doi: 10.36689/uhk/hed/2023-01-008

Four-Parameter Lognormal Curves Combined with the Quantile Method of Parameter Estimation as Models of Salary Distribution

Diana BÍLKOVÁ

Prague University of Economics and Business, Prague, Czechia; bilkova@vse.cz

Abstract: The purpose of this paper is to present the four-parameter lognormal curves used to construct salary distributions. The beginning of these curves was considered to be the amount of the minimum wage that was valid on January 1 of the previous year. The maximum sample value was considered as the endpoint. The quantile parameter estimation method was used to estimate the remaining two parameters. Salary distribution models were constructed separately for men and women and according to the employee's educational attainment. The results show that women's salary distributions are characterized by higher skewness and kurtosis and at the same time lower level and variability. Similarly, this is valid for the lowest categories of educational attainment, too. As the level of educational attainment increases, both the skewness and the kurtosis of salary distributions decrease with increasing level and variability. The data for this research come from the Czech Statistical Office.

Keywords: four-parameter lognormal curve; quantile parameter estimation method; salary distribution model

JEL Classification: E24; C51; C55

1. Introduction

The importance of the lognormal distribution as a model for sample distributions cannot be disputed. This model has found application in various fields, starting from astronomy, through technology, medicine, economics, and sociology. The characteristic features of the process described by the lognormal model are the gradual action of interdependent factors, the tendency to develop in a geometric sequence and the growth of random variability into systematic variability, i.e. differentiation. In the field of economics, among the many phenomena that the lognormal model allows to interpret are also salaries and wages of employees and household incomes.

The idea that the logarithms of the values of variables from the field of economics have a normal distribution is of older date and is based on the fact that the effects of a large number of different impulses, the result of which is the value of the observed quantity, are proportional to the state of this quantity at the relevant moment.

The main purpose of this paper is to present four-parameter lognormal curves in combination with the quantile method of parameter estimation as models of salary distribution. Salary distribution models were constructed separately for men and women and according to the education attainment, too. The objective is to specify the typical shapes of salary distribution models of relatively homogeneous groups of employees created in this way and to monitor the development of these shapes over time. The data for this research covered the period 2014–2020 and comes from the Czech Statistical Office.

2. Review to Literature

Four-parameter lognormal curves have historically been used in various fields of science and research. From a historical perspective, already Saving (1965) uses a four-parameter lognormal distribution to model the scale loss and the size distribution of manufacturing establishment. Siano and Metzler (1969) show that the lognormal distribution provides a convenient four-parameter empirical description of the structureless bands of the ultraviolet absorption spectra of hydroxypyridine derivatives. Lambert (1970) examines methods for estimating the parameters of the four-parameter lognormal distribution. Mahmood (1973) finds that the lognormal distribution of particle size is often applicable to particles in nature and in industrial processes. Wingo (1975) presents a procedure for overcoming anomalies in statistical modeling using the maximum likelihood method for theoretical parameter estimates of three-parametric and four-parametric lognormal distributions. Gentry (1978) used a four-parameter lognormal distribution for particle size and a normal distribution for particle charge, and reports applications of the distribution for modeling the optical diameter of asbestos fibers, the bimodal charge distribution of sodium chloride aerosols, and the size distribution of atmospheric aerosol.

From a more modern perspective, based on a probabilistic fatigue damage mechanics system created by combining statistical fatigue analysis and macroscopic damage mechanics, Zeng and Yu (1991) propose a simplified four-parameter lognormal model in the hope of finding applications in engineering. Wagner and Ding (1994) present two three-parameter and one four-parameter lognormal curves in describing the distribution of size of soil aggregates. Regalado and Ritter (2009) investigate soil water repellency, which can be characterized as the delayed infiltration time of a water droplet resting on the soil surface, which is the penetration time of the water droplet or the persistence of repellency. The authors fit a four-parameter lognormal distribution to both common patterns obtained using dynamic factor analysis and then additively combine them in a weighted multiple linear bimodal model. Malama and Kuhlman (2015) extend a three-parameter lognormal model for the unsaturated hydraulic conductivity of moisture retention using a slight modification of Mualem's theory, which is nearly exact for nonclay soils, to a four-parameter lognormal model by truncating the underlying distribution of pore size distribution to a physically permissible minimum and maximum pore radii.

Bílková (2020) deals with the use of four-parameter lognormal distribution in the field of economics for data on Czech employee wages. Wage distribution models were constructed according to the regions of the Czech Republic, and the different shape of these distributions was researched. Bílková (2019) deals with the comparison of the accuracy of parameter estimation of four-parameter lognormal curves with the accuracy of parameter estimation of three-parameter lognormal curves.

3. Methodology

3.1. Four-Parameter Lognormal Distribution

The random variable X has a four-parameter lognormal distribution with parameters μ, σ^2, θ and τ, where $-\infty < \mu < \infty, \sigma^2 > 0, -\infty < \theta < \tau < \infty$, if its probability density has the form

$$f(x;\mu,\sigma^2,\theta,\tau) = \frac{(\tau-\theta)}{\sigma\cdot(x-\theta)\cdot(\tau-\theta)\cdot\sqrt{2\pi}}\cdot\exp\left[-\frac{\left(\ln\frac{x-\theta}{\tau-x}-\mu\right)^2}{2\sigma^2}\right], \qquad \theta < x < \tau, \qquad (1)$$

$$= 0, \qquad else.$$

The probability density function of the four-parameter lognormal distribution can take different shapes depending on the values of the distribution parameters. The distribution can have two modes for a combination of parameter values $\sigma^2 > 2$ and $|\mu| < \sigma^2\cdot\sqrt{(1-2/\sigma^2)} - 2\,tanh^{-1}\sqrt{(1-2/\sigma^2)}$. Figures 1–3 present the different shapes of the probability density of the four-parameter lognormal distribution depending on the parameter values.

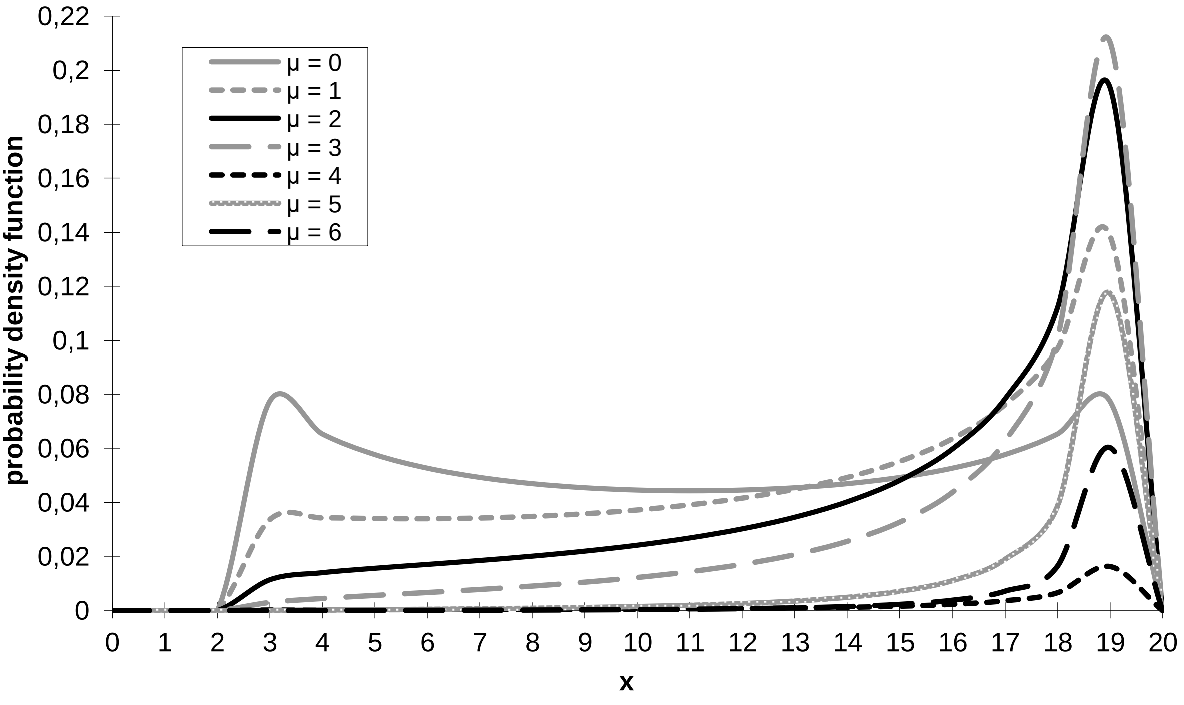

Figure 1. Probability density function shapes of the four-parameter lognormal distribution for parameter values σ = 2 (σ² = 4); θ = 2; τ = 20

If the random variable X has a four-parameter lognormal distribution with parameters μ, σ^2, θ, and τ, then the random variable

$$Y = ln\frac{X-\theta}{\tau-X} \qquad (2)$$

has a normal distribution with parameters μ and σ^2 and is a random variable

$$U = \frac{ln\frac{X-\theta}{\tau-X}-\mu}{\sigma} \qquad (3)$$

has a standardized normal distribution.

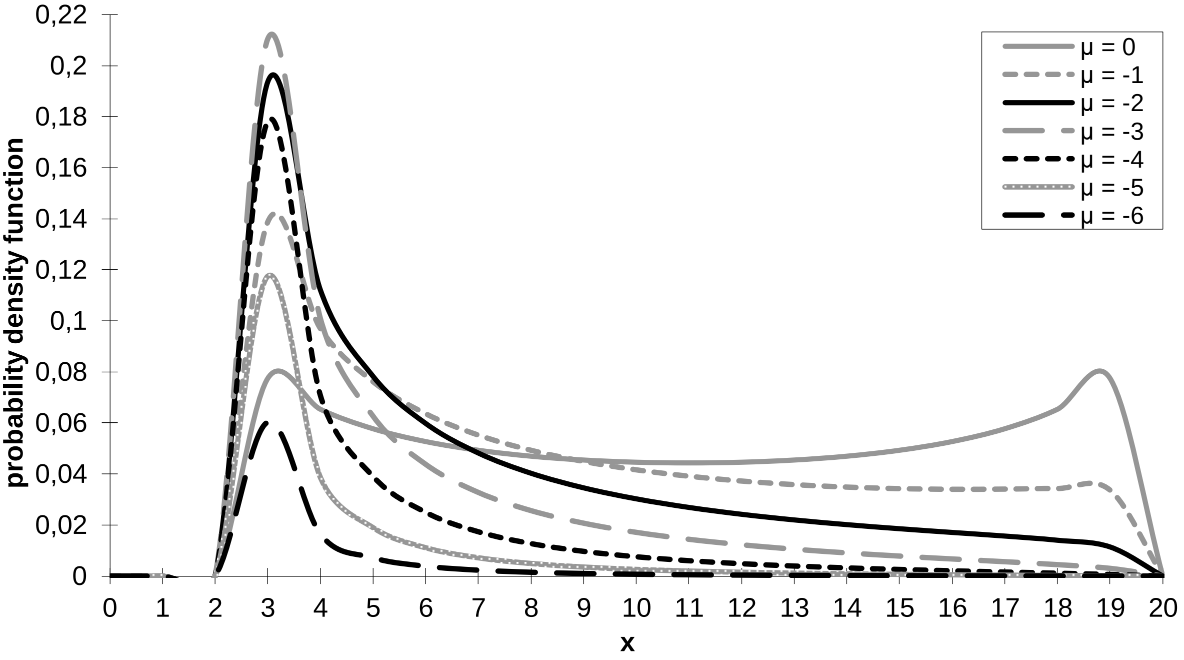

Figure 2. Probability density function shapes of the four-parameter lognormal distribution for parameter values σ = 2 (σ² = 4); θ = 2; τ = 20

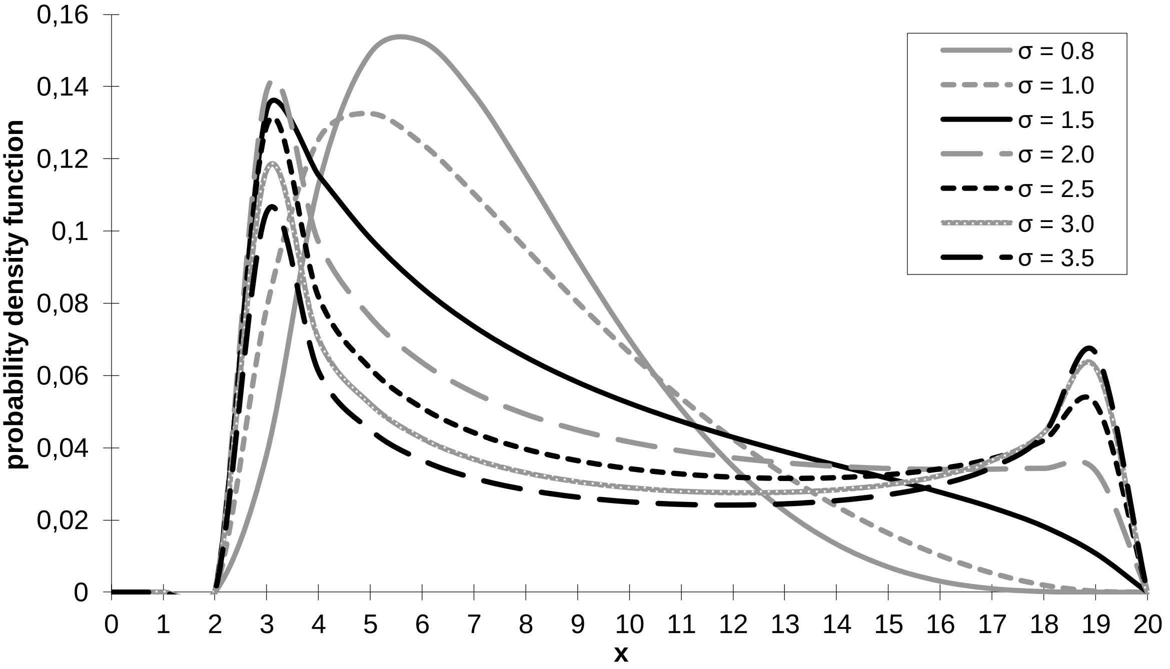

Figure 3. Probability density function shapes of the four-parameter lognormal distribution for parameter values μ = –1; θ = 2; τ = 20

Thus, the parameter μ represents the expected value of the random variable (2) and the parameter σ^2 is the variance of the random variable (2). The parameter θ represents the beginning (theoretical minimum) and the parameter τ represents the endpoint (theoretical maximum) of the four-parameter lognormal distribution of the random variable X. As mentioned above, the random variable (2) has a normal distribution with parameters μ and σ^2. The $100 * P\%$ quantile of the normal distribution with parameters μ and σ^2 of the random variable Y has the form

$$y_P = \mu + \sigma\mu_P, \tag{4}$$

where μ_P is the $100 * P\%$ quantile of the standardized normal distribution.

We assume the situation that we know the values of the parameters θ and τ, i.e. we consider the value of the parameter θ to be the monthly amount of the minimum wage valid on January 1 of the previous year (we assume that institutions do not respond quite flexibly to increases in minimum wage amounts, as the data suggests) and the value of the parameter τ to be the maximum sample value. We assume that institutions do not respond quite flexibly to increases in minimum wage amounts, as the data suggests.

3.2. Quantile Method of ParameterEstimation of Four-Parameter Lognormal Distribution for the Case of a Known Theoretical Minimum θ and Known Theoretical Maximum τ

We consider the value of the theoretical minimum θ to be the amount of the monthly minimum wage valid on January 1 of the previous year, and the value of the theoretical maximum to be the highest sample value of the employee's gross monthly salary. We therefore assume that we know the values of the parameters θ and τ and estimate the values of the remaining two parameters μ and σ^2.

We use the random variable (2), which has a normal distribution with parameters μ and σ^2 and the relation (4) for the $100 * P\%$ quantile of this variable. For estimation, we use the sample 50% and 75% quantiles (the sample median and sample upper quartile), which we calculate from the sample data set

$$\tilde{x}_{50} \; a \; \tilde{x}_{75},$$

which we substitute into the equations

$$y_{0.50} = ln\frac{\tilde{x}_{0.50} - \theta}{\tau - \tilde{x}_{0.50}} = \hat{\mu} + \hat{\sigma}\mu_{0.50} = \hat{\mu}, \tag{5}$$

$$y_{0.75} = ln\frac{\tilde{x}_{0.75} - \theta}{\tau - \tilde{x}_{0.75}} = \hat{\mu} + \hat{\sigma}\mu_{0.75} \tag{6}$$

We substitute equation (5) into equation (6) and after adjustment we get

$$\frac{ln\frac{\tilde{x}_{0.75} - \theta}{\tau - \tilde{x}_{0.75}} - \hat{\mu}}{\mu_{0.75}} = \frac{ln\frac{\tilde{x}_{0.75} - \theta}{\tau - \tilde{x}_{0.75}} - ln\frac{\tilde{x}_{0.50} - \theta}{\tau - \tilde{x}_{0.50}}}{\mu_{0.75}} = \hat{\sigma} \tag{7}$$

We estimate the remaining parameters μ and σ^2 using the formulas

$$\hat{\mu} = ln\frac{\tilde{x}_{0.50} - \theta}{\tau - \tilde{x}_{0.50}}, \tag{8}$$

$$\hat{\sigma} = \frac{ln\frac{\tilde{x}_{0.75}-\theta}{\tau-\tilde{x}_{0.75}} - ln\frac{\tilde{x}_{0.50}-\theta}{\tau-\tilde{x}_{0.50}}}{\mu_{0.75}} \qquad (9)$$

This theoretical part comes from Bílková (2019).

4. Results

Tables 1 and 2 include parameter estimations of four-parameter lognormal curves for models of salary distribution of men and women and for individual categories of educational attainment. Figures 4–13 characterize the development of models of salary distribution of men and women over time in separation for individual categories of educational attainment.

Table 1. Parameter estimations of the four-parameter lognormal distribution model – men

		Parameter estimation			
Year	Educational attainment	θ	τ	μ	σ
2014	Primary and incomplete	8,000	101,573	–2.118 330	0.592 896
	Secondary without A-level examination	8,000	162,393	–2.469 715	0.453 758
	Secondary with A-level examination	8,000	200,977	–2.204 078	0.352 670
	Post-secondary non-tertiary and bachelor´s	8,000	188,510	–1.967 703	0.405 963
	Higher	8,000	457,977	–2.794 689	0.575 774
2015	Primary and incomplete	8,500	117,100	–2.237 779	0.570 522
	Secondary without A-level examination	8,500	116,202	–2.036 623	0.482 337
	Secondary with A-level examination	8,500	120,376	–1.524 647	0.400 539
	Post-secondary non-tertiary and bachelor´s	8,500	183,333	–1.888 389	0.411 446
	Higher	8,500	468,488	–2.788 649	0.599 332
2016	Primary and incomplete	9,200	125,300	–2.355 467	0.621 345
	Secondary without A-level examination	9,200	123,490	–2.062 272	0.403 059
	Secondary with A-level examination	9,200	140,224	–1.695 371	0.384 522
	Post-secondary non-tertiary and bachelor´s	9,200	188,000	–1.881 127	0.413 765
	Higher	9,200	572,006	–2.927 412	0.564 757
2017	Primary and incomplete	9,900	99,407	–1.964 360	0.687 755
	Secondary without A-level examination	9,900	151,173	–2.196 462	0.399 940
	Secondary with A-level examination	9,900	234,055	–2.209 554	0.359 341
	Post-secondary non-tertiary and bachelor´s	9,900	161,971	–1.589 504	0.445 833
	Higher	9,900	594,331	–2.887 340	0.582 795
2018	Primary and incomplete	11,000	104,681	–1.914 383	0.663 423
	Secondary without A-level examination	11,000	154,303	–2.104 092	0.399 326
	Secondary with A-level examination	11,000	176,811	–1.733 450	0.394 345
	Post-secondary non-tertiary and bachelor´s	11,000	235,085	–1.905 981	0.418 620
	Higher	11,000	656,018	–2.894 681	0.583 026
2019	Primary and incomplete	12,200	159,518	–2.348 750	0.679 378
	Secondary without A-level examination	12,200	99,493	–1.482 924	0.472 031
	Secondary with A-level examination	12,200	244,879	–2.051 387	0.364 590
	Post-secondary non-tertiary and bachelor´s	12,200	174,077	–1.442 107	0.433 196
	Higher	12,200	654,822	–2.830 791	0.556 595
2020	Primary and incomplete	13,350	146,482	–2.118 705	0.726 013
	Secondary without A-level examination	13,350	146,462	–1.888 829	0.457 562
	Secondary with A-level examination	13,350	209,695	–1.804 389	0.382 311
	Post-secondary non-tertiary and bachelor´s	13,350	252,762	–1.824 119	0.411 046
	Higher	13,350	646,957	–2.766 533	0.566 719

Table 2. Parameter estimations of the four-parameter lognormal distribution model – women

Year	Educational attainment	Parameter estimation			
		θ	τ	μ	σ
2014	Primary and incomplete	8,000	90,593	–2.579 553	0.788 679
	Secondary without A-level examination	8,000	70,797	–2.158 153	0.670 754
	Secondary with A-level examination	8,000	137,337	–1.980 799	0.393 808
	Post-secondary non-tertiary and bachelor´s	8,000	114,623	–1.655 765	0.424 112
	Higher	8,000	297,764	–2.587 510	0.398 801
2015	Primary and incomplete	8,500	91,620	–2.580 129	0.841 298
	Secondary without A-level examination	8,500	73,521	–2.190 255	0.726 943
	Secondary with A-level examination	8,500	137,916	–1.963 475	0.420 828
	Post-secondary non-tertiary and bachelor´s	8,500	112,717	–1.603 680	0.461 233
	Higher	8,500	335,055	–2.692 874	0.398 890
2016	Primary and incomplete	9,200	76,701	–2.366 897	0.904 255
	Secondary without A-level examination	9,200	75,881	–2.209 633	0.742 280
	Secondary with A-level examination	9,200	122,007	–1.764 962	0.438 398
	Post-secondary non-tertiary and bachelor´s	9,200	127,865	–1.716 458	0.467 303
	Higher	9,200	318,661	–2.576 965	0.415 020
2017	Primary and incomplete	9,900	80,127	–2.251 009	0.927 585
	Secondary without A-level examination	9,900	86,369	–2.223 049	0.752 519
	Secondary with A-level examination	9,900	203,410	–2.302 699	0.431 193
	Post-secondary non-tertiary and bachelor´s	9,900	151,217	–1.828 363	0.472 733
	Higher	9,900	482,071	–2.946 159	0.413 054
2018	Primary and incomplete	11,000	117,410	–2.565 334	0.921 637
	Secondary without A-level examination	11,000	118,342	–2.443 786	0.739 067
	Secondary with A-level examination	11,000	149,598	–1.831 912	0.457 423
	Post-secondary non-tertiary and bachelor´s	11,000	163,050	–1.806 632	0.481 020
	Higher	11,000	375,665	–2.564 929	0.405 340
2019	Primary and incomplete	12,200	168,069	–2.862 998	0.878 699
	Secondary without A-level examination	12,200	349,411	–3.562 617	0.690 109
	Secondary with A-level examination	12,200	180,463	–1.956 706	0.434 243
	Post-secondary non-tertiary and bachelor´s	12,200	167,464	–1.712 724	0.497 725
	Higher	12,200	404,853	–2.501 197	0.349 863
2020	Primary and incomplete	13,350	158,029	–2.562 577	0.886 544
	Secondary without A-level examination	13,350	159,038	–2.483 049	0.721 258
	Secondary with A-level examination	13,350	240,005	–2.202 998	0.456 740
	Post-secondary non-tertiary and bachelor´s	13,350	240,173	–2.014 099	0.495 540
	Higher	13,350	413,894	–2.419 086	0.317 654

We succeeded to construct models of the salary distribution of Czech employees using four-parameter lognormal curves for the period 2014–2020, separately by gender and educational attainment, while the categories of educational attainment were broken down according to the website of the Czech Statistical Office into five categories: primary and incomplete education, secondary education without A-level examination, secondary education with A-level examination, post-secondary non-tertiary and bachelor´s education and higher education. The models of salary distribution of women are characterized by higher skewness and kurtosis with a lower level and variability compared to the models of salary distribution in the same year and in the corresponding category of educational attainment of men. It was also observed that the models of salary distribution for categories with the lowest educational attainment have the highest skewness and kurtosis and at the same time the lowest level and variability. As the educational attainment category increases, models of salary distribution tend to decrease in skewness and kurtosis while level and variability increase. For the salary distribution models of men and women for each category of educational attainment, it is true that the models of salary distribution at the beginning of the monitored period are characterized by the highest skewness and kurtosis and at the same time by the lowest level and variability. With the passage of time within the monitored period, the models of salary distribution for men and women and within each category of educational attainment have flattened, their peak also decreases and the level and variability increase.

5. Discussion

While Saving (1965) uses a four-parameter lognormal distribution to model the scale loss and the size distribution of manufacturing establishment, Bílková (2020) deals with application of four-parameter lognormal distribution and quantile method of parameter estimation in modelling of wage distribution, Bílková (2019) again deals with application of four-parameter lognormal distribution and quantile method of parameter estimation and this paper offers some comparison of the accuracy of four-parameter and three-parameter lognormal models. The added value of this paper is the application of four-parameter lognormal curves and the quantile method of estimating the parameters to the salary distribution of employees in the public sphere separated by educational attainment. Different shapes distribution models are typical for individual levels of educational attainment.

6. Conclusions

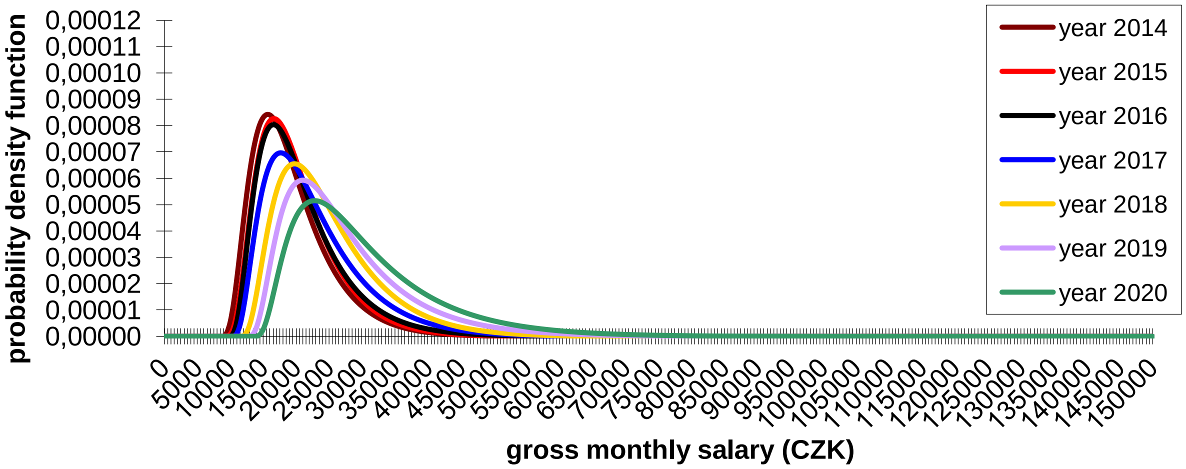

Figure 4. Development of the model distribution of the gross monthly salary of men in the period 2014–2020 for the category primary and incomplete education

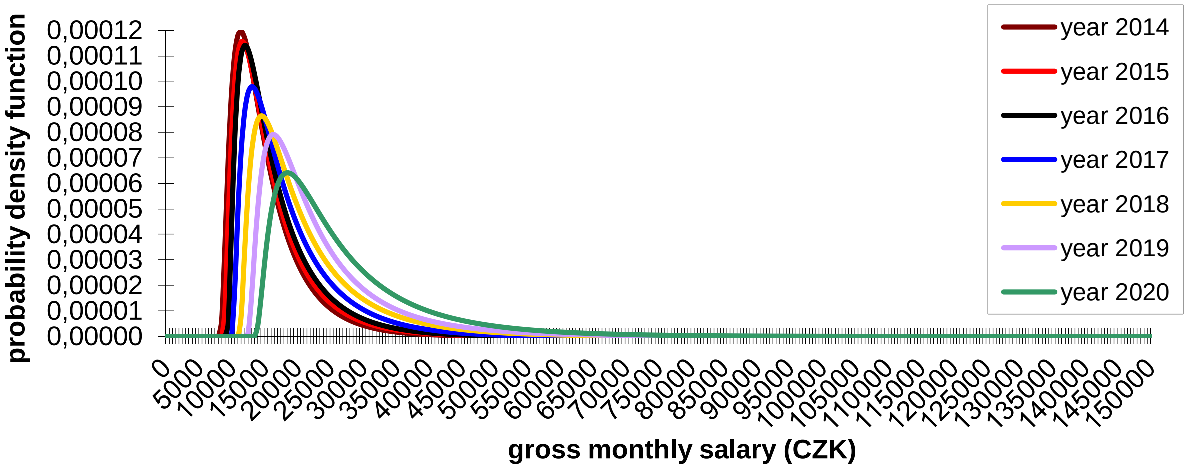

Figure 5. Development of the model distribution of the gross monthly salary of women in the period 2014–2020 for the category primary and incomplete education

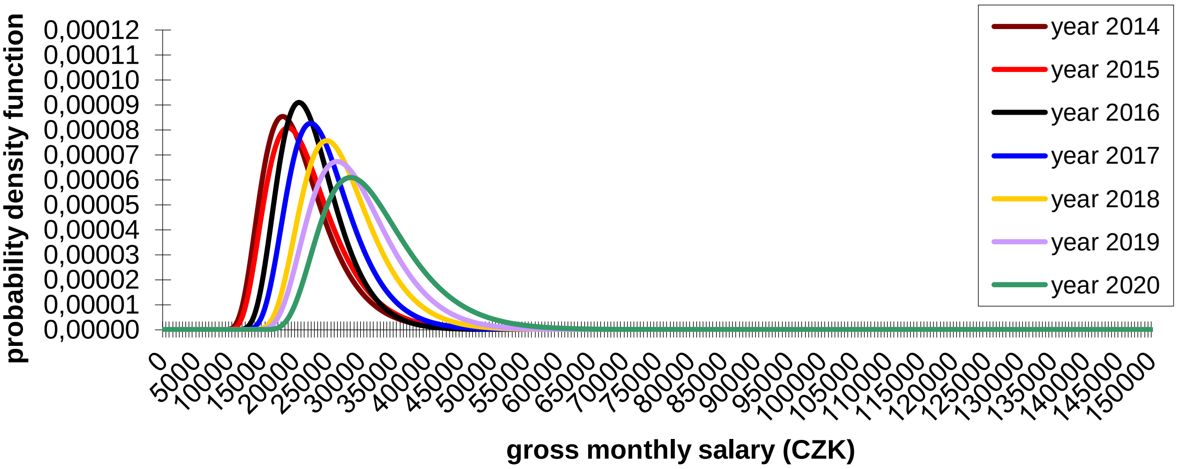

Figure 6. Development of the model distribution of the gross monthly salary of men in the period 2014–2020 for the category secondary education without A-level examination

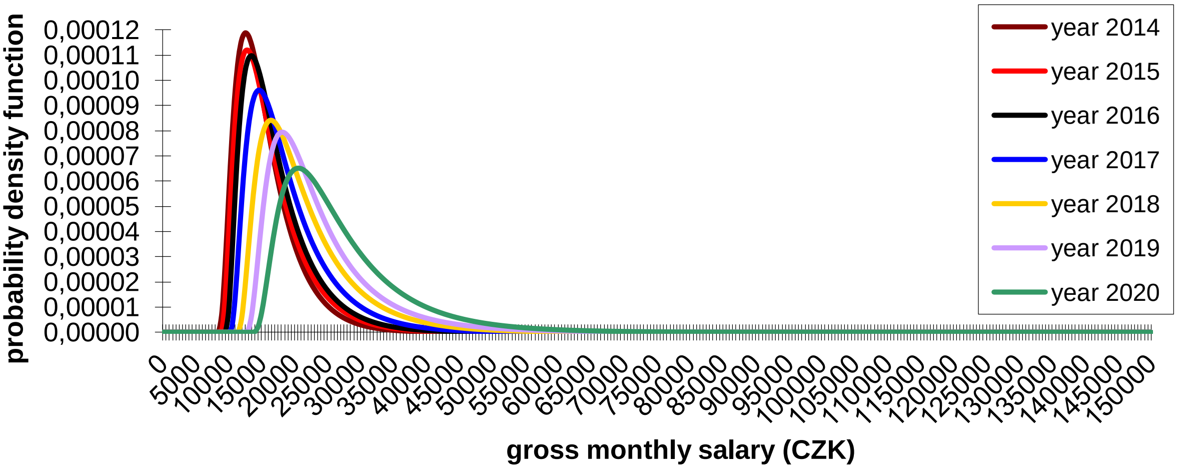

Figure 7. Development of the model distribution of the gross monthly salary of women in the period 2014–2020 for the category secondary education without A-level examination

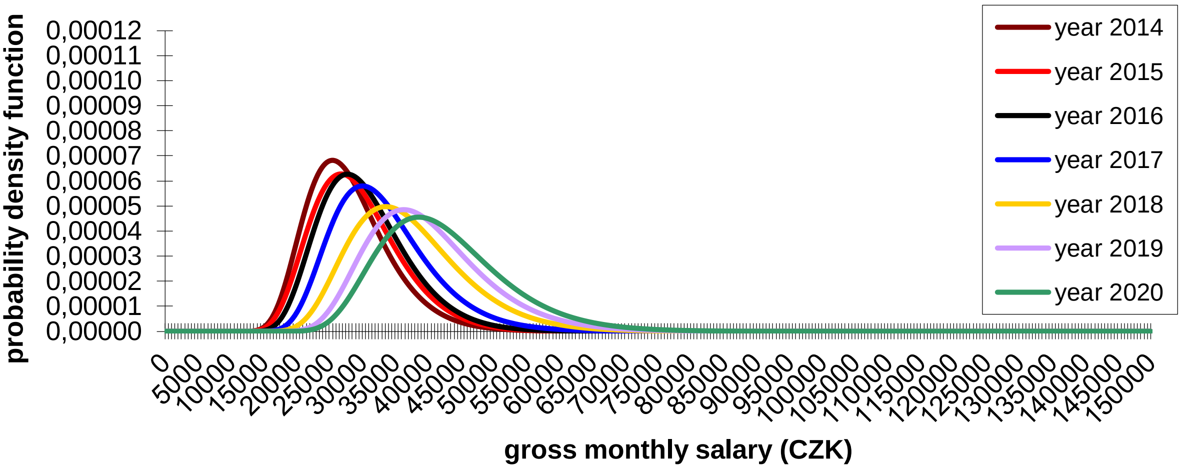

Figure 8. Development of the model distribution of the gross monthly salary of men in the period 2014–2020 for the category secondary education with A-level examination

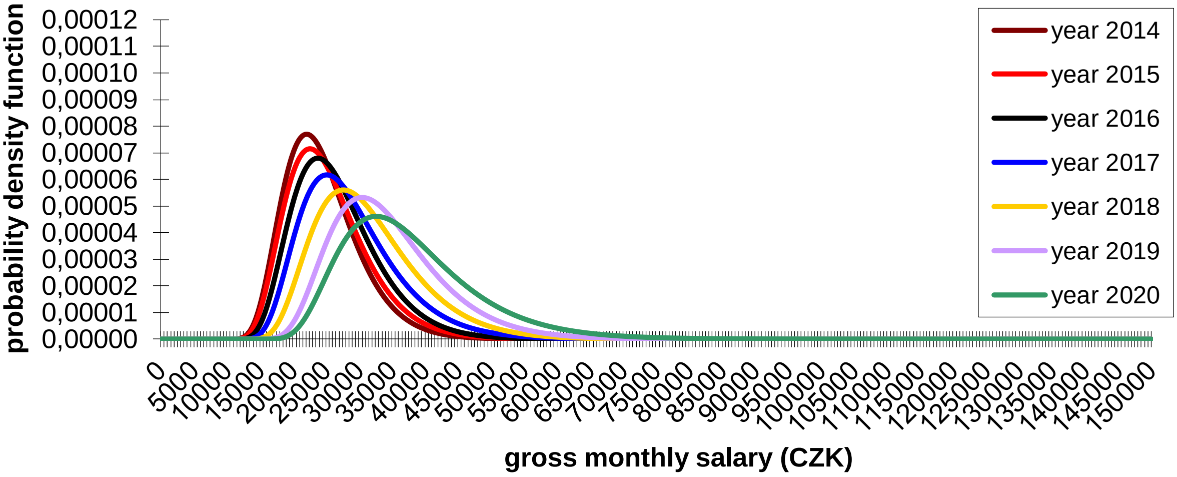

Figure 9. Development of the model distribution of the gross monthly salary of women in the period 2014–2020 for the category secondary education with A-level examination

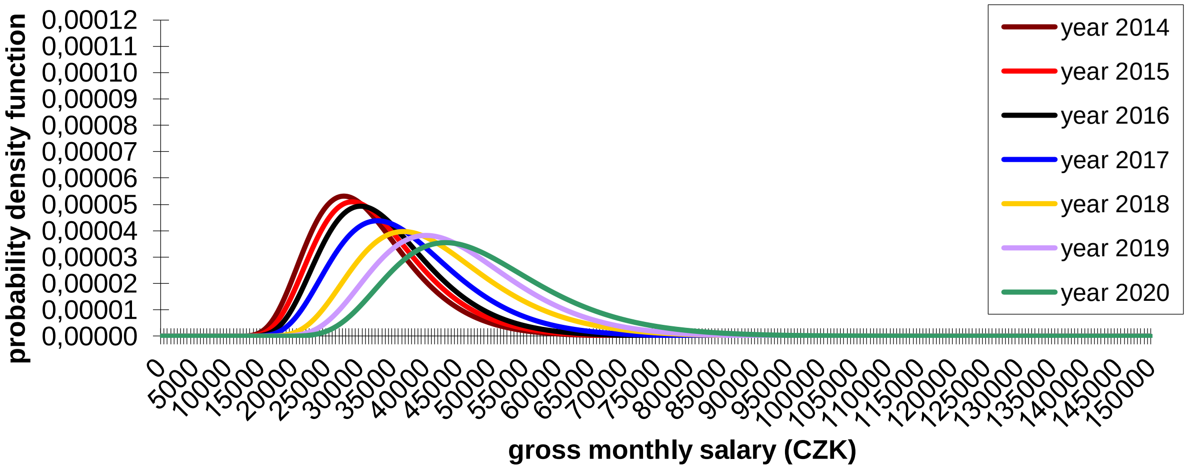

Figure 10. Development of the model distribution of the gross monthly salary of men in the period 2014–2020 for the category post-secondary non-tertiary and bachelor’s education

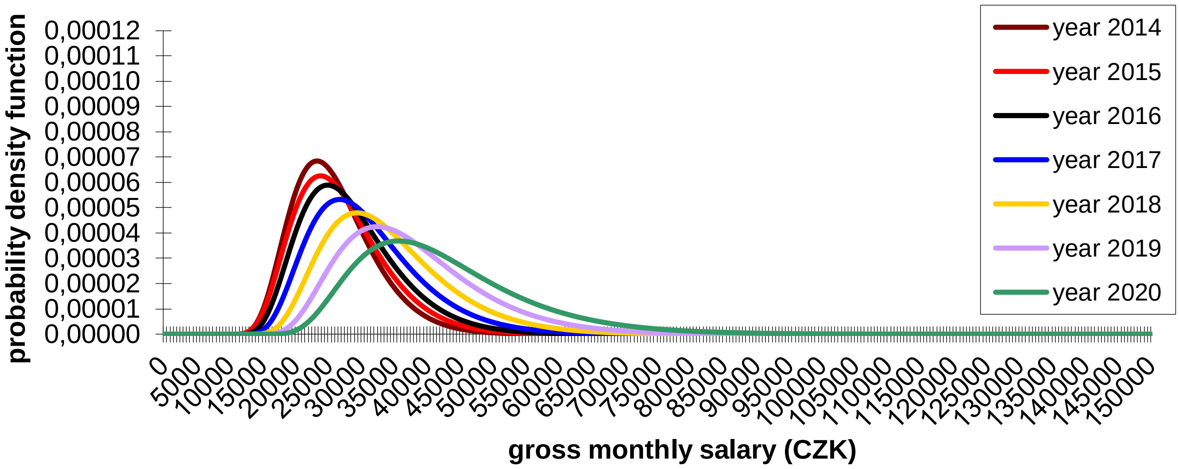

Figure 11. Development of the model distribution of the gross monthly salary of women in the period 2014–2020 for the category post-secondary non-tertiary and bachelor´s education

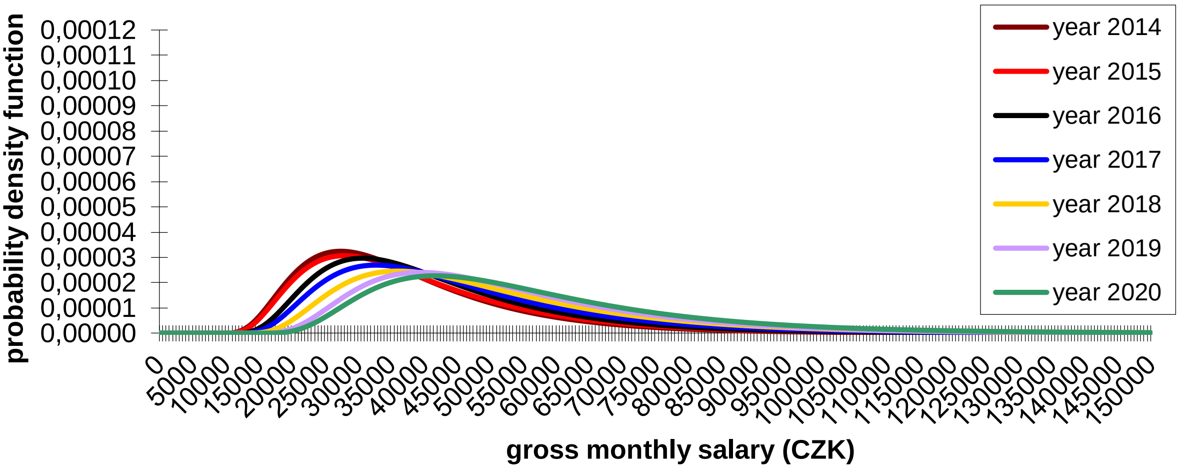

Figure 12. Development of the model distribution of the gross monthly salary of men in the period 2014–2020 for the category higher education

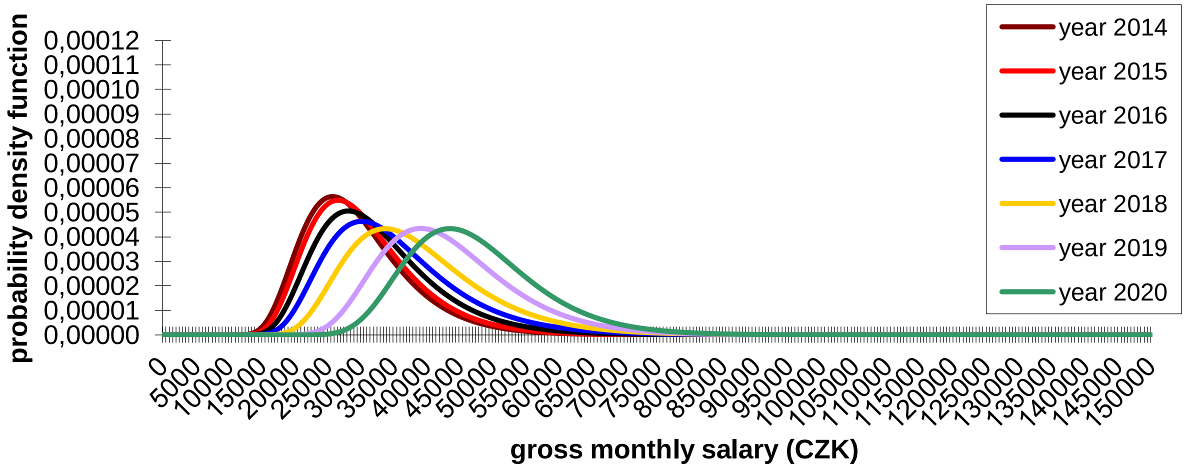

Figure 13. Development of the model distribution of the gross monthly salary of women in the period 2014–2020 for the category higher education

It was possible to construct models of the salary distribution of men and women in the period 2014–2020 according to the individual categories of educational attainment using four-parameter lognormal curves and quantile method of parameter estimation. It was also possible to identify the specific shapes of model curves of salary distribution for individual categories of educational attainment and capture their development over time. For future research, it will be interesting to compare the results using three-parameter curves.

Acknowledgments: This paper was subsidized by the funds of institutional support of a long-term conceptual advancement of science and research number IP400040 at the Faculty of Informatics and Statistics, University of Economics and Business, Czech Republic.

Conflict of interest: none.

References

Bílková, D. (2020). On Four-Parameter Lognormal Distribution and Application of Quantile Point Estimation Method: Application to Wage Distributions. In *13th International Scientific Conference on Reproduction of Human Capital - Mutual Links and Connection (RELIK 2020)* (pp. 11–21). Prague University of Economics and Business. https://relik.vse.cz/2020/download/pdf/325-Bilkova-Diana-paper.pdf

Bílková, D. (2019). Four-parameter Lognormal Curves in Wage Distribution Models: Comparison with Three-parameter Lognormal Curves. In *International Days of Statistics and Economics 2019*. https://doi.org/10.18267/pr.2019.los.186.14

Gentry, J. W. (1978). Applications of the four-parameter distribution to electrostatic charge and particle size distribution. *Powder Technology*, *20*(1), 115–126. https://doi.org/10.1016/0032-5910(78)80015-1

Lambert, J. A. (1970). Estimation of Parameters in the Four-Parameter Lognormal Distribution. *Australian Journal of Statistics*, *12*(1), 33–43. https://doi.org/10.1111/j.1467-842X.1970.tb00111.x

Mahmood, K. (1973). Lognormal Size Distribution of Particulate Matter. *Journal of Sedimentary Petrology*, 43(4), 1161–1166.

Malama, B., & Kuhlman, K. L. (2015). Unsaturated Hydraulic Conductivity Models Based on Truncated Lognormal Pore-Size Distributions. *Groundwater*, *53*(3), 498–502. https://doi.org/10.1111/gwat.12220

Regalado, C. M., & Ritter, A. (2009). A bimodal four-parameter lognormal linear model of soil water repellency persistence. *Hydrological Processes*, *23*(6), 881–892. https://doi.org/10.1002/hyp.7226

Saving, T. R. (1965). The Four-Parameter Lognormal, Diseconomies of Scale and the Size Distribution of Manufacturing Establishments. *International Economic Review*, *6*(1), 105–114. https://doi.org/10.2307/2525626

Siano, D. B., & Metzler, D. E. (1969). Band Shapes of the Electronic Spectra of Complex Molecules. *The Journal of Chemical Physics*, *51*(5), 1856–1861. https://doi.org/10.1063/1.1672270

Wagner, L. E., & Ding. D. (1994). Representing Aggregate Size Distributions as Modified Lognormal Distributions. *Transactions of the ASAE*, *37*(3), 815–821. https://doi.org/10.13031/2013.28145

Wingo, D. R. (1975). The use of interior penalty functions to overcome lognormal distribution parameter estimation anomalies. *Journal of Statistical Computation and Simulation*, *4*(1), 49–61. https://doi.org/10.1080/00949657508810109

Zeng, P., & Yu, X. (1991). Four-parameter model of PFDM on lognormal distribution. *Acta Aeronautica et Astronautica Sinica;(China)*, *12*. A69–A74.