

Práctica 2

Estadística descriptiva unidimensional: representación tabular y gráfica

Esta práctica va a estar dedicada a realizar con R Commander sencillos análisis descriptivos unidimensionales. Distinguiremos entre variables cualitativas y variables cuantitativas.

Contenido de la práctica

2.1. Variables cualitativas	13
2.1.1. Variables cualitativas nominales	14
2.1.1.1. Tabla de frecuencias	14
2.1.1.2. Diagrama de sectores	14
2.1.2. Variables cualitativas ordinales	16
2.1.2.1. Diagrama de barras	16
2.2. Variables cuantitativas	18
2.2.1. Variables cuantitativas discretas	18
2.2.1.1. Tabla de frecuencias	18
2.2.1.2. Diagrama de barras	19
2.2.2. Variables cuantitativas continuas	20
2.2.2.1. Histograma	20
2.2.2.2. Diagrama tallo-hoja	20
2.3. Ejercicios propuestos	21

2.1. Variables cualitativas

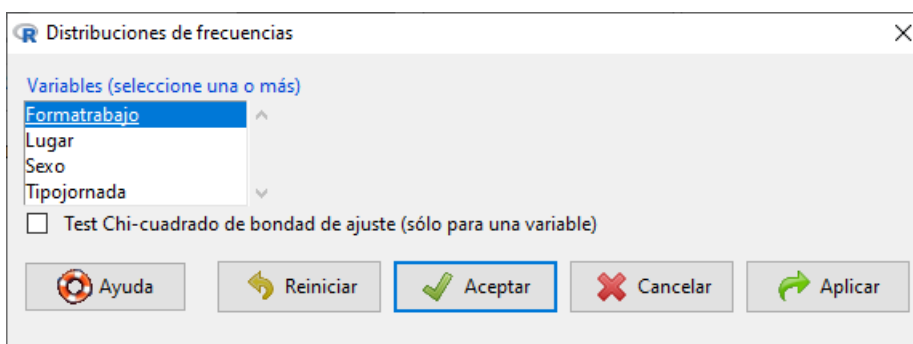
En este apartado, vamos a ver cómo describir con R Commander cualquier variable de tipo cualitativo (variables de tipo factor para R Commander). Las tablas de frecuencias, los diagramas de sectores y los diagramas de barras son las herramientas básicas para describirlas.

2.1.1. Variables cualitativas nominales

2.1.1.1. Tabla de frecuencias

En el fichero *Informacion.RData* tenemos la variable cualitativa nominal *Formatrabajo*, que indica las preferencias de los entrevistados sobre si preferirían trabajar por cuenta propia o por cuenta ajena. Para conocer cuántas personas preferirían trabajar por cuenta propia o por cuenta ajena, así como sus porcentajes en relación al total, tendremos que obtener las tablas de frecuencias absolutas y relativas (en porcentajes) de la variable *Formatrabajo*. Para ello:

1. Carga el archivo *Informacion.RData* con la opción *Datos > Cargar conjunto de datos*.
2. Selecciona en el menú principal *Estadísticos > Resúmenes > Distribución de frecuencias*. En el cuadro de diálogo que aparece, selecciona la variable *Formatrabajo*.



3. Tras pulsar *Aceptar*, en el cuadro de resultados aparecerán la tabla de frecuencias absolutas y la tabla de frecuencias relativas en porcentajes.

```
counts:
Formatrabajo
ajena propia
  97      40

percentages:
Formatrabajo
ajena propia
  70.8    29.2
```

A la vista de los resultados, de los 137 entrevistados, un 70,8% preferiría trabajar por cuenta ajena (97 entrevistados), mientras que sólo un 29,2% preferiría trabajar por cuenta propia (40 entrevistados).

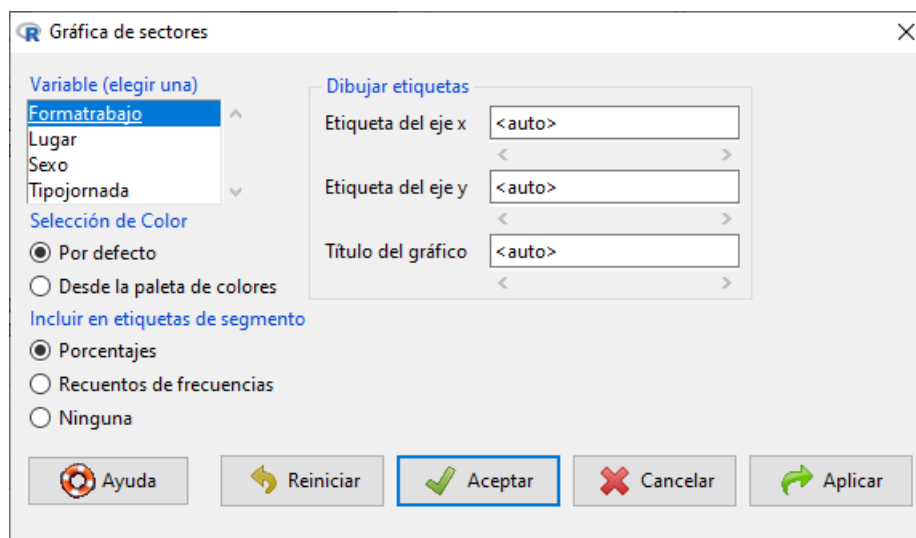
NOTA: se pueden obtener las tablas de frecuencias de varias variables cualitativas a la vez, sin más que seleccionarlas en el cuadro de diálogo anterior (utilizando si es necesario la tecla *<Ctrl>* y el ratón para seleccionarlas).

2.1.1.2. Diagrama de sectores

Además de la tabla de frecuencias, suele ser habitual añadir un gráfico cuando se está estudiando una variable de este tipo. El gráfico de sectores es uno de los más usados cuando se estudia una variable cualitativa nominal. El gráfico para la variable *Formatrabajo* se obtendría de la siguiente forma.

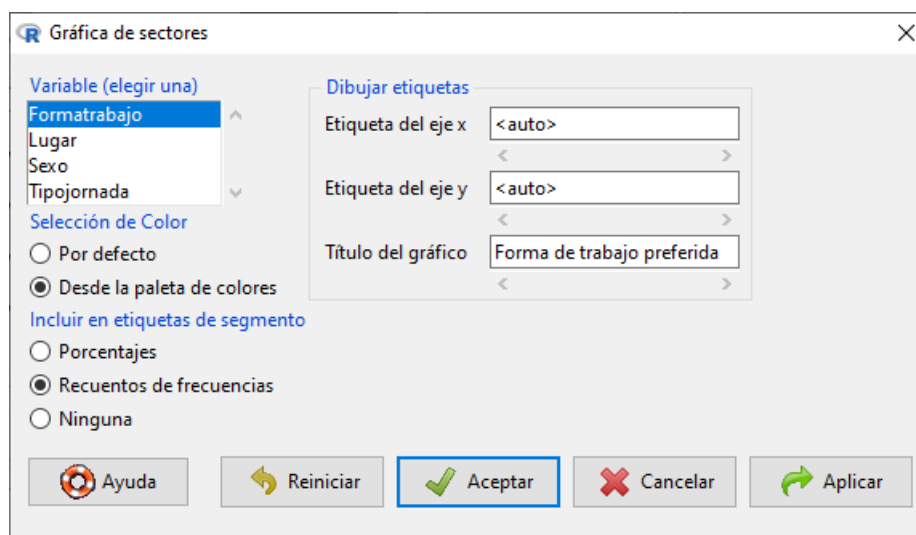
1. Selecciona en el menú principal *Gráficas > Gráfica de sectores*.

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



3. Tras pulsar *Aceptar*, te aparecerá en una nueva ventana el gráfico de sectores de la variable *Formatrabajo*.

El gráfico anterior se puede modificar un poco para adaptarlo a nuestro gusto, usando las opciones que aparecen en el cuadro de diálogo. Se pueden añadir opcionalmente tres rótulos: uno para el gráfico, otro para el eje x y otro para el eje y. También se pueden cambiar los porcentajes por las frecuencias absolutas. Además, se pueden cambiar los colores de los sectores con la opción Desde la paleta de colores. Por ejemplo, si completamos el cuadro de diálogo como se muestra a continuación:



, el gráfico de sectores que se obtiene utiliza los colores rojo (*indianred3*) y verde (*limegreen*), pues son los dos primeros colores definidos en la Paleta de colores asociada a la opción *Gráficas > Gama de colores*.

NOTA: R ofrece una amplia gama de herramientas y comandos para generar cualquier tipo de gráfico y añadir sobre él casi cualquier tipo de leyenda, rótulo, texto, etc. Sin embargo, debido a la complejidad del código R para realizar estas operaciones, nosotros no vamos a usarlos.

NOTA: la ventana de gráficos sólo admite un gráfico activo. Por tanto, una vez realizado un gráfico conviene copiarlo o guardarlo, pues al realizar el siguiente gráfico el anterior se pierde. Además de copiar un gráfico a *Word* con la opción *Archivo > Copiar para el área de transferencia > Como un Metafile*, R Commander permite guardar un gráfico en diferentes formatos (pdf, metaarchivo, png, jpeg, etc.) 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.

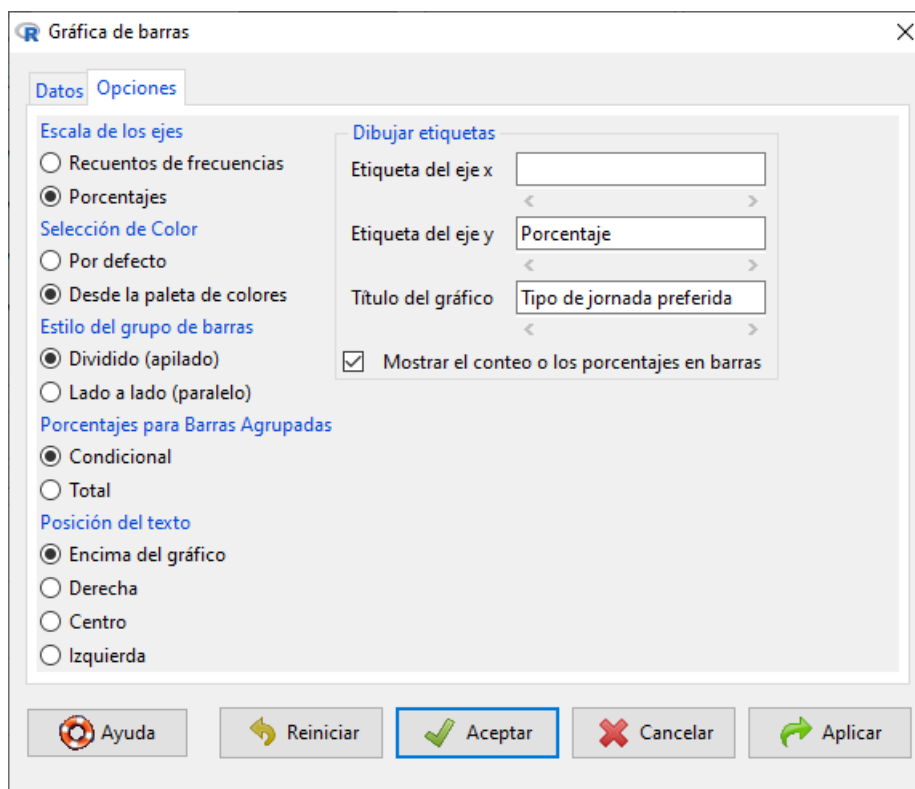
Ejercicio 11: Representa gráficamente la variable *Sexo* utilizando un diagrama de sectores, de manera que el título del gráfico sea *Sexo de los trabajadores*, aparezcan las frecuencias absolutas y los colores utilizados sean el marrón y el amarillo. Después, copia el gráfico obtenido y pégalo en un documento de *Word*.

2.1.2. Variables cualitativas ordinales

2.1.2.1. Diagrama de barras

Además de la tabla de frecuencias, para este tipo de variables el gráfico más utilizado para su representación es el gráfico de barras. La obtención de un gráfico de barras se obtiene con la opción *Gráficas > Gráfica de barras*. Supongamos que la variable *Tipojornada* que se encuentra en el fichero *Informacion.RData* es una variable de tipo cualitativa ordinal, y que el orden de las categorías es media jornada, jornada variable y jornada completa. Veamos cómo obtener el diagrama de barras de esta variable.

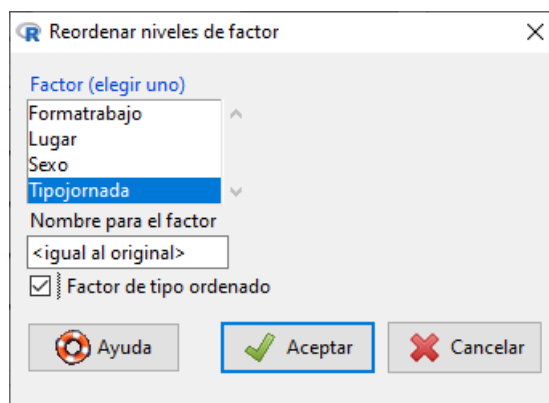
1. Selecciona en el menú principal *Gráficas > Gráfica de barras*.
2. Selecciona la variable *Tipojornada* en la pestaña *Datos* y rellena las diferentes opciones como se muestra a continuación.



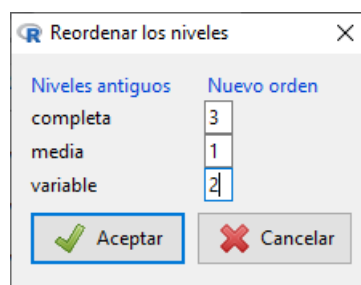
3. Tras pulsar *Aceptar*, te aparecerá el gráfico de barras de la variable *Tipojornada*.

Observa que el orden en el que aparecen las barras no es el orden natural que hemos supuesto para la variable (media, variable, completa) sino que aparecen en orden alfabético. Para cambiar el orden de las categorías de una variable de tipo factor, existe una opción en R Commander llamada *Reordenar niveles de factor*. Esta opción nos será de utilidad para establecer el orden natural de una variable cualitativa ordinal, cuando ésta no lo tenga por defecto. Por ejemplo, para establecer el orden natural de la variable *Tipojornada*, tendrías que hacer lo siguiente.

1. Selecciona en el menú principal *Datos > Modificar variables del conjunto de datos activo > Reordenar niveles de factor*.



2. En el cuadro de diálogo que aparece, elige la variable *Tipojornada*, selecciona la opción *Factor de tipo ordenado* y pulsa *Aceptar*. Al dejar *<igual al original>* en el campo *Nombre para el factor*, sobrescribiremos la variable *Tipojornada*, estableciendo una nueva ordenación para las categorías de dicha variable. Con la opción *Factor de tipo ordenado* le indicamos a R Commander que la variable es cualitativa ordinal.
3. Tras confirmar que deseas reescribir la variable *Tipojornada*, establece el nuevo orden de las categorías de la forma:



Tras pulsar *Aceptar*, ya se tendrá establecido el nuevo orden de las categorías para la variable *Tipojornada*. Si realizamos ahora el diagrama de barras, obtendremos en el gráfico la ordenación natural en lugar de la ordenación alfabética.

Notar que, si se obtuviesen las tablas de frecuencias para esta variable, en la salida también aparecerían las categorías ordenadas según el orden establecido. A través de las tablas de frecuencias vemos que, del total de entrevistados, 44 prefieren trabajar a media jornada (un 32.12%), 26 prefieren trabajar de forma flexible (un 18.98%) y 67 prefieren trabajar a jornada completa (un 48.91%).

2.2. Variables cuantitativas

Veamos a continuación qué tipo de tablas y gráficos se suelen utilizar, diferenciando entre variables cuantitativas discretas y variables cuantitativas continuas.

2.2.1. Variables cuantitativas discretas

Cuando una variable cuantitativa discreta toma pocos valores distintos, la tabla de frecuencias y el gráfico de barras nos pueden ser útiles para describir la variable.

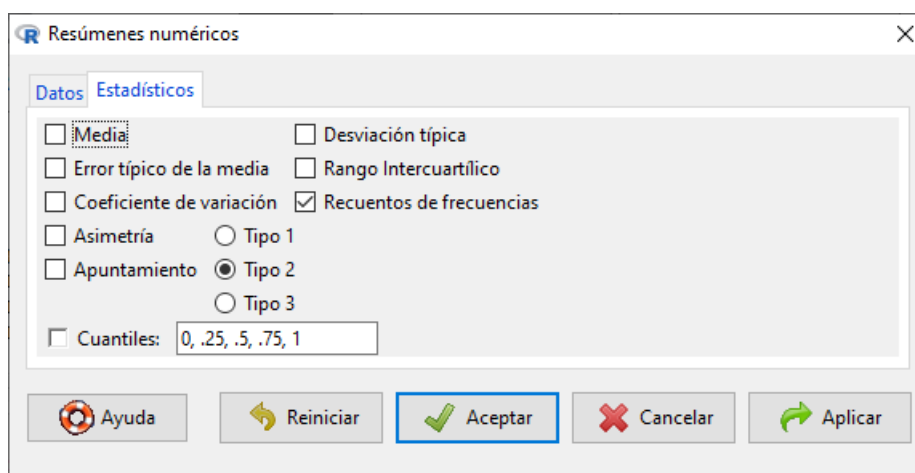
Vamos a analizar la variable *Hijos*, que se encuentra en el archivo *Hijos.xlsx*. Dicho archivo contiene el número de hijos de 297 familias, es decir, contiene 297 casos y una variable, *Hijos*, que es el número de hijos por familia.

Ejercicio 12: Importa el conjunto de datos *Hijos.xlsx*. ¿Cuántos valores distintos toma la variable *Hijos*? ¿De qué tipo es la variable? ¿Cómo considera R Commander la variable *Hijos*: numérica o factor?

2.2.1.1. Tabla de frecuencias

La variable cuantitativa discreta anterior toma pocos valores distintos. Al tratarse de una variable numérica, los pasos a seguir para obtener su tabla de frecuencias son:

1. Selecciona en el menú principal *Estadísticos > Resúmenes > Resúmenes numéricos*.
2. En la pestaña *Datos* del cuadro de dialogo que aparece, elije la variable *Hijos*.
3. En la pestaña *Estadísticos*, marca únicamente la opción *Recuentos de frecuencias*.



La tabla obtenida se muestra a continuación:

```
> discreteCounts(hijos["Hijos", drop=FALSE])
Distribution of Hijos
      Count Percent
0         37   12.46
1         76   25.59
2         68   22.90
3         53   17.85
4         42   14.14
5         15    5.05
6          6    2.02
Total    297  100.01
```

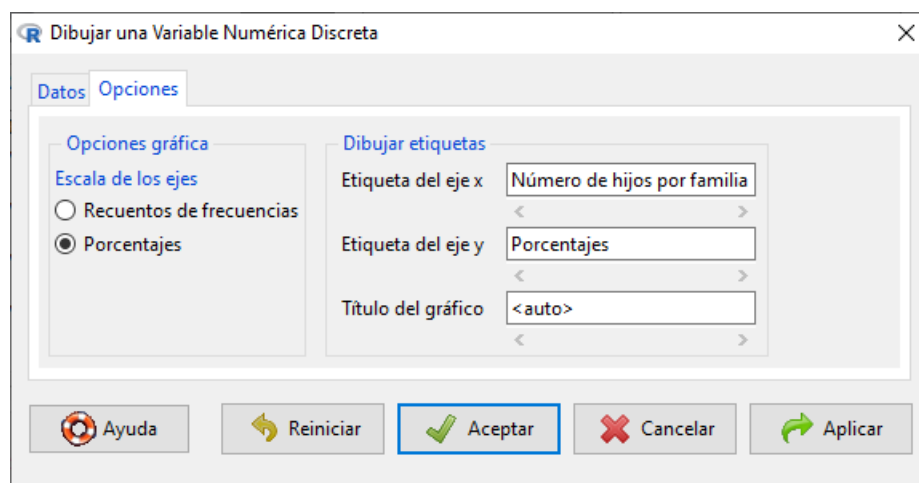
Con la información proporcionada por la tabla de frecuencias, podríamos decir que el número de familias estudiadas ha sido 297, que el número máximo de hijos ha sido 6 mientras que el mínimo ha sido 0, que el 25.59% de las familias tiene un único hijo siendo esta clase la mayoritaria de entre todas las familias estudiadas, y que el 39.06% de las familias estudiadas son numerosas (tienen tres hijos o más).

2.2.1.2. Diagrama de barras

Como complemento a la información anterior, se puede añadir el gráfico de barras de la variable. Recuerda que para obtener dicho gráfico, habría que usar la opción *Gráficas > Gráfica de Barras*. Sin embargo, en nuestro ejemplo, no podríamos usar directamente esa opción, pues R sólo permite obtener dichos gráficos de las variables de tipo factor, y la variable *Hijos* la tiene definida como variable numérica. La alternativa para obtener dicho gráfico sería crear una variable auxiliar de tipo factor a partir de la variable *Hijos*, tal y como se explicó en la práctica anterior cuando transformamos la variable numérica *sexo*, con valores 1 y 2, en otra de tipo factor, con valores *femenino* y *masculino*.

Ejercicio 13: Crea una nueva variable de tipo factor, llamada *HijosFactor*, que tome valores numéricos. Después, represéntala gráficamente mediante un diagrama de barras.

Para no tener que utilizar esta variable auxiliar, R Commander ofrece la opción *Gráficas > Dibujar una variable numérica discreta*, que permite obtener un gráfico similar al anterior donde las barras se sustituyen por líneas. A continuación, se muestra dicho gráfico para la variable *Hijos*, junto con la forma de rellenar la pestaña de opciones asociada a esta opción de R Commander.



Por último, cabe destacar, que cuando una variable cuantitativa discreta toma muchos valores, las tablas de frecuencias y los gráficos de barras 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.

2.2.2. Variables cuantitativas continuas

La tabla de frecuencias se obtiene de la manera que se ha descrito para una variable discreta. R Commander la pone en clases. Clases que coinciden con las del Histograma.

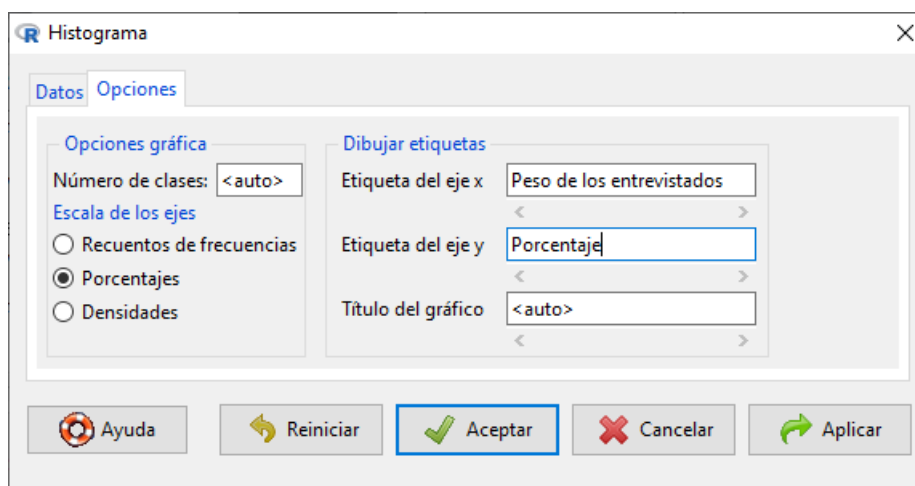
La herramienta básica para representar gráficamente una variable de tipo continuo es el histograma. Además, se pueden obtener el diagrama tallo hoja, como complemento al histograma.

Vamos a estudiar la variable *Peso* del archivo *Informacion.RData*, que contiene el peso de cada una de las personas entrevistadas.

2.2.2.1. Histograma

Para la obtención del histograma de la variable *Peso*, tienes que:

1. Seleccionar en el menú principal *Gráficas > Histograma*.
2. En la pestaña *Datos* del cuadro de diálogo que aparece, elige la variable *Peso*.
3. En la pestaña *Opciones*, puedes elegir cuántos intervalos (número de clases) quieres para el histograma o dejar que R Commander los calcule de forma automática (<auto>), puedes poner etiquetas en los ejes o título al gráfico, y puedes elegir la escala en el eje *y* (frecuencias, porcentajes o densidades).



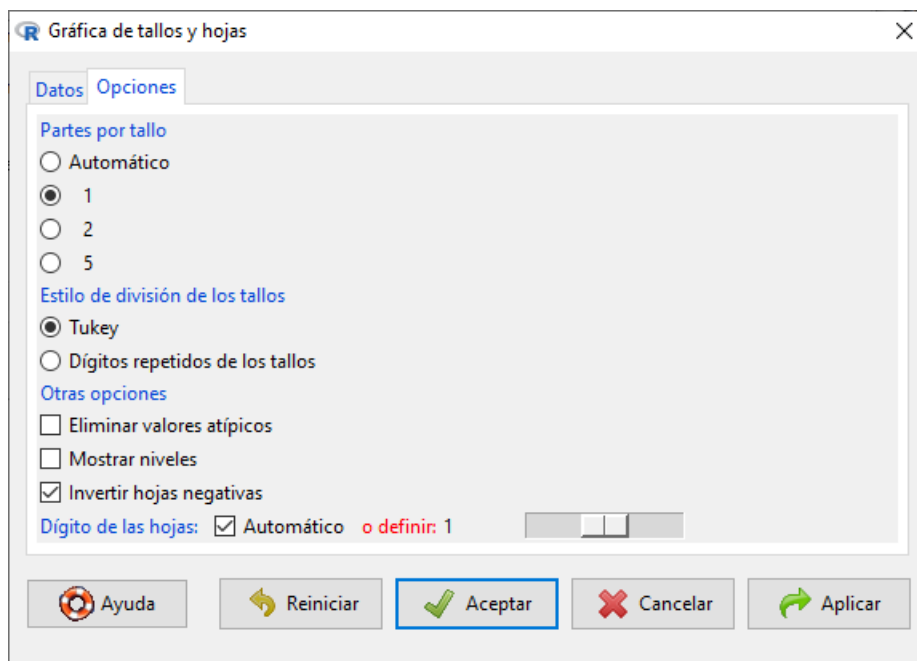
Al pulsar *Aceptar* con las opciones del cuadro de diálogo anterior obtendremos el histograma. Observa cómo la variable *Peso* parece tener cierta asimetría hacia la derecha.

2.2.2.2. Diagrama tallo-hoja

Cuando el número de casos con los que se trabaja es grande, el histograma es el gráfico más utilizado para describir una variable de tipo continuo. Si el número de datos no es muy grande, entonces se pueden

utilizar representaciones alternativas como el diagrama tallo-hoja. R Commander permite la obtención de este gráfico a través de la opción *Gráficas > Gráfica de tallos y hojas*.

A continuación, se muestran el diagrama tallo-hoja para la variable *Peso*, así como el cuadro de diálogo empleado.



```
> with(Información, stem.leaf(Peso, m=1, trim.outliers=FALSE, depths=FALSE, na.rm=TRUE))
1 | 2: represents 12
leaf unit: 1
      n: 137
 4 | 12367899
 5 | 0000113344555555666677778888888999
 6 | 0000000000012222333333344455555566667788889
 7 | 000000011233444455555566788888
 8 | 00002244555556778
 9 | 024568
10 | 25
11 | 00
```

Ejercicio 14: Representa gráficamente la variable *Peso* utilizando 16 intervalos en el caso del histograma y duplicando los tallos en el caso del diagrama de tallo y hojas. Compara los resultados.

2.3. Ejercicios propuestos

Ejercicio 15: En el fichero *Informacion.RData*, que recoge información sobre un conjunto de personas entrevistadas por un departamento de recursos humanos, se encuentran entre otras las variables *Sexo* (varon o mujer), *Lugar* (si el entrevistado prefiere trabajar en Madrid o no le importaría trabajar en otro lugar de España), y *Calif1* y *Calif2* (puntuaciones otorgadas a cada entrevistado por dos personas del departamento de recursos humanos).

1. ¿Qué porcentaje de varones hay entre los entrevistados? ¿Y de mujeres?
2. Representa los porcentajes anteriores mediante un diagrama de sectores.

3. ¿A cuántos de los entrevistados no les importaría trabajar fuera de Madrid? ¿Qué porcentaje representan sobre el total de entrevistados?
4. Representa los porcentajes de entrevistados que prefieren trabajar en Madrid y que no les importaría trabajar fuera de Madrid mediante un diagrama de barras.
5. Usando la variable *Calif1*, ¿cuál ha sido la calificación media obtenida por el conjunto de entrevistados?
6. ¿Cuáles han sido las calificaciones máxima y mínima de los entrevistados, según la variable *Calif1*?
7. Obtén el histograma de la variable *Calif1*. ¿Parece que existe algún tipo de asimetría? En caso afirmativo, ¿hacia dónde?
8. Repite los apartados 1.5 a 1.7 usando la variable *Calif2* y compara los resultados.

Ejercicio 16: En un seguimiento realizado a un grupo de 284 clientes de una empresa, se ha anotado el número de veces que los clientes han realizado reclamaciones a la empresa en el último año, para resolver alguna incidencia. Así, un cliente puede que no haya tenido que reclamar nunca a lo largo del año, y otro puede que haya tenido que reclamar 4 veces. Los datos se encuentran en el archivo *Veces.RData*.

1. Calcula el promedio de veces que un cliente ha reclamado a la empresa.
2. Calcula el número mínimo de veces que un cliente ha reclamado y el número máximo de veces que un cliente ha reclamado.
3. Obtén la tabla de frecuencias para el número de veces que un cliente ha reclamado. ¿Cuál es el valor más frecuente? Comenta la tabla a través de sus porcentajes.
4. Representa gráficamente, de varias formas distintas, el número de veces que se ha reclamado. Utiliza todos aquellos gráficos que tengan sentido.