Biometrics 72, 659–662
June 2016

DOI: 10.1111/biom.12534

BOOK REVIEWS

EDITOR:
DONNA PAULER ANKERST

MICHAEL FRIENDLY AND DAVID MEYER. **Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data**. Boca Raton, FL: CRC Press.

This book makes a very useful contribution by focusing on graphical methods for portraying discrete data and the results of fitting models to such data. Existing books on the analysis of discrete data pay relatively little attention to graphics. This book's main emphasis is on categorical data and logistic and loglinear models for them, but two chapters deal with count data, including dealing with the common existence of overdispersion and zero-inflation.

The authors show three types of plots: *Data plots*, *Model plots*, and *Data+Model plots* that combine the two, showing how well the model fits the data and portraying the uncertainty of estimated response means. The book has three sections: *Getting Started* introduces graphical methods for categorical data, working with the data in various forms (e.g., types of contingency tables), and fitting and graphing discrete distributions with useful displays, such as "rootograms," "Ord plots," and "Poissonness plots," and extensions due to Dave Hoaglin and John Tukey for other distributions. *Exploratory and Hypothesis-Testing Methods* presents displays and plots for two-way contingency tables, mosaic displays for multiway contingency tables, and plots such as biplots for correspondence analysis. *Model-Building Methods* presents plots relevant for standard models for discrete data, emphasizing logistic regression and its extensions for multinomial data, loglinear models for contingency tables, and generalized linear models for count data and their extensions. The 11 chapters each have exercises for practicing the methods and their graphic displays.

As the section outline suggests, the book covers the most popular statistical methods for analyzing discrete data. Among the types of graphical displays not presented are classification trees, graphical models for conditional independence structure, and depictions of estimates for models with large numbers of predictors (such as lasso estimates as functions of a smoothing parameter). Discrete modeling methods not covered include quasi-likelihood methods, such as generalized estimating equations for marginal models with multivariate responses, generalized linear mixed models, and Bayesian inference. But I believe it was sensible for the authors to emphasize graphics for basic methods, such as contingency table analysis and ordinary logistic regression, as the book already contains an impressive amount of material and will be very useful for most discrete-data analyses conducted by applied statisticians. Also, the authors consider many nonstandard models within these general classes, such as ordinal loglinear models and specialized models for square contingency tables.

Probably, the reason graphics have received relatively little attention in existing books for categorical data is because of the challenge of reducing even a bivariate association between qualitative variables to a simple graphic in which the key information is quickly clear to the eye. It is easier to do this in the context of logistic regression with quantitative explanatory variables, in which case many of the plots resemble standard ones from normal regression modeling. Examples of useful plots include influence and diagnostic plots (e.g., plotting studentized residuals against hat values, with Cook's distance values portrayed by the size of a bubble) and added-variable plots. Especially helpful for portraying practical implications of the model parameter estimates are "effect plots" that portray how the probability of an outcome varies across values

of an explanatory variable, together with confidence regions showing the plausible values, when other explanatory variables are fixed at typical values, such as their means.

For contingency tables, graphical presentation is more challenging, as graphics direct the eye to look cell-by-cell, rather than providing a summary that the eye can quickly absorb. The graphics are most helpful for two-way and three-way tables. For two-way tables, one can start with a grouped barchart that portrays results on a response variable at each category of an explanatory variable, or a "spineplot" that stacks them. A "tile plot" shows squares proportional in area to cell counts in the format of the table. A "mosaic plot" does this with contiguous rectangles to reflect a response/explanatory distinction. "Sieve diagrams" portray each cell by a number of rectangles equal to the cell count.

The mosaic plots and other association plots are more informative when entries show departures from a particular model. For example, the authors suggest adding color reflecting sizes of Pearson residuals, to highlight cells with strong evidence of departure from models, such as independence of variables in two-way tables and mutual independence and conditional independence in three-way tables. I have a minor criticism about these plots: I believe that it is more sensible to base color displays on standardized residuals than Pearson residuals. For example, a 2×2 table has only a single degree of freedom for checking independence; all four standardized residuals have the same absolute value, whereas the four Pearson residuals can all have different magnitudes (as they will when the cell expected frequencies all differ). The mosaic plots then suggest there are four distinct bits of information on lack of fit, which is not the case. In fact, later in the book (p. 356) in the context of loglinear models the authors mention that standardized residuals are preferable to Pearson residuals, so it seems a bit inappropriate to use the Pearson residuals for displays for simple contingency tables.

Besides presenting graphics, the book also has considerable discussion of the corresponding discrete-data models. The Preface mentions that Prof. Friendly uses this book as a main course textbook on discrete data, but indicates that he supplements it with a standard text on categorical data analysis for additional readings. As its title suggests, the book illustrates implementation of the methods using R software. As R has become more popular, recent years have seen publication of several books that focus on its use for categorical data (e.g., Kateri, 2014). Friendly has previously presented guidance on graphics for SAS users (e.g., Friendly, 2000). An alternative approach to the Friendly and Meyer book would have been to put main emphasis on the graphical methods, and supplement the presentation with information about all the major Statistics software packages in common use, for instance, including Stata and SPSS as well as R and SAS. But such an approach may well now be impractical, given the explosion in recent years in both methodology and software. I myself gave up on attempting this in the most recent edition of my general book on categorical data analysis (Agresti, 2013); I decided to focus on the methods themselves and put software information and examples at a supplementary text website. But those who mainly use R for their analyses will undoubtedly prefer the Friendly and Meyer approach of incorporating the R code in the book itself.

Overall, I found the material to be clearly presented and to fill an important niche. I would like to have seen graphic presentations also for likelihood and profile likelihood functions, and discussion of their usefulness and their behavior for some awkward situations, such as complete separation in logistic regression and consequent infinite maximum likelihood estimates. Some models also seem to merit more attention, such as baseline-category logit models with qualitative predictors, for which the large number of parameters can be intimidating. See Tutz and Schauberger (2013) for recent work on using star plots for effects in multinomial logit models. However, I think that this book is an important contribution and a strong addition to the existing discrete-data texts.

For many years, Prof. Friendly has been the most effective promoter in Statistics of graphical methods for categorical data. We owe thanks to Friendly and Meyer for promoting graphical methods and showing how easy it is to implement them in R. This impressive book is a very worthy addition to the library of anyone who spends much time analyzing categorical data.

References

Agresti, A. (2013). *Categorical Data Analysis*, 3rd ed. Hoboken, New Jersey: Wiley.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary, North Carolina: SAS Institute.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*, New York: Birkauser/Springer.

Tutz, G., and Schauberger, G. (2013). Visualization of categorical response models—From data glyphs to parameter glyphs. *Journal of Computational and Graphical Statistics* **22**, 156–177.

Alan Agresti
Department of Statistics
University of Florida
Gainesville, Florida
aa@stat.ufl.edu

ADRIAN BADDELEY, EGE RUBAK, AND ROLF TURNER, **Spatial Point Patterns: Methodology and Applications with R.** Boca Raton, FL: CRC Press

As possibly the first user of the software package now known as "spatstat," it is an honor and a pleasure to review Baddeley, Rubak, and Turner's wonderful new book entitled "Spatial point patterns: Methodology and applications with R."

Contrary to popular belief, the spatstat cradle did not stand in Western Australia. Indeed, the package's roots can be traced back at least to the first half of the 1990s when Adrian Baddeley, then based in Amsterdam, wrote routines for the computation of (Kaplan–Meier) estimates of functional summary statistics for stationary point processes. Since then, Baddeley and Turner, later joined by Rubak, with input from many leading experts in point process analysis, have turned the package into an indispensable tool for researchers and practitioners alike. In a reverse direction, new discoveries are being implemented quickly in spatstat.

The fruits of this decades' long process are very well documented in the book. As the authors point out in their preface, the book is not intended to be an introduction to point process theory for mathematicians. Rather, they aim to focus on the principles of statistical inference for spatial data and to help researchers in application domains with the practicalities of the analysis and the interpretation of the results. In this, they have succeeded brilliantly.

The book is written in a distinct, at times funny, always accessible style. General principles of every aspect of spatial point pattern analysis, from data collection to model validation, are discussed in great detail with pointers to the specialized literature for those who wish to gain a deeper understanding of the technicalities. The principles are illustrated by means of a wide collection of examples that can be reproduced by the reader in R. Moreover, a selection of frequently asked questions from spatstat users is answered at the end of each chapter.

The book is conveniently organized in four main parts. In the first part, an informal introduction to point patterns, R and the spatstat package is given. The second part is devoted to exploratory data analysis. After describing tools for plotting spatial data, attention is paid to classical summary statistics, such as the intensity, pair correlation function and distance based functionals for quantifying the degree of clustering or regularity in the data.

The longest and, perhaps, core part of the book is the third one in which statistical inference is considered. In this part, the authors present a coherent framework for model fitting, selection and validation. In Chapter 9, the main focus is on likelihood-based inference for Poisson models defined by a loglinear intensity function. The next chapter discusses hypothesis testing. It explains in detail how Monte Carlo envelopes can and cannot be used, before turning to residuals and other tools for model validation in Chapter 11. Inference for Cox, cluster and Markov point processes is discussed as well. For the latter class of models, the default technique is pseudo- or composite likelihood.

Throughout, connections to standard statistical techniques are emphasized and common misconceptions cleared up. The last three chapters concern more specialist and less developed topics, including spatio-temporal point processes and replication.

In summary, I warmly recommend the book to anyone who wishes to analyze point patterns professionally.

M. N. M. van Lieshout
CWI and University of Twente
Amsterdam and Enschede, The Netherlands
M.N.M.van.Lieshout@cwi.nl

TAN WAI-YUAN, **Stochastic Models with Applications to Genetics, Cancers, AIDS, and Other Biomedical Systems, 2nd ed.** Singapore: World Scientific Publishing Co.

The book is a rich source of examples of models in biology and medicine. It consists of 10 chapters. The author does not assume prior knowledge of Markov chains or diffusion theory. In the first half, the book develops the basics of Markov chain theory in discrete and continuous time, including stationary distributions, limiting behavior, and Markov chain Monte Carlo methods. The second half deals with diffusion models and state space models.

The strength of the book is the huge number of examples and applications. These include among others branching processes, the Wright model in population genetics, and the AIDS epidemiology in San Francisco. The book is a good source for an applied probability course for nonmathematicians with an emphasis on applications.

Silke Rolles
Department of Mathematics
Technical University Munich
Munich, Germany
srolles@ma.tum.de

KATIE HARRON, HARVEY GOLDSTEIN, AND CHRIS DIBBEN, eds, **Methodological Developments in Data Linkage**. New York: Wiley.

While maximization of the informative value of data by merging multiple sources has long been practiced by administrative sectors, census bureaus, and population registries, the science is receiving added boost in a new big data era that urges more publicly funded data to be made available for research. For example, a decade after the 7-year Prostate Cancer Prevention Trial completed, study statisticians were able to match patient records from the trial with individual long-term survival records from U.S. death indices in order to draw powerful conclusions as to the effect of chemo-prevention on prostate-cancer mortality, a long-term critical endpoint that was too expensive to follow as part of the original 7-year trial.

While the informatics field has long concerned itself with record linkage, what statisticians can bring to the table are the ubiquitous tools that honestly report the confidence of results, including quantitative assessment of error rates, optimal estimates of matches, and methods for dealing with biases due to data missing-not-at-random. For the statistician looking to enter this field, this book is the ideal place to start.

After a brief introduction of common terms in Chapter 1, Chapter 2 provides a comprehensive history, overview, notation, and definitions of the basic concepts of probabilistic linkage. Under probabilistic matching, two records from the different lists are assigned a probability of matching based on values of a set of defined matching variables, such as age, date of birth, and address. It is comforting to see all the standard supervised and unsupervised classification methods, the Expectation-Maximization (EM)-algorithm, Markov Chain Mote Carlo Methods (MCMC), as well as some of the standard statistical methods on dealing with missing data that are not-missing-at-random in use for these problems. Interesting examples of application from the 1990 U.S. Decennial Census that are clearly presented motivate the techniques. Further real case studies from Canada, Australia, and the United Kingdom are provided in Chapter 3. A comprehensive meta-analysis reflecting the low rates of adequate error reporting in the literature, with official definitions and recommenda-

tions for the future are given in Chapter 4. Chapter 5 covers the secondary analysis of linked data and comprehensively walks through the general estimating equation formulas necessary for multiple scenarios. The concrete exposition allows the reader to see exactly how the experts correct for noninformative nonlinkage mechanisms using sample audits or auxiliary information. Clarifying simulation studies are presented showing the importance of correcting for nonlinkage, and the analogies to nonignorable missing data are provided to encourage further research in this area.

Picking up where Chapter 5 left off, Chapter 6 postulates record linkage as a missing data problem, bringing it to familiar terrain for many statisticians. The advantages of prior-informed imputation over probabilistic record linkage in terms of reducing bias are shown in an illuminating example concerning electronic health records and a series of simulations. Chapter 7 introduces the world of graphical databases with a tour de force of graphics, coding, and efficiency advantages over the more widely used relational databases. Compelling examples of graphical methods are provided, including the commonly seen recommendations from Amazon: "People who bought this book, also bought these." This chapter was a real eye-opener, leaving no stone unturned. Beginning with case studies across several countries, Chapter 8 dives into large-scale linkage, guiding the reader through the nuts and bolts of hashing, link keys, similarity tables, and the role of anonymization. And finally, Chapter 9 covers the important issue of privacy-preserving record linkage, providing a layman's overview of encryption, hashing, salting, and Bloom filters, and all the techniques that clever computer scientists use to thwart attacks.

The danger with many edited books is that they may be uneven or noncohesive. And the presumption with any book covering the specifics of constructing data files is that it is bound to be dry and boring. This was not the case here. The editors clearly selected writers with practical experience who could tell an interesting story in the context of real examples. The chapters were carefully organized to flow from one topic to another, with each chapter sequentially building upon the prior chapter. I often forgot chapters were written by different authors. Essential concepts and formulas were given so as not to overwhelm the reader, plenty of references were provided for those who want more. As a statistician who wants to start research in big data and electronic health records, I went in knowing nothing about the field, could read an intriguing book during my leisure time, and came out knowing a lot. This book is a must-read for scientists working in all facets of data linkage.

Donna Pauler Ankerst
Department of Mathematics
Technical University Munich
Munich, Germany
ankerst@tum.de