Andrew Bray
University of Massachusetts, Amherst
Amherst, MA
760-519-5979
andrew.bray@gmail.com

# Implementing Reproducible Research

Victoria Stodden, Friedrich Leisch, Roger D. Peng.

Publisher Address: CRC Press, 2013. ISBN 978-1-4665-6159-5. xix+428 pp.

The challenge of reproducibility in the computational era is being dealt with across the sciences, with each field developing its own tools and best practices. This book is an important step in bringing together a broad group of scientists to share what has been learned. In the preface to this book, the editors clarify that their objective is not to convince the reader *why* reproducibility is important, but rather *how* it can be achieved.

Reproducibility is defined as the calculation of quantitiative scientific results by independent scientists using the original datasets and methods. The chapters from the various contributors are organized into three sections - Tools, Practices and Guidelines, and Platforms - that mirror the three directions of research in reproducibility.

One contributor usefully divides tools for reproducible research into three general categories: tools for literate programming (knitr in R), workflow management systems (VisTrails), and tools for environment capture (Sumatra and CDE). I was most familiar with knitr, but I still benefitted greatly from the discussion of the design elements that specifically enable reproducibility, including chunk re-use and conditional evaluation. The remaining tools were new to me and address the added challenges when faced with a more complex workflow and software that involves many dependencies. Note that these chapters are not full tutorials; you will not emerge as a competent tool-user, but hopefully with a sense of which tools you would like to invest time in learning.

Part two

As an user of R and the knitr package, I really enjoyed getting to learn tools and methods of the Python-verse, including IPython and org-mode.

Several of the chapters discuss how reproducibility can be facilitated by working with open data, open source software, open access journals. I was left wondering why the authors chose to publish their work as a printed book costing $60 (it appears that none of the authors used a SPARC license to allow for access to chapters on their personal websites, as is recommended in chapter ZZQ). As the

As a cover-to-cover read, the book can be repetative (LP, VCS,

Given the broad coverage of fields and range in technical detail, it is difficult to know who the target audience for this book is.

Although it was not noted anywhere in the text, full pdfs of each chapter can be downloaded for free through the Open Science Framework (implementingrr.org). This was a relief for me: several of the

In sum: do not buy this book. Do read the pdfs of the chapters of interest to you and distribute them far and wide. This is vitally important stuff.

Circular data refers to data that may be thought of as points on the unit circle, such as wind direction, or time of day. As the authors note, there are not many books available on the topic, the most recent being Mardia and Jupp (1999) and Jammalamadaka and SenGupta (2001), which are more theory-oriented texts. The authors state that they aimed to produce a short, modern, computer based introduction to the analysis of circular data which would be useful to both scientists and statisticians. They make extensive use of the R *circular* package, and include some code of their own. I came to this book as a long time R user, but having no experience analyzing circular data, so I was curious to see their approach.

The first chapter gives a brief introduction to circular statistics and R, sensibly directing those who don't know R to consult other resources, including internet sources such as the nearest CRAN mirror. It also introduces the R *circular* package and makes note of some of the default choices that affect circular data. The authors have established a website for code and data used in the book which also includes an R workspace containing those items. The text has a straightforward organization. Chapter 2 covers graphical methods for circular data, and Chapter 3 descriptive statistics. The authors warn the reader of the existence alternative definitions of variance, and helpfully suggest clues that one might use to guess which was used in a published paper that didn't explicitly state which definition was used. Chapter 4 presents definitions of moments and densities for circular data. Chapter 5 covers some basic elements of inference: tests of uniformity and symmetry, bootstrapping, and testing a null hypothesis about the mean. Chapter 6 covers maximum likelihood estimation for the unimodal distributions presented in Chapter 4. Chapter 7 covers the comparison of two samples, and Chapter 8 deals with regression models. Overall I found the presentation clear, if rather brief. I suspect that some sections would be challenging for scientists without a strong background in mathematics. In particular the definition of the trigonometric moments is completely abstract, with no examples given to help the neophyte. This brevity should not be a problem for the scientist

who cares only to acquire the R tools to analyze circular data, but might be a challenge for those who wish to achieve a deeper understanding. The authors do refer readers to other texts on circular data which cover the theory in greater detail.

As a long time R user, I had a few quibbles. In Chapter 1 the authors give instructions for installing the circular package that make use of a menu interface instead of the simple and direct `install.packages("circular")`, and they give examples using the dangerous shorthand `'T'` for `'TRUE'`, though the code in later chapters consistently uses `'TRUE'`. The code presented does not conform to any standard style guidelines for R code, for example there is no use of indentation, and often two commands are entered on a single line, separated by a semi-colon. These are minor irritations rather than serious failings.

I also question the analytical advice at the beginning of Chapter 5 where the authors suggest that one start with a test of uniformity, and if that null hypothesis is rejected, test for symmetry. They then suggest fitting the Jones-Pewsy family (no relation) if the null hypothesis of symmetry is not rejected, and something like an inverse Batschelet distribution if it is. This seems akin to testing for normality before fitting a normal distribution. As the authors later note, one can fit the inverse Batschelet and test null hypotheses about parameters corresponding to symmetry or reduction to the von Mises distribution via likelihood ratio tests. The authors also provide functions for quantile-quantile plots. In my opinion, fitting a plausible distribution followed by checking diagnostic plots should be the canonical practice, rather than a sequence of preliminary tests or a formal goodness-of-fit test.

The strong points of this text: how to use the R *circular* package, with datasets and examples to illustrate the methods. The authors have created a website (`http://circstatinr.st-andrews.ac.uk`) where one may find R code and datasets, as well as an R workspace containing those items. They have provided R functions for fitting recently developed distributions and other modern methods such as bootstrapping. One who already knows R can quickly get up to speed. I think the text will be useful for self-study by scientists and statisticians, and potentially useful in some applied courses or as a supplement to a more theoretical course. There is much it doesn't cover in any depth, including mixture distributions and Bayesian inference, but perhaps that is too much to ask of a concise introduction.

Andrew Bray
*University of Massachusetts, Amherst*