

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.3 Análisis descriptivo de datos con una variable. Ejercicios

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

UAB



Análisis descriptivo de datos con una variable. Ejercicios

1. Análisis descriptivo de una variable: variables cualitativas

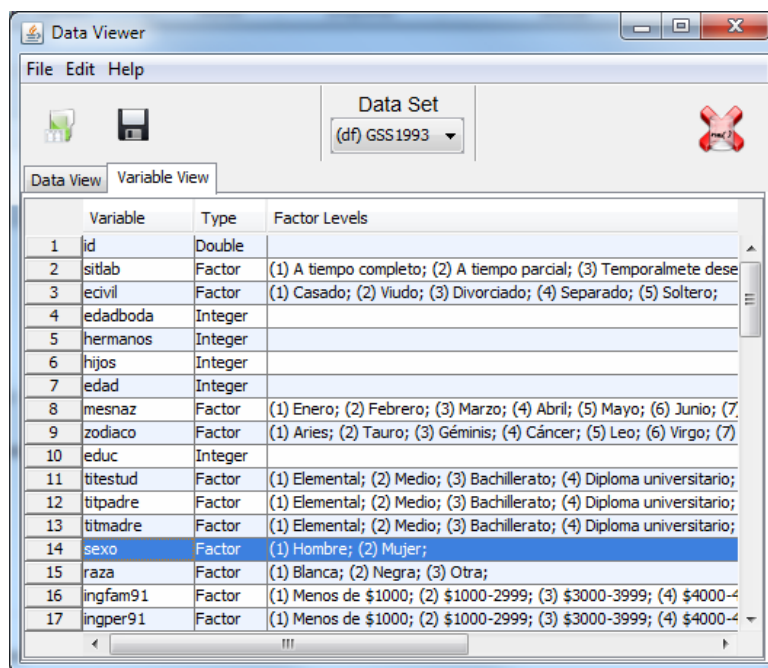
El objetivo de estos ejercicios prácticos es introducir el uso del software Deducer para el análisis descriptivo de una única variable medida a nivel nominal y ordinal (variables categóricas o cualitativas). Para realizar un análisis descriptivo básico procederemos de la siguiente manera: (1) obtendremos tablas de distribuciones de frecuencias, (2) crearemos gráficos para representar la información de las tablas (de barras y de sectores), y (3) estimaremos estadísticos de resumen adecuados a este tipo de variables, la moda para las variables nominales, y además la mediana y los percentiles para las variables ordinales.

► En esta práctica trabajaremos con los archivos **GSS1993.rda** y **Victimitzacio2008.rda** que se encuentran en la página web del capítulo III.3 del manual.

1.1. Análisis descriptivo de variables cualitativas nominales

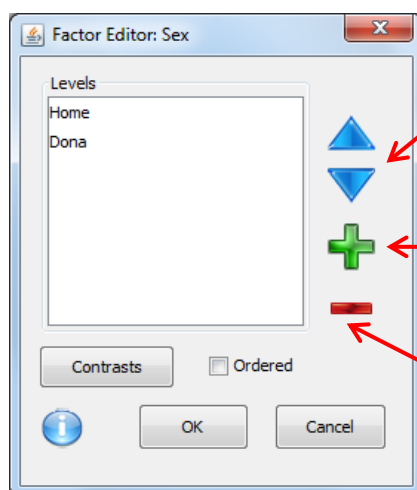
Consideremos la variable cualitativa nominal **SEXO** de la base **GSS1993.rda**. Deducer codifica automáticamente y en orden ascendente, a partir del número 1, las categorías de las variables según el orden en que son introducidas en la pestaña “*Data View*”¹. Por ejemplo, si en la matriz de datos la primera categoría introducida de la variable sexo es “hombre” y la segunda es “mujer”, Deducer codificará esta variable de la siguiente manera: (1) Hombre y (2) Mujer. Las etiquetas se pueden observar en la pestaña “*Variable View*”.

¹ Siempre y cuando la variable haya sido definida como FACTOR (variable categórica o cualitativa).



Sin embargo, es posible alterar este orden de codificación según el tipo de análisis que vayamos a realizar. Para ello tenemos dos opciones:

1. En la ventana *Data Viewer* (pestaña “*Variable Viewer*”) hacer click en la casilla correspondiente de la columna “*Factor Levels*” y aparecerá el editor con las siguientes opciones:

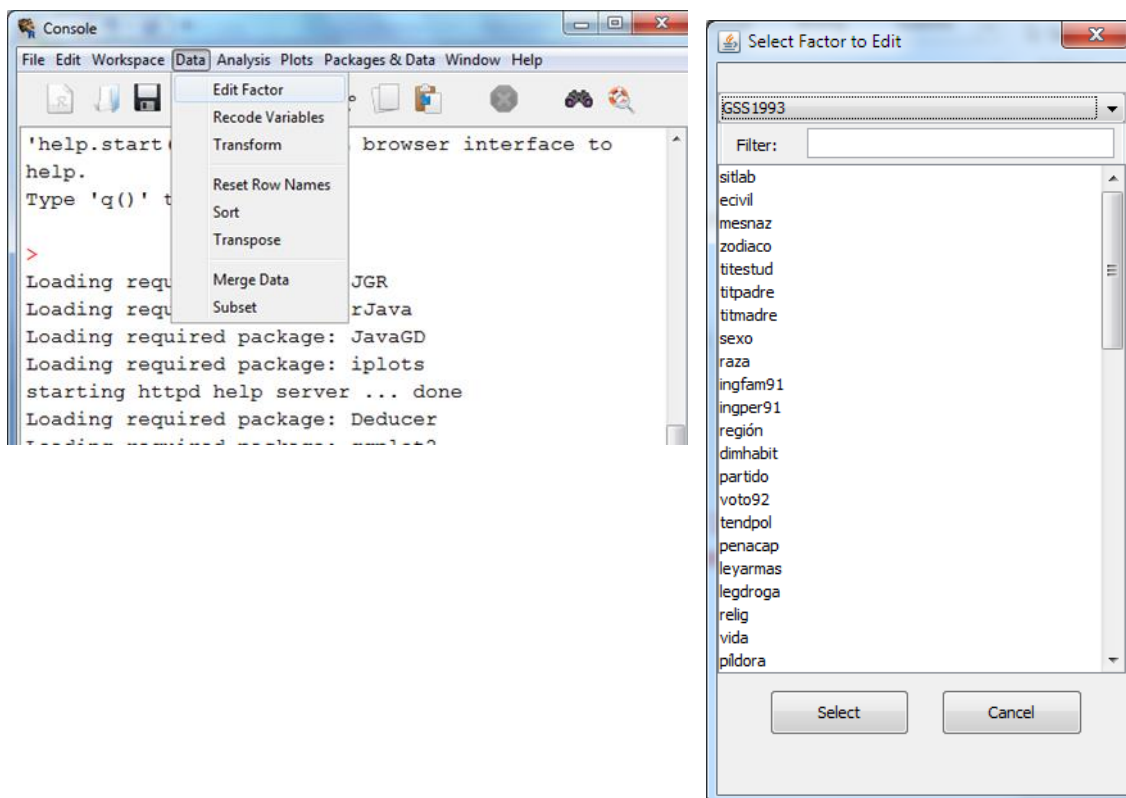


Para cambiar el orden de las categorías

Se puede agregar una categoría

Se puede quitar una categoría

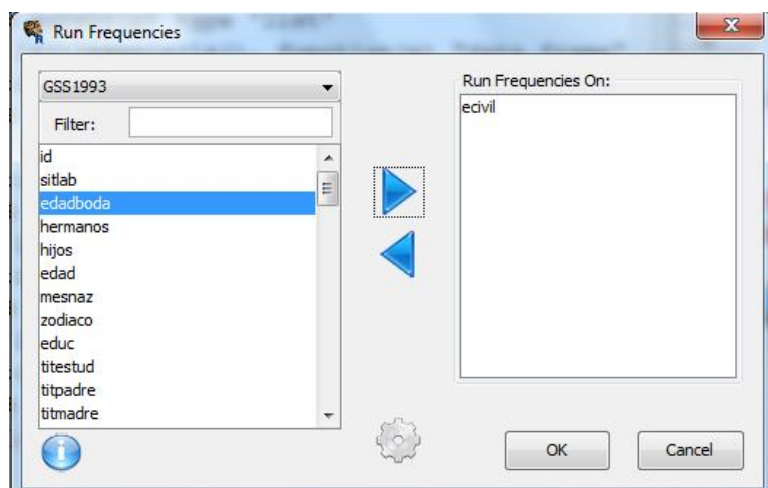
2. En la ventana *Console*, seleccionamos la opción: *Data/Edit Factor* y obtendremos un cuadro de diálogo con todas las variables. Por ejemplo, presionando sobre la variable *Sexo* obtendremos el mismo cuadro de la figura anterior.



► A continuación tomamos la variable Estado Civil (“ecivil”) de la base de datos **GSS1993.rda** y pediremos: (1) una tabla de distribución de frecuencias, (2) un gráfico de barras y (3) el estadístico de la moda.

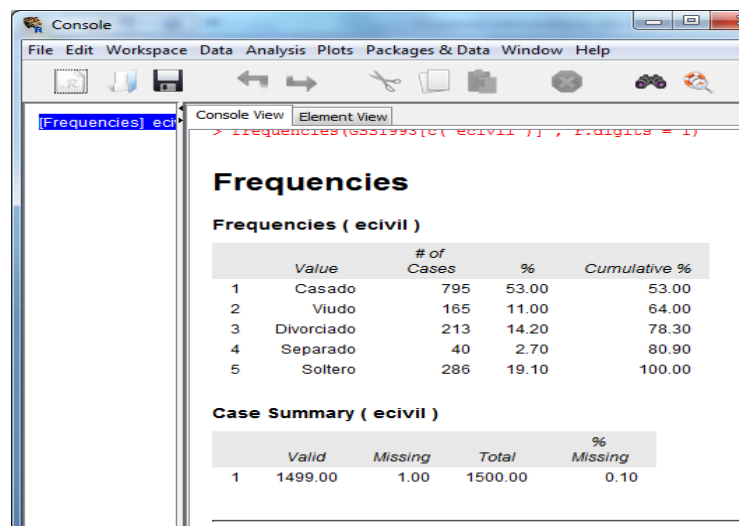
(1) Tabla de distribución de frecuencias. En la ventana “*Console*” ir a:

Analysis/Frequencies



Pasamos la variable “ecivil” al recuadro *Run Frequencies On*, tal y como aparece en la figura anterior (la opción *Filter* nos permite buscar las variables de una manera más sencilla escribiendo el nombre de la variable en este recuadro).

Luego presionamos *Ok* y si tenemos cargado el paquete *DeducerRichOutput* obtendremos la siguiente tabla de distribución de frecuencias:



The screenshot shows the R console window with the following output:

```
> frequencies(GSS1993[c("ecivil")], r.digits = 1)
```

Frequencies

Frequencies (ecivil)

	Value	# of Cases	%	Cumulative %
1	Casado	795	53.00	53.00
2	Viudo	165	11.00	64.00
3	Divorciado	213	14.20	78.30
4	Separado	40	2.70	80.90
5	Soltero	286	19.10	100.00

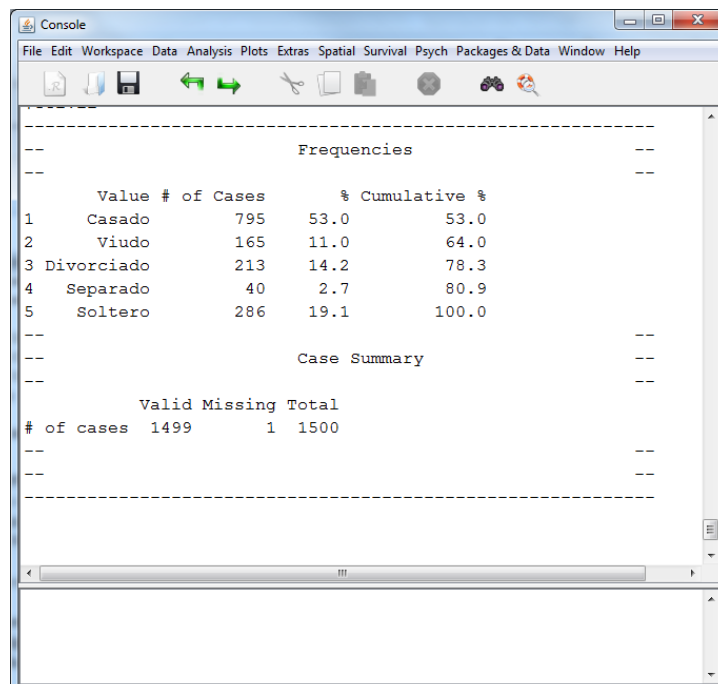
Case Summary (ecivil)

	Valid	Missing	Total	% Missing
1	1499.00	1.00	1500.00	0.10

NOTA: Otra manera de hacer lo anterior sería escribiendo en la consola la siguiente instrucción:

```
frequencies(GSS1993[c("ecivil")], r.digits = 1)
```

Si no tenemos cargado el paquete *DeducerRichOutput* obtendremos una tabla menos estética, pero con la misma información.

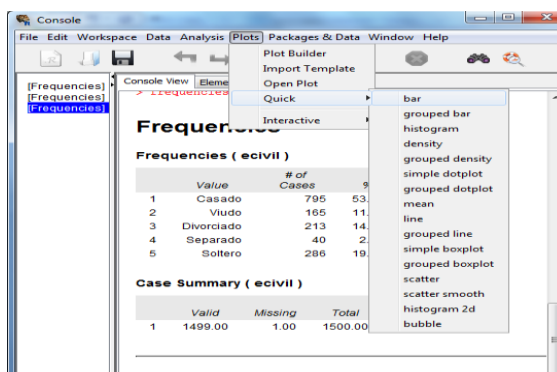


The screenshot shows the R console window with the following output:

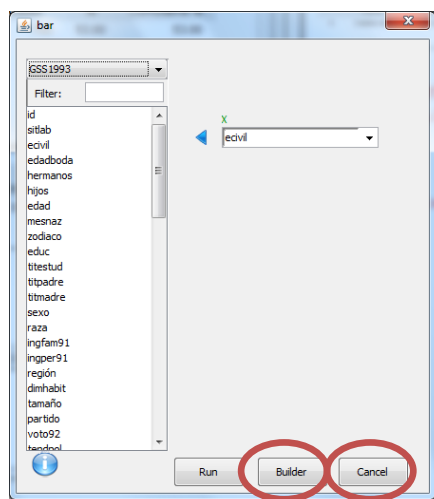
```
-----
--                               Frequencies                               --
--                               -----                               --
--                               Value # of Cases    % Cumulative %
1      Casado      795    53.0      53.0
2      Viudo      165    11.0      64.0
3  Divorciado    213    14.2      78.3
4      Separado    40     2.7      80.9
5      Soltero    286    19.1     100.0
-----
--                               Case Summary                               --
--                               -----                               --
--                               Valid Missing Total
# of cases 1499      1 1500
-----
```

(2) Gráfico de barras. En la ventana “Console” ir a:

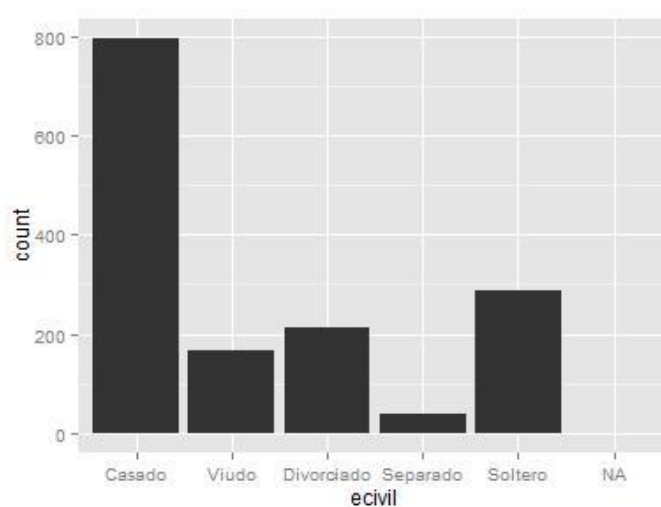
Plots/Quick (o template)/Bar



Aparecerá la siguiente ventana. Colocamos la variable “ecivil” en el recuadro de la X, y presionamos Run.



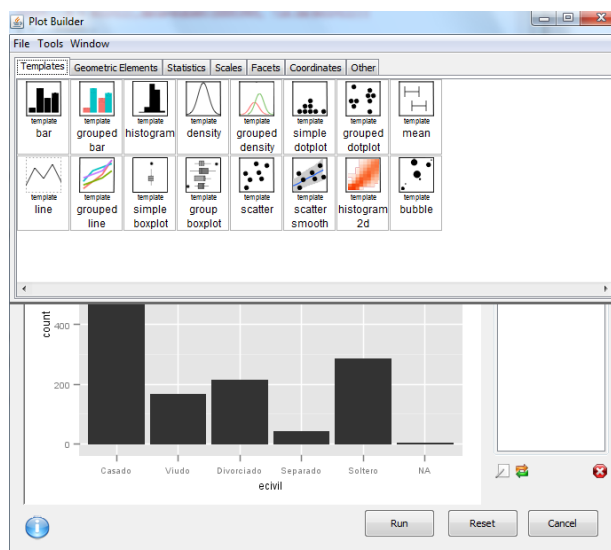
Obtendremos el siguiente gráfico:



NOTA: El procedimiento anterior genera automáticamente la siguiente instrucción en la ventana de la *Console*, que es otra forma de hacer lo anterior:

```
ggplot() +  
  geom_bar(aes(y = ..count..,x = ecivil),data=GSS1993)
```

Si en vez de presionar *Run*, presionamos *Builder* obtendremos la siguiente ventana:



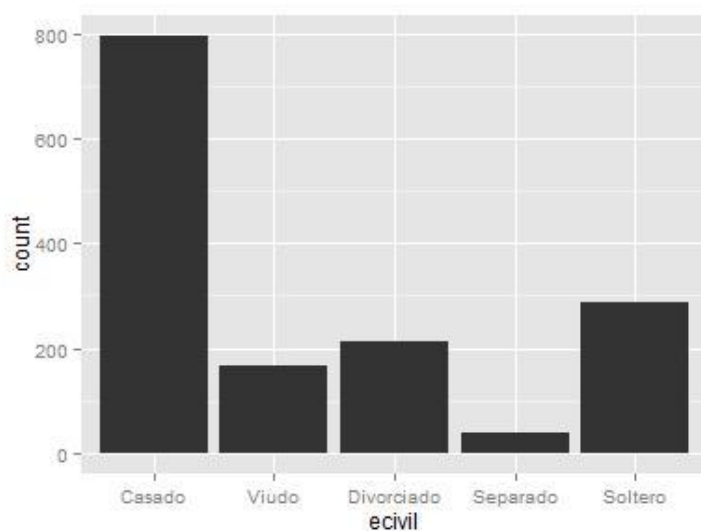
En ella, podemos ir construyendo tantos gráficos como deseemos dependiendo del tipo de variables en las que estemos interesados analizar.

Sin embargo, si observamos detenidamente el gráfico de barras de la variable “ecivil” notaremos que está representando los valores perdidos (NA). Para eliminarlos del gráfico podemos utilizar la siguiente instrucción:

```
ggplot() + geom_bar(aes(y = ..count..,x = ecivil),data=subset(GSS1993,  
  is.na(ecivil)))
```

NOTA: La instrucción resaltada en amarillo indica que se hará un gráfico con un subconjunto de datos (`data=subset`) de la base **GSS1993.rda** que NO tomará en cuenta los valores perdidos de la variable “ecivil” (`is.na(ecivil)`).

Se obtiene entonces el siguiente gráfico de barras:

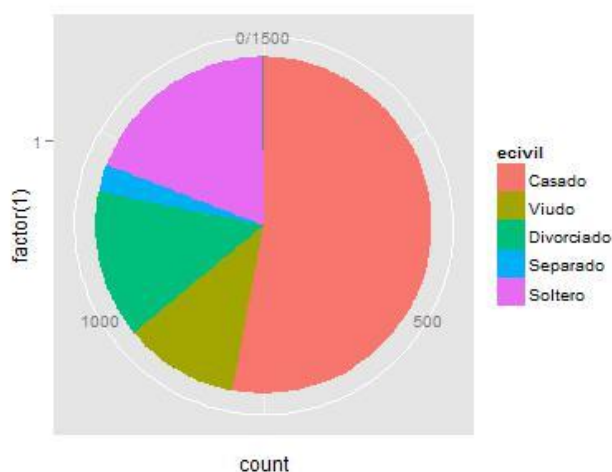


- (3) Gráfico de sectores. Dado que la opción para hacer un gráfico de sectores, o “Pie Chart”, no se encuentra disponible en Deducir a través de menús (al menos con los paquetes actualmente instalados) tenemos que hacer el gráfico utilizando las instrucciones de R en la “Console”.

Para obtener un gráfico de sectores de la variable “civil” escribimos la siguiente instrucción:

```
ggplot(GSS1993, aes(x = factor(1), fill = civil)) +  
geom_bar(width = 1) +  
coord_polar(theta = "y")
```

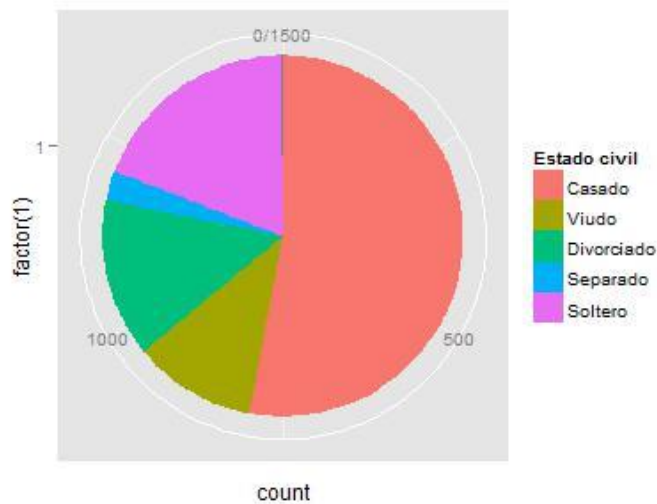
Y obtenemos el siguiente gráfico:



Para modificar el título de las etiquetas podemos aplicar la siguiente instrucción:

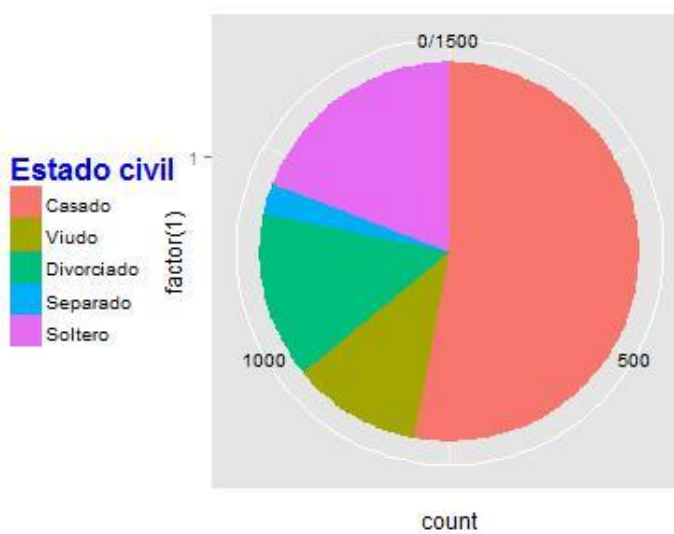
```
ggplot(GSS1993, aes(x = factor(1), fill = ecivil)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  scale_fill_discrete(name="Estado civil")
```

Y obtenemos:



También podemos editar el tamaño, el color y la ubicación de las etiquetas añadiendo más líneas de código (resaltadas en amarillo):

```
ggplot(GSS1993, aes(x = factor(1), fill = ecivil)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  scale_fill_discrete(name="Estado civil") +
  theme(
    legend.position="right",
    legend.title = element_text(colour="blue", size=16, face="bold"),
    axis.text.x=element_text(color="black"))
```



► Ejercicio 1

Utilizando la base de datos **GSS1993.rda** obtén una tabla de frecuencias y un gráfico de barras para las siguientes variables cualitativas nominales: situación laboral (“sitlab”), raza (“raza”) y sexo (“sexo”).

► Ejercicio 2

Con la matriz de datos **Victimitzacio2008.rda** ejecutar los procedimientos anteriores con las siguientes variables: sexo (“sexe_ps”), estudios (“estudis4”) y situación profesional (“sit_prof”).

1.2. Análisis descriptivo de variables cualitativas ordinales

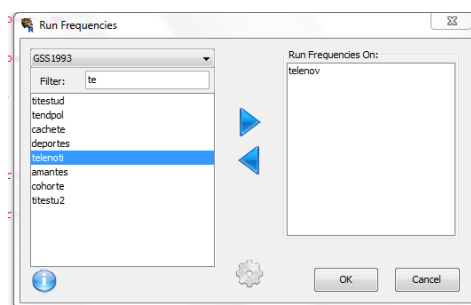
Consideremos la variable cualitativa ordinal **telenov** de la base de datos **GSS1993.rda** resultado de preguntar en el cuestionario de la encuesta: “¿Cuán frecuente ve Ud. comedias o dramas en la televisión?”.

Telenov una variable **cualitativa ordinal** codificada con valores del 1 al 5, según la frecuencia de consumo de este tipo de programas: (1) diariamente, (2) varios días a la semana, (3) varios días al mes, (4) raramente y (5) nunca.

Pediremos para esta variable: (1) una tabla de distribución de frecuencias, (2) un gráfico de barras y (3) el estadístico de la moda.

Para ello: A través del menú de la Console: *Analysis / Frequencies*

(1) Tabla de distribución de frecuencias: colocamos la variable **telenov** en el recuadro *Run Frequencies On:*



Se generará la tabla de frecuencias siguiente:

```
> frecuencies(GSS1993[c("telenov")], r.digits = 1)
```

Frecuencias

Frecuencias (telenov)

	Value	# of Cases	%	Cumulative %
1	Diariamente	313	21.00	21.00
2	Varios días a la semana	550	36.90	57.90
3	Varios días al mes	258	17.30	75.20
4	Raramente	275	18.50	93.70
5	Nunca	94	6.30	100.00

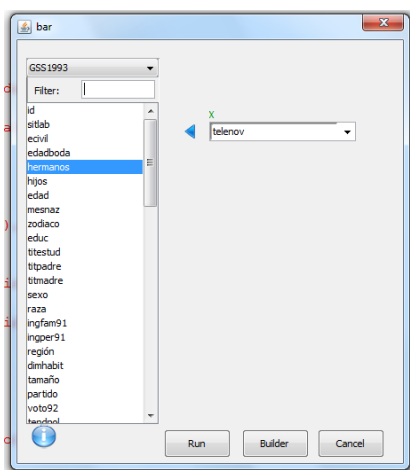
Case Summary (telenov)

	Valid	Missing	Total	% Missing
1	1490.00	10.00	1500.00	0.70

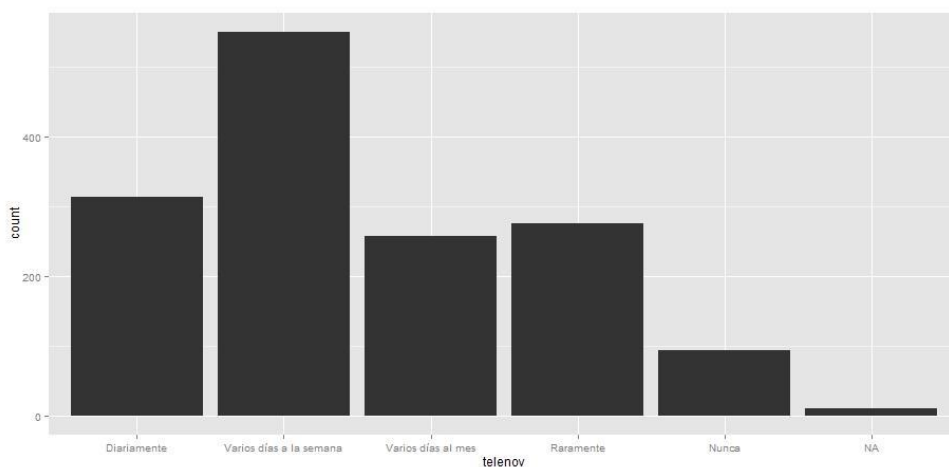
(2) Gráfico de barras. En la ventana “*Console*” ir a:

Plots/Quick (o Template)/Bar

Colocamos la variable **telenov** en el recuadro X.

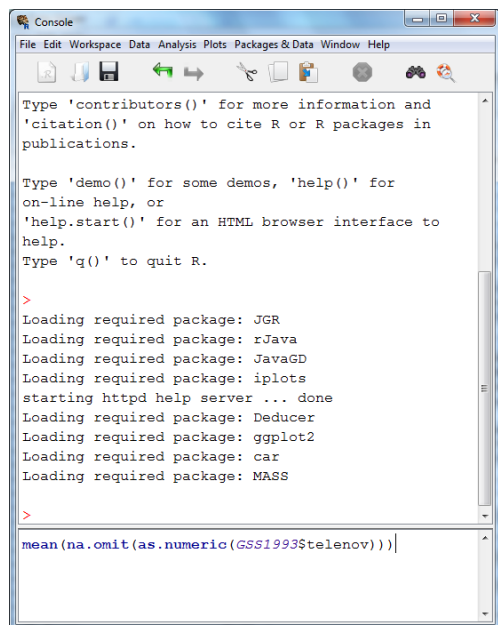


Se generará el siguiente gráfico:



- (4) La mediana: en el recuadro de instrucciones (ver siguiente figura) escribimos la siguiente instrucción y presionamos *enter*:

```
median(na.omit(as.numeric(GSS1993$telenov)))
```



Se generará el siguiente resultado:

```
> median(na.omit(as.numeric(GSS1993$telenov)))
[1] 2
```

Resumen de los resultados obtenidos de la variable **telenov**:

- La tabla de frecuencias de la variable muestra en la primera columna el código de la etiqueta de la variable, en la segunda el nombre de la etiqueta, en la tercera el número de casos por cada categoría, en la cuarta el porcentaje de cada categoría y en la última el porcentaje acumulado.
- En la tabla inferior aparecen el número total de casos contabilizados (1490) y seguidamente los valores perdidos (10).
- La moda, el valor más frecuente, corresponde a la categoría “varios días a la semana” (550 respuestas que corresponden al 36.9% de los casos).
- La mediana, el valor que acumula el 50% de los casos, es el valor 2. El 57,9% de las personas encuestadas ve programas de comedias o dramas en la televisión “varios días a la semana” o “diariamente”.

► Ejercicio 3

Repetir el análisis con las variables ordinales de la base **GSS1993.rda**: título escolar (“*titestud*”) y título escolar del padre (“*titpadre*”).

Podemos pedir los **cuartiles** a través de la siguiente instrucción:

```
> quantile(na.omit(as.numeric(GSS1993$ingper91), .25))
0% 25% 50% 75% 100%
1 9 14 17 22
>
```

Que omite los NAs	La tratamos como numérica solo para la ejecución de esta instrucción.	Dataframe en uso	La variable en cuestión	Cuartiles (25%)
-------------------	--	------------------	-------------------------	-----------------

Cabe destacar que para realizar todas estas operaciones hemos transformado la variable cualitativa a numérica (o cuantitativa) con la especificación *as.numeric*.

Para pedir **percentiles** específicos nos valdremos de la siguiente instrucción que presentamos en dos partes:

- (1) Primero generamos una variable local que llamaremos “**ingresos**”, escribiendo la siguiente instrucción:

```
ingresos <- as.numeric(na.omit(GSS1993$ingper91))
```

```
> ingresos <- as.numeric(na.omit(GSS1993$ingper91))
>
```

- (2) Una vez generada la variable escribimos la siguiente instrucción:

```
> quantile(ingresos, prob=.50)
50%
14
\
```

Ello nos devolverá el percentil 50, es decir, el valor de la mediana (prob=0.50). Con esta instrucción podemos pedir en Deducer el percentil que deseemos. Por ejemplo, si queremos conocer el percentil 60, sólo tenemos que cambiar en la consola **prob=0.50** por **prob=0.60** y tendremos:

```
> quantile(ingresos, prob=.60)
60%
15
>
```

Ahora nos preguntamos:

- ¿Cuál es el percentil 30? _____
- ¿Cuál es el percentil 60? _____
- ¿Cuál es el percentil 14? _____

A continuación se presentan otras instrucciones para realizar el mismo procedimiento con variables cualitativas *factor*. Primero aparece la que ejecuta *Deducer* con el comando *Descriptive* pero donde hemos añadido la especificación *as.numeric* (en caso contrario no se puede obtener pues se exige que sea *integer* o *double*) y a continuación presentamos otra que permite solicitar diversos percentiles a la vez con el comando *quantile* de R.

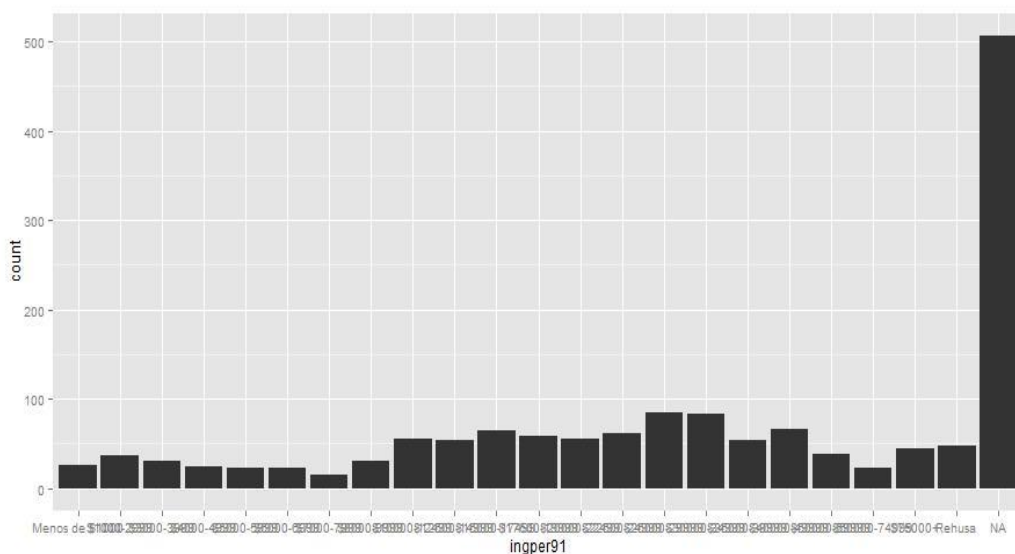
Comando “Descriptive” de **Deducer**:

```
descriptive.table(vars = d(as.numeric(ingper91)),data=GSS1993,
  func.names =c("Valid N","Median","25th Percentile","75th Percentile"))
```

Comando “Quantile” de **R**:

```
quantile(as.numeric(GSS1993$ingper91), c(0.14,0.25,0.30,0.50,0.60,0.75),
  na.rm=TRUE)
quantile(as.numeric(na.omit(GSS1993$ingper91)), c(.14,.25,.30,.50,.60,.75))
```

► El diagrama de barras lo pedimos a través de: “Console” *Plots/Quick (Template)/Bar*:



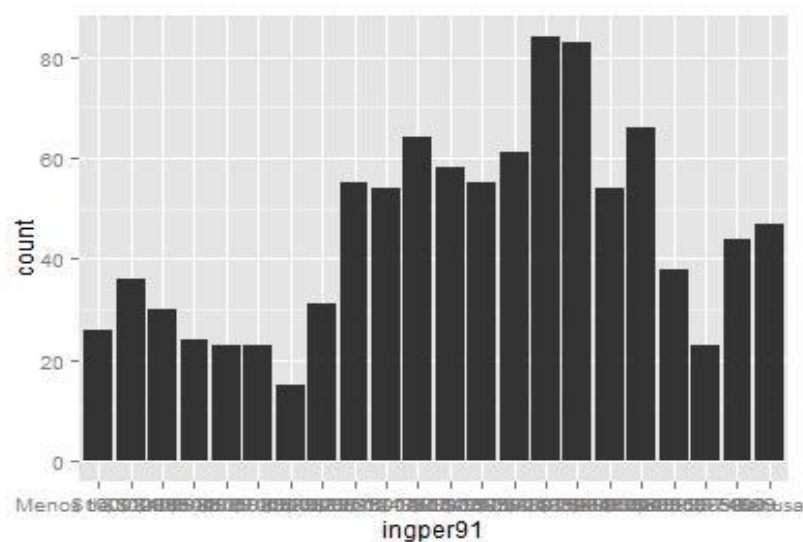
La instrucción que ha generado este gráfico de barras (que aparece en el recuadro de sintaxis inmediatamente después de su ejecución) es la siguiente:

```
ggplot() +
  geom_bar(aes(y = ..count..,x = ingper91),data=GSS1993)
```


Para quitar los NA del gráfico de barras añadimos la siguiente línea en la instrucción anterior (resaltada en amarillo):

```
ggplot() +  
  geom_bar(aes(y = ..count.., x = ingper91), data = subset(GSS1993,  
  lis.na(ingper91)))
```

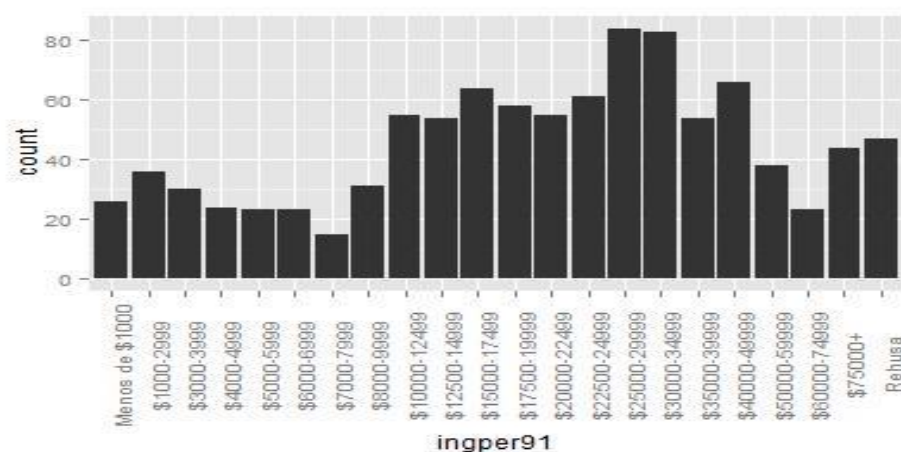
Y obtenemos:



Sin embargo, como se puede apreciar en el gráfico las etiquetas de las categorías están superpuestas lo cual las hace indistinguibles. Para solucionar este inconveniente añadimos a la instrucción anterior el siguiente código (resaltado en amarillo):

```
ggplot() +  
  geom_bar(aes(y = ..count.., x = ingper91), data = subset(GSS1993,  
  lis.na(ingper91))) +  
  theme(axis.text.x = element_text(angle = 90))
```

Lo que estamos haciendo con este código es decirle a Deducer que las etiquetas se presenten en un ángulo de 90 grados. Así obtenemos el siguiente gráfico:



► **Ejercicio 5**

Hacer un análisis similar con la variable *ingfam91* (los ingresos brutos familiares en el año 1991) de la base **GSS1993.rda**.

► **Ejercicio 6**

Con la matriz de datos **Victimitzacio2008.rda** se puede hacer el mismo tipo de ejercicio con las variables: *ps9_a*, *ps9_b* ó *osp6*.

2. Análisis descriptivo de una variable: variables cuantitativas o numéricas

Esta segunda parte complementa la anterior en el objetivo de introducir el uso del software R para el análisis descriptivo de una única variable, pero ahora referida a una variable que está medida a nivel numérico (variables cuantitativas, ya sean continuas o discretas, y que a R identifican como *integer* o *double*). Para realizar este análisis descriptivo procederemos a obtener tablas de distribuciones de frecuencias, gráficos para representar la información de las tablas (histogramas, polígonos de frecuencias, diagramas de caja y gráficos de tallo y hoja), así como los estadísticos de resumen adecuadas para este tipo de variables: de tendencia central (moda, mediana, media), de posición o tendencia no central (máximo, mínimo y percentiles), de dispersión (varianza, desviación típica, coeficiente de variación), y de forma de la distribución (asimetría y curtosis).

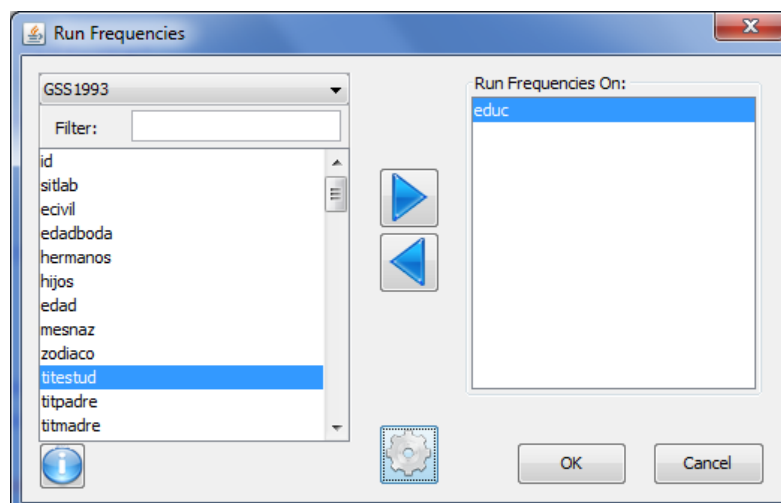
► Como antes trabajaremos con los archivos **GSS1993.rda** y **Victimitzacio2008.rda** que se encuentran en la página web del capítulo III.3 del manual.

2.1. Tabla de distribución de frecuencias, estadísticos e histograma


Consideraremos la variable cuantitativa discreta **EDUC**. Codificada de tipo "numérica", y con formato "integer" en R, con valores del 0 hasta el 20 que indican el número de años de escolarización de la persona entrevistada. Además, la variable contiene el valor perdido NA que corresponde a los "no sabe".

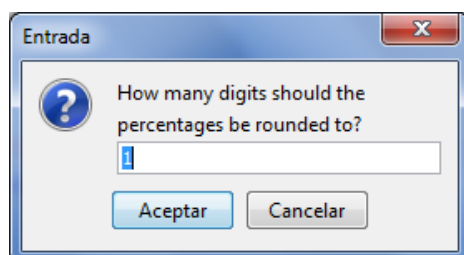
► A través del procedimiento *Frequencies* pediremos en primer lugar la tabla de distribución de frecuencias, a través del menú: **Analysis / Frequencies**

Nos aparece el cuadro de diálogo del procedimiento:



Seleccionamos la variable **educ** y la colocamos en el recuadro de "Run Frequencies On:". De esta manera obtenemos la tabla de distribuciones de frecuencias.

El icono  nos permite pedir el número de decimales a los que redondea los porcentajes, por defecto 1. No hay que cambiarlo y lo podemos dejar así:



Y clicamos finalmente sobre "OK" en el cuadro de diálogo principal y observamos los siguientes resultados:

Frecuencias

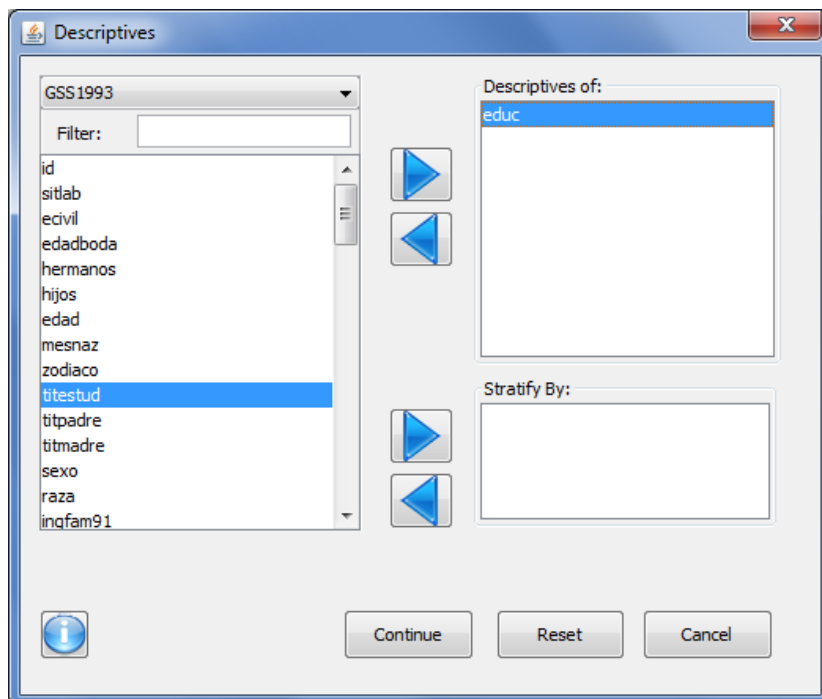
Frecuencias (educ)

	Value	# of Cases	%	Cumulative %
1	0	2	0.10	0.10
2	2	4	0.30	0.40
3	4	7	0.50	0.90
4	5	7	0.50	1.30
5	6	20	1.30	2.70
6	7	26	1.70	4.40
7	8	59	3.90	8.40
8	9	45	3.00	11.40
9	10	55	3.70	15.00
10	11	81	5.40	20.50
11	12	445	29.70	50.20
12	13	135	9.00	59.20
13	14	166	11.10	70.30
14	15	70	4.70	75.00
15	16	208	13.90	88.90
16	17	46	3.10	92.00
17	18	71	4.70	96.70
18	19	24	1.60	98.30
19	20	25	1.70	100.00

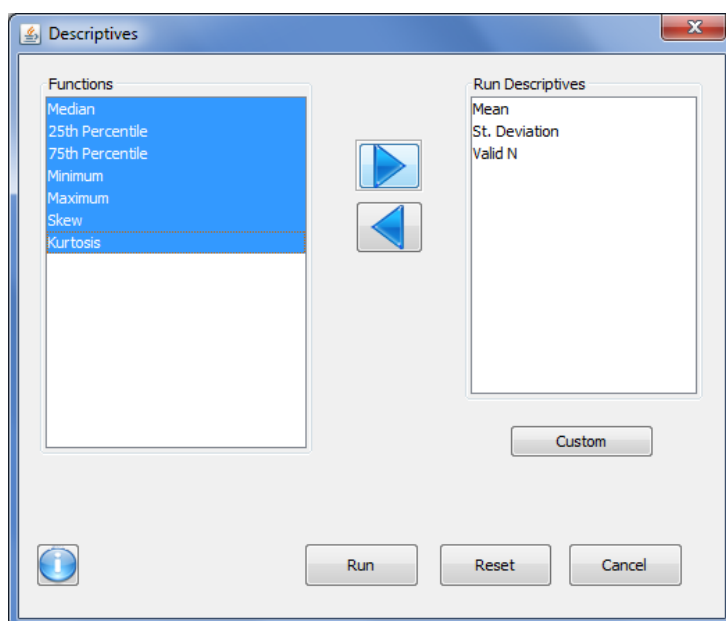
Case Summary (educ)

	Valid	Missing	Total	% Missing
1	1496.00	4.00	1500.00	0.30

► En segundo lugar pediremos los estadísticos descriptivos de tendencia central, los cuartiles, las medidas de dispersión y las de distribución a través del procedimiento "Descriptivas". De nuevo seleccionamos la variable **educ** y la colocamos en el recuadro de "Descriptivas of:":



Clicaremos luego sobre "Continue" y elegiremos los estadísticos. Por defecto nos saca la media (*mean*), la desviación típica (*standard deviation*) y el número de casos válidos (*valid n*). Lo que haremos será añadir el resto de funciones de la izquierda para obtener también la mediana (*median*), los percentiles 25 y 75, el máximo, el mínimo, la asimetría (*skew*) y la curtosis (*Kurtosis*):



Los resultados son estos:

Descriptive Statistics

	St		Valid N	Median	25th Percentile	75th Percentile	Minimum	Maximum	Skew	Kurtosis
	Mean	Deviation								
educ	13.04	3.07	1496	12.00	12.00	15.25	0	20	-0.309	0.708

- Las frecuencias se refieren a 1496 casos de los 1500 totales, los 4 casos que corresponden a los valores perdidos de "No sabe" y no se tendrán en cuenta en los cálculos.
- La distribución de frecuencias se ha resumido y expresado a través de varios estadísticos que nos informan de sus características:
 - Vemos en primer lugar que el conjunto de valores varía entre el **0** (el valor **mínimo**) y el **20** (valor **máximo**). Por lo tanto, el recorrido o el **rango** es **20**.
 - La **moda**, el valor más frecuente, no aparece, pero de la tabla de distribución de frecuencias se observa fácilmente que es **12** años de escolarización, y corresponde a un total de 445 casos, es decir, el 29,7% de la muestra ha sido escolarizado 12 años. Con variables cuantitativas este estadístico sólo es informativo cuando tenemos un número reducido de valores, como es el caso.
 - La **mediana**, valor tal que el 50% de observaciones son inferiores a él y el 50% superiores, es también **12**. El valor que corresponde al porcentaje acumulado del 50,2% es 12 años de escolarización. Es decir, el 50,2% de los individuos tienen una escolarización igual o inferior a 12 años, y, en particular, el 50% acumulado de la muestra corresponde también al 12.
 - El mínimo, el máximo, la moda y la mediana se pueden observar directamente sobre la tabla de frecuencias sin necesidad de consultar la tabla de estadísticos.
 - Los **cuartiles** primero y segundo (P_{25} y P_{50}) coinciden con el valor 12, lo que revela una concentración de efectivos en los 12 años de escolarización. El tercer cuartil (P_{75}) es 15,75, y se ha calculado interpolando un valor aproximado entre el 15 y el 16:

$$P_{75} = (0,75 \times 15) + (0,25 \times 16) = 15,25$$

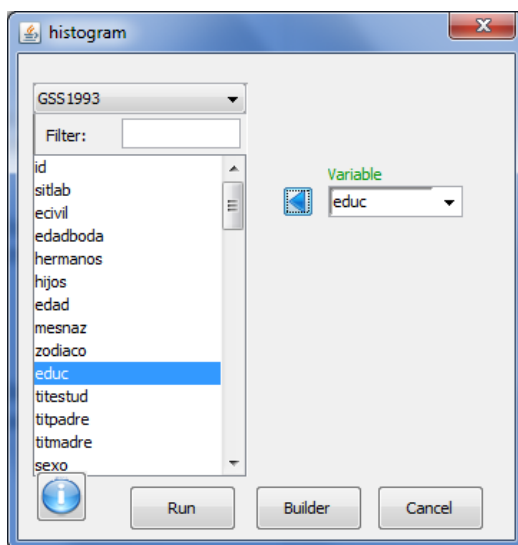
- La **media**, la suma de todas las observaciones dividida por número total de observaciones, es de **13,04**, un valor superior a los de la moda y la mediana. Este valor superior viene dado por la influencia de algunos individuos con muchos años de escolarización.
- Con el dato del estadístico de la **suma (19504)** podemos calcular la media:

$$\text{Media } \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^{1496} x_i}{1496} = \frac{19504}{1496} = 13,04$$

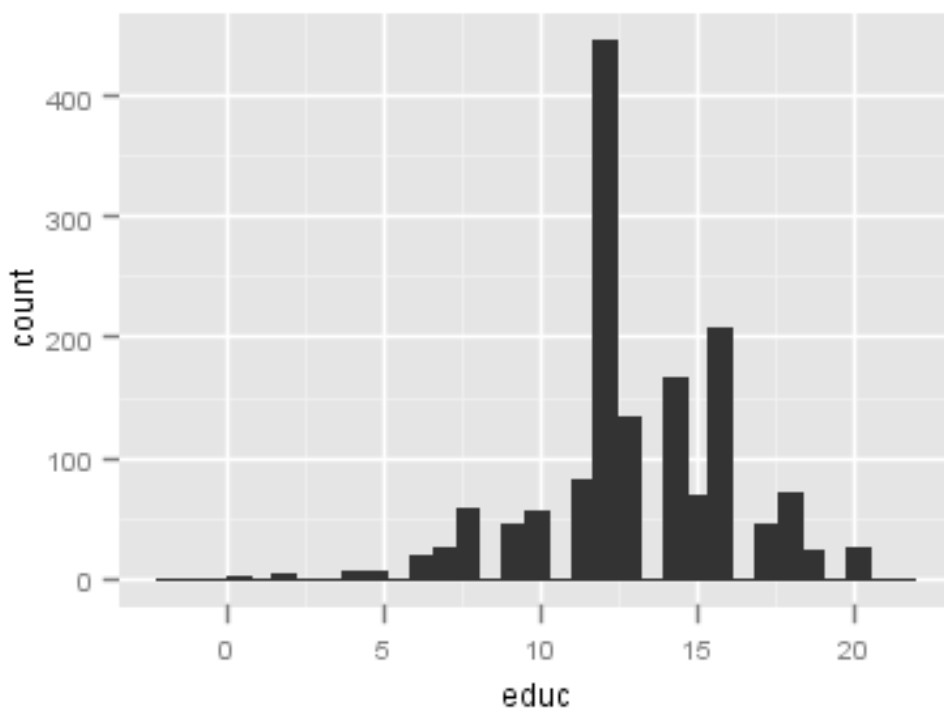
- La **varianza** es 9,45, y la **desviación típica** 3,07. Es decir, considerando las desviaciones de los valores respecto del valor medio (13,04), la media de todas estas desviaciones es de 3,07 años de escolarización. Puede comprobar con la calculadora que la desviación típica es la raíz cuadrada de la varianza.
- La medida de **asimetría** nos da el valor **-0,309**, que nos indica, por su valor negativo, la existencia de un cierto sesgo hacia la izquierda, es decir, la presencia de valores menos frecuentes a la izquierda y una mayor concentración de frecuencias en los valores medios o altos.

- La medida de **curtosis** de **0,708**, al ser positiva, nos informa de que se trata de una distribución apuntada en relación a la distribución normal (este aspecto se verá más adelante en el curso).

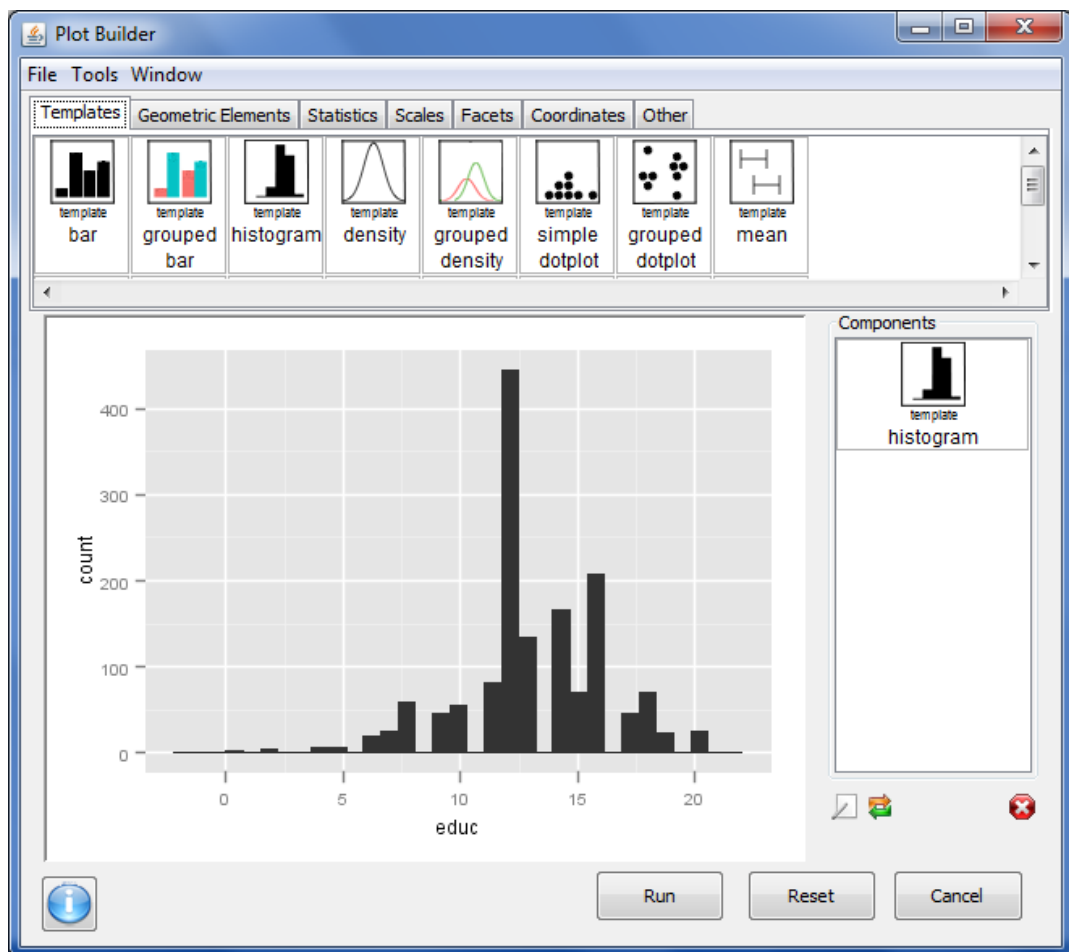
► En tercer lugar pediremos la representación gráfica adecuada para esta variable, el histograma. Desplegamos el menú *Plots* y elegimos, en primer lugar, *histogram* dentro del submenú *Templates*. Nos aparece el cuadro de diálogo donde seleccionaremos y pasaremos la variable **educ** en recuadro "Variable":



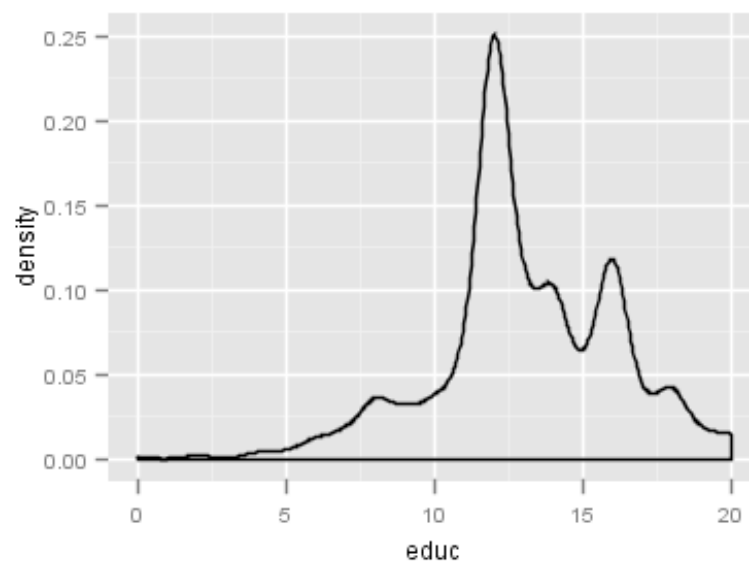
Clicamos sobre *Run* y se obtiene como resultado:



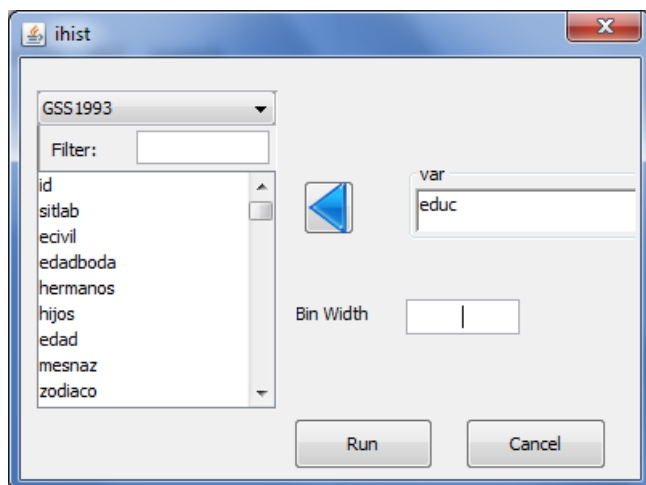
En lugar de ejecutar directamente la instrucción de obtención del gráfico podríamos optar por editarlo a través del *Plot Builder*:



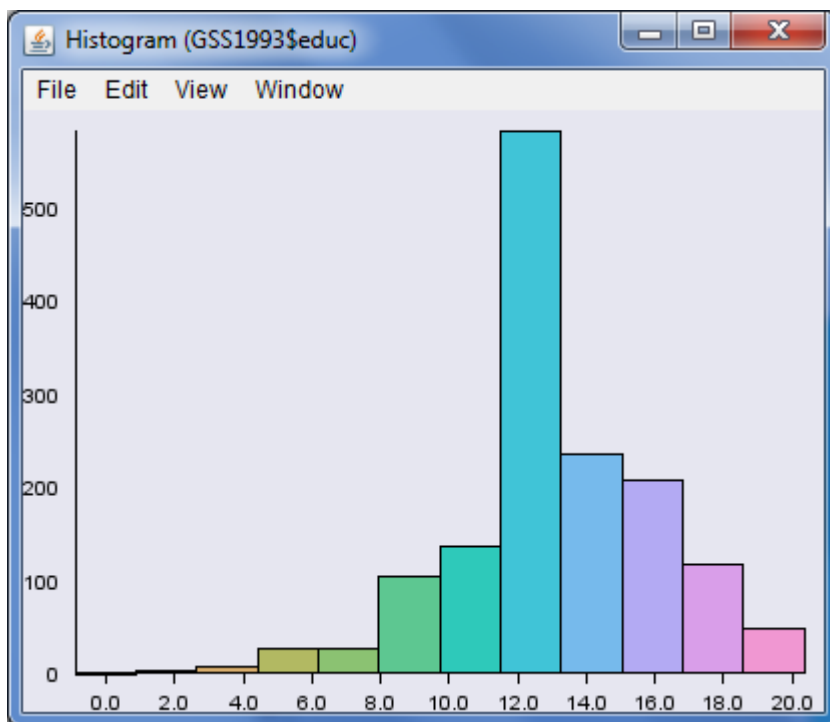
Un gráfico similar se puede obtener si pedimos un gráfico de "density" (menú *Plots / Templates / density*):



También podemos obtener un histograma interactivo a través del menú *Plots / Interactive / Histogram*:



La opción *Bin Width* nos permite elegir la anchura de los intervalos para dibujar los rectángulos del histograma. Lo podemos dejar en blanco y después interactivamente modificarla. Cuando se genere el gráfico a parecerá en color blanco y con una anchura que decide el software. Podemos cambiar la anchura con la flecha hacia arriba o la flecha hacia abajo, y también variar donde comienza el primer intervalo moviéndonos con las flechas de derecha e izquierda. Además, le podemos cambiar el color seleccionando a través del menú por ejemplo *View / Siete Colores (rainbow)*. Haciendo un ajuste a una anchura de 2 y con el color del arco iris se obtiene esta imagen:



► **Ejercicio 7**

Repetiremos el mismo análisis descriptivo con la variable *edad*.

Descriptive Statistics

	St.		Valid N	Median	25th Percentile	75th Percentile	Minimum	Maximum	Skew	Kurtosis
	Mean	Deviation								
edad	46.23	17.42	1495	43	32.50	59.00	18	89	0.50	-0.70

- Las frecuencias se refieren a 1495 casos de los 1500 totales, los 5 casos que corresponden a los valores perdidos de "No contesta", y no se tendrán en cuenta en los cálculos.
- Completa la información de las siguientes afirmaciones:
 - Vemos en primer lugar que el conjunto de valores varía entre el ____ (el valor **mínimo**) y el ____ (valor **máximo**). Por lo tanto, el recorrido o el rango es ____.
 - La **moda** (o modas), son los valores más frecuentes, ____ y ____ años, y corresponde a un total de ____ casos (el ____%). Recuerdese que con variables con muchos valores hay una tendencia a obtener frecuencias bajas y, por tanto, a encontrarnos más de una moda. En estos casos el estadístico deja de ser representativo de la distribución de la variable.
 - La **mediana**, valor tal que el 50% de observaciones son inferiores a él y el 50% superiores, es _____. Este valor corresponde al ____% acumulado y, por tanto, hasta ese valor hay acumulado el ____% de los individuos.
 - El **primer cuartil** es el valor ____ y el **tercer cuartil** es _____.
 - - ¿Cuál es el **percentil 10** (P_{10})? _____. Es decir, el ____% de la muestra tiene hasta ____ años.
 - - ¿Cuál es el **percentil 90** (P_{90})? _____. Es decir, el ____% de la muestra tiene hasta ____ años.
 - La **media**, la suma de todas las observaciones dividida por número total de observaciones, es de _____, un valor más _____ que la mediana, por la influencia de algunos individuos con _____ años. Puede comprobar que la media es la suma _____ dividida por el total de individuos _____.
 - La **varianza** es _____, y la **desviación típica** _____. ¿En qué unidad se expresa la desviación? _____ ¿Y la varianza? _____.
 - La medida de **asimetría** nos da el valor _____, que nos indica la existencia de un sesgo hacia _____.

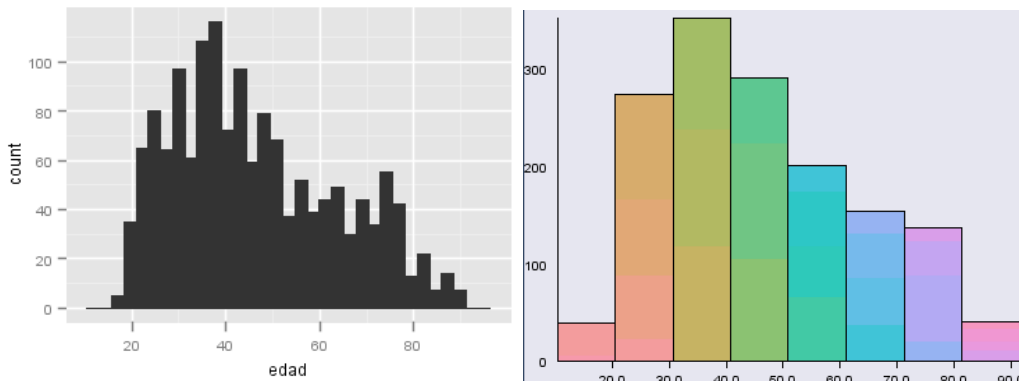
Frecuencias (edad)

	Value	# of Cases	%	Cumulative %						
	1	18	5	0.30	0.30	36	53	18	1.20	68.50
	2	19	17	1.10	1.50	37	54	19	1.30	69.80
	3	20	18	1.20	2.70	38	55	22	1.50	71.20
	4	21	22	1.50	4.10	39	56	12	0.80	72.00
	5	22	15	1.00	5.20	40	57	18	1.20	73.20
	6	23	28	1.90	7.00	41	58	25	1.70	74.90
	7	24	23	1.50	8.60	42	59	14	0.90	75.90
	8	25	30	2.00	10.60	43	60	16	1.10	76.90
	9	26	27	1.80	12.40	44	61	11	0.70	77.70
	10	27	22	1.50	13.80	45	62	17	1.10	78.80
	11	28	42	2.80	16.70	46	63	19	1.30	80.10
	12	29	30	2.00	18.70	47	64	13	0.90	80.90
	13	30	36	2.40	21.10	48	65	17	1.10	82.10
	14	31	31	2.10	23.10	49	66	19	1.30	83.30
	15	32	28	1.90	25.00	50	67	11	0.70	84.10
	16	33	33	2.20	27.20	51	68	16	1.10	85.20
	17	34	25	1.70	28.90	52	69	19	1.30	86.40
	18	35	41	2.70	31.60	53	70	9	0.60	87.00
	19	36	42	2.80	34.40	54	71	15	1.00	88.00
	20	37	37	2.50	36.90	55	72	19	1.30	89.30
	21	38	41	2.70	39.70	56	73	20	1.30	90.60
	22	39	38	2.50	42.20	57	74	18	1.20	91.80
	23	40	36	2.40	44.60	58	75	17	1.10	93.00
	24	41	36	2.40	47.00	59	76	13	0.90	93.80
	25	42	30	2.00	49.00	60	77	15	1.00	94.80
	26	43	39	2.60	51.60	61	78	14	0.90	95.80
	27	44	28	1.90	53.50	62	79	7	0.50	96.30
	28	45	30	2.00	55.50	63	80	6	0.40	96.70
	29	46	29	1.90	57.50	64	81	9	0.60	97.30
	30	47	32	2.10	59.60	65	82	10	0.70	97.90
	31	48	20	1.30	60.90	66	83	3	0.20	98.10
	32	49	27	1.80	62.70	67	84	3	0.20	98.30
	33	50	21	1.40	64.10	68	85	4	0.30	98.60
	34	51	26	1.70	65.90	69	86	5	0.30	98.90
	35	52	21	1.40	67.30	70	87	6	0.40	99.30
						71	88	3	0.20	99.50
						72	89	7	0.50	100.00

Case Summary (edad)

	Valid	Missing	Total	% Missing
1	1495.00	5.00	1500.00	0.30

- El **histograma** que genera el software se puede representar de estas formas:



► Cuando disponemos de la información de la dispersión (varianza y desviación típica) de dos variables que se calculan a partir medias diferentes, no tiene sentido comparar directamente con estos estadísticos. El índice adecuado es el **coeficiente de variación**, que es una medida de dispersión relativa que se define como el cociente entre la desviación típica y la media, multiplicado por 100:

$$CV = \frac{s}{\bar{x}} \times 100$$

Consideramos la comparación de las variables **educ** y **edad**. Podemos pedir los estadísticos necesarios de ambas variables:

Descriptive Statistics

	Mean	St. Deviation	Valid N
educ	13.04	3.07	1496
edad	46.23	17.42	1495

Los coeficientes de variación serán:

$$CV(educ) = \frac{s}{\bar{x}} \times 100 = \frac{3,074}{13,04} \times 100 = 23,54\%$$

$$CV(edad) = \frac{s}{\bar{x}} \times 100 = \frac{17,418}{46,23} \times 100 = 37,68\%$$

Con estos resultados podemos afirmar que la variable *edad* tiene una mayor dispersión relativa.

Con R lo podemos calcular con la siguiente instrucción en el caso de la **edad**:

```
>sd(GSS1993$edad,na.rm=TRUE)/mean(GSS1993$edad,na.rm=TRUE)*100
```

► Ejercicio 8

Se pueden repetir los análisis anteriores con otras variables cuantitativas de la matriz **GSS1993.rda**: *hijos*, *edadboda*, *indsocec*, *horastv*.

► Ejercicio 9

Con la matriz de datos **Victimitzacio2008.rda** se pueden reproducir los mismos análisis descriptivos con las variables: *edat_ps*, i amb *ideologi*, *ps1_a* i *osp8_a* consideradas como cuantitativas.

2.2. Anàlisis exploratorio de una variable cuantitativa

El anàlisis descriptivo que acabamos de ver se puede completar y complementar con otros estadísticos y gràficos que permiten en particular realizar un anàlisis exploratorio destinada a inspeccionar los datos, identificar valores atípicos, obtener descripciones, comprobar supuestos de las variables y caracterizar diferencias entre subpoblaciones (grupos de casos).

En esta pràctica utilizaremos algunos de estos instrumentos con una finalidad de descripci3n de las característic3s de la distribuci3n de una variable cuantitativa, y para obtener un tipo de representaci3n gràfica de inter3s: el **diagrama de caja**.

La exploraci3n de los datos se puede complementar con dos estadísticos adicionales a los que hemos visto: la **media recortada** y el **rango intercuartil** (tambi3n llamado amplitud o desviaci3n intercuartil).

- La **media recortada al 5%** que es la media aritm3tica calculada despu3s de haber eliminado el 5% de los casos m3s grandes y el 5% de los menores (los casos extremos), lo que proporciona una mejor estimaci3n de la tendencia central. Con las instrucciones de R:

```
> mean(GSS1993$educ,na.rm=TRUE,trim=5/100)
[1] 13.10163
```

```
> mean(GSS1993$edad,na.rm=TRUE,trim=5/100)
[1] 45.64959
```

Comprobamos c3mo la media de la variable *educ* (13,04) cambia a 13,10, un valor ligeramente inferior, cuando se recorta. En el caso de la variable *edad* baj de 46,25 a 45,67.

- La amplitud o el **rango intercuartil** (RI) que es una medida de la dispersi3n de los datos. Es la distancia entre el tercer cuartil (el percentil 75) y primer cuartil (el percentil 25).

Para recordar los valores de los cuartiles podemos ejecutar las instrucciones:

```
> summary(GSS1993$educ)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.   NA's
 0.00  12.00  12.00  13.04  15.25  20.00    4
```

```
> summary(GSS1993$edad)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.   NA's
18.00  32.50  43.00  46.23  59.00  89.00    5
```

Por tanto, se obtienen los resultados siguientes:

$$RI(educ) = P_{75} - P_{25} = 15,25 - 12 = 3,25$$

$$RI(edad) = P_{75} - P_{25} = 59 - 32,5 = 26,5$$

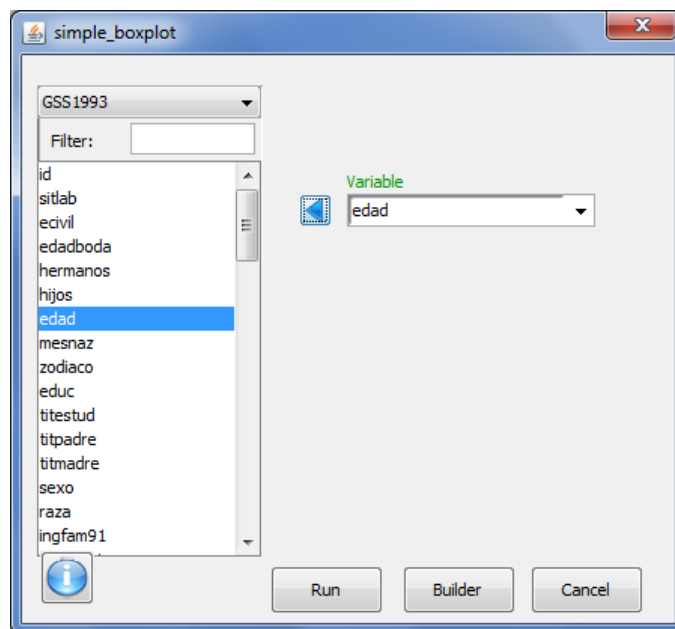
Con las instrucciones de R verificamos estos resultados:

```
> IQR(GSS1993$educ,na.rm=TRUE)
[1] 3.25
```

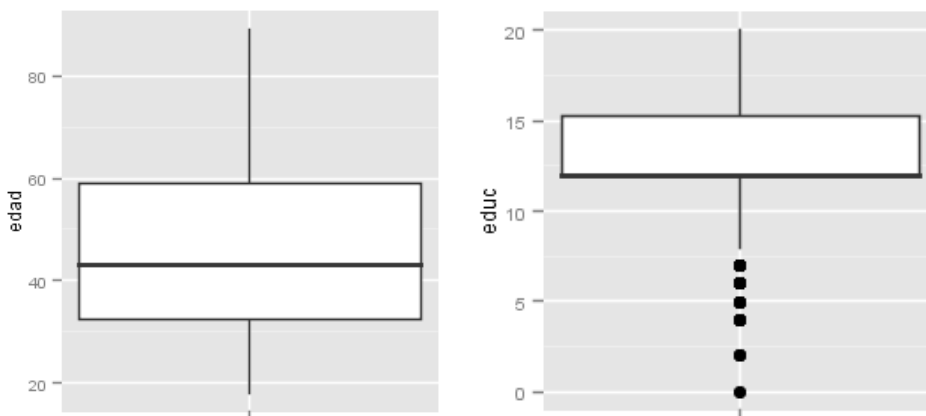
```
> IQR(GSS1993$edad,na.rm=TRUE)
[1] 26.5
```

- El **diagrama de caja** (*box plot*) es otra representación gráfica de interés que nos proporciona los rasgos característicos de la distribución de los datos, de posición, dispersión y simetría, en este caso referidos a las variables **educ** y **edad**. Para obtener un gráfico como este por el menú podemos ir a *Plots / Templates / simple boxplot*.

Nos aparece este cuadro de diálogo donde hemos elegido la variable **edad**:



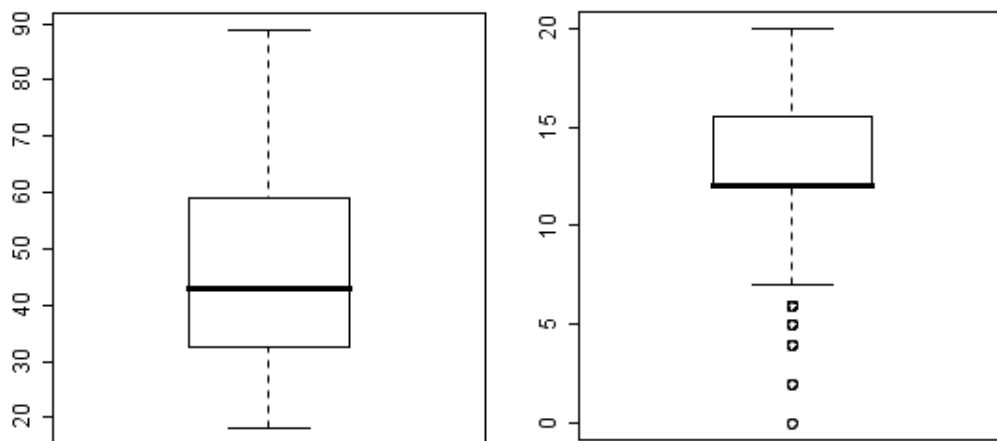
Si ejecutamos el procedimiento obtenemos estos resultados, también para la variable **educ**:



También podemos ejecutar directamente el comando `boxplot` de R desde la línea de instrucciones de la consola. Pedimos el gráfico de cada variable por separado con:

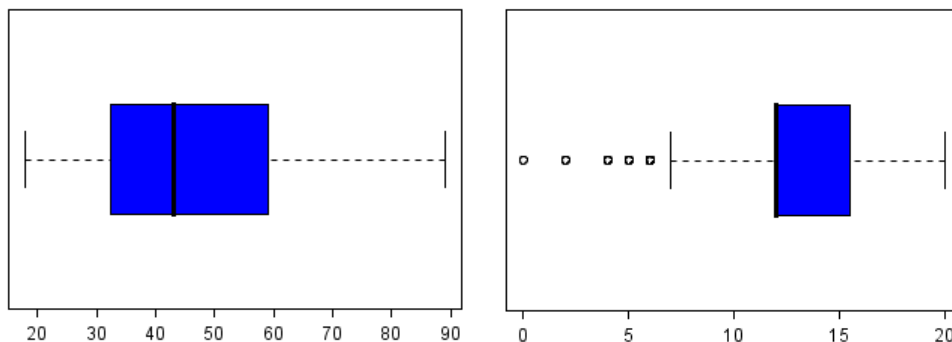
```
> boxplot(GSS1993$educ)
> boxplot(GSS1993$edad)
```

Con estos resultados²:



Podemos adicionalmente completar la instrucción para obtener el gráfico horizontal y darle color azul, por ejemplo:

```
> boxplot(GSS1993$edad,col="blue",horizontal=TRUE)
> boxplot(GSS1993$educ,col="blue",horizontal=TRUE)
```



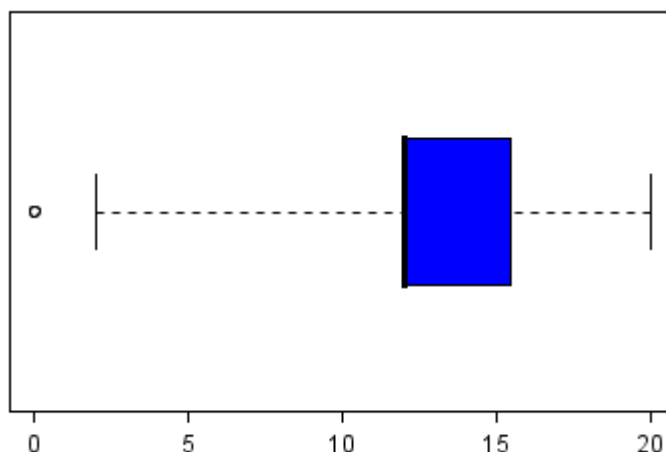
- En el primer diagrama podemos observar la falta de valores extremos o atípicos (*outliers*) que sí aparecen en el segundo. Los valores extremos que se identifican con el símbolo **o** corresponden a aquellos casos que se sitúan a una distancia de los límites de la caja superior a 1,5 veces la desviación intercuartil (diferencia entre el tercer y el primer cuartil, la distancia que determina la caja).
- Los *outliers* **severos** corresponden a los casos situados a una distancia por encima de 3 veces la desviación intercuartil. Para determinar qué valores extremos son

² Si las dos variables tuvieran la misma escala se podrían pedir conjuntamente poniendo ambas variables dentro del paréntesis separadas por coma.

severos hay que especificarlo cambiando el valor por defecto de 1,5 por 3, mediante la opción *range*. Para la variable **educ**:

```
> boxplot(GSS1993$educ,col="blue",horizontal=TRUE,range=3)
```

El nuevo gráfico muestra cómo hay un solo caso más allá de 3 veces el rango intercuartil:



- La mediana de la edad es el valor 43, y se identifica por la línea gruesa de la caja. En el caso de los años de escolarización, la mediana 12 se sitúa en el extremo inferior de la caja, que corresponde también al primer cuartil; esto nos muestra la concentración de efectivos en este valor.
- Se puede comprobar, por tanto, la mayor simetría de la variable relativa a la edad y el comportamiento más atípico de algunos valores bajos y altos de la variable de escolarización.

► Ejercicio 10

Repetir los análisis anteriores con otras variables cuantitativas de la matriz **GSS1993.rda**: *hijos*, *edadboda*, *indsocec*, *horastv*.

► Ejercicio 11

Con la matriz de datos **Victimitzacio2008.rda** se pueden reproducir los mismos análisis exploratorios con las variables: *edat_ps*, *ideologi*, *ps1_a* y *osp8_a*.