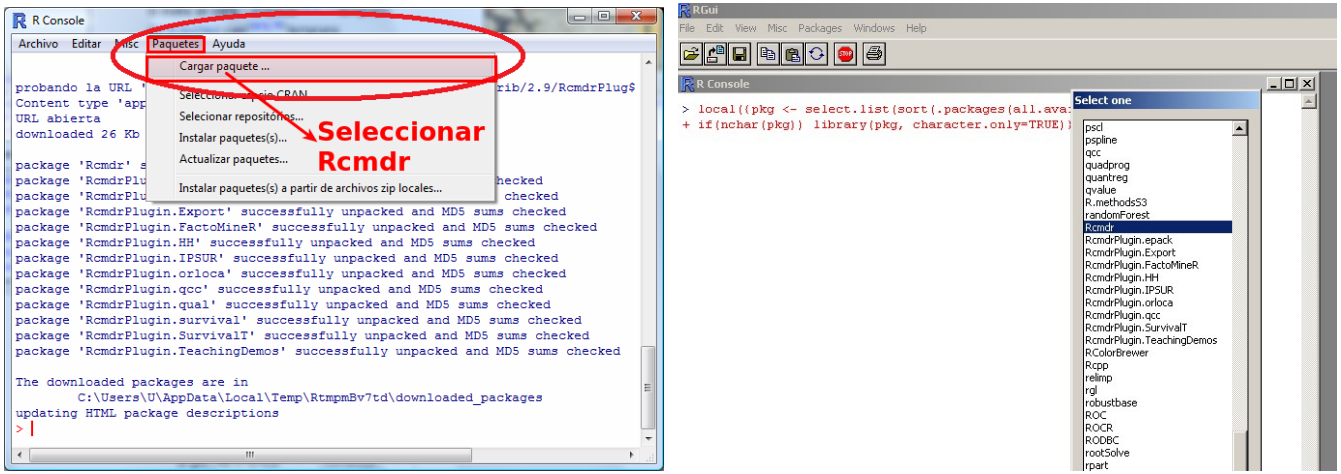
Práctica 4. Contrastes para dos muestras

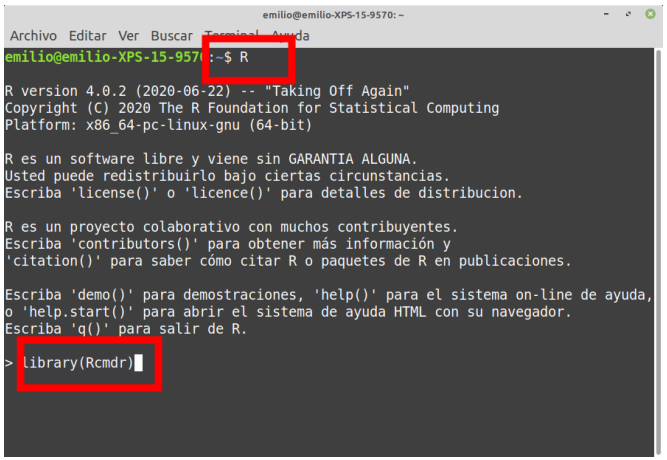
Departamento de Estadística
Universidad de Oviedo

Cargar el programa R y el paquete Rcommander

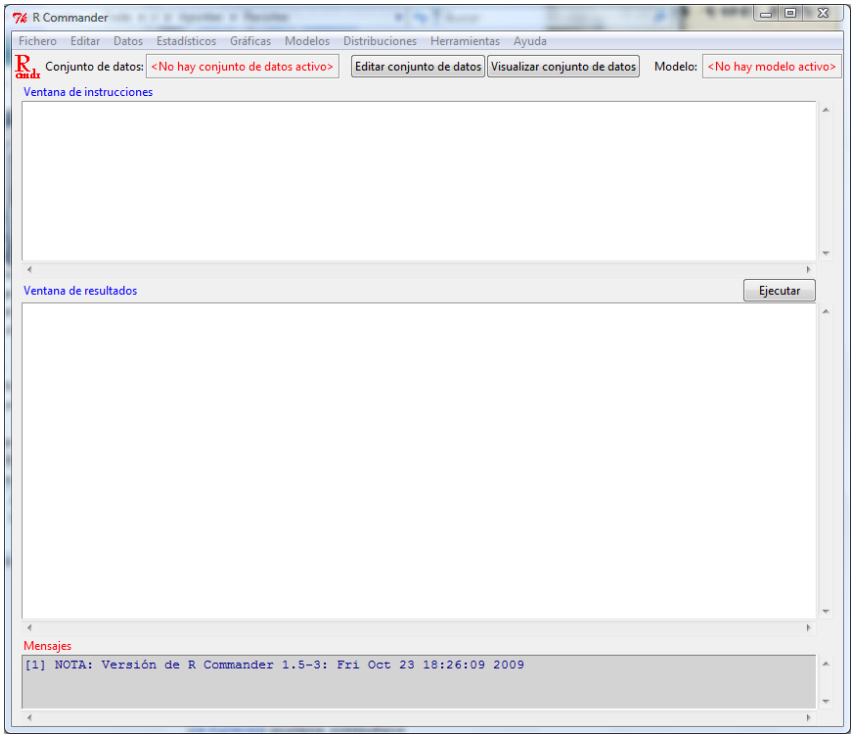
- Iniciamos el programa R.
- Cargamos el paquete RCommader. Dos opciones:
  1. Menú *Paquetes* → *Cargar paquete* → Seleccionamos **Rcmdr**.

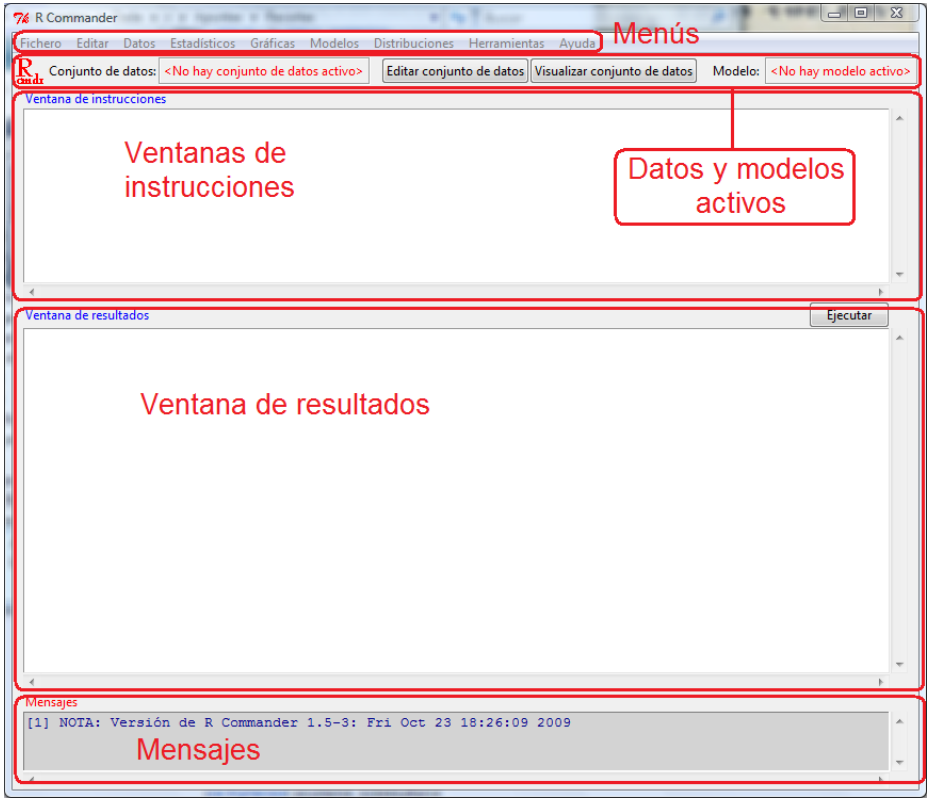


  2. Escribimos **library(Rcmdr)** en la consola y pulsamos retorno de carro.



Cargar Rcommander





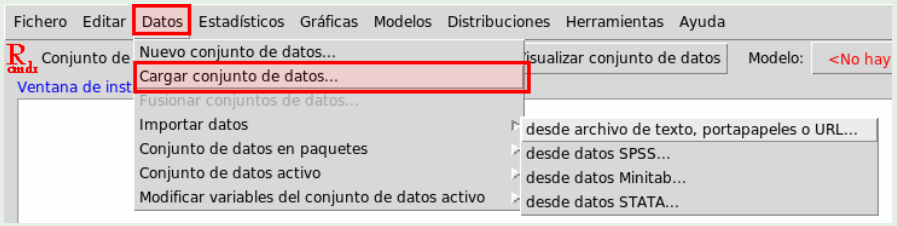
El fichero `acero.rda` se encuentra en el Campus Virtual. Hay que haberlo descargado previamente.

Cargar la base de datos `acero.rda`

Datos

➥ Cargar conjunto de datos

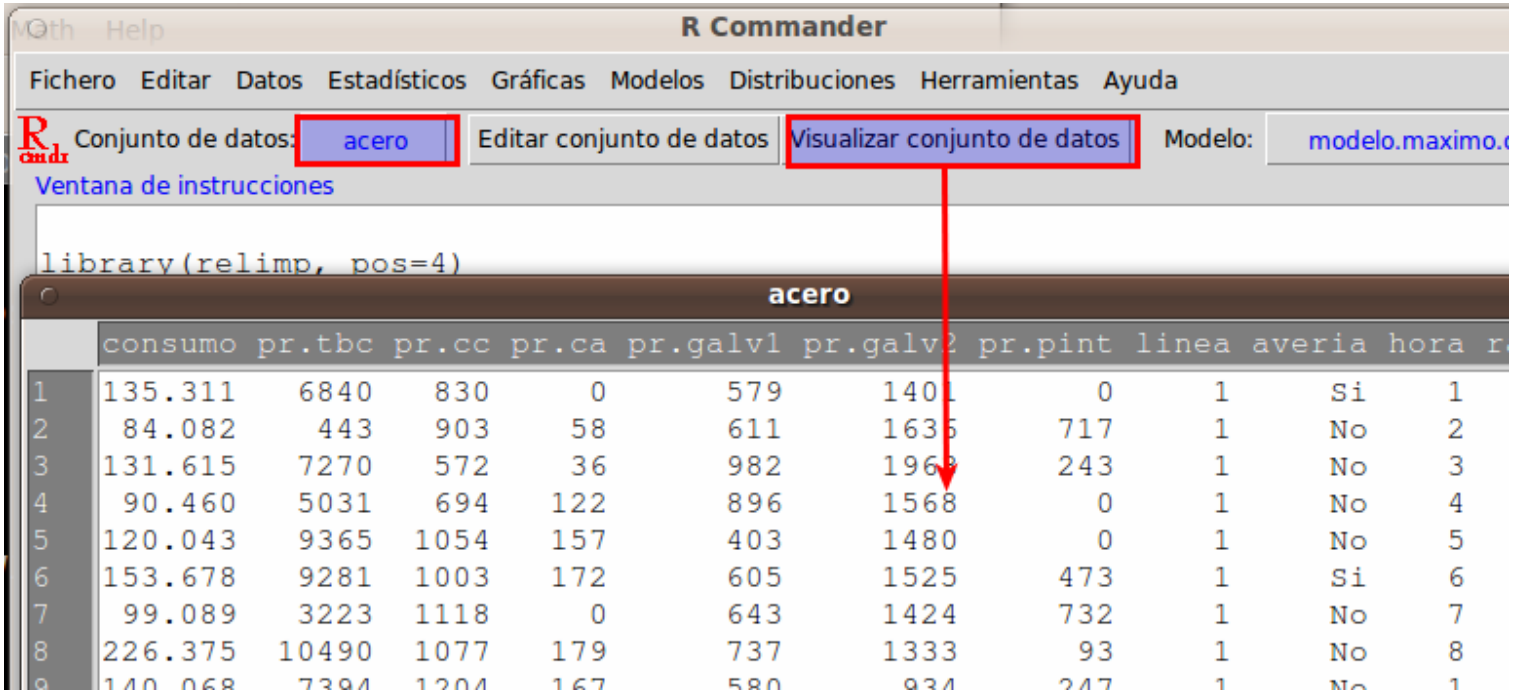
➥ Seleccionar **acero.rda**



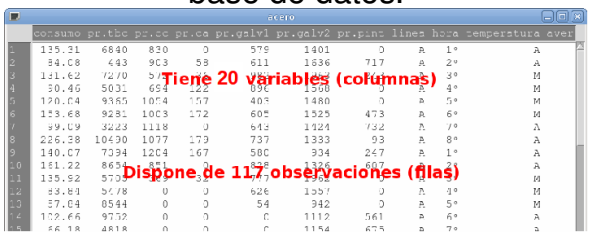

```
> load("/home/emilio/clases/acero.rda")
```

```
NOTA: El conjunto de datos acero tiene 117 filas y 20 columnas.
```

- Para visualizar la base de datos:



Aparece una ventana con los datos disponibles. Moviendo el cursor hacia la izquierda o hacia abajo podemos recorrer toda la base de datos.



Variables de la base de datos `acero`

1. `consumo` Consumo energético de la empresa (Megavatios/hora).
2. `pr.tbc` Producción del tren de bandas calientes (Toneladas de acero).
3. `pr.cc` Producción de colada continua (Toneladas de acero).
4. `pr.ca` Producción del convertidor de acero (Toneladas de acero).
5. `pr.galv1` Producción de galvanizado de tipo I (Tns. de acero).
6. `pr.galv2` Producción de galvanizado de tipo II (Tns. de acero).
7. `pr.pint` Producción de chapa pintada (Tns. de acero).
8. `linea` Línea de producción empleada (A o B).
9. `turno` Turno de mañana (M), tarde (T), noche (N).
10. `temperatura` Temperatura del sistema: Alta, Media y Baja.
11. `pres.aver` Presencia de averías: hubo Averías (A), no hubo averías (NoA).
12. `nun.aver` Número de averías detectadas.
13. `sistema` Activación de un sistema de detección de sobrecalientamiento: encendido (ON), apagado (OFF).
14. ...

Contraste de hipótesis

Toma de decisiones

H_0 (Hipótesis nula): Todo sigue igual
H_1 (Hipótesis alternativa): Se ha producido un cambio

Se rechaza, o no, la hipótesis nula: nunca se afirma que sea cierta.

- p-valor. Cuantifica el grado de compatibilidad de los datos con la H_0. Si es pequeño, se rechaza H_0.
- Intervalo de confianza. Cuantifica los límites de la variación. Permite estimar las consecuencias *prácticas* de esa variación.

Es mucho menos informativo decir que la proporción de niñas es significativamente distinta del 50%, que es lo que se deduce del p-valor, que decir que el intervalo de confianza de la proporción de niñas varía entre 48.45% y 48.77% al 95%.

Como ingenieros debemos plantear los distintos escenarios, las probabilidades asociadas y las consecuencias de los mismos.

Repaso

En la práctica anterior se hicieron tests para una muestra:

- Tests sobre la media poblacional:

¿El consumo medio de alcohol está entre 125 y 225 gramos semanales?

1. Aplicar el test de Shapiro-Wilk sobre normalidad de los datos.
2. ¿Siguen los datos una distribución normal?
   - Sí $\longrightarrow$ test t de Student para una muestra
   - No $\longrightarrow$ test de Wilcoxon para una muestra

- Tests sobre porcentajes:
  - Test de proporciones para una muestra.

Contraste de dos muestras

Contrastando dos grupos.

¿El consumo de alcohol es mayor entre hombres que entre mujeres?

- Tests sobre proporciones: Test de proporciones para dos muestras.
- Test sobre varianzas: test F (bajo condiciones de normalidad); Levene (no normalidad)
- Tests sobre promedios:
  1. Realice el test de Shapiro-Wilk sobre normalidad;
  2. ¿Existe normalidad?
     - Sí (test paramétrico); ¿Cómo son las muestras?
       Independientes $\longrightarrow$ Test t para muestras independientes
       Dependientes $\longrightarrow$ Test t para muestras pareadas
     - No (test no parámetrico); ¿Cómo son las muestras?
       Independientes $\longrightarrow$ Test de Wilcoxon para dos muestras
       Dependientes $\longrightarrow$ Test de Wilcoxon pareado

Contraste de proporciones

Las posibilidades para contrastar las proporciones de dos muestras son:

$H_0 : p_A = p_B$ $H_1 : p_A \neq p_B$	$H_0 : p_A \geq p_B$ $H_1 : p_A < p_B$	$H_0 : p_A \leq p_B$ $H_1 : p_A > p_B$
`two.sided`	`less`	`greater`

siendo p_A y p_B las proporciones en el grupo A y B, respectivamente.

Ejemplo: ¿El porcentaje de aprobados es el mismo para hombres y para mujeres?

$$\begin{array}{l} H_0 : p_{Hombres} = p_{Mujeres} \\ H_1 : p_{Hombres} \neq p_{Mujeres} \end{array}$$ o lo que es lo mismo $$\begin{array}{l} H_0 : p_{Hombres} - p_{Mujeres} = 0 \\ H_1 : p_{Hombres} - p_{Mujeres} \neq 0 \end{array}$$

Pista. Si la expresión *igual* aparece en la pregunta, entonces ponerla como H_0. En los demás casos, ponerla como H_1.

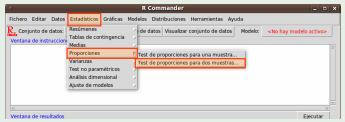
Regla de decisión (Interprete el intervalo de confianza)

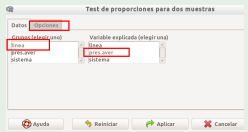
- Si p-valor es pequeño, rechazamos H_0. Por lo tanto, H_1 es cierta.
- Si p-valor es grande, no rechazamos H_0 (esto no implica necesariamente que la hipótesis nula sea cierta).

¿El porcentaje de horas en las que hubo averías es menor en la línea A que en la B?

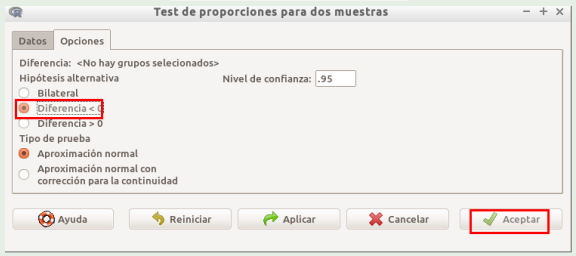
$$\boxed{\begin{array}{l} H_0 : p_A \geq p_B \text{ (prop. averías igual o mayor en la línea A)} \\ H_1 : p_A < p_B \end{array}} \iff \boxed{\begin{array}{l} H_0 : p_A - p_B \geq 0 \\ H_1 : p_A - p_B < 0 \end{array}}$$

- Estadísticos
  - ➥Proporciones
    - ➥Test de proporciones para dos muestras



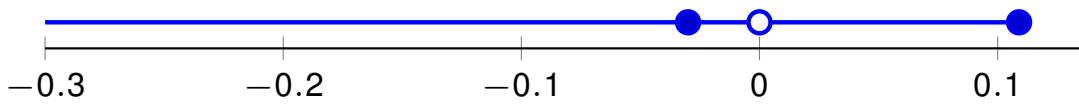


- Seleccionamos **linea** y **pres.aver**
- Opciones
  - ➥Hipótesis alternativa: **Diferencia** < 0 (debido a $p_A - p_B < 0$)
    - ➥OK




```
data:  .Table
X-squared = 0.0673, df = 1, p-value = 0.3976
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000  0.1095155
sample estimates:
   prop 1    prop 2
0.2295082 0.2500000
```

Intervalo de confianza de la diferencia



Como el p-valor (0.3976) es mayor que α no se rechaza la hipótesis nula; no hay evidencias de que vaya más averías en la línea A.

Contraste de varianzas

Test F para dos varianzas

Las hipótesis para este test son

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ (homocedasticidad)}$$
$$H_1 : \sigma_A^2 \neq \sigma_B^2 \text{ (heterocedasticidad)}$$

Regla de decisión (Interprete el intervalo de confianza.)

- Si p-valor es pequeño, rechazamos H_0. Por lo tanto, H_1 es cierta.
- Si p-valor es grande, no rechazamos H_0 (esto no implica necesariamente que la hipótesis nula sea cierta).

El consumo de alcohol en los hombres, ¿tiene la misma dispersión que el consumo de alcohol en las mujeres?

$$H_0 : \sigma^2_{Hombres} = \sigma^2_{Mujeres} \text{ (Varianzas iguales de consumo)}$$
$$H_1 : \sigma^2_{Hombres} \neq \sigma^2_{Mujeres} \text{ (Diferentes)}$$

$$\Longleftrightarrow \quad \boxed{\begin{array}{l} H_0 : \dfrac{\sigma^2_{Hombres}}{\sigma^2_{Mujeres}} = 1 \\ H_1 : \dfrac{\sigma^2_{Hombres}}{\sigma^2_{Mujeres}} \neq 1 \end{array}}$$

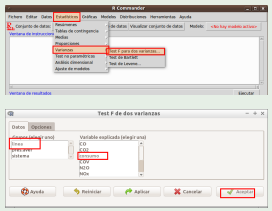
¿Son iguales las varianzas de los consumos de la línea A y B? (suponiendo normalidad)

$H_0 : \sigma_A^2 = \sigma_B^2$ (Varianzas iguales de consumo)
$H_1 : \sigma_A^2 \neq \sigma_B^2$ (Diferentes)

$$\iff \boxed{\begin{array}{l} H_0 : \dfrac{\sigma_A^2}{\sigma_B^2} = 1 \\ H_1 : \dfrac{\sigma_A^2}{\sigma_B^2} \neq 1 \end{array}}$$

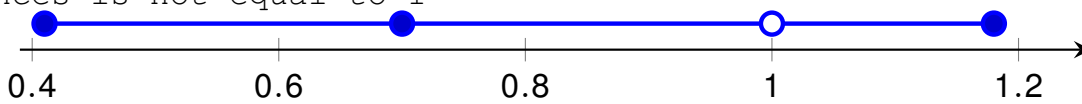
Estadísticos
➥Varianzas
  ➥Test F para dos varianzas

Seleccionar variables **linea** y **consumo**
➥OK



```
        F test to compare two variances

data:  consumo by linea
F = 0.7035, num df = 60, denom df = 55, p-value = 0.1834
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4158963 1.1828332
sample estimates:
ratio of variances
         0.7034893
```

Intervalo de confianza del cociente



Como el p-valor (0.1834) es mayor que α no se rechaza la hipótesis nula. Podemos suponer que no hay diferencias significativas entre las varianzas de las líneas, es decir, las dos muestras provienen de poblaciones que tienen la misma varianza.

Contraste de medias

Para contrastar dos medias, verificamos:

1. Si se puede admitir o no la normalidad.
2. La relación que hay entre los datos. Éstas pueden ser:
   - Independientes: Cuando deseamos comparar *dos grupos*. Ejemplo: ¿El consumo de alcohol es igual para los hombres y para las mujeres?

Sexo	...	Consumo	...
Hombre		10	
Hombre		12	
...		...	
Hombre		7	
Mujer		9	
Mujer		11	
...		...	
Mujer		8	

   - Pareadas: Cuando queremos comparar *dos variables*. En este caso, cada individuo tiene un valor asociado en cada una de las dos variables. Ejemplo: ¿El gasto en alcohol es igual que el gasto en telefonía?

Alcohol	Telefonía
14	10
12	12
...	...
5	7

Contraste de medias

El siguiente cuadro resume los diferentes contrastes de comparación de promedios que vamos a ver en las prácticas de laboratorio de esta asignatura:

	¿Distrib. Normal?	Independ.	Test - Contraste
Media de la diferencia	Sí	Sí	t para muestras independientes
Media de la diferencia	Sí	No	t para datos relacionados
Mediana de la diferencia	No	Sí	Wilcoxon para dos muestras
Mediana de la diferencia	No	No	Wilcoxon para muestras pareadas

Se compararán las medias o medianas de ambos grupos o variables.

$H_0 : \mu_A = \mu_B$ $H_1 : \mu_A \neq \mu_B$	$H_0 : \mu_A \geq \mu_B$ $H_1 : \mu_A < \mu_B$	$H_0 : \mu_A \leq \mu_B$ $H_1 : \mu_A > \mu_B$

Muestras independientes

- Si consideramos sólo a los hombres, ¿sigue la variable consumo de alcohol una distribución normal?
- Si consideramos sólo a las mujeres, ¿sigue la variable consumo de alcohol una distribución normal?

Sexo	...	Consumo	...
Hombre		10	
Hombre		12	
...		...	 ?
Hombre		7	
Mujer		9	
Mujer		11	
...		...	 ?
Mujer		8	

¿La variable *consumo* sigue una distribución normal en la línea A? ¿Y en la línea B?

Estadísticos
➥Resúmenes
  ➥Test de Normalidad

Seleccionamos: **consumo**
Pinchamos en el botón test por grupos:
➥Seleccionamos: **línea**
  ➥Aceptar
Aceptar

(para eliminar el test por grupo, pinchar en *restablecer*)

(Sigue $\longrightarrow$)

Test de Shapiro-Wilk:

H_0: los datos provienen de una distribución normal
H_1: los datos no provienen de una distribución normal

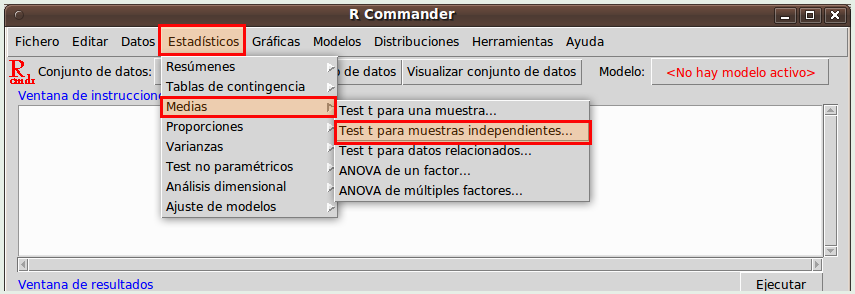
```
linea: A

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9708, p-value = 0.1534

------------------------------------------------------------------------------------------
linea: B

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9746, p-value = 0.2841
```

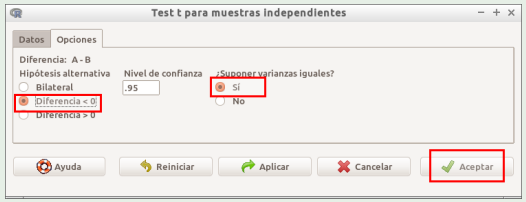
Para los datos de la línea A el p-valor es 0.1534 y para los datos de la línea B es 0.2841. En ambos casos suficientemente grande como para no rechazar la hipótesis nula (se puede admitir la normalidad de los datos).

¿El consumo medio es menor en la línea A que en la B?

$$\begin{array}{l} H_0 : \mu_A \geq \mu_B \\ H_1 : \mu_A < \mu_B \end{array} \iff \begin{array}{l} H_0 : \mu_A - \mu_B \geq 0 \\ H_1 : \mu_A - \mu_B < 0 \end{array}$$

Estadísticos
➥Medias
➥Test t para muestras independientes





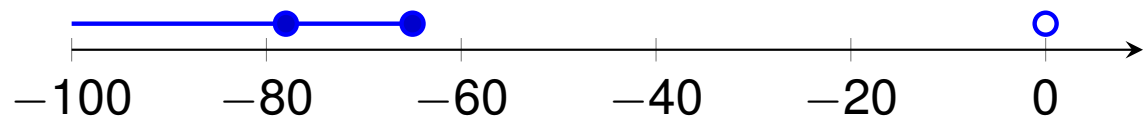
Seleccionar variables **linea** y **consumo**
➥Seleccionar: **Diferencia** $<$ **0**
➥**Varianzas iguales** (ver ejemplo previo)
➥OK

(Sigue ⟶)

$$\boxed{\begin{array}{l} H_0 : \mu_A \geq \mu_B \\ H_1 : \mu_A < \mu_B \end{array}} \iff \boxed{\begin{array}{l} H_0 : \mu_A - \mu_B \geq 0 \\ H_1 : \mu_A - \mu_B < 0 \end{array}}$$

```
        Two Sample t-test
data:  consumo by linea
t = -10.1697, df = 115, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -65.32647
sample estimates:
mean in group A mean in group B
       98.3182        176.3716
```

Intervalo de confianza



Como el p-valor es prácticamente cero, se rechaza la hipótesis nula. Por lo tanto, el consumo medio en la línea A es menor.

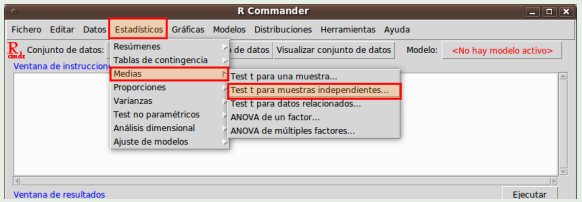
¿El consumo promedio es igual en ambas líneas?

$$\boxed{\begin{aligned} H_0 &: \mu_A = \mu_B \\ H_1 &: \mu_A \neq \mu_B \end{aligned}} \iff \boxed{\begin{aligned} H_0 &: \mu_A - \mu_B = 0 \\ H_1 &: \mu_A - \mu_B \neq 0 \end{aligned}}$$

Estadísticos

➥Medias

➥Test t para muestras independientes

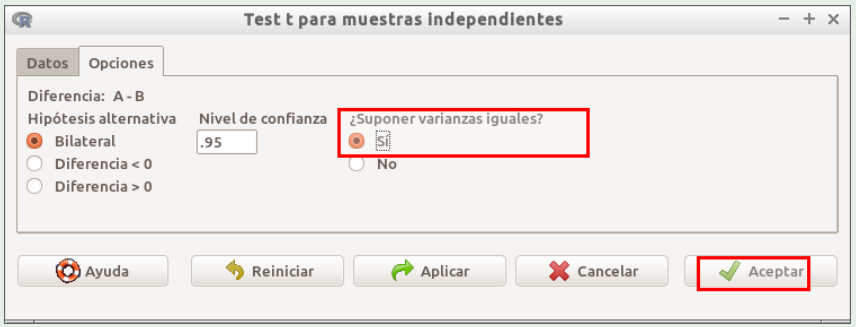


Seleccionar variables **linea** y **consumo**

➥Seleccionar: **Bilateral**

➥**Varianzas iguales**

➥OK



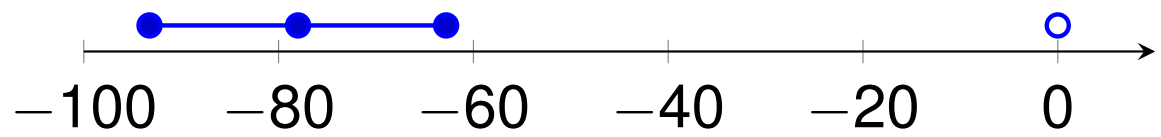
(Sigue ⟶)

¿El consumo promedio es igual en ambas líneas?

$$\boxed{\begin{array}{l} H_0 : \mu_A = \mu_B \\ H_1 : \mu_A \neq \mu_B \end{array}} \iff \boxed{\begin{array}{l} H_0 : \mu_A - \mu_B = 0 \\ H_1 : \mu_A - \mu_B \neq 0 \end{array}}$$

```
data:  consumo by linea
t = -10.1697, df = 115, p-value < 2.2e-16
alternative hypothesis: true difference in means is nc
95 percent confidence interval:
 -93.25631 -62.85051
sample estimates:
mean in group A mean in group B
        98.3182         176.3716
```

Intervalo de confianza



El consumo es distinto.

Muestras pareadas

Calcule la diferencia entre la variable V1 y V2. ¿Sigue esta diferencia una distribución normal?

V1	V2	V1-V2	
14	10	4	
12	12	0	
...		...	 ?
5	7	-2	

Calcule una nueva variable como la diferencia entre *pr.cc* y *pr.pint*. Llame a esta variable *diferencia*.

Datos

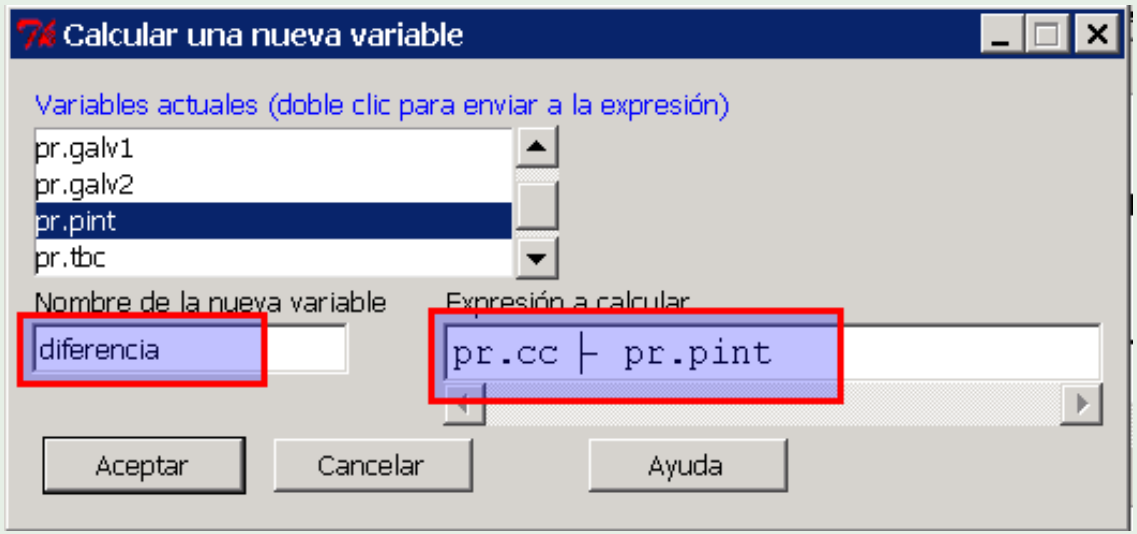
➥Modificar variables del conjunto de datos activo

➥Calcular una nueva variable

Nombre de la nueva variable: **diferencia**

Expresión a calcular: **pr.cc - pr.pint**

Aceptar

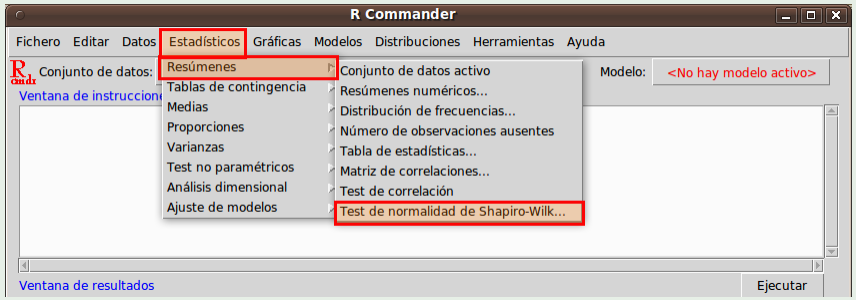



```
> acero$diferencia <- with(acero, pr.cc - pr.pint)
```

¿Sigue la variable *diferencia* una distribución normal?

Estadísticos
➥Resúmenes
➥Test de normalidad . . .

Seleccionar **diferencia**
➥OK




```
        Shapiro-Wilk normality test

data:  diferencia
W = 0.988, p-value = 0.3948
```

Como el p-valor es 0.39, no se rechaza la hipótesis de normalidad.

Datos pareados

Se ha preguntado a diversos jóvenes sobre su gasto mensual en telefonía y en consumo de alcohol:

	Gasto en (euros)		
Persona	Telefonía	Alcohol	diferencia
Juan	10	23	-13
María	7	12	-5
José	11	27	-16
...	...	...	

¿Existe relación entre lo que gasta Juan en Telefonía (10 euros) y en Alcohol (23 euros)?

- Sí, a Juan le corresponde el par de datos (10, 23).
- Son *datos pareados* (no son independientes).

¿El gasto medio es igual en telefonía y en alcohol?

$$H_0 : \mu_{Telefonia} = \mu_{Alcohol}$$
$$H_1 : \mu_{Telefonia} \neq \mu_{Alcohol}$$

Son datos pareados: hay que utilizar el test t *pareado* o el Wilcoxon *pareado*.

¿Son iguales las producciones medias de (*pr.cc*) y de (*pr.pint*)?

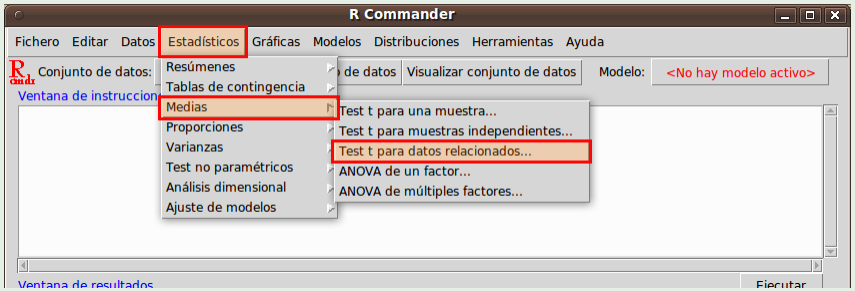
$$\begin{array}{l} H_0 : \mu_{pr.cc} = \mu_{pr.pint} \\ H_1 : \mu_{pr.cc} \neq \mu_{pr.pint} \end{array} \iff \begin{array}{l} H_0 : \mu_{pr.cc} - \mu_{pr.pint} = 0 \\ H_1 : \mu_{pr.cc} - \mu_{pr.pint} \neq 0 \end{array}$$

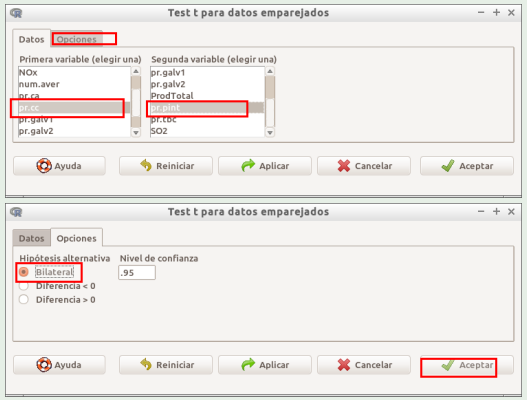
Como la variable *diferencia* (que es la diferencia entre ambas variables) sigue una distribución normal, utilizaremos el test t pareado.

Estadísticos
➥Medias
➥Test t para datos relacionados

Primera variable
➥Seleccionar **pr.cc**
➥Segunda variable
➥Seleccionar **pr.pint**
➥Hipótesis alternativa
➥**Bilateral**

Aceptar



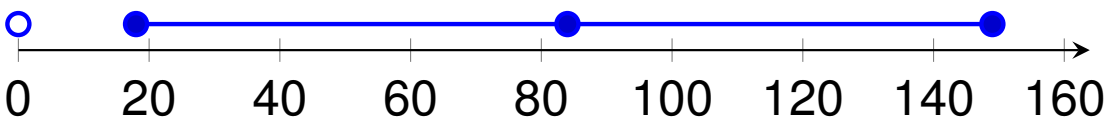


(Sigue ⟶)

$$\boxed{\begin{array}{l} H_0 : \mu_{pr.cc} = \mu_{pr.pint} \\ H_1 : \mu_{pr.cc} \neq \mu_{pr.pint} \end{array}} \iff \boxed{\begin{array}{l} H_0 : \mu_{pr.cc} - \mu_{pr.pint} = 0 \\ H_1 : \mu_{pr.cc} - \mu_{pr.pint} \neq 0 \end{array}}$$

```
        Paired t-test

data:  pr.cc and pr.pint
t = 2.5405, df = 116, p-value = 0.01239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  18.56348 149.91515
sample estimates:
mean of the differences
                84.23932
```

Intervalo de confianza de la diferencia



El p-valor es 0.012; por lo que las medias de `pr.cc` y `pr.pint` son significativamente distintas.

Muestras independientes

- Si consideramos sólo a los hombres, ¿sigue la variable consumo de alcohol una distribución normal?
- Si consideramos sólo a las mujeres, ¿sigue la variable consumo de alcohol una distribución normal?

Sexo	...	Consumo	...
Hombre		10	
Hombre		12	
...		...	 ?
Hombre		7	
Mujer		9	
Mujer		11	
...		...	 ?
Mujer		8	

¿Sigue la variable *pr.galv2* una distribución normal en cada línea?

Estadísticos
➥Resúmenes
➥Test de Normalidad

Seleccionamos: **pr.galv2**
Pinchamos en el botón test por grupos:
➥Seleccionamos: **línea**
➥Aceptar
Aceptar

(para eliminar el test por grupo, pinchar en *restablecer*)

(Sigue $\longrightarrow$)

```
linea: A

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8955, p-value = 7.985e-05

------------------------------------------------------------
linea: B

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9111, p-value = 0.0005496
```

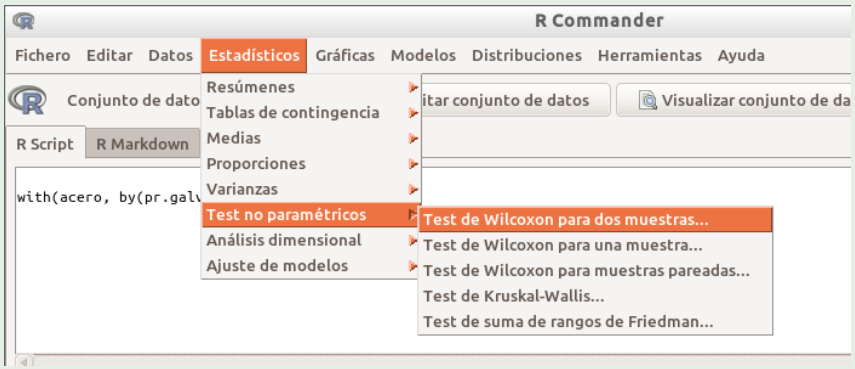
No hay condiciones de normalidad.

¿Son iguales las producciones medias de (*pr.galv2*) según la línea)?

$$H_0 : Mediana_A = Mediana_B$$
$$H_1 : Mediana_A \neq Mediana_B$$

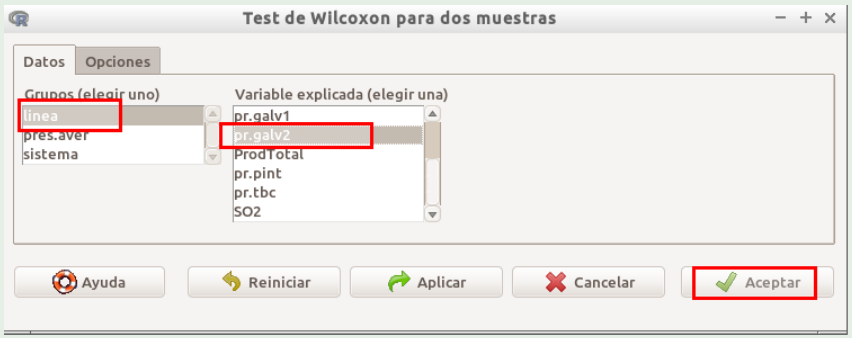
Como la variable *pr.galv2* no sigue una distribución normal, utilizaremos el test de Wilcoxon.

Estadísticos
➥Test no paramétricos
➥Test de Wilcoxon para dos muestras



Primera variable
➥Seleccionar **linea**
➥Segunda variable
➥Seleccionar **pr.galv2**
➥Hipótesis alternativa
➥**Bilateral**

Aceptar



(Sigue ⟶)

```
Wilcoxon rank sum test with continuity correction

data:  pr.galv2 by linea
W = 1431, p-value = 0.1314
alternative hypothesis: true location shift is not equal to 0
```

No se rechaza la hipótesis de igualdad de producción en ambas líneas.

Muestras pareadas

Calcule la diferencia entre la variable V1 y V2. ¿Sigue esta diferencia una distribución normal?

V1	V2	V1-V2	
14	10	4	
12	12	0	
...		...	 ?
5	7	-2	

Calcule una nueva variable como la diferencia entre *pr.galv1* y *pr.galv2*. Llame a esta variable *dif*.

Datos

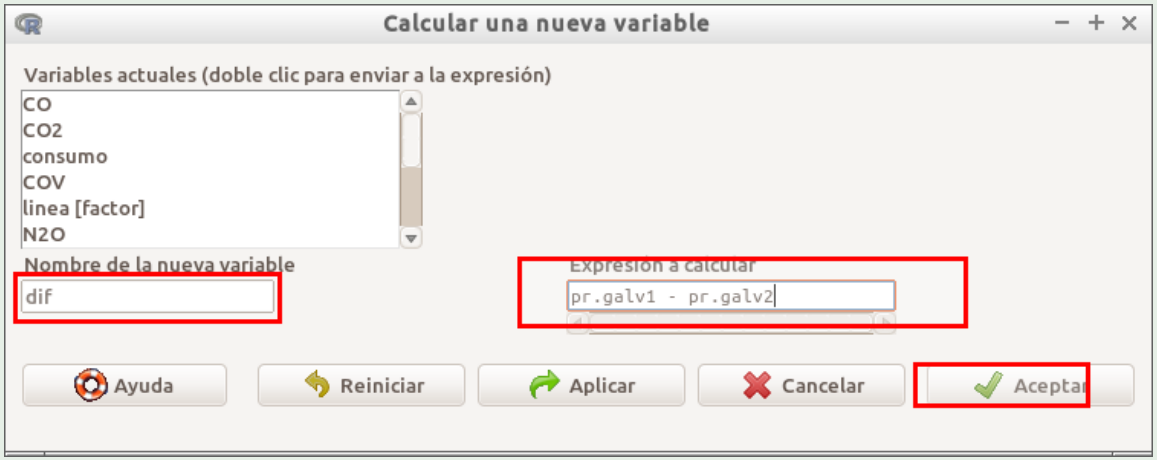
➥Modificar variables del conjunto de datos activo

➥Calcular una nueva variable

Nombre de la nueva variable: **dif**

Expresión a calcular: **pr.galv1 - pr.galv2**

Aceptar

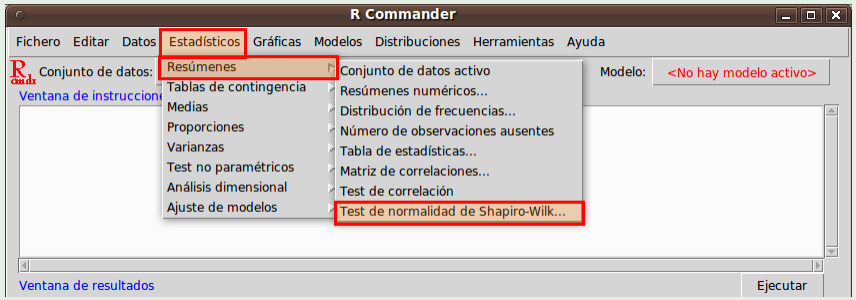



```
> acero$dif <- with(acero, pr.galv1 - pr.galv2)
```

¿Sigue la variable *dif* una distribución normal?

Estadísticos
➥Resúmenes
➥➥Test de normalidad...

Seleccionar **dif**
➥OK




```
        Shapiro-Wilk normality test

data:  dif
W = 0.9671, p-value = 0.005665
```

Como el p-valor es pequeño, se rechaza la hipótesis de normalidad.

Compárese, en promedio, la producción *pr.galv1* y *pr.galv2*

$$H_0 : Mediana_{pr.galv1-pr.galv2} = 0$$
$$H_1 : Mediana_{pr.galv1-pr.galv2} \neq 0$$

Como la variable *dif* (que es la diferencia entre ambas variables) no sigue una distribución normal, utilizaremos el test de Wilcoxon pareado.

Estadísticos
➥Test no paramétricos
➥Test de Wilcoxon para muestras pareadas

Primera variable
➥Seleccionar **pr.galv1**
➥Segunda variable
➥Seleccionar **pr.galv2**
➥Hipótesis alternativa
➥**Bilateral**
Aceptar

(Sigue ⟶)

```
        Wilcoxon signed rank test with continuity correction

data:  pr.galv1 and pr.galv2
V = 249, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Se rechaza la hipótesis de igualdad de producción en ambos tipos de galvanizado.