Práctica
Programación en R

Esta práctica fue realizada con el objetivo de que los estudiantes tengan un primer acercamiento a la manipulación de datos y a la generación de gráficos adecuados con R, que les será útil para la realización del Trabajo Práctico de Estadística Descriptiva. No es una introducción exhaustiva de la herramienta ni pretende serlo. A lo largo de la práctica se irá guiando al estudiante a la búsqueda y uso de funciones particulares que ayudarán a familiarizarse con el entorno.

Introducción

1. Una de las primeras tareas a realizar será leer una base de datos proveniente de un archivo. En general, los datos son muchos y no es práctico ingresarlos manualmente. En este caso, se trabajará con el archivo *anorexia.data*[1]. Dicha base de datos corresponde a una recolección de datos que hizo la "Asociación de Lucha contra la Bulimia y la Anorexia"durante el mes de octubre del año 2012. La información es acerca de 59 personas con síntomas de anorexia que se habían acercado a la institución en busca de ayuda durante los primeros nueve meses de ese año. Las variables utilizadas son "Sexo", "Cantidad de visitas", "Edad "y "Principal signo visible"(0 - dieta severa, 1 - hiperactividad, 2 - uso de laxantes, 3 - uso de ropa holgada).

a) Abra la base de datos con algún editor de texto.[2] ¿De qué tamaño es la muestra? ¿Cuántas variables se encuentran? ¿Cómo se llaman? ¿De qué tipo son? ¿Cómo está separada una columna de otra?

b) Lea la base de datos desde R.

> **Ayuda:** puede ser útil buscar la documentación de la función `read.table()`. ¿Qué argumentos toma? ¿De qué clase es lo que retorna?

2. Una vez cargada la base de datos, se deben proveer formas de trabajar con las distintas variables. Una manera muy sencilla de acceder a todos los valores de una variable es a través del operador `$`:.

```
> dataframe_name$variable_name
```

Sin embargo, la notación $ puede no resultar tan conveniente. La función `attach()` toma una base de datos (una lista o un *data frame*) como argumento y hace temporalmente visible sus componentes como variables. La función `detach()` revierte dicho proceso.

```
> attach(dataframe_name)
> variable_name
```

a) Teniendo en cuenta lo explicado anteriormente, acceda a los contenidos de cada una de las variables por separado. Por ejemplo, para la variable "Sexo"debería lograr un output similar al siguiente:

```
 [1] F F F F F F F F F F F F F F F F F F F F F F F F F
[26] F F F F F F F F F F F F M M M M M M M M M M M M M
```

[1]Disponible para descargar en el sitio de comunidades.
[2]Esto es recomendable hacerlo únicamente si el archivo tiene un tamaño razonable.

```
[51] M M M M M M M M M
Levels: F M
```

b) Utilizando la función `class()`, averigue de qué clase es cada variable.

c) Cuente la cantidad de mujeres y de hombres que hay en la base de datos.

Ayuda: puede ser útil buscar la documentación de la función `summary()`.

d) Calcule la edad máxima, la mínima y el promedio de todas las edades.

Ayuda: puede ser útil buscar la documentación de las funciones `summary()`, `sum()`, `length()`, `mean()`.

Tablas de frecuencia

3. En el caso de las variables cuantitativas, antes de presentar distintos gráficos, puede ser útil resumir los datos en tablas de frecuencias como las que se ejemplifican a continuación en la Figura 1 y en la Figura 2.

NÚMERO DE VISITAS DURANTE LOS PRIMEROS 9 MESES DE 2012 ARGENTINA

N° de Visitas	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
1	17	0.2881	17	0.2881
2	24	0.4068	41	0.6949
3	15	0.2542	56	0.9492
4	2	0.0339	58	0.9831
5	1	0.0169	59	1.0000
Total	**59**	**1**		

Fuente: Asociación de Lucha Contra la Bulimia y la Anorexia

Figura 1: Número de visitas

EDAD DE LOS PACIENTES CON ANOREXIA ARGENTINA, OCTUBRE 2012

Intervalo (edad)	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
[11;14)	6	0,1017	6	0,1017
[14;17)	20	0,3390	26	0,4407
[17;20)	18	0,3051	44	0,7458
[20;23)	7	0,1186	51	0,8644
[23;26)	2	0,0339	53	0,8983
[26,29)	3	0,0508	56	0,9492
[29;32)	2	0,0339	58	0,9831
[32;35)	1	0,0169	59	1,0000
Total	**59**	**1**		

Fuente: Asociación de Lucha Contra la Bulimia y la Anorexia

Figura 2: Edad de los pacientes

a) Realice la tabla de frecuencias para la variable "Número de visitas" como se muestra en la

Figura 1 .

Ayuda: puede ser útil buscar la documentación de las funciones `table()`, `cumsum()`, `cbind()`, `rbind()`, `round()`, `names()`.

b) Realice la tabla de frecuencias para la variable "Edad "como se muestra en la Figura 2. ¿Por qué en la fila de esta tabla se encuentran rangos etarios en lugar de valores únicos?

Ayuda: puede ser útil buscar la documentación de las funciones`cut()`, `table()`, `cumsum()`, `cbind()`, `rbind()`, `round()`, `names()`.

4. Muchas veces es interesante cruzar variables de modo de obtener análisis bivariados. Por ejemplo, en la Figura 3 se presenta una tabla del principal signo visible según el sexo. Observar que en este caso se trata de dos variables cualitativas.

PRINCIPAL SIGNO VISIBLE EN PACIENTES CON ANOREXIA ARGENTINA, OCTUBRE 2012

Principal signo visible	Mujeres	Hombres	Total
Dieta severa	6	4	**10**
Hiperactividad	9	10	**19**
Uso de laxantes	14	3	**17**
Uso de ropa holgada	8	5	**13**
Total	**37**	**22**	**59**

Fuente: Asociación de Lucha Contra la Bulimia y la Anorexia

Figura 3: Principal signo visible según el sexo

Encuentre la forma de cruzar las variables mencionadas de modo de obtener el contenido de la tabla anterior.

Ayuda: puede ser útil buscar la documentación de las funciones `table()`, `apply()`, `cbind()`, `rbind()`.

Gráficos

Las comodidades para la realización de gráficos son un componente importante y extremadamente versatil del entorno R. Es posible usarlas para producir una amplia variedad de gráficos estadísticos predeterminados así como para construir nuevas clases.
A modo general, puede decirse que los comandos de ploteo en R se dividen principalmente en tres grupos:

- Alto nivel: crean un nuevo gráfico, posiblemente con títulos, etiquetas, ejes, etc.
- Bajo nivel: agregan información a un gráfico ya existente. Por ejemplo: líneas, etiquetas o puntos.
- Interactivos: permiten agregar o extraer información interactivamente desde un gráfico ya existente. No serán vistos en esta práctica.

Además, R mantiene una lista de parámetros gráficos, los cuales pueden ser manipulados para personalizar los gráficos. A continuación se trabajará con estos conceptos.

5. Uno de los tipos de gráficos más utilizados para resumir variables cualitativas es el *diagrama de sectores circulares* o *gráfico de torta*. En la Figura 4 se presenta un ejemplo para la variable "Principal signo visible".

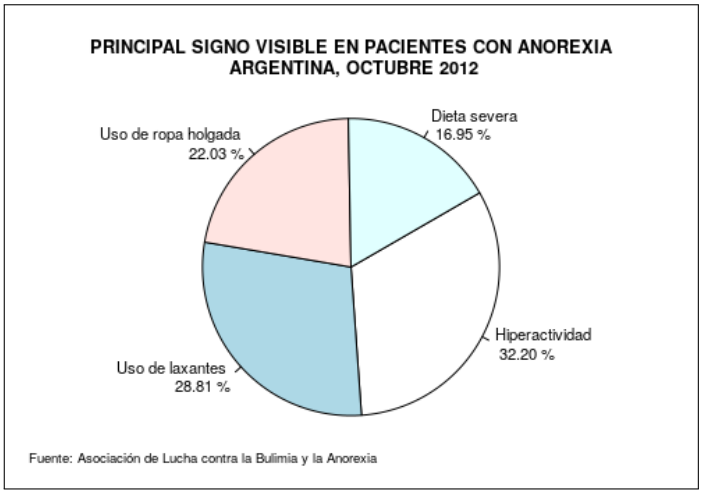


Figura 4: Principal signo visible

a) Lea la documentación de la función `pie()`. ¿Cuál es el significado de los argumentos `x, labels, clockwise, init.angle` y `main`?

b) Realice una primera versión del gráfico de sectores para la variable mencionada, utilizando los argumentos vistos anteriormente. No se preocupe si el gráfico aún no se ve como lo espera.

c) En este tipo de gráficos es importante incluir los porcentajes de cada categoría, porque no siempre es claro visualmente si un sector es mayor, menor o igual a otro. Realice una segunda versión del gráfico que tenga en cuenta esto.

Ayuda: puede ser útil buscar la documentación de la función `paste()`, y generar un nuevo vector de etiquetas que, además del nombre de cada categoría, contenga el porcentaje que corresponde a cada una.

d) Como se mencionó anteriormente, R provee funciones de graficado llamadas *de bajo nivel* que permiten agregar más información a un gráfico ya existente, como líneas, puntos o etiquetas. Una de ellas es `mtext`, utilizada para escribir texto en los márgenes de un gráfico. Busque la documentación de dicha función. ¿Cuaĺ es el significado de los argumentos `text, side, line, at, adj, cex` y `font`?

e) Realice una tercera versión del gráfico de sectores, utilizando la función estudiada anteriormente.

f) Por último, R provee una función `par()` utilizada para setear permanentemente o consultar todos los parámetros gráficos. Puede obtener una lista con sus nombres y valores simplemente haciendo:

```
> par()
```

Si bien estos parámetros son muchos y no es necesario conocerlos todos, puede ser útil buscar entre ellos cuando se quiera modificar un aspecto particular de los gráficos. Por ejemplo, es posible que

le sea conveniente cambiar los valores del parámetro `mar`, para definir márgenes adecuados en sus gráficos.

Aclaración: algunos parámetros específicos de `par()` pueden pasarse directamente como parámetros en la llamada a la función de alto nivel que grafica. Por ejemplo, es válido hacer lo siguiente:

```
> pie(table(Sexo), cex = 0.7,...)
```

Busque la documentación de la función `par()`. ¿Cuántos parámetros gráficos se pueden setear o consultar a través de ella? ¿Para qué sirve el parámetro `cex`?

g) Realice una última versión del gráfico de sectores que se vea similar al que se muestra en la Figura 4.

6. Otro de los tipos de gráficos más utilizados para variables cualitativas es el *gráfico de barras*. En la Figura 5 se presenta un ejemplo, nuevamente para la variable "Principal signo visible".

Figura 5: Principal signo visible

a) Lea la documentación de la función `barplot()`. ¿Cuál es el significado de los argumentos `height, horiz, xlab, ylab, xlim, cex.axis` y `cex.names`?

b) Realice una primera versión del gráfico de barras para la variable mencionada, utilizando todo lo visto anteriormente. No se preocupe si el gráfico aún no se ve como lo espera.

c) Para lograr una mejor percepción visual, las categorías en los gráficos de barras suelen ordenarse de forma creciente o decreciente. Realice una segunda versión del gráfico de barras teniendo en cuenta esto.

Ayuda: puede ser útil buscar la documentación de la función `order()`.

d) Setee los parámetros gráficos necesarios y realice una última versión del gráfico que se vea similar al que se presenta en la Figura 5.

7. Un gráfico de barras compuesto permite comparar resultados para diferentes grupos. Por ejemplo, en la Figura 6 se puede visualizar el principal signo visible según el sexo del paciente. Realice un gráfico similar, teniendo en cuenta cómo se utiliza la función `barplot()` cuando el argumento `height` es una matriz.

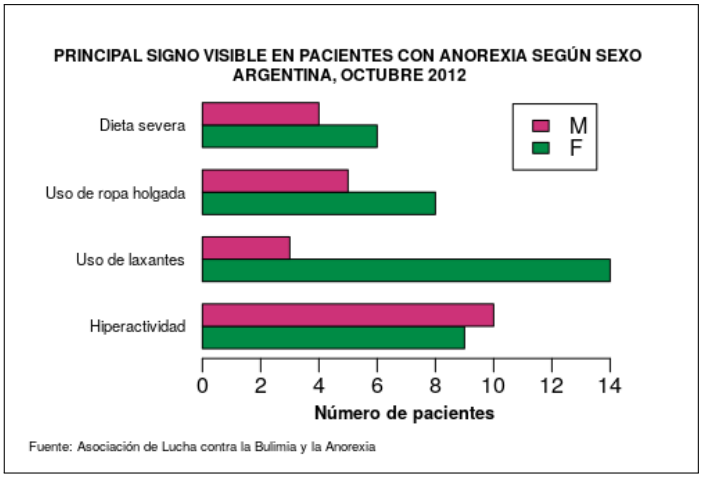


Figura 6: Principal signo visible

8. Para representar gráficamente datos de variables discretas, especialmente cuando toman pocos valores diferentes, se puede utilizar un *gráfico de bastones*. Por ejemplo, en la Figura 7 se muestra un gráfico de bastones correspondiente al número de visitas a la asociación por paciente.

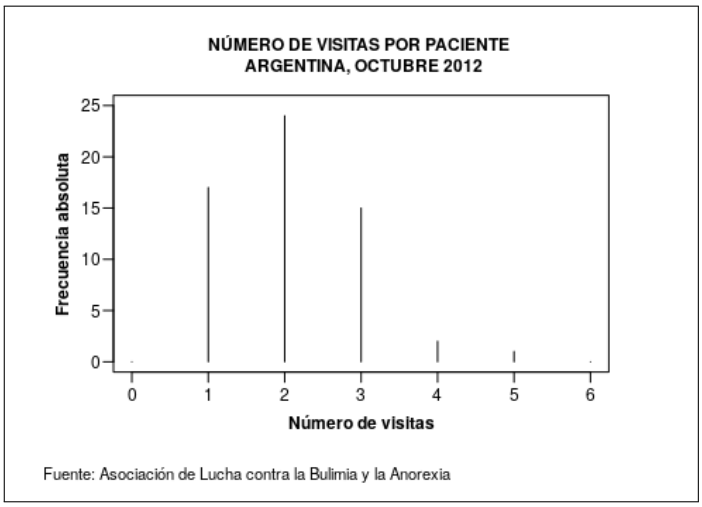


Figura 7: Número de visitas por paciente

a) Busque la documentación de la función `plot()`. Observe que la descripción dice "Función genérica para graficar objetos de R". La idea de *función genérica* implica, en este caso, que el tipo de gráfico generado por `plot()` dependerá del tipo o clase de su primer argumento. A continuación, se presentan algunos ejemplos de posibles llamadas:

- `plot(x, y)`: si `x` e `y` son vectores, el resultado será un diagrama de dispersión [3]
- `plot(f)`: si `f` es un factor, entonces generará un gráfico de barras.

[3] Un diagrama de dispersión es un gráfico que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos.

- `plot(f,y)`: si `f` es un factor e `y`, un vector numérico, la llamada producirá un *boxplot comparativo*[4].

A su vez, `plot()` cuenta con un parámetro `type`. Estudie los distintos gráficos que puede realizar a través de dicho argumento.

b) Realice un gráfico de bastones similar al de la Figura 7 utlizando la función `plot()` y lo visto anteriormente.

c) Los gráficos de bastones suelen ir acompañados por *gráficos escalonados* que se utilizan para mostrar la distribución acumulada de una variable discreta. En la Figura 8 se muestra el ejemplo correspondiente al caso trabajado. Confeccione dicho gráfico a través de la función `plot()`.

Ayuda: puede ser útil buscar la documentación de la función `abline()` para agregar líneas punteadas de referencia.

Figura 8: Número de visitas por paciente

9. En el caso de las variables continuas o discretas que asumen muchos valores distintos, suele utilizarse un *histograma* para representarlas gráficamente. En la Figura 9 se puede ver el histograma correspondiente a la edad de los pacientes.
Observación: los intervalos etarios utilizados en el gráfico coinciden con los de la correspondiente tabla de frecuencias.

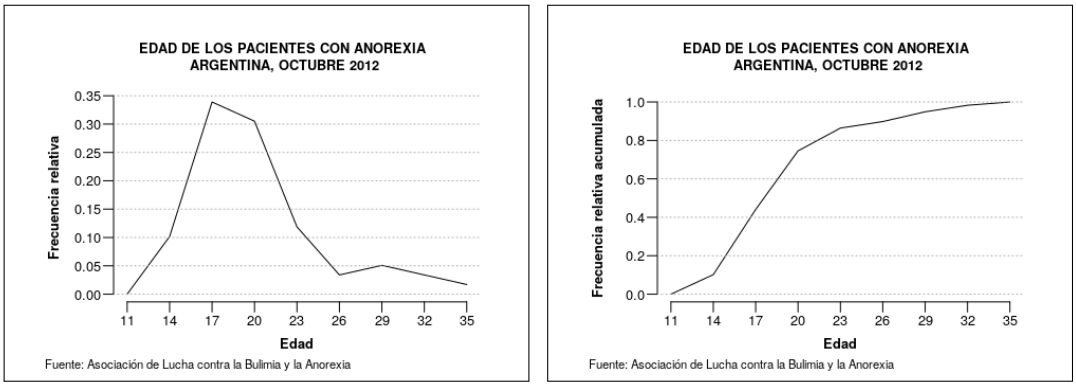
a) Busque la documentación de la función `hist()`. ¿Para qué sirven los argumentos `breaks` y `right`? Utilice dicha función para generar el histograma presentado anteriormente.

Ayuda: puede ser útil buscar la documentación de la función `axis()` para dibujar los ejes independientemente.

b) Los histogramas suelen ir acompañados de dos gráficos auxiliares: el *polígono de frecuencias* y el *polígono acumulativo*. En la Figura 10 se muestran dichos gráficos.
Teniendo en cuenta todas las funciones trabajadas anteriormente, utilice `plot()` para realizar cada uno de los gráficos. Piense qué valor tomará el argumento `type` en estos casos.

[4] Vea el ejercicio 10.

Figura 9: Edad de los pacientes



(a) Polígono de frecuencias (b) Polígono acumulativo

Figura 10: Edad de los pacientes

10. El *boxplot* es un gráfico basado en los cinco números resumen de un conjunto de datos: el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo. En la Figura 11 se muestra un ejemplo de boxplot para la variable "Edad ".

a) Busque la documentación de la función `summary()` si no lo hizo antes. Pruébela para cada columna de la base de datos. Compare los resultados para las variables "Sexo "y "Número de visitas ". ¿Por qué cree que calcula diferentes medidas en ambos casos? ¿Qué observa al calcular las medidas resúmenes de la variable "Principal signo visible"?

El comportamiento anterior se debe a que `summary()` es una *función genérica*. Como se vio en un ejercicio previo, en R, este tipo de funciones determina la clase de su argumento y usa dicha información para seleccionar el método apropiado. Así, por ejemplo, `summary()` calcula distintas medidas resúmenes según si el argumento es una variable cualitativa (en general, asociada a un factor) o cuantitativa. Sin embargo, es importante notar que siempre se deben analizar los resultados que devuelve.

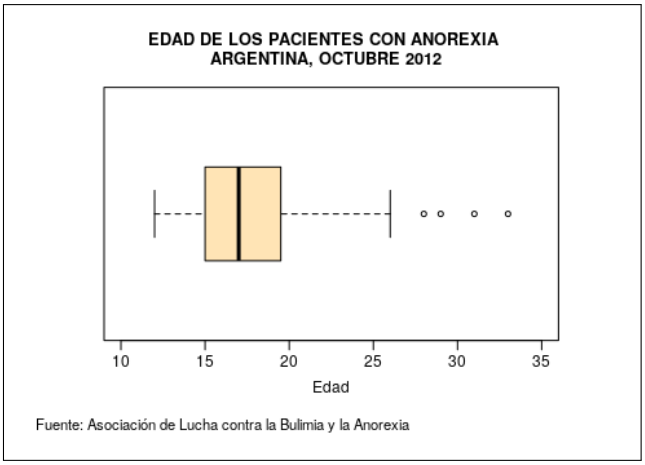


Figura 11: Edad de los pacientes

b) Busque la documentación de la función `boxplot()` y realice el gráfico de la Figura 11.

c) Los boxplots comparativos son muy útiles para la comparación gráfica de distribuciones en grupos distintos. En la Figura 12 se muestra un ejemplo para la edad según el sexo de los pacientes.

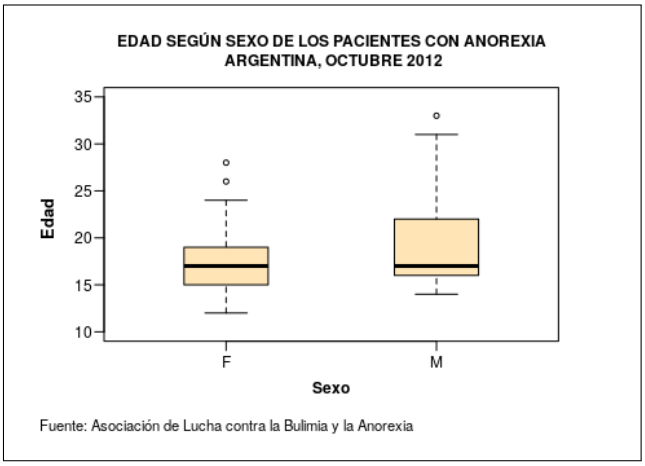


Figura 12: Edad de los pacientes según el sexo

El gráfico anterior puede ser generado nuevamente con la función `boxplot()`, utilizando el parámetro `formula`. Hágalo.

Ayuda: el parámetro `formula` toma algo de la forma $y \sim grp$, donde `y` es un vector numérico de datos que serán distribuidos en grupos, acordes a la variable `grp` (usualmente un factor).