Asymptotic confidence intervals for the length of the Shortt under random censoring

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 93-03

ASYMPTOTIC CONFIDENCE INTERVALS FOR THE LENGTH OF THE SHORTT UNDER RANDOM CENSORING.

Jan Beirlant
John H.J. Einmahl

Eindhoven, February 1993
The Netherlands

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariate: Dommelbuilding 0.03
Telephone: 040-47 3130

ISSN 0926 4493

ASYMPTOTIC CONFIDENCE INTERVALS FOR THE LENGTH OF THE SHORTT UNDER RANDOM CENSORING

Jan Beirlant[1]
Catholic University Leuven

John H.J. Einmahl
Eindhoven University of Technology

Dedicated to the memory of Nico Willems.

A shortt of a one dimensional probability distribution is defined to be an interval which has at least probability t and minimal length. The length of a shortt and its obvious estimator are significant measures of scale of a distribution and the corresponding random sample, respectively. In this note a non-parametric asymptotic confidence interval for the length of *the* (uniqueness is assumed) shortt is established in the random censorship from the right model. The estimator of the length of the shortt is based on the product-limit (PL) estimator of the unknown distribution function. The proof of the result mainly follows from an appropriate combination of the Glivenko-Cantelli theorem and the functional central limit theorem for the PL estimator.

AMS 1991 subject classifications. 62G15, 62G20, 62G35.
Key words and phrases. Confidence interval, length of shortt, random censorship.
Running head: Length of shortt under censoring.

---

[1]Research performed while the author was research fellow at the Eindhoven University of Technology

1 Introduction and main result

Let $X_1, X_2, \cdots, X_n$ be a random sample from a univariate distribution function (*df*) F. An outlier resistant scale estimator based on such a sample is defined as the length of a shortest closed interval containing at least fraction t (shortt) of the data. This estimator possesses many desirable robustness properties; in particular for the case $t = \frac{1}{2}$, the asymptotic breakdown point is 50%; see Rousseeuw and Leroy (1988) for more details. A functional (in t) central limit theorem for this estimator is established in Grübel (1988), see also Einmahl and Mason (1992). It turns out that the length of the shortt has the "good" $n^{-\frac{1}{2}}$ rate of convergence, whereas the most prominent *location* estimators based on the shortt have only a rate of $n^{-\frac{1}{3}}$; cf. Andrews et al. (1972) and Kim and Pollard (1990).

Let F be continuous and write

$$(1) \quad U(t) = \inf\{b - a : F(b) - F(a) \geq t\}, \; 0 < t < 1 ,$$

for the theoretical counterpart of the length of the shortt, i.e. for our parameter of interest. It is the purpose of this note to derive a simple asymptotic confidence interval for $U(t)$, in the more general case that the $X_i, 1 \leq i \leq n$, are randomly censored from the right.

In order to be more explicit, let us introduce some notation. Let $X_1, \cdots, X_n$ be as above and let $Y_1, \cdots, Y_n$ be an independent random sample from a *df* G, which we also assume to be continuous. In the random censorship from the right model we observe the independent pairs $(Z_i, \delta_i), 1 \leq i \leq n$, where $Z_i = X_i \wedge Y_i$ and $\delta_i = I_{\{X_i \leq Y_i\}}$. The *df* of the Z_i is denoted with H and is easily seen to be equal to $1 - (1-F)(1-G)$. The well-studied product-limit estimator F_n of F is given by

$$F_n(x) = 1 - \prod_{Z_{i:n} \leq x} \left(1 - \frac{\delta_{i:n}}{n - i + 1}\right), x \in I\!R ,$$

where $Z_{1:n} \leq \cdots \leq Z_{n:n}$ are the order statistics of the Z_i and $\delta_{i:n}$ are the corresponding δ's. Observe that trivially $F_n(x) = F_n(Z_{n:n})$ for $x > Z_{n:n}$. For $0 < t < 1$, let $U_n(t)$ be the empirical analogue of $U(t)$ based on F_n, i.e.

$$(2) \quad U_n(t) = \inf\{b - a : F_n(b) - F_n(a-) \geq t\} .$$

Write $[l_n, r_n]$ for the almost surely unique random interval pertaining to $U_n(t)$.

We also introduce the following empirical (sub-) *df*'s:

$$H_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x)}(Z_i), \; x \in I\!R ,$$

$$H_n^1(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_i 1_{(-\infty,x]}(Z_i), \; x \in I\!R ,$$

and write

$$D_n(x) = \int_{-\infty}^{x} \frac{1}{(1 - H_n(u))^2} dH_n^1(u), \; x < Z_{n:n} \; .$$

Furthermore, set

$$\hat{\sigma} = \left\{(1 - F_n(r_n))^2 (D_n(r_n) - D_n(l_n)) + t^2 D_n(l_n)\right\}^{\frac{1}{2}}$$

and let $c := c(\alpha)$ denote the $(1 - \frac{\alpha}{2})$-th quantile of the standard normal *df*. In order to establish our result we need the following mild regularity condition on F:

(3) F has a density f which is positive and continuous on its support (β, γ), $-\infty \leq \beta < \gamma \leq \infty$, strictly increasing on $(\beta, \eta]$ and strictly decreasing on $[\eta, \gamma)$ for some $\eta \in [\beta, \gamma)$.

Let $[l, r]$ be the now uniquely defined interval pertaining to $U(t)$.

THEOREM. Let $0 < t < 1$ be fixed, assume that (3) holds and that $H(r) < 1$. Then for any $0 < \alpha < 1$

$$\lim_{n \to \infty} P\left(U_n(t - \frac{c\hat{\sigma}}{n^{\frac{1}{2}}}) < U(t) < U_n(t + \frac{c\hat{\sigma}}{n^{\frac{1}{2}}})\right) = 1 - \alpha \; .$$

2 Proof of the result

For the proof we need the Glivenko-Cantelli theorem and the uniform central limit theorem for F_n.

FACT 1 (see e.g. Wang (1987)). As $n \to \infty$

$$\sup_{x \leq Z_{n:n}} |F_n(x) - F(x)| \to_P 0 \; .$$

A number of consequences of Fact 1 are stated in the next corollary. In the remainder of this proof I denotes a closed interval $[a, b]$. For a function g with left-hand limits, write

$$g(I) = g(b) - g(a-) \; .$$

Denoting with $|I|$ the length of I we define

$$\tilde{F}_n(y) = \sup_{\substack{|I| \leq y \\ I \subset (-\infty, Z_{n:n}]}} F_n(I) \; ,$$

$$\tilde{F}_{(n)}(y) = \sup_{\substack{|I|\le y \\ I\subset(-\infty, Z_{n:n}]}} F(I) .$$

COROLLARY 1. Under the conditions of the Theorem we have for small enough $\varepsilon > 0$, as $n \to \infty$,

(4) $\sup_{I\subset(-\infty,Z_{n:n}]} |F_n(I) - F(I)| \to_P 0,$

(5) $\sup_{|y-U(t)|\le\epsilon} |\tilde{F}_n(y) - \tilde{F}_{(n)}(y)| \to_P 0,$

(6) $U_n(t) \to_P U(t),$

(7) $\sup_{x\le Z_{n:n}-U_n(t)} |F_n([x, x + U_n(t)]) - F([x, x + U(t)])| \to_P 0,$

(8) $l_n \to_P l$ and $r_n \to_P r$.

PROOF. The statement in (4) trivially follows from Fact 1 and the assertion in (5) follows immediately from (4). Write

$$\tilde{F}(y) = \sup_{|I|\le y} F(I)$$

and note that U is continuous and strictly increasing on $(0, 1)$ (because of (3)) and that $\tilde{F}$ is its inverse. Now (5) implies that for all small $\delta > 0$

(9) $P(\tilde{F}_n(U(t - \delta)) < t \le \tilde{F}_n(U(t + \delta))) \to 1 \ (n \to \infty).$

Observe that

$$U_n(t) = \inf\left\{y : \tilde{F}_n(y) \ge t\right\},\ 0 < t < 1 ,$$

and hence from (9) we have

$$P(U(t - \delta) < U_n(t) \le U(t + \delta)) \to 1 \ (n \to \infty) ,$$

which yields (6).

The proof of (7) follows from the fact that its left-hand side is less than or equal to

$$\sup_{x\le Z_{n:n}-U_n(t)} |F_n([x, x + U_n(t)]) - F([x, x + U_n(t)])|$$

$$+ \sup_{x\le Z_{n:n}-U_n(t)} |F([x, x + U_n(t)]) - F([x, x + U(t)])| ,$$

in combination with (4), (6) and the uniform continuity of F. Finally, assertion (8) is a consequence of (7) and (6); see e.g. Kim and Pollard (1990, p.208). □

To present the uniform central limit theorem for F_n, which we state in an almost sure construction setting, write

$$H^1(x) = P(Z_i \leq x, \delta_i = 1) = \int_{-\infty}^{x} (1 - G(u))dF(u),\ x \in \mathbb{R}\ ,$$

$$D(x) = \int_{-\infty}^{x} \frac{1}{(1 - H(u))^2} dH^1(u),\ x < \sup\{y : H(y) < 1\}\ ,$$

and

$$\alpha_n(x) = n^{\frac{1}{2}}(F_n(x) - F(x)),\ x \in \mathbb{R}\ .$$

FACT 2 (see e.g. Shorack and Wellner (1986, p.308)). Let $R \in \mathbb{R}$ with $H(R) < 1$. Under the conditions of the Theorem there exist probabilistically equivalent versions $\tilde{\alpha}_n$ of α_n and a standard Wiener process W such that as $n \to \infty$

(10) $\sup_{x \leq R} |\tilde{\alpha}_n(x) - (1 - F(x))W(D(x))| \to 0$ a.s.

Remark 1. Throughout we will choose $R > r$.
Remark 2. To prove our result we will proceed on the probability space on which (10) holds. Without confusion, we shall henceforth drop the symbol $\tilde{}$ from the notation.

Write $V(x) = (1 - F(x))W(D(x))$.

COROLLARY 2. As $n \to \infty$

$$\sup_{I \subset (-\infty, R]} |\alpha_n(I) - V(I)| \to 0 \text{ a.s.}$$

Define $\tilde{\alpha}_n(y) = n^{\frac{1}{2}}(\sup_{|I| \leq y} F_n(I) - \tilde{F}(y))$; hence $\tilde{\alpha}_n(U(t)) = n^{\frac{1}{2}}(\sup_{|I| \leq U(t)} F_n(I) - t)$.

PROPOSITION. Under the conditions of the Theorem we have as $n \to \infty$

(11) $\tilde{\alpha}_n(U(t)) \to V([l, r])$ a.s.

Remark 3. It is readily checked that $V([l,r])$ is a centered normal random variable with variance $(1-F(r))^2(D(r)-D(l))+t^2D(l) =: \sigma^2$.

PROOF. Define

$$\tilde{F}_{n,R}(y) = \sup_{\substack{|I|\le y \\ I\subset(-\infty,R]}} F_n(I) \ ,$$

$$\tilde{F}_R(y) = \sup_{\substack{|I|\le y \\ I\subset(-\infty,R]}} F(I) \ ,$$

and

$$\tilde{\alpha}_{n,R}(y) = n^{\frac{1}{2}}(\tilde{F}_{n,R}(y) - \tilde{F}_R(y)) \ .$$

Observe that because of Remark 1 we have $\tilde{F}_R(U(t)) = \tilde{F}(U(t)) = t$. Also

$$(12)\quad \lim_{n\to\infty} P(\tilde{F}_{(n)}(U(t)) = \tilde{F}(U(t))) = 1.$$

Furthermore, we have for the empirical counterparts of these quantities that $\tilde{F}_n(U(t)) = \sup_{|I|\le U(t)} F_n(I)$, and because of (4), (5) and (12)

$$\lim_{n\to\infty} P(\tilde{F}_n(U(t)) = \tilde{F}_{n,R}(U(t))) = 1 \ .$$

Therefore it suffices to prove (11) with $\tilde{\alpha}_n(U(t))$ replaced by $\tilde{\alpha}_{n,R}(U(t))$. The proof of this will be given along the lines of the proof of Proposition 3.1 in Einmahl and Mason (1992).

First observe that

$$V([l,r]) - \tilde{\alpha}_{n,R}(U(t)) \le V([l,r]) - n^{\frac{1}{2}}(F_n([l,r]) - F([l,r])) \ .$$

Hence by Corollary 2

$$\limsup_{n\to\infty} V([l,r]) - \tilde{\alpha}_{n,R}(U(t)) \le 0 \ \text{ a.s.}$$

We also have

$$(13)\quad \tilde{\alpha}_{n,R}(U(t)) - V([l,r])$$

$$\le \left\{ n^{\frac{1}{2}}\Big(\sup_{\substack{|I|\le U(t) \\ I\subset(-\infty,R] \\ t-n^{-1/4}<F(I)\le t}} F_n(I) - t\Big) - V([l,r]) \right\}$$

$$\vee \left\{ n^{\frac{1}{2}} (\sup_{\substack{I \subset (-\infty, R] \\ F(I) \leq t - n^{-1/4}}} F_n(I) - t) - V([l,r]) \right\} .$$

The second term on the right of (13) is less than or equal to

$$\sup_{\substack{I \subset (-\infty, R] \\ F(I) \leq t}} n^{\frac{1}{2}} (F_n(I) - F(I)) + |V([l,r])| - n^{\frac{1}{4}}$$

$$\leq \sup_{I \subset (-\infty, R]} |\alpha_n(I) - V(I)| + 2 \sup_{I \subset (-\infty, R]} |V(I)| - n^{\frac{1}{4}} ,$$

which by Corollary 2 and the fact that $H(R) < 1$, converges to $-\infty$ almost surely. The first term on the right of (13) is less than or equal to

$$(14) \quad n^{\frac{1}{2}} \sup_{\substack{|I| \leq U(t) \\ I \subset (-\infty, R] \\ t - n^{-1/4} < F(I) \leq t}} (F_n(I) - F(I)) - V([l,r])$$

$$\leq \sup_{I \subset (-\infty, R]} |\alpha_n(I) - V(I)| + \left(\sup_{\substack{|I| \leq U(t) \\ I \subset (-\infty, R] \\ t - n^{-1/4} < F(I) \leq t}} V(I) \right) - V([l,r])$$

From again Corollary 2 and condition (3) in conjunction with the continuity of V we see that the right-hand side of (14) converges to 0 almost surely. □

We finally need the following

LEMMA. Under the conditions of the Theorem we have as $n \to \infty$

$$\hat{\sigma} \to_P \sigma .$$

PROOF. Easy; based on Fact 1, (8) and the well known fact that

$$\sup_{x \leq R} |D_n(x) - D(x)| \to_P 0 \ (n \to \infty) .$$

□

PROOF OF THE THEOREM. From the Proposition and the Lemma it is immediate that, as $n \to \infty$, $\tilde{\alpha}_n(U(t))/\hat{\sigma}$ converges weakly to a standard normal random variable, implying that

$$P\left(U_n(t - \frac{c\hat{\sigma}}{n^{\frac{1}{2}}}) \leq U(t) < U_n(t + \frac{c\hat{\sigma}}{n^{\frac{1}{2}}}) \right)$$

$$= P\left(t - \frac{c\hat{\sigma}}{n^{\frac{1}{2}}} \leq \sup_{|I| \leq U(t)} F_n(I) < t + \frac{c\hat{\sigma}}{n^{\frac{1}{2}}}\right)$$

$$= P(-c \leq \tilde{\alpha}_n(U(t))/\hat{\sigma} < c) \to 1 - \alpha \ .$$

A little reflection shows that $P(U_n(t - \frac{c\hat{\sigma}}{n^{\frac{1}{2}}}) = U(t)) = 0$. This completes the proof. □

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). *Robust Estimates of Location.* Princeton Univ. Press, Princeton, N.J.

Einmahl, J.H.J. and Mason, D.M. (1992). Generalized quantile processes. *Ann. Statist.* **20**, 1062-1078.

Grübel, R. (1988). The length of the shorth. *Ann. Statist.* **16**, 619-628.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Rousseeuw, P. and Leroy, A. (1988). A robust scale estimator based on the shortest half. *Statist. Neerlandica* **42**, 103-116.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Wang, J.-G. (1987). A note on the uniform consistency of the Kaplan-Meier estimator. *Ann. Statist.* **15**, 1313-1316.

Dept. of Mathematics
Catholic University Leuven
Celestijnenlaan 200 B
3001 Heverlee
Belgium

Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands

List of COSOR-memoranda - 1993

Number	Month	Author	Title
93-01	January	P. v.d. Laan C. v. Eeden	Subset selection for the best of two populations: Tables of the expected subset size
93-02	January	R.J.G. Wilms J.G.F. Thiemann	Characterizations of shift-invariant distributions based on summation modulo one.
93-03	February	Jan Beirlant John H.J. Einmahl	Asymptotic confidence intervals for the length of the shortt under random censoring.