R for undergraduate statisticians

Gonzlez-Arteaga, Teresa
University of Valladolid, Department of Statistic and Operational Research
Prado de la Magdalena s/n
47011 Valladolid, Spain
E-mail: teresag@eio.uva.es

Introduction

A relevant question is how to integrate education in computing into the statistics curriculum. The answer naturally changes with technology, different types of computer usage and over time. It is useful to look into how computing is taught in statistics programs. We find different programs with alternative approaches. There are those who consider that "statistical computing need not be relegated to courses devoted to statistical computing" (Hunter 2005) or "I have never believed that courses in statistical computing that just cover packages have a place in the statistics curriculum" (Gentle 2004). However, others think that "an undergraduate course covering programming and data management prepares students for most statistics courses" (Monahan 2004) and "we strongly endorse a course in programming and statistical computing with a heavy mix of exploratory data analysis and modern statistical methods" (Nolan and Temple 2010). Although the dilemma is bound to continue, it is beyond question that computing is an essential part of education in statistics. What statisticians are doing involves a significant computational component. Finally, we agree with Nolan and Temple 2010 when they conclude that explicit courses on computing must be introduced, but we also need to explicitly include computing as part of existing courses and treat the computational aspect as an important element.

We must clarify that the term statistical computing is used here as in the proposed approach of Nolan and Temple 2010 and it should not be confused with computational statistics. The latter includes topics like numerical optimization, Monte Carlo studies, MCMC, and so on. The term statistical computing involves the ability to reason about computational resources, work with large data sets and perform computationally intensive tasks.

It is possible to broadly justify why R should be taught to undergraduate students of Statistics, but perhaps this is not necessary nowadays. The R statistical computing environment and graphics (R Development Core Team, 2010) has already become an important part of statistical training. We therefore only point out some of its main features. It is well known that R is a free software, released under the GNU General Public License and that it has been developed and maintained by a strong team of renowned researchers in computational disciplines. Besides, many additional contributed packages are avalaible for a great number of purposes from the Comprehensive R Archive Network (CRAN http://cran.r-project.org/), while the needs of any user can be catered for by writing new packages. The concept of packages as extensions to the R base system is one of its greatest strengths.

R is a high-level programming environment which provides both command-line and graphical interfaces. It allows for interactive data analysis as well as scripting in order to process large amounts of data. R supports modern concepts like object-oriented programming. One of the outstanding features of R is that it provides a broad toolkit of statistical plots and allows for easy modification and very fine control over the appearance of plots.

The remainder of this article is devoted to outlining a computing course with R for undergraduate statisticians from my own teaching experience. This course is followed earlier on the program to facilitate the instruction of the upper division courses in sampling, regression, and so on.

A course in statistical computing

Our approach has been shaped by our experience teaching in a course called "Statistical Computing" at Valladolid University. Our students are second year undergraduates in statistics. Class size has been small and has always been taught in a computer lab. Students' computational and mathematical skills are variable and elementary and they hardly have previous experience with data. These students are supposed to learn more advanced statistical methods later and to perform applied statistics in the real world.
The main objective is to develop computing skills that are important for statistical practice. Although it is difficult to establish what exactly should be taught, some considerations are kept in mind. It is necessary to address statistical and computational reasoning in addition to the fundamentals of programming with data.
The overall structure of the course was designed to acquire a command of the R computing environment. We want our students to appreciate the benefits of R's command language for the practice of statistics, not only the effort and time needed to acquire it. Therefore, we put a lot of emphasis on the easiness of reproduce analysis and we consider it is essential for them to get used to managing the online and help resources available for R.
In spite of the fact that we do not explicitly teach specific statistical methods we try students to gain an insight into important concepts of probability and statistical inference by doing simulations. We also discuss the exploration of real data and see some statistical modeling in a heuristic manner rather than with formal theory. This allows us to focus on the computational ideas and on tools that are useful for many data analyse. In this way, the course is instrumental in the development of subsequent advanced statistical methods.

Topics

The choice of what topics to include was made considering our general objectives. More specifically, we are interested in the combination of statistical reasoning programming, visualization and reporting. Next we provide a succinct list of topics for the course that we teach.

Introduction to the R environment
We begin with an overview of the R software and environment, computing language, how to download and install R, contributed packages, finding help within.

Data structures and objects manipulation
We include handling vectors, factors, matrices, arrays, data frames, lists and usual functions. Issues such as subbsetting and vectorized operations are highlighted.

Visualization and graphics
We look at functions for traditional graphics and for lattice graphics. Basic plot types and graphics parameters are considered in addition to high-level graphics functions to produce complete plots and low-level graphics functions to add further drawing and annotation to a plot.

Programmning
This topic contains control flow, writing functions and debugging.

Digital input - output
We discuss the use of R to read and write files of different types of data. The record of plots in files and the redirection of the output to a file are considered as well.

Reporting tools

A short introduction is given to the command Sweave and the package R2HTML.

We must point out that we do not follow this order strictly, talking intermittently about the different topics while increasing the difficulty and revisiting them in greater detail. We encounter these topics in the context of exploring real data and doing simulations with the summaries and creative graphical displays that involved. So the computing topics are introduced by using them in a practical setting.
The use of R to generate graphs is covered in greater length than in typical courses of R because our goals are better achieved in this way. R has a very rich visualization environment that produces high-quality statistical plots. Murrell 2009 describes generally the graphics facilities of R. Briefly, there are three graphics systems: base graphics system, trellis graphics and grid graphics. However, we only address the graphics package and the lattice package, Murrell 2005, Sarkar 2007. This allows stydebts to be aware of the highly customizable R system.
Reporting tools in R have become widespread more recently (Leisch 2002, Kuhn 2006) and they have seldom been covered, despite being essential for statistical practice. Those tools take the idea of reproducing the analysis one stage further. Since R is a programming language, it can be used to write scripts which can then be run at any time to repeat the analysis. This is especially useful if the data changes or to apply the same analysis to a different data set. A further step is to integrate the R analysis into a text document. Rather than cutting and pasting, a better approach is to prepare the text document with a text editor such as HTML, OpenOffice or LATEX and incorporate executable R code into the document. This has been implemented for R through the Sweave function and some contributed packages (R2HTML, odfWeave).

Teaching methodology

As for how to teach statistical computing, we must remark that it is quite different from more traditional statistical topics. First of all, simply training students how to modify templates must be avoided. We need to teach how to think and reason about computing and express statistical tasks as computations.
The best way that the students learn a programming language is by actually using the language on problem sets. Short lectures are useful to introduce the concepts, but the exercises and projects are the main part of the class so that the students can gain the experience and knowledge required. They learn by trying things and figuring out what went wrong and why things did not work. Then, active learning arises naturally.
As learning R is an ongoing process, we encouraged our students to explore the numerous online resources. Various sites provide books, documentation, code and email lists. I recommend students to complement the lecture notes with a short list of books, such as Dalgaard 2008, Maindonald and Braun 2007, Rizzo 2008, Venables 2011, Vezani 2005. Although some of these are quite advanced they serve as a reference. They are essential as, after all, students have to learn to manage on their own. All course materials are available online so that students can use them in a variety of ways and at their own pace.

Final remarks

In this article we have shared our experience on the introduction of a course in statistical computing before integrating computing into traditional classes in our curricula in order to train undergraduates to be better prepared for modern data analysis.
Our course focuses on teaching the students how to develop new code rather than just calling up existing functions. At the same time, we improve their exploratory data analysis skills and intuition and teach them how to think computationally. These are critical competences for the applied statistician.

Finally, since we rely on learning by doing. we practice problem-solving based teaching. There is a relatively steep learning curve when you begin to use R, but it gets easier over time with practice. We have found there are some common problems that we should warn our students of, along with suggestions for helping them to avoid frustration. In the end, most people felt they liked R.

REFERENCES (RÉFERENCES)

Dalgaard, P. 2008. "Intoductory statistics with R". Springer.

Gentle J.E. 2004. Courses in Statistical Computing and Computational Statistics. The American Statistician, February 2004, Vol. 58, No. 1, 2 - 5.

Hunter D.R. 2005. Teaching Computing in Statistical Theory Courses. The American Statistician, November 2005, Vol. 59, No. 4, 327 - 333

Kuhn M. 2006. Sweave and the open document format - the odfWeave package. R News, 6(4):2-8, October 2006

Leisch F. 2002. Sweave, Part I: Mixing R and LaTeX: A short introduction to the Sweave file format and corresponding R functions. R News 2 (3): pp. 2831.

Maindonald, J. and Braun,J. 2007. "Data analysis and graphics using R. An exampled based approach". Cambridge University press.

Monahan J. 2004. Teaching Statistical Computing at North Carolina State University. The American Statistician, February 2004, Vol. 58, No. 1, 6 - 8.

Murrell P. 2005. R Graphics. Chapman and Hall/CRC.

Murrell P. 2009. R Graphics. Wiley Interdisciplinary Reviews: Computational Statistics, Volume 1, Issue 2.

Nolan D. and Temple Lang D. 2010. Computing in the Statistics Curricula. The American Statistician, May 2010, Vol. 64, No. 2, 97 - 107.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria3-900051-07-0. URL http://www.R-project.org.

Rizzo M.L. 2008. Statistical Computing with R. Chapman and Hall/CRC

Sarkar D. 2007. Lattice Multivariate Data Visualization with R. Springer. 2nd edition.

Venables W.N., Smith D.M. and the R Development Core Team 2011. An Introduction to R. URL http://www.R-project.org.

Vezani J. 2005 Using R for introductory Statistics. Chapman and Hall/CRC

RÉSUMÉ (ABSTRACT)

The R statistical computing environment has already become a very important part of statistical training. The software is constantly evolving and R is no exception. R allows a wide variety of statistical analysis and visualization of great sophistication. Therefore, we must address our attention to the real important issues if we want to succeed in preparing our students. We need to have clearly defined our objectives. We recognize the students need to be more computationally capable and literate. We are interested in the combination of programming, visualization, statistical reasoning and reporting.

The question is what a first course about R for future statisticians must contain. These students are going to use R to learn advanced statistical methods and to perform applied statistics in the real world. Then, we must go beyond an introductory statistics course with R for practitioners of other disciplines. We want our students to appreciate the benefits of R's command language for the practice of statistics and, not only the effort and time consuming required. Learning to use R is time well spent due to R remains at the cutting edge of statistical computing for quite some time. We need to rethink which topics to include and, what is even more important, how to teach them. The aim of this presentation is to think about all this carefully and to share my teaching experience.