Biostatistics with R

An Introductory Guide for Field Biologists

Biostatistics with R provides a straightforward introduction on how to analyse data from the wide field of biological research, including nature protection and global change monitoring. The book is centred around traditional statistical approaches, focusing on those prevailing in research publications. The authors cover *t* tests, ANOVA and regression models, but also the advanced methods of generalised linear models and classification and regression trees. Chapters usually start with several useful case examples, describing the structure of typical datasets and proposing research-related questions. All chapters are supplemented by example datasets and thoroughly explained, step-by-step R code demonstrating the analytical procedures and interpretation of results. The authors also provide examples of how to appropriately describe statistical procedures and results of analyses in research papers. This accessible textbook will serve a broad audience of interested readers, from students, researchers or professionals looking to improve their everyday statistical practice, to lecturers of introductory undergraduate courses. Additional resources are provided on www.cambridge.org/biostatistics.

Jan Lepš is Professor of Ecology in the Department of Botany, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and senior researcher in the Biology Centre of the Czech Academy of Sciences in České Budějovice. His main research interests include plant functional ecology, particularly the mechanisms of species coexistence and stability, and ecological data analysis. He has taught many ecological and statistical courses and supervised more than 80 student theses, from undergraduate to PhD.

Petr Šmilauer is Associate Professor of Ecology in the Department of Ecosystem Biology, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic. His main research interests are multivariate statistical analysis, modern regression methods and the role of arbuscular mycorrhizal symbiosis in the functioning of plant communities. He is co-author of multivariate analysis software Canoco 5, CANOCO for Windows 4.5 and TWINSPAN for Windows.

'We will never have a textbook of statistics for biologists that satisfies everybody. However, this book may come closest. It is based on many years of field research and the teaching of statistical methods by both authors. All useful classic and advanced statistical concepts and methods are explained and illustrated with data examples and R programming procedures. Besides traditional topics that are covered in the premier textbooks of biometry/biostatistics (e.g. R. R. Sokal & F. J. Rohlf, J. H. Zar), two extensive chapters on multivariate methods in classification and ordination add to the strength of this book. The text was originally published in Czech in 2016. The English edition has been substantially updated and two new chapters 'Survival Analysis' and 'Classification and Regression Trees' have been added. The book will be essential reading for undergraduate and graduate students, professional researchers, and informed managers of natural resources.'

Marcel Rejmánek,
Department of Evolution and Ecology, University of California, Davis, CA, USA

Biostatistics with R

An Introductory Guide for Field Biologists

JAN LEPŠ
University of South Bohemia, Czech Republic

PETR ŠMILAUER
University of South Bohemia, Czech Republic

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, 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

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108727341

DOI: 10.1017/9781108616041

First published 2020

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

ISBN 978-1-108-48038-3 Hardback
ISBN 978-1-108-72734-1 Paperback

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

Cambridge University Press & Assessment
978-1-108-72734-1 — Biostatistics with R: An Introductory Guide for Field Biologists
Jan Lepš , Petr Šmilauer
Frontmatter
More Information

Contents

Cambridge University Press & Assessment
978-1-108-72734-1 — Biostatistics with R: An Introductory Guide for Field Biologists
Jan Lepš , Petr Šmilauer
Frontmatter
More Information

Preface

Modern biology is a quantitative science. A biologist weighs, measures and counts, whether she works with aphid or fish individuals, with plant communities or with nuclear DNA. Every number obtained in this way, however, is affected by random variation. Aphid counts repeatedly obtained from the same plant individual will differ. The counts of aphids obtained from different plants will differ more, even if those plants belong to the same species, and samples coming from plants of different species are likely to differ even more. Similar differences will be found in the nuclear DNA content of plants from the same population, in nitrogen content of soil samples taken from the same or different sites, or in the population densities of copepods across repeated samplings from the same lake. We say that our data contain a random component: the values we obtain are random quantities, with a part of their variation resulting from randomness.

But what actually is this randomness? In posing such a question, we move into the realm of philosophy or to axioms of probability theory. But what is probability? A biologist is usually happy with a pragmatic concept: we consider an event to be random if we do not have a causal explanation for it. Statistics is a research field which provides recipes for how to work with data containing random components, and how to distinguish deterministic patterns from random variation. Popular wisdom says that statistics is a branch of science where precise work is carried out with imprecise numbers. But the term **statistics** has multiple meanings. The layman sees it as an assorted collection of values (football league statistics of goals and points, statistics of MP voting, statistics of cars passing along a highway, etc.). Statistics is also a research field (often called mathematical statistics) providing tools for obtaining useful information from such datasets. It is a separate branch of science, to a certain extent representing an application of probability theory. The term statistic (often in singular form) is also used in another sense: a numerical characteristic computed from data. For example, the well-known arithmetic average is a statistic characterising a given data sample.

In scientific thinking, we can distinguish deductive and inductive approaches. The **deductive approach** leads us from known facts to their consequences. Sherlock Holmes may use the facts that a room is locked, has no windows and is empty to deduce that the room must have been locked from the outside. Mathematics is a typical example of a deductive system: based on axioms, we can use a purely logical (deductive) path to derive further statements, which are always correct if the initial axioms are also correct (unless we made a mistake in the derivation). Using the deductive approach, we proceed in a purely logical manner and do not need any comparison with the situation in real terms.

The **inductive approach** is different: we try to find general rules based on many observations. If we tread upon 1-cm-thick ice one hundred times and the ice breaks each time, we can conclude that ice of this thickness is unable to carry the weight of a grown person. We conclude this using inductive thinking. We could, however, also employ the deductive approach by using known physical laws, strength measurements of ice and the known weight of a grown person. But usually, when treading on thin ice, we do not know its exact thickness and sometimes the ice breaks and sometimes it does not. Usually we find, only after breaking through it, that the ice was quite thin. Sometimes even thicker ice breaks, but such an event is affected by many circumstances we are not able to quantify (ice structure, care in treading, etc.) and we therefore consider them as random. Using many observations, however, we can estimate the probability of breaking through ice based on its thickness by using the methods of mathematical statistics. Statistics is therefore a tool of inductive thinking in such cases, where the outcome of an experiment (or observation) is affected by random variability.

Thanks to advances in computer technology, statistics is now available to all biologists. Statistical analysis of data is a necessary prerequisite of manuscript acceptance in most biological journals. These days, it is impossible to fully understand most of the research papers in biological journals without understanding the basic principles of statistics. All biologists must plan their observations and experiments, as only correctly collected data can be useful when answering their questions with the aid of statistical methods. To collect your data correctly, you need to have a basic understanding of statistics.

A knowledge of statistics has therefore become essential for successful enquiry in almost all fields of biology. But statistics are also often misused. Some even say that there are three kinds of lies: a non-intentional lie, an intentional lie and statistics. We can 'adorn' bad data by employing a complex statistical method so that the result looks like a substantial contribution to our knowledge (even finding its way into prestigious journals). Another common case of statistical misuse is interpreting statistical ('correlational') dependency as causal. In this way, one can 'prove' almost anything. A knowledge of statistics also allows biologists to differentiate statements which provide new and useful information from those where statistics are used to simply mask a lack of information, or are misused to support incorrect statements.

The way statistics are used in the everyday practice of biology changed substantially with the increased availability of statistical software. Today, everyone can evaluate her/his data on a personal computer; the results are just a few mouse clicks away. While your computer will (almost) always offer some results, often in the form of a nice-looking graph, this rather convenient process is not without its dangers. There are users who present the results provided to them by statistical programs without ever understanding what was computed. Our book therefore tries not only to teach you how to analyse your data, but also how to understand what the results of statistical processing mean.

What is **biostatistics**? We do not think that this is a separate research field. In using this term, we simply imply a focus on the application of statistics to biological problems. Alternatively, the term **biometry** is sometimes used in a similar sense. In our book, we place an emphasis on understanding the principles of the methods presented and the rules of their use, not on the mathematical derivation of the methods. We present individual methods in a way that we believe is convenient for biologists: we first show a few examples of biological problems that can be solved by a given method, and only then do we present its principles and assumptions. In our explanations we assume that the reader has attended an introductory undergraduate mathematical course, including the basics of the theory of probability. Even so, we try to avoid complex mathematical explanations whenever possible.

This book provides only basic information. We recommend that all readers continue a more detailed exploration of those methods of interest to them. The three most recommended textbooks for this are Quinn & Keough (2002), Sokal & Rohlf (2012) and Zar (2010). The first and last of these more closely reflect the mind of the biologist, as their authors have themselves participated in ecological research. In this book, we adopt some ideas from Zar's textbook about the sequence in which to present selected topics. After every chapter, we give page ranges for the three referred textbooks, each containing additional information about the particular methods. Our book is only a slight extension of a one-term course (2 hours lectures + 2 hours practicals per week) in Biostatistics, and therefore sufficient detail is lacking on some of the statistical methods useful for biologists. This primarily concerns the use of multivariate statistical analysis, traditionally addressed in separate textbooks and courses.

We assume that our readers will evaluate their data using a personal computer and we illustrate the required steps and the format of results using two different types of software. The program R lacks some of the user-friendliness provided by alternative statistical packages, but offers practically all known statistical methods, including the most modern ones, for free (more details at cran.r-project.org), and so it became *de facto* a standard tool, prevailing in published biological research papers. We assume that the reader will have a basic working knowledge of R, including working with its user interface, importing data or exporting results. The knowledge required is, however, summarised in Appendix A of this book, which can be found after the last chapter. The program Statistica represents software for the less demanding user, with a convenient range of menu choices and extensive dialogue boxes, as well as an easily accessible and modifiable graphical presentation of results. Instructions for its use are available to the reader at the textbook's website: www.cambridge.org/biostatistics.

Example data used throughout this book are available at the same website, but also from our own university's web address: www.prf.jcu.cz/biostat-data-eng.xlsx.

Note that in most of our 'use case examples' (and often also in the example data), the actual (or suggested) number of replicates is very low, perhaps too low to provide reasonable support for a real-world study. This is just to make the data easily tractable while we demonstrate the computation of test statistics. For real-world studies, we recommend the reader strives to attain more extensive datasets. If there is no citation for our example dataset, such data are not real.

In each chapter, we also show how the results derived from statistical software can be presented in research papers and also how to describe the particular statistical methods there.

In this book, we will most frequently refer to the following three statistical textbooks providing more details about the methods:

- J. H. Zar (2010) *Biostatistical Analysis*, 5th edn. Pearson, San Francisco, CA.
- G. P. Quinn & M. J. Keough (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.
- R. R. Sokal & E. J. Rohlf (2012) *Biometry*, 4th edn. W. H. Freeman, San Francisco, CA.

Other useful textbooks include:

- R. H. Green (1979) *Sampling Design and Statistical Methods for Environmental Biologists*. Wiley, New York.
- R. H. G. Jongmann, C. J. F. ter Braak & O. F. R. van Tongeren (1995) *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, Cambridge.
- P. Šmilauer & J. Lepš (2014) *Multivariate Analysis of Ecological Data Using Canoco 5*, 2nd edn. Cambridge University Press, Cambridge.

More advanced readers will find the following textbook useful:

- R. Mead (1990) *The Design of Experiments. Statistical Principles for Practical Application*. Cambridge University Press, Cambridge.

Where appropriate, we cite additional books and papers at the end of the corresponding chapter.

Acknowledgements

Both authors are thankful to their wives Olina and Majka for their ceaseless support and understanding. Our particular thanks go to Petr's wife Majka (Marie Šmilauerová), who created all the drawings which start and enliven each chapter.

We are grateful to Conor Redmond for his careful and efficient work at improving our English grammar and style.

The feedback of our students was of great help when writing this book, particularly the in-depth review from a student point of view provided by Václava Hazuková. We appreciate the revision of Section 2.7, kindly provided by Cajo ter Braak.