

Core Statistics

Based on two courses for new graduate students, *Core Statistics* provides concise coverage of the fundamentals of inference for parametric statistical models, including both theory and practical numerical computation. The book considers both frequentist maximum likelihood estimation and Bayesian stochastic simulation, focusing on general methods applicable to a wide range of models and emphasizing the common questions addressed by the two approaches.

This compact book aims to cover the core knowledge needed by beginning graduate students in statistical subjects, at a level suitable for those going on to develop new methods or undertaking novel applications of statistical modelling: Bayesian and frequentist approaches to modelling and inference; practical computational implementation, including numerical issues; brief coverage of some essential probability; and the essentials of R as a statistical programming language. Aimed also at any quantitative scientist who uses statistical methods, this book will deepen readers' understanding of why and when methods work and explain how to develop suitable methods for non-standard situations, such as in ecology, big data analysis and genomics.

SIMON N. WOOD works as a professor of statistics at the University of Bath, with particular interests in statistical computing, methodology of smoothing and environmental statistics.

INSTITUTE OF MATHEMATICAL STATISTICS
TEXTBOOKS

Editorial Board

D. R. Cox (University of Oxford)
B. Hambly (University of Oxford)
S. Holmes (Stanford University)
X.-L. Meng (Harvard University)

IMS Textbooks give introductory accounts of topics of current concern suitable for advanced courses at master's level, for doctoral students and for individual study. They are typically shorter than a fully developed textbook, often arising from material created for a topical course. Lengths of 100–290 pages are envisaged. The books typically contain exercises.

Other Books in the Series

1. *Probability on Graphs*, by Geoffrey Grimmett
2. *Stochastic Networks*, by Frank Kelly and Elena Yudovina
3. *Bayesian Filtering and Smoothing*, by Simo Särkkä
4. *The Surprising Mathematics of Longest Increasing Subsequences*, by Dan Romik
5. *Noise Sensitivity of Boolean Functions and Percolation*, by Christophe Garban and Jeffrey E. Steif

Core Statistics

SIMON N. WOOD
University of Bath



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/9781107415041

© Simon N. Wood 2015

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 2015

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

Library of Congress Cataloging-in-Publication data
 Wood, Simon N. Core statistics / Simon Wood, University of Bath.
 pages cm

Includes bibliographical references and index.

ISBN 978-1-107-07105-6 (hardback)

1. Statistics – Textbooks. 2. Statistics – Study and teaching (Graduate) I. Title.
 QA276.2.W66 2015
 519.5–dc23 2014039012

ISBN 978-1-107-07105-6 Hardback

ISBN 978-1-107-41504-1 Paperback

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>	viii
1 Random variables	1
1.1 Random variables	1
1.2 Cumulative distribution functions	1
1.3 Probability (density) functions	2
1.4 Random vectors	3
1.5 Mean and variance	7
1.6 The multivariate normal distribution	9
1.7 Transformation of random variables	11
1.8 Moment generating functions	13
1.9 The central limit theorem	13
1.10 Chebyshev, Jensen and the law of large numbers	14
1.11 Statistics	16
Exercises	17
2 Statistical models and inference	19
2.1 Examples of simple statistical models	20
2.2 Random effects and autocorrelation	23
2.3 Inferential questions	26
2.4 The frequentist approach	27
2.5 The Bayesian approach	37
2.6 Design	44
2.7 Useful single-parameter normal results	46
Exercises	48
3 R	49
3.1 Basic structure of R	50
3.2 R objects	51
3.3 Computing with vectors, matrices and arrays	54
3.4 Functions	63
3.5 Useful built-in functions	67

vi	<i>Contents</i>	
3.6	Object orientation and classes	68
3.7	Conditional execution and loops	70
3.8	Calling compiled code	73
3.9	Good practice and debugging	75
	Exercises	76
4	Theory of maximum likelihood estimation	79
4.1	Some properties of the expected log likelihood	79
4.2	Consistency of MLE	81
4.3	Large sample distribution of MLE	82
4.4	Distribution of the generalised likelihood ratio statistic	83
4.5	Regularity conditions	85
4.6	AIC: Akaike's information criterion	85
	Exercises	87
5	Numerical maximum likelihood estimation	89
5.1	Numerical optimisation	89
5.2	A likelihood maximisation example in R	100
5.3	Maximum likelihood estimation with random effects	105
5.4	R random effects MLE example	109
5.5	Computer differentiation	116
5.6	Looking at the objective function	127
5.7	Dealing with multimodality	130
	Exercises	130
6	Bayesian computation	133
6.1	Approximating the integrals	133
6.2	Markov chain Monte Carlo	135
6.3	Interval estimation and model comparison	149
6.4	An MCMC example: algal growth	154
6.5	Geometry of sampling and construction of better proposals	158
6.6	Graphical models and automatic Gibbs sampling	168
	Exercises	182
7	Linear models	184
7.1	The theory of linear models	185
7.2	Linear models in R	195
7.3	Extensions	207
	Exercises	210
	Appendix A Some distributions	213
A.1	Continuous random variables: the normal and its relatives	213
A.2	Other continuous random variables	215
A.3	Discrete random variables	217

<i>Contents</i>		vii
Appendix B Matrix computation		219
B.1 Efficiency in matrix computation		219
B.2 Choleski decomposition: a matrix square root		221
B.3 Eigen-decomposition (spectral-decomposition)		224
B.4 Singular value decomposition		230
B.5 The QR decomposition		231
B.6 Sparse matrices		232
Appendix C Random number generation		233
C.1 Simple generators and what can go wrong		233
C.2 Building better generators		237
C.3 Uniform generation conclusions		239
C.4 Other deviates		240
<i>References</i>		241
<i>Index</i>		245

Preface

This book is aimed at the numerate reader who has probably taken an introductory statistics and probability course at some stage and would like a brief introduction to the core methods of statistics and how they are applied, not necessarily in the context of standard models. The first chapter is a brief review of some basic probability theory needed for what follows. Chapter 2 discusses statistical models and the questions addressed by statistical inference and introduces the maximum likelihood and Bayesian approaches to answering them. Chapter 3 is a short overview of the R programming language. Chapter 4 provides a concise coverage of the large sample theory of maximum likelihood estimation, and Chapter 5 discusses the numerical methods required to use this theory. Chapter 6 covers the numerical methods useful for Bayesian computation, in particular Markov chain Monte Carlo. Chapter 7 provides a brief tour of the theory and practice of linear modelling. Appendices then cover some useful information on common distributions, matrix computation and random number generation. The book is neither an encyclopedia nor a cookbook, and the bibliography aims to provide a compact list of the most useful sources for further reading, rather than being extensive. The aim is to offer a concise coverage of the core knowledge needed to understand and use parametric statistical methods and to build new methods for analysing data. Modern statistics exists at the interface between computation and theory, and this book reflects that fact. I am grateful to Nicole Augustin, Finn Lindgren, the editors at Cambridge University Press and the students on the Bath course ‘Applied Statistical Inference’ and the Academy for PhD Training in Statistics course ‘Statistical Computing’ for many useful comments.