Robust estimators for Pareto and lognormal distributions in income inequalities evaluation

Daniel Kosiorowski[1], Damian Tracz[2]

Abstract

Considerations related to a nature of allocation of wealth within a populations have a central position in the economics and public debate related to social justice and social solidarity. In this paper we show selected aspects of robust estimation of the income distribution. We focus our attention on two well-known models for the income distribution namely on the Pareto and lognormal distributions and on popular income inequality measures namely on the Lorentz curve and the Gini coefficient. The presented arguments however are applicable to a wide class of over 100 models used for income distributions modelling which are by default estimated by means of maximal likelihood methodology. Our findings can be applied to the robust financial modelling as well. Theoretical considerations are illustrated by means of empirical examples.

Keywords: *Pareto distribution, log-normal distribution, Gini coefficient, robust estimation*

JEL Classification: C13, C14, D60

1. Introduction

A modeling of income and wealth distributions within populations originated over 100 years ago but is still in the spotlight of economists, politicians and social researchers. Proper knowledge about the structure of income in a particular country should imply a suitable taxation system and government aid programs. The debate on the relationship of economic growth, income distribution and social welfare has been intensified during first half of the twentieth century. When based on socialism and communism doctrines Eastern European countries - lead by Russia – carried out economic reforms intended to significantly reduce (or even eliminate) polarization between poor and rich citizens. According to this every member of society should have the same wealth and live in a country with exact same chances for everyone. Modified versions of socialism were also tried in some other countries all over the world, especially China and Scandinavian countries are worth mentioning because their assumption related to perfect income equality was relaxed. In 1955 Simon Kuznets in his article titled "Economic growth and income inequality" tried to deal with the relationship of income inequality to the economic growth of countries in a different stages of development.

[1] Corresponding author: Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: dkosioro@uek.krakow.pl.

[2] Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: damian.tracz@gmail.com.

He concluded that the underdeveloped countries income inequality was strong enough in the beginning, stabilized thereafter and reduced as the country went from developing to developed (Kuznets, 1955). This hypothesis was called “inverted U-shaped pattern of income inequality”. Subsequently many of important studies were made, some of them confirmed Kuznets findings and few criticized it for poor quality of used data and questionable methodology. Nevertheless as post-crisis (2009-2015) economic world still struggles with low economic growth problems, it is very important to run politics in order to find the optimum point of wealth inequality. Improper estimation of wealth distribution could lead to the conclusion that inequalities are too high and trigger some corrective action like raising taxes in high income bracket, and therefore smothering productivity and investment activities among well-paid citizens. On the other hand to liberal taxation system and insufficient public aid programs could lead to widening gap between low and well-paid people, that next can be a reason of social unrest or even rebellion.

The main objective of this paper is to present selected aspects of robust estimation of Pareto and log-normal distribution, based on solutions proposed by Brazauskas and Serfling (2000, 2001, 2004) and compare their results with maximal likelihood estimators that in fact are highly efficient but could be negatively affected by outliers. It is worth noting Economic Size Distributions including models with high probability in the upper tail like above mentioned Pareto and lognormal, as well as gamma distributions are appropriate in dealing with not only income/wealth topics but have been also successfully used in actuarial assumptions, risk management, city size analysis and file size distribution on the Internet (see Kleiber and Kotz, 2003). Moreover in this paper we focus our attention on popular and well-recognized income inequality measures namely Gini coefficient and the Lorenz curve. Theoretical considerations consisted in the paper are illustrated by some empirical examples, that mainly come from Canada income data. The rest of the paper is organized as follows: In Section 2, the basic properties of the lognormal distribution are presented. In Section 3 we describe Pareto distribution properties, Section 3 is the main part of the paper and here we present results of lognormal and Pareto parameters robust estimation. Paper ends with conclusions and selected references.

2. Lognormal distribution

In 1931 French economist and engineer R. Gibrat developed a commonly used lognormal model. Gibrat asserted that the income of an individual (or the size of a firm) may be considered the joint effect of a large number of mutually independent causes that have worked during a long period of time. It is well-known that there is a close relationship between normal and lognormal distributions. A random variable Y has a lognormal distribution LN(μ, σ) if X = log Y has the normal distribution N(μ, σ^2), where μ and σ^2 are the mean and variance of the underlying normal variable X, but become respectively the shape and scale parameters of the lognormal variable Y.

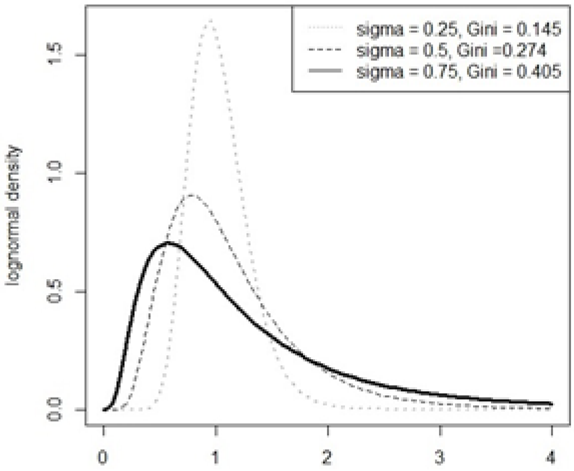

Fig. 1. The lognormal distribution density functions for a different scale parameters.

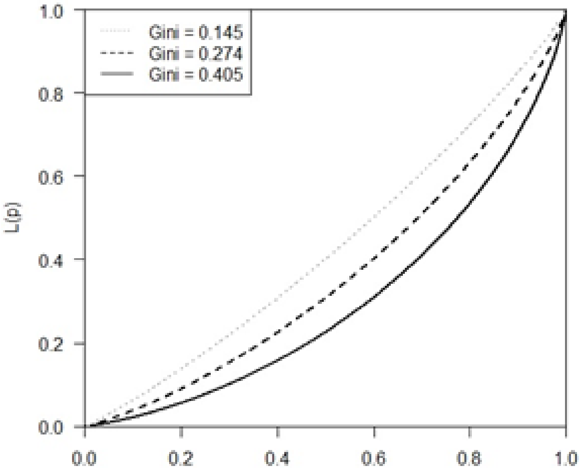

Fig. 2. Lorenz curve shapes for a different values of Gini coefficient.

In 1945 famous polish economist M. Kalecki applied lognormal distribution for United Kingdom personal incomes during 1938-1939 and revealed that lognormal distribution fits well to data only when certain part of the data was omitted. Therefore he extended LN(μ, σ) two parameter lognormal model to LN(τ, μ, σ) three parameter lognormal distribution of Y = τ + e^x where τ represent threshold value and X is a random variable with mean μ and standard deviation σ (the same as in two parameter version). When parameter $\theta = (\tau, \mu, \sigma)$ then probability density function is defined as follows:

$$f(y;\theta) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}(y-\tau)} \exp\left\{ -\dfrac{[\log(y-\tau)-\mu]^2}{2\sigma^2} \right\} & \tau < y < \infty, \sigma > 0, -\infty < \mu < \infty \\ 0 & \textit{otherwise} \end{cases} . \quad (1)$$

As estimation of three parameter lognormal distribution $LN(\tau, \mu, \sigma)$ is a problematic and computationally extensive (see Serfling, 2002). We focus our attention on estimation of two parameter model and well-known MLE estimators of the location μ, scale σ parameters and expected value:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \log Y_i, \tag{2}$$

$$\hat{\sigma}_{ML} = \left(\frac{1}{n}\sum_{i=1}^{n} \left(\log Y_i - \hat{\mu}_{ML} \right)^2 \right)^{\frac{1}{2}}, \tag{3}$$

$$\hat{E}(Y) = \exp\left\{ \hat{\mu}_{ML} + \frac{\hat{\sigma}^2_{ML}}{2} \right\}. \tag{4}$$

3. Pareto distribution

Rule 80-20 claims that roughly 80% of the effects come from 20% of the causes. This rule has been attributed to Italian scientist V. Pareto, who in nineteenth century studied the economic agents income data (reported for tax purposes) and his research became a pillar of statistical income distributions. He found a regularity of observed income distribution sourced from tax records - a stable linear relation of the form log $N(x) = A$ - α log x where $x \geq x_m > 0$, $\alpha > 1$ and $N(x)$ is the number of economic units with income $X > x$, where X denotes the income variable with range $[x_m, \infty)$. The Pareto type I model is the solution of that linear relationship.

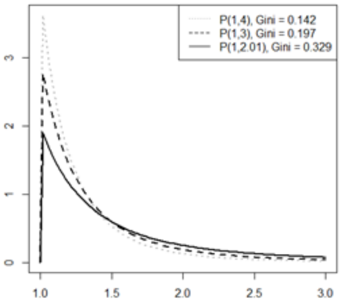

Fig. 3. Pareto distribution density functions for different shape parameters.

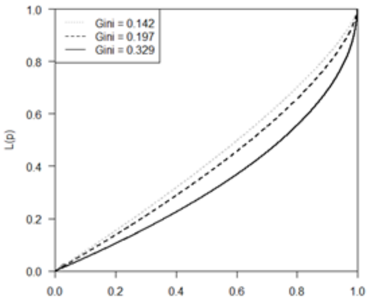

Fig. 4. Lorenz curve shapes for a different values of Gini coefficient.

The probability density function of the Pareto distribution is defined as follows:

$$f(x) = \begin{cases} \dfrac{\alpha x_m^{\alpha}}{x^{\alpha+1}} & x \geq x_m \\ 0 & x < x_m \end{cases}. \tag{5}$$

In estimation of Pareto α shape MLE estimator attains the minimum possible variance among a large class of competing estimators:

$$\hat{\alpha}_{ML} = \frac{n}{\sum_{i=1}^{n} \log(X_i / x_m)}. \tag{6}$$

Expected value in Pareto distribution is defined as:

$$E(X) = \begin{cases} \infty & \alpha \le 1 \\ \dfrac{\alpha x_m}{\alpha - 1} & \alpha > 1 \end{cases}, \tag{7}$$

and variance is defined as below:

$$D^2(X) = \begin{cases} \infty & \alpha \in (1,2] \\ \dfrac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \alpha > 2 \end{cases}, \tag{8}$$

for $\alpha \le 1$ variance does not exist.

4. Robust estimation of lognormal and Pareto distribution and properties of applied estimators

In parametric modeling of income distribution MLE estimators are most commonly used due to their high efficiency. At the same time these estimators are not robust, it means that are highly influenced by outliers in the upper/lower tail of the income distribution. Small relative errors in estimated model parameters could be a reason of a large relative errors in estimated quantiles or tail probabilities. Subsequently it leads to drawing improper conclusions about income inequalities and Gini coefficients. Measuring and comparing level of estimator robustness can be performed based on the Influence Function (IF) and the finite sample breakdown point (BP). The BP can be used as a measure of global robustness, while the IF captures local robustness of the estimator (for more details see Kosiorowski and Tracz, 2014).

Estimation of lognormal and Pareto model parameters on a robust and high efficient manner was extensively studied by Brazauskas and Serfling. They introduced and provide deep analysis of Generalized Median and Trimmed Mean estimators in above mentioned models. The **generalized median (GM)** statistics are defined by taking median of the $\binom{n}{k}$ evaluations of a given kernel $h(x_1,...,x_k)$ over all $k-$ sets of the data. In lognormal model GM estimator for location parameter is then calculated as follows:

$$\hat{\mu}_{GM}(k) = MED\left\{h\left(X_{i1},...,X_{ik}\right)\right\}, \tag{9}$$

and kernel for the location parameter estimator is then:

$$h(x_1,...,x_k)=\frac{1}{k}\sum_{i=1}^{k}\log x_i \ , \tag{10}$$

subsequently GM estimator for scale parameter is defined as:

$$\overset{\wedge}{\sigma}_{GM}(m)=\left(MED\left\{h\left(X_{i1},...,X_{im}\right)\right\}\right)^{1/2}, \tag{11}$$

with a respectively defined kernel:

$$h_2(x_1,...,x_m)=\frac{1}{mM_{m-1}}\sum_{1\le i\le j\le m}(\log x_i-\log x_j)^2 \tag{12}$$

where M_{m-1} denotes median of chi-square distribution with m-1 degrees of freedom. It was proved that kernel-type quantile estimator has a limiting normal distribution (see Veraverbeke, 1987). Moreover both estimators are characterized by smooth and bounded IF and $BP=1-\frac{1}{2}^{1/p}$ where p equals k variable for the μ location parameter and m variable for the σ scale parameter. Proposed GM estimator for the parameter α in the Pareto model in case of x_m known is defined as:

$$\hat{\alpha}_{GM}=MED\left\{h\left(X_{i1},...,X_{ik}\right)\right\}, \tag{13}$$

with a respectively defined kernel:

$$h(x_1,...,x_k;x_m)=\frac{1}{C_k}\frac{k}{\sum_{j=1}^{k}\log\left(x_j/x_m\right)} \tag{14}$$

where c_k is a multiplicative, the median – unbiasing factor i.e. chosen so that the distribution of $h(x_1,...,x_k;x_m)$ has median α - values of c_k for $k=2$, $c_2=1.1916$, $k=3$ $c_3=1.1219$.

The trimmed mean is formed by discarding the population β_1 lowest obs. and the proportion of β_2 uppermost obs., where β_1 and β_2 satisfying $0\le\beta_1,\beta_2<1/2$, and averaging the remaining ones. In particular, for estimating α, with known x_m Brasauskas and Serfling proposed following version of the estimator:

$$\hat{\alpha}_T=\left(\sum_{i=1}^{n}c_{ni}\log\left(X_{(i)}/x_m\right)\right)^{-1} \tag{15}$$

with $c_{ni}=0$ for $1\le i\le[n\beta_i]$, $c_{ni}=0$ for $n-[n\beta_2]+1\le i\le n$ and $c_{ni}=1/d(\beta_1,\beta_2,n)$ for $[n\beta_1]+1\le i\le n-[n\beta_1]$, where $[\cdot]$ denotes "greatest integer part" and $d\left(\beta_1,\beta_2,n\right)=\sum_{j=[n\beta_1]+1}^{n-[n\beta_2]}\sum_{i=0}^{j-1}(n-i)^{-1}$.

In order to investigate properties of the considered estimators MLE, GM and TM we performed intensive simulation studies involving datasets of size 500 observations sourced from the below described mixtures of distributions:

Lognormal model evaluation:

1. Mixture of LN(3,6) x 10% and LN(2,3) x 90%.
2. Mixture of normal distribution N(100,25) x 10% and LN(2.14,1) x 90%.
3. Mixture of gamma distribution G(10, 4) x 10% and LN(2.14,1) x 90%.

Pareto model evaluation:

1. Mixture of P(1,5) x 10% and P(10,5) x 90%.
2. Mixture of lognormal distribution LN(2.14,1) x 10% and P(7,2) x 90%.
3. Mixture of normal distribution N(3300, 500) x 10% and P(2500,4) x 90%.

Our output sets contain small relative share of distorted data but in fact mixtures like this can be a reason of high errors in estimation of parameters. In lognormal evaluation, notation of GM(k,m) was used.

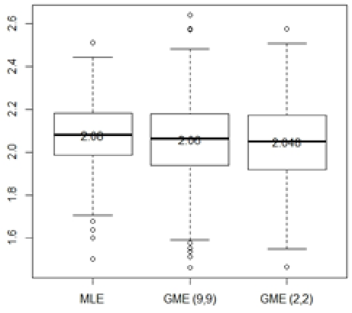

Fig. 5. Comparison of estimators for μ parameter in the first mixture.

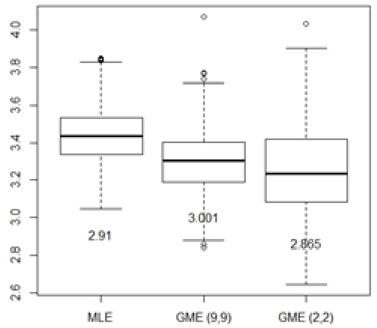

Fig. 6. Comparison of estimators for σ parameter in the first mixture.

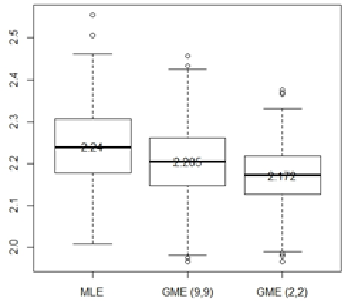

Fig. 7. Comparison of estimators for μ parameter in the second mixture.

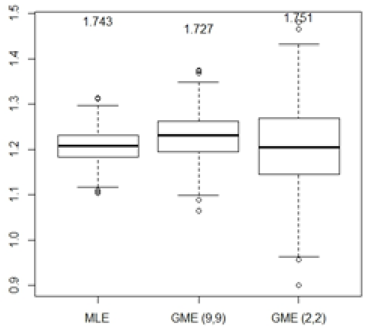

Fig. 8. Comparison of estimators for σ parameter in second mixture.

Based on Fig. 5-8 we can assert that estimation of u parameter using either MLE or robust GM(9,9) or GM(2,2) gives comparable results. Looking at the σ parameter evaluation it is easy to observe that all estimators are characterized by relatively high skewness and there is a visible difference between GM(9,9) and GM(2,2) in terms of efficiency.

In Pareto parameters evaluation we considered situations in which the x_m parameters was estimated as minimal value in a sample. However based on our previous research (see Kosiorowski and Tracz, 2014) it is worth considering the cases in which x_m is estimated x_m as quantile of order $\gamma \in (0, 0.3)$ (where γ parameter was optimizes with respect to a value of the standard Kolmogorov – Smirnov goodness of fit statistics).

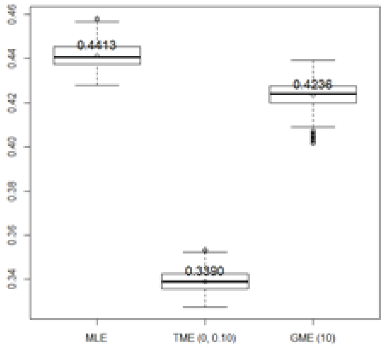

Fig. 9. Comparison of estimators for the first mixture and x_m taken as minimum.

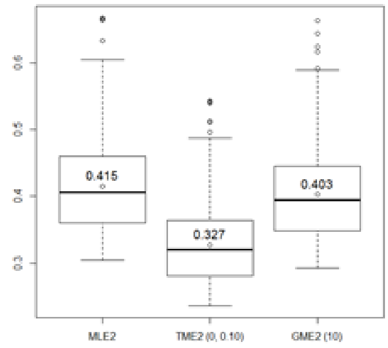

Fig. 10. Comparison of estimators for the second mixture and x_m taken as minimum.

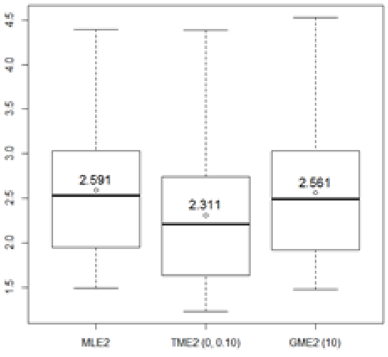

Fig. 11. Comparison of estimators for the third mixture and x_m taken as minimum.

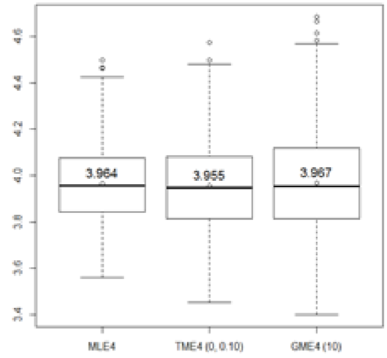

Fig. 12. Comparison of estimators for the clean data P(2500,4) and x_m taken as min.

Fig. 9-12 show comparison of estimators with respect to their dispersion. We can say that the robust estimators exhibit comparable properties to the MLE estimator for practical purposes For the TM and GM we observed bounded IF, the GM outperforms the TM however.

The estimated IF for MLE and Gini coefficient are unbounded and hence the estimators are sensitive to outliers. In a context of conducting a social politics basing on the estimated probability distribution of the income, we studied empirical example of TOTAL INCOME in Canada (2001) using census data from MINESSOTA POPULATION CENTER (https://international.ipums.org/international/).

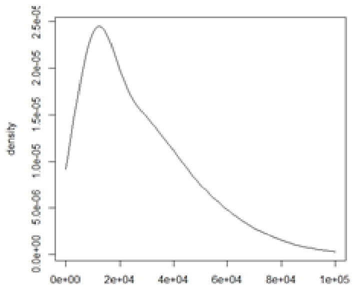

Fig. 13. Canada (2001) income distribution based on a near 600k sample size.

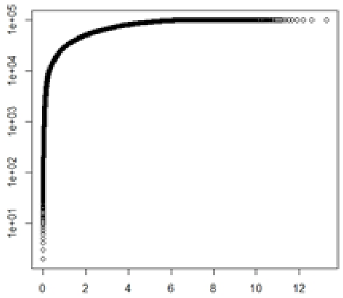

Fig. 14. Canada (2001) Q-Q Pareto chart.

Fig. 14 shows how important proper estimation of x_m in Pareto model is. If the data follows a Pareto model, observations on a Q-Q chart should form almost a straight line. The leftmost point of the fitted line can be used as an estimate of the threshold. In our estimation of Pareto and lognormal parameters we however used x_m taken as a minimum in order to enable better comparison of two models that are under investigation. Nonparametric Gini coefficient for data is 0.43 and positions Canada above the middle of world income inequalities ranking.

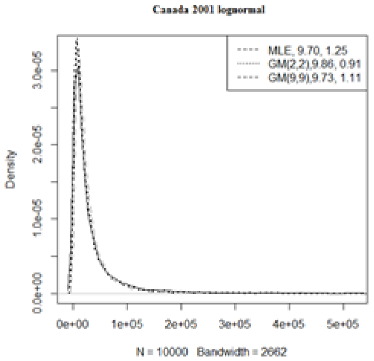

Fig. 15. Results of lognormal parameters estimation for Canada (2001).

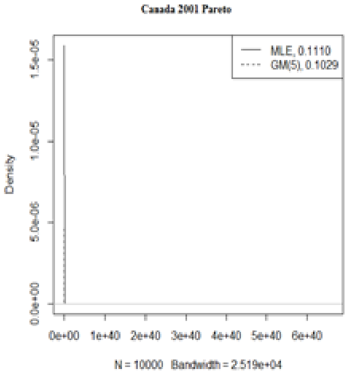

Fig. 16. Results of Pareto parameter estimation for Canada (2001).

Conclusion

The conducted simulations and an inspection of the empirical example lead us to the following conclusions. The MLE, TM, GM estimators crucially depend on the estimate of the scale parameter x_m in the Pareto model. Assuming that this parameter is known, we recommend the GM estimator for k = 5-7, which is a compromise between a need of high efficiency and high robustness. The GM(k,m) estimator in a lognormal model for k and m from the range 5-7 constituted to be a reasonable choice between highly efficient GM(9,9) and highly robust GM(2,2). The MLE, TM, GM estimators in Pareto and MLE and GM estimators in lognormal model strongly depend on the distributional assumption. Before estimation certain diagnostic procedure inspecting a sample should be performed. We recommend an usage of simple Q-Q plot based procedures. Moreover non-parametric method of estimation income distribution like a one based on local linear polynomial estimator can be perceived as complementary source of information.

References

Brazauskas, V., & Serfling, R. (2000). Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial Journal*, *4*(4), 12-27.

Brazauskas, V., & Serfling, R. (2000). Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. *Extremes*, *3*(3), 231-249.

Brazauskas, V., & Serfling, R. (2003). Favorable estimators for fitting Pareto models: A study using goodness-of-fit measures with actual data. *Astin Bulletin*, *33*(2), 365-382.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*, *470*. John Wiley & Sons.

Kosiorowski, D., & Tracz, D. (2014). On Robust Estimation of Pareto Models and Its Consequences for Government Aid Programs Evaluation. In: Lula, P., Rojek, T. (eds.), *KNOWLEDGE ECONOMY SOCIETY. Contemporary tools of organizational resources management*, Cracow: Foundation of the Cracow University of Economics, 253-266.

Kuznets, S. (1955). Economic growth and income inequality. *The American economic review*, 1-28.

Serfling, R. (2002). Efficient and robust fitting of lognormal distributions. *North American Actuarial Journal*, *6*(4), 95-109.

Veraverbeke, N. (1987). A kernel-type estimator for generalized quantiles. *Statistics & probability letters*, *5*(3), 175-180.