Psicothema 2009. Vol. 21, nº 4, pp. 652-655
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

¿Existe vida más allá del SPSS? Descubre R

Paula Elosua Oliden
Universidad del País Vasco

R es un entorno de programación, análisis estadístico y generación de gráficos distribuido bajo licencia GNU. Es un poderoso aliado para la investigación y una excepcional herramienta de trabajo para la docencia. Está constituido por más de 1.400 paquetes integrados con los que es posible ejecutar simples análisis descriptivos o aplicar los más complejos y novedosos modelos formales. Además, la incorporación a R de interfaces gráficas como Rcommander que crean entornos de trabajo amigables muy similares al entorno del SPSS permiten saltar la barrera de la accesibilidad, y utilizarlo sin ningún tipo de reparo en la docencia. ¿Existe algo mejor? Libre, gratuito, asequible, accesible y siempre a la vanguardia.

Is there life beyond SPSS? Discover R. R is a GNU statistical and programming environment with very high graphical capabilities. It is very powerful for research purposes, but it is also an exceptional tool for teaching. R is composed of more than 1400 packages that allow using it for simple statistics and applying the most complex and most recent formal models. Using graphical interfaces like the Rcommander package, permits working in user-friendly environments which are similar to the graphical environment used by SPSS. This last characteristic allows non-statisticians to overcome the obstacle of accessibility, and it makes R the best tool for teaching. Is there anything better? Open, free, affordable, accessible and always on the cutting edge.

Las tareas relacionadas con el tratamiento y análisis de datos en el campo de las ciencias sociales y de la salud pueden ejecutarse utilizando varias herramientas informáticas, de entre las cuales sobresalen, por su extendido uso en nuestro entorno, SPSS, SAS, STATISTICA, Systat, Stata o GenStat. Las diferencias entre ellas residen básicamente en su mayor o menor grado de generalidad, en el costo o en la facilidad de uso. Este triple criterio (generalidad, costo, facilidad de uso) además de la tradición de uso guían la elección de uno u otro software para la investigación y para la docencia. Sin duda, se trata, en todos los casos, de aplicaciones eficientes basadas en el uso de potentes interfaces gráficas (GUI Graphical User Interface) con sistemas de menús y ventanas desplegables que tienden un puente al usuario en el proceso de modelización estadística. Las *GUI*s ejercen de intermediarios entre los modelos formales y el usuario, que sin un entrenamiento excesivamente intensivo es capaz de explorar, analizar y modelizar sólo con un simple clic de ratón. Los entornos de trabajo amigables han simplificado enormemente los procesos relacionados con el tratamiento de datos y, como consecuencia, el análisis de datos se ha socializado. Aunque los efectos positivos de la generalización puedan a veces verse ensombrecidos con una aplicación incorrecta de los modelos o una torpe interpretación de los resultados, los beneficios superan sobradamente los estropicios con los que en ocasiones nos encontramos e incluso cometemos.

Fecha recepción: 19-12-08 • Fecha aceptación: 11-3-09
Correspondencia: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20018 San Sebastián (Spain)
E-mail: paula.elosua@ehu.es

Sin embargo, no todo son glosas hacia estos paquetes informáticos. Son económicamente costosos, por lo tanto difícilmente asequibles por el estudiante o por el usuario no integrado en grandes compañías o instituciones. Son además rígidos en sus procedimientos, en el sentido de que no ofrecen la posibilidad de acomodar o adaptar a las necesidades particulares de cada usuario ni sus algoritmos, ni el diseño matricial de entrada de datos (caso×variable), ni la presentación de resultados que en muchas ocasiones se torna interminable con una profusión de tablas y salidas que puede llegar a entorpecer y ralentizar la tarea. Son, por supuesto, programas eficientes y fiables en la aplicación de los modelos que implementan, pero no son flexibles y son costosos. Si dispusiéramos de una herramienta que reuniera las características positivas de estos programas, pero además, ofreciera ventajas añadidas, tal vez podría interesarnos conocerla.

Pues bien, esta herramienta existe; hablamos de R. El objetivo de estas líneas es mostrar las propiedades tanto para la docencia como para la investigación de un entorno de programación y análisis estadístico que es todavía algo desconocido por el público general. R posee una estructura versátil, fácilmente adaptable a las necesidades del usuario básico, medio o avanzado, del estudiante o del profesor, tiene una capacidad de análisis asombrosa, un entorno sorprendente para el desarrollo de gráficos y además se enmarca dentro de la filosofía de software libre, lo cual la convierte en gratuita. En R confluyen características que la hacen única, es libre, de código abierto, dispone de versiones para distintas plataformas (Microsoft Windows, Linux/UNIX o Macintosh) y está siempre a la vanguardia de los más avanzados modelos estadísticos. Intentaremos presentarla.

R es un entorno de programación y análisis estadístico y gráfico derivado del lenguaje de programación *S* (Becker, Chambers y

Wilks, 1988; Chambers, 1998; Chambers y Hastie, 1992; Venables y Ripley, 2000). Existe una versión de este lenguaje distribuida por *Insightful Corporation* bajo el nombre comercial de *S-Plus*, y una versión libre con código abierto conocida como R. Esta última fue desarrollada por **R**oss Ihaka y **R**obert Gentleman (ésta es una de las razones del nombre R; Ihaka y Gentleman, 1996) del Departamento de Estadística de la Universidad de Auckland (Nueva Zelanda). La primera versión de R se difundió rápidamente y la expansión es hoy irrefrenable. Desde su creación R se alimenta y crece con los trabajos de investigadores provenientes de prácticamente todas las ramas del conocimiento. Las aportaciones desinteresadas de funciones y librerías de propósito tanto general como específico hacen de R en un entorno dinámico formado por una comunidad en movimiento continuo y acelerado que se inscribe dentro de la filosofía del software libre.

R, en tanto en cuanto software libre, se inscribe dentro del proyecto GNU *General Public Licence* (Licencia Pública General, GNU). Se trata de una licencia creada por *Free Software Foundation* (Fundación para el software libre), organización fundada por Richard Matthew Stallmana en el año 1985. El principal propósito de la licencia GNU es declarar la libertad del uso, modificación y distribución del software y protegerlo de intentos de privatización que puedan de algún modo restringir su uso (el contenido de la licencia puede consultarse en el sitio http://www.gnu.org/copyleft/gpl. html). Dentro de esta licencia se distribuyen un sinfín de programas, muchos de los cuales son versiones libres del software informático generalista más utilizado. De entre ellos tal vez los más extendidos sean la suite ofimática *OpenOffice*, el navegador *Mozilla*, los artículos de *wikipedia,* el sistema operativo *GNU/Linux*, o el editor de textos *Emacs*.

Parte de la vasta información disponible sobre R es accesible a través de la web CRAN (*Comprehensive R Archive Network*; http://cran.r-project.org/), sitio oficial de R. Es la página base del proyecto R, desde la cual se puede descargar la última versión del programa (un equipo formado por unas doce personas, *R Development Core Team*; asumió en 1997 las labores de actualización semestral del código de R), consultar manuales sobre R, obtener ayuda sobre su funcionamiento a través de un sistema de ayuda on line, y, en definitiva, estar al corriente de los movimientos en este entorno de trabajo. La bibliografía sobre R es amplia, porque a la propiamente desarrollada para R se añade la bibliografía sobre S o S-plus. Entre los libros más importantes es imprescindible citar el manual de referencia que desarrolla y actualiza con cada versión de R el *R Development Core Team* (2008) y el ya clásico y excelente volumen de Venables y Ripley (2002) que se ha convertido en el libro de cabecera de los usuarios de R y S «*Modern applied statistics with S-plus*». Como obras introductorias, las más aconsejables serían el sencillo trabajo de Venables, Smith y the R Development Core Team (2007) publicado bajo el título «*An introduction to R*» y el volumen de Paradis (2005) titulado «*R for beginners*». Existen traducciones al castellano de ambas en el sitio http://cran.es.r-project.org/, donde también pueden consultarse los textos originales. La obra de Dalgaard (2002), «*Introductory statistics with R*», es un volumen compacto que cubre el contenido de un curso básico de estadística, y al mismo tiempo introduce al lector en los conceptos elementales de la programación en R. Son más recientes el libro de Braun y Murdoch (2007), que ofrece un primer curso en estadística utilizando R, y la obra publicada por Crawley (2008), que en sus 900 páginas ilustra el proceso de modelado estadístico con R. También es altamente aconsejable el libro de Fox (2002) titulado «*An R and S-plus companion to applied regression*». Para usuarios iniciados la última actualización de la obra de Chambers (2007), uno de los artífices del lenguaje S, es una excelente opción. La transición entre el SPSS o SAS al R puede acompañarse de la lectura del trabajo de Muenchen que con el título «*R for SAS and SPSS Users*» ofrece dos productos: un corto y accesible en red (http://rforsasandspssusers.com/), y otro de 470 páginas publicado recientemente por Springer (Muenchen, 2009).

R es más que un software para el análisis de datos, es un entorno de trabajo que va incrementando continuamente sus capacidades con la incorporación de paquetes y funciones que se integran perfectamente en el sistema R. En la actualidad el entorno R está compuesto por más de 1.400 paquetes integrados (Fox, 2008). Dispone de funciones básicas para los más elementales análisis descriptivos (media aritmética, desviación estándar, varianza…) y para los más complejos modelos formales derivados de los últimos avances en el campo de la estadística, psicometría, bioinformática, estadística bayesiana, bioestadística, minería de datos, econometría y finanzas, ecología, márketing, estadística robusta, sensometría, estadística espacial, estadística en las ciencias políticas, visualización y gráficos, análisis de redes neuronales y mucho más. Refiriéndonos a la psicometría, por ejemplo, podríamos señalar que incluye librerías para la estimación de modelos de respuesta al ítem, análisis de correspondencias, modelos de ecuaciones estructurales, escalamiento multidimensional, teoría clásica de tests, meta-análisis, modelos multinivel o análisis en micro-arrays. La revista *Journal of Statistical Software* dedicó un volumen especial a las últimas contribuciones en el campo de la teoría de tests (Leeuw y Mair, 2007). Sin embargo, R no es solamente una colección de paquetes integrables; es también un programa intérprete que permite al usuario definir funciones y utilizarlas según sus necesidades.

Aparte de las capacidades de análisis estadístico, R es un potentísimo generador de gráficos que cuenta con numerosas y variadas funciones y librerías diseñadas con esta finalidad. Es posible componer un simple plot, definir figuras extremadamente complejas e incluso crear animaciones (Maindonald y Braun, 2007; Murrel, 2005; Sarkar, 2008). Baste como prueba la figura 1. La primera figura es una imagen del volcán *Maunga Whau* creada por el Grupo de Desarrollo de R (el original tiene color) y la segunda, diseñada por Thomas Lumley perteneciente al Core Team, ilustra una prueba de hipótesis. Ambos gráficos han sido extraídos de una muestra más extensa que está disponible junto a los códigos para su generación en el sitio http://addictedtor.free.fr/graphiques/. Todos ellos evidencian la versatilidad y posibilidades de R.

Si las ventajas que ofrece R son tantas, ¿a qué se debe su escaso uso e incluso desconocimiento en las ciencias sociales y de la salud? No es un entorno que se utilice de forma generalizada en las universidades para la docencia de las asignaturas relacionadas con el análisis de datos, y tampoco observamos su uso extensivo en el ámbito de la investigación. Existen proyectos institucionales aislados como R UCA (http://softwarelibre.uca.es/node/787) o CRISOL desarrollados por la Universidad de Cádiz y por la Universidad Carlos III, respectivamente (http://crisol.uc3m.es/) cuyo objetivo es implementar el software libre. Algunos profesores, a título personal, hemos introducido R en la docencia de grado y de postgrado; sin embargo, constatamos su escasa difusión. Hasta hace relativamente poco tiempo el entorno R estaba limitado a personal especializado con conocimientos de análisis estadístico y

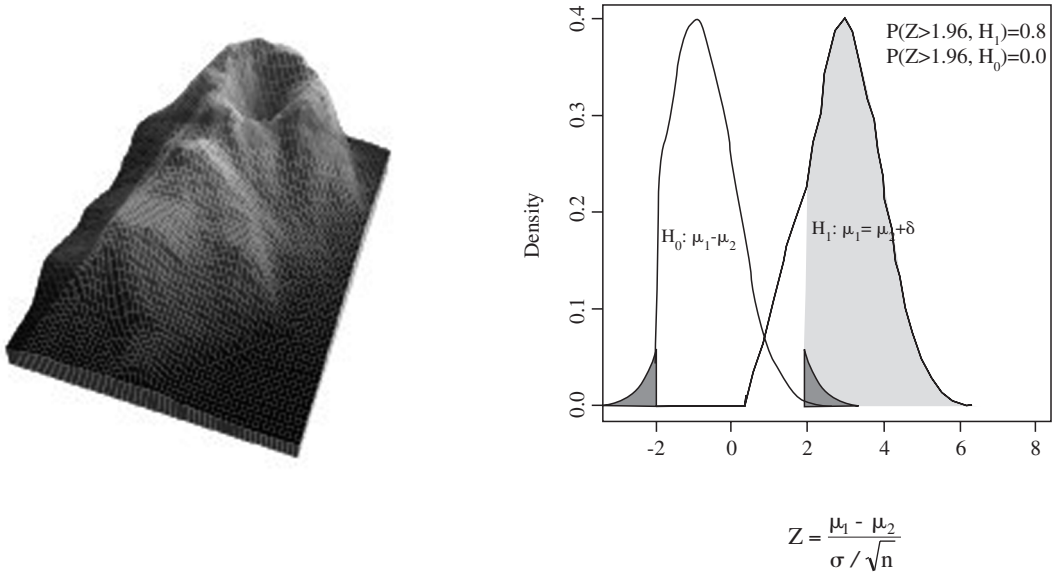

Figura 1. *Ejemplos de gráficos generados con R*

programación porque su entorno de trabajo natural es la construcción de códigos sobre una interfaz de comandos en línea que dista de las interfaces gráficas basada en ventanas emergentes y menús desplegables.

Sin embargo, existen interfaces gráficas desarrolladas para el análisis de datos bajo R; por ejemplo, *R.NET* (http://www.u.arizona.edu/~ryckman/RNet.php), *Poor Man's GUI* (http//www.math.csi.cuny.edu/pmg) o *Rcommander* (Fox, 2005). De entre ellas destacamos la última, tanto por sus prestaciones como por su simplicidad. *Rcommander* es un paquete adicional de R (conocido como Rcmdr) creado por John Fox —uno de los grandes dinamizadores del proyecto R—. El trabajo con R a través de *Rcommander* genera un entorno amigable similar al que utiliza SPSS. Es el intermediario perfecto para acercar al usuario habitual de paquetes comerciales al entorno de programación R, permitiendo una transición sencilla a la vez que una aproximación a esta filosofía de trabajo.

Rcommander incluye funciones para ejecutar análisis estadísticos y generar gráficos. Los menús desplegables permiten: importar/exportar datos, manipular variables (recodificar, calcular…), seleccionar casos, describir variables, generar gráficos, obtener estadísticos básicos, coeficientes de fiabilidad, contrastar hipótesis, aplicar el modelo lineal general o modelos de análisis multivariados como el análisis factorial, el análisis de correspondencias o el análisis de componentes principales. Las opciones que ofrece *Rcommander* cubrirían los contenidos de los cursos básicos de análisis de datos en las ciencias sociales y de la salud. Fox (2007) publicó un manual sobre esta librería que está disponible en el sitio http://socserv.mcmaster.ca/jfox/Courses/soc3h6/Getting-Started-with-the-Rcmdr.pdf. Familiarizarse con *Rcommander* permite además ir profundizando en este entorno de programación, porque entre otras funciones incluye la edición de comandos. Al respecto señala Fox «*I hope that the RCommander graphical user interface motivates users to explore the true power of R, which is best exploited through the standard command-line interface*» (Fox, 2008; comunicación personal), y yo, con su permiso, hago mías sus palabras.

R es gratuito, por lo tanto económicamente asequible, y la incorporación de *Rcommander* lo hace accesible al lego en programación y análisis. La asequibilidad y accesibilidad son características que junto con la fiabilidad y eficacia deberían de guiar la selección de un software para la docencia del análisis de datos. No es suficiente contar con un programa eficiente y de fácil manejo. Convendría que la labor docente y la formación del alumno se apoyaran en herramientas que no estén limitadas por licencias corporativas que restringen su uso a un número de puestos de trabajo determinado. El profesor debería, en la medida de lo posible, ofrecer al alumno la oportunidad de trabajar en el análisis de datos fuera de las aulas para las que la Universidad haya adquirido licencias de trabajo. La incorporación de R a la docencia permitiría asimismo satisfacer una de las demandas de la educación universitaria relacionada con la formación continua. Trabajar con R ofrece al alumno la posibilidad de disponer y dominar una herramienta que le permitirá ejercitar de forma autónoma cuanto ha aprendido a la par que profundizar en su formación. Son beneficios que difícilmente podría alcanzar con otro software para el tratamiento de datos.

A la postre, R viene altamente avalado por constituirse en el entorno de mayor implementación entre la población estadística, lo cual garantiza su validez; no en vano se reconoce a R como la *lingua franca* de la estadística computacional. R es versátil, libre y con la utilización de interfaces como *Rcommander* es fácil de utilizar. Estas características convierten a R en un poderoso aliado para la enseñanza del análisis de datos y para la adquisición por parte del alumno y personal interesado de una herramienta continuamente actualizada que le ofrece la posibilidad de una autonomía de trabajo no disponible bajo cualquier ningún otro entorno. ¿Existen obstáculos para su implementación generalizada más allá de la inercia y la tradición de uso de otros programas?

Es cierto que el primer contacto con R puede producir sensación de aspereza y algún rechazo derivado de la impresión de que trabajar con R requiere un entrenamiento intensivo y un alto nivel de conocimientos de estadística o programación. Esta primera impresión, sin embargo, queda superada a los pocos minutos, y el tiempo invertido en el aprendizaje de R se rentabiliza inmediata-

mente. La comunidad que trabaja en R y con R está formada por un grupo muy dinámico cada vez más numeroso y heterogéneo, que dispone de grupos de discusión en los que es posible consultar cualquier duda o problema. Además, son muchos los materiales que se generan en torno a R. Existen manuales, códigos ya escritos para ejecutar funciones, listas de distribución muy activas y útiles, páginas dedicadas a FAQ a las que se accede desde la página principal de CRAN, un boletín de noticias (*R News*) y conferencias internacionales con periodicidad anual (*useR!).* En definitiva, el usuario de R novel o avanzado entra a formar parte de un activo y entusiasta grupo de trabajo que le resolverá o tratará de resolver los problemas que pueda encontrar en el manejo de este extraordinario entorno de trabajo. Invitamos al lector, sea este investigador, profesor, alumno, lego o experto a adentrarse en R y descubrir las posibilidades que ofrece. Estamos seguros de que no quedará defraudado.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el Ministerio de Ciencia e Innovación (PSI2008-00856) y por la Universidad del País Vasco (GIU08/17).

Tabla 1
Notas para la instalación de R y *Rcommander*

La instalación de R arranca desde el sitio oficial de CRAN (http://cran.r-project.org/). Una vez seleccionada la plataforma utilizada habitualmente (Windows, Linux o Mac) se accede a una página en la cual se escogerá la opción base (*Binaries for base distribution*) que abrirá a su vez una nueva página con el archivo de instalación correspondiente a la última versión de R. En el momento de redacción de este trabajo R-2.8.0-win32.exe.

El proceso de instalación es estándar, y permite seleccionar el español como idioma base. Tras el proceso de instalación R está listo para ser utilizado. Una vez ejecutado, aparecerá el símbolo del sistema (>) en la consola de R (pantalla principal), lo cual indica que R está dispuesto para recibir comandos u órdenes. Pruebe el lector a introducir cualquier operación algebraica.

La instalación de *Rcommander* (Rcmdr) se realiza a través de la opción Paquetes>Instalar de la barra de tareas de R. La selección de esta opción abrirá una ventana en la que se seleccionará una imagen o espejo de CRAN desde la cual se descargará e instalará Rcmdr de forma automática. La instalación española viene acompañada de una traducción del manual.

Una vez instalado Rcmdr es necesario «cargarlo». Para ello, basta seleccionar las opciones Paquetes>Cargar de la barra de tareas y de entre el listado de librerías disponibles se elegirá Rcmdr. También es posible acceder a *Rcommander* directamente desde la consola tecleando tras el símbolo del sistema (>) library(Rcmdr).

La primera vez que se solicite cargar el paquete Rcmdr, R avisará de que para el correcto funcionamiento de Rcmdr es necesario instalar paquetes adicionales. El procedimiento de instalación es automático; busca las librerías necesarias y solicita al usuario permiso para instalarlas. El usuario sólo necesitará asentir a los requerimientos del proceso de instalación.

Referencias

Becker, R.A., Chambers, J.M., y Wilks, A.R. (1988). *The new S language: A programming environment for data analysis and graphics*. Pacific Grove, CA: Wadsworth.

Braun, W.J., y Murdoch, D.J. (2007). *A first course in statistical programming with R*. Cambridge: Cambridge University Press.

Chambers, J.M. (1998). *Programming with data: A guide to the S language*. New York: Springer.

Chambers, J.M. (2007). *Software for data analysis: Programming with R*. New York: Springer.

Chambers, J., y Hastie, T. (1992). *Statistical models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Crawley, M.J. (2008). *The R book.* Chichester: John Wiley & Sons.

Delgaard, P. (2002). *Introductory statistics with R.* New York: Springer.

Fox, J. (2002). *An R and S-plus companion to applied regression*. Thousand Oaks, CA: Sage.

Fox, J. (2005). The Rcommander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, *14*(9), 1-42.

Fox, J. (2007). *Getting Started With the R Commander*. Dirección URL: http://socserv.mcmaster.ca/jfox/Courses/soc3h6/Getting-Started-with-the-Rcmdr.pdf.

Fox, J. (2008). Editorial. *RNews*, *8*(2),1-2.

Ihaka, R., y Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314.

Leew, J., y Mair, P. (2007). An introduction to the special volumen on «Psychometrics in R». *Journal of Statistical Software*, *20*(1), 1-5.

Maindonald, J., y Braun, J. (2007). *Data analysis and graphics using R. Second edition. Cambridge University Press.* Dirección URL: http://wwwmaths.anu.edu.au/~johnm/r-book.html.

Muenchen, R.A. (2007). *R for SAS and SPSS users*. Dirección URL: http://rforsasandspssusers.com/.

Muenchen, R.A. (2009). *R for SAS and SPSS users*. New York: Springer.

Murrel, P. (2005). *R Graphics*. Boca Ratón, Florida: Chapman & Hall.

Paradis, E. (2005). *R for Beginners*. Institut des Sciences de l'Evolution: Montpellier, France.

R Development Core Team (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sarkar, D. (2008). *Lattice. Multivariate data visualization with R.* New York: Springer.

Venables, W.N., y Ripley, B.D. (2000). *S programming.* Springer: New York.

Venables, W.N., y Ripley, B.D. (2002). *Modern applied statistics with S*. Fourth edition. New York: Springer.

Venables, W.N., Smith, D.M., y the R Development Core Team (2007). *An introduction to R.* Vienna, Austria: R Foundation for Statistical Computing.