Práctica 5

Contrastes de hipótesis con una y dos muestras

Contenido

La Inferencia estadística busca extraer conclusiones generales sobre una característica de una población a partir del estudio de una muestra de esa población, a diferencia de la Estadística descriptiva, que solo busca describir y resumir el comportamiento de los individuos de la muestra.

Los objetivos más habituales de la Inferencia estadística son:

- **Estimar** el valor de una característica (**parámetro**) desconocido de la variable que caracteriza una población.
- Realizar **contrastes** (o **tests**) de **hipótesis**.

En este bloque se plantearán diferentes problemas relacionados con el contraste de hipótesis acerca de los parámetros o las distribuciones que caracterizan a las variables aleatorias. En primer lugar, se ilustrará el procedimiento general que permite resolver contrastes de hipótesis estadísticas. A continuación, se examinarán los contrastes de hipótesis más habituales con una y dos muestras.

1. Ideas principales sobre contrastes de hipótesis

En cualquier contraste de hipótesis hay que fijar lo que se conoce como hipótesis nula y su alternativa.

H_0:	Hipótesis nula
H_1:	Hipótesis alternativa

En todo el proceso estadístico se supone que la hipótesis nula es cierta y sólo la rechazamos cuando existan evidencias estadísticas para ello, es decir cuando lo que observamos en la muestra tenga una probabilidad muy pequeña de ocurrir de ser cierta dicha hipótesis nula.

Se denomina contraste de hipótesis al procedimiento estadístico que sirve para decidir, con cierto nivel de error, si se rechaza o no la hipótesis nula planteada a partir de los resultados obtenidos en una muestra extraída de la población (valor del estadístico de la muestra).

Para medir la discrepancia entre lo que hemos observado y lo que esperaríamos observar si la hipótesis nula fuese cierta, se utiliza un estadístico muestral. Un **estadístico** es una medida numérica, calculada a partir de la muestra, cuya distribución de probabilidad es conocida si la hipótesis nula es cierta.

En la práctica, la decisión que se adopta de rechazar o no la hipótesis nula se toma a partir del p-valor o significación del estadístico El **p-valor** es la probabilidad de obtener un valor del estadístico al menos tan alejado de la hipótesis nula como el que se ha observado, si H_0 es cierta. Es una probabilidad, por lo que toma valores entre 0 y 1.

Al plantear un contraste y tomar una decisión (rechazar o no rechazar H_0) podemos equivocarnos de dos formas:

- Error de tipo I: Rechazar H_0 cuando es cierta. Se denota por α y se denomina **nivel de significación** del contraste.
- Error de tipo II: No rechazar H_0 cuando es falsa (H_1 es cierta).

Lo ideal sería que los dos tipos de error fuesen lo más pequeños posible. Esto es imposible, ya que cuando el error de tipo I disminuye, el de tipo II aumenta y viceversa.

Por ello, en cualquier contraste se fija un nivel de significación α (se suele fijar $\alpha = 0.05$) y se toma una decisión siguiendo el siguiente criterio:

- Si el p-valor es muy pequeño, p-valor< α, se rechaza H_0, es decir, que al nivel de significación del 100 α% podemos afirmar H_1. Se dice que el resultado es (estadísticamente) significativo.
- Si el p-valor no es pequeño, p-valor $\geq \alpha$, no se rechaza H_0, es decir, que al nivel de significación del 100 α% no podemos afirmar H_1. Se dice que el resultado no es (estadísticamente) significativo.

Procedimiento general. Se han desarrollado muchos tipos de contrastes y en los siguientes apartados, veremos algunos de los más utilizados (implementados en R Commander y en la mayor parte de software estadístico).

Una ventaja de los contrastes de hipótesis es que todos funcionan igual y si hemos entendido bien la filosofía subyacente podremos aplicar correctamente cualquier contraste, siempre que tengamos claros los siguientes puntos:

- Saber cuál es la hipótesis nula (H_0) y la hipótesis alternativa (H_1) del contraste.
- Saber qué condiciones son necesarias para poder aplicarlo (por ejemplo que la muestra sea normal, o de tamaño grande, etc.).
- Tener un software que realice los cálculos y nos proporcione el p-valor.
- Aplicar la regla de decisión basada en el p-valor que conocemos, al nivel de significación α deseado.

2. Contrastes para la media poblacional con una muestra

Para resolver contrastes de hipótesis sobre una media poblacional, previamente hemos de tener en cuenta si la muestra procede de una población normal o no, para poder utilizar el test t para una muestra o no.

Nota: Si no se puede suponer que la muestra es normal, se podría utilizar el test no paramétrico de Wilcoxon para una muestra que realiza contrastes de hipótesis acerca de la mediana de la población.

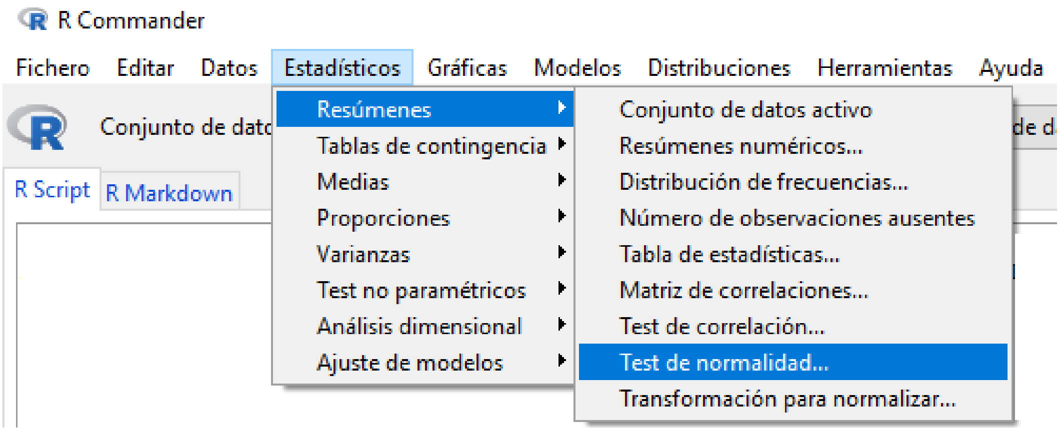
2.1. Test de normalidad

Por la especial importancia que tiene la distribución normal, existen varios contrastes específicos para estudiar la bondad de ajuste a esta distribución.

Se tiene una m.a.s. (X_1, ... , X_n) de una v.a. continua X. Se plantea el contraste de normalidad con las hipótesis

H_0:	X sigue distribución normal
H_1:	X NO sigue distribución normal

R Commander da acceso a los tests de normalidad mediante el menú **Estadísticos / Resúmenes / Test de normalidad ....**

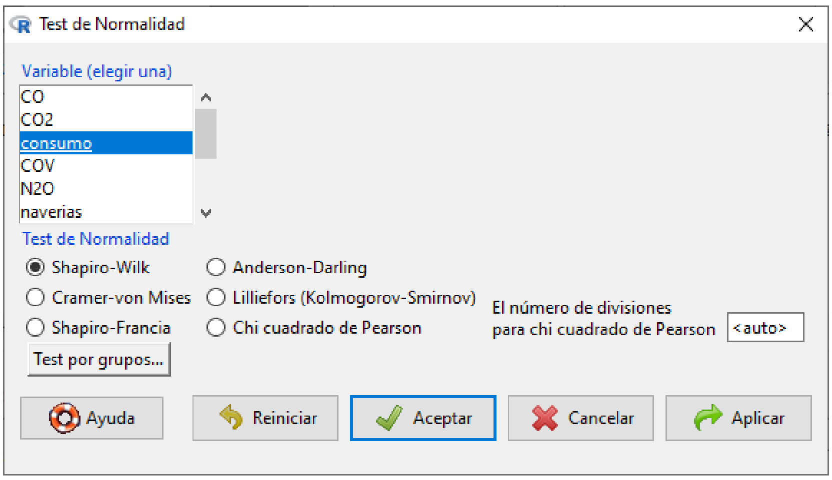


Por defecto, lleva a cabo el test de normalidad de Shapiro-Wilk (`shapiro.test`), pero también están disponibles los tests de Anderson-Darling (`ad.test`), Cramer-von Mises (`cvm.test`), Lilliefors (Kolmogorov-Smirnov) (`lillie.test`), Shapiro-Francia (`sf.test`) y el test Chi-cuadrado de Pearson (`chisq.test`).

Para cada test nos indica el valor del estadístico del contraste y el p-valor (**p-value**).

Ejemplo 1. Con los datos del fichero **acero2.rda**, estudia la normalidad de la variable *consumo*.

Solución: ***Estadísticos → Resúmenes → Test de normalidad ...***



Se pueden observar los resultados al solicitar en los tests de normalidad: el test de Shapiro-Wilk o el test de Lilliefors (Kolmogorov-Smirnov).

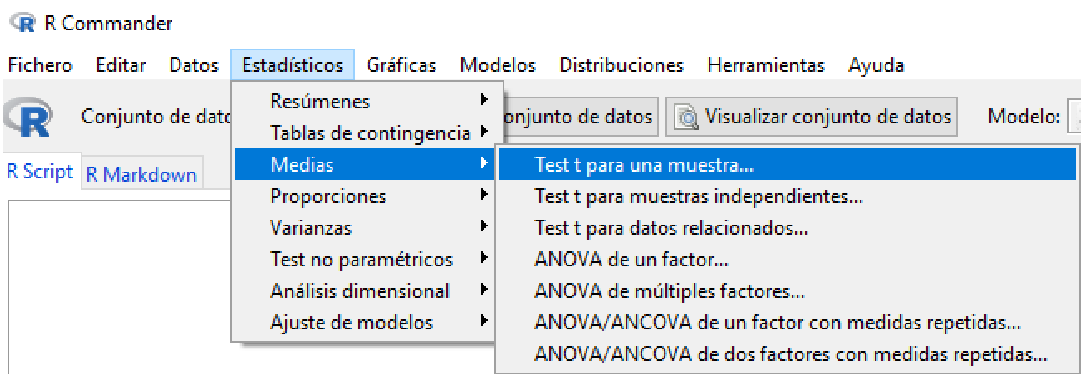
```
> normalityTest(~consumo, test="shapiro.test", data=acero2)

 Shapiro-Wilk normality test

data: consumo
W = 0.99252, p-value = 0.7829
```

```
> normalityTest(~consumo, test="lillie.test", data=acero2)

 Lilliefors (Kolmogorov-Smirnov) normality test

data: consumo
D = 0.041614, p-value = 0.8883
```

En el ejemplo se obtiene con el estadístico de Shapiro-Wilk un p-valor de 0.7829, mientras que con Lilliefors (Kolmogorov-Smirnov) el p-valor es 0.8883, por lo que con los niveles de significación habituales (α= 0.01, 0.05, 0.1) con ambos procedimientos no se rechaza la hipótesis nula de normalidad ya que el p-valor es mayor que el nivel de significación. Por lo tanto podemos suponer que la variable consumo sigue una distribución normal.

2.2. Test t para una muestra

Para llevar a cabo contrastes de hipótesis e intervalos de confianza acerca de la media de una población normal con varianza desconocida, se tiene la opción **Test t para una muestra ...** del menú **Estadísticos / Medias**.



Se considera una m.a.s. (X_1, ... , X_n) de una v.a. continua $X \equiv \text{norm}(\mu,\sigma)$. Dependiendo de la hipótesis alternativa que se plantee, se hablará de contrastes unilaterales o bilaterales:

C. bilateral	C. unilateral dcha.	C. unilateral izda.
H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$ `two.sided`	H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$ `greater`	H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$ `less`

El estadístico que se utiliza para resolver estos contrastes viene dado por

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \equiv t_{n-1}$$

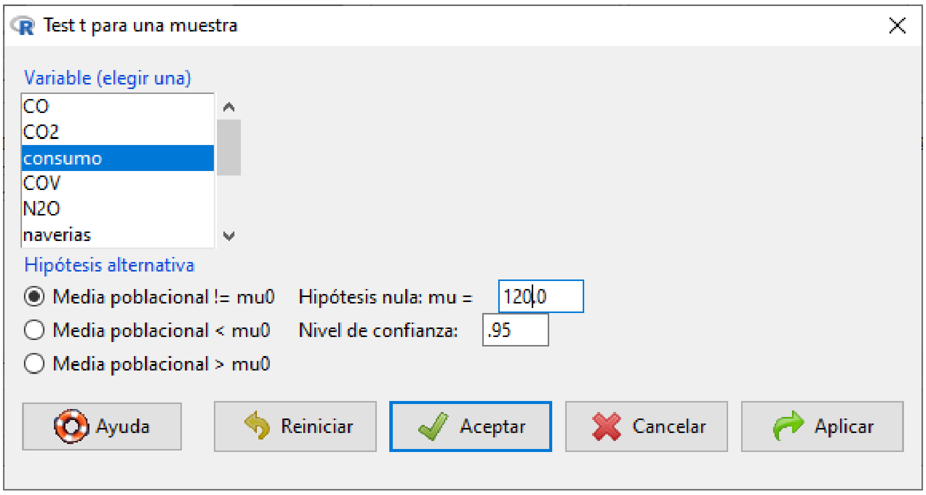
donde $\bar{X}$ es la media muestral y S es la cuasidesviación típica. Este estadístico sigue una distribución *t* de Student con n-1 grados de libertad.

Ejemplo 2. Con los datos del fichero **acero2.rda**, ¿es el consumo medio igual a 120 kilovatios/hora?

Solución: Ya que se puede admitir que el consumo sigue una distribución normal (Ejemplo 1), podemos suponer que *X* ≡ norm(μ,σ) con μ y σ desconocidos, por lo que el contraste adecuado en este caso es el **Test t para una muestra ...** Las hipótesis a contrastar son:

H_0:	μ = 120 (el consumo medio es de 120)
H_1:	μ ≠ 120 (el consumo medio no es de 120)

Estadísticos → Medias → Test t para una muestra



La salida del R Commander proporciona los elementos para resolver el contraste de hipótesis (el valor del estadístico, los grados de libertad de la distribución muestral del estadístico y el p-valor), la estimación puntual de la media poblacional (media muestral, $\bar{X}$) y un intervalo de confianza al nivel de confianza fijado en el cuadro de diálogo.

```
> with(acero2, (t.test(consumo, alternative='two.sided', mu=120.0, conf.level=.95)))

One Sample t-test

data: consumo
t = 3.8136, df = 116, p-value = 0.000221
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
129.3516    149.5614
sample estimates:
mean of x
139.4565
```

Hipótesis alternativa

Estimación puntual de la media poblacional

Si se toma α=0,05 entonces como el p-valor (bilateral) es 0.000221 < 0.05, se rechaza H_0, es decir, hay evidencias suficientes para afirmar que el consumo medio es diferente de 120 kilovatios/hora.

Ejemplo 3. Con los datos del fichero acero2.rda, ¿El consumo medio es menor de 140 kilovatios/hora?

Solución: En este caso, como vuelve a tratarse de la variable consumo y ya sabemos que se comporta de manera normal, se utiliza el Test t para la media. Ahora las hipótesis son:

H_0:	$\mu \geq 140$ (el consumo medio no es menor de 140)
H_1:	$\mu < 140$ (el consumo medio es menor de 140)

Estadísticos → Medias → Test t para una muestra

Selecciona la variable *consumo*, escribe 140 en la hipótesis nula, marca *Media poblacional* < mu0 y acepta.

La salida es:

```
One Sample t-test

data: consumo
t = -0.1065, df = 116, p-value = 0.4577
alternative hypothesis: true mean is less than 140
95 percent confidence interval:
-Inf   147.9159
sample estimates:
mean of x
139.4565
```

Como el p-valor (0.4577) supera a los niveles de significación α habituales (α = 0.01, 0.05, 0.1), no se rechaza H_0. No hay evidencia suficiente en la muestra para afirmar que el consumo medio es inferior a 140.

3. Contrastes para la proporción poblacional con una muestra

A menudo, interesa saber la proporción de individuos de una población que cumplen cierta condición o propiedad, *p*. Así, *p* representa la probabilidad de que un individuo de dicha población cumpla la condición. Supongamos que *p* es desconocido y se dispone de una m.a.s. de tamaño *n* para hacer inferencias acerca de *p*.

La situación anterior se puede modelar mediante una v.a. *Y* definida como

$$Y = \begin{cases} 1 & \text{si el individuo cumple la condición} \\ 0 & \text{si el individuo NO cumple la condición} \end{cases}$$

con lo que se dispone de una m.a.s. $(Y_1, \ldots, Y_n)$ de una v.a. $Y \equiv \text{binom}(1,p)$.

Para resolver contrastes de hipótesis acerca de la proporción poblacional, *p*, de la forma

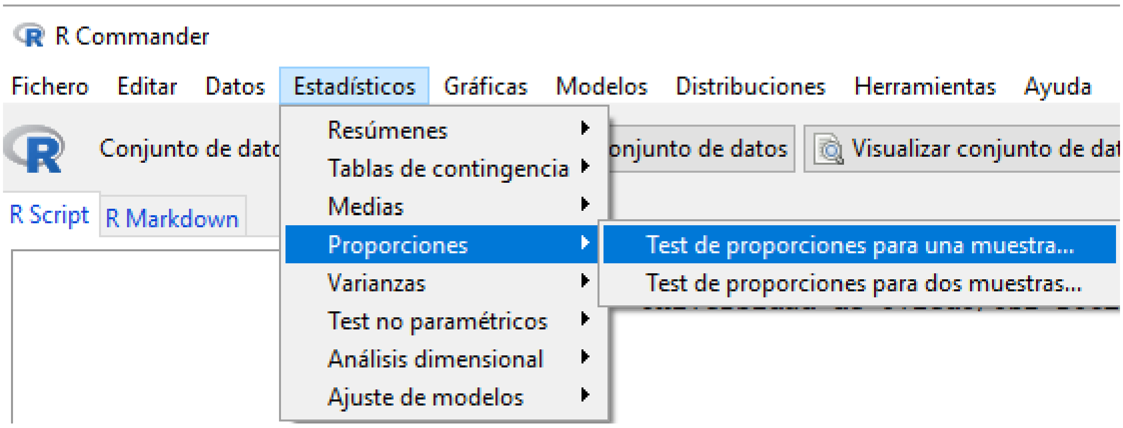
C. bilateral	C. unilateral dcha.	C. unilateral izda.
H_0: $p = p_0$ H_1: $p \neq p_0$ `two.sided`	H_0: $p \leq p_0$ H_1: $p > p_0$ `greater`	H_0: $p \geq p_0$ H_1: $p < p_0$ `less`

el estadístico habitual del contraste de una proporción es de la forma,

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \xrightarrow[n \to \infty]{} Z \equiv \mathrm{norm}(0{,}1)$$

donde $\hat{p}$ denota a la proporción muestral de individuos que cumplen la condición y que coincide con la media muestral de la variable *Y* por lo que este estadístico tiene distribución aproximadamente normal si el tamaño de muestra es suficientemente grande ($n \geq 30$), aplicando el Teorema Central del Límite.

R Commander dispone de la opción **Test de proporciones para una muestra** del menú **Estadísticos / Proporciones**.



Hay que tener en cuenta que

- R Commander permite aplicar este contraste a las variables dicotómicas (variables de tipo factor con exactamente dos niveles) del conjunto de datos.
- R Commander considera, por defecto, que la proporción *p* se refiere a la primera categoría por orden alfabético, salvo que las categorías de la variable hayan sido previamente ordenadas por el investigador. Esto se puede hacer con la opción del menú **Datos / Modificar variables del conjunto de datos activo / Reordenar niveles del factor...**

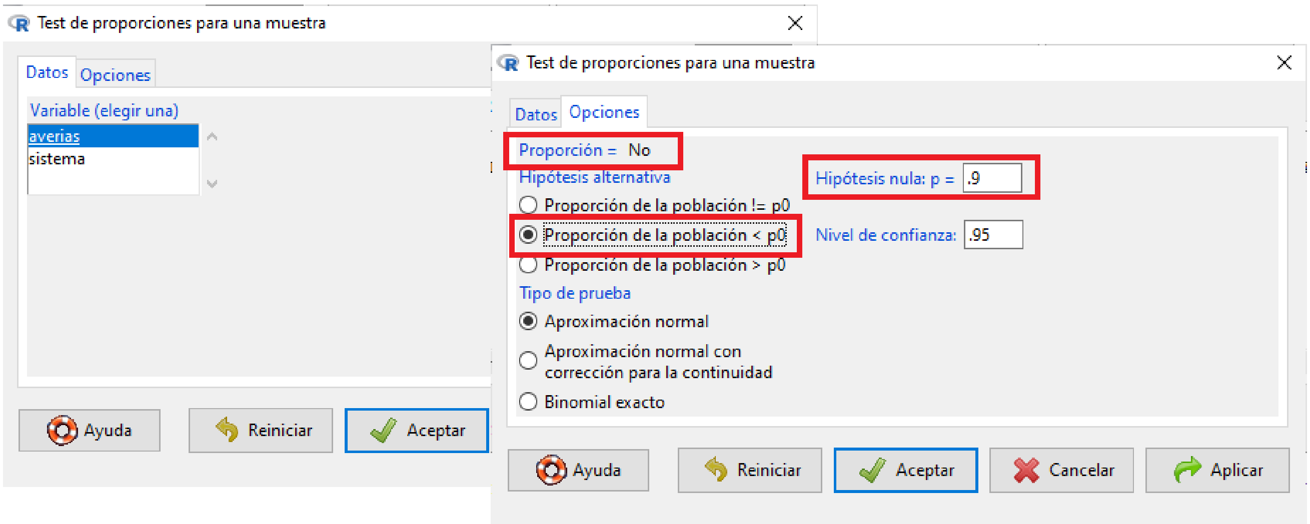
Ejemplo 4. Considera el fichero **acero2.rda**. Se quiere saber si el porcentaje de horas con avería es excesivo, considerándose excesivo si el porcentaje es mayor del 10%.

Solución: Sea p_{Si} = ``proporción poblacional de horas con avería''. En primer lugar debemos fijarnos en que el primer nivel de la variable tipo factor *averias* es el valor No. Por ello, resolveremos el contraste de hipótesis acerca de si la proporción de horas sin avería, $p_{No} = 1 - p_{Si}$, es menor de 0.9. Con lo que el contraste de hipótesis a plantear en el ejemplo será de la forma:

H_0:	$p_{No} \geq 0.9$ (proporción razonable de averías)
H_1:	$p_{No} < 0.9$ (proporción excesiva de averías)

Ahora solo tienes que hacer: ***Estadísticos → Proporciones → Test de proporciones para una muestra***

En este caso, después de seleccionar la variable *averia*, en la pestaña **Opciones** se debe especificar el valor 0.9 como **Hipótesis nula** y marcar la opción *Proporción de la población < p0 en* **Hipótesis alternativa**. El contraste se realiza, por defecto, con la *Aproximación normal* en **Tipo de prueba**.



El resultado es:

Frequency counts (test is for first level):
averias
No Si
89 28

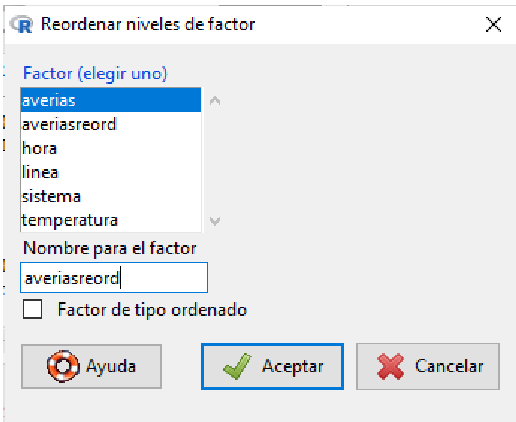
1-sample proportions test without continuity correction

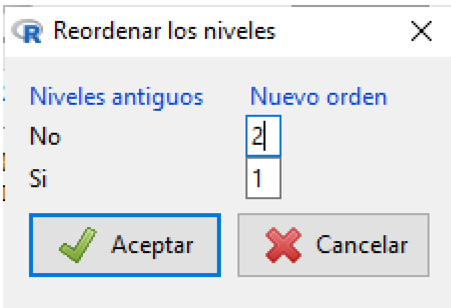
data: rbind(.Table), null probability 0.9
X-squared = 25.232, df = 1, **p-value = 0.0000002542**
alternative hypothesis: true p is less than 0.9
95 percent confidence interval:
 0.0000000 0.8192062
sample estimates:
    p
0.7606838

Estimación puntual de la proporción poblacional de horas sin averias

Como el p-valor (0.0000002542) es muy pequeño se rechaza la hipótesis nula, por lo que con un nivel de significación del 5% se puede afirmar que el porcentaje de horas con averías es excesiva.

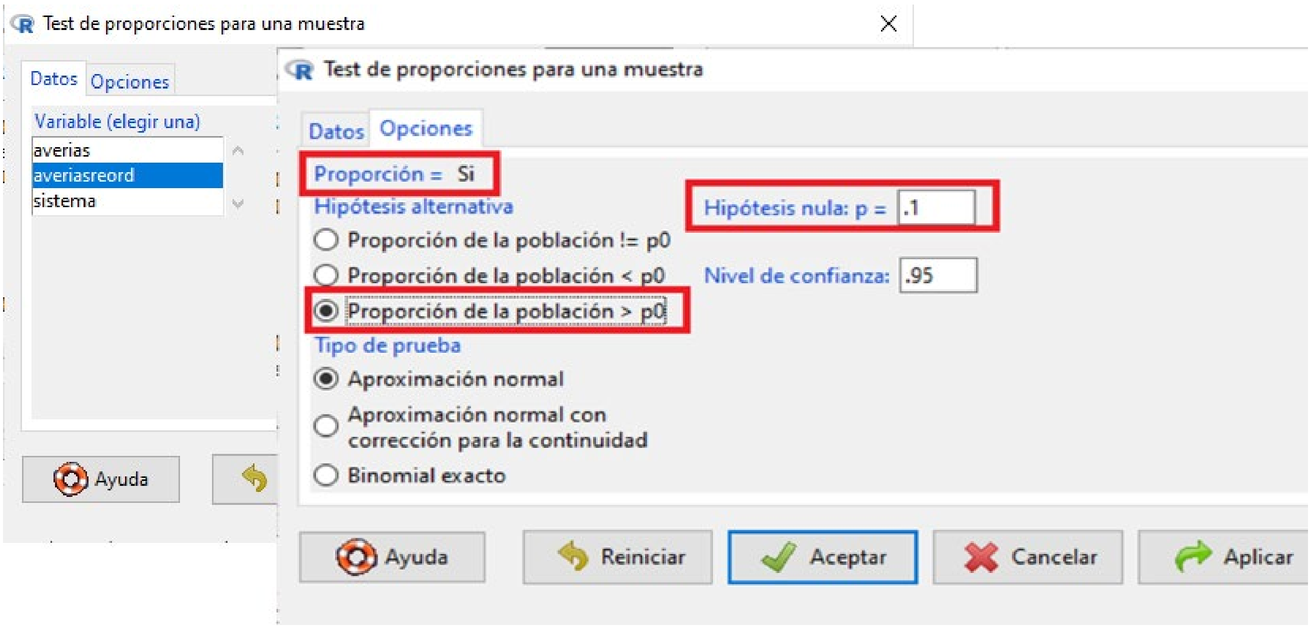
A la misma conclusión se llega si se reordenan los niveles de la variable *averias* (*averiasreord*) para que Si sea el primer nivel de esta nueva variable (Datos / Modificar variables del conjunto de datos activo / Reordenar niveles de factor).





Trabajando así, el contraste a plantear sería

H_0:	$p_{Si} \leq 0.1$	(proporción razonable de averías)
H_1:	$p_{Si} > 0.1$	(proporción excesiva de averías)



El resultado sería entonces

Frequency counts (test is for first level):
averiasreord
Si No
28 89

1-sample proportions test without continuity correction

data: rbind(.Table), null probability 0.1
X-squared = 25.232, df = 1, p-value = 0.0000002542
alternative hypothesis: true p is greater 9im 0.1
95 percent confidence interval:
0.1807938 1.0000000
9sample estimates:
p
0.2393162

Estimación puntual de la proporción poblacional de horas con averías

Con lo que la conclusión es la misma: con un nivel de significación del 5% se puede afirmar que el porcentaje de horas con averías es excesivo.

Nota 1: El estadístico proporcionado en la salida del R Commander, **X-squared**, coincide con el cuadrado del estadístico habitual del contraste de una proporción,

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

y es un caso particular (cuando *i*=2) del estadístico Chi-cuadrado que se utiliza en los contrastes de bondad de ajuste a una distribución discreta

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i} \xrightarrow[n\to\infty]{} {\chi^2}_{i-1}$$

donde se comparan son las frecuencias observadas en la muestra de los datos, $O_i=n_i$, con las frecuencias esperadas en una muestra con el mismo número de datos si la hipótesis nula fuera cierta, $E_i=n\, p_i^0$.

Nota 2: Bajo el supuesto de que ninguna frecuencia esperada E_i es inferior a 5, el estadístico X-squared se distribuye según una distribución χ^2. En el caso de que alguna frecuencia esperada sea inferior a 5 se suele utilizar la **corrección para la continuidad** de Yates, en la que el estadístico es

$$T = \sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i} \xrightarrow[n\to\infty]{} {\chi^2}_{i-1}$$

Nota 3: Si los tamaños muestrales son muy pequeños, conviene hacer el contraste con la opción **Binomial exacto**. La prueba binomial exacta parte del hecho de que, si la hipótesis nula fuera cierta, la distribución del número de casos favorables en la muestra sería binomial, y se obtiene así el p-valor correspondiente.

```
Frequency counts (test is for first level):
averiasreord
Si No
28 89

 Exact binomial test

data: rbind(.Table)
number of successes = 28, number of trials = 117, p-value = 0.00001002
alternative hypothesis: true probability of success is greater than 0.1
95 percent confidence interval:
 0.1757437 1.0000000
sample estimates:
probability of success
      0.2393162
```

4. Comparación de medias poblacionales con dos muestras independientes

Para trabajar con dos muestras independientes, R Commander necesita una variable cuantitativa en el archivo de datos que contenga los valores de la variable cuya media se desea comparar, y una variable cualitativa dicotómica (con solo dos niveles) que indique con sus valores la muestra o grupo a la que pertenece cada individuo.

Para elegir el procedimiento estadístico adecuado, en primer lugar hay que comprobar si la variable cuantitativa se puede considerar que tiene distribución normal en ambos grupos o muestras. En caso afirmativo, podemos utilizar el **Test t para muestras independientes** que requiere normalidad para comparar las medias de *X* e *Y*.

Nota: En el caso de que en alguno (o ambos) grupos no se pueda suponer normalidad para la variable numérica, se podría recurrir al contraste no paramétrico de Wilcoxon para dos muestras para comparar las medianas poblacionales de *X* e *Y*.

4.1. Test de normalidad por grupos

Ejemplo 5. Con los datos del fichero acero2.rda, ¿se puede suponer que el consumo según haya o no avería sigue distribución normal?

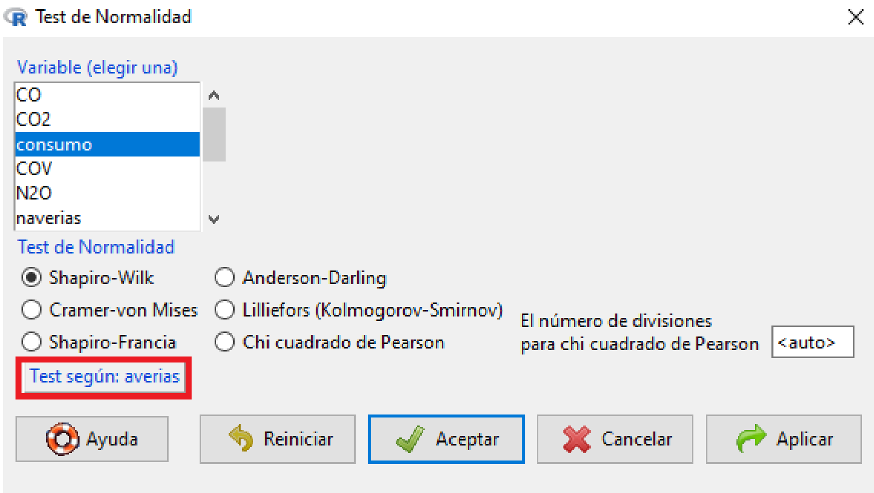
Solución: Sea *X* = ``consumo''. Consideramos la v.a. *X* en cada uno de los grupos que define la variable *averias*, denotando por $X_{No}$ = ``consumo cuando no hay avería'' y por X_{Si} = ``consumo cuando si hay avería''.

En primer lugar vamos a estudiar la normalidad de ambas variables. Para ello, podemos utilizar alguno de los test de normalidad disponibles en R Commander, pero en este caso el interés NO es que la variable consumo (*X*) tenga distribución normal, sino que X_{No} y X_{Si} tengan distribución normal:

H_0: X_{No} sigue distribución normal
H_1: X_{No} NO sigue distribución normal

H_0: X_{Si} sigue distribución normal
H_1: X_{Si} NO sigue distribución normal

En el menú **Test de normalidad** se dispone del botón **Test por grupos** para seleccionar averias como variable para formar los grupos: No y Si.



El resultado sería:

```
> normalityTest(consumo ~ averias, test="shapiro.test", data=acero2)

 --------
 averias = No

 Shapiro-Wilk normality test

data: consumo
W = 0.98686, p-value = 0.5137

 --------
 averias = Si
```

```
Shapiro-Wilk normality test

data: consumo
W = 0.9644, p-value = 0.4408
```

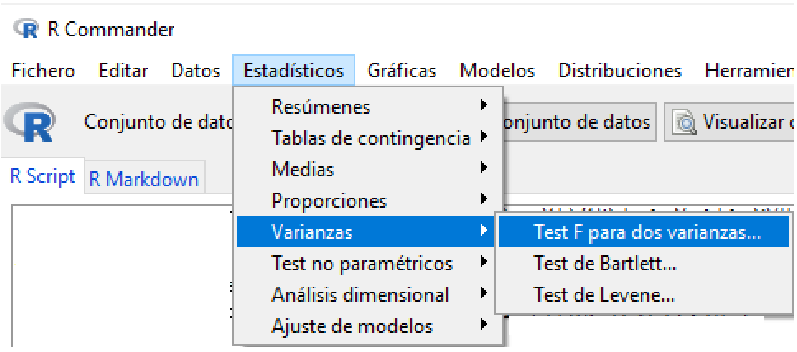
En ambos grupos se obtienen p-valores altos para la variable consumo (p-valor = 0.5137 cuando no hay avería y p-valor = 0.4408 cuando hay avería), por lo que podemos suponer que el consumo sigue distribución normal tanto cuando no hay avería como cuando sí la hay. Por lo tanto, podemos suponer que $X_{No} \equiv$ norm(μ_{No},σ_{No}) y $X_{Si} \equiv$ norm(μ_{Si},σ_{Si}).

4.2. Comparación de varianzas con dos muestras normales independientes

Un paso previo a un contraste de comparación de medias en muestras independientes de 2 poblaciones normales, *X* e *Y*, es determinar si las varianzas en ambas poblaciones pueden suponerse iguales o por el contrario podemos afirmar que son distintas.

C. bilateral
H_0: $\sigma_X^2 = \sigma_Y^2$ H_1: $\sigma_X^2 \neq \sigma_Y^2$ `two.sided`

R Commander proporciona entre las opciones del menú **Estadísticos / Varianzas** tres contrastes diferentes para comparar las varianzas. En este caso se utilizará el **test F para dos varianzas**, puesto que se están considerando dos poblaciones con distribución normal.



El estadístico utilizado es:

$$T = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \equiv F_{n_X - 1, n_Y - 1}$$

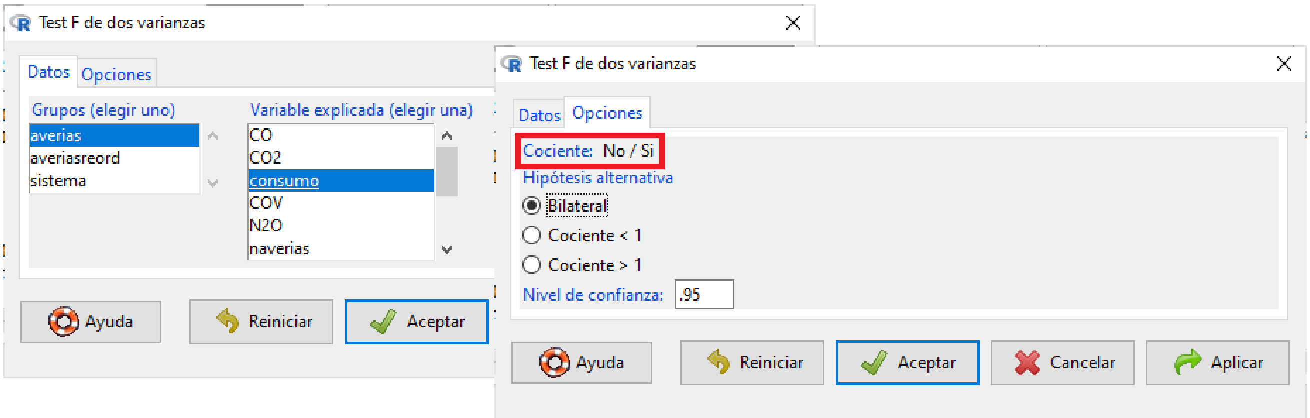
Según la conclusión que se obtenga de la comparación de las varianzas, se utilizarán estadísticos distintos en el contraste de comparación de las medias.

Nota: Este test se utiliza habitualmente como un contraste auxiliar en uno de los contrastes de igualdad de medias (el contraste t para dos muestras independientes). No obstante, también tiene importancia en sí mismo, puesto que una estrategia básica para la mejora de la calidad pasa por la identificación de las causas que producen la variabilidad, para intentar reducirla. Por esta razón en muchas ocasiones se realizan test de comparación de varianzas (normalmente con hipótesis alternativas unilaterales) para analizar si las medidas consideradas para reducir la variabilidad han surtido efecto.

Ejemplo 6. Con los datos del fichero **acero2.rda**, ¿son iguales las varianzas de los consumos con y sin averías?

Solución: Para resolver el contraste de igualdad de varianzas

H_0:	$\sigma_{No}^2 = \sigma_{Si}^2$	(varianzas iguales)
H_1:	$\sigma_{No}^2 \neq \sigma_{Si}^2$	(varianzas distintas)



El resultado es:

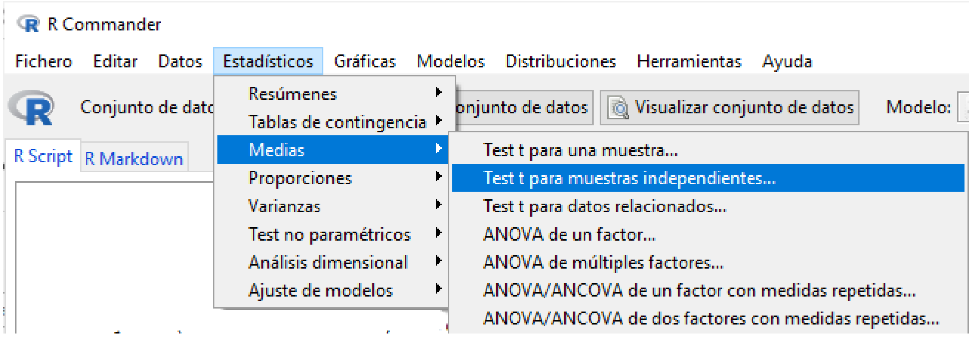
```
> Tapply(consumo ~ averias, var, na.action=na.omit, data=acero2) # variances by group
  No    Si
3123.748 2802.630

> var.test(consumo ~ averias, alternative='two.sided', conf.level=.95, data=acero2)

 F test to compare two variances

data: consumo by averias
F = 1.1146, num df = 88, denom df = 27, p-value = 0.7731
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5696427 1.9686748
sample estimates:
ratio of variances
    1.114577
```

Como el p-valor (0.7731) es alto (es mayor que los niveles de significación habituales) no se rechaza la hipótesis nula. Podemos suponer que no hay diferencias significativas entre las varianzas del consumo con o sin avería, es decir, las dos muestras provienen de poblaciones que tienen la misma varianza.

Por lo tanto, podemos suponer que $X_{No} \equiv norm(\mu_{No},\sigma)$ y $X_{Si} \equiv norm(\mu_{Si},\sigma)$. Vamos a comparar ahora el consumo medio cuando hay averías y cuando no las hay.

4.3. Test t para muestras independientes

Con el menú **Estadísticos / Medias / Test t para muestras independientes** se resuelven contrastes de hipótesis que comparan las medias de dos poblaciones normales independientes.



Los contrastes de hipótesis a contrastar serán de la forma:

C. bilateral	C. unilateral dcha.	C. unilateral izda.
H_0: $\sigma_X^2 = \sigma_Y^2$ H_1: $\sigma_X^2 \neq \sigma_Y^2$ `two.sided`	H_0: $\sigma_X^2 \leq \sigma_Y^2$ H_1: $\sigma_X^2 > \sigma_Y^2$ `greater`	H_0: $\sigma_X^2 \geq \sigma_Y^2$ H_1: $\sigma_X^2 < \sigma_Y^2$ `less`

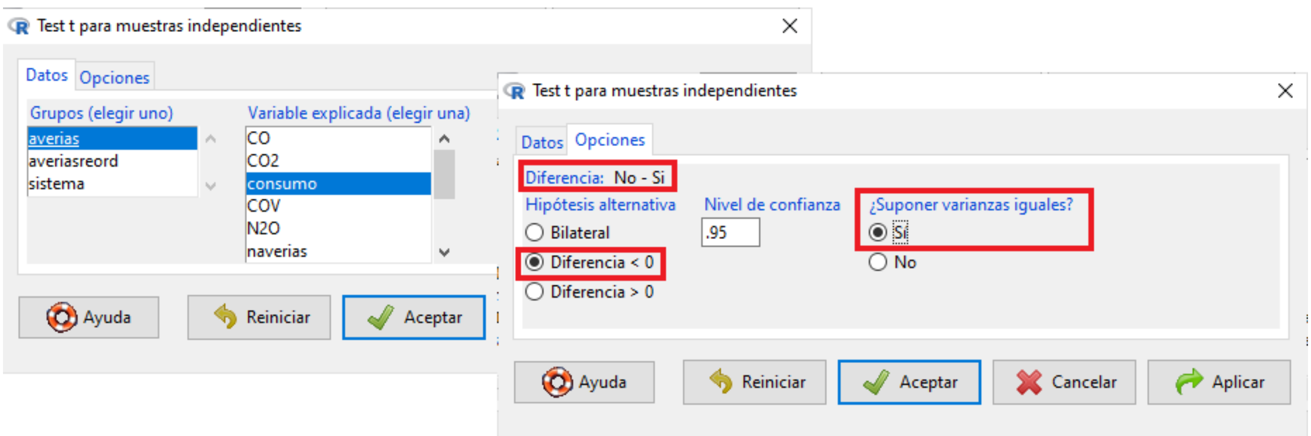
Ejemplo 7. Con los datos del fichero acero2.rda, ¿se puede afirmar que el consumo medio es mayor cuando hay avería que cuando no?

Solución: Se plantea el contraste unilateral

$H_0 : \mu_{NO} \geq \mu_{SI}$ (consumo menor o igual con avería)

$H_1 : \mu_{NO} < \mu_{SI}$ (consumo mayor con avería)

Siguiendo la secuencia: ***Estadísticos → Medias → Test t para muestras independientes***, se elige como Variable explicada: **consumo** y en Grupos se elige la variable *averias*, especificando en **Opciones** que la hipótesis alternativa es Diferencia < 0 (contraste unilateral a la izquierda) y en la pregunta **¿Suponer varianzas iguales?** se debe seleccionar Si.



Cuando se puede suponer que las varianzas son iguales, el estadístico que compara las medias de dos variables con distribución normal es de la forma

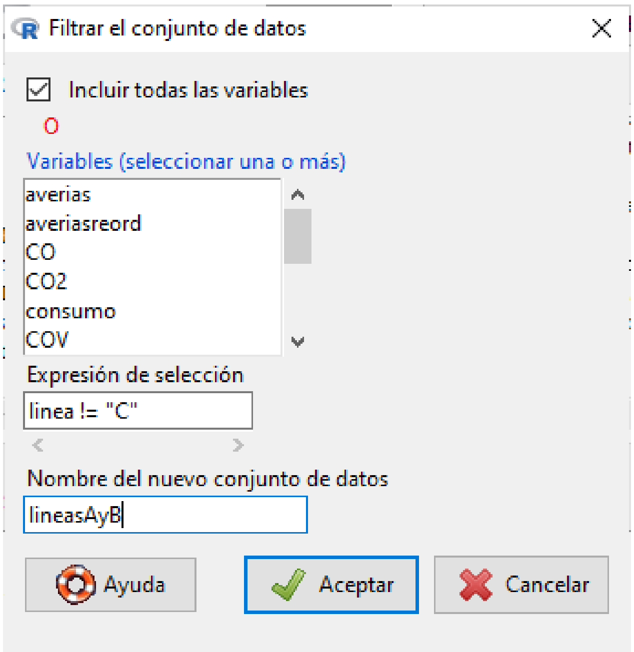
$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{(n_X - 1) + (n_Y - 1)}}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \equiv t_{n_X + n_Y - 2}$$

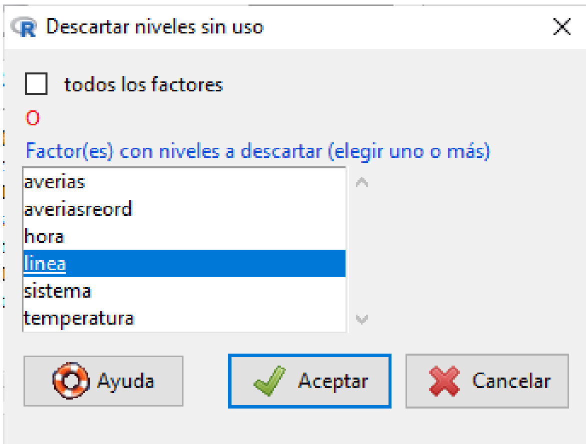
```
> t.test(consumo~averias, alternative='less', conf.level=.95, var.equal=TRUE, data=acero2)

 Two Sample t-test

data: consumo by averias
t = -0.94235, df = 115, p-value = 0.174
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 8.564113
sample estimates:
mean in group No mean in group Si
   136.7585     148.0321
```

En el ejemplo, como el p-valor es 0.174, mayor que los niveles de significación habituales no se rechaza la hipótesis nula, es decir, no hay evidencias para afirmar que el consumo medio es mayor cuando hay averías que cuando no las hay.

Ejemplo 8. Con los datos del fichero **acero2.rda**, ¿se puede afirmar, con un nivel de significación α = 0.05, que el consumo promedio es diferente en las líneas A y B?

Solución: En primer lugar debemos filtrar el fichero *acero2.rda* para seleccionar sólo las líneas A y B, creando un nuevo fichero que denominaremos *lineasAyB.RData*. Con la opción **Descartar niveles sin uso** del menú **Modificar variables del conjunto de datos activo** conseguimos que la variable linea tenga sólo los niveles A y B.





Consideramos la v.a. X = "consumo" en los grupos que define la variable "línea", denotando por X_A = "consumo en la línea A" y X_B = "consumo en la línea B".

Observando los resultados del **test de Shapiro-Wilk** para el *consumo* en ambos grupos de la variable *linea*, se obtienen en ambos casos p-valores por encima de 0.05 (0.1737 y 0.07302, respectivamente), por lo que podemos concluir al 5% de significación que el consumo sigue distribución normal tanto en la línea A como en la línea B, es decir, $X_A \equiv \text{norm}(\mu_A,\sigma_A)$ y $X_B \equiv \text{norm}(\mu_B,\sigma_B)$.

Si contrastamos la igualdad de varianzas entre ambas variables normales, H_0: $\sigma_A^2 = \sigma_B^2$, el estadístico del contraste de igualdad de varianzas (**test F para dos varianzas**) tiene un p-valor igual a 0.01369 < 0.05. Por lo tanto, SÍ se rechaza H_0 y para un nivel de significación α=0.05 concluimos que ambas varianzas no se pueden asumir iguales.

```
> var.test(consumo ~ linea, alternative='two.sided', conf.level=.95, data=lineasAyB)

 F test to compare two variances

data: consumo by linea
F = 0.44221, num df = 38, denom df = 38, p-value = 0.01369
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2318854 0.8432890
sample estimates:
ratio of variances
    0.4422063
```

El estadístico que compara las medias de dos variables X e Y independientes con distribución normal cuando las varianzas son distintas tiene distribución t de Student pero los grados de libertad se obtienen con la aproximación de Welch, así:

$$T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}} \equiv t_f \text{ con } f = \frac{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X-1}+\frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y-1}}$$

Con el menú **Estadísticos / Medias / Test t para muestras independientes** y la opción (por defecto) de varianzas distintas se obtiene un p-valor bilateral = 0.01851, con lo que se deduce que las medias son distintas y por ello planteamos un contraste unilateral para decidir si una de las medias es significativamente mayor o menor que la otra.

Con el test para Hipótesis alternativa: Diferencia < 0 al resolver el contraste

H_0: $\mu_A \geq \mu_B$ (consumo menor o igual en línea B)

H_1: $\mu_A < \mu_B$ (consumo mayor en línea B)

obtenemos la siguiente salida del R Commander:

```
> t.test(consumo~linea, alternative='less', ..., var.equal=FALSE, data=lineasAyB)

Welch Two Sample t-test

data: consumo by linea
t = -2.415, df = 66.111, p-value = 0.009255
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -8.567596
sample estimates:
mean in group A mean in group B
109.8410    137.5485
```

El p-valor del contraste de hipótesis unilateral es 0.009255 y por lo tanto al nivel de significación α= 0.05 se rechaza H_0, con lo que podemos afirmar que el consumo medio en la línea B es mayor que el consumo medio en la línea A.

5. Comparación de proporciones con dos muestras independientes

El contraste para la comparación de proporciones pretende decidir si hay diferencias o no para la proporción de individuos que cumplen cierta condición entre dos grupos diferentes de individuos.

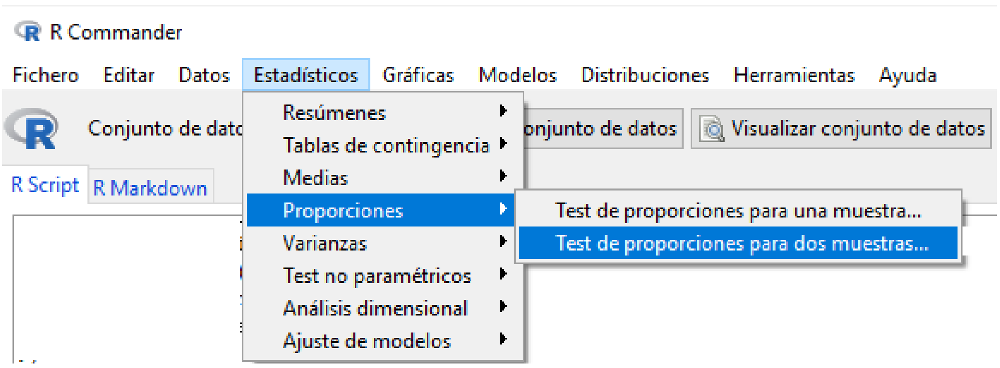
Si denotamos por p_X a la proporción de individuos de la población X que cumplen cierta condición y por p_Y a la proporción de individuos de la población Y que cumplen la misma condición, los contrastes a formular pueden ser de las siguientes formas:

C. bilateral	C. unilateral dcha.	C. unilateral izda.
H_0: $p_X^2 = p_Y^2$ H_1: $p_X^2 \neq p_Y^2$ `two.sided`	H_0: $p_X^2 \leq p_Y^2$ H_1: $p_X^2 > p_Y^2$ `greater`	H_0: $p_X^2 \geq p_Y^2$ H_1: $p_X^2 < p_Y^2$ `less`

Se seleccionan m.a.s. de las poblaciones X e Y de tamaños n_X y n_Y, respectivamente. Sean $\hat{p}_X$ y $\hat{p}_Y$, las correspondientes proporciones muestrales de individuos que cumplen cierta condición en X e Y. Si n_X y n_Y son suficientemente grandes (n_X, $n_Y \geq 30$), el estadístico del contraste viene dado por

$$T = \frac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \xrightarrow[n\to\infty]{} Z \equiv \text{norm}(0,1) \;\; \text{con}\; \hat{p} = \frac{n_X\hat{p}_X + n_Y\hat{p}_Y}{n_X + n_Y}$$

Este tipo de contrastes se obtienen en R Commander con la opción **Estadísticos / Proporciones / Test de Proporciones para dos muestras...**



Ejemplo 9. Con los datos del fichero acero2.rda, ¿el porcentaje de horas en las que no hubo averías es menor cuando está apagado el sistema que cuando está encendido?

Solución: Hay que tener en cuenta que para realizar estos contrastes, debemos tener organizados los datos de forma que una variable tipo factor (Variable explicada) identifique a los individuos que cumplen la condición de los que no y una segunda variable tipo factor indique a través de sus valores la muestra o grupo a la que pertenece cada individuo (Grupos). Además R Commander considera, por defecto, que las proporciones a comparar están asociadas en la primera variable con la primera categoría por orden alfabético.

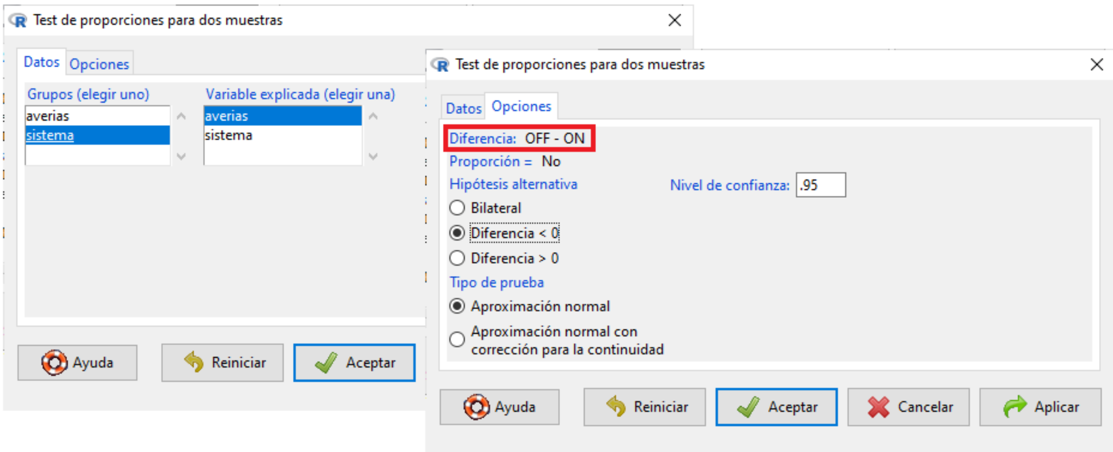
En este ejemplo, por lo tanto, la variable explicada seria *averias* y los grupos los daría la variable *sistema*. La proporción a comparar sería la de horas en las que no hay averías que corresponde al primer valor (No) de la variable averias, p_{No} y los grupos a comparar serían los dos valores de la variable *sistema* (OFF y ON).

Vamos a resolver con la ayuda de la salida de R Commander el contraste unilateral

H_0: $p_{\mathrm{No}}^{\mathrm{OFF}} \geq p_{\mathrm{No}}^{\mathrm{ON}}$ (igual o mejor con el sistema apagado)

H_1: $p_{\mathrm{No}}^{\mathrm{OFF}} < p_{\mathrm{No}}^{\mathrm{ON}}$ (peor con el sistema apagado)

Estadísticos →Proporciones →Test de proporciones para dos muestras




```
Percentage table:
   averias
sistema  No  Si    Total Count
  OFF   72.9 27.1  100  59
  ON    79.3 20.7  100  58

 2-sample test for equality of proportions without continuity correction

data: .Table
X-squared = 0.66405, df = 1, p-value = 0.2076
alternative hypothesis: less
95 percent confidence interval:
 -1.000000 0.065007
sample estimates:
  prop 1  prop 2
0.7288136 0.7931034
```

Como el p-valor = 0.2076 es mayor que los niveles de significación habituales, no podemos rechazar la hipótesis nula y por lo no podemos afirmar que el porcentaje de horas sin averías sea menor con el sistema apagado que con él encendido.

Nota: Los tests que utiliza R para este tipo de contrastes se basan en el uso del estadístico X-squared que coincide con el cuadrado del estadístico habitual T, mencionado anteriormente. El estadístico X-squared tiene distribución aproximada χ^2.

6. Ejercicios propuestos

Ejercicio 1.¿Apoyan estos datos la hipótesis de que el nivel medio de emisión de dióxido de carbono (CO2) supera las 100 t/h?

Ejercicio 2. ¿Apoyan estos datos la hipótesis de que el consumo promedio es menor de 130 megavatios-hora?

Ejercicio 3.¿Apoyan estos datos la hipótesis de que el consumo promedio es menor de 130 megavatios-hora en aquellas horas en las que la temperatura es alta? Realiza un diagrama de cajas de la variable consumo para cada una de las temperaturas consideradas y comenta los resultados.

Ejercicio 4.¿Apoyan estos datos la hipótesis de que el porcentaje de veces que se usa la línea A es mayor del 20%?

Ejercicio 5. La compra del sistema de sobrecalentamiento no ha sido rentable si, en general, éste es usado menos del 40% de las veces. Considerando los datos como una muestra aleatoria del comportamiento de esta empresa respecto a las variables en estudio, ¿permiten concluir que la compra del sistema no ha sido rentable?

Ejercicio 6. Se quiere comparar el consumo promedio con el sistema de detección de sobrecalentamiento encendido y apagado.

a) ¿Qué test sería adecuado para ello? Detalla los procedimientos que has llevado cabo para contestar a esta pregunta.

b) Si la hipótesis alternativa en el contraste es que el consumo medio con el sistema apagado es mayor que con el sistema encendido, ¿cuál es el p-valor asociado a dicho test? ¿qué se concluye?

Ejercicio 7. Realiza un contraste para determinar si el porcentaje de horas que el sistema de detección de sobrecalentamiento está apagado es mayor en la línea A que en la B. ¿Cuál es el p-valor asociado a dicho test?, ¿qué se concluye?

<u>Soluciones</u>

Ejercicio 1. ¿Apoyan estos datos la hipótesis de que el nivel medio de emisión de dióxido de carbono (CO2) supera las 100 t/h?

En primer lugar se debe contrastar la normalidad de la variable "nivel de emisión de dióxido de carbono (CO2)". Los resultados del test de normalidad de Shapiro-Wilk son:

```
> normalityTest(~CO2, test="shapiro.test", data=acero2)

Shapiro-Wilk normality test

data: CO2
W = 0.99238, p-value = 0.771
```

El p-valor del **test de normalidad** de Shapiro-Wilk aplicado a estos datos es 0.771. Como es claramente mayor que el nivel de significación (habitualmente $\alpha = 0.05$ y en general como máximo $\alpha = 0.1$), no se rechaza la hipótesis nula, es decir, no hay evidencias en contra de suponer la normalidad de la variable CO2.

Para contrastar si el nivel medio de emisión de dióxido de carbono supera las 100 t/h podemos plantearnos el contraste H_0: $\mu \leq 100$ frente a la alternativa H_1: $\mu > 100$ y se puede usar el **test t para una muestra** para realizar el contraste puesto que se puede asumir la normalidad de la variable CO2. Los resultados de dicho procedimiento con R son:

```
> with(acero2, (t.test(CO2, alternative='greater', mu=100.0, conf.level=.95)))

  One Sample t-test

data: CO2
t = 1.2095, df = 116, p-value = 0.1145
alternative hypothesis: true mean is greater than 100
…
sample estimates:
mean of x
 104.6281
```

y puesto que el p-valor es 0.1145, no se rechaza la hipótesis nula, es decir, no hay evidencias en contra de suponer que el nivel medio de emisión de dióxido de carbono supera las 100 t/h.

Ejercicio 2. ¿Apoyan estos datos la hipótesis de que el consumo promedio es menor de 130 megavatios-hora?

En primer lugar se debe contrastar la normalidad de la variable consumo. Los resultados del test de normalidad de Shapiro-Wilk son:

```
> normalityTest(~consumo, test="shapiro.test", data=acero2)

Shapiro-Wilk normality test

data: consumo
W = 0.99252, p-value = 0.7829
```

El p-valor del test de normalidad de Shapiro-Wilk aplicado a estos datos es 0.7829. Como es claramente mayor que el nivel de significación no se rechaza la hipótesis nula, es decir, no hay evidencias en contra de suponer la normalidad de la variable “consumo”.

Para contrastar si el consumo promedio es menor de 130 megavatios-hora podemos plantearnos el contraste $H_0: \mu \geq 130$ frente a la alternativa $H_1: \mu < 130$ y, puesto que se puede asumir la normalidad de la variable “consumo”, se puede usar el test t para una muestra para realizar el contraste.

Los resultados de dicho procedimiento con R son:

```
> with(acero2, (t.test(consumo, alternative='less', mu=130.0, conf.level=.95)))
```

```
  One Sample t-test

data: consumo
t = 1.8535, df = 116, p-value = 0.9668
alternative hypothesis: true mean is less than 130
…
sample estimates:
mean of x
 139.4565
```

y puesto que el p-valor es 0.9668, no se rechaza la hipótesis nula, es decir, no hay evidencias en contra de suponer que el consumo medio sea mayor o igual de 130 megavatios-hora.

Ejercicio 3.¿Apoyan estos datos la hipótesis de que el consumo promedio es menor de 130 megavatios-hora en aquellas horas en las que la temperatura es alta? Realiza un diagrama de cajas de la variable consumo para cada una de las temperaturas consideradas y comenta los resultados.

En este caso es necesario comenzar filtrando los datos y creando un nuevo conjunto de datos que llamaremos, por ejemplo, *acero_tempAlta*,

```
acero_tempAlta <- subset(acero2, subset=temperatura=="Alta")

NOTA: El conjunto de datos acero_tempAlta tiene 46 filas y 21 columnas.
```

A continuación debemos contrastar la normalidad de la variable consumo con temperaturas altas. Los resultados del test de normalidad de Shapiro-Wilk son:

```
> normalityTest(~consumo, test="shapiro.test", data=acero_tempAlta)

Shapiro-Wilk normality test

data: consumo
W = 0.97477, p-value = 0.4112
```

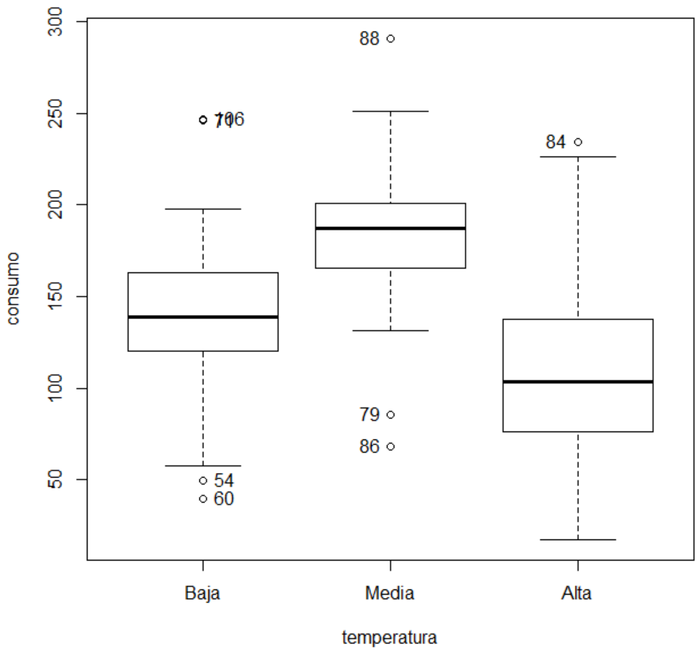
con lo que se puede suponer la normalidad y, por tanto, utilizar el test t para una muestra, de forma análoga al ejercicio anterior.

Los resultados de dicho test son:

```
> with(acero_tempAlta, (t.test(consumo, alternative='less', mu=130.0, conf.level=.95)))

One Sample t-test

data: consumo
t = -2.7268, df = 45, p-value = 0.004542
alternative hypothesis: true mean is less than 130
…
sample estimates:
mean of x
 109.4409
```

por lo tanto se rechaza la hipótesis nula ($H_0 : \mu \geq 130$) y se concluye que el consumo a temperaturas altas es menor de 130 megavatios-hora.

Si hacemos el correspondiente diagrama de cajas, lo primero debe ser volver a considerar el conjunto de datos acero2, puesto que debemos volver a los datos de origen. Hecho esto y el correspondiente diagrama de cajas para la variable consumo agrupada por temperatura obtenemos el gráfico:



donde se observa claramente como los consumos a temperaturas altas han sido los más bajos. De hecho, si calculamos la media global del consumo es de 139.4565 megavatios-hora y el consumo medio a temperatura alta ha sido de 109.4409 megavatios-hora. Este hecho explicaría que en general no se pueda concluir que el consumo es menor de 130, pero si nos restringimos a las horas con temperatura alta sí.

Ejercicio 4.¿Apoyan estos datos la hipótesis de que el porcentaje de veces que se usa la línea A es mayor del 20%?

En primer lugar se debe recodificar la variable línea que tiene tres valores o niveles en una variable dicotómica que tome solo los valores “A” y “no A”, la nueva variable se denominará *lineaAnoA*:

```
acero2 <- within(acero2, {
 lineaAnoA <- Recode(linea, '"A"="A"; "B"="noA"; "C"="noA"', as.factor=TRUE)
})
```

El primer nivel de la nueva variable es “A”, por lo que contraste H_0: $p \leq 0.20$ frente a la alternativa H_1: $p > 0.20$, donde p representa la proporción de veces que se usa la línea A, se resuelve con el procedimiento de R, **Test de proporciones para una muestra**:

```
Frequency counts (test is for first level):
lineaAnoA
 A noA
 39 78
```

```
  1-sample proportions test without continuity correction

data: rbind(.Table), null probability 0.2
X-squared = 13, df = 1, p-value = 0.0001557
alternative hypothesis: true p is greater than 0.2
…
sample estimates:
    p
0.3333333
```

Como el p-valor es 0.0001557 que es menor que los niveles de significación habituales, se rechaza la hipótesis nula y se concluye que existen evidencias significativas de que el porcentaje de veces que se usa la línea A es mayor del 20%.

Ejercicio 5. La compra del sistema de sobrecalentamiento no ha sido rentable si, en general, éste es usado menos del 40% de las veces. Considerando los datos como una muestra aleatoria del comportamiento de esta empresa respecto a las variables en estudio, ¿permiten concluir que la compra del sistema no ha sido rentable?

Si se denota p = Pr(OFF), con el procedimiento **Test de proporciones para una muestra** contrastaremos $H_0 : p \leq 0.6$ frente a la alternativa $H_1 : p > 0.6$, puesto que OFF < ON

```
Frequency counts (test is for first level):
sistema
OFF ON
 59 58

  1-sample proportions test without continuity correction

data: rbind(.Table), null probability 0.6
X-squared = 4.4672, df = 1, p-value = 0.9827
alternative hypothesis: true p is greater than 0.6
…
sample estimates:
    p
0.5042735
```

El p-valor es 0.9827, por lo que no hay evidencias de que Pr(OFF) > 0.6 o, lo que es lo mismo, que Pr(ON) < 0.4. Por lo tanto no hay evidencias significativas de que el sistema no es rentable.

Ejercicio 6. Se quiere comparar el consumo promedio con el sistema de detección de sobrecalentamiento encendido y apagado.

a) ¿Qué test sería adecuado para ello? Detalla los procedimientos que has llevado cabo para contestar a esta pregunta.

b) Si la hipótesis alternativa en el contraste es que el consumo medio con el sistema apagado es mayor que con el sistema encendido, ¿cuál es el p-valor asociado a dicho test? ¿qué se concluye?

Se consideran las variables
X = "consumo cuando el sistema de detección de sobrecalentamiento está apagado" e
Y = "consumo cuando el sistema de detección de sobrecalentamiento está encendido".

Los datos corresponden a dos muestras independientes.

- ¿Normalidad? **Test de normalidad de "consumo" por grupos de "sistema"**

```
> normalityTest(consumo ~ sistema, test="shapiro.test", data=acero2)
 --------
 sistema = OFF

Shapiro-Wilk normality test

data: consumo
W = 0.9904, p-value = 0.9237
 --------
 sistema = ON

Shapiro-Wilk normality test

data: consumo
W = 0.97637, p-value = 0.3156
```

Como ambos p-valores son mayores que los niveles de significación habituales, se pueden suponer muestras procedentes de poblaciones normales.

- ¿Varianzas iguales? **Test F para dos varianzas**

```
> var.test(consumo ~ sistema, alternative='two.sided', conf.level=.95, data=acero2)

F test to compare two variances

data: consumo by sistema
F = 0.85231, num df = 58, denom df = 57, p-value = 0.546
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5053356 1.4357938
sample estimates:
ratio of variances
    0.8523072
```

Puesto que el p-valor es 0.546, se puede suponer que las dos poblaciones normales tienen varianzas iguales.

- De todo lo anterior se deduce que el test adecuado es el **Test t para muestras independientes**, con la opción varianzas iguales activada.

```
> t.test(consumo~sistema, alternative='greater', conf.level=.95, var.equal=TRUE, data=acero2)

Two Sample t-test

data: consumo by sistema
t = 2.3668, df = 115, p-value = 0.009806
alternative hypothesis: true difference in means is greater than 0
…
```

```
sample estimates:
mean in group OFF mean in group ON
    151.1983     127.5122
```

El p-valor es 0.009806 con lo que se rechaza la hipótesis nula, es decir, el consumo medio con el sistema apagado es significativamente mayor que con el sistema encendido.

Ejercicio 7. Realiza un contraste para determinar si el porcentaje de horas que el sistema de detección de sobrecalentamiento está apagado es mayor en la línea A que en la B. ¿Cuál es el p-valor asociado a dicho test? ¿qué se concluye?

En primer lugar filtramos el archivo acero2 los datos correspondientes a las líneas A y B, es decir, seleccionamos los datos que no corresponden a la línea C:

```
acero2_lineaAyB <- subset(acero2, subset=linea!="C")
NOTA: El conjunto de datos acero2_lineaAyB tiene 78 filas y 22 columnas
```

A pesar del filtro, la variable linea sigue teniendo 3 niveles aunque el nivel C con 0 datos, por lo que descartamos niveles sin uso de la variable linea:

```
acero2_lineaAyB <- within(acero2_lineaAyB, {
 linea <- droplevels(linea)
})
```

Si p = Pr(OFF), contrastaremos H_0: $p_A \leq p_B$ frente a la alternativa H_1: $p_A > p_B$ con el **Test para proporciones para dos muestras**. Los resultados son los siguientes:

```
Percentage table:
   sistema
linea    OFF  ON Total Count
  A  46.2 53.8  100  39
  B  51.3 48.7  100  39

  2-sample test for equality of proportions without continuity correction

data: .Table
X-squared = 0.20526, df = 1, p-value = 0.6747
alternative hypothesis: greater
…
sample estimates:
  prop 1  prop 2
0.4615385 0.5128205
```

Como el p-valor es 0.6747, no hay evidencias significativas de que el porcentaje de horas que el sistema de detección de sobrecalentamiento está apagado sea mayor en la línea A que en la B.