VIII FESTIVAL INTERNACIONAL DE MATEMÁTICA
7 al 9 de junio de 2012. Sede Chorotega, Universidad Nacional, Liberia, Costa Rica

Aprendiendo estadística con R

M.Sc. José Andrey Zamora Araya
Universidad Nacional
andreyzamora@gmail.com

Licda. Rosibel Tatiana Vallejos Brenes
Liceo Mauro Fernández Acuña
ctaty@costarricense.cr

Resumen: Se introduce el software estadístico R y sus ventajas y desventajas en la enseñanza y aprendizaje de la estadística descriptiva. Se trabajará desde su instalación, el uso de bibliotecas y algunos comandos básicos y algunos ejemplos de su uso en la enseñanza de la estadística.

Palabras clave: Enseñanza de la Estadística, Software estadístico R, tecnologías de la información

Abstract: We introduce the R statistical software and its advantages and disadvantages in teaching and learning of descriptive statistics. It will work since its installation, use oflibraries and some basic commands and some examples of its use in statistical educational.

Key Word: Academic Performance, Mathematical Education, Diagnostic tests and Higher Education

Introducción

R es un software o más bien un lenguaje de comandos de manipulación y análisis estadístico basado en el lenguaje estadístico S desarrollado por AT&T, con la diferencia de que R es un programa de código abierto y gratis, lo que lo ha hecho muy popular en los ámbitos académicos.

Dado sus características, R tiene un gran potencial para ser usado en la educación pública, pues al no tener que pagar por el software y puede ser instalado en diversos sistemas operativos IOS de MAC, Linux o Windows. Quizá el mayor inconveniente que ven en un principio los nuevos usuarios de R es su interfaz gráfica que algunos dirían “poco amigable”, en el sentido de que hay que programar las funciones, pues a diferencia de Windows no hay botones o ventanas que despliegan menús donde el usuario puede elegir opciones.

No obstante, R es una herramienta sumamente útil ya que al hecho de ser un programa de código abierto y gratuito debe añadirse su capacidad de análisis y poder de cálculo estadístico, el proveer operaciones estadísticas y brindar un lenguaje de programación que puede ser usado para crear nuevas funciones o extender las actuales, creación de gráficos y la posibilidad de trabajar desde estadísticas simples hasta tópicos más avanzados como análisis multivariado, modelo complejos de estructura de covariancia entre otros.

Sus ventajas, superan con creces las desventajas que pueda tener y se convierte en una opción para el aprendizaje y enseñanza de la estadística. Además, dada su popularidad se han desarrollado interfaces gráficas de usuario GUI, por sus siglas en inglés (Grafical User Interface) de uso libre para R que hacen un poco más amigable la interacción con el usuario entre ellas están:

- RStudio, *http: //www.rstudio.org/*
- R commander, *http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html*

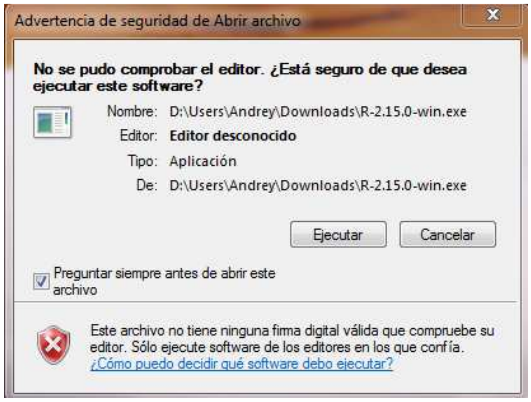
- ESS (Emacs Speaks Statistics), *http: //www.walware.de/goto/stated*
- JGR (Java GUI for R), *http://cran.r-project.org/web/packages/JGR/index.html*

Por ello, es que debe potencializarse el uso de R como herramienta versátil en la enseñanza y aprendizaje de la estadística e incluso de ciertos conceptos matemáticos que pueden ser atendidos con ayuda de este potente software.

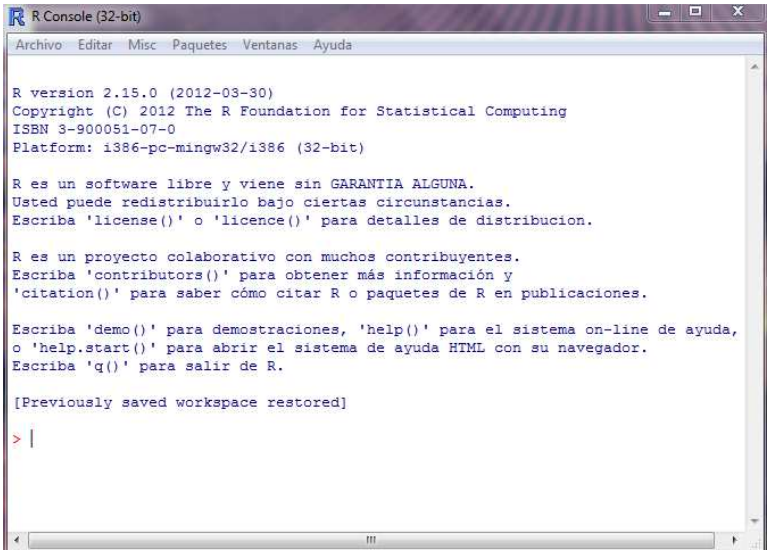
¿Cómo instalar R?

Instalar R es sumamente simple, solo hay que dirigirse a la página *http://www.r-project.org/* donde se le da click a la opción **download R**, luego se escoge un CRAN mirror, por ejemplo el de Chile *http://dirichlet.mat.puc.cl/* y se descargar la aplicación para el sistema operativo que el usuario tenga en su computadora.

Se seguirá el ejemplo como si se fuera a instalar R en una PC que usa Windows, en cuyo caso se elegirá la opción *Download R for Windows* y luego *install R for the first time*. Finalmente se descargará un archivo ejecutable, que al hacer doble click asobre él aparecerá la siguiente leyenda



Luego se le da ejecutar, se escoge el idioma y se instalará en la computadora. Una vez abierto el programa se presentará una consola como la siguiente



Tal y como aparece en el primer menú, R es un software libre y viene sin GARANTIA ALGUNA, aunque es posible redistribuirlo bajo ciertas circunstancias. Mediante citation () podemos saber cómo citar R o paquetes de R en publicaciones y así dar el crédito a la enorme cantidad de personas que desarrollan este proyecto.

Al escribir el comando citation () aparecerá la siguiente información

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Primeros pasos con R

Por defecto, R abre una sola ventana, la consola o ventana de comandos de R en el cual se introducen los comandos y será dónde se vean los resultados de los análisis. Justo después de la cabecera, aparece una línea en blanco con el símbolo > en el margen izquierdo. A partir de ese momento R espera que escriban COMANDOS e instrucciones para comenzar a trabajar (Conesa, 2011).

Para ejecutar un comando, basta con introducirlo y presionar la tecla ENTER al final, R devolverá inmediatamente el resultado; si lo que se desea es escribir un comentario, como por ejemplo “mi primer comando de R”, y por ende no se pretende que el programa lo ejecute, basta con poner el símbolo de numeral (#) al inicio del comentario y de esta manera R no trate de ejecutarlo.

Las órdenes elementales en R consisten en expresiones o en asignaciones, una orden consiste en una expresión, se evalúa, se imprime y su valor se pierde, en cambio una asignación evalúa una expresión, no la imprime y guarda su valor en una variable. Se puede hacer la asignación con el signo de igualdad (=) o bien con el símbolo < –. En cada línea sólo caben 128 caracteres, si se desea escribir más, una opción es utilizar otra línea que comienza con el símbolo + en el lado izquierdo de la pantalla en vez de >. Con la tecla ↑ se recupera las instrucciones utilizadas en la sesión inmediatamente anterior y con las teclas ←,→ se pueden corregir errores. Para separar expresiones se emplea el punto y coma (;), la combinación Ctrl + ^,[interrumpe la edición o ejecución en curso, finalmente q() es el comando para salir de R (Conesa, 2011).

R como calculadora

El uso más básico que tiene R es como calculadora, pues se pueden realizar cálculos aritméticos simples con los simbolos de +, –, *, / ^ para las operaciones básicas. Algunos ejemplos sencillos y funciones de uso común pueden verse en Verzani (2004), Por ejemplo

```
> 5+8
[1] 13
> 2*4
[1] 8
> 75-12
[1] 63
> (3-5)*4
[1] -8
> 7-1*2
[1] 5
> sqrt(16)# saca raiz cuadrada
[1] 4
> exp(1)# realiza las potencias con base e
[1] 2.718282
> log(3)# realiza logaritmo natural de 3
[1] 1.098612
> log(100,10) # realiza el logaritmo en base 10 de 100
[1] 2
> sin(pi)
[1] 1.224606e-16
> cos(pi)
[1] -1
> tan(pi)
[1] -1.224606e-16
> floor(4.3)# redondea hacia abajo
[1] 4
> ceiling(4.7)# redondea hacia arriba
[1] 5
> trunc(12.48)#trunca a cero decimales
[1] 12
> round(12.4815879,digits=2)# redondea al número de decimales indicados
[1] 12.48
> signif(12.4815879,digits=6)# da el número de cifras significativas indicadas
[1] 12.4816
>
```

También se pueden crear vectores, de hecho que está diseñado de forma la mayoría de operaciones y de funciones están definidas con carácter vectorial, es decir para operar componente a componente, por ejemplo si deseamos crear un vector x lo definimos de la siguiente manera

x= c (1,5,7,15,-3)

Los paréntesis () se emplean para los argumentos de las funciones y para agrupar expresiones algebraicas. Los corchetes [ ] o dobles corchetes [[ ]] para seleccionar partes de un objeto así como el símbolo $. Por ejemplo

x[3] representa la posición 3 del vector x

También se puede crear vectores que en lugar de números contengan caracteres, incluso se pueden nombrar las entradas como por ejemplo en una lista de clase.

lista=c("Andrey", "Rosibel", "Any", "Pedro")

names(lista)=c("Profesor","Profesora", "Estudiante 1", "Estudiante 2")

Estadísticas con R

Al ser R un programa estadístico, es posible calcular todas las estadísticas descriptivas que se requieran tanto para variables como para atributos, no obstante, para el cálculo de algunos estadísticos es necesario instalar bibliotecas que contienen funciones específicas para realizar dichos cálculos, dos ejemplos son las bibliotecas “modeest” y” fmsb” las cuales se utilizan para calcular la moda y los percentiles de una distribución, respectivamente. Para instalar las bibliotecas es necesario tener acceso a internet y posicionarse en la barra de estado en paquetes → instalar paquetes y luego elegir un mirror, para luego descargar las bibliotecas deseadas.

Por ejemplo, suponga que se quiere sacar las estadísticas descriptivas de las notas del primer parcial de un grupo de estudiantes, para ello se crea un vector denominado “notas” como se muestra a continuación:

```
notas=c(74,56,72,40,82,76,72,87,81,50, 65, 62) # se crea el vector de notas
sort(notas) # ordena los valores del vector
order(notas) # da la posición ordenada de menor a mayor
sum(notas) # suma los valores del vector
cumsum(notas) # da la frecuencia acumulada de los datos
length(notas) # da la longitud del vector
min(notas) # da el menor valor de la distribución
max(notas) # da el mayor valor de la distribución
mean(notas) # da el promedio de la distribución
median(notas) # da la mediana de la distribución
quantile(notas) # da los cuatro cuartiles de la distribución
sd(notas) # da de la desviación estándar de la distribución
var(notas) # da la variancia de la distribución
```

Como no hay una función estándar para la moda, para ello hay que bajar la biblioteca “modeest” y usar la función mfv(notas)

```
library(modeest) # carga la biblioteca modeest
mfv(notas) # calcula la moda
library(fmsb) # carga la biloteca fmsb
percentile(notas) # da el percentil que representan los datos del vector notas
```

También es posible tener acceso a ciertas bases de datos que vienen incorporadas en muchas bibliotecas, con el fin de ejemplificar el uso de las funciones que contienen. Por ejemplo Arriaza, Fernández, López, Muñoz, Pérez & Sánchez (2008) plantean un ejercicio muy interesante con la base iris del paquete datasets.

Gráficos con R

En R es posible hacer muchos tipos de gráficos histogramas, gráficos lineales, gráficos circulares y muchos otros más, de hecho la resolución es bastante buena y existen bibliotecas especializadas en gráficos. Los gráficos más comunes, como histogramas, gráficos de barras y gráficos de pastel se pueden trabajar directamente con los siguientes comandos hist(), barplot() y pie() respectivamente.

Ahora bien, lo más común es realizar gráficos como complemento de la información que se quiere presentar, por lo general se resume la información en forma tabular y luego se presenta un gráfico sobre dicha información. Como ejemplo se detallará la construcción de una tabla que relacione el nivel económico de los estudiantes con el género

x= matrix(c(15,10,25,45,10,20),nrow=2) # se define la tabla como una matriz

rownames(x)= c("Mujer", "hombre") # se colocan las etiquetas de las filas

colnames(x)=c("bajo", "medio", "alto") # se colocan las etiquetas de las columnas

```
       bajo medio alto
Mujer    15    25   10
hombre   10    45   20
> 
```

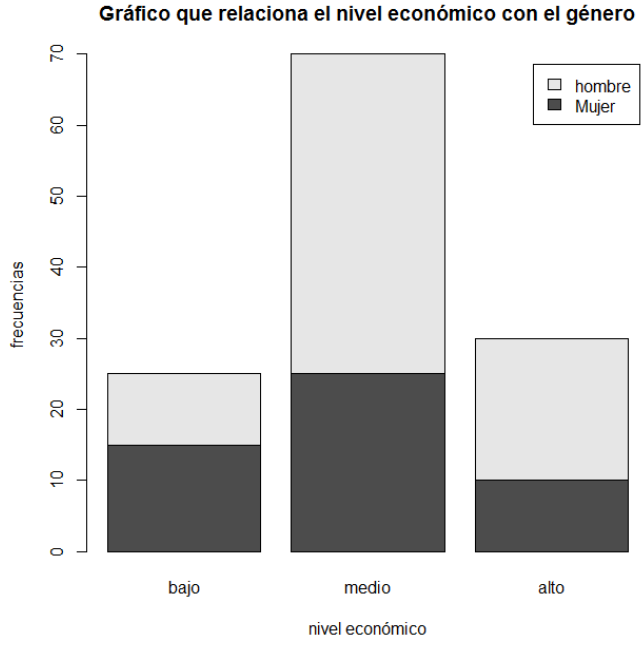
Luego para construir los gráficos que resuman este tipo de información se escriben los comandos

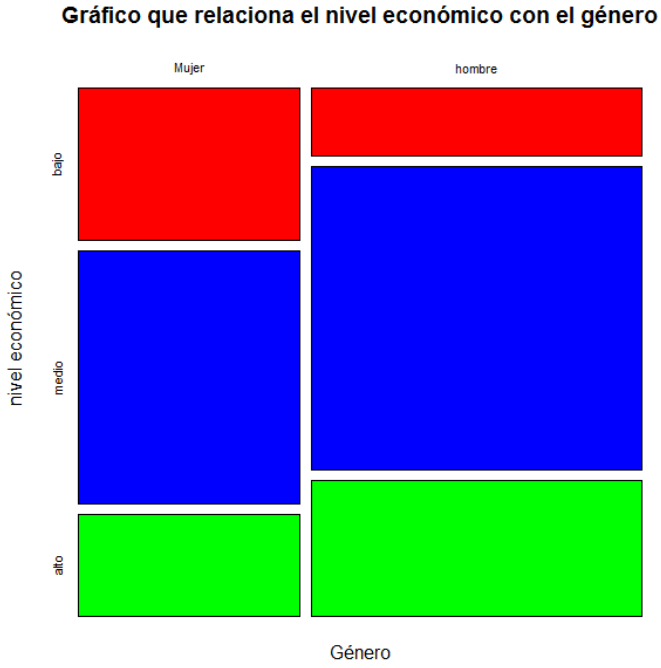
barplot(x, main="Gráfico que relaciona el nivel económico con el género",

xlab="nivel económico", ylab="frecuencias", legend = rownames(x))

mosaicplot(x, col=c("red","blue","green"),main="Gráfico que relaciona el nivel económico con el género", xlab="Género", ylab="nivel económico")

Como puede observarse, esto es tan solo una pincelada de lo que se puede hacer con R, tan solo se necesita un poco de motivación y estar dispuesto a dedicar unas horas de tiempo para poder experimentar una parte de lo que r puede ofrecer.





Actividades

Para poner en práctica lo aprendido hasta ahora se proponen las siguientes actividades

Actividad N°1

1. Realice los siguientes cálculos aritméticos, utilizando el programa R

a) $\frac{24\div-8+7}{9-6\cdot(2)}$

b) $2^3\cdot\ -7+4-\left(\frac{1}{3}+\frac{1}{2}\right)$

c) $\frac{\sqrt{16}+5\cdot4-3^{-2}}{3\cdot(4-8)+1}$

d) $\frac{1}{2}\left[\left(\frac{5}{4}-2^{-1}\right)\cdot\frac{13}{4}-\sqrt{7}+8\div\left(2-\frac{1}{5}\right)-3^2\right]$

e) $2\cdot\sin\left(\frac{\pi}{3}\right)+5\cdot\cos\left(\frac{\pi}{4}\right)-\frac{\tan\left(\frac{\pi}{6}\right)}{4}$

f) $4\cdot\log_2(5)-3\cdot\log_3(7)+\frac{1}{3}\ln^3(8)$

2. Para los ejercicios de la parte 1, redondee los resultados finales a una cifra decimal.

Actividad N°2

1. Suponga que se toma una muestra de 20 taxistas que laboran en el centro de la ciudad de Heredia durante setiembre de 2011. Los datos obtenidos para los 20 taxistas se describen a continuación:

ID	GC	NP	AS
1	8,0	11	9
2	7,6	9	7
3	9,7	9	1
4	12,6	10	7
5	12,9	8	1
6	10,2	9	9
7	14,2	10	6
8	8,4	8	5
9	14,0	13	3
10	13,9	12	3
11	8,7	12	5
12	9,4	12	1
13	7,4	7	8
14	13,4	11	4
15	11,1	13	7
16	13,5	10	9
17	8,6	10	1
18	13,5	11	4
19	9,7	10	7
20	11,6	11	5

ID: Identificación del taxista **NP**: Número de pasajeros transportados por turno

GC: Gasto diario en combustible (en miles de colones) **AS**: Años de servicio como taxista

a) Construya dos vectores y denótelos GC y NP para guardar los datos de Gasto de combustible y Número de pasajeros.
b) Calcule la moda, la mediana, la media, desviación estándar, varianza y los cuartiles, par los datos de GC y NP.
c) Represente gráficamente las variables GC y NP.

Actividad N°3

1. De acuerdo con MIDEPLAN durante el 2008 las distintas regiones de Costa Rica han presentado diferencias en cuanto al porcentaje de hogares pobres (PHP) y la tasa de mortalidad infantil (TMI) por cada mil nacidos vivos, como se muestra a continuación

Indicadores	**Central**	**Indicadores**	**Huetar Norte**	**Indicadores**	**Huetar Atlántica**
PHP	14,0%	PHP	24,7%	PHP	16,4%
TMI	8,8	TMI	10,2	TMI	8,0

Indicadores	**Chorotega**	**Indicadores**	**Pacífico Central**	**Indicadores**	**Brunca**
PHP	26,0%	PHP	25,7%	PHP	24,6%
TMI	9,1	TMI	8,2	TMI	9,4

a) Construya dos vectores y denótelos PHP y TMI para guardar los datos de porcentaje de hogares pobres y tasa de mortalidad infantil.
b) Calcule la moda, la mediana, la media, desviación estándar, varianza y los cuartiles, par los datos de PHP y TMI.
c) Represente gráficamente las variables PHP y TMI.

Actividad N°4

1. Considere el siguiente cuadro referido a un grupo de estudiantes universitarios en cuanto a su género y estado conyugal

	Estado conyugal		
Género	Casado	Soltero	otro
Masculino	8	14	5
Femenino	12	9	10

a) Construya la tabla anterior mediante el programa R.
b) Calcule las marginales y las proporciones basadas en el total de la muestra.
c) Represente gráficamente la tabla anterior (realice al menos dos gráficos diferentes).

Referencias Bibliográficas

Arriaza, A.J, Fernández, F, López, M.A, Muñoz, M, Pérez, S & Sánchez, A (2008). Estadística Básica con R y R- commander. Servivio de publicaciones de la Universidad de Cádiz. http:// knuth.uca.es/ebrcmdr.

Conesa, D. (marzo, 2011) Grup d'Estadística Espacial i Temporal en Epidemiologia i Medi Ambient Dept. d'Estadística i Investigació Operativa Universitat de València recuperado de http://www.uv.es/conesa/CursoR/material/handout-sesion1.pdf

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Verzani, J (2004). Using R for introductory Statistics. Chapman & Hall/CRC.