

Preface

Everything of importance has been said before by somebody who did not discover it.

Alfred North Whitehead

Everything that needs to be said has already been said. But since no one was listening, everything must be said again.

André Gide

True to the above quotes, most of this book's contents have appeared in print before, but often where biologists are unlikely to look – statistics journals and books, and methods papers in other fields. My task is to translate ideas known to statisticians into the language of experimental biology.¹ With a background in both biology (BSc, PhD, postdoc) and data analysis (MPhil in Computational Biology and over seven years working as a preclinical statistician in the pharmaceutical industry), hopefully I am fluent enough in both languages to perform a successful translation.²

The contents of this book have little overlap with other statistics-for-biologists books because they mostly focus on statistical analysis. Analysis is but one step of the scientific workflow (Figure 0.1), and before you can analyse data you need to do an experiment. This requires planning, good execution, and quality control checks. These critical topics are rarely taught to biologists, who are expected to learn them on their own. The consequence of this approach is predictable; some biologists obtain the necessary skills, but many do not. This book focuses on the first three steps of the scientific workflow, and data analysis is briefly discussed in Chapter 4.

This book was written to improve the quality of research conducted in academic, government, and industrial labs and institutions. Scientists and funders now recognise that bias and irreproducibility are undermining preclinical biomedical research [2, 5, 28, 30, 42, 80, 83, 84, 123, 172, 240, 251, 305, 316, 342]. There are many reasons why experiments cannot be reproduced (discussed in Chapter 1) and this book focuses on the role that experimental design and data analysis have on making results reproducible.

¹ The term *biology* refers to laboratory-based experimental biology throughout. 'Field biologists' also conduct experiments, and most statistics-for-biologists books target this audience.

² There are some novel ideas here, such as the distinction between front-aligned and end-aligned designs (Section 2.11) and the distinction between biological, experimental, and observational units, to replace the biological versus technical replicate distinction (all of Chapter 3).

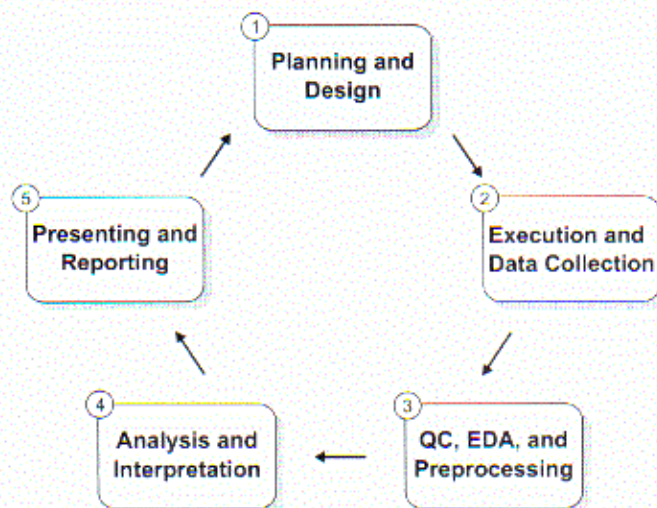


Fig. 0.1 The scientific workflow. This book focuses on steps 1–3. QC = quality control; EDA = exploratory data analysis.

Prerequisites

This book is for experimental biologists, at any level, conducting basic research or with an applied, clinical, or translational focus. Knowledge covered in an introductory statistics-for-biologists course is assumed, and concepts like the standard deviation and common statistical tests such as the *t*-test, analysis of variance (ANOVA), regression, and correlation should be familiar. It is fine if some time has passed since you formally covered these topics. Mathematical proofs are not included and equations are kept to a minimum, but given the subject, are unavoidable. The emphasis is on the ideas, concepts, and principles, and how to implement them. Hand calculations are unnecessary because statistical software is available.

Quantitative researchers who analyse biological data such as statisticians, bioinformaticians, and computational biologists might also find this book useful. Topics of interest include sources of heterogeneity and confounding in biological experiments (Section 2.12), quality control checks for biological data (Section 6.1), and understanding which types of replication address biologically interesting questions (Chapter 3).

The freely available R statistics language is used for data analysis and graphs.³ Prior knowledge is useful, but not required. The Appendix gives a brief introduction to R and the examples in the main text assume familiarity with this material. The topics however can be followed without learning or using R. The data sets can be found in the *labstats* package on CRAN⁴ and R code can be downloaded from GitHub.⁵

³ Available at www.r-project.com

⁴ <https://cran.r-project.org/web/packages/labstats/>

⁵ <https://stanlazic.github.io/EDLB.html>

The key prerequisite to derive maximum value from this book is experience conducting biological experiments and analysing the subsequent data – and the more experience the better!

How to read this book

Chapters 1–5 should be read in order as later material depends on earlier ideas, but Chapter 6 on Exploratory Data Analysis can be read at any time. Chapters 1–3 contain no R code, but for Chapters 4–6 sitting in front of a computer and running the code will reinforce the ideas.

Ideas or concepts discussed in detail later in the book will inevitably have to be mentioned earlier. To avoid excessive cross-referencing, the glossary lists the page where the main discussion of the entry is located (if there is one). For example, the term *experimental unit* is mentioned for the first time in this preface, but is discussed extensively in Section 3.2. The glossary entry for this term provides a short definition and indicates that further information can be found on page 96.

Typographical conventions

Constant width font is used for R code, R output, and when referring to R functions or objects. Lines of code entered by the user start with ‘>’ or ‘+’. These symbols do not need to be entered, only the code that follows them. A sign like the one in the margin draws attention to a warning, a key point, a subtlety with R, or a concept that is often misunderstood.

Acknowledgements

This book has benefited greatly from comments by Maarten van Dijk, Irmgard Amrein, and especially Lutz Slomianka. Pierre Farmer and Miguel Camargo also provided constructive feedback on earlier drafts. My wife, Brynn, has read every word in this book, which is beyond the call of duty, and her comments have improved it immensely. I also thank her for her support, well, at least until page 305, at which point she declared, ‘You should stop now; no one wants to read that much about statistics.’ I didn’t always follow everyone’s good advice, but I am grateful for their input.

Katrina Halliday and Jade Scard at Cambridge University Press were a pleasure to work with and made the whole process easy and enjoyable. I also thank Judith Shaw for her expert copy-editing. Finally, I would like to thank the developers and contributors of the free software R, Emacs, LaTeX, JabRef, knitr, and Inkscape, which I used to write this book.

S.E. Lazic
Cambridge, 2016