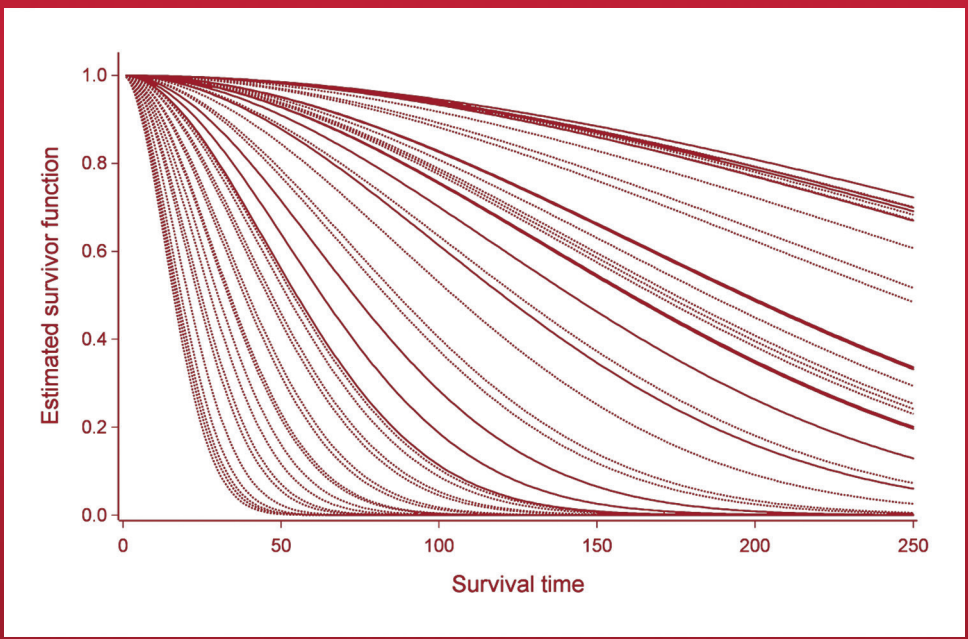
Texts in Statistical Science

Modelling Survival Data in Medical Research

Third Edition



David Collett

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Modelling Survival Data in Medical Research

Third Edition

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Statistical Theory: A Concise Introduction
F. Abramovich and Y. Ritov

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

Interpreting Data: A First Course in Statistics
A.J.B. Anderson

Introduction to Probability with R
K. Baclawski

Linear Algebra and Matrix Analysis for Statistics
S. Banerjee and A. Roy

Analysis of Categorical Data with R
C. R. Bilder and T. M. Loughin

Statistical Methods for SPC and TQM
D. Bissell

Introduction to Probability
J. K. Blitzstein and J. Hwang

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Second Edition
R. Caulcutt

The Analysis of Time Series: An Introduction, Sixth Edition
C. Chatfield

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Problem Solving: A Statistician's Guide, Second Edition
C. Chatfield

Statistics for Technology: A Course in Applied Statistics, Third Edition
C. Chatfield

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Third Edition
D. Collett

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Applied Statistics: Principles and Examples
D.R. Cox and E.J. Snell

Multivariate Survival Analysis and Competing Risks
M. Crowder

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Nonlinear Time Series: Theory, Methods, and Applications with R Examples
R. Douc, E. Moulines, and D.S. Stoffer

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

Linear Models with R, Second Edition
J.J. Faraway

A Course in Large Sample Theory
T.S. Ferguson

Multivariate Statistics: A Practical Approach
B. Flury and H. Riedwyl

Readings in Decision Analysis
S. French

Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Bayesian Data Analysis, Third Edition
A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin

Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Logistic Regression Models
J.M. Hilbe

Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects
J.S. Hodges

Statistics for Epidemiology
N.P. Jewell

Stochastic Processes: An Introduction, Second Edition
P.W. Jones and P. Smith

The Theory of Linear Models
B. Jørgensen

Principles of Uncertainty
J.B. Kadane

Graphics for Statistics and Data Analysis with R
K.J. Keen

Mathematical Statistics
K. Knight

Introduction to Multivariate Analysis: Linear and Nonlinear Modeling
S. Konishi

Nonparametric Methods in Statistics with SAS Applications
O. Korosteleva

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Statistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Design and Analysis of Experiments with R
J. Lawson

Design and Analysis of Experiments with SAS
J. Lawson

A Course in Categorical Data Analysis
T. Leonard

Statistics for Accountants
S. Letchford

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Statistical Theory, Fourth Edition
B.W. Lindgren

Stationary Stochastic Processes: Theory and Applications
G. Lindgren

The BUGS Book: A Practical Introduction to Bayesian Analysis
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

Time Series Analysis
H. Madsen

Pólya Urn Models
H. Mahmoud

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistics in Engineering: A Practical Approach
A.V. Metcalfe

Statistical Inference: An Integrated Approach, Second Edition
H. S. Migon, D. Gamerman, and F. Louzada

Beyond ANOVA: Basics of Applied Statistics
R.G. Miller, Jr.

A Primer on Linear Models
J.F. Monahan

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Elements of Simulation
B.J.T. Morgan

Probability: Methods and Measurement
A. O'Hagan

Introduction to Statistical Limit Theory
A.M. Polansky

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West, and J. Harrison

Statistics in Research and Development, Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Introduction to Statistical Process Control
P. Qiu

Sampling Methodologies with Applications
P.S.R.S. Rao

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Essential Statistics, Fourth Edition
D.A.G. Rees

Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists
F.J. Samaniego

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Bayesian Networks: With Examples in R
M. Scutari and J.-B. Denis

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Decision Analysis: A Bayesian Approach
J.Q. Smith

Analysis of Failure and Survival Data
P. J. Smith

Applied Statistics: Handbook of GENSTAT Analyses
E.J. Snell and H. Simpson

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Data Driven Statistical Methods
P. Sprent

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications
W. W. Stroup

Survival Analysis Using S: Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Understanding Advanced Statistical Methods
P.H. Westfall and K.S.S. Henning

Statistical Process Control: Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Generalized Additive Models: An Introduction with R
S. Wood

Epidemiology: Study Design and Data Analysis, Third Edition
M. Woodward

Practical Data Analysis for Designed Experiments
B.S. Yandell

Texts in Statistical Science

Modelling Survival Data in Medical Research

Third Edition

David Collett
NHS Blood and Transplant
Bristol, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published in 1994 by Chapman and Hall.

Second edition published in 2003 by Chapman and Hall/CRC.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150505

International Standard Book Number-13: 978-1-4987-3169-0 (eBook - PDF)

Contents

Preface **xv**

1 Survival analysis **1**
1.1 Special features of survival data 1
1.1.1 Censoring 2
1.1.2 Independent censoring 3
1.1.3 Study time and patient time 3
1.2 Some examples 5
1.3 Survivor, hazard and cumulative hazard functions 10
1.3.1 The survivor function 10
1.3.2 The hazard function 12
1.3.3 The cumulative hazard function 13
1.4 Computer software for survival analysis 14
1.5 Further reading 15

2 Some non-parametric procedures **17**
2.1 Estimating the survivor function 17
2.1.1 Life-table estimate of the survivor function 19
2.1.2 Kaplan-Meier estimate of the survivor function 21
2.1.3 Nelson-Aalen estimate of the survivor function 24
2.2 Standard error of the estimated survivor function 25
2.2.1 Standard error of the Kaplan-Meier estimate 26
2.2.2 Standard error of other estimates 27
2.2.3 Confidence intervals for values of the survivor function 28
2.3 Estimating the hazard function 31
2.3.1 Life-table estimate of the hazard function 31
2.3.2 Kaplan-Meier type estimate 32
2.3.3 Estimating the cumulative hazard function 35
2.4 Estimating the median and percentiles of survival times 36
2.5 Confidence intervals for the median and percentiles 38
2.6 Comparison of two groups of survival data 40
2.6.1 Hypothesis testing 41
2.6.2 The log-rank test 44
2.6.3 The Wilcoxon test 48
2.6.4 Comparison of the log-rank and Wilcoxon tests 49

2.7 Comparison of three or more groups of survival data 50
2.8 Stratified tests 52
2.9 Log-rank test for trend 54
2.10 Further reading 56

3 The Cox regression model 57
3.1 Modelling the hazard function 57
3.1.1 A model for the comparison of two groups 58
3.1.2 The general proportional hazards model 59
3.2 The linear component of the model 60
3.2.1 Including a variate 61
3.2.2 Including a factor 61
3.2.3 Including an interaction 62
3.2.4 Including a mixed term 63
3.3 Fitting the Cox regression model 65
3.3.1 Likelihood function for the model 67
3.3.2 Treatment of ties 69
3.3.3 The Newton-Raphson procedure 71
3.4 Confidence intervals and hypothesis tests 72
3.4.1 Confidence intervals for hazard ratios 73
3.4.2 Two examples 73
3.5 Comparing alternative models 76
3.5.1 The statistic $-2\log\hat{L}$ 77
3.5.2 Comparing nested models 78
3.6 Strategy for model selection 83
3.6.1 Variable selection procedures 84
3.7 Variable selection using the lasso 90
3.7.1 The lasso in Cox regression modelling 91
3.7.2 Data preparation 92
3.8 Non-linear terms 95
3.8.1 Testing for non-linearity 96
3.8.2 Modelling non-linearity 97
3.8.3 Fractional polynomials 98
3.9 Interpretation of parameter estimates 99
3.9.1 Models with a variate 99
3.9.2 Models with a factor 100
3.9.3 Models with combinations of terms 104
3.10 Estimating the hazard and survivor functions 107
3.10.1 The special case of no covariates 110
3.10.2 Some approximations to estimates of baseline functions 110
3.11 Risk adjusted survivor function 116
3.11.1 Risk adjusted survivor function for groups of individuals 117
3.12 Explained variation in the Cox regression model 120
3.12.1 Measures of explained variation 122
3.12.2 Measures of predictive ability 123

3.12.3 Model validation 124
3.13 Proportional hazards and the log-rank test 125
3.14 Further reading 128

4 Model checking in the Cox regression model 131
4.1 Residuals for the Cox regression model 131
4.1.1 Cox-Snell residuals 132
4.1.2 Modified Cox-Snell residuals 133
4.1.3 Martingale residuals 135
4.1.4 Deviance residuals 136
4.1.5 Schoenfeld residuals 137
4.1.6 Score residuals 138
4.2 Assessment of model fit 142
4.2.1 Plots based on the Cox-Snell residuals 142
4.2.2 Plots based on the martingale and deviance residuals 145
4.2.3 Checking the functional form of covariates 147
4.3 Identification of influential observations 152
4.3.1 Influence of observations on a parameter estimate 153
4.3.2 Influence of observations on the set of parameter estimates 155
4.3.3 Treatment of influential observations 158
4.4 Testing the assumption of proportional hazards 160
4.4.1 The log-cumulative hazard plot 161
4.4.2 Use of Schoenfeld residuals 163
4.4.3 Tests for non-proportional hazards 164
4.4.4 Adding a time-dependent variable 166
4.5 Recommendations 168
4.6 Further reading 169

5 Parametric proportional hazards models 171
5.1 Models for the hazard function 171
5.1.1 The exponential distribution 172
5.1.2 The Weibull distribution 173
5.2 Assessing the suitability of a parametric model 177
5.3 Fitting a parametric model to a single sample 178
5.3.1 Likelihood function for randomly censored data 180
5.4 Fitting exponential and Weibull models 181
5.4.1 Fitting the exponential distribution 182
5.4.2 Fitting the Weibull distribution 186
5.4.3 Standard error of a percentile of the Weibull distribution 188
5.5 A model for the comparison of two groups 192
5.5.1 The log-cumulative hazard plot 192
5.5.2 Fitting the model 194
5.6 The Weibull proportional hazards model 199
5.6.1 Fitting the model 200

5.6.2 Standard error of a percentile in the Weibull model 201
5.6.3 Log-linear form of the model 203
5.6.4 Exploratory analyses 205
5.7 Comparing alternative Weibull models 208
5.8 Explained variation in the Weibull model 215
5.9 The Gompertz proportional hazards model 216
5.10 Model choice 218
5.11 Further reading 219

6 Accelerated failure time and other parametric models 221
6.1 Probability distributions for survival data 221
6.1.1 The log-logistic distribution 222
6.1.2 The lognormal distribution 222
6.1.3 The gamma distribution 224
6.1.4 The inverse Gaussian distribution 225
6.2 Exploratory analyses 225
6.3 Accelerated failure model for two groups 227
6.3.1 Comparison with the proportional hazards model 228
6.3.2 The percentile-percentile plot 231
6.4 The general accelerated failure time model 232
6.4.1 Log-linear form of the accelerated failure time model 234
6.5 Parametric accelerated failure time models 236
6.5.1 The Weibull accelerated failure time model 236
6.5.2 The log-logistic accelerated failure time model 239
6.5.3 The lognormal accelerated failure time model 240
6.5.4 Summary 241
6.6 Fitting and comparing accelerated failure time models 243
6.7 The proportional odds model 250
6.7.1 The log-logistic proportional odds model 253
6.8 Some other distributions for survival data 255
6.9 Flexible parametric models 256
6.9.1 The Royston and Parmar model 259
6.9.2 Number and position of the knots 262
6.9.3 Fitting the model 262
6.9.4 Proportional odds models 266
6.10 Modelling cure rates 268
6.11 Effect of covariate adjustment 270
6.12 Further reading 272

7 Model checking in parametric models 275
7.1 Residuals for parametric models 275
7.1.1 Standardised residuals 275
7.1.2 Cox-Snell residuals 276
7.1.3 Martingale residuals 277
7.1.4 Deviance residuals 277

7.1.5 Score residuals 277
7.2 Residuals for particular parametric models 278
7.2.1 Weibull distribution 279
7.2.2 Log-logistic distribution 279
7.2.3 Lognormal distribution 280
7.2.4 Analysis of residuals 280
7.3 Comparing observed and fitted survivor functions 284
7.4 Identification of influential observations 287
7.4.1 Influence of observations on a parameter estimate 287
7.4.2 Influence of observations on the set of parameter estimates 288
7.5 Testing proportional hazards in the Weibull model 291
7.6 Further reading 292

8 Time-dependent variables 295
8.1 Types of time-dependent variables 295
8.2 A model with time-dependent variables 296
8.2.1 Fitting the Cox model 297
8.2.2 Estimation of baseline hazard and survivor functions 300
8.3 Model comparison and validation 302
8.3.1 Comparison of treatments 303
8.3.2 Assessing model adequacy 303
8.4 Some applications of time-dependent variables 304
8.5 Three examples 306
8.6 Counting process format 316
8.7 Further reading 317

9 Interval-censored survival data 319
9.1 Modelling interval-censored survival data 319
9.2 Modelling the recurrence probability in the follow-up period 322
9.3 Modelling the recurrence probability at different times 325
9.4 Arbitrarily interval-censored survival data 332
9.4.1 Modelling arbitrarily interval-censored data 332
9.4.2 Proportional hazards model for the survivor function 334
9.4.3 Choice of the step times 337
9.5 Parametric models for interval-censored data 342
9.6 Discussion 343
9.7 Further reading 344

10 Frailty models 345
10.1 Introduction to frailty 345
10.1.1 Random effects 346
10.1.2 Individual frailty 346
10.1.3 Shared frailty 347
10.2 Modelling individual frailty 348

10.2.1 Frailty distributions 349
10.2.2 Observable survivor and hazard functions 351
10.3 The gamma frailty distribution 352
10.3.1 Impact of frailty on an observable hazard function 353
10.3.2 Impact of frailty on an observable hazard ratio 354
10.4 Fitting parametric frailty models 356
10.4.1 Gamma frailty 357
10.5 Fitting semi-parametric frailty models 363
10.5.1 Lognormal frailty effects 363
10.5.2 Gamma frailty effects 365
10.6 Comparing models with frailty 366
10.6.1 Testing for the presence of frailty 366
10.7 The shared frailty model 372
10.7.1 Fitting the shared frailty model 373
10.7.2 Comparing shared frailty models 374
10.8 Some other aspects of frailty modelling 377
10.8.1 Model checking 377
10.8.2 Correlated frailty models 378
10.8.3 Dependence measures 378
10.8.4 Numerical problems in model fitting 378
10.9 Further reading 379

11 Non-proportional hazards and institutional comparisons 381
11.1 Non-proportional hazards 381
11.2 Stratified proportional hazards models 383
11.2.1 Non-proportional hazards between treatments 385
11.3 Restricted mean survival 389
11.3.1 Use of pseudo-values 391
11.4 Institutional comparisons 393
11.4.1 Interval estimate for the *RAFR* 397
11.4.2 Use of the Poisson regression model 400
11.4.3 Random institution effects 402
11.5 Further reading 403

12 Competing risks 405
12.1 Introduction to competing risks 405
12.2 Summarising competing risks data 406
12.2.1 Kaplan-Meier estimate of survivor function 407
12.3 Hazard and cumulative incidence functions 409
12.3.1 Cause-specific hazard function 409
12.3.2 Cause-specific cumulative incidence function 410
12.3.3 Some other functions of interest 413
12.4 Modelling cause-specific hazards 414
12.4.1 Likelihood functions for competing risks models 415
12.4.2 Parametric models for cumulative incidence functions 418

12.5 Modelling cause-specific incidence 419
12.5.1 The Fine and Gray competing risks model 419
12.6 Model checking 422
12.7 Further reading 428

13 Multiple events and event history modelling 429
13.1 Introduction to counting processes 429
13.1.1 Modelling the intensity function 430
13.1.2 Survival data as a counting process 431
13.1.3 Survival data in the counting process format 433
13.1.4 Robust estimation of the variance-covariance matrix 434
13.2 Modelling recurrent event data 435
13.2.1 The Anderson and Gill model 436
13.2.2 The Prentice, Williams and Peterson model 437
13.3 Multiple events 443
13.3.1 The Wei, Lin and Weissfeld model 443
13.4 Event history analysis 448
13.4.1 Models for event history analysis 449
13.5 Further reading 455

14 Dependent censoring 457
14.1 Identifying dependent censoring 457
14.2 Sensitivity to dependent censoring 458
14.2.1 A sensitivity analysis 459
14.2.2 Impact of dependent censoring 461
14.3 Modelling with dependent censoring 463
14.3.1 Cox regression model with dependent censoring 464
14.4 Further reading 470

15 Sample size requirements for a survival study 471
15.1 Distinguishing between two treatment groups 471
15.2 Calculating the required number of deaths 472
15.2.1 Derivation of the required number of deaths 474
15.3 Calculating the required number of patients 479
15.3.1 Derivation of the required number of patients 480
15.3.2 An approximate procedure 483
15.4 Further reading 484

A Maximum likelihood estimation 487
A.1 Inference about a single unknown parameter 487
A.2 Inference about a vector of unknown parameters 489

B Additional data sets **491**
B.1 Chronic active hepatitis 491
B.2 Recurrence of bladder cancer 492
B.3 Survival of black ducks 492
B.4 Bone marrow transplantation 495
B.5 Chronic granulomatous disease 495

Bibliography **499**

Index of Examples **521**

Preface

This book describes and illustrates the modelling approach to the analysis of survival data, using a wide range of examples from biomedical research. My experience in presenting many lectures and courses on this subject, at both introductory and advanced levels, as well as in providing advice on the analysis of survival data, has had a big influence on its content. The result is a comprehensive practical account of survival analysis at an intermediate level, which I hope will continue to meet the needs of statisticians in the pharmaceutical industry or medical research institutes, scientists and clinicians who are analysing their own data, and students following undergraduate or postgraduate courses in survival analysis.

In preparing this new edition, my aim has been to incorporate extensions to the basic models that dramatically increase their scope, while updating the text to take account of the wider availability of computer software for implementing these techniques. This edition therefore contains new chapters covering frailty models, non-proportional hazards, competing risks, multiple events, event history analysis and dependent censoring. Additional material on variable selection, non-linear models, measures of explained variation and flexible parametric models has also been included in earlier chapters.

The main part of the book is formed by Chapters 1 to 7. After an introduction to survival analysis in Chapter 1, Chapter 2 describes methods for summarising survival data, and for comparing two or more groups of survival times. The modelling approach is introduced in Chapter 3, where the Cox regression model is presented in detail. This is followed by a chapter that describes methods for checking the adequacy of a fitted model. Parametric proportional hazards models are covered in Chapter 5, with an emphasis on the Weibull model for survival data. Chapter 6 describes parametric accelerated failure time models, including a detailed account of their log-linear representation that is used in most computer software packages. Flexible parametric models are also described and illustrated in this chapter, while model-checking diagnostics for parametric models are presented in Chapter 7.

The remaining chapters describe a number of extensions to the basic models. The use of time-dependent variables is covered in Chapter 8, and the analysis of interval-censored data is considered in Chapter 9. Frailty models that allow differences between individuals, or groups of individuals, to be modelled using random effects, are described in Chapter 10. Chapter 11 summarises techniques that can be used when the assumption of proportional

hazards cannot be made, and shows how these models can be used in comparing survival outcomes across a number of institutions. Competing risk models that accommodate different causes of death are presented in Chapter 12, while extensions of the Cox regression model to cope with multiple events of the same or different types, including event history analysis, are described in Chapter 13. Chapter 14 summarises methods for analysing data when there is dependent censoring, and Chapter 15 shows how to determine the sample size requirements of a study where the outcome variable is a survival time.

All of the techniques that have been described can be implemented in many software packages for survival analysis, including the freeware package R. However, sufficient methodological details have been included to convey a sound understanding of the techniques and the assumptions on which they are based, and to help in adapting the methodology to deal with non-standard problems. Some examples in the earlier chapters are based on fewer observations than would normally be encountered in medical research programmes. This enables the methods of analysis to be illustrated more easily, as well as allowing tabular presentations of the results to be compared with output obtained from computer software. Some additional data sets that may be used to obtain a fuller appreciation of the methodology, or as student exercises, are given in an Appendix. All of the data sets used in this book are available in electronic form from the publisher's web site at `http://www.crcpress.com/`.

In writing this book, I have assumed that the reader has a basic knowledge of statistical methods, and has some familiarity with linear regression analysis. Matrix algebra is used on occasions, but an understanding of linear algebra is not an essential requirement. Bibliographic notes and suggestions for further reading are given at the end of each chapter, but so as not to interrupt the flow, references in the text itself have been kept to a minimum. Some sections contain more mathematical details than others, and these have been denoted with an asterisk. These sections can be omitted without loss of continuity.

I am indebted to Doug Altman, Alan Kimber, Mike Patefield, Anne Whitehead and John Whitehead for their help in the preparation of the current and earlier editions of the book, and to NHS Blood and Transplant for permission to use data from the UK Transplant Registry in a number of the examples. I also thank James Gallagher and staff of the Statistical Services Centre, University of Reading, and my colleagues in the Statistics and Clinical Studies section of NHS Blood and Transplant, for giving me the opportunity to rehearse the new material through courses and seminars. I am particularly grateful to all those who took the trouble to let me know about errors in earlier editions. Although these have been corrected, I would be very pleased to be informed (`d.collett@btinternet.com`) of any further errors, ambiguities and omissions in this edition. Finally, I would like to thank my wife Janet for her support and encouragement over the period that this book was written.

David Collett
September, 2014

Chapter 1

Survival analysis

Survival analysis is the phrase used to describe the analysis of data in the form of times from a well-defined *time origin* until the occurrence of some particular event or *end-point*. In medical research, the time origin will often correspond to the recruitment of an individual into an experimental study, such as a clinical trial to compare two or more treatments. This in turn may coincide with the diagnosis of a particular condition, the commencement of a treatment regimen or the occurrence of some adverse event. If the end-point is the death of a patient, the resulting data are literally survival times. However, data of a similar form can be obtained when the end-point is not fatal, such as the relief of pain, or the recurrence of symptoms. In this case, the observations are often referred to as *time to event* data, and the methods for analysing survival data that are presented in this book apply equally to data on the time to these end-points. The methods can also be used in the analysis of data from other application areas, such as the survival times of animals in an experimental study, the time taken by an individual to complete a task in a psychological experiment, the storage times of seeds held in a seed bank or the lifetimes of industrial or electronic components. The focus of this book is on the application of survival analysis to data arising from medical research, and for this reason much of the general discussion will be phrased in terms of the survival time of an individual patient from entry to a study until death.

1.1 Special features of survival data

We must first consider the reasons why survival data are not amenable to standard statistical procedures used in data analysis. One reason is that survival data are generally not symmetrically distributed. Typically, a histogram constructed from the survival times of a group of similar individuals will tend to be *positively skewed*, that is, the histogram will have a longer 'tail' to the right of the interval that contains the largest number of observations. As a consequence, it will not be reasonable to assume that data of this type have a normal distribution. This difficulty could be resolved by first transforming the data to give a more symmetric distribution, for example by taking logarithms. However, a more satisfactory approach is to adopt an alternative distributional model for the original data.

The main feature of survival data that renders standard methods inappropriate is that survival times are frequently *censored*. Censoring is described in the next section.

1.1.1 Censoring

The survival time of an individual is said to be censored when the end-point of interest has not been observed for that individual. This may be because the data from a study are to be analysed at a point in time when some individuals are still alive. Alternatively, the survival status of an individual at the time of the analysis might not be known because that individual has been *lost to follow-up*. As an example, suppose that after being recruited to a clinical trial, a patient moves to another part of the country, or to a different country, and can no longer be traced. The only information available on the survival experience of that patient is the last date on which he or she was known to be alive. This date may well be the last time that the patient reported to a clinic for a regular check-up.

An actual survival time can also be regarded as censored when death is from a cause that is known to be unrelated to the treatment. However, it can be difficult to be sure that the death is not related to a particular treatment that the patient is receiving. For example, consider a patient in a clinical trial to compare alternative therapies for prostatic cancer who experiences a fatal road traffic accident. The accident could have resulted from an attack of dizziness, which might be a side effect of the treatment to which that patient has been assigned. If so, the death is not unrelated to the treatment. In circumstances such as these, the survival time until death from all causes, or the time to death from causes other than the primary condition for which the patient is being treated, might also be subjected to a survival analysis.

In each of these situations, a patient who entered a study at time t_0 dies at time $t_0 + t$. However, t is unknown, either because the individual is still alive or because he or she has been lost to follow-up. If the individual was last known to be alive at time $t_0 + c$, the time c is called a censored survival time. This censoring occurs after the individual has been entered into a study, that is, to the right of the last known survival time, and is therefore known as *right censoring*. The right-censored survival time is then less than the actual, but unknown, survival time. Right censoring that occurs when the observation period of a study ends is often termed *administrative censoring*.

Another form of censoring is *left censoring*, which is encountered when the actual survival time of an individual is less than that observed. To illustrate this form of censoring, consider a study in which interest centres on the time to recurrence of a particular cancer following surgical removal of the primary tumour. Three months after their operation, the patients are examined to determine if the cancer has recurred. At this time, some of the patients may be found to have a recurrence. For such patients, the actual time to recurrence is less than three months, and the recurrence times of these patients is left-

censored. Left censoring occurs far less commonly than right censoring, and so the emphasis of this book will be on the analysis of right-censored survival data.

Yet another type of censoring is *interval censoring*. Here, individuals are known to have experienced an event within an interval of time. Consider again the example concerning the time to recurrence of a tumour used in the above discussion of left censoring. If a patient is observed to be free of the disease at three months, but is found to have had a recurrence when examined six months after surgery, the actual recurrence time of that patient is known to be between three months and six months. The observed recurrence time is then said to be interval-censored. We will return to interval censoring later, in Chapter 9.

1.1.2 Independent censoring

An important assumption that will be made in the analysis of censored survival data is that the actual survival time of an individual, t, does not depend on any mechanism that causes that individual's survival time to be censored at time c, where $c < t$. Such censoring is termed *independent* or *non-informative censoring*. This means that if we consider a group of individuals who all have the same values of relevant prognostic variables, an individual whose survival time is censored at time c must be representative of all other individuals in that group who have survived to that time. A patient whose survival time is censored will be representative of those at risk at the censoring time if the censoring process operates randomly. Similarly, when survival data are to be analysed at a predetermined point in calendar time, or at a fixed interval of time after the time origin for each patient, the prognosis for individuals who are still alive can be taken to be independent of the censoring, so long as the time of analysis is specified before the data are examined. However, this assumption cannot be made if, for example, the survival time of an individual is censored through treatment being withdrawn as a result of a deterioration in their physical condition. This type of censoring is known as *dependent* or *informative censoring*. The methods of survival analysis presented in most chapters of this book are only valid under the assumption of independent censoring, but techniques that enable account to be taken of dependent censoring will be described in Chapter 14.

1.1.3 Study time and patient time

In a typical study, patients are not all recruited at exactly the same time, but accrue over a period of months or even years. After recruitment, patients are followed up until they die, or until a point in calendar time that marks the end of the study, when the data are analysed. Although the actual survival times will be observed for a number of patients, after recruitment some patients may be lost to follow-up, while others will still be alive at the end of the study.

The calendar time period in which an individual is in the study is known as the *study time*.

The study time for eight individuals in a clinical trial is illustrated diagrammatically in Figure 1.1, in which the time of entry to the study is represented by a '•'.

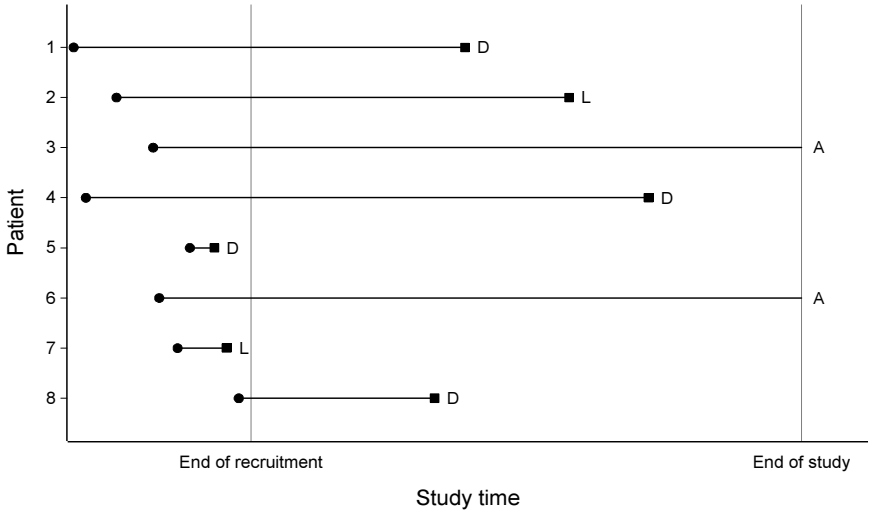


Figure 1.1 *Study time for eight patients in a survival study.*

This figure shows that individuals 1, 4, 5 and 8 die (D) during the course of the study, individuals 2 and 7 are lost to follow-up (L), and individuals 3 and 6 are still alive (A) at the end of the observation period.

As far as each patient is concerned, the trial begins at some time t_0. The corresponding survival times for the eight individuals depicted in Figure 1.1 are shown in order in Figure 1.2. The period of time that a patient spends in the study, measured from that patient's time origin, is often referred to as *patient time*. The period of time from the time origin to the death of a patient (D) is then the survival time, and this is recorded for individuals 1, 4, 5 and 8. The survival times of the remaining individuals are right-censored (C).

In practice, the actual data recorded will be the date on which each individual enters the study, and the date on which each individual dies or was last known to be alive. The survival time in days, weeks or months, whichever is the most appropriate, can then be calculated. Most computer software packages for survival analysis have facilities for performing this calculation from input data in the form of dates.

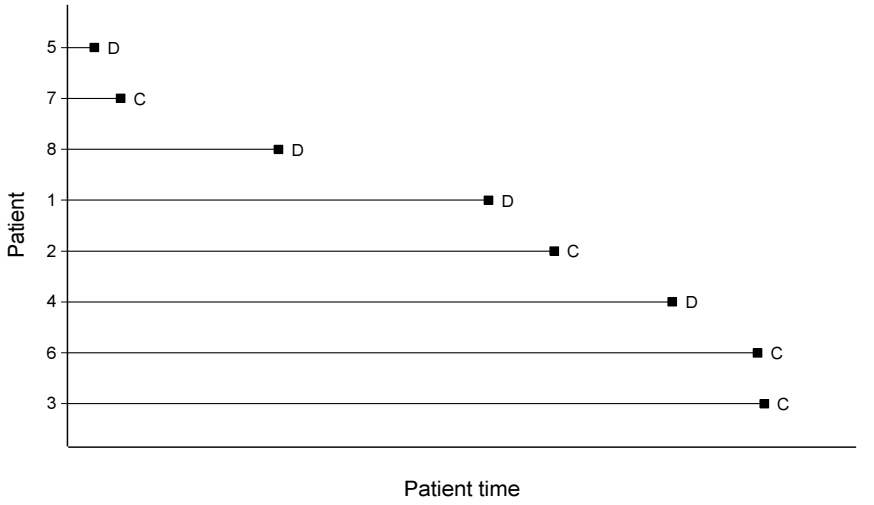


Figure 1.2 *Patient time for eight patients in a survival study.*

1.2 Some examples

In this section, the essential features of survival data are illustrated through a number of examples. Data from these examples will then be used to illustrate some of the statistical techniques presented in subsequent chapters.

Example 1.1 Time to discontinuation of the use of an IUD
In trials involving contraceptives, prevention of pregnancy is an obvious criterion for acceptability. However, modern contraceptives have very low failure rates, and so the occurrence of bleeding disturbances, such as amenorrhoea (the prolonged absence of bleeding), irregular or prolonged bleeding, become important in the evaluation of a particular method of contraception. To promote research into methods for analysing menstrual bleeding data from women in contraceptive trials, the World Health Organisation made available data from clinical trials involving a number of different types of contraceptive (WHO, 1987). Part of this data set relates to the time from which a woman commences use of a particular method until discontinuation, with the discontinuation reason being recorded when known. The data in Table 1.1 refer to the number of weeks from the commencement of use of a particular type of intrauterine device (IUD), known as the Multiload 250, until discontinuation because of menstrual bleeding problems. Data are given for 18 women, all of whom were aged between 18 and 35 years and who had experienced two previous pregnancies. Discontinuation times that are censored are labelled with an asterisk.

In this example, the time origin corresponds to the first day in which a woman uses the IUD, and the end-point is discontinuation because of bleed-

Table 1.1 *Time in weeks to discontinuation of the use of an IUD.*

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

* Censored discontinuation times.

ing problems. Some women in the study ceased using the IUD because of the desire for pregnancy, or because they had no further need for a contraceptive, while others were simply lost to follow-up. These reasons account for the censored discontinuation times of 13, 18, 23, 38, 54 and 56 weeks. The study protocol called for the menstrual bleeding experience of each woman to be documented for a period of two years from the time origin. For practical reasons, each woman could not be examined exactly two years after recruitment to determine if they were still using the IUD, and this is why there are three discontinuation times greater than 104 weeks that are right-censored.

One objective in an analysis of these data would be to summarise the distribution of discontinuation times. We might then wish to estimate the median time to discontinuation of the IUD, or the probability that a woman will stop using the device after a given period of time. Indeed, a graph of this estimated probability, as a function of time, will provide a useful summary of the observed data.

Example 1.2 Prognosis for women with breast cancer
Breast cancer is one of the most common forms of cancer occurring in women living in the Western world. However, the biological behaviour of the tumour is often unpredictable, and a number of studies have focussed on whether the tumour is likely to have metastasised, or spread, to other organs in the body. Around 80% of women presenting with primary breast cancer are likely to have tumours that have already metastasised to other sites. If these patients could be identified, adjunctive treatment could be focussed on them, while the remaining 20% could be reassured that their disease is surgically curable.

The aim of an investigation carried out at the Middlesex Hospital, documented in Leathem and Brooks (1987), was to evaluate a histochemical marker that discriminates between primary breast cancer that has metastasised and that which has not. The marker under study was a lectin from the albumin gland of the Roman snail, *Helix pomatia*, known as *Helix pomatia* agglutinin, or HPA. The marker binds to those breast cancer cells associated with metastasis to local lymph nodes, and the HPA stained cells can be identified by microscopic examination. In order to investigate whether HPA staining can be used to predict the survival experience of women who present with breast cancer, a retrospective study was carried out, based on the records of women who had received surgical treatment for breast cancer. Sections of the tumours of these women were treated with HPA and each tumour was subsequently classified as being positively or negatively stained, positive staining corresponding to a tumour with the potential for metastasis. The study was concluded in July

1987, when the survival times of those women who had died of breast cancer were calculated. For those women whose survival status in July 1987 was unknown, the time from surgery to the date on which they were last known to be alive is regarded as a censored survival time. The survival times of women who had died from causes other than breast cancer are also regarded as right-censored. The data given in Table 1.2 refer to the survival times in months of women who had received a simple or radical mastectomy to treat a tumour of Grade II, III or IV, between January 1969 and December 1971. In the table, the survival times of each woman are classified according to whether their tumour was positively or negatively stained.

Table 1.2 *Survival times of women with tumours that were negatively or positively stained with HPA.*

Negative staining	Positive staining	
23	5	68
47	8	71
69	10	76*
70*	13	105*
71*	18	107*
100*	24	109*
101*	26	113
148	26	116*
181	31	118
198*	35	143
208*	40	154*
212*	41	162*
224*	48	188*
	50	212*
	59	217*
	61	225*

* Censored survival times.

In the analysis of the data from this study, we will be particularly interested in whether or not there is a difference in the survival experience of the two groups of women. If there were evidence that those women with negative HPA staining tended to live longer after surgery than those with positive staining, we would conclude that the prognosis for a breast cancer patient was dependent on the result of the staining procedure.

Example 1.3 Survival of multiple myeloma patients
Multiple myeloma is a malignant disease characterised by the accumulation of abnormal plasma cells, a type of white blood cell, in the bone marrow. The proliferation of the abnormal plasma cells within the bone causes pain and the destruction of bone tissue. Patients with multiple myeloma also experience anaemia, haemorrhages, recurrent infections and weakness. Unless treated, the condition is invariably fatal. The aim of a study carried out at the Medical Center of the University of West Virginia, USA, was to examine

the association between the values of certain *explanatory variables* or *covariates* and the survival time of patients. In the study, the primary response variable was the time, in months, from diagnosis until death from multiple myeloma.

The data in Table 1.3, which were obtained from Krall, Uthoff and Harley (1975), relate to 48 patients, all of whom were aged between 50 and 80 years. Some of these patients had not died by the time that the study was completed, and so these individuals contribute right-censored survival times. The coding of the survival status of an individual in the table is such that zero denotes a censored observation and unity death from multiple myeloma.

At the time of diagnosis, the values of a number of explanatory variables were recorded for each patient. These included the age of the patient in years, their sex (1 = male, 2 = female), the levels of blood urea nitrogen (*Bun*), serum calcium (*Ca*) and haemoglobin (*Hb*), the percentage of plasma cells in the bone marrow (*Pcells*) and an indicator variable (*Protein*) that denotes whether or not the Bence-Jones protein was present in the urine (0 = absent, 1 = present).

The main aim of an analysis of these data would be to investigate the effect of the risk factors *Bun*, *Ca*, *Hb*, *Pcells* and *Protein* on the survival time of the multiple myeloma patients. The effects of these risk factors may be modified by the age or sex of a patient, and so the extent to which the relationship between survival and the important risk factors is consistent for each sex and for each of a number of age groups will also need to be studied.

Example 1.4 Comparison of two treatments for prostatic cancer
A randomised controlled clinical trial to compare treatments for prostatic cancer was begun in 1967 by the Veteran's Administration Cooperative Urological Research Group. The trial was double-blind and two of the treatments used in the study were a placebo and 1.0 mg of diethylstilbestrol (DES). The treatments were administered daily by mouth. The time origin of the study is the date on which a patient was randomised to a treatment, and the end-point is the death of the patient from prostatic cancer.

The full data set is given in Andrews and Herzberg (1985), but the data used in this example are from patients presenting with Stage III cancer, that is, patients for whom there was evidence of a local extension of the tumour beyond the prostatic capsule, but without elevated serum prostatic acid phosphatase. Furthermore, the patients were those who had no history of cardiovascular disease, had a normal ECG result at trial entry, and who were not confined to bed during the daytime. In addition to recording the survival time of each patient in the study, information was recorded on a number of other prognostic factors. These included the age of the patient at trial entry, their serum haemoglobin level in gm/100 ml, the size of their primary tumour in cm^2 and the value of a combined index of tumour stage and grade. This index is known as the Gleason index; the more advanced the tumour, the greater the value of the index.

Table 1.3 *Survival times of patients in a study on multiple myeloma.*

Patient number	Survival time	Status	Age	Sex	*Bun*	*Ca*	*Hb*	*Pcells*	*Protein*
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12.0	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14.0	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13.0	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14.0	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14.0	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10.0	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5.0	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11.0	16	1
34	5	1	77	1	23	8	9.0	29	0
35	10	1	61	1	13	10	14.0	19	0
36	91	1	58	2	27	11	11.0	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13.0	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	17	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

Table 1.4 gives the data recorded for 38 patients, where the survival times are given in months. The survival times of patients who died from other causes, or who were lost during the follow-up process, are regarded as censored. A variable associated with the status of an individual at the end of the study takes the value unity if the patient has died from prostatic cancer, and zero if the survival time is right-censored. The variable associated with the treatment group takes the value 2 when an individual is treated with DES and unity if an individual is on the placebo treatment.

The main aim of this study is to determine the extent of any evidence that patients treated with DES survive longer than those treated with the placebo. Since the data on which this example is based are from a randomised trial, one might expect that the distributions of the prognostic factors, that is the age of patient, serum haemoglobin level, size of tumour and Gleason index, will be similar over the patients in each of the two treatment groups. However, it would not be wise to rely on this assumption. For example, it could turn out that patients in the placebo group had larger tumours on average than those in the group treated with DES. If patients with large tumours have a poorer prognosis than those with small tumours, the size of the treatment effect would be overestimated, unless proper account was taken of the size of the tumour in the analysis. Consequently, it will first be necessary to determine if any of the covariates are related to survival time. If so, the effect of these variables will need to be allowed for when comparing the survival experiences of the patients in the two treatment groups.

1.3 Survivor, hazard and cumulative hazard functions

In summarising survival data, there are three functions of central interest, namely the *survivor function*, the *hazard function*, and the *cumulative hazard function*. These functions are therefore defined in this first chapter.

1.3.1 The survivor function

The actual survival time of an individual, t, can be regarded as the observed value of a variable, T, that can take any non-negative value. The different values that T can take have a *probability distribution*, and we call T the *random variable* associated with the survival time. Now suppose that this random variable has a probability distribution with underlying *probability density function* $f(t)$. The *distribution function* of T is then given by

$$F(t) = \mathrm{P}(T < t) = \int_0^t f(u)\,\mathrm{d}u, \tag{1.1}$$

and represents the probability that the survival time is less than some value t. This function is also called the *cumulative incidence function*, since it summarises the cumulative probability of death occurring before time t.

Table 1.4 *Survival times of prostatic cancer patients in a clinical trial to compare two treatments.*

Patient number	Treatment	Survival time	Status	Age	Serum haem.	Size of tumour	Gleason index
1	1	65	0	67	13.4	34	8
2	2	61	0	60	14.6	4	10
3	2	60	0	77	15.6	3	8
4	1	58	0	64	16.2	6	9
5	2	51	0	65	14.1	21	9
6	1	51	0	61	13.5	8	8
7	1	14	1	73	12.4	18	11
8	1	43	0	60	13.6	7	9
9	2	16	0	73	13.8	8	9
10	1	52	0	73	11.7	5	9
11	1	59	0	77	12.0	7	10
12	2	55	0	74	14.3	7	10
13	2	68	0	71	14.5	19	9
14	2	51	0	65	14.4	10	9
15	1	2	0	76	10.7	8	9
16	1	67	0	70	14.7	7	9
17	2	66	0	70	16.0	8	9
18	2	66	0	70	14.5	15	11
19	2	28	0	75	13.7	19	10
20	2	50	1	68	12.0	20	11
21	1	69	1	60	16.1	26	9
22	1	67	0	71	15.6	8	8
23	2	65	0	51	11.8	2	6
24	1	24	0	71	13.7	10	9
25	2	45	0	72	11.0	4	8
26	2	64	0	74	14.2	4	6
27	1	61	0	75	13.7	10	12
28	1	26	1	72	15.3	37	11
29	1	42	1	57	13.9	24	12
30	2	57	0	72	14.6	8	10
31	2	70	0	72	13.8	3	9
32	2	5	0	74	15.1	3	9
33	2	54	0	51	15.8	7	8
34	1	36	1	72	16.4	4	9
35	2	70	0	71	13.6	2	10
36	2	67	0	73	13.8	7	8
37	1	23	0	68	12.5	2	8
38	1	62	0	63	13.2	3	8

The survivor function, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t, and so from Equation (1.1),

$$S(t) = \mathrm{P}(T \geqslant t) = 1 - F(t). \tag{1.2}$$

The survivor function can therefore be used to represent the probability that an individual survives beyond any given time.

1.3.2 The hazard function

The *hazard function* is widely used to express the risk or hazard of an event such as death occurring at some time t. This function is obtained from the probability that an individual dies at time t, conditional on he or she having survived to that time. For a formal definition of the hazard function, consider the probability that the random variable associated with an individual's survival time, T, lies between t and $t+\delta t$, conditional on T being greater than or equal to t, written $\mathrm{P}(t \leqslant T < t + \delta t \mid T \geqslant t)$. This conditional probability is then expressed as a probability per unit time by dividing by the time interval, δt, to give a *rate*. The hazard function, $h(t)$, is then the limiting value of this quantity, as δt tends to zero, so that

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{\mathrm{P}(t \leqslant T < t + \delta t \mid T \geqslant t)}{\delta t} \right\}. \tag{1.3}$$

The function $h(t)$ is also referred to as the *hazard rate*, the *instantaneous death rate*, the *intensity rate* or the *force of mortality*.

From the definition of the hazard function in Equation (1.3), $h(t)$ is the event rate at time t, conditional on the event not having occurred before t. Specifically, if the survival time is measured in days, $h(t)$ is the approximate probability that an individual, who is at risk of the event occurring at the start of day t, experiences the event during that day. The hazard function at time t can also be regarded as the expected number of events experienced by an individual in unit time, given that the event has not occurred before then, and assuming that the hazard is constant over that time period.

The definition of the hazard function in Equation (1.3) leads to some useful relationships between the survivor and hazard functions. According to a standard result from probability theory, the probability of an event A, conditional on the occurrence of an event B, is given by $\mathrm{P}(A \,|\, B) = \mathrm{P}(AB)/\mathrm{P}(B)$, where $\mathrm{P}(AB)$ is the probability of the joint occurrence of A and B. Using this result, the conditional probability in the definition of the hazard function in Equation (1.3) is

$$\frac{\mathrm{P}(t \leqslant T < t + \delta t)}{\mathrm{P}(T \geqslant t)},$$

which is equal to

$$\frac{F(t + \delta t) - F(t)}{S(t)},$$

where $F(t)$ is the distribution function of T. Then,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}.$$

Now,

$$\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\}$$

is the definition of the derivative of $F(t)$ with respect to t, which is $f(t)$, and so

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.4}$$

Taken together, Equations (1.1), (1.2) and (1.4) show that from any one of the three functions, $f(t)$, $S(t)$, and $h(t)$, the other two can be determined.

1.3.3 The cumulative hazard function

From Equation (1.4), it follows that

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\{\log S(t)\}, \tag{1.5}$$

and so

$$S(t) = \exp\{-H(t)\}, \tag{1.6}$$

where

$$H(t) = \int_0^t h(u)\,\mathrm{d}u. \tag{1.7}$$

The function $H(t)$ features widely in survival analysis, and is called the *integrated* or *cumulative hazard function*. From Equation (1.6), the cumulative hazard function can also be obtained from the survivor function, since

$$H(t) = -\log S(t). \tag{1.8}$$

The cumulative hazard function, $H(t)$, is the cumulative risk of an event occurring by time t. If the event is death, then $H(t)$ summarises the risk of death up to time t, given that death has not occurred before t. The cumulative hazard function at time t can also be interpreted as the expected number of events that occur in the interval from the time origin to t.

It is possible for the cumulative hazard function to exceed unity. Using Equation (1.8), $H(t) \geqslant 1$, when $-\log S(t) \geqslant 1$, that is when $S(t) \leqslant e^{-1} = 0.37$. The cumulative hazard is then greater than unity when the probability of an event occurring after time t is less than 0.37, and means that more than one event is expected in the time interval $(0, t)$. The survivor function, $S(t)$, is then more correctly defined as the probability that one *or more* events occur after time t. The interpretation of a cumulative hazard function in terms of

the expected number of events is only reasonable when repetitions of an event are possible, such as when the event is the occurrence of an infection, migraine or seizure. When the event of interest is death, this interpretation relies on individuals being immediately resurrected after death has occurred! Methods for analysing times to multiple occurrences of an event are considered later in Chapter 13, and a more mathematical interpretation of the hazard and cumulative hazard functions when multiple events are possible is included in Section 13.1 of that chapter.

In the analysis of survival data, the survivor function, hazard function and cumulative hazard function are estimated from the observed survival times. Methods of estimation that do not require the form of the probability density function of T to be specified are described in Chapters 2 and 3, while methods based on the assumption of a particular survival time distribution are presented in Chapters 5 and 6.

1.4 Computer software for survival analysis

Most of the techniques for analysing survival data that will be presented in this book require suitable computer software for their implementation. Many computer packages for survival analysis are now available, but of the commercially available software packages, SAS (SAS Institute Inc.), S-PLUS (TIBCO Software Inc.) and Stata (StataCorp) have the most extensive range of facilities. In addition, the R statistical computing environment (R Core Team, 2013) is free software, distributed under the terms of the GNU General Public License. Both S-PLUS and R are modern implementations of the S statistical programming language, and include a comprehensive range of modules for survival analysis. Any of these four packages can be used to carry out the analyses described in subsequent chapters of this book.

In this book, the data sets used to illustrate the different methods of survival analysis have been analysed using SAS 9.4 (SAS Institute, Cary NC), mainly using the procedures `lifetest`, `lifereg` and `phreg`. Where published SAS macros have been used for more specialised analyses, these are documented in the 'Further reading' section of each chapter.

In some circumstances, numerical results in the output produced by software packages may differ. This is often due to different default methods of calculation being used. A particularly important example of this occurs when a data set includes two or more individuals with the same survival times. In this case, the SAS `phreg` procedure and the R package `survival` (Therneau, 2014) default to different methods of handling these tied observations, leading to differences in the output. The default settings can of course be changed, and the treatment of tied survival times is described in Section 3.3.2 of Chapter 3. Differences in numerical values may also result from different settings being used for parameters that control the convergence of certain iterative procedures, and different methods being used for numerical optimisation.

1.5 Further reading

An introduction to the techniques used in the analysis of survival data is included in a number of general books on statistics in medical research, such as those of Altman (1991) and Armitage, Berry and Matthews (2002). Machin, Cheung and Parmar (2006) provide a practical guide to the analysis of survival data from clinical trials, using non-technical language.

There a number of textbooks that provide an introduction to the methods of survival analysis, illustrated with practical examples. Lee and Wang (2013) provides a broad coverage of topics with illustrations drawn from biology and medicine, and Marubini and Valsecchi (1995) describe the analysis of survival data from clinical trials and observational studies. Hosmer, Lemeshow and May (2008) give a balanced account of survival analysis, with excellent chapters on model development and the interpretation of the parameter estimates in a fitted model. Klein and Moeschberger (2005) include many example data sets and exercises in their comprehensive textbook, and Kleinbaum and Klein (2012) provide a self-learning text on survival analysis. Applications of survival analysis in the analysis of epidemiological data are described by Breslow and Day (1987) and Woodward (2014). Introductory texts that describe the application of survival analysis in other areas include those of Crowder et al. (1991) who focus on the analysis of reliability data, and Box-Steffensmeier and Jones (2004) who give a non-mathematical account of time to event analysis in the social sciences.

Comprehensive accounts of the subject are given by Kalbfleisch and Prentice (2002) and Lawless (2002). These books have been written for the postgraduate statistician or research worker, and are usually regarded as reference books rather than introductory texts. A concise review of survival analysis is given in the research monograph of Cox and Oakes (1984), and in the chapter devoted to this subject in Hinkley, Reid and Snell (1991). The book by Hougaard (2000) on multivariate survival data incorporates more advanced topics, after introductory chapters that cover the basic features of survival analysis. Therneau and Grambsch (2000) base their presentation of survival analysis on the counting process approach, leading to a more mathematical development of the material. Harrell (2001) gives details on many issues that arise in the development of a statistical model not found in other texts, and includes an extensive discussion of two case studies.

There are many general books on the use of particular software packages for data analysis, and some that give a detailed account of how they are used in the analysis of survival data. Allison (2010) provides a comprehensive guide to the SAS software for survival analysis. Der and Everitt (2013) also include material on survival analysis in their text on the use of SAS for analysing medical data. Therneau and Grambsch (2000) give a detailed account of how SAS and S-PLUS are used to fit the Cox regression model, and extensions to it. This book includes a description of a number of SAS macros and S-PLUS functions that supplement the standard facilities available in these packages.

The use of S-PLUS in survival analysis is also described in Everitt and Rabe-Hesketh (2001) and Tableman and Kim (2004), while Broström (2012) shows how R is used in the analysis of survival data. Venables and Ripley (2002) describe how graphical and numerical data analyses can be carried out in the S environment that is implemented in both R and S-PLUS; note that S code generally runs under R. A similarly comprehensive account of the R system is given by Crawley (2013), while Dalgaard (2008) gives a more elementary introduction to R. The short introduction to R of Venables and Smith (2009) is also available from R Core Team (2013). The use of Stata in survival analysis is presented by Cleves et al. (2010), and Rabe-Hesketh and Everitt (2007) give a more general introduction to the use of Stata in data analysis.

Chapter 2

Some non-parametric procedures

An initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times for individuals in a particular group. Such summaries may be of interest in their own right, or as a precursor to a more detailed analysis of the data. Survival data are conveniently summarised through estimates of the survivor function and hazard function. Methods for estimating these functions from a single sample of survival data are described in Sections 2.1 and 2.3. These methods are said to be *non-parametric* or *distribution-free*, since they do not require specific assumptions to be made about the underlying distribution of the survival times.

Once the estimated survivor function has been found, the median and other percentiles of the distribution of survival times can be estimated, as shown in Section 2.4. Numerical summaries of the data, derived on the basis of assumptions about the probability distribution from which the data have been drawn, will be considered later in Chapters 5 and 6.

When the survival times of two groups of patients are being compared, an informal comparison of the survival experience of each group of individuals can be made using the estimated survivor functions. However, there are more formal procedures that enable two groups of survival data to be compared. Two non-parametric procedures for comparing two or more groups of survival times, namely the *log-rank test* and the *Wilcoxon test*, are described in Section 2.6.

2.1 Estimating the survivor function

Suppose first that we have a single sample of survival times, where none of the observations are censored. The survivor function $S(t)$, defined in Equation (1.2), is the probability that an individual survives for a time greater than or equal to t. This function can be estimated by the *empirical survivor function*, given by

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times} \geqslant t}{\text{Number of individuals in the data set}}. \qquad (2.1)$$

Equivalently, $\hat{S}(t) = 1 - \hat{F}(t)$, where $\hat{F}(t)$ is the *empirical distribution function*, that is, the ratio of the total number of individuals alive at time t to the total

number of individuals in the study. Notice that the empirical survivor function is equal to unity for values of t before the first death time, and zero after the final death time.

The estimated survivor function $\hat{S}(t)$ is assumed to be constant between two adjacent death times, and so a plot of $\hat{S}(t)$ against t is a step-function. The function decreases immediately after each observed survival time.

Example 2.1 Pulmonary metastasis
One complication in the management of patients with a malignant bone tumour, or osteosarcoma, is that the tumour often spreads to the lungs. This pulmonary metastasis is life-threatening. In a study concerned with the treatment of pulmonary metastasis arising from osteosarcoma, Burdette and Gehan (1970) give the following survival times, in months, of eleven male patients.

11 13 13 13 13 13 14 14 15 15 17

Using Equation (2.1), the estimated values of the survivor function at times 11, 13, 14, 15 and 17 months are 1.000, 0.909, 0.455, 0.273 and 0.091. The estimated value of the survivor function is unity from the time origin until 11 months, and zero after 17 months. A graph of the estimated survivor function is given in Figure 2.1.

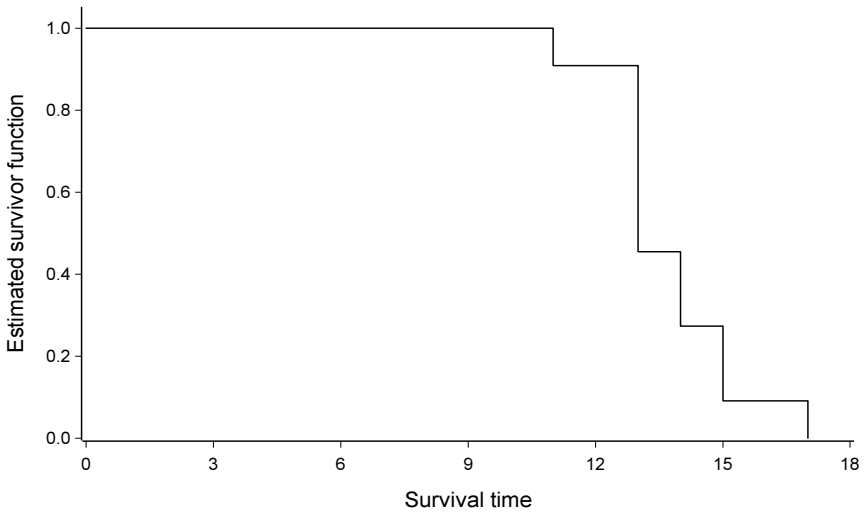


Figure 2.1 *Estimated survivor function for the data from Example 2.1.*

The method of estimating the survivor function illustrated in the above example cannot be used when there are censored observations. The reason for this is that the method does not allow information provided by an individual

whose survival time is censored before time t to be used in computing the estimated survivor function at t. Non-parametric methods for estimating $S(t)$, which can be used in the presence of censored survival times, are described in the following sections.

2.1.1 Life-table estimate of the survivor function

The *life-table estimate* of the survivor function, also known as the *actuarial estimate of survivor function*, is obtained by first dividing the period of observation into a series of time intervals. These intervals need not necessarily be of equal length, although they usually are. The number of intervals used will depend on the number of individuals in the study, but would usually be somewhere between 5 and 15.

Suppose that the jth of m such intervals, $j = 1, 2, \ldots, m$, extends from time t'_{j-1} to immediately before time t'_j, where we take $t_0 = 0$ and $t_m = \infty$. Also, let d_j and c_j denote the number of deaths and the number of censored survival times, respectively, in this interval, and let n_j be the number of individuals who are alive, and therefore at risk of death, at the start of the jth interval. We now make the assumption that the censoring process is such that the censored survival times occur uniformly throughout the jth interval, so that the average number of individuals who are at risk during this interval is

$$n'_j = n_j - c_j/2. \tag{2.2}$$

This assumption is sometimes known as the *actuarial assumption.*

In the jth interval, the probability of death can be estimated by d_j/n'_j, so that the corresponding survival probability is $(n'_j - d_j)/n'_j$. Now consider the probability that an individual survives beyond time t'_{j-1}, $j = 2, 3, \ldots, m$, that is, until some time after the start of the jth interval. This will be the product of the probabilities that an individual survives through each of the $j-1$ preceding intervals, and so the life-table estimate of the survivor function is given by

$$S^*(t) = \prod_{i=1}^{j-1} \left(\frac{n'_i - d_i}{n'_i} \right), \tag{2.3}$$

for $t'_{j-1} \leqslant t < t'_j$, $j = 2, 3, \ldots, m$. The estimated probability of surviving beyond the start of the first interval, t'_0, is of course unity, while the estimated probability of surviving beyond t'_m is zero. A graphical estimate of the survivor function will then be a step-function with constant values of the function in each time interval.

Example 2.2 Survival of multiple myeloma patients
To illustrate the computation of the life-table estimate, consider the data on the survival times of the 48 multiple myeloma patients given in Table 1.3. In this illustration, the information collected on other explanatory variables for each individual will be ignored.

The survival times are first grouped to give the number of patients who die, d_j, and the number who are censored, c_j, in each of the first five years of the study, and in the subsequent three-year period. The number at risk of death at the start of each of these intervals, n_j, is then computed, together with the adjusted number at risk, n'_j. Finally, the probability of survival through each interval is estimated, from which the estimated survivor function is obtained using Equation (2.3). The calculations are shown in Table 2.1, in which the time period is given in months, and the jth interval that begins at time t'_{j-1} and ends just before time t'_j, for $j = 1, 2, \ldots, m$, is denoted $t'_{j-1}{}^{-}$.

Table 2.1 *Life-table estimate of the survivor function for the data from Example 1.3.*

Interval	Time period	d_j	c_j	n_j	n'_j	$(n'_j - d_j)/n'_j$	$S^*(t)$
1	0–	16	4	48	46.0	0.6522	1.0000
2	12–	10	4	28	26.0	0.6154	0.6522
3	24–	1	0	14	14.0	0.9286	0.4013
4	36–	3	1	13	12.5	0.7600	0.3727
5	48–	2	2	9	8.0	0.7500	0.2832
6	60–	4	1	5	4.5	0.1111	0.2124

A graph of the life-table estimate of the survivor function is shown in Figure 2.2.

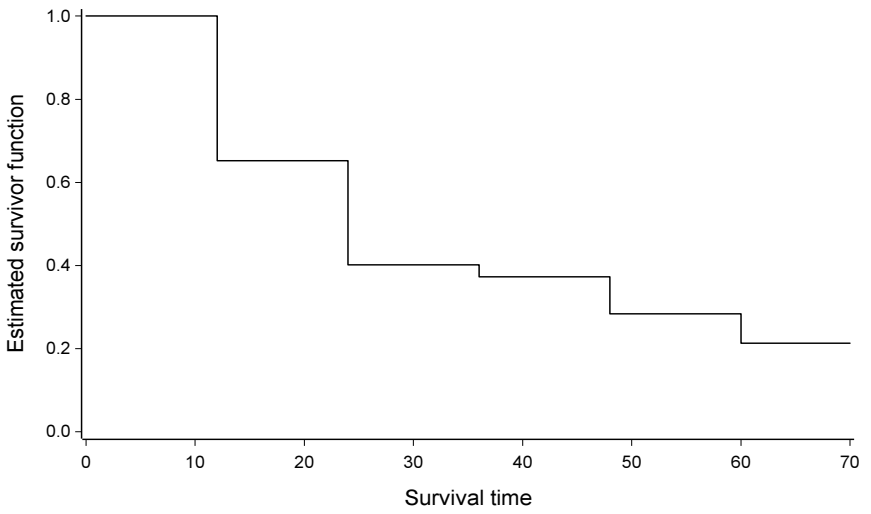


Figure 2.2 *Life-table estimate of the survivor function.*

The form of the estimated survivor function obtained using this method is sensitive to the choice of the intervals used in its construction, just as the

shape of a histogram depends on the choice of the class intervals. On the other hand, the life-table estimate is particularly well suited to situations in which the actual death times are unknown, and the only available information is the number of deaths and the number of censored observations that occur in a series of consecutive time intervals. In practice, such interval-censored survival data occur quite frequently.

When the actual survival times are known, the life-table estimate can still be used, as in Example 2.2, but the grouping of the survival times does result in some loss of information. Alternative methods for estimating the survivor function are then more appropriate, such as that leading to the Kaplan-Meier estimate.

2.1.2 Kaplan-Meier estimate of the survivor function

The first step in the analysis of ungrouped censored survival data is normally to obtain the *Kaplan-Meier estimate* of the survivor function. This estimate is therefore considered in some detail. To obtain the Kaplan-Meier estimate, a series of time intervals is constructed, as for the life-table estimate. However, each of these intervals is designed to be such that one death time is contained in the interval, and this death time is taken to occur at the start of the interval.

As an illustration, suppose that $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$ are three observed survival times arranged in rank order, so that $t_{(1)} < t_{(2)} < t_{(3)}$, and that c is a censored survival time that falls between $t_{(2)}$ and $t_{(3)}$. The constructed intervals then begin at times $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$, and each interval includes the one death time, although there could be more than one individual who dies at any particular death time. Notice that no interval begins at the censored time of c. Now suppose that two individuals die at $t_{(1)}$, one dies at $t_{(2)}$ and three die at $t_{(3)}$. The situation is illustrated diagrammatically in Figure 2.3, in which D represents a death and C a censored survival time.

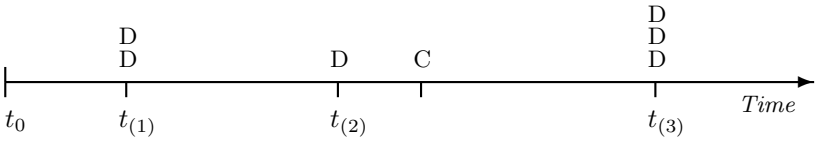


Figure 2.3 *Construction of intervals used in the derivation of the Kaplan-Meier estimate.*

The time origin is denoted by t_0, and so there is an initial period commencing at t_0, which ends just before $t_{(1)}$, the time of the first death. This means that the interval from t_0 to $t_{(1)}$ will not include a death time. The first constructed interval extends from $t_{(1)}$ to just before $t_{(2)}$, and since the second death time is at $t_{(2)}$, this interval includes the single death time at $t_{(1)}$. The second interval begins at time $t_{(2)}$ and ends just before $t_{(3)}$, and includes the death time at $t_{(2)}$ and the censored time c. There is also a third interval beginning at $t_{(3)}$, which contains the longest survival time, $t_{(3)}$.

In general, suppose that there are n individuals with observed survival times $t_1, t_2, \ldots, t_n$. Some of these observations may be right-censored, and there may also be more than one individual with the same observed survival time. We therefore suppose that there are r death times amongst the individuals, where $r \leqslant n$. After arranging these death times in ascending order, the jth is denoted $t_{(j)}$, for $j = 1, 2, \ldots, r$, and so the r ordered death times are $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. The number of individuals who are alive just before time $t_{(j)}$, including those who are about to die at this time, will be denoted n_j, for $j = 1, 2, \ldots, r$, and d_j will denote the number who die at this time. The time interval from $t_{(j)} - \delta$ to $t_{(j)}$, where δ is an infinitesimal time interval, then includes one death time. Since there are n_j individuals who are alive just before $t_{(j)}$ and d_j deaths at $t_{(j)}$, the probability that an individual dies during the interval from $t_{(j)} - \delta$ to $t_{(j)}$ is estimated by d_j/n_j. The corresponding estimated probability of survival through that interval is then $(n_j - d_j)/n_j$.

It sometimes happens that there are censored survival times that occur at the same time as one or more deaths, so that a death time and a censored survival time appear to occur simultaneously. In this situation, the censored survival time is taken to occur immediately after the death time when computing the values of the n_j.

From the manner in which the time intervals are constructed, the interval from $t_{(j)}$ to $t_{(j+1)} - \delta$, the time immediately before the next death time, contains no deaths. The probability of surviving from $t_{(j)}$ to $t_{(j+1)} - \delta$ is therefore unity, and the joint probability of surviving from $t_{(j)} - \delta$ to $t_{(j)}$ and from $t_{(j)}$ to $t_{(j+1)} - \delta$ can be estimated by $(n_j - d_j)/n_j$. In the limit, as δ tends to zero, $(n_j - d_j)/n_j$ becomes an estimate of the probability of surviving the interval from $t_{(j)}$ to $t_{(j+1)}$.

We now make the assumption that the deaths of the individuals in the sample occur independently of one another. Then, the estimated survivor function at any time, t, in the kth constructed time interval from $t_{(k)}$ to $t_{(k+1)}$, $k = 1, 2, \ldots, r$, where $t_{(r+1)}$ is defined to be ∞, will be the estimated probability of surviving beyond $t_{(k)}$. This is actually the probability of surviving through the interval from $t_{(k)}$ to $t_{(k+1)}$, and all preceding intervals, and leads to the Kaplan-Meier estimate of the survivor function, which is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{2.4}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, with $\hat{S}(t) = 1$ for $t < t_{(1)}$, and where $t_{(r+1)}$ is taken to be ∞. If the largest observation is a censored survival time, t^*, say, $\hat{S}(t)$ is undefined for $t > t^*$. On the other hand, if the largest observed survival time, $t_{(r)}$, is an uncensored observation, $n_r = d_r$, and so $\hat{S}(t)$ is zero for $t \geqslant t_{(r)}$. A plot of the Kaplan-Meier estimate of the survivor function is a step-function, in which the estimated survival probabilities are constant between adjacent death times and decrease at each death time.

Equation (2.4) shows that, as for the life-table estimate of the survivor function in Equation (2.3), the Kaplan-Meier estimate is formed as a product of a series of estimated probabilities. In fact, the Kaplan-Meier estimate is the limiting value of the life-table estimate in Equation (2.3) as the number of intervals tends to infinity and their width tends to zero. For this reason, the Kaplan-Meier estimate is also known as the *product-limit estimate* of the survivor function.

Note that if there are no censored survival times in the data set, $n_j - d_j = n_{j+1}$, $j = 1, 2, \ldots, k$, in Equation (2.4), and on expanding the product we get

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times \cdots \times \frac{n_{k+1}}{n_k}.$$

This reduces to n_{k+1}/n_1, for $k = 1, 2, \ldots, r-1$, with $\hat{S}(t) = 1$ for $t < t_{(1)}$ and $\hat{S}(t) = 0$ for $t \geqslant t_{(r)}$. Now, n_1 is the number of individuals at risk just before the first death time, which is the number of individuals in the sample, and n_{k+1} is the number of individuals with survival times greater than or equal to $t_{(k+1)}$. Consequently, in the absence of censoring, $\hat{S}(t)$ is simply the empirical survivor function defined in Equation (2.1). The Kaplan-Meier estimate is therefore a generalisation of the empirical survivor function that accommodates censored observations.

Example 2.3 Time to discontinuation of the use of an IUD
Data from 18 women on the time to discontinuation of the use of an intra-uterine device (IUD) were given in Table 1.1. For these data, the survivor function, $S(t)$, represents the probability that a woman discontinues the use of the contraceptive device after any time t. The Kaplan-Meier estimate of the survivor function is readily obtained using Equation (2.4), and the required calculations are set out in Table 2.2.

Table 2.2 *Kaplan-Meier estimate of the survivor function for the data from Example 1.1.*

Time interval	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0–	18	0	1.0000	1.0000
10–	18	1	0.9444	0.9444
19–	15	1	0.9333	0.8815
30–	13	1	0.9231	0.8137
36–	12	1	0.9167	0.7459
59–	8	1	0.8750	0.6526
75–	7	1	0.8571	0.5594
93–	6	1	0.8333	0.4662
97–	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

The estimated survivor function, $\hat{S}(t)$, is plotted in Figure 2.4. Note that since the largest discontinuation time of 107 days is censored, $\hat{S}(t)$ is not defined beyond $t = 107$.

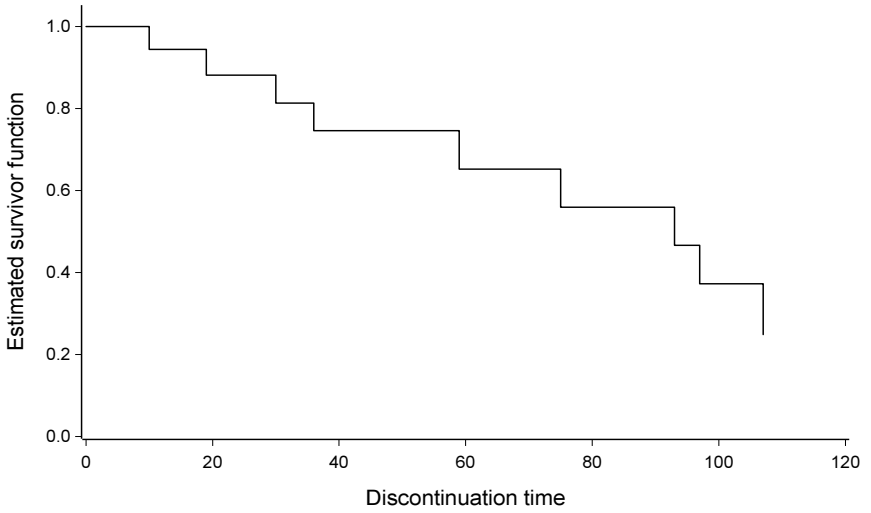


Figure 2.4 *Kaplan-Meier estimate of the survivor function for the data from Example 1.1.*

2.1.3 Nelson-Aalen estimate of the survivor function

An alternative estimate of the survivor function, which is based on the individual event times, is the *Nelson-Aalen estimate*, given by

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp(-d_j/n_j). \tag{2.5}$$

This estimate can be obtained from an estimate of the cumulative hazard function, as shown in Section 2.3.3. Moreover, the Kaplan-Meier estimate of the survivor function can be regarded as an approximation to the Nelson-Aalen estimate. To show this, we use the result that

$$e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots,$$

which is approximately equal to $1 - x$ when x is small. It then follows that $\exp(-d_j/n_j) \approx 1 - (d_j/n_j) = (n_j - d_j)/n_j$, so long as d_j is small relative to n_j, which it will be except at the longest survival times. Consequently, the Kaplan-Meier estimate, $\hat{S}(t)$, in Equation (2.4), approximates the Nelson-Aalen estimate, $\tilde{S}(t)$, in Equation (2.5).

The Nelson-Aalen estimate of the survivor function, also known as *Altshuler's estimate*, will always be greater than the Kaplan-Meier estimate at any given time, since $e^{-x} \geqslant 1 - x$, for all values of x. Although the Nelson-Aalen estimate has been shown to perform better than the Kaplan-Meier

estimate in small samples, in many circumstances, the estimates will be very similar, particularly at the earlier survival times. Since the Kaplan-Meier estimate is a generalisation of the empirical survivor function, the latter estimate has much to commend it.

Example 2.4 Time to discontinuation of the use of an IUD
The values shown in Table 2.2, which gives the Kaplan-Meier estimate of the survivor function for the data on the time to discontinuation of the use of an intrauterine device, can be used to calculate the Nelson-Aalen estimate. This estimate is shown in Table 2.3.

Table 2.3 *Nelson-Aalen estimate of the survivor function for the data from Example 1.1.*

Time interval	$\exp(-d_j/n_j)$	$\tilde{S}(t)$
0–	1.0000	1.0000
10–	0.9460	0.9460
19–	0.9355	0.8850
30–	0.9260	0.8194
36–	0.9200	0.7539
59–	0.8825	0.6653
75–	0.8669	0.5768
93–	0.8465	0.4882
97–	0.8187	0.3997
107	0.7165	0.2864

From this table we see that the Kaplan-Meier and Nelson-Aalen estimates of the survivor function differ by less than 0.04. However, when we consider the precision of these estimates, which we do in Section 2.2, we see that a difference of 0.04 is of no practical importance.

2.2 Standard error of the estimated survivor function

An essential aid to the interpretation of an estimate of any quantity is the precision of the estimate, which is reflected in the *standard error* of the estimate. This is defined to be the square root of the estimated variance of the estimate, and is used in the construction of an interval estimate for a quantity of interest. In this section, the standard error of estimates of the survivor function are given.

Because the Kaplan-Meier estimate is the most important and widely used estimate of the survivor function, the derivation of the standard error of $\hat{S}(t)$ will be presented in detail in this section. The details of this derivation can be omitted on a first reading.

2.2.1 Standard error of the Kaplan-Meier estimate*

The Kaplan-Meier estimate of the survivor function for any value of t in the interval from $t_{(k)}$ to $t_{(k+1)}$ can be written as

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j,$$

for $k = 1, 2, \ldots, r$, where $\hat{p}_j = (n_j - d_j)/n_j$ is the estimated probability that an individual survives through the time interval that begins at $t_{(j)}$, $j = 1, 2, \ldots, r$. Taking logarithms,

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j,$$

and so the variance of $\log \hat{S}(t)$ is given by

$$\text{var}\left\{\log \hat{S}(t)\right\} = \sum_{j=1}^{k} \text{var}\left\{\log \hat{p}_j\right\}. \tag{2.6}$$

Now, the number of individuals who survive through the interval beginning at $t_{(j)}$ can be assumed to have a *binomial distribution* with parameters n_j and p_j, where p_j is the true probability of survival through that interval. The observed number who survive is $n_j - d_j$, and using the result that the variance of a binomial random variable with parameters n, p is $np(1-p)$, the variance of $n_j - d_j$ is given by

$$\text{var}\,(n_j - d_j) = n_j p_j (1 - p_j).$$

Since $\hat{p}_j = (n_j - d_j)/n_j$, the variance of $\hat{p}_j$ is $\text{var}\,(n_j - d_j)/n_j^2$, that is, $p_j(1 - p_j)/n_j$. The variance of $\hat{p}_j$ may then be estimated by

$$\hat{p}_j(1 - \hat{p}_j)/n_j. \tag{2.7}$$

In order to obtain the variance of $\log \hat{p}_j$, we make use of a general result for the approximate variance of a function of a random variable. According to this result, the variance of a function $g(X)$ of the random variable X is given by

$$\text{var}\,\{g(X)\} \approx \left\{\frac{\mathrm{d}g(X)}{\mathrm{d}X}\right\}^2 \text{var}\,(X). \tag{2.8}$$

This is known as the *Taylor series approximation* to the variance of a function of a random variable. Using Equation (2.8), the approximate variance of $\log \hat{p}_j$ is $\text{var}\,(\hat{p}_j)/\hat{p}_j^2$, and using Expression (2.7), the approximate estimated variance of $\log \hat{p}_j$ is $(1 - \hat{p}_j)/(n_j \hat{p}_j)$, which on substitution for $\hat{p}_j$, reduces to

$$\frac{d_j}{n_j(n_j - d_j)}. \tag{2.9}$$

* Sections marked with an asterisk may be omitted without loss of continuity.

Then, from Equation (2.6),

$$\text{var}\left\{\log \hat{S}(t)\right\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}, \tag{2.10}$$

and a further application of the result in Equation (2.8) gives

$$\text{var}\left\{\log \hat{S}(t)\right\} \approx \frac{1}{[\hat{S}(t)]^2}\,\text{var}\left\{\hat{S}(t)\right\},$$

so that

$$\text{var}\left\{\hat{S}(t)\right\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.11}$$

Finally, the standard error of the Kaplan-Meier estimate of the survivor function, defined to be the square root of the estimated variance of the estimate, is given by

$$\text{se}\left\{\hat{S}(t)\right\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\right\}^{\frac{1}{2}}, \tag{2.12}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$. This result is known as *Greenwood's formula*.

If there are no censored survival times, $n_j - d_j = n_{j+1}$, and Expression (2.9) becomes $(n_j - n_{j+1})/n_j n_{j+1}$. Now,

$$\sum_{j=1}^{k} \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^{k}\left(\frac{1}{n_{j+1}} - \frac{1}{n_j}\right) = \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

which can be written as

$$\frac{1 - \hat{S}(t)}{n_1 \hat{S}(t)},$$

since $\hat{S}(t) = n_{k+1}/n_1$ for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$, in the absence of censoring. Hence, from Equation (2.11), the estimated variance of $\hat{S}(t)$ is $\hat{S}(t)[1 - \hat{S}(t)]/n_1$. This is an estimate of the variance of the empirical survivor function, given in Equation (2.1), on the assumption that the number of individuals at risk at time t has a binomial distribution with parameters $n_1, S(t)$.

2.2.2 Standard error of other estimates*

The life-table estimate of the survivor function is similar in form to the Kaplan-Meier estimate, and so the standard error of this estimator is obtained in a similar manner. In the notation of Section 2.1.1, the standard

error of the life-table estimate is given by

$$\text{se}\,\{S^*(t)\} \approx S^*(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n'_j(n'_j-d_j)}\right\}^{\frac{1}{2}}.$$

The standard error of the Nelson-Aalen estimator is

$$\text{se}\left\{\tilde{S}(t)\right\} \approx \tilde{S}(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n_j^2}\right\}^{\frac{1}{2}},$$

although other expressions have been proposed.

2.2.3 Confidence intervals for values of the survivor function

Once the standard error of an estimate of the survivor function has been calculated, a *confidence interval* for the corresponding value of the survivor function, at a given time t, can be found. A confidence interval is an interval estimate of the survivor function, and is the interval which is such that there is a prescribed probability that the value of the true survivor function is included within it. The intervals constructed in this manner are sometimes referred to as *pointwise confidence intervals*, since they apply to a specific survival time.

A confidence interval for the true value of the survivor function at a given time t is obtained by assuming that the estimated value of the survivor function at t is normally distributed with mean $S(t)$ and estimated variance given by Equation (2.11). The interval is computed from *percentage points* of the standard normal distribution. Thus, if Z is a random variable that has a standard normal distribution, the upper (one-sided) $\alpha/2$-point, or the two-sided α-point, of this distribution is that value $z_{\alpha/2}$ which is such that $\text{P}(Z > z_{\alpha/2}) = \alpha/2$. This probability is the area under the standard normal curve to the right of $z_{\alpha/2}$, as illustrated in Figure 2.5. For example, the two-sided 5% and 1% points of the standard normal distribution, $z_{0.025}$ and $z_{0.005}$, are 1.96 and 2.58, respectively.

A $100(1-\alpha)\%$ confidence interval for $S(t)$, for a given value of t, is the interval from $\hat{S}(t) - z_{\alpha/2}\,\text{se}\,\{\hat{S}(t)\}$ to $\hat{S}(t) + z_{\alpha/2}\,\text{se}\,\{\hat{S}(t)\}$, where $\text{se}\,\{\hat{S}(t)\}$ is found from Equation (2.12). These intervals for $S(t)$ can be superimposed on a graph of the estimated survivor function, as shown in Example 2.5.

One difficulty with this procedure arises from the fact that the confidence intervals are symmetric. When the estimated survivor function is close to zero or unity, symmetric intervals are inappropriate, since they can lead to confidence limits for the survivor function that lie outside the interval (0,1). A pragmatic solution to this problem is to replace any limit that is greater than unity by 1.0, and any limit that is less than zero by 0.0.

An alternative procedure is to transform $\hat{S}(t)$ to a value in the range $(-\infty, \infty)$, and obtain a confidence interval for the transformed value. The

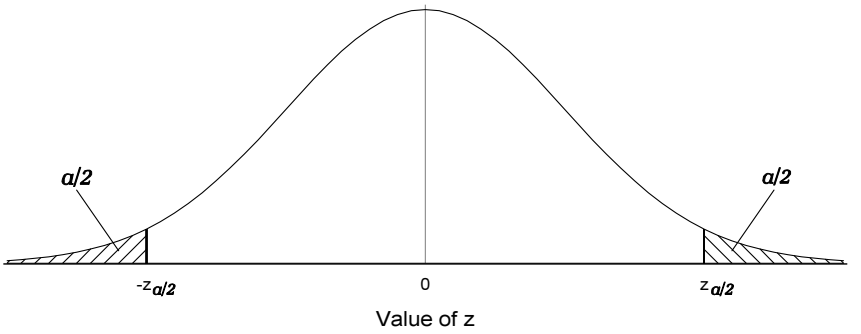


Figure 2.5 *Upper and lower $\alpha/2$-points of the standard normal distribution.*

resulting confidence limits are then back-transformed to give a confidence interval for $S(t)$ itself. Possible transformations are the logistic transformation, $\log[S(t)/\{1 - S(t)\}]$, and the complementary log-log transformation, $\log\{-\log S(t)\}$. Note that from Equation (1.8), the latter quantity is the logarithm of the cumulative hazard function. In either case, the standard error of the transformed value of $\hat{S}(t)$ can be found using the approximation in Equation (2.8).

For example, the variance of $\log\{-\log \hat{S}(t)\}$ is obtained from the expression for $\text{var}\{\log \hat{S}(t)\}$ in Equation (2.10). Using the general result in Equation (2.8),

$$\text{var}\{\log(-X)\} \approx \frac{1}{X^2}\,\text{var}\,(X),$$

and setting $X = \log \hat{S}(t)$ gives

$$\text{var}\left[\log\{-\log \hat{S}(t)\}\right] \approx \frac{1}{\{\log \hat{S}(t)\}^2} \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}.$$

The standard error of $\log\{-\log \hat{S}(t)\}$ is the square root of this quantity. This leads to $100(1 - \alpha)\%$ limits of the form

$$\hat{S}(t)^{\exp[\pm z_{\alpha/2}\,\text{se}\{\log[-\log \hat{S}(t)]\}]},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

A further problem is that in the tails of the distribution of the survival times, that is, when $\hat{S}(t)$ is close to zero or unity, the variance of $\hat{S}(t)$ obtained using Greenwood's formula can underestimate the actual variance. In these circumstances, an alternative expression for the standard error of $\hat{S}(t)$ may be used. Peto et al. (1977) propose that the standard error of $\hat{S}(t)$ should be obtained from the equation

$$\text{se}\,\{\hat{S}(t)\} = \frac{\hat{S}(t)\sqrt{\{1 - \hat{S}(t)\}}}{\sqrt{(n_k)}},$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, where $\hat{S}(t)$ is the Kaplan-Meier estimate of $S(t)$ and n_k is the number of individuals at risk at $t_{(k)}$, the start of the kth constructed time interval.

This expression for the standard error of $\hat{S}(t)$ is conservative, in the sense that the standard errors obtained will tend to be larger than they ought to be. For this reason, the Greenwood estimate is recommended for general use.

Example 2.5 Time to discontinuation of the use of an IUD
The standard error of the estimated survivor function, and 95% confidence limits for the corresponding true value of the function, for the data from Example 1.1 on the times to discontinuation of use of an IUD, are given in Table 2.4. In this table, confidence limits outside the range $(0, 1)$ have been replaced by zero or unity.

Table 2.4 *Standard error of $\hat{S}(t)$ and confidence intervals for $S(t)$ for the data from Example 1.1.*

Time interval	$\hat{S}(t)$	se $\{\hat{S}(t)\}$	95% confidence interval
0–	1.0000	0.0000	
10–	0.9444	0.0540	(0.839, 1.000)
19–	0.8815	0.0790	(0.727, 1.000)
30–	0.8137	0.0978	(0.622, 1.000)
36–	0.7459	0.1107	(0.529, 0.963)
59–	0.6526	0.1303	(0.397, 0.908)
75–	0.5594	0.1412	(0.283, 0.836)
93–	0.4662	0.1452	(0.182, 0.751)
97–	0.3729	0.1430	(0.093, 0.653)
107	0.2486	0.1392	(0.000, 0.522)

From this table we see that, in general, the standard error of the estimated survivor function increases with the discontinuation time. The reason for this is that estimates of the survivor function at later times are based on fewer individuals. A graph of the estimated survivor function, with the 95% confidence limits shown as dashed lines, is given in Figure 2.6.

It is important to observe that the confidence limits for a survivor function, illustrated in Figure 2.6, are only valid for any given time. Different methods are needed to produce confidence bands that are such that there is a given probability, 0.95 for example, that the survivor function is contained in the band for all values of t. These bands will tend to be wider than the band formed from the pointwise confidence limits. Details will not be included, but references to these methods are given in the final section of this chapter. Notice also that the width of these intervals is very much greater than the difference between the Kaplan-Meier and Nelson-Aalen estimates of the survivor function, shown in Tables 2.2 and 2.3. Similar calculations lead to confidence limits based on life-table and Nelson-Aalen estimates of the survivor function.

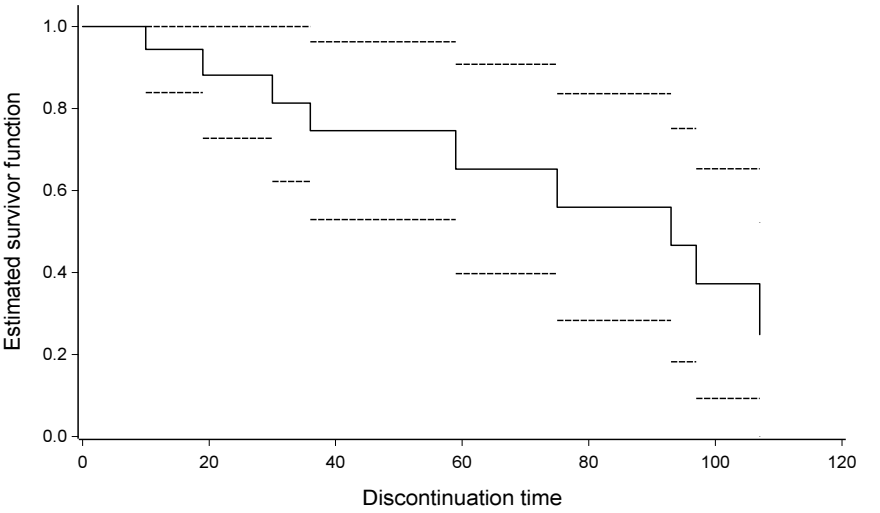


Figure 2.6 *Estimated survivor function and 95% confidence limits for* $S(t)$.

2.3 Estimating the hazard function

A single sample of survival data may also be summarised through the hazard function, which shows the dependence of the instantaneous risk of death on time. There are a number of ways of estimating this function, two of which are described in this section.

2.3.1 Life-table estimate of the hazard function

Suppose that the observed survival times have been grouped into a series of m intervals, as in the construction of the life-table estimate of the survivor function. An appropriate estimate of the average hazard of death per unit time over each interval is the observed number of deaths in that interval, divided by the average time survived in that interval. This latter quantity is the average number of persons at risk in the interval, multiplied by the length of the interval. Let the number of deaths in the jth time interval be d_j, $j = 1, 2, \ldots, m$, and suppose that n'_j is the average number of individuals at risk of death in that interval, where n'_j is given by Equation (2.2). Assuming that the death rate is constant during the jth interval, the average time survived in that interval is $(n'_j - d_j/2)\tau_j$, where τ_j is the length of the jth time interval. The life-table estimate of the hazard function in the jth time interval is then given by

$$h^*(t) = \frac{d_j}{(n'_j - d_j/2)\tau_j},$$

for $t'_{j-1} \leqslant t < t'_j$, $j = 1, 2, \ldots, m$, so that $h^*(t)$ is a step-function.

The asymptotic standard error of this estimate has been shown by Gehan (1969) to be given by

$$\text{se}\{h^*(t)\} = \frac{h^*(t)\sqrt{\{1 - [h^*(t)\tau_j/2]^2\}}}{\sqrt{(d_j)}},$$

and confidence intervals for the corresponding true hazard over each of the m time intervals can be obtained in the manner described in Section 2.2.3.

Example 2.6 Survival of multiple myeloma patients
The life-table estimate of the survivor function for the data from Example 1.3 on the survival times of 48 multiple myeloma patients was given in Table 2.1. Using the same time intervals as were used in Example 2.2, calculations leading to the life-table estimate of the hazard function are given in Table 2.5.

Table 2.5 *Life-table estimate of the hazard function for the data from Example 1.3.*

Time period	τ_j	d_j	n'_j	$h^*(t)$
0–	12	16	46.0	0.0351
12–	12	10	26.0	0.0397
24–	12	1	14.0	0.0062
36–	12	3	12.5	0.0227
48–	12	2	8.0	0.0238
60–	36	4	4.5	0.0444

The estimated hazard function is plotted as a step-function in Figure 2.7. The general pattern is for the hazard to remain roughly constant over the first two years from diagnosis, after which time it declines and then increases gradually. However, some caution is needed in interpreting this estimate, as there are few deaths two years after diagnosis.

2.3.2 Kaplan-Meier type estimate

A natural way of estimating the hazard function for ungrouped survival data is to take the ratio of the number of deaths at a given death time to the number of individuals at risk at that time. If the hazard function is assumed to be constant between successive death times, the hazard per unit time can be found by further dividing by the time interval. Thus, if there are d_j deaths at the jth death time, $t_{(j)}$, $j = 1, 2, \ldots, r$, and n_j at risk at time $t_{(j)}$, the hazard function in the interval from $t_{(j)}$ to $t_{(j+1)}$ can be estimated by

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \tag{2.13}$$

for $t_{(j)} \leqslant t < t_{(j+1)}$, where $\tau_j = t_{(j+1)} - t_{(j)}$. Notice that it is not possible to use Equation (2.13) to estimate the hazard in the interval that begins at the final death time, since this interval is open-ended.

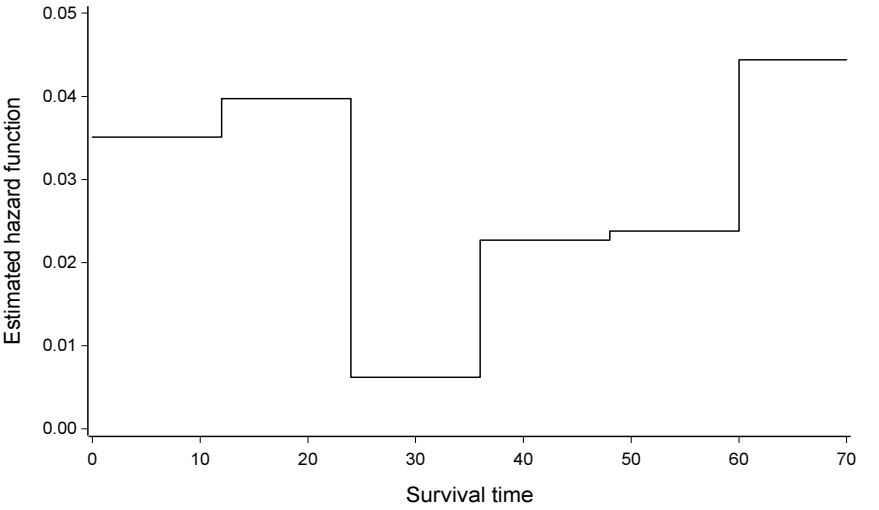


Figure 2.7 *Life-table estimate of the hazard function for the data from Example 1.3.*

The estimate in Equation (2.13) is referred to as a *Kaplan-Meier type estimate*, because the estimated survivor function derived from it is the Kaplan-Meier estimate. To show this, note that since $\hat{h}(t)$, $t_{(j)} \leqslant t < t_{(j+1)}$, is an estimate of the risk of death per unit time in the jth interval, the probability of death in that interval is $\hat{h}(t)\tau_j$, that is, d_j/n_j. Hence an estimate of the corresponding survival probability in that interval is $1-(d_j/n_j)$, and the estimated survivor function is as given by Equation (2.4).

The approximate standard error of $\hat{h}(t)$ can be found from the variance of d_j, which, following Section 2.2.1, may be assumed to have a binomial distribution with parameters n_j and p_j, where p_j is the probability of death in the interval of length τ. Consequently, $\text{var}\,(d_j) = n_j p_j(1-p_j)$, and estimating p_j by d_j/n_j gives

$$\text{se}\,\{\hat{h}(t)\} = \hat{h}(t)\sqrt{\frac{n_j - d_j}{n_j d_j}}.$$

However, when d_j is small, confidence intervals constructed using this standard error will be too wide to be of practical use.

Example 2.7 Time to discontinuation of the use of an IUD
Consider again the data on the time to discontinuation of the use of an IUD for 18 women, given in Example 1.1. The Kaplan-Meier estimate of the survivor function for these data was given in Table 2.2, and Table 2.6 gives the corresponding Kaplan-Meier type estimate of the hazard function, computed from Equation (2.13). The approximate standard errors of $\hat{h}(t)$ are also given.

Table 2.6 *Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.*

Time interval	τ_j	n_j	d_j	$\hat{h}(t)$	se $\{\hat{h}(t)\}$
0–	10	18	0	0.0000	–
10–	9	18	1	0.0062	0.0060
19–	11	15	1	0.0061	0.0059
30–	6	13	1	0.0128	0.0123
36–	23	12	1	0.0036	0.0035
59–	16	8	1	0.0078	0.0073
75–	18	7	1	0.0079	0.0073
93–	4	6	1	0.0417	0.0380
97–	10	5	1	0.0200	0.0179

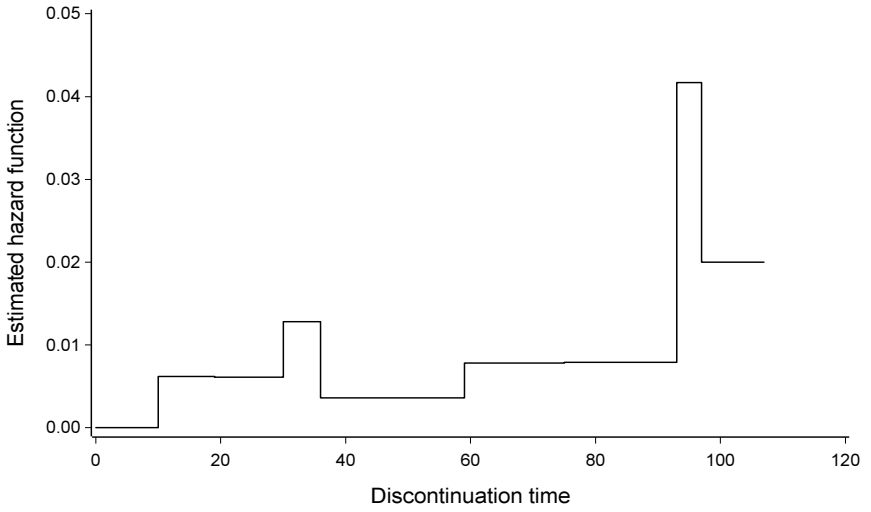


Figure 2.8 *Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.*

Figure 2.8 shows a plot of the estimated hazard function. From this figure, there is some evidence that the longer the IUD is used, the greater is the risk of discontinuation, but the picture is not very clear. The approximate standard errors of the estimated hazard function at different times are of little help in interpreting this plot.

In practice, estimates of the hazard function obtained in this way will often tend to be rather irregular. For this reason, plots of the hazard function may be 'smoothed', so that any pattern can be seen more clearly. There are a number of ways of smoothing the hazard function, that lead to a weighted average of values of the estimated hazard $\hat{h}(t)$ at death times in the neighbourhood of t.

For example, a *kernel smoothed* estimate of the hazard function, based on the r ordered death times, $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, with d_j deaths and n_j at risk at time $t_{(j)}$, can be found from

$$h^{\dagger}(t) = b^{-1} \sum_{j=1}^{r} 0.75 \left\{ 1 - \left(\frac{t - t_{(j)}}{b} \right)^2 \right\} \frac{d_j}{n_j},$$

where the value of b needs to be chosen. The function $h^{\dagger}(t)$ is defined for all values of t in the interval from b to $t_{(r)} - b$, where $t_{(r)}$ is the greatest death time. For any value of t in this interval, the death times in the interval $(t-b, t+b)$ will contribute to the weighted average. The parameter b is known as the *bandwidth* and its value controls the shape of the plot; the larger the value of b, the greater the degree of smoothing. There are formulae that lead to 'optimal' values of b, but these tend to be rather cumbersome. Fuller details can be found in the references provided in the final section of this chapter. In this book, the use of a modelling approach to the analysis of survival data is advocated, and so model-based estimates of the hazard function will be considered in subsequent chapters.

2.3.3 Estimating the cumulative hazard function

The interpretation of the cumulative hazard function in terms of the expected number of events that occur up to a given time, given in Section 1.3.3 of Chapter 1, means that this function is important in the identification of models for survival data, as will be seen later in Sections 4.4 and 5.2. In addition, since the derivative of the cumulative hazard function is the hazard function itself, the slope of the cumulative hazard function provides information about the shape of the underlying hazard function. For example, a linear cumulative hazard function over some time interval suggests that the hazard is constant over this interval. Methods that can be used to estimate this function will now be described.

The cumulative hazard at time t, $H(t)$, was defined in Equation (1.7) to be the integral of the hazard function, but is more conveniently found using Equation (1.8). According to this result, $H(t) = -\log S(t)$, and so if $\hat{S}(t)$ is the Kaplan-Meier estimate of the survivor function, $\hat{H}(t) = -\log \hat{S}(t)$ is an appropriate estimate of the cumulative hazard function to time t.

Now, using Equation (2.4),

$$\hat{H}(t) = -\sum_{j=1}^{k} \log \left(\frac{n_j - d_j}{n_j} \right),$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, and $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ are the r ordered death times, with $t_{(r+1)} = \infty$.

If the Nelson-Aalen estimate of the survivor function is used, the estimated cumulative hazard function, $\tilde{H}(t) = -\log \tilde{S}(t)$, is given by

$$\tilde{H}(t) = \sum_{j=1}^{k} \frac{d_j}{n_j}.$$

This is the cumulative sum of the estimated probabilities of death from the first to the kth time interval, $k = 1, 2, \ldots, r$, and so this quantity has immediate intuitive appeal as an estimate of the cumulative hazard.

An estimate of the cumulative hazard function also leads to an estimate of the corresponding hazard function, since the differences between adjacent values of the estimated cumulative hazard function provide estimates of the underlying hazard, after dividing by the time interval. In particular, differences in adjacent values of the Nelson-Aalen estimate of the cumulative hazard lead directly to the hazard function estimate in Section 2.3.2.

2.4 Estimating the median and percentiles of survival times

Since the distribution of survival times tends to be positively skewed, the median is the preferred summary measure of the location of the distribution. Once the survivor function has been estimated, it is straightforward to obtain an estimate of the *median survival time*. This is the time beyond which 50% of the individuals in the population under study are expected to survive, and is given by that value $t(50)$ which is such that $S\{t(50)\} = 0.5$.

Because the non-parametric estimates of $S(t)$ are step-functions, it will not usually be possible to realise an estimated survival time that makes the survivor function exactly equal to 0.5. Instead, the estimated median survival time, $\hat{t}(50)$, is defined to be the smallest observed survival time for which the value of the estimated survivor function is less than 0.5.

In mathematical terms,

$$\hat{t}(50) = \min\{t_i \mid \hat{S}(t_i) < 0.5\},$$

where t_i is the observed survival time for the ith individual, $i = 1, 2, \ldots, n$. Since the estimated survivor function only changes at a death time, this is equivalent to the definition

$$\hat{t}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) < 0.5\},$$

where $t_{(j)}$ is the jth ordered death time, $j = 1, 2, \ldots, r$.

In the particular case where the estimated survivor function is exactly equal to 0.5 for values of t in the interval from $t_{(j)}$ to $t_{(j+1)}$, the median is taken to be the half-way point in this interval, that is $(t_{(j)} + t_{(j+1)})/2$. When there are no censored survival times, the estimated median survival time will be the smallest time beyond which 50% of the individuals in the sample survive.

Example 2.8 Time to discontinuation of the use of an IUD
The Kaplan-Meier estimate of the survivor function for the data from Example 1.1 on the time to discontinuation of the use of an IUD was given in Table 2.2. The estimated survivor function, $\hat{S}(t)$, for these data was shown in Figure 2.4. From the estimated survivor function, the smallest discontinuation time beyond which the estimated probability of discontinuation is less than 0.5 is 93 weeks. This is therefore the estimated median time to discontinuation of the IUD for this group of women.

A similar procedure to that described above can be used to estimate other *percentiles* of the distribution of survival times. The pth percentile of the distribution of survival times is defined to be the value $t(p)$ which is such that $F\{t(p)\} = p/100$, for any value of p from 0 to 100. In terms of the survivor function, $t(p)$ is such that $S\{t(p)\} = 1-(p/100)$, so that for example the 10th and 90th percentiles are given by

$$S\{t(10)\} = 0.9, \quad S\{t(90)\} = 0.1,$$

respectively. Using the estimated survivor function, the estimated pth percentile is the smallest observed survival time, $\hat{t}(p)$, for which $\hat{S}\{\hat{t}(p)\} < 1-(p/100)$.

It sometimes happens that the estimated survivor function is greater than 0.5 for all values of t. In such cases, the median survival time cannot be estimated. It would then be natural to summarise the data in terms of other percentiles of the distribution of survival times, or the estimated survival probabilities at particular time points.

Estimates of the dispersion of a sample of survival data are not widely used, but should such an estimate be required, the *semi-interquartile range* (*SIQR*) can be calculated. This is defined to be half the difference between the 75th and 25th percentiles of the distribution of survival times. Hence,

$$SIQR = \frac{1}{2}\{t(75) - t(25)\},$$

where $t(25)$ and $t(75)$ are the 25th and 75th percentiles of the survival time distribution. These two percentiles are also known as the *first* and *third quartiles*, respectively. The corresponding sample-based estimate of the *SIQR* is $\{\hat{t}(75) - \hat{t}(25)\}/2$. Like the variance, the larger the value of the *SIQR*, the more dispersed is the survival time distribution.

Example 2.9 Time to discontinuation of the use of an IUD
From the Kaplan-Meier estimate of the survivor function for the data from Example 1.1, given in Table 2.2, the 25th and 75th percentiles of the distribution of discontinuation times are 36 and 107 weeks, respectively. Hence, the *SIQR* of the distribution is estimated to be 35.5 weeks.