Statistics for Biology and Health

Jianguo Sun

The Statistical Analysis of Interval-censored Failure Time Data

Springer

Jianguo Sun
Department of Statistics
University of Missouri
Columbia, MO 65211
USA
sunj@missouri.edu

Library of Congress Control Number: 2006922771

ISBN-10: 0-387-32905-6
ISBN-13: 978-0387-32905-5

Printed on acid-free paper.

Printed in the United States of America. (MP/POW)

9 8 7 6 5 4 3 2 1

springeronline.com

To Xianghuan, Ryan, and Nicholas

Preface

Interval censoring is a type of censoring that has become increasingly common in the areas that produce failure time data. In the past 20 years or so, a voluminous literature on the statistical analysis of interval-censored failure time data has appeared. The main purpose of this book is to collect and unify some statistical models and methods that have been proposed for analyzing failure time data in the presence of interval censoring.

A number of books have been written that provide excellent and comprehensive coverage of the statistical analysis of failure time data in the presence of right censoring. These include Cox and Oakes (1984), Fleming and Harrington (1991), Andersen et al. (1993), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003), and Lawless (2003). In general, right-censored failure time data can be treated as a special case of interval-censored data, and some of the inference approaches for right-censored data can be directly, or with minor modifications, applied to the analysis of interval-censored data. However, most of the inference approaches for right-censored data are not appropriate for interval-censored data due to the fundamental differences between right censoring and interval censoring. The censoring mechanism behind interval censoring is much more complicated than that behind right censoring. For right-censored failure time data, substantial advances in the theory and development of modern statistical methods are due to the theory of counting processes. Because of the complexity and special structure of interval censoring, the same theory is not applicable to interval-censored data. The goal of this book is to complement the literature on right-censored data by presenting statistical models and methods specifically developed for interval-censored failure time data.

This book is intended to provide an up-to-date reference for those who are conducting research on the analysis of interval-censored failure time data as well as those who need to analyze interval-censored data to answer substantive questions. It can also be used as a text for a graduate course in statistics or biostatistics that has basic knowledge of probability and statistics as a pre-

requisite. The main focus is on methodology, and applications of the methods that are based on real data are given along with numerical calculations.

To keep the book at a reasonable length, some topics are discussed only briefly at the end of each chapter in the Bibliography, Discussion, and Remarks section or in the last chapter. Also, although some asymptotic results are discussed, their technical derivations are not presented. Because the literature on interval-censored data is extensive, the choice of subject matter is difficult. The material discussed in detail is to some extent a reflection of the author's interests in this field. However, our attempt has been to present a relatively complete and comprehensive coverage of the fundamental concepts along with selected topics in the field.

Chapter 1 contains introductory material and surveys basic concepts and regression models for the analysis of failure time data. Examples of right- and interval-censored survival data are discussed, and several types of interval censoring commonly seen in practice are described. Before considering the nonparametric and semiparametric approaches, which are the focus of the book, some parametric models and methods are presented in Chapter 2. Also, in Chapter 2, some imputation approaches are briefly investigated for the analysis of interval-censored failure time data.

Chapters 3 to 10 concern nonparametric and semiparametric approaches for interval-censored data. Chapter 3 considers statistical procedures for nonparametric estimation of survival and hazard functions, and Chapter 4 deals with nonparametric comparisons of survival functions. Both rank-based and survival-based procedures are investigated. Regression analysis of current status data, or case I interval-censored data, is discussed in Chapter 5, and Chapter 6 considers regression analysis of general, or case II interval-censored failure time data. The analysis of bivariate interval-censored failure time data is the subject of Chapter 7, which considers both nonparametric and semiparametric approaches. Chapter 8 deals with doubly censored failure time data. In this situation, the survival time of interest is the duration between two related events and the observations on the occurrences of both events could be right- or interval-censored. The analysis of event history data in the presence of interval censoring, which are commonly referred to as panel count data, is considered in Chapter 9. Chapter 10 contains brief discussions of several other important topics in the field for which it is not feasible to give a detailed discussion. These include regression diagnostics, regression analysis with interval-censored covariates, Bayesian inference approaches, and informative interval censoring.

In all chapters except Chapter 10, we have used references sparsely except in the last section of each chapter, which provides bibliographical notes including related references.

Many persons have contributed directly and indirectly to this book. First, I want to thank Diane Finkelstein, Jian Huang, Linxiong Li, Liuquan Sun, Tim Wright, and Ying Zhang for their many critical comments and suggestions. I am especially indebted to Tim Wright, who patiently read all the

chapters and made numerous corrections in an early draft of the book. I owe my thanks to Do-Hwan Park, Xingwei Tong, Lianming Wang, Zhigang Zhang, Qiang Zhao, and Chao Zhu, who not only read parts of the draft and gave their important comments but also provided great computational help. Also, I would like to express my thanks to Nancy Flourney, our department chair, for her encouragement and support during this period, and Jack Kalbfleisch, Steve Lagakos, Jerry Lawless, and LJ Wei for their important influence on my academic life and their guidance in the early years of my research.

Finally, I thank my family and especially my wife, Xianghuan, for her patience and support during this project.

January 2006 *Jianguo Sun*

Contents

1

Introduction

1.1 Failure Time Data

By failure time data, we mean data that concern positive random variables representing times to certain events. Examples of the event, often referred to as the failure or survival event, include death, the onset of a disease or certain milestone, the failure of a mechanical component of a machine, or learning something. The occurrence of the event is usually referred to as a *failure*. Sometimes we also use the terminology survival data and refer to the variable of interest as survival time or the survival variable. Failure time data arise extensively in medical studies, but there are many other investigations that also produce failure time data. These include biological studies, demographical studies, economic and financial studies, epidemiological studies, psychological experiments, reliability experiments, and sociological studies.

The analysis of failure time data usually means addressing one of three problems. They are estimation of survival functions, comparison of treatments or survival functions, and assessment of covariate effects or the dependence of failure time on explanatory variables. We consider methods that can be used to deal with these problems for interval-censored data. A survival function, which is formally defined below, gives the probability that failure time is greater than a certain time and is of considerable interest in failure time analysis.

For a number of reasons, special methods are required to treat failure time data. One reason, which also is a major feature that distinguishes the analysis of failure time data from other statistical fields, is the existence of censoring, such as right censoring, which is discussed below. Censoring mechanisms can be quite complicated and thus necessitate special methods of treatment. The methods available for other types of data are usually simply not appropriate for censored data. Truncation is another feature of some failure time data that requires special treatments. We focus mainly on censoring and discuss only some special types of truncation. Before discussing censoring and truncation

Table 1.1. Remission times in weeks for acute leukemia patients

Group	Survival times in weeks
6-MP	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25* 32*, 32*, 34*, 35*
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

in more detail, we describe two examples to introduce failure time data and their features.

1.1.1 Remission Times of Acute Leukemia Patients

Table 1.1, reproduced from Freireich et al. (1963) and Gehan (1965), presents a typical set of failure time data arising from a clinical trial on acute leukemia patients. In the table, remission times in weeks are given for 42 patients in two treatment groups. One treatment is the drug 6-mercaptopurine (6-MP) and the other is the placebo treatment. The study was performed over a one-year period and the patients were enrolled into the study at different times. A primary concern is the comparison of the two treatments with respect to ability to maintain remission. In other words, it is of interest to know if the patients with drug 6-MP had significantly longer remission times than those given the placebo treatment.

For the observed information given in Table 1.1, the starred numbers are censoring times or censored remission times. That is, such an observation is the amount of time from when the patient entered the study to the end of the study. These remission times were censored because these patients were still in the state of remission at the end of the trial and thus their remission times were known only to be greater than the censoring times. For the other patients, their remission times were observed exactly. This situation commonly occurs in failure time studies, and the resulting data are usually referred to as right-censored failure time data. Note that for the comparison of the two treatments, a simple t-test is not applicable because it cannot handle the censored remission times, and certainly discarding these times is not desirable. For more discussion and the analysis of this data set, readers are referred to Kalbfleisch and Prentice (2002) in addition to Freireich et al. (1963) and Gehan (1965).

1.1.2 Times to the First Use of Marijuana

Turnbull and Weiss (1978) discussed a set of failure time data from a study on the use of marijuana by high school students, and the data are given in Table 1.2. In the study, 191 California high school boys were asked the question, "when did you first use marijuana?" As expected, some boys remembered the exact age when they first used it, and some boys used it but could not

Table 1.2. Ages in years to the first use of marijuana

Age	No. of exact observations	No. of left-censored observations	No. of right-censored observations
10	4	0	0
11	12	0	0
12	19	2	0
13	24	15	1
14	20	24	2
15	13	18	3
16	3	14	2
17	1	6	3
18	0	0	1
>18	4	0	0

remember when they first used marijuana. Also there were boys who never used it.

Corresponding with these three situations, there are three types of observations about the age when marijuana was first used. For the first situation, the age is known exactly. For the second and third situations, the age is known only to be smaller or greater than the current age of the boy, and these types of observations are usually referred to as left-censored or right-censored observations, respectively. For the data set, one question of interest is to estimate the probability of having used marijuana at certain ages for high school boys. It is apparent that the simple empirical estimate is not appropriate unless one disregards some of the left- and right-censored observations. Among others, Klein and Moeschberger (2003) and Turnbull and Weiss (1978) analyzed this data set.

1.1.3 Censoring and Truncation

As mentioned above, censoring is one of the unique features of failure time data. By censoring, we mean that an observation on a survival time of interest is incomplete, that is, the survival time is observed only to fall into a certain range instead of being known exactly. Note that censored data are different from missing data as censored observations still provide some partial information, whereas missing observations provide no information about the variable of interest. Different types of censoring arise in practice, but the one that receives most of the attention in the literature is right censoring.

By right censoring or right-censored failure time data, we mean that the failure time of interest is observed either exactly or to be greater than a censoring time. A typical situation that yields right-censored observations is one in which a survival study has to end due to, for example, time constraints or resource limitations. In this case, for subjects whose survival events have not occurred at the end of the study, their survival times are not observed exactly

but are known to be greater than the study end time, i.e., they are right-censored. For subjects who have already failed by the end of the study, their failure times are known exactly. Of course, the study end time could be different for different subjects, and some subjects may withdraw from the study before the end for some reasons. In a more general setting, which is appropriate in many applications, for each subject, there exists a censoring variable representing the right censoring time. If the survival variable is smaller than the censoring variable, the observation is exact and otherwise, it is right-censored. This is usually referred to as the random censorship model.

It is apparent that in general, one has to understand the way that right censoring occurs to analyze right-censored failure time data properly. To simplify the analysis, an independent right censoring mechanism is commonly assumed. By this, we mean that the failure rate or hazard is the same for the subjects who are still in the study and the subjects who have been censored out. More specifically, under independent right censoring, we have that

$$\lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t, Y(t) = 1)}{\Delta t}$$

(Kalbfleisch and Prentice, 2002), where T denotes the survival variable of interest, and $Y(t) = 1$ means that the subject has neither failed nor been censored prior to time t. Under the random censorship model, the above condition is equivalent to

$$\lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t, C \ge t)}{\Delta t},$$

where C denotes the censoring variable.

There exist different types of right censoring as well as other types of censoring. For example, the censoring mechanism that stops the study at the same fixed time point for all subjects is usually referred to as Type 1 censoring. Type 2 censoring means that the study stops if a prespecified number of individuals out of all study individuals have failed. In addition to right censoring, some observations may be left-censored, meaning that the failure time is known only to be less than certain time. Interval censoring, the focus of this book, is introduced in the next section.

Truncation refers to situations where a subject is included in a study only if the corresponding failure time satisfies certain conditions. A simple and common example that yields truncated failure time data is a cohort study in which subjects are included in the study only if they experience some initial event prior to the survival event. In this case, for all subjects in the study, their failure times are greater than the occurrence times of the initial event. This type of truncation is commonly referred to as left-truncation. Independent truncation can be defined similarly to independent right censoring and is usually assumed for the analysis of truncated failure time data. For a more detailed discussion of right censoring and truncation, among others,

see Kalbfleisch and Prentice (2002) and Lawless (2003). They give various statistical methods for the analysis of right-censored failure time data such as those discussed in Section 1.1.1.

1.2 Failure Time Data with Interval Censoring

As discussed in the previous section, failure time data occur in many ways and in many fields, and there are a number of reasons why special methods are needed for their analyses. One key feature of failure time data is censoring, and there exist many excellent books on right censoring. Here we focus on interval censoring, which is more challenging than right censoring, and for such data the methods developed for right censoring do not generally apply.

By interval censoring, we mean that study subjects or failure time processes of interest are not under continuous observation. As a consequence, the failure or survival time is not always exactly observed or right-censored. For an interval-censored observation, one only knows a window, that is, an interval, within which the survival event has occurred. Exact or right-censored failure times can be regarded a special case of interval-censored failure times as in such cases, the interval reduces to a single point or is unbounded on the right. More generally, one could define an interval-censored observation as a union of several nonoverlapping windows or intervals (Turnbull, 1976).

Interval-censored failure time data occur in many areas including demographical, epidemiological, financial, medical, and sociological studies. A typical example of interval-censored data occurs in medical or health studies that entail periodic follow-ups, and many clinical trials and longitudinal studies fall into this category. In such situations, interval-censored data may arise in several ways. For instance, an individual may miss one or more observation times that have been scheduled to clinically observe possible changes in disease status and then return with a changed status. Alternatively, individuals may visit clinical centers at times that are convenient to them rather than at predetermined observation times. In both situations, the data on change in status are interval-censored. Even if all study subjects follow exactly the predetermined observation schedule, one still cannot observe the exact time of the occurrence of the change of the status assuming that it is a continuous variable. In the last situation, one has grouped failure time data, that is, interval-censored data for which the observation for each subject is a member of a collection of nonoverlapping intervals. Grouped failure time data can be dealt with relatively easily. Among others, Lawless (2003) discussed this type of failure time data. In the following, we focus on interval-censored data that are not grouped failure time data.

We present several examples below to further illustrate some of the general concepts, definitions, common features, and the structure of interval-censored data. The first two examples concern univariate failure time variables representing the time from the beginning of a study to the occurrence of an event of

Table 1.3. Death times in days for 144 male RFM mice with lung tumors

Group	Tumor status	Death times
CE	With tumor	381, 477, 485, 515, 539, 563, 565, 582, 603, 616, 624, 650 651, 656, 659, 672, 679, 698, 702, 709, 723, 731, 775, 779 795, 811, 839
	No tumor	45, 198, 215, 217, 257, 262, 266, 371, 431, 447, 454, 459 475, 479, 484, 500, 502, 503, 505, 508, 516, 531, 541, 553 556, 570, 572, 575, 577, 585, 588, 594, 600, 601, 608, 614 616, 632, 632, 638, 642, 642, 642, 644, 644, 647, 647, 653 659, 660, 662, 663, 667, 667, 673, 673, 677, 689, 693, 718 720, 721, 728, 760, 762, 773, 777, 815, 886
GE	With tumor	546, 609, 692, 692, 710, 752, 773, 781, 782, 789, 808, 810 814, 842, 846, 851, 871, 873, 876, 888, 888, 890, 894, 896 911, 913, 914, 914, 916, 921, 921, 926, 936, 945, 1008
	No tumor	412, 524, 647, 648, 695, 785, 814, 817, 851, 880, 913, 942 986

interest. The third example is about a univariate failure time variable representing the duration between two related events. The fourth example contains two correlated failure times of interest.

1.2.1 Lung Tumor Data

Hoel and Walberg (1972) give a set of data for 144 male RFM mice in a tumorigenicity experiment that involves lung tumors. The data are presented in Table 1.3 and consist of the death time of each animal measured in days and an indicator of lung tumor presence (1) or absence (0) at time of death. The experiment involves two treatments, conventional environment (CE, 96 mice) and germ-free environment (GE, 48 mice). Lung tumors in RFM mice are predominantly nonlethal, meaning that the occurrence of a tumor does not change the death rate.

Tumorigenicity experiments are usually designed to determine whether a suspected agent or environment accelerates the time until tumor onset in experimental animals. In these situations, the time to tumor onset is usually of interest but not directly observable. Instead, only the death or sacrifice time of an animal is observed, and the presence or absence of a tumor at the time is known. If the tumor can be considered to be rapidly lethal, meaning that its occurrence kills the animal right away, it is reasonable to treat the time of death or sacrifice of an animal as an exact or right-censored observation of the tumor onset time. In this case, the data can be analyzed by methods developed for right-censored failure time data. On the other hand, if the tumor is nonlethal as that considered here, then the time to tumor onset is only known to be less than or greater than the observed time of death or sacrifice. In other words, only left- or right-censored observations on the tumor onset

time are available, and the tumor onset time is interval-censored. This type of interval-censored data is commonly referred to as current status data (see Section 1.3.1).

Among others, one common objective of tumorigenicity experiments is to investigate the effect of a suspected agent or environment on tumor prevalences or incidence rates. For the data in Table 1.3, for example, it is of interest to compare the lung tumor incidence rates of the two treatment groups. More discussion and the analysis of this data set are given in Sections 3.2, 4.5.1, 5.2.2, 5.3.2, 5.4.2, and 5.5.2.

1.2.2 Breast Cancer Study

Table 1.4 presents data from a retrospective study on early breast cancer patients who had been treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. The data are reproduced from Finkelstein and Wolfe (1985) and consist of 94 patients who were given either radiation therapy alone (RT, 46) or radiation therapy plus adjuvant chemotherapy (RCT, 48).

In the study, patients were supposed to be seen at clinic visits every 4 to 6 months. However, actual visit times differ from patient to patient, and times between visits also vary. At visits, physicians evaluated the cosmetic appearance of the patient such as breast retraction, a response that has a negative impact on overall cosmetic appearance. The goal of the study is to compare the two treatments, radiation therapy alone and radiation therapy plus adjuvant chemotherapy, with respect to their cosmetic effects. Adjuvant chemotherapy improves the relapse-free and overall survival for some patients. But there exists some experimental and clinical evidence that suggests that chemotherapy intensifies the acute response of normal tissue to radiation treatment.

The data contain information about the time to breast retraction. However, no exact time was observed. There are 38 patients who did not experience breast retraction during the study, giving right-censored observations

Table 1.4. Observed intervals in months for times to breast retraction of early breast cancer patients

Group	Observed intervals in months
RT	(45,], (25,37], (37,], (4,11], (17,25], (6,10], (46,], (0,5], (33,], (15,], (0,7], (26,40], (18,], (46,], (19,26], (46,], (46,], (24,], (11,15], (11,18] (46,], (27,34], (36,], (37,], (22,], (7,16], (36,44], (5,12], (38,], (34,] (17,], (46,], (19,35], (46,], (5,12], (9,14], (36,48], (17,25], (36,], (46,] (37,44], (37,], (24,], (0,8], (40,], (33,]
RCT	(8,12], (0,5], (30,34], (16,20], (13,], (0,22], (5,8], (13,], (30,36], (18,25] (24,31], (12,20], (10,17], (17,24], (18,24], (17,27], (11,], (8,21], (17,26], (35,] (17,23], (33,40], (4,9], (16,60], (33,], (24,30], (31,], (11,], (15,22], (35,39] (16,24], (13,39], (15,19], (23,], (11,17], (13,], (19,32], (4,8], (22,], (44,48] (11,13], (34,], (34,], (22,32], (11,20], (14,17], (10,35], (48,]

denoted by the intervals with no right end points. For the other patients, the observations are intervals, representing the time periods during which breast retractions occurred. The intervals are given by the last clinic visit time at which breast retraction had not yet occurred and the first clinic visit time at which breast retraction was detected. For example, the observation $(6, 10]$ means that at month 6, the patient had shown no deterioration in cosmetic state, but by the next visit at month 10, breast retraction was present. That is, we have interval-censored data for the time to breast retraction. The analysis of this data set is discussed in Sections 2.3.4, 2.4.3, 3.4.4, 4.5.2, 6.2.3, and 6.5.3.

1.2.3 AIDS Cohort Study

Table 1.5 gives a set of data arising from a cohort study of 257 individuals with Type A or B hemophilia and is reproduced from Kim et al. (1993). The subjects in the study were treated at two French hospitals beginning in 1978 and were at risk for infection of the human immunodeficiency virus (HIV) through contaminated blood factor received for their treatments. The table includes only 188 subjects who were found to be infected with HIV during the study period that lasted from 1978 to August 1988. Among these infected patients, 41 subsequently progressed during the study to the acquired immunodeficiency syndrome (AIDS) or related clinical symptoms, which will be simply referred to as an AIDS diagnosis. One variable of great interest in this study, and also in other similar studies, is the time from HIV infection (or more precisely HIV seroconversion) to AIDS diagnosis. It is often referred to as AIDS incubation or latency time. The AIDS latency time provides information about HIV infection progression and plays an important role in, for example, predicting HIV prevalences.

In this study of HIV infection times, only intervals that bracket the infection time for each study subject are available. This is because HIV infection status was determined by retrospective tests of stored blood sera, and thus the exact HIV infection time was not observed. The intervals given in Table 1.5 are formed by the times at which the last negative and first positive test results were obtained with a unit of six months. In terms of AIDS diagnosis times, they either were observed exactly (for 41 subjects with AIDS diagnosis before the collection of the data) or were right-censored (for the other subjects). This type of censored data is usually referred to as doubly censored failure time data. Note that in the original data set, there are a few subjects whose AIDS diagnosis times were given by narrow intervals, and these are not included in Table 1.5 for simplicity.

In addition to HIV infection and AIDS diagnosis times, Table 1.5 also includes information on a covariate that is a group indicator. The subjects in the study were classified into two groups according to the amount of blood factor that they received. The heavily treated group includes the individuals who received at least 1000 μg/kg of the blood factor for at least one year

Table 1.5. Observed intervals in 6-month scale given by $(L, R]$ for HIV infection time and observations (denoted by T with starred numbers being right-censored times) for AIDS diagnosis time for 188 HIV-infected patients (the numbers in parentheses are multiplicities)

L	R	T		L	R	T		L	R	T		L	R	T		L	R	T	
Lightly treated group																			
0	5	23*	(2)	0	11	23*	(2)	0	12	23*	(3)	0	14	23*		11	14	23*	(3)
0	15	23*	(9)	0	16	23*	(4)	0	17	23*		0	18	23*		11	15	23*	
2	10	23*		5	8	23*		6	10	23*		6	12	23*		13	16	23*	
7	12	23*		7	13	23*		7	15	23*		8	13	23*		5	13	21	
8	14	23*	(3)	9	12	23*	(2)	9	16	23*		10	14	23*	(4)	10	12	23*	(2)
11	13	23*	(4)	11	14	23*		12	14	23*	(4)	12	15	23*	(3)	10	15	23*	
13	15	23*	(4)	14	16	23*	(5)	0	3	8		0	12	15		10	16	23*	
5	12	16		9	11	20		9	12	21		10	12	20		2	16	21	
12	13	22		12	15	22		0	13	23*		6	13	17		12	14	20	
3	11	23*		4	11	23*		5	13	23*		7	16	23*		7	16	21	
8	12	23*		9	15	23*		11	13	23									
Heavily treated group																			
0	7	23*		0	11	23*		0	12	23*	(2)	0	13	23*		0	7	16	
0	14	23*	(3)	0	15	23*	(2)	0	16	23*		2	14	23*		8	11	18	
4	7	23*	(2)	6	9	23*		6	10	23*		7	10	23*		9	12	16	
8	10	23*	(2)	8	12	23*	(3)	9	11	23*	(7)	9	12	23*	(2)	9	14	16	
9	15	23*		10	12	23*		10	13	23*	(4)	11	13	23*	(7)	7	15	23*	
11	14	23*	(2)	12	15	23*	(3)	12	16	23*		13	15	23*	(8)	0	13	23*	
13	16	23*	(2)	14	16	23*	(5)	0	7	13		0	10	12		2	15	23*	
0	15	21		2	7	17		4	7	12		4	8	13		6	15	23*	
6	9	19		7	10	15		8	12	18		8	12	22		12	14	18	
8	13	15		8	13	18		9	11	15		9	11	16		12	15	18	(2)
9	12	17		9	12	23		11	13	20		12	14	20		12	14	21	
13	15	23																	

between 1982 and 1985, whereas the subjects in the lightly treated group received less than 1000 μg/kg in each year. Among others, one objective of interest in this type of study is to estimate the distribution of AIDS latency time. One could also be interested in investigating the effect of covariates on the distribution of the AIDS latency time. These are discussed in detail in Section 8.5.2.

1.2.4 AIDS Clinical Trial

Goggins and Finkelstein (2000) discussed a data set arising from an AIDS clinical trial, AIDS Clinical Trial Group (ACTG) 181, on HIV-infected individuals. The study is a natural history substudy of a comparative clinical trial of three anti-pneumocystis drugs and concerns the opportunistic infection cytomegalovirus (CMV). During the study, among other activities, blood

and urine samples were collected from the patients at their clinical visits and tested for the presence of CMV, which is also commonly referred to as shedding of the virus. These samples and tests provide observed information on the two variables, the times to CMV shedding in blood and in urine.

The observed information is presented in data set I of Appendix A and contains the observed intervals for the times to CMV shedding in blood and urine from 204 patients who provided at least one urine and blood samples during the study. Some intervals contain time zero, that is, the shedding times are left-censored because the shedding had already occurred for these patients when they entered the study. Some intervals have no right end points, that is, the shedding times are right-censored because the corresponding patients had not yet started shedding by the end of the study. For the other patients, their observed intervals are given by the last negative and first positive blood and urine tests, respectively. In summary, we have two possibly correlated failure times of interest, and observations on both of them are interval-censored.

In addition to the observed information about CMV shedding times in blood and in urine, the data set also includes information about the patient's baseline CD4 cell counts given by the indicator variable CD4.ind. In particular, the patients are classified into two groups with CD4.ind equal to 1 if the baseline CD4 cell count was less than 75 (cells/μl) and 0 otherwise. The CD4 cell count indicates the status of a person's immune system and is commonly used to measure the stage of HIV infection. For this data set, one problem of interest is to estimate the association between CMV shedding times in blood and in urine or the joint distribution of the times to CMV shedding in blood and in urine. It is also often of considerable interest to determine the relationship between the time to CMV shedding and the baseline CD4 cell count or whether the baseline CD4 cell count is predictive of CMV shedding in either blood or urine. The analysis of this data set is discussed in Sections 7.2.3 and 7.4.3l.

More examples of interval-censored failure time data and their analyses are given throughout the book. In the next section, we formally introduce several types of interval-censored data that are commonly seen in practice and their corresponding formulations. The methods for their analyses are discussed in the following chapters.

1.3 Types of Interval Censoring and Their Formulations

Let T be a nonnegative random variable representing the failure time of an individual in a failure time study. An observation on T is interval-censored if instead of observing T exactly, only an interval $(L\,,\,R]$ is observed such that

$$T \in (L\,,\,R]\,, \tag{1.1}$$

where $L \leq R$. In the following, we use the convention that $L = R$ means an exact observation, and $R = \infty$ represents a right-censored observation.

In this book, four types of interval censoring that commonly occur in practice and their analyses are considered in detail.

1.3.1 Case I Interval-censored Failure Time Data

The term *case I interval-censored data* is commonly used to refer to interval-censored failure time data in which all observed intervals "include" either time zero or infinity (Groeneboom and Wellner, 1992; Huang, 1996). In other words, the observation on each individual failure time is either left- or right-censored, that is, either $L = 0$ or $R = \infty$. Case I interval-censored data occur when each study subject is observed only once and the only observed information for the survival event of interest is whether the event has occurred no later than the observation time. Instead of the intervals in (1.1), a more convenient representation of case I interval-censored data is $\{ C , \delta = I(T \leq C) \}$, where C denotes the observation time and I is the indicator function. Note that case I interval-censored data differ from right-censored data or left-censored data, which usually include some failure times that are observed exactly.

Case I interval-censored data are also often referred to as *current status data*, a term originating from demographical studies. Cross-sectional studies and tumorigenicity experiments on nonlethal tumors are two types of studies that frequently produce case I interval-censored data. The former is commonly used in demographical studies, and the lung tumor study discussed in Section 1.2.1 provides an example of the latter. Note that there is a fundamental difference between the current status data arising from these two types of studies although they are analyzed in the same way. The current status data from the former occur mainly due to study designs, whereas those given in the latter are observed usually due to the inability to measure the variable directly and/or accurately.

1.3.2 Case II Interval-censored Failure Time Data

Interval-censored data that include at least one interval $(L , R]$ with both L and R belonging to $(0, \infty)$ are usually referred to as *general or case II interval-censored data* (Groeneboom and Wellner, 1992; Huang and Wellner, 1997; Sun, 1998, 2005). In other words, case II interval-censored data are interval-censored data that include some finite intervals away from zero.

Another way to represent a case II interval-censored observation is to use

$$\{ U , V , \delta_1 = I(T \leq U) , \delta_2 = I(U < T \leq V) , \delta_3 = 1 - \delta_1 - \delta_2 \} \quad (1.2)$$

assuming that each subject is observed twice, where U and V are two random variables satisfying $U \leq V$ with probability 1. This formulation is convenient and often used, for example, in a theoretical investigation of an inference procedure. Note that by taking $U = V = C$, case I interval-censored data can be described by (1.2). Yu et al. (2000) generalize this formulation to include exact

observations. Note that in the literature, the term case II interval-censored data is sometimes used to refer only interval-censored data that are given in representation (1.2).

Another generalization of the formulation (1.2) is to assume that there exists a set of observation time points, say $U_1 \leq U_2 ... \leq U_K$, for each study subject, where K is a random integer. The observed information then has the form

$$\{ (K , U_j , \delta_j = I(U_{j-1} < T \leq U_j)) , \ j = 1, ..., K \} , \tag{1.3}$$

where $U_0 = 0$. This formulation or type of failure time data is often referred to as case K or mixed case interval-censored data (Schick and Yu, 2000; Wellner, 1995). It is apparent that the above formulation includes the representation (1.2) as a special case and provides a natural representation of interval-censored failure time data arising from longitudinal studies with periodic follow-up.

All three representations, (1.1) to (1.3), give rise to the same likelihood function. Note that although both representations (1.2) and (1.3) seem natural, it is not common to have interval-censored data collected or given in these formats in practice. However, it is much easier and more natural to impose assumptions such as independence with T on them than on representation (1.1), which is often needed for derivation of the asymptotic properties of inference procedures. For data given in representation (1.2) or (1.3), one can easily obtain the corresponding data with representation (1.1). On the other hand, it is apparently impossible to transform representation (1.1) to (1.3) without extra information about observation process, and it is not straightforward to transform observations given in representation (1.1) to these in the representation (1.2). More discussion on this is given later. In the following chapters, we mainly focus on the first two representations and use them interchangeably.

1.3.3 Doubly Censored Failure Time Data

Consider a survival study involving two related events and let X and S denote the times of the occurrences of the two events with $X \leq S$. Define $T = S - X$ and suppose that T is the survival time of interest. By doubly censored failure time data, we mean that the observations on both X and S are interval-censored (De Gruttola and Lagakos, 1989; Sun, 2004). Specifically, suppose that instead of observing X and S exactly, one only observes two intervals $(L , R]$ and $(U , V]$ such that

$$X \in (L , R] \ , \ \ S \in (U , V] ,$$

where $L \leq R$ and $U \leq V$ with probability 1. In other words, the observations on T are doubly censored.

The special type of doubly censored data in which S is only right-censored occurs commonly, and in this case, one has either $U = V$ or $V = \infty$. Another

formulation for this special case that may be more natural is to assume that there exists a censoring variable C, which is often assumed to be independent of S. The observation on S then consists of $S^* = \min\{S, C\}$ and $\delta = I(S^* = S)$, where I is the indicator function as before.

One often sees doubly censored failure time data in disease progression studies where the two events may represent infection and subsequent onset of a certain disease, respectively, such as the example discussed in Section 1.2.3. In these situations, doubly censored observations occur mainly due to the nature of the disease and/or the structure of the study design. In the example given in Section 1.2.3, X and S represent HIV infection and AIDS diagnosis times, respectively, and T is the AIDS latency time. For most AIDS cohort studies, as in this example, because HIV infection usually is determined through periodic blood tests, observations on it are commonly interval-censored. Also, observations on the diagnosis of AIDS could be, for example, right-censored due to the end of the study, thus yielding doubly censored data on T.

Doubly censored failure time data include as special cases right-censored and interval-censored failure time data. For example, they reduce to interval-censored data if the time of occurrence of the first event, X, can be observed exactly ($L = R$). Furthermore, if the observation on the time of occurrence of the subsequent event, S, is exact or right-censored, we then have a right-censored observation on T. Note that for doubly censored data, if X is observed exactly, for inferences about T, one may relabel so that $X = 0$, which typically is done in failure time data analysis

In the literature, doubly censored data considered here are sometimes referred to as doubly interval-censored data (Sun, 1995) to distinguish them from another type of doubly censored failure time data. In the latter, the survival time of interest is observed exactly if it is within a window and left- or right-censored if it is to the left or right of the window (Cai and Cheng, 2004; Chen and Zhou, 2003; Turnbull, 1974). A key difference between the two types of data is that for the latter type of data, some exact failure times are observed, but if not, they become case I interval-censored data. The methods required for the analyses of these two types of doubly censored data are different.

1.3.4 Panel Count Data

Interval censoring occurs in a more general setting than survival studies. In failure time data analysis, the random variable of interest is always the time to an event, and the event is treated as an absorbing event. In other words, the event can occur only once such as fatal failure or death. In practice, however, there exist many situations where the event of interest can occur multiple times such as a tumor or disease symptom. In these situations, in addition to the time to the event or between the occurrences of the event, one may also want to study the occurrence process of the event. Without interval censoring, that is, if the process is observed continuously, then one

has what is commonly called *recurrent event data*, in which one knows all the exact occurrence times of the event (Cai and Schaubel, 2004; Chang and Wang, 1999). In the presence of interval censoring, which arises if the subject or occurrence process is observed only at discrete time points, one only knows the numbers of the occurrences of the event between observation times. In this case, the observed data are often referred to as *panel count data* (Kalbfleisch and Lawless, 1985; Sun and Wei, 2000). However, if the event can occur only once, then the data become interval-censored failure time data. Panel count data are also sometimes referred to as interval count data or interval-censored recurrent event data (Lawless and Zhan, 1998; Thall, 1988).

Panel count data frequently occur in long-term clinical, industrial, or animal studies. In a follow-up cancer study, for example, one could be interested in the recurrence rate of one or more types of tumors or of tumors at one or more locations. For such a study, it is usually impossible or impractical to follow study subjects continuously, and thus panel count data are obtained. Another example is longitudinal sociological studies on, for example, job changes.

Define a counting process $N(t)$ with $N(t)$ denoting the number of occurrences of a recurrent event up to and including time t. For usual survival problems, $N(t)$ is a 0-1 counting process, and the counting process formulation has been used extensively in the literature for the development of statistical methods for the analysis of right-censored failure time data. For more detailed discussion on this, one can read, for example, the book by Andersen et al. (1993). The methodology described there can also be used for the analysis of recurrent event data. In the case of panel count data, the values of $N(t)$ are known only at different observation time points, and we do not know the time points at which $N(t)$ jumps. In this book, for the analyses of panel count data, we focus on methods that allow observation times to vary from subject to subject.

1.3.5 Independent Interval Censoring, Notation, and Remarks

By *independent interval censoring*, as independent right censoring, we mean that the mechanism that generates the censoring is independent of the underlying variable of interest completely or given covariates. For current status data, this implies that C and T are independent. For interval-censored data given in representation (1.2) or (1.3), the independent interval censoring means that the joint distribution of U and V or the U_j's contains no parameters that are involved in the survival function of T. With respect to the data given in the format (1.1), the independent interval censoring assumes that an interval $(L, R]$ gives no more than the information that T is simply bracketed by the two observed values. In other words, we have

$$P(T \le t \,|\, L = l, R = r, L \le T < R) = P(T \le t \,|\, l \le T < r)$$

(Self and Grossman, 1986; Zhang et al., 2005), or

$$P(L < T \leq R \,|\, L = l, R = r) = P(l < T \leq r)$$

and the joint distribution of L and R is free of the parameters involved in the survival function of T. More remarks about this independent censoring mechanism are given in Section 10.5 along with discussion on situations where it does not hold. Under the independent interval censoring, one does not have to deal with the censoring mechanism in analyzing interval-censored data. Throughout the book, the independent interval censoring is assumed unless otherwise specified.

For presentation of an interval-censored observation, instead of $(L, R]$, one could also use $[L, R]$, $[L, R)$, or (L, R) (Peto, 1973; Turnbull, 1976). If T is continuous, it is apparent that there is no difference among them in the sense that they represent the same observed information about T. On the other hand, if T is discrete, care is needed because the information given by them can be different. Some discussion on this can be found in Ng (2002), and the notation $(L, R]$ is used throughout this book.

As mentioned above, for T, exact and right-censored observations can be seen as special cases of interval-censored observations. In practice, a set of interval-censored data may include both exact and purely interval-censored observations. Suppose that T is continuous. Then for an exact observation $T = t_0$, its likelihood contribution is $f(t_0)$, and for a purely interval-censored observation $(L, R]$, the likelihood contribution has the form $S(L) - S(R)$, where $f(t)$ and $S(t) = P(T > t)$ denote the density and survival functions of T, respectively. In the following, we mainly focus attention on purely interval-censored observations and the corresponding likelihood contribution in the construction of likelihood functions. In other words, for the construction of likelihood functions, we assume for convenience that no exact observations are present unless otherwise specified. The derivation and development of most likelihood-based inference procedures in this book hold when exact failure times are present and the corresponding likelihood contributions are included.

In addition to those described in the previous subsections, interval censoring can also occur in other formulations. For example, interval-censored data can arise from a multi-state model (Commenges, 2003). Also in a survival study, the variable that suffers interval censoring may be a covariate instead of the survival time of interest as discussed above (Goggins et al., 1999b). More generally, observations on both covariates and survival variables may be interval-censored (Zhao et al., 2005). More discussion on this can be found in Section 10.3. As in the case of right censoring, truncation may occur together with interval censoring. By truncation, as before, we mean that a subject is included in a study only if its failure time belongs to a certain window. Here truncation can occur for the same reasons as those for right-censored failure time data. For example, left-truncated and interval-censored data occur if the survival time T is observed only if T is greater than a certain value and only an interval to which T belongs can be observed. In the following, we focus mainly on situations without truncation.

We remark that in practice, interval-censored data are often collected and presented as discrete data, and this is especially the case when the data arise from follow-up studies with day, month, or year as the time unit. Therefore, it is natural and convenient to treat the underlying survival variables as discrete variables in the development of approaches for their analyses. Also it is reasonable and sometimes convenient to treat them as continuous variables as the measured values are often approximations to the true values due to, for example, measurement errors. This is especially the case for the investigation of large sample properties of the methods of analysis. In the following discussion, the two formulations are used interchangeably depending on convenience and purpose.

1.4 Concepts and Some Regression Models

Let T denote a nonnegative random variable representing the failure time of a subject, that is, the survival variable of interest. For inferences about T, the survival function and the hazard function are particularly useful for modeling. The survival function of T is defined as the probability that T exceeds a value t. Let $S(t)$ denote the survival function of T. Then one has

$$S(t) = P(T > t), \quad 0 < t < \infty.$$

The hazard function is defined differently for continuous and discrete survival variables and these definitions are given below. The probability density and distribution functions are often used too in survival analysis although not as frequently as the survival and hazard functions.

In addition to reviewing these functions along with their relationships, this section describes several continuous semiparametric regression models commonly used in survival analysis. These include the Cox or proportional hazards model, the proportional odds model, the additive hazards model, the accelerated failure time model, and the linear transformation model. Two discrete regression models are also presented. Some commonly used parametric models are discussed in the next chapter along with the corresponding inference procedures and the imputation approach for the analysis of interval-censored data.

1.4.1 Continuous Survival Variables

Assume that T is absolutely continuous and thus its probability density function $f(t)$ exists. By definition, it is easy to see that the density function and the survival function satisfy

$$f(t) = -dS(t)/dt$$

or

$$S(t) \;=\; \int_t^{\infty} f(s)\, ds\;.$$

The hazard function of T at time t is defined as

$$\lambda(t) \;=\; \lim_{\Delta t \to 0^+} \frac{P(\, t \le T < t + \Delta t \,|\, T \ge t\,)}{\Delta t}\;.$$

It represents the instantaneous probability that a subject fails at time t given that the subject has not failed before t. The survival, density, and hazard functions have one-to-one relationship. Specifically, given the density or survival function, we have

$$\lambda(t) \;=\; \frac{f(t)}{S(t)} \;=\; -\,\frac{d \log S(t)}{dt}\;.$$

On the other hand, it can be proved that

$$S(t) \;=\; \exp\left[\,-\int_0^t \lambda(s)\, ds\,\right] \;=\; \exp\left[\,-\,\Lambda(t)\,\right]$$

and

$$f(t) \;=\; \lambda(t)\, \exp[\,-\,\Lambda(t)\,]\,,$$

where $\Lambda(t) \;=\; \int_0^t \lambda(s)\, ds$, which is commonly referred to as the cumulative hazard function of T.

1.4.2 Discrete Survival Variables

Assume that T is a discrete survival variable taking values $s_1 < s_2 < ...$ with probability function $\{\, f(s_j) = P(T = s_j)\,;\, j = 1, 2, ...\}$. Then one has

$$S(t) \;=\; \sum_{j:t<s_j} f(s_j)\;.$$

In this case, the hazard of T at s_j is defined as

$$p_j \;=\; P(\, T = s_j \,|\, T \ge s_j\,) \;=\; \frac{f(s_j)}{S(s_j-)}\,,$$

the conditional probability that the failure occurs at s_j given that the failure has not occurred before s_j, $j \;=\; 1, 2, ...$

As in the continuous case, the survival, density, and hazard functions uniquely determine each other. Based on the above definitions, one can show that

$$S(t) \;=\; \prod_{j:t \ge s_j} (1 \,-\, p_j)$$

and

$$f(s_j) \;=\; p_j \prod_{l=1}^{j-1} (1 \,-\, p_l)\,.$$

In survival analysis, due to the special structure of the observed information and questions of interest, it is more convenient to model the hazard function or the survival function than other functions that determine the distribution of T. The remainder of this section discusses several such regression models.

1.4.3 The Proportional Hazards Model

Let $\boldsymbol{Z}$ be a vector of covariates including, for example, treatment indicator, age, and gender. As remarked before, a regression analysis provides an assessment of covariate effects on failure time, which is one of the important tasks in survival analysis. For this, a regression model is usually needed to specify how the covariates affect the failure time of interest. The proportional hazards (PH) or Cox model assumes that the hazard function of T has the form

$$\lambda(t; \boldsymbol{Z}) \;=\; \lambda_0(t)\, \exp(\boldsymbol{Z}'\boldsymbol{\beta}) \tag{1.4}$$

given covariates $\boldsymbol{Z}$ (Cox, 1972). In the above, $\lambda_0(t)$ is an arbitrary unspecified baseline hazard function, and $\boldsymbol{\beta}$ is the vector of regression parameters. This model specifies that the covariates act multiplicatively on the hazard function.

The model (1.4) says that the ratio of the hazard functions for two subjects with different covariates is constant. In particular, for the two-sample situation where $Z \;=\; 0$ or 1, one has

$$\frac{\lambda(t; Z = 1)}{\lambda(t; Z = 0)} \;=\; \exp(\beta)\;.$$

Under the PH model, the conditional density and survival functions of T given $\boldsymbol{Z}$ have the forms

$$f(t; \boldsymbol{Z}) \;=\; \lambda_0(t)\, \exp(\boldsymbol{Z}'\boldsymbol{\beta})\, \exp\big[\, -\Lambda_0(t)\, \exp(\boldsymbol{Z}'\boldsymbol{\beta})\,\big]$$

and

$$S(t; \boldsymbol{Z}) \;=\; \exp[\,-\Lambda_0(t)\, \exp(\boldsymbol{Z}'\boldsymbol{\beta})\,] \;=\; [\,S_0(t)\,]^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})}\;,$$

where

$$\Lambda_0(t) \;=\; \int_0^t \lambda_0(s) ds$$

and

$$S_0(t) \;=\; \exp\left[\, -\int_0^t \lambda_0(s) ds \right]$$

are the baseline cumulative hazard function and the baseline survival function. The conditional cumulative hazard function of T given $\boldsymbol{Z}$ has the form