

EACS-Estadística descriptiva univariante

Índice

1. Estudio descriptivo de variables cualitativas
 - Tablas de frecuencias
 - Gráficos de sectores y de barras
 - Tabla de frecuencias acumuladas
2. Estudio descriptivo de variables discretas
 - Medidas descriptivas y tabla de frecuencias
 - Gráfico de líneas y gráfico de barras
3. Estudio descriptivo de variables continuas
 - Medidas descriptivas
 - Gráficos para variables continuas: Diagrama de caja, Histograma, Estimación de la densidad, Diagrama de puntos, Diagrama de tallos y hojas
4. Ejercicios

1. Estudio descriptivo de variables cualitativas

Las variables cualitativas se tratan en R Commander como variables de tipo factor. Las herramientas básicas para describirlas son las tablas de frecuencias, los diagramas de sectores y los diagramas de barras.

El archivo *Pulso.RData*, que ya hemos utilizado anteriormente, contiene información sobre una muestra compuesta por datos de estudiantes, recogidos durante varios cursos académicos, que participaron en el siguiente experimento: Primero se tomaron su pulso cardíaco en reposo (*Pulso1*). Luego, lanzaron una moneda al aire. Si la moneda salía cara, debían correr durante un minuto, en otro caso, se quedaban sentados durante ese minuto. Al cabo de este minuto, todos los estudiantes se volvieron a tomar el pulso (*Pulso2*).

El archivo contiene también información adicional de los estudiantes, en particular las variables numéricas *Altura*, *Peso*, *Edad* y *Año* (que contiene el año del curso académico en el que se tomaron los datos) y las variables categóricas *Sexo* (que ahora es factor y toma valores “hombre” y “mujer”), *Fumador*, *Alcohol*, *Ejercicio* (que es ordinal, con las categorías “baja”, “moderada” y “alta”) y *Correr* (que contiene el resultado del lanzamiento de la moneda).

Notad que el nombre del fichero externo es *Pulso.RData*, mientras que el nombre interno de la base de datos (*data frame* en R) es *pulso*.

Un resumen de todas estas variables se puede obtener con la opción *Estadísticos > Resúmenes > Conjunto de datos activo*, que proporciona esta salida:

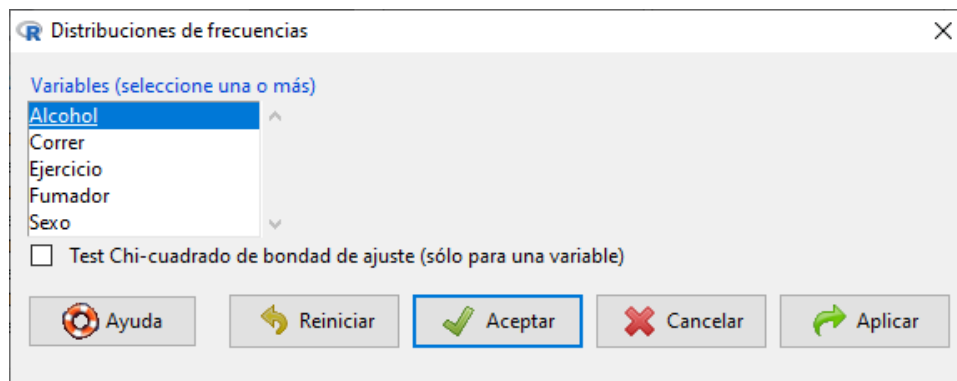
```
> summary(pulso)
  Altura      Peso      Edad      Sexo      Fumador Alcohol  Ejercicio Correr  Pulsol
Min.   : 68.0   Min.   : 27.00   Min.   :18.00   hombre:59   no:99   no:42   baja   :37   no:64   Min.   : 47.00
1st Qu.:165.2   1st Qu.: 56.25   1st Qu.:19.00   mujer :51   si:11   si:68   moderada:59   si:46   1st Qu.: 68.00
Median :172.5   Median : 63.00   Median :20.00                                     alta   :14   Median : 76.00
Mean   :171.6   Mean   : 66.33   Mean   :20.56                                     3rd Qu.: 82.00
3rd Qu.:180.0   3rd Qu.: 75.00   3rd Qu.:21.00                                     Max.   :145.00
Max.   :195.0   Max.   :110.00   Max.   :45.00                                     NA's   :1

  Pulso2      Año
Min.   : 56.0   Min.   :93.00
1st Qu.: 72.0   1st Qu.:95.00
Median : 84.0   Median :96.00
Mean   : 96.8   Mean   :95.63
3rd Qu.:125.0   3rd Qu.:97.00
Max.   :176.0   Max.   :98.00
NA's   :1
```

Tablas de frecuencias

En el fichero *Pulso.RData* tenemos la variable *Alcohol*, que indica si un estudiante bebe regularmente o no. Para conocer cuántas personas beben regularmente y cuántas no, así como sus porcentajes sobre el total de la muestra, podemos obtener las tablas de frecuencias absolutas y de porcentajes de la variable *Alcohol*. Para ello:

- 1) Carga el archivo *Pulso.RData* en R Commander con la opción *Datos > Cargar conjunto de datos*.
- 2) Selecciona *Estadísticos > Resúmenes > Distribución de frecuencias*. En el cuadro de diálogo que aparece, selecciona la variable *Alcohol*.



- 3) Tras pulsar *Aceptar*, en el cuadro de resultados (*Salida*) aparecerán la tabla de frecuencias absolutas y la tabla de porcentajes, con el siguiente aspecto:

```
counts:
Alcohol
no sí
42 68
```

```
percentages:
Alcohol
      no      sí
38.18 61.82
```

A la vista de los resultados, y redondeando a un decimal, podemos describir esta variable diciendo que, **de los 110 estudiantes de la muestra, un 61.8% refiere que bebe regularmente (68 estudiantes), mientras que sólo un 38.2% refiere que no lo hace (42 estudiantes)**.

Se pueden obtener las tablas de frecuencias de varias variables cualitativas a la vez, sin más que seleccionarlas en el cuadro de diálogo anterior (utilizando si es necesario la tecla *<Ctrl>* y el ratón para seleccionarlas).

NOTA 1.1: Para llevar los resultados de las tablas a un documento en Word no hay más que seleccionarlos con el ratón en el cuadro de resultados, presionar *<Ctrl>+<c>*, ir al documento Word y pegar los resultados con *<Ctrl>+<v>*. Cuando se copia una salida de R Commander a un editor de texto, conviene usar un tipo de letra monoespaciada (esto es, un tipo de letra en el que todos los caracteres ocupan la misma anchura) para que las tablas no se desconfiguren. El tipo *Courier New*, en el que están escritas las tablas de este documento, es monoespaciado.

RECOMENDACIÓN: Cuando redactéis un informe estadístico completo, no es recomendable copiar directamente la salida numérica que proporciona Remdr. Es mejor que seleccionéis, de toda esa información, la que es relevante para vuestro informe, y elaboréis vuestras propias tablas. Conviene que incluyáis los títulos de las columnas en español, el título de la tabla, un pie de tabla con explicaciones, si es pertinente, etcétera.

Gráficos de sectores y de barras

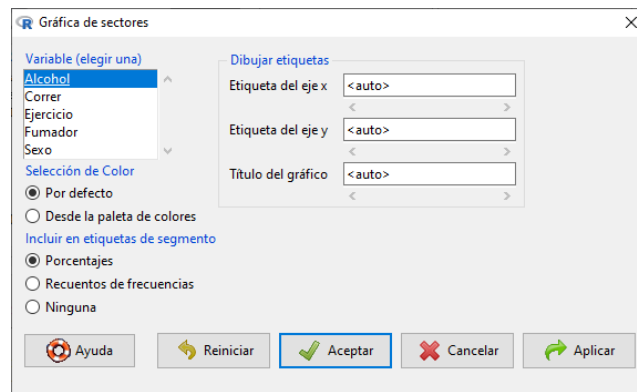
Cuando se está describiendo una variable cualitativa, puede ser conveniente añadir un gráfico, además de la tabla de frecuencias. El gráfico de sectores es uno de los más usados cuando se describe una variable de tipo nominal con pocas modalidades distintas, mientras que el de barras proporciona más información cuando la variable es de tipo ordinal.

Gráfico de sectores

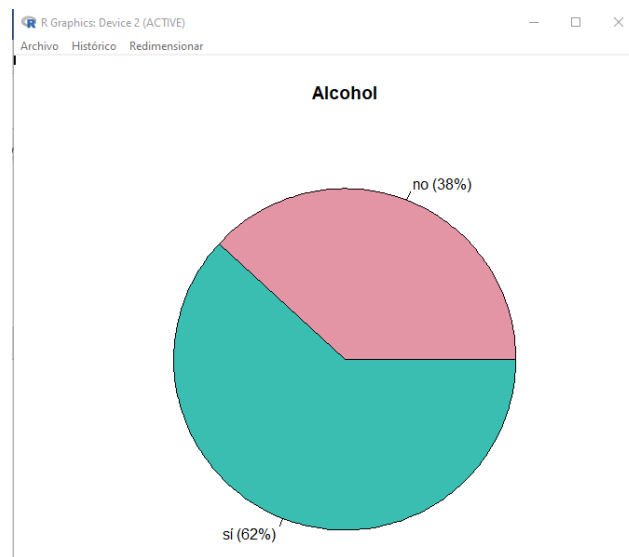
Para obtener el gráfico de sectores de la variable *Alcohol*, los pasos que seguiríais son:

- 1) Selecciona en el menú principal *Gráficas > Gráfica de sectores*.

2) Selecciona la variable *Alcohol* en el cuadro de diálogo que aparece.

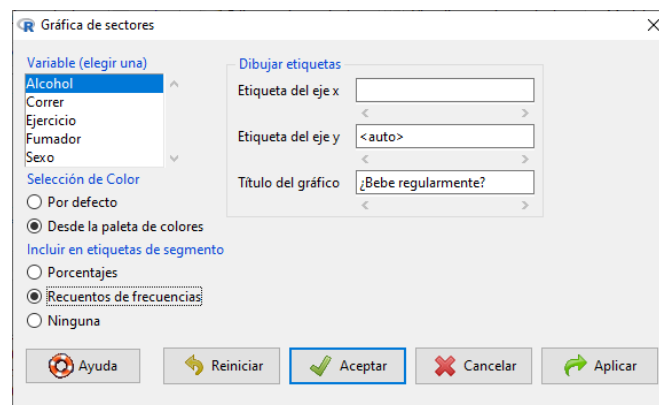


3) Tras pulsar *Aceptar*, aparecerá una nueva ventana (*R Graphics: Device 2*) con el gráfico de sectores:

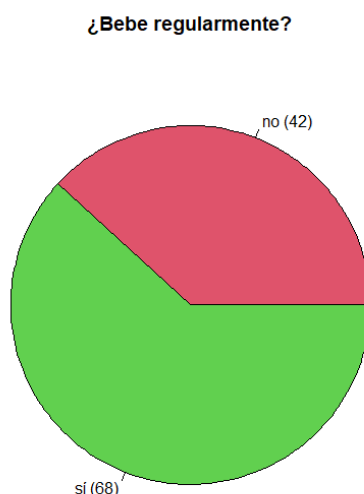


Las opciones “por defecto” se pueden modificar en el cuadro de diálogo. Se pueden cambiar los colores de los sectores con la opción *Desde la paleta de colores*. Además, podemos elegir que nos muestre las frecuencias absolutas asociadas a las diferentes categorías de la variable en lugar de los porcentajes, o que no muestre nada.

Por ejemplo, si completamos el cuadro de diálogo como se muestra a continuación,



el gráfico de sectores que se obtiene es el que se muestra más abajo, donde aparecen las frecuencias absolutas en lugar de los porcentajes, y donde los sectores tienen los colores rojo y verde, pues son los dos primeros definidos en la *Paleta de colores*.

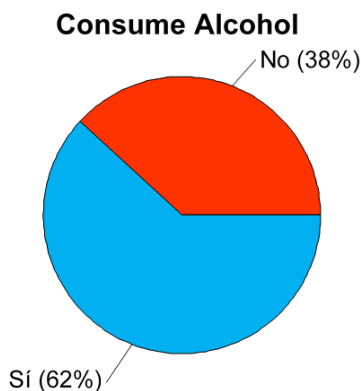


NOTA 1.2: Los colores de la paleta se pueden cambiar con la *Gráficas > Gama de colores*. R ofrece también una amplia gama de herramientas y comandos, así como paquetes específicos (por ejemplo, *ggplot2*) para generar cualquier tipo de gráfico y añadir sobre él casi cualquier tipo de leyenda, rótulo o texto. El estudio de estas posibilidades está más allá de los objetivos de nuestro curso. Os aconsejamos que utilizéis los gráficos por defecto o con alguna de las opciones de las ventanas de R Commander. En vuestros informes también podréis usar las utilidades que ofrece Word para el manejo de gráficos, si queréis modificar su tamaño o apariencia.

La ventana de gráficos sólo admite un gráfico activo. Por tanto, si queremos conservar un gráfico debemos copiarlo o guardarlo, pues al realizar el siguiente gráfico el anterior se pierde. Para copiarlo desde la ventana de gráficos a un documento Word lo más sencillo es utilizar el botón derecho del ratón, que permite copiar el gráfico como metaarchivo (*metafile*) o como dibujo (*bitmap*). Una vez copiado, podemos pegarlo utilizando `<Ctrl>+<v>` o *Archivo > Pegar*.

Si copiáis el gráfico como *bitmap*, observaréis que ese gráfico no se puede editar en Word. Si no necesitáis editar el gráfico es la opción más conveniente. Si necesitáis editarlo tenéis que copiarlo como *metafile*. En ambos casos podéis modificar el tamaño del gráfico en vuestro informe picando sobre el gráfico con el botón izquierdo del ratón para seleccionarlo; después podéis picar en el botón de la parte inferior derecha del gráfico y, sin levantar el dedo, moverlo hasta obtener el tamaño deseado. La opción *Tamaño y posición* del menú que se obtiene presionando el gráfico con el botón derecho del ratón permite modificar el tamaño de forma más precisa.

Si lo habéis copiado en Word como *metafile*, podéis elegir la opción *Editar o Modificar imagen* en el menú contextual que se obtiene presionando el gráfico con el botón derecho del ratón. A continuación, se muestra el gráfico de sectores obtenido tras realizar cambios (colores, tamaños de letra, etiquetas ...) usando la opción de *Editar*.



NOTA 1.3: Además de copiar el gráfico, R Commander también permite guardarlo en diferentes formatos (pdf, metaarchivo, png, jpeg, ...) para poder utilizarlo posteriormente con otras aplicaciones. Para guardar un gráfico, hay que utilizar la opción *Archivo > Guardar como*, disponible en la ventana de gráficos o la opción *Guardar* del menú contextual que se abre al picar en el gráfico de la ventana gráfica de R Commander con el botón derecho del ratón.

Gráfico de barras

La variable *Ejercicio* del conjunto de datos *pulso* representa la regularidad con la que un estudiante realiza ejercicio físico (de forma baja, moderada o alta). La opción *Estadísticos > Resúmenes > Distribución de frecuencias* proporciona la tabla:

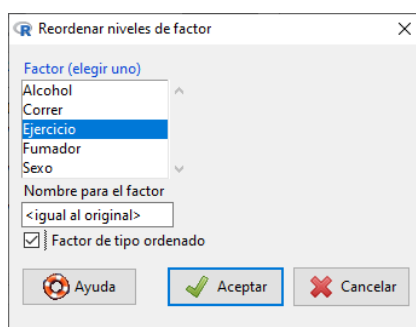
```
counts:
Ejercicio
  alta      baja moderada
    14      37      59

percentages:
Ejercicio
  alta      baja moderada
 12.73    33.64    53.64
```

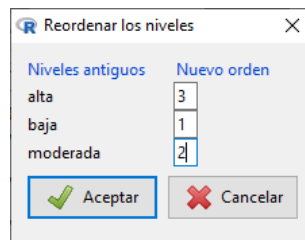
Observa que el orden en el que aparecen las tres categorías es alfabético (alta, baja, moderada), que no es el orden natural si queremos ordenarlas de menor a mayor intensidad (baja, moderada, alta).

Para cambiar el orden de las categorías de una variable de tipo factor usaremos la opción en R Commander llamada *Reordenar niveles de factor*.

- 1) Selecciona *Datos > Modificar variables del conjunto de datos activo > Reordenar niveles de factor*.
- 2) En el cuadro de diálogo que aparece, elige la variable *Ejercicio*, selecciona la opción *Factor de tipo ordenado* y pulsa *Aceptar*. Hemos dejado *<igual al original>* en el campo *Nombre para el factor*, para que sobrescriba la variable *Ejercicio* en lugar de crear una nueva variable. Al seleccionar la opción *Factor de tipo ordenado* le indicamos a R Commander que la variable *Ejercicio* es de tipo ordinal.



- 3) Tras confirmar que deseamos reescribir la variable *Ejercicio*, se puede establecer el nuevo orden de las categorías de la forma: baja => 1, moderada => 2 y alta =>3, utilizando la siguiente ventana:



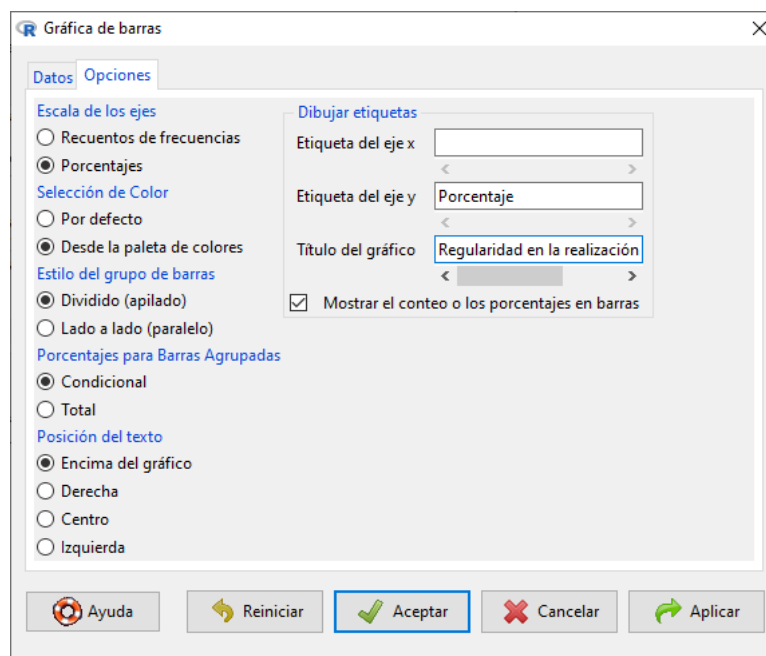
Tras pulsar *Aceptar*, ya se tendrá establecido el nuevo orden de las categorías para la variable *Ejercicio*. Si volvemos a ejecutar *Estadísticos > Resúmenes > Distribución de frecuencias* obtendremos

```
counts:
Ejercicio
  baja moderada  alta
    37      59    14

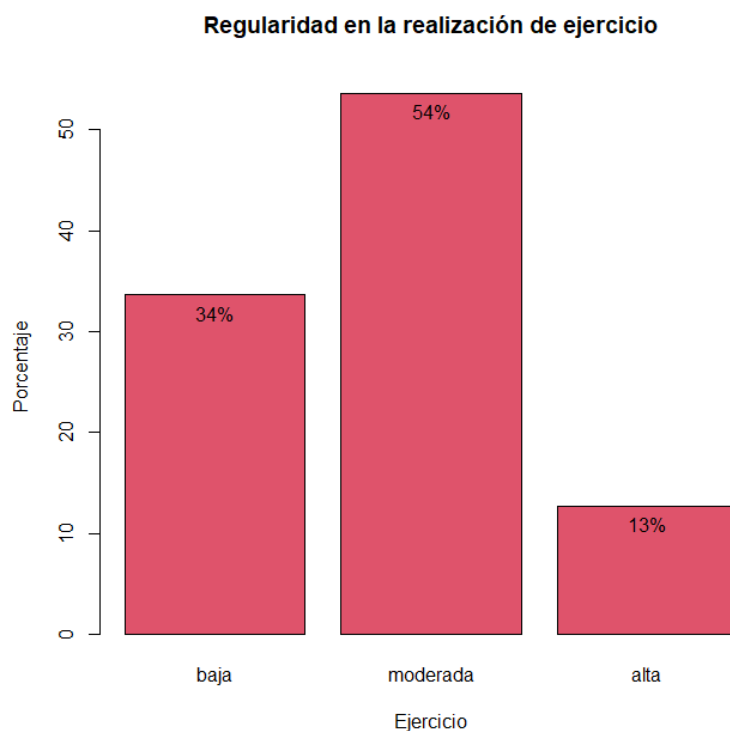
percentages:
Ejercicio
  baja moderada  alta
 33.64   53.64  12.73
```

Para obtener el diagrama de barras de esta variable ordinal haremos lo siguiente:

- 1) Selecciona en el menú principal *Gráficas > Gráfica de barras*.
- 2) Selecciona la variable *Ejercicio* en la pestaña *Datos*. En la pestaña de *Opciones* selecciona, en la *Escala de los ejes*, *Porcentajes* y en *Color*, *Desde la paleta de colores*. Modifica las etiquetas como se indica a continuación:



- 3) Tras pulsar *Aceptar*, te aparecerá en una nueva ventana el gráfico de barras de la variable *Ejercicio*.



Utilizando la tabla de frecuencias y el diagrama de barras podemos describir la variable *Ejercicio* diciendo que, **de los 110 estudiantes de la muestra, 37 realizan poco ejercicio (un 33.6%), 59 lo realizan de forma moderada (un 53.6%) y 14 lo realizan con una alta regularidad (un 12.8%). La distribución de estos porcentajes puede verse en el diagrama de barras.**

Tabla de frecuencias acumuladas

R Commander no proporciona una opción que permita obtener tablas de frecuencias acumuladas. Si necesitamos esas tablas las tendremos que crear por nuestra cuenta. La opción más sencilla es utilizar Excel, pero también se puede hacer con R Commander, usando un “*script*” R, que es un “programa” con una lista de órdenes, sentencias y comandos de R que se ejecutan secuencialmente y permiten obtener una salida determinada, en este caso una tabla de frecuencias acumuladas.

La construcción de “scripts” está más allá de los objetivos de este curso, pero vamos a emplear este ejemplo para ilustrar cómo son y cómo se pueden utilizar. Definiremos en primer lugar una función, llamada *Acumuladas* con dos argumentos: el nombre del conjunto de datos (*base*) y la variable de tipo factor contenida en ese conjunto de datos de la que queremos obtener tablas de frecuencias acumuladas (*variable*). La definición de dicha función figura a continuación.

```
Acumuladas <- function(base, variable) {
  Frecuencia<-eval(parse(text=paste("with(",base,",", table(",variable,")"))))
  Porcentaje <- round(100*Frecuencia/sum(Frecuencia), 3)
  F_acumulada <- cumsum(Frecuencia)
  P_acumulado <- round(100*F_acumulada/sum(Frecuencia), 3)
  tabla <- cbind(Frecuencia, Porcentaje, F_acumulada, P_acumulado)
  cat("\nTablas de frecuencias de la variable", variable,"\n\n")
  tabla}
```

Para poder utilizar esta función, la tenemos que ejecutar en el panel (o cuadro) de sintaxis, *R Script*. La podemos copiar de este mismo documento utilizando <Ctrl>+<c> y pegarla en *R Script* con <Ctrl>+<v>. También la podemos copiar del fichero de texto *Script acumuladas.txt*, o del archivo de instrucciones R *Acumuladas.R*. disponibles en esta práctica. Estos dos ficheros se pueden abrir utilizando el accesorio de Windows *Bloc de notas*.

Una vez ejecutado el script anterior, si ejecutamos *Acumuladas("pulso","Ejercicio")* (observa que el nombre de la base de datos y el nombre de la variable van entre dobles comillas) obtendremos la siguiente salida.

```

> Acumuladas <- function(base, variable) {
+ Frecuencia<-eval(parse(text=paste("with(",base,",", table(",variable,")"))))
+ Porcentaje <- round(100*Frecuencia/sum(Frecuencia), 3)
+ F_acumulada <- cumsum(Frecuencia)
+ P_acumulado <- round(100*F_acumulada/sum(Frecuencia), 3)
+ tabla <- cbind(Frecuencia, Porcentaje, F_acumulada, P_acumulado)
+ cat("\nTablas de frecuencias de la variable", variable,"\n\n")
+ tabla}

> Acumuladas("pulso","Ejercicio")

Tablas de frecuencias de la variable Ejercicio

```

	Frecuencia	Porcentaje	F_acumulada	P_acumulado
baja	37	33.636	37	33.636
moderada	59	53.636	96	87.273
alta	14	12.727	110	100.000

La base de datos, en este caso *pulso*, se tiene que haber cargado previamente en la sesión, aunque no es imprescindible que sea la base de datos activa.

NOTA 1.4: El script anterior se puede cargar directamente usando la opción *Fichero > Abrir archivo de instrucciones* y seleccionar *Acumuladas.R*.

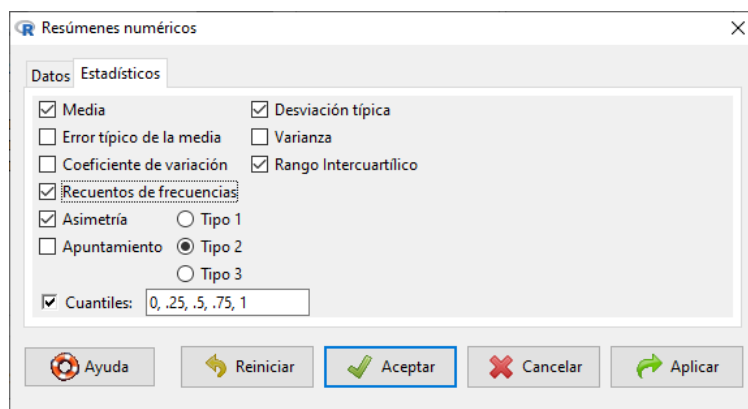
2. Estudio descriptivo de variables discretas

Las variables discretas son variables cuantitativas, para las que tiene sentido calcular las medidas descriptivas que usan operaciones aritméticas, como la media o la desviación típica. La tabla de frecuencias y el gráfico de líneas, que es un gráfico similar al gráfico de barras, también son útiles para describir su distribución.

Medidas descriptivas y tabla de frecuencias

La variable *Hijos*, que se encuentra en el archivo *hijos.RData*, contiene el número de hijos de 300 familias, es decir, contiene 300 casos de esa variable numérica discreta. Para obtener los estadísticos habituales, los pasos que seguiríais son:

- 1) Desde R Commander, carga el archivo *hijos.RData*.
- 2) Selecciona en el menú principal *Estadísticos > Resúmenes > Resúmenes numéricos*.
- 3) En la pestaña *Datos* del cuadro de diálogo que aparece, elige la variable *Hijos*.



- 4) En la pestaña *Estadísticos*, puedes elegir qué medidas descriptivas (estadísticos) quieres obtener. Por defecto, calcula la media de la variable, su desviación típica, el rango intercuartílico, los cuartiles (cuantiles 0.25, 0.5 y 0.75), el mínimo (cuantil 0) y el máximo (cuantil 1). Hemos seleccionado también *Recuento de frecuencias*, que proporciona una tabla de frecuencias de esta variable, y *Asimetría*, que proporciona el coeficiente de asimetría.

Al pulsar *Aceptar*, obtendremos los siguientes resultados.

```
> numSummary(hijos[,"Hijos", drop=FALSE], statistics=c("mean", "sd",
+ "quantiles","skewness"), quantiles=c(0,.25,.5,.75,1), type="2")
  mean      sd IQR skewness 0% 25% 50% 75% 100%  n
2.246667 1.60032  2 0.7252661  0  1  2  3    8 300

> discreteCounts(hijos[,"Hijos", drop=FALSE])
Distribution of Hijos
  Count Percent
0      37  12.33
1      76  25.33
2      68  22.67
3      53  17.67
4      42  14.00
5      15   5.00
6       6   2.00
8       3   1.00
Total   300 100.00
```

Los nombres de los estadísticos vienen en inglés (*mean* = media, *sd* = desviación típica, *IQR* = rango intercuartílico, *skewness* = coeficiente de asimetría). El percentil 0% es el mínimo (0) y el percentil 100% es el máximo (8). La tabla de frecuencias está bajo el título *Distribution of Hijos*.

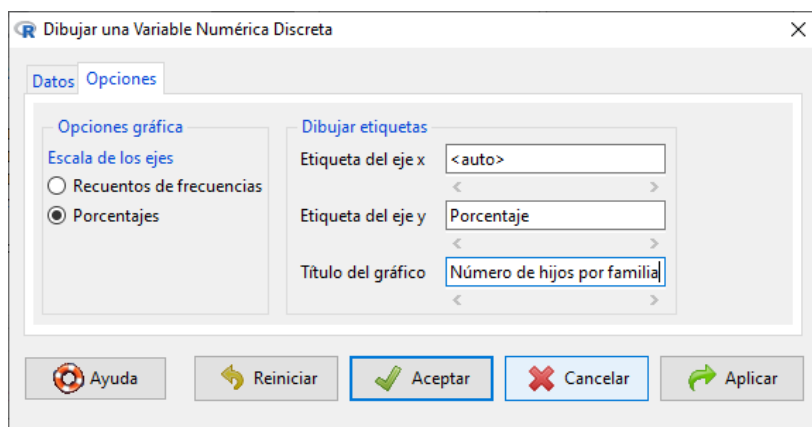
Ejercicio 2.1: Usando la tabla de frecuencias anterior, indica cuál es la moda de esta variable y qué porcentaje de familias tienen 3 o más hijos.

Gráfico de líneas y gráfico de barras

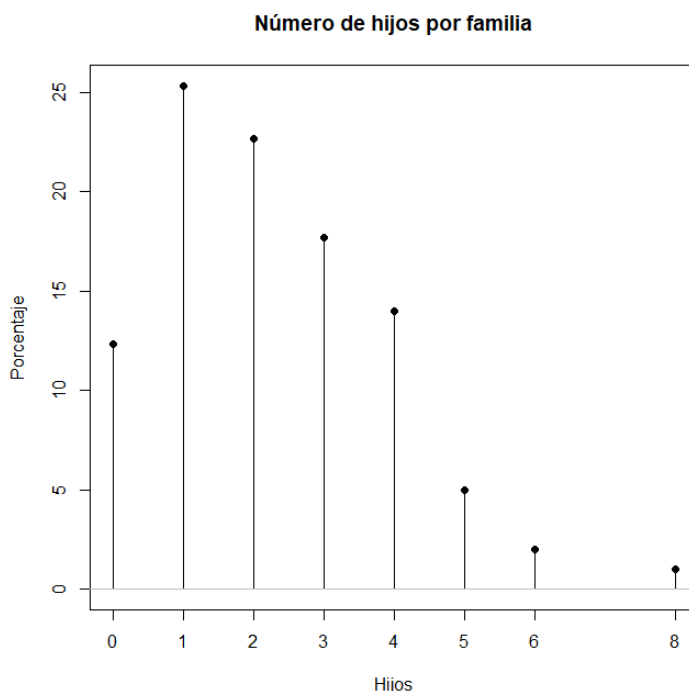
El gráfico de líneas es el que, por defecto, ofrece R Commander para describir variables numéricas discretas. Resulta particularmente útil si esta variable discreta tiene pocas categorías distintas. El gráfico de líneas es similar al diagrama de barras. Para cada modalidad de la variable, dibuja una línea vertical, y la altura de esta línea es proporcional a la frecuencia absoluta de esa modalidad en la muestra, y por lo tanto, también a la frecuencia relativa o al porcentaje.

Para la variable *Hijos*, este gráfico se obtiene con el siguiente procedimiento.

- 1) Con el conjunto de datos *hijos* activo, selecciona en el menú principal *Gráficas > Dibujar una variable numérica discreta*.
- 2) Selecciona la variable *Hijos* y rellena las opciones como se muestra en la siguiente ventana, donde estamos seleccionando que el eje y represente porcentajes y donde añadimos una al eje y un título al gráfico.

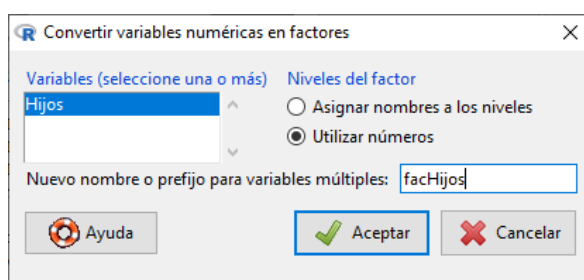


Tras pulsar *Aceptar*, obtendrás el siguiente gráfico, donde para cada posible valor de la variable *Hijos*, la altura de línea representa el porcentaje de familias con ese número de hijos.



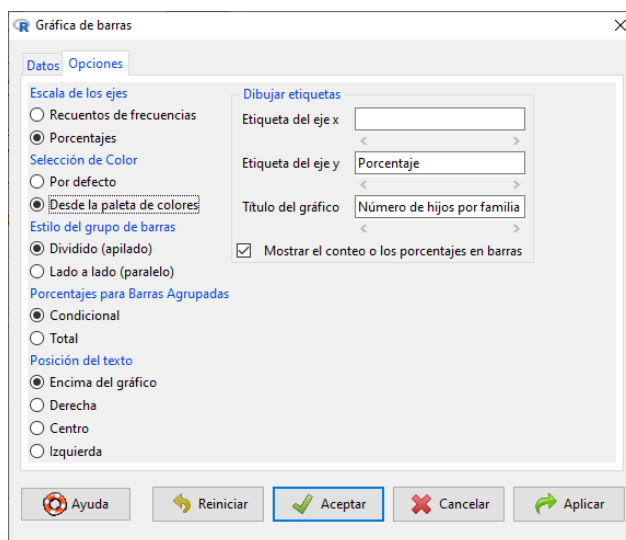
Esta misma información se podría representar también con un diagrama de barras, como el que ya hemos utilizado para describir variables categóricas ordinales, pero R Commander sólo permite obtener gráficos de barras para variables de tipo factor. Como la variable *Hijos* es de tipo numérico, si queremos obtener el gráfico de barras tendremos que construir una variable auxiliar de tipo factor en la que el número de hijos sea una “etiqueta”. Esto lo puedes realizar de la siguiente manera.

- 1) Selecciona *Datos > Modificar variables del conjunto de datos activo > Convertir variable numérica en factor*.
- 2) Rellena el cuadro de diálogo de la forma que se muestra a continuación y pulsa *Aceptar*. Al hacerlo estamos creando una nueva variable de tipo factor llamada *facHijos*, que contendrá las etiquetas “0”, “1”, “2”, ... como sus niveles o categorías. Al llamar *facHijos* a esta variable, no sobrescribe la variable *Hijos* de la base de datos.

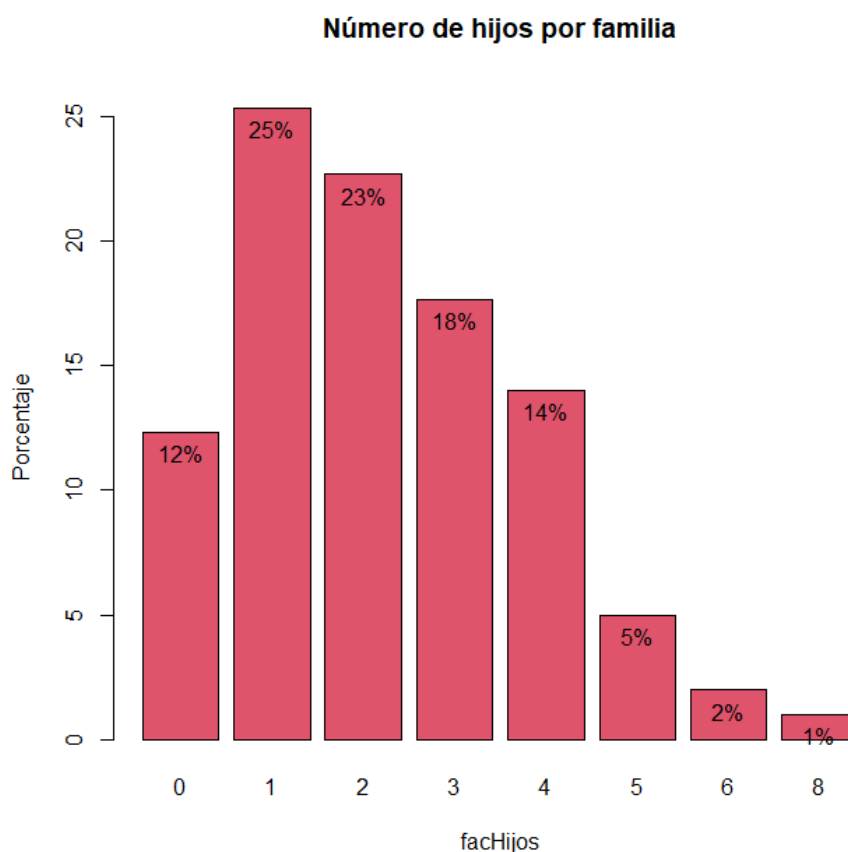


Si visualizamos cómo nos ha quedado el conjunto de datos, veremos que tenemos las dos variables, aparentemente iguales. Sin embargo, una es de tipo numérico (*Hijos*) y la otra de tipo factor (*facHijos*). Para la de tipo numérico, *Hijos*, podremos obtener medidas numéricas (media, máximo, mínimo, ...) pero no un gráfico de barras. Para la de tipo factor, *facHijos*, podremos obtener el gráfico de barras, pero no medidas numéricas.

Para obtener el gráfico de barras, utilizaremos la opción *Gráficas > Gráfica de barras* para la variable *facHijos*. Si seleccionamos las opciones tal y como aparecen en la siguiente ventana



obtendremos el gráfico de barras que se muestra a continuación.



Observa que el gráfico de barras sólo representa las categorías con frecuencia positiva, y por eso **no ha representado el 7**, ya que en la muestra no hay ninguna familia con 7 hijos. El gráfico de líneas, sin embargo, tiene en cuenta que la variable representada es numérica y ha dejado el hueco del 7, respetando la escala natural de la variable.

A modo de resumen, con la información proporcionada por las medidas descriptivas, la tabla de frecuencias y los gráficos, podríamos decir, por ejemplo, lo siguiente:

Entre las 300 familias la muestra el número máximo de hijos es 8 mientras que el mínimo es 0. En promedio, cada familia tiene 2.25 hijos, siendo la desviación típica 1.60 hijos. Al menos la mitad de las familias tienen como

mucho 2 hijos y al menos la mitad de las familias tienen 2 hijos o más. El 25.33% de las familias tiene un único hijo, siendo esta clase la mayoritaria (moda) de entre todas las familias estudiadas. El 39,67% de las familias estudiadas son numerosas (tienen tres hijos o más).

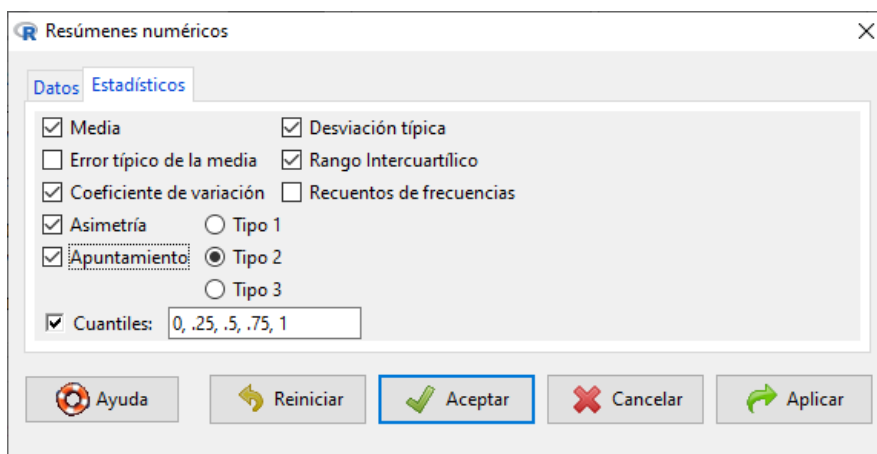
Cuando una variable discreta toma muchos valores distintos, las tablas de frecuencias y el gráfico de líneas ya no son herramientas útiles para describir la variable. En ese caso, la variable discreta se suele estudiar descriptivamente como si fuese una variable de tipo continuo. Las herramientas para analizar variables continuas se van a describir a continuación.

3. Estudio descriptivo de variables continuas

Las herramientas básicas para describir una variable de tipo continuo son las medidas descriptivas habituales (estadísticos de posición, dispersión y forma). R Commander proporciona también, en el menú *Gráficas*, muchas opciones gráficas para describir la distribución de variables continuas: diagrama de puntos, histograma, estimar densidad, diagrama de tallos y hojas, y diagrama de caja.

Medidas descriptivas

Vamos a estudiar la variable *Pulso1* del archivo *Pulso.RData*, que recoge el pulso en reposo de una muestra de 110 estudiantes. Las medidas descriptivas se obtienen como se ha explicado en la sección anterior, a través de la opción *Estadísticos > Resúmenes > Resúmenes numéricos*. En este caso, conviene añadir en la salida el coeficiente de variación (como medida de dispersión relativa) y las medidas de forma (asimetría y apuntamiento). R Commander tiene tres opciones para el cálculo de los coeficientes de asimetría y apuntamiento. Se recomienda dejar Tipo 2, que es la opción por defecto.



La salida que se obtiene se muestra a continuación.

```
> numSummary(pulso[,"Pulso1", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles",
+ "cv", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2")
  mean      sd IQR      cv skewness kurtosis 0% 25% 50% 75% 100%  n NA
75.68807 13.29766 14 0.1756903 1.511977 6.712752 47 68 76 82 145 109 1
```

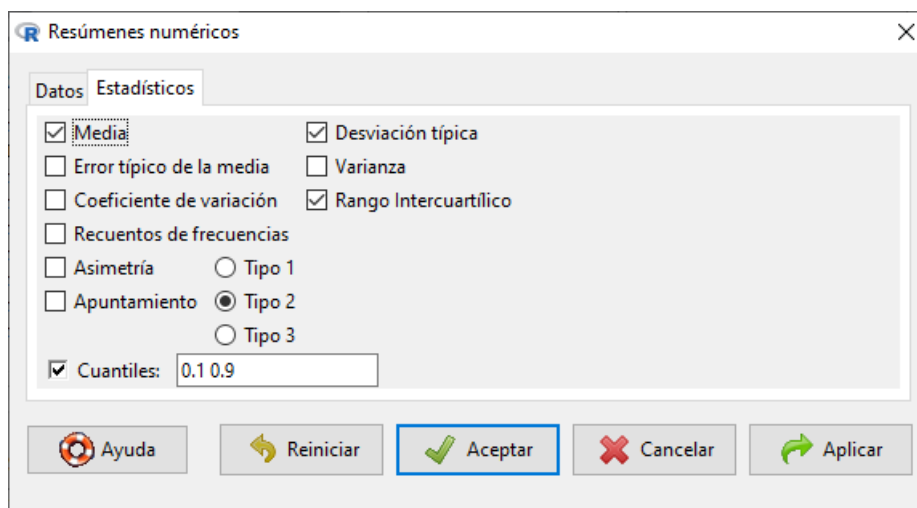
Estos estadísticos nos permiten hacer una primera descripción de la variable, en la que hemos redondeado los estadísticos a dos decimales: **El pulso medio de los individuos de esta muestra es de 75,69 pulsaciones por minuto (ppm), con una desviación típica de 13,30 ppm. La mediana, 76 ppm, es casi igual que la media, lo que suele ser indicio de simetría. Pero en este caso, el coeficiente de asimetría es 1.51, positivo y grande, lo que indica asimetría hacia la derecha. El rango de variación, 98 ppm, también es grande si lo comparamos con la media. Los datos registrados van desde un mínimo de 47 hasta un máximo de 145 ppm. La cuarta parte de los individuos de la muestra no superan las 69 ppm, la mitad no superan las 76 ppm, pero al menos una cuarta parte han registrado 82 ppm o más. El coeficiente de apuntamiento, 6.71, es positivo y grande, lo que indica que la distribución de esta variable tiene mayor apuntamiento que la normal.** Las consideraciones sobre la forma de la distribución (en particular, sobre la

simetría y el apuntamiento) se tendrán que completar con el análisis gráfico. Veremos más adelante que estos valores elevados de asimetría y de apuntamiento están muy influidos por valores erróneos de la variable.

NOTA 3.1: Los coeficientes de asimetría y de apuntamiento no tienen unidades (son adimensionales) y, por lo tanto, se puede evaluar si son “suficientemente grandes” sin tener en cuenta las unidades de la variable, teniendo en cuenta solamente el tamaño de la muestra. Utilizaremos la siguiente regla empírica:

- Si el coeficiente de asimetría se encuentra fuera del intervalo $(-2 * \sqrt{6/\sqrt{n}}, 2 * \sqrt{6/\sqrt{n}})$, la variable se puede considerar claramente asimétrica. Si el valor es positivo, la simetría es a la derecha, y si es negativo, la asimetría es a la izquierda.
- Si el coeficiente de apuntamiento se encuentra fuera del intervalo $(-4 * \sqrt{6/\sqrt{n}}, 4 * \sqrt{6/\sqrt{n}})$, se puede considerar que la variable tiene distinto apuntamiento que la normal. Si el valor es positivo, la variable es más apuntada, y si es negativo, es menos apuntada.

Podemos utilizar la opción *Resúmenes numéricos* de R Commander para calcular otros cuantiles, en lugar de los que proporciona por defecto. Por ejemplo, si queremos obtener los percentiles 10 y 90 de *Pulsol*, que son los cuantiles 0.1 y 0.9, bastará con incluir 0.1 y 0.9 en el campo *Cuantiles*.



Al Aceptar, obtendremos la siguiente salida, que nos dice que el percentil 10 es 61.8 ppm y el 90 es 88.4 ppm.

```
> numSummary(pulso[, "Pulsol", drop=FALSE], statistics=c("mean", "sd", "quantiles", "CV"), quantiles=c(0.1, 0.9))
  mean      sd IQR  10%  90%   n NA
75.68807 13.29766 14 61.8 88.4 109 1
```

Cuando la variable toma muchas modalidades distintas, la opción *Recuento de frecuencias* del cuadro de diálogo anterior crea una serie de intervalos de clase para esa variable, y realiza una tabla de frecuencias para esos intervalos de clase. Si hubiésemos marcado esa opción en el cuadro de diálogo anterior, habríamos obtenido la siguiente tabla de frecuencias, donde los intervalos creados comienzan en 40 y tienen amplitud 10.

```
> binnedCounts(pulso[, "Pulsol", drop=FALSE])
Binned distribution of Pulsol
      Count Percent
 [40, 50]      3   2.75
 (50, 60]      7   6.42
 (60, 70]     30  27.52
 (70, 80]     40  36.70
 (80, 90]     22  20.18
 (90, 100]     3   2.75
 (100, 110]    2   1.83
 (110, 120]    1   0.92
 (120, 130]    0   0.00
 (130, 140]    0   0.00
 (140, 150]    1   0.92
 Total       109  99.99
```

El algoritmo que ha utilizado para seleccionar estos intervalos de clase es el mismo que utilizará, por defecto, para calcular los intervalos de clase cuando le pidamos un histograma.

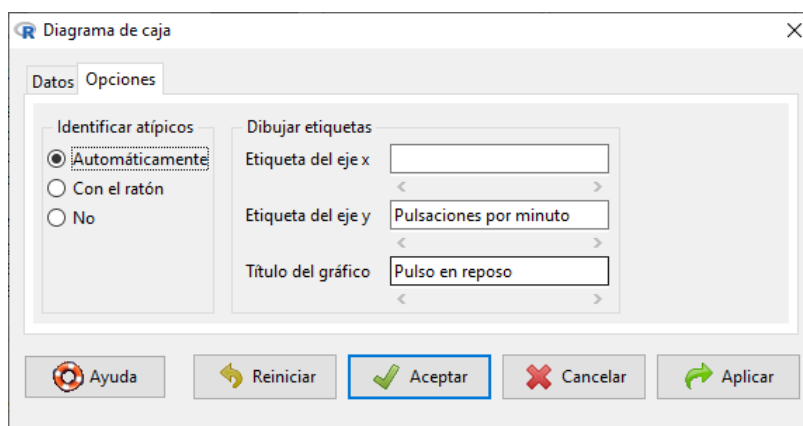
Gráficos para variables continuas

La información sobre la distribución de la variable que proporcionan los estadísticos descriptivos se puede ilustrar, contrastar o completar utilizando gráficos. En el menú *Gráficas*, R Commander proporciona varias opciones: diagrama de puntos, histograma, estimar densidad, diagrama de tallos y hojas y diagrama de caja. Es conveniente empezar obteniendo el diagrama de caja, que muestra gráficamente la relación entre los cuartiles de la variable y es además una buena herramienta para descubrir datos atípicos.

Diagrama de caja

Para la obtención del diagrama de caja de la variable *Pulsol*, tendrías que:

- 1) Selecciona en el menú principal *Gráficas > Diagrama de caja*.
- 2) En la pestaña *Datos* del cuadro de diálogo que aparece, elige la variable *Pulsol*.
- 3) En la pestaña *Opciones*, puedes elegir la identificación automática de los datos atípicos, la identificación manual con el ratón o que no indique las etiquetas de los datos atípicos. Se recomienda dejar la opción *Automáticamente*. También se pueden escribir etiquetas para los ejes y un título para el gráfico.



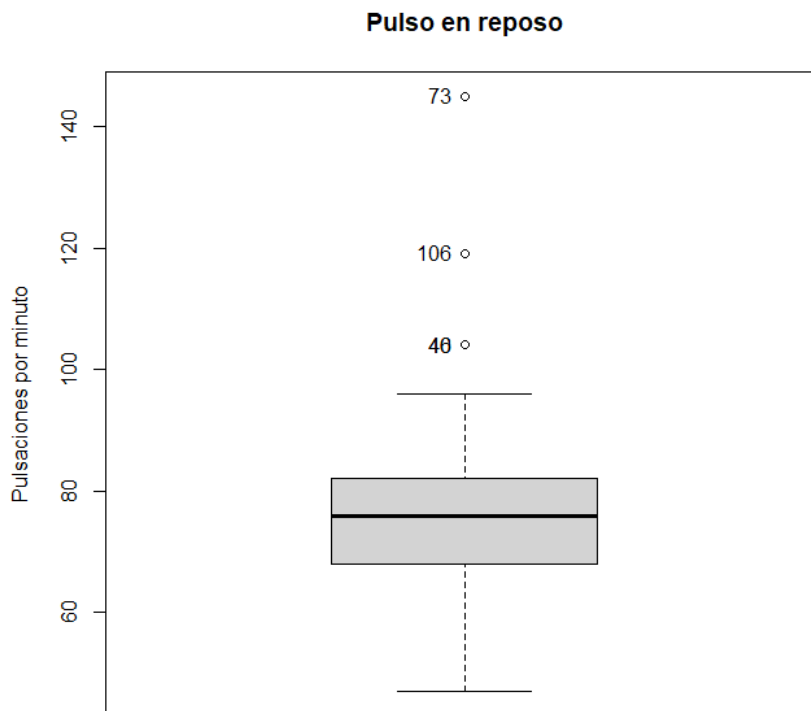
Al pulsar *Aceptar* con las opciones del cuadro de diálogo anterior, obtendremos el diagrama de caja.

Los estudiantes con etiquetas “40”, “46”, “73” y “106” son identificados como casos atípicos, como puede verse en el cuadro de resultados *Salida*:

```
> Boxplot( ~ Pulsol, data=pulso, id=list(method="y"), ylab="Pulsaciones por minuto",
+ main="Pulso en reposo")
```

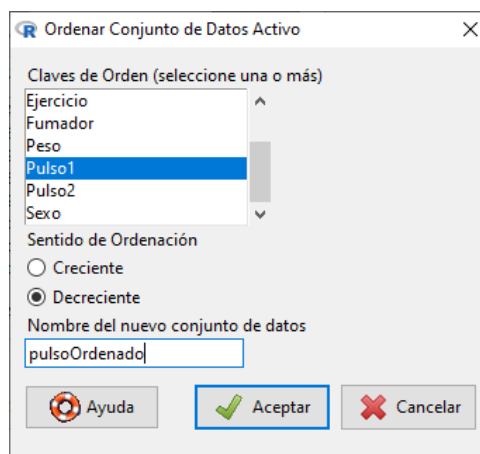
[1] "40" "46" "73" "106"

Las etiquetas de los casos atípicos también aparecen en el gráfico, aunque puede verse que las etiquetas “40” y “46” están solapadas, ya que corresponden a dos individuos con las mismas ppm, en este caso 104.



Los datos atípicos suelen estar en los extremos de la distribución. Para identificarlos sería útil disponer de los datos ordenados de la variable. Esto se puede hacer con la opción *Datos > Conjunto de datos activo > Ordenar el conjunto de datos activo*.

Si usamos esa opción, y rellenamos su cuadro de diálogo como se muestra a continuación, crearemos un nuevo conjunto de datos, llamado *pulsoOrdenado*, donde los casos del conjunto de datos *pulso* aparecerán ordenados de forma decreciente con respecto a la variable *Pulso1*.



Usando el editor de datos con *pulsoOrdenado*, veremos que en las primeras filas aparecen los estudiantes que tienen los valores de *Pulso1* más grandes (que corresponde a los datos atípicos en la parte de arriba del diagrama de caja).

Podemos comprobar que los casos con etiquetas (*rowname*) 40, 46, 73 y 106 se corresponden con los valores 104, 104, 145 y 119 de *Pulso1*, respectivamente. Las etiquetas de *rowname* indican cual era el orden original de estos casos en el conjunto de datos *pulso*.

	1	2	3	4	5	6	7	8	9	10	11	
	rowname	Altura	Peso	Edad	Sexo	Fumador	Alcohol	Ejercicio	Correr	Pulso1	Pulso2	Año
1	73	179	80.0	20	hombre	no	sí	moderada	sí	145	155	97
2	106	93	27.0	19	mujer	no	no	baja	no	119	120	98
3	40	155	49.0	18	mujer	no	sí	moderada	no	104	92	95
4	46	164	46.0	18	mujer	no	no	moderada	no	104	96	95
5	3	167	62.0	18	mujer	no	sí	alta	sí	96	176	93
6	50	167	70.0	22	hombre	sí	sí	baja	no	92	84	96
7	103	170	63.0	20	mujer	no	sí	baja	sí	92	120	98
8	5	173	64.0	18	mujer	no	sí	baja	no	90	88	93
9	89	162	50.0	19	mujer	no	sí	moderada	sí	90	160	97
10	94	170	58.0	21	hombre	sí	sí	moderada	no	90	84	98
11	107	161	43.0	19	mujer	no	no	baja	no	90	89	98

Depuración de datos

Los valores 104, 119 y 145 de *Pulso1* son muy elevados, por lo que convendría comprobar si son errores.

Si no son errores, pueden proporcionar información interesante sobre los individuos que no siguen la misma pauta que el resto de la muestra para esta variable.

Si son errores, los tendremos que corregir y, de no ser posible recuperar el valor correcto, los tendremos que eliminar del análisis de esta variable. Si la variable es esencial para nuestro estudio, se suele recomendar eliminar el caso.

En este ejercicio asumiremos que son errores, que no los podemos corregir y que esta variable es esencial para nuestro estudio, por lo que los vamos a eliminar de la muestra. El resultado de eliminar estos cuatro casos es el conjunto de datos *pulsoNuevo.RData*.

Ejercicio 3.1: Carga el conjunto de datos *pulsoNuevo.RData*. Obtén las medidas descriptivas de la variable *Pulso1* de este nuevo conjunto de datos y compáralas con las obtenidas para el conjunto de datos *pulso*. Ahora, los valores de los coeficientes de asimetría y apuntamiento son pequeños. La asimetría y apuntamiento que habíamos observado en el primer análisis eran consecuencia de valores erróneos.

```

mean      sd IQR      CV      skewness      kurtosis      0% 25% 50% 75% 100%      n NA
75.68807 13.29766 14 0.1756903 1.511977      6.712752      47 68 76 82 145 109 1 < CON ERRORES
74.07619 10.06775 12 0.1359107 -0.3005143 -0.06616746 47 68 75 80 96 105 1 < SIN ERRORES
    
```

Ejercicio 3.2: Obtén el diagrama de caja de *Pulso1* en este conjunto de datos *pulsoNuevo*. ¿Se detecta algún dato atípico nuevo, después de eliminar los cuatro casos con datos erróneos?

COMENTARIO: Los pasos que tendríamos que haber seguido para eliminar los cuatro datos erróneos y obtener el conjunto de datos *pulsoNuevo.RData*, utilizando el editor de R Commander, son los siguientes:

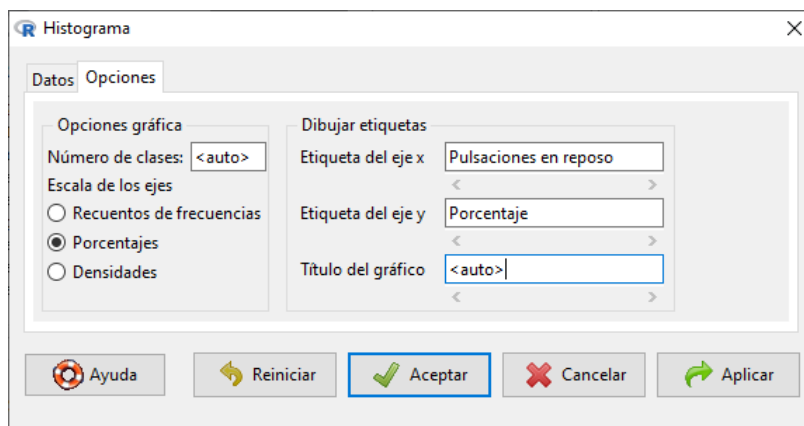
- 1) Seleccionamos *pulso* como *Conjunto de datos activo*.
- 2) Utilizamos el botón *Editar conjunto de datos* para eliminar los datos erróneos. Seleccionaremos las filas que llevan las etiquetas 106, 73, 46 y 40 en la columna *rowname* y las borramos, una a una, usando, en el menú *Editar* la opción *Borrar la fila actual*.
- 3) Guardamos los datos pulsando en el botón *Aceptar*, o en *Fichero > Salir y guardar*.
- 4) Cambiamos el nombre de este conjunto de datos, ejecutando en el cuadro *R Script* la orden *pulsoNuevo=pulso*.
- 5) Seleccionaremos *pulsoNuevo* como *Conjunto de datos activo*.
- 6) Cuando se utiliza el Editor para modificar un conjunto de datos, las variables factor se convierten en variables *character*. Podemos anular este cambio utilizando *Datos > Conjunto de datos activo > Convertir todas las variables de caracteres en factores*. También hará falta volver a ordenar las etiquetas de *Ejercicio*.

- Guardaremos este conjunto de datos en nuestra carpeta con la opción *Datos > Conjunto de datos activo > Guardar el conjunto de datos activo*, con el nombre *pulsoNuevo.RData*.

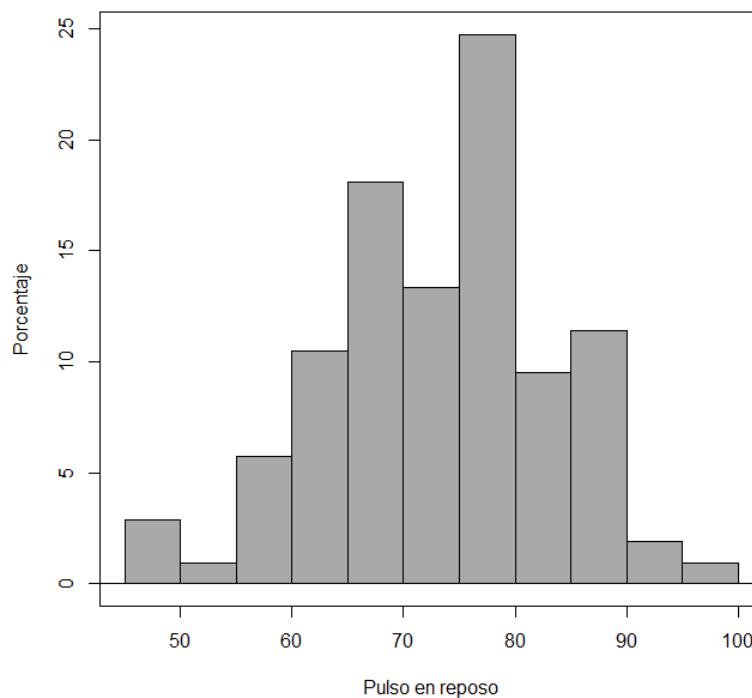
Histograma

Para la obtención del histograma de la variable *Pulso1*, en el conjunto de datos *pulsoNuevo*, tendrás que:

- Seleccionar en el menú principal *Gráficas > Histograma*.
- En la pestaña *Datos* del cuadro de diálogo, elige la variable *Pulso1*.
- En la pestaña *Opciones*, puedes dejar que R Commander calcule el número de intervalos de clase de forma automática (<auto>) o elegir cuántos intervalos quieres para el histograma (que serán todos de igual amplitud). También puedes poner etiquetas en los ejes y un título al gráfico, y puedes elegir la escala en el eje y (frecuencias, porcentajes o densidades).



Al pulsar *Aceptar* con las opciones del cuadro de diálogo anterior, obtendremos el siguiente histograma.



Si te fijas en el cuadro de sintaxis (*R Script*) verás que el orden R que se ejecuta al pedir el histograma desde el menú de *Gráficas* y con estas opciones es:

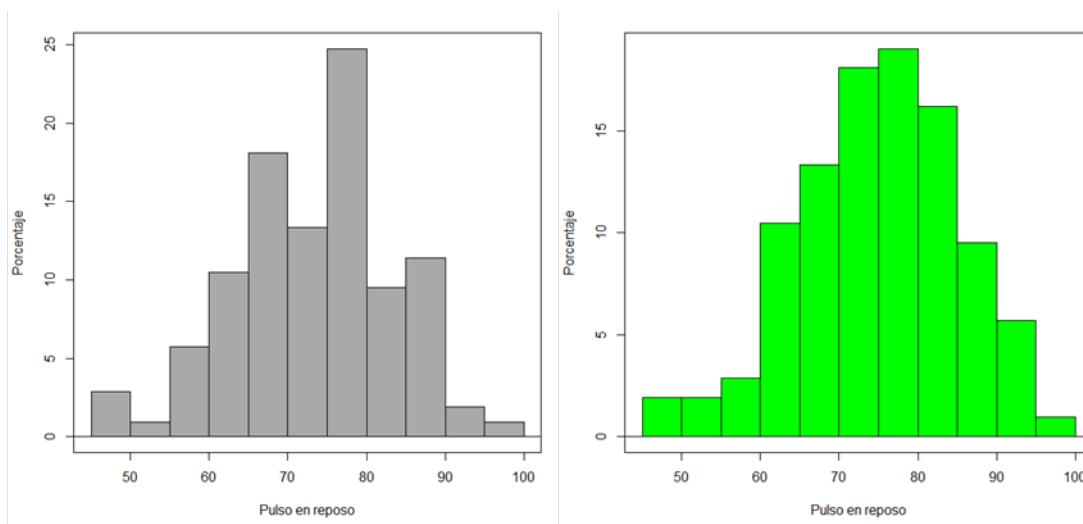
```
with(pulsoNuevo, Hist(Pulso1, scale="percent", breaks="Sturges", col="darkgray",
  xlab="Pulso en reposo", ylab="Porcentaje"))
```

El comando utilizado por R Commander es *Hist*. Detrás del nombre de la variable, *Pulso1*, viene una lista de opciones, que se pueden cambiar.

Por ejemplo, el color por defecto de las barras del histograma es “*darkgray*”. Cambiando esta etiqueta por “*green*” y volviendo a ejecutar la orden desde el cuadro de sintaxis, conseguiríamos que las barras apareciesen de color verde. Los intervalos elegidos por defecto para realizar el histograma son de la forma (a,b]. Si quisiéramos utilizar intervalos de la forma [a,b), habría que añadir en la lista de opciones la opción *right=FALSE*. Con estos dos cambios, la orden quedaría

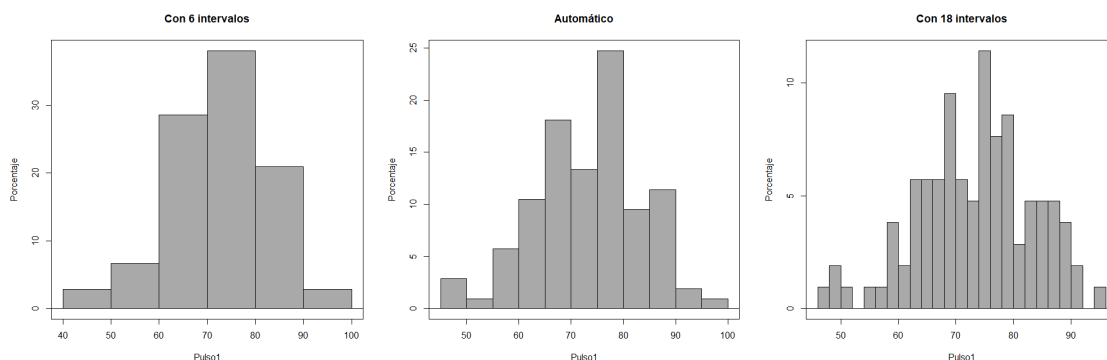
```
with(pulsoNuevo, Hist(Pulso1, scale="percent", breaks="Sturges", col="green",
  xlab="Pulso en reposo", ylab="Porcentaje", right=FALSE))
```

Si la ejecutamos, en *R Script*, obtendremos un nuevo histograma. Lo podemos comparar con el obtenido con las opciones por defecto:



Se puede ver que el histograma cambia si cambiamos el tipo de intervalo. De todas formas, en los dos gráficos se puede observar una cierta simetría y la presencia de un grupo relativamente importante de valores muy bajos, que podrían ser atípicos.

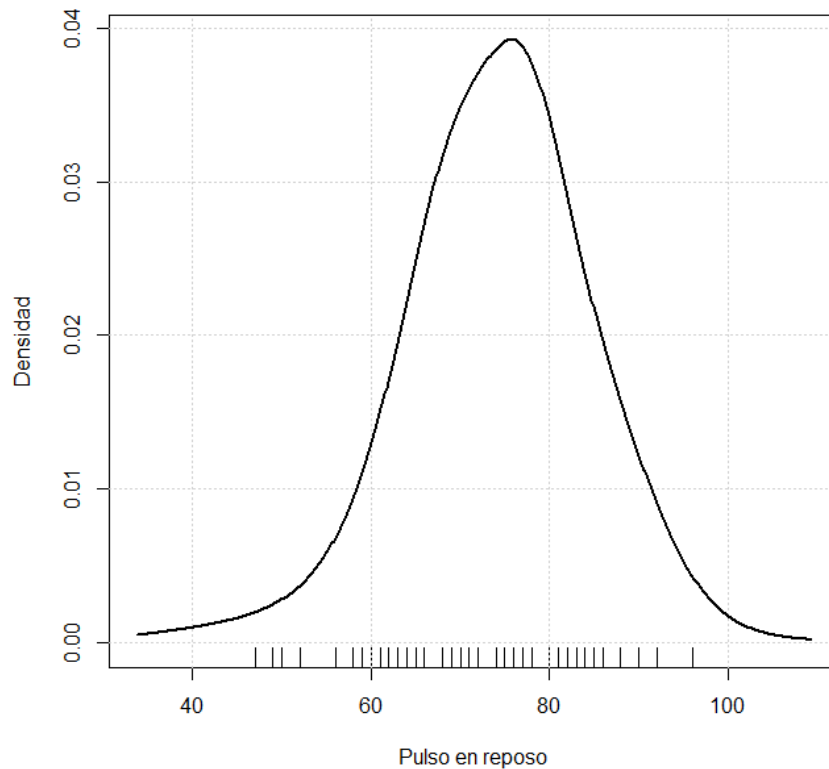
La forma del histograma depende mucho del número de intervalos seleccionado. El algoritmo por defecto (Sturges) propone 11 intervalos para *Pulso1*. Vamos a compararlo a continuación con los histogramas con 6 y con 18 intervalos.



Estimar densidad

El histograma puede ser muy distinto, para los mismos datos, si cambiamos el número y la amplitud de los intervalos. R Commander proporciona un procedimiento alternativo para visualizar la forma de la distribución que no depende de la elección de estos intervalos: *Estimar densidad*.

Este gráfico se puede entender como un histograma “suavizado”, en el que el número de clases se ha hecho muy grande y los intervalos de clase se han hecho muy estrechos. Al seleccionar la variable *Pulso1* en el cuadro de diálogo que aparece con la opción *Gráficas > Estimar densidad*, el gráfico que obtendrás sería el siguiente.



Las marcas en el eje de abscisas indican la posición de los datos. Si comparamos este gráfico con los histogramas, la simetría de los datos resulta algo más evidente.

Diagrama de puntos

El diagrama de puntos representa los datos individualizados como puntos, en lugar de utilizar la frecuencia en los intervalos de clase. Por lo tanto, puede ser más informativo que el histograma o la estimación de la densidad cuando el tamaño muestral es pequeño.

En el menú de *Gráficas*, verás que aparece dos veces la opción *Diagrama de puntos*. Usando la primera, para la variable *Pulso1*, en el conjunto de datos *pulsoNuevo*, obtenemos:

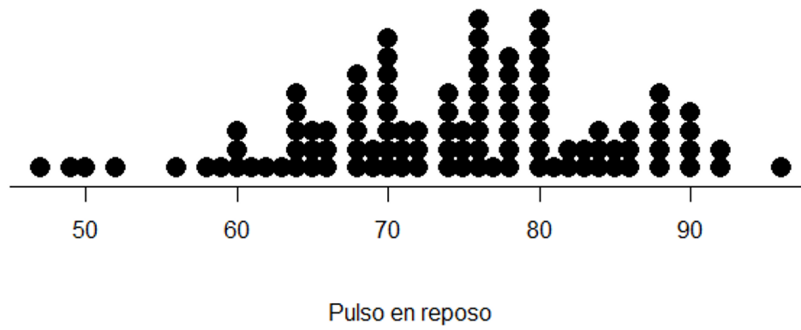


Diagrama de tallos y hojas

El diagrama de tallo y hojas se obtiene con la opción *Gráficas > Gráfica de tallos y hojas*, pero no aparece en la ventana de gráficos sino en el cuadro de resultados *Salida*. Para la variable Pulso1, con las opciones por defecto, se obtiene:

```
> with(pulsoNuevo, stem.leaf(Pulso1, na.rm=TRUE))
[1] "Warning: NA elements have been removed!!"
1 | 2: represents 12
leaf unit: 1
n: 105
LO: 47 49
 4 5* | 02
 7 5. | 689
18 6* | 00012344444
32 6. | 55566688888899
51 7* | 000000011122244444
(20) 7. | 55566666666678888888
34 8* | 00000000012233444
17 8. | 55666888888
 7 9* | 000022
 1 9. | 6
```

Selección de gráficos a utilizar

Por último, podemos plantearnos cuál es el gráfico que tenemos que utilizar en nuestros informes estadísticos cuando analizamos una variable numérica. La decisión dependerá de la naturaleza y las características de la variable concreta que estemos estudiando. Como regla general:

- Para variables discretas con pocas modalidades distintas, el gráfico de líneas es una buena opción y permite visualizar la información de la tabla de frecuencias.
- Para variables continuas, y para variables discretas con muchas modalidades distintas, el diagrama de caja da una primera aproximación a la distribución de la variable, indicando la posición de los cuartiles, y además permite detectar datos atípicos. Conviene obtener el histograma o la estimación de la densidad, que complementan la información que proporciona el diagrama de caja y que permiten confirmar, o algunas veces poner en cuestión, la información proporcionada por los estadísticos descriptivos. En particular, pueden ser útiles para visualizar la simetría o asimetría de la distribución.
- Cuando el tamaño muestral no es lo suficientemente grande, el histograma o la estimación de la densidad pueden ser poco informativos y en este caso es preferible utilizar el diagrama de puntos.

COMENTARIO: El diagrama de caja deja fijados los porcentajes de muestra que quedan entre las medidas que dibuja, es decir, mínimo, Q1, mediana, Q3 y máximo. Así vemos las distancias que hay entre ellos y esto aporta información. Además, permite detectar datos atípicos. El histograma deja fijadas las distancias entre los valores de la variable, es decir, fija los intervalos y son los porcentajes de la muestra lo que se distribuye en esos intervalos. Así, además de ver la forma de la variable, vemos los intervalos en los que la variable es más o menos frecuente.

4. Ejercicios

1. En un servicio de Traumatología, con objeto de realizar una correcta planificación, interesa conocer la distribución de las patologías (lesión de rodilla, de cadera, de tobillo, de cráneo, otras) de los pacientes atendidos en Urgencias durante los últimos seis meses. Los datos se encuentran en el archivo *patologia.RData*. Explica la distribución de las patologías. [Sugerencia: utiliza de la tabla de frecuencias y el diagrama de sectores.]
2. En una encuesta sobre hábitos televisivos, se ha preguntado a 80 participantes el número de series de televisión que siguen actualmente. Los resultados obtenidos se encuentran en el archivo *series.RData*. Realizar el estudio descriptivo de la variable *número de series de TV seguidas actualmente* [Sugerencia: comenta su tabla de frecuencias, su diagrama de líneas y sus medidas de posición y dispersión.]
3. Un servicio de medicina interna, con objeto de planificar debidamente sus recursos, estudia mediante muestreo aleatorio el número de urgencias atendidas por día, extrayendo una muestra de 60 días al azar. Las urgencias atendidas por día se encuentran en el archivo *urgencias.RData*. Realiza un estudio descriptivo de la variable *número de urgencias atendidas* [Sugerencia: comenta sus medidas de posición, dispersión y forma, el diagrama de caja y, entre el resto de gráficas, la que te resulte más fácil de comentar para completar la información de las medidas descriptivas].
4. El archivo *dvd.RData* contiene el número de dvd que poseen 56 estudiantes. Describe esta variable con los estadísticos y gráficas que consideres más convenientes.
5. Usando el archivo *Pulso.RData*, estudia descriptivamente la variable numérica *peso* y las variables categóricas *Fumador* y *Año* (recuerda que la variable *Año* hace referencia al curso académico en el que se obtuvieron los datos, y es por lo tanto una variable categórica ordinal).
6. El archivo *trabajo.RData* contiene una parte pequeña de una encuesta realizada para estudiar la satisfacción laboral de un grupo de trabajadores de una cierta empresa. En concreto, contiene el número de horas de trabajo semanales por cada entrevistado, los años de antigüedad en la empresa y la edad de los entrevistados. Obteniendo los diferentes estadísticos que necesitéis, completar las siguientes preguntas.
 - a) Si dividimos la muestra en cinco partes iguales, vemos que el 20% de los empleados más antiguos llevan más de ____ años en la empresa, mientras que el 20% de los empleados con menos antigüedad en la empresa llevan ____ años o menos trabajando en la misma. Por otro lado, hay un 20% de individuos que declara trabajar más de ____ horas semanales. El 20% central de la muestra trabaja entre ____ y ____ horas a la semana.
 - b) ¿Qué porcentaje de los casos válidos de la muestra trabaja exactamente 40 horas semanales? ____% ¿Qué porcentaje, también sobre las respuestas válidas, trabaja menos de 40 horas semanales? ____% ¿Y 41 o más horas semanales? ____%.
 - c) Sorprende que hay ____ personas (un ____% del total de casos válidos) que declaran trabajar 98 horas a la semana. Otros coeficientes de interés de la variable horas de trabajo a la semana son: un apuntamiento de ____, una desviación típica de ____, y una media de _____. Desde el punto de vista de la simetría, este coeficiente, con un valor de _____ nos da idea de un ligero sesgo positivo.
 - d) El ____% de los casos válidos de la muestra llevan 3 o menos años en la empresa, en tanto que solamente un ____% llevan 40 o más años de antigüedad. Sorprende una desviación estándar de ____ años para una media de antigüedad de ____ años.
 - e) La persona con más edad que ha contestado la encuesta tiene ____ años, pero hay en la muestra la nada despreciable cifra de un ____% de personas que tienen 80 años o más.
 - f) La edad mínima para comenzar a trabajar en la empresa era de 12 años. Conocido ese dato, ____ de los entrevistados han proporcionado mal sus datos. Lista esos casos.

Para pensar: Analiza la afirmación de un candidato que dice: *prometo que, si salgo elegido, subiré todos los sueldos de forma que nadie cobre por debajo de la media nacional.*