Statistics for Biology and Health

Odd O. Aalen
Ørnulf Borgan
Håkon K. Gjessing

Survival and Event History Analysis

A Process Point of View

Springer

Statistics for Biology and Health

Statistics for Biology and Health

Aalen/Borgan/Gjessing: Survival and Event History Analysis: A Process Point of View
Bacchieri/Cioppa: Fundamentals of Clinical Research
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints
Duchateau/Janssen: The Frailty Model
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
Martinussen/Scheike: Dynamic Regression Models for Survival Data
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
Nielsen: Statistical Methods in Molecular Evolution
O'Quigley: Proportional Hazards Regression
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software
Proschan/LanWittes: Statistical Monitoring of Clinical Trials: A Unified Approach
Siegmund/Yakir: The Statistics of Gene Mapping
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
Sun: The Statistical Analysis of Interval-censored Failure Time Data
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model
Ting: Dose Finding in Drug Development
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL
Zhang/Singer: Recursive Partitioning in the Health Sciences
Zuur/Ieno/Smith: Analysing Ecological Data

Odd O. Aalen • Ørnulf Borgan • Håkon K. Gjessing

Survival and Event History Analysis

A Process Point of View

Springer

Odd O. Aalen
Department of Biostatistics
Institute of Basic Medical Sciences
University of Oslo
Oslo, Norway
o.o.aalen@medisin.uio.no

Ørnulf Borgan
Department of Mathematics
University of Oslo
Oslo, Norway
borgan@math.uio.no

Håkon K. Gjessing
Norwegian Institute of Public Health
Oslo, Norway
and
Section for Medical Statistics
University of Bergen
Bergen, Norway
hakon.gjessing@fhi.no

Series Editors

M. Gail
National Cancer Institute
Rockville, MD 20892
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

K. Krickeberg
Le Chatelet
F-63270 Manglieu
France

W. Wong
Department of Statistics
Stanford University
Stanford, CA 94305-4065
USA

J. Samet
Department of Epidemiology
School of Public Health
Johns Hopkins University
615 Wolfe Street
Baltimore, MD 21205-2103
USA

ISBN 978-0-387-20287-7 e-ISBN 978-0-387-68560-1
DOI: 10.1007/978-0-387-68560-1

Library of Congress Control Number: 2008927364

9 8 7 6 5 4 3 2 1

springer.com

To Marit, John, and Margrete

To Tone, Eldrid, and Yngve

To Liên, Eli, Einar, Wenche, and Richard

Preface

Survival and event history analysis have developed into one of the major areas of biostatistics, with important applications in other fields as well, including reliability theory, actuarial science, demography, epidemiology, sociology, and econometrics. This has resulted in a number of substantial textbooks in the field. However, rapidly developing theoretical models, combined with an ever-increasing amount of high-quality data with complex structures, have left a gap in the literature. It has been our wish to provide a book that exposes the rich interplay between theory and applications. Without being unnecessarily technical, we have wanted to show how theoretical aspects of statistical models have direct, intuitive implications for applied work. And conversely, how apparently disparate and complex features of data can be set in a theoretical framework that enables comprehensive modeling.

Most textbooks in survival analysis focus on the occurrence of single events. In actual fact, much event history data consist of occurrences that are repeated over time or related among individuals. The reason for this is, in part, the development toward increased registration and monitoring of individual life histories in clinical medicine, epidemiology, and other fields. The material for research and analysis is therefore more extensive and complex than it used to be, and standard approaches in survival analysis are insufficient to handle this.

It is natural to view such life histories as stochastic processes, and this is the basic idea behind our book. We start with the now classical counting process theory and give a detailed introduction to the topic. Leaving out much mathematical detail, we focus on understanding the important ideas. Next, we introduce standard survival analysis methods, including Kaplan-Meier and Nelson-Aalen plots and Cox regression. We also give a careful discussion of the additive hazards model. Then we extend the use of the counting process framework by counting several events for each individual. This yields very fruitful models, especially when combined with the additive hazards model. We further include time-dependent covariates, or marker processes, to define what we term dynamic path analysis, an extension of classical path analysis to include time. This allows us to explicitly analyze how various processes influence one another. Most of this is new, or very recently developed, material.

Another new aspect of the book is the explicit connection drawn between event history analysis and statistical causality. The focus is on causal formulations where time is explicitly present, including ideas like Granger causality, local independence, and dynamic path analysis.

Unique to this book is the emphasis on models that give insight into the rather elusive concept of hazard rate, and the various shapes the hazard rate can have. The effect of unobserved heterogeneity, or frailty, is broadly discussed with focus on a number of artifacts implied by the frailty structure as well as on applying these models to multivariate survival data. A new class of process-based frailty models is introduced, derived from processes with independent increments. We also study models of underlying processes that, when hitting a barrier, lead to the event in question. It is emphasized that frailty is an essential concept in event history analysis, and wrong conclusions may be drawn if frailty is ignored.

The applied aspect of the book is supported by a range of real-life data sets. Most of the data are well known to us through active collaboration with research groups in other fields. By necessity, some of the examples based on them have been simplified to achieve a more pedagogical exposition. However, we have made a strong effort to keep close to the relevant research questions; in particular, our three case studies are intended as more or less full-blown analyses of data of current interest.

Hence, the book contains a lot of new material for both the practicing statistician and those who are more interested in the theory of the subject. The book is intended primarily for researchers working in biostatistics, but it should also be found interesting by statisticians in other fields where event history analysis is of importance. The reader should have some theoretical background in statistics, although the mathematics is kept at a "user" level, not going into the finer theoretical details.

The book should be well suited as a textbook for graduate courses in survival and event history analysis, in particular for courses on the analysis of multivariate or complex event history data. Some exercises are provided at the end of each chapter, and supplementary exercises, as well as some of the datasets used as examples, may be found on the book's Web page at www.springer.com/978-0-387-20287-7.

A number of friends and colleagues have read and commented on parts of this book, have participated in discussion on themes from the book, or have provided us with data used as examples, and we want to thank them all: Per Kragh Andersen, Vanessa Didelez, Ludwig Fahrmeir, Johan Fosen, Axel Gandy, Jon Michael Gran, Nina Gunnes, Robin Henderson, Nils Lid Hjort, Niels Keiding, Bryan Langholz, Stein Atle Lie, Bo Lindqvist, Torben Martinussen, Sven Ove Samuelsen, Thomas Scheike, Finn Skårderud, Anders Skrondal, Halvor Sommerfelt, Hans Steinsland, Hans van Houwelingen, and Stein Emil Vollset.

Hege Marie Bøvelstad and Marion Haugen did an invaluable job making graphs for a number of the examples. They also fearlessly took on the daunting task of preparing all our references, saving us long hours of tedious work.

Preliminary versions of the book have been used as the text for a graduate course in survival and event history analysis at the University of Oslo and for a Nordic course organized as part of a Nordplus program for master-level courses in

biostatistics. We thank the students at these courses, who gave valuable feedback and pointed out a number of typos.

Most of the work on the book has been done as part of our regular appointments at the University of Oslo (Odd O. Aalen and Ørnulf Borgan) and the Norwegian Institute of Public Health (Håkon K. Gjessing), and we thank our employers for giving us time to work on the book as part of our regular duties. In addition, important parts of the work were done while the authors were participating in an international research group at the Centre for Advanced Study at the Norwegian Academy of Science and Letters in Oslo during the academic year 2005–2006. We express our gratitude to the staff of the center for providing such good working facilities and a pleasant social environment.

Oslo
January 2008

Odd O. Aalen
Ørnulf Borgan
Håkon K. Gjessing

Contents

Chapter 1
An introduction to survival and event history analysis

This book is about survival and event history analysis. This is a statistical methodology used in many different settings where one is interested in the occurrence of events. By events we mean occurrences in the lives of individuals that are of interest in scientific studies in medicine, demography, biology, sociology, econometrics, etc. Examples of such events are: death, myocardial infarction, falling in love, wedding, divorce, birth of a child, getting the first tooth, graduation from school, cancer diagnosis, falling asleep, and waking up. All of these may be subject to scientific interest where one tries to understand their cause or establish risk factors. In classical survival analysis one focuses on a single event for each individual, describing the occurrence of the event by means of survival curves and hazard rates and analyzing the dependence on covariates by means of regression models.

The connecting together of several events for an individual as they occur over time yields event histories. One might, for instance, be interested in studying how people pass through diseases. Clearly, the disease might have a number of stages. In parallel with the development, there will typically be a number of blood tests and other diagnostics, different treatment options may be attempted, and so on. The statistical analysis of such data, trying to understand how factors influence each other, is a great challenge.

Event histories are not restricted to humans. A sequence of events could also happen to animals, plants, cells, amalgam fillings, hip prostheses, light bulbs, cars – anything that changes, develops, or decays. Although a piece of technical equipment is very different from a human being, that does not prevent statistical methods for analyzing event history data to be useful, for example, in demography and medicine as well as in technical reliability. Motivated by our own research interests, the focus of this book is on applications of event history methodology to medicine and demography. But the methodology we present also should be of interest to researchers in biology, technical reliability, econometrics, sociology, etc., who want to apply survival and event history methods in their own fields.

Survival and event history analysis is used as a tool in many different settings, some of which are:

- Proving or disproving the value of medical treatments for diseases.

O.O. Aalen et al., *Survival and Event History Analysis: A Process Point of View*, Statistics for Biology and Health, DOI 10.1007/978-0-387-68560-1_1,

- Understanding risk factors, and thereby preventing diseases.
- Evaluating reliability of technical equipment.
- Understanding the mechanisms of biological phenomena.
- Monitoring social phenomena like divorce and unemployment.

Even though the purpose of a statistical analysis may vary from one situation to another, the ambitious aim of most statistical analyses is to help understand causality.

The purpose of this introductory chapter is twofold. Our main purpose is to introduce the reader to some basic concepts and ideas in survival and event history analysis. But we will also take the opportunity to indicate what lies ahead in the remaining chapters of the book. In Section 1.1 we first consider some aspects of classical survival analysis where the focus is on the time to a single event. Sometimes the event in question may occur more than once for an individual, or more than one type of event is of interest. In Section 1.2 such event history data are considered, and we discuss some methodological issues they give rise to, while we in Section 1.3 briefly discuss why survival analysis methods may be useful also for data that do not involve time. When events occur, a natural approach for a statistician would be to count them. In fact, counting processes, a special kind of stochastic process, play a major role in this book, and in Section 1.4 we provide a brief introduction to counting processes and their associated intensity processes and martingales. Finally, in Section 1.5, we give an overview of some modeling issues for event history data.

The counting process point of view adopted in this book is the most natural way to look at many issues in survival and event history analysis. Therefore the mathematical tools of counting processes and martingales should not be considered pure technicalities, but the reader should make an effort to understand the concepts and ideas they express. Pure technicalities, like regularity assumptions, will not be emphasized much in this book; our aim is to explain the concepts involved.

1.1 Survival analysis: basic concepts and examples

We shall start by considering classical survival analysis which focuses on the time to a single event for each individual, or more precisely the time elapsed from an initiating event to an event, or endpoint, of interest. Some examples:

- Time from birth to death.
- Time from birth to cancer diagnosis.
- Time from disease onset to death.
- Time from entry to a study to relapse.
- Time from marriage to divorce.

As a generic term, the time from the initiating event to the event of interest will be denoted a *survival time*, even when the endpoint is something different from death.

1.1.1 What makes survival special: censoring and truncation

Superficially, one might think that a survival time is just a measurement on a scale, and that an analysis of the survival times for a sample of individuals could be handled by the well-developed statistical methods for analyzing continuous, or possibly discrete, data. So, why not use ordinary linear regression and other standard statistical methods? The reason this will not usually work is a fundamental problem that one almost always meets when studying survival times. The point is that one actually has to wait for the event to happen, and when the study ends and the analysis begins, one will typically find that the event in question has occurred for some individuals but not for others. Some men had a testicular cancer during the period of observation, but most, luckily, had no such cancer. Those people might develop it later, but that is something we will not have knowledge about. Or, in a study of divorce, some couples were divorced during the time of the study, while others were not. Again, they may divorce later, but that is unknown to us when we want to analyze the data. Hence, the data come as a mixture of complete and incomplete observations. This constitutes a big difference compared to most other statistical data. If a doctor measures your blood pressure, it is done in one sitting in a couple of minutes. Most measurements are like that: they are made on a single occasion. But measuring survival times is altogether a different story, and this is what requires a different statistical theory.

The incomplete observations are termed *censored* survival times. An illustration of how censored survival times may arise is given in Figure 1.1. The figure illustrates a hypothetical clinical study where 10 patients are observed over a time period to see whether some specific event occurs. This event, or endpoint, could be death, remission of disease, relapse of disease, etc. The left panel shows the observations as they occur in calendar time. The patients enter the study at different times and are then followed until the event occurs or until the closure of the study after 10 years. For statistical analysis one often focuses on the time from entry to the event of interest. Each individual will then have his or her own starting point, with time zero being the time of entrance into the study. The right panel of Figure 1.1 shows the observations in this *study time* scale. For patients number 1, 2, 4, 5, 6, and 9 the event was observed within the period of the study, and we have complete observation of their survival times. Patients 3, 7, 8, and 10 had not yet experienced the event when the study closed, so their survival times are censored. Notice that the censoring present in these data cut off intervals on the right-hand side, and so one talks about *right-censoring*. In the simple example of Figure 1.1, closure of the study was the only reason for censoring. However, in real-life clinical studies, right-censored observations will also occur when an individual withdraws from the study or is lost to follow-up.

As indicated in Figure 1.1, the study time scale used in a statistical analysis will usually not be calendar time. There are commonly many possible different time scales that may be used. In a clinical study the initiating event, which corresponds to time zero in the study time scale, could be time of diagnosis, time of entry into the study, time of admission to hospital, time of remission, etc. The choice of time

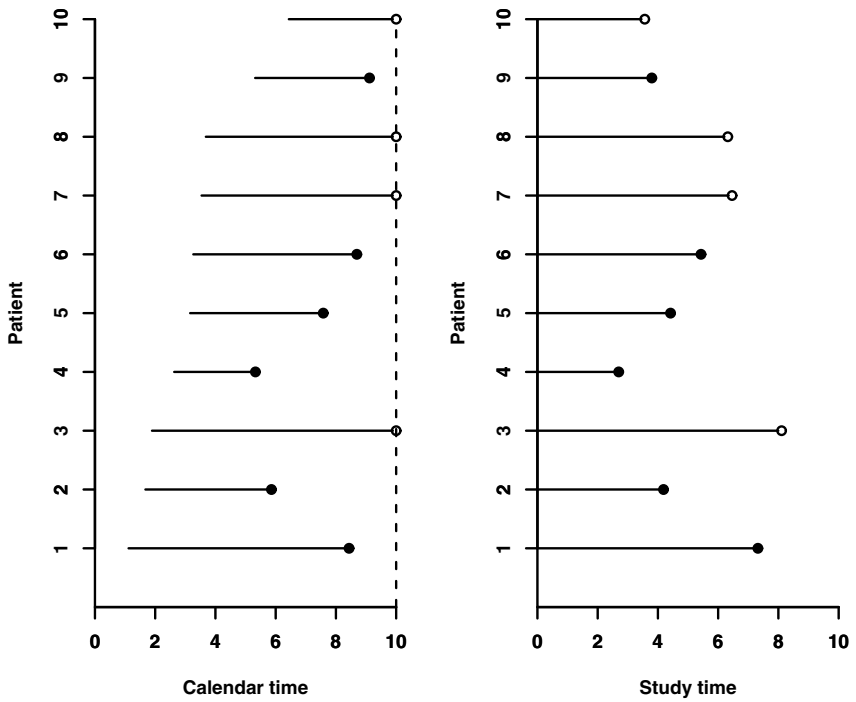

Fig. 1.1 *Patient follow up in a hypothetical clinical study with 10 patients. Left panel shows the actual calendar time. Right panel shows the same observations in the study time scale, where time 0 for each individual is his or her entry into the study. A filled circle indicates occurrence of the event, while an open circle indicates censoring. In the left panel the dotted vertical line indicates the closing date of the study.*

scale to use in a statistical analysis is usually a pragmatic one: what will make the analysis most relevant and clear with respect to the issues we study.

The set of individuals for which the event of interest has not happened before a given time t (in the chosen study time scale), and who have not been censored before time t, is termed the *risk set* at time t. In the right-hand panel of Figure 1.1 the risk set starts with 10 individuals, and then gradually declines to one and finally zero individuals.

A concept related to right-censoring is that of *left-truncation*. In a clinical study it may be that the patients come under observation some time after the initiating event. For instance, in a study of myocardial infarction only those who survive the initial phase and reach the hospital will be included in the study. Time zero for an individual patient may be the time of infarction, and if the patient reaches the hospital and is included in the study it may happen at different times for different patients. If

the patient dies prior to reaching the hospital, he or she will never be entered in the study. The data arising here are left-truncated, and one may also use the term *delayed entry* about such data. This implies that the risk set will not only decline over time, but will also increase when new individuals enter the study. Originally, many methods of survival analysis were developed only for right-censored survival data. But the counting process approach, which we shall focus on in this book, clearly shows that most methods are equally valid for data with a combination of left-truncation and right-censoring.

As a simple example of survival data with right-censoring, we may consider the data on the 10 patients from the hypothetical clinical study of Figure 1.1. The data from the right-hand panel of the figure are, with an asterisk indicating a right-censored survival time,

$$7.32,\ 4.19,\ 8.11^*,\ 2.70,\ 4.42,\ 5.43,\ 6.46^*,\ 6.32^*,\ 3.80,\ 3.50^*.$$

This example illustrates well that right-censored survival times cannot be handled by ordinary statistical methods. In fact, even a simple mean cannot be calculated due to the censoring. If we cannot calculate the mean, then we cannot find a standard deviation or perform a t-test, a regression analysis, or almost anything else.

1.1.2 Survival function and hazard rate

Even though ordinary statistical methods cannot handle right-censored (and left-truncated) survival data, it is in fact quite simple to analyze such data. What one needs is the right concepts; and there are two basic ones that pervade the whole theory of survival analysis, namely the *survival function* and the *hazard rate*. One starts with a set of individuals at time zero and waits for an event that might happen. The survival function, $S(t)$, which one would usually like to plot as a *survival curve*, gives the expected proportion of individuals for which the event has not yet happened by time t. If the random variable T denotes the survival time, one may write more formally:

$$S(t) = P(T > t). \tag{1.1}$$

Remember that we use the term survival time in a quite general sense, and the same applies to the term survival function. Thus the survival function gives the probability that the event of interest has not happened by time t (in the study time scale), and it does not have to relate to the study of death. Often the survival function will tend to zero as t increases because over time more and more individuals will experience the event of interest. However, since we use the terms survival time and survival function also for events that do not necessarily happen to all individuals, like divorce or getting testicular cancer, the random variable T may be infinite. For such situations, the survival function $S(t)$ will decrease toward a positive value as t goes to infinity.

The survival function (1.1) specifies the unconditional probability that the event of interest has not happened by time t. The hazard rate $\alpha(t)$, on the other hand, is

defined by means of a conditional probability. Assuming that T is absolutely continuous, that is, that it has a probability density, one looks at those individuals who have not yet experienced the event of interest by time t and considers the probability of experiencing the event in the small time interval $[t, t+dt)$. Then this probability equals $\alpha(t)dt$. To be more precise, the hazard rate is defined as a limit in the following way:

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leqslant T < t + \Delta t \mid T \geq t). \tag{1.2}$$

Notice that while the survival curve is a function that starts at 1 and declines over time, the hazard rate can be essentially any nonnegative function. Note also that the hazard rate is the model counterpart of the incidence rate that is commonly computed in epidemiological studies.

The concepts are illustrated in Figure 1.2. The left-hand panels show two typical hazard rates, one that reaches a maximum and then declines and one that increases all the time. The right-hand panels show the corresponding survival curves. Notice that it is not so easy to see from the survival curves that the hazard rates are actually very different. In fact, the shape of the hazard rate is an issue we will return to several times in the book. The hazard rate may seem like a simple concept, but in fact it is quite elusive and, as we will see in Chapters 6, 10, and 11, it hides a lot of complexities.

From censored survival data, we can, as described in Section 3.2, easily estimate a survival curve by the *Kaplan-Meier estimator*. The estimation of a hazard rate is more tricky. But, as explained in Section 3.1, what can easily be done is to estimate the cumulative hazard rate

$$A(t) = \int_0^t \alpha(s)ds \tag{1.3}$$

by the *Nelson-Aalen estimator*. The increments of a Nelson-Aalen estimate may then be smoothed to provide an estimate of the hazard rate itself.

There are two fundamental mathematical connections between the survival function and the (cumulative) hazard rate that should be mentioned here. First note that by (1.2) and (1.3), we have

$$A'(t) = \alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{S(t) - S(t+\Delta t)}{S(t)} = -\frac{S'(t)}{S(t)}. \tag{1.4}$$

Then by integration, using that $S(0) = 1$, one gets

$$-\log\{S(t)\} = \int_0^t \alpha(s)ds,$$

and it follows that

$$S(t) = \exp\left\{-\int_0^t \alpha(s)ds\right\}. \tag{1.5}$$

In Appendix A.1 we describe how we may define the cumulative hazard rate for arbitrary survival times, which need be neither absolutely continuous nor discrete,

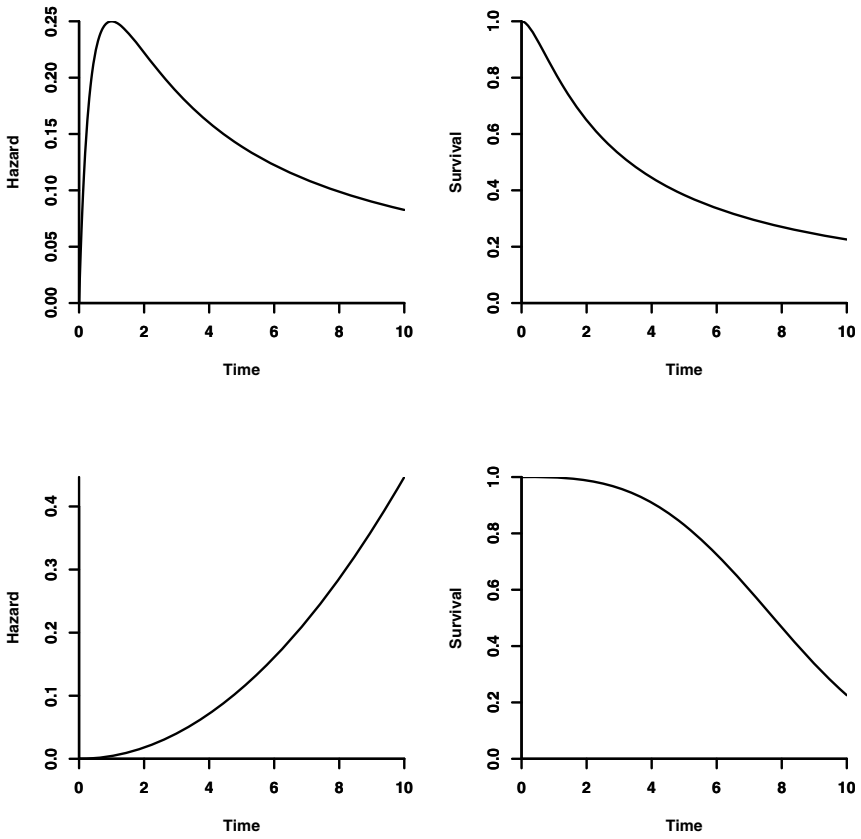

Fig. 1.2 *Illustrating hazard rates and survival curves. The hazard rates on the left correspond to the survival curves on the right.*

and we show how formula (1.5) is generalized to this situation. These results turn out to be useful when we study the properties of the Kaplan-Meier estimator and its relation to the Nelson-Aalen estimator; cf. Section 3.2.

1.1.3 Regression and frailty models

A main purpose of many studies is to assess the effect of one or more *covariates*, or explanatory variables, on survival. If there is only one categorical covariate, like gender or the stage of a cancer patient, the effect of the covariate may be assessed by

estimating a survival curve for each level of the covariate and testing whether any observed differences are significant using the tests of Section 3.3. However, in most studies there will be a number of covariates, and some of them will be numeric. Then, as in other parts of statistics, regression modeling is called for. A number of regression models have been suggested for censored survival data, and we will consider some of them in Chapters 4 and 5.

The most used regression model for censored survival data is Cox's regression model. For this model it is assumed that the hazard rate of an individual with covariates $x_1, \ldots, x_p$ takes the form

$$\alpha(t|x_1, \ldots, x_p) = \alpha_0(t) \exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}. \tag{1.6}$$

Here $\alpha_0(t)$ is a *baseline hazard* that describes the shape of the hazard rate as a function of time, while $\exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}$ is a *hazard ratio* (sometimes denoted a *relative risk*) that describes how the size of the hazard rate depends on covariates. Note that (1.6) implies that the hazard rates of two individuals are proportional (Exercise 1.5).

An alternative to Cox's model is the additive regression model due to Aalen, which assumes that the hazard rate of an individual with covariates $x_1, \ldots, x_p$ takes the form

$$\alpha(t|x_1, \ldots, x_p) = \beta_0(t) + \beta_1(t)x_1 + \cdots + \beta_p(t)x_p. \tag{1.7}$$

For this model $\beta_0(t)$ is the baseline hazard, while the *regression functions* $\beta_j(t)$ describe how the covariates affect the hazard rate at time t.

In (1.6) and (1.7) the covariates are assumed to be fixed over time. More generally, one may consider covariates that vary over time. In Sections 4.1 and 4.2 we will discuss more closely Cox's regression model and the additive regression model, including their extensions to time-varying covariates.

The regression models (1.6) and (1.7) may be used to model differences among individuals that may be ascribed to *observable* covariates. Sometimes one may want to model *unobservable* heterogeneity between individuals that may be due, for example, to unobserved genetic or environmental factors. One way of doing this is to assume that each individual has a *frailty* Z. Some individuals will have a high value of Z, meaning a high frailty, while others will have a low frailty. Then, conditional on the frailty, the hazard rate of an individual is assumed to take the form

$$\alpha(t|Z) = Z \cdot \alpha(t). \tag{1.8}$$

In Chapter 6 we will take a closer look at such proportional frailty models where each individual has its own unobserved frailty. Frailty models may also be used to model the dependence of survival times for individuals within a family (or some other cluster). Then, as discussed in Chapter 7, two or more individuals will share the same frailty or their frailties will be correlated.

1.1.4 The past

Another concept we shall introduce is that of a *past*. The definition of the hazard rate in (1.2) conditions with respect to the event of interest not having occurred *before* time t. It therefore makes an assumption about how information on the past influences the present and the future. This is a main reason why the hazard rate is a natural concept for analyzing events occurring in time. So even if we did not have the complication of censoring, the hazard rate would be a concept of interest. The idea of defining a past, and hence also a present and a future, is a fundamental aspect of the methods that shall be discussed in this book.

In (1.2) the only information we use on the past is that the event has not occurred before time t. When working with left-truncated and right-censored survival data, the information we use on the past will be more involved and will include information on all events, delayed entries, and censorings that occur *before* time t (in the study time scale), as well as information on the time-fixed covariates. (The handling of time-varying covariates requires some care; cf. the introduction to Chapter 4.) When working with event history data, where more than one event may occur for an individual, the information on the past becomes even more involved.

It turns out that it is extremely fruitful to introduce some ideas from the theory of stochastic processes, since parts of this theory are designed to precisely understand how the past influences the future development of a process. In particular counting processes and martingales turn out to be useful. We will give a brief introduction to these types of stochastic processes in Section 1.4, while a more thorough (although still fairly informal) treatment is given in Chapter 2.

As introduced in Section 1.1.1, right-censoring is a rather vague notion of lifetimes being incompletely observed. It is necessary to be more precise about the censoring idea to get mathematically valid results, and we will see in Sections 1.4 and 2.2.8 that the concept of a past and the use of counting processes help to provide a precise treatment of the censoring concept.

1.1.5 Some illustrative examples

The concepts and methods mentioned earlier play an increasingly important role in a number of applications, as the following examples indicate. Most of these examples, with associated data sets, will be used for illustrative purposes throughout the book. In connection with the examples, we will take the opportunity to discuss general problems and concepts that are the topics of later chapters in the book. We start with examples that are simple, seen from a methodological and conceptual point of view, and then move on to examples that require more complex concepts and methods.

Example 1.1. Time between births. The Medical Birth Registry of Norway was established in 1967; it contains information on all births in Norway since that time.

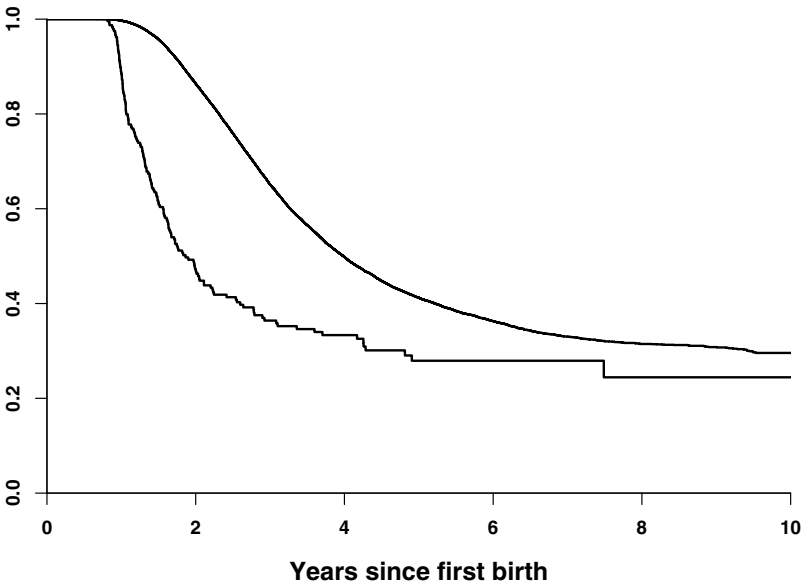

Fig. 1.3 *Empirical survival curves of the time between first and second birth. Upper curve: First child survived one year. Lower curve: First child died within one year. (Based on data from the Medical Birth Registry of Norway.)*

The registry is an invaluable source for epidemiological research of perinatal health problems as well as on demographic research related to fertility. Focusing on the latter, we will use information from the Birth Registry on the time between births for a woman. In particular, we will study the time between the first and second births of a woman (by the same father), and how this is affected if the first child dies within one year of its birth. Figure 1.3 shows Kaplan-Meier survival curves (cf. Section 3.2.1) for the time to the second birth for the 53 296 women for whom the first child survived one year and for the 262 women who lost their first child within one year of its birth. From the survival curves we see, for example, that it takes less than two years before 50% of the women who lost their first child will have another one, while it takes about four years before this is the case for the women who do not experience this traumatic event. We note that the survival curves give a very clear picture of the differences between the two group of women.

A more detailed description of the data on time between first and second births is given in Example 3.1. In addition to this example, the data will be used for illustration in Examples 3.8, 3.10, and 10.2. We will use data from the Birth Registry also to study the time between the second and third births of a woman and how this is affected by the genders of the two older children; cf. Examples 3.2, 3.9, and 3.14. □

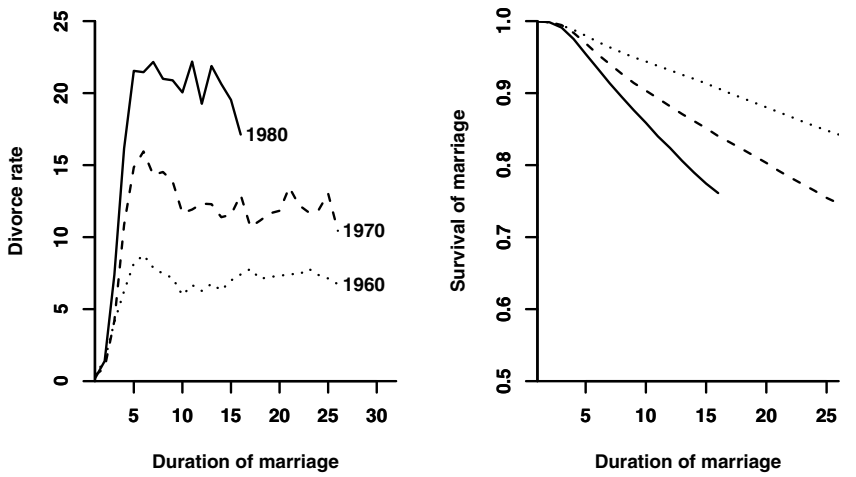

Fig. 1.4 *Rates of divorce per 1000 marriages per year (left panel) and empirical survival curves (right panel) for marriages contracted in 1960, 1970, and 1980. (Based on data from Statistics Norway.)*

Example 1.2. Divorce in Norway. The increase in divorce rates is a general feature of Western societies. The phenomenon is illustrated well by using the concepts of hazard rates and survival curves. In Figure 1.4 we show empirical hazard rates (rates of divorce) and empirical survival curves for marriages contracted in Norway in 1960, 1970, and 1980. The increase in divorce risk with marriage cohort is clearly seen. Furthermore, the hazard rates show an expected increase with duration of marriage until about five years where a slight decline and then leveling out occurs. The survival curves show how the proportions still married are decreasing to various degrees in the different cohorts. The concepts of survival curve and hazard rate are very suitable for describing the phenomenon, as opposed to the rather uninformative descriptions common in the popular press, comparing, for instance, divorces in a given year with the number of marriages contracted in the same year.

The rates of divorce and survival curves of Figure 1.4 are computed from data from Statistics Norway. The data are given in Example 5.4, where we also explain how the divorce rates and survival curves are computed. The divorce data are also used for illustrative purposes in Examples 11.1 and 11.2. □

Example 1.3. Myocardial infarction. Clinical medicine is probably the largest single area of application of the traditional methods of survival analysis. Duration is an important clinical parameter: to the severely ill patient it is of overriding importance how long he may expect to live or to stay in a relatively good state. An example

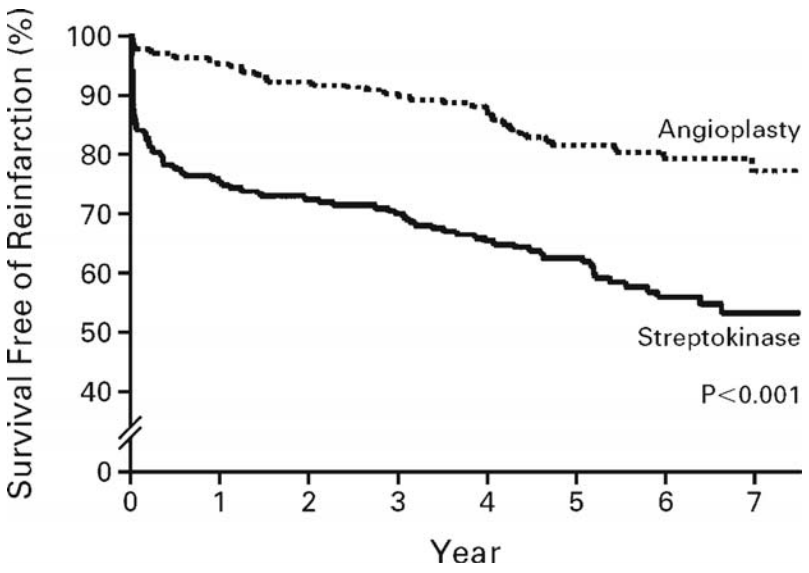

Fig. 1.5 *Survival curves for two treatments of myocardial infarction. Figure reproduced with permission from Zijlstra et al. (1999). Copyright The New England Journal of Medicine.*

is shown in Figure 1.5 comparing reinfarction-free survival of patients with myocardial infarction when randomized to treatment with angioplasty or streptokinase (Zijlstra et al., 1999). The latter treatment is a medication, while angioplasty consists of inflating a tiny balloon in the blood vessel to restore the passage of blood. The plots in the figure certainly give a clear conclusion: many lives would be spared if all patients were given angioplasty.

A closer look tells us that the main difference between the groups comes very early, in the first few weeks. A contentious issue in survival analysis is how to describe in the most fruitful way how survival depends on treatment and other covariates. Most commonly Cox's regression model (1.6) is used. This was done in the paper by Zijlstra et al. (1999); adjusting for a number of factors the hazard ratio of dying in the streptokinase group versus the angioplasty group was estimated to 2.31. Cox's regression model assumes proportionality of hazards over time, and when this assumption fails, the estimated hazard ratio will be an average measure that does not consider the change in effect over time. If one considered a shorter time interval the hazard ratio might be much larger. One focus in the current book is how to analyze changes in effects over time, and it will be shown that ideas from frailty theory (cf. Chapter 6) are essential to understand changes in hazard rates and hazard ratios. □

Example 1.4. Carcinoma of the oropharynx. Another example from clinical medicine is given by Kalbfleisch and Prentice (2002, section 1.1.2 and appendix A).

They present the results of a clinical trial on 195 patients with carcinoma of the oropharynx carried out by the Radiation Therapy Oncology Group in the United States. The patients were randomized into two treatment groups ("standard" and "test" treatment), and survival times were measured in days from diagnosis. A number of covariates were recorded for each patient at the entry to the study:

- x_1 = sex (1 = male, 2 = female),
- x_2 = treatment group (1 = standard, 2 = test),
- x_3 = grade (1 = well differentiated, 2 = moderately differentiated, 3 = poorly differentiated),
- x_4 = age in years at diagnosis,
- x_5 = condition (1 = no disability, 2 = restricted work, 3 = requires assistance with self-care, 4 = confined to bed),
- x_6 = T-stage (an index of size and infiltration of tumor ranging from 1 to 4, with 1 indicating a small tumor and 4 a massive invasive tumor),
- x_7 = N-stage (an index of lymph node metastasis ranging from 0 to 3, with 0 indicating no evidence of metastases and 3 indicating multiple positive nodes or fixed positive nodes).

The main purpose of such a clinical trial is to assess the effect of treatment. In addition, one would like to study the effects of the other covariates on survival. As these covariates are measured at entry and do not change over time, they are *fixed* covariates.

We will use the oropharynx data for illustration in Examples 4.6, 4.7, 4.8, 4.10, 4.13, and 10.1. In all these examples, we will exclude the two patients (numbers 136 and 159) who are missing information on some of the covariates. □

Example 1.5. Hip replacements. The Norwegian Arthroplasty Registry was started in 1987. At first only total hip replacements were registered, but from 1994 it includes replacements of other joints. By the end of 2005 the registry included information on more than 100 000 total hip replacements and about 30 000 replacements of other joints. The Arthroplasty Registry has provided valuable information on the epidemiology of joint replacements, in particular total hip replacements.

From September 1987 to February 1998 almost 40 000 patients had their first total hip replacement operation at a Norwegian hospital (Lie et al., 2000). In this book we will use a random sample of 5000 of these patients – 3503 females and 1497 males – to study the survival of patients who have had a total hip replacement operation. In particular we will compare the mortality among these patients to the mortality of the general Norwegian population. Figure 1.6 shows the Kaplan-Meier survival curve for the hip replacement patients (as a function of time since operation) together with the survival curve one would expect to get for a group of individuals from the general population with the same age and sex distribution as the hip replacement patients (see Section 3.2.5 for details). It is seen that the hip replacement patients have a slightly better survival than a comparable group from the general population. The main reason for this is that a patient needs to be in a fairly good health condition to be eligible for a hip replacement operation.