ESTIMATES OF QUANTILE CHARACTERISTICS BASED ON RIGHT CENSORED DATA

Ivana Malá – Václav Sládek

Abstract

In survival analysis, we usually apply positively skewed, heavy-tailed probability distributions and the analysed datasets contain not only complete but also censored values. In the contribution, random samples with right censored data are treated and properties of quick estimators of central tendency – weighted means of selected quantiles – are of interest. We use both parametric, under an assumption of the probability distribution, and non-parametric distribution-free method based on Kaplan-Meier estimator of a survival function (for data with censored values) and linearized inverse empirical function and kernel estimator of the density for complete data. The Monte Carlo simulation study is performed in the R program. The impact of censoring is of interest, we use datasets with all values complete, and with 10 to 50 per cent of censored values. The dependence on the sample size is discussed; samples with 50, 100, and 500 observations are used. Random samples are generated from the lognormal distribution, independent right censoring is used.

Keywords: right censored data, Monte Carlo simulation, robust moments

JEL Code: C24, C63, C14

Introduction

There are basically three main possible approaches in statistics: classic parametric, nonparametric and Bayesian. In this contribution, we compare the parametric and non-parametric approach to the estimation of sample quantiles from datasets that includes the right (independently) censored values. The frequently used characteristics of the level are the mean and the median. In the survival analysis, we usually use highly skewed probability distributions to model the data and for this reason, we prefer quantile characteristics to the moment characteristics. Robust moments such as the L-moments, TL-moments or LQ-moments (Hosking, 1990, Mudholkar & Hutson, 1998, Šimková & Picek, 2017) are considered robust because they are less affected by outliers or sensitive to the contamination of the data than classical moments. The L-moments use expected values of order statistics, in the LQ moments, the expected values are substituted by a weighted mean of selected quantiles – quick estimates of location (Shabri & Jemain, 2007, Mudholkar & Hutson, 1998). The sample quick estimators of the location estimate the population characteristics based on a consistent estimator of the population quantiles. We can mention the Tukey´s trimean or the Gastwirth estimator (Pearson, 2011) of the central tendency, we use more general characteristics defined by selected quantiles and their weighted mean.

For the proper use of the LQ moments, we need a deeper knowledge of properties of these estimators of quantiles. The lognormal distribution is frequently used for the modeling of incomes, robust moments are applied in the moment matching method in (Bílková, 2014). The main task of this contribution is to describe properties of three and five quantiles quick estimators of the central tendency for this distribution. We use a Monte Carlo simulation to show properties of these estimators. Based on the simulation study, we study also the impact of the censoring (0-50%) on the estimated values.

1 Methods

1.1 Quick location estimators

Denote *f* a density, *F* a cumulative distribution function, *S* a survival function, and *Q* a quantile function of a positive value continuous random variable *X*. In the survival analysis, we prefer quantile characteristics to the classical moments, as we usually work with skewed distributions. Distribution-free estimation of classical moments is not straightforward in case of the presence of censored values in the data.

Instead of a median, we will analyze more sophisticated estimators of the location, where more quantiles are included in the statistics. The statistics trimean is given as

$$0.25\,Q(0.25) + 0.5\,Q(0.5) + 0.25\,Q(0.75). \tag{1}$$

In (2) we define a more general three-point quantile estimator (Mudholkar & Hutson, 1998, Šimková & Picek, 2017).

$$\tau_{a,p,3}(X) = p\,Q(a) + (1-2p)\,Q(0.5) + p\,Q(1-a), \tag{2}$$

for $0 \le p \le 0.5, 0 < a \le 0.5.$ The inclusion of two additional quantiles in the weighted mean gives a strong emphasis on the center (median), but the two quantiles also bring in significant representation from the edges (the role depends on *p* and corresponding weights 2*p* for quantiles and 1–2*p* for the median).

From (2) we obtain the median for $p = 0$ and the trimean for $p = 0.25$ and $a = 0.25.$ The formula includes also the Gastwirth quick estimator of the location (Pearson, 2011)

$$0.3\,Q(1/3)+0.4\,Q(0.5)+0.3\,Q(2/3). \qquad (3)$$

Now, select $0 \le p \le 0.25, 0 \le a \le 0.1.$ Instead of using three quantiles (a median and two symmetric quantiles corresponding to the selected probabilities $P = a$ and $1-P = 1-a$) a quick five-point quantile estimator of the location $\tau_{a,p,5}$ based on a median and two pairs of symmetric quantiles (defined by probabilities $P = a$, $1-a$, $5a$, and $1-5a$) in the form

$$\tau_{a,p,5}(X) = pQ(a) + pQ(5a) + (1-4p)Q(0.5) + pQ(1-5a) + pQ(1-a) \qquad (4)$$

is used to evaluate LQ-moments in (Shabri & Jemain, 2007). The estimator is a weighted average of 5 quantiles corresponding to the probabilities $a, 5a, 0.5, 1-5a,$ and $1-a$ with the weighs p for $Q(a), Q(5a), Q(1-5a),$ and $Q(1-a)$ and the complementary probability $1-4p$ for the median $Q(0.5)$.

It is obvious, that for all symmetric distributions the values of $\tau_{a,p,3}(X)$ and $\tau_{a,p,5}(X)$ are equal to the median $Q(0.5)$. The quantile function is defined for all distributions without any assumption of the existence of the moments. For this reason, estimators (2) and (4) can be used even for distributions like a Cauchy distribution. If the expected value exists, then $E(X) = Q(0.5)$ and $\tau_{a,p,5}(X) = \tau_{a,p,3}(X) = Q(0.5) = E(X)$ for all acceptable values of parameters a and p.

In this contribution, we use these characteristics for the positively skewed lognormal distribution. We analyze not only the theoretical values but also the possibility to obtain their sample values or estimates. For this reason, we use both non-parametric and parametric counterparts. In order to estimate characteristics (2) and (4), the estimates of the quantile function are substituted into formulas.

For the data with all completed values, we can estimate quantiles by any definition of a sample quantile function. For example, there are 9 possible definitions of sample quantiles in the program R (R Core Team, 2017, funkce *quantile* in the package *stats*). All estimators (based on the empirical distribution) asymptotically approach the unknown theoretical quantile function. In Sheather and Marron (1990), these methods that interpolate and smooth the order statistics are discussed. We use a linearized sample quantile function (default type 7 in R) based on the formula (for the probability $0 < P < 1$) and $h = (n-1)P + 1$

$$\hat{Q}_{lin}(P) = X_{(\lfloor h \rfloor)} + \left(h - \lfloor h \rfloor\right)\left(X_{(\lfloor h \rfloor + 1)} - X_{(\lfloor h \rfloor)}\right), \qquad (5)$$

where $\lfloor . \rfloor$ is a floor function and the ordered sample is denoted by $\left(X_{(1)}, X_{(2)}, ..., X_{(n)}\right)$.

The second estimate is a kernel type estimate of quantiles $\hat{Q}_{\ker}$ based on the kernel estimator of the density of *X*. Estimates are smooth quantiles based on the quasi-inverse of the kernel estimate of the cumulative distribution function (Racine & Hayfield, 2018; Padgett, 1986 for censored data).

The estimators $\hat{Q}_{lin}$ and $\hat{Q}_{\ker}$ are distribution-free nonparametric estimators, maximum likelihood estimator $\hat{Q}_{MLE}$ is based on the assumption of the lognormal distribution.

All computations and simulations are performed in R (R Core Team, 2017), for the evaluation of quantiles we use the packages *np* (Racine & Hayfield, 2018) and *survival* (Therneau, 2015, Therneau & Grambsch, 2000).

1.2 Estimates of quantiles from censored data

In order to estimate theoretical quantiles, if censored values are included in our data, we are able to estimate quantiles using non-parametric methods to avoid any assumption of the probability distribution. In this text, we use estimates based on Kaplan-Meier estimator of the survival function (Kaplan & Meier, 1958).

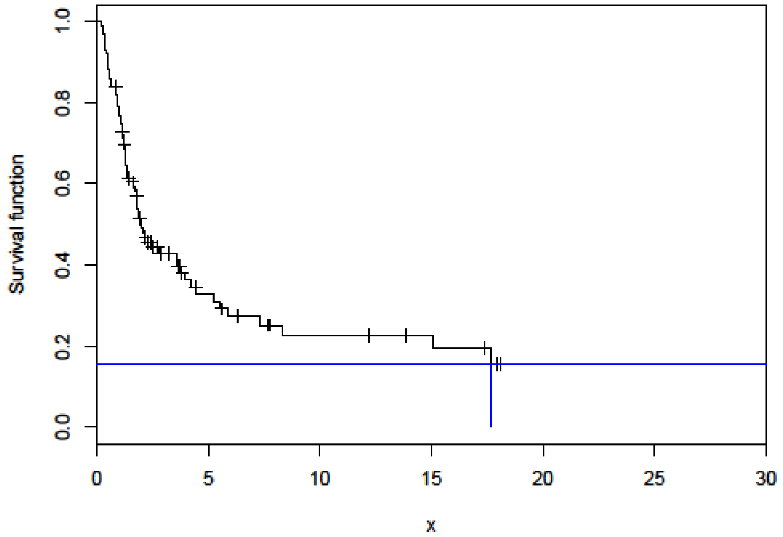
The Kaplan-Meier estimate of the survival function S $\left(S = 1 - F\right)$ is constructed (for positive values of *x*) as

$$\hat{S}(x) = \prod_{i,\, x_i \le x} \left(1 - \frac{d_i}{n_i}\right), \tag{6}$$

where d_i is a number of events that happened at x_i and n_i denotes the number of all observations under the risk at x_i.

Then the P^{th} quantile for a survival curve *S*(*x*) is estimated as the location at which a horizontal line at the height 1–*P* intersects the plot of $\hat{S}(x)$. If the survival curve does not fall to$1 - P,$ then that quantile is undefined. This problem is discussed in this text, as in case of the heavy censoring, the sample quantiles for too high probabilities are not defined. This problem can be solved by the use of parametric approach, but we have to take into account insufficient information about tails of the distribution. In Figure 1 the estimated survival function is shown for a sample of 100 values with 28% (28 censored, 72 non-censored values) of right censored values (in the random censoring we choose the censoring variable to produce 30% of censored data in the mean). The highest non-censored value is 17.7 (the vertical line in blue) enables to estimate quantiles for probabilities up to 0.845. The value 0.155 is shown by the horizontal blue line. In the figure, non-censored data are given by plus signs. In the sample, the maximum is equal to 39.6 and 4 values in the generated sample are higher than 17.7. From this curve, we can estimate only quick estimators, that use quantiles to *P* = 0.845.

Fig. 1: Kaplan-Meier estimate of survival function, *n*=100, *LN*(1;1.5²)



Source: own computations

1.3 Quick estimators for the lognormal distribution

Applying definitions (2) and (4) on the symmetric distribution, we obtain the centre of the symmetry. In the contribution, we use skewed two-parametric lognormal distributions $LN(\mu;\sigma^2)$ with low and high coefficient of skewness. For the lognormal distribution we obtain

$$\begin{aligned}\tau_{a,p,3} &= pe^{\mu+\sigma u_\alpha} + (1-2p)\,e^{\mu} + pe^{\mu-\sigma u_\alpha} = e^{\mu}\left[p\left(e^{\sigma u_\alpha}+e^{-\sigma u_\alpha}\right)+1-2p\right]= \\ &= e^{\mu} + p\left[e^{\mu}\left(e^{\sigma u_\alpha}+e^{-\sigma u_\alpha}-2\right)\right],\end{aligned} \tag{7}$$

$$\begin{aligned}\tau_{a,p,5} &= e^{\mu}\left[p\left(e^{\sigma u_\alpha}+e^{-\sigma u_\alpha}\right)+p\left(e^{\sigma u_{5\alpha}}+e^{-\sigma u_{5\alpha}}\right)+1-4p\right]= \\ &= e^{\mu} + p\left[e^{\mu}\left(e^{\sigma u_\alpha}+e^{-\sigma u_\alpha}+e^{\sigma u_{5\alpha}}+e^{-\sigma u_{5\alpha}}-4\right)\right].\end{aligned} \tag{8}$$

Both estimators are the product of the median (e^{μ}) and a term depending on the parameter σ^2 and selected parameters *a* and *p* of the quick estimator of the location.

1.4 Simulation study

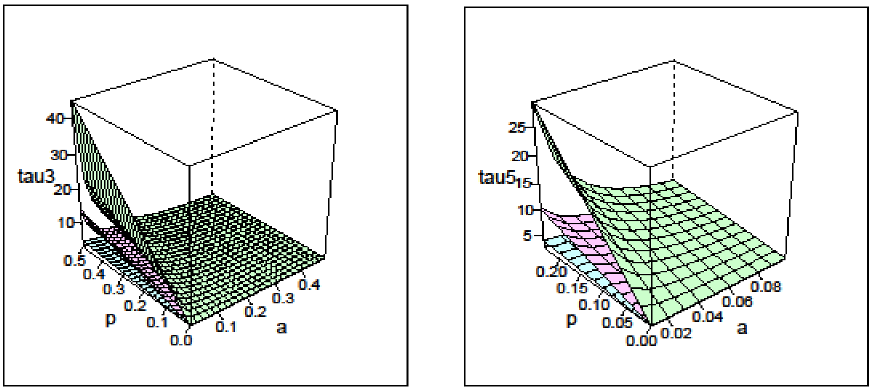
In the simulation, we generate 10,000 samples with 50, 100, and 500 observations from two lognormal distributions with parameters $\mu_1 = 1,\ \sigma^2 = 0.5$ and $\mu_2 = 1,\ \sigma^2 = 1.5.$ We selected one distribution with a low skewness (coefficient of skewness 1.06) and one distribution relatively skewed (coefficient of skewness 33.47) with the same median $x_{0.5} = \mathrm{e}^1 = 2.71$. The characteristics of these distributions are

$$E(X) = 3.08,\ \sqrt{D(X)} = 1.64,\ \ trimean = 2.80, Gastwirth\ location = 2.77, \tag{9}$$

$$E(X) = 8.37, \sqrt{D(X)} = 24.4,\ \ trimean = 3.48, Gastwirth\ location = 3.07. \tag{10}$$

In Figure 2 the values of quick estimators for three distributions (the middle surface is constructed for the lognormal distribution with $\mu_1 = 1, \sigma^2 = 1$) with the same median 2.71, standard deviations 1.64, 5.87, 24.4 and coefficients of skewness 1.06, 6.18 and 33.47.

Fig. 2: 3D plot of *τ*3 and *τ*5 for *LN*(1;0.5), *LN*(1;1), *LN*(1;1.5)



Source: own computations

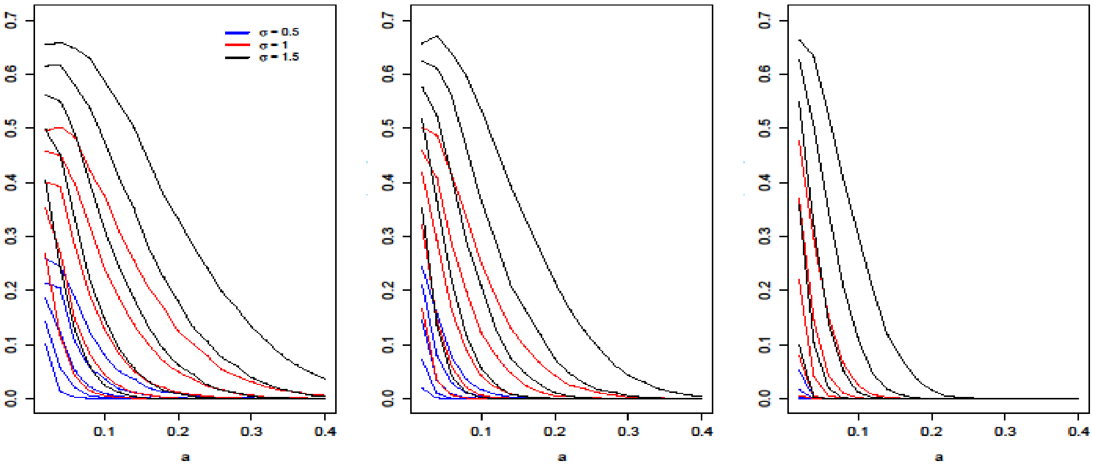
The complete generated samples (referred as 0% censored) were then transformed to censored data with approximatively (in the mean) 10%, 20%, 30%, 40%, and 50% of right censored data. The independent random censoring was used. The *n* independent values of a censoring random variable *C* with the lognormal distribution with $\sigma_C = 1$ and μ_C were generated independently of the sample from the distribution of *X*. The parameter μ_C was found to obtain the chosen percentage of censored data by the formula

$$\mu_C = \mu + \sqrt{\sigma_C^2 + 1} \cdot u_{1-P}$$

for 100*P*% of censored data in the censored sample. The observed censored data are then given by *min*(*X*, *C*); the value is censored, if *X* > *C*. We use $\mu_C = 2.432, 1.941, 1.586, 1.284, \text{and}\, 1$ in order to obtain (in the expected value) 10, 20, 30, 40, and 50% of censored data for the first distribution and $\mu_C = 3.310, 2.517, 1.945, 1.456, \text{and}\, 1$ for the second distribution in the simulation study ((9) and (10)).

In Figure 3, the relative frequencies of failed estimates due to the censored data are shown. In case of the heavy censoring, it is not possible to find upper quantiles, in our problem usually $\hat{Q}(1-a)$ for small values a. Curves for relative frequencies for 10 – 50 per cent of censored data are ordered from right to left and they seem to be shifted with different rate. It means, that in case of larger random sample we can estimate high quantiles even in case of heavy censoring.

Fig. 3: Percentage of failed estimates, sample sizes *n*=50 (left), *n*=100 and *n*=500 (right)

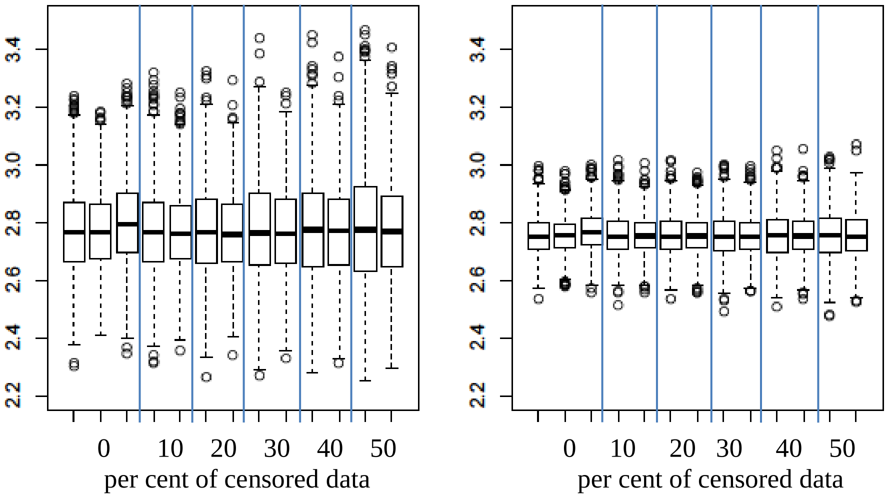


Source: own computations

In Figure 4 we present box-plots of the Gastwirth estimator for 100 and 500 observations and variance of lognormal distribution equal to 0.5 (to avoid problems with estimation of quantiles, see middle and right part of Figure 3, blue lines). Three estimates of the quantile function are used for the complete data; we compare the distribution for the non-parametric estimate, the maximum likelihood estimate and the kernel estimate. For datasets with censored data, an estimator based on Kaplan-Meier estimator and maximum likelihood estimates are given. The symmetric distributions are shown, the variances increase with the percentage of censored data (and decreases with the sample size). The variances are smaller for the parametric estimates (left box from the pair), but the difference is not large.

In Table 1 we present estimated bias and squared standard error (MSE) for Gastwirth estimator (3) based on the Monte Carlo simulation. We compare these values for sample sizes 100 and 500 and three lognormal distributions with increasing parameter σ (values 0.5, 1, 1.5). The data are not contaminated; for this reason, the parametric approach is probably optimal. But the performance of distribution-free methods is comparable to the parametric method. In our simulation, the kernel estimate is not superior to the usual sample quantiles. Its efficiency decreases with the parameter of skewness, we register more outliers in estimates for large values of the parameter σ.

Fig. 4: Box plots for Gastwirth estimate; complete data: non-parametric (left), parametric (middle), kernel (right)), for censored data non-parametric (left), parametric (right).



Source: own computations

Conclusion

In the contribution, the quick location estimates of the location are treated. Their properties are analyzed with the use of simulations for the lognormal distribution.

The quick estimates of the location are based on quantiles, instead of means. For their evaluation from a sample, sample quantiles should be estimated from the random sample. In case of heavy right censoring, the problem of estimation of upper quantiles usually arises because of the lack of high non-censored values. In the contribution, we illustrate the problem for the lognormal distribution. In Figure 2, the percentage of failed estimation of (1–*a*) % quantile is shown for sample sizes 50 to 500 and for mild to heavy censoring from 10 to 50% of censored data. The maximum likelihood parametric fit is applicable even for small samples and heavy censoring, but the quality of estimates is expected to be poor as we have information only on a small left part of the distribution.

From our simulations, it follows that it is necessary to pay attention to data and to sufficient sample size in relation to the percentage of censored data. The choice of the parameter *a* should be large enough to be able to estimate (1–*a*) % quantile. Small values of *a* seem to be theoretically interesting, by the application is restricted to large samples or only a small part of censored data.

Tab. 1: Gastwirth quick estimator results (censored 0%-50%, methods for $\hat{Q}$ lq lin, par MLE, Kq kernel, KM Kaplan-Meier)

		n=100, *σ*=0.5		*n*=500, *σ*=0.5		*n*=100, *σ*=1		*n*=500, *σ*=1		*n*=100, *σ*=1.5		*n*=500, *σ*=1.5	
%	Est	bias	MSE	bias	MSE	Bias	MSE	bias	MSE	Bias	MSE	bias	MSE
0	Lq	0.0132	0.0242	-0.0020	0.0049	0.0186	0.1104	0.0017	0.0223	0.0493	0.2810	0.0084	0.0567
	par	0.0122	0.0190	-0.0023	0.0040	0.0240	0.0863	0.0032	0.0171	0.0402	0.2216	0.0045	0.0412
	Kq	0.0403	0.0251	0.0089	0.0049	0.0186	0.1149	0.0223	0.0222	0.0493	0.2915	0.0084	0.0595
10	KM	0.0147	0.0229	-0.0022	0.0051	0.0263	0.1124	0.0089	0.0215	0.0617	0.2819	0.0077	0.0577
	par	0.0115	0.0205	-0.0020	0.0041	0.0238	0.0895	0.0057	0.0175	0.0420	0.2318	0.0042	0.0433
20	KM	0.0155	0.0249	-0.0023	0.0055	0.0310	0.1178	0.0141	0.0228	0.0681	0.3083	0.0132	0.0608
	par	0.0102	0.0215	-0.0026	0.0045	0.0251	0.0931	0.0093	0.0194	0.0433	0.2548	0.0062	0.0472
30	KM	0.0177	0.0275	-0.0022	0.0061	0.0365	0.1322	0.0087	0.0245	0.0815	0.3454	0.0152	0.0667
	par	0.0129	0.0250	-0.0031	0.0050	0.0270	0.1064	0.0051	0.0212	0.0561	0.2811	0.0084	0.0544
40	KM	0.0258	0.0311	-0.0025	0.0069	0.0481	0.1571	0.0167	0.0282	0.1123	0.4308	0.0222	0.0742
	par	0.0176	0.0275	-0.0028	0.0057	0.0227	0.1265	0.0095	0.0227	0.0660	0.3247	0.0092	0.0622
50	KM	0.0286	0.0388	0.0007	0.0082	0.0831	0.2244	0.0252	0.0356	0.2039	0.9528	0.0410	0.0961
	par	0.0200	0.0331	-0.0017	0.0062	0.0346	0.1400	0.0149	0.0284	0.0676	0.4015	0.0167	0.0724

Source: own computations

Acknowledgment

The support of the funds of institutional support of a long-term conceptual advancement of science and research number IP 400 040 at the Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic is gladly acknowledged.

References

Bílková, D. (2014). Alternative Means of Statistical Data Analysis: L-Moments and TL-Moments of Probability Distributions, *Statistika*, 94, 77-94.

Hosking, J. R. M. (1990). L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society*, 52, 105–124.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from incomplete. *Journal of the American Statistical Association*, 53, 457-481.

Lio,Y.L., Padgett,W.J., & Yu, K.F. (1986). On the asymptotic properties of a kernel type quantile estimator from censored samples. *Journal of Statistical Planning and Inference*, 14, 169-177.

Mudholkar G.,S., & Hutson A., D. (1998). LQ-moments: Analogs of L-moments, *Journal of Statistical Planning and Inference*, 71, 191-208.

Racine, J.S., & Hayfield, T. (2018). Package np.
https://cran.rproject.org/web/packages/np/np.pdf

Padgett, W. J. (1986). A kernel-type estimator of a quantile function from right censored data. *Journal of the American Statistical Association*, 81, 215-222.

Pearson R. (2011). *Exploring Data in Engineering, the Sciences, and Medicine*. Oxford University Press Inc.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Shabri, A., & Jemain, A.A. (2007). LQ-Moments for Statistical Analysis of Extreme Events. *Journal of Modern Applied Statistical Methods*, 6, 228-238

Sheather, S., & Marron, J.S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85, 410-416.

Šimková, T., & Picek, J. (2017). A comparison of L-, LQ-, TL-moment and maximum likelihood high quantile estimates of the GPD and GEV distribution. *Communications in Statistics - Simulation and Computation*, 46, 5991-6010.

Therneau, T.M., & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Therneau, T.M. (2015). A Package for Survival Analysis in S.
https://CRAN.R-project.org/package=survival

Contact

Ivana Malá
University of Economics, Prague
W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic
malai@vse.cz

Václav Sládek
University of Economics, Prague
W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic
vaclav.sladek@vse.cz