TEACHING WITH R – A CURSE OR A BLESSING?

Dulce Gomes and Bruno de Sousa
University of Coimbra, Faculty of Psychology and Education Sciences, CINEICC
bruno.desousa@fpce.uc.pt

Although the advantages of R are well-known (free, open source, continually updated by experts), it is not the first choice among college students, especially those not majoring in mathematics or statistics. A problem that appears when teaching R is that once the great potential of the software is understood, the temptation is to focus immediately on more advanced analysis, which adds frustration for beginning learners of R. Our online module begins with basic visualizations and summaries of data, followed by how mathematical functions can be used to model data, and concludes with some inferential statistics, such as linear regression. This approach will not only help R beginners, but can also be used by school teachers in a global plan to bring R to the youngest learners.

INTRODUCTION

Teaching R to both undergraduate and graduate students in Mathematics or Social Sciences, in our experience is rarely a stress-free task as it can very easily become a burden for both teachers and students.

Many of our students begin a course in Statistics with little or no knowledge in programming languages and have never heard of the R software. R is often, at a first approach, not very well accepted by learners, including those who already have some programming skills in other languages. One of the possible reasons for this resistance to R, whether the learners are undergraduate or graduate students or even other teachers from a variety of areas of science, is the fact that, teachers tend to emphasize a wide range of commands and programming lines from very early on, making learning R a slow and frustrating task.

Learning R may be seen as similar to learning a foreign language, and if talent and immersion are two attributes that matter less when learning a language, as Chris Lonsdale states in his TEDx talk (2013), then a student with no skills or practice in statistics or programming languages should be able to learn R. Relevance, tolerance of ambiguity, and comprehension are some of the principles (Lonsdale, 2013; Krashen, 2017) that might guarantee success in learning R.

This work will not only present some of the experiences in teaching R to graduate and undergraduate students from different backgrounds, but will also report the results of a pilot study with high school students. By including this latter group, the main objective was to determine students' receptivity to learning R and what motivates young learners when working in R. Based on Linnenbrink and Pintrich (2003), who state that students are motivated when they feel excited about a task or feel that their activity is worthwhile, some thoughts will be given on the differences and similarities experienced in these 3 groups of students.

Finally, the design of an online module will be proposed, intended not only for those beginning to use R, but also for school teachers in order to bring R to the youngest learners. Our goal is to offer a proposal to overcome some of the difficulties encountered when learning R and to extend the potential use to R to learners of a younger age.

OUR STUDENTS

Three of the five principles proposed by Chris Lonsdale (2013) were selected as they were a direct reflection of our students' experiences. As mentioned above, they are: relevance in terms of how the students viewed the concepts taught as significant; tolerance of ambiguity, in the students' acceptance that time to be creative and constructive when learning R is needed before clarity is achieved; and comprehension, meaning that a student only becomes an R practitioner when he/she is able to understand how R language works. Although these 3 principles were present in the three types of student groups studied, their intensity varied accordingly to each group.

For the first year undergraduate students, R was introduced during a 15-week Statistics class given in Science and Technology programs such as Chemistry, Biotechnology, and Biochemistry, among others in which over 90% of the students had never had any contact with a

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.*
Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

programming language. The main difficulties encountered in this group were the motivation and relevance to their field of study and their future careers. Most of the students were reluctant learners. Reluctant learners are content with just getting by. One common thread among reluctant learners is their perception of themselves, known as self-efficacy (Sanacore, 2008). They simply avoid doing any of the proposed tasks and continue to think that the statistics is not going to be useful, admitting "This does not interest me very much." or "What I would prefer is to start working in my field of study."

For graduate students, the problematic issues focused more on their comprehension of the R language. A total of 20 students from different PhD programs in Psychology, Medicine and Education Sciences were part of an 8 hour workshop entitled "Good (And Not So Good) Statistical Practice: Common Errors in Quantitative Research and How to Avoid Them" where an introduction to R was part of the program. Relevance to their research area was not an issue, but impatience to becoming comfortable with the program was very present, especially when most of them work with more user-friendly statistical software. Another issue that arose among PhD students was the individuals' varying degree of knowledge, both technological and scientific, and in particular their overall knowledge of certain statistical principles. When faced with a very easy task, they tended to devalue or even feel devalued, and when the task was more difficult, only the most technologically capable and the most knowledgeable in statistics felt confident in trying to complete the tasks.

In addition to college students, a pilot study with high school kids was carried out where RStudio was presented to 11th graders during a 2-day program. The activities were performed in two classes, one with 20 students and the other with 12 students. The students who participated in this pilot study attend a school in Mértola, Portugal, one of the most depopulated regions in Alentejo, in the interior of country. This school was selected based on the high level of motivation demonstrated by teachers and students in welcoming such an approach to statistics. With some exceptions, these students offered no resistance to programing in R, and thus they gained an excellent understanding of how R works and appreciated the potential of such a tool in their math classes. The lack of enthusiasm for R and the reluctance to carry out the tasks shown initially by a few students quickly disappeared during the course of the training.

Creating activities that students enjoy can positively affect their motivation, but it is only one of the many challenges that teachers face. In addition, teachers need to be able to transmit rigorous scientific knowledge (in many cases within a very extensive list of topics) and to prepare students for their final exams, leaving very little time to go beyond the curricula which covers the material for these exams, and finally motivating students to learn. Only that way will the knowledge acquired be retained and useful for the great challenge faced by students when entering the unpredictable labor market.

To overcome some of the difficulties and challenges encountered in these three different groups of students, in the next section we will present the design of an online module for learning statistics with R that is focused on reasoning, calculations and critical thinking, but also with special attention given to applying the theoretical knowledge to real situations and the reality of everyday life. The advancement and the relevance of statistics in the world today is undeniable, but more important is the recognition that informed knowledge of a country's statistics and statistical literacy will improve the promotion and participation of young people and adults in building countries that are more inclusive and socially fair. This notion must be urgently transmitted to the students as quickly as possible during their academic careers. Thus, by proposing this online module through social-networks such as Facebook, these issues will be addressed within a student-centered learning environment, making use of technology to motivate students to learn statistics and to retain the knowledge learnt, rather than simply memorizing material for an exam.

AN ONLINE MODULE DESIGN FOR TEACHING STATISTICS WITH R

With Facebook being one of the most popular social media platforms, with almost 90% of individuals under the age of 30 using it (numbers for USA from the Pew Research Center Report, 2016), it became the natural choice for our proposed online module of teaching statistics with R.

Our approach addressed the 3 principles of Relevance, Tolerance of Ambiguity, and Comprehension by creating the following modules.

Module 1 addressed some introductory mathematical functions and how they can be used when modeling data, allowing students to feel comfortable with the new language, especially with the creation and the manipulation of objects. These more mechanical procedures are introduced in the context of the manipulation of a dataset, allowing them to create subgroups of data regarding (for example men and women, diseased versus healthy subjects), or to separate a database by different classes in a school. At this point, students start to see the relevance of the more unexciting mathematical operations used to manipulate data.

Module 2 followed with some basic visualizations and summaries of data. The HELP search is explored in order to expose the students to the richness of the R language and how it is structured. The dimension of the more than 10,000 packages is presented and how the idea of fully understanding all its functionalities can easily be seen as a 'mission impossible'. The goal here is to encourage the students to be tolerant of ambiguity, stressing that clarity will come once they know what and how to get the functions needed for their research goals.

Module 3 was dedicated to inferential statistics. When students start to use R in modules 1 and 2 for their calculations, avoiding having to resort to the traditional pen, pencil and calculator tools, they gain some free time to understand and interpret statistical concepts as well as to explore their own datasets.

These modules were implemented on a Facebook page where different groups were created to upload the different activities proposed for the 3 modules. Using Facebook created a sense of virtual community, one in which students and teachers are connected and can participate via ongoing online discussions.

The following three different activities were proposed to our groups of students:

Sample Activity 1: For PhD students

Through the two hours of introduction to R, students were able to manipulate their databases and to perform some visualizations. The activity was focused on how to use the Scatter Plot Matrix in order to visualize the form of the distribution of the different variables and the relationship between two variables from a dataset. In addition, regression models were performed, separating the results by sex. All the graphs were saved in a format ready for publication.

Sample Activity 2: For first year undergraduate students

During the 15-week course, and as part of their evaluation, students had to do two practical assignments consisting of writing an R script and a report. The first assignment consisted in the idealization of a study in which they would have to simulate 4 distinct variables: a continuous one following a normal distribution; a discrete one; a nominal one with two categories; and finally an ordinal variable with 4 categories. Notice that a story behind these 4 variables needed to be created. After the preparation of their databases, unique to each of the 285 students enrolled in the course, it was next analyzed in a univariate and bivariate descriptive perspective. For the second assignment, statistical inference techniques needed to be applied, namely confidence intervals and parametric and non-parametric hypothesis tests (for one parameter and for comparison of parameters which included one-way ANOVA), and linear regression.

Sample Activity 3: For high school students

The activity proposed was divided into two parts. First, an introduction to R was given where students were introduced to simple mathematical calculations, manipulations of vectors and matrices, construction of frequency tables, several types of graphs, and the calculation of statistical measures. In the second part, an activity entitled *C.S.I. Mértola* was done in which each team of two students had to respond to several statistical challenges in order to discover a fictional killer. In the 5 different stages of the investigations students needed to sort a list of 26 suspects based on statements such as "Evaluating all the clues left by the murderer, we conclude that the suspect cannot have an height less than or equal to the first quartile; not more than the third quartile; and not higher than the 90th percentile" or "After performing a linear regression of the weight on the height of an individual, you should eliminate as possible suspects all the individuals that are more discordant among the suspects." Passing successfully all the 5 stages of the investigation should lead students to identify the murderer.

FINAL COMMENTS

While high school students found it very easy to start working with R, their motivation to continue working on it is a challenge for teachers. Changing the colors of the graphs or the orientation of a boxplot was not enough to keep students engaged. Preparing activities such as *C.S,I. Mértola* is time-consuming and requires commitment from teachers.

For 1st year college students and PhD students, a clear increase was seen in terms of difficulty understanding R. Frustration with the R language was more present in the latter, but in both cases understanding the lines of commands and interpreting the results were particularly difficult.

The approach of beginning to learn R through simple mathematical operations and manipulation of vectors and matrices, followed by simple descriptions of data and visualizations and ending with inferential statistics, allows teachers to adjust the different activities that need to be proposed in order for students to be able to achieve the 3 main principles of learning R: relevance, tolerance of ambiguity, and comprehension. Our experience tells us that working on tolerance of ambiguity is a main issue for older students, while relevance is a key factor for the younger ones.

Introducing our pilot study to high school students served to encourage greater implementation of the use of free software, namely R, in the pedagogical practices of teachers and in their daily life. Presenting it in a Facebook environment contributed to promoting greater student participation in the learning process, one that is more focused not only on their autonomous work, but also on the development of their abilities to work as a team.

Starting this process even at an early stage in school will most certainly contribute not only to the improvement of teaching methodologies, but also to the promotion of statistical literacy among students and teachers.

ACKNOWLEDGMENTS

Dulce Gomes has been partially supported by Research Centre for Mathematics and Applications, UID/MAT/04674/2013, University of Évora, supported by FCT- Foundation for Science and Technology, Portugal.

Bruno de Sousa has been partially supported by CINEICC, Cognitive and Behavioural Center for Research and Intervention, UID/PSI/00730/2014, University of Coimbra, supported by FCT- Foundation for Science and Technology, Portugal.

We would also like to thank the school (*Agrupamento de Escolas de Mértola*), the teachers (Emília Rocha and Madalena Rodrigues) and the students who participated in the present study.

REFERENCES

Granito, M., & Chernobilsky, E. (2012). The Effect of Technology on a Student's Motivation and Knowledge Retention. *NERA Conference Proceedings, 17*. http://digitalcommons.uconn.edu/nera_2012/17

Krashen, S. (2017). The Case for Comprehensible Input. *Language Magazine*, http://www.sdkrashen.com/content/articles/case_for_comprehensible_input.pdf

Linnebrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly*, *19*(2), 119-137. DOI: 10.1080/10573560390143076

Lonsdale, C. (2013). How to learn any language in six months. TEDx Lingnan University on YouTube, https://www.youtube.com/watch?v=d0yGdNEWdn0

Pew Research Center (2016). Social Media Update 2016, http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Sanacore, J. (2008). Turning Reluctant Learners into Inspired Learners. Clearing House. *A Journal of Educational Strategies, Issues and Ideas, 82*(1), 40-44.