rkTeaching: un paquete de R para la enseñanza de Estadística

rkTeaching: an R package for teaching Statistics

Autor
correo electrónico
Institución

Resumen

En la Universidad San Pablo CEU se ha venido utilizando programas como Excel, Statgraphics o SPSS para la enseñanza de Estadística en el ámbito de las Ciencias de la Salud. Sin embargo, en los últimos años, se ha hecho una apuesta firme por el software libre en las aulas, introduciendo el uso de R para la enseñanza de Estadística. Conscientes de que el principal punto débil de R es la falta de una interfaz gráfica de usuario amigable, se ha desarrollado el paquete rkTeaching, basado en la interfaz gráfica RkWard. La valoración por parte de los alumnos refleja claramente su mayor facilidad de uso y aprendizaje frente a SPSS.

Abstract

At the San Pablo CEU University programs like Excel, Statgraphics or SPSS has been used for teaching of Statistics in the field of Health Sciences. However, in recent years, it has made a firm commitment to use free software in the classroom, introducing the use of R for teaching Statistics. Aware that the main weakness of R is the lack of a friendly graphical user interface, it has been developed the rkTeaching package, based on the graphical interface RkWard. The assessment by the students clearly reflects its ease of use and learning compared to SPSS.

Palabras clave: Estadística, Enseñanza, R, Interfaz Gráfica de Usuario, RKWard, rkTeaching.
Keywords: Statistics, Teaching, R, Graphical User Interface, RKWard, rkTeaching.

Introducción

El uso de programas informáticos para el tratamiento y análisis de datos se ha convertido en una herramienta imprescindible para la docencia de la Estadística. En la última década, en la Universidad San Pablo CEU se han utilizado programas como Excel[1], Statgraphics[2] o SPSS[3] para la docencia de la Estadística en titulaciones de Ciencias de la Salud como Medicina, Farmacia, Psicología, Fisioterapia, Enfermería, Óptica y Nutrición. Estos programas tienen distintas ventajas e inconvenientes:

- Excel: Es posiblemente el programa más conocido y extendido por ir integrado en el paquete Office de Microsoft. Resulta bastante fácil de manejar para análisis de datos

1 http://office.microsoft.com/es-es/excel/
2 http://www.statgraphics.net/
3 http://www-01.ibm.com/software/es/analytics/spss/

sencillos, pero no incorpora procedimientos para análisis más complejos.

- Statgraphics: Es un programa bastante pedagógico relativamente fácil de aprender a usar, por lo que está muy extendido en el ámbito universitario, pero muy poco en el ámbito hospitalario.
- SPSS: Es el programa más extendido en el ámbito hospitalario por su potencia para realizar casi cualquier tipo de análisis, ya que incorporar su propio lenguaje de programación. Su contrapartida es que tiene una dificultad de aprendizaje bastante alta.

Sin embargo, todos estos programas tienen el serio inconveniente de que no son software libre, lo que supone, en primer lugar, que hay que pagar por su uso, y algunas son bastante caras como SPSS, lo que es un serio inconveniente para que los alumnos puedan acceder a ellas para uso doméstico. Y en segundo lugar, no permiten acceder al código fuente, por lo que no son fácilmente adaptables a las necesidades docentes.

Por tal motivo, en el 2008 el departamento de Matemáticas de la Universidad San Pablo CEU se planteó la introducción del software libre en la enseñanza de la Estadística y se optó por el uso de R[4] (R Development Core Team 2001). R es una implementación de código abierto del lenguaje de análisis de datos S. Es, por tanto, software libre que además es multiplataforma (existen versiones para Unix/Linux, Windows y Mac) y está desarrollado y mantenido por una enorme comunidad de programadores en todo el mundo. Al ser, en el fondo, un lenguaje de programación especialmente pensado para el tratamiento y análisis de datos, es fácilmente ampliable mediante nuevas funciones y procedimientos que suelen distribuirse en forma de paquetes también de código abierto. A finales de julio de 2012 el repositorio Comprehensive R Archive Network (CRAN) disponía de casi 4000 paquetes[5] que implementan los procedimientos más habituales para el análisis estadístico pero también los más avanzados, llegando a superar incluso a SPSS. Esto ha hecho de R el software libre de análisis de datos más extendido entre la comunidad científica.

Sin embargo, R presenta aún serios inconvenientes ya que, al ser un lenguaje de comandos, su dificultad de aprendizaje es bastante alta y tampoco dispone de una interfaz gráfica de usuario (GUI) lo suficientemente sencilla y madura para facilitar su uso a los usuarios noveles. Así pues, conscientes de que esta era el principal punto débil de R para su uso en el ámbito de la enseñanza, nos marcamos como objetivo desarrollar una interfaz gráfica de usuario sencilla para facilitar la enseñanza de la estadística y reducir así la curva de aprendizaje de R.

La interfaz gráfica de usuario RKWard

Actualmente existen varias interfaces gráficas de usuario para R[6] pero la mayoría están pensadas para usuarios avanzados o programadores de R.

En un primer momento se optó por R Commander (Fox, J. 2005), que fue la primera GUI orientada a usuarios no expertos, multiplataforma y ampliable mediante un sistema de plugins, lo que permitió crear el plugin RcmdrPluginTeachingExtras que se utilizó durante dos años para impartir las prácticas de Estadística en las titulaciones de Medicina, Farmacia y Psicología. Aunque la experiencia fue buena, R Commander presentaba todavía algunos

4 http://www.r-project.org/
5 http://cran.r-project.org/web/packages/
6 http://www.sciviews.org/_rgui/

inconvenientes, como que al estar basada en librerías gráficas Tcl/Tk bastante anticuadas, su aspecto visual era distinto en las diferentes plataformas, que la salida era bastante pobre en texto plano, o que los cuadros de diálogo no recordaban loas opciones marcadas en análisis previos.

Mientras tanto, en 2002 Thomas Friedrichsmeier había desarrollado RKWard[7] (Rödiger, S. et al. 2012), otra GUI de código abierto[8] basado en las librerías KDE y Qt, mucho más modernas y con un aspecto visual más homogéneo. Desde el principio, RKWard fue desarrollado pensando tanto en usuarios noveles como en usuarios experimentados. Para los usuarios noveles proporciona una serie de menús y cuadros de diálogos que permitieran que cualquier persona con conocimientos de estadística pudiera realizar fácilmente los análisis de datos más comunes con RKWard sin necesidad de aprender los comandos de R. Los cuadros de diálogo van acompañados generalmente de un asistente que ayuda al usuario a ir seleccionando las distintas opciones de cada análisis (figura 1). También proporciona un método de entrada y manipulación de datos en forma de hoja de cálculo muy cómoda e intuitiva (figura 2). Para los usuarios que quieren aprender el lenguaje R para explotar todo su potencial y automatizar análisis, RKWard permite generar el código R asociado a los menús y las opciones de los cuadros de diálogos seleccionados, facilitando la compresión del código generado (figura 3). Finalmente, para los usuarios experimentados, RKWard proporciona también un entorno de desarrollo integrado para programar en R con su propia consola de ejecución y depuración (figura 4). Por otro lado, la salida de RKWard es html, lo cual permite aplicar un formato mucho más rico a los resultados de los análisis, así como insertar gráficos fácilmente (figura 5). Además, al final de cada salida aparece un enlace que permite invocar de nuevo el cuadro de diálogo con los parámetros que originaron dicha salida, lo cual facilita la reproducción de los análisis.

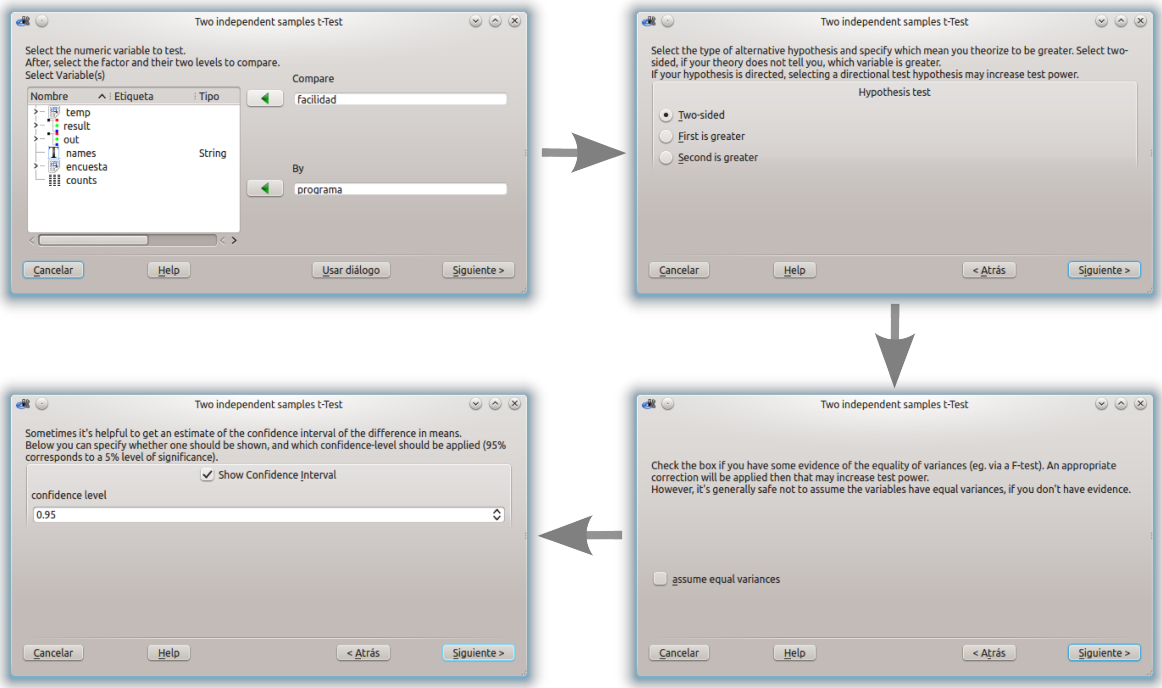

Figura 1. Cuadros de diálogo del asistente para realizar un contraste de hipótesis de comparación de medias.

7 http://rkward.sourceforge.net/

8 RkWard es distribuido bajo licencia GNU GPL versión 2 o superior.

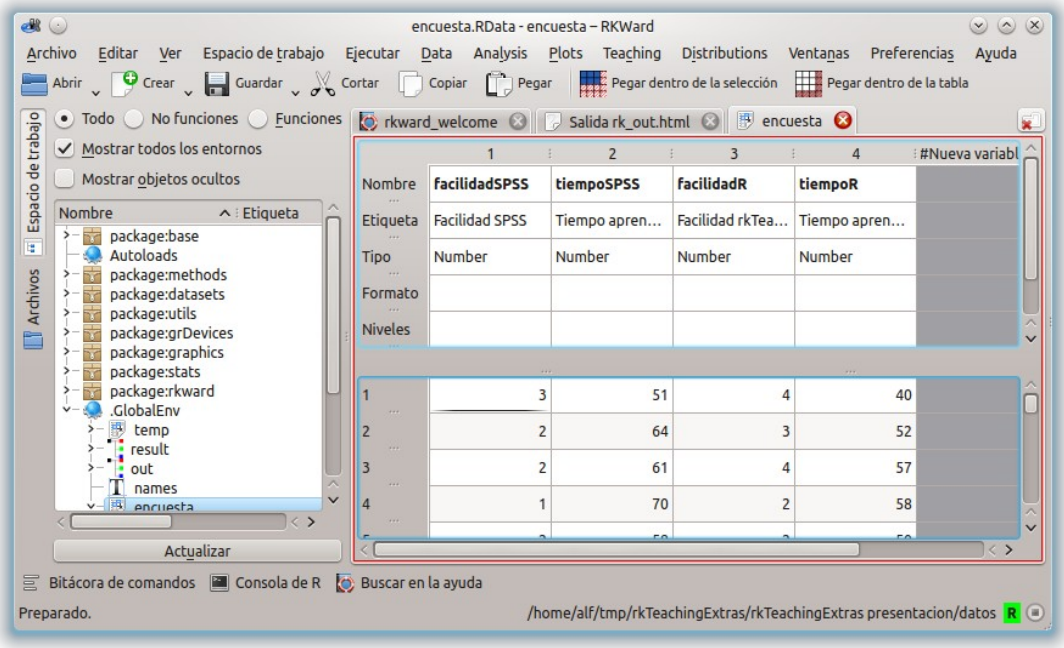

Figura 2. Tabla de entrada de datos. Cada columna representa una variable y cada fila un individuo. Cada variable debe tener un nombre y un tipo, y adicionalmente puede tener una etiqueta que aparecerá en la salida de los análisis, el formato (número de decimales, alineación, etc.) y los niveles que tiene en caso de ser un factor.

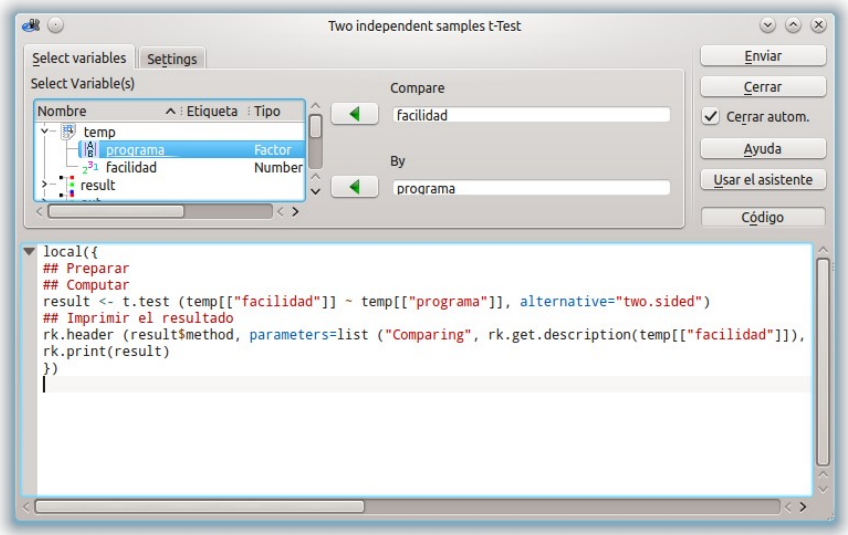

Figura 3. Código generado por el cuadro de diálogo para realizar un contraste de hipótesis de comparación de medias.

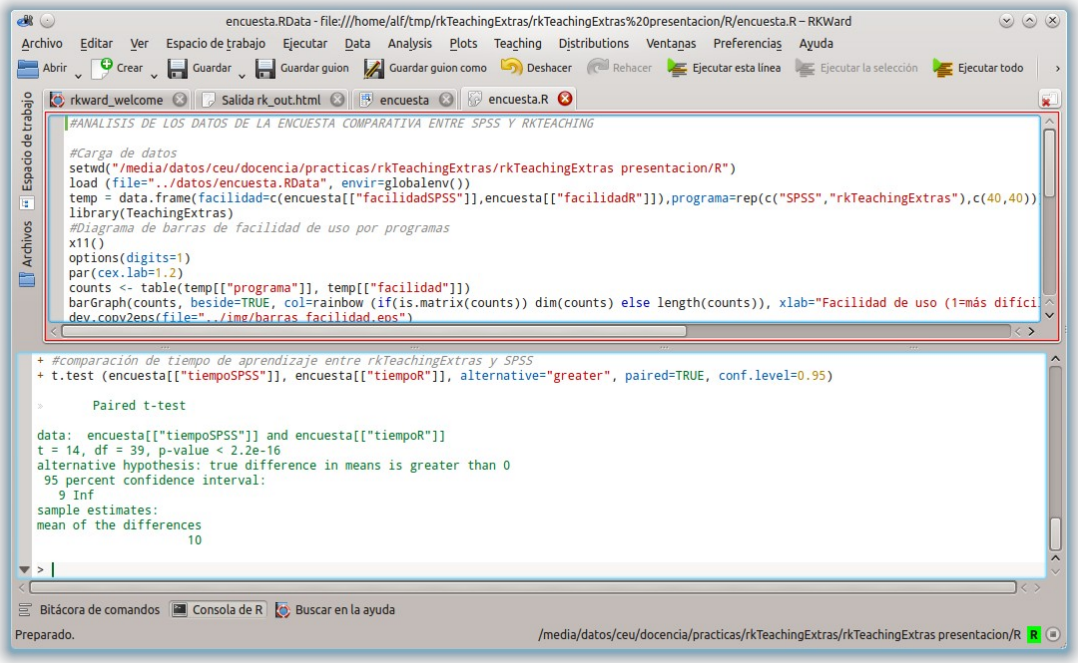

Figura 4. Entorno de desarrollo en R integrado en RKWard.

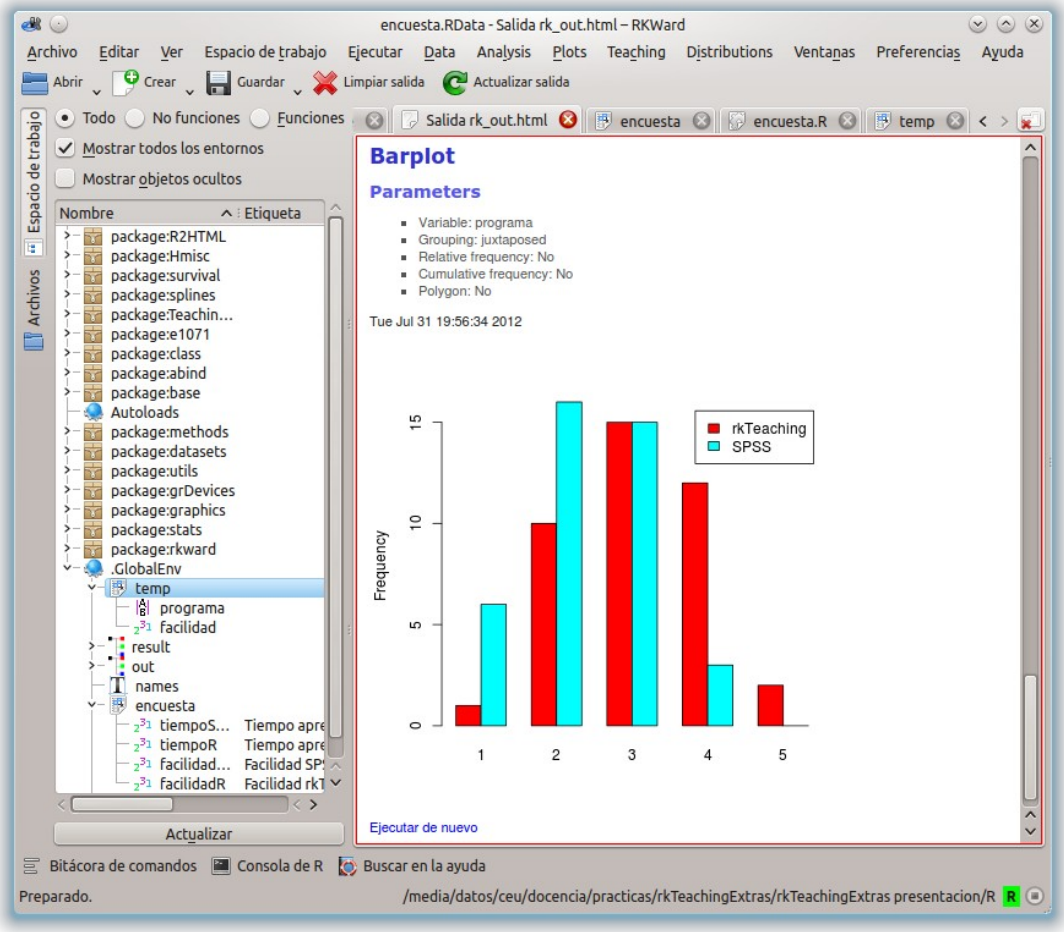

Figura 5. Salida de RKWard en html.

El principal inconveniente de RKWard era que sólo estaba disponible para plataformas Linux, lo cual limitaba enormemente su aplicación a la docencia. Pero en 2010 se lanza la versión 0.5.5 RKWard que ya es multiplataforma y además incorpora la posibilidad de ampliar la GUI mediante plugins, por lo que rápidamente se decide migrar el plugin desarrollado para R Commander a RKWard.

El sistema de plugins que se incorpora a partir de esta versión permite crear nuevos menús y cuadros de diálogo de manera bastante sencilla, sin necesidad de tener grandes conocimientos de programación. En esencia, un plugin de RKWard se compone de tres ficheros: un fichero XML que describe los componentes que configuran el cuadro de diálogo (botones, cuadros de selección, cuadros de entrada de texto, etc), un fichero con código javascript que se encarga de generar el código de R correspondiente a las opciones seleccionadas en el cuadro de diálogo, y un fichero de ayuda opcional en formato html (Friedrichsmeier, T. y Michalke, M. 2011).

El paquete rkTeaching

Sobre la base de RKWard se ha desarrollado el paquete rkTeaching como un plugin para esta GUI. Puesto que el objetivo perseguido era disponer de una GUI amigable para que los alumnos aprendieran a utilizar R para realizar análisis estadísticos, los principios que han guiado el diseño de este plugin han sido:

- Simplicidad: Esto ha supuesto eliminar de los cuadros de diálogo las opciones más complejas que se han considerado prescindibles en los procedimientos estadísticos habituales, y organizar el resto por bloques lógicos. Por ejemplo, en los procedimientos gráficos, las opciones comunes a los gráficos siempre aparecen en una pestaña separada.
- Asistencia al usuario. Todos los cuadros de diálogo incorporan un asistente de ayuda al usuario que le dirige paso a paso a través de las distintas opciones que debe seleccionar para realizar cada análisis.
- Adaptado a la programación de la enseñanza la Estadística en la USP CEU. El plugin contiene cuadros de diálogo para los procedimientos habituales en un primer curso de estadística aplicada a las Ciencias de la Salud. Así, por ejemplo, se incorpora un menú específico para sacar tablas de frecuencias, tal y como se explican en las clases de teoría.
- Salidas orientadas a facilitar la comprensión del alumno. Las salidas de los distintos procedimientos están formateadas para facilitar la comprensión del alumno. Algunas de ellas incluso se han adaptado al formato presentado en las clases de teoría para que les resulte más familiar. Por otro lado, se ha incorporado la librería de javascript MathJax[9] para facilitar la visualización de fórmulas matemáticas en la salida html.

El paquete rkTeaching añade un nuevo menú a los de RKWard con la etiqueta Teaching. Bajo este menú se despliegan los siguientes submenús:

- Frequency Tabulation. Este menú contiene los procedimientos para la construcción de tablas de frecuencias con frecuencias absolutas, frecuencias relativas, frecuencias absolutas acumuladas y frecuencias relativas acumuladas. Contiene los siguientes

9 http://www.mathjax.org/

submenús:

- Frequency Tabulation, para construir tablas con datos no agrupados.
- Frequency Tabulation (Grouped Data), para tablas con datos agrupados en intervalos. Permite la construcción de los intervalos de acuerdo a diferentes criterios.

- Plots. Este menú contiene los procedimientos gráficos más habituales para la descripción de una variable y para la descripción de relaciones entre variables. Contiene los siguientes submenús:
  - Bars, para dibujar diagramas de barras. Las barras pueden representar frecuencias absolutas y acumuladas y es posible dibujar el polígono de frecuencias sobre las barras. Además, las barras pueden agruparse de acuerdo a algún factor.
  - Histogram, para dibujar histogramas. Las barras pueden representar frecuencias absolutas y acumuladas, así como densidades, y también es posible dibujar el polígono de frecuencias sobre las barras. Al igual que para las tablas de frecuencias, los intervalos de las clases pueden construirse de acuerdo a diferentes criterios.
  - Scatterplot, para dibujar diagramas de dispersión. Se pueden clasificar los puntos de acuerdo a algún factor para representar diferentes nubes de puntos. También permite dibujar una recta de ajuste de regresión por mínimos cuadrados.

  No se han incorporado otros procedimientos gráficos habituales como el diagrama de cajas por estar ya incorporados en los menús de RKWard con la simplicidad requerida.
- Descriptive Statistics. Este menú contiene los procedimientos para el cálculo de los estadísticos muestrales. Contiene los submenús:
  - Statistics, para el cálculo de los estadísticos más comunes: Tendencia central (media, mediana y moda), dispersión (varianza, cuasivarianza, desviación típica, cuasidesviación típica, coeficiente de variación, rango y recorrido), forma (coeficiente de asimetría y coeficiente de apuntamiento) y cuantiles.
  - Detailed calculation, para el cálculo detallado de los estadísticos. Este menú es especialmente útil para que los alumnos comprendan el procedimiento que se sigue para calcular cada estadístico. También permite a los alumnos contrastar sus cálculos manuales (figura 6).
- Regression. Este menú contiene los procedimientos para el cálculo de modelos de regresión simple. Contiene los submenús:
  - Linear Regression, para el ajuste de modelos lineales. Permite guardar el modelo como un objeto de R que puede utilizarse después para hacer comparativas o predicciones.
  - Non Linear Regression, para el ajuste de modelos no lineales. Incorpora los modelos cuadrático, cúbico, potencial, exponencial, logarítmico, inverso y sigmoidal. También permite guardar los modelos.
  - Model Comparison, para hacer una comparativa de modelos de regresión. Los modelos aparecen ordenados de mayor a menor coeficiente de determinación.

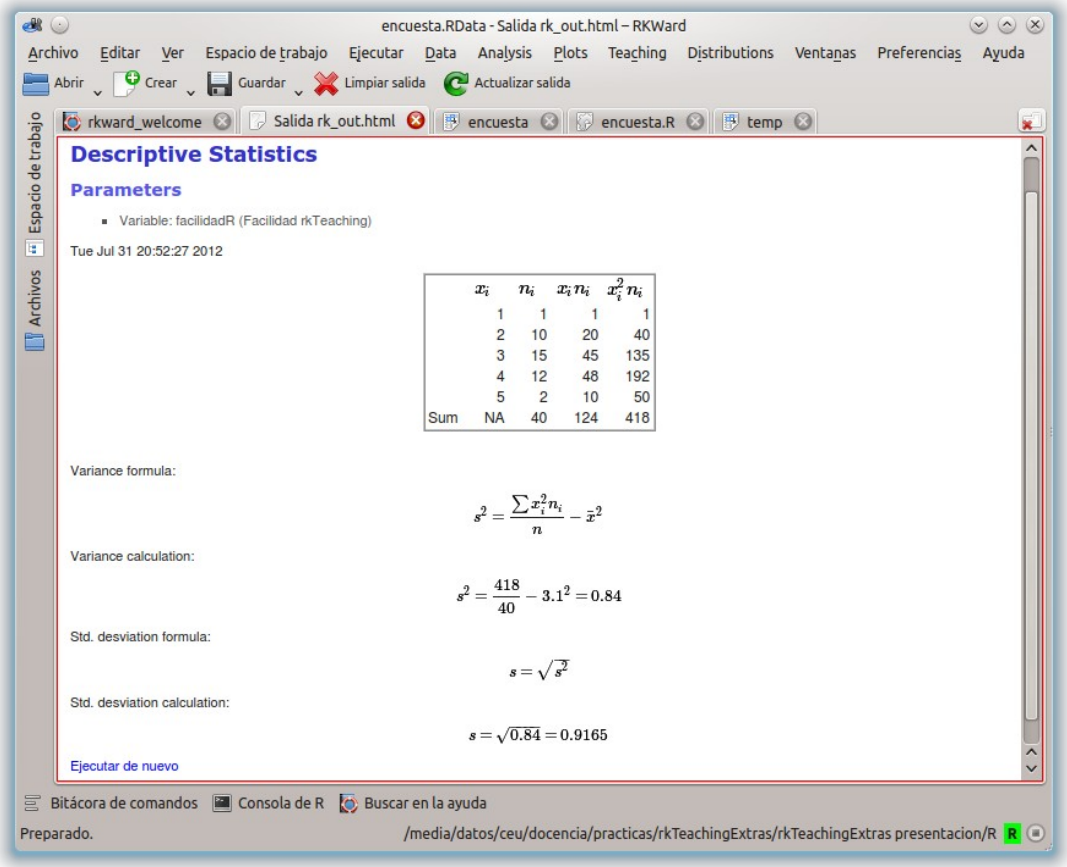

Figura 6. Cálculo detallado de estadísticos muestrales.

- Preditions, para hacer predicciones con un modelo de regresión. Incorpora además la opción para obtener los intervalos de confianza para las predicciones.

- Parametric Test. Este menú contiene los procedimientos para realizar los contrastes de hipótesis paramétricos más comunes. Contiene los submenús:
  - Mean comparison. Contiene los procedimientos para los contrastes relacionados con las medias. A su vez contiene los submenús:
    - One Sample T Test, para el contraste de hipótesis de la media de una población.
    - Two Independent Samples T Test, para el contraste de comparación de medias de dos poblaciones independientes.
    - Two Paired Samples T Test, para el contraste de comparación de medias de dos poblaciones pareadas.
    - ANOVA. Para realizar contrastes de análisis de la varianza para uno o varios factores.
    - Mean Sample Size, para el cálculo de tamaños muestrales para contrastes de medias.
  - Variances Comparison: Contiene los procedimientos para los contrastes de

comparación de varianzas. A su vez, contiene los submenús:

- F Test, para la comparación de las varianzas de dos poblaciones.
- Levene's test para la comparación de la variabilidad de dos o más poblaciones.

- Proportions Comparison: Contiene los procedimientos para los contrastes de proporciones. A su vez, contiene los submenús:
  - One Proportion Test, para el contraste de hipótesis de la proporción de una población.
  - Two Proportion Test, para el contraste de hipótesis de las proporciones de dos poblaciones independientes.
  - Proportion Sample Size, para el cálculo de tamaños muestrales para los contrastes de proporciones.

- Non Parametric Tests. Este menú contiene los procedimientos para realizar los contrastes de hipótesis no paramétricos más habituales. Contiene los submenús:
  - Normality Test, para el contraste de normalidad de la muestra. Incorpora el contraste de Shapiro-Wilk y el de Kolmogorov.
  - U Mann-Whitney Test, para el contraste de la U de Mann-Whitney.
  - Wilcoxon Test, para el contraste de Wilcoxon.
  - Kruskal-Wallis, para el contraste de Kruskal-Wallis. Incorpora la comparación por pares.
  - Friedman Test, para el contraste de Friedman.
  - Chi-square Test, para los contrastes de independencia basados en la distribución Chi-cuadrado.
- Concordance. Este menú contiene los procedimientos para el análisis de concordancia. Contiene los submenús:
  - Intraclass Correlation Coeficient, para el cálculo de coeficiente de correlación intraclase.
  - Cohen's Kappa, para el cálculo del coeficiente kappa de Cohen.
- Simulations. Este menú contiene varios procedimientos para realizar la simulación de experimentos aleatorios. Estos experimentos son muy útiles para que los alumnos comprendan algunos de los conceptos probabilísticos más abstractos. Contiene los submenús:
  - Coins Tosses, para simular el lanzamiento de monedas.
  - Dice Roll, para simular el lanzamiento de dados.
  - Sample Generation, para generar muestras aleatorias de cualquier tamaño a partir de una distribución conocida.
  - Small Numbers Law, para mostrar visualmente cómo la distribución binomial se aproxima a la distribución de Poisson a medida que aumenta el número de

repeticiones n y disminuye la probabilidad de éxito p.

- Central Limit Theorem, para ver cómo las suma de variables independientes converge a una distribución normal.

Otros procedimientos habituales en un primer curso de estadística, como el cálculo de probabilidades de distintas distribuciones no se han incorporado por estarlo ya en RKWard con las sencillez requerida.

Comparativa docente de rkTeaching con SPSS

Para valorar la sencillez de manejo de rkTeaching, se realizó un estudio comparativo de la facilidad de uso y el tiempo de aprendizaje entre rkTeaching y SPSS. Para ello se tomó una muestra de 40 alumnos de medicina que habían aprobado el curso básico de estadística pero nunca habían manejado rkTeaching ni SPSS. Los alumnos recibieron una pequeña clase introductoria sobre la introducción de datos con cada uno de los programas y a continuación se les pidió que realizasen una serie de ejercicios con ambos programas. Los ejercicios consistieron en introducir los datos de una muestra, dibujar un histograma, calcular varios estadísticos descriptivos, calcular un modelo de regresión lineal y dibujarlo, calcular una probabilidad de una variable normal y hacer un contraste de hipótesis de comparación de medias. Para que le orden de los programas no influyese, los alumnos se dividieron aleatoriamente en dos grupos de 20 alumnos, de manera que los primeros empezaron con rkTeaching y luego con SPSS y los segundos al revés. Al final se les pasó una sencilla encuesta sobre la facilidad de uso y también se midió el tiempo que tardaron en hacer la tarea con cada programa. Las variables medidas fueron: Tiempo de realización de la tarea con rkTeaching (en min), tiempo de realización de la tarea con SPSS (en min), facilidad de uso de rkTeaching (escala discreta de 1=más difícil a 5=más fácil) y facilidad de uso de SPSS (escala discreta de 1=más difícil a 5=más fácil).

La comparativa de los tiempos de aprendizaje mostró que el tiempo de aprendizaje de SPSS fue significativamente mayor que el tiempo de aprendizaje con rkTeaching (figura 7) con un p-valor $4{,}8e{-}17$ y un intervalo de confianza del 95% para diferencia de medias $(9{,}06, \infty)$, lo que indica que el tiempo medio para realizar las tareas con SPSS fue al menos 9 minutos mayor que el de rkTeaching, lo que supone una reducción del tiempo de al menos un 17%.

De igual modo, la comparativa de la facilidad de uso reveló que la facilidad de uso de rkTeaching fue significativamente mayor que la de SPSS (figura 8) con un p-valor $1{,}7e{-}06$ y un intervalo de confianza del 95 % para diferencia de medias $(0{,}4993, \infty)$, lo que indica que la facilidad de manejo de rkTeaching es de al menos medio punto más que con SPSS, lo que supone un aumento de la facilidad de al menos un 10%.

Conclusiones y trabajo futuro

Con el objetivo de introducir el uso del software libre en la enseñanza de la estadística en la Universidad San Pablo CEU se ha desarrollado el paquete rkTeaching sobre la base de la GUI RKWard. El paquete rkTeaching se ha utilizado para impartir las prácticas de Estadística en las titulaciones de Medicina, Farmacia y Psicología con éxito. Para valorar la facilidad de uso de rkTeaching frente al SPSS, que era el software utilizado hasta el momento, se realizó un experimento que reveló que el aprendizaje con rkTeaching por parte de los alumnos es más rápido e intuitivo que con SPSS.

Como trabajo futuro se plantea la traducción de rkTeaching al castellano, mejorar aún más la salida para incluir interpretaciones de los resultados e incorporar cuadros de diálogos para análisis más complejos como el análisis multivariante.

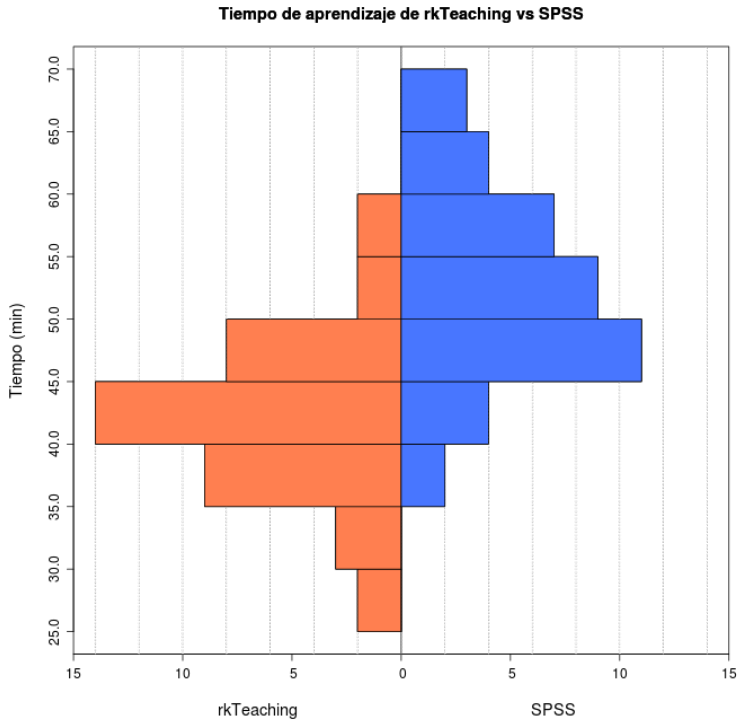

Figura 7. Distribución de frecuencias del tiempo de aprendizaje de rkTeaching con respecto a SPSS.

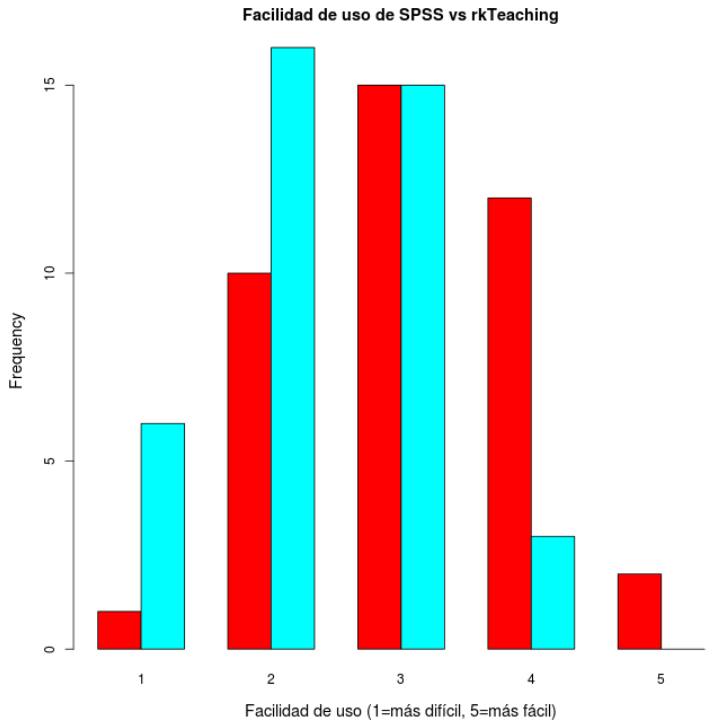

Figura 8. Distribución de frecuencias de la facilidad de uso (de 1=más difícil a 5=más fácil) de rkTeaching con respecto a SPSS.

Bibliografía

Fox, J. (2005). The R Commander: A Basic-Statistsics Graphical User Interface to R. *Journal of Statistical Software*, 14(9), 1-42. URL: http://www.jstatsoft.org/v14/i09/.

Friedrichsmeier, T. y Michalke, M. (2011). Introduction to Writting Plugins for RKWard. Recuperado el 15 de julio de 2012, de http://rkward.sourceforge.net/documents/devel/-plugins/index.html.

R Development Core Team (2001). *R: A Language and Environment for Statistical Computing*. Viena: R Foundation for Statistical Computing.

Rödiger, S.; Friedrichsmeier, T.; Kapat, P. y Michalke, M. (2012). RKWard: A Comprehensive Graphical User Interface and Integrated Development. Environment for Statistical Analysis with R. *Journal of Statistical Software*, 49(9), 1-34.URL: http://www.jstatsoft.org/v49/i09.