

# Introducción a la estadística descriptiva – Uso de Rcommander

*Material correspondiente a la Asignatura: Probabilidad y Estadística*

*Docentes: Dra. Valeria Leoni -Lic. Luciana Chiapella*

## **Introducción:**

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. R forma parte de un proyecto colaborativo y abierto. Sus usuarios pueden publicar paquetes que extienden su configuración básica. Esto lo ha vuelto uno de los softwares más utilizados para el análisis de datos, ya que es de libre acceso y cuenta con infinidad de opciones disponibles para la ejecución de las más diversas tareas.

Si bien el trabajo habitual con R requiere del conocimiento del lenguaje de programación, Rcommander es una interfaz de R que permite realizar tareas básicas de análisis de datos sin necesidad de escribir algoritmos en un lenguaje específico. Rcommander es útil para ejecutar estas tareas a través de su sistema de menús y submenús, disponibles para ser seleccionados realizando clicks.

Para poder utilizar Rcommander, es necesario instalar dos cosas: a) R como software de análisis y cálculo estadístico y b) Rcommander como interfaz de usuario.

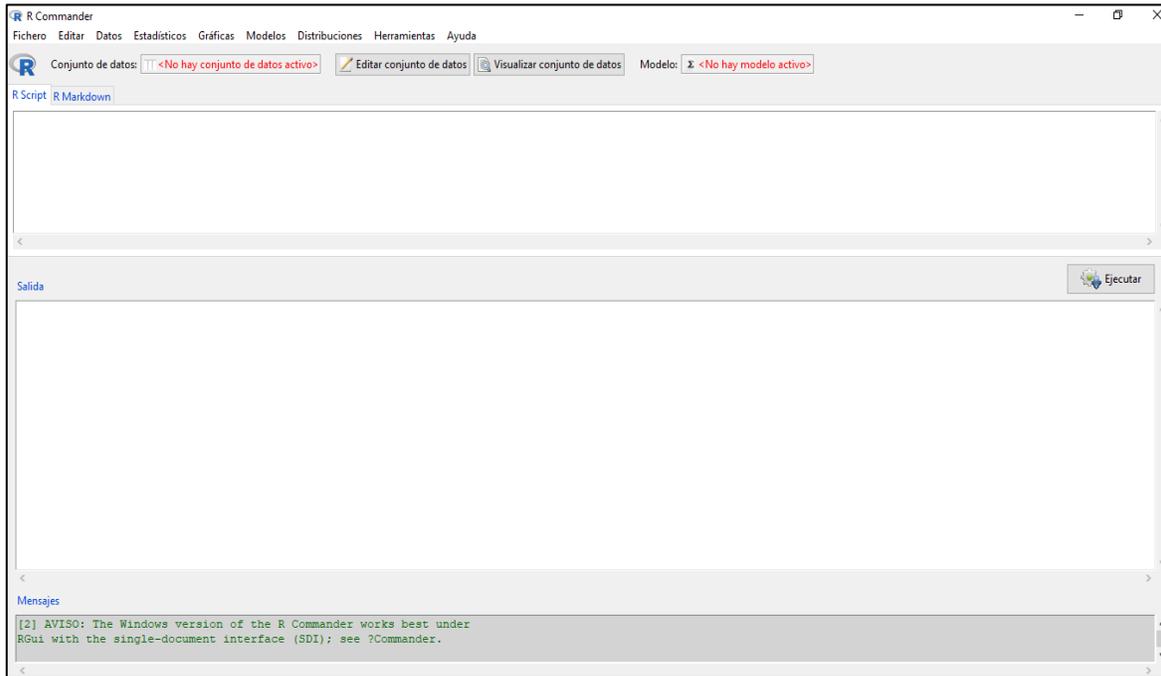
## **Instalación:**

Para instalar R, simplemente hay que ingresar a <https://cran.r-project.org/> y seleccionar la opción de descarga correspondiente al sistema operativo que utilizamos (por ejemplo, *Download R for Windows*. Si se instala por primera vez, bastará con elegir la opción *base*).

Habitualmente, se suele utilizar una interfaz más “amigable” para el trabajo en R, RStudio. Puede ser descargada en su sitio oficial (<https://www.rstudio.com/>) y también es de libre acceso. Sin embargo, en este curso será suficiente con tener instalado R.

Una vez en R, vamos a instalar el paquete “Rcmdr” que nos permitirá utilizar Rcommander. Para ello, tenemos que ir al menú superior que dice *Paquetes* y seleccionar *Instalar paquete(s)*. La ventana que se abre nos pide elegir el servidor desde donde descargaremos todo lo necesario. A continuación, elegimos alguna opción y, al aceptar, aparecerá una lista de paquetes disponibles. Para estar seguros de contar con todo lo necesario, seleccionamos todos los paquetes que comienzan con *Rcmdr*. Luego de darle OK, esperamos que finalice su instalación (suele tardar algunos minutos).

Para poder utilizar Rcommander, vamos a *Paquetes*, elegimos *Cargar paquete* y, de la lista que nos aparece, seleccionamos *Rcmdr*. Nos aparecerá la siguiente ventana de trabajo:



## Estadística descriptiva utilizando Rcommander:

Vamos a utilizar Rcommander para realizar un estudio descriptivo de variables cualitativas, cuantitativas discretas y cuantitativas continuas. Para ello, utilizaremos los datos correspondientes a las tres situaciones descritas en el apunte *Introducción a la estadística descriptiva* (Katz, R.D., disponible en <https://usuarios.fceia.unr.edu.ar/~valeoni/>).

### a) Variables cualitativas

En la Situación 1, se observa el tipo de falla que presenta cada caja de cartón. Dado que el tipo de falla puede *variar* en cada caja de cartón que observamos, diremos que se trata de una *variable*. Además, las observaciones que se obtienen son atributos o cualidades de cada unidad elemental (la caja de cartón), por lo que estamos en presencia de una *variable cualitativa*.

Primero, vamos a cargar las observaciones registradas. Para ello, vamos a *Datos* y luego a *Nuevo conjunto de datos*. Ingresamos un nombre para el conjunto de datos (supongamos, *Cajas*) y aceptamos. En la tabla que aparece, la primera columna (*rowname*) contiene el número de orden de cada observación. En la segunda columna (*V1*) podemos cargar nuestras observaciones. Podemos reemplazar su encabezado, *V1*, por el nombre de la variable, por ejemplo, *Tipo de falla*. Luego, en esa columna, a partir de la fila 1, cargaremos nuestras observaciones, una debajo de la otra y aceptamos una vez que finalizamos. Vale aclarar que es posible importar datos a Rcommander que se encuentren cargados en otros formatos (.csv, .xlsx, etc.)

Como primer acercamiento a los datos, podemos realizar una tabla de frecuencias. Para ello, vamos a *Estadísticos*, *Resúmenes* y elegimos *Distribución de frecuencias*. Elegimos la variable *Tipo de falla* y aceptamos. En la parte inferior de la pantalla, dentro de *Mensajes*, podemos

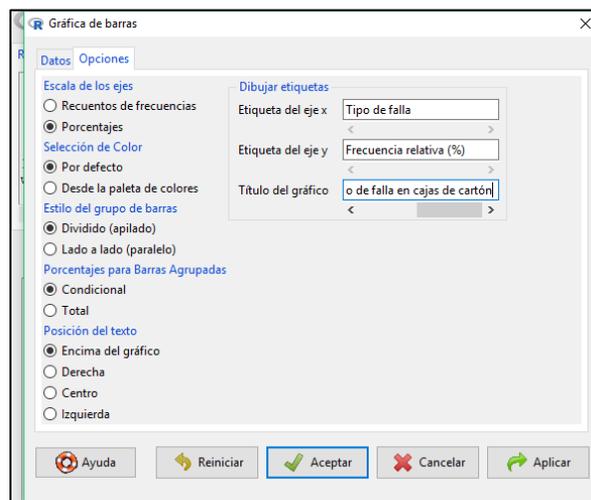
verificar si el conjunto de datos cargado tiene tantas observaciones como debe tener (en este caso, 50). Obtenemos dos tablas: la primera, corresponde a las frecuencias absolutas para cada tipo de falla. Así, diremos que hay tres cajas con cartón roto (falla A), 18 con cartón abultado (falla B), etc. La segunda tabla corresponde a las frecuencias relativas porcentuales. Observamos que el 6% de las cajas tiene el cartón roto (falla A), el 36% tiene cartón abultado (falla B), etc. ¿Cuál es la falla más frecuente? ¿Y la menos frecuente?

```

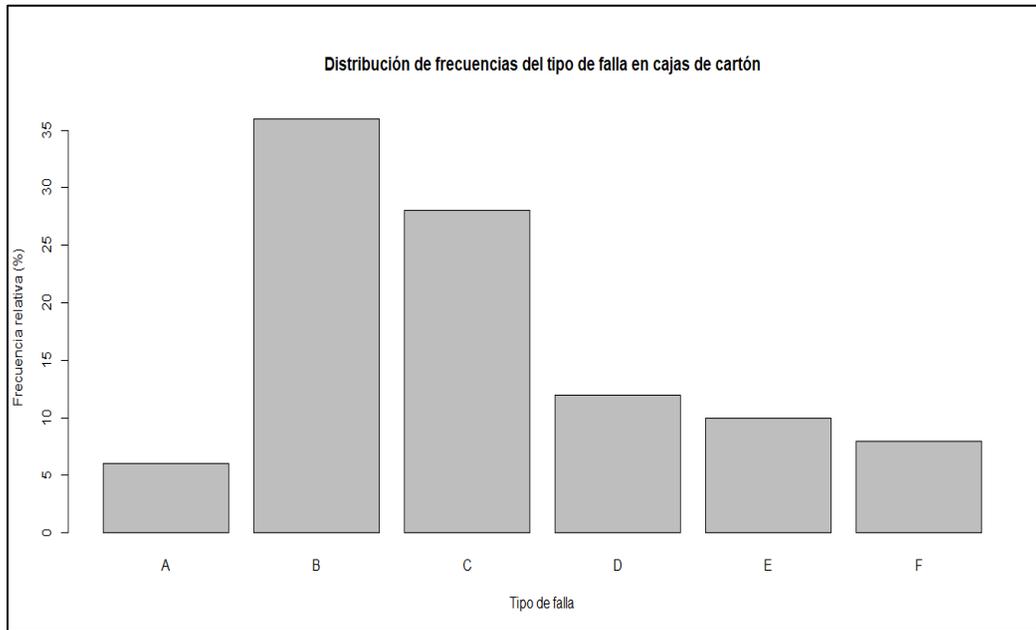
R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Conjunto de datos: Cajas Editar conjunto de datos Visualizar conjunto de datos
R Script R Markdown
summary(Cajas)
local({
  .Table <- with(Cajas, table(Tipo.de.falla))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
Salida
> local({
+ .Table <- with(Cajas, table(Tipo.de.falla))
+ cat("\ncounts:\n")
+ print(.Table)
+ cat("\npercentages:\n")
+ print(round(100*.Table/sum(.Table), 2))
+ })
counts:
Tipo.de.falla
A B C D E F
3 18 14 6 5 4
percentages:
Tipo.de.falla
A B C D E F
6 36 28 12 10 8
Mensajes
RGui with the single-document interface (SDI); see ?Commander.
[3] NOTA: El conjunto de datos Cajas tiene 50 filas y 1 columnas.

```

Vamos a hacer un diagrama de barras. Para esto, vamos a *Gráficas, Gráficas de barras*. Observe que son pocas las opciones disponibles dentro del menú *Gráficas*, ¿a qué cree que se debe? Dentro de la pestaña *Datos*, elegimos la variable correspondiente. En la pestaña *Opciones*, podemos optar si representar las frecuencias absolutas (*Recuentos de frecuencias*) o las relativas porcentuales (*Porcentajes*). Además, podemos elegir los nombres de los ejes.



Y así obtenemos el siguiente gráfico:

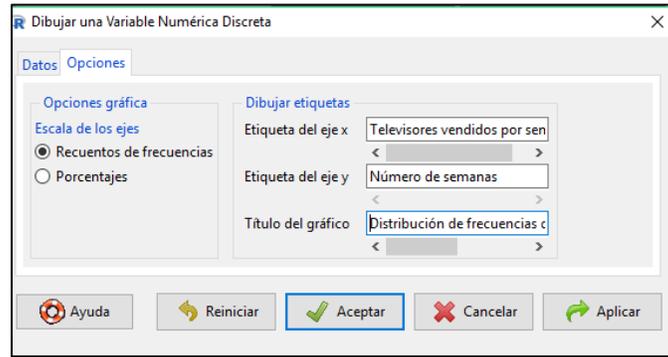


## b) Variables cuantitativas discretas

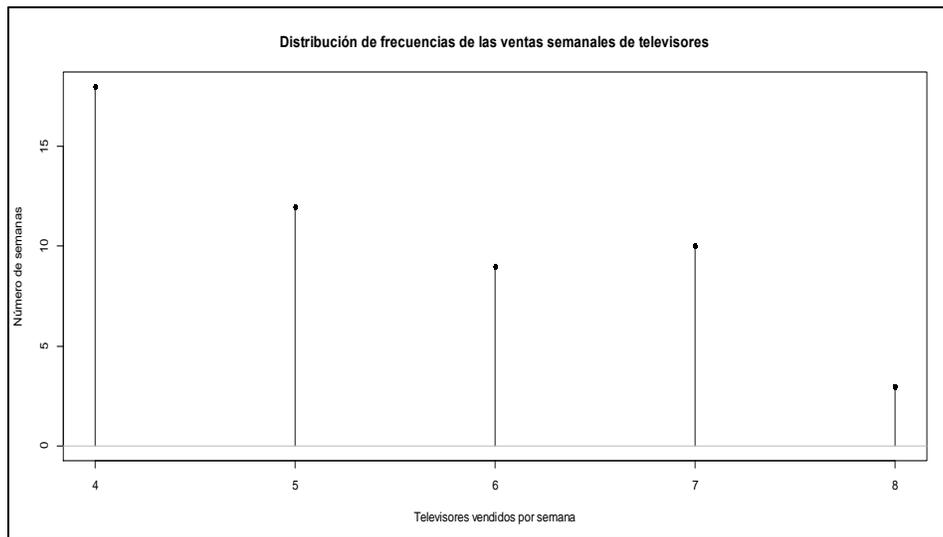
Primero, vamos a cargar los datos de la Situación 2 de manera similar a la descrita para la Situación 1. Podemos llamar a la variable de interés *Televisores vendidos*. Comenzamos calculando algunas medidas descriptivas de interés. Vamos a *Estadísticos, Resúmenes* y elegimos *Conjunto de datos activo*. En la salida encontramos los valores mínimo y máximo de nuestro conjunto de datos, y el valor de la media, mediana, primer cuartil y tercer cuartil. Tener estos datos nos permite, entre otras cosas, chequear algunos problemas de carga, como podría ser haber ingresado, por error de tipeo, un número muy grande.

```
Salida
> summary(Ventas)
Televisores.vendidos
Min.   :4.000
1st Qu.:5.000
Median :6.000
Mean   :5.731
3rd Qu.:6.250
Max.   :8.000
```

Podemos continuar realizando un diagrama de frecuencias absolutas para representar la distribución de frecuencias de la variable. Para esto, vamos a *Gráficas* y seleccionamos *Dibujar una variable numérica discreta*. En el menú que aparece, seleccionamos el conjunto de datos donde cargamos nuestra variable y luego tenemos opciones similares a las vistas para el diagrama de barras para variables cualitativas. Si queremos representar las frecuencias absolutas, ponemos:



y obtenemos el siguiente gráfico:

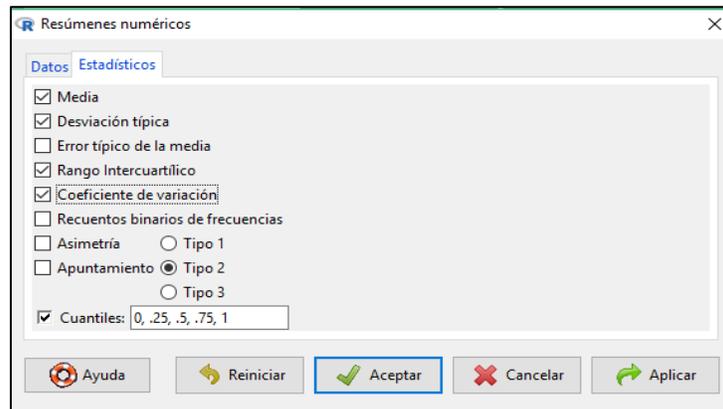


### c) Variables cuantitativas continuas

Vamos a considerar ahora los datos de la Situación 3. Los mismos deben ser cargados de la forma ya mencionada.

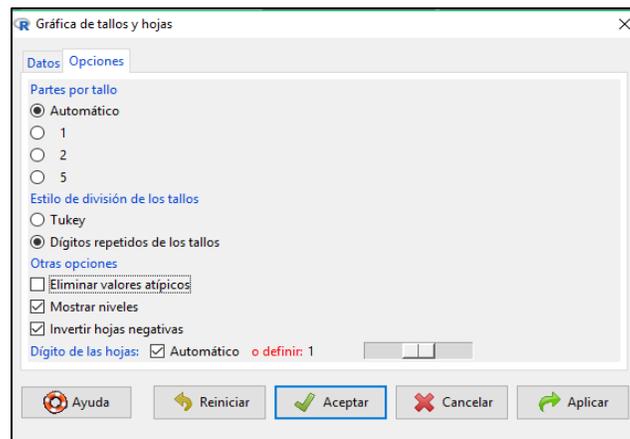
Igual que hicimos con los datos de la Situación 2, vamos a explorar los datos del contenido de las bolsas de azúcar mediante el cálculo de medidas descriptivas. Podemos hacerlo yendo a *Estadísticos, Resúmenes, Conjunto de datos activo* al igual que antes o bien yendo a *Estadísticos, Resúmenes* y allí seleccionamos *Resúmenes numéricos* (este camino también se puede hacer cuando estamos frente a variables cuantitativas discretas). En la pestaña *Datos* elegimos nuestra variable y en *Estadísticos* podemos elegir las medidas descriptivas que nos interesan. En el siguiente ejemplo, se pide el cálculo del promedio (media), el desvío estándar (desviación típica), el rango intercuartílico, el coeficiente de variación, el valor mínimo (cuantil 0), el máximo (cuantil 1), la mediana (cuantil correspondiente al 0.50 de los datos), el primer cuartil (cuantil correspondiente al 0.25 de los datos) y el tercer cuartil (cuantil correspondiente al 0.75

de los datos). Repasaremos más adelante la interpretación de estos valores en el contexto del problema.



```
> numSummary(Azúcar["Peso..en.g.", drop=FALSE], statistics=c("mean", "sd", "IQR",
+ "quantiles", "cv"), quantiles=c(0,.25,.5,.75,1))
  mean      sd  IQR      cv 0% 25% 50% 75% 100% n
504.18 22.58922 30.75 0.04480389 457 488 505 518.75 554 50
```

Una primera aproximación a la distribución de los datos se puede obtener mediante el gráfico de tallo y hojas. En R Commander, podemos solicitarlo yendo a *Gráficas, Gráficas de tallos y hojas*. La configuración de la ventana de opciones suele ser:

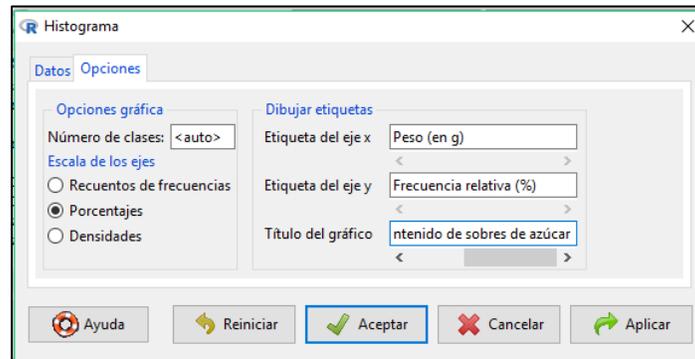


Para los datos de la Situación 3, obtenemos:

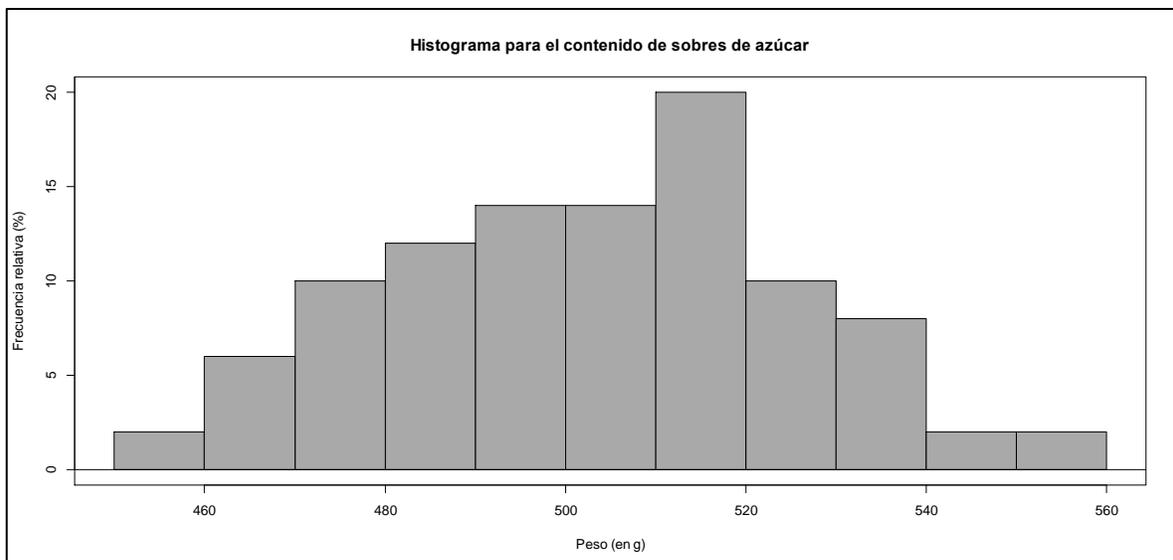
```
> with(Azúcar, stem.leaf(Peso..en.g., style="bare", trim.outliers=FALSE, na.rm=TRUE))
1 | 2: represents 12
leaf unit: 1
      n: 50
 1  45 | 7
 3  46 | 89
 9  47 | 013488
15  48 | 456889
21  49 | 247799
(8) 50 | 02337889
21  51 | 112346689
12  52 | 068
 9  53 | 000125
 3  54 | 07
 1  55 | 4
```

Veremos en clase cómo interpretamos este diagrama.

Otro gráfico muy utilizado y que es específico para la representación de datos correspondientes a variables cuantitativas continuas es el histograma. Para realizarlo, vamos a *Gráficas*, y elegimos *Histograma*. Podemos optar por representar las frecuencias absolutas o las relativas porcentuales. Por ejemplo:



Y obtenemos:

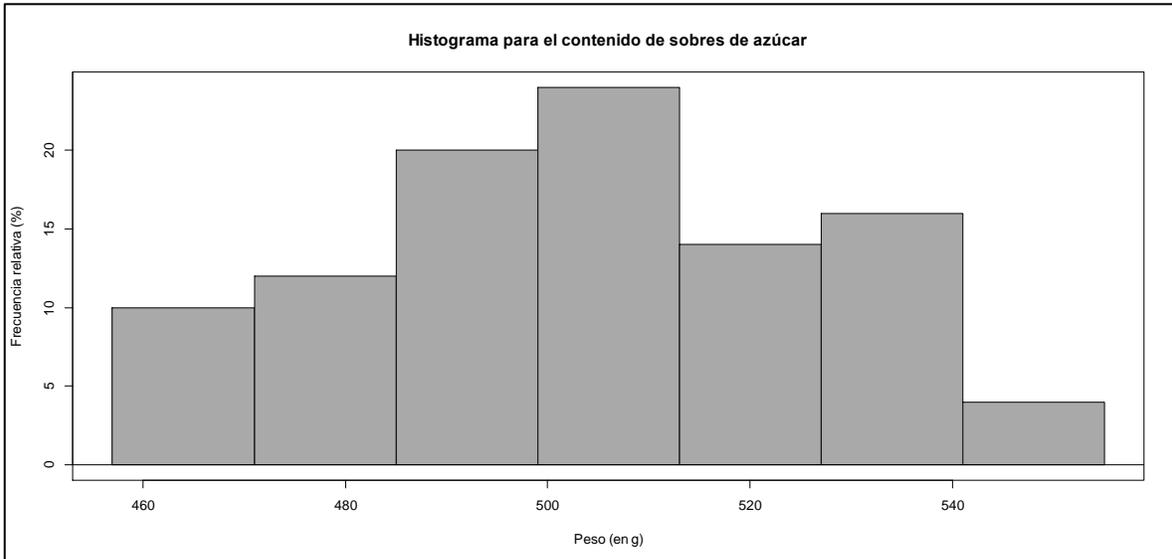


En el caso anterior, dejamos que Rcommander calcule en forma automática el número de intervalos a utilizar (Número de clases: <auto>), pero esto puede ser modificado. Sin embargo, si cambiamos el número de clases desde la ventana de configuración, Rcommander sigue eligiendo el número de intervalos de modo que los límites de cada uno de ellos ocurran en números “redondos”. Entonces es necesario realizar algunas modificaciones desde la consola R script. Por ejemplo, si consideramos el criterio de hacer tantos intervalos como la raíz cuadrada de la cantidad de datos, en este caso tenemos  $\sqrt{50} \cong 7$  intervalos. El menor valor observado es 457 g y el mayor es 554, dando una amplitud de 97 unidades, por lo que cada uno de nuestros intervalos debería tener una amplitud de  $97/7 \cong 14$  unidades. Dado que por defecto los intervalos son cerrados por izquierda y abiertos por derecha, para que se incluya la mayor

observación entonces debemos hacer llegar el último intervalo hasta 555. La sentencia debe quedar así:

```
with(Azúcar, Hist(Peso..en.g., scale="percent", breaks=seq(457,555,by=14), col="darkgray",  
  xlab="Peso (en g)", ylab="Frecuencia relativa (%)",  
  main="Histograma para el contenido de sobres de azúcar"))
```

Seleccionamos y al ejecutar obtenemos:



Finalmente, también podemos optar por realizar un diagrama de caja. Vamos a *Gráficas*, *Diagrama de caja* y seleccionamos nuestra variable, así obtendremos:

