Modeling Discrete Time-to-Event Data

Gerhard Tutz, Matthias Schmid

Springer International Publishing, 2016, x + 247 pages, $89.99 hardcover

ISBN: 978-3-319-28156-8

Readership: Applied statisticians, students of statistics and researchers.

The analysis of failure times is an important field in statistical research with applications in various areas, including manufacturing, demography, econometrics and epidemiology. There are many publications which deal with the analysis of failures in continuous time. However, failure times are often discrete in practice. This book presents statistical methods for the analysis of discrete failure times. It aims to introduce readers to basic and advanced concepts of discrete hazard modelling, to demonstrate the relationship between generalized linear models and hazard models and to consider applications from the social sciences, economics and biomedical sciences. The book also provides functions to apply the statistical methodology presented in the statistical software R.

This book offers 10 chapters which cover a range of areas, from well-known topics such as the life table, basic regression models and competing risk models, to special topics such as nonparametric modelling, tree-based approaches, predictor selection for high-dimensional models and methods for evaluating discrete time-to-event model fits. Throughout the book, the presented methods are demonstrated through real-world examples and applications, and relationships to survival analysis in continuous time are explained. At the end of each chapter, references to the literature are provided for further reading, along with a set of exercises for self-teaching. The relevant functions from the R package discSurv are listed and explainedin detail.

This book is indeed well-structured and is easy to access mathematically. It starts from basic concepts and leads to several modern extensions that allow for more flexible modelling of discrete time data. It would be a welcome improvement if further detailed illustrations of the R code and output are presented for some or all of the real-world examples. The book is certainly very useful for applied statisticians, students or researchers from various areas who are looking for a comprehensive overview of the statistical methods used in discrete failure time analysis, and for examples and ideas applicable to their specialised problems.

Shuangzhe Liu: shuangzhe.liu@canberra.edu.au

Program of Mathematics and Statistics

University of Canberra, Canberra, ACT 2602, Australia

Principles of Copula Theory

Fabrizio Durante and Carlo Sempi
CRC Press, 2015, 332 pages, £72.99, hardcover
ISBN: 978-1-439-88442-3

Readership: PhD-level researchers in Probability, Operations Research, Statistics and Biostatistics interested in mathematical foundations of copula theory.

This book represents a rigorous introduction to the theory of copula models, the biggest and the most thorough yet at that. The level of detail and rigor targets mathematicians working in probability theory. The exposition starts with an overview of the history of the subject followed by eight chapters laying out the theory of Copula models. The idea of the book was to present modern theoretical foundations for Copula models now that the field has seen over 50 years of research that has greatly accelerated recently with the development of applications in Finance, Operations Research, Statistics and Biostatistics. While the applied content is not given much attention in this book, theoretical concepts, including many that have not been covered in any of the other books on the subject, are the focus. Alternative proofs are given for key foundational results such as Sklar theorem, and a discussion that follows sheds further light on the interesting ideas and advantages that each particular angle of view reveals.

Chapter 1 is a housekeeping one that gives an overview of basic probability theory facts, definitions and notation that are useful in the rest of the book and focused on description of dependence. The basic theory of dependence is presented in chapter 2 with the Sklar theorem being the key. In the sequel, the book gradually prepares the reader and the necessary theoretical toolbox for modern ways to construct flexible families of copulas. Chapter 3 introduces Copulas from a measure-theoretic prospective. Chapter 4 presents a view of approximation theory where flexible copula subclasses are defined as approximating a bigger class through convergence to the limit models. Chapter 6 lays out specific flexible families of copulas. The last two chapters are devoted to generalizations such as quasi- and semi-copulas. The last chapter presents one of the few developments that explicitly targets applied areas of research, multivariate ageing for vectors of exchangeable lifetimes.

For a mathematically-oriented researcher in Statistics, Biostatistics and Applied Probability, and further in the specific applied fields of science, the book will serve as a reference for theoretical ideas potentially inspiring applied theory development. However, for the most part such developments remain beyond the scope of this book, which would necessitate further reading on the subject.

Alex Tsodikov *tsodikov@umich.edu*
Department of Biostatistics, School of Public Health, University of Michigan
Ann Arbor, MI 48109, USA

Principal Component Analysis Networks and Algorithms

Xiangyu Kong, Changhua Hu, Zhansheng Duan
Springer, 2017, XXII + 323 pages, £112.00, hardcover
ISBN: 978-9-811-02913-4

Readership: Applied statisticians.

This book is a unique collection of various algorithms for principal component analysis (PCA) and minor component analysis (MCA), respectively taken to mean the identification of larger and smaller eigenvalues of a covariance matrix. The authors thoroughly explore various aspects and extensions of related problems with emphasis on neural networks and automated learning. Devising fast and efficient algorithms for updating PCA and MCA as new data arrive without directly computing the new variance covariance matrix is certainly an important problem in signal processing and authors provide, in a single volume, a comprehensive account of these techniques in a variety of different contexts. All three authors are control engineers and thus unsurprisingly this book targets those audiences with implementation in neural network problems in mind. Even within that group, the book seems to be still highly specialized.

The book is not an easy read. Anyone willing to read it will require a serious effort and time commitment. The writing is terse and the book expects a lot from its readers in terms of mathematical expertise and competency in neural networks and control theory. Consequently, it does not qualify as an ideal textbook. That said, the book certainly presents a comprehensive account of the current state of art in this important area and thus as a research monograph, it has an immense value for the interested specialized readers.

The book is of limited interest to most statisticians, especially since no traditional statistical issues are addressed within this book. Most of the terminology used is also foreign to statisticians (for example, and rather curiously, the authors use the phrase “statistically irrespective” to refer to statistical independence). There are no examples, data, analyses or statistical comparative evaluations of techniques presented. Due to these major hurdles to overcome, the book is unlikely to attract the attention of traditional statisticians. This disconnect is not a reflection on these engineer authors since to begin with, statisticians were possibly not the target audience they had in mind when they wrote the volume.

The subject matter aside, unfortunately, in other non-technical matters, the book is not very reader friendly. The background material given in Chapter 2, although comprehensive, is essentially a list of results, stated without any clear context or motivation. Even statements of many of these results assume a lot in terms of prerequisites. There is no subject index and this makes the book inaccessible as a reference source. Separate reference list appears for each chapter and unfortunately these references are not in alphabetical order. Further, (especially in Chapter 2) some of the references from the list are not cited in the text at all. Unless one has read the entire book cover to cover, one would be at a loss to search for a specific topic, result, reference or contribution within the book. A subject index in such a specialized book is extremely important and I would have additionally preferred to see alphabetically listed references all in one place.

There are many (forgivable) typographical and grammatical errors throughout the book. These errors could have been avoided easily with a tighter editorial supervision. There are also a few technical errors/slips and these make certain mathematical statements either incorrect or ambiguous. Many of these errors are not obvious and I hope the authors will make available errata perhaps on their personal and/or publisher's websites.

Ravindra Khattree: *khattree@oakland.edu*
Department of Mathematics and Statistics

Oakland University, Rochester, Michigan 48309, USA

Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry

Susmita Datta and Bart J. A. Mertens,
Springer Frontiers in Probability and the Statistical Sciences, 2017, 295 pages, £82.00, hardback
ISBN: 978-3-319-45807-6

Readership: Graduate students, data analysts and statistics researchers.

While much attention has been given to the analysis of various biological data types, such as microarrays and next-generation sequencing data, less has been focused on molecules further downstream , such as those that arise from mass spectrometry experiments. In this edited collection, Datta and Mertens have assembled a collection of chapters from leading statisticians working in the areas of proteomics, metabolomics and lipidomics. Having such a reference available will be enormously benefcial to analysts encountering data from these platforms for the very first time.

Each chapter begins with a description of the technological platform being used, followed by the particular analytical challenge being addressed and its application to motivating datasets. There is a vast repertoire of techniques being utilized in these problems, ranging from linear model-based normalization and adjustment strategies to Bayesian modeling to compositional data modeling and to functional data analysis. The chapters are fairly self-contained, and the authors provide many references at the end of each chapter. Some of the highlights include a functional data analytic approach to mass spectrometric alignment by Srivastava, an exposition of the popularly used PeptideProphet algorithms by Gil and Datta, a discussion of the use of cross-validated predictors for omics data by Rodriguez-Girondo et al., and the first chapter on batch effects for mass spectrometric data by Mertens.

There is now an emerging pipeline for the analysis of high-throughput noisy biological data that roughly proceeds as

Preprocessing -> Normalization -> Differential expression -> Clustering/Classification -> Downstream analyses

What could have aided the book was an overview chapter describing such a pipeline and how each of the chapters fell into one or more of these categories. In addition, I anticipate that there will be analytical challenges common across the differing platforms that could be synthesized in such an overview chapter for the reader.

Finally, as with many bioinformatics books and research monographs, articles that describe and/or disseminate software will have major practical impact in the field. There is a nice chapter on MALDIquant by Gibb and Strimmer, which serves as an all-purpose tool for mass spectrometric data.

In conclusion, I expect that this book will soon become a go-to reference for any analyst who is seeking exposure to the analysis of data from mass spectrometric experiments.

Debashis Ghosh *DEBASHIS.GHOSH@ucdenver.edu*
Colorado School of Public Health, University of Colorado Anschutz Medical Campus
13001 East 17th Place, Aurora, Colorado 80045, USA

International Statistical Review

International Statistical Review (2017), 85, 2, 371–380

Book Reviews

Editor: Ananda Sen

Introduction to Data Science
Laura Igual and Santi Seguí
Springer, 2017, xiv + 218 pages, €48.14, softcover
ISBN: 978-3-319-50016-4

Readership: Undergraduate students in computer sciences and statistics.

This compact book is designed as an entry level text covering the area of modern computational statistics that is frequently referred to as data analytics or data science. The field is relatively new with yet not well-defined scope. In fact, it lies somewhere at the interface between computer science and statistics. Independently of varying scope that is seen in literature on data science, two aspects are frequently evident nowadays. Firstly, one deals with massive data with inherent complexity in structure. Secondly, the methods of analysis heavily depend on cleverly designed learning algorithms. Due to the complex structure of the data, typically, a significant amount of time is spent on visualisation techniques that allow extraction of patterns. All this is the premise of the current book. For the methods, a catchy phrase, 'Big Data' techniques, was coined to popularise the emerging field.

In the text, it is suggested that upper-level undergraduate and beginner graduate students in technical disciplines, including computer science and statistics, constitute the target audience. I would dispute this claim on two grounds. While the book discusses statistical inference in a few chapters, it uses very few methodological tools that are available to any undergraduate student who has taken a basic course in statistics and probability. Additionally, when the algorithmic aspects that are central for the methodology are presented, little is said about their foundations or about the principles that are implemented. I conclude that any upper-level undergraduate student in statistics or in computer science will be disappointed at the insufficient depth of presentation and the lack of formalism that they should have learnt at that stage of education. Thus, in my opinion, the best target audience of data analytics is entry level students-level undergraduate students in statistics and/or computer science, and I discuss the text with this proposition in mind.

As seen in other attempts to deliver data analytics to entry level students, one faces a few major challenges. Firstly, in order to demonstrate computational methods, one has to choose a software platform in which rather advanced algorithms can be implemented. There are unfortunately not so many choices available, and traditional presentation of many of them is oriented towards experts with advanced programming skills. The authors opted for 'Python', and such a choice, while justified (R-package being another natural choice), puts a lot of burden on a lecturer to introduce students to the principles of programming in this software. However, one can assume that students in statistics or computer science should not have an aversion to raw computing codes, and with proper guidance using 'Python' should not be an obstacle in understanding the material. As a matter of fact, the text is interlaced with a lot of code lines. It is done at the cost of explaining mathematical formalism. Clearly, such an approach will have limited appreciation outside the computer science community. The second challenge is a choice of data sets and examples that should motivate and illustrate concepts. In this respect, the book slightly disappoints since there is no well-documented library of the data while examples are somewhat hidden in the text. Most of the examples rely on data from various sources, but a dedicated website or at least a roadmap for datasets is necessary for a student willing to practise the techniques and programs. Finally, any text aimed at beginners must limit its methodological contents either in range or in depth. The authors have chosen the second option by giving quite a wide range of topics that are explained in a rather sketchy manner. In fact, the graphics together with the code take up more space than the text.

From a general point of view, the book provides an interesting account of a relatively wide range of data science methods to new undergraduate students. The choice of topics represents the subject quite well. The book has a good list of references, including several influential contributions in the field. A big drawback for its usefulness as a textbook is the lack of exercises or well-formulated examples for practising the introduced techniques. Any book in this area would greatly benefit from a list of projects that could test the ability of comprehensive utilisation of the methods. Overall, due to its limited methodological content, the book is more of a survey of methods accompanied by numerical tools available in 'Python'. Nevertheless, it can be utilised as an in-classroom textbook if sufficiently complemented by examples, descriptions of methods and computer labs.

Krzysztof Podgorski: *Krzysztof.Podgorski@stat.lu.se*
Department of Statistics
Lund University, Sweden

Educational Measurement for Applied Researchers
Margaret Wu, Hak Ping Tam and Tsung-Hau Jen
Springer Nature Singapore Pte Ltd., 2016, xiv + 306 pages, £82.00, hardcover
ISBN: 978-981-10-3300-1

Readership: Students, applied researchers and practitioners interested in educational measurement.

The main motivation of the book is to provide key concepts of educational and psychological measurement. As the authors stated in the preface, the book '*is not a comprehensive text on measurement*', and proper references to more advanced topics are provided in text. Instead, the ambitious goal of the book is to explain '*complex statistical analyses to a layperson*', so that the underlying statistics can be also accessible to non-mathematicians. Some basic statistics background (descriptive statistics, hypothesis testing and elements of probability) is however recommended.

The book starts with a preliminary chapter that introduces basic nuances of measurement and, hence, focuses on various aspects related to the use of classical test theory (CTT) and item response theory (IRT) that are popular in different fields ranging from psychology to education. It also presents two separate chapters on Bayesian and multidimensional IRT models. Whenever convenient, comparisons between models are made explicitly.

There are various aspects of the book that I found interesting. At the end of each chapter, the reader can find various exercises that may help in applying the introduced methods in some real case studies. In particular, there are some 'discussion points' that can potentially help the reader to think more in depth. The bibliography, which is arranged by chapter, provides in addition to the references cited in the text, a list of further readings for the readers interested in more advanced topics. In some parts of the text, basic instructions are also given on how to implement the given methodologies with popular computational software.

In my opinion, the book is suitable as a textbook at a graduate level for students and a reference guide for applied researchers interested in educational measurement.

Fabrizio Durante: *fabrizio.durante@unisalento.it*
Dipartimento di Scienze dell'Economia
Università del Salento
73100 Lecce, Italy

Algorithms for Data Science
Brian Steele, John Chandler and Swarna Reddy
Springer, 2016, xxiii + 430 pages, £49.99/$66.99, hardcover
ISBN: 978-3-319-45795-6

Readership: Students training to work on problems with a statistical content where there are large computational demands.

This 430-page book is divided into three parts. The first part is on data reduction, whereas information extraction and predictive analytics are the focus of the second and the third parts, respectively. The book has 12 chapters in addition to an appendix that contains the solutions to the exercises as well as pointers to accessing the Twitter API.

Part I of this text is primarily devoted to accessing data from the web, to the use of Hadoop for distributed processing across computer clusters and to low-level programming using Python. Parts II and III mix the use of Python with the use of R. Examples, which are worked through in detail, work with sets of data that are available from the web. These include US Behavioral Risk Factor Surveillance System (BFRSS) data, NASDAQ stock market data and data collected from Twitter.

Algorithms are, for purposes of this text, functions or a series of functions, with 'very little reliance on existing data analytic algorithms'. In Part I, matrix operations are used to handle correlation and least squares calculations, as part of a process that is designed to give students experience in Python coding, and in order to develop a deep understanding of 'fundamental, workhorse algorithms …'. While this may be useful as a training exercise, it does need to be made clear that for practical purposes, it is more preferable to use packaged functions that have been extensively tested in day-to-day use by experienced analysts. There is no attention, in Part I, to the checks that scatterplots, and residual and other diagnostic plots, can provide.

Parts II and III show a greater willingness to use existing functions from Python and R packages and do pay attention to the simple uses of residual plots. Adjusted R-square is used to measure model adequacy. Information statistics are not mentioned. Cross-validation is explained and used with nearest neighbour prediction. Attention is largely limited to relatively standard types of model—multiple regression, k-nearest neighbour prediction, k-means and multinomial naïve Bayes prediction. Parts II and III are a useful source of tutorials on relevant mathematical theory, followed by worked examples and exercises.

Code appears as code scripts. Especially in Part I, it would be helpful to structure substantial parts of the code into functions, making it easier to follow and reuse. I was unable to find any mention of the help that IPython (for Python), or RStudio (for R), provide, both for managing and documenting code and for structuring work into projects.

Overall, this is a useful book on mining of Big Data combining the fundamental principles, algorithms and data.

John H. Maindonald: *john@statsresearch.co.nz*
40 Futuna Close,
Wellington 6012
New Zealand

Equivalence: Elizabeth L. Scott at Berkeley
A.L. Golbeck
CRC Press, 2017, xxvi + 596 pages, $68, softcover
ISBN: 978-1-4822-4944-6

Readership: Persons interested in the history of statistics and in gender equity issues in academia.

Does your university campus have a childcare centre? Does it permit women to access the biggest and most powerful pieces of research equipment? Does it employ married couples in the same department? Does the University Staff Club admit both men and women? If your answer to these four questions are 'yes of course and why wouldn't it be so', then you owe something to the subject of this massive biography.

Elizabeth L. Scott (1917–1988) started her research career in astronomy, but made the transition to statistics as a result of wartime employment with the *Stat Lab* of Princeton University, working with Jerzy Neyman at the University of California, Berkeley. Over the next 30 years or so, she brought her quantitative skills to bear on tough and important problems in disciplines ranging from climatology (the effectiveness of cloud seeding) to health research (the effect of ozone depletion on skin cancer rates). But it is as a researcher and advocate for gender equity that Scott is most known and celebrated. The issue of gender equity was brought to a head at the University of California, Berkeley, in the late 1960s, where women were discriminated against in a systematic way on many fronts, including in particular the four issues raised at the start of this review. Scott's skills were in demand from the University administration in terms of collecting the data needed for an evidence base, fitting statistical models and advocating in behalf of women identified as experiencing salary discrimination.

Scott kept immaculate records, with every piece of paper that crossed her desk being dated and filed. This pattern of record-keeping brings to mind another great female statistician, Florence Nightingale. This ethic was clearly a huge bonus to the author, who provides excellent footnotes recording where each piece of evidence in the text may be found. She also quotes from the letters both to and from Scott, to vary the pace of the text.

The biography is academic rather than popular in style, being set in LaTex and therefore having the look and feel of a statistical monograph rather than a general interest biography. That said, the book does an excellent job at conveying the exciting and invigorating (and sometimes frustrating!) atmosphere that existed for students and academics in the 1960s at Berkeley. Scott's relationship with another giant of statistics, Jerzy Neyman, is handled skilfully and sensitively. The chapter on the exchange of letters between them while Scott was studying in Paris in 1945–55 was one of my favourites to read. The author chose to refer to Scott as Betty throughout the book, while referring to everyone else by their surname. The author did justify her choice at the beginning of the book, and I can follow the arguments that it makes sense to be consistent and to encourage a sense of identification of the reader with the subject.

There are over 500 pages of biography, organised in nine sections and 34 chapters. The main flow of chronological, with some gathering together of topics under themes, such as the chapter on students. Though this chapter sits at number 9, it covers Scott's entire 40-year career. There are a number of photographs gathered in the centre of the book.

I think this book will be of most interest to academics, as many of the issues are quite specific to the university sector. The way in which Scott was able to continue her research while simultaneously serving the University system through her gender discrimination work is exemplary and should be inspirational to the academic women of today. Women are still recognised as being under-represented at higher levels of academia, particularly in science, even though it is now 50 years after Scott commenced her investigations! It is still all too easy for those few women to spend time serving their institution at the expense of their own research careers.

However, some young female academics may find it either puzzling or disheartening to discover that Scott never married and lived at home with her mother her entire life. Such a lifestyle would have been rare at the time and would still be rare now. I do not think this aspect of Scott's life would find much resonance with 21st century young women.

Men and women who are interested in the history of statistics and in the history of gender equity in universities will want to own this book. There is inspiration to be gained and lessons to be learnt by those who still face gender inequity in academia today.

Alice Richardson: *alice.richardson@anu.edu.au*
ANU College of Medicine, Biology and Environment
Canberra, Australia

A Panorama of Statistics: Perspectives, Puzzles and Paradoxes in Statistics
Eric Sowey and Peter Petocz
John Wiley & Sons, 2017, xii + 313 pages, softcover
ISBN: 978-1-119-07582-0

Readership: Everybody.

The authors write in the acknowledgments section as follows:

'Our longest-standing debt is to Martin Gardner (1914–2010), whose monthly "Mathematical Games" columns appeared in *Scientific American* from 1956 to 1981. We were still in high school when we first discovered these short, lively and often quirky essays, which have drawn tens of thousands of people (us included) in to the pleasures and fascinations of mathematics.

Among the approximately 300 essays (which were collected into 15 books between 1959 and 1997), there are fewer than a dozen on probability, and just a handful in which a statistical concept is mentioned. Each of us had, in the past, wondered why no-one had yet brought Gardner's vibrant approach to statistics. When we discovered our common interest, it became an irresistible challenge to do it ourselves'.

Between 2003 and 2015, the authors published in *Teaching Statistics* 36 columns entitled 'Statistical Diversions'. The book under review is a revision, reorganisation and extension of this material. It contains six parts (Introduction, Statistical description, Preliminaries to inference, Statistical inference, Some statistical byways and Answers), divided in 26 chapters. Chapters 1–25 have a common structure: an overview of the theme and five questions relating to the theme. The final chapter provides the 'Answers', which is in effect a wide (107 pages) commentary on the questions. Its reader 'will be surprised by the variety of ways in which statistics can capture and hold your interest' (as the authors say in the preface).

The questions are presented in an interesting way that attracts attention. I give five examples (parts from Questions 1.1, 1.3, 9.3, 17.4 and 22.5) all somewhat differently formulated where and why did John Kerrich toss a coin for 10000 times? Do most people in London have more than the average number of legs? Who studied in 1872 statistically the efficacy of prayer, how and with which result? What do you think about the success of Paul the Octopus in predicting outcomes of games in the Soccer World Cup 2010? Why did Gosset use the pseudonym Student?

According to the back cover: ‘This book is a stimulating panoramic tour—quite different from a textbook journey—of the world of statistics in both its theory and practice, for teachers, students and practitioners’. This tour has 25 lively written stages, containing interesting and useful knowledge in statistics and its history, also providing challenges and even entertainment.

Jorma K. Merikoski: *jorma.merikoski@uta.fi*
Faculty of Natural Sciences
FI-33014 University of Tampere
Finland

Improving Population Health Using Electronic Health Records. Methods for Data Management and Epidemiological Analysis
Neal D. Goldstein
CRC Press, 2017, xix + 254 pages, £45.99, softcover
ISBN: 978-11381-9637-7

Readership: Novice and experienced researchers, healthcare professionals and public health practitioners interested in working with electronic health records or similar secondary data.

Over the years, the amount of routinely collected health data including electronic health records (EHR) has increased considerably. There has also been an increasing interest in using these data for research purposes. As these data had not been collected originally for research purposes, the research utilising such so called secondary data has some special challenges to be considered. The author of the book states in the preface that a hands-on guide to performing research with EHR data (from clinical epidemiology point of view) did not exist. This book aims to fulfil that need.

The two initial chapters provide an introduction to the basic concepts and the structure of the presentation in the book. The idea is to give a view of the whole research process involving EHR and to describe each step of this process separately.

The book is divided into three parts. First part of the book (Chapters 3–6) focuses on understanding the data, the second part (Chapters 7–12) describes how to conduct research and the third part (Chapters 13–15) delves into interpretation and implementation of the results.

Chapter 3 deals with planning of research that uses EHR and includes a brief discussion about setting the research questions, research ethics and data availability. In Appendix 1, there is a ‘secondary data research planner’ that helps to document several essential issues encountered during the research process. Privacy issues (and informed consents) receive only a very limited attention, although these may be the biggest threats to EHR-based research in the near future. Chapters 4–6 are more technically oriented and describe how to export, organise, link and preprocess the data. These include many of the essential tasks encountered in practice, and I admire the author’s ability to give such a comprehensive presentation of the selected issues. Also, example code for data manipulations in R is given in Appendix 2 (unfortunately, I was not able to find those codes in a separate file available online at the time of writing this review) The most obvious limitation of these chapters is that more complex preprocessing including data abstraction and theoretical ideas related to enrichment of data using external information is almost entirely missing, even though all readers should familiarise themselves with these possibilities documented in the earlier literature (Sund et al. 2014).

Part 2 of the book is an introduction to basic epidemiological concepts and research designs and in that sense is not limited to secondary EHR data. Chapter 7 describes cross-sectional, case–control and cohort study designs. Chapter 8 focuses on basic types of epidemiological measures. Chapter 9 warns the reader against the most common types of bias and confounding. Again, the author has succeeded in providing extremely clear presentation of complex issues that works very well in introductory-level text. Chapters 10–12 contain more hands-on type material encountered in statistical and epidemiological analyses.

Chapter 13 discusses the interpretation of the results obtained from the statistical analyses for causal inference. Short notes about testing, adjustment, logistic and proportional hazards regression and p-values are certainly interesting and important things to know, but it may be somewhat difficult to see the bigger picture. Chapters 14 and 15 are very general ones presenting some ideas related to presentation, publication and dissemination of ideas. Those are certainly beneficial and good-to-know practices for junior researchers, but the actual connection to EHR data in general is not clear in these chapters.

In summary, this is an excellent book for novice researchers who want to work with secondary EHR data and an interesting read to even more experienced researchers, statisticians and data analysts. The style of writing is clear, and the selection of issues to be presented is excellent. This is an introductory-level book. Pragmatic analysis requires more advanced skills and acquired knowledge, but the ‘vocabulary’ presented in the book is already a great initiation into the research with EHR data and opens up avenues for efficient collaboration.

Reijo Sund: *reijo.sund@helsinki.fi*
Centre for Research Methods, Faculty of Social Sciences
University of Helsinki, P.O. Box 18, FI-00014, Finland &
Kuopio Musculoskeletal Research Unit (KMRU), Institute of Clinical Medicine
University of Eastern Finland (UEF), Kuopio, Finland

References

Sund, R., Gissler, M., Hakulinen, T. & Rosen, M. (2014). Use of health registers. In *Handbook of Epidemiology*, 2nd edition, Eds. Ahrens, W. & Pigeot, I. New York: Springer, pp. 707–730.

Survival Analysis in Genetics and Medicine
Jialiang Li and Shuangge Ma
CRC Press, 2013, 381 pages, £78.99, hardcover
ISBN: 978-1-43989-311-1

Readership: Researchers working on areas in survival analysis, specifically interested in applications of time-to-event studies in genetics.

Modelling and analysis of lifetime data is an important aspect of statistical work in a wide variety of scientific and technological fields. The field has expanded rapidly in recent years and applications of statistical procedures ranging from investigations of the durability of manufactured items to studies of human diseases and their treatment. The objective of the book is to present statistical procedures useful for the analysis of lifetime data with a focus on application in genetics.

The book is organised into six chapters. Chapter 1 presents basic concepts in survival analysis. Specific features of survival studies such as censoring and truncation were also discussed. In Chapter 2, non-parametric estimation of survival function is presented. Comparison of two lifetime distributions using non-parametric inference procedure is discussed. The chapter also discusses parametric and semiparametric regression models in survival studies. Unlike other textbooks in survival analysis, the chapter does not discusses basic principles of regression analysis, but it focuses on special models employed in survival studies.

Chapter 3 is devoted to modelling and analysis of interval censored lifetime data. The non-parametric estimation procedures of interval censored data are discussed. Various semi-parametric regression models for the analysis of interval censored data are also explained.

Chapter 4 gets into advanced topics such as non-parametric regression, multivariate survival models, cure rate model and Bayesian framework. The non-parametric regression in survival studies are somewhat non-standard and are typically not discussed in other textbooks. The analysis of competing risks data using cause-specific hazards and cumulative incidence function are discussed. The basic tools for Bayesian analysis of lifetime data are also explained.

The diagnostic procedures useful in medical studies are discussed in Chapter 5. The elementary procedures such as sensitivity-specificity analysis is presented. In addition, time-dependent diagnostic methods under censoring are discussed. This is quite different from the traditional diagnostic procedures based on hazard rates employed in regression models. Several clinical examples are presented to illustrate the utility of the procedures.

Chapter 6 is dedicated to study applications of survival analysis techniques in genetics, which is one of objectives of the book. The set of statistical genetics procedures for censored survival outcome data are explained. Various topics like marginal association measures, multivariate prediction models and hierarchical structures in survival analysis with high-dimensional covariates are presented. The book ends with a comprehensive bibliography and index.

The great strength of the book lies in its comprehensive treatment of both classical and novel methods, covering almost all aspects of survival analysis that biostatisticians are confronted with in everyday practice. The text is very well organised, and both writing style and notation are remarkably homogeneous. The readers will appreciate the inclusion of clinical studies as applications in the book. Anyone already familiar with analysis of survival data should own a copy of this text, as it serves as a wonderful reference for the most recent advances in the field. This book is a great reference tool for both researchers applying the current survival analysis methods and for statisticians developing new methodologies in statistical genetics. This book is an excellent collection on current survival analysis methods and can lead the audience to learn about them and discover appropriate literature. Practitioners can find easy access to many advanced survival methods through this book.

Frailty models are extensively employed in survival studies to explain heterogeneity in the population as well as to model dependence among variables. The effect of unobserved covariates on survival time, which is common in statistical genetics, is usually studied by frailty models. The only weakness of the book is the exclusion of the topic frailty models.

Despite some minor limitations, the authors have succeeded in providing a well-written and well-organised survival analysis textbook useful in statistical genetics. I will definitely recommend this book to researchers and practitioners working in genetics and medicine.

P. G. Sankaran: *Sankaran.p.g@gmail.com*
Department of Statistics
Cochin University of Science and Technology
Cochin 682022, India

Statistical Rethinking: A Bayesian Course with Examples in R and Stan
Richard McElreath
CRC Press, 2015, 469 pages, £67.99, hardcover
ISBN: 978-1-482-25344-3

Readership: Individuals with an understanding of Bayesian principles and readers interested in principles of modern Bayesian statistical modelling.

Statistical rethinking is an introduction to applied Bayesian data analysis. The principal audience comprises of masters and PhD students and researchers in natural and social sciences, with some knowledge of calculus, linear algebra, probability, statistics and R programming. The book follows a computational approach that combines explanation with R and Stan code examples. It uses more computational mathematics than formal mathematical statistics. The main topic of the book is generalised linear multilevel modelling (GLMMs) model following a Bayesian approach. It covers the basics of regression analysis through multilevel models, missing data and Gaussian process models. Interesting real data sets and examples and exercises can be found at the end of each chapter. The book contains a good selection of extension activities, which are labelled according to difficulty. There are occasional paragraphs labelled 'rethinking' or 'overthinking' that contain finer details. The presentation is replete with metaphors ranging from the 'statistical Golems' in Chapter 1 through 'Monsters and Mixtures' in Chapter 11 and 'Adventures in Covariance' in Chapter 13.

The Chapter 1—The Golem of Prague—is intended to set the framework of modes that represent natural or social phenomena. Chapter 2—Small Worlds and Large Words—and Chapter 3—Sampling the Imaginary—introduce Bayesian statistics and the tools to perform Bayesian calculations. Chapter 4—Linear Models—introduces simple linear regression, while Chapter 5 contains an overview of multiple regression tackling subjects such as multicollinearity. Omitted variables and post-treatment bias are covered in the same chapter. Chapters 6 and 7 cover overfitting and interactions. Chapter 8 provides an overview of Markov chain Monte Carlo methods with considerable insight into some challenges encountered (such as problems emerging from using flat priors used with variance parameters). The MCMC methods in this chapter are used to fit non-linear models in Chapter 10.

Chapter 9—Big Entropy and Generalized Linear Models—introduces maximum entropy, while Chapter 10—Counting and Classification—Chapter 11—Monsters and Mixtures—Chapter 12—Multilevel Models—Chapter 13—Adventures in Covariance—and Chapter 14—Missing Data and Other opportunities—provide an overview of a wide range of practically important methods ranging from standard generalised linear models through to mixtures, multilevel models, missing data and spatial correlation modelled using Gaussian process. Chapter 15—the final chapter is titled *Horoscopes* and presents some reflections on how to design, conduct and document research.

The book is accompanied by a R package, rethinking, which is available on the author's website and GitHub repository. The core of this package consists of two functions, map and map2stan, that allow many different statistical models to be built up from standard model formulas. The function map2stan builds a Stan model that can be used to fit the model using MCMC sampling. Some of the more advanced models in the last chapter are written directly in Stan code. There is also a technical manual with additional documentation. This book makes a valuable contribution to the literature of Bayesian statistical modelling combining theory, explanation and code.

Diego Andrés Pérez Ruiz: *diego.perezruiz@manchester.ac.uk*
School of Mathematics
University of Manchester, Oxford Road, Manchester M13 9PL, UK