Modern Statistics for Modern Biology

If you are a biologist and want to get the best out of the powerful methods of modern computational statistics, this is your book. You can visualize and analyze your own data, apply unsupervised and supervised learning, integrate datasets, apply hypothesis testing, and make publication-quality figures using the power of R/Bioconductor and ggplot2.

This book will teach you "cooking from scratch", from raw data to beautiful illuminating output, as you learn to write your own scripts in the R language and to use advanced statistics packages from CRAN and Bioconductor. It covers a broad range of basic and advanced topics important in the analysis of high-throughput biological data, including principal component analysis and multidimensional scaling, clustering, multiple testing, unsupervised and supervised learning, resampling, the pitfalls of experimental design, and power simulations using Monte Carlo, and it even reaches networks, trees, spatial statistics, image data, and microbial ecology. Using a minimum of mathematical notation, it builds understanding from well-chosen examples, simulation, visualization, and above all hands-on interaction with data and code.

- **R package msmb** contains complete code and the example datasets, allowing students to recreate all examples, figures, and results in the book
- **Solutions, slides, and dynamic material** available on the course website
- Introduces **methods on a "need to know" basis**, so students tackle biological questions immediately and understand motivation for the methods
- **Real-life examples** done from scratch, guiding students through realistic complexities and building practical intuition
- Includes a wrap-up chapter that explains the complete workflow from design of experiments to analysis of results, identifying **common pitfalls with big data**
- All figures and results generated by the code in the book, demonstrating how **reproducible research** works

SUSAN HOLMES is Professor of Statistics at Stanford University, California. She specializes in exploring and visualizing multidomain biological data, using computational statistics to draw inferences in microbiology, immunology and cancer biology. She has published over 100 research papers, and has been a key developer of software for the multivariate analyses of complex heterogeneous data. She was the Breiman Lecturer at NIPS 2016, has been named a Fields Institute fellow, and is currently a fellow at the Center for the Advances Study of the Behavioral Sciences.

WOLFGANG HUBER is Research Group Leader and Senior Scientist at the European Molecular Biological Laboratory, where he develops computational methods for new biotechnologies and applies them to biological discovery. He has published over 150 research papers in functional genomics, cancer and statistical methods. He is a founding member of the open-source bioinformatics software collaboration Bioconductor and has co-authored two books on Bioconductor.

Modern Statistics for Modern Biology

Susan Holmes
Stanford University, California

Wolfgang Huber
European Molecular Biology Laboratory

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108705295
DOI: 10.1017/9781108551441

First published 2018

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-108-70529-5 Paperback

Additional resources for this publication at www.cambridge.org/msmb

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Image credits for chapter openers: Chapter 1, Wikicommons;
Chapter 4, xkcd.com/1347; Chapter 5, mikedabell/iStock/Getty Images;
Chapter 6, extract from xkcd.com/882/; Chapter 7, The Matrix: scene 291 Close on Computer Screen © Warner Bros.; Chapter 8, xkcd.com/1725;
Chapter 9, Robert Orchard/Moment/Getty Images;
Chapter 13, University of Adelaide Library: Rare Books and Special Collections, R.A. Fisher Digital Archive,
http://hdl.handle.net/2440/81670.

For Sonia, Sara, Agnès, Johnny, Camille
. . . and the “girls” who make me love the life sciences

For Alexander

Contents

Expanded Contents

Cambridge University Press
978-1-108-70529-5 — Modern Statistics for Modern Biology
Susan Holmes , Wolfgang Huber
Frontmatter
More Information

Cambridge University Press
978-1-108-70529-5 — Modern Statistics for Modern Biology
Susan Holmes , Wolfgang Huber
Frontmatter
More Information

Cambridge University Press
978-1-108-70529-5 — Modern Statistics for Modern Biology
Susan Holmes , Wolfgang Huber
Frontmatter
More Information

Introduction

What is happening in biological data analysis?

The two instances of *modern* in the title of this book reflect the two major recent revolutions in biological data analysis:

- Biology, formerly a science with sparse, often only qualitative data, has turned into a field whose production of quantitative data is on par with high energy physics or astronomy and whose data are wildly more heterogeneous and complex.
- Statistics, a field that in the 20th century had become an application ground for probability theory and calculus, often taught loaded with notation and a perceived heavy emphasis on hypothesis testing, has been transformed by the ubiquity of computers and of data in machine-readable form. Exploratory data analysis, visualization, resampling, simulations, pragmatic hybridizations of Bayesian ideas and methods with frequentist data analysis have become parts of the toolset.

The aim of this book is to enable scientists working in biological research to quickly learn many of the important ideas and methods that they need to make the best of their experiments and of other available data. The book takes a hands-on approach. The narrative in each chapter is driven by classes of questions or by certain data types. Methods and theory are introduced on a need-to-know basis. We don't try to systematically deduce from first principles. The book will often throw readers into the water and help them to swim to their destinations despite missing details.

"Watersnood in Groningen, 1686", Jan Luyken, 1698. Rijksmuseum Amsterdam.

By no means will this book replace systematic training in underlying theory: probability, linear algebra, software engineering, databases, multivariate statistics. Such training takes many semesters of coursework. Perhaps the book will whet your appetite to engage more deeply with one of these fields.

The challenge: heterogeneity

Any biological system or organism is composed of tens of thousands of components, which can be in different states and interact in multiple ways. Modern biology aims to understand such systems by acquiring comprehensive – and this means high dimensional – data in their temporal and spatial context, with multiple covariates and interactions. Dealing with this complexity will be our primary challenge. We face real, biological complexity as well as the complexities and heterogeneities of the data we are able to acquire with our always imperfect instruments.

Biological data come in all sorts of shapes: nucleic acid and protein sequences, rectangular tables of counts, multiple tables, continuous variables, batch factors, phenotypic images, spatial coordinates. Besides data measured in lab experiments, there are clinical data, longitudinal information, environmental measurements, networks, lineage trees, annotation from biological databases in free text or controlled vocabularies, . . .

> "Homogeneous data are all alike;
> all heterogeneous data are heterogeneous in their own way."
> The Anna Karenina principle

It is this heterogeneity that motivates our choice of R and Bioconductor as the computational platform for this book – more on this below.

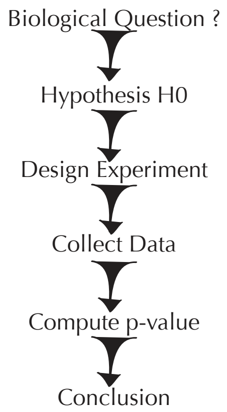


Figure 1: The hypothesis testing paradigm recommended by R.A. Fisher starts with the formulation of a null hypothesis and the design of an experiment before the collection of any data. We could think in a similarly schematic way about model fitting – just replace *Hypothesis H0* by *Parametric Model* and *Compute p-value* by *Fit Parameters*.

What's in this book?

Figure 1 outlines a sequential view of statistical data analysis. Motivated by the groundbreaking work on significance and hypothesis testing in the 1930s by Fisher (1935) and Neyman and Pearson (1936), it is well amenable to mathematical formalism, especially the part where we compute the distribution of test statistics under a hypothesis (null or alternative), or where we set up distributional assumptions and search for analytical approximations.

Real scientific discovery rarely works in the caricature manner of Figure 1. Tukey (1977) emphasized two separate approaches. The first he termed **exploratory data analysis (EDA)**. EDA uses the data themselves to decide how to conduct the statistical analysis. EDA is built on simple tools for plotting data. EDA is complemented by **confirmatory data analysis** (CDA): robust inferential methods that do not rely on complex assumptions to reach scientific conclusions. Tukey recommended an iterative approach, schematized in Figure 2, that enables us to see the data at different resolutions and from different perspectives. This enables the refinement of our understanding of the data.

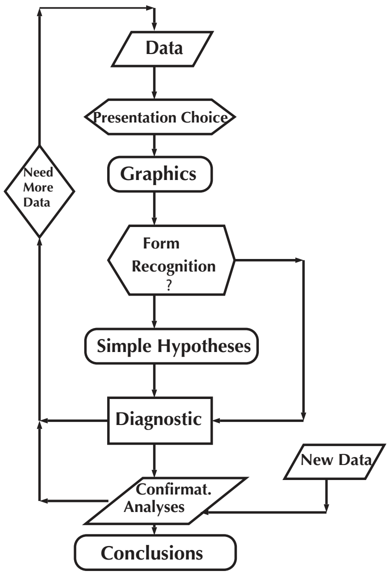


Figure 2: J.W. Tukey recommended starting any analysis with the data and wrote: "No catalogue of techniques can convey a willingness to look for what can be seen, whether or not anticipated" (Holmes Junca, 1985).

Biology in the late 1990s raised the **large-*p* small-*n* problem**: consider a gene expression dataset for $n = 200$ patient samples on $p = 20{,}000$ genes. If we want to construct a regression or classification model that "predicts" a clinical variable, for instance the disease type or outcome, from the 20,000 genes, or features, we immediately run into problems,[1] since the number of model parameters would have to be orders of magnitudes larger than the number of replicate measurements n. At least, this is the case for common models, say, an ordinary linear model. Statisticians realized that they could remedy the situation by requiring sparsity through the use of regularization techniques (Hastie et al., 2008), i.e., by requiring many of the potential parameters to be either zero or at least close to it.

A generalization of the sparsity principle is attained by invoking one of the most

[1] Called *non-identifiability* or *overfitting*.

powerful recent ideas in high-dimensional statistics, which goes by the name **empirical Bayes**: we don't try to learn the parameters associated with each feature from scratch, but rather use the fact that some or all of them will be similar, or even the same, across all features, or across groups of related features. There are several important book-long treatments (Efron, 2010) of the subject of large scale inference so essential in modern estimation and hypotheses testing.

This icon signals that we are using a Monte Carlo approximation method, so-called because it harnesses randomness, similar to the randomness of casino games. Ironically, for many casino games the probability of winning is not known analytically, and casinos use their own empirical data to evaluate the odds.

Simulations play an essential role in this book, as *many of the results we need* escape the reach of standard analytic approaches. In other words, simulations liberate us from only considering methods that are analytically tractable, and from worrying about the appropriateness of simplifying assumptions or approximations.

In this book, we try to cover the full range of these developments and their applications to current biological research. We cover many different types of data that modern biologists have to deal with, including RNA-Seq, flow cytometry, taxa abundances, imaging data and single-cell measurements. We assume no prior training in statistics. However, you'll need some familiarity with R and willingness to engage in mathematical and analytical thinking.

Generative models are our basic building blocks. In order to draw conclusions about complicated data *it tends to be useful* to have simple models for the data generated in this or that situation. We do this through the *top-down* use of probability theory and deduction, which we introduce in Chapter 1. We will use examples from immunology and DNA analysis to describe useful generative models for biological data: binomial, multinomial and Poisson random variables.

Once we know how data would look under a certain model, we can start working our way backwards: given some data, what model is most likely able to explain it? This *bottom-up* approach is the core of **statistical inference**, and we explain it in Chapter 2.

We saw the primary role of **graphics** in Tukey's scheme (Figure 2), and so we'll learn how to visualize our data in Chapter 3. We'll use the grammar of graphics and *ggplot2*.

Real biological data often have more complex distributional properties than what we could cover in Chapter 1. We'll use **mixtures** that we explore in Chapter 4; these enable us to build realistic models for heterogeneous biological data and provide solid foundations for choosing appropriate variance-stabilizing transformations.

The large, matrix-like ($n \times p$) datasets in biology lend themselves to **clustering**: once we define a distance measure between matrix rows (the features), we can cluster and group the genes by similarity of their expression patterns, and similarly, for the columns (the patient samples). We'll cover clustering in Chapter 5. Since clustering relies only on distances, we can even apply it to data that are not matrix-shaped, as long as there are objects and distances defined between them.

Further following the path of EDA, we cover the most fundamental unsupervised analysis method for simple matrices – **principal component analysis** – in Chapter 7. We turn to more heterogeneous data that combine multiple data types in Chapter 9. There, we'll see nonlinear unsupervised methods for counts from single-cell data. We'll also address how to use generalizations of the multivariate approaches covered in Chapter 7 to combinations of categorical variables and multiple assays recorded on the same observational units.

The basic **hypothesis testing** workflow outlined in Figure 1 is explained in Chapter 6. We take the opportunity to apply it to one of the most common queries to $n \times p$ datasets: which of the genes (features) are *associated with* a certain property of the samples, say, disease type or outcome? However, conventional significance thresholds would lead to lots of spurious associations: with a false positive rate of $\alpha = 0.05$ we expect $p\alpha = 1000$ false positives if none of the $p = 20{,}000$ features has a true association. Therefore we also need to deal with multiple testing.

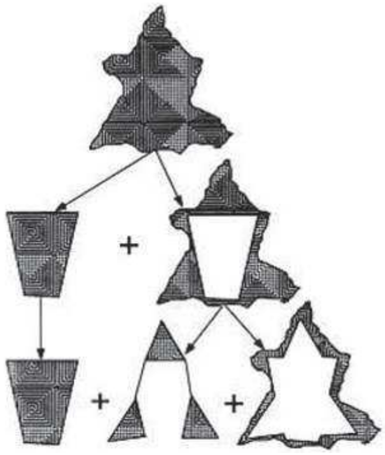


Figure 3: Analyzing data is not a one-step process. Each step involves visualizing and decomposing some of the complexity in the data. Tukey's iterative data structuration can be conceptualized as Total = $V_1 + V_2 + V_3$.

One of the most fruitful ideas in statistics is that of **variance decomposition**, or analysis of variance (ANOVA). We'll explore this, in the framework of linear models and generalized linear models, in Chapter 8. Since we'll draw our example data from an RNA-Seq experiment, this gives us also an opportunity to discuss models for such count data and concepts of *robustness*.

Nothing in biology makes sense except in the light of evolution,[2] and evolutionary relationships are usefully encoded in phylogenetic trees. We'll explore **networks and trees** in Chapter 10.

A rich source of data in biology are **images**, and in Chapter 11 we reinforce our willingness to do EDA on all sorts of heterogeneous data types by exploring feature extraction from images and spatial statistics.

Finally in Chapter 12, we will look at **statistical learning**, i.e., training an algorithm to distinguish between different types of objects depending on their multidimensional feature vector. We'll start simple with low-dimensional feature vectors and linear methods, and then explore classification in high-dimensional settings.

We wrap up in Chapter 13 with considerations on **good practices** in the design of experiments and in data analysis. For this we'll use and reflect on what we have learned in the course of the preceding chapters.

Computational tools for modern biologists

As we'll see over and over again, the analysis approaches, tools and choices to be made are manifold. Our work can only be validated by keeping careful records in a reproducible script format. **R and Bioconductor** provide such a platform.

Although we are tackling many different types of data, questions and statistical methods hands-on, we maintain a consistent computational approach by keeping all the computation under one roof: the R programming language and statistical environment, enhanced by the biological data infrastructure and specialized method packages from the Bioconductor project. The reader will have to start by acquiring some familiarity with R before using the book.

[2] Theodosius Dobzhansky – see *Nothing in Biology Makes Sense Except in the Light of Evolution* on Wikipedia.

R code is a major component of this book. It is how we make the textual explanations explicit. Virtually every data visualization in the book is produced with code that is shown to equip the reader to replicate all of these figures, and any other results shown (as in Figure 4).

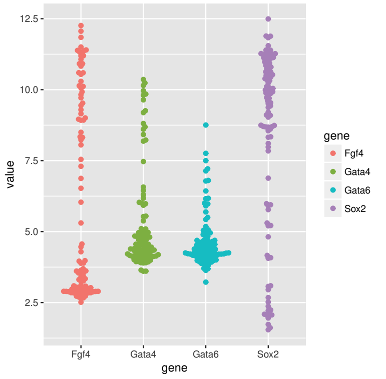


Figure 4: Comparison of the expression levels of four developmentally important genes in the mouse embryo. Each dot represents the measurement from one single cell; the y-axis is on a logarithmic scale (arbitrary units). The code that produces this plot is given in Chapter 3.

Even if you have a basic familiarity with R, don't worry if you don't immediately understand every line of code in the book. Although we have tried to keep the code explicit and give tips and hints at potentially challenging places, there will be instances where

- there is a function invoked that you have not seen before and that does something mysterious, or
- there is a complicated R expression that you don't understand (perhaps involving `apply` functions or data manipulations from the *dplyr* package).

Don't panic. For the mysterious function, have a look at its manual page. Open up RStudio and use the object explorer to look at the variables that go into the expression, and those that come out. Split up the expression to look at intermediate values.

In Chapters 1 and 2, we use *base* R functionality for light doses of plotting and data manipulation. As we successively need more sophisticated operations, we introduce the *ggplot2* way of making graphics in Chapter 3. Besides the powerful grammar of graphics concepts that enable us to produce sophisticated plots using only a limited set of instructions, this implies using the *dplyr* way of data manipulation. Sometimes, we have traded in what would be convoluted loop and `lapply` constructs for elegant *dplyr* expressions, but this requires you to get acquainted with some novelties such as *tibbles*, the `group_by` function and pipes (`%>%`).

Why R and Bioconductor?

There are many reasons why we have chosen to present all analyses on the R (Ihaka and Gentleman, 1996) and Bioconductor (Huber et al., 2015) platforms.

Download R and Rstudio to follow the code in the book.

Cutting edge solutions The availability of over 10,000 packages ensures that almost all statistical methods are available, including the most recent developments. Moreover, there are implementations of or interfaces to many methods from computer science, mathematics, machine learning, data management, visualization and internet technologies. This puts thousands of person-years of work by experts at your fingertips.

Open source and community-owned R and Bioconductor have been built collaboratively by a large community of developers. They are constantly tried and tested by thousands of users.

Data input and wrangling Bioconductor packages support the reading of many of the data types and formats produced by measurement instruments used in modern biology, as well as the needed technology-specific “preprocessing” routines. The community is actively keeping these up-to-date with the rapid developments in the instrument market.

Cambridge University Press
978-1-108-70529-5 — Modern Statistics for Modern Biology
Susan Holmes , Wolfgang Huber
Frontmatter
More Information

Simulation There are random number generators for every known statistical distribution and powerful numeric routines for linear algebra, optimization, etc.

Visualization and presentation R can make attractive, publication-quality graphics. We've dedicated Chapter 3 to this, and practice data visualization extensively throughout the book.

Easy-to-use interactive development environment RStudio is easy and fun to use and helps with all aspects of programming in R. It is an essential tool in following the iterative approach to data analysis schematized in Figure 2.

Reproducibility As an equivalent to the laboratory notebook that is standard good practice in labwork, we advocate the use of a computational diary written in the R markdown format. We use the *knitr* package to convert R markdown into easy-to-read and shareable HTML or PDF documents. These can even become full-fledged scientific articles or supplements. Together with a version control system, R markdown helps with tracking changes.

Collaborative environment R markdown enables the creation of websites containing code, text, figures and tables with a minimum of work.

Rich data structures The Bioconductor project has defined specialized data containers to represent complex biological datasets. These help to keep your data consistent, safe and easy to use.

Interoperability and distributed development Bioconductor in particular contains packages from diverse authors that cover a wide range of functionalities but still interoperate because of the common data containers.

Documentation Many R packages come with excellent documentation in their function manual pages and vignettes. The vignettes are usually the best starting point in a package, as they give you a high-level narrative account of what the package does, whereas the manual pages give detailed information on input, output and inner workings of each function. There are online tutorials, forums and mailing lists for many aspects of working with R and Bioconductor.

High-level language R is an interpreted high-level language. Its roots in LISP and its functional programming features mean that code is data and can be computed on, which enables efficient programming and is fun. These features facilitate constructing powerful domain-specific languages.[3] R is not a fixed language – throughout its history, it has been actively evolving and is constantly improving.

How to read this book

The printed version of this book is supplemented by an online version in HTML at `http://bios221.stanford.edu/book/` and `http://www.huber.embl.de/msmb/`.

The online sites:

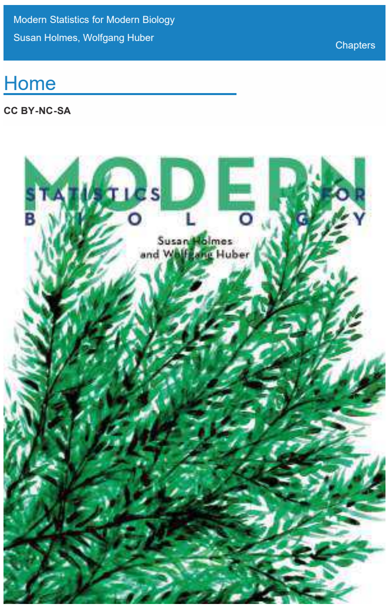
- provide the `.R` files and all needed input data files;
- are constantly updated to fix typos and make clarifications;

[3] Examples include R's formula interface, the grammar of graphics in *ggplot2*, the data manipulation functionality of *dplyr* and R markdown.

- have up-to-date code that will run with contemporary versions of R, CRAN packages and Bioconductor.

Please do not despair if code in the printed version of the book is not working with your version of R and all the packages. Please do not despair if code on the website is not working with an older version of R or packages. This is fully to be expected and no reason for worries, surprises or even comments. We recommend following the installation instructions – which includes getting the right, matching versions of everything – on the webpage.

The chapters in the book build upon each other, but they are reasonably self-contained, so they can also be studied selectively. Each chapter starts with a section on motivations and goals. Questions in the text help you check whether you are following along. The text contains extensive R code examples throughout. You don't need to scrape R code from the HTML or manually copy it from the book. Use the R files (extension `.R`) on the book's website. Each chapter concludes with a summary of the main points and a set of exercises. The book ends with an index and a concordance section, which should be useful when looking for specific topics.



Modern Statistics for Modern Biology
Susan Holmes, Wolfgang Huber
Chapters
Home
CC BY-NC-SA

Figure 5: The online version provides the text in HTML, data files and up-to-date code.

Notes and extra information appear under the devil icon: this is the devil who looks after the details.

Cambridge University Press
978-1-108-70529-5 — Modern Statistics for Modern Biology
Susan Holmes , Wolfgang Huber
Frontmatter
More Information