

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 & Assessment

978-1-108-72734-1 — Biostatistics with R: An Introductory Guide for Field Biologists

Jan Lepš, Petr Šmilauer

Frontmatter

[More Information](#)



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

101 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

© Jan Lepš and Petr Šmilauer 2020

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press & Assessment.

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 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.

Contents

	<i>Preface</i>	<i>page</i> xiii
	<i>Acknowledgements</i>	xvii
1	Basic Statistical Terms, Sample Statistics	1
	1.1 Cases, Variables and Data Types	1
	1.2 Population and Random Sample	3
	1.3 Sample Statistics	4
	1.4 Precision of Mean Estimate, Standard Error of Mean	9
	1.5 Graphical Summary of Individual Variables	10
	1.6 Random Variables, Distribution, Distribution Function, Density Distribution	10
	1.7 Example Data	13
	1.8 How to Proceed in R	13
	1.9 Reporting Analyses	17
	1.10 Recommended Reading	18
2	Testing Hypotheses, Goodness-of-Fit Test	19
	2.1 Principles of Hypothesis Testing	19
	2.2 Possible Errors in Statistical Tests of Hypotheses	21
	2.3 Null Models with Parameters Estimated from the Data: Testing Hardy–Weinberg Equilibrium	26
	2.4 Sample Size	26
	2.5 Critical Values and Significance Level	27
	2.6 Too Good to Be True	29
	2.7 Bayesian Statistics: What is It?	30
	2.8 The Dark Side of Significance Testing	32
	2.9 Example Data	35
	2.10 How to Proceed in R	35
	2.11 Reporting Analyses	37
	2.12 Recommended Reading	37
3	Contingency Tables	39
	3.1 Two-Way Contingency Tables	39
	3.2 Measures of Association Strength	44
	3.3 Multidimensional Contingency Tables	46
	3.4 Statistical and Causal Relationship	47
	3.5 Visualising Contingency Tables	49
	3.6 Example Data	50
	3.7 How to Proceed in R	50

3.8	Reporting Analyses	54
3.9	Recommended Reading	54
4	Normal Distribution	55
4.1	Main Properties of a Normal Distribution	55
4.2	Skewness and Kurtosis	56
4.3	Standardised Normal Distribution	57
4.4	Verifying the Normality of a Data Distribution	58
4.5	Example Data	60
4.6	How to Proceed in R	60
4.7	Reporting Analyses	63
4.8	Recommended Reading	64
5	Student's t Distribution	65
5.1	Use Case Examples	65
5.2	t Distribution and its Relation to the Normal Distribution	66
5.3	Single Sample Test and Paired t Test	67
5.4	One-Sided Tests	70
5.5	Confidence Interval of the Mean	72
5.6	Test Assumptions	73
5.7	Reporting Data Variability and Mean Estimate Precision	74
5.8	How Large Should a Sample Size Be?	77
5.9	Example Data	79
5.10	How to Proceed in R	79
5.11	Reporting Analyses	82
5.12	Recommended Reading	83
6	Comparing Two Samples	84
6.1	Use Case Examples	84
6.2	Testing for Differences in Variance	85
6.3	Comparing Means	87
6.4	Example Data	88
6.5	How to Proceed in R	88
6.6	Reporting Analyses	91
6.7	Recommended Reading	91
7	Non-parametric Methods for Two Samples	92
7.1	Mann–Whitney Test	93
7.2	Wilcoxon Test for Paired Observations	95
7.3	Using Rank-Based Tests	97
7.4	Permutation Tests	97
7.5	Example Data	99
7.6	How to Proceed in R	99
7.7	Reporting Analyses	102
7.8	Recommended Reading	103

8	One-Way Analysis of Variance (ANOVA) and Kruskal–Wallis Test	104
8.1	Use Case Examples	104
8.2	ANOVA: A Method for Comparing More Than Two Means	104
8.3	Test Assumptions	105
8.4	Sum of Squares Decomposition and the F Statistic	106
8.5	ANOVA for Two Groups and the Two-Sample t Test	108
8.6	Fixed and Random Effects	108
8.7	F Test Power	109
8.8	Violating ANOVA Assumptions	110
8.9	Multiple Comparisons	111
8.10	Non-parametric ANOVA: Kruskal–Wallis Test	115
8.11	Example Data	116
8.12	How to Proceed in R	117
8.13	Reporting Analyses	127
8.14	Recommended Reading	128
9	Two-Way Analysis of Variance	129
9.1	Use Case Examples	129
9.2	Factorial Design	130
9.3	Sum of Squares Decomposition and Test Statistics	132
9.4	Two-Way ANOVA with and without Interactions	134
9.5	Two-Way ANOVA with No Replicates	135
9.6	Experimental Design	135
9.7	Multiple Comparisons	137
9.8	Non-parametric Methods	138
9.9	Example Data	139
9.10	How to Proceed in R	139
9.11	Reporting Analyses	149
9.12	Recommended Reading	150
10	Data Transformations for Analysis of Variance	151
10.1	Assumptions of ANOVA and their Possible Violations	151
10.2	Log-transformation	153
10.3	Arcsine Transformation	156
10.4	Square-Root and Box–Cox Transformation	156
10.5	Concluding Remarks	157
10.6	Example Data	158
10.7	How to Proceed in R	158
10.8	Reporting Analyses	163
10.9	Recommended Reading	163
11	Hierarchical ANOVA, Split-Plot ANOVA, Repeated Measurements	164
11.1	Hierarchical ANOVA	164
11.2	Split-Plot ANOVA	167
11.3	ANOVA for Repeated Measurements	169

11.4	Example Data	171
11.5	How to Proceed in R	171
11.6	Reporting Analyses	181
11.7	Recommended Reading	182
12	Simple Linear Regression: Dependency Between Two Quantitative Variables	183
12.1	Use Case Examples	183
12.2	Regression and Correlation	184
12.3	Simple Linear Regression	184
12.4	Testing Hypotheses	187
12.5	Confidence and Prediction Intervals	190
12.6	Regression Diagnostics and Transforming Data in Regression	190
12.7	Regression Through the Origin	195
12.8	Predictor with Random Variation	197
12.9	Linear Calibration	197
12.10	Example Data	198
12.11	How to Proceed in R	198
12.12	Reporting Analyses	204
12.13	Recommended Reading	205
13	Correlation: Relationship Between Two Quantitative Variables	206
13.1	Use Case Examples	206
13.2	Correlation as a Dependency Statistic for Two Variables on an Equal Footing	206
13.3	Test Power	209
13.4	Non-parametric Methods	212
13.5	Interpreting Correlations	212
13.6	Statistical Dependency and Causality	213
13.7	Example Data	216
13.8	How to Proceed in R	216
13.9	Reporting Analyses	218
13.10	Recommended Reading	218
14	Multiple Regression and General Linear Models	219
14.1	Use Case Examples	219
14.2	Dependency of a Response Variable on Multiple Predictors	219
14.3	Partial Correlation	223
14.4	General Linear Models and Analysis of Covariance	224
14.5	Example Data	225
14.6	How to Proceed in R	226
14.7	Reporting Analyses	237
14.8	Recommended Reading	238

15	Generalised Linear Models	239
15.1	Use Case Examples	239
15.2	Properties of Generalised Linear Models	240
15.3	Analysis of Deviance	242
15.4	Overdispersion	243
15.5	Log-linear Models	243
15.6	Predictor Selection	244
15.7	Example Data	245
15.8	How to Proceed in R	246
15.9	Reporting Analyses	250
15.10	Recommended Reading	251
16	Regression Models for Non-linear Relationships	252
16.1	Use Case Examples	252
16.2	Introduction	253
16.3	Polynomial Regression	253
16.4	Non-linear Regression	255
16.5	Example Data	256
16.6	How to Proceed in R	256
16.7	Reporting Analyses	259
16.8	Recommended Reading	260
17	Structural Equation Models	261
17.1	Use Case Examples	261
17.2	SEMs and Path Analysis	261
17.3	Example Data	265
17.4	How to Proceed in R	265
17.5	Reporting Analyses	272
17.6	Recommended Reading	272
18	Discrete Distributions and Spatial Point Patterns	274
18.1	Use Case Examples	274
18.2	Poisson Distribution	274
18.3	Comparing the Variance with the Mean to Measure Spatial Distribution	276
18.4	Spatial Pattern Analyses Based on the K-function	279
18.5	Binomial Distribution	280
18.6	Example Data	283
18.7	How to Proceed in R	283
18.8	Reporting Analyses	289
18.9	Recommended Reading	289
19	Survival Analysis	290
19.1	Use Case Examples	290
19.2	Survival Function and Hazard Rate	291

19.3	Differences in Survival Among Groups	293
19.4	Cox Proportional Hazard Model	293
19.5	Example Data	295
19.6	How to Proceed in R	295
19.7	Reporting Analyses	302
19.8	Recommended Reading	302
20	Classification and Regression Trees	303
20.1	Use Case Examples	303
20.2	Introducing CART	304
20.3	Pruning the Tree and Crossvalidation	306
20.4	Competing and Surrogate Predictors	307
20.5	Example Data	308
20.6	How to Proceed in R	309
20.7	Reporting Analyses	316
20.8	Recommended Reading	316
21	Classification	317
21.1	Use Case Examples	317
21.2	Aims and Properties of Classification	317
21.3	Input Data	319
21.4	Similarity and Distance	319
21.5	Clustering Algorithms	320
21.6	Displaying Results	320
21.7	Divisive Methods	321
21.8	Example Data	322
21.9	How to Proceed in R	322
21.10	Other Software	324
21.11	Reporting Analyses	325
21.12	Recommended Reading	325
22	Ordination	326
22.1	Use Case Examples	327
22.2	Unconstrained Ordination Methods	327
22.3	Constrained Ordination Methods	330
22.4	Discriminant Analysis	331
22.5	Example Data	333
22.6	How to Proceed in R	333
22.7	Alternative Software	340
22.8	Reporting Analyses	341
22.9	Recommended Reading	341
Appendix A:	First Steps with R Software	343
A.1	Starting and Ending R, Command Line, Organising Data	343
A.2	Managing Your Data	349

Table of Contents		xi
A.3	Data Types in R	351
A.4	Importing Data into R	357
A.5	Simple Graphics	359
A.6	Frameworks for R	360
A.7	Other Introductions to Work with R	362
<i>Index</i>		363

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.