

Testing for an ignorable sampling bias under random double truncation

Jacobo de Uña-Álvarez¹

CINBIO, Universidade de Vigo, Spain

Correspondence

Jacobo de Uña-Álvarez, CINBIO,
Universidade de Vigo, 36310 Vigo, Spain.
Email: jacobo@uvigo.es

Funding information

Ministerio de Ciencia e Innovación,
Grant/Award Number:
PID2020-118101GB-I00

In clinical and epidemiological research doubly truncated data often appear. This is the case, for instance, when the data registry is formed by interval sampling. Double truncation generally induces a sampling bias on the target variable, so proper corrections of ordinary estimation and inference procedures must be used. Unfortunately, the nonparametric maximum likelihood estimator of a doubly truncated distribution has several drawbacks, like potential nonexistence and nonuniqueness issues, or large estimation variance. Interestingly, no correction for double truncation is needed when the sampling bias is ignorable, which may occur with interval sampling and other sampling designs. In such a case the ordinary empirical distribution function is a consistent and fully efficient estimator that generally brings remarkable variance improvements compared to the nonparametric maximum likelihood estimator. Thus, identification of such situations is critical for the simple and efficient estimation of the target distribution. In this article, we introduce for the first time formal testing procedures for the null hypothesis of ignorable sampling bias with doubly truncated data. The asymptotic properties of the proposed test statistic are investigated. A bootstrap algorithm to approximate the null distribution of the test in practice is introduced. The finite sample performance of the method is studied in simulated scenarios. Finally, applications to data on onset for childhood cancer and Parkinson's disease are given. Variance improvements in estimation are discussed and illustrated.

KEYWORDS

bootstrap, goodness-of-fit, interval sampling, nonparametric statistics, survival analysis

1 | INTRODUCTION

The issue of random double truncation is ubiquitous in clinical and epidemiological research, among other fields. Double truncation appears when the observation of the target variable is restricted by two random limits. An example is found in Survival Analysis, when the target is an event time, and the data correspond to all the events between two specific dates.¹ Such sampling design has been often referred to as interval sampling.² Double truncation can be regarded as an extension of left-truncation, a well-known issue that affects cross-sectional data or registries with delayed entries.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Author. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Autopsy-confirmed diseases is a remarkable example of double truncation in such left-truncated settings, since undergoing the event of interest (death) before the end of the study becomes then a prerequisite for recruitment; this induces truncation from the right and, thus, results in doubly truncated event times.³

Unlike for one-sided (left or right) truncation, the nonparametric maximum likelihood estimator (NPMLE) under random double truncation does not have a closed form. Efron and Petrosian⁴ introduced the NPMLE of a cumulative distribution function (CDF) under random double truncation; the motivation was found in the analysis of quasar luminosities that are subject to random detection limits. These authors exploited the distinctive features of double truncation to introduce a self-consistency equation for Turnbull's⁵ estimator and, consequently, to propose an iterative algorithm for its calculation. The Efron-Petrosian NPMLE was further investigated by Shen,⁶ Moreira and de Uña-Álvarez,¹ Emura et al⁷ and, more recently, by de Uña-Álvarez and Van Keilegom,⁸ who developed asymptotic theory to cope with double truncation in the presence of covariates. On the other hand, several R packages implementing the NPMLE for doubly truncated data have been launched along the last decade, so the application of the Efron-Petrosian estimator has become easy; available packages include `DTDA`,⁹ `SurvTrunc`¹⁰ and `double.truncation`.¹¹ Importantly, the consistency of the Efron-Petrosian estimator depends on the assumption of quasi-independence between the target variable and the truncating variables. Testing procedures for a quasi-independence assumption in this setting have been investigated; see for instance Martin and Betensky.¹² On the other hand, Moreira et al¹³ introduced a copula-based extension of the Efron-Petrosian estimator to cope with possibly dependent truncation. Throughout this article it is assumed that the truncating variables are independent of the target.

The sampling bias induced by double truncation has been discussed and illustrated in a number of research papers. For instance, Zhu and Wang² investigated the sampling bias when estimating the birth rate from cancer registries obtained with interval sampling. In particular, they found that the plausible linear trend for the birth process was converted into a rather unrealistic bell-shaped curve due to the truncation effects. Mandel et al¹⁴ discussed the biases that may arise in the proportional hazards regression model when fitted from interval sampling data too. Specifically, when looking for a relationship between genetic information and age at onset for Parkinson's disease, they found substantial differences between the estimated coefficients with and without the correction for double truncation. The sampling bias for the Parkinson's disease study was further explored in de Uña-Álvarez,¹⁵ who concluded the oversampling of patients with intermediate ages at diagnosis. Rennert and Xie³ described the issue in the setting of autopsy-confirmed neurodegenerative diseases. Interestingly, they illustrated how by ignoring the double truncation one may overestimate or underestimate the target survival, depending on the particular situation. The impact of the sampling bias in the Mann-Whitney two-sample test was pointed out by these authors too. On the contrary, Moreira and de Uña-Álvarez¹ discussed how, with interval sampling, the sampling bias may vanish when the truncating variables are uniformly distributed; this was indeed the case for the childhood cancer registry investigated in the referred paper.

To sum up, random double truncation may, or may not, induce a sampling bias on the target variable. It is convenient to identify in practice whether such a potential sampling bias exist, at least due to the following reasons:

- When there is no sampling bias the empirical cumulative distribution function (ECDF) is a consistent (and efficient) estimator of the target CDF, so there is no need to look for alternative estimators;
- The NPMLE which corrects for double truncation may not exist, or may be nonunique;¹⁶ and
- The estimation of a distribution from doubly truncated data given by the corresponding NPMLE, when it exists and is unique, usually entails a large variance.

In other words, when there is no sampling bias one will have a strong motivation to apply the ECDF instead of the Efron-Petrosian NPMLE. And this is why testing for an ignorable sampling bias under double truncation becomes relevant. Note that this is irrelevant for one-sided truncation, where a sampling bias is always present: left-truncation results in an oversampling of relatively large event times; the opposite situation occurs under right-truncation.

The potential sampling bias under double truncation can be assessed from the joint graphical display of the ECDF and the Efron-Petrosian NPMLE. When both curves are close together one gets support for the hypothesis of ignorable sampling bias. Similarly, one may plot the NPMLE for the sampling probability to check whether it is roughly constant on the support of the target variable; see for instance Example 2.1.4 in de Uña-Álvarez et al.¹⁷ Formal testing methods that guarantee a given significance level are however missing in the literature at the time of writing. A related reference is Moreira et al,¹⁸ who introduced and investigated through simulations several statistics to test for a parametric model for the pair of truncating variables. Still, a key difference in the current setting is that the null hypothesis of ignorable sampling bias does not characterize the (bivariate) truncation distribution; see Section 2 for further details. This complicates the

application of the ideas in Moreira et al¹⁸ to test for no sampling bias. In particular, a new resampling plan to approximate the null distribution of the introduced test statistic will be needed.

The interest in the random double truncation model has rapidly increased in the last years. Recent methods to handle doubly truncated outcomes include, among other topics, smoothing methods,^{19,20} proportional hazards regression,^{14,21} rank regression for linear models,²² competing risks,²³ two-sample problems,²⁴ the estimation of a bivariate distribution,^{2,25,26} or maximum likelihood theory for parametric models.^{27,28} Formally testing for an ignorable sampling bias is important in all these settings with double truncation because, when there is no sampling bias, ordinary methods apply and the estimation variance can be reduced.

The rest of the article is organized as follows. In Section 2 we introduce the required notation and a test statistic for the null hypothesis of ignorable sampling bias. The proposed test is defined as the supremum distance between the Efron-Petrosian NPMLE and the ECDF of the doubly truncated outcomes. The asymptotic null distribution of the test statistic is obtained, and a bootstrap algorithm to approximate the null distribution in practice is proposed. In Section 3 a simulation study to investigate the finite sample performance of the proposed test is conducted. In Section 4 illustrative applications of the proposed methods to data on childhood cancer and Parkinson's disease are given. A final discussion that mentions some possible extensions of the introduced methods and alternative testing approaches for an ignorable sampling bias is given in Section 5.

2 | METHODS

2.1 | Notation and preliminary remarks

Let X be the target variable with CDF F , assumed to be supported on an interval $[a_X, b_X] \subset \mathcal{R}$ or, more generally, on a set \mathcal{S}_X with lower and upper limits a_X and b_X . Let (U, V) be a couple of truncating variables independent of X with (bivariate) CDF K , so the triplet (X, U, V) is observed only when $U \leq X \leq V$. It is assumed that $\alpha \equiv P(U \leq X \leq V)$ is strictly positive. In this setting, the observation of X is limited by the condition $a_U \leq X \leq b_V$, where a_U and b_V are the lower and upper endpoints of the supports of U and V , respectively. Then, it is natural to impose the restrictions $a_U \leq a_X$ and $b_X \leq b_V$.²⁹ If these conditions are violated, no consistent estimation of F can be provided. Similarly, one must assume $P(U \leq V) = P(U \leq b_X) = P(V \geq a_X) = 1$ for the identifiability of K .

The sampling information is a set of independent and identically distributed (iid) observations (X_i, U_i, V_i) , $1 \leq i \leq n$, with the conditional distribution of (X, U, V) given $U \leq X \leq V$. In general, the ECDF of the X_i 's, $F_n^*(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$, is not consistent for $F(x)$. This is because the CDF of X_1 is given by

$$F^*(x) = \alpha^{-1} \int_{a_X}^x G(t) dF(t) \quad (1)$$

with $G(t) = P(U \leq t \leq V) = \int_{u \leq t \leq v} dK(u, v)$ and, thus, it does not coincide with F unless G is constant. Note that the function G reports the sampling probability for the values of the target X ; that is, $G(t)$ is the probability of observing $X = t$. When G is strictly positive on \mathcal{S}_X , (1) admits the reverse formulation

$$F(x) = \alpha \int_{a_X}^x G(t)^{-1} dF^*(t) \quad (2)$$

and $\alpha = 1 / \int_{a_X}^{b_X} G(t)^{-1} dF^*(t)$ holds. Assumption (C1) below ensures that G is strictly positive, provided that the densities of the truncating variables exist and have convex support; this implies in particular that there are no regions within $[a_U, b_V]$ where F is not identifiable. Equation (2) reveals that F can be estimated from F_n^* and a consistent estimator for G . Actually, this is the classical idea behind inverse-probability-weighting estimation, where each datum X_i is weighted according to its inverse sampling probability $1/G(X_i)$. The Efron-Petrosian NPMLE⁴ F_n can be derived in this way, since it corresponds to (2) with F^* replaced by F_n^* and G replaced by its NPMLE G_n .⁶ The asymptotic properties of F_n and G_n were recently detailed by de Uña-Álvarez and Van Keilegom,⁸ including uniform consistency and weak convergence results.

Unfortunately, G_n does not have a closed-form; besides, the NPMLE F_n may be nonunique, or even nonexistent.¹⁶ This motivates the interest in testing for the null hypothesis $\mathcal{H}_0 : G(x) = \alpha, x \in \mathcal{S}_X$. Under \mathcal{H}_0 the ECDF F_n^* is consistent

for F , so there is no need to work with F_n ; this is in agreement with the fact that, when G is constant, there is no sampling bias on X , and the X_i 's are representative for the target population. Note that, with one-sided truncation, G is monotone increasing (left-truncation) or monotone decreasing (right-truncation); therefore, testing for \mathcal{H}_0 is irrelevant in that case. Another way to write the null hypothesis is $\mathcal{H}_0 : F(x) = F^*(x), x \in S_X$; that both formulations are equivalent immediately follows from (1) or (2). Certainly, by using (1), it is easy to see that the variance of $G(X)$ is zero whenever F^* equals F and, hence, $G(X)$ is constant almost surely in that case. The reverse implication is obviously true. Therefore, it is natural to test for \mathcal{H}_0 by comparing F_n to F_n^* . This is formalized in the following subsection.

2.2 | Test statistic

A possible test statistic for \mathcal{H}_0 is given by the L_∞ norm of $F_n - F_n^*$:

$$D_n \equiv D_n(F_n, F_n^*) = \sup_{x \in S_X} |F_n(x) - F_n^*(x)|. \tag{3}$$

Large values of D_n lead to the rejection of the hypothesis of ignorable sampling bias; of course, here the vague meaning of “large” must be formalized by using the null distribution of D_n , so the given significance level is respected. When (3) leads to the acceptance of \mathcal{H}_0 one may say that there is no significant deviation between the Efron-Petrosian NPMLE F_n and the ECDF F_n^* or, alternatively, that the sampling bias is ignorable. In such a case, the simple estimator F_n^* can be used to estimate the target F , and variance improvements with respect to F_n can be obtained in this way. Moreover, under \mathcal{H}_0 the estimator F_n^* is the NPMLE of F ; this can be easily seen by decomposing the full likelihood (see (6) below) as a product of the marginal likelihood of the X_i 's and the conditional likelihood of the (U_i, V_i) 's given the X_i 's.

Since both F_n^* and F_n are concentrated on the X_i 's, a simple expression for (3) can be derived. Specifically, the Efron-Petrosian NPMLE can be written as

$$F_n(x) = \frac{\alpha_n}{n} \sum_{i=1}^n G_n(X_i)^{-1} I(X_i \leq x) \tag{4}$$

where $\alpha_n = \int G_n dF_n = n / \sum_{i=1}^n G_n(X_i)^{-1}$. Then,

$$D_n = \max_{1 \leq i \leq n} |F_n(X_i) - F_n^*(X_i)| = n^{-1} \alpha_n \max_{1 \leq i \leq n} \left| \sum_{j=1}^n \left[\frac{1}{G_n(X_j)} - \frac{1}{\alpha_n} \right] I(X_j \leq X_i) \right|. \tag{5}$$

Equation (5) indicates that D_n is essentially the maximum cumulative difference between the inverse sampling probabilities $1/G_n(X_i)$ and $1/\alpha_n$. There exists no explicit formula for G_n ; indeed, the couple (F_n, G_n) is implicitly defined as the maximizer of the full likelihood of the (X_i, U_i, V_i) 's, which is given by

$$\mathcal{L}_n = \left(\int \int_{u \leq x \leq v} dK(u, v) dF(x) \right)^{-n} \prod_{i=1}^n dF(X_i) dK(U_i, V_i). \tag{6}$$

The assumed independence between X and (U, V) is critical to justify (6). Maximization of \mathcal{L}_n with respect to both F and K leads to their NPMLEs, F_n and K_n respectively. Finally, the NPMLE of G is computed from K_n as $G_n(x) = \int_{u \leq x \leq v} dK_n(u, v)$.

The complicated structure of (6) results in a circularity, in the sense that F_n depends on G_n , as indicated by (4), while G_n itself is depending on F_n . Actually, F_n and G_n solve the couple of equations

- (i) $F_n(x) = \left(\sum_{i=1}^n G_n(X_i)^{-1} \right)^{-1} \sum_{i=1}^n G_n(X_i)^{-1} I(X_i \leq x)$,
- (ii) $G_n(x) = \left(\sum_{i=1}^n (F_n(V_i) - F_n(U_i-))^{-1} \right)^{-1} \sum_{i=1}^n (F_n(V_i) - F_n(U_i-))^{-1} I(U_i \leq x \leq V_i)$

(Shen⁶). In practice, iterative algorithms that aim the maximization of \mathcal{L}_n are used to compute the Efron-Petrosian NPMLE F_n and/or the corresponding empirical sampling bias G_n . See de Uña-Álvarez et al¹⁷ for a thorough review of the NPMLE with doubly truncated data.

The limiting null distribution of D_n is given in the following result. We will refer to the following conditions, where $[a_{U_1}, b_{U_1}]$ and $[a_{V_1}, b_{V_1}]$ denote the supports of U_1 and V_1 :

- (C1) F has only a finite number of discontinuity points and is continuous everywhere else; $a_{U_1} < a_X < a_{V_1}$, $b_{U_1} < b_X < b_{V_1}$, $b_X - a_X < \infty$ and $P(V_1 - U_1 \geq \nu) = 1$ for some $\nu > 0$
- (C2) The marginal densities of U_1 and V_1 have convex support and are bounded on S_X

Condition (C1) admits continuous and discrete distributions, provided that the number of point masses of F is finite. Besides, this condition (C1), together with the convexity condition in (C2), implies that the sampling probability for X is bounded away from zero, which is important in proofs. We mention that the inequalities $a_{U_1} < a_X < a_{V_1}$ and $b_{U_1} < b_X < b_{V_1}$ and the equality $P(V_1 - U_1 \geq \nu) = 1$ in (C1) are only slightly stronger versions of the aforementioned identifiability conditions for F and K . Finally, (C2) guarantees that the operator \mathcal{A} appearing the asymptotic representation of F_n is bounded; this is needed for obtaining the asymptotic properties of F_n . See de Uña-Álvarez and Van Keilegom⁸ for further details and discussion.

Theorem 1. *Assume conditions (C1) and (C2). Then, under \mathcal{H}_0 , it holds $\sqrt{n}D_n \rightarrow \sup_{x \in S_X} |\mathcal{B}(x)|$ in distribution, where \mathcal{B} is a zero-mean Gaussian process.*

Theorem 1 follows from de Uña-Álvarez and Van Keilegom,⁸ Theorem 2.1, where an asymptotic representation for the centered Efron-Petrosian NPMLE $F_n(x) - F(x)$ as a sum of zero-mean iid random variables is given. Under \mathcal{H}_0 that representation still holds for $F_n(x) - F_n^*(x)$, although the extra term $F^*(x) - F_n^*(x)$ naturally appears. This however does not introduce new difficulties in proofs, due to the simple structure of the ECDF; note that the class of indicator functions is Donsker, so weak convergence is obtained. This, together with the continuous mapping theorem, is enough to conclude. The straightforward details are omitted.

The distribution of the limiting variable $\sup_{x \in S_X} |\mathcal{B}(x)|$ in Theorem 1 is difficult to handle; in general it will depend on an operator which does not have a closed-form. To be specific, the process $\mathcal{B}(x)$ is the limit of $S_{1n}(x) + S_{2n}(x)$, $x \in S_X$, where $S_{1n}(x) = n^{-1/2} \sum_{i=1}^n \mathcal{A}[h_i](x)$ for a certain linear operator \mathcal{A} and certain iid random functions h_i , and $S_{2n}(x) = -n^{-1/2} \sum_{i=1}^n (I(X_i \leq x) - F(x))$. Moreover, \mathcal{A} equals the sum of an infinite series, $\sum_{r=0}^{\infty} A^r$, where A is an operator depending on population parameters in a complicated way.⁸ Thus, it is unclear how the asymptotic law arising from the term $S_{1n}(x) + S_{2n}(x)$ can be used for practical purposes. Interestingly, both $S_{1n}(x)$ and $S_{2n}(x)$ are zero-mean under the null hypothesis of ignorable sampling bias; however, in general the expectation of $S_{2n}(x)$ is given by $n^{1/2}(F(x) - F^*(x))$, which does not vanish when \mathcal{H}_0 is false. This guarantees the consistency of D_n as the sample size n grows.

Due to the aforementioned complexity of the asymptotic null distribution of D_n , a bootstrap approximation is proposed in the following subsection.

2.3 | Bootstrap approximation

Bootstrap methods for randomly truncated data have been widely investigated in the literature. In the left-truncated setting, Gross and Lai³⁰ discussed and compared two different bootstrap resampling plans: the simple bootstrap and the obvious bootstrap. For doubly truncated data, the simple and the obvious bootstrap were used in Moreira and de Uña-Álvarez¹ to approximate the sampling distribution of the Efron-Petrosian NPMLE F_n . The simple bootstrap resamples from the (X_i, U_i, V_i) 's with replacement. On the other hand, the obvious bootstrap independently resamples the target variable X and the truncating couple (U, V) from F_n and K_n respectively; then, an acceptance/rejection rule is implemented, so only the triplets satisfying $U \leq X \leq V$ are retained. As discussed in the aforementioned works, these two bootstraps are not equivalent. In particular, with the obvious bootstrap triplets (X_i, U_j, V_j) , $i \neq j$, are possible as long as $U_j \leq X_i \leq V_j$ is satisfied, while only triplets belonging to the original sample are allowed by the simple bootstrap. This makes a difference with respect to the right-censored setting, in which the simple bootstrap and the obvious bootstrap induce the same probability law.³¹

In the testing framework, bootstrap methods aim the estimation of the null distribution of the test statistic. For this, the bootstrap usually incorporates the null hypothesis at some stage of the resampling algorithm. Unfortunately, it is unclear how this can be done in the setting of this article, since the null hypothesis does not determine the distribution of the random variables X and (U, V) . A possibility is to apply the obvious bootstrap described above but with F_n replaced by

F_n^* when resampling the target variable, so the hypothesis $F = F^*$ is mimicked. Preliminary simulations indicate however that such bootstrap approximation may be conservative (results not shown). An alternative route is to modify the structure of the bootstrap test statistic in such a way that the simple bootstrap becomes applicable; see Martínez-Cambor and Corral³² for a deeper insight. We will follow this approach in order to propose a bootstrap approximation for the null distribution of D_n .

Introduce $D_n^{(1)} = \sup_{x \in S_X} |F_n(x) - F(x) + F^*(x) - F_n^*(x)|$. Note that equality $D_n = D_n^{(1)}$ holds under \mathcal{H}_0 , while $D_n \neq D_n^{(1)}$ under the alternative. Given the bootstrap triplets (X_i^b, U_i^b, V_i^b) , $1 \leq i \leq n$, sampled with replacement from the (X_i, U_i, V_i) 's, define $D_n^b = \sup_{x \in S_X} |F_n^b(x) - F_n(x) + F_n^*(x) - F_n^{*,b}(x)|$. Here, F_n^b and $F_n^{*,b}$ denote the estimators F_n and F_n^* when computed from the bootstrap resample. For left-truncated data, Gross and Lai³⁰ established under certain conditions the validity of the simple bootstrap to approximate the distribution of the NPMLE F_n , a result that can be readily extended to $F_n(x) - F_n^*(x)$. This proves that the bootstrap process $F_n^b(x) - F_n(x) + F_n^*(x) - F_n^{*,b}(x)$, $x \in S_X$, consistently approximates the distribution of $D_n^{(1)}$ in the left-truncated setting. The formal extension of such result to double truncation is not immediate, however. For instance, Edgeworth expansions as those invoked by Gross and Lai³⁰ have not been developed for the Efron-Petrosian estimator; the complex nature of F_n discussed in Section 2.2 makes this difficult. Importantly, simulation results in the following section, see also the Supporting Information, reveal that the proposed approximation works well when testing for \mathcal{H}_0 ; this suggests that the simple bootstrap is consistent under double truncation too. On the other hand, our results are in agreement with Moreira and de Uña-Álvarez,¹ who showed that the simple bootstrap performs satisfactorily in the construction of pointwise confidence intervals for a doubly truncated CDF. Note that the bootstrap approximation proposed for the test statistic D_n does not draw resamples under the null hypothesis; rather, the bootstrap version of the test statistic D_n^b is centered around $F_n(x) - F_n^*(x)$. This moves the testing problem close to the confidence interval setting.

In practice, D_n^b can be implemented as a maximum along the X_i 's, since (with the simple bootstrap) the X_i^b 's are by force a subset of the original data. We propose to compute D_n^b , as defined here, for a large number of bootstrap replicates B ; then, the null distribution of D_n is approximated from its bootstrap evaluations D_n^b , $1 \leq b \leq B$. Equivalently, the bootstrap P -value is computed as

$$p^B = B^{-1} \sum_{b=1}^B I(D_n^b \geq D_n). \quad (7)$$

The null hypothesis of ignorable sampling bias is rejected when p^B is smaller than or equal to the nominal significance level for the test. Note that, since the distribution of the D_n^b 's approximate that of $D_n^{(1)}$, and since $D_n = D_n^{(1)}$ under \mathcal{H}_0 , the P -value in (7) is conjectured to follow a uniform distribution under the null hypothesis. On the other hand, in general the distribution of D_n is shifted to the right when compared to $D_n^{(1)}$, the shift increasing as F^* departs from F . Therefore, a natural conjecture is that the proposed P -value is stochastically dominated by a uniform random variable under the alternative, anticipating the consistency of the proposed bootstrap approach.

The performance of the test statistic D_n with the given bootstrap approximation is investigated by simulations in the following section.

3 | SIMULATION STUDY

In this section we investigate the finite sample performance of the test statistic D_n through simulations. Two different models are considered: the first one corresponds to interval sampling (Model 1), while in the second one the truncation limits are simulated independently (Model 2). Specifically, for a target variable X supported on the unit interval $S_X = [0, 1]$ and certain parameter values $\zeta > 0$ and $\rho > 0$ we consider the following scenarios:

- M1 (Model 1). $U = (1 + \zeta)Z^\rho - \zeta$ and $V = U + \zeta$, where $Z \sim U(0, 1)$ is independent of X ;
- M2 (Model 2). $U = (1 + \zeta)Z_1 - \zeta$ and $V = \zeta(Z_2^{-\rho} - 1)$, where $Z_i \sim U(0, 1)$, $i = 1, 2$, are independent random variables and independent of X

In Model 1 (interval sampling) the ζ parameter stands for the width of the sampling interval. In this Model 1, the left-truncation limit U is supported on $(-\zeta, 1)$, while V is supported on $(0, 1 + \zeta)$; this implies that both F and K are identifiable. In Model 2 (independent truncation limits), U is again supported on $(-\zeta, 1)$ but V is now supported on $(0, \infty)$.

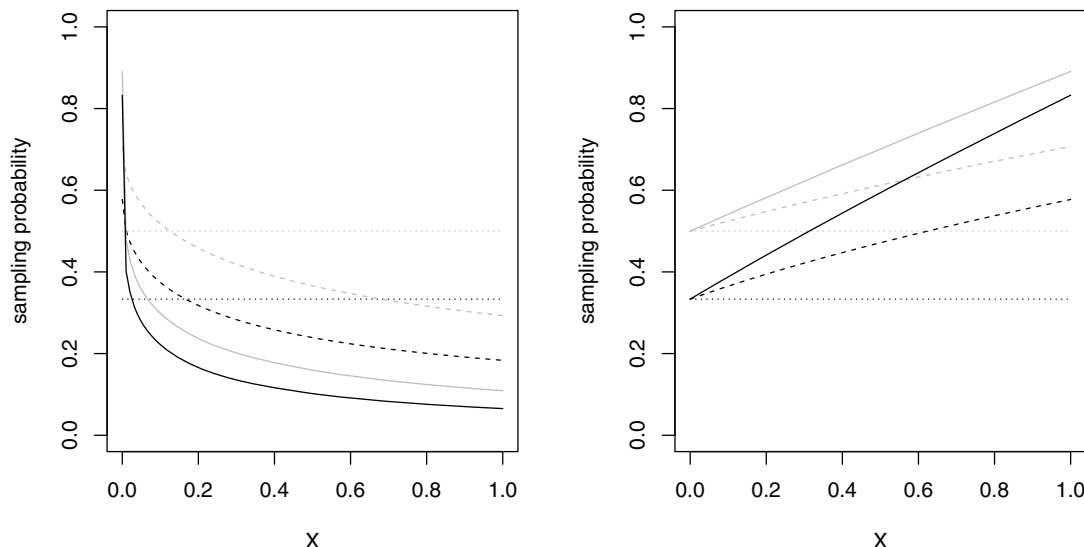


FIGURE 1 Sampling bias for simulated Model 1 (left) and Model 2 (right): $\rho = 1$ (dotted line), $\rho = 2$ (dashed line) and $\rho = 6$ (solid line). Black and grey lines correspond to $\zeta = 1/2$ and $\zeta = 1$ respectively.

In this case, F is identifiable, but K is not; this is because $V < U$ may occur (the couple (U, V) should be redefined as conditionally on $U \leq V$ if the identifiability of K is aimed). In both models M1 and M2 the sampling bias is ignorable only when $\rho = 1$. Specifically, for Model 1 and $x \in (0, 1)$ it holds

$$G(x) = (1 + \zeta)^{-1/\rho}[(x + \zeta)^{1/\rho} - x^{1/\rho}], \tag{8}$$

so when $\rho = 1$ the sampling probability $G(x)$ is free of X and the truncation proportion is $(1 + \zeta)^{-1}$. On the other hand, for Model 2 one gets

$$G(x) = \zeta^{1/\rho}(1 + \zeta)^{-1}(x + \zeta)^{1-1/\rho}; \tag{9}$$

again, (9) is constant when $\rho = 1$, giving the same truncation rate as in Model 1. Smaller values of ζ result in a larger proportion of truncated data. For instance, in Model 1 the truncation rate for $\rho = 1$, $\rho = 2$ and $\rho = 6$ is 50%, 61% and 81% ($\zeta = 1$), or 67%, 74% and 87% ($\zeta = 1/2$). For Model 2, the truncation rate is 39% ($\rho = 2, \zeta = 1$), 30% ($\rho = 6, \zeta = 1$), 53% ($\rho = 2, \zeta = 1/2$) or 41% ($\rho = 6, \zeta = 1/2$). We will investigate the influence of the ζ parameter in the performance of the test; specifically, values $\zeta = 1$ and $\zeta = 1/2$ will be considered.

Besides the null case $\rho = 1$ we will consider weak ($\rho = 2$) and strong ($\rho = 6$) deviations from the hypothesis of ignorable sampling bias under Model 1 and Model 2. The sampling probabilities (8) and (9) for the several choices of ρ and ζ are depicted in Figure 1. From this Figure 1 it is seen that, when $\rho \neq 1$, the sampling probability decreases (Model 1) or increases (Model 2) as X grows, with a more clear violation of \mathcal{H}_0 when $\rho = 6$. Finally, the target variable X will be taken as uniformly distributed on the unit interval or, alternatively, distributed as a $Beta(a, b)$ model with shape parameters $(a, b) = (1, 1/2)$ and $(a, b) = (1/2, 1)$. This will allow for the investigation of the sensitiveness of D_n to changes in F .

In Table 1 the proportion of rejections of \mathcal{H}_0 performed by D_n along 1000 Monte Carlo replicates is given; the results corresponds to a uniformly distributed target X . We considered sample sizes $n \in \{100, 200\}$ and nominal significance levels $\gamma \in \{0.1, 0.05, 0.01\}$. The number of bootstrap replicates was $B = 500$. From Table 1 it is seen that the proposed test respects the nominal level fairly well. On the other hand, the power of the test increases with the sample size and the degree of departure from the hypothesis of ignorable sampling bias, as expected. Finally, the influence of ζ is more subtle. For Model 1, the statistical power increases with ζ , and this is well connected to the meaning of this parameter, which is the width of the sampling interval under M1. For Model 2, the opposite occurs. This can be explained from the fact that, when $\zeta = 1/2$, the sampling probability G under M2 departs more from a constant than with $\zeta = 1$ (Figure 1, right). Violations of the level occurred in some of the scenarios when using a smaller number of bootstrap replicates ($B = 300$, results not shown); this sensitivity should be taken into account when using the proposed bootstrap approximation in practice.

TABLE 1 Proportion of rejections of \mathcal{H}_0 performed by D_n along 1000 Monte Carlo trials, Model 1 (interval sampling) and Model 2 (independent truncation limits). The nominal level of the test is γ . The null hypothesis corresponds to $\rho = 1$, while $\rho = 2$ and $\rho = 6$ represent weak and strong deviation from \mathcal{H}_0 , respectively. The number of trials that were discarded for $\rho = (1, 2, 6)$ because the Efron-Petrosian NPMLE did not exist or was not unique is indicated between brackets at the right of the ζ value. The number of bootstrap replicates was 500.

γ	$\rho = 1$			$\rho = 2$			$\rho = 6$		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
Model 1, $n = 100$									
$\zeta = 1$ (18, 49, 118)	0.0866	0.0346	0.0031	0.5342	0.4038	0.1672	0.9524	0.8980	0.6474
$\zeta = 1/2$ (36, 63, 153)	0.0788	0.0342	0.0062	0.2241	0.1249	0.0395	0.5502	0.3601	0.0874
Model 1, $n = 200$									
$\zeta = 1$ (6, 25, 79)	0.0755	0.0493	0.0131	0.8585	0.7713	0.5149	0.9957	0.9946	0.9739
$\zeta = 1/2$ (18, 49, 97)	0.0876	0.0407	0.0051	0.4585	0.3291	0.1367	0.8959	0.7697	0.4363
Model 2, $n = 100$									
$\zeta = 1$ (18, 18, 6)	0.0804	0.0397	0.0071	0.5193	0.4012	0.1670	0.9789	0.9447	0.7736
$\zeta = 1/2$ (33, 23, 16)	0.0889	0.0465	0.0062	0.6888	0.5333	0.2651	0.9990	0.9878	0.9258
Model 2, $n = 200$									
$\zeta = 1$ (7, 9, 4)	0.0886	0.0483	0.0030	0.8214	0.7215	0.4480	1.000	1.000	0.9970
$\zeta = 1/2$ (16, 8, 5)	0.0884	0.0467	0.0071	0.9234	0.8569	0.6522	1.000	1.000	1.000

The simulations were repeated with a different distribution for the target variable X . As mentioned, $Beta(1, 1/2)$ and $Beta(1/2, 1)$ models were considered to this end. The results, see Web Tables 1 and 2 in the Supporting Information, went in the expected direction; that is, the power of D_n decreased as the target distribution concentrated in areas along which G was relatively flat. To be specific, the power of D_n for $X \sim Beta(1, 1/2)$ (which shifts the uniform density to the right) was smaller compared to the figures in Table 1. Naturally, the situation was just the opposite with $X \sim Beta(1/2, 1)$. This is in agreement with the fact that, for both Model 1 and Model 2, the absolute value of the first derivative of $G(x)$ is monotone decreasing.

In the simulation study an issue related to the possible nonexistence or nonuniqueness of the Efron-Petrosian NPMLE occurred. Specifically, for a small number of trials the estimator F_n could not be computed in a reliable way. These trials were eliminated when computing the empirical rejection levels attached to D_n . In Table 1 the exact number of discarded samples is provided; it is seen that the problem occurs less frequently when increasing the sample size. The same issue appeared in the bootstrap resamples. In this case, the bootstrap P -value p^B in (7) was computed from the available evaluations of D_n^b ; anyway, these were the total amount of $B = 500$ evaluations for most of the simulated trials. In practice, potential issues with the computation of the NPMLE can be avoided through the preliminary fitting of a parametric truncation model; of course, this may induce some estimation bias, particularly when the parametric family is misspecified. To be explicit, in the setting of parametric truncation the weights $G_n(X_i)$ in (4) are replaced by $G(X_i; \hat{\theta})$, where $G(x; \hat{\theta}) = \int_{u \leq x \leq v} dK(u, v; \hat{\theta})$ is the sampling probability for $X = x$ under a given parametric family $K(\cdot, \cdot; \theta)$ for the truncation CDF. This leads to an alternative estimator for F , of semiparametric nature. The θ parameter can be estimated by maximizing the conditional likelihood of the (U_i, V_i) 's given the X_i 's. See Moreira et al¹⁸ for more on this. Indeed, the parametric truncation setup provides an alternative test for ignorable sampling bias because, when $G(\cdot; \theta_0)$ is flat for some θ_0 in the parametric space, a test for the null hypothesis $\theta = \theta_0$ is valid for that aim. A limitation of this alternative approach is that it is not omnibus, since it may fail to detect departures from the null hypothesis when the parametric family is misspecified.

4 | REAL DATA APPLICATIONS

4.1 | Childhood cancer data

We consider all the children diagnosed from cancer in the region of North Portugal (which includes the districts of Porto, Braga, Bragança, Vila Real and Viana do Castelo) between January 1999 and December 2003.¹ The number of cases was

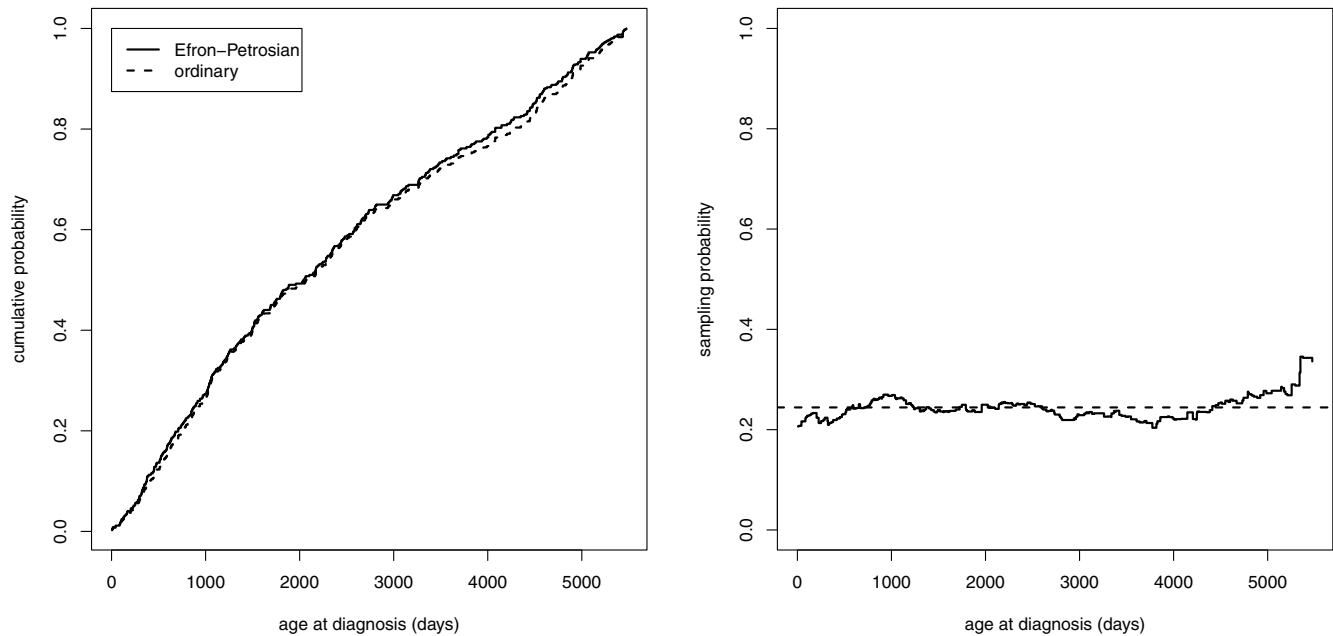


FIGURE 2 Left: Efron-Petrosian NPMLE (solid line) and ECDF (dashed line). Right: NPMLE of the sampling probability (solid line) and sampling probability under the hypothesis of ignorable sampling bias (dashed line). Childhood cancer data.

409. The target variable is the age at diagnosis X (in days) and, thus, it is doubly truncated due to the interval sampling. The truncating values (U, V) are determined by the birth dates. Specifically, V is defined as age (in days) at December 2003, while U is defined as $U = V - 1825$ (interval width of 5 years). Note that, with this notation, the individual is recruited when $U \leq X \leq V$ holds. On the other hand, no censoring on X is present, since only children diagnosed from cancer are included in the sample. Information on X was missing for three cases; hence, the sample size is $n = 406$. The data are available within the data frame `ChildCancer` of the R package `DTDA`.

The value of the test statistic D_n was 0.0206, with corresponding P -value 0.9120 based on $B = 500$ bootstrap resamples. Thus the test largely accepts the null hypothesis of ignorable sampling bias. This is in agreement with the informal analysis of Moreira and de Uña-Álvarez,¹ who obtained a truncation distribution close to uniform for this dataset.

In Figure 2 the goodness-of-fit plots for \mathcal{H}_0 corresponding to the processes $F_n(x) - F_n^*(x)$ (left) and $G_n(x) - \alpha_n$ (right) are provided; the estimated proportion of truncation is, in this case, $1 - \alpha_n = 0.7557$. From this Figure 2 it is seen that the ECDF is close to the Efron-Petrosian estimator, and that the sampling probability is roughly constant. The fact that an ignorable sampling bias holds in this case can be explained from the homogeneity of the birth process for the individuals who will develop cancer along their childhood.

Variance improvements when using $F_n^*(x)$ instead of $F_n(x)$ to analyze the data on childhood cancer are depicted in Figure 3. Specifically, the ratio $\sigma_n(x)/\sigma_n^*(x)$ for $x \in \{X_i, 1 \leq i \leq n\}$ is displayed, where $\sigma_n^*(x) = (F_n^*(x)(1 - F_n^*(x))/n)^{1/2}$ is the usual empirical standard error of $F_n^*(x)$ and $\sigma_n(x)$ is the standard error of the Efron-Petrosian estimator based on the simple bootstrap ($B = 500$ replicates). From Figure 3 it is seen that $\sigma_n(x)$ is several times larger than $\sigma_n^*(x)$ in most of the support of X ; interestingly, the standard error of $F_n(x)$ relative to $F_n^*(x)$ may be as large as 3.5 at particular quantiles x . This illustrates how testing for \mathcal{H}_0 may help to reduce the estimation variance and, hence, to facilitate the statistical inference from the doubly truncated outcomes.

4.2 | Parkinson’s disease data

Clark et al³³ studied the association between genetic information and age of onset of Parkinson’s disease. In that study, in order to eliminate potential biases coming from different survival profiles, the selected patients were those with the DNA sample taken 8 years at maximum after the onset of Parkinson. Therefore, the age at onset X is truncated from the right by the age at blood sampling V , and left-truncated by $U = V - 8$ (age in years). Two different groups of patients were considered: early onset, with ages at onset ranging between 35 and 55 years ($n = 99$); and late onset, for which the

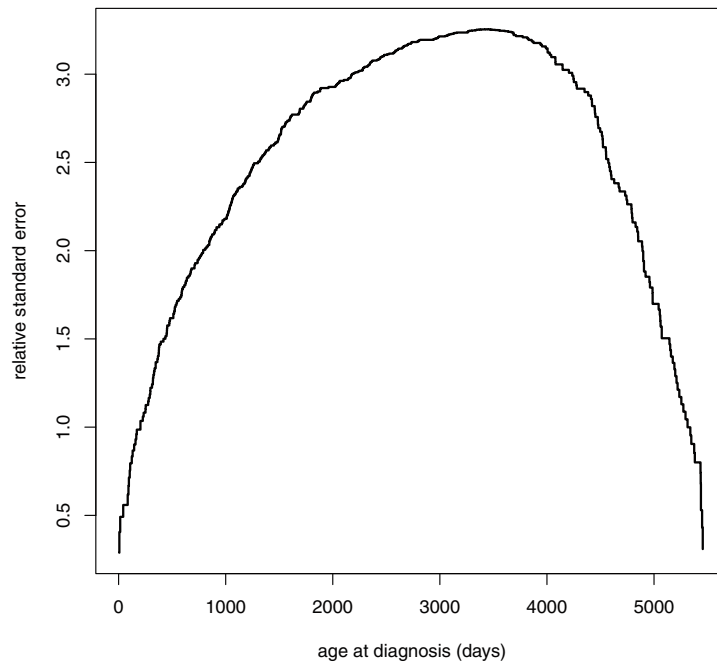


FIGURE 3 Standard error of the Efron-Petrosian NPMLE relative to the ECDF function. Childhood cancer data.

ages range between 63 and 87 years ($n = 100$). These two datasets are available from `PDearly` and `PDlate` objects in the package `DTDA`. Truncation values are missing for two patients in the early onset group, and these cases were removed for our analyses; the final sample size in this group is thus $n = 97$.

The informal graphical assessment for ignorable sampling bias for the early and late onset groups is given in Figures 4 and 5 respectively. In both cases a sampling bias is revealed; this is much clearer for the late onset group, in which the sampling probabilities are extremely low for ages below 69 years (Figure 5, right). This results in a clear departure between F_n and F_n^* at the left tail (Figure 5, left). The evidences against \mathcal{H}_0 in the early onset group are weaker, although the empirical sampling probability G_n exhibits a clearly increasing shape. The formal testing of \mathcal{H}_0 through D_n gave the following results (p -values p computed from 500 replicates; 30 resamples were removed for the late onset group due to the nonexistence/nonuniqueness of the NPMLE): $D_n = 0.2612$ and $p = 0.0520$ (early onset), and $D_n = 0.7929$, $p = 0.0022$ (late onset). Therefore, at significance level 0.05 the test accepts the null for the early onset group, but it rejects it for the late onset.

For completeness, the accuracy of F_n relative to F_n^* in both groups was calculated (Web Figures 1 and 2 in the Supporting Information), even when one would probably be in favour of using the Efron-Petrosian NPMLE for the Parkinson's disease study, according to the attained P -values. Similarly as in the childhood cancer study, the standard error of the Efron-Petrosian estimator was several times that of $F_n^*(x)$ for most of the x -values (the X_i 's), with a larger relative deficiency of F_n at the left tail of the distribution. The situation for the late onset group was critical, in the sense that the standard error of F_n was more than ten times larger at specific quantiles. Unfortunately, the formal tests performed by D_n give few chances to work with F_n^* in this case.

5 | DISCUSSION

Random truncation induces most of the times a sampling bias on the target variable. This is always the case with left- or right-truncation, often encountered in the analysis of time-to-event data, where proper corrections are needed. However, with double truncation the situation may be different, since the truncation limits may compensate each other so the sampling bias becomes negligible. If that is the case, ordinary statistical procedures are applicable, thus simplifying the estimation and inference.

In this article a formal test for the null hypothesis of ignorable sampling bias under random double truncation has been proposed. The test is based on the maximum departure between the Efron-Petrosian NPMLE and the ECDF. The

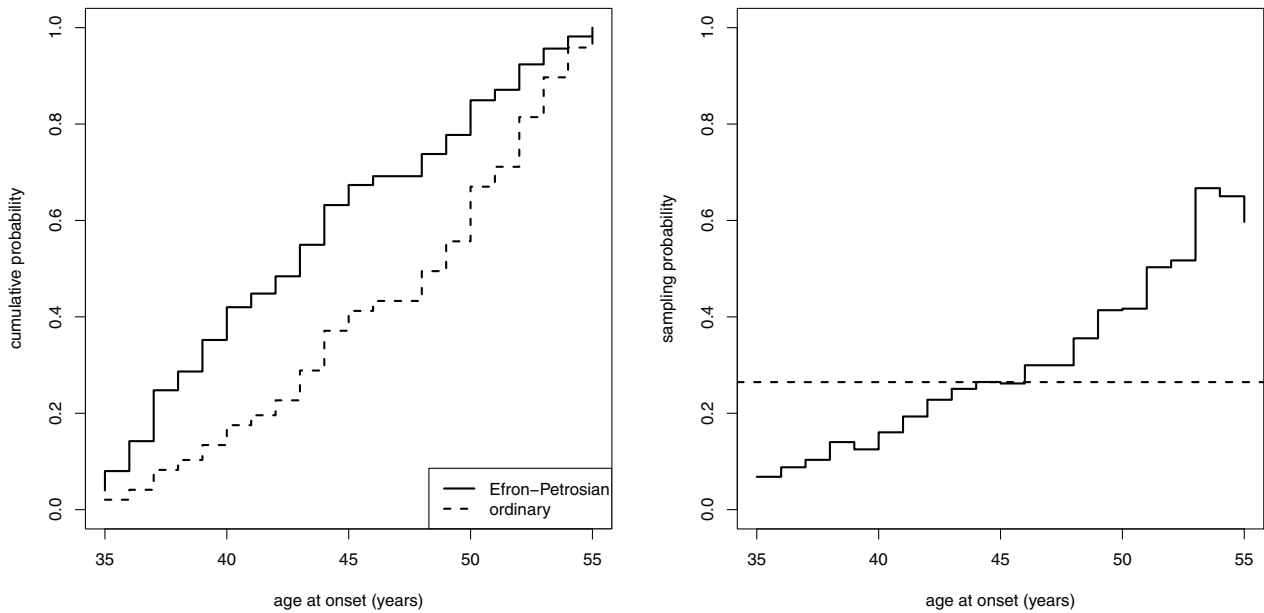


FIGURE 4 Left: Efron-Petrosian NPMLE (solid line) and ECDF (dashed line). Right: NPMLE of the sampling probability (solid line) and sampling probability under the hypothesis of ignorable sampling bias (dashed line). Parkinson's disease data, early onset.

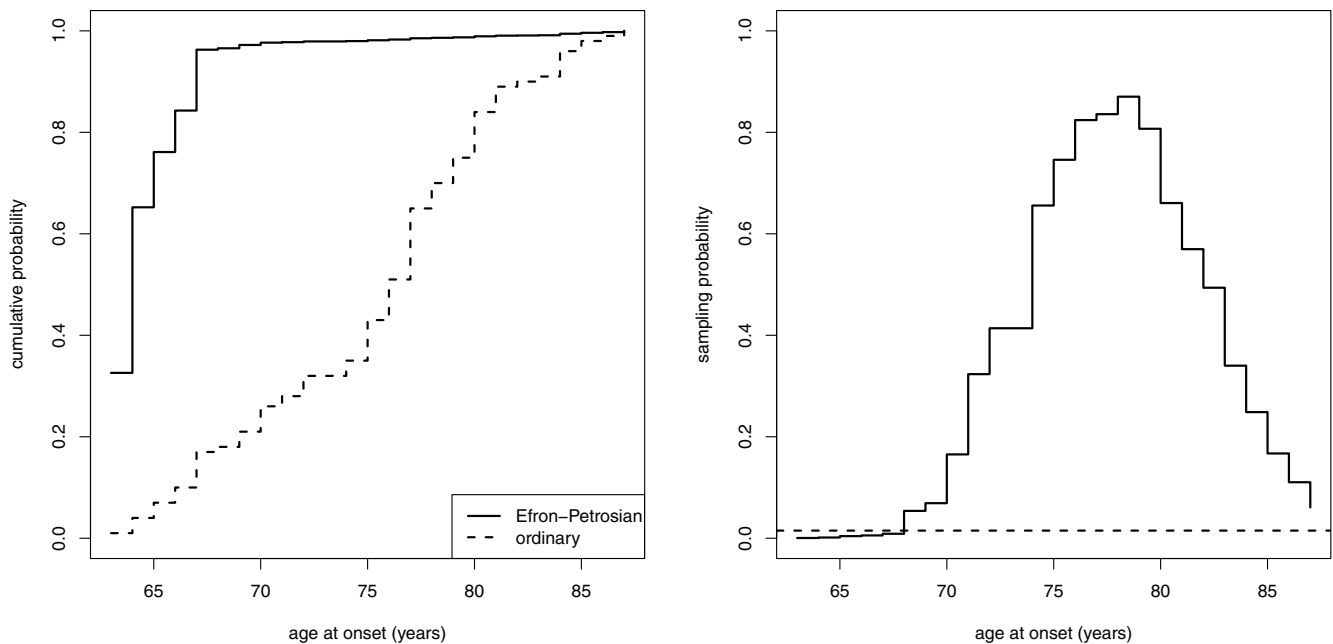


FIGURE 5 Left: Efron-Petrosian NPMLE (solid line) and ECDF (dashed line). Right: NPMLE of the sampling probability (solid line) and sampling probability under the hypothesis of ignorable sampling bias (dashed line). Parkinson's disease data, late onset.

asymptotic null distribution of the test statistic has been established, and a bootstrap procedure for the practical application of the test has been designed. Simulation studies have been conducted in order to investigate the finite sample performance of the test. The method has been found to respect the nominal level well, while exhibiting a power that increases with the sample size and the degree of violation of the null hypothesis. Applications to data on childhood cancer and Parkinson's disease have served to further illustrate the proposed method, including the variance improvements entailed by the acceptance of H_0 when estimating the target distribution F .

When the null hypothesis of ignorable sampling bias is accepted, there is no evidence against the validity of the ECDF for the estimation of the target CDF. The decision of using the ECDF based on the proposed test may entail, however,

a variance increase compared to that provided by the standard formula $F(x)(1 - F(x))/n$, even when \mathcal{H}_0 is true; this should be acknowledged in any subsequent analysis. The variance increase will be more evident as the significance level employed to test for \mathcal{H}_0 grows. For instance, under \mathcal{H}_0 , for Model 1 in the simulation study, case $\zeta = 1$, the variance increase at the three quartiles of X ranged between 9% and 17% when $\gamma = 0.05$, rising to between 19% and 32% when $\gamma = 0.1$. The issue must be taken into account in statistical practice. See for instance Shuster,³⁴ who investigated the impact of hybrid testing strategies in clinical trials.

Test statistics for \mathcal{H}_0 based on the sampling probability process $G_n(x) - \alpha_n$, $x \in S_X$, could be considered too. Preliminary simulations performed by the author (results not shown) indicate that the test based on the supremum norm of this alternative process, D_n^G say, does not dominate (nor is dominated by) D_n in the sense of the power. Importantly, since the jump points of $G_n(x)$ correspond to the truncation values, the practical implementation and interpretation of D_n^G requires some care. It is also possible to consider distances other than the supremum (Kolmogorov-Smirnov type) norm to redefine the test statistic D_n , such as Cramér-von Mises or Anderson-Darling type distances. This is currently under study and the corresponding results will be presented elsewhere.

Another possible route to explore when looking for powerful testing procedures is that given by smooth tests. With smooth tests density functions, rather than cumulative distributions, are compared. This may result in power improvements when the bandwidth factor is properly chosen; see for instance Martínez-Cambor and de Uña-Álvarez.³⁵ Recently, tests based on the comparison of empirical characteristic functions have been investigated in a variety of settings.^{36,37} Such approach could be brought here too in order to construct a test for \mathcal{H}_0 .

The null hypothesis of ignorable sampling bias can be written as $\mathcal{H}_0 : a(x) = 1$, $x \in S_X$, where $a(x) = \alpha^{-1}G(x)$ is the normalized sampling probability. A generalization of this testing problem is the one in which the null states $a(x) = a_0(x)$, $x \in S_X$, for a fully specified function $a_0(x)$. This is relevant when there exists information on the sampling bias other than ignorability; for instance, with interval sampling such information could be given by general population registries reporting birth rates for the process of interest. Under this generalized null hypothesis, the NPMLE of F is just the inverse-probability-weighted estimator, F_n^0 say, which attaches weight $a_0(X_i)^{-1}$ to X_i , $1 \leq i \leq n$. Obviously, D_n can be generalized for this problem, becoming the maximum deviation between the $F_n(X_i)$'s and the $F_n^0(X_i)$'s. Formal theory can be derived similarly as in Theorem 1, although regularity conditions on the function $a_0(x)$ must be imposed. When the fully specified function $a_0(x)$ in \mathcal{H}_0 is replaced by a parametric family things are more complicated; the fact that the function $a(x)$ does not characterize the truncation distribution is responsible for this. Minimum-distance and pseudo-likelihood approaches are possible; the practical performance of such estimators and the development of the corresponding asymptotic theory are interesting topics for our future research.

ACKNOWLEDGEMENTS

Thanks to two anonymous reviewers and to an Associate Editor for comments and suggestions that helped to improved the presentation of the article. Work supported by the Grant PID2020-118101GB-I00, Ministerio de Ciencia e Innovación. Funding for open access charge: Universidade de Vigo/CISUG.

DATA AVAILABILITY STATEMENT

Code to reproduce the simulation results in Section 3 and the real data analyses in Section 4 is available from the Web Supporting Information.

ORCID

Jacobo de Uña-Álvarez  <https://orcid.org/0000-0002-4686-8417>

REFERENCES

1. Moreira C, de Uña-Álvarez J. Bootstrapping the NPMLE for doubly truncated data. *J. Nonparametric Stat.* 2010;22:657-583.
2. Zhu H, Wang MC. Analysing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika.* 2012;99:345-361.
3. Rennert L, Xie SX. Bias induced by ignoring double truncation inherent in autopsy-confirmed survival studies of neurodegenerative diseases. *Stat. Med.* 2019;38:3599-3613.
4. Efron B, Petrosian V. Nonparametric methods for doubly truncated data. *J. Am. Stat. Assoc.* 1999;94:824-834.
5. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc. Ser. B.* 1976;38:290-295.
6. Shen PS. Nonparametric analysis of doubly truncated data. *Ann. Inst. Stat. Math.* 2010;62:835-853.
7. Emura T, Konno Y, Michimae H. Statistical inference based on the nonparametric maximum likelihood estimator under double-truncation. *Lifetime Data Anal.* 2015;21:397-418.

8. de Uña-Álvarez J, Van Keilegom I. Efron-Petrosian integrals for doubly truncated data with covariates: an asymptotic analysis. *Bernoulli*. 2021;27:249-273.
9. Moreira C, de Uña-Álvarez J, Crujeiras R. DTDA: Doubly Truncated Data Analysis. 2022 R package version 3.0.1, <https://CRAN.R-project.org/package=DTDA>
10. Rennett L. SurvTrunc: Analysis of Doubly Truncated Data. 2018 R package version 0.1.0, <https://CRAN.R-project.org/package=SurvTrunc>
11. Emura T, Hu YH, Huang CY. double.truncation: Analysis of Doubly Truncated Data. 2020 R package version 1.7, <https://CRAN.R-project.org/package=double.truncation>
12. Martin EC, Betensky RA. Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *J. Am. Stat. Assoc.* 2005;100:484-492.
13. Moreira C, de Uña-Álvarez J, Braekers R. Nonparametric estimation of a distribution function from doubly truncated data under dependence. *Comput. Stat.* 2021;36:1693-1720.
14. Mandel M, de Uña-Álvarez J, Simon DK, Betensky RA. Inverse probability weighted Cox regression for doubly truncated data. *Biometrics*. 2018;74:481-487.
15. de Uña-Álvarez J. R packages for the statistical analysis of doubly truncated data: a review. *arXiv 2004.08978*, 1–23. 2020a.
16. Xiao J, Hudgens MG. On nonparametric maximum likelihood estimation with double truncation. *Biometrika*. 2019;106:989-996.
17. de Uña-Álvarez J, Moreira C, Crujeiras RM. *The Statistical Analysis of Doubly Truncated Data: With Applications in R*. Hoboken, NJ: John Wiley; 2021.
18. Moreira C, de Uña-Álvarez J, Van Keilegom I. Goodness-of-fit tests for a semiparametric model under random double truncation. *Comput. Stat.* 2014;29:1365-1379.
19. Moreira C, Van Keilegom I. Bandwidth selection for kernel density estimation with doubly truncated data. *Computat. Stat. Data Anal.* 2013;61:107-123.
20. Moreira C, de Uña-Álvarez J, Meira-Machado L. Nonparametric regression with doubly truncated data. *Computat. Stat. Data Anal.* 2016;93:294-307.
21. Rennett L, Xie SX. Cox regression model with doubly truncated data. *Biometrics*. 2018;74:725-733.
22. Ying Z, Yu W, Zhao Z, Zheng M. Regression analysis of doubly truncated data. *J. Am. Stat. Assoc.* 2020;115:810-821.
23. de Uña-Álvarez J. Nonparametric estimation of the cumulative incidences of competing risks under double truncation. *Biom. J.* 2020b;62:852-867.
24. Shen PS. A class of rank-based tests for doubly-truncated data. *TEST*. 2013;22:83-102.
25. Zhu H, Wang MC. Nonparametric inference on bivariate survival data with interval sampling: association estimation and testing. *Biometrika*. 2014;101:519-533.
26. Zhu H, Wang MC. A semi-stationary copula model approach for bivariate survival data with interval sampling. *Int. J. Biostat.* 2015;11:151-173.
27. Emura T, Hu YH, Konno Y. Asymptotic inference for maximum likelihood estimators under the special exponential family with double-truncation. *Stat. Pap.* 2017;58:877-909.
28. Dörre A, Emura T. *Analysis of Doubly Truncated Data. An Introduction*. Singapore: Springer; 2019.
29. Woodroffe M. Estimating a distribution function with truncated data. *Ann. Stat.* 1985;13:163-177.
30. Gross ST, Lai TL. Bootstrap methods for truncated and censored data. *Stat. Sin.* 1996;6:509-530.
31. Efron B. Censored data and the bootstrap. *J. Am. Stat. Assoc.* 1981;76:312-319.
32. Martínez-Cambor P, Corral N. A general bootstrap algorithm for hypothesis testing. *J. Stat. Plann. Inference*. 2012;142:589-600.
33. Clark J, Reddy S, Zheng K, Betensky RA, Simon DK. Association of PGC-1 α polymorphisms with age of onset and risk of Parkinson's disease. *BMC Med. Genet.* 2011;12:69.
34. Shuster JJ. Diagnostics for assumptions in moderate to large simple clinical trials: do they really help? *Stat. Med.* 2005;24:2431-2438.
35. Martínez-Cambor P, de Uña-Álvarez J. Non-parametric k-sample tests: Density functions vs distribution functions. *Computat. Stat. Data Anal.* 2009;53:3344-3357.
36. Cousido-Rocha M, de Uña-Álvarez J, Hart J. A two-sample test for the equality of univariate marginal distributions for high-dimensional data. *J. Multivar. Anal.* 2019;174:104537.
37. Henze N, Jiménez-Gamero MD. A test for Gaussianity in Hilbert spaces via the empirical characteristic functional. *Scand. J. Stat.* 2021;48:406-428.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: de Uña-Álvarez J. Testing for an ignorable sampling bias under random double truncation. *Statistics in Medicine*. 2023;1-13. doi: 10.1002/sim.9828