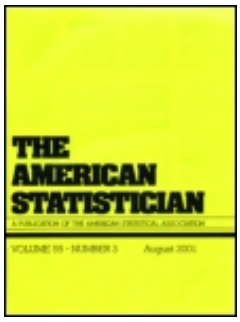

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Reviews of Books and Teaching Materials

To cite this article: (2016) Reviews of Books and Teaching Materials, The American Statistician, 70:4, 424-433, DOI: 10.1080/00031305.2016.1234902

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1234902

one particular design that has been decided on ahead of time. It would have been nice to see an example that walked the reader through the decision-making process of choosing between, for example, a cluster randomized cross-sectional design, a longitudinal design, and a pseudo cluster randomized design.

One major limitation of the book is its focus on presenting sample size equations that rely on critical values computed from the normal distribution (rather than, for instance, critical values from a t distribution). This, of course, allows for a simple presentation of sample size formula. However, if computer software had been introduced earlier in the book (rather than being relegated to a very short chapter at the end of the book), it could have been used to illustrate computations using a t distribution. More focus on using software for power computations would also have been more in line with the preferred approach of most researchers, who are more likely seek out software to compute sample sizes rather than plug numbers into an equation to compute sample sizes by hand.

The authors justify the large-sample approach by claiming that multilevel models should not be used for studies with fewer than 20 clusters. However, an alternative approach is not suggested, and it is certainly not uncommon to see cluster randomized trials with fewer than 20 clusters. Community intervention trials are a prime example. As it is, important issues regarding the computation of degrees of freedom in mixed effects models are ignored entirely.

The second important limitation of the book is its treatment of dichotomous outcomes. One problem is that the authors never settle on a consistent procedure for estimating the standard errors needed to compute power, and they do not clearly connect the test statistics used to derive sample size formulas to particular statistical models. In the chapter on cluster randomized designs, sample size formulas are provided for both the risk difference and the odds ratio. The first seems to be based on a literature that corrects tests for independent data using design effects, whereas the second is based on asymptotic arguments associated with generalized linear mixed models. The standard error in each case depends on an ICC, the definition of which is potentially ambiguous. However, the reader is given no guidance regarding when a particular sample size formula or ICC definition should be chosen. In the chapter on multisite designs only the formula based on logistic modeling of the odds ratio is provided. In Chapter 7, an entirely different model (a hierarchical binomial model) is used to determine standard errors.

Another problem is that the authors never address whether and how researchers should use covariates in conjunction with binary outcomes. The literature is confusing on this point, with some articles showing that covariate adjusting treatment effect estimates harms power and others showing that it helps. So I grant that this is a difficult issue to discuss in a book at this level. However, I do not agree that the book should ignore the issue entirely.

Schochet (2013) showed that a generalized estimating equations approach to modeling binary data ensures that covariates function the same way in models for binary outcomes as they do in linear models for continuous outcomes. However, the generalized estimating equations approach is ignored in the manuscript.

The book has its share of typographical errors, and occasionally these contribute to difficulties in following the argument. For instance, the discussion on page 104 seems mistaken until one realizes that the wrong variance components are in the numerator of the two equations defining intracluster correlation coefficients.

Despite the above critiques, there is more good than bad in this book. Simple sample size formulas are provided for standard designs and for a few more advanced designs as well. Optimal designs using cost constraints are defined for a wider variety of designs than in other books I am aware of. Despite its flaws, I think this book deserves a place on the bookshelf of both researchers who plan experimental studies and statisticians who advise them.

References

Berger, M. P. F., and Wong, W. K. (2009), *An Introduction to Optimal Designs for Social and Biomedical Research*, Chichester, UK: Wiley. [428]

Bloom, H. S. (1995), "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs," *Evaluation Review*, 19, 547–556. [428]

Donner, A., and Klar, N. (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*, London: Edward Arnold. [428]

Schochet, P.Z. (2013), "Statistical Power for School Based RCTs with Binary Outcomes," *Journal of Research on Educational Effectiveness*, 6, 263–294. [429]

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., and Raudenbush, S. (2011), *Optimal Design Plus Empirical Evidence: Documentation for the Optimal Design Software*. Available at *http://hlmsoft.net/od/od-manual-20111016-v300.pdf* [428]

Christopher H. Rhoads
University of Connecticut

Spatio-Temporal Methods in Environmental Epidemiology.

Gavin Shaddick and James V. Zidek. Boca Raton, FL: Chapman & Hall/CRC Press, 2015, xxxi + 365 pp., $89.95 (H), ISBN: 978-1-48-223703-0.

The past two decades have seen a rapid expansion in the development of spatial and spatio-temporal statistical methods and a corresponding expansion in the number of books and textbooks outlining the theoretical and applied sides of this growing analytic toolbox. Within the field of environmental epidemiology, such methods have grown to address spatial, temporal, and spatio-temporal prediction of exposures for given locations in space and/or time, and the association to health outcomes observed in individuals living, working, or moving within the same area and experiencing the predicted exposures. The specific area of air pollution epidemiology provides multiple challenges in this regard: first, air pollution levels for multiple pollutants are monitored at fixed locations providing temporal and spatial snapshots of a complex, multidimensional, dynamic environmental process; and, second, health outcome data often occur as aggregate outcomes of individuals living in enumeration districts often collected for nonepidemiological purposes

(e.g., billing). The epidemiologic, statistical, and geographical challenges in linking these data and providing epidemiologic insight are numerous, compounded in the space-time setting, and require thoughtful application of challenging methods to provide insight into underlying disease processes. The authors of this text, both accomplished researchers in the area, provide a much-needed consolidation of spatio-temporal modeling methods with an overall goal "to promote the interface (of environmental epidemiology and spatio-temporal modelling) between statisticians and practitioners to allow rapid advances in the field of spatio-temporal statistics to be fully exploited in assessing risk to health" (from the preface).

The textbook condenses many complex topics into accessible and manageable chapters addressing key elements of modern spatio-temporal analyses of environmental epidemiologic data. The book sets the stage by giving an overview of a very general hierarchical framework for the analysis of environmental data, a framework popularized in climate science by Berliner (1996) and in general environmental statistics by Wikle (2003). This broad hierarchy consists of the observation process (also referred to in the literature by the data process or the measurement process), the underlying environmental process driving space-time dynamics, and the (prior) distributions of parameters for both of the processes. The framework provides a solid conceptual setting for the methods in the book, as well as the Bayesian framework for analysis described throughout. The authors build on this framework to outline a very helpful list of elements of a strong spatio-temporal model, nicely preparing the reader for the details that follow. Next, the authors provide a whirlwind tour of epidemiologic study designs (e.g., cohort, case–control), generalized linear models (including smoothers and splines), and detailed definitions of Poisson and logistic models common in standard, nonspatial epidemiology. The authors provide helpful R examples throughout.

Chapters 3 and 4 provide a readable but thought-provoking overview of theoretical (and philosophical) concepts of uncertainty and Bayesian statistics. The two chapters together provide a valuable primer of concepts and terms often overlooked by readers seeking to jump immediately into complex modeling algorithms, but I plan to recommend them to students and colleagues just entering the field as a great source for the underlying concepts, described with examples relevant to the environmental epidemiology setting. Chapter 5 quickly follows illustrating the use of Markov chain Monte Carlo and integrated nested Laplace approximation implementations for implementing Bayesian modeling concepts, demonstrated with WinBUGS and R-INLA examples.

Chapters 6 and 7 detail strategies and challenges for modeling large environmental datasets. Topics include variable selection methods with a thoughtful discussion of the role of *p*-values (mirroring recent discussions in the American Statistical Association's consensus statement on *p*-values, Wasserstein and Lazar 2016), model averaging, and model comparisons (e.g., through Bayes factors). Analytic challenges such as missing data, measurement error, and preferential sampling often arise in environmental epidemiology and are each described in detail along with focused data examples and accompanying code.

Chapters 8–10 separately examine methods for spatio-temporal estimation of local disease risk, methods for spatio-temporal prediction of continuous exposure fields, and methods for modeling temporal variation in exposures, respectively. Chapter 8 introduces spatio-temporal models for the small area estimation of health outcome risks and rates (i.e., disease mapping). The authors provide accessible introductions and examples for concepts ranging from "borrowing strength" and shrinkage estimation through Markov random fields and the use of conditional autoregressive prior distributions, again, with multiple examples and implementation in both WinBUGS and R-INLA. Chapter 9 follows with a focus on prediction of the exposure field from observations from a network of monitors. Methods include estimation of spatial correlation, standard, and model-based kriging, and, again, the authors provide relevant examples including a detailed assessment of NO_2 exposures across all of Europe. Chapter 10 provides an overview of time series methods applied to environmental pollution data including forecasting, filtering, and the spectral domain.

Chapters 11 and 12 bring exposure and health outcome data together, with specific focus on studies assessing health impacts of air pollution in data measured over space and time. Chapter 11 provides the methodological framework, again with many relevant and helpful examples and accompanying code, while Chapter 12 provides an important review of the many things that can (and often do) go wrong in large-scale epidemiologic analyses including aggregation (ecologic) bias and hidden exposure pathways. The authors outline mechanisms for acknowledging ecological bias and models for estimating personal exposures for individuals moving through various micro-environments.

Chapter 13 provides an in-depth look at design criteria for spatial and spatio-temporal monitoring of exposure fields to improve prediction of continuous exposures and aid in estimation of resulting health effects. The authors describe several existing large-scale monitoring networks and include elements of sampling and optimal design theory as applied to these networks. Examples include air pollution, temperature, acid deposition, and sampling in stream networks.

Chapter 14 concludes by pushing the envelope into emerging applications of spatial and spatio-temporal environmental statistics. The authors review methods for assessing nonattainment of regulatory limits, methods for modeling infectious disease dynamics in space and time, spatial deformation to improve statistical analysis, and extending methods to address networks of multivariate outcomes where individual monitors may or may not measure every pollutant.

The text covers a remarkable number of topics in its 318 pages (including many full color graphics and examples of code and output). The structure outlined above provides excellent coverage of many areas of recent development, held together with compelling examples and illustrations. Not all topics are easy and some may not be immediately accessible to all epidemiologic practitioners (e.g., spectral methods and Bochner's lemma in Chapter 10), but advanced topics are clearly marked and available in context for readers interested in digging deeper but not overly distracting for those seeking to move along to the next topic. Most (but not all) examples directly relate to air pollution epidemiology so other topics in environmental epidemiology (e.g., toxicology) are not stressed in detail. That said, the focus of the text is on spatio-temporal methods of analysis and air pollution epidemiology's combination of

complex, multivariate, and dynamic space-time fields of exposure and space-time-specified health outcomes provide the quintessential example for space-time studies in the field.

Overall, I found the book a comprehensive overview placing many different topics into a logical perspective with focused, helpful examples. I enjoyed reading the book, am already recommending it to colleagues, and anticipate referring to it often in my future work.

References

Berliner, M. (1996), "Hierarchical Bayesian Time Series Models," in *Maximum Entropy and Bayesian Methods*, eds. K. Hanson and R. Silver, Boston, MA: Kluwer Academic Publishers, pp. 15–22. [430]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [430]

Wikle, C. K. (2003), "Hierarchical Models in Environmental Science," *International Statistical Review*, 71, 181–199. [430]

Lance A. Waller
Emory University
http://orcid.org/0000-0001-5002-8886

Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery. Walter W. Piegorsch. New York: Wiley, 2015, xv + 470 pp., $115.00 (H), ISBN: 978-1-11-861965-0.

In *Statistical Data Analytics*, Piegorsch aims to teach the statistical analysis skills that form the foundations for data mining, informatics, and knowledge discovery. However, it should be emphasized that the goal of the book is not to explore computational methods for data mining or informatics, but rather to provide a broad base of statistical knowledge for data analysis, with a threefold focus on data summarization and statistical inference, supervised learning, and unsupervised learning. The book is written to an audience familiar with multivariable calculus and linear algebra, though a concise review of the latter is given in an appendix. *Statistical Data Analytics* is well-written and could easily serve as a graduate course textbook, an instructional resource, or a statistical reference.

The material in *Statistical Data Analytics* is presented similarly in each chapter. For every statistical method that is discussed, the author introduces and develops the underlying mathematical model as well as important details associated with a data analysis. An illustrative example is given for each method, with R code and output provided when applicable. The author usually uses the examples effectively to elucidate the purpose and ideas behind the statistical techniques. For most of the chapters, multiple statistical methods are discussed and practical examples are supplied from a variety of fields (finance, economics, medicine, biology, genetics, astronomy, etc.). While most of the R code is based on built-in functions from a variety of packages, some user-defined functions are supplied. For those unfamiliar with R, an appendix is given that teaches the basics of R programming. At the end of each chapter, a number of applied and theoretical exercises are provided. An online repository contains the datasets needed for the chapter exercises, and solutions to the exercises are available in a separate manual.

In the first chapter, Piegorsch makes a brief case that the focus of the book—statistical learning and analytics—is a necessary starting point for understanding how to effectively handle "big data." Thus, while the remaining chapters of the book do not describe strategies specific to data mining and informatics, the author makes it clear that mastering the statistical techniques for analyzing smaller datasets is essential for the overarching process of knowledge discovery for any size data. In addition to establishing the philosophical approach of the book, Chapter 1 is used to discuss problems that occur with data collection and to define the differences in statistical description and modeling.

Chapters 2 through 5 of the book provide a review of probability, data manipulation, data visualization, and statistical inference. These chapters are mainly included so that, theoretically, someone with minimal prior statistical knowledge could understand the statistical learning methods presented in the latter two-thirds of the book. Specifically, Chapter 2 reviews the most important concepts of probability, random variables, and statistical distributions. Chapters 3 and 4 explain a variety of classical and modern data summarization and graphical procedures, including measures of location and variability, histograms, boxplots, and a variety of other techniques. Chapter 5 tackles statistical inference, with an emphasis on the use of likelihoods, confidence intervals, and hypothesis tests. Also included in this chapter is a discussion of multiple testing and the false discovery rate. While Chapters 3 and 4 are very accessible even to a statistically naive reader, the material in Chapters 2 and 5 is fairly dense and would be challenging for most readers not previously familiar with the content. To the author's credit, the majority of important concepts are clearly and efficiently presented, and when details are lacking, other sources are referenced.

Chapters 6 through 9 of the text discuss what are arguably the most important methods of supervised learning: linear regression, generalized linear models, and discriminant analysis. Given that the reader has a solid understanding of matrix algebra and the material in Chapters 2–5, these chapters are relatively straightforward and provide a nice balance between theory and application. Chapters 6 and 7 offer a wide-ranging coverage of simple linear regression and multiple linear regression, respectively. In Chapter 7, there is a strong emphasis on model building and related procedures such as ridge regression, LASSO, and cross-validation. Disappointingly, here and throughout the remainder of the book, only quantitative predictors are considered, with a few exceptions. A brief discussion of analysis of variance (ANOVA) models is given at the end of Chapter 7, but strategies for handling categorical predictors using methods such as dummy variables are only peripherally mentioned in the context of regression. Chapter 8 introduces generalized linear models for potentially nonnormal responses. Special attention is given to logistic regression, log-linear models for contingency tables, and gamma regression, with real-world examples provided for each procedure. Chapter 9 completes the section of the book dedicated to supervised learning with a discussion of