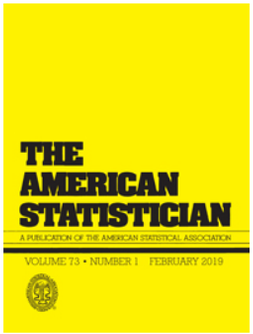
Taylor & Francis
Taylor & Francis Group



The American Statistician

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

REVIEWS OF BOOKS AND TEACHING MATERIALS

To cite this article: (2019) REVIEWS OF BOOKS AND TEACHING MATERIALS, The American Statistician, 73:1, 94-104, DOI: 10.1080/00031305.2018.1538846

To link to this article: https://doi.org/10.1080/00031305.2018.1538846

in this book, and yet, the exposition was clear and written in a nontechnical way to avoid overwhelming novices. In fact, delving into these historical statistical landmarks was reminiscent of a weekend visit to the local museum, which is perhaps the ideal tone for which this book should strive.

A Panorama of Statistics is an ambitious book meant to introduce a wide audience to an extensive array of professional perspectives and puzzling peculiarities about statistics, but, in my opinion, it does not quite achieve the necessary balance of nontechnical exposition and sufficient depth in order to appeal to both statistical novices and experts. Perhaps this book might have been considerably improved through the removal of Sowey and Petocz's requirement of concluding each chapter with exactly five questions, and by moving the detailed and engaging historical anecdotes from the solutions section to the body of the text. Additionally, perhaps Sowey and Petocz may also have been too ambitious with the amount of material covered; reducing the number of topics covered and devoting more space to each topic may help in obtaining the desired balance between including sufficient depth and avoiding technical exposition necessary to reach a broad audience. There is substantial potential in this book, and, as a fan of puzzles and paradoxes, I believe that there is an eager and enthusiastic audience for the type of book that Sowey and Petocz set out to write. And these shortcomings being stated, this book certainly has many elements of interest throughout, but sometimes you have to dig for them in places you might not naturally expect, such as the references or solutions.

Michael J. Higgins
Department of Statistics, Kansas State University, Manhattan, KS

Statistical Modeling and Machine Learning for Molecular Biology. Alan M. Moses. Boca Raton, FL: Chapman & Hall/CRC Press, 2016, xvi+264 pp., $72.95(P), ISBN: 978-1-48-225859-2.

Since joining UCLA in 2013, I have been the instructor of "Statistical Methods in Computational Biology," a core course in the inter-departmental Bioinformatics Ph.D. program. Due to the diverse background of enrolled students, it has been a great challenge for me to find an appropriate textbook. In the end, I designed my own set of lecture notes, which in my opinion still have much room for polishing up. Therefore, when I was invited to review *Statistical Modeling and Machine Learning for Molecular Biology* by Alan Moses, my first impression was that it filled a much-needed niche for bioinformatics and computational biology graduate education. Having now finished reading the book, this impression has been confirmed. This book is unique in that it attempts to explain comprehensive statistical modeling and machine learning concepts to molecular biologists. This is a tremendously challenging mission: statisticians and computer scientists are good at explaining models and concepts using mathematical formulas, but explaining them in words? Surely, this will require great skill and effort on the part of the author to make the explanations accurate and concise. Dr. Moses has made this brave attempt, and I think he has done an exceptional job, probably thanks to his interdisciplinary training and background. I would strongly recommend this book to computational biologists who would like to learn statistical modeling and machine learning methods, and I plan to use this book for my teaching in Spring 2019. This book is also a useful reference for graduate students in statistics and machine learning who would like to develop powerful and robust bioinformatics methods, because the author has discussed multiple genomic data analyses and the corresponding popular bioinformatics tools throughout this book.

This book has four sections. The first section provides an overview of the foundation of statistical modeling. The next three sections cover three main topics in statistical machine learning: unsupervised learning (clustering) and two types of supervised learning (regression and classification). Each section is a comprehensive collection of concepts and methods that have been widely used in computational biology. I greatly appreciate that the author specifically designed the first chapter as a guideline for readers. In this chapter, he summarized the book contents, talked about the prerequisites, listed the uncovered topics, and explained his motivation for writing this book. This chapter is so informative that I would strongly recommend interested readers to read it first and decide whether this book fits their needs. In fact, I think it would be a good idea for the publisher to make the first chapter open-access, if possible, to increase the visibility and impact of this book. Regarding the prerequisites, this book does not assume readers to have any prior knowledge of statistics, but it does require a strong background in molecular biology and familiarity with multivariate calculus and some linear algebra. This book focuses on introducing concepts and methods and does not provide any computer code. Hence, readers are expected to implement the methods using R, Python, or MATLAB by following other textbooks or online tutorials. I like the author's candid sharing of his own learning experience and opinions on other classic statistics and bioinformatics books. These reflections are extremely valuable for graduate students just entering the computational biology field.

After Chapter 1, the book moves into technical materials, starting from the statistical foundation in Chapters 2–4. Unlike in many other statistics textbooks that cover similar materials, the writing is fun and witty, infused with the author's intuition and understanding. Connecting classic statistical concepts (e.g., multiple testing) to modern biological applications (e.g., eQTL analysis) is also unique. Regarding the difficulty level, although no statistics background is assumed, my teaching experience tells me that readers should have an introductory level of probability and mathematical statistics to fully understand the materials presented in this book. One key challenge for beginners to statistics is the understanding of uncertainty. There are good online courses such as the seven-week long *Introduction to Probability* taught by Dr. Joseph Blitzstein (Harvard University) available at www.edx.org. Readers who are interested in the mathematical details behind the statistical foundation in this book may refer to classic undergraduate textbooks such as Ross' (2018) *A First Course in Probability* and Rice's (2009) *Mathematical Statistics and Data Analysis* for alternative explanations of the same concepts.

The remaining eight chapters are evenly divided into the three statistical machine learning topics: clustering (Section II), regression (Section III), and classification (Section IV). Each topic is comprehensive, and the author has managed to give a broad overview of the methods and algorithms that have been most widely used in bioinformatics. I appreciate the comprehensiveness, but I also worry that the technical difficulty level of some topics is probably too high for beginners. For example, for model-based clustering (Chapter 6), the author introduced the expectation–maximization algorithm for estimating multivariate Gaussian mixture model parameters, which is an advanced topic even for a reader who has taken undergraduate-level statistics classes. Other advanced topics include generalized linear models (Chapter 7.8) and regularization in multiple regression (Chapter 9), each of which requires multiple lectures for a thorough explanation based on my teaching experience, but they only have two and fourteen pages in this book, respectively. Moreover, classification is a comprehensive and complex topic that includes diverse statistical and machine learning methods, some of which are known to be difficult for beginners to understand. For these advanced topics, adding more graphical illustration may help beginning readers understand the main ideas. It would also be helpful to add references to some other textbooks or online lecture notes to fill in the technical gaps for readers. Two excellent reference books are James et al.'s (2013) *An Introduction to Statistical Learning with Applications in R* ("the ISL book") and Murphy's (2012) *Machine Learning: A Probabilistic Perspective*, both of which offer alternative explanations to many of the advanced topics covered in this book, and they provide computer codes in R and MATLAB, respectively. I like the author's idea of including bioinformatics examples at the end of several chapters to demonstrate the application of the methods introduced in each chapter. However, from a biologist's perspective, it may make more sense to see the biological relevance first, at the beginning of each chapter, before they dive into the statistical methods.

As a statistician, I noticed that the notation does not distinguish between coefficient parameters and their estimators. For example, b_1 denotes the (unobserved) true slope in p. 145, and it was also used to denote the maximum likelihood estimate of the slope in p. 148. This leads to confusion: for example, during Chapter 7.4 on hypothesis testing, the null hypothesis is $H_0 : b_1 = 0$. Distinguishing parameters and estimators is critical for understanding statistical inference, and this ambiguity in the notation may present a barrier for readers to understand important concepts.

The author chose not to cover two important statistical topics for bioinformatics research: exploratory data analysis (EDA) and dimension reduction. The concept of EDA was promoted by Tukey (1962) as an essential step for data analysis before practitioners formulate any statistical assumptions or apply any statistical methods. The four-week online course "Statistics for Genomic Data Science" taught by Dr. Jeff Leek (Johns Hopkins University) at www.coursera.org includes an excellent introduction to EDA techniques used for genomic research, including data normalization and batch effect removal. In addition to the clustering techniques covered in Section II, dimension reduction is another critical unsupervised learning topic. Although clustering can also be considered as a dimension reduction technique, as noted in Chapter 5.10, the most popular dimension reduction/visualization methods are not covered in this book, including principal component analysis (PCA), multidimensional scaling (MDS), nonnegative matrix factorization (NMF), and t-distributed stochastic neighbor embedding (t-SNE). Beginner readers may refer to Chapter 10 in the ISL book or Chapter 8 in Irizarry and Love's (2017) *Data Analysis for the Life Sciences with R* book for an introduction to PCA and MDS. For NMF and t-SNE, advanced readers may refer to the original research papers by Lee (2001) and Maaten and Hinton (2008), respectively.

To summarize, I think this book serves well as a concise and comprehensive introduction to popular statistical modeling and machine learning methods used in bioinformatics research. Its unique feature is the integration of Dr. Moses' insights into those methods and the cutting-edge genomic applications. The most suitable readers, in my opinion, are graduate students who would like to pursue methodological research in bioinformatics and have an introductory-level foundation in probability, mathematical statistics, multivariate calculus, and linear algebra. For a future edition of this book, I hope that the author can consider adding more references to related textbooks, more graphical illustration, and more exercises directly related to biological data analysis. In addition, I think that adding more clarification on statistical inference, for example, confidence intervals in the frequentist paradigm vs. credible intervals in the Bayesian paradigm, as well as some discussion on EDA and dimension reduction, will make this book even more appealing to bioinformatics readers.

Jingyi Jessica Li
Department of Statistics, University of California, Los Angeles, CA

References

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 1–426. [104]

Irizarry, R. A., and Love, M. I. (2017), *Data Analysis for the Life Sciences With R*. Boca Raton, FL: CRC Press, 1–376. [104]

Lee, D. D., and Seung, H. S. (2001), "Learning the Parts of Objects by Nonnegative Matrix Factorization," *Nature*, 401, 788–791. [104]

Maaten, L., and Hinton, G. (2008), "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, 9, 2579–2605. [104]

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press, 1–1104. [104]

Rice, J. A. (2009), *Mathematical Statistics and Data Analysis* (3rd ed.), Belmont, CA, Cengage Learning, 1–684. [103]

Ross, S. (2018), *A First Course in Probability* (10th ed.), Boston, MA, Pearson, 1–528. [103]

Tukey, J. W. (1962), "The Future of Data Analysis," *The Annals of Mathematical Statistics*, 33, 1–67. [104]