

Preface

The R programming language is becoming the lingua franca of computational statistics. It is the usual statistical software platform used by statisticians, economists, engineers and scientists both in corporations and in academia. Established international companies use R in their data analysis. R has gained its popularity for two reasons. First, it is an OS independent free open-source program which is popularised and improved by hundreds of volunteers all over the world. A plethora of packages are available for many scientific disciplines. Second, common analysts can do complicated analyses without deep computer programming knowledge. This book on the basic elements of computational statistics presents the tools and concepts of univariate and multivariate data analyses with a strong focus on applications and implementations. The aim of this book is to present data analysis in a way that is understandable for non-mathematicians and practitioners who are confronted by statistical data analysis. All practical examples may be recalculated and modified by the reader: all data sets and programmes (Quantlets) used in the book are downloadable from the publisher's home page of this book (www.quantlet.de). The text contains a wide variety of exercises and covers the basic mathematical, statistical and programming problems.

The first chapter introduces the reader to the basics of the R language, taking into account that only minimal prior experience in programming is required. Starting with the developing history and R environments under different operating systems, the book discusses the syntax. We start the description of the syntax with the classical 'Hello World!!!' program. The use of R as an advanced calculator, data types, loops, if then conditions, own function construction and classes are the topics covered in this chapter. As in statistical analysis one deals with data, special attention is paid to work with vectors and matrices.

The second part deals with the numerical techniques which one needs during the analysis. A short excursion into matrix algebra will be helpful in understanding multivariate techniques provided in the further sections. Different methods of numerical integration, differentiation and root finding help the reader to get inside the core of the R system.

Chapter 3 highlights set theory, combinatoric rules, plus some of the main discrete distributions: binomial, multinomial, hypergeometric and Poisson.

Different characteristics, cumulative distribution functions and density functions of the continuous distributions: uniform, normal, t , χ^2 , F , exponential and Cauchy will be explained in detail in Chapter 4.

The next chapter is devoted to univariate statistical analysis and basic smoothing techniques. The histogram, kernel density estimator, graphical representation of the data, confidence intervals, different simple tests as well as tests that need more computations, like the Wilcoxon, Kruskal–Wallis, sign tests, are the topics of Chapter 5.

The sixth chapter deals with multivariate distributions: their definition, characteristics and application of general multivariate distributions, multinormal distributions, as well as classes of copulas. Further, Chapter 7 discusses linear and nonlinear relationships via regression models.

Chapter 8 partially extends the problems solved in Chapter 5, but also considers more sophisticated topics, such as multidimensional scaling, principal component, factor, discriminant and cluster analysis. These techniques are difficult to apply without computational power, so they are of special interest in this book.

Theoretical models need to be calibrated in practice. If there is no data available, then Monte Carlo simulation techniques are necessary parts of each study. Chapter 9 starts from simple sampling techniques from the uniform distribution. These are further extended to simulation methods from other univariate distributions. We also discuss simulation from multivariate distributions, especially copulae.

Chapter 10 describes more advanced graphical techniques, with special attention to three-dimensional graphics and interactive programmes using packages `lattice`, `rgl` and `rpanel`.

This book is designed for the advanced undergraduate and first-year graduate student as well as for the inexperienced data analyst who would like a tour of the various statistical tools in a data analysis workshop. The experienced reader with a good knowledge of statistics and programming will certainly skip some sections of the univariate models, but hopefully enjoy the various mathematical roots of the multivariate techniques. A graduate student might think that the first section on description techniques is well known to him from his training in introductory statistics. The programming, mathematical and the applied parts of the book will certainly introduce him into the rich realm of statistical data analysis modules.

A book of this kind would not have been possible without the help of many friends, colleagues and students. For many suggestions, corrections and technical support, we would like to thank Aymeric Bouley, Xiaofeng Cao, Johanna Simone Eckel, Philipp Gschöpf, Gunawan Gunawan, Johannes Haupt, Uri Yakobi Keller, Polina Marchenko, Félix Revert, Alexander Ristig, Benjamin Samulowski, Martin Schelisch, Christoph Schult, Noa Tamir, Anastasija Teterova, Tatjana Tissen-Diabaté, Ivan Vasylenko and Yafei Xu. We thank Alice Blanck and Veronika Rosteck from Springer Verlag for continuous support and valuable

suggestions on the style of writing and the content covered. Special thanks go to the anonymous proofreaders who checked not only the language but also the statistical, programming and mathematical content of the book. All errors are our own.

Berlin, Germany
Dresden, Germany
Augsburg, Germany
April 2017

Wolfgang Karl Härdle
Ostap Okhrin
Yarema Okhrin

Basic Elements of Computational Statistics

Härdle, W.; Okhrin, O.; Okhrin, Y.

2017, XXI, 305 p. 97 illus., 66 illus. in color., Hardcover

ISBN: 978-3-319-55335-1