BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 3. 258–259
Advance Access publication March 11, 2008
doi:10.1093/bib/bbn012

Book Review

Data Analysis and Graphics Using R: An Example-Based Approach, Second Edition.

John Maindonald and John Braun
Cambridge University Press;
ISBN 978-0-521-86116-8;
502pp.; 2007; $80.00.

A volume in the 'Cambridge Series in Statistical and Probabilistic Mathematics', 'Data Analysis and Graphics Using R' is presented as a gentle tour guide for new R users, aiming to help them navigate through many powerful tools that the open source R system offers. As the authors point out in the Preface, the book is 'aimed at scientists who wish to do statistical analysis on their own'. Every effort has been made to ensure the book is useful for practical data analysis. This book is particularly useful for researchers in the life science realm who have limited exposure to statistical methodology or theory.

The R system, an open source, free software package for data analysis and graphics, has gained substantial popularity among statisticians over the years. Unlike commercial software packages, R is specifically designed such that it is easy for regular users to create contributed packages and share them with other users. The openness generated surprisingly rich resources in all areas of statistics and beyond. The R system has evolved into an all-purpose software platform for scientific computing and graphics. Researchers in genomics and bioinformatics fields routinely face challenges in analyzing large datasets generated from high-throughput assays such as DNA microarray, genotyping and sequencing technologies. The Bioconductor package, built on top of R, offers a great variety of tools for analyzing -omics data. However, current books on Bioconductor require knowledge of the R system. This book can serve as an introductory guide or reference for these users, especially researchers who are only interested in using R as a data analysis tool.

There are 14 chapters in this book. The first chapter provides a brief introduction to the R system. Topics include defining a variable, reading in data from an external file. This chapter also introduces basic R programming, such as writing functions, loops. The chapter ends with a brief introduction to R graphics tools. Chapter 2 illustrates basic functions in R for data visualization and summary, and how to use them to address basic data analysis questions. Chapter 3 introduces basic concepts in statistical models including probability distributions and random samples. Chapter 4 provides a rather intuitive yet accurate introduction on statistical inference including point estimation, confidence intervals construction and hypothesis testing. There is also a starred section that briefly survey basic theories behind maximum likelihood estimate and Bayesian inference. The next four chapters discuss basic regression models including simple linear regression, multiple linear regression, ANOVA, generalized linear models and basic survival models. Practical issues such as model fitting assessment, model assumption diagnostics and model comparison are discussed in detail. Although the authors started from the most basic models, they also briefly covered advanced topics such as polynomial regression and smoothing splines techniques in nonlinear and nonparametric models. Chapter 9 covers basic time series models. Chapter 10 offers a detailed introduction of random effect and multi-level models, experimental design and repeated measure methodology. This chapter may be of particular interest to biomedical researchers who conduct clinical studies since the models discussed in this chapter are commonly used in clinical data analysis. Chapter 11 provides a nice introduction to tree-based supervised learning techniques. Powerful tools such as regression tree and classification tree are explained in detail. Chapter 12 and 13 cover multivariate data analysis. Important techniques such as principle component analysis, multi-dimensional scaling and discriminant analysis are described. The authors also survey graphical representation

tools that are useful for exploring multi-dimensional data. The final chapter discusses practical issues using the R software and provides more details about its powerful graphical display environment.

Instead of the regular references, the authors provide separate reference lists for methodology, datasets, R packages and websites. And in addition to regular term and author index, the authors also provide an index of R symbols and functions. These resources are a great help for R users who can use the book as a reference guide for finding information they need fast.

The authors should be complimented for their attention-to-details writing style. Each chapter is ended with a Recap section to summarize the material covered; a Further Reading section, which listed reference for special, in-depth and advanced topics; an Exercise section for readers to practice using the R tools discussed. Solutions to selected exercises, together with R scripts, codes, figures and additional notes can be found at a website maintained by the authors.

As the R system is rapidly evolving (a new version every half year), it is important to keep the content of the book up-to-date. It seems that the authors are aware of this issue. They have made updates from the previous edition and are planning for a new edition. At the mean time, I think it would be great for the readers if the authors can provide continued and detailed updates on their website. Another place the book can be improved is the organization of the many datasets used. There are so many different datasets discussed in this book, while a great strength, I found it difficult to keep track of them. It would be nice if an index of all dataset as well as some background information of them can be provided. Such that the users can quickly identify an example dataset of their interest or a dataset that resembles the one on their hand.

The material presented in this volume made it an ideal textbook for an applied statistics course designed for students who need training in data analysis. Given the comprehensiveness, this book can also serve as a reference for general R users. Although most of the examples in the book are not directly relevant for scientists in genomics and bioinformatics fields, I do think they will benefit from the examples and in-depth coverage of essential data analysis techniques and methods.

Zhaohui Steve Qin
Center for Statistical Genetics
Department of Biostatistics
School of Public Health
University of Michigan
Ann Arbor, MI 48109
E-mail: qin@umich.edu