Statistics for Biology and Health

David G. Kleinbaum
Mitchel Klein

Survival Analysis

A Self-Learning Text

Second Edition

生存分析自学教材

第2版

Springer

世界图书出版公司
www.wpcbj.com.cn

David G. Kleinbaum
Mitchel Klein

Survival Analysis

A Self-Learning Text

Second Edition

Springer

David G. Kleinbaum
Department of Epidemiology
Rollins School of Public Health at
Emory University
1518 Clifton Road NE
Atlanta GA 30306
Email: dkleinb@sph.emory.edu

Mitchel Klein
Department of Epidemiology
Rollins School of Public Health at
Emory University
1518 Clifton Road NE
Atlanta GA 30306
Email: mklein@sph.emory.edu

Library of Congress Control Number: 2005925181

ISBN-13: 978-0-387-23918-7 e-ISBN: 978-0-387-29150-5

Printed on acid-free paper.

springer.com

Statistics for Biology and Health

Statistics for Biology and Health

Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations.
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints.
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS.
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction. 2nd ed.
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions using R and Bioconductor
Hougaard: Analysis of Multivariate Survival Data.
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text. 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations.
Nielsen: Statistical Methods in Molecular Evolution.
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators.
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software.
Salsburg: The Use of Restricted Significance Tests in Clinical Trials.
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations.
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case.
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model.
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Zhang/Singer: Recursive Partitioning in the Health Sciences.

To *Rosa Parks*

Nelson Mandela

Dean Smith

Sandy Koufax

And

countless other persons, well-known or unknown,
who have had the courage to stand up for their beliefs for the benefit of humanity.

Preface

This is the second edition of this text on survival analysis, originally published in 1996. As in the first edition, each chapter contains a presentation of its topic in "lecture-book" format together with objectives, an outline, key formulae, practice exercises, and a test. The "lecture-book" format has a sequence of illustrations and formulae in the left column of each page and a script in the right column. This format allows you to read the script in conjunction with the illustrations and formulae that high-light the main points, formulae, or examples being presented.

This second edition has expanded the first edition by adding three new chapters and a revised computer appendix. The three new chapters are:

Chapter 7. Parametric Survival Models
Chapter 8. Recurrent Event Survival Analysis
Chapter 9. Competing Risks Survival Analysis

Chapter 7 extends survival analysis methods to a class of survival models, called parametric models, in which the distribution of the outcome (i.e., the time to event) is specified in terms of unknown parameters. Many such parametric models are acceleration failure time models, which provide an alternative measure to the hazard ratio called the "acceleration factor". The general form of the likelihood for a parametric model that allows for left, right, or interval censored data is also described. The chapter concludes with an introduction to frailty models.

Chapter 8 considers survival events that may occur more than once over the follow-up time for a given subject. Such events are called "recurrent events". Analysis of such data can be carried out using a Cox PH model with the data layout augmented so that each subject has a line of data for each recurrent event. A variation of this approach uses a stratified Cox PH model, which stratifies on the order in which recurrent events occur. The use of "robust variance estimates" are recommended to adjust the variances of estimated model coefficients for correlation among recurrent events on the same subject.

Chapter 9 considers survival data in which each subject can experience only one of several different types of events ("competing risks") over follow-up. Modeling such data can be carried out using a Cox model, a parametric survival model or a model which uses *cumulative incidence* (rather than survival).

The Computer Appendix in the first edition of this text has now been revised and extended to provide step-by-step instructions for using the computer packages STATA (version 7.0), SAS (version 8.2), and SPSS (version 11.5) to carry out the survival analyses presented in the main text. These computer packages are described in separate self-contained sections of the Computer Appendix, with the analysis of the same datasets illustrated in each section. The SPIDA package used in the first edition is no longer active and has therefore been omitted from the appendix and computer output in the main text.

In addition to the above new material, the original six chapters have been modified slightly to correct for errata in the first edition, to clarify certain issues, and to add theoretical background, particularly regarding the formulation of the (partial) likelihood functions for the Cox PH (Chapter 3) and extended Cox (Chapter 6) models.

The authors' website for this textbook has the following weblink: http://www.sph.emory.edu/~dkleinb/surv2.htm

This website includes information on how to order this second edition from the publisher and a freely downloadable zip-file containing data-files for examples used in the textbook.

Suggestions for Use

This text was originally intended for self-study, but in the nine years since the first edition was published, it has also been effectively used as a text in a standard lecture-type classroom format. The text may also be use to supplement material covered in a course or to review previously learned material in a self-instructional course or self-planned learning activity. A more individualized learning program may be particularly suitable to a working professional who does not have the time to participate in a regularly scheduled course.

In working with any chapter, the learner is encouraged first to read the abbreviated outline and the objectives and then work through the presentation. The reader is then encouraged to read the detailed outline for a summary of the presentation, work through the practice exercises, and, finally, complete the test to check what has been learned.

Recommended Preparation

The ideal preparation for this text on survival analysis is a course on quantitative methods in epidemiology and a course in applied multiple regression. Also, knowledge of logistic regression, modeling strategies, and maximum likelihood techniques is crucial for the material on the Cox and parametric models described in chapters 3–9.

Recommended references on these subjects, with suggested chapter readings are:

Kleinbaum D, Kupper L, Muller K, and Nizam A, **Applied Regression Analysis and Other Multivariable Methods, Third Edition,** Duxbury Press, Pacific Grove, 1998, Chapters 1–16, 22–23

Kleinbaum D, Kupper L and Morgenstern H, **Epidemiologic Research: Principles and Quantitative Methods,** John Wiley and Sons, Publishers, New York, 1982, Chapters 20–24.

Kleinbaum D and Klein M, **Logistic Regression: A Self-Learning Text, Second Edition,** Springer-Verlag Publishers, New York, Chapters 4–7, 11.

Kleinbaum D, **ActivEpi-A CD Rom Electronic Textbook on Fundamentals of Epidemiology,** Springer-Verlag Publishers, New York, 2002, Chapters 13–15.

A first course on the principles of epidemiologic research would be helpful, since all chapters in this text are written from the perspective of epidemiologic research. In particular, the reader should be familiar with the basic characteristics of epidemiologic study designs, and should have some idea of the frequently encountered problem of controlling for confounding and assessing interaction/effect modification. The above reference, **ActivEpi,** provides a convenient and hopefully enjoyable way to review epidemiology.

Acknowledgments

We wish to thank Allison Curry for carefully reviewing and editing the previous edition's first six chapters for clarity of content and errata. We also thank several faculty and MPH and PhD students in the Department of Epidemiology who have provided feedback on the first edition of this text as well as new chapters in various stages of development. This includes Michael Goodman, Mathew Strickland, and Sarah Tinker. We thank Dr. Val Gebski of the NHMRC Clinical Trials Centre, Sydney, Australia, for providing continued insight on current methods of survival analysis and review of new additions to the manuscript for this edition.

Finally, David Kleinbaum and Mitch Klein thank Edna Kleinbaum and Becky Klein for their love, support, companionship, and sense of humor during the writing of this second edition.

Contents

1 Introduction to Survival Analysis

Introduction

This introduction to survival analysis gives a descriptive overview of the data analytic approach called **survival analysis.** This approach includes the type of problem addressed by survival analysis, the outcome variable considered, the need to take into account "censored data," what a survival function and a hazard function represent, basic data layouts for a survival analysis, the goals of survival analysis, and some examples of survival analysis.

Because this chapter is primarily descriptive in content, no prerequisite mathematical, statistical, or epidemiologic concepts are absolutely necessary. A first course on the principles of epidemiologic research would be helpful. It would also be helpful if the reader has had some experience reading mathematical notation and formulae.

Abbreviated Outline

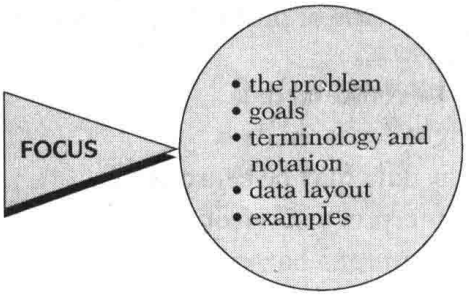
The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing the chapter, the learner should be able to:

1. Recognize or describe the type of problem addressed by a survival analysis.
2. Define what is meant by censored data.
3. Define or recognize right-censored data.
4. Give three reasons why data may be censored.
5. Define, recognize, or interpret a survivor function.
6. Define, recognize, or interpret a hazard function.
7. Describe the relationship between a survivor function and a hazard function.
8. State three goals of a survival analysis.
9. Identify or recognize the basic data layout for the computer; in particular, put a given set of survival data into this layout.
10. Identify or recognize the basic data layout, or components thereof, for understanding modeling theory; in particular, put a given set of survival data into this layout.
11. Interpret or compare examples of survivor curves or hazard functions.
12. Given a problem situation, state the goal of a survival analysis in terms of describing how explanatory variables relate to survival time.
13. Compute or interpret average survival and/or average hazard measures from a set of survival data.
14. Define or interpret the hazard ratio defined from comparing two groups of survival data.

Presentation



This presentation gives a general introduction to survival analysis, a popular data analysis approach for certain kinds of epidemiologic and other data. Here we focus on the problem addressed by survival analysis, the goals of a survival analysis, key notation and terminology, the basic data layout, and some examples.

I. What Is Survival Analysis?

We begin by describing the type of analytic problem addressed by survival analysis. Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs.*

Outcome variable: **Time until an event occurs**

Start follow-up → TIME → Event

By **time,** we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the **age** of an individual when an event occurs.

Event: death
disease
relapse
recovery

By **event,** we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual.

Assume 1 event

> 1 event → Recurrent event or Competing risk

Although more than one event may be considered in the same analysis, we will assume that only one event is of designated interest. When more than one event is considered (e.g., death from any of several causes), the statistical problem can be characterized as either a recurrent events or a **competing risk** problem, which are discussed in Chapters 8 and 9, respectively.

Time ≡ survival time

Event ≡ failure

In a survival analysis, we usually refer to the time variable as **survival time,** because it gives the time that an individual has "survived" over some follow-up period. We also typically refer to the event as a **failure,** because the event of interest usually is death, disease incidence, or some other negative individual experience. However, survival time may be "time to return to work after an elective surgical procedure," in which case failure is a positive event.