Práctica 2

Estadística descriptiva unidimensional

Contenido

Los ejemplos de esta práctica se realizarán con el conjunto de datos *acero2.*

1. Frecuencias y porcentajes

Para obtener las frecuencias y los porcentajes de las modalidades de una variable utilizaremos el menú *Estadísticos*.

R commander sólo calcula frecuencias y porcentajes de variables factor (cualitativas). No podemos obtener directamente la distribución de frecuencias de variables numéricas, antes tenemos que convertirlas en factor.

Ejemplo 1. Halla la distribución de frecuencias de la variable estadística *averias*.

Solución: Procede de la siguiente forma

Estadísticos → Resúmenes → Distribución de frecuencias

selecciona la variable *averias* y acepta.

Los pasos anteriores proporcionan el siguiente resultado:

```
> .Table <- table(acero2$averias)

> .Table # counts for averias

No  Si
89  28

> round(100*.Table/sum(.Table), 2) # percentages for averias

  No   Si
76.07 23.93
```

Así, se han obtenido el número de casos (frecuencia absoluta) y el porcentaje de cada modalidad de la variable *averias* dentro de la muestra.

Ejemplo 2. Halla la distribución de frecuencias de la variable estadística *naverias*.

Solución: En este caso, al tratarse una variable numérica, debes crear en primer lugar una nueva variable, *fnaverias*, de tipo factor, te recuerdo el proceso a seguir:

Datos → Modificar variables del conjunto de datos activo → Convertir variable numérica en factor

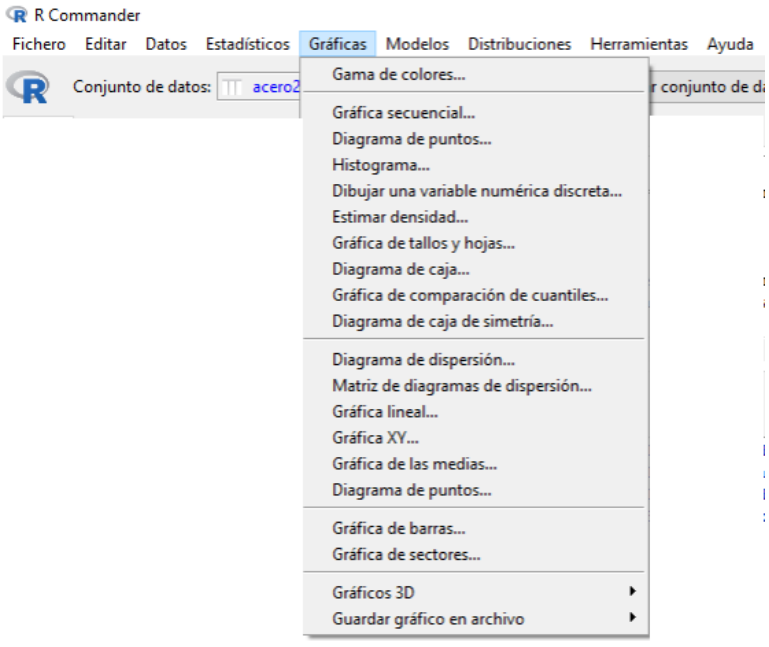
y en la ventana emergente, selecciona la variable *naverias*, *Utilizar números*, escribe el nuevo nombre (*fnaverias)* y *acepta.*

Una vez hecho esto, siguiendo los pasos del ejemplo anterior (*Estadísticos → Resúmenes → Distribución de frecuencias,* selecciona la variable *fnaverias* y acepta), obtienes

```
> acero2$fnaverias <- as.factor(acero2$naverias)

> .Table <- table(acero2$fnaverias)

> .Table # counts for fnaverias

 0  1  2  3  4
89  2  9  9  8


> round(100*.Table/sum(.Table), 2) # percentages for fnaverias

   0    1    2    3    4
76.07 1.71 7.69 7.69 6.84
```

2. Gráficas

Para realizar representaciones gráficas emplearemos la opción del menú *Gráficas*.



R Commander sólo ejecuta los gráficos de sectores y de barras para variables de tipo factor (cualitativas). No podemos obtener directamente esa clase de gráficos para variables de tipo numérico, antes tenemos que convertirlas en factor.

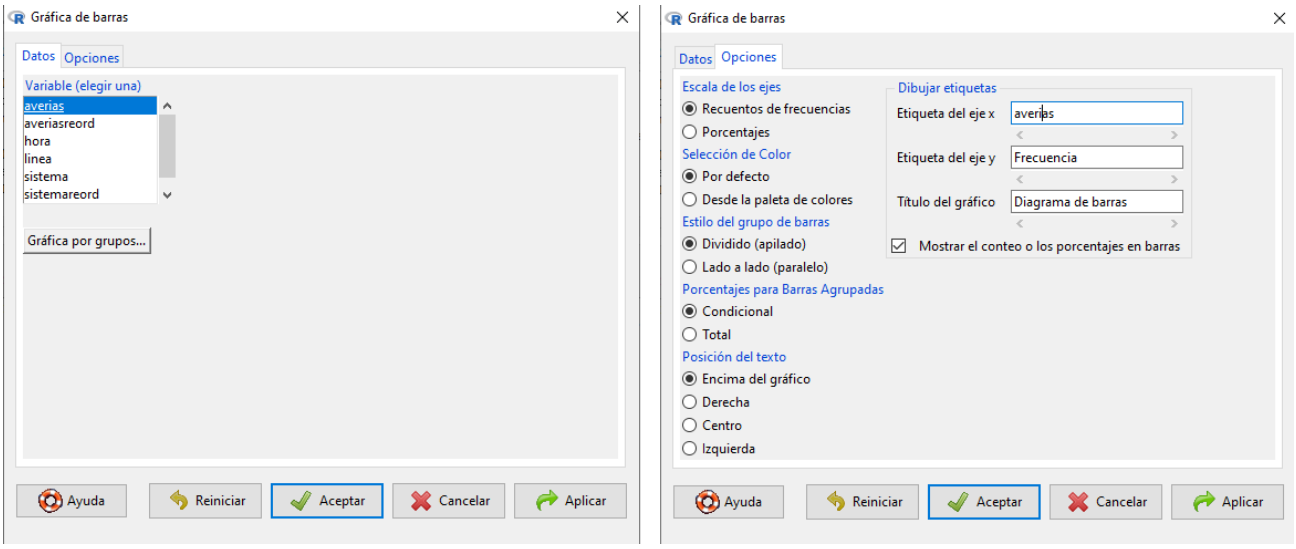
2.1. Diagrama de barras

Ejemplo 3. Representa gráficamente la distribución de la variable *averias* mediante una gráfica de barras.

Solución: Se trata de una variable cualitativa, por lo que una forma adecuada de representarla gráficamente sería utilizar un diagrama de barras. Para obtener la gráfica sigue la ruta

Gráficas→Gráficas de barras

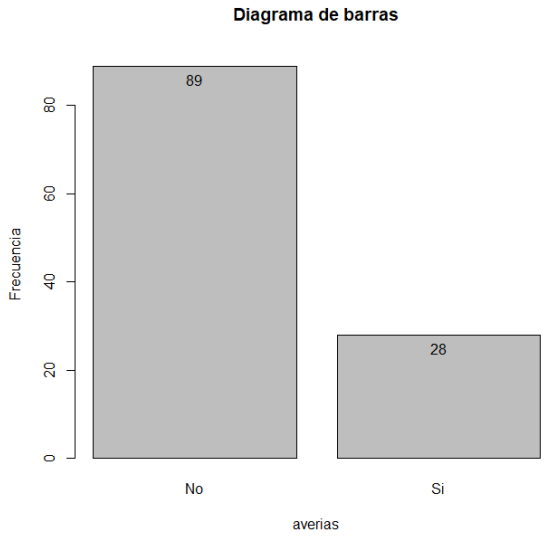
y rellena las pestañas de *Datos* y *Opciones* en la ventana emergente



En las ventanas de instrucciones y salida aparece la instrucción *R* para dibujar el gráfico de barras:

```
Barplot(averias, xlab="averias", ylab="Frecuencia", main="Diagrama de barras", label.bars=TRUE)
```

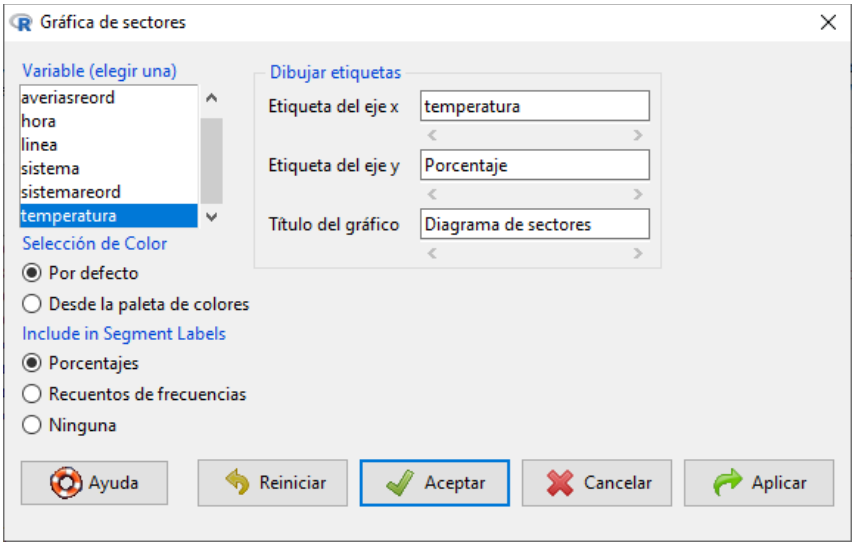
La gráfica obtenida es



2.2. Diagrama de sectores

Ejemplo 4. Representa gráficamente la distribución de la variable *temperatura* mediante un gráfico de sectores en el que figuren como etiquetas los porcentajes.

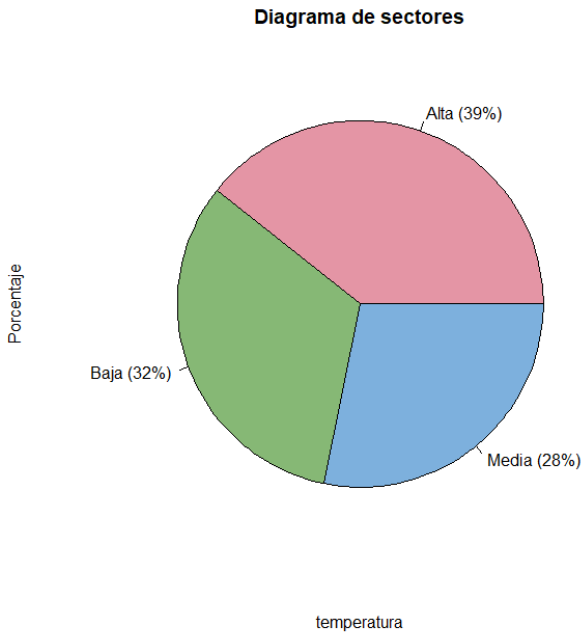
Solución: Sigue la ruta **Gáficas → Gráfica de sectores**, selecciona la variable temperatura, rellena la ventana emergente y acepta.



O bien, escribe en la ventana de instrucciones:

```
piechart(temperatura, xlab="temperatura", ylab="Porcentaje", main="Diagrama de sectores", col=rainbow_hcl(3), scale="percent")
```

y pincha en *ejecutar*.

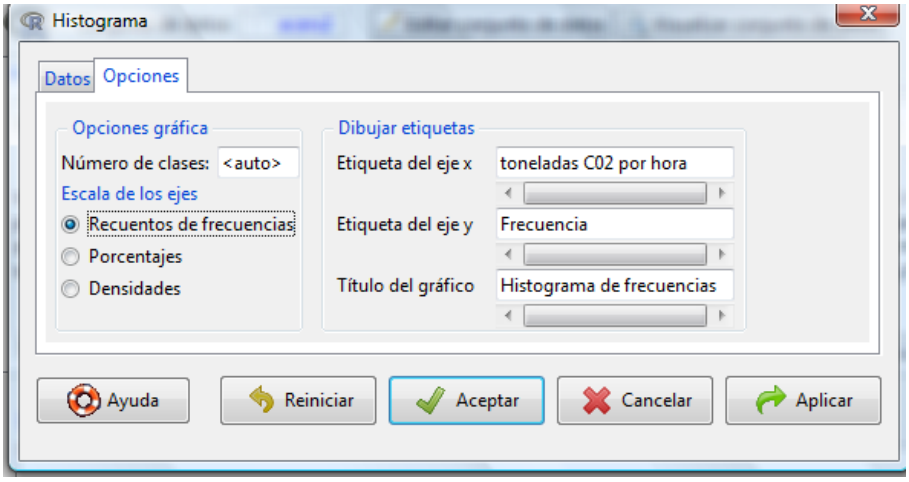


2.3. Histograma

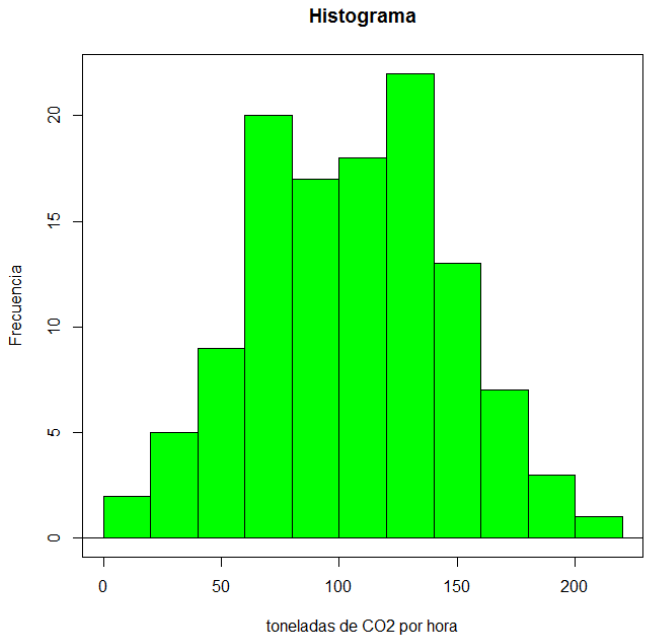
Sirve para representar variables estadísticas numéricas continuas.

Ejemplo 5. Representa el histograma de la variable *CO2*.

Solución: Sigue la secuencia **Gráficas → Histograma**, en la pestaña *Datos* de la ventana *Histograma* selecciona la variable *CO2*, pincha en la pestaña **opciones** y rellena la ventana de la forma siguiente



Y pincha en *Aceptar*.



Escribiendo en la ventana de instrucciones

```
Hist(acero2$CO2, scale="frequency", breaks="Sturges", col="green", xlab="toneladas de CO2 por hora", ylab="Frecuencia", main="Histograma")
```

conseguimos el mismo resultado.

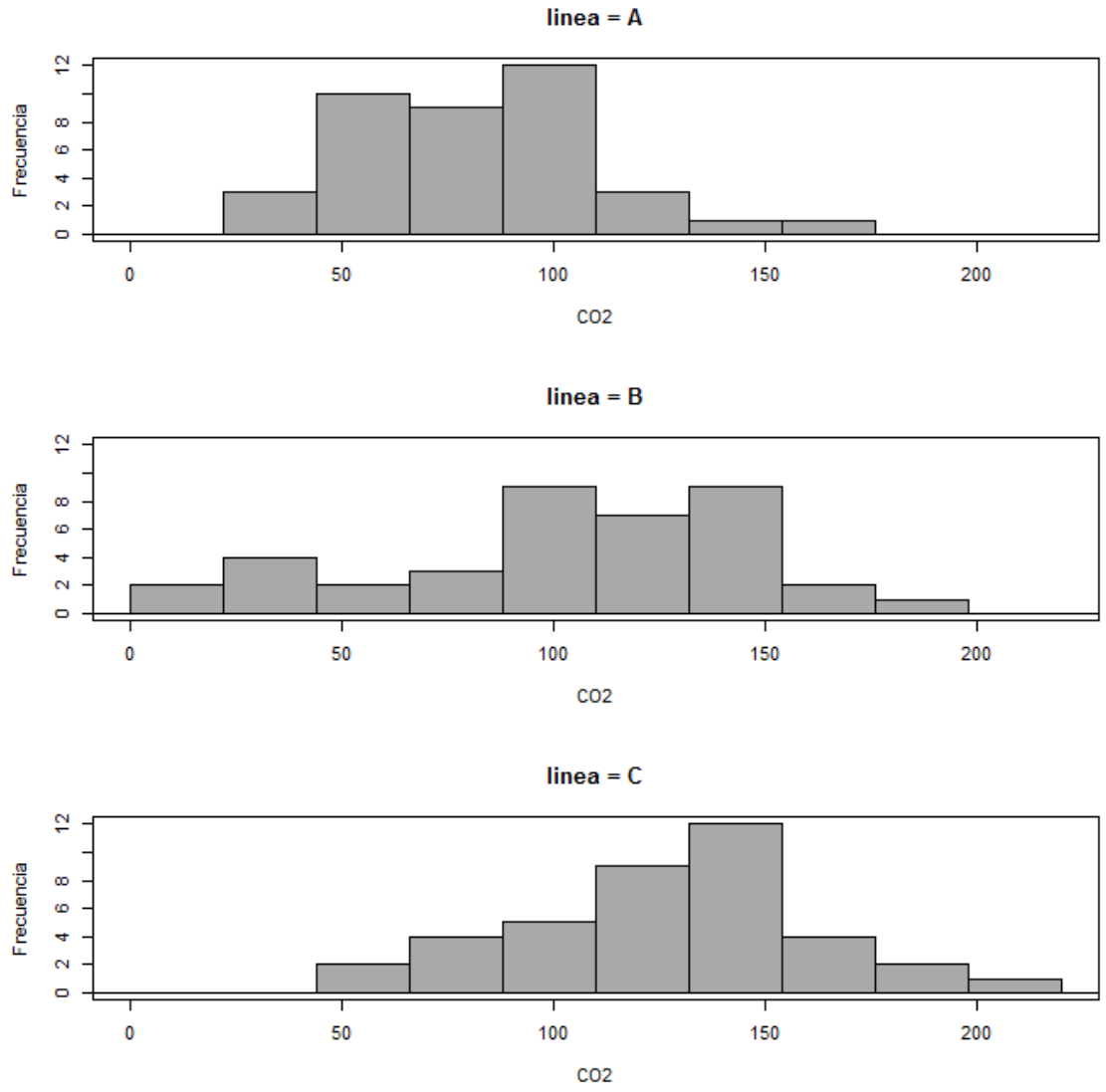
Notas:

- Por defecto, el número de barras del histograma se calcula automáticamente. Puede indicarse un número concreto en la opción *Número de clases*; sin embargo, *Rcmdr* considera ese número como una sugerencia (el cálculo del número de barras lo hace de forma que los hitos de los ejes queden entre dos barras y sean números redondos).
- Por defecto, la altura de las barras está expresada en frecuencias. Se pueden escoger también porcentajes o densidades, en la opción *Escala de los ejes*.

También podemos obtener los histogramas de una variable para cada nivel de un factor lo que nos permite realizar comparaciones.

Ejemplo 6. Obtén los histogramas de la variable *CO2* para cada nivel de la variable *linea*.

Solución: Si en la pestaña *Datos* de la ventana *Histograma* tras seleccionar la variable *CO2*, pinchas *gráfica por grupos* y en la ventana *grupos* seleccionas la variable *linea*, obtendrás los histogramas de la variable CO2 según cada nivel de la variable *linea.*

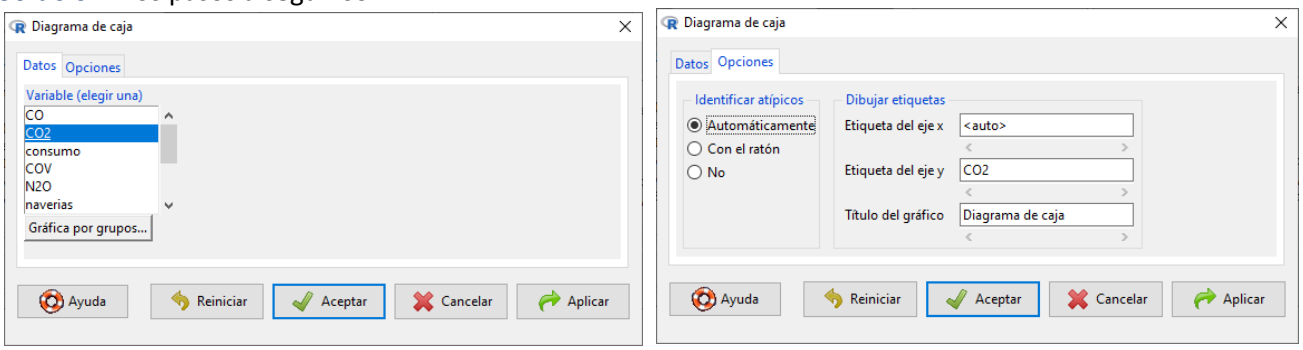


2.4. Diagrama de caja

Ahora vamos a ver otra representación gráfica de las variables estadísticas cuantitativas continuas, los gráficos de caja, útiles para detectar valores atípicos y comparar la distribución de una variable estadística en distintas muestras.

Ejemplo 7. Obtén el diagrama de caja de la variable *CO2*.

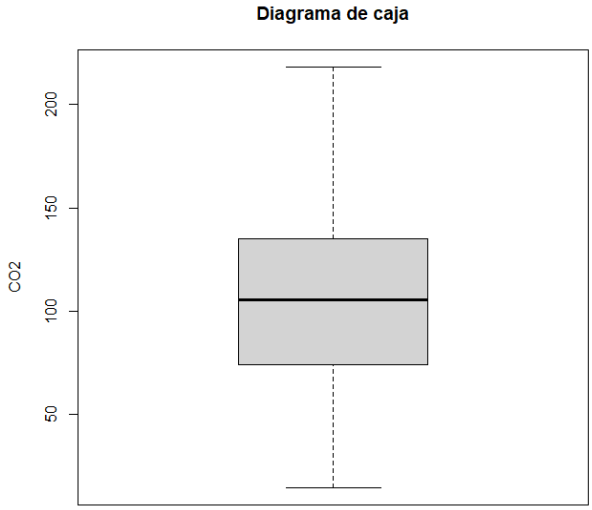
Solución: Los pasos a seguir son



En la ventana de instrucciones aparece:

```
Boxplot( ~ CO2, data=acero2, id=list(method="y"), ylab="CO2", main="Diagrama de caja")
```

El resultado es



A partir de dicho diagrama se observa, por ejemplo, que no existen datos atípicos para la variable *CO2* en esta muestra.

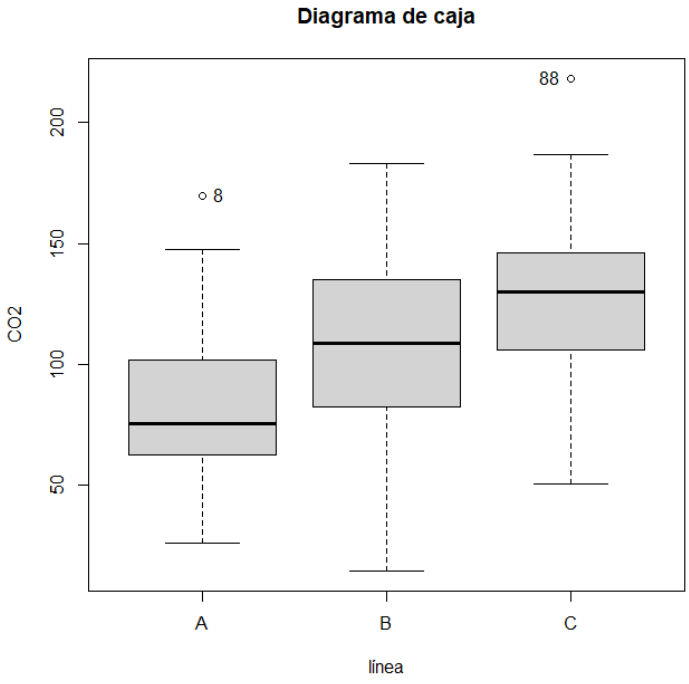
Ejemplo 8. Obtén el diagrama de caja de la variable *CO2* para cada nivel de la variable *linea*.

Solución: Sigue los pasos del ejercicio anterior, pero en la pestaña *Datos* de la ventana *Diagrama de caja* pincha en *Gráfica por grupos*, en la ventana *Grupos* selecciona la variable *linea* y acepta.

En la ventana de instrucciones figura:

```
Boxplot(CO2~linea, data=acero2, id=list(method="y"), xlab="línea", ylab="CO2", main="Diagrama de caja")
```

El gráfico de caja por grupos es:



En el gráfico se observan dos valores atípicos:

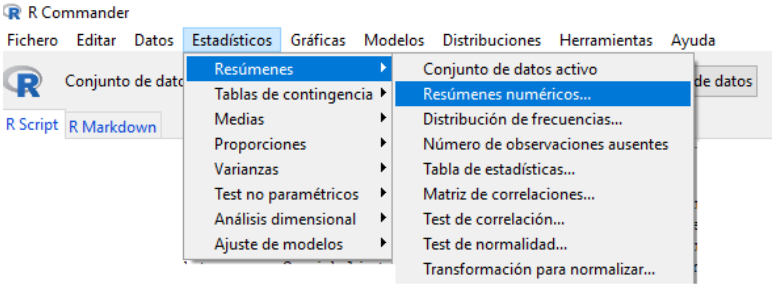
- en la observación 8, 8ª hora del primer día en la línea A, en esa hora se emiten 169.8100 toneladas de CO2.
- en la observación 88, 2ª hora del segundo día en la línea C, en esa hora se emiten 218.3075 toneladas de CO2

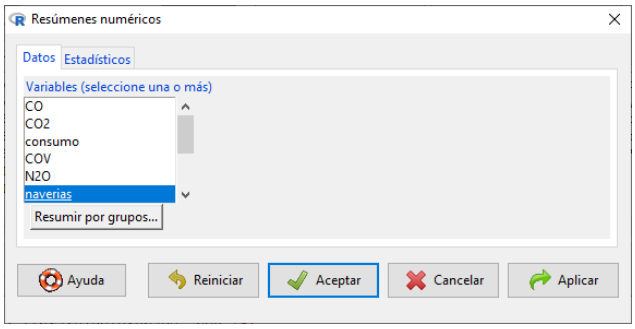
3. Medidas de centralización y dispersión

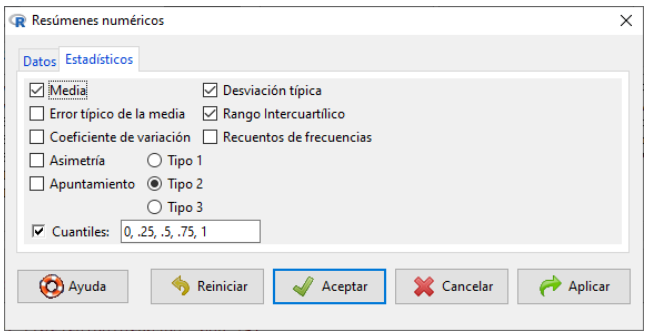
En esta sección vamos a ver cómo podemos calcular medias, desviaciones típicas y percentiles de variables numéricas. Utilizaremos el menú *Estadísticos.*

Ejemplo 9. Calcula la media, la desviación típica, los cuartiles, el rango y el recorrido intercuartílico de la variable *naverias*.

Solución: Sigue los pasos








```
numSummary(acero2[,"naverias", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"),
quantiles=c(0,.25,.5,.75,1))

      mean       sd IQR 0% 25% 50% 75% 100%   n
 0.6752137 1.292078   0  0   0   0   0    4 117
```

Los resultados nos indican que la media es de 0.6752137 averías por hora, con una desviación típica de 1.292078 averías por hora. Los tres cuartiles (25%, 50% y 75%) son iguales a 0. El número de averías varía desde 0 (el 0% es el valor mínimo de la muestra) hasta 4 (el 100% es el valor máximo de la muestra) por tanto el rango es 4. El recorrido intercuartílico (= IQR) es igual a 0.

Se observa que en, al menos, el 75% de las horas observadas no hubo averías (el percentil 75 es igual a 0).

Ejemplo 10. Calcula los principales estadísticos descriptivos de la variable *CO2*.

Solución: Sigue los pasos del ejercicio anterior pero selecciona la variable *CO2*.

```
numSummary(acero2[,"CO2", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"),
quantiles=c(0,.25,.5,.75,1))
     mean       sd   IQR     0%     25%      50%      75%     100%   n
 104.6281 41.38971 61.16 14.285 73.7175 105.3075 134.8775 218.3075 117
```

La emisión media de CO2 es de 104.63 toneladas/hora, con una desviación típica de 41.39 toneladas/hora. La emisión mínima de CO2 es 14.285 toneladas/hora y la máxima es 218.3075 toneladas/hora. El 25% de los casos analizados emiten 73.7175 toneladas de CO2/hora o menos; el 50 %, como mucho 105.3075 toneladas/hora y un 25% emite al menos 134.8775 toneladas de CO2/hora.

El botón *Resumir por grupos* nos permite obtener los resúmenes numéricos para cada modalidad de una variable factor. En este caso es conveniente marcar el estadístico "Coeficiente de variación" para poder comparar adecuadamente la dispersión respecto a la media. Por ejemplo, las medidas descriptivas de la variable CO2 según la línea de producción son:

```
numSummary(acero2[,"CO2", drop=FALSE], groups=acero2$linea, statistics=c("mean", "sd", "IQR",
"quantiles", "cv"), quantiles=c(0,.25,.5,.75,1))
       mean       sd      IQR        cv      0%      25%      50%      75%     100% CO2:n
A  82.46763 29.70361 39.39625 0.3601850 25.8125  62.3125  75.2800 101.7088 169.8100    39
B 103.37429 44.83468 52.70250 0.4337121 14.2850  82.0750 108.4175 134.7775 183.2675    39
C 128.04250 35.61952 40.27375 0.2781851 50.3175 105.6412 129.7475 145.9150 218.3075    39
```

Como puede observarse, en términos relativos es menor la dispersión del CO2 en la línea C que en la A, aunque su desviación típica sea mayor en la línea C que en la A.

4. Ejercicios propuestos

Lee el conjunto de datos *mtcars* del paquete *datasets* de *R.*

Ejercicio 1. Representa mediante un gráfico de sectores y un gráfico de barras la variable *carb*. ¿Qué representación gráfica es más adecuada?

Ejercicio 2. ¿Qué gráfico es apropiado para representar el consumo (*mpg*)? Represéntalo usando porcentajes en el eje de ordenadas.

Ejercicio 3. ¿Cuántos automóviles tienen cambio manual (*am* = 1)? ¿Cuál es el porcentaje de vehículos con cambio manual?

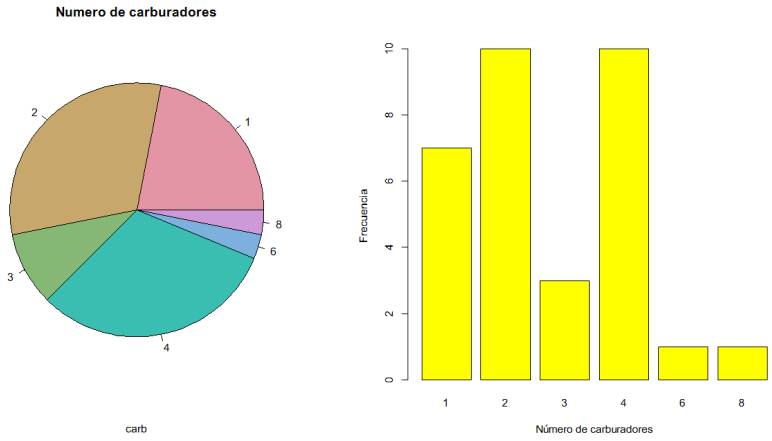
Ejercicio 4. Observa la distribución del peso (*wt*). ¿Cuánto vale el peso medio? ¿Cuánto vale la mediana? Calcula el rango y el recorrido intercuartílico.

Ejercicio 5. Del total de la muestra, ¿puede asegurarse que en al menos un 25% de los datos el consumo ha sido mayor de 21 mpg?, ¿podemos asegurar que más del 50% de los vehículos consumen más de 15'5 mpg y menos de 20 mpg?

Ejercicio 6. Realiza un gráfico para el consumo de los automóviles en el que aparezcan dos diagramas de cajas, uno para el consumo cuando el vehículo tiene cilindros en v y otro cuando tiene cilindros en serie. Comenta dicho gráfico. ¿Cuánto vale el consumo medio y el rango en cada uno de estos dos casos?

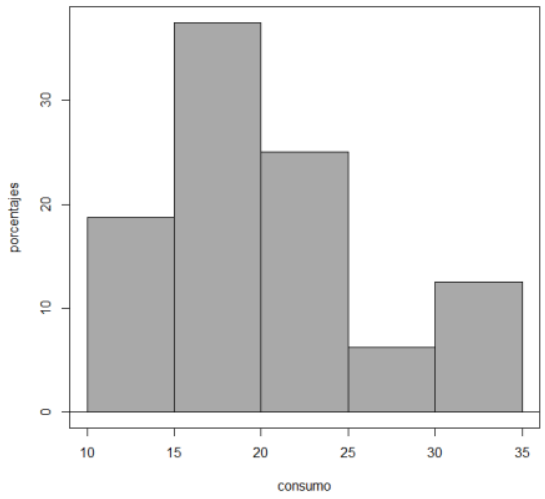
Soluciones

Ejercicio 1. Como *carb* es numérica debes convertirla primero en factor, las gráficas pedidas para *fcarb* (*carb* convertida en factor)son:



En este caso es más adecuado el gráfico de barras, porque las modalidades de la variable son cantidades, y la relación de orden entre ellas queda diluida por la representación circular.

Ejercicio 2. Como *mpg* es una variable cuantitativa continua podemos utilizar un histograma o un diagrama de caja. Ya que nos dicen que lo representemos usando porcentajes en el eje de ordenadas, dibujaremos un histograma



Ejercicio 3. Distribución de frecuencias de *fam* (*am* convertida en factor)
counts:
fam
0 1
19 13
¿Cuántos automóviles tienen cambio manual (*am* = 1)? 13.
percentages:
fam
0 1
59.38 40.62
¿Cuál es el porcentaje de vehículos con cambio manual ? 40.62%.

Ejercicio 4.

mean sd IQR 0% 25% 50% 75% 100% n
3.21725 0.9784574 1.02875 1.513 2.58125 3.325 3.61 5.424 32

Peso medio = 3.21725 libras/1000
Mediana (50%) = 3.325 libras/1000
Rango = 5.424 – 1.513 = 3.911
RIC = 1.02875

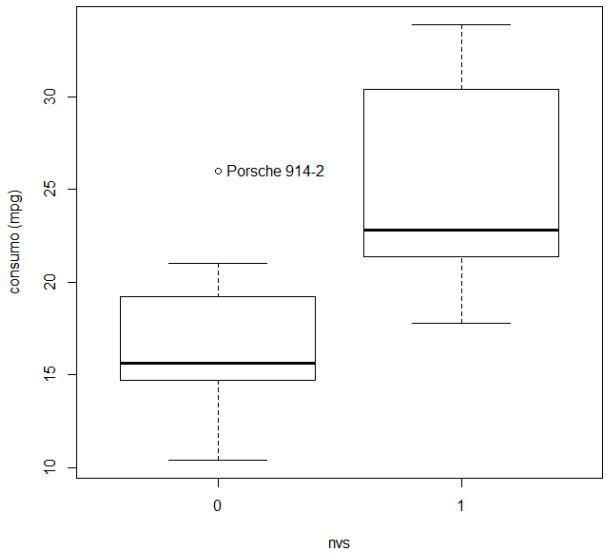
Ejercicio 5.

mean sd IQR 0% 25% 50% 75% 100% n
20.09062 6.026948 7.375 10.4 15.425 19.2 22.8 33.9 32

¿Puede asegurarse que en al menos un 25% de los datos el consumo ha sido mayor de 21 mpg? Si ya que el percentil 75 es 22.8.

¿Podemos asegurar que más del 50% de los vehículos consumen más de 15’5 mpg y menos de 20 mpg? No porque el intervalo (15’5,20) está contenido en el intervalo [primer cuartil, tercer cuartil] = [15.425, 22.8], menos del 50% de los automóviles consumen más de 15’5 mpg y menos de 20 mpg.

Ejercicio 6. Diagrama de caja por grupos:



Resúmenes numéricos por grupos

	mean	sd	IQR	0%	25%	50%	75%	100%	data:n
0	16.61667	3.860699	4.300	10.4	14.775	15.65	19.075	26.0	18
1	24.55714	5.378978	8.225	17.8	21.400	22.80	29.625	33.9	14

Consumo medio si vs = 0: 16.61667 mpg y Rango si vs = 0: 26.0 – 10.4 = 15.6
Consumo medio si vs = 1: 24.55714 mpg y Rango si vs =1: 33.9 – 17.8 = 16.1