Utilización de software para el análisis de datos

Probabilidades y Estadística

En este apunte encontrarán las indicaciones básicas sobre cómo realizar mediante el software *R* los ejercicios correspondientes al práctico de estadística descriptiva. Este apunte no busca ser exhaustivo en el desarrollo de las bondades de los comandos que se explican. En el caso de querer profundizar en el uso de *R* y Rcommander, existe mucha bibliografía disponible online para hacerlo.

Comenzando...

En primer lugar deben realizar la instalación del software *R*. Para esto deben acceder al sitio `cran.r-project.org/` y descargar el software según su sistema operativo. Una vez que terminaron la instalación de *R* aconsejamos instalar RStudio accediendo a `www.rstudio.com/products/rstudio/download2/`. Esto no es necesario, pero nos ofrecerá un entorno amigable para trabajar con *R*.

De todas las librerías que ofrece *R*, vamos a utilizar la librería RCommander cuando queramos realizar análisis estadísticos de datos. Si instalaron RStudio, en la ventana de la derecha abajo, hacen click donde dice Packages. Y luego donde dice Install. En el buscador que les abre ponen Rcmdr. Dejan cliqueada la casilla que les pregunta si quieren instalar las dependencias y lo ponen a instalar.

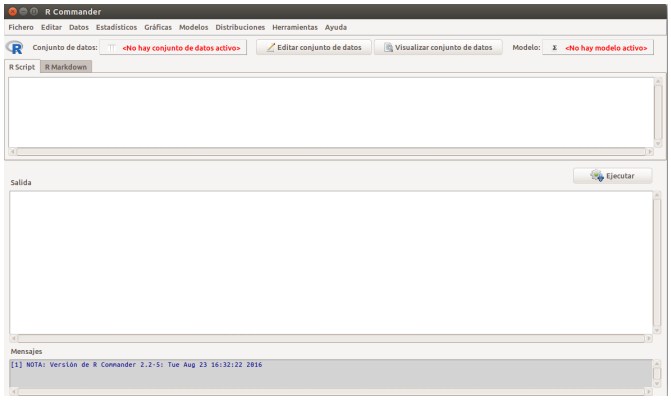
En caso de que no hayan instalado RStudio, abren el *R* y ejecutan install.packages(Rcmdr).

Para comenzar a usar la librería RCommander, debemos abrir *R* y escribir la siguiente sentencia

> library("Rcmdr")

> Commander()

Esto nos abrirá un entorno gráfico como el de la figura.



Cargar datos

Para comenzar a trabajar debemos cargar un conjunto de datos. Para esto, hacemos click en el menú datos y hacemos lo siguiente

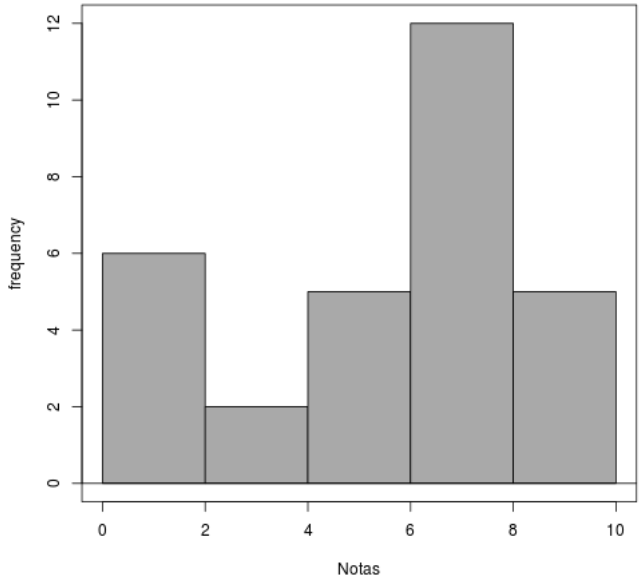
Datos → Importar datos → Elegimos el formato en los que tenemos nuestros datos.

Esto nos abrirá un menú donde pondremos el nombre de nuestro conjunto de datos. Luego debemos indicar el directorio donde se encuentra el archivo que contiene nuestros datos. Una vez que hicimos esto, ya tenemos nuestros datos cargados. Para ver si los hemos cargado bien, podemos hacer click en Visualizar conjunto de datos

Si cargamos el conjunto de datos del ejemplo 1 del práctico de Estadística descriptiva y queremos realizar un histograma o un gráfico de barras haríamos lo siguiente:

Para realizar el histograma

Vamos al menú Gráficas y hacemos click en Histograma. Esto nos abrirá un menú donde tendremos que elegir nuestra variable a graficar y las propiedades del histograma. Damos aceptar y obtenemos un histograma como el de la figura



Para realizar el gráfico de barras

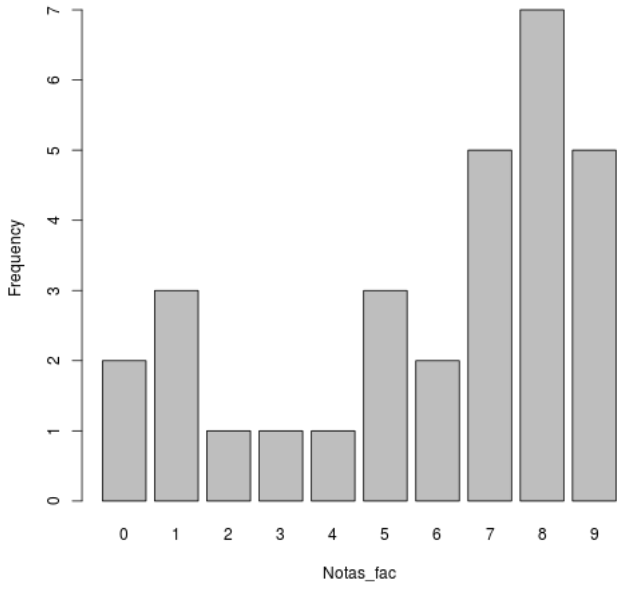
Cuando realizamos gráficos de barras, trabajamos con variable de tipo categóricas, aunque estemos hablando de números enteros. Por lo tanto, debemos pasar las variables numéricas a categóricas (factor). Para esto, hacemos lo siguiente:

Datos → Modificar variables... → Convertir variable numérica en factor.

Esto nos abre una ventana, donde debemos seleccionar la variable que queremos convertir y el nombre que le asignaremos, en este caso puede ser Notas_fac. Debemos seleccionar la casilla Utilizar números. Con esto ya tendremos en nuestro Dataset una variable de tipo categórica. En el caso de querer hacerlo desde la consola de R, debemos hacer [1]

> Dataset$Notas_fac <− as.factor(Dataset$Notas)

Entonces ahora hacemos click en Gráficas y luego en Gráficas de barra y tendremos un gráfico como el que sigue



[1] Con el símbolo <− asignamos la expresión de la derecha a la variable de la izquierda y con el símbolo $ accedemos a la columna de nuestro archivo de datos

Para guardar los gráficos debemos ir a Gráficas → Guardar gráfico en archivo En caso de que nuestra versión de Rcommander no cuente con esta opción, vamos a hacerlo con la función Barplot.

```
> Barplot(Dataset$Notas_fac)
```

Manipulación de histogramas

Como pudimos ver antes, cuando creamos el histograma tenemos la opción de indicarle en cuántas clases queremos agrupar nuestros datos. Muchas veces queremos indicarle no el número de clases, sino los límites inferiores y superiores de cada una. Para eso tenemos que manipular el comando Hist con el cual realizamos los histogramas.

Este comando tiene un parámetro que se llama breaks donde uno puede cargar los límites de los intervalos. Si tenemos cargado en Dataset los datos del ejemplo 3 del práctico, para realizar un histograma con 5 clases que represente la tabla de frecuencia por intervalos deberíamos hacer [2]

```
> Hist(Dataset$Tiempo, breaks=seq(9,18,(18-9)/5), col = 'green')
```

Tablas de frecuencia. Variable cualitativa

RCommander sólo permite, mediante menú, la realización de tablas de frecuencias para variables cualitativas. Para esto, hacemos click en Estadísticos, Resúmenes, Distribución de frecuencias. En ese menú elegimos la variable cualitativa. Por ejemplo, utilizando la variable categórica Notas_fac que creamos antes conseguiremos lo siguiente

```
counts:
factor
0 1 2 3 4 5 6 7 8 9
2 3 1 1 1 3 2 5 7 5

percentages:
factor
    0     1     2     3     4     5     6     7     8     9
 6.67 10.00  3.33  3.33  3.33 10.00  6.67 16.67 23.33 16.67
```

Muchas veces, como en el caso de las Notas, nuestra variable no es cualitativa. Si bien podemos transformarla como hicimos en el caso de los gráficos de barra, no siempre es necesario hacer esto.

Tablas de frecuencia. Variable cuantitativa

Si se quiere hacer una tabla de frecuencias para una variable cuantitativa discreta lo tenemos que hacer mediante código de R, usando el comando table. Para el caso de las Notas, tendríamos que

```
# Para distribución de frecuencias absolutas  table(Dataset$Notas)

# Para frecuencias absolutas acumuladas  cumsum(table(Dataset$Notas))

# Para distribución de frecuencias relativas  table(Dataset$Notas)/sum(table(Dataset$Notas))

# Para distribución de porcentajes  100* table(Dataset$Notas)/sum(table(Dataset$Notas))
```

Al ejecutar los comandos anteriores, los valores se muestran en la pantalla pero no se guardan y la presentación de los resultados no es muy amigable. Podemos mejorar eso mediante los siguientes comandos

```
# Creamos un data frame [3] donde guardar la tabla de frecuencias absolutas. Lo llamamos tabla  tabla <-as.data.frame(table(Datas
```

[2] El comando seq(inicio, fin, amplitud) genera una secuencia de número de *inicio* a *fin* con un paso de *amplitud*

[3] Un data frame es uno de los tipos de datos de R

Agregamos una columna FR en la tabla para las frecuencias relativas tabla$FR <- tabla$Freq/sum(tabla$Freq)

Agregamos una columna FAc en la tabla para las frecuencias acumuladas tabla$FAc <- cumsum(tabla$Freq)

Por último agregamos la frecuencia relativa porcentual FRP tabla$FRP <- 100*tabla$FR

```
> tabla
   Var1 Freq FAc         FR       FRP
1     0    2   2 0.06666667  6.666667
2     1    3   5 0.10000000 10.000000
3     2    1   6 0.03333333  3.333333
4     3    1   7 0.03333333  3.333333
5     4    1   8 0.03333333  3.333333
6     5    3  11 0.10000000 10.000000
7     6    2  13 0.06666667  6.666667
8     7    5  18 0.16666667 16.666667
9     8    7  25 0.23333333 23.333333
10    9    5  30 0.16666667 16.666667
```

Con estos códigos se puede ampliar la distribución de frecuencias que el software arroja para el caso de variable cualitativas.

En casa, hacer con estos comandos la tabla de frecuencia del Ejemplo 2 del práctico

Medidas de resumen

Para calcular las medidas de resumen de un conjunto de datos cuantitativos, hacemos click en Estadísticos → Resúmenes → Resúmenes numéricos.

En la pestaña que se abre, debemos elegir la variable del conjunto de datos, y los estadísticos que queremos calcular.

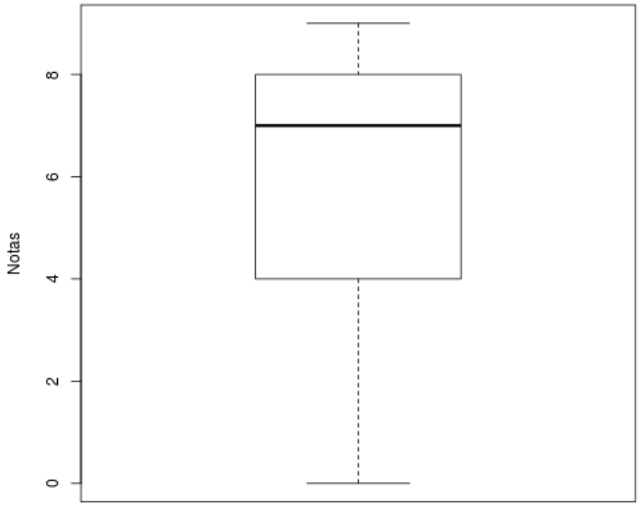
```
    mean       sd  IQR 0%  25% 50% 75% 100%  n
5.833333 2.972092 3.75  0 4.25   7   8    9 30
```

Esto también se puede hacer fácilmente con un comando que se llama summary.

> summary(Dataset$Notas)

Boxplot

Para realizar un Boxplot de las variables cuantitativas del conjunto de datos, debemos ir a Gráficas → Diagrama de cajas. En la pestaña que se abre, debemos elegir la variable del conjunto de datos y tenemos algunas opciones referidas a los datos atípicos. Para el caso de la variable Notas el boxplot sería


Notas
0
2
4
6
8