Graficando 1

David H. Duncan

January 20, 2016

¡Muy buenas! En esta leccion Ud importara un conjunto de datos desde un archivo de datos guardado como .CSV, y realizar varias graficas con ellos. ¿List@?

Ud va a cargar datos que ya hubiera visto en la visita a la galería, los que resumen el número de hijos y hijas que tiene las mujeres entre 14 y 50 años de las varias provincias de El Ecuador. Si quiere explorar el base de datos, desde esta página http://www.ecuadorencifras.gob.ec/sistema-integrado-de-consultas-redatam/ se escoge el enlace para el VII Censo de Poblacion y VI de Vivienda 2010.

Copie y pegue el siguiente en la consola:

```
pathtofile <<- function(course_, lesson_, file_){
  if(as.character(packageVersion("swirl")) > "2.2.21"){
    file.path(get_swirl_option("courses_dir"), course_, lesson_, file_)
  } else {
    file.path(find.package("swirl"), "Courses", course_, lesson_, file_)
  }
}

# Make path to xlsx available to user
path2csv1 <- file.path(path.package('swirl'), 'Courses',
                       'ConoceR',
                       'Graficando_1',
                       'hijosResumen.csv')
```

Con el último paso hubiera creado un objeto denominada 'path2csv1' que contiene la dirección completa para el tramo de datos para que funcione para cada persona sin saber donde haya almacenado su directorio del curso. Llame a la función read.csv() con el argumento path = path2csv1 (para dirección al archivo csv), y asígnelo al resultado el nombre 'resumen'.

```
resumen <- read.csv(path2csv1)
```

Antes de nada, hechar un vistazo a la estructura del conjunto en la ventanilla ENVIRONMENT, por tocar la flechita azul al lado del nombre del objeto 'resumen'. Allí se ve que la variable región tiene tipo 'Factor', mientras que la variable hijos es 'num' (para numérica). En este caso, los dos son correctos. R suele reconocer el tipo correctamente por defecto.

En la última lección usted aprendió algunas funciones más para explorar un conjunto nuevo de datos como str(), names(), head(), y summary(). Ya se ve la información de estructura en la ventanilla ENVIRONMENT, pero eche un vistazo al resumen de los variables ahora con summary(resumen). Acuérdese que en cualquier momento de cualquier lección, siempre y cuando vea el indicador (>) se puede teclear play() para salir y experimentar. Luego, después de jugar, para continuar la lección, se teclea nxt().

```
summary(resumen)
```

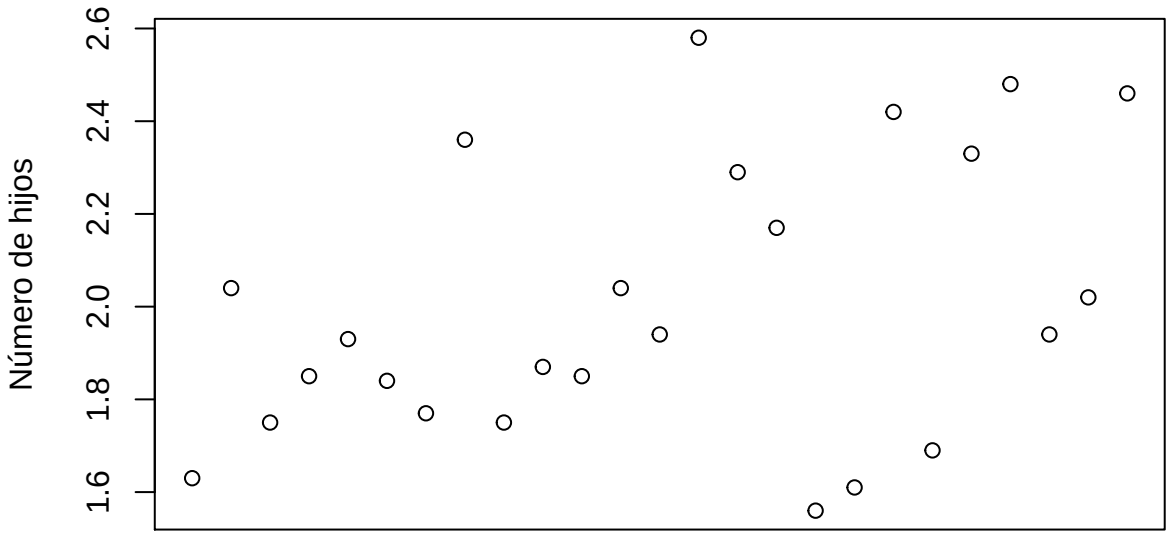
```
##         region       hijos
##  Azuay     : 1   Min.   :1.560
##  Bolivar   : 1   1st Qu.:1.770
```

```
##  Cañar     : 1   Median :1.940
##  Carchi    : 1   Mean   :2.007
##  Chimborazo: 1   3rd Qu.:2.290
##  Cotopaxi  : 1   Max.   :2.580
##  (Other)   :19
```

De esta vista de summary() parece que tenemos un conjunto con un valor por Provincia del país con el valor del promedio del número de hijos y hijas por mujer en cada una.

Muy bien, graficamos pues. Tiramos todas juntas para esta primera vista. Escriba plot(resumen$hijos, ylab='Número de hijos', xaxt='n')

```
plot(resumen$hijos, ylab='Número de hijos', xaxt='n')
```



Ya con esta gráfica vemos la dispersión de los promedios del número de hijos y hijas por mujer.

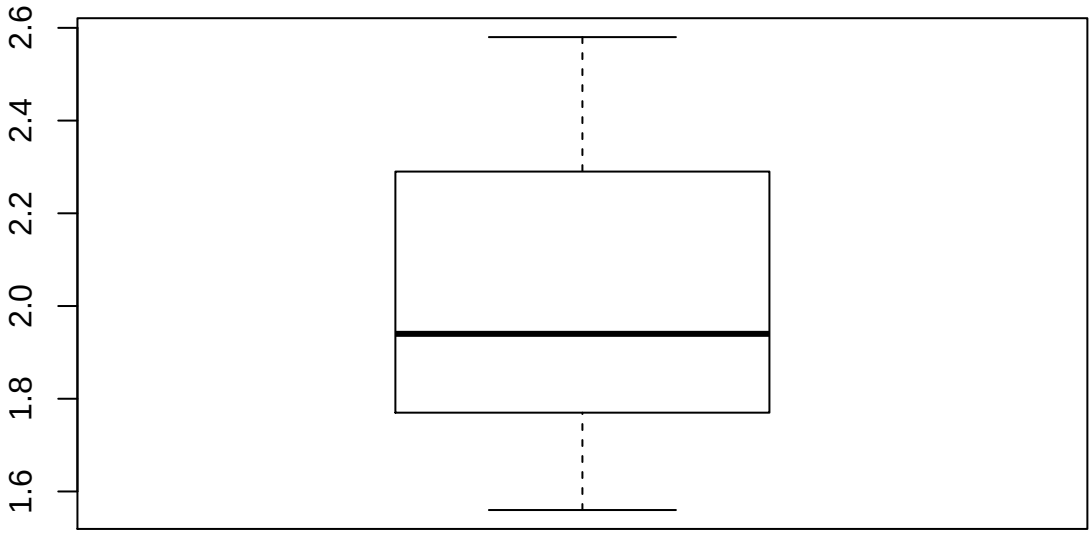
Antes de seguir, entonces, ¿según la gráfica aproximadamente qué es el rango de los valores?

- 1.6-2.6
- 1.6-2.5
- 1-3
- no se lo puede estimar

Una gráfica sencilla de puntos mediante plot() es un primer paso excelente para familiarizarse con los datos. En ella se puede ver si hay valores raros o no creibles.

Sin embargo, una gráfica mejor para resumir una sola variable cuantitativa como el número de hijos sea una diagrama de caja, también conocido como diagrama de caja y bigote, o boxplot en inglés. La función tomo el nombre del inglés, teclee boxplot(resumen$hijos).

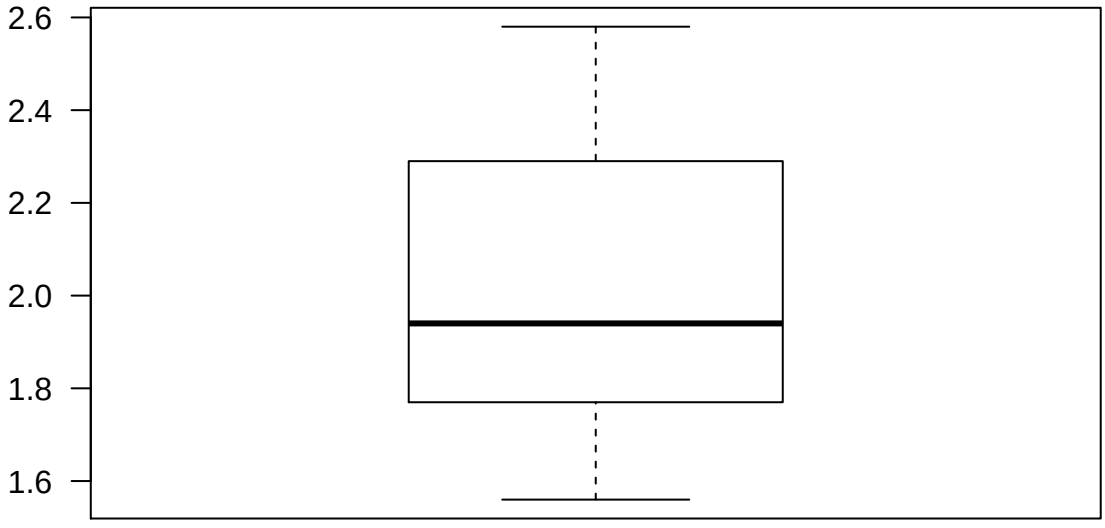
```
boxplot(resumen$hijos)
```



Un encanto del sistema base de gráficas en R es que es algo parecido al modo de trabajar con papel y bolígrafo. Una vez que estemos satisfechos con la base de la gráfica, podemos añadir elementos como títulos, leyendas, y muchos elementos más. Pero, justo igual a dibujar con un bolí en papel, no podemos borrar un elemento una vez trazado. Hay que volver a dibujarla, y en esto se ve la gran ventajo de un sistema repetible como lo que ofrece R. Ahora, vamos a girar las etiquetas del eje Y para que se lea mejor los valores.

Uno debe esperar pasar por varios borradores de una gráfica, y como le explicaba en el último dialogo, es por eso que guardamos código en un archivo. Así se puede volver a repetir al instante. Ahora, para girar los valores del eye Y, la manera más eficaz pueda ser recuperar el último llamada a boxplot(resumen$hijos) con la flecha arriba, y insertar ‘, las=1’ antes del último cierre de paréntesis. Siempre y cuando introducimos un argumento nuevo en una función, hay que separarlo de los de más con una coma. Introduzca el argumento con ‘las’ ahora...

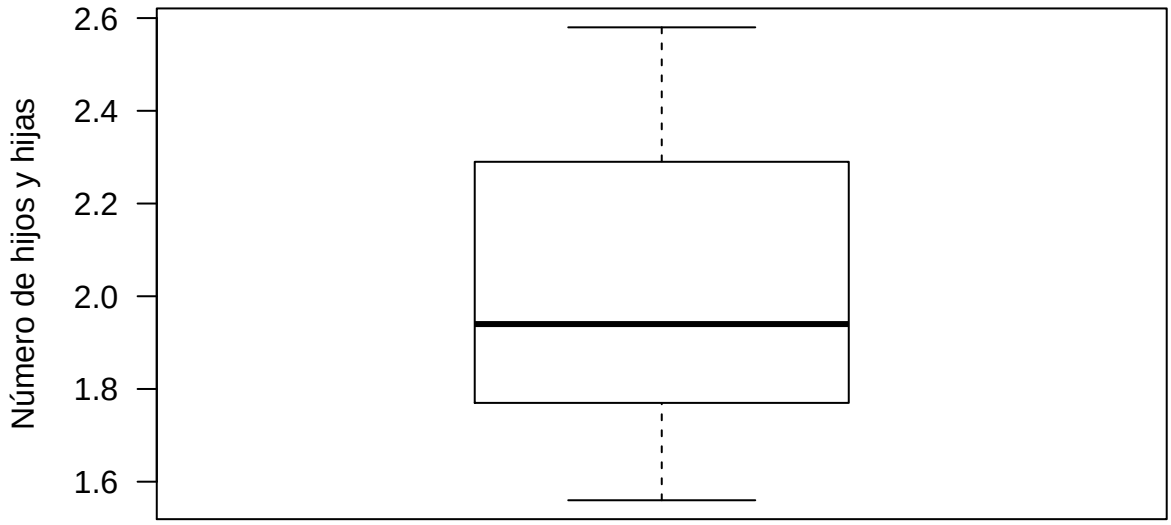
```
boxplot(resumen$hijos, las=1)
```



¿No se siente inmediatamente más comodo leyendo el contenido así? Por ejemplo, la pregunta sobre el rango de valores hubiera sido mucho más fácil con la gráfica en esta orientación, ¿verdad?

Ahora añada una etiqueta para el eje Y. Tiene usted dos opciones. (1) con title(ylab= 'Número de hijos y hijas'), o (2) por recuperar el último comando con boxplot y insertar, antes del último cierre de paréntesis ylab='Número de hijos y hijas'.

```
boxplot(resumen$hijos, las=1, ylab="Número de hijos y hijas")
```



¿Se acuerde en la visita a la galería cuando le expliqué las partes de una digrama de cajas? Basicamente, la diagrama divide los datos en cuartos. Esto nos permite hacer observaciones sencillas como 'en 75% de todas las provincias de El Ecuador las mujeres por promedio tienen más de N niños.' En este caso N sería el primer cuartil o el punto que indica un cuarto de todas las observaciones. Antes de seguir, entonces, ¿Cuál de los siguientes valores sería más cercano a N según esta gráfica?

- 1.75
- 1.95
- 2.30
- 1.55

Hacemos una modificación que a mi me encanta. Se puede superimponer dos gráficas una por encima de otra para comunicar aspectos distintos de los datos. A esta diagrama de cajas vamos a añadir un tipo denominado 'stripchart', y su función es stripchart(). Usted va a llamar a esta función con 6 argumentos. Acá está el código entero, deténgase un ratito para observarlos stripchart(resumen$hijos, method = 'jitter', jitter = 0.1, add = TRUE, vertical = TRUE, pch = 19).Lecomentoloselementosenseguida.'''stripchart(resumen$hijos, method = 'jitter', jitter = 0.1, add = TRUE, vertical = TRUE, pch=19) '''

En este último llamada empezamos con el objeto x, lo que era la variable resumen$hijos. El segundo argumento 'method' define el método de acomodar los puntos 'jitter'. Jitter no tiene ninguna tradución fácil, pero se puede pensar en agitación causada por inestablidad nerviosa o algo así. El siguiente argumento jitter adjuste el grado máximo de desplacimiento horizontal de los puntos. Luego, el argumento 'add' (añadir) es imprescindible. Es ello que especifica que los puntos sean por encima de la gráfica actual en lugar de en una diagrama nueva. El argumento 'vertical' por defecto es 'FALSE', es decir horizontal, así que tuvimos que cambiarlo a 'TRUE'. Por último, el argumento 'pch' es lo que controla el tipo de punto.

No me soprendería si usted esta pensando que todo esto es demasiado detallado. Pero, fíjese, con estos dos elementos gráficos elementos juntos se aprovecha del mensaje de resumen que ofrece la caja y bigote, pero a

la vez se ve el número de valores y su distribución cruda. En efecto combine lo bueno de las dos primeras vistas que hemos visto. Podemos dejar esta gráfica en su estado actual, una vez que acabemos de añadir las etiquetas necesarias.

Necesita un título, ¿verdad?. Ponga title(main= 'Promedio del número de hijos y hijas de mujeres ecuatorianas 2010') por ejemplo. Fuera del contexto de una lección de swirl, hay opciones para romper el título rompe en dos o más lineas.

```
title(main= 'Promedio del número de hijos y hijas de mujeres ecuatorianas 2010')
```

Si no cabe en la ventanilla, haga click en ZOOM, justa arriba de la gráfica para verla más bonita. Podríamos declarar nuestra fuente de datos como subtítulo, otra vez con la función title. Trate esto title(sub = 'Fuente > VII Censo de Poblacion y VI de Vivienda 2010').

```
title(sub = 'Fuente > VII Censo de Poblacion y VI de Vivienda 2010')
```

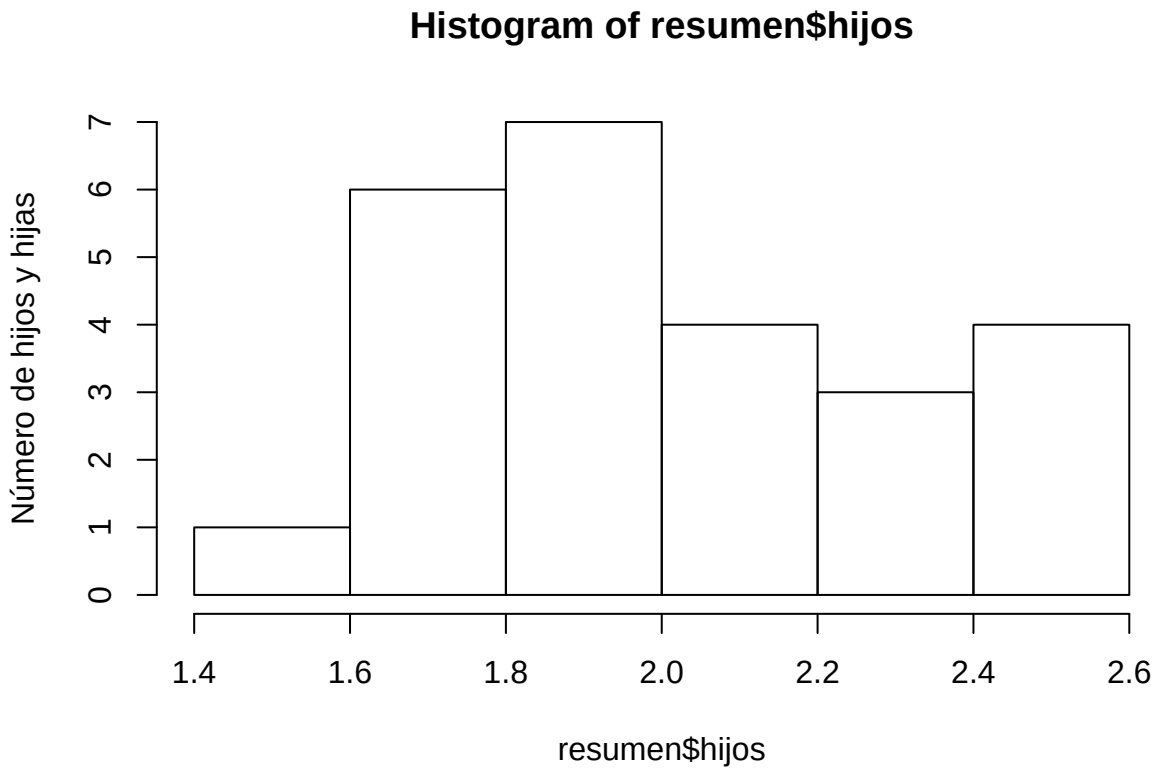
Ahora yo diría que usted tiene una gráfica linda en su simplicidad que comunica mucho de este conjunto de datos. También provoca muchas preguntas más, como ¿cuáles Provincias sobresaltan del conjunto con máximo o mínimo promedio de número de niños por mujer? Preguntas así podemos explorar en la siguiente lección através del conjunto de datos desglosados.

Otra pregunta de interpretación de esta gráfica antes de cambiar el tipo. La mediana es un estadístico que indica el valor en la mitad de todos en el conjunto impar, o en el caso de un número par de valores, divide los dos valores centrales. Entonces, según la gráfica ¿qué cree usted es la mediana del número promedio de hijos y hijas que tuvieron las mujeres ecuatorianas entre 14 - 50 años de edad en 2010?

- 1.95
- 2.10
- 2.30

Antes de salir, debemos realizar una histrograma ¿verdad?, tal vez la gráfica más famosa para resumenes de una variable cuantitativa. Bueno, así se lo hace hist(resumen$hijos, ylab='Número de hijos y hijas').

```
hist(resumen$hijos, ylab='Número de hijos y hijas')
```



En la lección actual, solo le queda a usted un momento de reflexión en lo que haya aprendido. Según su experiencia, ¿cree usted que se podría haber añadido el título, subtítulo, y etiqueta del eje, todos a la vez?

- Si, la función title() tiene argumentos para cada tipo de título o etiqueta
- No, solo acepta un argumento a la vez

De hecho, se pudiera haber añadido todos los elementos de título y etiqueta en la llamad original a boxplot(), sin llamar a la función title(). Le animo colocar todos argumentos en una sola llamada en su álbum de recortes. Es buena practica romper lineas con INTRO para mantener código fácil de leer y editar. La función stripchart() para superimponer los puntos por encima tendrá que ser una llamada aparte, despues de lo de boxplot().

Con eso, usted ha terminado la lección. No se olvide de trasladar código que le parece útil a su álbum de recortes. En esta lección hemos visto una manera muy buena de resumir una variable cuantitativa de forma atractiva y informativa. En la siguiente usted jugará con opciones para revelar la identidad de las Provincias, y hacer contrastes con factores de agrupación que puedan ayudar en explicar los patrones en el número de hijos y hijas entre las mujeres ecuatorianas.

Hasta la proxima lección, tenga usted un buen día.