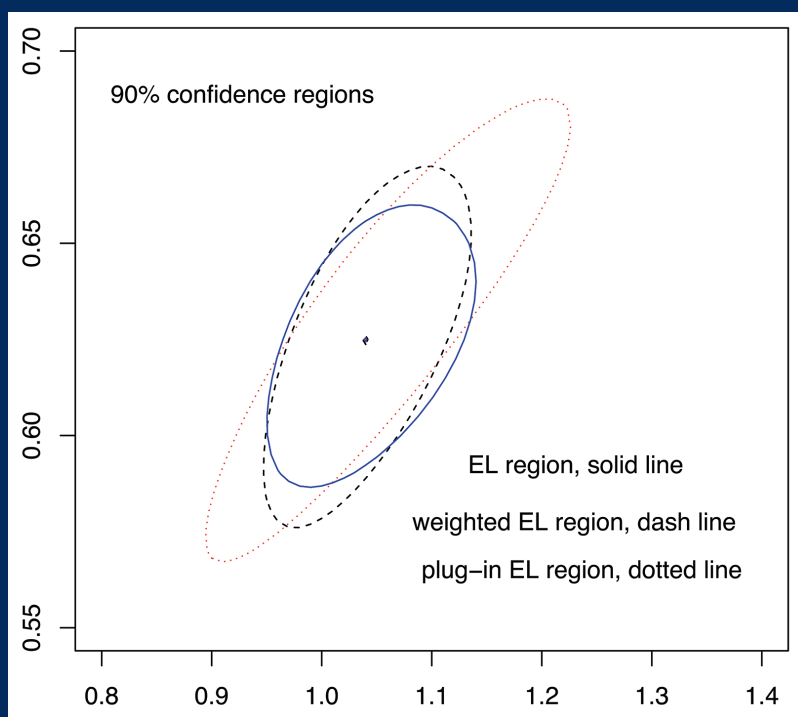


Chapman & Hall/CRC Biostatistics Series

Empirical Likelihood Method in Survival Analysis



Mai Zhou



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Empirical Likelihood Method in Survival Analysis

Chapman & Hall/CRC Biostatistics Series

Editor-in-Chief

Shein-Chung Chow, Ph.D., Professor, Department of Biostatistics and Bioinformatics,
Duke University School of Medicine, Durham, North Carolina

Series Editors

Byron Jones, Biometrical Fellow, Statistical Methodology, Integrated Information Sciences,
Novartis Pharma AG, Basel, Switzerland

Jen-pei Liu, Professor, Division of Biometry, Department of Agronomy,
National Taiwan University, Taipei, Taiwan

Karl E. Peace, Georgia Cancer Coalition, Distinguished Cancer Scholar, Senior Research Scientist
and Professor of Biostatistics, Jiann-Ping Hsu College of Public Health,
Georgia Southern University, Statesboro, Georgia

Bruce W. Turnbull, Professor, School of Operations Research and Industrial Engineering,
Cornell University, Ithaca, New York

Published Titles

**Adaptive Design Methods in
Clinical Trials, Second Edition**
Shein-Chung Chow and Mark Chang

**Adaptive Designs for Sequential
Treatment Allocation**
Alessandro Baldi Antognini and
Alessandra Giovagnoli

**Adaptive Design Theory and
Implementation Using SAS and R,
Second Edition**
Mark Chang

**Advanced Bayesian Methods for Medical
Test Accuracy**
Lyle D. Broemeling

Advances in Clinical Trial Biostatistics
Nancy L. Geller

Applied Meta-Analysis with R
Ding-Geng (Din) Chen and Karl E. Peace

**Basic Statistics and Pharmaceutical
Statistical Applications, Second Edition**
James E. De Muth

**Bayesian Adaptive Methods for
Clinical Trials**
Scott M. Berry, Bradley P. Carlin,
J. Jack Lee, and Peter Muller

**Bayesian Analysis Made Simple: An Excel
GUI for WinBUGS**
Phil Woodward

**Bayesian Methods for Measures of
Agreement**
Lyle D. Broemeling

Bayesian Methods in Epidemiology
Lyle D. Broemeling

Bayesian Methods in Health Economics
Gianluca Baio

**Bayesian Missing Data Problems: EM,
Data Augmentation and Noniterative
Computation**
Ming T. Tan, Guo-Liang Tian,
and Kai Wang Ng

Bayesian Modeling in Bioinformatics
Dipak K. Dey, Samiran Ghosh,
and Bani K. Mallick

**Benefit-Risk Assessment in
Pharmaceutical Research and
Development**
Andreas Sashegyi, James Felli, and
Rebecca Noel

**Biosimilars: Design and Analysis of
Follow-on Biologics**
Shein-Chung Chow

Biostatistics: A Computing Approach
Stewart J. Anderson

**Causal Analysis in Biomedicine and
Epidemiology: Based on Minimal
Sufficient Causation**
Mikel Aickin

**Clinical and Statistical Considerations
in Personalized Medicine**
Claudio Carini, Sandeep Menon,
and Mark Chang

Clinical Trial Data Analysis using R

Ding-Geng (Din) Chen and Karl E. Peace

Clinical Trial Methodology

Karl E. Peace and Ding-Geng (Din) Chen

Computational Methods in Biomedical Research

Ravindra Khattree and Dayanand N. Naik

Computational Pharmacokinetics

Anders Källén

Confidence Intervals for Proportions and Related Measures of Effect Size

Robert G. Newcombe

Controversial Statistical Issues in Clinical Trials

Shein-Chung Chow

Data and Safety Monitoring Committees in Clinical Trials

Jay Herson

Design and Analysis of Animal Studies in Pharmaceutical Development

Shein-Chung Chow and Jen-pei Liu

Design and Analysis of Bioavailability and Bioequivalence Studies, Third Edition

Shein-Chung Chow and Jen-pei Liu

Design and Analysis of Bridging Studies

Jen-pei Liu, Shein-Chung Chow,
and Chin-Fu Hsiao

Design and Analysis of Clinical Trials for Predictive Medicine

Shigeyuki Matsui, Marc Buyse,
and Richard Simon

Design and Analysis of Clinical Trials with Time-to-Event Endpoints

Karl E. Peace

Design and Analysis of Non-Inferiority Trials

Mark D. Rothmann, Brian L. Wiens,
and Ivan S. F. Chan

Difference Equations with Public Health Applications

Lemuel A. Moyé and Asha Seth Kapadia

DNA Methylation Microarrays: Experimental Design and Statistical Analysis

Sun-Chong Wang and Arturas Petronis

DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments

David B. Allison, Grier P. Page,
T. Mark Beasley, and Jode W. Edwards

Dose Finding by the Continual Reassessment Method

Ying Kuen Cheung

Elementary Bayesian Biostatistics

Lemuel A. Moyé

Empirical Likelihood Method in Survival Analysis

Mai Zhou

Frailty Models in Survival Analysis

Andreas Wienke

Generalized Linear Models: A Bayesian Perspective

Dipak K. Dey, Sujit K. Ghosh,
and Bani K. Mallick

Handbook of Regression and Modeling: Applications for the Clinical and Pharmaceutical Industries

Daryl S. Paulson

Inference Principles for Biostatisticians

Ian C. Marschner

Interval-Censored Time-to-Event Data: Methods and Applications

Ding-Geng (Din) Chen, Jianguo Sun,
and Karl E. Peace

Introductory Adaptive Trial Designs: A Practical Guide with R

Mark Chang

Joint Models for Longitudinal and Time-to-Event Data: With Applications in R

Dimitris Rizopoulos

Measures of Interobserver Agreement and Reliability, Second Edition

Mohamed M. Shoukri

Medical Biostatistics, Third Edition

A. Indrayan

Meta-Analysis in Medicine and Health Policy

Dalene Stangl and Donald A. Berry

Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools

Marc Lavielle

Modeling to Inform Infectious Disease Control

Niels G. Becker

Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms, and Case Studies

Mark Chang

Multiple Testing Problems in Pharmaceutical Statistics

Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz

Noninferiority Testing in Clinical Trials: Issues and Challenges

Tie-Hua Ng

Optimal Design for Nonlinear Response Models

Valerii V. Fedorov and Sergei L. Leonov

Patient-Reported Outcomes: Measurement, Implementation and Interpretation

Joseph C. Cappelleri, Kelly H. Zou, Andrew G. Bushmakina, Jose Ma. J. Alvir, Demissie Alemayehu, and Tara Symonds

Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting

Qi Jiang and H. Amy Xia

Randomized Clinical Trials of Nonpharmacological Treatments

Isabelle Boutron, Philippe Ravaud, and David Moher

Randomized Phase II Cancer Clinical Trials

Sin-Ho Jung

Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research

Chul Ahn, Moonseong Heo, and Song Zhang

Sample Size Calculations in Clinical Research, Second Edition

Shein-Chung Chow, Jun Shao and Hansheng Wang

Statistical Analysis of Human Growth and Development

Yin Bun Cheung

Statistical Design and Analysis of Stability Studies

Shein-Chung Chow

Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis

Kelly H. Zou, Aiyi Liu, Andriy Bandos, Lucila Ohno-Machado, and Howard Rockette

Statistical Methods for Clinical Trials

Mark X. Norleans

Statistical Methods in Drug Combination Studies

Wei Zhao and Harry Yang

Statistics in Drug Research: Methodologies and Recent Developments

Shein-Chung Chow and Jun Shao

Statistics in the Pharmaceutical Industry, Third Edition

Ralph Buncher and Jia-Yeong Tsay

Survival Analysis in Medicine and Genetics

Jialiang Li and Shuangge Ma

Theory of Drug Development

Eric B. Holmgren

Translational Medicine: Strategies and Statistical Methods

Dennis Cosmatos and Shein-Chung Chow

Chapman & Hall/CRC Biostatistics Series

Empirical Likelihood Method in Survival Analysis

Mai Zhou

University of Kentucky

Lexington, USA



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2016 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20150506

International Standard Book Number-13: 978-1-4665-5493-1 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Contents

List of Figures	xi
List of Tables	xiii
Preface	xv
1 Introduction	1
1.1 Survival Analysis	1
1.1.1 Hazard Function	2
1.1.2 Censored Observations	3
1.1.3 Parametric and Nonparametric Models	4
1.1.4 Parametric Maximum Likelihood Estimator	5
1.1.5 The Nelson–Aalen and Kaplan–Meier Estimators	6
1.2 Empirical Likelihood	11
1.2.1 Likelihood Function	11
1.2.2 Empirical Likelihood, Uncensored Sample	12
1.3 Empirical Likelihood for Right Censored Data	15
1.3.1 Likelihood Function in Terms of Hazard	17
1.4 Confidence Intervals Based on the EL Test	19
1.5 Datasets	20
1.6 Historical Notes	20
1.7 Exercises	21
2 Empirical Likelihood for Linear Functionals of Hazard	23
2.1 Empirical Likelihood, Poisson Version	23
2.2 Feasibility of the Constraints (2.5)	25
2.3 Maximizing the Hazard Empirical Likelihood	27
2.4 Some Technical Details	31
2.4.1 The Constrained Hazard Under the Null Hypothesis	35
2.5 Predictable Weight Functions	36
2.5.1 One-Sample Log Rank Test	37
2.6 Two-Sample Tests	39
2.7 Hazard Estimating Equations	42
2.7.1 The Semi-Parametric AFT model	43
2.7.2 Overdetermined Equations	46
2.8 Empirical Likelihood, Binomial Version	50

2.9	Poisson or Binomial?	55
2.10	Some Notes on Counting Process Martingales	56
2.10.1	Compensated Counting Process as Martingale	58
2.10.2	Censoring as Splitting/Thinning of a Counting Process	60
2.11	Discussion, Remarks, and Historical Notes	60
2.12	Exercises	61
3	Empirical Likelihood for Linear Functionals of the Cumulative Distribution Function	63
3.1	One-Sample Means	63
3.2	Proof of Theorem 23	69
3.3	Illustration	73
3.4	Two Independent Samples	78
3.5	Equality of k Medians	81
3.6	Functionals of the CDF and Functionals of Hazard	84
3.6.1	Alternative Proof of the Wilks Theorem for Means	87
3.7	Predictable Mean Function	87
3.8	Discussion, Historic Notes, and Remarks	88
3.9	Exercises	89
4	Empirical Likelihood Analysis of the Cox Model	91
4.1	Introduction	91
4.2	Empirical Likelihood Analysis of the Cox Model	92
4.2.1	Profile out the Baseline	94
4.2.2	Inference Involving Baseline	96
4.2.3	Parameters θ and λ	97
4.2.4	Confidence Intervals for $h(\beta, \theta)$ or $h(\beta, \lambda)$	101
4.3	Confidence Band for Baseline	104
4.4	An Alternative Empirical Likelihood Approach	107
4.5	Yang and Prentice Extension of the Cox Model	109
4.6	Historical Notes	116
4.7	Some Known Results about the Cox Model	116
4.8	Exercises	118
5	Empirical Likelihood Analysis of Accelerated Failure Time Models	119
5.1	AFT Models	119
5.2	AFT Regression Models	121
5.3	The Buckley–James Estimator	122
5.4	An Alternative EL for the B–J Estimator	125
5.5	Rank Estimator for the AFT Regression Model	128
5.6	AFT Correlation Models	128
5.7	EL Analysis of AFT Correlation Models	130
5.7.1	Quantile Regression Models	133
5.8	Discussion and Historical Remarks	135
5.9	Exercise	136

6	Computation of Empirical Likelihood Ratio with Censored Data	137
6.1	Empirical Likelihood for Uncensored Data	137
6.2	EL after Jackknife	139
6.3	One- or Two-Sample Hazard Features	143
6.4	EL testing concerning mean functions	145
6.4.1	EM Algorithm	146
6.4.2	A Recursive Algorithm	147
6.5	EL Test for Cox Models and Yang–Prentice Models	151
6.5.1	Yang and Prentice Model	151
6.6	Testing for AFT Models	152
6.6.1	The AFT Correlation Model	152
6.6.2	The AFT Regression Model	152
6.7	EL for Overdetermined Estimating Equations	153
6.8	Testing Part of the Parameter Vector	154
6.9	Intermediate Parameters	154
6.10	Lorenz Curve and Trimmed Mean	156
6.11	Confidence Intervals	163
6.12	Historic Note and Generalizations	167
6.13	Exercises	168
7	Optimality of Empirical Likelihood and Plug-in Empirical Likelihood	169
7.1	Pseudo Empirical Likelihood Ratio Test	169
7.2	Tests Based on Empirical Likelihood	172
7.3	Optimal Confidence Region	173
7.4	Illustrations	175
7.5	Adjustment of the Pseudo Empirical Likelihood Test	175
7.6	Weighted Empirical Likelihood	178
7.6.1	A Final Comment	182
7.7	Discussion and Historic Notes	182
8	Miscellaneous	183
8.1	Smoothing	183
8.2	Exponential Tilted Likelihood	183
8.3	Confidence Bands	185
8.3.1	One Sample	186
8.3.2	Two Independent Samples	187
8.3.3	Band within a Cox Model	188
8.4	Discussion and Historic Notes	189
8.5	Exercise	190
	References	191

List of Figures

1.1	Plot of the Kaplan–Meier curve, with 95% pointwise confidence intervals by log transformation.	8
2.1	A plot illustrating the feasibility of theta.	26
2.2	The -2LLR vs. β . Just determined case (Gehan).	48
2.3	The -2LLR vs. β . Overdetermined case.	49
2.4	Three -2 log empirical likelihood ratios vs. parameter θ .	50
2.5	Time changed and jump size changed Poisson process and its intensity function.	57
3.1	Q–Q plot of -2 log EL ratio under the null hypothesis.	70
3.2	Minimizing over the piecewise constant -2 log empirical likelihood ratios.	84
4.1	Confidence region of possible (β, λ) values and η contour lines over it.	103
4.2	Calculation of the baseline survival function in the Yang and Prentice model.	112
5.1	The -2 log empirical likelihood ratio contour plot, Stanford heart transplant data, 152 cases.	125
5.2	Alternative EL analysis of B–J. Contours of -2 log EL ratio. Compare to Fig. 5.1. Stanford heart transplant data, 152 cases.	127
6.1	Profile for minimizing over τ . Minimum happens near $\tau = 2.14$.	159
7.1	Size and shape of joint confidence regions for (a_1, a_2) from (1) EL and (2) plug-in EL.	176
7.2	Size and shape of EL regions and plug-in EL regions. Both types of regions center at the same location.	177
7.3	Three confidence regions by (1) EL, (2) plug-in EL and (3) weighted EL.	181

List of Tables

2.1	Rank Estimators and Empirical Likelihood 90% Confidence Intervals for AFT Model β with the Myeloma Data	45
2.2	The Choice of Two Versions of EL and Type of Survival Distribution	56

Preface

The first book on empirical likelihood was published in 2001 (by Owen and also from CRC), thirteen years after Owen published his first paper on empirical likelihood in 1988 [78].

This fascinating methodology attracted a lot of researchers and has been under rapid development ever since. Numerous papers have been published since then and the list is getting longer every day.

It is now fourteen years since the publication of Owen's 2001 book on empirical likelihood. I feel the time is perhaps ripe for another book on empirical likelihood.

Aside from the obvious accumulation of research progress in the fourteen years since, another obvious development is the vastly improved computing power and universal availability of computers. No longer are expensive workstations in the labs available only to a few. They are everywhere and in every student's backpack.

During the last 14 years, the software R became the most popular choice of language among statistics researchers, and went from version 1.x.x to 3.x.x. I feel the easy-to-use, widely available R software for calculating empirical likelihood will boost the everyday use of empirical likelihood and in turn stimulate more research in this area.

This book includes many worked out examples with the associated R code. You can copy and paste them into an R command window.

The R packages used in this book include `emplik`, `survival`, `KMsurv`, `ELYP`. We also briefly mention related packages `km.ci`, `kmc`, `gmm`, `rms`, `el.convex`. The latter group of packages is not crucial when reading this book. The package `emplik` is version 1.01 at the time of writing this book. This package has over 12 years of history. On the other hand, the package `ELYP` is less refined and is version 0.72.

We use the `survival` package for both its datasets and some of the estimation functions (for example, to obtain the Cox partial likelihood estimate). The package `KMsurv` contains only datasets, and we use them in several examples. The package `emplik` is the main package for calculations related to empirical likelihood, except for those related to the Cox model, which are in the package `ELYP`. The package `kmc` contains the functions that implement the recursive algorithm we describe in Chapter 6. It is still quite fluid and should eventually be integrated into `emplik` in the future. Finally, the package `km.ci` provides empirical likelihood confidence intervals and confidence bands for the Kaplan–Meier survival probabilities.

I will keep updating and uploading the package `emplik` and `ELYP` to the public

repository CRAN after the publication of the book, and maintain a Web page for any updates:

<http://www.ms.uky.edu/~mai/EmpLik.html>

The empirical likelihood method has its root in survival analysis. The very first paper originating the empirical likelihood method [112] is about empirical likelihood with the Kaplan–Meier estimator. So it seems to me that the empirical likelihood method naturally fits in with survival analysis. Also, over the years I have worked mostly on the empirical likelihood applications in survival analysis. So I chose to concentrate on this area.

Owen [81] covers much wider topics and also contains several sections about empirical likelihood with various censored, truncated, or other incomplete data. He discussed many forms of censoring, including interval censoring and double censoring. This book deals only with right censored data. Also, we do not discuss high-order asymptotic results for the empirical likelihood ratio. I feel the practical usefulness of high-order results in survival analysis concerning empirical likelihood ratio is not clear at this moment.

The core content of the book is Chapters 1, 2, 3, 4 (less Section 4.5), and 6. Chapter 1 discuss the empirical likelihood for right censored data, Chapter 6 for some computational tricks for censored data empirical likelihood, and the rest of the materials are pretty much standard survival analysis topics, treated with empirical likelihood. Basic knowledge of survival analysis is assumed.

Chapter 5 covers semi-parametric accelerated failure time models. This subject has a long history, but somehow standard software does not usually include this model and it is less used in practice (compared to the Cox regression model).

Section 4.5 discusses a recent extension of the Cox model by Yang and Prentice [129]. I include it here because I believe the empirical likelihood method is particularly suited for the statistical inference of this model.

Chapter 7 is about the optimality of confidence regions derived from empirical likelihood ratio tests or plug-in empirical likelihood ratio tests. It is a bit of a surprise that confidence regions can be so different in shape and orientation based on censored data. Chapter 8 collects mainly several empirical likelihood confidence band results, among other things.

There is a long list of people to whom I want to say THANK YOU. I am afraid the list is so long that I won't be able to stop for a long time. So instead of all the names, I shall list several categories.

First, all my colleagues. I have benefited tremendously over the years by reading your work and writings, by personal interactions, and in some cases, by collaborating on research. Some of the names appear in the reference list at the end of this book, but there are many more whose names do not appear in the references. THANK YOU!

I am also grateful to my students. I enjoyed working with you all.

I also want to acknowledge the support of an NSF grant.

I want to thank the many people who helped me put together this manuscript, correcting numerous typos and awkward grammar. All the remaining errors are my own.

Finally, I want to thank my family; they helped in this book project in numerous ways.

Mai Zhou

Introduction

Survival analysis has long been a classic area of statistical study. For example, the famous Kaplan–Meier estimator got its name from a paper published in the year 1958. Many textbooks on survival analysis are available and the list is still growing. The main survival analysis procedures are available in all major statistical software packages. On the other hand, empirical likelihood is a methodology that has only recently been developed. The name “empirical likelihood” seems to appear first in Owen’s 1988 paper. Only one book so far is available on empirical likelihood and most commercial statistical software does not yet include empirical likelihood procedures.

1.1 Survival Analysis

What is survival analysis? One might say “survival analysis is the statistical analysis of failure time data.” In fact, some books are titled exactly as such. It is certainly correct, but it begs people to ask “what is failure time data?,” which then takes longer to explain.

One might also say that “survival analysis is Kaplan–Meier estimator + log-rank test + Cox proportional hazards model.” This description is too simplistic, but certainly very specific and constructive.

Perhaps we should be asking: what is the difference between survival analysis and regular statistical analysis? Or what are the unique features of survival analysis not seen in other branches of statistics?

We can list several features unique to survival analysis:

1. In survival analysis, the parameters of interest are often the “hazard” instead of cumulative distribution function (CDF) or mean.
2. In survival analysis, the available data is subject to censoring.
3. In survival analysis, nonparametric procedures are more common.

Empirical likelihood is a nonparametric method and thus fits into the third point above for survival analysis. We also point out that the Kaplan–Meier estimator, log-rank test and Cox model are all nonparametric procedures. Let us discuss the above features in more detail.

1.1.1 Hazard Function

Let $F(t)$ denote the CDF of the random variable X of interest; then the cumulative hazard function is defined as

$$\Lambda(t) = \int_{-\infty}^t \frac{dF(s)}{1 - F(s-)} . \quad (1.1)$$

We comment that this definition is valid for either continuous or discrete CDF $F(t)$, and for X that can take negative values. If $F(t)$ is discrete, the integration is the Stieljes integral. When the CDF is continuous, $F(s-) = F(s)$, the integration on the right-hand side can be simplified to $-\log(1 - F(t))$ and thus we have (for the continuous case)

$$\Lambda(t) = -\log(1 - F(t)) . \quad (1.2)$$

If the CDF has a density $f(s)$, then

$$\Lambda(t) = \int_{-\infty}^t \frac{f(s)}{1 - F(s)} ds$$

and

$$\frac{\partial}{\partial t} \Lambda(t) = \frac{f(t)}{1 - F(t)} .$$

If we define the *hazard function* $h(t)$ as

$$h(t) = \frac{f(t)}{1 - F(t)} ,$$

then the relation between $\Lambda(t)$ and $h(t)$ is similar to that of CDF $F(t)$ to the density $f(t)$, i.e., $\Lambda(t) = \int_{-\infty}^t h(s) ds$ and $\frac{\partial}{\partial t} \Lambda(t) = h(t)$.

The probabilistic interpretation of hazard $h(t)$ is that $h(t)dt$ is the conditional probability of the random variable taking a value in $[t, t + dt)$, given it is larger than or equal to t :

$$h(t)dt = \frac{f(t)dt}{1 - F(t-)} = P(t \leq X < t + dt | X \geq t) .$$

Compare this to the similar interpretation for the density $f(t)$:

$$f(t)dt = P(t \leq X < t + dt) .$$

The hazard $h(t)$ must be nonnegative but does not have an upper bound. The cumulative hazard function $\Lambda(t)$ must be nonnegative and nondecreasing, but again can be unbounded. In fact, if the CDF is continuous, then $\Lambda(t)$ must be unbounded. This can be seen from $\Lambda(t) = -\log(1 - F(t))$. On the other hand, if the CDF is discrete, then $\Lambda(t)$ does not increase to infinity as t increases, but the last jump is always of size one. This can be seen by (supposing t^* is the last jump point)

$$\Delta\Lambda(t_{last}) = \Delta\Lambda(t^*) = \frac{\Delta F(t^*)}{1 - F(t^*-)} = 1 ,$$

because $\Delta F(t^*) = F(t^*+) - F(t^*-) = 1 - F(t^*-)$.

The inverse formula that recovers the CDF given a cumulative hazard is a bit awkward in the sense that the continuous and discrete versions look quite different: if the CDF/cumulative hazard is continuous, then

$$1 - F(t) = e^{-\Lambda(t)} . \quad (1.3)$$

If the CDF/cumulative hazard is purely discrete, then we have

$$1 - F(t) = \prod_{s \leq t} (1 - \Delta\Lambda(s)) , \quad (1.4)$$

where $\Delta\Lambda(s) = \Lambda(s+) - \Lambda(s-)$. We notice that there are at most a countable many terms in the product, because there are at most a countable number of jumps in a monotone function.

In the case of a partly continuous, partly discrete CDF/cumulative hazard, we have to combine the two formulae:

$$1 - F(t) = e^{-\Lambda_c(t)} \prod_{s \leq t} (1 - \Delta\Lambda(s)) \quad (1.5)$$

where $\Lambda_c(t)$ is the continuous part of the cumulative hazard:

$$\Lambda_c(t) = \Lambda(t) - \sum_{s \leq t} \Delta\Lambda(s) .$$

Our discussion later in this book will mostly focus either on the continuous case or the purely discrete case, not on the mixed case.

The nonparametric estimation of the cumulative hazard function leads to the Nelson–Aalen estimator. The two sample log-rank test can be viewed as comparing the two hazard functions from two samples. The Cox proportional hazards model is a regression model which models how the ratio of hazards relates to the covariates. We shall discuss the Cox model in Chapter 4 and review the Nelson–Aalen estimator and log-rank test in subsections later in the present chapter.

Remark: At first glance, it is not clear how a rather innocent looking transformation of CDF to hazard has such an influence on survival analysis. For one thing, it removed the constraint that the jumps of a CDF must sum to one. Second, by working on conditional probabilities, it localized the parameters and made the estimation problem easier with censoring. This also leads to the application of martingales in survival analysis.

1.1.2 Censored Observations

The random variable of interest in survival analysis is “time to failure,” denoted by X . Typically this is a positive, continuous random variable. Between the start and the end of a “life,” a lot can happen, and often there are some conditions that prevent us from following up the “life” to its eventual failure. This leads to censoring.