Conditional Distributions of Incomes and their Characteristics

Ivana Malá
University of Economics in Prague
W.Churchill sq.4
Prague, Czech Republic
E-mail: malai@vse.cz

In the text conditional distributions of the incomes per capita of the Czech households in 2008 are analyzed. In a lot of papers ([1], [2]) households are divided into subgroups according to social or demographic characteristics as education, age or job of a head of a household or size of municipality. In this paper information of another type is of interest. We will suppose that we know whether the income per capita for a household is less or greater than a given quantity. Moreover in various questionnaires intervals are offered for a quantification of a level of income. If the probability distribution of income is chosen, it is easy to find conditional distributions given quantities mentioned above. In the first part of the text formulas for densities, cumulative distribution and quantile functions of conditional distributions are given for a random variable with continuous distribution. Two income distributions (three parametric lognormal and three parametric Dagum distributions, [3]) were used for the modelling of income distribution per capita of the Czech households. Maximum likelihood estimates of the parameters are taken from [4]. Basic formulas for these two distributions are introduced in the part two of this paper.

If the parameters of chosen probability distribution are estimated by maximum likelihood estimates, characteristics of the distribution (such as expected value, median or mode) could be evaluated with the use of known theoretical formulas and estimated values of parameters by substituting estimates into the theoretical formulas. These estimates have optimal limit properties of maximum likelihood estimates. But the choice of a suitable distribution is crucial, in case of the distribution that cannot fit data well, all conclusions are misleading. For a successful model of wages or incomes a flexible, skewed distribution with high variability is necessary. Two distributions that are used in this paper (lognormal and Dagum) are considered to be able to describe well such data. Lognormal distribution is used for incomes (or wages) in the Czech Republic in [1], [2] and [4], Dagum distribution in [4]. In [5] generalized lambda distribution is successfully used for incomes in the Czech Republic. According to Akaike`s criterion, the fit of Dagum distribution was found to be superior to lognormal distribution [4] for analysed years 2005-2008. For this paper only last year (2008) was selected. The appropriateness of a chosen distribution can also be analysed by comparing sample and theoretical (calculated with the use of estimated parameters) characteristics. In this paper only theoretical properties of both fits are studied. We can notice that characteristics of location and variability are similar (Table 1) and even conditional distributions and their characteristics of both distributions are similar, but of course not equal.

Methods

Suppose that X is a random variable with continuous distribution with the density function f, distribution function F and $100P\%$ quantile x_P (or quantile function F^{-1}). Suppose that $X > z$ for a given real z. Then the distribution of X given $X > z$ is described by the cumulative distribution function

$$F(x|X>z) = P(X \le x|X>z) = \begin{cases} 0 & x \le z, \\ \dfrac{P(z<X\le x)}{P(X>z)} = \dfrac{F(x)-F(z)}{1-F(z)} & x > z, \end{cases} \tag{1}$$

and from (1) we obtain the density function of the conditional distribution

$$f(x|X>z)=\frac{f(x)}{1-F(z)} \quad x>z, \qquad = 0 \quad x\le z. \tag{2}$$

On the other hand if we know that the income of a household is less (or equal to) than given *z*, we obtain for conditional density and distribution function formulas (3) and (4)

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

and

$$f\left(x|X\le z\right)=\frac{f\left(x\right)}{F\left(z\right)} \quad x<z, \qquad =0 \quad x\ge z. \tag{4}$$

The expected value of conditional distributions with densities (2) and (4) can be written as a function of *z* in the form

$$\frac{1}{1-F(z)}\int_z^\infty xf(x)dx \quad \left(resp.\frac{1}{F(z)}\int_0^z xf(x)dx\right). \tag{5}$$

From (1) a 100*P*% quantile of *X* given $X>z$ $x_{Iz,P}$ can be written as

$$x_{Iz,P}=F^{-1}\left(P(1-F(z))+F(z)\right)=F^{-1}\left(P-F(z)(1-P)\right), \quad 0<P<1. \tag{6}$$

Taking into account (3) we obtain by easy calculation 100*P*% quantile $x_{IIz,P}$ of *X* given $X<z$ as

$$x_{IIz,P}=F^{-1}\left(PF(z)\right), \quad 0<P<1. \tag{7}$$

Similar formulas can be found for the conditional distributions of *X* given $z_1<X\le z_2$, for real $z_1<z_2$ *:*

$$F(x|z_1<X\le z_2)= \begin{cases} 0 & x\le z_1, \\ \dfrac{F(x)-F(z_1)}{F(z_2)-F(z_1)} & z_1<x<z_2, \\ 1 & x\ge z_2, \end{cases} \tag{8}$$

$$f(x|z_1<X\le z_2)=\frac{f(x)}{F(z_2)-F(z_1)} \quad z_1<x<z_2, \qquad = 0 \quad \text{otherwise}, \tag{9}$$

and

$$x_{IIIz_1z_2,P}=F^{-1}\left(P(F(z_2)-F(z_1))+F(z_1)\right)=F^{-1}\left(PF(z_2)+\left(F(z_1)-PF(z_1)\right)\right), \quad 0<P<1. \tag{10}$$

From formulas (8) - (10) previous results for conditions $X>z$ and $X\le z$ can be derived as special cases for obvious choices of limits z_1 or z_2.

In the text two probability distributions frequently used for the modelling of income distributions are employed. Three parametric lognormal distribution with positive parameter σ^2 and real parameters μ and θ can be described by the density function

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma(x-\theta)}exp\left(-\frac{(ln(x-\theta)-\mu)^2}{2\sigma^2}\right) \quad x>\theta, \qquad = \quad 0 \quad x\le\theta, \tag{11}$$

and 100*P*% quantile

$$x_{LN,P} = \theta + exp\left(\mu + \sigma u_P\right) \quad 0 < P < 1, \tag{12}$$

where u_P is a 100*P* percent quantile of a standard normal distribution. Dagum distribution has a probability density function

$$\begin{aligned} f(x) &= \frac{\alpha p y^{\alpha p-1}}{\left(\beta^{\alpha p}(1+(x/\beta)^{\alpha}\right)^{p+1}} \quad x > 0, \\ &= \quad 0 \qquad\qquad\qquad\quad x \le 0, \end{aligned} \tag{13}$$

and quantile function

$$x_{\text{Dagum},P} = \sqrt[\alpha]{\frac{\beta^{\alpha}}{\sqrt[p]{P^{-1}} - 1}}. \tag{14}$$

Both distributions are unimodal with the modes given by the formulas

$$x_{\text{LN,mode}} = \theta + e^{\mu-\sigma^2} \quad \text{and} \quad x_{\text{Dagum,mode}} = \beta\left(\frac{\alpha p - 1}{\alpha + 1}\right)^{1/\alpha}. \tag{15}$$

These formulas can be used in searching for the mode of a conditional distribution. In both formulas (2) and (4) the original density is divided by a constant. We can derive that if an original mode is included in the interval of the condition, the original mode is equal to the mode of the conditional distribution. If it is not true, the mode of the conditional distribution coincides with the value *z*. Expected values of the conditional distributions were evaluated from the definition, all integrals were computed numerically.
All computations were performed in R or Excel.

Results

Conditional distributions are evaluated for the net annual income per capita of the Czech households in 2008 in CZK (in 2008 approximately 25CZK/EUR). Data from the Statistics on Income and Living Conditions survey (EU- SILC) organized by the Czech Statistical Office [5] was used. In the dataset 11,294 households are included in 2008 and the income is given as a ratio of the total net annual income and the number of members of the household. In [4] three parametric lognormal and three parametric Dagum distributions were fitted into these data. Maximum likelihood estimates of parameters are given in the Table 1 together with the estimated characteristics of the location and variability. Quartile deviation is taken as a half of an interquartile range.

Table 1: Estimated parameters, estimates of characteristics of the level and variability of fitted distributions (CZK)

Distribution	Lognormal (3 parameters)			Dagum (3 parameters)		
parameter	μ	σ	θ	α	θ	p
estimate	11.703	0.421	-171.167	4.330	113,878.9	1.159
***E*(*X*)**	131,969			130,540		
mode	101,120			106,687		
median	120,762			119,267		
$\sqrt{D(X)}$	58,191			59,808		
quartile deviation	34,807			29,347		

In the table very similar values of characteristics can be seen. Lognormal distribution has greater value of expected value and median, the mode is less for this distribution than for Dagum distribution. Moment characteristic of variability (standard deviation) is greater for Dagum distribution, quantile characteristic (quartile deviation) for lognormal distribution.

In the Figures 1 and 2 characteristics of the location of conditional distributions are given as a function of a condition *z*. Left figure illustrates characteristics of location (level of incomes) for lognormal distribution, right figure for Dagum distribution, in the horizontal axis values of the condition *z* are given. The scale of both figures is the same and results can be compared for both distributions. In the Figure 1 characteristics for the condition $X \leq z$ are shown. Curves on both parts of the figure (left and right) seem to be similar and they increase to the unconditional characteristics of *X* with high *z* (as it was expected). The value of mode increases as a line with the slope one to the mode of the original distribution and then it is constant and equal to this mode. Conditional median and expected value approach unconditional values more slowly (approximately 500,000 CZK). In the Figure 2 the results for condition $X > z$ is shown. Behavior of characteristics is more complicated. It is obvious that all values should increase from unconditional values with increasing quantity *z*. The course of conditional mode is again by part linear, for the expectation and median the situation is different. We can notice different slopes of lines for both distributions, almost parallel for lognormal distribution only.

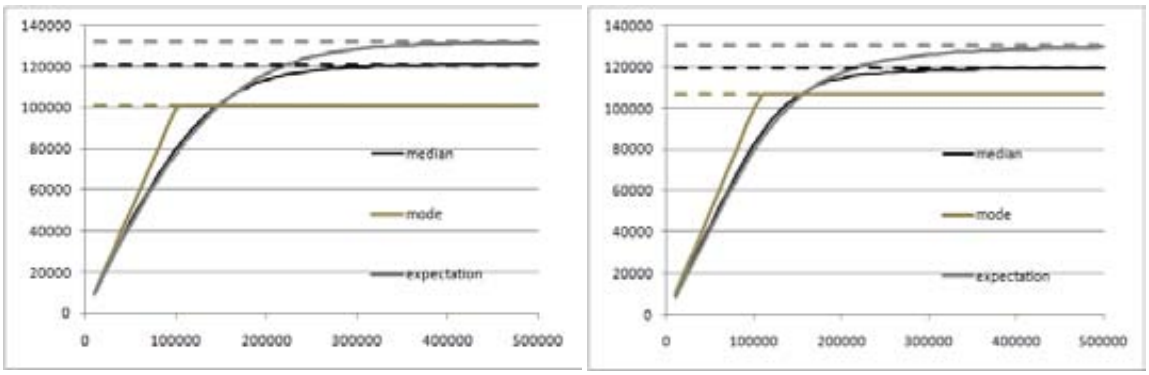

Figure 1: Characteristics of location (expected value, median, mode in CZK) of original distribution (dotted lines) and conditional distributions given $X \leq z$ (solid lines) for conditions z on the horizontal axis. (lognormal distribution left, Dagum distribution right)

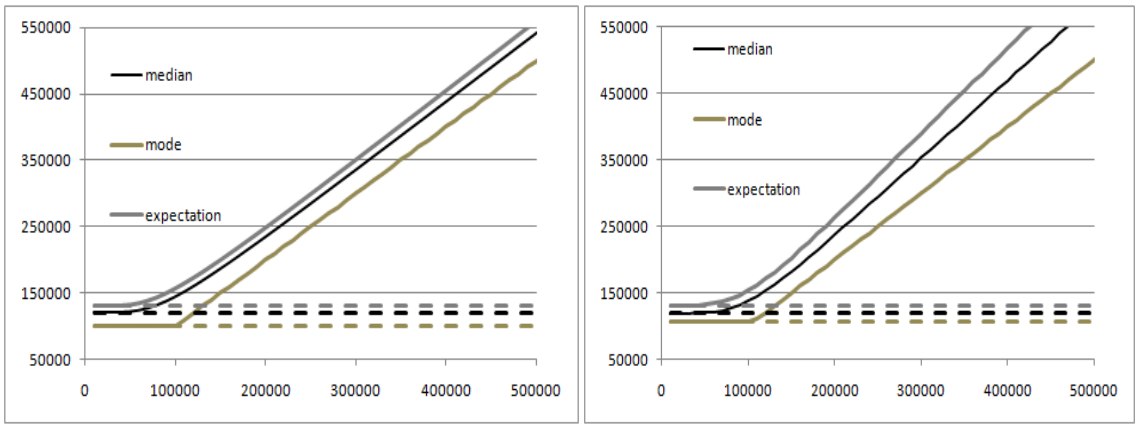

Figure 2: Characteristics of location (expected value, median, mode in CZK) of original distribution (dotted lines) and conditional distributions given $X > z$ (solid lines) for conditions z on the horizontal axis, (lognormal distribution left, Dagum distribution right)

In the Figures 3 estimated densities of the unconditional distributions (grey solid lines) and conditional densities for two conditions $X \leq z$ (z = 100,000 and z = 200,000 CZK) are shown. We can notice lower value of the density in the mode for original lognormal than for Dagum distribution. The same relation occurs for the conditional density for z = 100,000. This value is less than modes of distributions (Table 1), the density is increasing for incomes less than 100,000 (and 0 for greater incomes). The value 200,000 was chosen greater than both modes, the conditional density is unimodal (conditional and unconditional modes are equal). For high z, conditional density is similar to unconditional density. In the Figure 4 conditional probability densities for z = 100,000, 200,000, 300,000 and 400,000 CZK are shown for the condition $X > z$. The curves move with increasing z from the left (unconditional density or X > 100,000 CZK) to the right (X > 400,000 CZK).

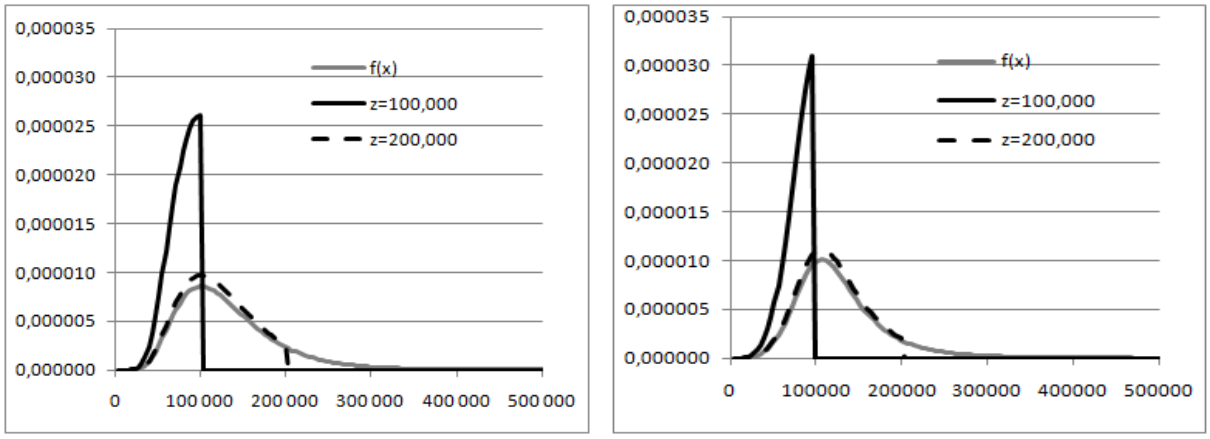

Figure* 3*: Probability density of original distribution (lognormal distribution left, Dagum distribution right) and densities of conditional distributions given $X < z$ for z = 100,000 and 200,00 CZK

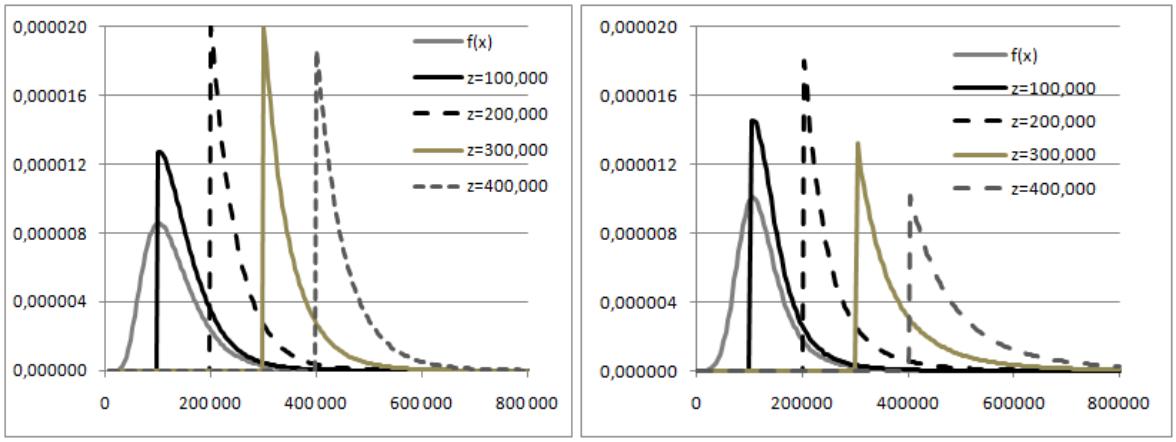

Figure* 4*: Probability density of original distribution (lognormal distribution left, Dagum distribution right) and densities of conditional distributions given $X > z$ for z = 100,000, 200,000, 300,000, 400,000 CZK

Conclusions

In the paper conditional probability distributions of a continuous random variable under the condition of the type $X > z$, $X \leq z$ (for given z) and $z_1 < X \leq z_2$ are studied. This situation can be frequently met in an analysis of incomes (or wages), it means that income is not known exactly, but we only know that its value is less or greater than a given amount (or is included in a given interval). With the use of simple computation conditional distributions can be described and the impact of the side information can be evaluated.

REFERENCES

[1] Bartošová, J., Bína, V. Modelling of Income Distribution of Czech Households in Years 1996 – 2005. *Acta Oeconomica Pragensia*, Vol. 17, pp. 3 − 18. ISSN 0572-3043. 2009

[2] Bílková, D. Application of Lognormal Curves in Modeling of Wage Distributions. *Journal of Applied Mathematics*, Vol. 1, Iss. 2, pp. 341 − 352. ISSN 1337-6365. 2008.

[3] Kleiber, C., Kotz, S. *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley-Interscience, New York. ISBN 0-471-15069-9. 2003.

[4] Malá, I. Distribution of Incomes per Capita of the Czech Households from 2005 to 2008. *In CD: Proceedings of 10th International Conference Aplimat 2011, Bratislava, Slovak Republic*, pp. 1583-1588. ISBN 978-80-89313-51-8. 2011.

[5] Pacáková, V., Sipková, L. Generalized Lambda Distributions of Household's Incomes. *E + M Ekonomie a Management*, roč. X, č. 1, s. 98 – 107. ISSN 1212-3609. 2007.

[6] www.czso.cz