https://doi.org/10.24275/AQRF5869

La estadística computacional: una propuesta didáctica

Computational statistics: a teaching approach

*Isabel Quintas Pereira**

Resumen

Se presenta la propuesta de enseñanza de temas de matemáticas, en particular la estadística postulada por Nolan, Temple Lang, Kaplan, Pruim y Horton, quienes muestran que la manera moderna de enseñar estadística requiere del uso intensivo de la computación y la capacidad gráfica actual. Se propone el uso del lenguaje R, ya que en licenciaturas como economía, sociología o biología, en las que se necesita analizar cantidades crecientes de datos, la programación está ausente de los currículos. La propuesta se acompaña de ejemplos del material diseñado utilizando *software* libre para la enseñanza de estadística, cálculo y álgebra lineal del Proyecto mosaico, y de la adaptación hecha sobre éstos.

Palabras clave: enseñanza de estadística, *software* libre, lenguaje R, computación, Proyecto mosaico.

Abstract

This is a proposal to teach some mathematical topics, statistics in particular, that Nolan, Temple Lang, Kaplan, Pruim and Horton developed as part of the Mosaic Project, trying to show a modern approach to statistical education, making an intense use of computational and graphics utilities using the R computing language. But the student of these different subjects needs to have some computational topics in the curriculum, which do not presently exist. Their proposal came with materials for instructors and students, using R to teach statistics and calculus. Some examples of that material are included here.

Key words: statistical education, free software, R language, computing, Mosaic project.

Artículo recibido: 12/11/19
Apertura del proceso de dictaminación: 17/04/20
Artículo aceptado: 20/04/20

* Profesora-investigadora, departamentos de Política y Cultura y Producción Económica, UAM Xochimilco, México [iquintas@correo.xoc.uam.mx].

No hay enseñanza sin investigación ni investigación sin aprendizaje [...] Investigo para constatar, constatando intervengo, interviniendo educo y me educo. Investigo para conocer lo que aún no conozco y comunicar lo novedoso.

Paulo Freire

Cuando en vez de maquilador seamos un país creador, se requerirá de una educación superior amplia, compleja, profunda y humanística.

Víctor Luis Porter Galetar

INTRODUCCIÓN

El nacimiento de esta propuesta se debe a dos comunidades absolutamente diferentes en su quehacer cotidiano, que se encontraron en un espacio común: la educación.

Mosaic Project[1] es una comunidad educativa innovadora –surgida en 2001, en la región de la bahía de San Francisco– que proponía métodos alternativos de educación, propiciando la inclusión racial de una sociedad muy heterogénea: niños de color, hijos de inmigrantes, todos. Proponían una educación basada en proyectos de arte, investigación al aire libre, con actividades lúdicas, enfocándose en las ciencias ambientales y un aprendizaje social emocional que propiciara la inclusión y la tolerancia.

El proyecto inicialmente estaba dirigido a niños de 10 a 12 años, pero en poco tiempo se extendió a jóvenes, con un importante componente artístico: crear mosaicos colectivamente, para desembocar en un proyecto que no sólo incluye arte, sino que se extiende a la investigación científica y especialmente al entrenamiento y educación de instructores y facilitadores de distintas disciplinas en distintos lugares, con el objetivo de apoyar a las comunidades locales.

En 2011, los profesores Randall Pruim, Nicholas Horton y Daniel Kaplan organizaron talleres para ayudar a instructores y profesores de estadística de

[1] Proyecto mosaico [https://mosaicproject.org/es/] [https://mosaicproject.org], fecha de consulta: enero de 2019.

nivel medio y superior a introducir el lenguaje R –y una serie de herramientas tecnológicas disponibles en el ambiente de R– en sus cursos de estadística.[2] Estos talleres se ofrecieron antes de la Conferencia sobre la Enseñanza de la Estadística en Estados Unidos (United States Conference on Teaching Statistics –USCOTS, 2011). Los talleres fueron exitosos, se mejoró el material y se repitieron para USCOTS 2013, eCOTS 2014 (Electronic Conference on Teaching Statistics), ICOTS 2015 (International Conference on Teaching Statistics) y USCOTS 2015. El material desarrollado inicialmente para estos talleres, titulado "Enseñar estadística usando R", ha sido difundido, ampliado y hecho público por los autores, asociados con el Proyecto mosaico, bajo una licencia pública que promueve compartir, copiar, adaptar y transmitir todo lo publicado.[3]

Esta propuesta de una nueva didáctica de la estadística computacional surge de estas dos comunidades alternativas: por un lado, la comunidad del Proyecto mosaico y, por otro, la comunidad del *software* libre, la cual opta por el lenguaje R y que se ve reflejada en un nuevo Proyecto r mosaico para la enseñanza de la estadística y las matemáticas.

El comienzo de la comunidad del *sotfware* libre puede encontrarse a principios de la década de 1980, cuando un grupo de programadores, científicos de la computación, la mayoría pertenecientes a universidades, se organizan para oponerse a la naciente industria del *software* y, liderados por Richard Stallman, instauran el concepto de *copyleft* (licencia GNU), esto es, que aquellos programas o códigos liberados con *copyleft* pasan a ser de dominio público, incluso el código fuente, y pueden ser utilizados, modificados, adaptados y distribuidos a otros usuarios, siempre y cuando los productos realizados con éstos sean también liberados bajo el mismo tipo de licencia.[4]

[2] N. Horton, R. Pruim y D. Kaplan, *Start Teaching With R*, ed. 1.1, Mosaic Project, 2015 [https://github.com/ProjectMOSAIC/LittleBooks].

[3] La leyenda dice: "This material is copyrighted by authors under a Creative Commons Attribution 3.0 Unported License. You are free to *Share* (to copy, distribute and transmit the work) and *Remix* (to adapt the work) if you attribute our work. More detailed information about the licence is available at this web page: http://mosaic-web.org/go/teachingRLicencse.html".

[4] El proyecto GNU se anunció durante 1983, cuando un grupo de programadores proponían el regreso al espíritu de colaboración imperante durante las décadas de 1960 y 1970, cuando la programación era privativa de los círculos científicos –en los que se trabajaba en grupos– y de muy pocas grandes empresas que tenían máquinas y programas exclusivos. Liderado por Richard Stallman, este grupo escribió el sistema operativo GNU, compatible con UNIX (GNU viene de NO es UNIX) y promueven la licencia pública general de código abierto para que la comunidad lo pueda utilizar, mejorar y depurar para el beneficio de todos. A partir de entonces han aparecido diferentes grupos que impulsan el *software* libre, aunque variando en sus limitaciones y alcances. Todavía está en vigor en esa comunidad un debate con respecto

El lenguaje R surge en esta comunidad en la década de 1990; fue creado por Ross Ihaka y Robert Gentleman como un dialecto del lenguaje S de los laboratorios ATT, pero con licencia GNU. Este lenguaje fue escrito especialmente para la estadística y ha sido adoptado por esta comunidad, que lo ha enriquecido; se han desarrollado módulos para resolver problemas específicos de diferentes disciplinas, así como incrementar las capacidades gráficas. Actualmente, R es un estándar en estadística en casi todas las disciplinas; se han escrito varios ambientes de desarrollo integrados que facilitan la interacción con el lenguaje, el manejo de los datos, el trabajo por proyectos, la instalación de paquetes externos, la graficación, la depuración de código, etcétera. La propuesta del Proyecto r mosaico usa RStudio[5] como su entorno de desarrollo integrado (IDE).

EL NUEVO PARADIGMA DE LA ENSEÑANZA DE LA ESTADÍSTICA

En su artículo de 2010, "Computing in the Statistics Curricula",[6] los doctores Deborah Nolan y Duncan Temple Lang proponen la necesidad de cambiar significativamente la forma de enseñar estadística, así como el contenido de los programas. Debido al uso generalizado de la computación y el acceso a una cantidad ilimitada de datos disponibles en las redes, el uso y la naturaleza de

---

a la fina línea entre qué restricciones pueden aplicarse y qué puede llamarse *libre* todavía. El siguiente hito fue el sistema operativo Linux: Ken Thomson junto con Dennis Ritchie decidieron escribir nuevamente el sistema operativo en lenguaje ensamblador para incrementar su velocidad, lo que los llevó a desarrollar el sistema UNIX, que en 1991 se transformó en Linux, para PC, con licencia GNU. El lenguaje R surge como una alternativa de *software* libre al lenguaje S de los laboratorios Bell de AT&T, desarrollado para realizar análisis de datos. Fue desarrollado inicialmente por Gentleman e Ihaka de la Universidad de Auckland en Nueva Zelanda (es importante notar cómo el *software* libre nace en las universidades, en un ambiente de colaboración entre colegas). Fue liberado bajo licencia GNU y es un ejemplo de desarrollo por colaboración de tal vez miles de programadores. En la última década se ha convertido en el lenguaje más utilizado en investigación científica, aunque inicialmente fue desarrollado para la estadística: en realidad, R es un lenguaje de programación. Esto ha permitido que se le anexen módulos (*packages*) para áreas como la minería de datos, la bioestadística, la bioinformática o la inteligencia artificial. Como programa, se trata de un lenguaje orientado a objetos, de tipo intérprete, con una gran capacidad gráfica.

[5] RStudio posee versiones libres tanto para laptops como para servidores, y versiones bajo licencia comercial tanto para computadoras personales como para servidores.

[6] Deborah Nolan y Duncan Temple Lang, "Computing in the Statistics Curricula", *The American Statistician*, vol. 64:2, 2010, pp. 97-107 [https://doi.org/10.1198/tast.2010.09132].

la estadística han cambiado; actualmente, es un insumo en todas las disciplinas y, por supuesto, las ciencias sociales no son ajenas a su impacto. Es necesario entonces cambiar el enfoque, se necesita una alfabetización en computación y programación simultánea a las matemáticas, que incluyan mucho más que el uso de la computadora como una herramienta que hace los cálculos; es necesario lograr habilidades para el manejo de bases de datos grandes, el procesamiento computacional intensivo para el análisis de estos datos y la comprensión de las limitaciones de las inferencias que se realicen a partir de éstos. Los autores proponen la enseñanza de la estadística "en combinación con problemas científicos y métodos estadísticos modernos" para enseñar a *pensar con los datos*.

La filosofía de este nuevo paradigma para la enseñanza de la estadística se enfoca en la modelación de problemas y situaciones, la graficación de datos, la inferencia obtenida por el re-muestreo,[7] haciendo uso tanto de modelos estadísticos como de técnicas gráficas multivariadas que permite el lenguaje R; ellos sostienen que la estadística y el análisis de datos es imposible sin la computación, y ésta apenas aparece en los currículos de varias disciplinas y está ausente en las ciencias sociales.

Pero esta nueva pedagogía de la estadística requiere de un esfuerzo adicional de capacitación para que los maestros puedan adoptarla. Varios son los autores que en dicho sentido han escrito artículos, impartido talleres y producido material didáctico, todos ellos profesores de departamentos de estadística y computación en diversas universidades y colegios. Uno de los más prolíficos es N. Horton, del Amherest College, quien propone un currículum para nivel medio superior con la premisa de "Pensar con los datos";[8] por su parte, Ben Baumer[9] propone un curso para el nivel medio superior[10] que sigue la misma filosofía.

Desde 2011, Pruim, Kaplan y Horton, unidos en el Proyecto r mosaico, como una comunidad de educadores trabajando para desarrollar una nueva manera de introducir las matemáticas, la estadística, la computación y el

[7] Se llama *re-muestreo* a la obtención de distintas muestras aleatorias a partir de la base de datos original para introducir los conceptos de confiabilidad y variabilidad.

[8] N. Horton y J. Hardin, "Teaching the next Generation of Statistics Students to 'Think with Data': Special Issue on Statistics and the Undergraduate Curriculum", *The American Statistician*, vol. 69:4, 2015, pp. 259-265.

[9] B. Baumer, "A Data Science Course for Undergraduates: Thinking with Data", *The American Statistician*, vol. 69:4, 2015, pp. 334-342.

[10] Que corresponde al nivel de los cursos de estadística para licenciatura en las ciencias sociales.

modelado a los estudiantes de nivel medio superior y de las universidades, se dedicaron a la capacitación de profesores e instructores de estadística mediante cursos y talleres. Para ello escogieron al lenguaje R como su herramienta computacional, y para facilitar el manejo del lenguaje a los instructores legos en programación, diseñaron un paquete que provee cierto número de funciones necesarias en los cursos de estadística y cálculo, así como una estandarización de la sintaxis del lenguaje y la producción de gráficas; se trata del paquete *mosaic* (*mosaic package*).[11] En 2015 se publicó el material utilizado y mejorado durante dichos talleres: *Start Teaching With R* para los instructores y *Student's Guide to R* para los estudiantes, ambos bajo licencia Creative Commons Attribution 3.0, para difundir su propuesta de enseñanza de la estadística.

LA PROPUESTA DE CAPACITACIÓN DEL PROYECTO R MOSAICO

El material de los profesores Pruim, Kaplan y Horton –preparado inicialmente para los talleres de la Conferencia sobre Enseñanza de la Estadística de 2011– se llamó "Enseñar estadística usando R"; estaba dirigido a personas con conocimientos suficientes de estadística pero, en la mayoría de los casos, sin una preparación en programación, ya que esta asignatura fue retirada de los currículos de casi todas las disciplinas una vez que apareció Windows y que los usuarios se acostumbraron a aplicaciones que sólo requerían un clic del ratón para seleccionar una opción de las disponibles en un menú diseñado por otra persona. Las más de dos décadas de usar menús y ratones produjo un temor a utilizar la programación, incluso un rechazo, tildando a las herramientas que lo requieren de "no actualizadas". Para el caso particular del lenguaje R, dentro de la comunidad académica usuaria de la estadística, pero sin formación en programación, se dice que es un lenguaje "con una curva de aprendizaje muy lenta".

Para contrarrestar este problema, los autores escribieron el paquete *mosaic*, que facilita el uso de R; el paquete *mosaic* accesa a un buen número de otros paquetes de R, que llevan a cabo funciones específicas sin que los principiantes tengan que conocer los detalles engorrosos de éstas. Incluye cerca de 100 funciones para graficar, realizar análisis básico de datos, modelos estadísticos, pruebas estadísticas, calcular parámetros, realizar operaciones del álgebra

[11] El *package* Mosaic forma parte de las librerías actuales de R y se baja del CRAN como cualquier otro paquete.

lineal y del cálculo, ajustar datos, encontrar raíces de ecuaciones, trabajar con mapas de Google, incluir videos, hacer conversiones, etcétera. También incluye varias bases de datos que se utilizan en los ejemplos demostrativos; como ambiente de desarrollo utilizan el IDE RStudio.

El material para este curso se presenta en los libros *Start Teaching With R*[12] y *A Student's Guide to R*;[13] el primero es una guía para el maestro y el segundo es para el estudiante; ambos fueron publicados bajo licencia libre Creative Commons Attribution 3.0 Unported License.[14]

La guía para el maestro, desarrollada para capacitar a los instructores de estadística de nivel inicial e intermedio, contiene capítulos introductorios para dotarlos de la experiencia básica necesaria del lenguaje R y de RStudio. Esta introducción habilita al instructor para que tenga desde el comienzo herramientas suficientes y pueda desarrollar su propio material, aunque el dominio del lenguaje sea un proceso gradual.

Propone una serie de estrategias que pueden seguirse como guía para los diferentes cursos:

1. Hacer algo con R desde la primera clase.
2. Mostrar con ejemplos qué hace R antes de pedirle a los estudiantes que hagan algo.
3. Enseñar las particularidades de R como lenguaje, por ejemplo, la necesidad de la corrección sintáctica explícitamente. En particular, insistir en la sintaxis de las funciones de R, su estructura, qué hace y qué información necesita para hacerlo.
4. Textualmente dice "Menos volumen más creatividad" y propone que se enseñen menos métodos pero aplicados a distintas situaciones, para que el estudiante pueda transferir ese conocimiento a otros problemas.
5. Tratar de tener exámenes en computadora para motivar el aprendizaje del lenguaje.

[12] N. Horton, R. Pruim y D. Kaplan, *Start Teaching With R*, *op. cit.*

[13] N. Horton, R. Pruim y D. Kaplan, *A Student's Guide to R*, ed. 1.2, Mosaic Project, 2015 [https://github.com/ProjectMOSAIC/LittleBooks].

[14] La leyenda dice textualmente: "This material is copyrighted by the authors under a Creative Commons Attribution 3.0 Unported License. You are free to Share (to copy, distribute and transmit the work) and to Remix (to adapt the work) if you attribute our work. More detailed information about the licensing is available at this web page: http://www.mosaic-web.org/teachingRlicense.html".

6. Repensar los cursos: deben desaparecer las tablas estadísticas; no enseñar distintas fórmulas para facilitar el cálculo en tablas, sólo las fórmulas más intuitivas, enfocadas a los conceptos, no en el cálculo.
7. No complicar los comandos, utilizar los más simples posibles.

Y para esto propone dos tácticas:

1. Introducir inmediatamente las gráficas e insistir en su interpretación.
2. Introducir la generación de datos simulados y el muestreo repetido sobre una misma población.

Para llevar a cabo esta estrategia, el material presenta en ocho capítulos una serie de actividades que se proponen para los cursos introductorios. Las distintas actividades utilizan bases de datos reales sobre demografía, epidemiología y datos de estadísticas deportivas que han sido incorporados al paquete *mosaic*. De manera transparente para el usuario, el paquete carga otras utilerías de R que se necesitan para su funcionamiento. El contenido de los capítulos es el siguiente:

1. Introducción al proyecto.
2. Cómo utilizar RStudio y hacer una gráfica.
3. Ejemplos para comenzar un curso con estadística descriptiva.
4. Algo de sintaxis del lenguaje y sobre graficación.
5. Simulación e inferencia.
6. Lo que los alumnos deben saber de R.
7. Lo que los instructores deben saber de R.
8. Promoción del trabajo interactivo.

A continuación se presentan tres de las situaciones didácticas propuestas por el grupo de *R Mosaic* para un curso para principiantes; cada situación está pensada para introducir algún tema del programa de estudios. El instructor puede elaborar su clase a partir de estas prácticas.

PRIMERA SITUACIÓN: DISTRIBUCIÓN DE PROBABILIDAD

Los autores hacen referencia a un experimento incluido por R.A. Fisher[15] en su texto de 1925 sobre la metodología estadística, en el cual plantea –a partir de una situación ocurrida años antes, con una señora inglesa que se dice catadora de te–[16] un ejemplo pedagógico para comenzar con los conceptos de probabilidad y distribución de probabilidad. Para abordar este tema se propone un ejemplo similar: arrojar una moneda. El resultado puede ser cara o cruz, o éxito o fracaso, igual que en el ejemplo de la catadora.

Se arrojan 10 monedas y se cuentan los "éxitos"; se dirá que es éxito si cae cara. Se vuelven a tirar 10 monedas muchas veces y se cuenta el número de éxitos en cada tirada. Se cuentan las veces que se obtuvo un éxito, dos éxitos, tres éxitos... ¿Cuántas veces habrá que repetirlo para tener una buena representación? *Mosaic* nos ofrece la función *rflip*( ) para emular a la moneda; basta con teclear

```
require (mosaic)
>  rflip()

Flipping 1 coin [ Prob(Heads) = 0.5 ] ...

T

Number of Heads: 0 [Proportion Heads: 0]

> rflip (10)

Flipping 10 coins [ Prob(Heads) = 0.5 ] ...

T T H T H T H H H H

Number of Heads: 6 [Proportion Heads: 0.6]
```

El texto propone primero hacer y observar el resultado, sólo después explicar las instrucciones; se trata de aprender haciendo, y se aprende la estadística y simultáneamente la programación.

[15] R.A. Fisher (1890-1962), biólogo y matemático, considerado el creador de la estadística moderna.

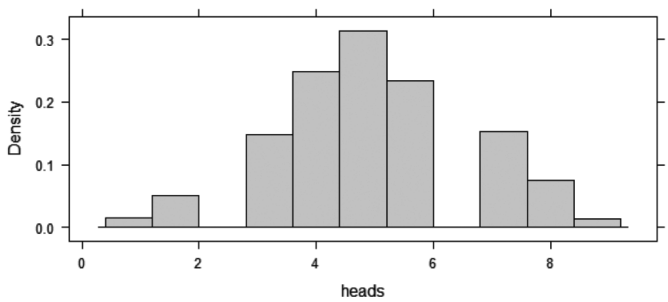
[16] Cuenta que en una reunión la señora aseguró ser capaz de reconocer cómo había sido preparado el te. Fisher propuso entonces que la señora probara tal aseveración presentándole varias tazas de te. En su texto de 1925 lo convirtió en una situación didáctica: ¿cuántas tazas debería probar?, ¿cuántas acertó?, ¿a qué conclusión se llega?

Entonces se agrega una opción para repetirlo *n* veces, arrojar 10 monedas varias veces, para lo que se teclea

```
> do (3) * rflip(10)
  n heads tails prop
1 10     5     5  0.5
2 10     6     4  0.6
3 10     4     6  0.4
```

Pero para realizar el experimento mil veces habrá que guardar los resultados, en este caso en la variable *salidas*:

```
# se guardan los resultados de 1000 tiradas de 10 monedas
>   salidas <- do (1000) * rflip(10)
>
> tally(~ heads, data = salidas)

heads
  1   2   3   4   5   6   7   8   9
 12  40 118 199 251 187 122  60  11
> histogram (~heads, data = salidas, whith =1)
```



Este sencillo ejercicio permite observar que sólo aproximadamente 25% de las veces se obtienen cinco caras y cinco cruces, pero obtener ocho caras apenas ocurrió en 6% de los casos, etcétera. Como la función *rflip*( ) utiliza un método *random*, cada estudiante tendrá valores diferentes. La función *tally*( ) da un resumen de los datos y la función *histogram*( ) grafica el histograma de

los datos. El estudiante puede deducir qué hace cada instrucción al observar el resultado. Cada pequeño *script* va agregando nuevas instrucciones del lenguaje. El experimento podrá repetirse el número de veces que se desee.

SEGUNDA SITUACIÓN: CÁLCULO DE PARÁMETROS DESCRIPTIVOS

Para el estudio descriptivo de variables cuantitativas interesa la media, la mediana, la desviación estándar, el rango intercuartílico, etcétera; en este caso, se utiliza una base de datos de un Centro de Investigación Sobre Depresión (CESD), que entre otras muchas variables –como edad, sexo, adicción, tipo de tratamiento, etcétera– incluye un índice *cesd*, indicativo de la gravedad de la depresión, que varía de 0 a 60.

El siguiente código carga los datos

```
> require (mosaic)
> require(mosaicData)
> options(digits=4)
> head(HELOrct,3)
```

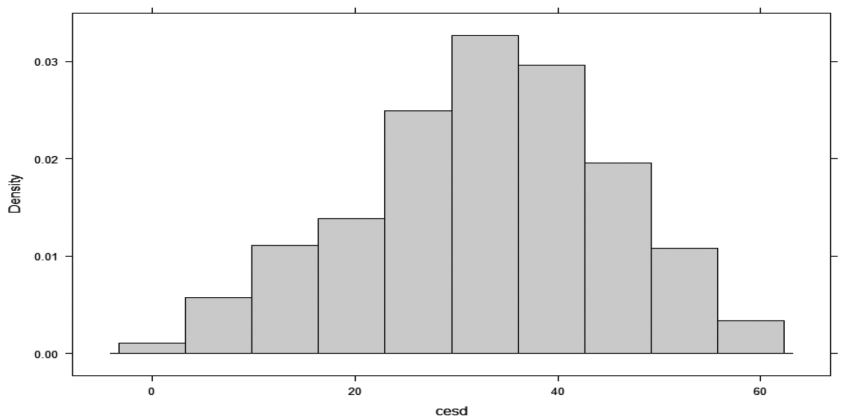
Para obtener los valores de los parámetros de interés simplemente debe teclearse:

```
> # valor medio
> mean( ~cesd, data = HELPrct)
[1] 32.85
> mean(HELPrct$cesd)
[1] 32.85
>
> #desviación estandar
> sd(~cesd, data = HELPrct)
[1] 12.51
> sd(HELPrct$cesd)
[1] 12.51
>
> # mediana
> median(~cesd, data = HELPrct)
[1] 34
```

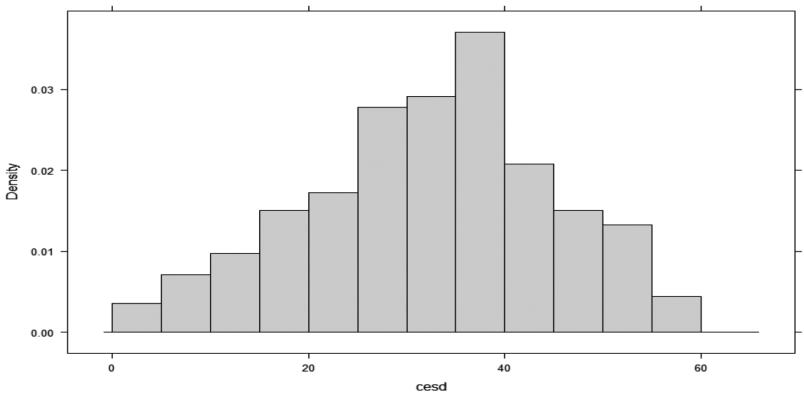
```
> # cuartiles
> with(HELPrct, quantile(cesd))
0%  25%  50%  75% 100%
1   25   34   41   60
>
> with(HELPrct, quantile(cesd, c(.025,  .975)))
2.5% 97.5%
6.3  55.0

> # resumen  Ésta es una función de Mosaic. En R se llama summary()
>
> favstats(~cesd, data = HELPrct)
min Q1 median Q3 max  mean    sd   n missing
1 25     34 41  60 32.85 12.51 453       0
```

La función *histogram*( ) se utiliza para observar la distribución de frecuencias de los datos; cuenta con opciones para indicar el ancho de las barras y el centrado de éstas.

```
> histogram(~cesd, data = HELPrct)


>
```



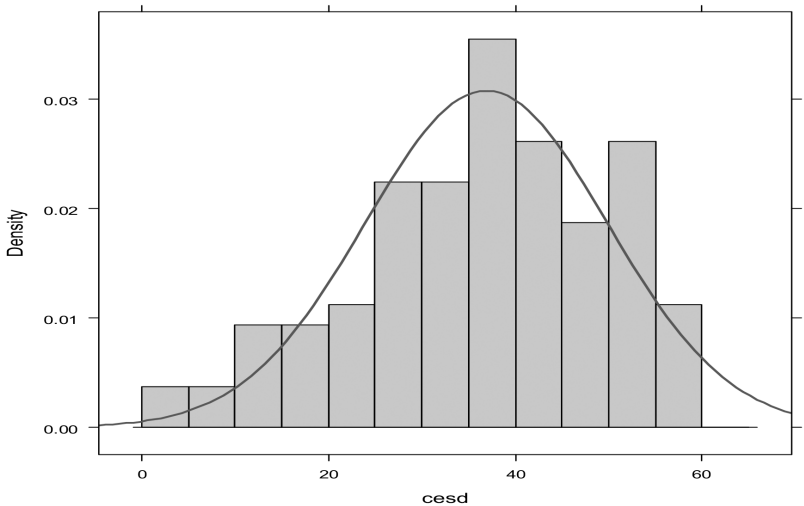
```
> histogram(~cesd, width=5, center=2.5, data = HELPrct)
```



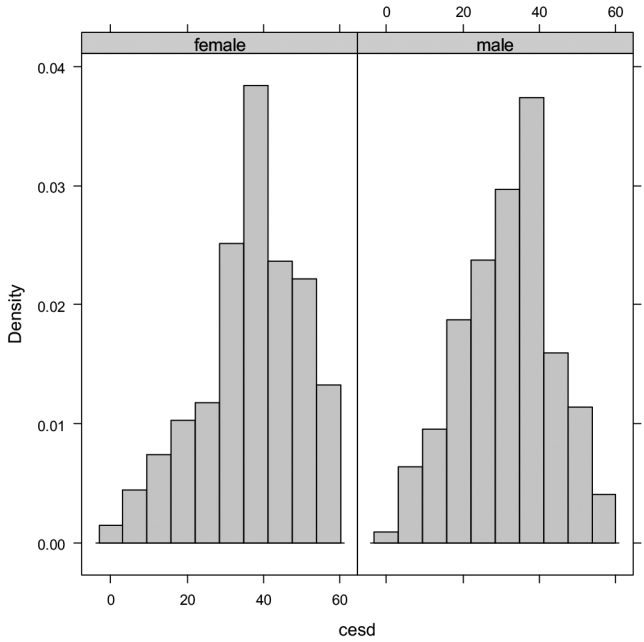

```
> # detalle por sexo
> tally(~ sex, data = HELPrct)
sex
female   male
   107    346

>
```

Si se desea separar el archivo en uno de hombres y otro de mujeres, puede utilizarse la función *filter*( ).

```
> mujeres  <- filter(HELPrct, sex=='female')
> hombres  <- filter(HELPrct, sex=='male')
>

> histogram(~ cesd, width=5, center=2.5, data = mujeres)
> histogram(~ cesd, width=5, center=2.5, fit ="normal",data = mujeres)


> histogram(~ cesd | sex, data = HELPrct)
```



También se pueden tener gráficas simultáneas para propósitos de comparación.



TERCERA SITUACIÓN: ANÁLISIS DE DATOS

Se utiliza la base de datos *Births78*, que contiene el número de nacimientos ocurridos en Estados Unidos cada día de 1978. El tipo de información de dicha base puede conocerse observando su encabezado. Un diagrama de dispersión puede servir para buscar algún patrón interesante en dichos datos; por ejemplo, el número de niños nacidos cada día, ¿sigue algún patrón?

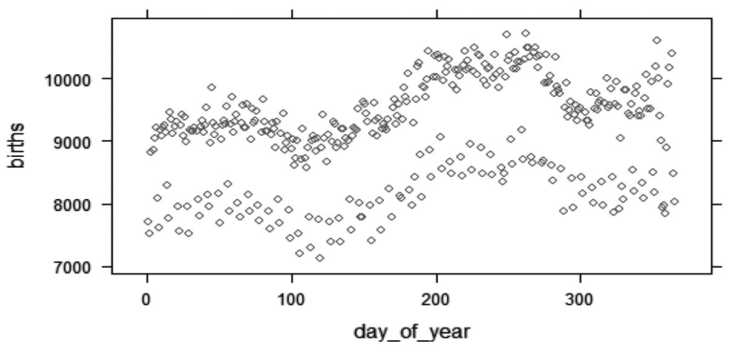
```
> require (mosaic)
> head (Births78)
```

	date	births	wday	year	month	day_of_year	day_of_month	day_of_week
1	1978-01-01	7701	Sun	1978	1	1	1	1
2	1978-01-02	7527	Mon	197	1	2	2	2
3	1978-01-03	8825	Tue	1978	1	3	3	3
4	1978-01-04	8859	Wed	1978	1	4	4	4
5	1978-01-05	9043	Thu	1978	1	5	5	5
6	1978-01-06	9208	Fri	1978	1	6	6	6

Al cargar el paquete *mosaic* varias bases de datos están disponibles, y entre éstas se encuentra *Births78*. La instrucción *head*( ) permite conocer cómo se llama cada uno de los campos o variables del archivo y muestra los primeros registros. Las variables del archivo son: la fecha, el número de nacimientos, el día de la semana, el año y el mes, y luego el día del año, del mes y de la semana.

Para obtener una gráfica que muestre el número de nacimientos ocurridos cada día de 1977 se traza una gráfica de dispersión; para ello se utiliza la función *xyplot*( ).

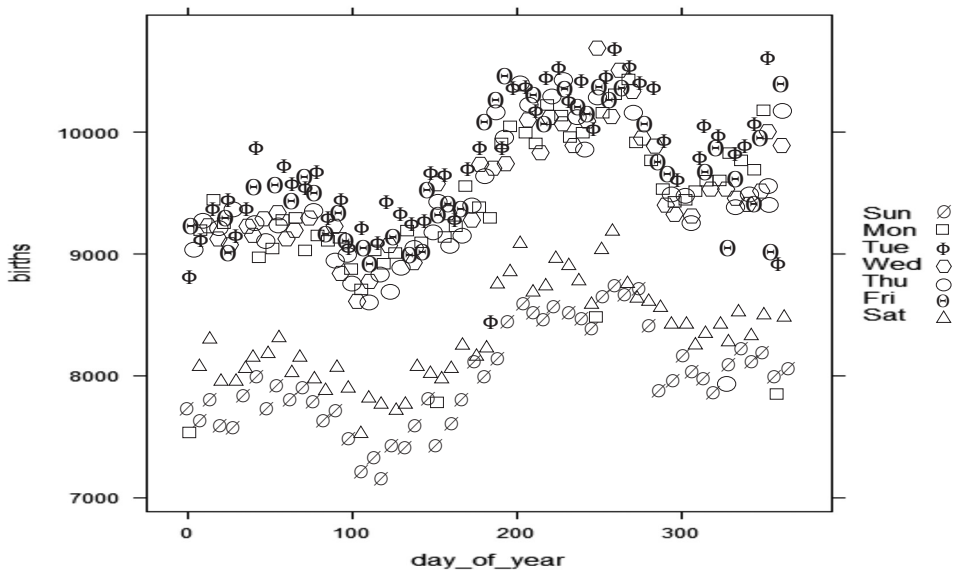
```
> xyplot (births ~day_of_year, data = Births78)
```



La gráfica muestra un patrón que difiere de lo que podría esperarse, por ejemplo, que el número de nacimientos se mantuviera aproximadamente constante; por el contrario, se ve que hay una oscilación en el año, y una tendencia creciente. Además, se distinguen dos bandas claramente diferenciadas. Puede pedirse a los estudiantes que la describan y traten de explicar el comportamiento.

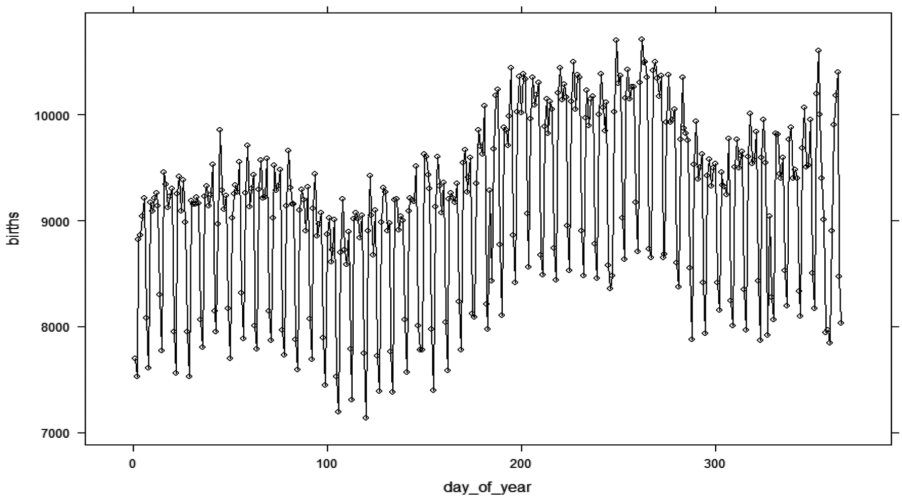
Una opción es modificar la gráfica para que permita distinguir los días de la semana:

```
xyplot(births ~ day_of_year,data =Births78, groups=wday,
       auto.key=list(space="right"))
```



Otra opción puede ser unir los datos consecutivos con un simple cambio de parámetros en la instrucción *xyplot*( ).

```
xyplot(births ~ day_of_year,data =Births78, typ ="b")
```



El ejemplo permite ver la capacidad de análisis gráfico que ofrece el lenguaje R con apenas unas instrucciones, las cuales pueden dominarse en un par de clases. La propuesta es presentar datos para el análisis e interpretación, no quedarse en el cálculo de parámetros o estimadores estadísticos de las variables. El objetivo es aprender con los datos.

ADAPTACIÓN DE LA PROPUESTA

Se creyó necesario repensar el modelo propuesto por el Proyecto r mosaico por varios motivos: el primero, la importancia de tener el material en castellano;[17] si bien las instrucciones de R –como las de cualquier otro lenguaje de programación– están en inglés, el contexto será más claro en nuestro idioma. El segundo motivo es utilizar bases de datos relacionadas con las temáticas que sean interesantes para los alumnos a las que están dirigidas, por ejemplo, de biología, economía, demografía o sociología. Para los ejemplos que se presentarán se utilizan datos obtenidos del Instituto Nacional de Estadística y Geografía (Inegi).

La propuesta que presentamos no hace uso del paquete *r mosaic*; se usan directamente las instrucciones del lenguaje, pues para los cursos del

[17] El grupo *r mosaic* ya tiene una traducción preliminar de la guía para el estudiante, la cual puede descargarse de https://github.com/jarochoeltrocho/MOSAIC-LittleBooks-Spanish; fue traducida por Francisco Javier Jara Ávila como una colaboración de la Universidad de Costa Rica.

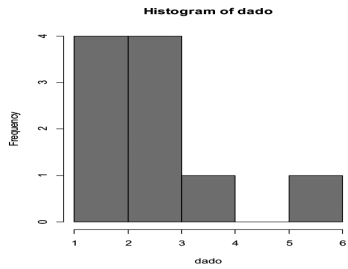
nivel para principiantes los requisitos del lenguaje R pueden obtenerse del paquete básico sin agregar complejidad a la sintaxis, como se mostrará en las siguientes situaciones didácticas propuestas.

PRIMERA SITUACIÓN BIS: DISTRIBUCIÓN DE PROBABILIDAD

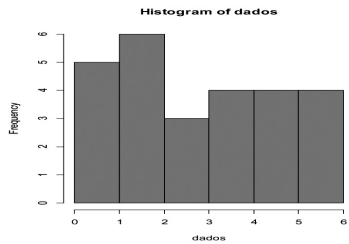
El ejercicio que se plantea es similar al presentado para ver el tema de distribución de probabilidad; se trata de simular que se tira un dado 10 veces y se cuenta cuántas veces se obtiene 1, cuántas 2, 3, 4, 5 o 6 y se presenta una gráfica con las frecuencias obtenidas. Para esto sólo se utilizan dos instrucciones que habrá que explicar posteriormente: *ceiling(6*runif(10)*, que simula nuestro dado, y la función que grafica histogramas.

Se repite el ejercicio, pero tirando el dado 25 veces y se vuelve a graficar; se repite para un número mayor y se analizan los resultados. ¿Qué pasa con la gráfica de las frecuencias? Se trata de observar que la distribución de probabilidad tiende a ser uniforme.

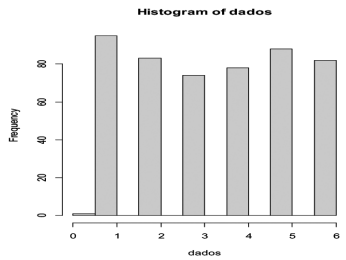
```
> # ejemplo 1
>
> ceiling(6*runif(10))
 [1] 5 6 2 1 3 6 2 3 3 6
```




```
> dado <- ceiling(6*runif(10))
> hist(dado,col=2)
>
> # ejemplo 2
>
> dado <- ceiling(6*runif(25))
> hist(dado,col=3)
```




```
>  # ejemplo 3
> dado <- ceiling(6*runif(500))
> hist(dado, col=5)
```



Para introducir otra distribución de probabilidad podemos proponer que se arrojen dos dados simultáneamente y que la variable sea la suma de lo obtenido. El único cambio es que la variable será suma = dado 1 + dado 2.

```
> suma <- ceiling( 6*runif(20))+ ceiling( 6*runif(20))
> suma
 [1]  6  4  7 12  8  6 11  8  6  9  9  7  7  6  6  7 10  5  3  5
```

Esta situación didáctica permite introducir la generación de números aleatorios, la simulación de situaciones para la experimentación y el tema de distribución de probabilidad, mostrando un caso en que la distribución es uniforme y otro que tiende a una normal, con apenas un par de instrucciones del lenguaje.

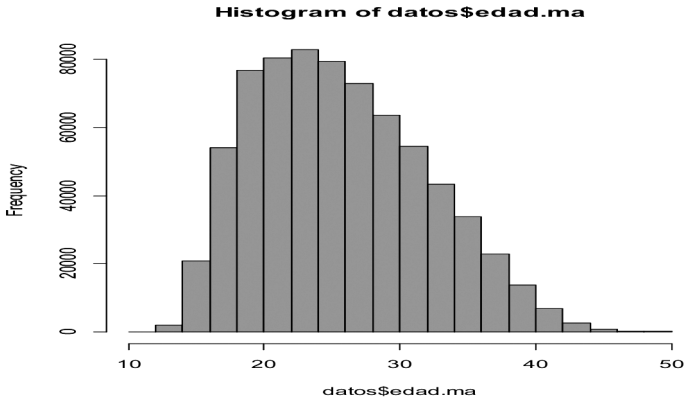
SEGUNDA SITUACIÓN BIS: CÁLCULO DE PARÁMETROS DESCRIPTIVOS

Para esta situación se obtuvieron datos de natalidad del Inegi; se trata de un archivo en formato *.csv* (separado por comas), con información de las edades de las madres y los padres de casi 750 mil nacimientos que ocurrieron y fueron registrados durante 2018.

El primer paso es leer el archivo para cargar los datos, lo que crea un cuadro de datos (*data frame*). El archivo posee la edad de la madre y el padre de todos los bebés nacidos y registrados durante 2018. En esta situación didáctica se aprenderá a leer un archivo externo, observar la información que contiene, calcular los parámetros descriptivos como media, mediana, desviación, etcétera; se presentan gráficas que permiten visualizar de distintas maneras alguna característica –como la edad– para el conjunto de todos los padres.

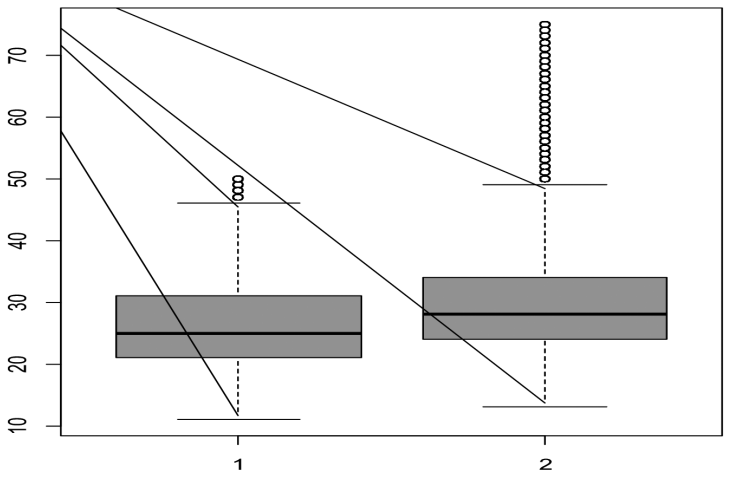
```
> datos < read_csv("C:/Users/52551/Desktop/Articulo/Nacimientos/nat.2018.csv")
Parsed with column specification:
cols(
  No = col_character(),
  sexo = col_double(),
  edad.ma = col_double(),
  edad.pa = col_double(),
  dia = col_double(),
  mes = col_double(),
  anio = col_double()
)
```

```
|=========================================| 100%  13 MB
> head (datos)       # muestra encabezado y primeras líneas del archivo
# A tibble: 6 x 7
     No    sexo edad.ma edad.pa   dia   mes  anio
     <chr> <dbl>   <dbl>   <dbl> <dbl> <dbl> <dbl>
     1 1       2      37      38    14     3  2018
     2 2       2      27      30    13     8  2018
     3 3       2      21      22    27     4  2018
     4 4       2      20      NA    16     5  2018
     5 5       1      31      33    13     2  2018
     6 7       1      21      29    12     5  2018

> # cálculo de los parámetros descriptivos de la edad de las madres
> mean(datos$edad.ma, na.rm=TRUE)
[1] 26.13255
> digits(4)
Error in digits(4) : no se pudo encontrar la función "digits"
> mean(datos$edad.pa, na.rm=TRUE)
[1] 29.35676
> median(datos$edad.ma,na.rm=TRUE )
[1] 25
> sqrt(var(datos$edad.ma,na.rm=TRUE ))
[1] 6.315919
> range(datos$edad.ma,na.rm=TRUE )
[1] 11 50
> #
> # también se puede obtener el resumen
> summary(datos$edad.ma)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  11.00   21.00   25.00   26.13   31.00   50.00     860
> # y la gráfica de los datos
> hist (datos$edad.ma, col = 6)
```



Si interesa conocer la edad del 1% más joven y el 1% mayor de las mujeres, puede calcularse con la función *quartile*( ). También es interesante comparar con la edad de los hombres, para lo que se utilizan los diagramas de caja.

```
>  quantile (datos$edad.ma, 0.99, na.rm=TRUE)
99%
 41
> quantile (datos$edad.ma, 0.01, na.rm=TRUE)
1%
15
>
> boxplot(datos$edad.ma, datos$edad.pa, col ="coral")
```



Si bien este caso requiere presentar más instrucciones, ocurre lo mismo en el caso de utilizar el paquete *r mosaic*, pues éste llama a las mismas funciones estadísticas. Acá se agrega algún detalle de la sintaxis del lenguaje que el paquete *mosaic* oculta, pero se gana en el dominio del lenguaje, al introducir por ejemplo `datos$edad.ma` para referirnos a una variable particular de la base de datos. Esto abre paso a que se introduzcan los tipos de datos básicos del lenguaje.

Esta base de datos puede utilizarse para una situación didáctica en la que se introduzca el tema de la regresión lineal, primero graficando, con un diagrama de dispersión, la edad del padre en relación con la edad de la madre para, a partir de ahí, introducir la regresión lineal simple para observar, por ejemplo, si se puede encontrar alguna relación entre las edades de los progenitores.

NOTAS FINALES

En estos últimos años, en las diversas disciplinas de las ciencias sociales se habla de sistemas complejos, de trabajo cooperativo en redes, de *big data*, de modelos estructurales, de modelos multivariados, de modelos dinámicos o de juegos cooperativos y no cooperativos, pero nada de eso puede hacerse sin una formación básica en matemáticas, estadística y computación; esta formación también es fundamental para trabajar con los modelos econométricos ya incorporados a los programas de estudios en economía, pero en los que se utilizan paqueterías cerradas. Se requiere modificar los currículos, pero además es necesario capacitar a los instructores con esta nueva filosofía de cómo enseñar, modificando radicalmente métodos arraigados por la costumbre.

Efectuar este cambio no es fácil, y lo que aquí se presentó es una opción que, por supuesto, debe ser trabajada para adaptarla a cada una de las disciplinas, pero hay que empezar a hacerlo ya. El primer paso es desarrollar material básico para los temas comunes de los primeros cursos de estadística a nivel licenciatura.

Sin duda, el lenguaje R es la opción que más se está difundiendo y su característica de ser *software* libre y abierto está convirtiéndolo en el estándar, al menos, de la próxima década.