

## Preface

R is a powerful and flexible statistical and graphical environment that is freely distributed under the GNU Public Licence<sup>a</sup> for all major computing platforms (Windows, MacOSX and Linux). This open source licence along with a relatively simple scripting syntax has promoted diverse and rapid evolution and contribution. As the broader scientific community continues to gain greater instruction and exposure to the overall project, the popularity of R as a teaching and research tool continues to accelerate.

It is now widely acknowledged that R proficiency as a scientific skill set is becoming increasingly more desirable and useful throughout the scientific community. However, as with most open source developments, the emphasis of the R project remains on the expansive development of tools and features. Applied documentation still remains somewhat sparse and somewhat incomprehensible to the average biologist. Whilst there are a number of excellent texts on R emerging, the bulk of these texts are devoted to the R language itself. Any featured examples therein are used primarily for the purpose of illustrating the suite of commonly used R features and procedures, rather than to illustrate how R can be used to perform common biostatistical analyses.

Coinciding with the increasing interest in R as both a learning and research tool for biostatistics, has been the success of a relatively new major biostatistics textbook (Quinn and Keough, 2002). This text provides detailed coverage of most of the major statistical concepts and tests that biologists are likely to encounter with an emphasis on the practical implementation of these concepts with real biological data. Undoubtedly, a large part of the appeal of this book is attributable to the extensive use of real biological examples to augment and reinforce the text. Furthermore, by concentrating on the information biologists need to implement their research, and avoiding the overuse of complex mathematical descriptions, the authors have appealed to those biologists who don't require (or desire) a knowledge of performing or programming entire analyses from scratch. Such biologists tend to use statistical software that is already available and specifically desire information that will help them achieve reliable statistical and biological outcomes. Quinn and Keough (2002) also advocate a number of alternative

<sup>a</sup> This is an open source licence that ensured that the application as well as its source code is freely available to use, modify and redistribute.

texts that provide more detailed coverage of specific topics and that also adopt this real example approach.

Typically, most biostatistical texts focus on the principles of design and analysis without extending into the practical use of software to implement these principles. Similarly, R/S-plus texts tend to concentrate on documenting and showcasing the features of R without providing much of a biostatistical account of the principles behind the features or illustrating how these tools can be extended to achieve comprehensive real world analyses. Consequently, many biological students and professionals struggle to translate the theoretical advice into computational outcomes. Although some of these difficulties can be addressed after extensively reading through a number of software references, many of the difficulties remain. The inconsistency and incompatibility between theory texts and software reference texts is mainly the result of differing intentions of the two genres and is a source of great frustration.

The reluctance of biostatistical texts to promote or instruct on any particular statistical software (except for extremely specialized cases where historically only a single dedicated program was available) is in part an acknowledgment of the diversity of software packages available (each of which differs substantially in the range of features offered as well as the user interface and output provided). Furthermore, software upgrades generally involve major alternations to the way in which preexisting tasks are performed and thus being associated with a single software package tends to restrict the longevity and audience of the text. In contrast, although contributors are constantly extending the feature set of R environments, overall the project maintains a consistent user interface. Consequently, there is currently both a need and opportunity for a text that fills the gap between biostatistics texts and software texts, so as to assist biologists with the practical side of performing statistical analysis.

Many biological researchers and students have at one stage or another used one or other of the major biostatistics texts and gained a good understanding of the principles. However, from time to time (and particularly when preparing to generate a new design or analyse a new data set), they require a quick refresher to help remind them of the issues and principles relevant to their current design and/or analysis scenarios. In most cases, they do not need to re-read the more discursive texts and in many cases express a reluctance to invest large amounts of valuable research time doing so. Therefore, there is also a need for a quick reference that summarizes the key concepts of contemporary biostatistics and leads users step-wise through each of the analysis procedures and options. Such a guide would also help users to identify their areas of statistical naivete and enable them to return to a more comprehensive text with a more focused and efficient objective.

Therefore, the intended focus of this book will be to highlight the major concepts, principles and issues in contemporary biostatistics as well as demonstrate how to use R (as a research design, analysis and presentation tool) to complete examples from major biostatistics textbooks. In so doing, this proposed text acknowledges the important role that statistical software and real examples play in reinforcing statistical principles and practices.

Hence in summary, the intentions of the book are three-fold

- (i) To provide very brief refresher summaries of the main concepts, issues and options involved in a range of contemporary biostatistical analyses
- (ii) To provide key guides that steps users through the procedures and options of a range of contemporary biostatistical analyses
- (iii) To provide detailed R scripts and documentation that enable users to perform a range of real worked examples from statistics texts that are popular among biological and environmental scientists

## Worked examples

Where possible and appropriate, this book will make use the same examples that appear in the popular biostatistical texts so as to take advantage of the history and information surrounding those examples as well as any familiarity that users may have with those examples. Having said this however, access to these other texts will not be necessary to get good value out of the materials.

## Website

This book is augmented by a website (<http://www.wiley.com/go/logan/r>) which includes:

- raw data sets and R analysis scripts associated with all worked examples
- the *biology package* that contains many functions utilized in this book
- an R reference card containing links to pages within the book

## Typographical conventions

Throughout this book, all R language objects and functions will be printed in courier (monospaced) typeface. Commands will begin with the standard R command prompt (`<`) and lines continuing on from a previous line will begin with the continuation prompt (`+`). In syntax used within the chapter keys, *dataset* is used as an example and should be replaced by the name of the actual data frame when used. Similarly, all vector names should be replaced by the names used to denote the various variables in your data set.

## Acknowledgements

The inspiration for this book came primarily from Gerry Quinn and Mick Keough towards whom I am both indebted and infuriated (in equal quantities). As authors of a statistical piece themselves, they should know better than to encourage others

to attempt such an undertaking! I also wish to acknowledge the intellectualizing and suggestions of Patrick Baker and Andrew Robinson, the former of whom's regular supply of ideas remains a constant source of material and torment. Countless numbers of students and colleagues have also helped refine the materials and format of this book. As almost all of the worked examples in this book are adapted from the major biostatistical texts, the contributions of these other authors cannot be overstated. Finally, I would like to thank Nat, Kara, Saskia and Anika for your support and tolerance while I wrote this "extremely quite boring book with rid-ic-li-us pictures" (S. Logan, age 7).