Wiley Series in Probability and Statistics

THE STATISTICAL ANALYSIS OF DOUBLY TRUNCATED DATA

WITH APPLICATIONS IN R

JACOBO DE UÑA-ÁLVAREZ
CARLA MOREIRA
ROSA M. CRUJEIRAS

WILEY

The Statistical Analysis of Doubly Truncated Data: With Applications in R

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

The Statistical Analysis of Doubly Truncated Data: With Applications in R

Jacobo de Uña-Álvarez, Carla Moreira and Rosa M. Crujeiras

WILEY

This edition first published 2022

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data applied for

[ISBN: 9781119951377]

Cover Design: Wiley

Set in 9.5/12.5pt STIXTwoText by Straive, Chennai, India

10 9 8 7 6 5 4 3 2 1

To María Soledad, Marcos, Paula and Miguel, for their love, support and inspiration.
To Teo, Sabela and Andrés, for their infinite patience.

Contents

Preface

This book is the result of a long-standing collaboration among the three authors, which began when Carla Moreira was a PhD student under the supervision of Jacobo de Uña-Álvarez. Carla successfully defended her thesis, entitled 'The Statistical Analysis of Doubly Truncated Data: New Methods, Software Development, and Biomedical Applications', at the Universidade de Vigo in July 2010. At that time, only a small number of people seemed to be aware of the importance of random double truncation. Research papers on this topic were scarce before 2010, with the contribution by Bradley Efron and Vahe Petrosian in 1999 as the most relevant one. And, of course, no software was available. So, for us, it was a risky and exciting research exercise to embrace such an initiative.

We launched version 1.1 of our `R` package `DTDA` in September 2009. To our knowledge this was the first software library implementing the Efron–Petrosian estimator. The package included Efron and Petrosian's data on quasar luminosities, and we are very thankful to both scientists for sharing them. `DTDA` has been downloaded more than 45 thousand times up to now. We have taken the opportunity of writing this book to update and enhance `DTDA`, feeding it with new illustrative real datasets and enabling new functions and capabilities. We are confident in that the update of the package and the guidance provided by this book will exponentially increase the applications involving doubly truncated data, and also raise awareness about the implications of double truncation on inferential procedures.

Over these years, several researchers have collaborated with us in the fascinating adventure of investigating double truncation. Among them, we would like to mention Ingrid Van Keilegom, Micha Mandel, Rebecca Betensky, Luis Meira-Machado and Roel Braekers. We have enjoyed co-authoring a number of research papers with them. We also learned a lot about double truncation by studying real data problems posed by applied researchers; here we thank María José Bento, David Keith Simon, Zhi-Sheng Ye, Ana Cristina Santos and Henrique Barros for fruitful discussions and cooperation.

Nowadays, there is a considerable statistical community doing research on exploratory and inferential methods for doubly truncated data, partly motivated by new emerging applications in Biomedicine, Economics and Engineering, among other fields. At the time of writing the activity in this area of research is much more intense than ever before, as is evident from the number of papers on the topic published in the last couple of years. And the interest in double truncation is growing faster and faster!

This book aims to serve as a companion for those ones interested in learning about doubly truncated data analysis and inference, presenting a wide range of tools for estimating distribution and regression models. All the methods presented in this book are accompanied by real data and simulated examples and, at the end of each chapter, the reader will find the *do-it-yourself* code, mostly based on the `DTDA` package. This book is not written with the aim of being just read: its main purpose is to invite the reader to think, explore and experience.

This volume is also self-contained, providing a general overview on the main results. Further technical details and some omitted proofs can be consulted in the original references. It is also in our intention to leave several take-home messages. First, that the correction of the potential sampling bias arising from double truncation may be critical in estimation and inference. Second, that, even when the Efron–Petrosian estimator is conceptually complicated and its asymptotic theory may be overwhelming, its practical application is relatively simple from the available software packages and the good performance of resampling algorithms. Third, that external information on the sampling bias should be used whenever available, since the Efron–Petrosian estimator may be very noisy or even non-existing, particularly when the sample size is small to moderate.

We frankly hope that the reader will enjoy (and experience!) the book, at least as much as we have enjoyed writing it! Comments and suggestions from the readers on this edition are welcome; please send them to `jacobo@uvigo.es` to help us to improve the book.

Parts of this book were written while the authors were supported by the Grants MTM2017-89422-P (MINECO/AEI/FEDER, UE) (first author), UIDB/00013/2020 and UIDP/00013/2020 (second author), and MTM2016-76969-P (MINECO/AEI/FEDER, UE) (third author). This is acknowledged.

May 2021 *Jacobo de Uña-Álvarez, Carla Moreira and Rosa M. Crujeiras*
Vigo, V. N. Famalicão and Santiago de Compostela

List of Abbreviations

- AFT: accelerated failure time
- AIDS: acquired immunodeficiency syndrome
- AMISE: asymptotic mean integrated square error
- AMSE: asymptotic mean square error
- bcv: biased cross-validation
- BMISE: bootstrap mean integrated square error
- Boot: bootstrap
- cdf: cumulative distribution function
- CIF: cumulative incidence function
- cv: cross-validation
- ecdf: empirical cumulative distribution function
- DNA: deoxyribonucleic acid
- DPI: direct plug-in
- DPI_1: direct plug-in in one stage
- DPI_2: direct plug-in in two stages
- FGM: Farlie–Gumbel–Morgenstern
- HIV: human immunodeficiency virus
- ICCC: International Classification of Childhood Cancer
- iid: independent and identically distributed
- IPWE: inverse probability weighted estimator
- IQR: interquartile range
- ISE: integrated square error
- LSCV: least-squares cross-validation
- MISE: mean integrated square error
- MLE: maximum-likelihood estimator
- MSE: mean square error
- NP: nonparametric
- NPMLE: nonparametric maximum-likelihood estimator
- NR: normal reference

- pdf: probability density function
- OB: obvious bootstrap
- PB: percentile bootstrap
- PD: Parkinson's disease
- SB: simple bootstrap
- SBoot: smoothed bootstrap
- SD: standard deviation
- SEF: special exponential family
- SJ-dpi: Sheather–Jones direct plug-in
- SJ-ste: Sheather–Jones solve-the-equation plug-in
- SNPs: single nucleotide polymorphisms
- SP: semiparametric
- SPMLE: semiparametric maximum-likelihood estimator
- ucv: unbiased cross-validation

Notation

- X: target variable, supported on $[a_X, b_X]$; (U, V): truncating variables; the respective supports of U and V are $[a_U, b_U]$ and $[a_V, b_V]$
- (X, U, V): population triplet; (X_i, U_i, V_i), $1 \leq i \leq n$: observed independent triplets such that $(X_i, U_i, V_i) \stackrel{\mathrm{d}}{=} (X, U, V)|U \leq X \leq V$, where $\stackrel{\mathrm{d}}{=}$ stands for the equality in distribution
- ς: for interval sampling, width of the sampling interval, so $V = U + \varsigma$; it holds $\varsigma = d_1 - d_0$, where d_0 and d_1 are the dates that determine the sampling interval
- F, f and λ: cumulative distribution function, probability density function and hazard function of X, respectively
- $X \sim F$: variable X follows the cumulative distribution function F
- $Y \sim U(a, b)$: variable Y follows a uniform distribution on the interval (a, b)
- $\mathrm{supp}(F)$: support of the cumulative distribution function F
- $F(x-) = P(X < x)$: left-continuous version of the cumulative ditribution function F; for a continuous F it holds that $F(x) = F(x-)$ for each x
- K, k: bivariate cumulative distribution function and bivariate probability density function of (U, V), respectively
- K_U, K_V, k_U, k_V: marginal cumulative distribution functions and marginal probability density functions of U and V
- ϕ_i, ψ_i: probability masses attached to X_i and (U_i, V_i), respectively, appearing in the likelihood function
- $I(A)$: indicator of the event A
- $J_{ij} = I(U_i \leq X_j \leq V_i)$
- $(X^{\mathrm{b}}, U^{\mathrm{b}}, V^{\mathrm{b}})$: bootstrap resample of (X, U, V)
- G: weighting function, or biasing function, which reports the sampling probabilities for X; $G(x) = P(U \leq X \leq V|X = x) = \int_{u \leq x \leq v} \mathrm{d}K(u, v)$; last equality holds if X and (U, V) are independent
- $\alpha = P(U \leq X \leq V)$: proportion of non-truncated data

- $a = \alpha^{-1}G$: normalized biasing function
- $F^*(x) = P(X \leq x | U \leq X \leq V)$: truncated cumulative distribution function of X
- $K^*(u, v) = P(U \leq u, V \leq v | U \leq X \leq V)$: truncated bivariate cumulative distribution function of (U, V)
- $K_U^*, K_V^*, k_U^*, k_V^*$: truncated marginal cumulative distribution functions and truncated marginal probability density functions of U and V
- $\{K(\cdot, \cdot; \theta), \theta \in \Theta\}$, $\Theta \subset \mathcal{R}^d$: parametric family of distribution functions for the truncation couple, with parameter space Θ; $\theta = (\theta_1, \ldots, \theta_d)^t \in \Theta$
- θ_0: true value of θ
- $k(\cdot, \cdot; \theta)$: probability density function attached to $K(\cdot, \cdot; \theta)$
- $G(x; \theta) = \int_{u \leq x \leq v} \mathrm{d}K(u, v; \theta)$: weighting function, or biasing function, under the parametric truncation family
- $\alpha(\theta)$: probability of no truncation inherited from the parametric truncation family
- $F(\cdot; \theta)$: semiparametric version of F, inherited from the parametric truncation family
- $\{F(\cdot; \gamma), \gamma \in \Gamma\}$, $\Gamma \subset \mathcal{R}^d$: parametric distribution family for F
- γ_0: true value of γ
- $f(\cdot; \gamma)$: probability density function attached to $F(\cdot; \gamma)$
- $\mathcal{I}(\theta)$, $\mathcal{I}^{(1)}(\gamma)$: Fisher information matrices
- L: kernel function; h: bandwidth or smoothing parameter; $L_h(\cdot) = L(\cdot/h)/h$: rescaled kernel function
- $\mu_2(L) = \int t^2 L(t) \mathrm{d}t$
- $R(L) = \int L(t)^2 \mathrm{d}t$
- W_g: rescaled kernel function with pilot bandwidth g
- Z: vector of covariates
- $F(\cdot | z)$: conditional cumulative distribution function of X given $Z = z$
- F_{XZ}: joint cumulative distribution function of (X, Z)
- β: vector of regression coefficients; β_0: true value of β
- $\lambda(\cdot | z)$: conditional hazard function of X given $Z = z$
- $\alpha(z)$: conditional probability of no truncation given $Z = z$
- $m(z) = \mathrm{E}(X | Z = z)$: regression function
- $F_{(r)}$: cumulative incidence function of X, event type $\eta = r$
- $G_{(r)}$: conditional biasing function for X given $\eta = r$
- $\lambda_{(r)}$: cause-specific hazard function, or transition intensity, of X, event type $\eta = r$
- $\beta_{(r)}$: vector of regression coefficients for a regression model on $\lambda_{(r)}$
- C_θ: parametric copula family
- For two sequences a_n and b_n of real numbers, $a_n = o(b_n)$ means that a_n/b_n converges to zero
- For two sequences A_n and B_n of random variables, $A_n = o_P(B_n)$ means that A_n/B_n converges to zero in probability, and $A_n = O_P(B_n)$ means that A_n/B_n is bounded in probability

1

Introduction

1.1 Random Truncation

Random truncation generally refers to a situation in which a number of individuals of the target population cannot be sampled because a certain random event precludes them. When this random event is unrelated to the variables of interest standard statistical methods apply, with the only inconvenience of using a smaller sample size. In many practical cases, however, the truncation event is related to the variables under study, and specific methods to overcome the sampling bias must be considered.

This book is focused on random truncation phenomena that arise (usually, but not only) when sampling time-to-event data. That is, the variable of interest is the time X elapsed from a well-defined origin to another well-defined end point. In this setting, a truncated sample of X is a set of independent and identically distributed (iid) random variables $X_1, \ldots, X_n$ with the conditional distribution of X given $X \in B$, where B is a random set. Since the truncation event $\{X \in B\}$ is obviously related to X, standard statistical methods applied to the truncated sample may be systematically biased. For example, the ordinary empirical cumulative distribution function (ecdf) of $X_1, \ldots, X_n$ at point x, $F_n^*(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$, converges to $F^*(x) = P(X \leq x | X \in B)$ rather than to the target cumulative distribution function (cdf) $F(x) = P(X \leq x)$. This problem has received remarkable attention since the seminal paper by Turnbull (1976). Special forms of truncation when sampling time-to-event data are reviewed in Sections 1.2 and 1.3.

Time-to-event data are relevant in fields like Survival Analysis and Reliability Engineering, in which random truncation often occurs. Random truncation is found in Astronomy too, where X represents the luminosity of an stellar object that is subject to observation limits. Examples from these areas will be introduced and analysed throughout this book.

The Statistical Analysis of Doubly Truncated Data: With Applications in R, First Edition.
Jacobo de Uña-Álvarez, Carla Moreira and Rosa M. Crujeiras.