S-Plus for the Analysis of Biological Data

Rhondda Jones
Robin Gilliver
Simon Robson
&
Will Edwards

This manual was first produced in 2009. It was made possible by a Teaching and Learning Development Grant from James Cook University. The authors are grateful for the support provided by Dr Nick Szorenyi-Reischl, Director of Teaching and Learning Development.

Contents

Preface

This manual is designed to to teach people to use the statistical software S-Plus and to support the process of learning statistical concepts and methods. It is most useful as a workbook to accompany Whitlock and Schluter's *The Analysis of Biological Data*, published by Roberts & Company, Colorado. Although we include enough statistical background to put the procedures being demonstrated in context, we assume that readers will be acquiring most of their understanding of statistical concepts elsewhere.

Several of the authors of this manual have been teaching introductory biostatistics to undergraduate and postgraduate students on two campuses in Australia for more than a decade (in fact one of us, who would prefer not to be identified, taught a biostatistics course for the first time more than three decades ago). In 2008 we discovered the textbook *The Analysis of Biological Data* (referred to in this manual as ABD). We liked everything about the book: its explanations were beautifully clear and aimed at students much like our own; it used a wide variety of real biological examples; it emphasized concepts and procedures important to biologists and explained how they worked; and it introduced some newer computer-intensive techniques that almost all beginning researchers find themselves needing sooner rather than later. We immediately adopted the book as a text for our own introductory biostatistics course. But this adoption acted as a trigger for making some other changes to our teaching—and in particular, to the way we introduced students to statistical software.

To statistical novices, no statistical software is 'user-friendly', and its use needs to be introduced in a structured way which runs in parallel with their acquisition of statistical understanding. At the same time, teaching effort needs to stay focused on statistics rather than software, so that students do not come to see learning to use the software as their primary goal. This manual is intended to allow users to learn to use the software on their own, while keeping a focus on the concepts and procedures which it supports.

We have followed the ABD approach and layout very closely—indeed, we started out with the intention of simply demonstrating in S-Plus every example used in the body of that text. In the end, because everyone has a slightly different view of what should be included in a first statistics course, we added a number of other examples, mostly using our own data, to demonstrate software capabilities that would not otherwise have been covered.

Why S-Plus?

There are a lot of statistical software options, and most of them will execute all the procedures needed in an introductory course. In choosing a software package, we had four criteria beyond its ability to execute procedures taught in the course:

- **It should have little or no cost to students, and should run on operating systems that students are likely to use on their own machines.** Some of us (OK, one of us) remembered teaching statistics in the days when the only computing aid available to students was a hand calculator (the rest of us at least remember being taught that way). While we did not wish to return to those days, they had one huge advantage—students could work on the material anywhere and any time—not just in computer laboratories provided by the university.

Many of our students are part-time, and some are in remote locations. While we can now reasonably expect that students will have access to a computer at home, we cannot reasonably expect them to buy expensive software for themselves. That meant that if we wanted students to work off-campus, we needed to choose software which was either free or very cheap, or which gave students access on their own machines as part of the university's site licence.

- **It should be useful beyond the course.** We wanted students to use professional-quality software that they would not 'grow out of': providing access to all or most of the techniques they were likely to use throughout their careers; and able to import and export data in a wide range of formats (including text files, databases, spreadsheets, and other statistical software).

- **It should have a very strong graphics capability.** We wanted students to realize as quickly as possible that nothing substitutes for an intimate familiarity with the data they are analysing—and easily-usable graphics allow the data to be explored more quickly and thoroughly than anything else. We wanted the graphics capability to cover the whole range from quick-and-dirty exploratory plots to presentation and publication-quality graphs.

- **It should reinforce the statistical concepts we wanted students to grasp, and not get in the way of learning them.** We wanted to avoid both excessive or inappropriate output, and too much difficulty in using the software itself. Excessive output is often a problem with menu-driven software, which may be relatively easy to use[1], but often provides pages of output that users neither asked for nor know what to do with. Especially for novices, our preference was for software that gives users exactly what they request and offers warnings (or refuses to perform) when what they request is questionable. We believe that someone learning to use statistical procedures should also learn to think about what they are doing and work out exactly what it is they want, rather than making guesses about what button to click in the hope that something useful will happen. On the other hand, if software is too difficult to use, students will inevitably concentrate on learning the mechanics of how to use it rather than developing more fundamental understanding.

In the end we chose S-Plus as the best fit to our needs. That choice committed us to producing this manual: there are some excellent introductory books available for S-Plus, but none that we investigated is targeted at undergraduates who begin as complete statistical novices. S-Plus is very powerful and flexible, has superb editable graphics, and its site licence for universities gave enrolled students permission to use the software on their own computers. It had the additional advantage that it provides both a professional-quality graphical user interface (GUI), and an easily-accessible command language which is very similar to that used by the free open-source software R. Mostly we use the GUI, but we also decided to provide a parallel introduction to the command line and to writing basic scripts. Because the command language of S-Plus and R is so similar, students who use this manual should be able to move fairly painlessly to R by the end of it, if they need to do so.

[1] *Ease of use* is relative. We previously used software whose main selling feature was ease of use. But unless we spent considerable time teaching students to use it, they did not cope well. In our experience, the use of any statistical software needs to be introduced to beginners in a structured way in parallel with their statistical understanding.

How to use the manual

If you are a student using ABD as a text, and you have access to S-Plus, you can use S-Plus to work through each chapter of the manual independently. Every example is demonstrated in enough detail for you to carry it out on your own after reading the ABD chapter and/or covering the statistical concepts in class. You should execute every example yourself to make sure that you can carry out the procedures correctly and get to the right result. A set of exercises is provided at the end of each chapter for you to test your skills. You should make sure that you can do them all—you may require some assistance from your instructor to complete some of them successfully. All the data and scripts required for each chapter are available on the CD provided with the manual.

The first chapter of the manual is a basic introduction to S-Plus, and is one of a few chapters whose content is not linked to ABD. The second chapter introduces you to S-Plus graphics. The remaining chapters can be covered in several different orders, but you need to work through these two first. Not all the material in later chapters will necessarily be included in an introductory course.

In most chapters, we show how to execute statistical procedures using both the GUI and the command line. A few procedures require the use of scripts (short programs written in the S language). Where this is the case, we provide the scripts on the CD—and we show you how to load and use them (but we also encourage you to learn to write your own). In many cases, there are more efficient or elegant ways to write scripts than we have used here—in general, we have tried to produce scripts whose logic can be easily understood by beginners, rather than trying for maximum computational efficiency. Appendix 1 provides a more extensive summary of the S language and S-Plus functions relevant to each chapter of the manual.

You will notice a scattering of these shaded boxes throughout the manual. In general they contain material we think you will need around that point, but which is not immediately essential to the procedure being demonstrated. This is a practical manual, so where there are no shaded boxes, there are wide margins where you should not hesitate to add your own notes.

To the instructor

We believe that learning statistics is like learning to play the piano—there is no substitute for practice. Consequently, in our own teaching, we provide a lot of incentives for students to practice.

In the introductory course that we teach, we expect students to have worked through the appropriate chapter(s) in the manual and attempted the exercises *before* they arrive at the relevant practical class or tutorial—and the first 20 minutes of each 2-hour practical class includes a simple open-book practical test, marked in class, which requires them to analyse some new data using techniques covered in the chapter. (By the end of the course, most students score full marks on most of these tests.) We also run formal (but also open-book) practical exams twice during the course, where the emphasis is on demonstrating that students can make sensible decisions about what to do as well as demonstrating that they can do it. These are also graded immediately. Because students can take this manual—or anything else—into practical tests and exams, we are explicitly *not* testing how well they remember what buttons to press.

When we first changed to this very assessment-oriented approach to the acquisition of practical skills, one unexpected result was that the average grade on the theory exam at the end of the course (which was in the same format and covered the same material as previously) was significantly higher than that achieved by any previous class). Perhaps the development of practical skills really does improve theoretical understanding.

Acknowledgements

As noted above, the structure and content of this manual owes a huge debt to Whitlock and Schluter's text, which provides the best introduction we know of to statistical methods for biology students. We are also very grateful to the students in our 2009 biometrics class, and especially to the practical class tutors (Clwedd Burns, Gavin Coombes, Rie Hagihara, and Philip Newey) whose combined input and feedback improved the manual immensely. Finally, for all his help our thanks to Kris Angelovski of SolutionMetrics Pty Ltd, the Australian distributor of S Plus.

Typesetting conventions used in the manual

To make it easier to use the manual, there are several conventions you should note.

Navigating – The instructions about how to navigate around S-Plus are always given in a particular typeface. For example, the way to navigate from the drop-down menus looks like:

> **Statistics > Regression > Log-linear (Poisson)...**

Similarly, this typeface is used to indicate the parts of a dialogue box you need to change.

Entering new information – Where you are required to enter a name or value in a dialogue box or change an existing name, such as perhaps the name of a column (vector) in a data set, the instruction might look like:

> Right click at the top of the `No.deaths` column, select **Properties**, and enter `Number of cases (Frequency)` in the **Description** box.

`This typeface` represents a name or value that you can either enter or change, whereas **the navigation typeface** cannot be altered.

Coding – You learn in Chapter 1 the significance of single lines of code. When programming, pressing the **Enter** key always means something. Where we have reproduced lines of code that you will see on your computer screen, it has often been necessary to spread them over more than one line because the printed page is narrower than your screen. Where this has happened, the subsequent lines of that instruction have been indented. The following example shows two separate instructions (in a lighter shade) and the screen output.

```
titanic.array = table(titanicAdults$Gender,
   titanicAdults$Outcome)
titanic.array
        Died Survived
Female   109      316
  Male  1329      338
```

Quotation marks when used in code should be straight ("") and not curly (“ ”), sometimes called "smart" quotes. In S-Plus code, curly quotes will generate an error message.

These conventions will become clear as you work through the chapters.