Estadística descriptiva bidimensional con R Commander

Tablas estadísticas bidimensionales

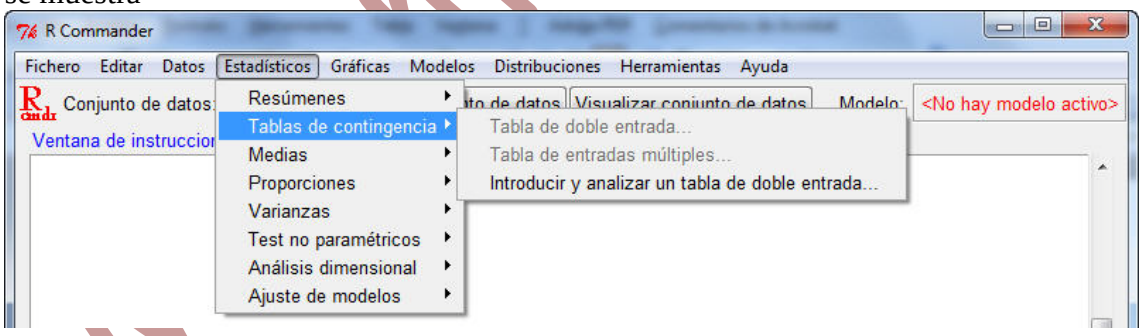
En el menú

Estadísticos
Tablas de contingencia…

podemos seleccionar
Tabla de doble entrada…
…
Introducir y analizar una tabla de doble entrada…

La opción **"Tabla de doble entrada"** puede no estar disponible si el conjunto de datos activo no contiene al menos dos variables de tipo carácter.
Si se desea ejecutar esta opción con variables numéricas de tipo discreto, debemos convertir la variable numérica en factor de la siguiente forma:
Datos
Modificar variables del conjunto de datos activo
Convertir variable numérica en factor …

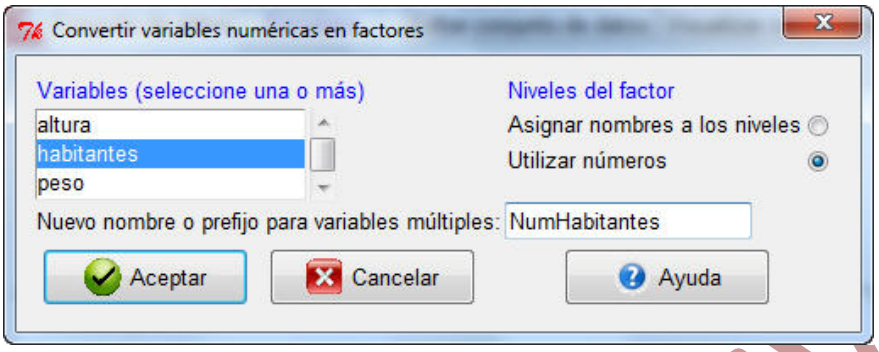
Practicamos con el siguiente ejemplo:
Utilizando el conjunto de datos activo "Encuesta" proveniente de haber importado el fichero *encues.txt*, si ejecutamos
Estadísticos
Tablas de contingencia…
se muestra



es decir, no nos permite realizar una Tabla de doble entrada con dicho fichero ya que sólo contiene una variable de tipo carácter, "estudios" (nivel de estudios). Si quisiéramos hacer una tabla de doble entrada con las variables "estudios" y "habitantes" (variable numérica que corresponde al número de personas con las que se convive y que toma los valores 2, 3, 4, 5, 6 y 8) hay que reconvertir primero la variable "habitantes" en una de tipo carácter ejecutando

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

y seleccionando la variable “habitantes” en la primera ventana, marcando la opción “Utilizar números” en Niveles de factor y escribiendo un nuevo nombre para la variable. En este caso, hemos elegido “NumHabitantes”:



Ahora al ejecutar

Estadísticos

Tablas de contingencia…

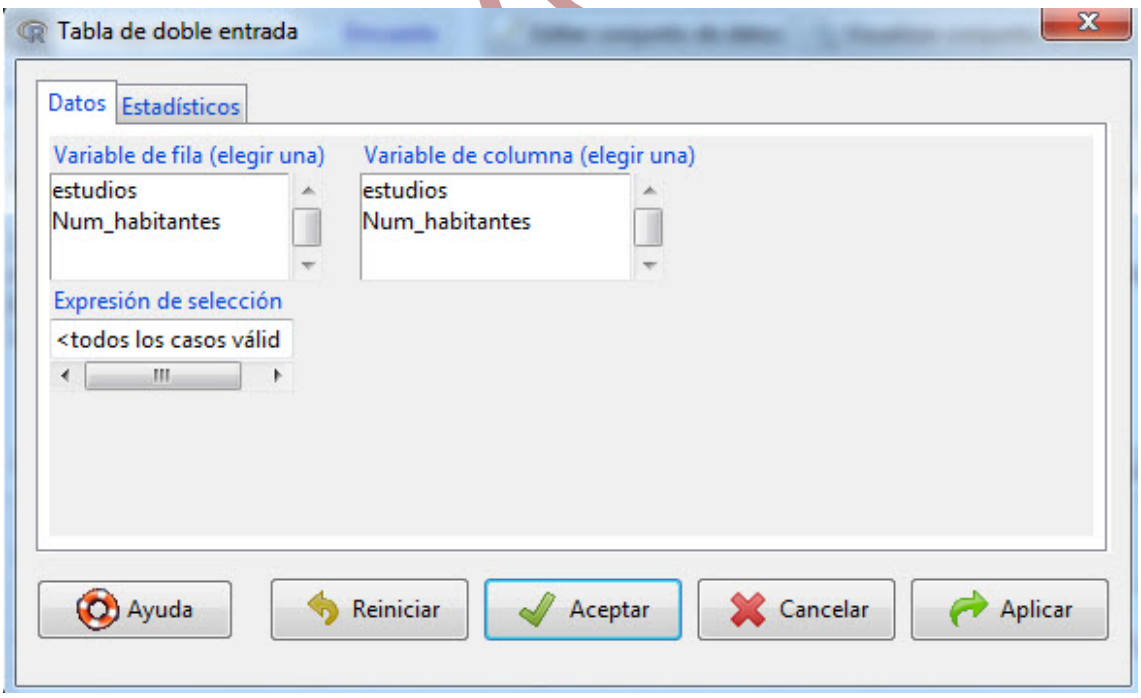
sí aparece activa la opción *Tabla de doble entrada*.

El menú

Estadísticos

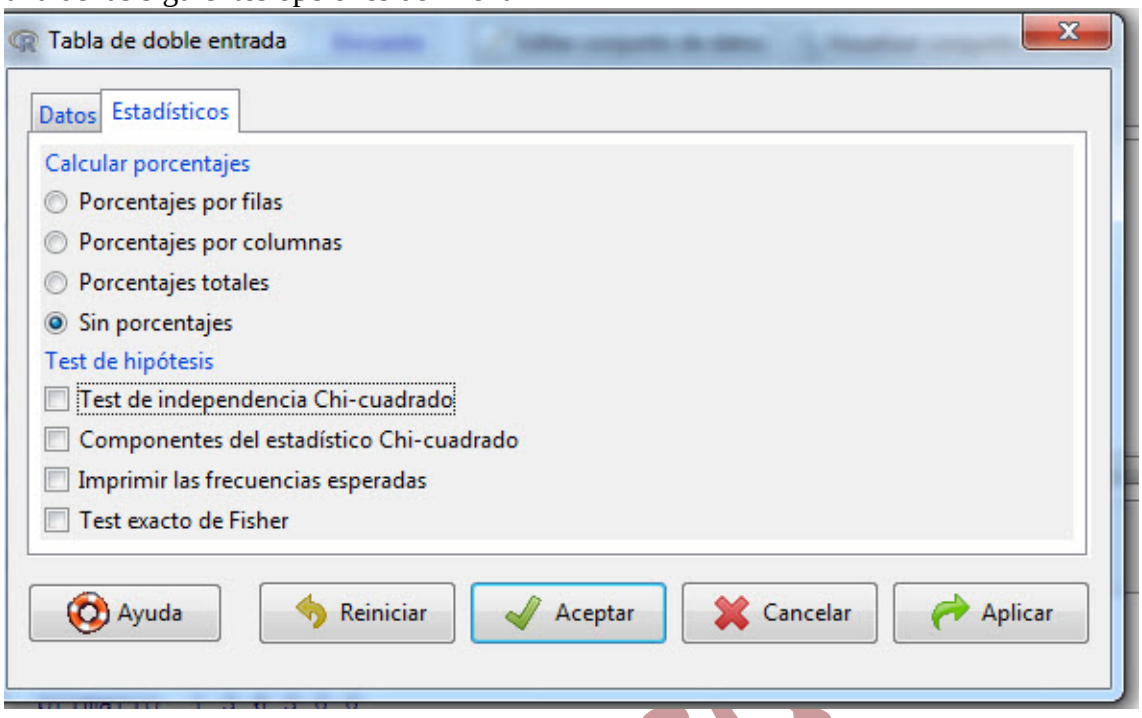
Tablas de contingencia…

Tabla de doble entrada…



permite la realización de una tabla de doble entrada seleccionando en la primera ventana la variable que queramos incluir por filas y en la segunda la que se quiera incluir por columnas

Si pinchamos en “Estadísticos” se muestra la siguiente ventana donde se puede marcar una de las siguientes opciones del menú



Calcular porcentajes:

- **Porcentaje por filas**: Proporciona una tabla de doble entrada con las frecuencias absolutas y una tabla que incluye la distribución marginal de la variable que se ha incluido por filas junto con los porcentajes (100*frecuencia relativa) de la distribución de la variable incluida por columnas condicionada a la incluida por filas.
- **Porcentaje por columnas**: Proporciona una tabla de doble entrada con las frecuencias absolutas y una tabla que incluye la distribución marginal de la variable que se ha incluido por columnas junto con los porcentajes (100*frecuencia relativa) de la distribución de la variable incluida por filas condicionada a la incluida por columnas.
- **Porcentajes totales**: Proporciona una tabla de doble entrada con las frecuencias absolutas y una tabla con los porcentajes (100*frecuencia relativa) de la distribución conjunta.
- **Sin porcentajes**: Proporciona una tabla de doble entrada con las frecuencias absolutas.

Veamos los resultados de ejecutar cada una de las opciones con el conjunto de datos activo *Encuesta* y las variables “estudios” y “NumHabitantes”. Notemos que se ha deshabilitado la opción *Test de independencia Chi-Cuadrado* del menú Test de hipótesis.

Opción *Sin porcentajes*

```
> .Table
                NumHabitantes
```

estudios	2	3	4	5	6	8
bachiller	3	4	6	4	3	0
diplomado	1	2	3	3	1	0
fp	0	3	0	1	0	1
primario	1	3	6	3	0	0
superior	0	1	1	0	0	0

En esta opción (ni en las siguientes) no mostramos la tabla de doble entrada con las frecuencias absolutas, que aparece en todas las opciones

Opción *Porcentajes totales*

> totPercents(.Table) # Percentage of Total

	2	3	4	5	6	8	Total
bachiller	6	8	12	8	6	0	40
diplomado	2	4	6	6	2	0	20
fp	0	6	0	2	0	2	10
primario	2	6	12	6	0	0	26
superior	0	2	2	0	0	0	4
Total	10	26	32	22	8	2	100

Opción *Porcentaje por filas*

> rowPercents(.Table) # Row Percentages

	NumHabitantes							
estudios	2	3	4	5	6	8	Total	Count
bachiller	15.0	20.0	30.0	20.0	15	0	100.0	20
diplomado	10.0	20.0	30.0	30.0	10	0	100.0	10
fp	0.0	60.0	0.0	20.0	0	20	100.0	5
primario	7.7	23.1	46.2	23.1	0	0	100.1	13
superior	0.0	50.0	50.0	0.0	0	0	100.0	2

Opción *Porcentaje por columnas*

> colPercents(.Table) # Column Percentages

	NumHabitantes					
estudios	2	3	4	5	6	8
bachiller	60	30.8	37.5	36.4	75	0
diplomado	20	15.4	18.8	27.3	25	0
fp	0	23.1	0.0	9.1	0	100
primario	20	23.1	37.5	27.3	0	0
superior	0	7.7	6.2	0.0	0	0
Total	100	100.1	100.0	100.1	100	100
Count	5	13.0	16.0	11.0	4	1

¡Atención a los errores de redondeo!

* Se ha comentado antes que el menú *Tabla de doble entrada* no permite la consideración de variables numéricas; sin embargo, puede hacerse una tabla de doble entrada con variables de este tipo escribiendo directamente el código de R. Por ejemplo, en la opción *Sin porcentajes* el código usado ha sido:

```
.Table <- xtabs(~estudios+NumHabitantes, data=Encuesta)
.Table
```

Si ejecutamos el código

```
.Table <- xtabs(~estudios+habitantes, data=Encuesta)
.Table
```

donde se ha cambiado la variable "NumHabitantes" por "habitantes", el resultado es

	habitantes					
estudios	2	3	4	5	6	8
bachiller	3	4	6	4	3	0
diplomado	1	2	3	3	1	0
fp	0	3	0	1	0	1
primario	1	3	6	3	0	0
superior	0	1	1	0	0	0

Esta es una alternativa a la solución planteada antes de *Convertir una variable numérica en factor*

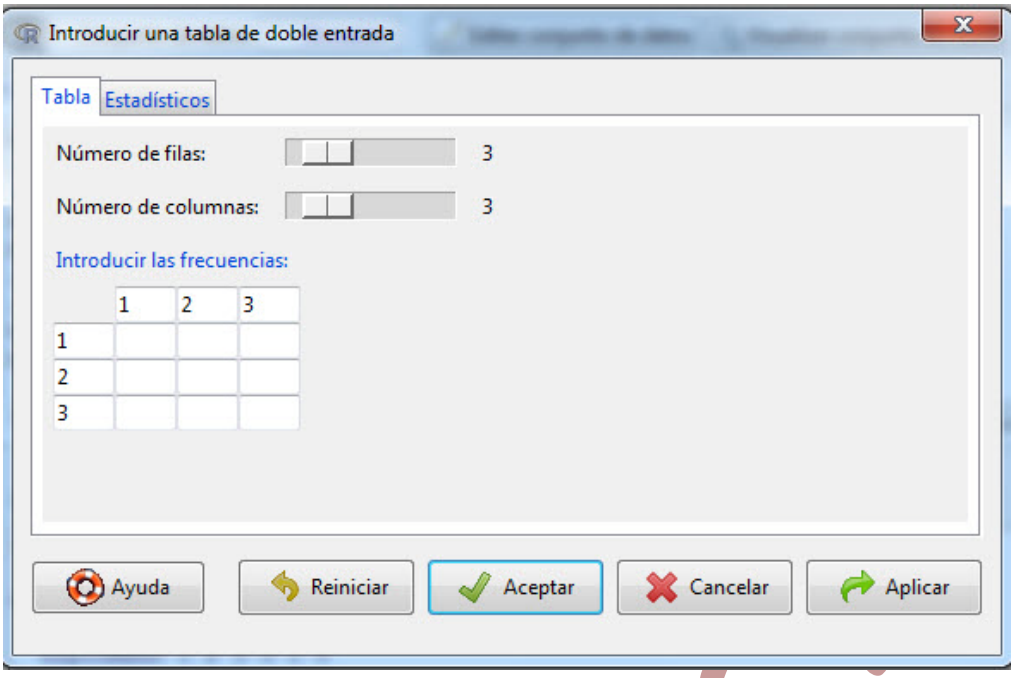
El menú

Estadísticos

Tablas de contingencia…

Introducir y analizar una tabla de doble entrada…

permite el mismo análisis del menú comentado con anterioridad pero con una tabla de doble entrada que se introduce por pantalla. La ventana que resulta de ejecutar este menú es



se deben incluir los datos relativos a la tabla que queramos analizar y en la opción "Estadísticos" tenemos las mismas posibilidades que en el caso anterior.

Los menús

Estadísticos
Resúmenes
Resúmenes numéricos

Estadísticos
Resúmenes
Distribuciones de frecuencias

y

Gráficas

estudiados en el caso de distribuciones unidimensionales, se pueden aplicar a distribuciones marginales y condicionadas.
Para ello, si los datos proceden de una tabla de doble entrada, se deben reconvertir en un fichero que R puede analizar. Esto ya se vio en los apuntes previos.

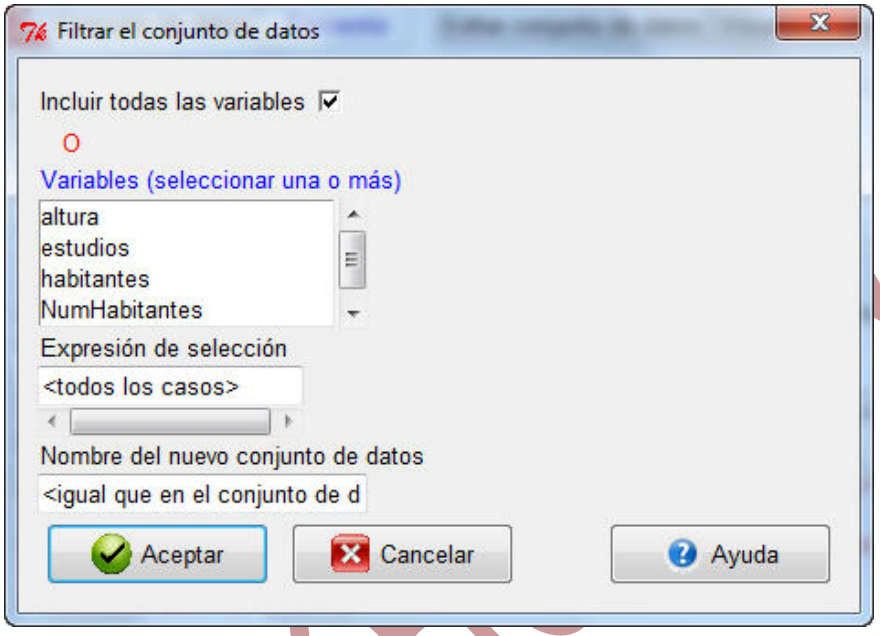
A partir de un fichero de datos de R donde los datos asociados a cada individuo se muestran por filas y las variables por columnas:

- Para hacer un análisis sobre una variable concreta (distribución marginal), basta seleccionar la variable correspondiente.

- Para hacer un análisis sobre una distribución condicionada, debemos ejecutar en primer lugar el menú

Datos
Conjunto de datos activo
Filtrar el conjunto de datos activo…

Por ejemplo, al ejecutarlo con el conjunto de datos activo *Encuesta*, se muestra la siguiente ventana



en la que debemos
- marcar o desmarcar la opción *Incluir todas las variables*
- seleccionar la variable (o variables) a la que se quiere condicionar
- incluir la expresión de la selección, que debe ser una expresión lógica. Para ello se deben usar los operadores lógicos

<	menor
<=	menor o igual
>	mayor
>=	mayor o igual
==	igual
!=	distinto
&	intersección (“y”)
|	unión (“o”)
!	negación (“no”)

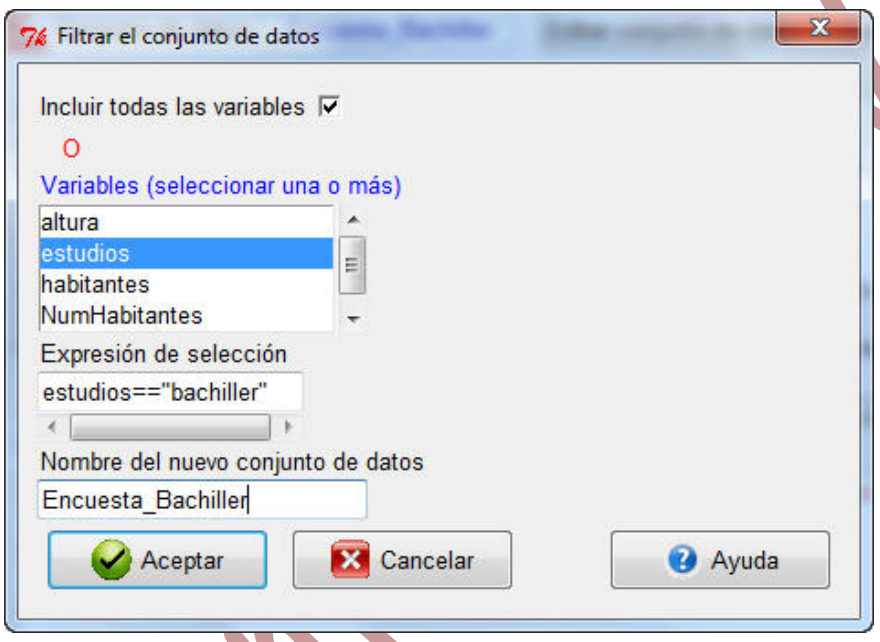
Además, si la variable es de tipo carácter su valor debe ir incluido entre comillas y ser idéntico a una de las posibles modalidades (¡atención al uso de mayúsculas, minúsculas, acentos, etc!)
- escribir un nuevo nombre para el nuevo conjunto de datos (si se mantiene el mismo nombre, se perderá el fichero anterior).

Con el nuevo fichero, cuando se realice el análisis numérico de una variable se estará haciendo considerando la distribución condicionada de dicha variable.

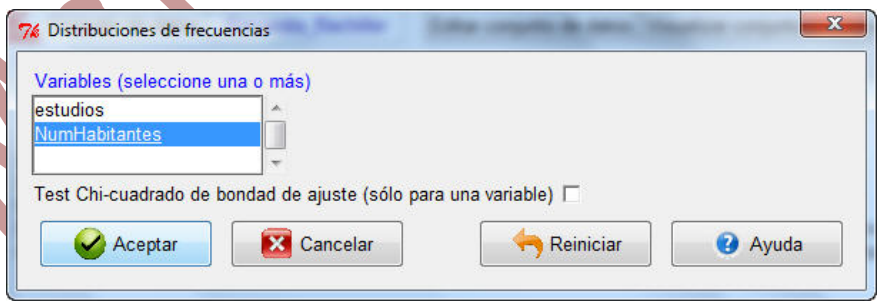
Ejemplo

Con el conjunto de datos activo *Encuesta*, calcular tabla de frecuencias, media, desviación típica y cuartiles de la variable Número de habitantes para aquellos individuos cuyo nivel de estudios es bachiller.
Los pasos a seguir serían

Datos
Conjunto de datos activo
Filtrar el conjunto de datos activo…



y con el conjunto de datos activo *Encuesta_Bachiller* ir a **Resúmenes, Distribuciones de frecuencias**



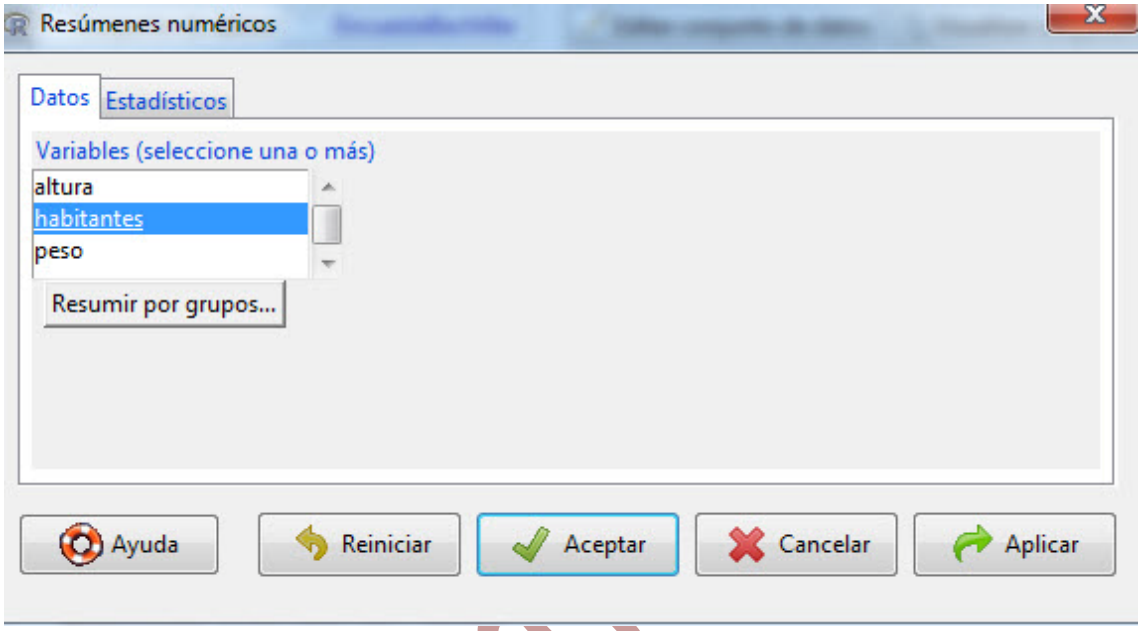
cuyo resultado es

```
> .Table <- table(Encuesta_Bachiller$NumHabitantes)

> .Table  # counts for NumHabitantes

2  3  4  5  6  8
3  4  6  4  3  0
```

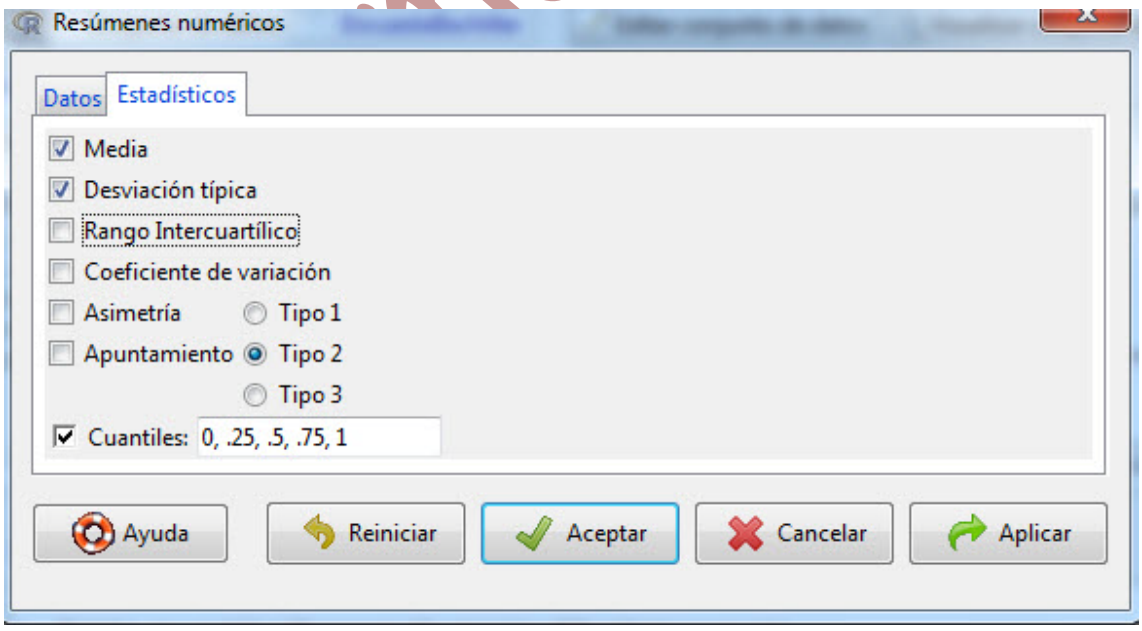
```
> round(100*.Table/sum(.Table), 2)  # percentages for NumHabitantes

 2  3  4  5  6  8
15 20 30 20 15  0
```

y

Estadísticos, Resúmenes, Resúmenes numéricos



Pinchar en Estadísitcos



cuyo resultado es

mean	sd	0%	25%	50%	75%	100%	n
4	1.297771	2	3	4	5	6	20

Notemos que en un caso se ha usado la variable “habitantes” y en otro “NumHabitantes”, dependiendo de la exigencia del tipo de variable. La tabla de frecuencias se podría haber hecho con la variable “habitantes” usando código.

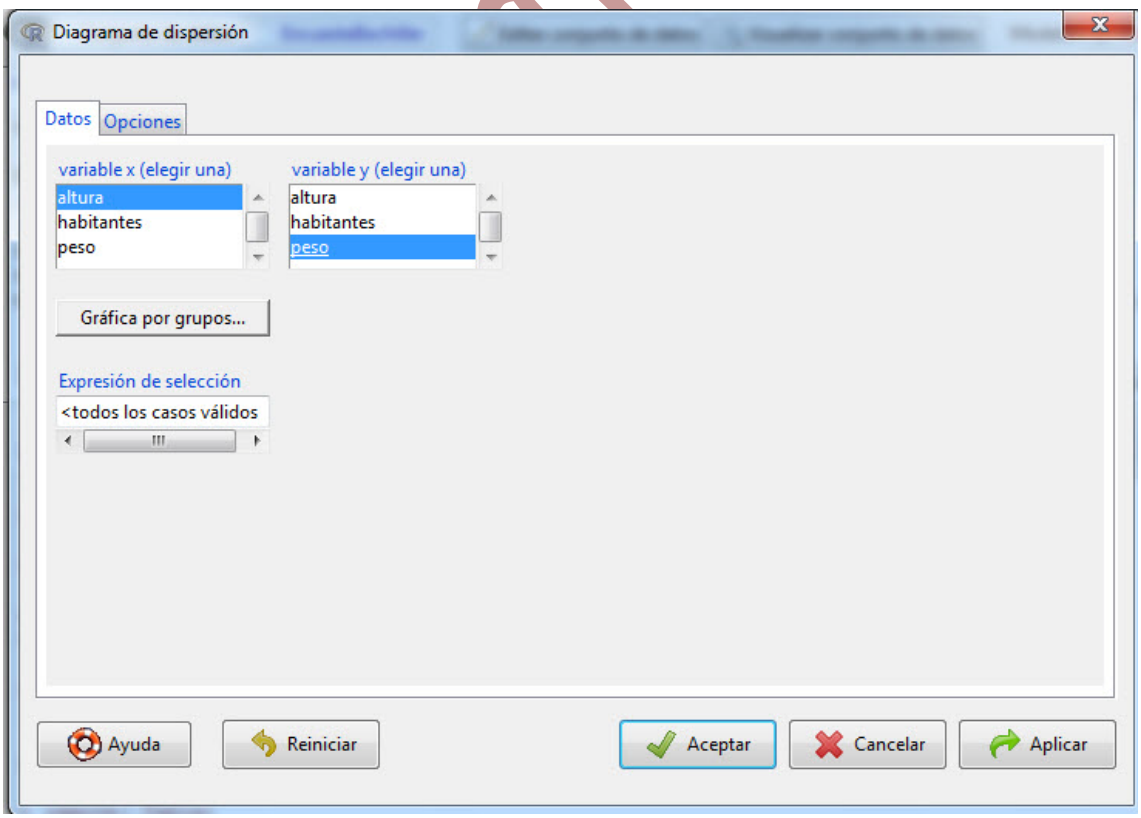
Recordemos también que los Resúmenes Numéricos se pueden realizar con la opción *Resumir por grupos…* y se haría un estudio de una o varias variables numéricas para cada una de las modalidades de la variable carácter seleccionada. El procedimiento previo permite el condicionamiento al valor de una variable numérica y a condiciones más complejas, por ejemplo, habitantes<=3 & habitantes >=6.

Gráficos bidimensionales

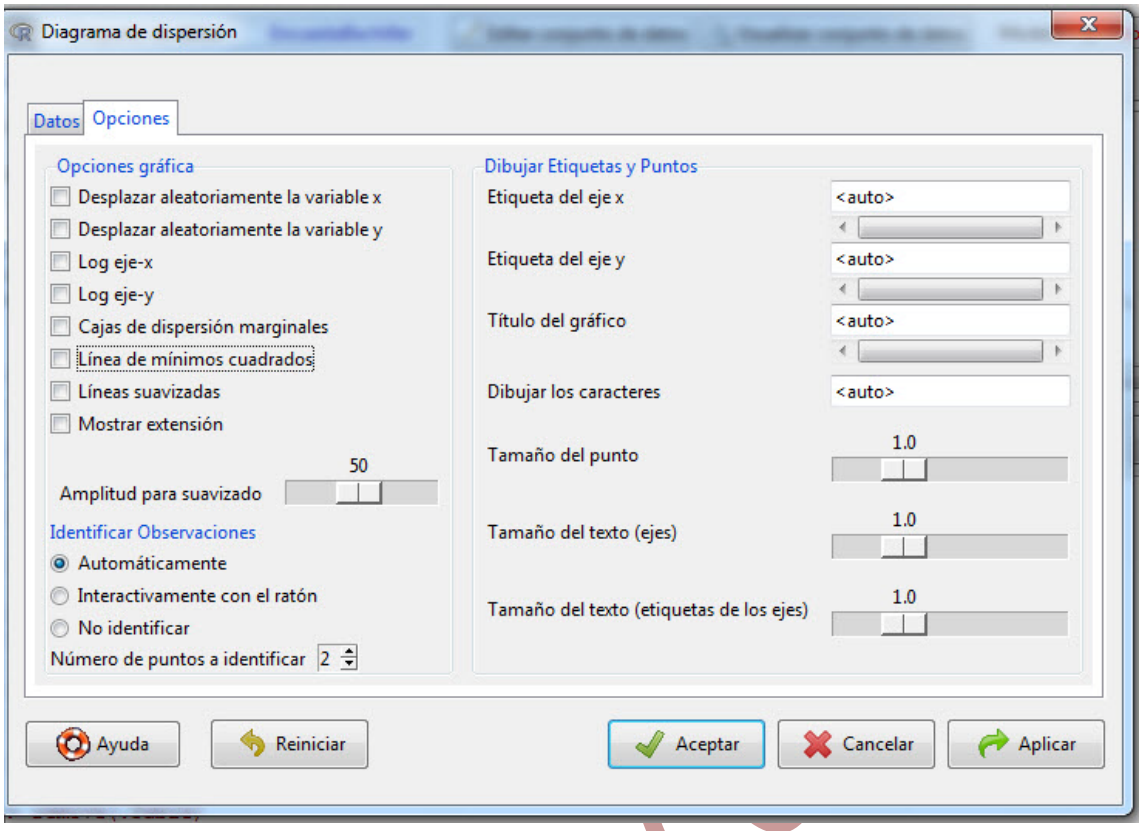
Gráficas

Diagrama de dispersión …

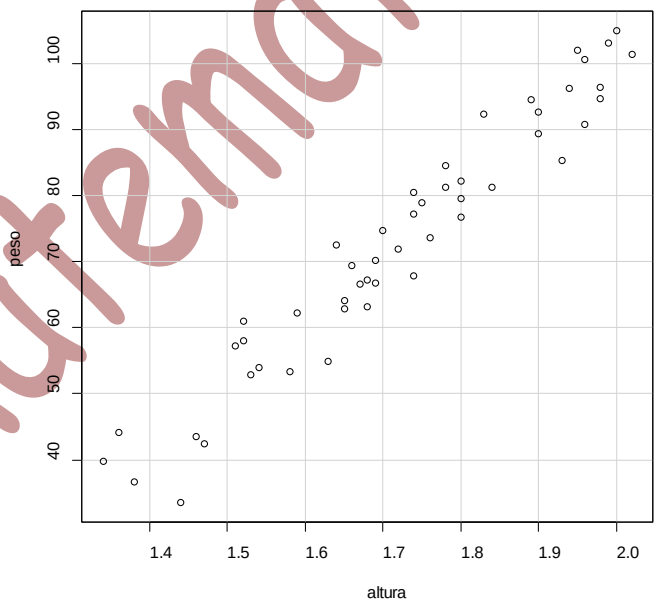
En la ventana que se muestra nos permite elegir, de entre las variables numéricas disponibles en el conjunto de datos activo, una variable para el eje de abscisas (x) y otra para el eje de ordenadas (y).
Por ejemplo, con el conjunto de datos activo *Encuesta*, seleccionando las variables “altura” y “peso”, respectivamente, y desmarcando todas las opciones



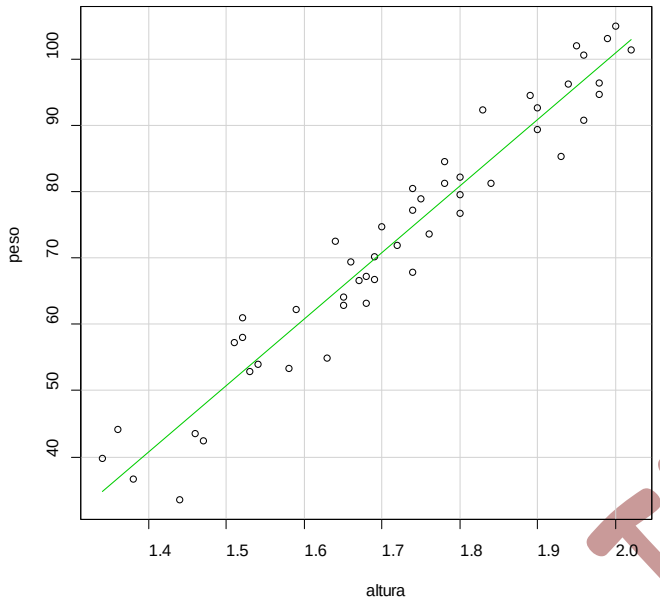
A continuación pinchar en opciones



se obtiene



Si se marca la opción *Línea de mínimos cuadrados* se muestra



Regresión y correlación

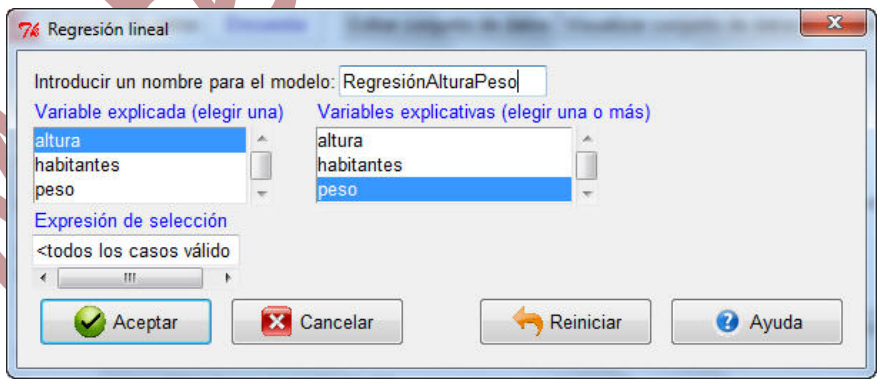
Cálculo de la recta de regresión mínimo cuadrática

Estadísticos

Ajuste de modelos

Regresión lineal …

En la ventana que se muestra se debe escribir un nombre para el modelo que se va a construir, seleccionar una variable numérica como variable explicada y una o más como variables explicativas (en este curso, sólo se ha estudiado el caso de una variable explicativa). Por ejemplo, para las variables “altura” y “peso” del conjunto de datos activo *Encuesta*, al ejecutar



se obtiene

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.090895 -0.029947 -0.001555 0.033607 0.094071

Coefficients:
```

```
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  1.0382612  0.0265428   39.12    <2e-16 ***
peso         0.0093557  0.0003521    26.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04679 on 48 degrees of freedom
Multiple R-squared: 0.9363,  Adjusted R-squared: 0.935
F-statistic: 705.9 on 1 and 48 DF,  p-value: < 2.2e-16
```

Algunos de los resultados que se muestran requieren conocimientos más avanzados que los estudiados en este curso para su interpretación. Si la recta que se ajusta es del tipo

$$y=a+bx$$

donde y es la "altura" y x el "peso"
Nos van a interesar

- el valor del coeficiente **a** (intercept o punto de corte con el eje de ordenadas).
- el valor del coeficiente **b** (la pendiente de la recta).
- el valor de $\mathbf{R^2}$ (coeficiente de determinación).

Así, la recta de mínimos cuadrados viene dada por

$$y= 1.0382612 + 0.0093557\ x$$

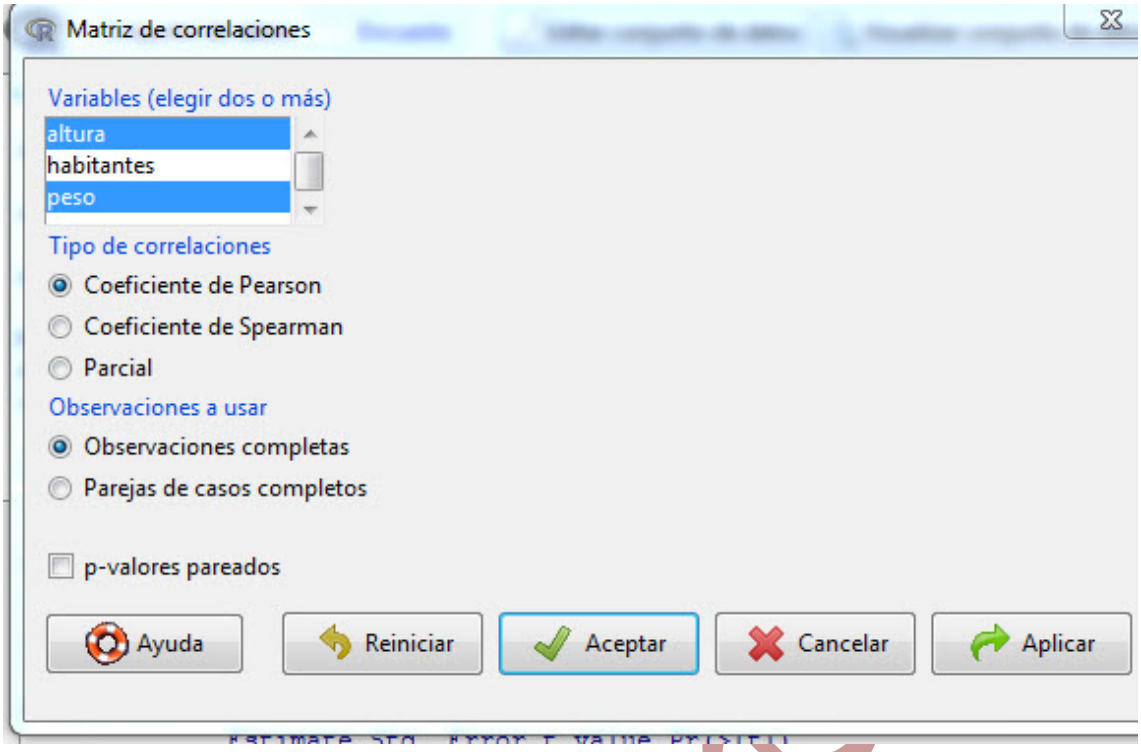
con un valor de $R^2 = 0.9363$.

Cálculo del coeficiente de correlación lineal

Dicho coeficiente se puede calcular como la raíz cuadrada del coeficiente de determinación con el mismo signo del coeficiente b. Se puede calcular directamente e independientemente del cálculo de la recta de regresión en el menú:

Estadísticos
Resúmenes
Matriz de correlaciones …

En este caso se nos muestra una ventana donde se deben elegir al menos dos variables numéricas. Para el conjunto de datos activo *Encuesta*, si ejecutamos



el resultado es

	altura	peso
altura	1.0000000	0.9676415
peso	0.9676415	1.0000000

de donde se deduce que el coeficiente de correlación lineal entre la altura y el peso es 0.96764145 que, salvo errores de redondeo, es la raíz cuadrada positiva del valor del coeficiente de determinación obtenido antes. Debemos notar que, obviamente, el coeficiente de correlación lineal de una variable consigo misma es 1.