A new non-parametric estimator of the cumulative distribution function under time-and random-censoring

N. Balakrishnan, Christian Paroissin, Magdalena Pereda Vivo

A new non-parametric estimator of the cumulative distribution function under time- and random-censoring

N. Balakrishnan[a], Ch. Paroissin[b], M. Pereda Vivo[c]

[a] *Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada, bala@mcmaster.ca*
[b] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France, christian.paroissin@univ-pau.fr*
[c] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France, mpvivo@univ-pau.fr*

Abstract

In this paper, we first provide a review of different non-parametric estimators for the cumulative distribution function under left-censoring. We then propose a new estimator based on a non-parametric likelihood approach using reversed hazard rate. Finally, we conclude with an application to a real data.

Keywords: Left-censoring, Limit of dection (LOD), Non-parametric likelihood method, Reversed hazard rate
2000 MSC: 62G05, 62N01

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945416.

1. Introduction

When dealing with data analysis, we often have to deal with some censoring situations which arise when, for some units, one has only partial information. For instance, when dealing with lifetime data, some duration may not be observed exactly since the event occurs later than a certain time point. A typical case is when one perform a medical study over a given period: all the lifetimes longer than this period then get censored. Such a situation is known as right-censoring and it has been investigated rather extensively in the literature. Sometimes, censoring may occur on the left. For instance,

when dealing with concentration measurements with an analytical method, one will observe an exact measurement only if it is larger than a certain threshold, called limit of detection; otherwise, one has only the information that the concentration lies between zero and this limit. Such a situation is called left-censoring. However, statistical methods and models for data subject to left censoring have received comparatively less attention in the literature.

In this paper, we consider both time- and random left-censoring schemes. For n experimental units, the quantity of interest (lifetime, concentration, etc.) is observed exactly only if it is greater than a certain threshold value. It is assumed that the observations are independent and are drawn from the same unknown distribution. Let $T_1, \ldots, T_n$ be a sample from an unknown underlying distribution F. Let $C_1, \ldots, C_n$ be the censoring values: these could be either deterministic in the case of time left-censoring or random in the case of random left-censoring. In the former case, the censors may be all equal or may not be equal (for instance, if there is multiple sources of censoring). In the latter case, $C_1, \ldots, C_n$ are assumed to be sample from an unknown underlying distribution, and are independent of $T_1, \ldots, T_n$. Thus, observations are $(X_1, \Delta_1), \ldots, (X_n, \Delta_n)$, where

$$\forall i \in \{1, \ldots, n\}, \quad X_i = \max(T_i, C_i) \quad \text{and} \quad \Delta_i = \mathbb{I}_{T_i \geqslant C_i}.$$

To deal with left-censored data, an easy, but a naive, approach involves replacing the censored data by anything between 0 and this censored value. To fix the idea, let us consider the case of measuring concentrations. The used instrumentation may not provide exact value if it is below a certain known level, called limit of detection (LOD). Assume that there is a single LOD. Then, the main substitution methods are the following ones: replace any observation below the LOD by 0, by LOD/2, by LOD/$\sqrt{2}$ or by LOD (see, for instance, Hornung and Reed [8]). Of course, if one wishes to estimate the mean concentration, the first method will clearly under-estimate it, while the last method will over-estimate it. In 2010, Helsel [7] recommended against using such an approach.

Since most of the papers deal with right-censoring, some alternative solutions have been proposed in the literature by transforming the data in order to switch left- to right-censoring. For instance, the following transformations have been considered: (a) $Y_i = A - X_i$ with A large enough; (b) $Y_i = 1/X_i$; (c) $Y_i = -X_i$. Here, we prefer to consider a direct analysis

of data with left-censoring. For the right-censoring case, the most popular non-parametric estimator of the survival function has been the one due to Kaplan and Meier [9]. Mimicking the construction of this estimator, several authors have proposed the so-called product-limit estimator for the cumulative distribution function under left-censoring situations. As we will see, some of these papers contain some mistakes. Further, some researchers have derived a non-parametric estimator for the cumulative distribution function by using a counting process approach.

The rest of this paper proceeds as follows. In Section 2, we introduce several notations that will be used subsequently. Section 3 is devoted to a review of some non-parametric estimators for the cumulative distribution function under time- and random left-censoring. A pointwise estimation of the variance of the estimator is also provided. In Section 4, we introduce a new non-parametric estimator for the cumulative distribution function based on non-parametric likelihood function. This estimator is then compared to the existing ones. Finally, in Section 6, a real-life data is analysed using the proposed estimators.

2. Notations

In this section, we introduce some notations that we will be used in the sequel.

- n is the number of observations (exact or left-censored);
- $T_1, \ldots, T_n$ are the exact measurements (not always observed);
- $C_1, \ldots, C_n$ are the censoring values (not always observed);
- $X_1, \ldots, X_n$ are the observed values (exact or left-censored);
- $\delta_1, \ldots, \delta_n$ are the indicators of the observation of exact values;
- m is the number of distinct (exact or not) observations;
- $x_{(1)} < \cdots < x_{(n)}$ are the ordered distinct (exact or not) observations:

$$x_{(1)} = \min\{x_i\} < x_{(2)} < \cdots < x_{(m)} = \max\{x_i\};$$

- for any $k \in \{1, \ldots, m\}$, d_k is the number of exact and observed measurements equal to $x_{(k)}$:

$$d_k = \#\{i \in \{1, \ldots, n\} : x_i = x_{(k)} \text{ and } \delta_i = 1\};$$

- for any $k \in \{1, \ldots, m\}$, q_k is the number of left-censored and observed measurements equal to $x_{(k)}$:

$$q_k = \#\{i \in \{1, \ldots, n\} : x_i = x_{(k)} \text{ and } \delta_i = 0\};$$

- for any $k \in \{1, \ldots, m\}$, y_k is the number of observations less than or equal to $x_{(k)}$:

$$y_k = \#\{i \in \{1, \ldots, n\} : x_i \leqslant x_{(k)}\} = \sum_{j=1}^{k} (d_j + q_j);$$

- l is the number of distinct exact observations;

- $x^\star_{(1)} < \cdots < x^\star_{(l)}$ are the ordered distinct exact observations:

$$x^\star_{(1)} = \min\{x_i \ ; \ \delta_i = 1\} < x^\star_{(2)} < \cdots < x^\star_{(l)} = \max\{x_i \ ; \ \delta_i = 1\};$$

- for any $k \in \{1, \ldots, l\}$, $d^\star_k$ is the number of exact measures equal to $x^\star_{(k)}$:

$$d^\star_k = \#\{i \in \{1, \ldots, n\} : x_i = x^\star_{(k)} \text{ and } \delta_i = 1\};$$

  note that $d^\star_k > 0$ since, by definition, there is at least one exact observation equal to $x^\star_{(k)}$;

- for any $k \in \{1, \ldots, l\}$, $y^\star_k$ is the number of observations less than or equal to $x_{(k)}$:

$$y^\star_k = \#\{i \in \{1, \ldots, n\} : x_i \leqslant x^\star_{(k)}\};$$

- for any $k \in \{1, \ldots, l\}$, $\tilde{d}^\star_k$ is the number of (exact or not) measures equal to $x^\star_{(k)}$:

$$\tilde{d}^\star_k = \#\{i \in \{1, \ldots, n\} : x_i = x^\star_{(k)} \text{ and } \delta_i = 1\} \geqslant d^\star_k.$$

3. Review of some non-parametric estimators of cumulative distribution function

For time- and random-censored data, many different non-parametric estimators have been discussed for the cumulative distribution function (CDF) in the literature. Here, we present a brief review of estimators detailing two different approaches, with the first one being based on the chain rule and the second being based on counting processes. As we will see, these two approaches lead to the same estimator.

3.1. Estimator(s) based on the chain rule

Using chain rule, we can express the cumulative distribution function at point $x_{(j)}$ with $j \in \{1, \ldots, m-1\}$, based on points $x_{(j+1)}, \ldots, x_{(m)}$, as follows:

$$F(x_{(j)}) = \mathbb{P}[T \leqslant x_{(j)}] = \prod_{k=j}^{m-1} \mathbb{P}[T \leqslant x_{(k)} | T \leqslant x_{(k+1)}] = \prod_{k=j}^{m-1} p_k.$$

Hence, a natural estimator can be obtained by replacing $p_1, \ldots, p_{m-1}$ by some estimators $\widehat{p}_1, \ldots, \widehat{p}_{m-1}$ (since $p_1 + \cdots + p_m = 1$, $\widehat{p}_m = 1 - \widehat{p}_1 - \cdots - \widehat{p}_{m-1}$). Following the seminal work of Miller [11], Blackwood [2] proposed the following estimator for p_j:

$$\widehat{p}_j = \left(1 - \frac{d_j}{y_j}\right)^{\delta_{(j)}},$$

where y_j and d_j are as defined in Section 2, and

$$\delta_{(j)} = \begin{cases} 1 & \text{if at least one observation at time } x_{(j)} \text{ is uncensored} \\ 0 & \text{if all observations at time } x_{(j)} \text{ are censored.} \end{cases}$$

Notice that $\delta_{(j)} = 1$ (resp. $\delta_{(j)} = 0$) if, and only if, $d_j \geqslant 1$ (resp. $d_j = 0$). Blackwood [2] claimed that $\widehat{p}_j$ is the maximum likelihood estimator of p_j (since it is the case for the right-censoring situation). Following what is classically done for the right-censoring case (see, for instance, [11]), we can consider d_j to be a realization of a random variable D_j, assumed to be binomially distributed with parameters y_j (observed) and p_j (unobserved and unknown). It leads to the following estimator of the CDF:

$$\forall t \geqslant 0, \quad \widehat{F}(t) = \prod_{j; x_{(j)} > t} \left(1 - \frac{d_j}{y_j}\right)^{\delta_{(j)}}. \tag{1}$$

As $\delta_{(j)} = 0$ is equivalent to $d_j = 0$, then this estimator can be expressed only by considering unique uncensored observations. Using the notations introduced earlier in Section 2, we have

$$\forall t \geqslant 0, \quad \widehat{F}(t) = \prod_{k; x^{\star}_{(k)} > t} \left(1 - \frac{d^{\star}_k}{y^{\star}_k}\right). \tag{2}$$

Blackwood [2] also presented an estimator of the variance by applying the same approach as the one leading to the Greenwood formula for the right-censoring case. He provided the following expression:

$$\widehat{var}\left[\widehat{F}(t)\right] = \left[\widehat{F}(t)\right]^2 \sum_{j; x_{(j)} > t} \frac{d_j \delta_{(j)}}{y_j(y_j - d_j)}. \tag{3}$$

It should be mentioned that Blackwood [2] also proposed non-parametric estimators for the means and the quantiles based on the non-parametric estimator of the CDF.

Despite this work of Blackwood [2], some researchers from other fields such as environmental science and physics have come up with the same estimator. Besides, Pajek *et al.* [12] have also considered another estimator based on a non-parametric estimator of the cumulative hazard function (CHF), such as the Nelson-Aalen and the Harrington-Flemming estimators. Unfortunately, this work contains some mistakes. It starts with an imprecise definition of the cumulative hazard function Λ (bounds of the integral are not given and a confusion caused by using the same variable for the function Λ and the integrand). According to Equation (5) in [12] providing a non-parametric estimator $\widehat{\Lambda}$ of Λ, it seems that, in fact, they are rather considering the cumulative reversed hazard function (CRHF). It can be seen through the fact that $\widehat{\Lambda}$ is a decreasing function, which is not the case for the CHF, but is true in fact for the CRHF. As a consequence, the estimator proposed in Equation (6) in [12] is incorrect. Indeed, they have used the relationship between the CHF and the survival function (as is usually done in the right censoring case, from Nelson-Aalen estimator to Harrington-Flemming estimator). But, since they have in fact an estimator of the CRHF, they should have used the relationship between the CRHF and the CDF. This way, the correct estimator should be (using our notations) as

$$\forall t \geqslant 0, \quad \widehat{F}(t) = \prod_{k; x^{\star}_{(k)} > t} \exp\left(-\frac{d^{\star}_k}{y^{\star}_k}\right).$$

3.2. Estimator based on counting processes

To the best of our knowledge, it seems that Gomez *et al.* [5] (see also [6]) were the first to propose an estimator based on counting processes. Quite surprisingly, they have derived a non-parametric estimator for the survival function of T (and not for the cumulative distribution function which is more natural to consider when dealing with left censored data). They have expressed the survival function as an integral equation. By replacing other unknown functions involved in this integral equation by their empirical counterparts, they defined an estimator of the survival function of T. In this way, this estimator appears to be the solution of a backward Doléans equation for which the solution can be determined explicitly in the present case. Later on, Tressou [16] proposed a new formulation of this estimator using the formalism developed by Gill and Johansen [3]. Tressou [16] considered only the case of random left-censoring and denotes by G the CDF for the censoring part. For any $t \geqslant 0$, let

$$\mathbb{H}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{x_i \leqslant t} \quad \text{and} \quad \mathbb{H}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{x_i \leqslant t;\delta_i=1}$$

be, respectively, the empirical versions of $H(t) = \mathbb{P}[X \leqslant t]$, the CDF of the (uncensored or censored) observations, and $H_1(t) = \mathbb{P}[X \leqslant t; \Delta = 1]$, the CDF of the uncensored observations. The reversed hazard rate can be defined as

$$R(t) = \int_{]t,\infty]} \frac{\mathrm{d}F}{F} = \int_{]t,\infty]} \frac{\mathrm{d}H_1}{H}.$$

Now, using the product integral function Ψ, we have $F = \Psi(R)$; see [3]. It then follows that a non-parametric estimator of F is given by

$$\widehat{F} = \prod_{]\cdot,\infty]} \left(1 - \mathrm{d}\widehat{R}\right) = \prod_{]\cdot,\infty]} \left(1 - \frac{\mathrm{d}\mathbb{H}_{1,n}}{\mathbb{H}_n}\right).$$

From the above expression for $\mathbb{H}_n$ and for $\mathbb{H}_{1,n}$, we get

$$\forall t \geqslant 0, \quad \widehat{F}(t) = \prod_{k=1}^{l} \left(1 - \frac{d_k^\star}{y_k^\star}\right)^{\mathbb{I}_{x_{(k)}^\star > t}}.$$

Notice that this estimator coincides with the one given by Patilea and Rolin [14] under the double censoring scheme when there are no right-censored

observations. Using the framework of Gill and Johansen [3], we can derive an estimate for the variance of $\widehat{F}(t)$ as

$$\widehat{\text{var}}\left(\widehat{F}(t)\right) = \left[\widehat{F}(t)\right]^2 \sum_{k=1}^{l} \frac{d_k^\star \mathbb{I}_{x_{(k)}^\star > t}}{y_k^\star (y_k^\star - d_k^\star)}.$$

Observe that this estimator based on counting processes is indeed the same as the one obtained by using the chain rule.

4. A new non-parametric estimator

Our goal here is to develop an estimator of the CDF based on non-parametric likelihood function; it has been done for the survival function in the case of right-censoring, but not an estimator for the CDF based on left-censoring. After providing an expression of the non-parametric likelihood function in terms of reversed hazard rate (RHR), we derive an estimator for the CDF.

Let data $= \{(x_1, \delta_1), \ldots, (x_1, \delta_n)\}$ be the set of all observations. We can then write the likelihood function as

$$L(p_1, \ldots, p_m; \text{data}) = \prod_{i=1}^{n} p_{x_i}^{\delta_i} F_{x_i}^{1-\delta_i},$$

where $F_k = p_1 + \cdots + p_k$ is the CDF of the discrete distribution. Let us recall that $F_0 = 0$ and $F_m = 1$. Using the notations introduced in Section 2, we can now express L as

$$L(p_1, \ldots, p_m; \text{data}) = \prod_{k=1}^{m} p_k^{d_k} F_k^{q_k} = \prod_{k=1}^{m} p_k^{d_k} \left(\sum_{j=1}^{k} p_j\right)^{q_k}. \tag{4}$$

However, this expression is not tractable for optimizing with respect to $p_1, \ldots, p_m$. Instead of expressing the likelihood function in term of mass probabilities, we will rather use the notion of RHR defined as in [1]:

$$\forall k \in \{1, \ldots, m\}, \quad r_k = \frac{\mathbb{P}[T = x_{(k)}]}{\mathbb{P}[T \leqslant x_{(k)}]} = \frac{p_k}{F_k}.$$

Note that we have $r_1 = 1$ and $r_m = p_m$. As we have $p_k = F_k - F_{k-1}$, with the convention that $F_0 = 0$, we have the following relationship:

$$\forall k \in \{1, \ldots, m\}, \quad r_k = \frac{F_k - F_{k-1}}{F_k} = 1 - \frac{F_{k-1}}{F_k}.$$

By induction, we then obtain

$$F_{k-1} = (1 - r_k)F_k = (1 - r_k)(1 - r_{k+1})F_{k+1} = \cdots = \prod_{j=k}^{m}(1 - r_j)$$

since $F_m = 1$. We can now rewrite the likelihood function with respect to $r_2, \ldots, r_m$ (keeping in mind that $r_1 = 1$). Because $p_k = r_k F_k$, Equation (4) turns to be

$$L(r_2, \ldots, r_m; \text{data}) = \prod_{k=2}^{m} r_k^{d_k} F_k^{d_k+q_k} = \prod_{k=2}^{m-1} r_k^{d_k} \left[\prod_{j=k+1}^{m} (1 - r_j)\right]^{d_k+q_k} = \prod_{k=2}^{m} r_k^{d_k}(1-r_k)^{y_{k-1}}.$$

Hence, the log-likelihood function takes os the following form:

$$\ell(r_2, \ldots, r_m; \text{data}) = \sum_{k=2}^{m}\{d_k \log r_k + y_{k-1} \log(1 - r_k)\}.$$

It turns that the value of r_k that maximizes the log-likelihood function to be

$$\forall k \in \{2, \ldots, m\}, \quad \widehat{r}_k = \frac{d_k}{d_k + y_{k-1}} = \frac{d_k}{y_k - q_k}$$

since $y_k = y_{k-1} + d_k + q_k$. We thus obtain the following estimator for F_{k-1}:

$$\widehat{F}_{k-1} = \prod_{j=k}^{m}\left(1 - \frac{d_j}{d_j + y_{j-1}}\right) = \prod_{j=k}^{m}\frac{y_{j-1}}{d_j + y_{j-1}} = \prod_{j=k}^{m}\left(1 - \frac{d_j}{y_j - q_j}\right),$$

using which we obtain the following non-parametric estimator of the CDF:

$$\forall t \geqslant 0, \quad \widehat{F}^{(1)}(t) = \prod_{j; x_{(j)} > t}\left(1 - \frac{d_j}{d_j + y_{j-1}}\right) \tag{5}$$

with the convention that $\prod_{\emptyset} = 0$ (meaning that if $t < x_{(1)}$, then $\widehat{F}^{(1)}(t) = 0$). Observe that for all j such that $d_j = 0$ (and, of course, $q_j > 0$), then $1 - d_j/(d_j + y_{j-1}) = 1$. Thence, the estimator can be defined only at points $x^\star_{(1)}, \ldots, x^\star_{(l)}$, and we then obtain

$$\forall t \geqslant 0, \quad \widehat{F}^{(1)}(t) = \prod_{k; x^\star_{(k)} > t}\left(1 - \frac{d^\star_k}{d^\star_k + y^\star_{k-1}}\right) = \prod_{k; x^\star_{(k)} > t}\left(1 - \frac{d^\star_k}{y^\star_k - q^\star_k}\right). \tag{6}$$

The last expression can be interpreted as follows. As claimed in some papers cited in Section 2 (see [4] and [15], for instance), if there is a tie between a censored and an uncensored observations, then it is assumed that the censored value is slightly smaller than the uncensored value. In such a case, when considering $y_k^\star - q_k^\star$, we remove these censored observations of the set of individuals at-risk.

Let us compare this new estimator with the one reviewed in the last section (recall that the two approaches considered previously lead to the same estimator). Let $t \geqslant 0$ be fixed and let us then consider the ratio

$$\frac{\widehat{F}^{(1)}(t)}{\widehat{F}(t)} = \prod_{k;x_{(k)}^\star > t} \left(1 - \frac{d_k^\star}{d_k^\star + y_{k-1}^\star}\right) / \left(1 - \frac{d_k^\star}{y_k^\star}\right)$$

Because $y_{k-1}^\star + d_k^\star = y_k^\star - q_k^\star \leqslant y_k^\star$ for any $k \in \{1, \ldots, l\}$, one can easily see that $\widehat{F}^{(1)}(t) \leqslant \widehat{F}(t)$. Let us consider the special case when there are no censored measurements. In such a case, we have $q_k^\star = 0$ for all $k \in \{1, \ldots, l\}$ (and $l = m$). In this case, the two estimators, $\widehat{F}^{(1)}$ and $\widehat{F}$ are identical.

The factors in the two products defining the former estimator and this new estimator differ only at points where there is both censored and uncensored measurements. It can be expected that this may occur essentially when dealing with random censoring (and with rounded values). Because products are defined from right to left, these two estimators will be different only on the lower tail. However, for left-censoring, the main issue is to estimate accurately the left tail.

We now seek an estimator for the variance of $\widehat{F}_T^{(1)}(t)$ for a given value of t. For this, we assume that $(\widehat{r}_2, \ldots, \widehat{r}_m)$ is an asymptotically normal estimator of $(r_2, \ldots, r_m)$, with asymptotic covariance matrix equal to the inverse of the Fisher information. For every $k \in \{2, \ldots, m\}$, we have

$$\frac{\partial^2 \ell}{\partial r_k^2}(\widehat{r}_2, \ldots, \widehat{r}_m; \text{data}) = -\frac{(y_k - q_k)^3}{d_k(y_k - d_k - q_k)}.$$

So, we can conclude that

$$\text{var}[\widehat{r}_k] \approx \frac{d_k(y_k - d_k - q_k)}{(y_k - q_k)^3}.$$

Using classical approximations for the variance based on the delta method,

one can get that

$$\widehat{\text{var}}\left[\widehat{F}^{(1)}(t)\right] = \left[\widehat{F}^{(1)}(t)\right]^2 \sum_{k:X_{(k)}>t} \frac{d_k}{y_{k-1}(y_k - q_k)}.$$

Note that, as for the Nelson-Aalen estimator of the CHF, one can use the above results to derive a non-parametric estimator of the CRHF and deduce another non-parametric estimator of the CDF (corresponding to Harrington-Flemming estimator in the right-censoring case) thanks to the relation between CRHF and CDF.

5. Application to a real-life data

In this section, we use the different estimators discussed in the previous sections to analyze a real data relating to pollutants in water, with measurements being subject to left-censoring with one or multiple limit of detection (LOD) values. Here, we consider copper concentrations in shallow groundwater samples from a Basin-Through zone in the San Joaquin Valley, California (see [10]), while studying groundwater quality. This dataset includes five different limits of detection: 1, 2, 5, 10 and 15. There are multiple limits of detection because it depends on the method used for measuring the amount of dilution and also because it may be decreasing over time as measurement gets improved. In Table 1, we have reported pointwise estimation of the CDF and its standard deviation, for the Blackwood estimator and for the newly proposed estimator. We observe that at some points, the two estimators are slightly different, these points corresponding to values with both censored and uncensored measurements. As we can observe, the two estimators are the same on the right part and differ from point 15, which is the largest value corresponding to both an exact measurement and a LOD. Below this point, the newly proposed estimator is slightly lower than the Blackwood estimator. This means that the estimate of the mean concentration will be less than the one obtained with the Blackwood estimator. As 15 is the largest value corresponding to a LOD, the two estimators of the variance are equal for the same reason as stated above. Below this point, the estimate of the variance of the newly proposed estimator is slightly larger than the estimate of the variance of the Blackwood estimator, except for $t = x^{\star}_{(2)}$.

t	$\widehat{F}(t)$	$\widehat{F}^{(1)}(t)$	$\widehat{\sigma}(\widehat{F}(t))$	$\widehat{\sigma}(\widehat{F}^{(1)}(t))$
1	0.2981959	0.2799105	0.07438262	0.07541081
2	0.4066308	0.4043151	0.07924497	0.07922304
3	0.6235005	0.6199498	0.07582786	0.07644654
4	0.7590441	0.7547215	0.06362657	0.06510580
5	0.7820455	0.7816759	0.06125617	0.06159916
6	0.8280481	0.8276568	0.05555525	0.05598826
8	0.8510495	0.8506473	0.05211982	0.05261188
9	0.8970522	0.8966282	0.04362071	0.04428404
12	0.9179138	0.9174800	0.03933148	0.03953237
14	0.9387755	0.9383319	0.03424881	0.03449597
15	0.9591837	0.9591837	0.02826635	0.02826635
17	0.9795918	0.9795918	0.02019884	0.02019884

Table 1: Pointwise estimation of the CDF and its standard deviation, for the Blackwood estimator and for the newly proposed estimator.

References

[1] Asha, G., Elbatal, I., Rejeesh, C.J., 2016. Further results on discrete mean past lifetime. Communications in Statistics – Theory and Methods 45, 1081–1098.

[2] Blackwood, L.G., 1991. Analyzing censored environmental data using survival analysis: Single sample techniques. Environmental Monitoring and Assessment 18, 25–40.

[3] Gill, R.D., Johansen, S., 1990. A survey of product integration with a view toward application in survival analysis. The Annals of Statistics 18, 1501–1555.

[4] Gillespie, B.W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., Adriaens, P., Demond, A., Luksemburg, W., Garabrant, D.H., 2010. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. Epidemiology 21, S64–70.

[5] Gomez, G., Julià, O., Utzet, F., Moeschberger, M.L., 1992. Survival Analysis For Left Censored Data. Springer Netherlands, Dordrecht.

[6] Gómez, G., Julià, O., Utzet, F., 1994. Asymptotic properties of the left Kaplan-Meier estimator. Communications in Statistics - Theory and Methods 23, 123–135.

[7] Helsel, D., 2010. Much ado about next to nothing: incorporating nondetects in science. The Annals of Occupational Hygiene 54, 257–262.

[8] Hornung, R.W., Reed, L.D., 1990. Estimation of average concentration in the presence of nondetectable values. Applied Occupational and Environmental Hygiene 5, 46–51.

[9] Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481.

[10] Millard, S.P., Deverel, S.J., 1988. Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits. Water Resources Research 24, 2087–2098.

[11] Miller, J., Rupert, G., 1981. Survival analysis. Wiley, New York.

[12] Pajek, M., Kubala-Kukuś, A., Banaś, D., Braziewicz, J., Majewska, U., 2004a. Random left-censoring: a statistical approach accounting for detection limits in x-ray fluorescence analysis. X-Ray Spectrometry 33, 306–311.

[13] Pajek, M., Kubala-Kukuś, A., Braziewicz, J., 2004b. Censoring: a new approach for detection limits in total-reflection x-ray fluorescence. Spectrochimica Acta Part B 59, 1091–1099.

[14] Patilea, V., Rolin, J.M., 2006. Product-limit estimators of the survival function with twice censored data. The Annals of Statistics 34, 925–938.

[15] Popovic, M., Nie, H., Chettle, D.R., McNeill, F.E., 2007. Random left censoring: a second look at bone lead concentration measurements. Physics in Medicine & Biology 52, 5369.

[16] Tressou, J., 2006. Nonparametric modeling of the left censorship of analytical data in food risk assessment. Journal of the American Statistical Association 101, 1377–1386.

Supplementary material of "New non-parametric estimators of the cumulative distribution function under time- and random-censoring"

N. Balakrishnan[a], Ch. Paroissin[b], M. Pereda Vivo[c]

[a] *Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada, bala@mcmaster.ca*
[b] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France, christian.paroissin@univ-pau.fr*
[c] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France, mpvivo@univ-pau.fr*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945416.

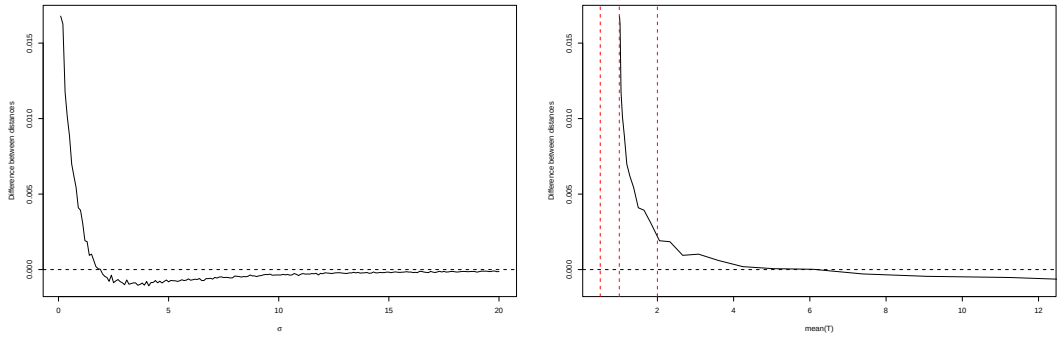
1. Empirical evaluation

We conduct here a numerical comparison of the estimators, based on Monte Carlo simulations. We consider two cases, here. The first one deals with multiple deterministic limit of detection (LOD) values corresponding to time left-censoring scheme. The second one corresponds to random left-censoring scheme in which censored values are assumed to be realizations of random variables.

1.1. Time left-censoring

For the exact measurements $T_1, \ldots, T_n$, we have used the log-normal distribution with parameters (μ, σ), and several values of the parameters are then considered. For the LOD C, three possible levels, namely 0.5, 1, and 2, are used in order to consider the case of multiple LODs. For each unit in the sample, one of these LODs have been selected randomly with equal probability. The sample size n has been fixed to be 50.

First, we fixed μ and then considered different values of σ. For each set of parameters, we computed the estimator in Equation (**??**) of the CDF of T



(a) (b)

Figure 1: (a) Time-censoring on the left with the choices of $\mu = 0$ and $\sigma \in (0, 20)$; (b) Time-censoring on the left with the choices of $\mu = 0$ and $\sigma \in (0, 2.25)$

and the one in Equation (**??**). In order to compare the performance of these two estimators, we computed the Kolmogorov-Smirnov distance between each estimator and the exact CDF as follows :

$$d_{ks} = \sup_{t \in \{x^\star_{(1)}, \cdots, x^\star_{(l)}\}} |F(t) - \widehat{F}(t)|$$

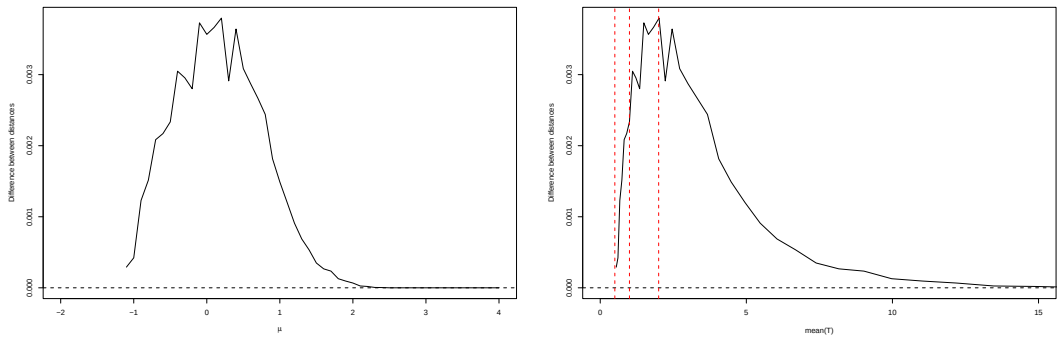
and

$$d^{(1)}_{ks} = \sup_{t \in \{x^\star_{(1)}, \cdots, x^\star_{(l)}\}} |F(t) - \widehat{F}^{(1)}(t)|,$$

where F is the CDF of the log-normal distribution with parameters (μ, σ). We repeated these steps $m = 10,000$ times and then calculated the average of the difference $d_{ks} - d^{(1)}_{ks}$ resulting from the m simulations. If this average difference is positive (resp. negative), it means that globally the newly proposed estimator of F is more (resp. less) accurate than the Blackwood estimator. To observe the behaviour, we first fixed μ to a given value and let σ vary and, next, we fixed σ to a given value and let μ vary.

In Figure 1, we have plotted this average difference when $\mu = 0$ and σ is varying between 0.1 and 20. We observe that the average difference is positive, next turns to be negative and then converges to zero as σ increases.

In Figures 2 and 3, we have fixed σ respectively to 1 and 2, while μ varies between -2 and 4. We observe that when $\sigma = 1$ (see Figure 2), the average



(a) (b)

Figure 2: (a) Time-censoring on the left with the choices of $\sigma = 1$ and $\mu \in (-2, 4)$; (b) Time-censoring on the left with the choices of $\sigma = 1$ and $\mu \in (-2, 2)$

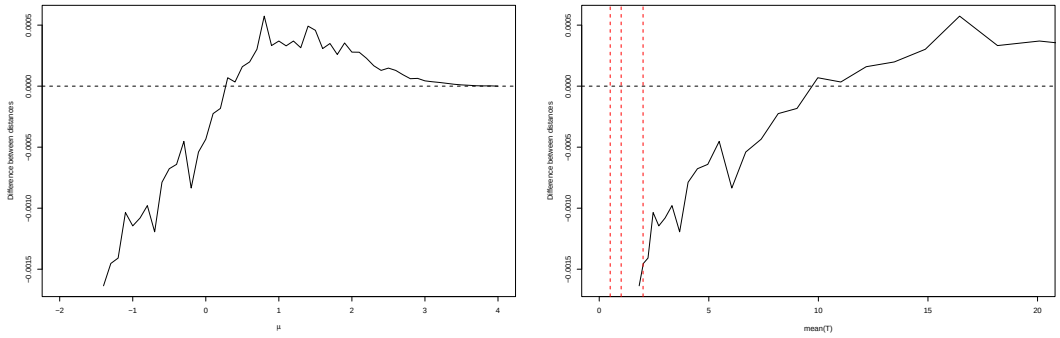
difference is always positive and goes to zero as μ increases. The situation is quite different when $\sigma = 2$ (see Figure 3). In this case, the average difference is negative, turns to be positive and then goes to zero as μ increases.

We thus see that when one of the two parameters of the log-normal distribution is fixed, the average difference tends to zero as the other parameter increases. This can be explained as follows. The expectation of T is equal to $\exp(\mu + \sigma^2/2)$. Hence, when μ and/or σ is increasing, the expectation of T is increasing and thus one will observe more exact measurements (in other words, the probability of censoring decreases). It will make the two estimators to be closer since there will be less values under the LOD.

We obtained similar plots for $n = 100$, and the sample size does not seem to affect this behaviour.

1.2. Random left-censoring

As explained above, in the case of random left-censoring, we assume that the censoring value C is also random. We consider log-normal distribution with parameters $\mu_C = 0$ and $\sigma_C = 1$ for the variable C. As in the previous case, we will observe how the average difference evolves according to μ and σ. In Figure 4, with $\mu = 0$ and $\sigma \in (0, 20)$, as in the first situation of time

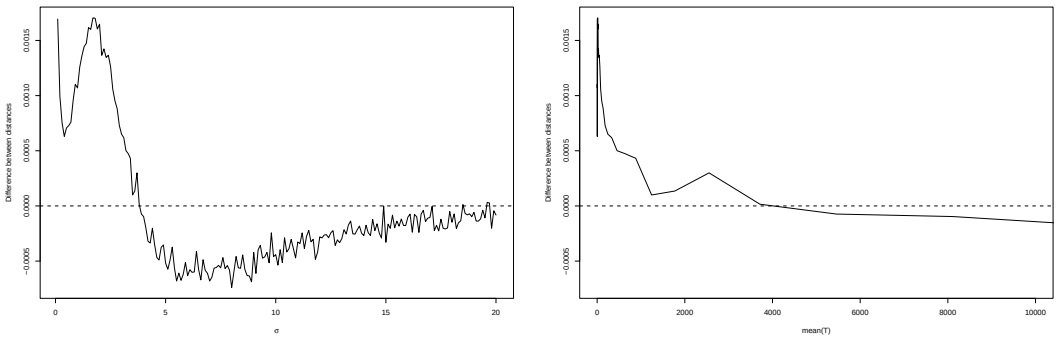


(a) (b)

Figure 3: (a) Time-censoring on the left with the choices of $\sigma = 2$ and $\mu \in (-2, 4)$; (b) Time-censoring on the left with the choices of $\sigma = 2$ and $\mu \in (-2, 2.5)$

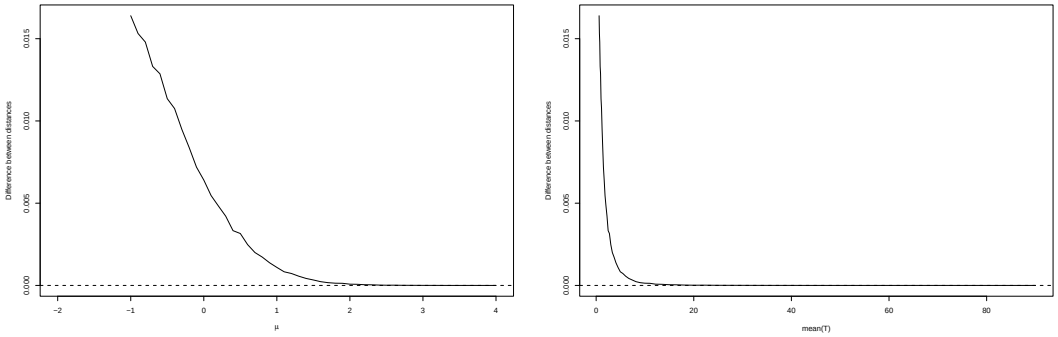
left-censoring, we observe exactly the same behaviour (positive, negative and limit to zero).

In Figure 5, we have set $\sigma = 1$ and $\mu \in (-2, 4)$. It should be noted that too small values of μ will lead to a data with only censored values in which case no estimator can be proposed. In this case, we observe the difference to be positive (still with a limit to zero).



(a) (b)

Figure 4: (a) Random left-censoring with the choices of $\mu = 0$ and $\sigma \in (0, 20)$; (b) Random left-censoring with the choices of $\mu = 0$ and $\sigma \in (0, 4.5)$.



(a) (b)

Figure 5: (a) Random left-censoring with the choices of $\sigma = 1$ and $\mu \in (-2, 4)$; (b) Random left-censoring with the choices of $\sigma = 1$ and $\mu \in (-2, 4)$.