

Using R in an Undergraduate Mathematical Statistics Sequence

William J. Owen¹

Department of Mathematics and Computer Science
University of Richmond, Richmond, VA

Key Words: freeware, probability, programming language, software, statistical inference

Abstract

Mathematical statistics, the two semester calculus-based introduction to probability and statistical inference, lays the mathematical foundation for statistics. Significant computational elements can be integrated into the course, so statistical software is often used to facilitate certain topics. Just as there are many choices for textbooks for the course, there are many choices for software as well. The purpose of this article is to outline my experiences with the software program R, a freeware package for statistical analysis, computation, and graphics. Based on my experiences from using the software for two years in this course, I wrote a guidebook for the software to address many student concerns. This guidebook is discussed, and a link to this manuscript is given at the end of the article.

1. Introduction

In the undergraduate curriculum, mathematical statistics is typically populated by mathematics majors/minors (or statistics majors if the program exists). This year-long sequence covers the theory of *probability* (first part) and *statistical inference* (second part). In probability, the fundamentals of random variables, distribution theory, expectation, simulation, and limit theorems are developed. Statistical inference entails such data analytic methods as estimation, hypothesis testing, and Bayesian methods. A key feature of this advanced course is that no knowledge of statistics is assumed prior to the start – the topics are built on a mathematical foundation of (usually) three semesters of calculus and some linear algebra. While it is possible to cover the course topics without the use of the computer (for years it was done so), a more modern approach integrates technology into the course to complement the subject matter. This is particularly important if this course turns out to be a student's only exposure to statistics. Instructors have many software options; these include computer algebra systems (CAS) such as *Maple* and *Mathematica* (students may have seen these used in their calculus courses), statistical packages such as Minitab, SPSS, and SAS, or even a spreadsheet program like Excel. The choice of which program to use is often based on 1) pragmatic reasons like the available university software site licenses and computer lab availability and 2) preferential reasons of the instructor.

If a site-licensed software program is used, there is no cost to the student. However, it is important that computer labs are widely available and documentation is accessible. If students

¹ email: wowen@richmond.edu

want to run the software on their own computer, there are a few options to consider. Some universities allow students to obtain copies – that expire after a short period of time – of site-licensed software for a nominal fee (or free in some cases). But, documentation is often not provided. Many software companies produce *student editions* of their popular software titles, and these can be “bundled” with the textbook. However, these versions usually lack the full features of the professional editions and/or include little documentation. If the professional version of a software title is desired, it may be possible for students to purchase temporary licenses (usually on the order of 4 to 12 months) to essentially “rent” the software from the manufacturer. Either of these options can often add a considerable cost to the textbook price, and not all operating systems may be supported.

There are some no-cost alternatives to purchasing statistical software. StatCrunch (West, Wu, and Heydt 2004), maintained by Integrated Analytics LLC, is a Java-based program for data analysis for the World Wide Web (WWW). The program runs in a Java-capable Web browser, and extensive online help is available. Other alternatives include XLISP-STAT (Tierney 1990), a statistical computing environment based on the Lisp language, and MacANOVA (downloaded at <http://www.stat.umn.edu/macanova/>). Another freeware choice is the program R (Ihaka and Gentleman 1996), and this is the subject matter of this article. I have used this program for three years in a mathematical statistics sequence, beginning in the Fall 2002 school year. For the first two years, I provided numerous handouts and online instruction material for using R; however, it was often the case that handouts would get tattered and lost, and any online material, unless printed out, would be referenced and later forgotten. Prior to the third year, I wrote a handbook called *The R Guide* (henceforth referred to as the “Guide”) that incorporated the fundamentals of the software that I wanted my students to be familiar with. This paper is about the Guide, and a WWW link to obtaining the Guide is given at the end of the paper. The remainder of the article is organized as follows: Section 2 introduces and describes the software, and in Section 3 some benefits of its use are illustrated. In Section 4, I outline my experiences with using the software and I address many of the obstacles that have been expressed by my students. In Section 5, the Guide is described in detail and Section 6 gives a brief assessment of other reference material on R that is available. Section 7 gives the conclusions.

2. What is R?

R is an integrated suite of software facilities for data manipulation, simulation, calculation and graphical display of data. It is an independent, open-source, and free implementation of the S programming language. This language, which was written in the mid-1970s, was a product of Bell Labs (of AT&T and now Lucent Technologies) and was originally a program for the Unix operating system. R is available in Windows and Macintosh versions, as well as in various flavors of Unix and Linux. A commercial product called S-PLUS is distributed by the Insightful Corporation. Although there are some differences between R and S-PLUS (mostly in the graphical user interface), they are essentially identical. The language manages and analyzes data very effectively, and it contains a suite of operators for calculations on arrays and matrices. It has impressive graphical capabilities for very sophisticated graphs and data displays. In addition, it can be used as an effective object-oriented programming language. The software itself comes with several manuals (both introductory and advanced) in electronic form, and the

program contains detailed help files for functions and a search engine to find functions apropos to a search term.

Over the last several years, R has gained a lot of attention as a research tool. Several texts have recently been published on the use of R in many different areas of statistics. In addition, users have contributed numerous packages and libraries to augment the capabilities of R. A quick glance at recent issues of the *Journal of Statistical Software* (<http://www.jstatsoft.org>) illustrates the attention and focus on R in the research arena: it is clear that R has become an important platform for new statistical algorithm development.

3. Why R?

In this section, I highlight my reasons for using R in mathematical statistics, although I admit that some could be considered as subjective. Two recent articles have explored its use from other perspectives: Horton, Brown, and Qian (2004) view R as a platform to explore statistical concepts and support the theory in mathematical statistics – both at the graduate and undergraduate level; the article by Hodgess (2004) focuses on its use in an undergraduate course on time series and compares features in R to those in Minitab. These articles provide an excellent exposition on R and the authors give validations for using R in college teaching. To add to those, I emphasize some key points relevant to this article below:

3.1 Cost and Availability

For the many choices of software available, the ultimate decision boils down to cost and availability. As mentioned previously, R is free and available on all computer platforms. It can be installed on campus computers, networks, and personal computers. In addition, it is not necessary for the computer to be connected to the internet when using the software.

3.2 A Powerful Toolbox

R is more than simply a program for calculating statistics and data management. There are several unique features in R that students can utilize, not only in mathematical statistics but also in other classes and later in life:

- Its use as a calculator. From the command line, complicated expressions can be evaluated. These include terms with exponents, trigonometric and transcendental functions, sums, and products. In addition, functions for vector and complex arithmetic forms are included. This is extremely useful for quick calculations for a variety of situations.
- Outstanding graphics. The graphical abilities in R are astounding and are publication quality. Early on, I have students to view the graphics demos included with the R software and give simple assignments of graphing common functions.

- R is a programming language. For students who are interested in computer science, this is a powerful feature; in addition, the possibility to integrate compiled code written in other languages, like C and Fortran, is extremely valuable.

3.3 Other promotional features

Lastly, there are “selling points” of the program that I try to impress on students during the semester:

- A marketable skill. Although it does take some time to learn the software, the computing skills that are learned can be “put on the resume” as an acquired skill. Furthermore, I view it as a terrific launch for students interested in pursuing graduate studies in statistics.
- Fewer mistakes. It is clear that R requires a higher degree of perceptiveness than other competing point-and-click software programs. However, I have found that this requires students to be more aware of what they are doing – and they often attain a superior understanding of the course material
- Undergraduate research projects. Many universities offer summer research programs for undergraduates. When I work with a student on a research subject, computational power is usually required, and my software choice for research is also R. By integrating R into the curriculum, students already have an introduction (granted, this is a subjective reason).

4. Common Obstacles for Students

Any computer program has a learning curve, and most users admit that R can be slow at the onset. The graphical user interface is rudimentary and those students with little computer programming experience can feel somewhat overwhelmed by the unadorned “>” prompt. During the first two years of using the software, there were some reoccurring struggles that students had with the software. In no particular order, they were

- 1) not very user friendly and no menu for routines
- 2) too many functions to “memorize”
- 3) confusion with data entry and reading in data from other sources
- 4) understanding and knowing when to use each data construct

Students who have used other statistical software, especially menu-driven packages like Minitab or SPSS, were prone to mention 1). For issue 2), while it is true that the number of functions that are used during the semester is large (probably near 100), it is important for students to understand that it is not necessary to memorize *every* function – admittedly, I refer the manuals routinely or at least a reference card like the excellent example written by [Tom Short](#). Issues 3) and 4) tend to be problems that appear more often in the second semester of the course – when data is analyzed – especially if an external data file is required by R for a homework problem. To avoid these issues listed above, it is feasible that some instructors do not consider using R and

they choose another program. However, because of my high regard for the software, these obstacles were my motivation for the construction of the Guide. It is described next.

5. Features of the Guide

The Guide is designed for users to start learning the software without any knowledge of statistics, as it is assumed in the mathematical statistics sequence. Later, when statistical methods are covered, no notation is included since it is assumed the reader can reference his or her textbook. This avoids any conflicts among differences in textbooks. The Guide is written in a manner that those students with little or no computer programming experience should find it and the software more user friendly. To keep things less complicated, I do not include any discussions on functions or libraries that are not included with the base software, even though there are a great deal of available routines that could be considered relevant to “mathematical statistics”; see, for example the library included in Venables and Ripley (2002). In addition, to keep the Guide a reasonable length, some details on functions are omitted and the reader is referred to the corresponding help files.

The Guide is divided into eight chapters, is fifty pages in length, and in the rear is an index for quick function reference. At my university, I have Campus Printing Services professionally bind the Guide in a notebook with a hard cover, and I require every student to purchase it at the bookstore for the printing cost of less than \$10. Although this is not necessary, I find that students prefer a hard copy to reference while using the software and the bound document is more secure than loose pages.

The key features of the Guide are:

5.1 A kick-start approach

Students begin learning the fundamentals of the software independently of the course material. The first four chapters cover the basics of algebraic and matrix computations, data types and data entry, functions, and graphics, so at a very early stage students can perform many operations they have already seen in other mathematics courses. I require my students to read these beginning chapters during the first few weeks of the course, and I assign exercises for homework to ensure they are prepared for the future (statistical) material in R.

5.2 Simple examples and exercises are included

Not only are several examples included that demonstrate functions throughout the Guide, I also give small sets exercises at the end of most chapters. They are usually quite simple and are in place for either assignment or student practice.

5.3 A chapter on data entry and data frames

I devoted an entire chapter to the subject of data entry – including entering from the keyboard, accessing R’s internal datasets and reading in data from external sources. In addition, since most

datasets are multivariate in nature and stored as in R as *data frames*, this is discussed here as well.

R contains functions for reading in text files, e.g. `scan()` and `read.table()`; the filename is specified in the function but this can get problematic when the path is lengthy. To simplify this, files can be searched interactively by using the `file.choose()` function within `scan()` or `read.table()`, and this very useful procedure is detailed in the Guide.

5.4 Inclusion of additional material

As well as the software instruction, I include a few important topics that are often not included in textbooks on mathematical statistics. For example, information on bin selection methods used for histograms is given in the graphical descriptions section in Chapter 5. Also, I include the simulation application of Monte Carlo integration in Chapter 6. I spend a fair amount of time on simulation in my course, and I find that many students find this subject to be a clever application (since students recall the other approximation methods from integral calculus). The Guide ends with a chapter on control flow and function writing that can be referenced by students with programming experience. For flexibility, any of these topics can either be assigned for reading or skipped entirely.

6. Alternate sources of reference

There are other resources (some free, some not) that can be considered for a companion text on R. None are designed specifically for mathematical statistics per se, but they contain relevant themes and topics. For a cost, three options (all in paperback form) are

- *Introductory Statistics with R* by Peter Dalgaard, Springer-Verlag, New York, 2002 (267 pages, \$47.95).
- *Using R for Introductory Statistics* by John Verzani, Chapman and Hall, New York, 2004 (432 pages, \$44.95).
- *Statistics: An Introduction Using R* by Michael Crawley, John Wiley and Sons, New York, 2005 (342 pages, \$40.00).

Note that most of these books are intended for introductory statistics courses, but each contains more applied material than is usually seen in introductory statistics – *i.e.* multiple comparisons, logistic regression, ANCOVA, etc. These are all well-written, self-contained texts and include exercises for the reader.

Other users of R have written short handbooks for the software and have made them freely available on the CRAN website (<http://cran.us.r-project.org>) in the “contributed documentation” section. Some are specific in their statistical focus (*e.g.* linear models, distribution fitting, or time series), and others are focused on specific features of the program (*e.g.* compiling source code). One choice that gives a broad overview of the software is

- *R for Beginners* by Emmanuel Paradis, 2005 (72 pages).

This reference thoroughly describes objects and data types in R and contains an excellent chapter on graphical commands. A short chapter on statistical analyses demonstrates a simple ANOVA model fit, so this publication is probably best utilized by readers with some statistical experience. Other free resources that are quite valuable are the “short reference cards” – like the one mentioned in Chapter 4 – that are useful for quick function reference. Aside from these mentioned, other reference material written by users can surely be found online via search engines.

7. Conclusions

Over the last decade, R has been shown to be an exciting vehicle for both statistical analysis and research. Certainly a student who has an interest in graduate study in statistics can benefit from an early introduction to the software. However, despite the significant help files and online support, it can be difficult to use the software at the undergraduate level without some supporting documentation. The Guide is designed to complement the material in a typical mathematical statistics sequence, and most topics found in common texts like Hogg and Tanis (2006) or Wackerly et al. (2002) are included from the computational standpoint. Since using the Guide in my class, I have observed an improvement in my students’ proficiency with the software. In the spirit of R being a free software product, I make the Guide freely available by accessing one of the links below.

Personal website: <http://www.mathcs.richmond.edu/~wowen/TheRGuide.pdf>

CRAN: <http://cran.us.r-project.org/other-docs.html>

References

Hodgess, E. M. (2004), “A Computer Evolution in Teaching Undergraduate Time Series,” *Journal of Statistics Education*, 12, Number 3 [online].

Hogg, R. V. and Tanis, E. A. (2006), *Probability and Statistical Inference*, 7th Edition. Prentice Hall, New York.

Horton, N. J., Brown, E. R., and Quin, L. (2004), “Use of R as a Toolbox for Mathematical Statistics Exploration,” *The American Statistician*, 58, 343-357.

Ihaka, R., and Gentleman, R. (1996), “R: A Language for Data Analysis and Graphics,” *Journal of Computational and Graphical Statistics*, 5, 299-314.

Tierney L (1990). *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.

Venables, W.N. and Ripley, B.D. (2002), *Modern Applied Statistics with S, 4th Edition*. Springer-Verlag, New York.

Wackerly, D. D., Mendenhall, W., and Scheaffer, R. L. (2002). *Mathematical Statistics with Applications, 6th Edition*, Thomson Learning, CA.

West, R, Wu, Y. and Heydt, D. (2004), “An Introduction to StatCrunch 3.0,” *Journal of Statistical Software*, 9, Issue 5 [online].