El entorno estadístico *R*: ventajas de su uso en la docencia y la investigación

por Marcos RUIZ SOLER y Emelina LÓPEZ GONZÁLEZ
Universidad de Málaga

El uso de la estadística como herramienta de análisis en la investigación es una práctica consolidada en la mayoría de las disciplinas científicas. Es por ello que la inclusión de conocimientos estadísticos ocupa un lugar permanente en muy diversos planes de estudio, ya sea con el propósito de permitirnos analizar los datos de nuestras investigaciones o de posibilitarnos la comprensión de los datos de investigaciones publicadas. Pero conviene recordar que la difusión de los desarrollos estadísticos —especialmente de aquellos más complejos— sólo ha tenido lugar de forma paralela al desarrollo de programas informáticos que los implementaran. Por muy apropiado e interesante que resultara un análisis, si éste no se encontraba en algún programa, el modelo de análisis en cuestión prácticamente no se utilizaba (pues su coste temporal era excesivo). Esta situación —junto a otras de índole comercial que no vamos a tratar— ha provocado durante las últimas dos décadas el nacimiento de toda una familia de paquetes estadísticos (*SAS, S-plus, Systat, Minitab, BMDP, SPSS, EQS, GLIM, Statistica, Statgraphics, ViSta, LatentGold*, etc.) y con ella el problema de decidir cuál seleccionar para fines docentes y de investigación.

El propósito de este artículo no es realizar un análisis comparativo de software estadístico (para ello véase Burns, 2007), sino más bien presentar un entorno de trabajo que —aun habiendo conocido una amplia difusión en estos últimos años— todavía resulta bastante inusual dentro del contexto de las ciencias humanas y sociales. Nuestra experiencia docente con este software, así como la de otros docentes, nos ha convencido del gran potencial de esta herramienta informática dentro del campo de las investigaciones educati-

vas. Por tal motivo, comenzaremos exponiendo las características esenciales del entorno de computación *R* (Ihaka y Gentleman, 1996), cómo se desarrolló, dónde encontrarlo, para qué sirve, etc. Seguidamente realizaremos un rápido paseo por el mismo, mostrando mediante algunos ejemplos ciertas características propias de su modo de funcionamiento —un tanto alejadas de las de otros paquetes convencionales— y cómo su filosofía se adapta magníficamente a los diversos enfoques del análisis de datos. A continuación pasaremos a mostrar su aplicación en la docencia, resaltando aquellos aspectos que pueden facilitar la labor del docente y la comprensión del discente. También mostraremos cómo este entorno suministra herramientas inexistentes —o muy difícilmente disponibles— en los paquetes estadísticos habituales para el análisis y representación de datos. Tras este recorrido por sus características generales y aplicaciones en docencia e investigación, resumiremos las ventajas sobresalientes del entorno y miraremos hacia dónde se dirige su desarrollo futuro.

¿Qué es *R* y para qué sirve?

Tal vez una de las mejores síntesis sea la de Brian D. Ripley, del Departamento de Estadística de Oxford: "*R* es un avanzado sistema de computación de estadística con gráficos de enorme calidad que se encuentra disponible gratuitamente para la mayoría de plataformas [sistemas operativos]". En realidad *R* es eso y mucho más. Pero antes, veamos un poco sobre su historia, su origen y desarrollo hasta nuestros días.

El proyecto *R* nació en 1992, en Auckland (Nueva Zelanda), como un experimento de Ross Ihaka y Robert Gentleman, al intentar usar métodos de LISP —aquel lenguaje de programación de Inteligencia Artificial que dio origen a LOGO, tan popular entre educadores— como "banco de pruebas" con el que comprobar ideas estadísticas. Las primeras copias estuvieron disponibles en 1993 y Martin Mächler, del EHT de Zürich, fue uno de los primeros en utilizarlo y animó a Ihaka y Gentleman a que lo distribuyeran como software libre bajo los términos de GPL (*Free Software Foundation*, 1991), hecho que ocurriría en 1995. De este modo, el interés por *R* creció rápidamente y los errores y sugerencias reportados por correo electrónico llevaron a la actualización de las versiones cada cierto tiempo. Dado que las universidades de Nueva Zelanda cobran a sus departamentos en función del tráfico de internet que generan (y era mucho el tráfico generado), Mächler se ofreció voluntariamente en 1996 para que los servidores del EHT mantuvieran la lista de correo *R-testers*, la cual se subdividió en 1997 en tres listas: *R-announce, R-help y R-devel*, que siguen activas hasta hoy. Al ser muy numerosos los mensajes recibidos, se optó finalmente por dar acceso público al código de *R* para que cada usuario pudiera realizar sus modificaciones. Sin embargo, con la idea de garantizar una cierta uniformidad se precisaba de un equipo que supervisara el trabajo. Esto dio lugar al "core group", formado por un equipo de una docena de personas, algunas de ellas reconocidos estadísticos, que constantemente han ido incorporando mejoras.

Con el crecimiento de la comunidad de usuarios de *R* se hizo necesario crear un lugar desde el que organizar y hacer disponible todo el material relacionado con *R*, conocido como CRAN, acrónimo de "Comprehensive R Archives Network" (http://CRAN.R-project.org). CRAN es una colección de websites con idéntico material: el programa (base), las extensiones (programas específicos) y la documentación sobre *R*. Pero más que explicar aquí en detalle sus características, invitamos al lector a que entre en la página web y se entretenga en explorar sus contenidos; seguro que descubre más de lo que había imaginado.

¿Cómo conseguir e instalar *R*?

Como expondremos algo más adelante, una de las principales ventajas de *R* es que se trata de software gratuito. Por consiguiente, para descargarlo únicamente precisa de una conexión a Internet. La página principal (*Home page*) de *R* puede localizarse en la dirección http://www.r-project.org, donde el usuario encontrará el programa, documentación y material relacionado. Una vez en ella, únicamente debe seleccionar la plataforma de uso (Windows, Mac o Linux), elegir el subdirectorio base y, finalmente, pinchar sobre el fichero ejecutable R-2.5.1-win32.exe (el número que sigue a *R* puede variar, dado que se refiere a la última versión existente). Llegados a este punto comenzará a descargarse el programa. El tamaño del mismo es relativamente pequeño (no llega a 30 Mb), especialmente si consideramos todos los análisis que incorpora. Una vez sepamos cómo manejarnos con el programa, podremos añadir muy fácilmente módulos o extensiones (denominadas *contributed packages*) según nuestras necesidades particulares de análisis; en cualquier caso, el programa base contiene sobradamente todo lo que se suele precisar en una asignatura convencional con contenidos de estadística descriptiva e inferencial.

Ventajas de *R* respecto a otros programas estadísticos

1. *R* es un software de calidad. Se trata de un lenguaje de ordenador muy bien diseñado, tanto desde la perspectiva analítica (estadística) como computacional (informática), según el decir de los expertos (Fox y Andersen, 2005). Esta afirmación se encuentra avalada, en cierto modo, por haber recibido John Chambers el *Software System Award* en 1998 por parte de la *Association for Computing Machinery* —la principal asociación profesional de informáticos— como premio al desarrollo de *S* [1] (lenguaje sobre el que se sustenta *R*). Chambers es actualmente un miembro del equipo de desarrollo (*R core*); algunas de las figuras más destacadas en análisis estadístico se encuentran asimismo estrechamente relacionadas con el desarrollo de *R* (por ejemplo, Douglas Bates, Brian Ripley o Luke Tierney).

2. *R* es gratuito. La asociación entre calidad-coste parece estar tan consolidada en nuestra mente que ante un producto gratuito surge automáticamente la sospecha: ¿será realmente bueno? Afortunadamente en el mundo de la informá-

tica esta asociación se ha debilitado como consecuencia de iniciativas de desarrollo de código abierto de excelente calidad. *R* se encuentra entre ese conjunto y no tiene nada que envidiar -más bien lo contrario- a otros productos comerciales de elevado precio [2]. Mas no solamente el software es gratuito, sino también una gran documentación, tanto general como específica (véase el apartado de *Contributed documentation*, con manuales en inglés y otros muchos idiomas: francés, español, italiano, polaco, húngaro o incluso croata o vietnamita, lo que es una muestra indirecta de su gran difusión). Disponer de programa y documentación de elevada calidad y a coste cero es con *R* una realidad.

3. *R* es multiplataforma. Esto significa que puede utilizarse prácticamente en cualquier tipo de ordenador. Hace algunos años los usuarios de Windows™ no tenían acceso a los programas de MacOS™, ni viceversa. Aunque este problema se ha solucionado en gran medida, parece que los usuarios de Linux -otro sistema operativo en creciente aumento- han de conformarse con otros programas tan sólo similares a los desarrollados para las plataformas anteriores. El entorno *R* no impone restricciones en este sentido y dispone de las mismas opciones en todas las plataformas, permitiendo así el acceso a usuarios con distintas plataformas operativas.

4. *R* admite programación. En realidad *R* incorpora un verdadero lenguaje de programación y no simplemente un conjunto de comandos que permite —siguiendo unas cuantas reglas sintácticas— formar ficheros de instrucciones (tal cual sucede en algunos paquetes estadísticos). Esto ofrece unas posibilidades enormes, dado que potencialmente no existen limitaciones en el tipo de operaciones a realizar. En ocasiones existen tareas o procedimientos que, sin ser específicamente de análisis estadístico, resultan necesarios en la investigación; éste es el caso, por ejemplo, de la generación de secuencias de contrabalanceo en diseños experimentales con variables intrasujeto, con el fin de neutralizar posibles efectos de error progresivo (fatiga, aprendizaje o influencias residuales). La elaboración de un programa en *R* fácilmente puede automatizar esta tarea que resultaría muy laboriosa cuando se dispone de numerosas condiciones (Ruiz Soler, 2005).

5. *R* crece continuamente. El crecimiento de *R* resulta espectacular, pues desde su primera versión aparecida el 29 de febrero de 2000 hasta la fecha se han creado nuevas versiones prácticamente con periodicidad mensual. Probablemente ello se deba a que muchas de las posibilidades que ofrece el programa provienen de las necesidades de enseñanza de unos pocos (Ripley, 2003). La lista de campos de conocimiento en los que se aplica *R* parece no tener fin y sigue aumentando (Burns, 2007). En el momento presente *R* dispone de más de 400 "contributed packages", que son módulos de diversas áreas. En la Tabla 1 pueden verse algunos de los módulos actualmente disponibles que pueden ser de interés para el investigador en ciencias de la educación o en ciencias del comportamiento.

TABLA 1: *Algunos paquetes de extensión del entorno R [3]*

	Paquete	Contenidos
Estadística básica		
	BSDA	Basic Statistics and Data Analysis
	distr	Object orientated implementation of distributions
	IPSUR	Introduction to Probability and Statistics Using R
	IS wR	Introductory Statistics with R
	moments	Moments, cumulants, skewness, kurtosis and related tests
	nortest	Tests for Normality
	npmc	Nonparametric Multiple Comparisons
	prettyR	Pretty descriptive stats.
Est. avanzada y multivariante		
	asuR	Advanced statistics using R
	glmmAK	Generalized Linear Mixed Models
	gnm	Generalized Nonlinear Models
	mvoutlier	Multivariate outlier detection based on robust methods
	robustbase	Basic Robust Statistics
	sem	Structural Equation Models
	tsDyn	Time series analysis based on dynamical systems theory
Regresión avanzada		
	dynlm	Dynamic Linear Regression
	lpridge	Local Polynomial (Ridge) Regression
	nlstools	Tools for nonlinear regression diagnostics
	quantreg	Quantile Regression
Datos categóricos		
	cat	Analysis of categorical-variable datasets with missing values
	catspace	Special models for categorical variables
	Design	Design Package
	drm	Regression and association models for repeated categorical data
	gllm	Generalised log-linear model
Experimentación		
	conf.design	Construction of factorial designs
	blockrand	Randomization for block random clinical trials
	experiment	experiment: R package for designing and analyzing randomized experiments
	psyphy	Functions for analyzing psychophysical data in R
	ROCR	Visualizing the performance of scoring classifiers
Psicometría		
	concord	Concordance and reliability
	eRm	Extended Rasch Modeling.

	irr	Various Coefficients of Interrater Reliability and Agreement
	MiscPsycho	Miscellaneous Psychometrics
	polycor	Polychoric and Polyserial Correlations
	Proxy	Distance and Similarity Measures
	psy	Various procedures used in psychometry
	psychometric	Procedures for Personality and Psychological Research
Gráficos		
	gplots	Various R programming tools for plotting data
	graph	graph: A package to handle graph data structures
	gRbase	A package for graphical modelling in R
	mimR	A package for graphical modelling in R
	misc3d	Miscellaneous 3D Plots
	RGraphics	Data and Functions from the book R Graphics
	scatterplot3d	3D Scatter Plot
	vcd	Visualizing Categorical Data
Varios		
	arm	Data Analysis Using Regression and Multilevel/Hierarchical Models
	boost	Boosting Methods for Real and Simulated Data
	boot	Boosting Methods for Real and Simulated Data
	bootstrap	Functions for the Book "An Introduction to the Bootstrap"
	corpora	Statistics for corpus linguists
	Epi	A package for statistical analysis in epidemiology.
	epitools	Epidemiology Tools
	FactoMineR	Factor Analysis and Data Mining with R
	fdim	Functions for calculating fractal dimension.
	gmodels	Various R programming tools for model fitting
	meta	Meta-Analysis
	Rcmdr	R Commander
	rmeta	Meta-analysis
	sampling	Survey Sampling
	sna	Tools for Social Network Analysis
	survey	analysis of complex survey samples
	zipfR	Statistical models for word frequency distributions
Utilidades		
	colorspace	Colorspace Manipulation
	DBI	R Database Interface
	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat,
	gdata	Various R programming tools for data manipulation
	gtools	Various R programming tools
	R.matlab	Read and write of MAT files together with R-to-Matlab connectivity
	R2HTML	HTML exportation for R objects
	TeachingDemos	Rcmdr Teaching Demos Plug-In
	xtable	Export tables to LaTeX or HTML

6. *R* dibuja excelentes gráficos. Posiblemente sea ésta una de las características más sobresalientes de este entorno, pues la variedad y calidad de gráficos que pueden conseguirse no tiene competencia alguna (para quienes deseen comprobar nuestra afirmación pueden consultar la página web: http://addictedtor.free.fr/graphiques/, en la que encontrarán cientos de gráficos). Para los científicos sociales creemos que son especialmente importantes los gráficos para datos categóricos, aunque lamentablemente todavía no se han difundido demasiado, pese a sus destacadas ventajas para la interpretación de datos procedentes de tablas de contingencia (Friendly y Ruiz-Soler, 2000; Ruiz-Soler y Pérez-García, 2003).

7. *R* facilita pensar matemáticamente. Una interfaz basada en comandos obliga a los estudiantes a pensar sobre lo que están haciendo (Ripley, 2003) y diversos autores han mostrado cómo puede emplearse *R* en la enseñanza de cursos introductorios de estadística (por ejemplo, Nolan y Speed, 2000); incluso existen trabajos que muestran cómo usar *R* como herramienta de trabajo para la exploración de conceptos estadísticos (Horton, Brown y Qian, 2004). En el apartado sobre la aplicación de *R* en la docencia expondremos algunos ejemplos que ilustrarán esta idea.

Un rápido paseo por el entorno *R*

A diferencia de otros paquetes estadísticos que están diseñados para trabajar con interfaces de usuario gráficos ("GUIs"), *R* es un entorno de comandos. Por tanto, iniciarse en *R* puede suponer un esfuerzo mayor que manejar un paquete estadístico convencional, pero el esfuerzo se ve muy rápidamente recompensado. Un interaface de comandos, sin embargo, tiene algunas importantes ventajas para los usuarios habituales, a saber (Fox y Andersen, 2005): 1) es más fácil corregir, modificar y repetir análisis (aunque programas como SPSS posibilitan también trabajar con comandos, en la práctica la mayoría de usuarios no emplean esa posibilidad cuando pueden usar el sistema de ventanas); 2) resulta más natural crear un registro permanente del trabajo realizado que puede combinarse de un modo simple con textos explicativos (Leisch, 2002); 3) frecuentemente los GUIs para sistemas estadísticos complejos son, o bien incompletos (requieren finalmente el uso de comandos) o bien 'bizantinos' (requieren la selección de un sinfín de menús, ventanas y opciones). De todos modos, conviene tener en cuenta el tipo de usuario que vaya a realizar el análisis. Como afirma Burns (2007), podemos dicotomizar los usuarios de la estadística en dos grandes grupos: 1) los de "necesito una prueba estadística que me dé una $p < .05$" y 2) los de "me encantaría saber qué dicen mis datos". Si pertenece al primer grupo, *R* no está pensado para Vd. Si pertenece al segundo grupo, *R* es muy probablemente la solución que estaba esperando.

Tal vez la mejor forma de explicar qué es *R* sea viendo cómo funciona mediante ejemplos reales. No obstante, conviene

advertir al lector que el objetivo aquí no es presentar una introducción a *R* sino simplemente exponer algunos conceptos básicos de su modo de operar que permitan formarse una idea sobre el estilo de trabajo del entorno y ofrecer, asimismo, información suficiente para comprender los ejemplos posteriores.

El entorno *R* funciona interpretando comandos, es decir, que los ejecuta inmediatamente (después de 'Enter'). Esto resulta muy claro con las operaciones aritméticas elementales.

```
> 2007+3    # año de inicio general de los estudios según el EEES
  2010
> 2010-2007  # años hasta la implantación de los estudios de grado
  3
> 6*25   # número de horas de una asignatura ECTS
  150
> 60/10   # ¿número de créditos de las asignaturas del futuro grado?
  6
2^10  # bits de información de un fichero de 1 kbyte
1024
```

Generar secuencias de números es tremendamente fácil.

```
> seq(1,10)
  1 2 3 4 5 6 7 8 9 10
> 1:10   #de forma abreviada
  1 2 3 4 5 6 7 8 9 10
> seq(0,100,5)   # secuencias crecientes en más de una unidad
[1]   0   5  10  15  20  25  30  35  40  45  50  55  60  65  70  75
[17]  80  85  90  95 100 105 110 115 120 125 130 135 140 145 150 155
[33] 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235
[49] 240 245 250 255 260 265 270 275 280 285 290 295 300
> 100:90   # generando una secuencia decreciente
  100  99  98  97  96  95  94  93  92  91  90
```

Pero para que todos los resultados que se generen no se "pierdan", conviene almacenarlos en variables (vectores u otras estructuras de datos). La sintaxis es muy simple: crear un nombre (por ejemplo, 'edad'), escribir el símbolo de asignación (<-), después la letra *c* (de concatenar) y, finalmente, entre paréntesis los valores (números). Escrito es más fácil:

```
edad <- c(18, 19, 20, 17, 18, 21, 19, 18, 43, 45).
```

Ahora ya podemos trabajar con la variable 'edad'. Supongamos que desea-

mos saber el número aproximado de días que ha vivido cada una de estas personas; para ello escribiremos:

```
> edad*365
  6570  6935  7300  6205  6570  7665  6935  6570 15695 16425
```

Lectura de datos

Cuando se trata de pequeños conjuntos de datos, éstos pueden introducirse directamente por teclado o por medio de un editor tipo hoja de cálculo del propio *R*; sin embargo, suele ser más habitual leer los datos desde ficheros externos: binarios, del portapapeles, de internet, de otros paquetes o incluso de otros dispositivos externos. Prácticamente no existe limitación en el tipo de ficheros a leer e incluso existe un *package* dedicado a importar datos de la mayoría de paquetes estadísticos (*foreign*). Veamos un ejemplo:

Técnicas.Estudio <read.table ("c:/MisDatos/Notas.txt").

En este ejemplo, 'Técnicas.Estudio' es el nombre de la estructura en la que se van a almacenar los datos; no existen apenas limitaciones en el modo de asignar los nombres a las variables: pueden tener una extensión considerable, admiten minúsculas y mayúsculas, permiten la inclusión de tildes y puntos separadores, etc. Esto facilita disponer de nombres fácilmente comprensibles sin tener que asignar etiquetas a cada uno de ellos, debido a las limitaciones habituales en los paquetes estadísticos convencionales que no permiten superar un máximo de ocho caracteres.

Aplicación de *R* en la docencia

El programa *R* podría ser —que lo es— magnífico, pero si no pensáramos que una de sus grandes ventajas es que puede aportar a la docencia elementos diferenciales respecto a otros productos, entonces no tendría demasiado sentido dedicar un artículo en una revista como ésta. Es por ello que en este apartado comenzaremos mostrando cómo solicitar mediante *R* algunos cómputos estadísticos simples, para después pasar a ver algunos procedimientos de especial interés en la explicación de conceptos estadísticos.

La elección de un paquete estadístico es normalmente una decisión institucional motivada principalmente por razones económicas. Pero esta decisión es importante y debería estar guiada por una serie de criterios (Dell'Omodarme y Valle, 2006): 1) el programa tendría que ser accesible al alumnado, es decir, permitir una copia para ser utilizada en sus ordenadores personales; esto supone la necesidad de que las licencias sean de bajo coste (o incluso mejor: gratuitas); 2) el programa debería permitir su aprendizaje en un periodo de tiempo razonablemente breve; 3) debería existir documentación substancial sobre el programa que se encontrara fácilmente disponible (tutoriales *on-line*, manuales, foros de discusión sobre dudas,

etc.); 4) el programa debería permitir realizar análisis de cualquier tipo, ofreciendo considerables herramientas estadísticas y gráficas; 5) el programa debería funcionar sobre diferentes sistemas operativos (Windows, GNU/Linux, etc.).

De las características mencionadas, algunos de los programas más usuales en ciencias sociales cumplen sólo parcialmente con ellas. Así, por ejemplo, el SPSS —por citar sólo uno de los programas más habituales— cumple dudosamente con el primer criterio, dado que no todas las Universidades disponen de licencias para ofrecer copias a sus alumnos; el segundo criterio se cumple aparentemente, dado que no es necesario mucho más que unas pocas horas para iniciarse en su uso, pero el viaje a través de innumerables ventanas dificulta considerablemente recordar en ocasiones subsiguientes cómo realizar una determinada tabla de contingencia o un gráfico no convencional; el tercer criterio se cumple en teoría (existe abundante bibliografía), pero los escasos recursos económicos del alumno normalmente se orientan hacia otro tipo de libros (más sustantivos que metodológicos); el cuarto criterio intenta cumplirse (hay que reconocer el esfuerzo por incorporar en cada versión nuevos análisis), pero sigue existiendo un desfase temporal todavía grande entre lo que existe y lo que hay; el quinto criterio solamente ha comenzado a cumplirse muy recientemente.

Imaginemos ahora un profesor universitario que desea explicar algunos conceptos elementales de estadística descriptiva. Para ello recurre a un ejemplo simple: las notas finales de curso de los alumnos de una clase. Las introduce y después obtiene los dos valores estadísticos que caracterizan a ese conjunto de valores:

```
> notas <- c(5,7,4,8,3,5,1,7,5,8,10)
> mean(notas)
[1] 5.727273
> sd(notas)
[1] 2.572583
```

También podría estar interesado en solicitar otros datos,

```
> min(notas)
  1
> max(notas)
  10
```

o la correlación entre la nota en Matemáticas y la nota en Física (como supondrá el lector, las notas de ambas asignaturas se han introducido ya previamente, pero no las mostramos en este ejemplo por su extensión, dado que se trata de 100 valores para cada una de las variables):

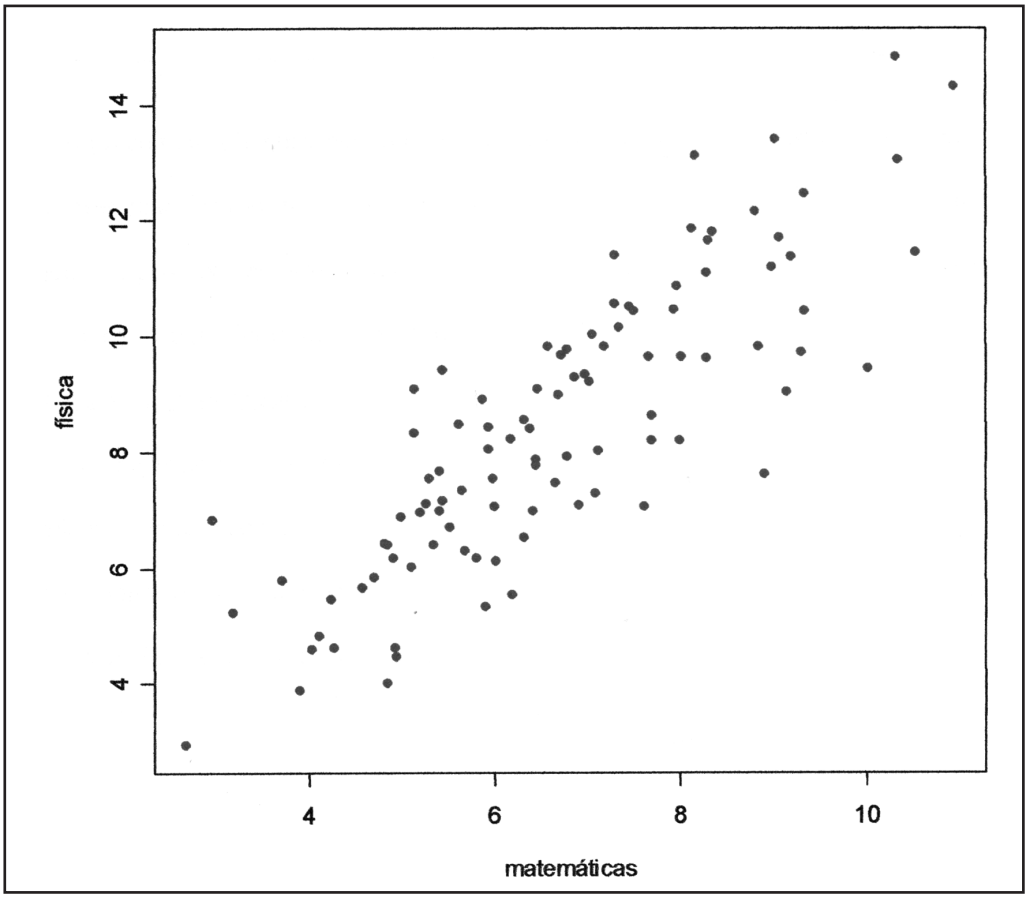
```
> matemáticas <- c(notas)
> cor(matemáticas, física)
  0.873451
```

Puede que incluso desee una representación gráfica, es decir, un diagrama de dispersión. En este caso, al tratarse de dos variables cuantitativas métricas tan sólo debería escribir:

```
plot(matemáticas,física)
```

Como puede contemplarse en los ejemplos anteriores, una primera ventaja de este entorno es que las salidas son

GRÁFICO 1: *Ejemplo de diagrama de dispersión*



—digámoslo así— muy limpias. No aparece nada innecesario, tan sólo aquello que se ha solicitado, sin complicados ni extensos listados de resultados con datos que enmascaran en muchas ocasiones la información buscada. Esto no significa que haya que teclear mucho más: *R* dispone de comandos capaces de aglutinar conjuntos de resultados. Veámoslo.

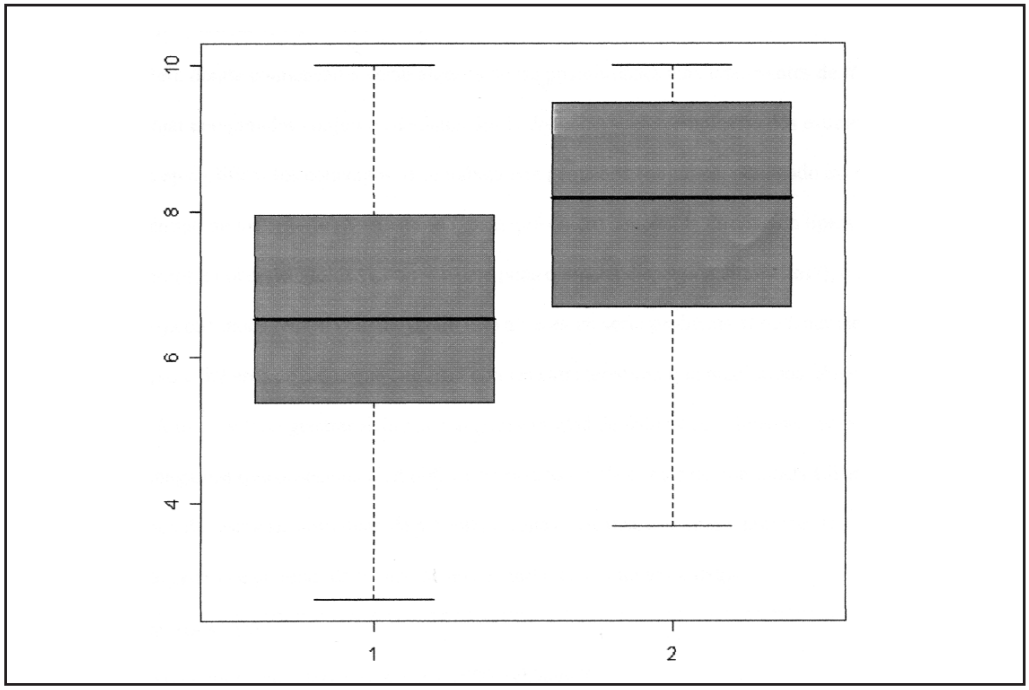
```
summary(notas)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.500   5.000   5.727   7.500  10.000
```

Aquí con un solo comando obtenemos información de los valores mínimo y máximo, de los tres cuartiles (por tanto, de la mediana) y de la media aritmética. Esto es sólo un ejemplo, pero existen otros diversos comandos que también ofrecen conjuntos de resultados (p. ej. para cualquier modelo de regresión, tanto lineal como no lineal).

Datos de un experimento

Supongamos que queremos representar gráficamente los datos de un experimento sobre didáctica de las matemáticas, en el que se ha manipulado la técnica de enseñanza (tradicional vs. innovadora). Para ello deberíamos teclear simplemente plot (técnica, notas) [asumiendo que ésos son los nombres de las variables con las que hemos trabajado].

GRÁFICO 2: *Ejemplo de gráfico de cajas*



¡Interesante!: aunque el comando es el mismo que hemos empleado con anterioridad, el programa selecciona automáticamente un gráfico de cajas (*boxplot*) porque "sabe" que la primera variable (técnica) es cualitativa (en realidad esta información la conoce porque dicha variable se ha definido previamente como factor). No ha sido necesario cambiar el nombre del gráfico (aunque también podríamos forzar al programa a ello mediante el correspondiente comando). Esto sucede igualmente con otros comandos (como *summary*) que ofrecen resultados distintos dependiendo del objeto al que se aplican, lo que supone, en la práctica, que son realmente muy pocos los comandos que es necesario conocer, aunque éstos se apliquen sobre objetos distintos.

Pero donde comienzan a verse algunas de las posibilidades más interesantes de *R* es al trabajar con grandes conjuntos de datos. En la docencia suelen emplearse dos estrategias para ejemplificar los conceptos: o se trabaja con pequeños conjuntos, perdiendo así en parte el sentido de los criterios estadísticos (¿para qué quiero conocer la desviación típica de 10 números, si observándolos ya veo si existe poca o mucha dispersión entre ellos?), o se trabaja con datos reales, constituyendo esto a veces un serio problema al no tener siempre acceso a ficheros de datos que cumplan con las características que precisamos. Pues bien, con *R* es muy fácil generar ficheros con gran cantidad de datos y con diversas restricciones. Supongamos que deseamos ejemplificar cómo analizar los datos de una escala Likert de 5 puntos de respuesta administrada a 1000 personas. Las instrucciones siguientes se encargarían de generar, en menos de un segundo, tal conjunto de datos.

```
> valores <- 1:5
> respuestas <- sample(valores, size=1000, replace=T)
> respuestas
   [1] 2 5 2 5 5 2 4 1 4 2 5 3 1 4 4 4 4 2 4 4 5 2 2 4 4 3 2 5 2 1 5 3
  [33] 4 1 1 2 2 1 2 2 2 4 1 2 4 3 5 1 5 5 2 4 4 2 4 4 3 4 4 5 1 2 1 1
  [65] 4 2 1 5 3 1 3 2 5 5 1 5 4 2 4 4 5 4 5 4 3 2 2 4 2 1 2 4 4 1 2 2
  ......
 [993] 2 5 1 2 4 2 5 5
```

Como se observa, los datos se generan así inmediatamente y se almacenan en la variable 'respuestas'. Al preguntar al programa sobre 'respuestas' nos aparece el listado con los valores (el número que aparece entre corchetes es un indicador de la posición que ocupa el primer elemento de cada fila dentro del vector).

Estadística inferencial: estimación

Siguiendo con la ventaja que acabamos de ver, es posible generar de manera tremendamente fácil grandes conjuntos de datos que se ajusten a los parámetros de una determinada distribución. Supongamos que deseamos —con fines didácticos— generar una distribución de

los resultados de aplicar un test de inteligencia a 1500 niños escolarizados. Tan sólo deberíamos teclear lo siguiente:

rnorm (1500, 100, 15).

En este caso se generaría un vector con 1500 elementos, con media 100 y desviación típica 15. Pero igualmente fácil sería su uso con otras conocidas distribuciones, como la logística (*rlogis*), la multinomial (*rmultinom*), la de Poisson (*rpoi*), la geométrica (*rgeom*), etc. Enseñar conceptos típicos de cualquier curso introductorio a la Estadística, tales como las proporciones y áreas de la curva normal, resulta así mucho más claro, pues por cada caso/ejemplo expuesto podemos ilustrar la idea inmediatamente con la representación gráfica correspondiente. Esto es de especial interés ante las preguntas que formulan ciertos alumnos, pues la respuesta de "imagínate que la zona crítica se desplaza hacia..." se convierte en una respuesta precisa y clara, con los valores exactos y su representación gráfica adecuada.

Asimismo, se abren infinidad de posibilidades para mostrar interactivamente los principios del muestreo aleatorio sin tener que acudir a tablas de números aleatorios o utilizar siempre los mismos ejemplos "prefabricados" de un manual de texto. Conceptos como la distribución muestral de la media, estimaciones puntuales o por intervalos, pueden ilustrarse muy bien mediante el empleo de *R* (algunos manuales de Estadística que emplean *R* en la docencia muestran esto muy claramente).

Estadística inferencial: contrastes

Tal vez lo mejor sea comenzar con un ejemplo. Imaginemos que un pedagogo desea averiguar hasta qué punto existe una relación entre el número de horas realizando ejercicios de análisis de datos y el rendimiento académico en la asignatura correspondiente de Estadística o Metodología. Para ello solicita a sus alumnos que realicen un autorregistro durante el cuatrimestre. Tras el examen final de la asignatura, el análisis en *R* podría ser éste:

```
> mi.modelo <- lm(Estadística~1+horas.estudio)
> mi.modelo
Call:
lm(formula = Estadística ~ 1 + horas.estudio)
Coefficients:
  (Intercept)  horas.estudio
      -1.5855         0.1788
```

¿Qué significa lo que acabamos de escribir? En este caso resulta claro que los valores de nuestras variables se han almacenado en 'Estadística' y 'horas.estudio'. El comando lm es el empleado para especificar que vamos a probar un modelo lineal. El símbolo ~ indica que a partir de ahí comienza la especificación del modelo y el 1 que deseamos que se calcule el valor del punto de corte (b_0). Aunque con esto sería suficiente para el cálculo, como no deseamos que los valores se pierdan tras el análisis, almacenamos éstos en un "objeto" que se nos ha ocurrido llamar 'mi.modelo'. Una vez más, observamos la pulcritud de la salida: únicamente nos encontramos con los valores de los dos parámetros estimados: b_0 y b_1.

Sin embargo, si necesitamos más información siempre es posible solicitarla con algún comando adicional. En este caso, mediante el comando *summary* obtenemos información de los intervalos de confianza de los parámetros y su correspondiente significación estadística, el coeficiente de determinación, el valor del estadístico *F*, etc.

Ésta es una característica fundamental de *R*, tal y como se señala en el manual *Introducción a R* del R Developement Core Team (2004, 3):

"Existe una diferencia fundamental en la filosofía que subyace a *R* y a la de otros sistemas estadísticos. En *R* un análisis estadístico se realiza en una serie de pasos, con unos resultados intermedios que se van almacenando como objetos para ser observados o analizados posteriormente, produciendo unas salidas mínimas. Sin embargo, en SAS o SPSS se obtendría de modo inmediato una salida copiosa para cualquier análisis".

```
> summary(mi.modelo)  # para obtener más información sobre el modelo
Call:
lm(formula = Estadística ~ 1 + horas.estudio)
Residuals:
    Min      1Q  Median      3Q     Max
-1.74528 -0.67885  0.09253  0.58263  2.10913
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.58547    0.51670  -3.068  0.00296  **
horas.estudio  0.17880    0.01016  17.602  < 2e-16 ***
—-
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.862 on 78 degrees of freedom
Multiple R-Squared: 0.7989,     Adjusted R-squared: 0.7963
F-statistic: 309.8 on 1 and 78 DF,  p-value: < 2.2e-16
```

En el Gráfico 3 representamos el modelo obtenido resultante del ejemplo.

Representaciones gráficas

Es precisamente éste uno de los aspectos donde *R* parece no tener prácticamente limitaciones. Si esta afirmación resultara exagerada, recomendamos la visita y exploración de los contenidos de la siguiente página web:

http://addictedtor.free.fr/graphiques/allgraph.php

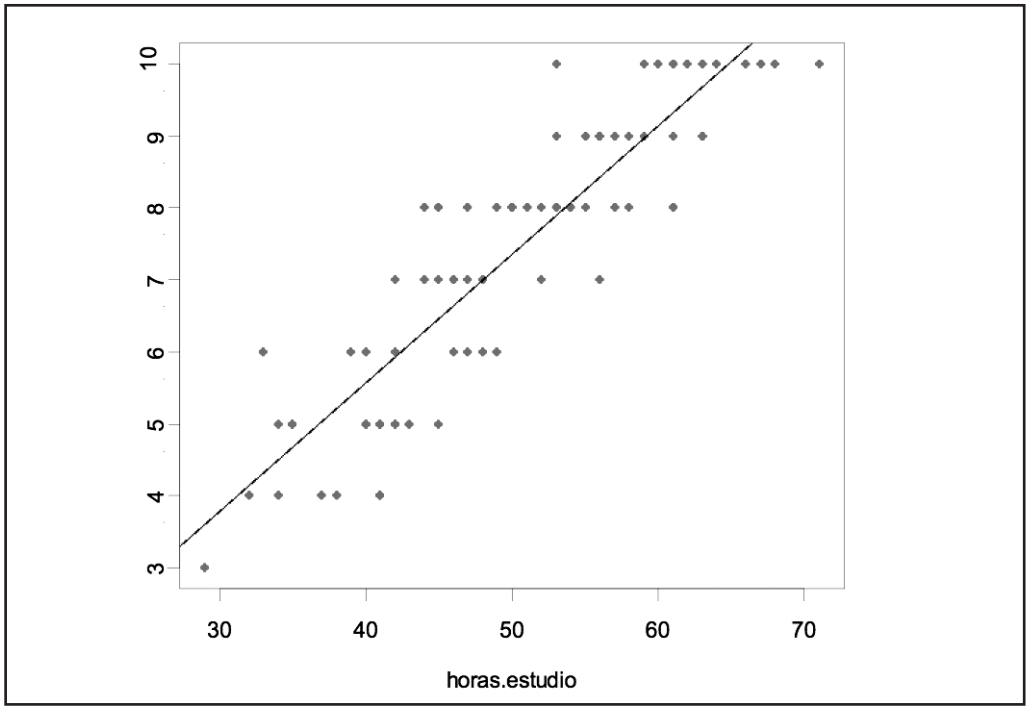
Muchas de estas gráficas son todavía desconocidas por numerosos investigadores, pero su aplicación en Ciencias Sociales y Humanas podría resultar de gran ayuda en la interpretación de resultados, especialmente aquellas referidas a datos categóricos (Friendly y Ruiz-Soler, 2000). Dado que una exposición de ellas excedería los propósitos de este artículo, dejamos su presentación para otro trabajo.

Finalmente, señalar que *R* dispone, además de su presentación habitual de uso en forma de comandos, una versión en forma de menús con ventanas, muy organizada, y que puede resultar muy interesante para la docencia (corresponde al paquete *Rcmdr*).

Aplicación de *R* en la investigación

Si aquello de lo que uno dispone ya sirve para conseguir sus fines, entonces podría parecer no muy pragmático cualquier cambio. Sin embargo, esta resisten-

GRÁFICO 3: *Ejemplo de ajuste del modelo lineal de regresión a los datos observados*



cia puede deberse quizá a que uno se ha acostumbrado a mirar las cosas de cierta manera. Existe una frase que refleja metafóricamente muy bien esta idea: "Cuando uno sólo dispone de un martillo, todos los problemas tienen forma de clavo".

Hemos visto algunas de las ventajas de *R* en la docencia, pero sería injusto no señalar, asimismo, su gran utilidad en la investigación. Y esto, al menos, por tres razones. En primer lugar, porque es posible explorar mucho mejor los datos, pues además de las técnicas estadísticas habituales destinadas a ello, posibilita el manejo de la matriz de datos de múltiples maneras, dado que existe una amplia gama de comandos para manejarla. En segundo lugar, porque existe una gran variedad de paquetes muy específicos (actualmente más de 400) que permiten adaptar *R* a casi cualquier campo (hay aplicaciones para genómica, geología, redes neuronales, epidemiología, econometría, psicometría, etc.), donde se incluyen comandos con operaciones de especial interés en estos campos. Aunque la multitud de paquetes de extensión disponibles puede resultar apabullante, en realidad debería observarse como la posibilidad de encontrar un paquete que se ajuste muy específicamente a nuestras necesidades de investigación. Y en tercer lugar, porque ofrece grandes posibilidades en la importación y exportación, no sólo de ficheros de datos, sino también de resultados a tablas, en HTML, Latex, MatLab, etc., lo cual facilita la difusión de los resultados obtenidos.

Conclusión

A través de los apartados anteriores hemos visto que *R* no solamente es capaz de dar cumplimiento a nuestras necesidades en la docencia y en la investigación, sino que abre nuevas posibilidades en la forma de trabajar que difícilmente podemos encontrar en otros paquetes estadísticos. Por tal motivo, consideramos que sería muy interesante un mayor uso en el ámbito educativo. En este sentido, algunas obras recientes pueden ayudar en su introducción, como los trabajos de Ugarte y Militino (2002); Chambers y Hastie (1997); Venables y Ripley (2003); Paradis (2002); Venables y cols. (2000); Muenchen (2007); Heirberger y Holland (2004); Dalgaard (2002) o Versan (2000).

Dirección de los autores: Marcos Ruiz Soler. Departamento de Psicobiología y Metodología de las CC. Comportamiento. Facultad de Psicología. Campus de Teatinos. 29071 Málaga. ruizsoler@uma.es. Emelina López González. Departamento de Métodos de Investigación e Innovación Educativa. Facultad de CC. Educación. Campus de Teatinos. 29071 Málaga. emelopez@uma.es.

Fecha de recepción de la versión definitiva de este artículo: 5.V.2008

Notas

[1] *R* es un entorno de código abierto para la computación estadística no diferente del sistema *S* desarrollado en los Laboratorios Bell (AT&T), el cual es la base del sistema comercial S-Plus (Ripley, 2003). Es por ello que toda la documentación existente para S-Plus es utilizable por los usuarios de *R* (salvo mínimas diferencias en algunos comandos).

[2] El entorno *R* no es simplemente un programa con posibilidades didácticas interesantes sino una genuina opción para el procesamiento de análisis de datos a gran escala. En este sentido, ya ha sido usado para la estimación y predicción de voto en las elecciones de Austria y de Reino Unido, lo que demuestra la robustez del programa y la confianza generada.

[3] El orden dentro de cada grupo no indica prioridad (siguen la ordenación alfabética); los más importantes se señalan con un asterisco. Los contenidos de esta tabla deben contemplarse únicamente como un ejemplo ilustrativo de la variedad de recursos disponibles para realizar distintos tipos de análisis de datos; en ningún caso se trata de una clasificación exhaustiva, pues su inclusión excedería los propósitos introductorios de este trabajo.

Bibliografía

BURNS, P. (2007) *R Relative to Statistical Packages* [Stata, SAS, and SPSS]. Technical Report 1. (UCLA Academic Technology Services).

CHAMBERS, J. M. y HASTIE, T. J. (1997) *Statistical models in S*. (London, Chapman & Hall).

DALGAARD, P. (2002) *Introductory Statistics with R* (New York, Springer).

DELL'OMODARME, M. y VALLE, G. (2006) Teaching Statistics with Excel and R. Ver http://arxiv.org/ftp/physics/papers/0601/0601083.pdf (Consultado el 21.I.2008).

FRIENDLY, M. y RUIZ-SOLER, M. (2000) Nuevos procedimientos gráficos para datos categóricos: de la representación a la cognición, en OÑATE, GARCÍA-SICILIA y RAMALLO (coords.) *Métodos numéricos en Ciencias Sociales* (Barcelona, CIMNE) pp. 83-96.

HEIBERGER, R. M. y HOLLAND, B. (2004) *Statistical Analysis and Data Display. An Intermediate course with Examples in S-Plus, R, and SAS* (New York, Springer).

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

LEISCH, F. (2002) Mixing R and LaTex, *R News,* 2:3, pp. 28-31.

MUENCHEN, B. (2007) *R for SAS y SPSS users* (PDF) (SAS Institute; SPSS, Inc. y Mathworks, Inc).

NOLAN, D. y SPEED, T. P. (2000) *Stat Labs: Mathematical Statistics Through Applications*. Springer. Ver: http://www.stat.Berkeley.edu/users/statlabs/ (Consultado el 20.I.2008).

NORTON, N. J.; BROWN, E. R. y QUIAN, L. (2004) Use of *R* as a Toolbox for Mathematical Statistics Exploration, *The American Statistician,* 58:4, pp. 343-357.

PARADIS, E. (2002) *R para principiantes*. Versión en español de "*R* for Beginners", traducido por AHUMADA, J. A. Ver: http://cran.r-project.org/doc/rdebuts_es.pdf (Consultado el 20.I.2008).

R DEVELOPMENT CORE TEAM (2004) *R: A language and environment for statistical computing* (Vienna, *R* Foundation for Statistical Computing).

RUIZ-SOLER, M. (2005) *Contrabalanceo de secuencias de estímulos. Una implementación de funciones en R.* (Granada, Actas del IX Congreso de Metodología de las Ciencias Sociales y de la Salud).

RUIZ-SOLER, M. y PÉREZ-GARCIA, A. (2003) *Representaciones gráficas para tablas de contingencia* (Valencia, Actas del VIII Congreso de Metodología de las Ciencias Sociales y de la Salud).

UGARTE, M. D. y MILITINO, A. F. (2002) *Estadística Aplicada con S-Plus* (2ª ed.) (Pamplona, Universidad Pública de Navarra).

VENABLES, W. N. y RIPLEY, B. D. (2003) *Modern Applied Statistics with S* (4ª ed.) (New York, Springer-Verlag).

VENABLES, B.; SMITH, D.; GENTLEMAN, R.; IHAKA, R. y MACHLER, M. (2000) *Notas sobre R. Un entorno de programación para análisis de datos y gráficos*. Traducción de GONZÁLEZ, A. y GONZÁLEZ, S. Ver http://cran.r-project.org/doc/ (Consultado el 20.I.2008).

VERSAN, J. (2004) *Using R for introductory statistics* (London, Taylor & Francis).

Resumen:
El entorno estadístico R: ventajas de su uso en la docencia y la investigación

Se presenta *R* como un programa informático de análisis de datos que puede ser de gran interés, tanto en su uso docente como de investigación. Tras una breve introducción histórica, se explican los elementos diferenciales del mismo que suponen ventajas respecto a otros paquetes estadísticos convencionales. A continuación ofrecemos una visión general

sobre el modo de operar del programa, ejemplificándolo a partir de tareas habituales en la docencia (cálculo de índices descriptivos, estimación de parámetros, contrastes estadísticos, etc.). Por último, se avanzan algunas de sus ventajas en la investigación y se concluye indicando el modo de introducirse en su uso.

Descriptores: *R*; educación estadística; análisis de datos; análisis estadístico; modelado estadístico; enseñanza de la estadística.

Summary: The statistical environment R: some advantages for teaching and research

A computer program is introduced for data analysis, which could be very interesting in order to be used as a teaching or research tool. After a short historical introduction, the specific elements of this one —in relation to other traditional statistical packages— are explained. Then we offer a whole view about how the program works, and this is illustrated from some usual tasks in teaching statistics (computation of descriptive indexes, parameters estimation, statistical contrasts, etc.). Finally, some its advantages for research are commented and the paper concludes pointing out how oneself could begin to use this program.

Key Words: *R*; statistics education; data analysis; statistical analysis; statistical modelling; statistics teaching.