Introduction to Statistics and Data Analysis

Christian Heumann · Michael Schomaker
Shalabh

Introduction to Statistics and Data Analysis

With Exercises, Solutions and Applications in R

Springer

Christian Heumann
Department of Statistics
Ludwig-Maximilians-Universität München
München
Germany

Michael Schomaker
Centre for Infectious Disease Epidemiology
and Research
University of Cape Town
Cape Town
South Africa

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Kanpur
India

ISBN 978-3-319-46160-1 ISBN 978-3-319-46162-5 (eBook)
DOI 10.1007/978-3-319-46162-5

Library of Congress Control Number: 2016955516

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The success of the open-source statistical software "*R*" has made a significant impact on the teaching and research of statistics in the last decade. Analysing data is now easier and more affordable than ever, but choosing the most appropriate statistical methods remains a challenge for many users. To understand and interpret software output, it is necessary to engage with the fundamentals of statistics.

However, many readers do not feel comfortable with complicated mathematics. In this book, we attempt to find a healthy balance between explaining statistical concepts comprehensively and showing their application and interpretation using *R*.

This book will benefit beginners and self-learners from various backgrounds as we complement each chapter with various exercises and detailed and comprehensible solutions. The results involving mathematics and rigorous proofs are separated from the main text, where possible, and are kept in an appendix for interested readers. Our textbook covers material that is generally taught in introductory-level statistics courses to students from various backgrounds, including sociology, biology, economics, psychology, medicine, and others. Most often, we introduce the statistical concepts using examples and illustrate the calculations both manually and using *R*.

However, while we provide a gentle introduction to *R* (in the appendix), this is not a software book. Our emphasis lies on explaining statistical concepts correctly and comprehensively, using exercises and software to delve deeper into the subject matter and learn about the conceptual challenges that the methods present.

This book's homepage, http://chris.userweb.mwn.de/book/, contains additional material, most notably the software codes needed to answer the software exercises, and data sets. In the remainder of this book, we will use grey boxes

```
R-command()
```

to introduce the relevant *R* commands. In many cases, the code can be directly pasted into *R* to reproduce the results and graphs presented in the book; in others, the code is abbreviated to improve readability and clarity, and the detailed code can be found online.

Many years of teaching experience, from undergraduate to postgraduate level, went into this book. The authors hope that the reader will enjoy reading it and find it a useful reference for learning. We welcome critical feedback to improve future editions of this book. Comments can be sent to christian.heumann@stat.uni-muenchen.de, shalab@iitk.ac.in, and michael.schomaker@uct.ac.za who contributed equally to this book.

We thank Melanie Schomaker for producing some of the figures and giving graphical advice, Alice Blanck from Springer for her continuous help and support, and Lyn Imeson for her dedicated commitment which improved the earlier versions of this book. We are grateful to our families who have supported us during the preparation of this book.

München, Germany — Christian Heumann
Cape Town, South Africa — Michael Schomaker
Kanpur, India — Shalabh
November 2016

Contents

About the Authors

Prof. Christian Heumann is a professor at the Ludwig-Maximilians-Universität München, Germany, where he teaches students in Bachelor and Master programs offered by the Department of Statistics, as well as undergraduate students in the Bachelor of Science programs in business administration and economics. His research interests include statistical modeling, computational statistics and all aspects of missing data.

Dr. Michael Schomaker is a Senior Researcher and Biostatistician at the Centre for Infectious Disease Epidemiology & Research (CIDER), University of Cape Town, South Africa. He received his doctoral degree from the University of Munich. He has taught undergraduate students for many years and has written contributions for various introductory textbooks. His research focuses on missing data, causal inference, model averaging and HIV/AIDS.

Prof. Shalabh is a Professor at the Indian Institute of Technology Kanpur, India. He received his Ph.D. from the University of Lucknow (India) and completed his post-doctoral work at the University of Pittsburgh (USA) and University of Munich (Germany). He has over twenty years of experience in teaching and research. His main research areas are linear models, regression analysis, econometrics, measurement error models, missing data models and sampling theory.