Estimating the mean of a heavy-tailed distribution under random censoring

Louiza Soltane*
Mohamed Khider University, Biskra, Algeria - louiza_stat@yahoo.com

Djamel Meraghni
Mohamed Khider University, Biskra, Algeria - djmeraghni@yahoo.com

Abdelhakim Necir
Mohamed Khider University, Biskra, Algeria - necirabdelhakim@yahoo.fr

Abstract

The central limit theorem introduced by Stute (1995) does not hold for some class of heavy-tailed distributions. In this work, we make use of the extreme value theory to propose an alternative estimating approach of the mean ensuring the asymptotic normality property. A simulation study is carried out to evaluate the performance of this estimation procedure and, as an application, confidence bounds to the mean of the survival time of Australian male Aids patients are provided.

Keywords: Extreme values; Hill estimator; Kaplan-Meier estimator; Random censoring.

1. Introduction

Let $X_1, ..., X_n$ be $n \geq 1$ independent copies of a non-negative random variable (rv) X, defined over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with cumulative distribution function (cdf) F. These rv's are censored to the right by a sequence of independent copies $Y_1, ..., Y_n$ of a non-negative rv Y, independent of X, with cdf G. At each stage $1 \leq j \leq n$, we can only observe the rv's $Z_j := \min(X_j, Y_j)$ and $\delta_j := \mathbf{1}\{X_j \leq Y_j\}$, with $\mathbf{1}\{\cdot\}$ denoting the indicator function. The latter rv indicates whether there has been censorship or not. If we denote by H the cdf of the observed $Z's$, then, by the independence of X and Y, we have $1 - H = (1 - F)(1 - G)$. Throughout this Chapter, we will use the notation $\overline{\mathcal{S}}(x) := \mathcal{S}(\infty) - \mathcal{S}(x)$, for any function $\mathcal{S}$. Assume further that F and G are heavy-tailed or, in other words, that $\overline{F}$ and $\overline{G}$ are regularly varying at infinity with negative indices $-1/\gamma_1$ and $-1/\gamma_2$ respectively. That is

$$\lim_{z\to\infty} \frac{\overline{F}(xz)}{\overline{F}(z)} = x^{-1/\gamma_1} \text{ and } \lim_{z\to\infty} \frac{\overline{G}(xz)}{\overline{G}(z)} = x^{-1/\gamma_2}, \tag{1}$$

for any $x > 0$. Consequently, H is heavy-tailed too, with tail index $\gamma := \dfrac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}$. Examples of censored data with apparent heavy tails can be found in Gomes and Neves (2011). The convergence rates of the limits (1) are formulated by the well-known second-order condition of regularly varying functions. In other words, there exist constants $\rho_j < 0$ and functions A_j, $j = 1, 2$ tending to zero, not changing sign near infinity and having regularly varying absolute values with indices ρ_j, such that for any $x > 0$

$$\lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma_1}}{A_1(t)} = x^{-1/\gamma_1} \frac{x^{\rho_1/\gamma_1} - 1}{\gamma_1\rho_1}, \tag{2}$$

and

$$\lim_{t\to\infty} \frac{\overline{G}(tx)/\overline{G}(t) - x^{-1/\gamma_2}}{A_2(t)} = x^{-1/\gamma_2} \frac{x^{\rho_2/\gamma_2} - 1}{\gamma_2\rho_2}. \tag{3}$$

The class of heavy-tailed distributions, satifying the second-order condition, takes a significant role in extreme value theory. It includes distributions such as Burr, Fréchet, Benktander, generalised Pareto, the log-logistic, log-gamma and α-stable $(0 < \alpha < 2)$, known to be appropriate models for fitting large insurance claims, log-returns, large fluctuations of prices, etc ... (see, e.g., Resnick, 2007).

The nonparametric maximum likelihood estimator of cdf F is given by Kaplan and Meier (1958) as the product limit estimator

$$\widehat{F}_n(x) := \begin{cases} 1 - \prod\limits_{Z_{j:n} \leq x} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} & \text{for } x < Z_{n:n}, \\ 1 & \text{for } x \geq Z_{n:n}, \end{cases} \tag{4}$$

where $Z_{1:n} \leq ... \leq Z_{n:n}$ denote the order statistics pertaining to the sample $Z_1, ..., Z_n$ with the corresponding concomitants $\delta_{[1:n]}, ..., \delta_{[n:n]}$ satisfying $\delta_{[j:n]} = \delta_i$ if $Z_{j:n} = Z_i$. The aim of this paper is to propose an asymptotically normal estimator for the mean $\mu = \mathbf{E}[X] := \int_0^\infty \overline{F}(x)dx$. By substituting $\widehat{F}_n$ for F in the previous equation, Stute (1995) defined the empirical mean for censored data by

$$\widetilde{\mu}_n := \sum_{i=1}^{n} \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{i:n}, \tag{5}$$

and established, in Corollary 1.2, its asymptotic normality. Explicitly, the author showed that

$$\sqrt{n}\left(\widetilde{\mu} - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right) \text{ as } n \to \infty,$$

where $\sigma^2 := \mathbf{Var}\left[Z_1\Gamma_0\left(Z_1\right)\delta_1 + \Gamma_1\left(Z_1\right)\left(1-\delta_1\right) - \Gamma_2\left(Z_1\right)\right]$, with

$$\Gamma_0(x) := \exp\left\{\int_0^x \frac{dH^{(0)}(s)}{\overline{H}(s)}\right\}, \tag{6}$$

$$\Gamma_1\left(x\right) := \int_0^x \frac{s\Gamma_0\left(s\right)}{\overline{H}\left(s\right)} dH^{(1)}\left(s\right) \text{ and } \Gamma_2\left(x\right) := \int_x^\infty \frac{\int_s^\infty t\Gamma_0\left(t\right) dH^{(1)}\left(t\right)}{\left[\overline{H}(s)\right]^2} dH^{(0)}\left(s\right),$$

provided that

$$I_1 := \int_0^\infty x^2\Gamma_0^2(x)dH^{(1)}(x) \text{ and } I_2 := \int_0^\infty x\left(\int_0^x \frac{dH^{(0)}(y)}{\left[\overline{H}(y)\right]^2}\right)^{1/2} dF(x), \tag{7}$$

be finite, where $H^{(j)}\left(v\right) := \mathbb{P}\left(Z_1 \leq v,\ \delta_1 = j\right), j = 0, 1$, are two functions defined on $\mathbb{R}_+$, that will play a prominent role in this work. However, assumptions (7) may be violated by a class of heavy-tailed distributions. We show that when F and G satisfy the second order conditions (2)-(3) with $\gamma_1 > \gamma_2/\left(1+2\gamma_2\right)$, then both I_1 and I_2 are infinite. In other words,the range

$$\mathcal{R} := \left\{\gamma_1, \gamma_2 > 0 : \frac{\gamma_2}{1+2\gamma_2} < \gamma_1 < 1\right\}, \tag{8}$$

is not covered by the central limit theorem established by Stute (1995). As an example of censored real datasets with indices belonging to $\mathcal{R}$, we may cite the Australian Aids data that will be described and analyzed in Section 4. After noting that these medical observations exhibit a heavy right tail (see Einmahl et al., 2008), we estimate, in Section 4, the corresponding extreme value index (EVI) γ_1 and the proportion $p := \gamma_2/\left(\gamma_1+\gamma_2\right)$ by 0.29 and 0.90, respectively, leading to a γ_2 estimate equal to 0.37. These values of (γ_1, γ_2) clearly lie in the range $\mathcal{R}$ where Stute's central limit theorem is not valid and thus no confidence interval could be constructed for the mean of this dataset. Consequently, we need to handle this situation by adopting an approach that is different from that of Stute (1995). This problem has already been addressed by Peng (2001) and Johansson (2003) for sets of complete data from heavy-tailed distributions with tail indices lying between $1/2$ and 1. A bias reduced version of Peng's estimator is provided in Brahimi et al. (2013). Note that in the non censoring case, we have $\gamma_1 = \gamma$ meaning that $\gamma_2 = \infty$, consequently $\mathcal{R}$ reduces to Peng's range.

2. Main Result

To define our estimator, we introduce an integer sequence $k = k_n$, representing a fraction of extreme order statistics, satisfying

$$1 < k < n,\ k \to \infty \text{ and } k/n \to 0 \text{ as } n \to \infty, \tag{9}$$

and we set $h = h_n := H^{-1}(1 - k/n)$, where $K^{-1}(y) := \inf\{x : K(x) \geq y\}$, $0 < y < 1$, denotes the quantile function of a cdf K. We start by decomposing μ into the sum of two terms as follows: $\mu = \int_0^h \overline{F}(x)dx + \int_h^\infty \overline{F}(x)dx =: \mu_1 + \mu_2$, then we estimate each term separately. Integrating the first integral by parts and changing variables in the second respectively yield

$$\mu_1 = h\overline{F}(h) + \int_0^h x dF(x) \text{ and } \mu_2 = h\overline{F}(h)\int_1^\infty \frac{\overline{F}(hx)}{\overline{F}(h)}dx.$$

By replacing h and $F(x)$ by $Z_{n-k:n}$ and $\widehat{F}_n(x)$ of formula (4) respectively, we get

$$\widehat{\mu}_1 = \prod_{j=1}^{n-k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{n-k:n} + \sum_{i=1}^{n-k}\frac{\delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{i:n}, \tag{10}$$

as an estimator to μ_1. Regarding μ_2, we apply the well-known Karamata theorem (see, for instance, de Haan and Ferreira, 2006, page 363), to write

$$\mu_2 \sim \frac{\gamma_1}{1-\gamma_1} h\overline{F}(h), \text{ as } n \to \infty,\ 0 < \gamma_1 < 1. \tag{11}$$

The quantities h and $\overline{F}(h)$ are, as above, naturally estimated by $Z_{n-k:n}$ and

$$\widehat{\overline{F}}_n(Z_{n-k:n}) = \prod_{j=1}^{n-k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}},$$

respectively. Now, it is clear that to derive an estimator to μ_2, one needs to estimate the tail index γ_1. The general existing method, which first appeared in Beirlant et al. (2007) and then developed in Einmahl et al. (2008), is to consider any consistent estimator of the extremal index γ based on the Z-sample and divide it by the proportion of observed observations in the tail. For instance, Einmahl et al. (2008) adapted Hill's estimator to introduce an estimator $\widehat{\gamma}_1^{(H,c)} := \widehat{\gamma}^H/\widehat{p}$ to the tail index $\gamma_1 = \gamma/p$ under random right censorship, where

$$\widehat{\gamma}^H := \frac{1}{k}\sum_{i=1}^{k}\log\frac{Z_{n-i+1:n}}{Z_{n-k:n}} \quad \text{and} \quad \widehat{p} := \frac{1}{k}\sum_{i=1}^{k}\delta_{[n-i+1:n]},$$

are the classical Hill estimator and the proportion of upper non-censored observations respectively. Further results of this last nature can be found in Ndao et al. (2014), Worms and Worms (2014), Brahimi et al. (2015), Ndao et al. (2016) , Stupfler (2016) and Beirlant et al. (2016). Let us now continue with the construction our new estimator. By replacing, in (11) , F and γ_1 by their respective empirical counterparts $\widehat{F}_n$ and $\widehat{\gamma}_1^{(H,c)}$, we obtain

$$\widehat{\mu}_2 := \frac{\widehat{\gamma}_1^{(H,c)}}{1-\widehat{\gamma}_1^{(H,c)}} Z_{n-k:n}\prod_{j=1}^{n-k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}, \quad \text{for } \widehat{\gamma}_1^{(H,c)} < 1, \tag{12}$$

as an estimator for μ_2. Finally, with (10) and (12), we construct our estimator $\widehat{\mu}$ of the mean μ as follows:

$$\widehat{\mu} := \sum_{i=1}^{n-k}\frac{\delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}} Z_{i:n} + \prod_{j=1}^{n-k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}\frac{Z_{n-k:n}}{1-\widehat{\gamma}_1^{(H,c)}}.$$

Our main result consists in the asymptotic normality of the newly introduced estimator $\widehat{\mu}$. It is stated in the following theorem which results in a corollary that is very useful in the practical construction of an asymptotic confidence interval for the expected value μ.

Theorem 1 *Assume that both second-order conditions of regular variation* (2) *and* (3) *hold with* $(\gamma_1, \gamma_2) \in \mathcal{R}$. *Let* $k = k_n$ *be an integer sequence satisfying* (9) *and* $h = h_n := H^{-1}(1 - k/n)$ *such that* $\sqrt{k}A_1(h) \to \lambda$, $\sqrt{k}A_2(h) = O(1)$ *and* $\sqrt{k}h\overline{F}(h) \to \infty$. *Then*

$$\frac{\sqrt{k}\left(\widehat{\mu} - \mu\right)}{h\overline{F}(h)} \xrightarrow{\mathcal{D}} \mathcal{N}\left(m, \mathcal{V}^2\right), \text{ as } n \to \infty,$$

where

$$m := \frac{\lambda}{(1 - p\rho_1)(1 - \gamma_1)^2} + \frac{\lambda}{(\gamma_1 + \rho_1 - 1)(1 - \gamma_1)},$$

and

$$\mathcal{V}^2 = \mathcal{V}^2(p, \gamma_1) := \frac{2p\gamma_1\left(\gamma_1 - p^2\gamma_1^2 + p^2 + 2p\gamma_1^2 - 3p\gamma_1\right)}{(\gamma_1 - 1)^2(1 - 2p + 2p\gamma_1)(1 - p + p\gamma_1)} - \frac{4\gamma_1^2}{(1 - \gamma_1)^3(1 - 2p + 2p\gamma_1)} + \frac{2(1 + 2p)\gamma_1^2}{p(1 - \gamma_1)^4}.$$

Corollary 1 *Under the assumptions of Theorem 1, with* $\lambda = 0$, *we have*

$$\sqrt{k}\left(\widehat{\mu} - \mu\right)/\sigma_{n,k} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \text{ as } n \to \infty,$$

where

$$\sigma_{n,k} := Z_{n-k:n} \prod_{j=1}^{n-k} \left(\frac{n - j}{n - j + 1}\right)^{\delta_{[j:n]}} \mathcal{V}\left(\widehat{p}, \widehat{\gamma}_1^{(H,c)}\right).$$

3. Simulation Study

We carry out a simulation study to illustrate the performance of our estimator, through two sets of censored and censoring data, from Burr model

$$F(x) = 1 - \left(1 + x^{1/\eta}\right)^{-\eta/\gamma_1}, \; G(x) = 1 - \left(1 + x^{1/\eta}\right)^{-\eta/\gamma_2}, \; x \geq 0,$$

where $\eta, \gamma_1, \gamma_2 > 0$. We fix $\eta = 1/4$ and choose the value 0.3 for γ_1. For the proportion of the really observed extreme values, we take $p = 0.40$ and 0.70. For each couple (γ_1, p), we solve the equation $p = \gamma_2/(\gamma_1 + \gamma_2)$ to get the pertaining γ_2-value. We vary the common size n of both samples $X_1, ..., X_n$ and $Y_1, ..., Y_n$, then for each size, we generate 1000 independent replicates to take our overall results as the empirical means of the results obtained through all the repetitions. To determine the optimal number (that we denote by k^*) of upper order statistics used in the computation of $\widehat{\gamma}_1^{(H,c)}$, we apply the algorithm of automatic selection given in page 137 of Reiss and Thomas (2007). The performance of the newly defined estimator $\widehat{\mu}$ is evaluated in terms of absolute bias (abs bias), mean squared error (mse) and confidence interval (conf int) accuracy via length and coverage probability (cov prob). The results, summarized in the Table 1. As expected, the sample size influences the estimation in the sense that the larger n gets, the better the estimation is. On the other hand, it is clear that the estimation accuracy increases when the censoring percentage decreases, which seems logical.

4. Application to AIDS Survival Data

In this Section, we apply our estimation procedure to the dataset known as Australian Aids data and provided by Dr P.J. Solomon and the Australian National Centre in HIV Epidemiology and Clinical Research. It consists in medical observations on 2843 patients (among whom 2754 are male) diagnosed with Aids in Australia before July 1^{st}, 1991. The datafile is available under the name "Aids2" in the package MASS of the statistical software R. In the literature, these data were analyzed with different prospects by several authors like, for instance, Ripley and Solomon (1994) and Venables and Ripley (2002) (pages $379 - 385$), Einmahl et al. (2008), Ndao et al. (2014) and Stupfler (2016). We apply the algorithm of Reiss and Thomas (2007) to obtain $k^* = 162$ as the optimal k-value and the corresponding estimates $\widehat{\gamma}_1^{(H,c)} = 0.90$ and $\widehat{p} = 0.29$. The mean survival time of male patients is estimated to be 1083.61 days with a 95%-confidence interval of $1082.58 - 1084.64$.

5. Conclusions

$\gamma_1 = 0.3 \rightarrow \mu = 1.298$						
$p = 0.40$						
n	$\widehat{\mu}$	biais abs	mse	bor. de conf.	prob.couv.	long.
500	1.247	0.052	0.021	$1.043 - 1.450$	0.88	0.407
1000	1.244	0.054	0.020	$1.099 - 1.389$	0.88	0.291
1500	1.233	0.065	0.005	$1.119 - 1.346$	0.80	0.227
2000	1.231	0.067	0.005	$1.135 - 1.328$	0.74	0.193
$p = 0.70$						
500	1.265	0.033	0.003	$1.069 - 1.460$	0.97	0.391
1000	1.269	0.029	0.002	$1.123 - 1.415$	0.96	0.291
1500	1.279	0.019	0.001	$1.162 - 1.395$	0.98	0.233
2000	1.278	0.020	0.001	$1.178 - 1.377$	0.96	0.199

Table 1: Absolute bias, mean squared error and 95-confidence interval accuracy of the mean estimator based on 1000 right-censored samples from Burr model with shape parameter 0.3

- The estimation of the mean of censored heavy-tailed distributions requires special methods because of their specific characteristics:
  - Rare observations in the tail.
  - Presence of incomplete data.
- Estimation by the non-parametric method is not applicable because there exist heavy-tailed distributions for which the Stute conditions are not satisfied.
- In practice, the Fréchet extreme-value type is the most interesting as it corresponds to heavy-tailed distributions.
- The main task in extreme value theory is the estimation of the EVI, which leads to solve problems related to the extremes of a random variable.

References

Beirlant, J., Bardoutsos, A., de Wet, T., & Gijbels, I. (2016). Bias reduced tail estimation for censored Pareto type distributions. Statist. Probab. Lett., 109, 78-88.

Beirlant, J., Guillou, A., Dierckx, G., & Fils-Villetard, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring. Extremes, 10(3), 151-174.

Brahimi, B., Meraghni, D., Necir, A., & Yahia, D. (2013). A bias-reduced estimator for the mean of a heavy-tailed distribution with an infinite second moment. J. Statist. Plann. Inference, 143(6), 1064-1081.

Brahimi, B., Meraghni, D., & Necir, A. (2015). Gaussian approximation to the extreme value index estimator of a heavy-tailed distribution under random censoring. Math. Methods Statist., 24(4), 266-279.

Einmahl, J. H., Fils-Villetard, A., & Guillou, A. (2008). Statistics of extremes under random censoring. Bernoulli, 14(1), 207-227.

Gomes, M. I., & Neves, M. M. (2011). Estimation of the extreme value index for randomly censored data. Biometrical Lett., 48(1), 1-22.

de Haan, L. & Ferreira, A. (2006). Extreme Value Theory: An Introduction. Springer-Verlag, New York.

Johansson, J. (2003). Estimating the mean of heavy-tailed distributions. Extremes, 6(2), 91-109.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc., 53(282), 457-481.

Ndao, P., Diop, A., & Dupuy, J. F. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. Comput. Statist. Data Anal., 79, 63-79.

Ndao, P., Diop, A., & Dupuy, J. F. (2016). Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. J. Statist. Plann. Inference, 168, 20-37.

Peng, L. (2001). Estimating the mean of a heavy tailed distribution. Statist. Probab. Lett., 52(3), 255-264.
Reiss, R.D., & Thomas, M. (2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser, Basel.
Resnick, S.I. (2007). Heavy-Tail Phenomena, probabilistic and statistical modeling. Springer.
Ripley, B. D., & Solomon, P. J. (1994). A note on Australian AIDS survival. University of Adelaide, Department of Statistics.
Stupfler, G. (2016). Estimating the conditional extreme-value index under random right-censoring. J. Multivariate Anal., 144, 1-24.
Stute, W. (1995). The central limit theorem under random censorship. Ann. Statist., 422-439.
Venables, W.N., & Ripley, B.D. (2002). Modern Applied Statistics with S, 4th edition. Springer.
Worms, J. & Worms, R. (2014). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes, 17, 337-358.