□ 应用统计学丛书

Survival Analysis:
Models and Applications

生存分析：模型与应用

刘 宪

高等教育出版社
HIGHER EDUCATION PRESS

□ 应用统计学丛书

Survival Analysis:
Models and Applications

生存分析：模型与应用

刘 宪

SHENGCUN FENXI
MOXING YU YINGYONG

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目 (C I P) 数据

生存分析：模型与应用 = Survival Analysis : Models and Applications : 英文 / 刘宪著. -- 北京：高等教育出版社, 2012. 7
ISBN 978-7-04-034826-2

Ⅰ. ①生… Ⅱ. ①刘… Ⅲ. ①生存率-统计分析（数学）-高等学校-教材-英文 Ⅳ. ①R195.3

中国版本图书馆 CIP 数据核字 (2012) 第 156244 号

策划编辑　王丽萍　　责任编辑　王丽萍　　封面设计　王凌波　　责任印制　朱学忠

出版发行	高等教育出版社	咨询电话	400-810-0598
社　　址	北京市西城区德外大街 4 号	网　　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　　刷	涿州市星河印刷有限公司	网上订购	http://www.landraco.com
开　　本	787 mm×1092 mm　1/16		http://www.landraco.com.cn
印　　张	28.75	版　　次	2012 年 7 月第 1 版
字　　数	740 千字	印　　次	2012 年 7 月第 1 次印刷
购书热线	010-58581118	定　　价	79.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

物 料 号　34826-00

Preface

Survival analysis concerns sequential occurrences of events governed by probabilistic laws. Recent decades have witnessed many applications of survival analysis in various disciplines. The primary objective of this book is to provide an introduction to the many specialized facets on survival analysis. Scope-wise, this book is expected to appeal to a wide variety of disciplinary areas. Given my multidisciplinary background in training and research, this book of survival analysis covers techniques and specifications applied in medicine, biostatistics, demography, mathematical biology, sociology, and epidemiology, with practical examples associated with each of those disciplines. The celebrated Cox model, used in almost all applied areas, is paid special attention to in this book, with three chapters devoted to this innovative perspective. I also describe counting processes and the martingale theory in considerable detail, in view of their flexibility and increasing popularity in survival analysis, particularly in the field of biostatistics. Regression modeling, mathematical simulation, and computing programming, which attach to different phases of survival analysis, are described and applied extensively in this book, so scientists and professors of various disciplines can benefit from using it either as a useful reference book or as a textbook in graduate courses.

In this book, a large number of survival functions, models, and techniques are introduced, described, and discussed with empirical examples. The presentation of those survival perspectives starts with the most basic specifications and ends with some more advanced techniques in the literature of survival analysis. With a considerable volume of empirical illustrations, I attempt to make the transition from the introductory to the advanced levels as coherent and smooth as possible. Almost for every major survival method or model, step-by-step instructions are provided for leading the reader in to learning how to perform the techniques, supplemented by empirical practices, computing programs, and detailed interpretations of analytic results.

Given the focus on application and practice in this book, the audience includes professionals, academics, and graduate students who have some experience in survival analysis. A fair number of illustrations on various topics permits professionals to learn new methods or to improve their professional skills in performing survival analysis. As it covers a wide scope of survival techniques and methods, from the introductory to the advanced, this book can be used as a useful reference book for planners, researchers, and professors who are working in settings involving various lifetime events. Scientists interested in survival analysis should find it a useful guidebook for the incorporation of survival data and methods into their projects.

Graduate students of various disciplines constitute another important component of the audience. Social science students can benefit from the application of survival concepts and methods to the solution of problems in sociology, economics, psychology, geography, and political science. This book provides a useful framework and practical examples of applied social science, especially at a time when more survival-related questions are raised. The accessibility of many observational, longitudinal data in the public domain since the 1980s will facilitate interested students to practice further the methods and techniques learned from this book. Graduates students of biology, medicine, and public health, who are interested in doing research for their future careers, can learn plenty of techniques from this book for performing mathematical simulation, clinical trials, and competing risks analyses on mortality and disease. Survival analysis and some other related courses have long been recognized as essential components for graduate students training in mathematical biology, epidemiology, and some of the biomedical departments. In medical schools, for example, this book can have wide appeal among medical students who want to know how to analyze data of a clinical trial for understanding the effectiveness of a new medical treatment or of a new medicine on disease.

If the reader attempts to understand the entire body of the methods and techniques covered by this book, the prerequisites should include calculus, matrix algebra, and generalized linear modeling. For those not particularly familiar with that required knowledge, they might want to skip detailed mathematical and statistical steps, and place their focus upon empirical illustrations and computer programming skills. By doing so, they can still command how to apply various survival techniques effectively, thereby adding new dimensions to their professional, research, or teaching activities. Therefore, this book can be read selectively by the reader who is not extremely competent with high-level mathematics and statistics.

The reviewers of the proposal and an example chapter for this book were: Kenneth Land, Duke University; David Swanson, University of California, Riverside; and Jichuan Wang, George Washington University. Additionally, a number of other colleagues and friends have enriched, supported and refined the intellectual development of this book, including Lyn Albrecht, Kristie Gore, Albert I. Hermalin, James Edward McCarroll, Robert Ursano, Lois Verbrugge, Anatoli I. Yashin, and Chu Zhang. Sincere thanks are given to Paul T. Savarese of the SAS Institute for letting me use some of his personal SAS programs in Chapters 4 and 8. Part of the work in Chapter 8 was initiated at the Population Studies Center, the Institute for Social Research at the University of Michigan, and the mentorship of Albert I. Hermalin is specially acknowledged.

I owe special thanks to Charles C. Engel, whose consistent support and help has made completion of this book possible. The staff of the Deployment Health Clinical Center, Walter Reed National Military Medical Center, provides tremendous dedication, competence, and excellence in the course of the preparation of this book. Malisa Arnold's and Phoebe McCutchan's assistance in editing the text and some of the graphs was vital.

Finally, I would like to thank my wife, Ming Dong, for her support and encouragement throughout the entire period of preparing, writing, and editing this book.

Xian Liu

本书由高等教育出版社和 Wiley 公司合作出版，由 Wiley 负责编辑加工和排版。故书中量和单位等难免有不符合我国编辑规范的地方，特此说明。

Contents

Preface xi

1 Introduction 1

1.1 What is survival analysis and how is it applied? 1
1.2 The history of survival analysis and its progress 2
1.3 General features of survival data structure 3
1.4 Censoring 4
1.4.1 Mechanisms of right censoring 5
1.4.2 Left censoring, interval censoring, and left truncation 6
1.5 Time scale and the origin of time 7
1.5.1 Observational studies 8
1.5.2 Biomedical studies 9
1.5.3 Health care utilization 9
1.6 Basic lifetime functions 10
1.6.1 Continuous lifetime functions 10
1.6.2 Discrete lifetime functions 12
1.6.3 Basic likelihood functions for right, left, and interval censoring 14
1.7 Organization of the book and data used for illustrations 16
1.8 Criteria for performing survival analysis 17

2 Descriptive approaches of survival analysis 20

2.1 The Kaplan–Meier (product-limit) and Nelson–Aalen estimators 21
2.1.1 Kaplan–Meier estimating procedures with or without censoring 21
2.1.2 Formulation of the Kaplan–Meier and Nelson–Aalen estimators 24
2.1.3 Variance and standard error of the survival function 27
2.1.4 Confidence intervals and confidence bands of the survival function 29
2.2 Life table methods 36
2.2.1 Life table indicators 37
2.2.2 Multistate life tables 40
2.2.3 Illustration: Life table estimates for older Americans 44
2.3 Group comparison of survival functions 46
2.3.1 Logrank test for survival curves of two groups 48
2.3.2 The Wilcoxon rank sum test on survival curves of two groups 51
2.3.3 Comparison of survival functions for more than two groups 55
2.3.4 Illustration: Comparison of survival curves between married and unmarried persons 58
2.4 Summary 61

3 Some popular survival distribution functions 63
3.1 Exponential survival distribution 63
3.2 The Weibull distribution and extreme value theory 68
3.2.1 Basic specifications of the Weibull distribution 68
3.2.2 The extreme value distribution 72
3.3 Gamma distribution 73
3.4 Lognormal distribution 77
3.5 Log-logistic distribution 80
3.6 Gompertz distribution and Gompertz-type hazard models 83
3.7 Hypergeometric distribution 89
3.8 Other distributions 91
3.9 Summary 92

4 Parametric regression models of survival analysis 93
4.1 General specifications and inferences of parametric regression models 94
4.1.1 Specifications of parametric regression models on the hazard function 94
4.1.2 Specifications of accelerated failure time regression models 96
4.1.3 Inferences of parametric regression models and likelihood functions 99
4.1.4 Procedures of maximization and hypothesis testing on ML estimates 101
4.2 Exponential regression models 103
4.2.1 Exponential regression model on the hazard function 103
4.2.2 Exponential accelerated failure time regression model 106
4.2.3 Illustration: Exponential regression model on marital status and survival among older Americans 108
4.3 Weibull regression models 113
4.3.1 Weibull hazard regression model 114
4.3.2 Weibull accelerated failure time regression model 115
4.3.3 Conversion of Weibull proportional hazard and AFT parameters 117
4.3.4 Illustration: A Weibull regression model on marital status and survival among older Americans 121
4.4 Log-logistic regression models 127
4.4.1 Specifications of the log-logistic AFT regression model 127
4.4.2 Retransformation of AFT parameters to untransformed log-logistic parameters 129
4.4.3 Illustration: The log-logistic regression model on marital status and survival among the oldest old Americans 131
4.5 Other parametric regression models 135
4.5.1 The lognormal regression model 136
4.5.2 Gamma distributed regression models 137
4.6 Parametric regression models with interval censoring 138
4.6.1 Inference of parametric regression models with interval censoring 138
4.6.2 Illustration: A parametric survival model with independent interval censoring 139
4.7 Summary 142

5 The Cox proportional hazard regression model and advances 144
5.1 The Cox semi-parametric hazard model 145

5.1.1 Basic specifications of the Cox proportional hazard model 145
5.1.2 Partial likelihood 147
5.1.3 Procedures of maximization and hypothesis testing on partial likelihood 150
5.2 Estimation of the Cox hazard model with tied survival times 154
5.2.1 The discrete-time logistic regression model 154
5.2.2 Approximate methods handling ties in the proportional hazard model 155
5.2.3 Illustration on tied survival data: Smoking cigarettes and the mortality of older Americans 157
5.3 Estimation of survival functions from the Cox proportional hazard model 161
5.3.1 The Kalbfleisch–Prentice method 162
5.3.2 The Breslow method 164
5.3.3 Illustration: Comparing survival curves for smokers and nonsmokers among older Americans 165
5.4 The hazard rate model with time-dependent covariates 169
5.4.1 Categorization of time-dependent covariates 169
5.4.2 The hazard rate model with time-dependent covariates 171
5.4.3 Illustration: A hazard model on time-dependent marital status and the mortality of older Americans 173
5.5 Stratified proportional hazard rate model 176
5.5.1 Specifications of the stratified hazard rate model 177
5.5.2 Illustration: Smoking cigarettes and the mortality of older Americans with stratification on three age groups 178
5.6 Left truncation, left censoring, and interval censoring 183
5.6.1 The Cox model with left truncation, left censoring, and interval censoring 184
5.6.2 Illustration: Analyzing left truncated survival data on smoking cigarettes and the mortality of unmarried older Americans 185
5.7 Qualitative factors and local tests 191
5.7.1 Qualitative factors and scaling approaches 191
5.7.2 Local tests 193
5.7.3 Illustration of local tests: Educational attainment and the mortality of older Americans 195
5.8 Summary 199

6 Counting processes and diagnostics of the Cox model 201

6.1 Counting processes and the martingale theory 202
6.1.1 Counting processes 202
6.1.2 The martingale theory 204
6.1.3 Stochastic integrated processes as martingale transforms 207
6.1.4 Martingale central limit theorems 208
6.1.5 Counting process formulation for the Cox model 211
6.2 Residuals of the Cox proportional hazard model 213
6.2.1 Cox–Snell residuals 213
6.2.2 Schoenfeld residuals 214
6.2.3 Martingale residuals 216

6.2.4 Score residuals 218
6.2.5 Deviance residuals 219
6.2.6 Illustration: Residual analysis on the Cox model of smoking cigarettes and the mortality of older Americans 220
6.3 Assessment of proportional hazards assumption 222
6.3.1 Checking proportionality by adding a time-dependent variable 225
6.3.2 The Andersen plots for checking proportionality 227
6.3.3 Checking proportionality with scaled Schoenfeld residuals 228
6.3.4 The Arjas plots 229
6.3.5 Checking proportionality with cumulative sums of martingale-based residuals 230
6.3.6 Illustration: Checking the proportionality assumption in the Cox model for the effect of age on the mortality of older Americans 232
6.4 Checking the functional form of a covariate 236
6.4.1 Checking model fit statistics for different link functions 236
6.4.2 Checking the functional form with cumulative sums of martingale-based residuals 237
6.4.3 Illustration: Checking the functional form of age in the Cox model on the mortality of older Americans 239
6.5 Identification of influential observations in the Cox model 243
6.5.1 The likelihood displacement statistic approximation 244
6.5.2 LMAX statistic for identification of influential observations 247
6.5.3 Illustration: Checking influential observations in the Cox model on the mortality of older Americans 248
6.6 Summary 253

7 Competing risks models and repeated events 255

7.1 Competing risks hazard rate models 256
7.1.1 Latent failure times of competing risks and model specifications 256
7.1.2 Competing risks models and the likelihood function without covariates 259
7.1.3 Inference for competing risks models with covariates 261
7.1.4 Competing risks model using the multinomial logit regression 263
7.1.5 Competing risks model with dependent failure types 266
7.1.6 Illustration of competing risks models: Smoking cigarettes and the mortality of older Americans from three causes of death 268
7.2 Repeated events 282
7.2.1 Andersen and Gill model (AG) 283
7.2.2 PWP total time and gap time models (PWP-CP and PWP-GT) 286
7.2.3 The WLW model and extensions 288
7.2.4 Proportional rate and mean functions of repeated events 291
7.2.5 Illustration: The effects of a medical treatment on repeated patient visits 294
7.3 Summary 308

8 Structural hazard rate regression models **310**

8.1 Some thoughts about the structural hazard regression models 310
8.2 Structural hazard rate model with retransformation of random errors 313
8.2.1 Model specification 314
8.2.2 The estimation of the full model 317
8.2.3 The estimation of reduced-form equations 318
8.2.4 Decomposition of causal effects on hazard rates and survival functions 323
8.2.5 Illustration: The effects of veteran status on the mortality of older Americans and its pathways 327
8.3 Summary 344

9 Special topics **347**

9.1 Informative censoring 347
9.1.1 Selection model 348
9.1.2 Sensitivity analysis models 351
9.1.3 Comments on current models handling informative censoring 352
9.2 Bivariate and multivariate survival functions 352
9.2.1 Inference of the bivariate survival model 353
9.2.2 Estimation of bivariate and multivariate survival models 355
9.2.3 Illustration of marginal models handling multivariate survival data 359
9.3 Frailty models 359
9.3.1 Hazard models with individual frailty 360
9.3.2 The correlated frailty model 364
9.3.3 Illustration of frailty models: The effect of veteran status on the mortality of older Americans revisited 366
9.4 Mortality crossovers and the maximum life span 376
9.4.1 Basic specifications 378
9.4.2 Relative acceleration of the hazard rate and timing of mortality crossing 381
9.4.3 Mathematical conditions for maximum life span and mortality crossover 383
9.5 Survival convergence and the preceding mortality crossover 384
9.5.1 Mathematical proofs for survival convergence and mortality crossovers 385
9.5.2 Simulations 387
9.5.3 Explanations for survival convergence and the preceding mortality crossover 393
9.6 Sample size required and power analysis 398
9.6.1 Calculation of sample size required 399
9.6.2 Illustration: Calculating sample size required 401
9.7 Summary 403

Appendix A The delta method **405**

Appendix B **Approximation of the variance–covariance matrix for the predicted probabilities from results of the multinomial logit model** **407**

Appendix C **Simulated patient data on treatment of PTSD ($n = 255$)** **410**

Appendix D **SAS code for derivation of φ estimates in reduced-form equations** **417**

Appendix E **The analytic result of $\kappa^*(x)$** **422**

References **424**

Index **438**

1

Introduction

1.1 What is survival analysis and how is it applied?

'What is survival analysis?' Before starting discussion on this topic, think about what 'survives.' In the cases considered here, we are talking about things that have a life span, those things that are 'born,' live, change status while they live, and then die. Therefore, 'survival' is the description of a life span or a living process before the occurrence of a status change or, using appropriate jargon, an *event*.

In terms of 'survival,' what we think of first are organisms like various animal species and other life forms. After birth, a living entity grows, goes through an aging process, and then decomposes gradually. All the while, they remain what they are – the same organisms. The gradual changes and developments over a life course reflect the survival process. For human beings in particular, we survive from death, disease, and functional disablement. While biology forms its primary basis, the significance of survival is largely social. At different life stages, we attend school, get married, develop a professional career, and retire when getting old. In the meantime, many of us experience family disruption, become involved in social activities, cultivate personal habits and hobbies, and make adjustments to our daily lives according to physical and mental conditions. These social facets are things that are not organisms but their life span is *like* that of a living being: things that *live*, things that have beginnings, transformations, and then *deaths*. In a larger context, survival can also include such events as an automobile breakdown, the collapse of a political system in a country, or the relocation of a working unit. In cases such as these and in others, existence dictates processes of survival and their status change, indicated by the occurrence of events.

The practice of survival analysis is the use of reason to describe, measure, and analyze features of events for making predictions about not only survival but also 'time-to-event processes' – the length of time until the change of status or the occurrence of an event – such as from living to dead, from single to married, or from healthy to sick. Because a life span, genetically, biologically, or mechanically, can be cut short by illness, violence, environment, or other factors, much research in survival analysis involves making comparisons among

Survival Analysis: Models and Applications, First Edition. Xian Liu.

groups or categories of a population, or examining the variables that influence its survival processes. As they have come to realize the importance of examining the inherent mechanisms, scientists have developed many methods and techniques seeking to capture underlying features of various survival processes. In the academic realm, survival analysis is now widely applied in a long list of applied sciences, owing considerably to the availability of longitudinal data that records histories of various survival processes and the occurrences of various events. At present, the concept of survival no longer simply refers to a biomedical or a demographic event; rather, it expands to indicate a much broader scope of phenomena characterized by time-to-event processes.

In medical research, clinical trials are regularly used to assess the effectiveness of new medicines or treatments of disease. In these settings, researchers apply survival analysis to compare the risk of death or recovery from disease between or among population groups receiving different medications or treatments. The results of such an analysis, in turn, can provide important information with policy implications.

Survival analysis is also applied in biological research. Mathematical biologists have long been interested in evolutionary perspectives of senescence for human populations and other species. By using survival analysis as the underlying means, they delineate the life history for a species' population and link its survival processes to a collection of physical attributes and behavioral characteristics for examining its responses to its environment.

Survival data are commonly collected and analyzed in social science, with topics ranging widely, from unemployment to drug use recidivism, marital disruption, occupational careers, and other social processes. In demography, in addition to the mortality analysis, researchers are concerned with such survival processes as the initiation of contraceptive use, internal and international migration, and the first live birth intervals.

In the field of public health, survival analysis can be applied to the analysis of health care utilization. Such examination is of special importance for both planners and academics because the health services system reflects the political and economic organization of a society and is concerned with fundamental philosophical issues involving life, death, and the quality of life.

Survival analysis has also seen wide applications in some other disciplines such as engineering, political science, business management, and economics. For example, in engineering, scientists apply survival analysis to perform life tests on the durability of mechanical or electric products. Specifically, they might track a sample of products over their life course for assessing characteristics and materials of the product's designed life and for predicting product reliability. Results of such studies can be used for the quality improvement of the products.

1.2 The history of survival analysis and its progress

Originally, survival analysis was used solely for investigations of mortality and morbidity on vital registration statistics. The earliest arithmetical analysis of human survival processes can be traced back to the 17th century, when the English statistician John Graunt published the first life table in 1662 (Graunt, 1939, original edition, 1662). For a long period of time, survival analysis was considered an analytic instrument, particularly in biomedical and demographical studies. At a later stage, it gradually expanded to the domain of engineering to describe/evaluate the course of industrial products. In the past forty years, the scope of survival analysis has grown tremendously as a consequence of rapid developments in computer science, particularly the advancement of powerful statistical software packages. The convenience of using computer software for creating and utilizing complex statistical models has led scientists of many disciplines to begin using survival models.

As applications of survival analysis have grown rapidly, methodological innovation has accelerated at an unprecedented pace over the past several decades. The advent of the Cox model and the partial likelihood perspective in 1972 triggered the advancement of a large number of statistical methods and techniques characterized by regression modeling in the analysis of survival data. The major contribution of the Cox model, given its capability of generating simplified estimating procedures in analyzing survival data, is the provision of a flexible statistical approach to model the complicated survival processes as associated with measurable covariates. More recently, the emergence of the counting processes theory, a unique counting system for the description of survival dynamics, highlights the dawning of a new era in survival analysis due to its tremendous inferential power and high flexibility for modeling repeated events for the same observation and some other complicated survival processes. In particular, this modern perspective combines elements in the large sample theory, the martingale theory, and the stochastic integration theory, providing a new set of statistical procedures and rules in modeling survival data. To date, the counting process system and the martingale theory have been applied by statisticians to develop new theorems and more refined statistical models, thus bringing a new direction in survival analysis.

1.3 General features of survival data structure

In essence, a survival process describes a life span from a specified starting time to the occurrence of a particular event. Therefore, the primary feature of survival data is the description of a change in status as the underlying outcome measure. More formally, a status change is the occurrence of an *event* designating the end of a life span or the termination of a survival process. For instance, a status change occurs when a person dies, gets married, or when an automobile breaks down. This feature of a status 'jump' makes survival analysis somewhat similar to some more conventional statistical perspectives on qualitative outcome data, such as the logistic or the probit model. Broadly speaking, those traditional models can also be used to examine a status change or the occurrence of a particular event by comparing the status at the beginning and the status at the end of an observation interval. Those statistical approaches, however, ignore the timing of the occurrence of this lifetime event, and thereby do not possess the capability of describing a time-to-event process. A lack of this capability can be detrimental to the quality of analytic results, thereby generating misleading conclusions. The logistic regression, for example, can be applied to estimate the probability of experiencing a particular lifetime event within a limited time period; nevertheless, it does not consider the time when the event occurs and therefore disregards the length of the survival process. Suppose that two population groups have the same rate of experiencing a particular event by the end of an observation period but members in one group are expected to experience the event significantly later than do those in the other. The former population group has an advantaged survival pattern because its average life is extended. Obviously, the logistic regression ignores this timing factor, therefore not providing precise information.

Most survival models account for the timing factor on a status jump. Given this capacity, the second feature of survival data is the description of a time-to-event process. In the literature of survival analysis, time at the occurrence of a particular event is regarded as a random variable, referred to as event time, failure time, or survival time. Compared to statistical techniques focused on structures, the vast majority of survival models are designed to describe a time course from the beginning of a specific time interval to the occurrence of a particular event. Given this feature, data used for survival analysis are also referred to as time-to-event data, which consist of information both about a discrete 'jump' in status as well as about the time passed until the occurrence of such a jump.

The third primary feature of survival data structure is censoring. Survival data are generally collected for a time interval in which the occurrences of a particular event are observed. As a result, researchers can only observe those events that occur within a surveillance window between two time limits. Consequently, complete survival times for many units under examination are not observed, with information loss taking place either before the onset or beyond the end of the study interval. Some units may be lost to observation in the middle of an investigation due to various reasons. In survival analysis, such missing status on event times is called *censoring*, which can be divided into a variety of types. For most censoring types, a section of survival times for censored observations are observable and can be utilized in calculating the risk of experiencing a particular event. In survival analysis, this portion of observed times is referred to as censored survival times. As censoring frequently occurs, the majority of survival analysis literally deals with incomplete survival data, and accordingly scientists have found ways to use such limited information for correctly analyzing the incomplete survival data based on some restrictive assumptions on the distribution of censored survival times. Given the importance of handling censoring in survival analysis, a variety of censoring types are delineated in Section 1.4.

As survival processes essentially vary massively based on basic characteristics of the observations and environmental conditions, a considerable body of survival analysis is conducted by means of censored data regression modeling involving one or more predictor variables. Given the addition of covariates, survival data structure can be viewed as consisting of information about three primary factors, otherwise referred to as a 'triple:' survival times, censoring status, and covariates. Given a random sample of n units, the data structure for survival analysis actually contains n such triples. Most survival models, as will be described extensively in later chapters, are built upon such a data structure.

Given different emphases on the variety of features, survival analysis is also known as duration analysis, time-to-event analysis, event histories analysis, or reliability data analysis. In this book, these concepts are used interchangeably.

1.4 Censoring

Methodologically, censoring is defined as the loss of observation on the lifetime variable of interest in the process of an investigation. In survival data, censoring frequently occurs for many reasons. In a clinical trial on the effectiveness of a new medical treatment for disease, for example, patients may be lost to follow-up due to migration or health problems. In a longitudinal observational survey, some baseline respondents may lose interest in participating in subsequent investigations because some of the questions in a previous questionnaire are considered too sensitive.

Censoring is generally divided into several specific types. If an individual has entered a study but is lost to follow-up, the actual event time is placed somewhere to the right of the censored time along the time axis. This type of censoring is called *right censoring*. As right censoring occurs far more frequently than do other types and its information can be included in the estimation of a survival model, the focus of this section is on the description of right censoring. For analytic convenience, descriptions of right censoring are often based on the assumption that an individual's censored time is independent of the actual survival time, thereby making right censoring noninformative. While this assumption does not always hold, the issue of informative censoring and the related estimating approaches are described in Chapter 9. Other types of censoring, including *left censoring* and *interval censoring*, are also described in this section. Additionally, I briefly discuss the impact of left truncation on survival analysis, a type of missing data that is different from censoring.

1.4.1 Mechanisms of right censoring

Right censoring is divided into several categories: Type I censoring, random censoring, and Type II censoring. In *Type I censoring*, each observation has a fixed censoring time. Type I censoring is usually related to a predetermined observation period defined according to the research design. Generally, a specific length of time is designed with a starting calendar date and an ending date. In most cases, only a portion of observations would experience a particular event of interest during this specified study interval and some others would survive to the endpoint. For those who survive the entire observation period, the only information known to the researcher is that the actual survival time is located to the right of the endpoint of the study period along the time axis, mathematically denoted by $T > C$, where T is the event time and C is a fixed censored time. Therefore, lifetimes of those survivors are viewed as right censored, with the length of the censored time equaling the length of the observation period.

Right censoring also occurs randomly at any time during a study period, referred to as *random censoring*. This type of censoring differs essentially from Type I censoring because the censored time is not fixed, but, rather, behaves as a random variable. Some respondents may enter the study after a specified starting date and then are right censored at the end of the study interval. Such observations are also listed in the category of random censoring because their *delayed entry* is random. Statistically, time for random censoring can be described by a random variable C_i (the subscript i indicates variation in C among randomly censored observations), generally assumed to be independent of survival time T_i. Mathematically, for a sample of n observations, case i ($i = 1, 2, \ldots, n$) is considered randomly censored if $C_i < T_i$ and $C_i < C$, where C is the fixed Type I censored time. The censored survival time for random censoring is measured as the time distance from the time of entry into the study to the time when random censoring occurs.

Figure 1.1 graphically displays the occurrences of Type I and random censoring. In this figure, I present data for six individuals who participate in a study of mortality at older ages, noted by, respectively, persons 1, 2, 3, 4, 5, and 6. The study specifies an observation period from 'start of study' to 'end of study.' The sign '×' denotes the occurrence of a death, whereas the sign '+' represents right censoring.

In Figure 1.1, person 1 enters the study at the beginning of the study and dies within the interval. Therefore, this case is an event, with time-to-event T_1 counted as the time elapsed from the start of the study to the time of death. Person 2 also enters the study at the beginning of the study, but at the end of the study, this person is still alive. Therefore, person 2 is a typical case of Type I right censoring, with the censored survival time equaling the full

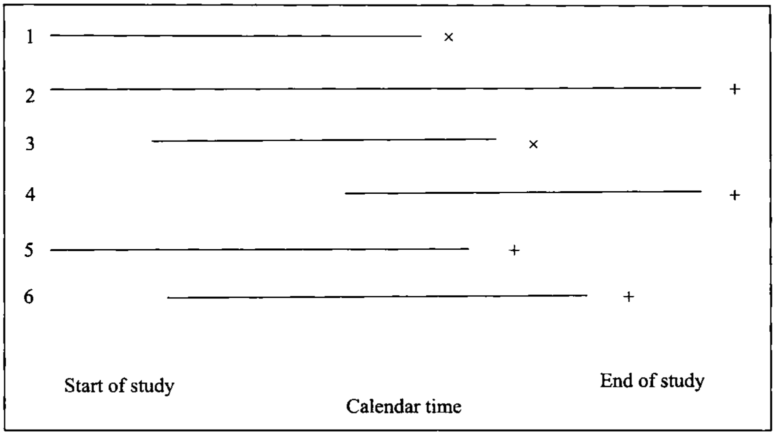

Figure 1.1 Illustration of Type I and random censoring.

length of the study interval. Persons 3 and 4 both enter the study after the start of the study, with person 3 deceased during the interval and person 4 alive throughout the rest of the interval. Consequently, person 3 has an event whose survival time is the distance from the time of the delayed entry to the time of death, whereas person 4 is a case of random censoring with the censored survival time measured as the length of time between the delayed entry and the end of the study. Entering the study later than expected, person 4 can also be considered a left truncated observation, which will be described in Subsection 1.4.2. Finally, persons 5 and 6 are lost to follow-ups before the termination of the study, with person 5 entering the investigation at the start and person 6 entering during the period of investigation. Both persons are randomly censored. Their censored times, denoted by C_5 and C_6, respectively, measured as the time elapsed between the starting date of the study and the censored time for person 5, or between the time of the delayed entry and the censored time for person 6. Unlike person 2, censored times for persons 4, 5, and 6 differ from each other and are smaller than C.

Type II right censoring refers to the situation in which a fixed number of events is targeted for a particular study. When the designed number of events is observed, a study would terminate automatically and all individuals whose survival times are beyond the time of termination are right censored. For those individuals, the censored survival time is measured as the distance from the start of observation to the time at which the study terminates. Type II right censoring is not related to a fixed ending time; rather, it is associated with a time determined by a date when a targeted number of events are observed. Given this restriction, surveys or clinical trials associated with Type II right censoring are much rarer than those with other types of right censoring.

1.4.2 Left censoring, interval censoring, and left truncation

Left censoring refers to a data point, known to be prior to a certain date but unknown about its exact location. This type of censoring frequently occurs in a study design involving two separate study stages. Individuals who enroll in the first selection process but are not eligible for the second process are viewed as left censored. For example, in a study of the initiation of first contraceptive use after marriage, if a couple marries but has already used