YOU R, SO YOU ARE !

Koo J. Rijpkema
Department of Mathematics and Computer Science
Eindhoven University of Technology, Eindhoven, The Netherlands
j.j.m.rijpkema@tue.nl

Until recently, commercially available software packages for statistical analysis were the norm. However, a real revolution is taking place nowadays: software R, freely available and supported by an active community is becoming increasingly important. This cannot remain without consequences for our way of teaching, as other skills will be needed and other techniques come within reach. To prepare our engineering students for this we have integrated the use of R and R Commander in our service courses Statistics and Design of Experiments for several years now. From the experiences gained we have developed best practices for deploying community-supported open source software and have developed a course "Data Analysis 2.0", incorporating principles of blended learning. They will be discussed during this presentation.

INTRODUCTION

Future engineers will have to deal with constantly changing expectations and challenges in their professional career. On the one hand, there is a continuous development of subject-specific knowledge, on the other hand, available methods of tooling and disclosure of information change, and evidence-based action and decision making become increasingly important. This requires a structurally different approach in the formation of engineers, as they have to learn how to cope in a responsible and critical way with the growing and changing availability of information. In the field of Engineering Statistics, Applied Data Analysis and Design of Experiments similar developments take place. Methods for advanced data analysis but also for the efficient planning of experiments are no longer only reserved for specialists in this field. Due to the availability of software they become within the reach of a large group of professionals. Whereas until recently this mainly concerned commercially available software packages, such as SAS, JMP, SPSS, Stata, Minitab or Statgraphics, now open source and freely available platforms under GNU license, supported by an active community of users and developers, take an increasingly prominent place. The statistical analysis and development environment R (www.r-project.org) has been gaining an increasingly important place in recent times and is becoming a standard in this area both within the academic world and within the industry.

On the one hand, these developments offer users the possibility to have easy access to implementations of recently developed methods and techniques. In recent years, for example, the number of packages available within R has grown exponentially (https://cloud.r-project.org/web/packages) to more than 12.000 at present. Moreover, because of the open source character of the software, these packages are relatively easy to adapt to specific wishes and circumstances of the user and can thus be made more tailor-made. On the other hand, users are required to have sufficient insight, experience and critical judgment to be able to select and evaluate adequate methods, weigh alternatives against each other and, if necessary, make adjustments. This has implications for the vision, ambition and structure of future education in the field of Engineering Statistics, Applied Data Analysis and Design of Experiments. However, within the active R-community, these aspects receive only limited attention because the main interest is focused on statistical-methodological aspects.

To develop an integral vision on the structural consequences of these developments for education in the field of Engineering Statistics, Applied Data Analysis and Design of Experiments I have carried out an Advanced Academic Teaching AAT-project "You R, so you are!" at the Eindhoven University of Technology, TU/e, in the Netherlands. In this contribution further details of this project are described, findings from newly developed courses 'Engineering Statistics 2.0' are reported, conclusions are summarized, whereas ambitions and future plans in this dynamic and challenging field are unfolded!

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]

THE PROJECT "YOU R, SO YOU ARE"

Aim of the project "You R, so You Are!" (named with a wink to René Descartes' saying 'Cogito Ergo Sum'!) is to develop and implement an integral educational vision for implementing the use of R in Engineering Statistics, Applied Data Analysis and Design or Experiments courses at the Eindhoven University of Technology, TU/e, in the Netherlands. It concerns courses offered for regular students from the various engineering disciplines, as well as courses provided for PDEng-trainees, participating in the Professional Doctorate in Engineering program at the TU/e, and for PhD students. Activities, developed within the framework of Continuing Education for Professional Engineers, are also included.

State of the Art

Within the framework of the "You R, so You Are" project, a number of existing initiatives in the field of Engineering Statistics, Applied Data Analysis and Design of Experiments education with integrated use of R were studied. This involved examining the content and layout of courses at both introductory and more specialized levels.

At introductory level we paid attention to bachelor-level introductory courses on applied statistics and data analysis with integrated use of R. In detail, the initiatives of Albert et al. (2012), Dalgaard (2008), Diez (2015), Field et al. (2012), Fox et al. (2010), Heiberger (2009) , Hothorn et al. (2014), Kerns (2012), Schumacker et al. (2013) and Verzani (2014) were studied. It is striking that, thanks to the availability of R, extra attention can be paid to the principles of more advanced and computationally intensive methods, such as resampling methods, which are applicable to the analysis of actual research data, also in situations where for example classical methods have their limitations. It is also striking that the approach is more focused on gaining insight and overview in the underlying principles and method of approach and less on formal mathematical or theoretical details. Finally, an important starting point is that the chosen software should be supportive and not leading in the course: it is a good understanding of and insight into the backgrounds of the statistical techniques discussed, not the 'technical' control of the software used that is important.

Special attention was given to initiatives, developed under GNU and CC-license, such as OpenIntro Statistics (www.openintro.org/stat), SWIRL (www.swirlstats.com) and MOSAIC (www.mosaic-web.org). They offer free and open source available educational material that can be easily adapted to detailed needs within specific courses. For our introductory courses, which are described in more detail below, we use Dietz et al. (2015), developed within the OpenIntro Statistics framework, as primary reference, adapting parts of it to our specific Engineering Statistics needs.

We also looked in more detail at courses with integrated use of R in more specialized areas. These included:

- Biostatistics & Clinical Trials: Chen et al. (2014), Shababa (2011)
- Chemometrics: Varmuza et al. (2009), Wehrens (2011),
- Design of Experiments, SPC & 6-Sigma Quality Control: Cano et al. (2012, 2015), Groemping (2011), Kenett et al. (2014),
- Multivariate Data Analysis & Data Mining: Everitt et al. (2009), Ledolter (2013), Torgo (2017), Williams (2011),
- Time series analysis: Cowpertwaith et al. (2009), Hyndman et al. (2018).

Here, too, thanks to freely available R-packages, new opportunities for analyses come within reach and will have influence on the actual contents of the course as already has been concluded in Rijpkema et al. (1991).

R Interfaces: R Commander and R Studio

R is known to have a steep learning curve for practical use. This can be settled for first time users through the use of Graphical User Interfaces for R. Within the projects' context, we gained experience with Deducer (http://www.deducer.org), JGR (https://www.rforge.net/JGR), PMG (https://CRAN.R-project.org/package=RPMG), Rattle (https://rattle.togaware.com), R Commander (www.rcommander.com), RExcel (http://rcom.univie.ac.at) and R Studio (http://www.rstudio.com). Based on this, it was decided to use R Commander for our introductory courses. This interface creates a natural bridge between a menu-driven and the script line-oriented

use of R. Furthermore, through so called plug-ins, R Commander can easily be extended to access specialized statistical methods, implemented within R-packages, in a user-friendly and menu-driven way. For courses at a more advanced level, R Studio is preferred, because of the wider possibilities it offers as Integrated Development Environment, IDE.

ENGINEERING STATISTICS 2.0: COURSES AND TEACHING MATERIALS

In the context of the "You R, so You Are"-project several courses with integrated use of R and R Commander were developed, realized and implemented. These include:

- Regular Courses within the Bachelor College, such as a course 'Applied Statistics' for 2nd years bachelor students in Chemical Engineering (150 participants/year, code: 6A6X0), a course 'Biostatistics' for 1st years bachelor students in Biomedical Engineering (300 participants/year, code 2DM80) and a training 'Analysis of Measurement Data' for 1st years bachelor students in Mechanical Engineering (300 students/year, code 4TR08).
- Elective Courses, such as 'Applied Data Analysis with R' for bachelor students (100 students/year, code 2AS00), and, for master students, 'Time Series Analysis and Forecasting with R' (50 students/year, code 2DD23), 'Design of Experiments with R' (40 students/year, code 2DMN00), 'Biostatistics 2' (40 students/year, code 2DBM90), and 'Multivariate Data Analysis with R' (first edition upcoming, code: JBM220).
- Courses for PDEng-trainees and PhD-students, such as the modular course 'Applied Statistics for Technological Designers', with three modules, viz. 'Engineering Statistics', 'Design of Experiments' and 'Reliability' (30 participants/year) and the course 'Practical Data Analysis for Researchers' (60 participants/year).
- Continuing Education courses for professional engineers, such as 'Practical Data Science with R' (4 days, once a year), 'Time Series Analysis and Forecasting with R' (3 days, once a year), 'Data Mining and Business Analytics' (4 days, once a year) and 'Multivariate Data Analysis with R' (4 days, once a year).

For each of these courses supporting teaching materials have been developed and made available to the participants. They include lecture handouts and a detailed study guide, with organizational details and a mix of additional information sources and assignments, which are complementary and supportive. Principles of Blended Learning are included, in line with our general "Statistics 2.0"-philosophy. The material is adapted to the subjects, level and intended learning outcomes of the relevant course where necessary and desirable!

CONCLUSIONS, AMBITIONS AND FUTURE PLANS

The AAT project "You R, so You Are!" has offered the opportunity to implement the use of R and R Commander to large scale courses in the field of Engineering Statistics, Applied Data Analysis and Design of Experiments. On the one hand insights and experiences gained from our "State of the Art"-exploration were operationalized, whereas, on the other hand, long-term personal experiences with the integration of technology and software in applied mathematics and statistics courses were used (Rijpkema, 1991).

Within the context of the project, priority was given to the realization of courses with integrated use of R and R Commander, in order to allow a large number of students from the TU/e Bachelor College and TU/e Graduate School to be introduced at an early stage with the expected future developments in this area and in their professional career. They are also introduced to the starting points of literate reporting, aimed at reproducible analyzes of obtained research data, and implemented in special R-packages, such as Sweave (http://leisch.userweb.mwn.de/Sweave), knitr (https://yihui.name/knitr/) and Markdown (https://rmarkdown.rstudio.com/).

When developing specific courses special attention was paid to existing communities and initiatives in which experiences and developed educational materials are shared. On the one hand this offers the possibility to efficiently adapt course materials to specific needs of courses, on the other hand it also means that on the basis of experience gained and own developed or adapted material within such communities an active role can be played.

For the future it is foreseen to further improve and elaborate the previously described courses with integrated use of R and R Commander and to keep them up-to-date. I have targeted

plans to contribute to the OpenIntro Statistics initiative with additional material, focusing on specific needs of and applications for engineering students. Furthermore, I hope to publish material developed for my Design of Experiments course with R and the R Commander Plugin for Design of Experiments in the near future.

Within the TU/e Graduate School and the TU/e-Continuing Education initiative, it seems useful to identify the needs for further use and training in the field of R, for example for the development of further training and introduction to technical aspects of handling and graphical presentation of data as well as the development of programs and packages within R. Special attention also deserves the way in which (future) users can gain insight and overview on the almost exponentially growing collection of available R-packages and the way these contributions can be validated and verified by the user.

Finally, it has a high priority to continue to closely follow the developments within R and the exponentially growing collection of related packages, both in terms of subject matter and in terms of direct implications to continue to offer up-to-date courses, consistent with the motto of the project: "You R, so You Are".

REFERENCES

Albert, J., & Rizzo, M. (2012). *R by Example*, New York: Springer.

Cano, E., Moguerza, J., & Concoba, M. (2015). *Quality Control with R*, New York: Springer.

Cano, E., Moguerza, J., & Redchuk, A. (2012). *Six Sigma with R*, New York: Springer.

Chen, D., & Peace, K. (2010). *Clinical Trial Data Analysis Using R and SAS, (2nd edition)*, Boca Raton, Florida: CRC Press.

Cowpertwait, P., & Metcalfe, A. (2009). *Introductory Time Series with R*, New York: Springer.

Dalgaard, P., (2008). *Introductory Statistics with R (2nd edition)*, New York: Springer.

Diez, D., Barr, C., & Çetinkaya-Rundel, M. (2015). *OpenIntro Statistics, (3rd edition)*, available from: http://www.openintro.org/stat/textbook.php

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*, New York: Springer.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*, Thousand Oaks, California: SAGE.

Fox, J., & Weisberg, H. (2010). *An R Companion to Applied Regression, (2nd edition)*, Thousand Oaks, California: SAGE.

Grömping, U. (2011). *Tutorial for designing experiments using the R package RcmdrPlugin.DoE*, available from: http://www1.beuth-hochschule.de/FB_II/reports/Report-2011-004.pdf

Heiberger, R., & Neuwirth, E. (2009). *R through Excel*, New York: Springer.

Hothorn, T., & Everitt, B., (2014). *A Handbook of Statistical Analyses Using R, (3rd edition)*, London: Chapman and Hall/CRC.

Hyndman, R., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice, (2nd edition)*, available from: https://otexts.org/fpp2/

Kenett, R., & Zacks, S. (2014). *Modern Industrial Statistics, (2nd edition)*, Chichester: Wiley.

Kerns, G. (2012) *Introduction to Probability and Statistics Using R*, available from http://ipsur.org/

Ledolter, J. (2013); *Data Mining and Business Analytics with R*, Chichester: Wiley.

Rijpkema, J., Simons, F., & Smits, J. (1991). Mathematics courses with a PC. *International Journal of Mathematical Education in Science and Technology*, *22*(5), 791-798.

Schumacker, R., & Tomek, S. (2013). *Understanding Statistics Using R*, New York: Springer.

Shahbaba, B. (2011). *Biostatistics with R*, New York: Springer.

Torgo, L. (2017). *Data Mining with R: Learning with Case Studies, (2nd edition)*, London: Chapman and Hall/CRC.

Varmuza, K., & Filzmoser, P. (2009); *Introduction to Multivariate Statistical Analysis in Chemometrics*, Boca Raton, Florida: CRC Press.

Verzani, J. (2014). *Using R for Introductory Statistics, (2nd edition)*, London: Chapman and Hall/CRC.

Wehrens, R. (2011). *Chemometrics with R*, New York: Springer.

Williams, G. (2011). *Data Mining with Rattle and R*, New York: Springer.