

R-Medic. Un programa de análisis estadísticos sencillo e intuitivo

R-Medic. A simple and intuitive statistical analysis software.

Mangeaud A¹, Elías Panigo DH¹.

Resumen

Se presenta el “soft” estadístico R-Medic, un programa que se utiliza directamente en la web sin necesidad de instalarlo o descargarlo en computadoras o dispositivos móviles. El programa posee cinco prestaciones generales: realiza un control que permite evaluar si las variables de la base de datos están correctamente cargadas, realiza tablas, gráficos, pruebas estadísticas y presenta herramientas didácticas de distribución de variables. Su uso es intuitivo ya que se debe definir si se poseen una o dos variables a estudiar y el “soft” propone los posibles análisis a realizar. Ofrece las pruebas de supuestos y responde con los estadísticos, valor p e interpretaciones de los resultados..

Palabras claves: Programa. Estadística básica. Inferencia estadística.

Abstract

The statistical software R-Medic is presented, a program that is used directly on the web without the need to install it or download it to computers or mobile devices. The program has five general features: it performs a control that allows to evaluate if the variables of the database are correctly loaded, makes tables, graphs, statistical tests and presents didactic tools for the distribution of variables. Its use is intuitive since it must be defined if one or two variables are to be studied and the software proposes the possible analyzes to be carried out. It offers the assumption tests and responds with the statistics, p-value and interpretations of the results.

Key words: Program. Basic statistics. Statistical inference.

1. Cátedra de Bioestadística Posgrado. Facultad de Ciencias de la Salud. Universidad Católica de Córdoba
Correspondencia: Secretaria de Posgrado. Jacinto Ríos 571 X5004FHP- Córdoba, Argentina – Córdoba.
Tel.: 54-351-4517299. e-mail: amangeaud@yahoo.com.ar

Introducción

Los usuarios que necesitan de análisis estadísticos tienen la oportunidad de acceder a innumerables programas que realizan tablas, gráficos, así como estimaciones y pruebas de hipótesis. Los hay desde los que son extremadamente onerosos, pasando por precios accesibles hasta gratuitos, desde pequeños que hacen sólo algunos análisis hasta otros muy complejos, los que se deben instalar en la computadora hasta los que funcionan directamente en la nube (web). La mayoría de ellos tienen excelentes prestaciones, resuelven análisis con algoritmos extensamente probados y sólo el gusto personal del operador, la recomendación de un colega o la posibilidad de acceder a ellos es lo que motiva al usuario a la elección de uno por sobre otros.

Por estas latitudes algunos programas se hicieron más conocidos, dependiendo del área de aplicación.

Podemos mencionar, solo a modo de ejemplo, SAS¹, SPSS², Statistica³, EpiInfo⁴, Minitab⁵, InfoStat⁶, entre otros. Pero el hito histórico que cambió el uso de estas herramientas ha sido la aparición del “soft” R⁷.

R es un “soft” libre, gratuito, cooperativo, que crece día a día por el aporte de personas que programan de manera altruista para este proyecto. Pero R tiene un inconveniente, R no es “amigable”. El usuario tiene que aprender a programar una sentencia (script) para cada orden que se le da. Es decir, hace falta una formación en la lógica y la formalidad de la programación que muchos usuarios no están dispuestos a intentar.

Han aparecido algunas interfaces que “conectan” al usuario con R y lo hacen un poco más amigable, ejemplos de ello son RCommander⁸, RStudio⁹, entre otros.

De todos modos, amigable significa que estas interfaces facilitan el uso, pero siempre basado desde una lógica que se aplica: el usuario debe saber exactamente a priori qué debe hacer. A modo de ejemplo, en cierta clínica es una hipótesis tácita que las médicas pediatras poseen una dulzura especial en el trato con los pacientes, por lo tanto, para probar esto se desea comparar si es diferente el valor medio de glucosa en sangre de una muestra de médicas pediatras con respecto a una muestra de médicas traumatólogas.

Luego de realizado el diseño y el protocolo pertinente el investigador posee un grupo de “datos” y debe saber que:

1. La base de datos a completar debe estar ordenada en filas y columnas de un modo particular: pacientes en filas, variables en columnas y sólo una fila (la primera) como nombre de las variables. (Esto es consensuado en todos los “soft”)
2. Cada columna debe tener sólo números (variables cuantitativas) o sólo letras (variables cualitativas), si se las mezcla cada “soft” decide qué hacer al respecto.
3. Debe buscar la solapa que diga algo similar a: Prueba t de diferencia de medias o prueba “t”. Pero si no se cumplen los supuestos estadísticos requeridos para esta prueba debe buscar alternativas, por ejemplo, buscar otra solapa que diga algo similar a Prueba de Mann-Whitney (o prueba de Wilcoxon).
4. Luego de realizar el análisis debe interpretar el valor p en función del conocimiento de la hipótesis nula (es decir, que el valor p sea grande o pequeño?, qué se rechaza o no se rechaza?)

Para una persona con experiencia en investigación, los 4 puntos anteriormente descriptos son irrelevantes, pero para estudiantes en formación o personas sin entrenamiento puede ser una barrera difícil de sortear.

Ante esta lógica de funcionamiento, nos propusimos crear un “soft” estadístico que facilite al usuario sin entrenamiento estadístico el acceso a los análisis adecuados y pertinentes. Por ello creamos R-Medic, una interfaz del programa R, pero de un uso sencillo e intuitivo.

Fundamento

Una de las ideas básicas y fundamental de la Estadística, que se imparte en las primeras clases es la definición de variable y los tipos de variables que se presentan. Para utilizar R-Medic los usuarios sólo eso deben conocer qué tipo de variables poseen (si son variables cuantitativas o cualitativas).

R-Medic permite trabajar con una o dos variables (columnas) a la vez. Estas, a su vez pueden ser cualitativas o cuantitativas. Entonces a partir de las cinco combinaciones posibles lleva al usuario a los análisis más comúnmente utilizados de la Estadística Básica aplicada especialmente a Ciencias de la Salud.

Una vez llegado a ese lugar el usuario obtiene los resultados estadísticos, tal como ocurre con otros softwares.

Además, encontrará las interpretaciones básicas de los resultados obtenidos en los análisis y frases que ayudarán a redactar un trabajo, presentación o publicación científica. Luego, debe realizar una búsqueda bibliográfica para tener la certeza de estar trabajando con el análisis adecuado para sus datos.

Prestaciones

R-Medic trabaja directamente en la web (www.r-medic.com), por lo que no debe ser instalado en la computadora o dispositivo móvil. El usuario sólo debe “subir” su base de datos (archivos csv, xls ó xlsx) y comenzar a trabajar en cualquier dispositivo móvil o computadora con acceso a internet, de un modo libre y gratuito.

Todos los resultados, tablas y gráficos obtenidos en R-Medic pueden ser descargados completamente.

Los siguientes cinco puntos principales son los que definen las prestaciones:

1. Control.

Esta pestaña, es tal vez la pestaña más importante y novedosa de R-Medic. “Control” corrobora si los datos han sido cargados de manera correcta, si se presenta un posible error y cuál es la ubicación del mismo.

El “Control” será ejercido sobre la variable que sea elegida. Lo correcto es realizar un control sobre cada variable de la base de datos antes de obtener tablas, gráficos o realizar un test estadístico.

Este algoritmo detecta si en la columna seleccionada se encuentran sólo números o si también hay caracteres alfanuméricos. Si la variable posee sólo números, devuelve la ubicación de la celda donde están el mínimo valor y el máximo valor como una manera de controlar si ambos extremos se encuentran entre los valores posibles de la variable. Si son caracteres alfanuméricos (sólo letras o letras combinadas con números) devuelve el nombre de las categorías y detecta diferencias que pueden no ser perceptibles sencillamente. Por ejemplo:

en más de una oportunidad al tipear la variable sexo, se suele cargar “Fem”, “FEM” “Fem” o “Fem”, y para un “soft” estadístico son cuatro sexos diferentes (nótese los espacios en las dos últimas palabras). R-Medic devuelve la ubicación de la celda donde se encuentra cada categoría. El usuario debe volver a su base original para cambiar dichos errores para luego ingresar nuevamente la base modificada a R-Medic. Con esta acción se obliga a tener una sola base original correctamente cargada.

2. Tablas

R-Medic realiza tablas para una variable y para dos variables a la vez, arrojando tablas de distribuciones de frecuencia o tablas con Estadística descriptiva (medidas de posición y dispersión), dependiendo si corresponde a variables cuantitativas o cualitativas.

3. Gráficos

R-Medic realiza gráficos de barras, de tortas, histograma de distribución de frecuencias, gráfico de cajas (boxplot), de puntos, de media +/- desvío estándar o error estándar y de dispersión (ejes XY).

Posee opciones para utilizar paletas de colores y aumentar el tamaño de los gráficos para ser utilizados en tamaños mayores a una hoja (por ejemplo, para no perder fidelidad en tamaños requerido para la confección de poster) y los gráficos pueden ser copiados y pegados en otro “soft”.

4. Pruebas de hipótesis

R-Medic contiene los test estadísticos más comúnmente utilizados¹⁰, que se aplican a una o a dos variables. Para acceder a las baterías de test se deben decidir si se trabaja con una o dos variables, si éstas son cuantitativas o cualitativas y cuál es el nivel de significación solicitado (alfa).

Cada test estadístico es acompañado de un análisis de los supuestos estadísticos. Para el caso de los análisis paramétricos el “soft” realiza previamente las pruebas de los supuestos estadísticos pertinentes.

Por ejemplo, en la prueba “t” para diferencia de medias, realiza las pruebas de la normalidad y homogeneidad de varianzas¹¹, y de no cumplirse estos requerimientos, recomienda realizar una prueba de Mann Whitney. Sin embargo, para la prueba de independencia chi cuadrado, de no cumplirse con los requisitos sobre el valor mínimo de esperados¹² realiza directamente una prueba Chi cuadrado exacta¹³.

El “soft” arroja los valores de los estadísticos pertinentes y valor p. Asimismo ofrece una interpretación de los resultados con frases que ayudan a extrapolar los resultados al pool de datos seleccionados.

El siguiente es un listado de los test estadísticos que el usuario puede realizar en R-Medic:

A) Solapa Pruebas de hipótesis

1. Una variable, cualitativa:
 - 1.a. Test de proporciones.
2. Una variable, cuantitativa:
 - 2.a. Prueba t para una Media
 - 2.b. Prueba de Normalidad (Shapiro-Wilks)
3. Dos variables, ambas cualitativas
 - 3.a. Diferencia de proporciones
 - 3.b. Regresión logística simple (opción 1)
 - 3.c. Test Chi Cuadrado
 - 3.d. Odds ratio, Riesgo relativo, Valores predictivos, sensibilidad y especificidad
4. Dos variables, una cuantitativa, otra cualitativa
 - 4.a. Prueba t de dos medias (muestras independientes)
 - 4.b. Prueba de Mann Whitney (muestras independientes)
 - 4.c. Análisis de la varianza a un factor (Anova)
 - 4.d. Análisis de Kruskal-Wallis
 - 4.e. Test de homogeneidad de varianzas de Bartlett (opción 1)
 - 4.f. Regresión logística simple (opción 2)
5. Dos variables, ambas cuantitativas
 - 5.a. Correlación lineal
 - 5.b. Regresión lineal simple

- 5.c. Prueba t (muestras apareadas)
- 5.d. Prueba de Wilcoxon para muestras apareadas
- 5.e. Test de homogeneidad de varianzas de Bartlett (opción 2)
- 5.f. Regresión logística simple (opción 3)

B) Solapa Sobrevida

Análisis de sobrevida de Kaplan Meier y comparación de grupos por Log Rank

5. Herramientas

Son herramientas didácticas para apoyar la enseñanza de la Estadística. Se presentan diferentes opciones para la comprensión de Distribuciones de Probabilidad: Distribuciones Binomial, Poisson, Normal, t, Chi Cuadrado y F. Cambiando los valores de la variable y los parámetros de la distribución, cambia la forma de la misma.

Unas líneas para entendidos en R

R-Medic utiliza la versión 3.3.3 (2017-03-06) de R.

Los principales paquetes de R utilizados son: agricolae, dplyr, fmsb, gdata, ggplot2, ggvis, gplots, gtools, plotly, Rcpp, splines, stats y survival, entre otros.

Conclusiones

La expectativa de los autores es que estudiantes y usuarios en general de áreas relacionadas con la Salud, puedan encontrar en R-Medic una herramienta útil en su iniciación a la estadística aplicada. Este “soft” no pretende reemplazar a los ya existentes, sino que cubrir un nicho vacío en lo referido a los aspectos: la utilización de un “soft” estadístico, libre, gratuito, directamente en la web, con dispositivos móviles o PC, con la utilización de una opción “Control” de la carga de los datos y la interpretación de las salidas de los resultados.

Bibliografía

1. https://www.sas.com/es_ar/software/analytics/stat.html
2. <https://www.ibm.com/ar-es/marketplace/spss-statistics>
3. <http://statistica.io/>
4. <https://www.cdc.gov/epiinfo/index.html>
5. <http://www.minitab.com/>
6. <http://www.infostat.com.ar/>
7. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016 URL <https://www.R-project.org/>.
8. <http://www.rcommander.com/>
9. <https://www.rstudio.com/>

Mangeaud A, Panigo DE. R-Medic. Un programa de análisis estadísticos sencillo e intuitivo.

10. Mangeaud, A..Bioestadística I. Ed Ferreyra. Córdoba, Argentina. 2014. 237 pp. ISBN: 978-987-1742-59-2
11. Fay, Michael P.; Proschan, Michael A.. Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*. 2010 4: 1–39. doi:10.1214/09-SS051.
12. Agresti, A. *Categorical data*. John Wiley & Sons, Inc., Hoboken, New Jersey. 2002 710 pp.
13. Mehta, C.R. ; Patel, N.R. Exact Inference for Categorical Data. In P. Armitage and T. Colton, eds., *Encyclopedia of Biostatistics*, Chichester: John Wiley, 1998 pp. 1411–1422..

