Comments on "Future of Statistical Computing"

Dianne Cook
Iowa State University

Hadley Wickham
Rice University

Technology has changed science, but it also bounds the way people work and think. As Lee Wilkinson points out: almost half a century ago John Tukey pioneered radical new computationally intensive ideas in statistical graphics and exploratory data analysis. One would expect that these ideas to have permeated into mainstream statistics to a much larger extent than they actually have today. A big reason for this is that the technology available to Tukey was not available to the masses. The majority of statisticians could not experiment in the same way and could only watch and admire from a distance.

Today everyone has the technology. We can all explore data, perform complex simulation, crunch large numbers, construct beautiful informative plots, develop statistical thinking and easily communicate their results in a variety of media. We are no longer bound by pencil and paper mathematical methods. We can build virtual prototypes of theory and methods, play with them, test them in a million ways, refine the design, and easily share them with others. Today's computational explosion helps us to be curious again, in a similar spirit to Tukey, and helps us better communicate our work.

Statistical computing is inextricably intertwined with other areas of statistics. In this discussion we focus on the challenges and opportunities in four areas: support for research in industry and academia, availability of software, changes in statistics education and outlets for communication.

The support of statistical computing research in academia is necessary for progress. Perhaps this is a good time to look into the past to see how far the field has come. For us, particularly notable highlights are the development of Orion by John McDonald, XLispStat by Luke Tierney, and the S language by John Chambers and colleagues. Orion provided the first implementation of linked brushing [**?**], on equipment that was hundreds of times more expensive than today's laptops, millions of times slower, but remarkably, with very similar screen resolution. Development of the S language [Becker and Chambers, 1984] transformed the practice of statistics, and was fostered by a research environment flush with monopolistic business money and no short-term profit pressures. This unique environment has since disappeared but the work has survived and evolved into the open source, every statistician's software, R (`www.R-project.org`), in an academic environment. XLispStat [Tierney, 1991], which was ground-breaking in its seamless integration of models and interactive plots, arose in an academic environment.

From a classical hackers birth, R has evolved from Ross Ihaka and Rob Gentleman's [Ihaka and Gentleman, 1996] baby, to a global project with a sophisticated management by R Core [R Development Core Team, 2003]. R's packaging system allows others to play too, and each year more and more packages implement more and more cutting edge research. This is a major achievement, but it comes at a cost: some of the best statistical computing brains of our generation are being used for software maintenance. Must we be prepared to pay this cost so that our work can have impact in the wider world?

R's impact on development time and cost has been particularly important. In the past, the time from method development to general distribution and usage may have been on the order of a decade, but with an R package, global distribution is trivial and the time from development to use has shortened to an order of months. Thanks to R and other open source systems, the cost to set yourself up with the latest statistical computing research is now essentially zero, except for the cost of the hardware to run it and the internet connection to download it. In contrast to Lee, we see little impact of commercial software on progress in statistical computing.

Algorithm development, while still important, is a little old school today. The work of Friedman, Breiman, and the other computational statistics pioneers mentioned by Wilkinson, are an important foundation, and

continue to be fundamentally important to today's challenges. Statistical "hacking", though, is the new frontier: cobbling disparate existing resources together to solve new problems, taking technological advances from other fields and trying them out for statistics, tapping into online feeds, processing and reporting live. This style of work is exemplified by the output of many young researchers in statistical computing, for example using music and sound to represent chat room data [Rubin and Hansen, 2003].

Electronic publishing makes research accessible and reproducible [Gentleman and Temple Lang, 2007], because as well as a written description of your work, you can open up your workbook include data, software, and videos. It is a growing expectation that others should be able to easily replicate your analyses and verify the results. We are hopeful that this trend will continue, allowing reproducible research to eliminate both scientific fraud and genuine errors. We expect more open-access, electronic outlets to emerge in the coming years. The technical cost of online publication is approaching zero, with both academic (e.g. `http://pkp.sfu.ca/ojs`) and non-academic (e.g. `http://wordpress.org/`) toolkits making it easy to get up and going. Several statisticians manage blogs where items related to statistics can be posted and debated. These are particularly interesting reading when the material crosses traditional disciplinary boundaries.

Statistical computing bridges statistics with other disciplines. It makes forays into other people's backyards, such as data mining (e.g. [Ripley, 1996]), while providing readily available software that allows others to venture into the backyard of statistics. The growing number of internet resources makes it possible for the masses to learn about statistics, and do their own statistical computing.

Education is a specialized form of communication and involves equipping students with an understanding of statistical thinking. Statistics is much more than a handful of statistical tests — with statistics we are curious about the variation in measurements. What we can't explain, we model stochastically. Computers make it possible to study variation much more fluidly than ever before: to simulate, sample, and permute to incorporate variability into analyses. We can show the variability using plots, instead of a small set of summary statistics created for their mathematical tractability. Tight coupling of numerical techniques with graphics software allows us to diagnose fits, and understand quirks in algorithms, so that it is possible to discuss diagnostics and algorithms in detail in classes. Graphical, numerical and algebraic presentation go hand in hand to help a wider range of students understand how statistics functions.

Statistical methods need to be taught with a support set of skills such as data management and programming. We will increasingly expect graduate students to have computational skills at least equal to their mathematical skills. The fundamental goal is to help students be better problem solvers using data and statistical thinking. A technology-savvy, science-educated workforce is critical for innovation in business. Many companies are sponsoring graduate research, and often the contract allows free-distribution of software and methodology, while protecting data. This is an important development for driving statistical computing research. Industry/academia collaboration will drive the education process.

In summary, it is an exciting time for statistical computing: open source statistical software allows us to quickly and cheaply track the frontiers of research; open source web software allows us to communicate easily and reliably; and randomization-based approaches allow us to move beyond traditional pencil-and-paper parametric statistics to explore graphical non-parametric approaches. We are particularly excited about the opportunities for enhancing our ability to be curious about data and our ability to communicate to our students and the general public.

References

[Becker and Chambers, 1984] Becker, R. A. and Chambers, J. M. (1984). *S: An Environment for Data Analysis and Graphics*. Wadsworth, Belmont, CA.

[Gentleman and Temple Lang, 2007] Gentleman, R. and Temple Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16:1–23(23).

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

[R Development Core Team, 2003] R Development Core Team (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

[Ripley, 1996] Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

[Rubin and Hansen, 2003] Rubin, B. and Hansen, M. (2003). Listening post. `http://www.earstudio.com/projects/listeningpost.html`.

[Tierney, 1991] Tierney, L. (1991). *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics*. John Wiley & Sons, New York.