Estatística con R

Salvador Naya Fernández · Escola Politécnica Superior.
Departamento de Matemáticas. Universidad de A Coruña. Grupo MODES. http://dm.udc.es/modes/

Resumo

O programa de Software libre R, é unha linguaxe para o cálculo estatístico e a xeración de gráficos. R foi desenrolado inicialmente por Gentleman e Ross Ihaka do departamento de estatística da Universidade de Auckland, na actualidade a lista de colaboradores deste software libre foi crecendo exponencialmente con aplicacións a tódolos campos da estatística, considerándose hoxe en día como a lingua franca da estatística.

Entre as razóns para elixir o R como paquete estatístico poderíase argumentar a estabilidade, baixo consumo de recursos, dinamismo e gratuidade. Neste traballo expoñerase algunhas das razóns para a elección do programa R para uso estatístico, tanto para docencia como para investigación.

Introdución

Nos últimos anos, houbo unha enorme expansión na utilización de tecnoloxías da información e da comunicación (TICs) no ensino e aprendizaxe da estatística en tódolos niveis educativos. Unha clase especial desta tecnoloxía, que se atopa en pleno proceso de expansión e que recentemente chamou a atención de moitos educadores e investigadores nesta área, é o uso de programas que no necesiten dunha licenza de usuario, é dicir, programas de Software libre. O fácil acceso a Internet fai posible a instalación de software libre o tempo que facilita tamén o acceso a grandes bases de datos, libros electrónicos e interactivos sobre o desenrolo de dito Software, grupos de discusión sobre la problemática dos programas o temas de discusión concretos. No caso de elixir un programa para uso estatístico é claro que o R é o máis axeitado por distintas razóns que se describen a continuación.

Entre as características do programa R poden destacarse as seguintes: é unha linguaxe orientado a obxectos, a capacidade de combinar análises "pre-empaquetados" con análises ad-hoc, específicos para cada situación determinada, a posibilidade de crear gráficos de alta calidade, as extensións específicas a áreas emerxentes como a bioinformática, a bioestatística ou a minería de datos.

O R como alternativa para docencia

O paquete R é unha linguaxe de entorno de programación libre o que permite que os usuarios o estendan definindo as súas propias funcións. A comunidade científica internacional elixiu o R como a "lingua franca" da análise de datos.

O R é un programa baseado sobre comandos, no que se pode acceder a tódolos procedementos e opcións a través de sintaxis computacional. Foi oficialmente presentado en 1997 e como software libre rexese pola licencia xeral pública ("General Public License" o GPL) da fundación de software libre ("Free Software Foundation" o GNU, http://www.gnu.org/).

Aínda que inicialmente foi utilizado maioritariamente pola comunidade estatística como programa para traballos de investigación a alternativa para uso docente no está tan estendida na docencia sendo a súa maior competencia o Software comercial, como Matlab, Statgraphics, S-Plus, SPSS, Minitab, SAS o Statistica. Recentemente estanse a crear interfaces gráficas co gallo de favorecer e simplificar o seu manexo, como é o caso do Rcomander, que permite competir cos programas comerciais, o facelo máis amigable ao usuario. A súa versatilidade vese favorecida pola posibilidade de cargar diferentes librerías ou paquetes con finalidades específicas de cálculo gráfico. Ademais, ten a vantaxe de que está dispoñible para os sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

Outra das vantaxes para uso docente do programa R está, ademais de que ofrece un uso gratuíto dun software de primeiro nivel, a posibilidade dun maior control das análises, extensa documentación, e un ambiente de programación desenrolado para aplicacións estatísticas e con capacidade para ser usado noutras áreas cuantitativas de diversas disciplinas.

Interacción do programa R coas follas de cálculo

As follas de cálculo, especialmente o Excel, son una excelente ferramenta para crear ambientes de aprendizaxe que enriquezan a representación (modelado), comprensión e solución de problemas matemáticos. Desafortunadamente, a maioría de docentes e estudantes nos limitamos a utilizar só funcións básicas delas, como tabular información e realizar cálculos mediante formulas, descoñecendo que ofrece funcionalidades que van más alá da tabulación, cálculo de fórmulas e gráficas de datos, permitindo crear e facer uso de simulacións que posibilitan aos usuarios construír pontes entre as ideas intuitivas e os conceptos formais. Así, a

profesora Pamela Lewis, autora do libro "A Maxia da Folla de Cálculo", considera que esta es unha ferramenta de aprendizaxe poderosa argumentando ao seu favor que desenrola nos estudantes unha gran cantidade de habilidades que non se conseguen mediante o uso doutro paquete comercial. Debe recordarse que para os partidarios del Software libre existen follas de cálculo non comerciais como o caso do Gnumeric que incorporan menús de ferramentas estatísticas.

Como proposta para o uso nas aulas, o Rexcel ou o RGnumeric resulta un complemento que pode ser instalado de forma gratuíta, incorporando un novo menú na folla de cálculo que permite a utilización conxunta nun mesmo programa das dúas ferramentas o R e folla de cálculo. Esta interacción permite a comodidade de traballar cos datos mediante a folla de cálculo engadindo todo o poder dun programa estatístico como o R, o que permite executar R desde a propia folla de maneira que as opcións de R aparecen incorporadas nun menú dentro da propia folla de cálculo.

O Rexcel pode instalarse mediante o paquete RExcelInstaller de CRAN ou tamén dende a propia Web do programa (http://rcom.univie.ac.at/) tendo ademais a posibilidade de seguir un curso interactivo para o seu cómodo aprendizaxe (http://rcom.univie.ac.at/).

Opcionalmente pódense mostrar os menús do R Comander como un menú de folla de cálculo. Tamén é posible integrar os plugins de R Commander dentro destes menús.

Como exemplo de aplicación, pode verse o caso do uso do Rexcel para a estimación non paramétrica no traballo de Cao e Naya (2010). Neste caso proponse o uso destas ferramentas para a estimación da densidade e da regresión aplicándoas ao estudo dun conxunto de datos dispoñibles no propio R. Tamén é de interese, no contexto do ension de técnicas non paramétricas, os traballos da profesora Müller (2010).

Bibliografía

Cao, R. e Naya, S. (2010). The use of Statistical Software to teach nonparametric curve estimation: from Excel to R. ICOTS 8. Ljubliana. Slovenia.

Dalgaard, P. (2008). Introductory Statistics with R. New York: Springer.

Fox, J. (2005) The R Commander: "A Basic-Statistics Graphical User Interface to R", Journal of Statistical Software, 14.

Heiberger, R.M. y Neuwirth, E. (2009). R Through Excel: A Spreadsheet

Interface for Statistics, Data Analysis, and Graphics. Series: Use R. New York: Springer.

Lewis, P. (2006). Spreadsheet Magic. Washington: International Society for Technology in Education.

Lind, D.A., Marchal, W.G., and Wathen, S.A. (2004). Basic Statistics Using Excel For Office Xp. McGraw-Hill.

Müller, M. (2010). "Exploring data with non and semiparametric models", ICOTS 8. Müller's webpage: http://www.marlenemueller.de/nspm.html.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Teixeira, A., Rosa, A. e Calapez, T. (2009). "Statistical power analysis with microsoft excel: normal tests for one or two means as a prelude to using non-central distributions to calculate power", Journal of Statistical Education, 17.

Ugarte, M.D., Militino, A.F. e Arnholt, A.T. (2008). Probability and Statistics with R. Boca Raton: Chapman & Hall.

Verzani, J. (2005). Using R for Introductory Statistics. Boca Raton: Chapman & Hall.

Wild, C. e Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). International Statistical Review, 67 (3), 223 – 265.