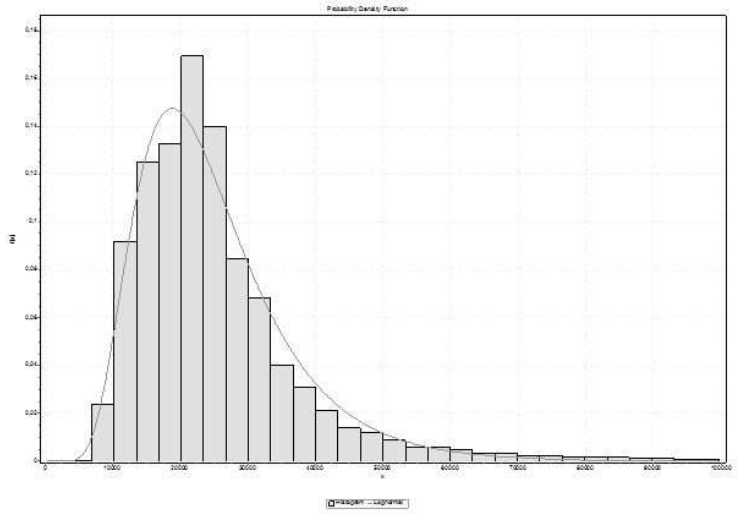

Figure 1: Two parameters log-normal distribution for wages from the year 2015. Here $\mu = 10,032$ and $\sigma = 0,43343$.

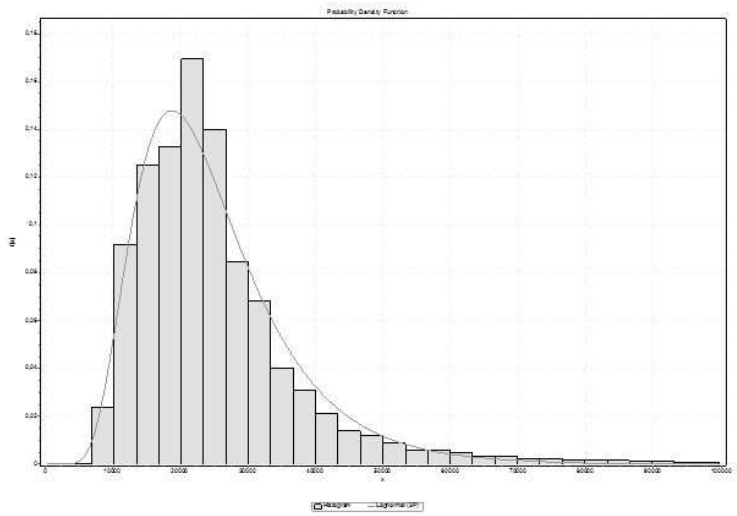

Figure 2: Three parameters log-normal distribution for wages from the year 2015. Here $\mu = 10,02$, $\sigma = 0,43841$ and $\gamma = 250$.

References

[1] N. L. Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(3/4):297–304, 1949.

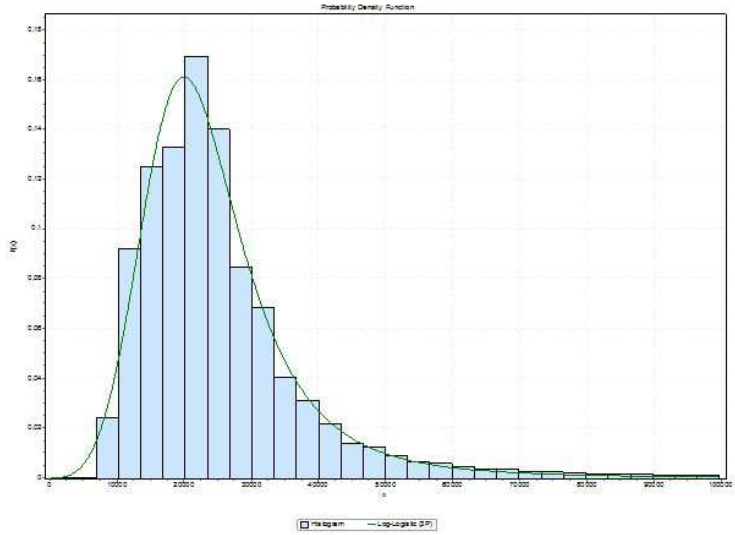

Figure 3: Log-logisic distribution for wages from the year 2015.

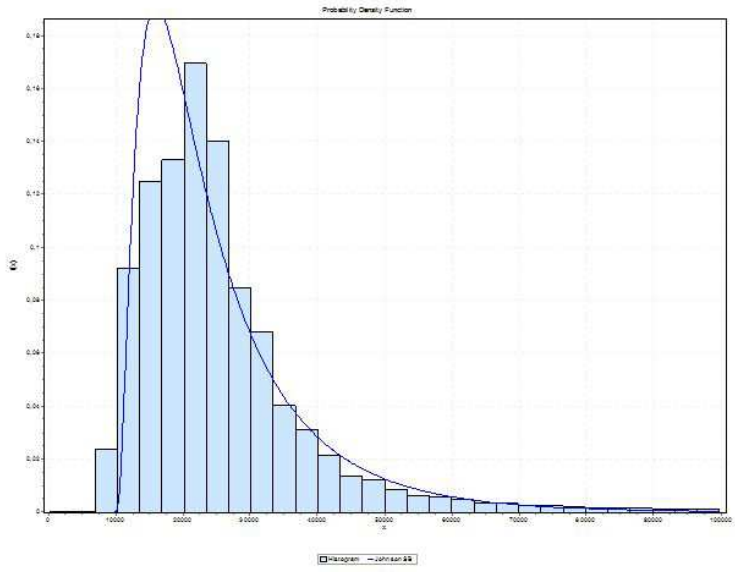

Figure 4: Johnson SB distribution for wages from the year 2015. Here $\mu = 9818,9$, $\sigma = 1,8883E+5$, $\gamma = 3,1927$ and $\delta = 1,1751$.

[2] L. Marek and M. Vrabec. Forecast of the income distribution in the czech republic in 2011. In *International Conference on Applied Business Research ICABR 2010*, page 142, 2010.

[3] M. Matějka and K. Duspivová. The czech wage distribution and the minimum wage impacts: an empirical analysis. *Statistika*, 93(2):61–75, 2013.

[4] M. Vrabec and L. Marek. Model for distribution of wages. In *Proceedings of the Applications of Mathematics and Statistics in Economics AMSE 2016*, pages 378–386, Banska Bystrica, 2016.