OXFORD

THE NEW STATISTICS WITH R

AN INTRODUCTION FOR BIOLOGISTS

ANDY HECTOR

The New Statistics with R

An Introduction for Biologists

ANDY HECTOR

Professor of Ecology
Department of Plant Sciences
University of Oxford

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2014949047

ISBN 978–0–19–872905–1 (hbk.)
ISBN 978–0–19–872906–8 (pbk.)

Printed and bound in Great Britain by
Bell & Bain Ltd., Glasgow

The New Statistics with R

I dedicate this book to the memory of Christine Müller.

Acknowledgements

First, I would like to thank Drew Purves, Steve Emmett, and their groups at Microsoft Research in Cambridge, as I made a substantial start to this book while on sabbatical as a visiting researcher in the computational ecology group there at the end of 2011.

Several people were instrumental in helping cultivate my interest in statistical analysis. I was first introduced to experiments during my final-year project with Phil Grime and colleagues at the Unit of Comparative Plant Ecology at Sheffield University. Shortly afterwards, one of the most rewarding parts of my PhD at Imperial College was learning statistics (and GLIM) from Mick Crawley. Bernhard Schmid shared this interest and enthusiasm and taught me a lot while I was a post-doc on the BIODEPTH project and later when we worked together at the Institute for Environmental Sciences at the University of Zurich (sorry for forsaking GenStat for R Bernhard!).

I benefitted from discussions with several statisticians during training courses or after visiting lectures including Douglas Bates, Martin Mächler, John Nelder, José Pinheiro, Bill Venables, and Hadley Wickham.

Many PhD students and post-docs helped me delve further into statistics with R, including some of the material covered in this book (and commented on draft chapters). I would like to thank all current and past group members, but particularly Robi Bagchi, Juliette Chamagne, Stefanie von Felten, Yann Hautier, Mikey O'Brien, Chris Philipson, Matteo Tanadini, and Sean Tuck.

The content of this book is based on teaching materials developed at the University of Zurich and the University of Oxford where I currently teach much of the statistical content for the Quantitative Methods for Biology course—my thanks to the course participants at both institutions and to the Oxford QM tutors particularly Yvonne Griffiths for her fine-tooth comb!

I learned a lot from collaborating on papers on statistical analysis with several colleagues including Tom Bell, Jarrett Byrnes, John Connelly, Forest Isbell, Marc Kéry, Michel Loreau, Owen Petchey, and Alain Zuur. Thanks to Ben Bolker and Vincent Calcagno for discussions on GLMMs and multimodel inference. I would also like to thank Maja Weilenmann and especially Lindsay Turnbull. Thanks to Lucy and Ian at OUP. This project would not have been possible without the generous work of the many people who have helped develop R. Finally, thank you—and sorry—to anyone who has slipped my mind as I rush to meet the book deadline!

Contents

1

Introduction

Unlikely as it may seem, statistics is currently a sexy subject. Nate Silver's success in out-predicting the political pundits in the last US election drew high-profile press coverage across the globe. Statistics may not remain sexy but it will always be useful. It is a key component in the scientific toolbox and one of the main ways we have of describing the natural world and of finding out how it works. In most areas of science, statistics is essential. In some ways this is an odd state of affairs. Mathematical statisticians generally don't require skills from other areas of science in the same way that we scientists need skills from their domain. We have to learn some statistics in addition to our core area of scientific interest. Obviously there are limits to how far most of us can go. This book is intended to introduce some of the most useful applied statistical analyses to researchers, particularly in the life and environmental sciences.

1.1 The aim of this book

My aim is to get across the essence of the statistical ideas necessary to intelligently apply linear models (and some of their extensions) within relevant areas of the life and environmental sciences. I hope it will be of use to students at both undergraduate and post-graduate level and researchers interested in learning more about statistics (or in switching to the software package used here). The approach is therefore not mathematical. I have minimized the number of equations—they are in numerous statistics textbooks and on the internet if you want them—and the

The New Statistics with R. Andy Hector. © Andy Hector 2015.
Published 2015 by Oxford University Press.

statistical concepts and theory are explained in boxes to try and avoid disrupting the flow of the main text. I have also kept citations to a minimum and concentrated them in the text boxes and final chapter (there is no Bibliography). Instead, the approach is to learn by doing through the analysis of real data sets. That means using a statistical software package, in this case the R programming language for statistics and graphics (for the reasons given below). It also requires data. In fact, most of us only start to take an interest in statistics once we have (or know we soon will have) data. In most science degrees that comes late in the day, making the teaching of introductory statistics more challenging. Students studying for research degrees (Masters and PhDs) are generally much more motivated to learn statistics. The next best thing to working with our own data is to work with some carefully selected examples from the literature. I have used some data from my own research but I have mainly tried to find small, relevant data sets that have been analysed in an interesting way. Most of them are from the life and environmental sciences (including ecology and evolution). I am very grateful to all of the people who have helped collect these data and to develop the analyses. For convenience I have tried to use data sets that are already available within the R software (the data sets are listed at the end of the book and described in the relevant chapter).

1.2 The R programming language for statistics and graphics

R is now the principal software for statistics, graphics, and programming in many areas of science, both within academia and outside (many large companies use R). There are several reasons for this, including:

- R is a product of the statistical community: it is written by the experts.
- R is free: it costs nothing to download and use, facilitating collaboration.
- R is multiplatform: versions exist for Windows, Mac, and Unix.

- R is open-source software that can be easily extended by the R community.
- R is statistical software, a graphics package, and a programming language all in one.

1.3 Scope

Statistics can sometimes seem like a huge, bewildering, and intimidating collection of tests. To avoid this I have chosen to focus on the linear model framework as the single most useful part of statistics (at least for researchers in the environmental and life sciences). The book starts by introducing several different variations of the basic linear model analysis (analysis of variance, linear regression, analysis of covariance, etc). I then introduce two extensions: generalized linear models (GLMs) (for data with non-normal distributions) and mixed-effects models (for data with multiple levels and hierarchical structure). The book ends by combining these two extensions into generalized linear mixed-effects models. The advantage of following the linear model approach (and these extensions) is that a wide range of different types of data and experimental designs can be analysed with very similar approaches. In particular, all of the analyses covered in this book can be performed in R using only three main classes of function; one for linear models (the lm() function), one for GLMs (the glm() function), and one for mixed-effects models (the lmer() and glmer() functions).

1.4 What is not covered

Statistics is a huge subject, so lack of space obviously precluded the inclusion of many topics in this book. I also deliberately left some things out. Many biological applications like bioinformatics are not covered. For reasons of space, the coverage is limited to linear models and GLMs, with nothing on non-linear regression approaches nor additive models (generalized additive models, GAMs). Because of the focus on an estimation-based approach I have not included non-parametric statistics. Experimental design is covered briefly and integrated into the relevant chapters. Information theory and information criteria are briefly introduced, but the relatively new and developing area of multimodel inference turned out to be largely beyond the scope of this book. Introducing Bayesian statistics is also a book-length project in its own right.

1.5 The approach

There are several different general approaches within statistics (frequentist, Bayesian, information theory, etc) and there are many subspecies within these schools of thought. Most of the methods included in this book are usually described as belonging to 'classical frequentist statistics'. However, this approach, and the probability values that are so widely used within it, has come under increasing criticism. In particular, statisticians usually accuse scientists of focusing far too much on *P*-values and not enough on effect sizes. This is strange, as the effect sizes—the estimates and intervals—are directly related to what we measure during our research. I don't know any scientist who studies *P*-values! For that reason I have tried to take an estimation-based approach that focuses on estimates and confidence intervals wherever possible. Styles of analysis vary (and fashions change over time). Because of this I will be frank about some of my personal preferences used in this book. In addition to making wide use of estimates and intervals I have also tried to emphasize the use of graphs for exploring data and presenting results. I have tried to encourage the use of *a priori* contrasts (comparisons that were planned in advance) and I avoid the use of corrections for multiple comparisons (and discourage their use in many cases). The most complex approaches in the book are the mixed-effects models. Here I have stuck closely to the approaches advocated by the software writers (and their own books). Finally, at the end of each chapter I try to summarize both the statistical approach and what it has enabled us to learn about the science of each example. It is easy to get lost in statistics, but for non-statisticians the analysis should not become an end in its own right, only a method to help advance our science.

1.6 The new statistics?

What is the 'new' statistics of the title? The term is not clearly defined but it appears to be used to cover both brand new techniques (e.g. meta-analysis, an approach beyond the scope of this book—I recommend the 2013 book by Julia Koricheva and colleagues, *Handbook of meta-analysis in ecology and evolution*) and a fresh approach to long-established methods. I use the term to refer to two things. First, the book covers some relatively new methods in statistics, including modern mixed-effects models (and their generalized linear mixed-effects model extensions) and the use of information criteria and multimodel inference. The new statistics also includes a back to basics estimation-based approach that takes account of the recent criticisms of *P*-values and puts greater emphasis on estimates and intervals for statistical inference.

1.7 Getting started

To allow a learning-by-doing approach the R code necessary to perform the basic analysis is embedded in the text along with the key output from R (the full R scripts will be available as support material from the R café at <http://www.plantecol.org/>). Some readers may be completely new to R, but many will have some familiarity with it. Rather than start with an introduction to R we will dive straight into the first example of a linear model analysis. However, a brief introduction to R is provided at the end of the book and newcomers to the software will need to start there.

2

Comparing Groups: Analysis of Variance

2.1 Introduction

Inbreeding depression is an important issue in the conservation of species that have lost genetic diversity due to a decline in their populations as a result of over-exploitation, habitat fragmentation, or other causes. We begin with some data on this topic collected by Charles Darwin. In *The effects of cross and self-fertilisation in the vegetable kingdom*, published in 1876, Darwin describes how he produced seeds of maize (*Zea mays*) that were fertilized with pollen from the same individual or from a different plant. Pairs of seeds taken from self-fertilized and cross-pollinated plants were then germinated in pots and the height of the young seedlings measured as a surrogate for their evolutionary fitness. Darwin wanted to know whether inbreeding reduced the fitness of the selfed plants. Darwin asked his cousin Francis Galton—a polymath and early statistician famous for 'regression to the mean' (not to mention the silent dog whistle!)—for advice on the analysis. At that time, Galton could only lament that, 'The determination of the variability . . . is a problem of more delicacy than that of determining the means, and I doubt, after making many trials whether it is possible to derive useful conclusions from these few observations. We ought to have measurements of at least fifty plants in each case'. Luckily we can now address this question using any one of several closely related

Published 2015 by Oxford University Press.