Biometrics 65, 326–339
March 2009

BOOK REVIEWS

EDITOR:
THOMAS M. LOUGHIN

TSIATIS, A. A. **Semiparametric Theory and Missing Data**. Springer, New York, 2006. xvi + 383 pp. $95.00/£74.85. ISBN 9780387324487.

In most studies, the intended (full) data, i.e., the data that the study investigators wish to collect, are inevitably incompletely observed. In modern studies, the full data are typically high dimensional, usually comprising many baseline and time-varying variables. Scientific interest, however, often focuses on some low-dimensional parameter of the distribution of the full data. Specification of realistic parametric models for the mechanism generating high-dimensional data is most often very challenging, if not impossible. Nonparametric and semiparametric models, i.e., models in which the data-generating process is characterized by parameters ranging over a large, non-Euclidean, space and, possibly, also a few meaningful real-valued parameters, meet the analytic challenge posed by these high-dimensional data because they do not make assumptions about the components of the full data distribution that are of little scientific interest. Analytic strategies based on semiparametric models avoid the possibility of incorrect inferences due to misspecification of models for the secondary parts of the full data law.

Drawing from the modern theory of semiparametric efficient inference developed since the 1980s, Robins and Rotnitzky (1992) derived a general estimating equations methodology in coarsened, i.e., incompletely observed, data models under non- or semiparametric models for arbitrary full data configurations. This methodology, based on the geometry of scores and influence functions, applies when the

data are coarsened at random (CAR) and the coarsening, i.e., censoring or missingness, mechanism is either known or correctly modeled. Since the publication of Robins and Rotnitzky (1992), there has been a proliferation of research papers that extended the methodology to non-CAR models and applied it to a number of important semiparametric problems with CAR and non-CAR data.

Tsiatis's book provides a very nice and complete introduction to the general estimating function methodology for inference in semiparametric models with CAR data. It focuses primarily on methods for missing data problems, but it also provides some discussion of how the general methodology can be applied to estimation in semiparametric censored data models. It covers methodology suitable for settings in which the probability of observing full data is positive. This setting is quite broad and applies to a number of important practical data structures including data arising from nonresponse in surveys, drop-out in longitudinal studies, and two-stage designs. However, it does not cover inference when the probability of observing full data is null as is the case, for example, with current status data structures. Also, the book does not cover methodology for semiparametric inference with non-CAR data and the associated sensitivity analysis methodology in non-CAR models.

The level of the book is more introductory than the seminal book of van der Laan and Robins (2003), the only other existing book on the topic. The book is accessible to a wider readership and in fact, it provides great preparation for those who wish to read van der Laan and Robins, which is mathematically more advanced and broader in scope. Tsiatis succeeds in presenting in a conversational, orderly, logical, and easy-to-read style an elaborate and technically challenging theory. The book skips a number of technical subtleties but it nevertheless retains sketches of the derivation of essentially all the key results of the theory with substantial mathematical detail. The technical developments in the book are accompanied by a few working examples that are carried through the entire book and which clearly illustrate the relevance of the theoretical results of each chapter. One small drawback is that in some chapters the bibliography is scarce.

The first five chapters cover background material on the theory of efficient estimation in arbitrary semiparametric models for full data. These chapters provide a nice heuristic and intuitive introduction to the geometric ideas that underlie the theory, especially as it applies to the construction of unbiased estimating equations. These chapters in themselves are a great resource to help readers prepare for the more advanced books on the topic: Bickel et al. (1993), van der Vaart (1998, chapter 25), and Kosorok (2007).

Chapter 1 introduces the notions of semiparametric models and estimators and illustrates these ideas with two widely used models, the restricted moment model (popularized by Liang and Zeger, 1986, with their Generalized Estimating Equations methodology) and the proportional hazards model (Cox, 1972). Chapter 2 summarizes key elements of the theory of Hilbert spaces that are used in the development of the theory of semiparametric estimation, in particular the notions of orthogonality, projection, and the projection theorem. Chapter 3 illustrates how the geometric ideas of chapter 2 are applied in the development of the theory of asymptotically efficient inference in parametric models. The chapter introduces the notions of regular and asymptotically linear (RAL) estimators, and of their influence functions. It gives a key result characterizing geometrically the influence functions of RAL estimators and of the efficient influence function and it illustrates how this characterization can be used to construct unbiased estimating equations. Chapter 4 builds on the ideas of chapter 3 to derive geometric results about influence functions of RAL estimators of parameters of semiparametric models and to indicate how these results can be applied to construct estimating functions. The chapter defines the notions of parametric submodels, tangent and nuisance tangent spaces, semiparametric efficient score and influence function, and semiparametric variance bound. The chapter ends with a detailed derivation of these objects, and their application to locally efficient estimation in, the restricted moment model. Chapter 5 is entirely devoted to the application of the results of chapter 4 to estimation in two popular semiparametric models: the location shift model and the Cox proportional hazards model.

The remaining nine chapters are devoted to inference with missing data. Chapter 6 nicely uses the problem of estimating the mean of an incompletely observed outcome in the presence of always-observed auxiliary variables, to overview the taxonomy of the missing data mechanisms and the different available estimation approaches: maximum likelihood under a parametric model, imputation, inverse weighted probability, and double-robust estimation. Chapters 7 to 12 cover the abovementioned estimating function methodology of Robins and Rotnitzky for estimation in semiparametric models with missing at random data. Chapter 7 defines coarsened data (albeit not in entire generality as only discrete coarsening variables are considered) and views missing data as a special case. It then defines CAR and derives the likelihood factorization under CAR. It then uses this factorization to characterize the nuisance tangent space, its orthogonal complement, and the set of influence functions of all RAL estimators in models in which the missingness probabilities are known. It then shows how these characterizations lead to inverse probability weighted (IPW) and augmented IPW estimating equations. The ideas are illustrated in the restricted mean model. Chapters 8 and 9 extend the results of chapter 7 to the case in which the missingness probabilities are unknown but estimated under a parametric model. These chapters focus on the more technically tractable case in which the nonresponse patterns are monotone, as in longitudinal studies with dropout. Chapters 10 and 11 deal with methods for constructing locally efficient and double-robust augmented IPW estimators, nicely and gradually explaining the difficulties arising in estimation with nonmonotone missing data structures and/or in inference in regression models with missing covariates. Chapter 12 describes a method, based on optimal linear combination of a set of estimating functions, to compute estimators with improved, albeit not optimal, efficiency. Chapter 14 deviates somewhat from semiparametric methods, to discuss the asymptotic properties of multiple imputation estimators for parameters of parametric models. This chapter is essentially based on the papers of Wang and Robins (1998) and Robins and Wang (2000).

In summary, the book presents an excellent introduction to the new developments in methodology for estimation in semiparametric models with missing data. The first five chapters

even serve as a nice introduction to more advanced books in the general theory of efficient estimation in semiparametric models. The material covered in the book will appeal to statisticians who wish to pursue research in methods for missing and censored data but it will also be valuable for anyone wishing to sample the area. It will be a good textbook for a one-semester introductory course in semiparametric estimating function methodology for missing data. In conclusion, the book is highly valuable.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 305–334.

Kosorok, M. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Liang, K. Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology—Methodological Issues*, N. Jewell, K. Dietz, and V. Farewell (eds), 297–331. Boston: Birkhäuser.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87,** 113–124.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer Verlag.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85,** 935–948.

Andrea Rotnitzky
Department of Economics
Universidad Di Tella
Buenos Aires, Argentina
Department of Biostatistics
Harvard School of Public Health
Boston, U.S.A.

DANIELS, M. J. and HOGAN, J. W. **Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis**. Chapman & Hall/CRC, Boca Raton, Florida, 2008. xx + 303 pp. $79.95/£42.99. ISBN 9781584886099 (Hardcover).

The introduction of SAS Proc Mixed in the early 1990s sparked a revolution in statistical practice. For the first time, statisticians could readily estimate a wide range of models from longitudinal data sets, even those beset with arbitrary patterns of missingness. The price of admission to this world is the assumption (typically untestable) that the dropout mechanism is missing at random (MAR), failure of which could cause likelihood-based inferences to go badly wrong. Thus a major thrust of research in the intervening years has been to develop methods that accommodate departures from MAR.

The new monograph by Daniels and Hogan offers a timely and thorough review of this maturing research area. The book comprises three parts: An introductory section describes a Bayesian strategy for longitudinal modeling with complete data; the second part adapts the modeling to MAR incomplete data; and the final section offers a paradigm for analysis of sensitivity to departures from MAR. The authors illustrate the methods in an array of examples drawn from recent clinical trials and observational studies in psychiatry, geriatrics, smoking cessation research, and HIV medicine.

The book is comprehensive in covering models for both continuous and discrete outcomes from both the pattern mixture and selection modeling perspectives. Analyses of sensitivity to nonignorability are organized around the identification of a sensitivity parameter—i.e., a parameter whose value affects the prediction of missing observations but not the fit of the model. Thus a complete expression of the authors' perspective in the analysis of a particular set of data would involve positing a model for the underlying complete data and dropout process; estimating it under MAR; then isolating the sensitivity parameter and constructing a prior for it, executing an adjusted analysis by averaging over this source of uncertainty.

The book's composition offers much to admire. The writing is clear and direct, the notation is sensible and consistent, and tables and figures are simple and uncluttered. Typos are mercifully rare; I estimate about one every eight pages.

I have only two criticisms worth mentioning: First, I am less sanguine than the authors about the possibility of using "expert opinion" and "historical data" to inform the prior distributions of nonignorability parameters in Bayesian sensitivity analysis. The problem is that relevant "historical data" seldom exist, and without such data, it is difficult to regard the opinions of even renowned experts as anything more than conjecture. Second, with the methodology's reliance on sampling-based inference procedures, readers may question the authors' decision to limit discussion of computing to an 11-page overview in Chapter 3. By the time one reaches the novel data analyses, there is seldom more than a perfunctory reference to the lengths of the Markov chains and the burn-in periods. Moreover, I was unable to find a link to the example WinBUGS code that the preface claims is posted on Professor Hogan's website.

Because this is a research monograph, the descriptions of background material take the form of brief but thoughtful synopses, and there are no exercises. The book would be ideal for self-study or as the text for an advanced course in longitudinal modeling. A more applied or basic course could use it as a supplement.

Dropout of uncertain provenance evidently is no less common a feature of longitudinal studies today than it was in the early 1990s. Fortunately we now possess a much better understanding of how to extract valid statistical inferences from such data. Biostatisticians who seek a clear and thorough overview of the state of knowledge in this area would do well to make this excellent book their first stop.

Daniel F. Heitjan
Department of Biostatistics & Epidemiology
University of Pennsylvania
Philadelphia, Pennsylvania, U.S.A.

BROEMELING, L. D. **Bayesian Biostatistics and Diagnostic Medicine**. Chapman and Hall/CRC Press, Boca Raton, Florida, 2007. xii + 198 pp. $79.95/£42.99. ISBN 9781584887676.

Drawing on his collaborative experiences with medical researchers and his long-standing interests in Bayesian methods, the author of this book shows how the Bayesian approach can be used to advantage when medical diagnosis is based on data with uncertainty. There is much emphasis on diagnosis via medical imaging, though other diagnostic modalities are considered as well.

After a brief chapter 1 which previews the book, chapter 2 describes the use of various imaging technologies in diagnostic medicine and the attendant statistical issues of diagnostic accuracy and interobserver agreement. Special attention is paid to diagnostic procedures in clinical trials. Indeed, a general strength of the book is careful discussion of study designs and protocols, which is a bonus relative to many biostatistical books written from a more narrow theory and methods perspective. A very brief third chapter describes some specific questions of accuracy or agreement arising in three studies.

Armed with motivating material, the author introduces Bayesian statistics in chapter 4. The ambitious attempt here is to cover the basic core of Bayesian inference and computation in about 30 pages. Noninformative and conjugate priors, WinBUGS and Minitab examples of inference in binomial sampling models, and a discussion of hypothesis testing feature prominently. An addition relative to most Bayes introductions is material on sample size and power (though this is limited to one of multiple possible views on this topic). The chapter concludes with discussion of Markov chain Monte Carlo (MCMC) methods, which would give an uninitiated reader some clue about what happens "under the hood" in the WinBUGS examples earlier in the chapter. As a point of historical curiosity, and in contrast with other introductions to Bayes, in this chapter the author cites the 1923 work of Lhoste as important in the historical development of Bayesian statistics.

Chapter 5 applies Bayesian methods to inferring diagnostic accuracy. More particularly, Bayesian models are fit to data on binary, ordinal, and continuous diagnostic scores, with receiver operating characteristic curves playing a prominent role in the latter two cases. Attention then turns to a multivariate scenario with a reader scoring multiple regions of each subject's image, and a setting with two diagnostic scores arising from different technologies. The chapter closes with a detailed look at sample size determination for inferring test accuracy. A short chapter 6 follows up with extensions to settings where test accuracy depends on subject covariates.

Chapter 7 broaches the topic of multiple raters, with a primary example being multiple radiologists scoring the same images. Well-entrenched frequentist procedures revolving around measures of agreement, such as the kappa statistic and intraclass correlation, are reworked from the Bayesian point of view.

Chapter 8 turns to more specialized issues of diagnostic tests within clinical trial settings. The author's extensive applied experience in this area shows through in the discussion of phase I to III trials, and the use of Bayesian sequential stopping rules. Various other advanced topics are collected in chapter 9, including the analysis of data not including gold-standard measurements, verification bias, and lead-time modeling.

The author adopts a lively tone throughout the book, moving briskly from topic to topic, and model to model. A real strength is the strong integration between models and concepts on the one hand, and real studies on the other hand. The inclusion of WinBUGS code is also a plus. It has the practical benefit of allowing the author to focus on modeling issues and the interplay between models and science, without undue distraction concerning MCMC implementation details.

One can always find things to quibble with. A minor distraction is that some of the figures could be more polished (Greek letters "spelled" on axis labels, undersmoothed and non-normalized histograms of MCMC output representing posterior distributions), but it is relatively easy to see past this.

My impression is that this book can and will serve multiple roles. Practicing statisticians working with diagnostic test data will find much to imitate and improvise upon should they wish to utilize Bayesian technology. Academic statisticians will be brought up to speed on both basic and advanced topics. And with the inclusion of exercises at the end of each chapter, the book could serve as a text for a fairly specialized graduate course.

Of course there is a line of thought that diagnostic testing is inherently a Bayesian pursuit, because germane quantities such as positive and negative predictive values, which describe the distribution of disease status given test data, arise via direct application of Bayes' theorem. Whether or not one subscribes to this view, this book is highly recommended for anyone whose interests touch on the statistical side of diagnostic medicine.

Reference

Lhoset, E. (1923). Le calcul des probabilities appliqué a l'artillerie, lois de probabilitie a prior. *Revu d'artillerie*, Mai, 405.

Paul Gustafson
Department of Statistics
University of British Columbia
Vancouver, British Columbia, Canada

BUNCHER, C. R. and TSAY, J.-Y. (eds). **Statistics in the Pharmaceutical Industry**, 3rd edition. Chapman and Hall/CRC Press, Boca Raton, Florida, 2006. 504 pp. $119.95/£51.99. ISBN 9780824754693.

This is the third edition of the book that first appeared in 1981. The second edition was published in 1994 and this new edition was published in 2006. Since the original edition was published much has changed in both the pharmaceutical industry as well as the field of statistics. This new edition attempts to catch up with these changes. On the whole, the editors succeed in compiling a set of chapters that give the reader a broad overview of many of the statistical issues that are faced when doing research in the pharmaceutical industry and therefore it is recommended to individuals who work or consult in this area of statistics. We now describe in more detail some of the elements of this book.

The goal of this book appears to be to provide to a diverse audience information about statistical issues that arise in studies conducted in the pharmaceutical industry. It is written at a level of sophistication that is appropriate for graduate students in statistics or biostatistics while maintaining a generality that allows applied researchers, without advanced

statistical training, to still benefit from its contents. This third edition contains 24 chapters contributed by 35 authors. The second edition had 25 chapters, but the differences between editions amount to more than one chapter. The new edition has added chapters including some that now discuss such issues as testosterone replacement therapy, active controlled noninferiority/equivalence trials, global harmonization of drug development, bridging strategies in global drug development, and reference intervals. In addition, there are now three chapters that all examine issues in interim analyses (from slightly different perspectives) and this is in contrast to the previous edition where there was one chapter dedicated to this topic. Overall, the contents of the 24 chapters in this edition cover all the major statistical issues that confront researchers during the implementation of pharmaceutical trials.

One of the more useful aspects of this book is that it describes nicely the different stages of drug development. These descriptions include an overview of the types of trials needed at each stage of drug development as well as nice flow charts depicting the interaction between the drug sponsor and the regulatory agencies involved in evaluating the drug. Within these descriptions, the authors present all of the abbreviations that are often used in the pharmaceutical industry and give clear definition of them (e.g., IND, NDA, CDER, ICH, etc.). While this may seem a trivial point, in fact it is not. Because most of the actual statistical methods used in the pharmaceutical industry can be considered "standard" (with the exception of the expansion of the use of adaptive designs to be discussed later in this review) the barrier that many statisticians face when interacting in this area of research is understanding all of the jargon that is used in casual conversations on the matter. Having a nice reference book that covers most of the "insider" jargon is a useful tool.

Another positive feature of this edition of the book is that the contributors for this edition have a variety of backgrounds. There is a good mix of academic statisticians, regulatory or government statisticians, and industry statisticians from both the United States and abroad. One could argue that there may be an underrepresentation of international statisticians (7 out of 35) given that one of the themes of this new edition is its attempt to deal with "global" issues in drug development and not merely ones related to the process and system within the United States. Still, the diversity of the backgrounds of the authors provides an opportunity to examine different chapters from different perspectives. For instance, the three chapters that examine interim analyses have different focuses and approaches. The third of these chapters describes "A regulatory perspective on data monitoring and interim analysis" and is very useful because rather than focusing specifically on statistical methods for interim analyses it describes issues that the regulatory agencies must consider during this process. Understanding these issues is often as important as understanding the statistical methods used to perform an interim analysis.

The new chapter on noninferiority/equivalence trials is a very important one because clearly the use of such trials has been growing rapidly in the pharmaceutical industry. The discussion and description in this chapter is good and the reference list for this chapter is very helpful. Given the difficulty of using placebo controlled trials in the future of pharmaceutical research, the use of noninferiority and/or equivalence trials will increase. One issue related to this area of research that is mentioned but probably could have been expanded upon is the potential drawback in the use of noninferiority trials without proper oversight. The problem one can see is that, by definition, a new trial can be designed where the new drug has a point estimate for treatment efficacy that is inferior to that of an existing drug, but still does not cross a noninferiority boundary upon statistical testing. Such instances can occur, and without proper oversight one can imagine that a series of such noninferiority trials could be designed such that over time the original standard of care is reduced. This book describes the concept of "fraction of active control effect preserved by the test treatment" and the "putative placebo" approach, which address in some way the issue of a diminishing treatment effect over time; however, the chapter could have been clearer on the potential drawbacks of such trial designs.

The new chapters on global harmonization and global drug development are useful mainly because they provide some real world context for issues surrounding the development of pharmaceutical agents in a global community. These chapters do not contain statistical issues per se, but still provide valuable insight to the applied statistician who needs to understand the "big picture" when working in pharmaceutical industry research.

Despite the good features described above, there are still some areas of growing concern in pharmaceutical statistics that this book does not cover with enough depth. First, there are growing concerns in the medical community about whether there is enough postmarketing surveillance being performed to carefully identify and monitor possible long-term safety concerns for approved drugs or agents. This book discusses briefly the issue of phase IV postmarketing studies, but the issue is larger than what is described. Due to the discovery of serious adverse events linked to approved drugs that have in some cases led to drugs being taken off the market, the statistical issue of signal detection in large databases is becoming a real question that needs to be examined. How can one monitor/examine large health-care databases in such a way to find "safety signals" when they exist when the data collected in these databases were not designed for this purpose?

The second issue is a more global statistical concern, but it clearly affects statistical research in the pharmaceutical industry. This issue is that currently many clinical researchers are performing studies (or meta-analyses) examining pharmaceutical agents using data that are derived from administrative databases, and not clinical trials. The use of these data to make inferences about the safety and efficacy of existing agents is seen monthly in medical publications ranging from disease/organ specific journals to more global journals such as *JAMA*. How should such retrospective, observational data analysis be performed?

While both of these topics could be books on their own, they both are clearly ones that face the statistician involved in pharmaceutical industry research on a daily basis. Neither issue is covered in any depth in this book, which is unfortunate.

Still, on the whole this book does provide the researcher a good reference for many of the statistical issues that confront a statistician working in the pharmaceutical world. It is useful

to the student learning about the techniques found in the industry, to the academic or independent statistician who may consult with regulatory or pharmaceutical industry, to the industry statistician who may design and analyze these types of studies regularly, to the regulatory statistician who evaluates results from such studies, and to the applied researcher who may not be a statistician but may be someone who works closely in some aspect of drug development. In conclusion, this book is a useful reference and is recommended for researchers/statisticians who work on problems in the pharmaceutical industry with some regularity.

Ralph B. D'Agostino Jr.
Department of Biostatistical Sciences
Wake Forest University School of Medicine
Winston-Salem, North Carolina, U.S.A.

MITRA, S., DATTA, S., PERKINS, T., and MICHAILIDIS, G. **Introduction to Machine Learning and Bioinformatics**. Chapman & Hall/CRC, Boca Raton, Florida, 2008. 366 pp. US$79.95/£39.99. ISBN 9781584886822.

Bioinformatics is an interdisciplinary field that involves dealing with large, diverse, and continually growing biomedical data sets. These sets often require intelligent learning tools to turn the raw data into scientific knowledge and benefit the medical, biological, and even the computational fields. Machine learning, data mining, and computational intelligence have become immediately useful computational tools to tackle the associated challenges. Although important publications including text books and web resources are available today, one of the remaining challenges is the great diversity of the interested learners and researchers with various backgrounds and disciplines, which includes computer scientists, statisticians, mathematicians, and most importantly, medical researchers.

The authors of the book intend to provide a good text/reference book that summarizes the latest developments in the interface between bioinformatics and machine learning, and to offer a thorough introduction to each field. One nice feature of this book is that it provides detailed comparisons of this book to other related published books and edited volumes, including strengths and limitations, which provide guidelines for the audience to decide whether to use it as a text book or a reference book. The authors also provide user-friendly and sufficient background material in each of the first five chapters that may be suitable for students with various backgrounds and different levels up to the advanced Ph.D. level.

After a general introduction, the book starts with a biological introduction in chapter 2. From chapter 3 to chapter 7 the authors provide detailed introductions of computational tools including, but not limited to, machine learning domains. In chapter 3, probabilistic learning and model-based learning are given. Supervised learning and classification methods are provided in chapter 4, while unsupervised learning and clustering methods are discussed in chapter 5. These five chapters describe the fundamental concepts and key algorithms of machine learning with many realistic, classical examples, which were drawn from bioinformatics as well as nonbioinformatics applications. Some material from these five chapters may overlap with standard data mining and statistical methods.

Advanced computational intelligence tools start from chapter 6, which include artificial neural networks, fuzzy systems, evolutionary computation, and rough sets. The authors first provide the definitions of these soft computing techniques and then turn to providing bioinformatics applications, including protein structure predictions, gene-network inferences, and sequence alignment. In chapter 7, the connection between machine learning and bioinformatics is established by further describing bioinformatics problems and machine learning applications. Other applications and examples of bioinformatics are given from the following fields: 3D protein images in chapter 8; gene expression data with soft computing in chapter 9; Bayesian machine learning in chapter 10; and statistical methods for proteomic data in chapters 11 and 12.

Statistics, computer science, and information science are gradually merging toward a more powerful computing field due to their strong medical applications, as it has become evident by the development of bioinformatics. One of the strengths of this book is the clear notation with a mathematical and statistical flavor, which will be attractive to *Biometrics* readers, especially to those new to statistical learning and data mining. It is also very readable for a variety of interested learners, researchers, and audiences from various backgrounds and disciplines.

One limitation of this book is that it did not mention computing software, although there is a wide variety of choice. The examples and exercises are taken from classical domains and are sometimes not directly related to the bioinformatics domain, which makes the introduction of machine learning techniques only moderately connected to bioinformatics. Because this is an introductory level of machine learning and bioinformatics book, some most recent research developments of these fields, such as embedded learning, semisupervised learning, and the applications of machine learning for the human genome project and single nucleotide polymorphism study are not included. Therefore, some advanced readers may feel that the book is not comprehensive enough and less updated; however, this book is undoubtedly a valuable resource for the general public.

Yulan Liang
Department of Family and Community Health
University of Maryland
Baltimore, Maryland, U.S.A.

SIEGMUND, D. and YAKIR, B. **The Statistics of Gene Mapping**. Springer, New York, 2007. xvii + 331 pp. $79.95/€69.50. ISBN 9780387496849.

Statistics plays an important role in the field of genetics, especially for the identification of genes responsible for many traits in diverse organisms. Many books have been published to date covering the statistical concepts and methods in gene mapping, a field that has been undergoing rapid developments recently due to major technological advances. What sets the book by Siegmund and Yakir apart from others is its focus on a number of key topics and concepts that have not received much discussion in other books and its emphasis on hands-on

learning through its extensive use of R. As such, any reader of this book who has spent time and effort to thoroughly digest the materials will be able to have a good grasp of the genetics problems presented to statisticians and a deep understanding of various designs and statistical tools developed over the past century in this field.

Although the book is aimed at a broad audience, a solid background at the undergraduate level of probability and statistics is needed to benefit most from this book. The first chapter reviews relevant topics in probability distributions and statistical inference, and the next two chapters introduce key concepts in genetics problems discussed later. These three chapters, comprising part I, provide the starting point for the rest of the book and enough opportunities to experience R.

Part II (chapters 4 to 8) focuses exclusively on mouse genetics involving crosses between inbred mice. The content evolves naturally from single marker analysis in chapter 4 to genome-wide analysis, with an excellent coverage on the issue of controlling false-positive rates at the genome level in chapter 5. Noncentrality parameters and local alternatives are introduced in chapter 4 and these concepts are used repeatedly in later chapters to evaluate the statistical power of a specific design and the effects of various factors, e.g., types of crosses and recombination fractions on the statistical power. The use of noncentrality parameters throughout this book is a distinct feature and it brings together naturally the consideration and statistical treatments of different designs and data. The discussion of multimarker analysis in chapter 5 and the determination of global threshold are excellent. Although no mathematical details are provided, the authors give a very lucid discussion of how the test statistics across different markers are correlated and how this knowledge can be used to approximate the distribution of the global maximum statistic. Chapter 6 is devoted to statistical power and confidence intervals in mapping genes. Although the mathematical formulas presented are not trivial, the authors succeed in conveying the key insights from the equations and supplement them with many examples that the readers can try out following the R codes in the book. The topic of interval mapping is thoroughly treated in chapter 7, from missing genotype imputation to statistical power. Although the specific missing data problem discussed in this chapter can be largely overcome by high-density genotyping platforms these days, it does teach the readers valuable statistical tools and thinking to deal with missing data problems that can be applied to other problems. The second part ends with chapter 8 that touches upon a number of more advanced topics, many of which are still not well addressed and under vigorous developments. For example, the issue of associating a phenotype with a number of markers presents significant statistical and computational challenges, and this chapter serves as a nice point from which interested readers can explore potential research topics on their own.

The last part of this book discusses a few selected topics in human genetics. Affected sib-pair design is the focus of chapter 9, which covers all key aspects of this widely used design to map genes for complex traits in humans. Admixture mapping is a relatively new research topic and the authors provide a succinct overview of the key ideas in this area in chapter 10. Hidden Markov models are introduced in this chapter as a tool to infer population origin of a specific locus in the genome. After a brief detour in chapter 10, chapter 11 returns to linkage analysis on more complex human pedigrees than affected sib pairs. No R codes are provided, possibly due to the complexity of the content in this chapter. This is probably the most dense chapter to read in this book. Chapter 12 gives an introductory account of the three most important topics in association analysis: case–control studies, population stratification, and family-based association studies. This is an area that is most actively pursued in current human genetics research as exemplified by weekly success stories from genome-wide association studies in top journals and media. The last chapter is devoted to statistical methods for the analysis of haplotypes, the focus of the International HapMap Project. Both haplotype inference and association are discussed, and the content of this chapter should provide a good starting point to understand more sophisticated methods that have been developed recently for haplotype analysis.

What I like most about this book is that the authors have made great efforts to explain the genetics problems in detail and discuss the key statistical concepts and ideas without using complicated notation and equations. As major contributors to this field, they share many of their insights with the readers in this book. As stated by the authors, this book does not cover all topics of the field, which is impossible, both due to the very large scope of the problems that have been tackled over the past many years and the active developments in this area. All the readers will appreciate the R codes given by the authors and the well-thought-out exercises that can reinforce the concepts discussed in the book. Any serious reader, new or old to the field, will come away with a deeper understanding of the many subjects discussed, and I highly recommend this book to anyone who has interest in learning the field and even those who are in the field.

Hongyu Zhao
Division of Biostatistics
Yale School of Public Health
New Haven, Connecticut 06520, U.S.A.

WANG, S.-C. and PETRONIS, A. **DNA Methylation Microarrays: Experimental Design and Statistical Analysis**. Chapman & Hall/CRC, Boca Raton, Florida, 2008. 256 pp. $79.95/£41.99. ISBN 9781420067279.

We have entered an exciting era, with microarray technologies allowing the study of innumerable molecular features. The first widely adopted technology were gene expression arrays, which allowed the rapid profiling of thousands of genes. Later, single nucleotide polymorphism genotyping arrays gained popularity for linkage analyses, association studies to identify common variants associated with common diseases, and quantifying copy number changes in the human genome. Subsequent advances have led to the development of DNA methylation microarrays. The experimental design and statistical analysis of such array technologies are the focus of this book. DNA methylation is a chemical modification of the DNA occurring, in mammals, at CpG sites; it is sometimes called the fifth base. DNA methylation has normal roles in

X-chromosome inactivation and genomic imprinting, and has been linked with many diseases. As a relative newcomer to the genomic revolution, its contribution to human disease is currently a matter of intense investigation.

This book is a helpful guide for researchers and students with an interest in performing genomic studies using high-throughput microarrays. It is written in three parts. Part I, chapters 1 and 2, provides the basic introduction to statistics and to the hybridization-based microarray technologies. Part II, chapters 3–10, contains the details of experimental design and statistical analysis. Chapter 3 describes various experimental designs, including a reference design, loop, factorial and others, and shows how to determine the number of arrays and samples necessary to test an a priori hypothesis. Chapter 4 describes methods for data normalization/preprocessing and chapters 5–10 statistical analysis of the preprocessed data. The first two of these chapters deal with identifying differentially methylated loci, and chapter 8 with statistical classification. In chapters 7, 9, and 10, exploratory analyses related to class discovery and to dependence and regulatory networks are presented. In the final part, three chapters cover the topics of online databases for genomic annotation, public repositories for microarray data, and open source software for data analysis. The figures in the book are in black and white, but an accompanying CD provides figures in color. Many of the figures do not require color for interpretation; however, those showing red-green heatmaps bring to attention a challenge faced by our colleagues with red-green color blindness.

The book will be most useful for the person who will be analyzing microarray data. Statisticians are likely to understand the most; however, biologists will surely benefit as well. A wide range of useful data analysis tools are covered without providing too much mathematical detail. Key formulas are provided, but the principal focus is on illustrating methods through the use of examples and figures. On occasion, an example is given without sufficient detail in order to follow the implementation of the methods described, or an assertion is made without supporting justification, but for the most part, this happens rarely. Other strengths throughout the book include the discussion of experimental design, the mention of software for certain analyses, and the inclusion of more advanced methods such as wavelets and genetic algorithms.

Many of the statistical methods described originated in the area of gene expression analysis (e.g., data preprocessing/normalization and volcano plots) and would be covered by other textbooks. What this book offers in addition is a few new methods, e.g., the biplot, and the many examples that use DNA methylation data for illustration. It is a bit of a disappointment that there is not more highlight of methods or issues that are particular to epigenetic data. For example, the chapter on experimental design does not mention designs specific to epigenetics, such as the use of identical twins to study environmental epigenetics, an area in which the authors are active. Also along these lines is the lack of mention of other microarray technologies, such as Illumina's widely used BeadArray technology (Illumina, San Diego, California). Illumina's BeadArray may not strictly fall under the hybridization technologies described in this book, and the data produced are qualitatively different. In particular, Illumina measures a beta-value, which is measured on range of 0 to 1. A statistical property of the beta-value distribution is that the variance is related to the mean. For such data, novel statistical methods are under development, e.g., Houseman et al. (2008). However, to be fair to the authors, the Illumina BeadArray technology is not an unbiased method like those that are discussed in the text, and the novel statistical methods developed for it were unpublished before their book went to press. Nonetheless, it is worth a reminder that the experimental platform is very important to the analysis method, and that the platforms available are constantly changing.

Overall, this book gives a nice summary of methods used for the analysis of hybridization-based microarray data. In light of the continued development of statistical methods (e.g., see Down et al., 2008), the book is not a substitute for a collaborator with expert knowledge in the area. Even so, the methods are highly relevant for technologies in use today, and will help researchers answer important biological questions.

References

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E. M., Thorne, N. P., Backdahl, L., Herberth, M., Howe, K. L., Jackson, D. K., Miretti, M. M., Marioni, J. C., Birney, E., Hubbard, T. J., Durbin, R., Tavare, S., and Beck, S. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology* **26,** 779–785.

Houseman, E. A., Christensen, B., Yeh, R.-F., Marsit, C., Karagas, M., Wrensch, M., Nelson, H., Wiemels, J., Zheng, S., Wiencke, J., and Kelsey, K. (2008). Model-based clustering of DNA methylation array data: A recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9,** 365.

Kimberly D. Siegmund
Department of Preventive Medicine
University of Southern California Keck School of Medicine
Los Angeles, California 90089, U.S.A.

DUDOIT, S. and van der LAAN, M. J. **Multiple Testing Procedures with Applications to Genomics**. Springer, New York, 2008. xxxiii + 588 pp. $84.95/€74.85. ISBN 9780387493169.

Hypothesis testing problems in genomics applications are challenging, because they often involve thousands and even millions (!) of hypotheses, with complex dependencies between the test statistics. For example, in microarray studies the expression levels of thousands of genes are simultaneously measured and the gene expression profiles are then tested for the purpose of identifying the differentially expressed and co-expressed genes. With high-speed computers, modern statisticians can perform a large number of statistical tests in a very short time. Performing these tests without adjusting for the multiple testing may lead to hundreds or thousands of false positives. Therefore, it is absolutely necessary to adjust for the multiple testing. The burgeoning field of genomics has raised the need for new multiple testing methods that will control for false positives in appropriate ways. Recognizing this need, the authors set out to develop and implement (in both `R` and `SAS`) multiple testing methods.

The book focuses on the methodology for multiple hypothesis testing developed by the authors and their collaborators

in a collection of technical reports. The methods suggested in this book are designed to control a broad class of error rates, including the false discovery rate (FDR) and the tail probability of the false discovery proportion, which may be more appropriate (and may lead to more powerful procedures) than the classical familywise error rate (FWER) in large-scale problems. The book advocates resampling methods that exploit the joint distribution of the test statistics to gain power. Similar to Westfall and Young (1993), the book suggests that bootstrap procedures provide the most appropriate solutions. However, while in Westfall and Young (1993) the procedures are designed to control the FWER, this book suggests a broad class of error measures in addition to the FWER. The suggested bootstrap procedures are compared to procedures in Westfall and Young (1993) as well as to a few other methods, such as the step-up BH procedure of Benjamini and Hochberg (1995). While the methods compare favorably to existing FWER controlling procedures, they may be conservative when controlling less stringent error rates, such as the FDR.

Chapter 1 introduces the multiple hypothesis framework, and discusses the main ingredients of a multiple hypothesis problem, including the choice of a type I error rate and the notion of adjusted p-values. The notation introduced is rigorous but somewhat cumbersome. Chapter 2 is already quite technical. It addresses the general characterization and explicit construction of proper null distributions for the test statistics. Chapter 3 provides a clear overview of multiple testing procedures, emphasizing the distinction between marginal procedures that are solely based on the marginal distribution of the test statistics and joint procedures that take into account the dependence structure of the test statistics. The joint procedures are examined in detail in chapters 4–7: chapter 4 proposes general joint single step procedures; chapter 5 proposes joint step-down procedures; chapter 6 proposes augmentation to the procedures suggested so far to control various type I error rates; and chapter 7 proposes new joint resampling-based empirical Bayes procedures. Chapters 4–7 are the most technical, and it is possible to postpone reading them till after reading chapters 8–12, which illustrate the suggested methodology very nicely. These illustrative examples give a sense of the advantage of the various procedures and demonstrate the complexity and importance of the problems that need to be addressed in genomics. Chapter 8 presents in detail simulation studies to assess the performance of the suggested procedures for problems concerning correlation and regression coefficients. Chapters 9–12 introduce detailed examples including code for the following problems: identification of differentially expressed and coexpressed genes in high-throughput gene expression experiments, such as microarray experiments; tests of association between gene expression measures and biological annotation metadata (e.g., gene ontology); sequence analysis; and the genetic mapping of complex traits using single nucleotide polymorphisms (code in Appendix C). Finally, chapter 13 discusses the software implementations. Appendix A gives a useful summary of the multiple testing procedures, and Appendix B summarizes known mathematical results that are used in the proofs. The book website includes supplementary material as well as the code and the example datasets. Unfortunately, at present the stated URL leads to a password-protected location.

The book is a useful resource for teaching an M.S. or a Ph.D. level course on multiple testing, in conjunction with a text on the classical approach and additional texts that introduce other modern approaches. The theoretical foundations of a general methodology for multiple hypothesis testing set out in the book can benefit statisticians developing their own methodology. The specific methods suggested can be useful for applied scientists encountering high-dimensional testing problems in their subject matter area.

In conclusion, this book is an important contribution to the literature on multiple comparisons. It proposes modern resampling methods to handle the large-scale problems in genomics applications. These problems are difficult or impossible to solve by classical means. Readers will have no difficulty applying the methods by using the available software and following the detailed examples in the book.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57,** 289–300.

Westfall, P. and Young, S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. New York: Wiley.

Ruth Heller
Department of Statistics, Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania, U.S.A.

LAZAR, N. A. **The Statistical Analysis of Functional MRI Data**. Springer, New York, 2008. xiv + 299 pp. $84.95/€64.15. ISBN 978-0-387-78190-7 (hardcover).

The *Statistical Analysis of Functional MRI Data* is a timely introduction to the current state of the field. As the author states in the preface, "The primary intended audience is statisticians who are interested in this growing field, and who wish to gain an understanding of the major problems and current solutions. A secondary audience is cognitive psychologists and other neuroscientists who use functional MRI (fMRI) as a research tool." This reviewer (a statistician) read the book in a study group filled largely with cognitive psychologists; the consensus for our group was that the book was mostly successful in its stated intentions, if somewhat too brief. In short, the book served as a useful introduction and summary for both groups of researchers to the statistical analysis of fMRI, and also pointed to the relevant literature for those interested in learning these methods more thoroughly.

The book begins by describing the science of fMRI, including brain anatomy, magnetic resonance, and acquisition of images. This is aimed at researchers who have only glancing familiarity with MRI (or neuroscience in general), and would probably be best followed up with the perusal of a more detailed textbook on the subject (e.g., Huettel, Song, and McCarthy, 2004). Likewise, chapter 2 is a brief introduction to the design of fMRI experiments, including the setting of image acquisition parameters and the presentation of experimental stimuli (e.g., block or event related). Both of these chapters should serve to familiarize statisticians with major

topics in the practice of fMRI to the extent that they can research the published literature more thoroughly or initiate discussions with experienced fMRI scientists. The first portion of the book concludes with chapter 3, which is an overview of sources of noise and variation in the fMRI signal and the typical preprocessing steps the data are subjected to before a statistical analysis is usually attempted. Choices made at this stage can have a huge impact on downstream analyses of the data and it is crucial for statisticians involved in the analysis of fMRI data to have familiarity with the preprocessing stream, even if they do not participate directly at this stage.

Chapters 4 and 5 begin the more statistically oriented portion of the book, considering voxel thresholding, multiple subjects, regional versus whole-brain analyses, block- and event-related experimental designs, and the oft-used general linear model. These topics are introduced in generality, though the author also highlights some selected recent developments that many researchers knowledgeable about the field may not be aware of.

Much of the last half of the book (chapters 6–9) covers advanced topics that reflect more recent advances in the field. These include temporal and spatiotemporal modeling of fMRI data in chapter 6, multivariate approaches (e.g., independent components analysis) in chapter 7, basis function approaches in chapter 8, and Bayesian methods in chapter 9. Many of these methods have yet to be adopted by the vast majority of practicing fMRI researchers and are very active areas of development from a methods perspective. Statisticians interested in methods development should get a good sense of what areas have attracted recent attention from this part of the book.

Chapter 10 covers the problem of multiple testing, including false discovery rate methods. Chapter 11 is a brief overview of miscellaneous issues (including the rapidly expanding area of functional connectivity analysis), and chapter 12 concludes with a case study utilizing eye motion data.

No book of this length can hope to cover all (or even most) of the topics related to the statistical analysis of fMRI in detail. However, major areas of development for statistical methodology are clearly described and much of the relevant literature is cited. This serves the useful purpose of giving researchers (both statisticians and practicing fMRI scientists) a good idea of what the major topics of interest are and where to start looking to learn them more thoroughly. This book is a good first step in bridging the gap between statisticians and practicing fMRI scientists, and will hopefully serve to spur better statistical practices among fMRI researchers and the acceleration of methodological development among statisticians.

Reference

Huettel, S. A., Song, A. W., and McArthy, G. (2004). *Functional Magnetic Resonance Imaging*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Wesley K. Thompson
Department of Psychiatry
University of Calfornia, San Diego
La Jolla, California, U.S.A.

IACUS, S. M. **Simulation and Inference for Stochastic Differential Equations with R Examples**. Springer, New York, 2008. xv + 284 pp. $79.95/€69.50. ISBN 978-0-387-75838-1.

Stochastic differential equations (SDEs) are used to model time series, where the time evolution of the process is altered with stochastic noise, forcing the system to stray from an otherwise deterministic path. While the topic of SDE's has been studied to great depth through abundant variations, to condense this topic into an introductory book, only first-order differential equation models with Gaussian noise at the level of the derivative are considered. Processes involving measurement error are beyond the scope of this text. This book succeeds at giving an overview of a complicated topic through a mix of simplified theory and examples, while pointing the reader in the right direction for more information. But like any introductory text it will leave some readers following the references to satisfy cravings for more detail. Consequently, this would be a good introductory or reference text for a graduate level course, where the instructor's knowledge extends substantially beyond the book.

The `R` examples advertised in the title play a prominent role in the development of the introduction to SDEs. Readers unfamiliar with the `R` language will appreciate the readable code and may consider it as an algorithmic outline that complements the discussion. For those new to `R`, the first appendix is filled with general lessons in using and programming in this language. The second appendix is a set of help files for the accompanying, free `SDE` software package. The `SDE` software for `R` is as simple to use as the book suggests and is general enough to apply to a wide range of model variations.

Chapter 1 gives an overview of the mathematical background required for the text and introduces a few examples of SDE models. The book uses some measure theory and readers with this background will certainly get more from the book; however, the use of sigma algebras and martingales are kept to a minimum to keep the book accessible. Measure theory is mostly used to hint at the greater mathematical depth of the topic, which is explored in some of the references. Although some knowledge about stochastic processes would be a preferred prerequisite, the background in the introductory chapter is sufficient such that a statistics master's student or mathematically mature undergraduate student should be able to understand the book.

Chapter 2 describes numerical methods for simulating the trajectory of an SDE model. Chapter 3 gives the basics of parameter estimation in SDE models from the perspectives of likelihood inference, method of moments, discretizations of continuous time estimators, and more. In this chapter Bayesians will find their two pages of book space packed with an explanation of the Bayesian paradigm, a description of Markov chain Monte Carlo methods and how they can be applied to SDE models, along with some keywords for your next literature search. Chapter 4 introduces the Akaike information criterion, the use of nonparametric density estimation, and change-point detection. While the sole real data example is confined to the final page and a half of the book, simulated data examples are abundant and give the book the feeling of

being practical while showcasing when methods succeed and fail.

Dave Campbell
Department of Statistics and Actuarial Sciences
Simon Fraser University
Surrey, British Columbia, Canada

MARKOVICH, N. **Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice**. John Wiley and Sons, Chichester, 2007. xxi + 310 pp. $120.00/€82.90. ISBN 9780470510872.

This book adopts a pragmatic approach to the analysis of data with spare observations in the tail domain of distributions revealing slower than exponential decay to 0. The author mixes parametric and nonparametric methods, and makes incursions in the realm of multivariate data analysis, whenever advisable.

In the early stages of mathematical statistics, Wilks and Gumbel, in a series of papers that culminated in their books (Gumbel, 1958; Wilks, 1962), extensively used extreme-order statistics, but the underlying extremal limit theorem had a limited scope, with stringent identical and independent distribution assumptions. Multivariate extensions and the development of probabilistic results for weakly dependent structures expanded the field, providing the appropriate background for applications. De Haan brought in Karamata's theory of regular variation; soon after, the *POT* (peaks over thresholds) methodology enhanced the intimate relations of the *GEV* (general extreme value) with the *GP* (generalized Pareto) distributions.

Regular and extended regular variation have been leading tools for the understanding of extreme value data, and the tail index, the inverse of the regular variation exponent, plays a prominent role in the analysis of heavy-tailed data. The Hill estimator inaugurated a glorious era of semiparametric and nonparametric developments, cf. the first chapter of Markovich's book, as well as Drees (2008), Gomes et al. (2008), and Hüsler and Peng (2008), for coordinated overviews of the state of the art in the field.

In chapter 1, after a brief presentation of classes of heavy-tailed distributions, several estimators of the tail index are compared; there is an interesting discussion on the optimal number of top-order statistics when using Hill's estimator, namely, using bootstrap techniques. "Rough" detection of tail heaviness, dependence in univariate and bivariate data, and long-range dependence are discussed, and detailed exemplification with web traffic and transmission control protocol flow data are provided.

Chapters 2–5 in the book are an overview of probability density estimation. Chapter 2 discusses the principles, an excellent starting point to make it useful for all readers; it goes far beyond the basic facts, discussing for instance estimation using dependent data. In chapter 3, parametric methodologies to estimate the tail are discussed, mixed with nonparametric methodologies to fit the body of the distribution, as well as variable-bandwidth kernel estimators. The transformation–detransformation technique, briefly presented in chapter 3 as an alternative to kernel, projection, and spline nonparametric estimators, is further addressed in chapter 4, where fixed and adaptive transformations are discussed. An interesting application of detransformation techniques to classification, using an empirical Bayes' algorithm, is described in chapter 5.

Chapter 6 is a nice round-up of high quantile estimation and distribution theory; simulated data and an application to web traffic data enhance the quality of the presentation. Chapter 8 discusses nonparametric estimation of the renewal function, with an inspiring application to transmission control protocol (TCP) flow data.

Chapter 7 focuses on the estimation of the hazard rate function, being of direct interest for those working in survival issues. There is an interesting application to hormesis detection, a bit difficult to follow for nonspecialists.

Technical proofs are postponed to appendices. At the end of each chapter notes and comments enhance important points, and a selection of exercises, with appropriate suggestions, is useful for a full grasp of the matter. Those wishing to use (part of) this text in advanced courses will find slides in the companion website http://www.wiley.com/go/nonparametric.

This book heavily relies on the authors' many achievements in the field, and in this sense there is no direct competitor. However, the recent books by de Haan and Ferreira (2006) and Resnick (2007) provide complimentary details. The long list of references is useful for readers wishing to seek further knowledge in some of the many areas covered in this monograph. For those interested in applications of extreme value data analysis in data networks, Resnick (2003), and references therein, is a nice alternative.

This book can be recommended to researchers in life sciences for the wealth of information about statistics of extremes and density function estimation. The application to hormesis is interesting for those concerned with this strange positive effect that low doses of toxic substances can have in living organisms.

References

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. New York: Springer.

Drees, H. (2008). Some aspects of extreme value statistics under serial dependence. *Extremes* **11,** 35–53.

Gomes, M. I., Canto e Castro, L., Fraga Alves, M. I., and Pestana, D. (2008). Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. *Extremes* **11,** 3–34.

Gumbel, E. J. (1958). *Statistics of Extremes*. New York: Columbia University Press.

Hüsler, J. and Peng, L. (2008). Review of testing issues in extremes: In honor of Professor Laurens de Haan. *Extremes* **11,** 99–111.

Resnick, S. I. (2003). Modeling data networks. In *Extreme Values in Finance, Telecommunications, and the Environment*, B. Finkenstadt and H. Rootzen (eds), 287–371. Boca Raton, Florida: Chapman and Hall.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.

M. Ivette Gomes
Departamento de Estatística e Investigação Operacional
(DEIO) and CEAUL
Faculdade de Ciências, Universidade de Lisboa
Lisboa, Portugal

CRYER, J. D. and CHAN, K.-S. **Time Series Analysis with Applications in R**, 2nd edition. Springer, New York, 2008. xiii + 491 pp. US$84.95/€74.85. ISBN 9780387759586.

This second edition of this book on time series analysis includes new material on time series regression models, spectral analysis, threshold models, and models of heteroscedasticity; the latter of which are heavily used in econometrics and have traditionally been left out of books on time series. The new chapters on heteroscedasticity and threshold models, in my opinion, are what set this book apart from others.

The book is written for a broad audience of students and is most appropriate for a semester course at the first year master's level or for upper division undergraduates. I would consider it most useful at the first year master's level for students in business, social sciences, and economics, and as an introductory course to time series for advanced juniors and seniors in statistics and the physical sciences. The authors attempt to mix applications with theory; however, they only assume minimal calculus (minimization) and most of the technical details appear in chapter appendices. Also included in an appendix is a review of the statistical concepts needed; such as expectation, variance, covariance, correlation, and conditional expectation. As such the intent of the book is for an applied course in time series methods.

The first chapter of the book introduces various time series by way of examples that are used throughout the book. Chapters 2 and 3 cover basic concepts of time series and trend in time series. Chapters 4 to 10 cover the traditional modeling of stationary, nonstationary, and seasonal time series. Chapters 11 to 15 contain more complicated modeling of time series regression models, heteroscedasticity, spectral methods, and threshold models.

The authors have chosen to use the R software package (R Development Core Team, 2007) for all examples in the book. The R code used to generate each table and figure is given immediately after the table or figure. There is an extensive appendix giving an introduction to the R software package. My experience is that students deeply appreciate having code for examples and so this aspect of the book will be well received by students. The authors have also extended the time series functionality in R in a package named TSA that is available on the R project's website.

Overall, the book is well laid out and well written. The TSA package easily loaded on my Mac and the software and example code ran without any problems. I would not recommend this text if one is looking for a theoretical development of time series. However, I have no reservations recommending it as the text for an applied course, which is the intended use of the book.

Reference

R Development Core Team (2007). R*: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Timothy D. Johnson
Department of Biostatistics
University of Michigan
Ann Arbor, Michigan 48109, U.S.A.

BRIEF REPORTS BY THE EDITOR

DONCASTER, C. P. and DAVEY, A. J. H. **Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences**. Cambridge University Press, Cambridge, United Kingdom, 2007. xiii + 288 pp. $120.00/£65.00. ISBN 9780521865623.

As the title suggests, the target audience for this book is researchers in the life sciences. The authors, an ecologist and an environmental scientist, aim to present a wide range of ANOVA models so that readers can appropriately select from among them in analyzing their own research data. Having taught courses to this target audience many times, and not yet found a text that I am completely happy with, I was eager to see what this new book had to offer.

My assessment is that I like *what* the authors say. I am very dissatisfied, however, with *how* they say it. The subjects that they cover are pertinent and well chosen to meet the needs of many field researchers. They begin with a long introduction to many of the concepts that underlie ANOVA models and the effects that they are designed to test. They then work their way through one-factor designs and into many different constructs of multifactor designs (crossed and nested designs, block designs, and split plots). They spend relatively longer discussing repeated measures than I have seen in any other text. They surround the main body of their text with advice and direction that has great potential to be useful to their intended readers.

Unfortunately, the writing is sufficiently imprecise (or just incorrect) in places that students will be confused or misled with disappointingly high probability. The very first sentence of the book describes analysis of variance as "a powerful statistic." On p. 10, α is described as a "threshold for rejecting the null hypothesis of insignificant explained variation." Other inaccuracies include stating that the F-distribution "assumes that the two mean squares are sampled from normally distributed populations" (p. 15); noting that a transformation of "Y or X, or both [can] rectify problems of heterogeneity of variances" (p. 16); and claiming that in the analysis of a randomized block experiment, "to omit block will result in falsely inflated error degrees of freedom, and consequently an increased likelihood of falsely rejecting a true null hypothesis" (p. 27). Ironically, they contradict themselves two pages later by pointing out that partitioning out block variation *increases* power (p. 29).

In the end, this book cannot be recommended unless it undergoes a thorough revision. If this were to happen, then the book might become a useful resource for graduate-level service courses.

MARTINEZ, W. L. and MARTINEZ, A. R. **Computational Statistics Handbook with MATLAB®**, 2nd edition. Chapman & Hall/CRC Press, Boca Raton, Florida, 2008. xxiii + 767 pp. $89.95/£39.99. ISBN 9781584885665.

The first edition of this book was reviewed by P. K. Dunn in *Biometrics* in 2003 (pp. 462–463). Dunn was generally positive about the book, predicting that it would be "an oft-used reference for the practitioner," but also noting that the authors spent more pages showing how to do basic statistics in MATLAB and fewer covering traditional topics in computational statistics than the title would imply. The authors explain their goals for the new edition in the preface: (1) to update the book to more current versions of MATLAB (R2007a) and its Statistics Toolbox (6.0) and (2) to respond to criticisms by bolstering its coverage of topics in computational statistics. To that end, they have added material, in particular in various aspects of multivariate statistics: multivariate probability calculations, exploratory analysis via graphics and dimension-reduction techniques, and clustering and classification (learning) methods.

My own brief assessment of the book leaves me impressed with the number of subjects covered, regardless of whether they are "computational" statistics or data analysis methods. I fully concur with Dunn that the book can be a valuable reference to practicing statisticians (or statistical researchers) using MATLAB as their computing engines.

STAPLETON, J. H. **Models for Probability and Statistical Inference: Theory and Applications**. John Wiley and Sons, New York, 2008. xiii + 440 pp. $116.95/€84.90. ISBN 9780470073728.

This book is intended to be a text for a two-semester sequence in probability and statistics at the master's level. The first six chapters provide coverage of standard subjects in probability—univariate and multivariate, discrete and continuous, and a variety of specific distributional families—before finishing with moment generating functions and convergence. The style of presentation is *not* theorem proof, although theorems are given and occasionally proved. Rather, Stapleton takes a much more descriptive approach, explaining, justifying, and demonstrating stated results, often using simulations. For example, asymptotic approximations are often compared to their exact distribution counterparts. There are far more graphics than I am accustomed to seeing in a standard theory text.

The last seven chapters on inference are presented in the same style, carrying on through estimation and testing, linear models, and categorical data. A final chapter briefly introduces subjects such as sampling, bootstrapping, and censoring. The prose throughout the book is clear and well aimed at a first-year master's student who is intelligent but not yet statistically sophisticated. Examples are clear and well chosen. Exercises are given at the end of each section, and answers to selected exercises are available at the back of the book.

INDRAYAN, A. **Medical Biostatistics**, 2nd edition. Chapman and Hall/CRC Press, Boca Raton, Florida, 2008. xlvi + 771 pp. $99.95/£52.99. ISBN 9781584888871.

This book is aimed primarily at "students, researchers, and professionals of medicine and health." The first edition was reviewed by F. D. J. Dunstan in *Biometrics* in 2002 (pp. 475–476). Dunstan pointed out "many cases where sloppiness has led to incorrect statements," and concluded by saying, "I don't think that the book will be found very useful by its target audience." Unfortunately, this new edition is no better. For example, in the chapter on regression, assumptions such as independence and homoscedasticity are stated improperly in terms of the residuals (defined clearly enough in the customary way) rather than in terms of model errors. On p. 535, the author suggests that having residuals follow a Gaussian distribution somehow helps ordinary least squares work better, and insists that homoscedasticity is important for both quantitative and qualitative response variables (how does one even measure variance on qualitative responses?). The worst error in this chapter is a bizarre discussion of residual diagnostics in which a plot of residuals is shown depicting an increasing trend—which of course will not happen in a linear regression unless the intercept is omitted—yet this pattern is interpreted as an indication of a need for an x^2 term!

The book does take a much more holistic approach to the subject of biostatistics than is typical in a text on the subject, spending a good bit of effort discussing study types and measurement types that are particular to medical and health applications. In principle, this could be a very worthwhile book for members of its target audience. Unfortunately, I am in full agreement with the reviewer of the previous edition. Considering the fundamental errors that are made, I would not use this book for any class, nor could I recommend it to a medical student, researcher, or professional.

KHATTREE, R. and NAIK, D. N. (eds). **Computational Methods in Biomedical Research**. Chapman and Hall/CRC Press, Boca Raton, Florida, 2008. xvii + 408 pp. $99.95/£54.99. ISBN 9781584885771.

This edited volume covers a broad array of topics of modern relevance in biomedical research. There are 12 chapters, each written by different authors, covering topics in statistics that are distinctly beyond the elementary level. Some of those topics covered include analysis of microarrays, classification and diagnosis methods, various forms of censored data and mixed models, and several kinds of correlated data. A list of articles and authors can be found on the publisher's website.

Most of the articles are overviews of their subjects, although one is mostly a case study, and two are essentially presentations of the authors' own work. The overviews generally seem like very reasonable places to begin research into a given subject: they are written discursively and conclude with numerous pages of references. The main exception is the article on classification rules for repeated measures data, which is almost entirely mathematical. Each article is completely self-contained, except perhaps for the very occasional reference to other chapters in the book, so that a reader may choose to read chapters in any number or order.

The book's primary strength is its breadth of topics. A beginning researcher or a reasonably experienced analyst

looking to learn about one of the covered topics could find value in reading the corresponding chapter of this book. This breadth is also possibly its greatest weakness, as the topics are generally not related to one another and overlap very little. A reader who is interested in any given chapter will find that most of the rest of the book offers little more toward their interests. The editors acknowledge this in the preface, pointing out that a more comprehensive treatment of the subject would require several volumes. The end result, though, is that this book should be in every library that supports biostatistical and biomedical research, but it may not necessarily be an essential addition to any one researcher's bookshelf.