International Statistical Review

International Statistical Review (2015), 83, 1, 160–174 doi:10.1111/insr.12095

Book Reviews

Editor: Ananda Sen

Bayesian Phylogenetics: Methods, Algorithms and Applications
Edited by Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis
Chapman & Hall/CRC, 2014, xxx + 365 pages, £63.99/$99.95, hardcover
ISBN: 978-1-4665-0079-2

Readership: Quantitatively oriented evolutionary biologists, statistics graduate and advanced undergraduate students, as well as researchers in the field.

Phylogenetics is involved with the estimation of evolutionary relationships among biological organisms, typically using molecular sequence data from living organisms, but also sometimes morphological characteristics from both living organisms and fossils. In many situations, the primary biological questions of interest are encapsulated in the estimation of a tree (part of a great Tree of Life) as a graphical structure that represents the past evolutionary history of speciation and descent of a group of organisms. Sometimes the biological questions are not directly related to the relationships among species, but an estimated phylogeny is critical to account for dependence due to shared history from common ancestors in data measured across different species. The history of phylogenetic methods traces back to the middle of the previous century. Statistical methods based on likelihood models for molecular sequence evolution originated in the early 1980s, and Bayesian approaches began in the mid-1990s, exploding in popularity throughout the 2000s following the release of powerful software packages specifically designed to carry out Bayesian phylogenetic analyses. There are many previous books on the topic of phylogenetics; some of these are statistical in point of view, and some of these devote a chapter or so to the Bayesian approach. However, *Bayesian Phylogenetics* has the distinction of being the first published text devoted solely to the Bayesian approach to phylogenetics.

The editors describe *Bayesian Phylogenetics* as providing a 'snapshot of current trends in Bayesian phylogenetic research' with a 'heavy emphasis on model selection' and other important themes including new approaches to improve MCMC mixing, divergence time estimation, the movement towards more biologically realistic models and the interface with population genetics. This is a fairly accurate portrayal of the book, which highlights both its strengths and weaknesses. The included chapters are generally well-written and informative about their topics, all of which are current, but *Bayesian Phylogenetics* has gaping holes of omission with regard to an informative introduction to the field and a broader survey of methods and topics that are outside the direct interests of the editors.

The first chapter provides an introduction to the contents of the book but fails to provide an overview of phylogenetics or the Bayesian approach to phylogenetics. A statistician who seeks an introduction to the field without previous exposure to phylogenetics will need to look elsewhere before finding this book to be useful. The closest resemblance to an introduction is curiously located in Chapter 7. This chapter contains a nice history of the earliest development of Bayesian phylogenetic methods and brief descriptions of tree representations and models of sequence evolution. An expansion of this part of the chapter as a separate chapter that introduced both the foundations and the history of the development of methods, as

well as locating this introduction at the beginning of the book (right after the short overview) would have improved the usefulness of the book immensely. Unfortunately, the remainder of Chapter 7 elects to compare a relatively new sequential stochastic approximation Monte Carlo approach to early conventional software implementations that have either been in stasis for more than a decade or have been replaced by a newer version much more recently than the 2001 version used.

Chapter 2 contains a nice summary of recent literature results that demonstrate unrecognised effects on branch length and divergence time estimation from what are supposedly only vaguely informative prior distributions. Chapters 3–6 extend for 100 pages and provide a wonderful description of many modern computational methods for estimating marginal likelihoods in a phylogenetic context. Readers interested in marginal likelihood calculations, even in other contexts, will find a wealth of helpful information. Readers interested in the broad array of current research in Bayesian phylogenetics will be disappointed that more than a quarter of the total content of the book is so narrowly focused. Chapter 8 describes a sequential Monte Carlo approach that can also be of interest beyond phylogenetics. Chapters 9–13 each outline an application of Bayesian inference for specific phylogenetic questions.

Bayesian Phylogenetics makes no claim to being comprehensive, but the list of topics that are excluded is both lengthy and contains areas under intensive active research by multiple groups. The omission of these topics lessens the impact that the book have might otherwise had. Missing topics include: co-speciation and coevolution, ancestral state estimation, detection of positive selection, joint estimation of phylogeny and alignment, gene-tree/species-tree estimation, species networks, estimation of the rate of speciation, trait evolution and the effects of protein structure on models of molecular evolution.

Despite its weaknesses, *Bayesian Phylogenetics* will be a useful resource for many researchers in the field and for statisticians interested in joining the game. For this latter group, Felsenstein's *Inferring Phylogenies* (2004) is a better place to get started for an overview of statistical phylogenetics, albeit from a markedly non-Bayesian point of view, before trying to digest the material here. Furthermore, the primary scientific literature will continue to be the only available resource for many important and interesting topics in Bayesian phylogenetics.

Bret Larget: *brlarget@wisc.edu*
Departments of Statistics and of Botany
1300 University Avenue, University of Wisconsin, Madison, WI 53706, USA

References

Felsenstein, J. (2004). *Inferring Phylogenies.* Sunderland, MA, USA: Sinauer Associates, Inc.

Biostatistics Decoded
A. Gouveia Oliveira
Wiley, 2013, x + 346 pages, £40.00/€50.10, paperback
ISBN: 978-1-119-95337-1

Readership: Biostatisticians, clinical researchers, health professionals, biostatistics and statistics graduate students.

As stated in the preface, the purpose of *Biostatistics Decoded* is to present statistical theory and biostatistical methods and applications through a different approach. In the next discussions are the four claims along with a review regarding how well those claims are met.

Claim 1: Integrates topics usually covered in separate books (sampling, study design, statistical methods)

This integration begins with the introduction in Chapter 1 as it gets directly at key terms of population, study design, sampling and statistical inference with the 'BIG IDEA' approach. There is good emphasis up front on the need for a clear definition of the population and the need for representativeness of sample.

This book is organised in short chapters, providing the reader with smaller digestible content. After the introduction, basic concepts and statistical inference chapters are reviewed as needed; the reader can be more selective in exploring the sampling methods, study design and statistical methods according to the area of application of interest.

Claim 2: Strictly non-mathematical approach

A cursory flip through pages does show a few formulas. If you wish to lessen the math, you might avoid some of the shaded grey boxes that appear throughout, which do go into more detailed discussion that is not often needed to grasp the bigger picture. One example is Section 2.8, which provides a detailed justification for the use of the $n-1$ divisor in the sample standard deviation formula. This book also includes statistical tables and explains how to use them to compute probabilities and p-values. If the focus is not on math, then use a statistical computer package to give the test statistic and p-value, and instead provide more focus on understanding what methods to apply, what these results are measuring and how to interpret them in context.

Chapters 2 and 3 do cover many key introductory statistical topics and do avoid nearly all mathematical expressions, but both could be enhanced with some nice visual pictures to better convey the descriptions, and both could also use more examples with a clinical context (versus generic variable A and B). In general, the book should promote graphing data first before 'compressing' data into just a few numbers.

Claim 3: Focus on topics for clinical researchers (simple and advanced)

Indeed the later chapters do provide more context examples and overviews to better understand the various studies, designs, tests and issues that arise in the medical field. The focus is on breadth rather than theoretical depth, covering from simple inference on means and proportions to logistic regression, factor analysis and meta-analysis: from sample size estimation to measurement issues, adaptive clinical trials and secondary analyses.

Some inappropriate usage of data exists that can be misleading, especially to the clinical researcher. For example, in Section 2.6, the variable patient chart number was treated as a numerical (quantitative variable), a histogram was made and summary measures (median, etc) were computed, which are not appropriate. Using a medical variable (e.g. pulse rates) would be better to show the main idea in this section.

The most disconcerting inaccuracy lies in the illustration of interpretation of confidence interval. On p. 44, the following statement appears:

If sample mean is 210, then P[210 – 2(std errors) <mu <210 +2(std errors)] = 0.95.

This is clearly incorrect as there is nothing random in this probability statement (the probability is 0 or 1). Throughout, the book continues to use the probability idea even when using the estimated standard error in making a confidence statement.

Claim 4: Theory is built on basic concepts so the reader can understand the conditions required and limitations
This claim is achieved to some extent, but the accuracy of the presentation has sometimes been hampered due to oversimplification. A case in point appears on p. 42, where a couple of bullets state the following:

- *Sample means have a normal distribution, regardless of the distribution of the attribute, but on the condition that THEY are large* (THEY refers to the sample size, not that sample mean, which is unclear).
- *Small samples have a normal distribution only if the attribute has a normal distribution* (the authors meant to state this for sample means).

This book starts by presenting basic concepts, such as the properties of means and variances, the properties of the normal distribution and the central limit theorem as building blocks, but much of the building block pieces are presented generically, without context, few visuals, and some statistical inaccuracies. However, because of its breadth, this book can be a great reference and contains very important pieces that are not always well emphasised or addressed in other texts.

Brenda Gunderson: *bkg@umich.edu*
Department of Statistics, University of Michigan
1085 South University, Ann Arbor, MI 48109-1107, USA

Nonparametric Statistics: A Step-by-Step Approach
Gregory W. Corder and Dale I. Foreman
Wiley, 2014, xiv + 267 pages, £63.50/€76.20, hardcover
ISBN: 978-1-118-84031-3

Readership: Students of statistics and researchers in other disciplines, especially in social sciences.

The nine chapters in the book cover the basic introductory topics in non-parametric statistics. The focus is on comparing two or more populations measured on nominal/ordinal as well as interval/ratio scales. The final chapter discusses a test for randomness based on the runs statistic. The authors start off by defining their parameters as the assumptions of traditional introductory statistical methods. This unusual use of language tripped me up from the very beginning and I never really regained my footing. Many other loose uses of language remain from Corder & Foreman (2010), for instance the distinction between data and variables, samples and observations and critical values and p-values. Columns of squared, cubed and fourth power deviations for calculating skewness and kurtosis in Chapter 2 are followed by hand calculations for many of the non-parametric methods. While I do think there is some value in at least discussing the hand calculations to provide insight into the working of test statistics, it is unlikely that this is how most students will be tackling their own datasets.

An eight-step procedure is used for each test, typically comparing the test statistic to a tabulated critical value. The sections on uses of each test in the literature are helpful, with references from the 1990s and 2000s in general.

A handful of drill-style questions accompany each chapter. Solutions follow immediately. The authors present plenty of SPSS 21 output, but they are silent on the issue of how to get confidence intervals even though they are recommended by the APA.

The cover says that this is the second edition (2014). This journal carried my review (Richardson, 2010) of Corder & Foreman (2010). The book has gained 20 pages, due largely to the addition of two tests. Appendices cover SPSS basics and 10 tables of critical values. The preface refers to web-based tools including a decision tree in the form of a Prezi that links to YouTube videos of SPSS 21 use. Unless you are desperate for the extra tests or this online material, you do not need to buy the second edition if you already own the first.

Alice M. Richardson: *Alice.Richardson@canberra.edu.au*
Faculty of Education, Science, Technology & Mathematics
University of Canberra, ACT 2601, Australia

References

Corder, G.W. & Foreman, D.I. (2010). *Nonparametric Statistics for Non-statisticians: A Step-by-Step Approach.* Hoboken, NJ: Wiley.

Richardson, A.M. (2010). Review of "Nonparametric Statistics for Non-Statisticians". *Int. Stat. Rev.*, **78**, 451–452.

Circular Statistics in R

Arthur Pewsey, Markus Neuhäuser, and Graeme D. Ruxton
Oxford University Press, 2013, 208 pages, $49.95, paperback
ISBN: 978-0-19-967113-7

Readership: Graduate students, researchers and professionals across biological, social, physical, medical and earth sciences.

The book is an easy to access guide suitable for all readers with varying experiences with circular statistics and R.

The book has eight chapters. It covers a wide range of circular statistics topics from graphics and summary statistics to circular distribution theory, and from inference to correlation and regression, with an emphasis on fitting a broad range of models and hypothesis testing. It introduces R's circular package, web-based R code and resources. It uses R not only as a data analysis method but also as a learning tool. It is not a textbook; however, it features many examples and case studies with intensive R illustrations and graphics. It also includes an appendix, which briefly comments on and features additional reading from the only six other books available. The book has an accompanying webpage http://circstatinr.st-andrews.ac.uk/, which provides access to resources, including text files with R codes and data sets discussed.

Indeed, this book is a much needed up-to-date guide to the theory and practice of circular statistics, for graduate students, researchers and professionals working in a wide range of scientific disciplines, and for applied statisticians.

Shuangzhe Liu: *shuangzhe.liu@canberra.edu.au*
Faculty of Education, Science, Technology & Mathematics
University of Canberra, ACT 2601, Australia

Displaying Time Series, Spatial and Space-Time Data with R
Oscar Perpinan Lamigueiro
CRC Press, 2014, vii + 200 pages, £49.99, hardback
ISBN: 978-1-4665-6520-3

Readership: Statistics undergraduate and graduate students, researchers in spatial statistics, environmental and geo-sciences, and computer science students.

A well-known deficiency of R–a popular free software environment for statistical computing and graphics–is its limited graphical capabilities, especially when compared with MATLAB, its main commercial competitor. A need for high-end graphical visualisation is particularly needed when working with space-time data. Thus, a researcher in this field will particularly benefit from this compact account of some elegant visualisation techniques implemented in R for time, space and timespace data.

The focus of the book is not on methodological data analysis but rather on purely graphical descriptive presentation of the data. While no attempt has been made to present a complete set of graphical methods, selected examples are very informative in presenting visualisation capacities of different R packages that have been developed and are currently available from CRAN. The author is knowledgeable on the subject and has ability to present essentials without obstructing them with technical details. Descriptions of the methods are somewhat sketchy, which sometimes makes it difficult to fully appreciate the associated features. Colorful illustrations that nicely complement the descriptive aspects of the book help in the process of understanding.

The book is organised into three chapters: *Time Series*, *Spatial Data* and *Space-Time Data*. Each chooses several visualisation techniques that are applied to present examples of relevant real life data. The use of real data makes the presentation more convincing. The book would, however, benefit from having a set of computer labs that would allow a reader to perform her or his own hands-on exploration of the presented computational tools. The text contains a substantial amount of references to the material that should help in enhancing the knowledge on the selected applications and methods. The book website and its repository freely provide the complete codes needed. All datasets are available without restriction for public use either from the repository or from other publicly available websites.

To summarise, the book is a valuable source of graphical visualisation analyses in R that would be very much appreciated by anyone considering non-commercial alternatives to MATLAB for her or his work on space and/or time data.

Krzysztof Podgorski: *Krzysztof.Podgorski@stat.lu.se*
Department of Statistics
Lund University, Box 743, 220 07 Lund, Sweden

A Chronicle of Permutation Statistical Methods: 1920–2000, and Beyond
Kenneth J. Berry, Janis E. Johnston, and Paul W. Mielke Jr.
Springer, 2014, xix + 517 pages, €116.59, hardcover
ISBN: 978-3-319-02743-2

Readership: Anyone with an interest in the history of statistics and a good knowledge of statistical methods.

The authors of this book state that the purpose is 'to chronicle the birth and development of permutation statistical methods over the approximately 80-year period from 1920 to 2000' (p. viii). To this end, the authors have consulted an impressive number of articles, books and other sources, with the reference list running over 56 pages and containing 1 498 entries. After a brief introductory chapter that explains the basics of permutation statistical methods and gives an overview of the contents of the remaining chapters, the main part of the book consists of four chapters devoted to detailing the development of permutation statistical methods during each of the approximately 20-year periods 1920–1939, 1940–1959, 1960–1979 and 1980–2000. The final chapter gives a brief overview of the development of permutation statistical methods beyond 2000.

The text details the important discoveries in the area of permutation-based statistical methods, the persons who made the discoveries, as well as the historical and social context in which the discoveries were made. This is accompanied by fact boxes, mainly containing short biographies of individuals mentioned in the text. The permutation statistical methods are described in such detail, including necessary formulas, that it often should be possible for the reader to apply the methods directly on his or her own datasets. Many of the early important discoveries in the area of permutation statistical methods were made by people trained in and primarily working in other disciplines than statistics. This means that many of the presented methods also had practical motivations and applications based on real world datasets that the researcher had to handle. Especially, research in the agricultural area seems to have given rise to real world problems that were solved by newly discovered permutation statistical methods. Some of these real world data sets are also presented in the text and the calculations given in detail. Because the practical implementation of permutation statistical methods is very much dependent on access to cheap high-speed computing, the development of computing is an integral part of the text. Although this is often interesting and gives valuable insight into the practicalities of the application of permutation statistical methods to real world datasets, sometimes it goes too far outside the scope of the book's subject, like when it provides fact boxes about Google and with biographies of the Google founders Larry Page and Sergey Brin. One wonders what these have achieved in the area of permutation statistical methods.

Although this is a very impressive work, there are some minor flaws, such as repeating blocks of texts, indicating that the text has not been properly edit-checked. For example, on p. 12, at the end of the text in one paragraph, it says: 'Table 1.2 contains the 1831 per capita relief expenditures, in shillings, for 36 parishes in two counties: Oxford and Hertford.' Then, the next paragraph on the same page starts: 'The relief expenditure data from Oxford and Hertford counties are listed in Table 1.2.' Likewise, on p. 32, it says: 'Geary investigated the ways that cancer mortality rates varied with the consumption of potatoes in Ireland [. . .]'. Then, just a few rows down on the same page, it says: '[. . .] Geary considered potato consumption and the incidence of cancer deaths in Ireland.'

Despite these shortcomings, it should be emphasised that this book is a very impressive work with a unique, comprehensive in-depth coverage of the birth and development of permutation-based statistical methods that should prove to be invaluable for anyone interested in this topic. It is, however, hard to agree with the back cover's assertion that its 'non-mathematical approach makes the text accessible to readers of all levels'. On the contrary, a good knowledge of statistical methods is necessary to fully appreciate the text. But for anyone with this knowledge and an interest in permutation statistical methods and the history of statistics, this is a must-have book. It should also be useful for practitioners who are using permutation-based statistical methods and are interested in understanding the foundations and the motivations behind the usage of a specific method. In short, it is a highly recommended book for the audience indicated in the readership category.

Andreas Rosenblad: *andreas.rosenblad@ltv.se*
Center for Clinical Research Västerås, Uppsala University
Västmanland County Hospital Västerås, S-721 89 Västerås, Sweden

Statistical Methods for Survival Data Analysis
Elisa T. Lee and John Wenyu Wang
Wiley, 2013, 513 pages, £86.95/€104.40, hardcover
ISBN: 978-1-118-09502-7

Readership: Biostatistics and Statistics graduate students, researchers in the area of survival analysis.

This fourth edition of the original 1980 text by Elisa Lee enhances a book that has been a solid introductory survival analysis text for over three decades. Interim editions in 1992 and 2003, when co-author John Wenyu Wang was added, have each made substantial improvements to previous editions. Much of the book remains accessible to readers with college algebra but not calculus. As in previous editions, examples with real data are abundant, and concepts are clearly explained. The layout of the text and equations, and frequent graphs and tables, all contribute to a pleasing and welcoming look.

The fourth edition adds several new topics and notably has added examples using the R software package in addition to the existing SAS and SPSS examples. Chapter 4 on non-parametric methods of estimating survival functions now includes the Nelson–Aalen estimator. Chapters 6 and 7 on parametric survival distributions and estimation of their parameters, respectively, have added the parameterization for the generalised gamma distribution used in SAS software. Chapter 7 has also added the Gompertz distribution. Chapter 11 on the Cox model has added substantial new material on model assessment, including methods for checking the proportional hazards assumption and the functional form of continuous covariates. The methods presented include plots based on Schoenfeld residuals in R and plots based on martingale residuals using the ASSESS option of the PHREG procedure in SAS. Chapter 12 on identifying prognostic factors in non-proportional hazards models has added several pages on methods for correlated data, specifically frailty models, both based on the Cox model and based on parametric accelerated failure time models. New exercises have been added at the end of each chapter. Some material has been dropped, including the bibliographic remarks at the end of each chapter.

Although much of the material presented is standard, I would take issue with a few items in the book: First, the presentation of the Kaplan–Meier (KM) estimator using interpolation between failure point estimates of S(t) rather than the maximum-likelihood-based step function is inappropriate. In Section 4.1, an example dataset of time to death in months for 10 lung cancer patients (with no censoring) is presented with both the step function version and the 'connect the dots' version of the KM plot as two viable alternatives The text goes on to calculate the sample median, which equals eight by taking the mean of the fifth and sixth ordered observations (both equal to eight), and which is the same as the median observed from the step function KM plot. However, the interpolated version of the KM yields a sample median of 7.3, which is clearly wrong; an unfortunate misplacement of a point on the KM graph (Figure 4.1b) further confuses the issue. This improper suggestion again appears in Figure 4.3(b). Second, the short discussion on missing data in Section 10.1 gives poor advice, including the suggestion to replace missing values of continuous variables with the mean and to replace missing categorical variables with a new indicator variable for a category of missing data. Multiple imputation deserves mention, and the other methods, albeit useful expedients in cases

of a few missing values deserve a healthy dose of caution with non-trivial amounts of missing data. Third, Chapter 13 gives useful methods to treat survival data as a dichotomous (e.g. alive and dead) or polychotomous outcome (e.g. complete, partial or no remission) but barely mentions that these methods can only be used when all subjects have the same potential follow-up time (e.g. survival to 1 year [yes/no] among those with potential follow-up through 1 year or 1-month evaluation of cancer remission). If this requirement is not met, such analyses are not appropriate.

Two other survival analysis texts at a similar level deserve mention–one by Hosmer *et al.* (2008), & another by Kleinbaum & Klein (2011). Although all three texts cover much of the same material, there are differences in presentation and coverage of side issues that may distinguish them for different purposes. More advanced texts by Klein & Moeschberger (2005) and Collett Collett (2014) may also be of interest.

In summary, this book continues to improve, and the fourth edition is a welcome addition to the available books on survival analysis. The expanded sections on modelling and the addition of R software examples are particularly helpful.

Brenda W. Gillespie: *bgillesp@umich.edu*
Center for Statistical Consultation and Research
University of Michigan, Ann Arbor, MI, USA

References

Hosmer, D.W. Jr., Lemeshow, S. & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* Hoboken, NJ: Wiley.

Kleinbaum, D.G. & Klein, M. (2011). *Survival Analysis: A Self-Learning Text*, 3rd ed. New York, NY: Springer.

Klein, J.P. & Moeschberger, M.L. (2005). *Survival Analysis: Techniques for Censored and Truncated Data.* New York, NY: Springer.

Collett, D. (2014). *Modelling Survival Data in Medical Research*, 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.

Sample Size Determination and Power
Thomas P. Ryan
Wiley, 2013, 374 pages, £73.50/€88.20, hardcover
ISBN: 978-1-118-43760-5

Readership: Readers interested in methodology, upper-level undergraduate and graduate students.

For nearly all scientific studies, it is important to determine a large enough sample size that will give adequate statistical power to detect a desired effect size. Sample size determination and power consideration therefore can arise in any scientific endeavours across a variety of fields. It is also an active research area moving along with expanding scientific fields and new design considerations. Because sample size consideration is a broad subject, it is impossible to have an excellent reference book that can cover all areas adequately and provide appropriate examples. The aim of this book is to provide a general purpose book on the subject of sample size determination. This text is a useful modern addition to the literature, and as intended, covers a wide range of topics and applications that require sample size considerations.

This very comprehensive book is well structured. It starts with an introductory chapter briefly reviewing hypothesis testing and confidence intervals, the core statistical concepts needed for

understanding sample size determination and power, and presents a comprehensive approach to sample size determination. It then covers a wide range of specific topics from comparison of means and proportions to more complicated designs and settings. Broad topic and application areas discussed include clinical trials, survival analysis, reliability, epidemiology, microarrays and quality improvement. Each chapter ends with a section covering currently available commercial and non-commercial software for sample size determination considered in that chapter.

This book gives a comprehensive overview with an extensive list of topics. For specific topics, however, the discussion is inevitably a quick overview, and attention is given to providing extensive literature coverage for many specific topics. For example, a chapter titled Clinical Trials provides a good overview and discussion of recent developments and issues surrounding clinical trials, including the particular importance of adequate sample size when sampling human subjects. This chapter includes many sections including phase II trials, cluster randomised trials, longitudinal clinical trials, non-inferiority trials, repeated measurements and more; however, each is given less than half a page. The discussions, therefore, are brief, and a more unified presentation rather than a brief mention of published articles is desired at times. For example, if anyone considering a cluster randomised trial is looking to carry out a sample size calculation, this book gives references, but no clear guidance or an example showing how to do it.

In summary, this book succeeds in the author's aim of providing a general purpose text for readers interested in methodology without much technical fuss. In general, the book focuses on providing wealth of discussions and reviews for sample size and power and is an excellent source for up-to-date software and references for a wide range of topics related to sample size determination. It is written with enough mathematical sophistication for inquisitive readers and can be a supplementary textbook for upper-level undergraduate and graduate-level special topics courses. Because of the wide coverage, for many topics, this book chooses to provide extensive references with brief discussions about important issues in lieu of the details of the specifics of the sample size calculation. This book, however, is an appealing modern reference book for statisticians as well as scientific investigators interested in understanding general issues and theoretical backgrounds in sample size consideration.

Hyungjin Myra Kim: *myrakim@umich.edu*
Center for Statistical Consultation and Research and Department of Biostatistics
University of Michigan, 1085 South University, Ann Arbor, MI 48109-1107, USA

Analysis of Multivariate and High-Dimensional Data
Inge Koch
Cambridge University Press, 2013, 526 pages, CAD$92.95, hardcover
ISBN: 978-0-5218-8793-9

Readership: Researches in multivariate and high dimensional statistics, graduate students with a significant theoretical statistical background who are interested in the area.

At the outset, I must mention that this book is not for the uninitiated and requires a serious reading with lots of concentration. That said, I must highly commend the author for writing an excellent comprehensive review of multivariate and high dimensional statistics albeit on the

highly selected chosen topics. The lucid treatment and thoughtful presentation are two additional attractive features that make this book very attractive. I have been an admirer of books written by G. A. F. Seber, and this book, again coming from that part of world, in terms of rigour as well as its presentation, reminds me of those books. The book concentrates on various topics nicely classified in three parts as Classical Methods, Factors and Groups, and Non-Gaussian Analysis. The first part covers the standard techniques of principal component, canonical and discriminant analyses. The second part is about Clustering, Factoring and Multidimensional Scaling, which arguably are also the classical methods. It is, however, the third part that is relatively new and neoclassical dealing with Independent Component Analysis, Projection Pursuits, Kernel Component Analyses and Feature Selection.

The first part, the *Classical Methods*, is very well written and at a much higher level, in terms of mathematical rigour as well as understanding, than the other recent books on similar topics. Discussion of techniques with all mathematical details, interpretations and asymptotics makes the discussion very comprehensive and nearly complete. Especially attractive is the discussion of principal component analysis for various cases when dimensionality increases. This makes a strong justification for the current revival of interest in principal component analysis as an exploratory tool for high dimensional data. The canonical correlation analysis chapter includes much more material than what is normally discussed in any multivariate analysis course. Clarity of presentation for a topic such as this, and inherently crowded by long tedious mathematical expressions, is a difficult task, but the author seems to have tamed the monster successfully. Addition of topics like partial least squares is very appropriate, taking the mystery out of the algorithms as well as the approach and bringing it into the mainstream statistics from a specialised discipline like chemometrics. The chapter on discriminant analysis is well written. It is, however, a vast topic and receives less attention in this book than it rightly must deserve. Logistic regression has become an standard technique for classification. Yet, it receives only a cursory treatment. The support vector machines are relatively newer, and their inclusion with more details along with a discussion of their comparison with other techniques could really add considerable value for learners. Discrimination with repeated measures data, where information through covariance structures could be incorporated, is another important missing section.

As for the second part, dealing with factorings and groupings, the chapters again provide a wealth of useful information in a succinct manner. However, some chapters, especially the one on cluster analysis could have used some more updating by including the newer techniques.

It is the third part of the book dealing with non-Gaussian data where the book really stands out at its best. All three topics (Independent Component Analysis, Projection Pursuit and Kernel Methods) are relatively very new, and the author has taken pain to address each of them very carefully. She has nicely extracted the relevant details from the recent literature on these subjects and sequenced them in a such way that a serious reader will be able to appreciate the importance of these methods and will understand how to apply them. I must say that by reading this part of the book, I felt not only a sense of satisfaction but also a sense of accomplishment. I believe there are very few people in the statistics domain, who could claim to be an authority on these topics having full understanding and appreciation of these methods. In that sense, this book does a great service by providing a concise account of these more recent techniques. Without any hesitation and with admiration, I would give the author a 10 out of 10 for undertaking and accomplishing this task.

A book of reasonable size, aiming to cover topics in reasonable details, will have limitations on what to include, and I cannot fault the author for her selection of topics. However, in my personal preference, I would have preferred to see an adequate coverage of Gaussian-based

analyses. The author has completely stayed away from multivariate regression, multivariate analysis of variance and covariance structures. Multivariate distribution theory is minimally included and important topics such as multivariate t or Wishart distributions are completely avoided. The multivariate Gaussian theory is elegant in its own way but that aside, given that enough (and important enough) asymptotics are covered in this book, it would have been nice to provide some basic results. The absence of Gaussian theory somewhat takes the punch out of the third part of the book titled 'Non-Gaussian Analysis' as a reader is missing an appropriate baseline for comparison. An average reader is likely to miss out on the importance of this part solely because of his or her lack of appreciation due to missing discussion of Gaussian distribution. Correspondence analysis is another topic, which I think deserves a place in a book similar to this one. In terms of some minor quibbles, the use of 'we' as well as 'I' (e.g. see Section 1.3.2) even within the same paragraph is somewhat annoying; notational convention is somewhat new (but the author has a reason for it), parallel coordinates as a graphical technique are inadequately covered, and some of the figures, in my opinion, are too tiny for what they are trying to depict, especially when they correspond to large datasets with a large number of points. Small pictures can unintentionally hide some important features, which the author may see but a reader may not.

As I already have said, I will not hold these points as a criticism of the book–these were the choices made by the author for *her* book. What I see is an excellent collection of information presented in a very succinct and organised way. The feat she has accomplished successfully for this difficult area of statistics is something very few could accomplish. The wealth of information is enormous and a motivated student can learn a great deal from this book. This book is not for everyone but I believe that it is going to work as a model for many more multivariate books to appear in the future. I highly recommend this excellent book to researchers working in the field of high dimensional data and to motivated graduate students.

Ravindra Khattree: *khattree@oakland.edu*
Department of Mathematics and Statistics, Oakland University
Rochester Hills, Michigan, USA

Statistical Analysis of Network Data with R
Eric D. Kolaczyk and Gabor Csardi
Springer, 2014, xiii + 207 pages, €52.99, softcover
ISBN: 978-1-4939-0982-7

Readership: Anyone interested in analyzing network data in R.

Network data is quite trendy at the moment as it is available in abundance, especially on the internet. This type of data is in many ways quite different from traditional data and therefore requires special methods for handling and analyzing it. This book is a quite practical guide to get started with analyzing networks using the statistical software R. For the reader, it is definitely beneficial to be already familiar with this software. The book starts in Chapter 2 with a definition of network graphs, how they are best represented and how to operate on them. As in all chapters in the book, the terminology is always well explained and everything is easy to follow. Chapter 3 concerns the visualisation of graphs, which is not a trivial task as networks are usually large and many different concepts and options are available. Chapter 4

introduces descriptive measures to describe the network and concludes with graph partitioning. The next six chapters concern modelling network data where the foundation is laid in Chapter 5 with a short introduction to underlying mathematical models. Chapter 6 introduces the three common statistical models, the exponential random graph model, network block models and latent network models. More specialised topics like inference on network topology or processes on network graphs as well as analysis of the flow in a network are discussed in Chapters 7–9. The final chapter considers very briefly dynamic networks.

Network data occurs in different fields with many different questions of interest and the book reflects that by having many real data sets coming from quite a range of disciplines, such as social science or biology. I would have preferred if the data description was a bit more detailed. Nonetheless, the book demonstrates well the methods and representation that can be used for the different data sets and illustrates their analysis well. Naturally, there are many R packages for the analysis available. The book, which is accompanied by its own package, relies here especially on the igraph package and other specialised packages for specific tasks.

A nice feature of the accompanying package is that it installs all other packages needed to follow the analysis in the book (and there are many) and contains all the codes from the book. There are also references to alternative packages available in R.

While Chapters 2–4 are quite self-containing and do not require additional reading, Chapters 5–10 are all quite short introductions to the specific methods, and additional reading is recommended there. Relevant references are conveniently provided at the end of each chapter.

While in my opinion this book is not useful as a text book, it is a very nice hands-on introduction to the analysis of network data that gives a good overview suitable for applied scientists and statisticians. If certain models and methods need deeper understanding, the book provides the necessary references for further reading.

Klaus Nordhausen: *klaus.nordhausen@utu.fi*
Department of Mathematics and Statistics
University of Turku, 20014 Turku, Finland

Statistical Methods for Ranking Data
Mayer Alvo and Philip L. H. Yu
Springer, 2014, xi + 273 pages, €90.09, hardcover
ISBN: 978-1-4939-1470-8

Readership: Basic and non-parametric statistics advanced undergraduate and graduate students, and practitioners and researchers interested in foundations and theories of statistical methods for ranking data.

This book treats the most important aspects of non-parametric ranking methods. Situations where it is desired to rank a set of individuals or objects in accordance with some criterion are studied. Additionally, the authors are also ranking data arising when transforming continuous or discrete data in the context of non-parametric analysis.

The interest of Mayer Alvo in this research was sparked by a problem emanating from the School of Nursing at the University of Ottawa; it was desired to test if patients who were mobile differed from those who were not with respect to how they ranked certain sets of situations.

Following an introductory chapter starting the book off, Chapter 2 is dedicated to exploratory analysis of data. Correlation is studied in Chapter 3, followed by tests for randomness, agreement and interaction in Chapters 4 and 5. General theory of ranking is discussed in Chapter 6. Hypothesis testing for ordered alternatives is discussed in Chapter 7, following a description of general probability models in Chapter 8; more advanced topics such as probit and factor models and decision trees are investigated in Chapters 9 and 10. The book concludes in Chapter 11 with a description of some extensions of distance-based models.

The theory deals with preferences between objects and judges of the objects, which can be highly subjective. This possibility would limit greatly the objectivity of the elements of the methods for ranking data. However, some studied elements in the book are perfectly objective. The associated theory based on probabilistic and asymptotic models serves as a rough approximation useful to carry out non-parametric inference.

For these reasons, we suggest the study of this content with more complete objective methods based on objective real data. The investigation would be more interesting and applicable in practice, especially in the context of real data, when the subsequent ranking produces consequences of justice, for example, respect to the life, in health assistance professional recognition and economic retribution. Other subjective methods would be more productive in an academic sense but are less appropriate for objective practical purposes.

Mariano Ruiz Espejo: *mruiz033@alu.ucam.edu*
Departamento de Ciencias Humanas y Religiosas, UCAM
Campus de Los Jerónimos 135, 30107 Guadalupe, Murcia, Spain

Statistical Studies of Income, Poverty and Inequality in Europe: Computing and Graphics in R Using EU-SILC
Nicholas T. Longford
Chapman & Hall/CRC Press, 2015, xxii + 354 pages, £49.99, hardcover
ISBN: 978-14665-6832-7

Readership: Researchers in specific methods of analysis of economic poverty and inequality.

Poverty, according to Atkinson (1998), is about as widespread in our societies currently as it was a few decades ago when, admittedly, our standards for what amounts to prosperity were somewhat more modest. The author of this reviewed book wrote that 'by definition, regardless of its details, the poor are poorly integrated in societies in which prosperity dominates, and substantial reduction of poverty would be generally hailed as a great social and economic achievement'.

The poor population is studied with surveys. The design and analysis of such surveys are statistical and computational. The book focuses on statistical computing.

The European Union Statistics on Income and Living Conditions (EU-SILC) is a programme of annual national surveys that collect data related to poverty and social exclusion by interviews with adult members of randomly selected households. In this book, the analyses of surveys conducted by EU-SILC are carried out using the statistical language R. With growing popularity of other software such as Stata or MATLAB, it would be useful to include implementations in alternative platforms.

There are nine chapters altogether that go back and forth between introducing concepts associated with poverty analysis and statistical framework. One noteworthy section is Section 2.4

that is devoted to Horvitz–Thompson estimation and is methodologically solid. Other statistical techniques presented elsewhere in the book are often subjective or approximate in nature.

Other books exist in the literature that provide a more objective guide to an appropriate statistical technique for survey data, but methodological development of statistical technique is not the focus of this book. On the other hand, the books open up a list of analytical agenda in the context of EU-SILC surveys. The presented methods are not exhaustive for the range of analysis suitable for EU-SILC but are illustrative in the use of software codes, figures, tables and graphics.

Mariano Ruiz Espejo: *mruiz033@alu.ucam.edu*
Departamento de Ciencias Humanas y Religiosas, UCAM
Campus de Los Jerónimos 135, 30107 Guadalupe, Murcia, Spain

References

Atkinson, A.B. (1998). *Poverty in Europe.* Oxford: Blackwell.

Statistical Theory and Inference
David J. Olive
Springer, 2014, xii + 434 pages, €62.99, hardcover
ISBN: 978-3-319-04971-7

Readership: Statistics undergraduate, graduate and doctoral students and teachers.

This book describes the most important aspects of subjective classical statistical theory and inference similar to the treatment in Rohatgi (1984). Topics that are revised cover probability and expectations, multivariate distributions and transformations, exponential families, estimation and hypothesis testing, large sample theory and Bayesian methods. The book also includes specific pointers for students.

The book can be considered as a guide for teachers and students in the first or second courses in classical statistical methods, which many universities around the world teach as part of the core programme in statistics, both at a Masters and a PhD level. The interest in the book, however, would be limited by its subjectivity, as the book does not put an emphasis on the model diagnostics. This creates a gap between the mathematical model and its application to a real problem. A further shortcoming of the book is the omission of any reference to the issue of identifiability of the population units.

The book has been written with careful details and can serve as a good reference on the topics it covers. It is highly recommended from a didactic point of view but is of limited use to statistical practitioners.

Mariano Ruiz Espejo: *mruiz033@alu.ucam.edu*
Departamento de Ciencias Humanas y Religiosas, UCAM
Campus de Los Jerónimos 135, 30107 Guadalupe, Murcia, Spain

References

Rohatgi, V.K. (1984). *Statistical Inference.* New York, NY: Wiley.