The R Project in Statistical Computing

Brian D. Ripley, University of Oxford,

ripley@stats.ox.ac.uk

R is an Open Source environment for statistical computing and graphics `not unlike' version 3 of the S system developed at Bell Laboratories (formerly AT&T, now Lucent), which is the basis of the commercial system S-PLUS® [1] from the Insightful Corporation.[2] This article describes the R philosophy and current status: it is definitely not a review as the author is a member of the R core team as well as the S-PLUS Advisory Board.

What is R?

Executive summary: R is an advanced statistical computing system with very high quality graphics that is freely available for most computing platforms. More details are available on-line at `www.r-project.org`.

R currently exists as source code, as well as binary distributions for most distributions of Linux, for Windows (95 and later: see Figure 1) and for the PowerPC Macintosh. It can be downloaded from *CRAN*, with a master site at `cran.r-project.org` and a series of mirrors including one at Bristol, `cran.uk.r-project.org`. It is distributed under the GNU Public Licence, which means that it is freely available, both in the sense that it is available for only the distribution costs and that the source code is open to view and modification.

The R project was started by Ross Ihaka and Robert Gentlemen at the University of Auckland, for teaching on Macintoshes which by today's standards were very limited. R still runs well on low-end systems: a basic installation takes around 20Mb of disc space and runs in 16Mb RAM. A usable Windows distribution fits on three floppy discs, although these days people are more likely to distribute CD-Rs or ask students to download it. The usefulness of the system is greatly enhanced by extension packages, eight of which ship with R and about 100 are available for download from CRAN. This download can be done easily from within R: for example

```
install.packages("VR",
  CRAN="cran.uk.r-project.org")
```

will download (if you are on-line) and install the R support software for Venables & Ripley (1999).

R provides a command-line-driven interpreter for the S language. The dialect supported is pretty close to that implemented in S-PLUS 2000, close enough that many users will not notice the differences. They will notice some differences in output: the R developers have different preferences.

What is it being used for?

With a freely-distributed system it can be hard to know what it is being used for and how extensively. One clue comes from requests for support. A free system such as R is supported by an email list, and that suggests that R is used extensively in elementary teaching, for student projects and surprisingly often by researchers including those in companies.

At present R is almost entirely driven from a command line, even on versions with GUI consoles (Windows, Mac, GNOME under Linux). It is hotly debated whether this is a disadvantage in teaching, as students are increasingly used to filling in a series of point-and-click forms with buttons to click for help. R is not like that, and although there are ways to customize it with such interfaces (by programming in C, Tcl/Tk or Java), these are laborious to use. My own department decided not to use it for undergraduate classes because a GUI interface was thought easier to support. The obverse side is that a command-line interface is a much richer environment, and students are forced to think about what they are doing when they use it. That debate will run and run.

Nolan & Speed (2000) show how R can be used in teaching an introductory statistics course. They integrates the theory of statistics with the practice of statistics through a collection of case studies, and use R (or S-PLUS) to analyse the data. This is an inspiring book that I highly recommend to those planning to teach real statistics.

R is an advanced system: it provides statistical methods up to M.Sc. level and beyond. Many research students use it as a research tool. It provides access to state-of-the-art methods in robust methods, density estimation, smoothing, multivariate analysis, neural networks, time series …. Classical statistics is also covered; one of the extension packages is devoted to classical tests, for example.

[1] Versions up to S-PLUS 2000 are based on S version 3; the current Unix/Linux version 6.0 of S-PLUS and the upcoming version 6.0 for Windows are based on S version 4.

[2] Until January 2001 known as MathSoft Inc.

There are now many examples of R being used for serious large-scale data analysis. It has been used for election forecasting in Austria and will be used in the UK; it is used for a number of systems in the processing of gene expression data, and my group have used it to process time series of brain images.

There are projects providing a Web-based interface to R. Rcgi (`http://stats.mth.uea.ac.uk/Rcgi/`) by M. J. Ray at the University of East Anglia provides a form interface for students to submit data or code to a copy of R running on the web server and get back numeric and graphical output.

R sounds a lot like S-PLUS!

To make a comparison more precise, let us confine attention to S-PLUS 2000, the current Windows version. R and S-PLUS 2000 are similar, and will be used in very similar ways in the teaching context. Their languages are similar enough that students who use S-PLUS in the university can practise with R at home (and it is clear from the email support list that many do). A few comparisons:

(a) S-PLUS 2000 is commercial and comes with commercial support. R is much easier to distribute.

(b) R is much smaller and runs on less powerful machines.

(c) S-PLUS 2000 is monolithic: R has a very small core and many extensions.

(d) S-PLUS 2000 does have a `menus and dialog boxes' interface as well as a command-line interface. It also has ways to edit graphs.

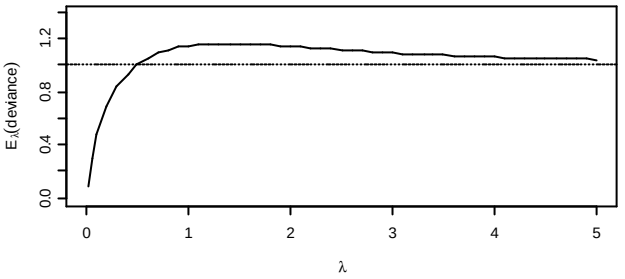
(e) There are conflicting claims about performance. I find them about equal, but R is more tolerant of badly-written code which can make S-PLUS slow to a crawl.

(f) There is not much to choose in quality. S-PLUS is a more mature system. I suspect that R has more bugs, but that these will get fixed much faster and by subject-matter experts.

(g) Both have 2D graphics of much higher quality than most commercial statistical systems. R is particularly good at annotating with mathematical notation (see Figure 2). S-PLUS has much richer facilities for multi-panel plots (Trellis™).

(h) Both are weak on `graphics for the video games generation', that is 3D and dynamic graphics, but R is weaker.

Figure 2: Mathematical annotation



This figure shows the expected value of the residual deviance for a sample from Poisson(λ), and hence that apparent under- or over-dispersion can often be attributable to bias.

Why Open Source?

R may be free to the end user, but its development has involved man-years of volunteer effort by the core team (about a dozen people, not all active at all times) and contributors. We are sometimes asked why we do it. One motivation for me has been to make available a first-class statistical system to those with fewer resources, in particular students and researchers in the third world. Now when students from developing countries ask Bill or me for a copy of S-PLUS, as they often do, we can point them towards R in the knowledge that they will be getting a very useful product for an unbeatable price. (When they ask for free, open source versions of our books as well, we can only shrug.)

An advantage for me is to be able to study all the source code, so I can find out (and by now probably have) precisely which variation on a statistical method has been implemented. There used to be an adage that to really learn a method you had to know how to do it by hand (and students were made to do so). The current analogue is to understand what the source code does.

Where is R going?

R has been developing very rapidly. Until the first non-beta release, version 1.0.0 on 29 Feb 2000, there was about a release a month. The aim is now for a full release about twice a year with intermediate bug-fix releases. We are still adding statistical features (most recently MANOVA and factor analysis at the request of psychologist users), and we have been addressing performance issues.

As a volunteer project, where R goes depends entirely on what its volunteers want to do. My impression is that it will now move fastest as a research tool in statistical computing, and in particular as one component in a distributed computing system.

Another possibility is that commercial users will want new features enough to pay for them to be developed; some aspects of graphics have been mentioned.

Using R as a computational engine in other systems is relatively easy and for open systems has no licence complications. There are ways to embed R into a, say, Visual Basic front-end and to call R functionality from Excel. This provides a rich vein to be explored, and now has Unix/Linux analogues, too.

Much of what is now available to all in R comes from the teaching needs of a few. Contributions will be welcomed with open arms.

References

Nolan, D. & Speed, T. P. (2000) *Stat Labs: Mathematical Statistics Through Applications*, Springer. Support material is available at `.http://www.stat.Berkeley.edu/users/statlabs/.`

Venables, W. N. & Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS*. Springer. Support material is available at `.http://www.stats.ox.ac.uk/pub/MASS3/`, including Complements describing its use with R.

Figure 1: Part of an R session on Windows