Springer Series in Statistics

Per Kragh Andersen
Ørnulf Borgan
Richard D. Gill
Niels Keiding

Statistical Models Based on Counting Processes

基于计数过程的统计模型

Springer-Verlag
世界图书出版公司

Per Kragh Andersen
Ørnulf Borgan
Richard D. Gill
Niels Keiding

Statistical Models Based on Counting Processes

With 128 Illustrations

Springer-Verlag
世界图书出版公司
北京 · 广州 · 上海 · 西安

书　　名：Statistical Models Based on Counting Processes
作　　者：P.K.Andersen, O. Borgan, et al.
中 译 名：基于计数过程的统计模型
出 版 者：世界图书出版公司北京公司
印 刷 者：北京世图印刷厂
发　　行：世界图书出版公司北京公司（北京朝阳门内大街 137 号　100010）
开　　本：大 32 开　850 × 1168　印　张：24.5
版　　次：1998 年 8 月第 1 版　1998 年 8 月第 1 次印刷
书　　号：7-5062-3817-9/ O•227
版权登记：图字 01-98-0367
定　　价：116.00 元

Springer Series in Statistics

Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Andrews/Herzberg: Data: A Collection of Problems from Many Fields for the Student and Research Worker.
Anscombe: Computing in Statistical Science through APL.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Bolfarine/Zacks: Prediction Theory for Finite Populations.
Brémaud: Point Processes and Queues: Martingale Dynamics.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Daley/Vere-Jones: An Introduction to the Theory of Point Processes.
Dzhaparidze: Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models.
Farrell: Multivariate Calculation.
Federer: Statistical Design and Analysis for Intercropping Experiments.
Fienberg/Hoaglin/Kruskal/Tanur (Eds.): A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
Fisher/Sen: The Collected Works of Wassily Hoeffding.
Good: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.
Goodman/Kruskal: Measures of Association for Cross Classifications.
Grandell: Aspects of Risk Theory.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Hartigan: Bayes Theory.
Heyer: Theory of Statistical Experiments.
Jolliffe: Principal Component Analysis.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.
Kres: Statistical Tables for Multivariate Analysis.
Le Cam: Asymptotic Methods in Statistical Decision Theory.
Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts.
Longford: Models for Uncertainty in Educational Testing.
Manoukian: Modern Concepts and Theorems of Mathematical Statistics.
Miller, Jr.: Simultaneous Statistical Inference, 2nd edition.
Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of *The Federalist Papers.*
Pollard: Convergence of Stochastic Processes.
Pratt/Gibbons: Concepts of Nonparametric Theory.
Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.

(continued after index)

Per Kragh Andersen/Niels Keiding
Statistical Research Unit
University of Copenhagen
Copenhagen
Denmark

Ørnulf Borgan
Institute of Mathematics
University of Oslo
Oslo
Norway

Richard D. Gill
Faculty of Mathematics
University of Utrecht
Utrecht
The Netherlands

Mathematics Subject Classification (1991): 60G07, 60G12, 60J25

Library of Congress Cataloging-in-Publication Data
Statistical models based on counting processes / Per Kragh Andersen
... [et al.].
p. cm.—(Springer series in statistics)
Includes bibliographical references and index.
ISBN 0-387-94519-9
1. Point processes. 2. Mathematical statistics. I. Andersen,
Per Kragh. II. Series.
QA274.42.S75 1992
519.2'3—dc20 92-11354

Printed on acid-free paper.

Reprinted in China by Beijing World Publishing Corporation, 1998

9 8 7 6 5 4 3 2 1

ISBN 0-387-94519-9 Springer-Verlag New York Berlin Heidelberg

Preface

One of the most remarkable examples of fast technology transfer from new developments in mathematical probability theory to applied statistical methodology is the use of counting processes, martingales in continuous time, and stochastic integration in *event history analysis*. By this (or generalized survival analysis), we understand the study of a collection of individuals, each moving among a finite (usually small) number of states. A basic example is moving from alive to dead, which forms the basis of *survival analysis*. Compared to other branches of statistics, this area is characterized by the dynamic temporal aspect, making modelling via the intensities useful, and by the special patterns of incompleteness of observation, of which *right-censoring* in survival analysis is the most important and best known example.

In this book our aim is to present a theory for handling such statistical problems, when observation is assumed to happen in continuous time. In most of the book, a minimum of assumptions are made (non- and semiparametric theory), although one chapter considers parametric models. The detailed plan of the book is described in Chapter I.

Our desire has been to give an exposition in terms of the necessary mathematics (including previously unfamiliar tools that have proved useful), but at the same time illustrate most methods by concrete practical applications, primarily from our own biostatistical experience. We also wanted to indicate some recent lines of research. As a result, the book moves along at varying levels of mathematical sophistication and also at varying levels of subject-matter discussion. We realize that this will frustrate some readers, in particular those who had expected to find a smooth textbook. In Section I.1, we suggest a reading guide for various categories of readers, and in Section I.2, we give a synopsis of the developments of the field up to now.

The suggestion of writing a monograph on this material came from Klaus Krickeberg in 1982. In assembling a group of authors, we wanted to include our colleague and friend Odd O. Aalen, who started the theory and keeps contributing to it; however, Odd wanted to concentrate on building up medical statistics in Oslo, Norway. The work on the book has become something of a life-style for the four of us, involving drafting sections individually and then meeting, usually in Copenhagen, to discuss in great detail the merits and otherwise of the latest contributions. These meetings always lasted for three full days, with discussions going on at the dinner table to the wonder of our wives and children, none of whom believed we really wanted (were able to?) finish the project. We will miss these opportunities for very positive professional and human interaction, and the concept of "Vol. II" has become our consolation. (Do not worry, this will take a while!)

We have enjoyed extremely positive interest from all colleagues. First, our Nordic colleagues who participated in getting the theory developed have always been ready to follow up with good advice. Besides Odd Aalen, we have interacted particularly with Martin Jacobsen, as well as with Elja Arjas, Nils Lid Hjort, Søren Johansen, and Gert Nielsen. Overseas, we owe particular thanks to Dorota Dabrowska, who read several chapters in great detail and kept us informed of several fascinating new developments. We have been given good advice and been entrusted with prepublication versions of new results by many colleagues, including Chang-Jo F. Chung, Hani Doss, Priscilla Greenwood, Piet Groeneboom, Jean Jacod, John P. Klein, Tze Leung Lai, Ian McKeague, Jens Perch Nielsen, Michael Sørensen, Yangin Sun, Aad van der Vaart, Bert van Es, Mark van Pul, Wolfgang Wefelmeyer, and Jon A. Wellner.

As with any effort of this kind, countless drafts have been produced, some before the age of mathematical word processing. Although later phases have sent (most of) the authors themselves to the TeX keyboards (in Copenhagen with much support from Peter Dalgaard), many versions were done by our secretaries in Copenhagen (in particular, Susanne Kragskov and Kathe Jensen), Oslo (Dina Haraldsson and Inger Jansen), Amsterdam (Carolien Pol-Swagerman), and Utrecht (Karin Berlang and Diana van Doorn). Most calculations and graphs were done (in S) by Lars Sommer Hansen and Klaus Krøier, and Mark van Pul and Kristian Merckoll did some additional calculations.

Our work with the book has been done as part of our appointments at the University of Copenhagen (Per Kragh Andersen, Niels Keiding, the latter for part of the period as Danish Ministry of Education Research Professor), the University of Oslo (Ørnulf Borgan), and Centre for Mathematics and Computer Science, Amsterdam, and the University of Utrecht (Richard D. Gill). Some work was done during visits to the Department of Biostatistics, University of Washington; the Medical Research Council Biostatistics Unit, Cambridge, England; the Department of Mathematics, University of Western Australia; and the Department of Statistics, The Ohio State University.

We acknowledge the hospitality and interest everywhere. Travel costs and secretarial and computing assistance have been financed by our institutions, with supplements from the Danish Medical Research Council (grant no. 12-9229), the Norwegian Council for Research in Science and the Humanities, and Johan and Mimi Wessmann's foundation.

February 1992

København	Per Kragh Andersen
Oslo	Ørnulf Borgan
Utrecht	Richard D. Gill
København	Niels Keiding

Contents

CHAPTER I

Introduction

I.1. General Introduction to the Book

I.1.1. Event History Analysis

As explained in the Preface, we study a collection of individuals, each moving among a finite (usually small) number of states. The exact transition times in continuous time form the modelling basis of the phenomena, although often these times are only incompletely observed. The best known example of such incomplete observation is *right-censoring* in classical survival analysis; here, not all of a set of independent lifetimes are observed, so that for some of them it is only known that they are larger than some specific value. We shall mention other kinds of censoring such as interval-censoring and left-censoring as well as *truncation* and *filtering*. Left-truncation in survival analysis means that an individual is included only if its lifetime is larger than some value, whereas by filtering, we mean that the individual is not under observation all the time, but only when a suitable indicator process is switched on.

During the 1960s, an extensive theory of martingales in continuous time and stochastic integration was developed, primarily in France. In the early 1970s, these tools were applied in control theory and electrical engineering.

I.1.2. Counting Processes, Martingales, Stochastic Integrals, Product-Integration

For event history analysis, a major contribution was the 1975 Berkeley Ph.D. dissertation "Statistical Inference for a Family of Counting Processes" by

O.O. Aalen. The basic nonparametric statistical problems for censored data were studied in terms of the conditional *intensity* (or *intensity process*) of the *counting process* that records the (uncensored) events as time proceeds. Formulating many statistical models of interest in terms of a multiplicative decomposition of the intensity process of this counting process into a product of a purely deterministic factor (*hazard*) and an observable factor, Aalen's main contribution was to show that the models may then be analyzed in a unified manner by these tools. The difference between the counting process and the integrated intensity process is a *martingale*. Exact (that is, nonasymptotic) results such as unbiasedness, as well as estimators of variability, are obtained by applying results on *stochastic integration* with respect to this martingale. Asymptotic statistical theory follows using *martingale central limit theory*. Finally, to convert the results to the conventional distribution function (or *survival function*), the general tool of *product-integration* is useful (Aalen and Johansen, 1978). This tool, however, has been very slow in penetrating statistical theory; for a review see Gill and Johansen (1990).

I.1.3 General Aim of This Book

Our aim in this book is, first, to present the above theory, completing some of its mathematical details, as well as surveying a range of topics developed since the first journal publication of the theory (Aalen, 1978b; Aalen and Johansen, 1978). A brief survey of the historical development of the subject is given in Section I.2. Second, we include some very concrete applications, providing detailed empirical discussions of a number of practical examples, mainly from our own experience. We here also consider methods for assessing goodness of fit in detail, displaying results graphically, and adapting the methods of statistical inference to the occasionally complicated observational patterns that one meets in practice. Third, the full mathematical exposition of the theory requires some mathematical tools that we have found useful to summarize.

At the end of this section, we list some of the aims that we do *not* seek to fulfill in this work. Before that, let us first explain the plan of the book, and then suggest how different categories of readers could navigate through the somewhat heterogeneous material presented.

I.1.4. Plan of the Book

To emphasize the practical purpose of the methods developed here, brief presentations are given in Section I.3 of 19 practical examples that will serve as illustrations in the following chapters. The examples are chosen to illustrate various patterns of event history often met in practice, focusing in particular on the various relevant time scales (calendar time, age, duration since onset of disease, etc.).

Chapter II presents the mathematical tools necessary for a full appreciation of the development in later chapters. A textbook-type treatment would have exceeded the limits of the book considerably; instead, we attempt a brief exposition that nevertheless contains precise statements of the definitions and results used later. Section II.1 explains informally but also briefly the key concepts in the basic example of classical survival analysis.

In Chapter III, we provide a rather detailed discussion of the specification of the statistical models in terms of the intensity process. The bulk of this chapter is a mathematical discussion of possible patterns of censoring and other kinds of incomplete observation (truncation, filtering), with extensive discussion of specific examples, mostly motivated by the empirical studies presented in Section I.3.

Chapter IV is an exposition of *nonparametric estimation* in Aalen's multiplicative intensity model. This includes the "Nelson–Aalen" estimator of the cumulative intensity, kernel smoothing methods for the intensity itself, the "Kaplan–Meier" estimator of the survival function and its generalization: the estimator of the transition matrix of a nonhomogeneous Markov process proposed by Aalen and Johansen (1978). The methods are illustrated and further specified in relation to the empirical studies. Chapter V treats k-sample *hypothesis tests* of the relevance of suitable stratifications, including one-sample tests for comparison with a standard. Linear rank tests are emphasized and illustrated with several empirical examples, while there are brief introductions to various specialized rank tests, tests based on the complete intensity process, and sequential tests.

Several basic aspects of *parametric* inference in life testing and epidemiology are usefully studied in the present framework. Chapter VI develops maximum likelihood and M-estimators and illustrates these methods on several examples. Asymptotic statistical theory is given, as well as methods for assessing goodness of fit, including generalization of the "total time on test" methodology from reliability.

Often, stratification is not sufficient and more specific *regression models* are required to obtain the desired insight into the influence of covariates on the intensities of occurrence of the relevant events. In survival analysis, the "semiparametric" regression model and the associated tool of partial likelihood, both due to Cox (1972), were seminal. Cox's model postulates the intensity to be a product of a parametric function of the covariates and an arbitrary function of time, and this model generalizes directly to counting processes. Chapter VII contains an extensive mathematical and applied statistical discussion of this approach. Other regression models are also discussed, particularly a (multivariate) linear intensity model due to Aalen (1980).

The methods developed in Chapters IV–VII are motivated by reasonable though sometimes rather ad hoc mathematical–statistical arguments (moment equations and similar estimating equations, unorthodox "maximum likelihood" constructions). Results on optimality or efficiency of the proposed methods are not directly available from the derivations; in fact, only a few scattered results on these topics are available in the literature. Chapter VIII provides a brief introduction to LeCam's theory of local asymptotic normality and contiguity, as well as the recent general mathematical theory of semiparametric and nonparametric estimation (Bickel et al., 1993; van der Vaart, 1988). It is shown that this provides a sufficiently general mathematical framework to allow efficiency results to be derived, which may sometimes be linked to generalized maximum likelihood properties.

Many further developments are possible within the general framework of statistical models based on counting processes. We restrict ourselves to brief indications of two lines of developments. First, heterogeneity among individuals is not always usefully described by stratification or regression modelling; sometimes a description as random variation may be necessary if not all relevant covariates are measured or to describe correlation among individuals. *Random effect models* are very natural and well known for the normal distribution, but, as shown in Chapter IX, it is still possible to define models and suggest estimation and hypothesis testing methods for event history models with random individual variation (often called random *frailty* or "dependent failure times"). In Chapter X, we report briefly on the fast growing area of generalization of the methods of *multivariate time scales* (calendar time and age; calendar time and duration on trial; duration on trial and duration on a particular, possibly intermittent, treatment). Specific surveys are given of sequential analysis of censored data with staggered entry and of multivariate hazard measures and multivariate Kaplan–Meier estimators for dependent survival times.

I.1.5. Reading Guides

It is obvious from the above synopsis of the book that various categories of readers may be interested in different parts of the material presented.

The *biostatistician* or *reliability engineer* interested primarily in the practical methods should survey our empirical examples in Section I.3 and study the introductory Section II.1, but should be able to follow the specific descriptions in Chapters IV–VII, IX, and X without a detailed study of the mathematical prerequisites in the rest of Chapter II and the detailed mathematical discussion in Chapter III. However, a cursory reading of the first half of Chapter III would be necessary. It is perhaps appropriate to emphasize that this book is *not* intended as a textbook in biostatistics—the mathematical theory has guided our selection of methods, and sometimes other approaches will be preferable in practice.

The *statistician* wanting an introduction to this area should go through Chapter I to see which substantive areas motivated the methods described, become familiar with the underlying mathematics in Chapter II, and understand the general points of censoring, truncation, and filtering in Chapter III. From this viewpoint, one may prefer a more cursory reading of Chapters IV–VII including several illustrative examples.

The *mathematical–statistical specialist* (in this field) looking for "what's new" would, in principle, not need Chapter II, though we hope that at least some from this group will appreciate our summary. The density of unfamiliar material may be highest in Chapters III, IV.4, VI.3, VII (later sections), VIII, IX, and X.3, but there are innovations throughout.

Finally, some readers may be *probabilists*, wanting an explanation of the statistical application of tools (cf. Chapter II) well known to them. They should read Chapter III carefully and study the empirical examples in Chapter I and selected cases about how the empirical problems raised there are handled in the statistical methodology Chapters IV–VII and IX.

I.1.6. Some Areas Not Covered

The book stops short of covering a number of important related theoretical and practical issues.

First, throughout the book we use only continuous-time methods. In most chapters, we study non- and semiparametric methods which are most relevant for small and medium-sized data sets with detailed information. For larger data sets, it will often be preferable to use the "parametric" log–linear intensity models deriving from assumptions of piecewise constant intensities, see Chapter VI and Section VII.6. The data also may be available only in grouped time, in which case generalized linear models based on the binomial distribution may be useful; these are not discussed here. Such methods will often be relevant in *epidemiological, demographic*, and *actuarial* applications. Second, our discussion has been influenced by our background in biostatistics; further applications in *reliability and industrial life testing* are also relevant at the theoretical level, as Arjas (1989) showed in his survey. Finally, although we have developed (and/or extended) many *computer algorithms* in connection with the practical examples, we have not carried any of these through to a state where we should like others to run the risk of using them. The graphs were made in S (Becker et al., 1988).

We should finally mention that as we finalized this book, the excellent textbook by Fleming and Harrington (1991) appeared. These authors provided a comprehensive introduction to the basic mathematical theory needed for the counting process approach and surveyed the by now classical statistical theory for censored survival data, using very well-documented examples from their own experience as illustration. It is likely that many readers will appreciate having access to Fleming and Harrington's book before and during the study of the present monograph.

I.2. Brief Survey of the Development of the Subject

In this section, we outline some general tendencies in the historical development of statistical inference for event history data, based on counting processes. Specific details are contained in the relevant chapters and their associated bibliographic remarks.

I.2.1. Aalen's Ph.D. Dissertation and Its Background

The point of departure for this monograh was O.O. Aalen's Ph.D. dissertation (1975). Here, concepts from classical demographical/actuarial life table analysis were studied using continuous-time martingale theory (Meyer, 1966), stochastic integration (Kunita and Watanabe, 1967), and counting process theory [Brémaud, 1972; Jacod, 1973, 1975; surveyed by Brémaud and Jacod (1977)] such as applied in control theory and electrical engineering (Boel, Varaiya, and Wong, 1975a,b). The asymptotic results may then be obtained from martingale central limit theory, in Aalen's (1975, 1977) original version based on the discrete-time results by McLeish (1974).

There were several lines in practical statistics leading up to this work. The classical life table analysis was interpreted statistically by Kaplan and Meier (1958) (Kaplan and Meier "almost simultaneously prepared separate manuscripts that were in many respects similar"; cf. the Notes about Authors on pp. 562–563 of the *JASA* issue). Sverdrup (1965) and Hoem (1969a,b, 1971) embedded well-established techniques from demography and actuarial mathematics into statistical inference for Markov processes. Nelson (1969) had presented a "hazard plot for incomplete failure time data" which had also been studied by Altshuler (1970) in the context of competing risks in animal experiments. Aalen (1972) in his Norwegian cand.real. (master's) thesis independently discussed this estimator (in this monograph we call it the Nelson–Aalen estimator). In survival analysis, nonparametric rank tests for censored data had begun to appear (Gehan, 1965; Mantel, 1966; Efron, 1967; Breslow, 1970; Peto and Peto, 1972).

For the classical competing risks model, Aalen (1976) gave a systematic exact and asymptotic theory for the cumulative hazard estimator, whereas Aalen (1978a), further elaborated by Fleming (1978a,b), gave estimators for various transition probabilities in this model. The asymptotic theory here was primarily based on the approach by Pyke and Shorack (1968) as utilized in the important paper by Breslow and Crowley (1974) on weak convergence of the Nelson–Aalen and Kaplan–Meier estimators.

I.2.2. Early Publications on Statistical Inference for Counting Processes

The journal article by Aalen (1978b) (based on his Ph.D. thesis) presented the general counting process framework, the Nelson–Aalen estimator of the cu-