Lección 3

Contrastes de hipótesis

En esta lección explicamos algunas instrucciones de `R` que permiten llevar a cabo de manera rápida contrastes de hipótesis.[1] En concreto, las funciones que explicamos son las siguientes:

- `z.test`, para realizar Z-tests para contrastar una media
- `t.test`, para realizar t-tests para contrastar una o dos medias (tanto de poblaciones independientes como emparejadas)
- `sigma.test`, para realizar χ^2-tests para contrastar una desviación típica
- `var.test`, para realizar F-tests para contrastar dos desviaciones típicas
- `binom.test`, para realizar tests binomiales para contrastar una proporción o dos proporciones de poblaciones emparejadas
- `prop.test`, para realizar prop-tests para contrastar una proporción o dos proporciones de poblaciones independientes grandes
- `fisher.test`, para realizar el test exacto de Fisher para contrastar dos proporciones de poblaciones independientes
- `mcnemar.test`, para realizar el test de McNemar para contrastar dos proporciones de poblaciones emparejadas

[1] **¡Atención!** Las fórmulas que hemos explicado en el curso son las más sencillas posibles. En la mayoría de los tests, `R` usa versiones optimizadas y más sofisticadas de estas fórmulas, que no siempre dan el mismo resultado que si hacemos el test "a mano", usando la fórmula vista en el curso. Por lo tanto, no os conviene usar las instrucciones que explicamos en esta lección al hacer los tests de contenidos de la asignatura, donde suponemos que empleáis exactamente las fórmulas explicadas en clase. Naturalmente, *sí* que tenéis que usar las instrucciones explicadas aquí para hacer el test de esta lección, en los talleres y en la vida real.

Como veremos, todas estas funciones tienen una sintaxis similar (salvo particularidades debidas al propio test, como por ejemplo a qué información se aplica). El grueso del capítulo está dedicado a ejemplos de uso de estas funciones.

Recordemos en lo que sigue que el *nivel de significación* α de un contraste es la probabilidad de cometer un *error de tipo I*, es decir, la probabilidad de rechazar la hipótesis nula si es verdadera. El *nivel de confianza* es el complementario del nivel de significación, $1-\alpha$, y por lo tanto es la probabilidad de no rechazar la hipótesis nula si es verdadera.

3.1. Contrastes para medias

Consideremos en primer lugar la situación más sencilla: tenemos una población que sigue una distribución normal de valor medio μ desconocido y desviación típica σ conocida. Si queremos contrastar la hipótesis nula $H_0 : \mu = \mu_0$, para un valor dado μ_0, con una hipótesis alternativa $H_1 : \mu > \mu_0$, $H_1 : \mu < \mu_0$ o $H_1 : \mu \neq \mu_0$, realizamos lo que llamamos un *Z-test*. Este test usa el estadístico

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}},$$

que sigue una ley normal estándar si la hipótesis nula es verdadera, y consiste básicamente en los pasos siguientes:

(1) Calculamos el valor z_0 del estadístico Z sobre la muestra.

(2) Calculamos el p-valor del contraste: $P(Z \geqslant z_0)$ si $H_1 : \mu > \mu_0$; $P(Z \leqslant z_0)$ si $H_1 : \mu < \mu_0$; o $2P(Z \geqslant |z_0|)$ si $H_1 : \mu \neq \mu_0$.

(3) Si este p-valor es menor que el nivel de significación α, rechazamos la hipótesis nula, y si no, no.

El Z-test está implementado en `R` en una función `z.test` que forma parte del paquete `TeachingDemos`; por lo tanto, para usarla primero hace falta instalar y cargar este paquete. La estructura básica de esta instrucción es

```
z.test(x, mu=..., sd=..., alternative=..., conf.level=...)
```

donde

- `x` es el vector de datos que forma la muestra que analizamos.
- El valor del parámetro `mu` es el valor μ_0 que queremos contrastar.
- El valor del parámetro `sd` es la desviación típica (conocida) de la población.
- El parámetro `alternative` puede tomar tres valores: `"two.sided"`, que representa la hipótesis alternativa $H_1 : \mu \neq \mu_0$; `"less"`, que corresponde a $H_1 : \mu < \mu_0$; y `"greater"`, que corresponde a $H_1 : \mu > \mu_0$. En este test, y en todos los que siguen, su valor por defecto, que no hace falta especificar, es `"two.sided"`.

- El valor del parámetro `conf.level` es el nivel de confianza $1-\alpha$. En este test, y en todos los que siguen, su valor por defecto, que no es necesario especificar, es `0.95`.

Veamos un ejemplo de su uso.

Ejemplo 3.1. Tenemos una muestra formada por los 25 números siguientes:

> 2.2, 2.66, 2.74, 3.41, 2.46, 2.96, 3.34, 2.16, 2.46, 2.71, 2.04, 3.74, 3.24, 3.92, 2.38, 2.82, 2.2, 2.42, 2.82, 2.84, 4.22, 3.64, 1.77, 3.44, 1.53.

Supongamos que esta muestra ha sido extraída de una población normal con desviación típica $\sigma = 0.8$. Creemos que el valor medio μ de la población no es 2. Para confirmarlo, vamos realizar el contraste

$$\begin{cases} H_0 : \mu = 2 \\ H_1 : \mu \neq 2 \end{cases}$$

con nivel de significación $\alpha = 0.05$:

```
> x=c(2.2,2.66,2.74,3.41,2.46,2.96,3.34,2.16,2.46,2.71,2.04,
  3.74,3.24,3.92,2.38,2.82,2.2,2.42,2.82,2.84,4.22,3.64,1.77,
  3.44,1.53)
> #Instalamos y cargamos "TeachingDemos"
...
> z.test(x, mu=2, sd=0.8, alternative="two.sided",
   conf.level=0.95)

         One Sample z-test

data:  x
z = 5.03, n = 25.00, Std. Dev. = 0.80, Std. Dev. of the sample
    mean = 0.16,
p-value = 4.905e-07
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 2.491206 3.118394
sample estimates:
mean of x
   2.8048
```

(Como los parámetros `alternative="two.sided"` y `conf.level=0.95` eran los que toma `R` por defecto, en realidad no hacía falta especificarlos.) Observad la información que obtenemos con esta instrucción:

- Información sobre la muestra x: longitud $n = 25$, la media de la muestra (`mean of x`) $\overline{x} = 2.8048$, y el error estándar $\sigma/\sqrt{n}$ del estadístico $\overline{X}$ (`Std. Dev. of the sample mean`), en este caso $0.8/\sqrt{25} = 0.16$.
- La hipótesis alternativa (`alternative hypothesis`), en este caso `true mean is not equal to 2`: la media verdadera, o poblacional, μ es diferente de 2.

- El valor `z` que toma el estadístico $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ sobre la muestra, en este caso 5.03.
- El p-valor (`p-value`) de nuestro test, en este caso `4.905e-07`, es decir, $4.905 \cdot 10^{-7}$.
- Un intervalo de confianza de nivel $100(1-\alpha)\,\%$ (en nuestro caso, `95 percent confidence interval`) para la μ en este test: en nuestro ejemplo, $[2.491206, 3.118394]$.

Lo único que no nos dice directamente es si tenemos que rechazar o no la hipótesis nula, pero esto lo deducimos del p-valor: como es más pequeño que el nivel de significación (de hecho, es *muy* pequeño), podemos rechazar la hipótesis nula, $\mu = 2$, en favor de la alternativa $\mu \neq 2$. Es decir, hay evidencia muy significativa para afirmar que $\mu \neq 2$. Otra manera de decidir si rechazamos o no la hipótesis nula es mirar si el valor que contrastamos pertenece al intervalo de confianza. Puesto que $2 \notin [2.491206, 3.118394]$, podemos rechazar la hipótesis nula.

Repitamos ahora el mismo test cambiando la hipótesis alternativa por $H_1 : \mu < 3$, es decir,

$$\begin{cases} H_0 : \mu = 3 \\ H_1 : \mu < 3 \end{cases}$$

y tomando como nivel de significación $\alpha = 0.1$:

```
> z.test(x, mu=3, sd=0.8, alternative="less", conf.level=0.9)

          One Sample z-test

data:  x
z = -1.22, n = 25.00, Std. Dev. = 0.80, Std. Dev. of the sample
    mean = 0.16, p-value = 0.1112
alternative hypothesis: true mean is less than 3
90 percent confidence interval:
     -Inf 3.009848
sample estimates:
mean of x
   2.8048
```

En este caso, el p-valor es 0.1112, mayor que 0.1, así que no podemos rechazar la hipótesis nula. El intervalo de confianza es ahora $]-\infty, 3.009848[$ (`Inf` representa ∞). El hecho que contenga el 3 (aunque por muy poco) también indica que no podemos rechazar la hipótesis nula.

En la mayoría de situaciones de la vida real, el Z-test es inútil, porque la desviación típica poblacional σ suele ser desconocida (de hecho, es por este motivo que `R` no trae definida por defecto una función para efectuarlo). En tal caso hay que utilizar el contraste más popular de `R`, el *t-test*. Este test usa diferentes estadísticos según que el contraste sea de una muestra o de dos, y en este último caso, según se comparen muestras independientes de poblaciones con varianzas diferentes, muestras independientes de poblaciones con varianzas

iguales, o muestras emparejadas. Aunque el test sólo es exacto (en el sentido de que da la conclusión con el nivel de significación requerido) cuando las poblaciones involucradas siguen distribuciones normales, también da buenos resultados en el caso de muestras grandes de poblaciones no normales, por lo que en este caso también se recomienda su uso. En la práctica, el t-test se usa como test "de talla única" para contrastar una o dos medias en cualquier situación, pero hay que tener claro que *si la muestra es pequeña y las poblaciones no siguen leyes normales, el resultado del test no es de ninguna manera fiable.*

Este test se efectúa con `R` con la instrucción `t.test`, cuya estructura básica es

```
t.test(x, y, mu=..., alternative=..., conf.level=..., paired=...,
  var.equal=..., na.omit=...)
```

donde:

- `x` es el vector de datos que forma la muestra que analizamos.
- `y` es un vector opcional; si lo entramos, `R` entiende que estamos haciendo un contraste de las dos medias, con hipótesis nula la igualdad de estas medias.
- Podemos sustituir los vectores `x` e `y` por una fórmula `variable1~variable2` que indique que separamos la variable numérica `variable1` en dos vectores definidos por los niveles de un factor `variable2` de dos niveles (o de otra variable asimilable a un factor de dos niveles, como por ejemplo una variable numérica que sólo tome dos valores diferentes). Con esta construcción, `R` entenderá estos vectores como ordenados por el orden natural de los niveles de `variable2`: `x` será el vector correspondiente al primer nivel, e `y` el correspondiente al segundo. Hay que tener esto en cuenta a la hora de especificar la hipótesis alternativa si es unilateral. Si las dos variables de la fórmula son columnas de un *dataframe*, se puede usar el parámetro `data=...` para especificarlo. Veremos varios ejemplos más adelante.
- El valor del parámetro `mu` sólo lo tenemos que especificar si hemos entrado una sola muestra, y en este caso es el valor μ_0 que queremos contrastar.
- El parámetro `alternative` puede tomar los tres mismos valores que en `z.test`, y su significado dependerá del tipo de test que efectuemos:
  - Si el test es de una sola muestra, el significado es el mismo que en `z.test`.
  - Si hemos entrado dos muestras y llamamos μ_x y μ_y a las medias de las poblaciones de las que hemos extraído las muestras x e y, respectivamente, entonces `"two.sided"` indica que la hipótesis alternativa es $H_1 : \mu_x \neq \mu_y$; `"less"`, indica que la hipótesis alternativa es $H_1 : \mu_x < \mu_y$; y `"greater"`, que la hipótesis alternativa es $H_1 : \mu_x > \mu_y$. Como en `z.test`, su valor por defecto, que no es necesario especificar, es `"two.sided"`.
- El valor del parámetro `conf.level` es el nivel de confianza $1 - \alpha$. Como en `z.test`, su valor por defecto, que no hace falta especificar, es `conf.level=0.95`

- El parámetro `paired` sólo lo tenemos que especificar si llevamos a cabo un contraste de dos medias. En este caso, si entramos `paired=TRUE`, estamos diciendo que las muestras son emparejadas, mientras que si entramos `paired=FALSE` (que es su valor por defecto), estamos diciendo que las muestras son independientes. Si se trata de muestras emparejadas, los vectores `x` e `y` tienen que tener la misma longitud, naturalmente.

- El parámetro `var.equal` sólo lo tenemos que especificar si llevamos a cabo un contraste de dos medias de poblaciones independientes, y en este caso sirve para indicar si queremos considerar las dos varianzas iguales (igualándolo a `TRUE`) o diferentes (igualándolo a `FALSE`, que es su valor por defecto).

- El parámetro `na.action` sirve para especificar qué queremos hacer con los valores `NA`. Es un parámetro genérico que se puede usar en casi todas las funciones de análisis de datos, y sus posibles valores son:

  - `na.omit`, su valor por defecto, elimina las entradas `NA` de los vectores (o los pares que contengan algún `NA`, en el caso de muestras emparejadas).
  - `na.exclude`, es como `na.omit`, pero guarda la información de las posiciones de los `NA` eliminados y la usa en algunas funciones donde esto es necesario (no en estos tests).
  - `na.fail` hace que la ejecución pare si encuentra algún `NA`.
  - `na.pass` no hace nada con los `NA` y permite que las operaciones internas de la función sigan su curso y los manejen como les corresponda.

  Por ahora, la opción por defecto `na.action=na.omit` es la adecuada, por lo que no hace falta usar este parámetro, pero es bueno saber que hay alternativas.

La función `t.test` tiene otros parámetros que se pueden consultar con el `help`. Veamos varios ejemplos de uso.

Ejemplo 3.2. Supongamos que sabemos que el vector del Ejemplo 3.1 proviene de una población normal, pero ahora desconocemos su varianza. Seguimos queriendo contrastar si la media de la población es 2 o no. En este caso, podemos hacer:

```
> t.test(x, mu=2) #No incluimos los parámetros con valores por
    defecto

        One Sample t-test

data:  x
t = 5.912, df = 24, p-value = 4.232e-06
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 2.523844 3.085756
sample estimates:
```

```
mean of x
   2.8048
```

Observemos que la estructura de la respuesta es similar a la de `z.test`:

- Nos da el valor del estadístico T sobre la muestra, `t = 5.912`, sus grados de libertad (*degrees of freedom*, en inglés), `df = 24`, y el p-valor del contraste, `p-value = 4.232e-06`.
- Con `alternative hypothesis: true mean is not equal to 2` nos recuerda que la hipótesis alternativa es $\mu \neq 2$.
- Nos da el intervalo de confianza al 95 % de este test para μ, `95 percent confidence interval`, en este caso $[2.523844, 3.085756]$, y la media $\overline{x}$ del vector x, en este caso 2.8048.

De esta respuesta nos interesa el p-valor, que es $4.232 \cdot 10^{-6}$, muy pequeño, que nos permite rechazar la hipótesis nula $H_0 : \mu = 2$ en favor de la hipótesis alternativa $H_1 : \mu \neq 2$. También nos puede interesar el intervalo de confianza del 95 % para la media poblacional.

El p-valor y el intervalo de confianza se pueden obtener directamente, añadiendo a la instrucción `t.test` los sufijos `$p.value` o `$conf.int`, respectivamente.

```
> t.test(x, mu=2)$p.value
[1] 4.231586e-06
> t.test(x, mu=2)$conf.int
[1] 2.523844 3.085756
attr(,"conf.level")
[1] 0.95
> t.test(x, mu=2)$conf.int[1]
[1] 2.523844
> t.test(x, mu=2)$conf.int[2]
[1] 3.085756
```

Podéis consultar los sufijos necesarios para obtener las otras componentes del resultado en el `help` de la función. Estos mismos sufijos se pueden usar con `z.test` y con cualquiera de los otros tests que explicaremos en esta lección.

Ejemplo 3.3. Tenemos una muestra del nivel de colesterol en plasma de 9 individuos, en mg/dl. Los datos son

$$203, 229, 215, 220, 223, 233, 208, 228, 209.$$

Queremos contrastar si el valor medio del nivel de colesterol en la población es de 220 mg/dl o no, a un nivel de significación del 10 %. Suponemos que este nivel de colesterol en plasma sigue una ley normal.

En este caso, podemos realizar un t-test:

```
> colesterol=c(203,229,215,220,223,233,208,228,209)
> t.test(colesterol, mu=220, alternative="two.sided",
   conf.level=0.9)

        One Sample t-test

data:  colesterol
t = -0.3801, df = 8, p-value = 0.7138
alternative hypothesis: true mean is not equal to 220
90 percent confidence interval:
 212.1435 225.1898
sample estimates:
mean of x
 218.6667
```

El p-valor es 0.7138, muy grande y en particular superior a 0.1, por lo tanto no podemos rechazar la hipótesis nula de que el valor medio sea 220 mg/dl. Además, el intervalo de confianza del 90 % del contraste es $[212.1435, 225.1898]$, y contiene el valor 220.

Ejemplo 3.4. Recordad el *dataframe* `iris`, que recoge datos de las flores de 50 ejemplares de cada una de tres especies de iris.

```
> str(iris)
'data.frame': 150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5
   ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1
   ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1
   1 1 1 1 1 1 1 1 1 ...
```

Queremos estudiar si la longitud media μ_v de los sépalos de las *Iris virginica* es mayor que la longitud media μ_s de los sépalos de las *Iris setosa*. Para ello realizamos el contraste

$$\begin{cases} H_0 : \mu_s = \mu_v \\ H_1 : \mu_s < \mu_v \end{cases}$$

En este caso, se trata de un contraste de dos muestras independientes. Como las muestras son grandes, podemos usar con garantías el t-test.

Como no sabemos nada de las varianzas, y no nos supone apenas esfuerzo realizar los tests, llevaremos a cabo el contraste en los dos casos: varianzas iguales y varianzas diferentes.[2]

[2] En realidad, se sabe que si las dos muestras provienen de poblaciones normales y son del mismo tamaño, el t-test tiende a dar la misma conclusión tanto si se supone que las dos varianzas son iguales

```
> S=iris[iris$Species=="setosa",]$Sepal.Length
> V=iris[iris$Species=="virginica",]$Sepal.Length
> t.test(S, V, alternative="less", var.equal=TRUE) #varianzas
    iguales

         Two Sample t-test

data:  S and V
t = -15.3862, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.411263
sample estimates:
mean of x mean of y
    5.006     6.588

> t.test(S, V, alternative="less", var.equal=FALSE) #varianzas
    diferentes

         Welch Two Sample t-test

data:  S and V
t = -15.3862, df = 76.516, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.410804
sample estimates:
mean of x mean of y
    5.006     6.588
```

En los dos casos el p-valor es prácticamente 0 y por lo tanto podemos rechazar la hipótesis nula: tenemos evidencia muy significativa de que, en promedio, las flores de la especie setosa tienen sépalos más cortos que las de la especie virginica. El intervalo de confianza del 95 % para la diferencia de medias $\mu_s - \mu_v$ en este contraste es, en ambos casos, aproximadamente $]-\infty, -1.41]$ y no contiene el 0, que es la hipótesis nula para la diferencia de medias.

Observad que la función `t.test` en el caso de dos muestras independientes con varianzas poblacionales diferentes usa una distribución t de Student con un número no entero de grados de libertad, y por lo tanto, en particular, no usa la fórmula que hemos explicado en clase.

Ejemplo 3.5. Para comprobar la eficacia de un cierto tratamiento del eccema atípico, se tomaron 10 pacientes con un eccema de más de 9 meses y se los trató durante 3 semanas

como si se supone que son diferentes (véase C. A. Markowski y E. P. Markowski, "Conditions for the Effectiveness of a Preliminary Test of Variance," *The American Statistician* 44 (1990), pp. 322–326). Por lo tanto, si en este caso supiéramos que estas longitudes tienen distribuciones normales, bastaría realizar uno de los dos tests.

con un placebo y a continuación se los trató durante 3 semanas más con el tratamiento bajo estudio. Después de cada periodo de tratamiento, se evaluó la importancia del eccema en una escala de 0 (no eccema) a 10 (máximo eccema), y se anotaron los resultados. Los datos que se obtuvieron fueron los siguientes:

Placebo	6	8	4	8	5	6	5	6	4	5
Tratamiento	5	6	4	5	3	6	6	2	2	6

Queremos determinar si el tratamiento es más eficaz que el placebo. Una manera de estudiarlo es comparando las medias de las valoraciones de la importancia del eccema después de los dos tratamientos. Digamos μ_p y μ_t a las medias después del placebo y del tratamiento, respectivamente. Tomaremos como hipótesis nula $H_0 : \mu_p = \mu_t$ (que indica que el tratamiento es tan ineficaz como el placebo) e hipótesis alternativa $H_1 : \mu_p > \mu_t$ (que representa que el tratamiento ha sido eficaz: la importancia media del eccema es menor después del tratamiento que después del placebo). Si podemos rechazar la hipótesis nula en favor de la alternativa, significará el tratamiento ha demostrado ser eficaz.

Observad que se trata de un contraste de dos muestras emparejadas, porque los datos refieren a los mismos 10 pacientes. Vamos a suponer que las valoraciones del eccema en ambos supuestos siguen leyes normales y que, por lo tanto, el resultado de un t-test es fiable.

```
> placebo=c(6,8,4,8,5,6,5,6,4,5)
> tratamiento=c(5,6,4,5,3,6,6,2,2,6)
> t.test(placebo, tratamiento, alternative="greater",
   paired=TRUE)

          Paired t-test

data:  placebo and tratamiento
t = 2.25, df = 9, p-value = 0.0255
alternative hypothesis: true difference in means is greater than
    0
95 percent confidence interval:
 0.2223398       Inf
sample estimates:
mean of the differences
                   1.2
```

El p-valor es 0.0255, menor que 0.5, y por lo tanto podemos rechazar la hipótesis nula con un nivel de significación del 5 %. El intervalo de confianza del 95 % para la diferencia de medias $\mu_p - \mu_t$ es $[0.222, \infty[$ y está a la derecha del 0. Por lo tanto, hay evidencia significativa de que el tratamiento es mejor que el placebo.

Ejemplo 3.6. Veamos un ejemplo de aplicación de `t.test` a una fórmula. Queremos contrastar si es cierto que fumar durante el embarazo implica un peso menor del recién nacido. Si llamamos μ_n y μ_f al peso medio de un recién nacido de madre no fumadora y

fumadora, respectivamente, el contraste que queremos realizar es

$$\begin{cases} H_0 : \mu_n = \mu_f \\ H_1 : \mu_n > \mu_f \end{cases}$$

Vamos a usar los datos recogidos en la tabla de datos `birthwt` incluida en el paquete `MASS`, que recoge algunos datos sobre una muestra de embarazadas y sus hijos.

```
> #Instalamos y cargamos el paquete MASS, si aun no lo está
...
> str(birthwt)
'data.frame': 189 obs. of  10 variables:
 $ low   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age   : int  19 33 20 21 18 21 22 17 29 26 ...
 $ lwt   : int  182 155 105 108 107 124 118 103 123 113 ...
 $ race  : int  2 3 1 1 1 3 1 3 1 1 ...
 $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
 $ ptl   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ht    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ui    : int  1 0 0 1 1 0 0 0 0 0 ...
 $ ftv   : int  0 3 1 2 0 0 1 1 1 0 ...
 $ bwt   : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665
    ...
> help(birthwt)
```

Con `help(birthwt)` nos enteramos de que la variable `smoke` indica si la madre ha fumado durante el embarazo (1) o no (0), y que la variable `bwt` da el peso del recién nacido en gramos. Lo primero que haremos será mirar si las muestras de madres fumadoras y no fumadoras contenidas en esta tabla son lo suficientemente grandes como para que el resultado del t-test sea fiable.

```
> table(birthwt$smoke)
  0   1
115  74
```

Sí que lo son.

Para entrar en la instrucción `t.test` los vectores de pesos de hijos de fumadoras y no fumadoras, usaremos la fórmula `bwt~smoke` especificando que `data=birthwt`. Fijaos que los valores de `smoke` son 0 y 1, y que R los considera ordenados en este orden (basta ver el resultado del `table` anterior). Por consiguiente, `bwt~smoke` representa, en este orden, el vector de pesos de recién nacidos de madres no fumadoras (`smoke=0`) y el vector de pesos de recién nacidos de madres fumadoras (`smoke=1`). Como la hipótesis alternativa es $\mu_n > \mu_f$, deberemos especificar en el t-test que `alternative="greater"`.

Como en el Ejemplo 3.4, vamos a realizar el t-test suponiendo que las varianzas son iguales y que son diferentes, y cruzaremos los dedos para que la conclusión sea la misma en ambos casos.

```
> t.test(bwt~smoke, data=birthwt, alternative="greater",
  paired=FALSE, var.equal=TRUE)

          Two Sample t-test

data:  bwt by smoke
t = 2.6529, df = 187, p-value = 0.004333
alternative hypothesis: true difference in means is greater than
    0
95 percent confidence interval:
 106.9528      Inf
sample estimates:
mean in group 0 mean in group 1
       3055.696        2771.919

> t.test(bwt~smoke, data=birthwt, alternative="greater",
   paired=FALSE, var.equal=FALSE)

          Welch Two Sample t-test

data:  bwt by smoke
t = 2.7299, df = 170.1, p-value = 0.003501
alternative hypothesis: true difference in means is greater than
    0
95 percent confidence interval:
 111.8548      Inf
sample estimates:
mean in group 0 mean in group 1
       3055.696        2771.919
```

En ambos casos hemos obtenido un p-valor inferior a 0.05, lo que nos permite afirmar que, en efecto, las madres no fumadoras tienen en promedio hijos más grandes que las fumadoras.

En vez de especificar los vectores de pesos con `bwt~smoke,data=birthwt`, hubiéramos podido usar `birthwt$bwt~birthwt$smoke`. Por ejemplo:

```
> t.test(birthwt$bwt~birthwt$smoke, alternative="greater",
  paired=FALSE, var.equal=TRUE)

          Two Sample t-test

data:  birthwt$bwt by birthwt$smoke
t = 2.6529, df = 187, p-value = 0.004333
alternative hypothesis: true difference in means is greater than
    0
95 percent confidence interval:
```

```
 106.9528      Inf
sample estimates:
mean in group 0 mean in group 1
       3055.696        2771.919
```

3.2. Contrastes para varianzas

El paquete `TeachingDemos` también lleva una función que permite efectuar contrastes sobre el valor de la desviación típica σ de una población normal; este tipo de contraste usa el estadístico

$$\chi^2_{n-1} = \frac{(n-1)\widetilde{S}^2_X}{\sigma_0^2}$$

que, si la hipótesis nula $H_0 : \sigma = \sigma_0$ es verdadera, sigue una distribución χ^2_{n-1}, y en consecuencia se le llama χ^2-test. La función de `TeachingDemos` para realizar un χ^2-test es `sigma.test`, y su estructura básica es

```
sigma.test(x, sigma=..., alternative=..., conf.level=...)
```

donde

- `x` es el vector de datos que forma la muestra que analizamos.
- `sigma` es el valor de la desviación típica que queremos contrastar.
- El significado de `alternative` y `conf.level`, y sus posibles valores, son los usuales.

Ejemplo 3.7. Se ha realizado un experimento para estudiar el tiempo X (en minutos) que tarda un lagarto del desierto en llegar a los 45º partiendo de su temperatura normal mientras está a la sombra. Los tiempos obtenidos (en minutos) en una muestra aleatoria de lagartos fueron

$$10.1, 12.5, 12.2, 10.2, 12.8, 12.1, 11.2, 11.4, 10.7, 14.9, 13.9, 13.3.$$

Supongamos que estos tiempos siguen una ley normal. ¿Aporta este experimento evidencia de que la desviación típica σ de X es inferior a 1.5 minutos? Para responder esta pregunta, hemos de realizar el contraste

$$\begin{cases} H_0 : \sigma \geqslant 1.5 \\ H_1 : \sigma < 1.5 \end{cases}$$

```
> #Cargamos TeachingDemos si no lo está
> TL45=c(10.1,12.5,12.2,10.2,12.8,12.1,11.2,11.4,10.7,14.9,13.9,
   13.3)
> sigma.test(TL45, sigma=1.5, alternative="less")

        One sample Chi-squared test for variance
```

```
data:  TL45
X-squared = 10.6885, df = 11, p-value = 0.5303
alternative hypothesis: true variance is less than 2.25
95 percent confidence interval:
 0.000000 5.256863
sample estimates:
var of TL45
   2.186288
```

El p-valor que obtenemos es 0.5303, muy grande, por lo que no tenemos evidencia que nos permita rechazar la hipótesis nula en favor de $\sigma < 1.5$.

El χ^2-test no se usa mucho en la práctica. En parte, porque realmente es poco interesante ya que suele ser difícil conjeturar la desviación típica a contrastar, y en parte porque su validez depende fuertemente de la hipótesis de que la variable aleatoria poblacional sea normal. En cambio, el contraste de las desviaciones típicas de dos muestras sí que es muy utilizado. Por ejemplo, en un contraste de dos muestras de tamaños diferentes de poblaciones normales independientes, nos puede interesar conocer *a priori* si las varianzas poblacionales son iguales o diferentes, en lugar de realizar el test bajo ambas suposiciones (imaginaos que los tests para varianzas iguales y varianzas diferentes dan conclusiones diferentes...). Si no las conocemos, ¿cómo podemos saber cuál es el caso? Usando el *var-test*, basado en el estadístico

$$F = \frac{\widetilde{S}^2_{X_1}}{\widetilde{S}^2_{X_2}}$$

que, si las dos poblaciones son normales, sigue una distribución F de Fisher. Por desgracia, este test es también muy sensible a la no normalidad de las poblaciones objeto de estudio: a la que una de ellas se aleja un poco de la normalidad, el test deja de dar resultados fiables.[3]

La instrucción para efectuar este test es `var.test` y su estructura básica es

```
var.test(x, y, alternative=..., conf.level=...)
```

donde

- `x` e `y` son los dos vectores de datos. Se pueden especificar mediante una fórmula como en el caso de `t.test`.
- El parámetro `alternative` puede tomar los tres mismos valores que en los tests anteriores. Si llamamos σ_x^2 y σ_y^2 a las varianzas de las poblaciones de las que hemos extraído las muestras x e y, respectivamente, entonces `"two.sided"` indica que la hipótesis alternativa es $H_1 : \sigma_x^2 \neq \sigma_y^2$; `"less"`, indica que la hipótesis alternativa es $H_1 : \sigma_x^2 < \sigma_y^2$; y `"greater"`, que la hipótesis alternativa es $H_1 : \sigma_x^2 > \sigma_y^2$. Su valor

[3] Véanse: E. S. Pearson, "The analysis of variance in cases of non-normal variation," *Biometrika* 23 (1931), pp. 114–133; G. E. P. Box, "Non-normality and tests on variances," *Biometrika* 40 (1953), pp. 318–335.

por defecto es `"two.sided"`, que es el que nos permite contrastar si las varianzas son iguales o diferentes.

- El valor del parámetro `conf.level` es, como siempre, el nivel de confianza $1-\alpha$, y su valor por defecto es `0.95`.

Para conocer otros parámetros, podéis consultar el `help`.

Ejemplo 3.8. Suponiendo que las longitudes de los sépalos de las flores de las diferentes especies de iris siguen leyes normales, ¿hubiéramos podido considerar de entrada iguales o diferentes las varianzas de las dos muestras en el Ejemplo 3.4? Veámoslo:

```
> S=iris[iris$Species=="setosa",]$Sepal.Length
> V=iris[iris$Species=="virginica",]$Sepal.Length
> var.test(S,V)

        F test to compare two variances

data:  S and V
F = 0.3073, num df = 49, denom df = 49, p-value = 6.366e-05
alternative hypothesis: true ratio of variances is not equal to
   1
95 percent confidence interval:
 0.1743776 0.5414962
sample estimates:
ratio of variances
         0.3072862
```

El p-valor es $6.366 \cdot 10^{-5}$, muy pequeño. Por lo tanto podemos rechazar la hipótesis nula de que las dos varianzas son iguales, en favor de la hipótesis alternativa de que las dos varianzas son diferentes. Esto nos hubiera permitido haberlas considerado de entrada diferentes en el t-test, y sólo haber realizado el `t.test` con `var.equal=FALSE`.

Observemos también que `var.test` nos da el intervalo de confianza al nivel especificado (o al 95 % si usamos el nivel de confianza por defecto) para el cociente de las varianzas σ_x^2/σ_y^2. En este caso, el intervalo de confianza al 95 % es $[0.1743776, 0.5414962]$ y como no contiene el 1, confirma la evidencia de que $\sigma_x^2 \neq \sigma_y^2$.

Ejemplo 3.9. Queremos contrastar si, de media, los gatos macho «grandes» (digamos, de más de 2 kg) son mayores que los gatos hembra «grandes». Para ello usaremos los datos recogidos en el *dataframe* `cats` que viene con el paquete `MASS` y que contiene información sobre el peso de una muestra de gatos de más de 2 kg, separados por su sexo.

```
> #Cargamos el paquete MASS
> str(cats)
'data.frame': 144 obs. of  3 variables:
 $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
```

```
 $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
> table(cats$Sex)
 F  M
47 97
> help(cats)
```

Consultando `help(cats)` nos enteramos de que la variable `Bwt` contiene el peso de cada gato en kg, y la variable `Sex` contiene el sexo de cada gato: F para hembra (*female*) y M para macho (*male*). Como vemos en la tabla de contingencia, los números de ejemplares de cada sexo son diferentes y grandes.

Así pues, si llamamos μ_m al peso medio de un gato macho de más de 2 kg y μ_h al peso medio de un gato hembra de más de 2 kg, el contraste que vamos a realizar es

$$\begin{cases} H_0 : \mu_m = \mu_h \\ H_1 : \mu_m > \mu_h \end{cases}$$

y para ello antes vamos a contrastar si las varianzas de ambas poblaciones son iguales o diferentes, para luego poder aplicar el `t.test` con el valor de `var.equal` adecuado. Vamos a suponer que los pesos en ambos sexos siguen leyes normales. Para que el contraste de las varianzas sea fiable es necesario que esta suposición sea cierta; para el de los pesos medios, no, ya que ambas muestras son grandes.

Contrastemos la igualdad de las varianzas:

```
> var.test(Bwt~Sex, data=cats)

    F test to compare two variances

data:  Bwt by Sex
F = 0.3435, num df = 46, denom df = 96, p-value = 0.0001157
alternative hypothesis: true ratio of variances is not equal to
   1
95 percent confidence interval:
 0.2126277 0.5803475
sample estimates:
ratio of variances
         0.3435015
```

El p-valor es muy pequeño, y por lo tanto podemos rechazar la hipótesis nula de que las varianzas son iguales y concluir que son diferentes. Así que en el t-test las consideraremos diferentes.

Recordemos ahora que la hipótesis alternativa que queremos contrastar es $H_1 : \mu_m > \mu_h$. En el factor `cats$Sex`, la `F` (hembra) va antes que la `M` (macho), y, por tanto, si entramos los vectores de pesos mediante `Bwt~Sex,data=cats`, el primer vector corresponderá a las gatas y el segundo a los gatos. Así pues, la hipótesis alternativa que hemos de especificar es que la media del primer vector es inferior a la media del segundo vector: `alternative="less"`.

```
> t.test(Bwt~Sex, data=cats, alternative="less",var.equal=FALSE)

    Welch Two Sample t-test

data:  Bwt by Sex
t = -8.7095, df = 136.838, p-value = 4.416e-15
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.4376663
sample estimates:
mean in group F mean in group M
       2.359574        2.900000
```

Observamos que el p-valor es prácticamente 0. Por lo tanto, podemos concluir que, efectivamente, de media, los gatos macho de más de 2 kg pesan más que los gatos hembra de más de 2 kg.

3.3. Contrastes para proporciones

Cuando tenemos que efectuar un contraste sobre una proporción, podemos usar el test binomial `binom.test`, que en realidad contrasta si, en una secuencia de experimentos de Bernoulli, el número de éxitos sigue una ley binomial con probabilidad de éxito fijada. Su sintaxis es

```
binom.test(x, n, p=..., alternative=..., conf.level=...)
```

donde

- `x` es un número natural, el número de éxitos en la muestra.
- `n` es un número natural, el número total de intentos en la muestra.
- `p` es la probabilidad de éxito que queremos contrastar.
- El significado de `alternative` y `conf.level`, y sus posibles valores, son los usuales.

Ejemplo 3.10. En una serie de 5 lanzamientos de una moneda, no he obtenido ninguna cara. ¿Puedo concluir que la moneda está trucada, en el sentido de que, en promedio, da cara en menos de la mitad de los lanzamientos? Para decidirlo, si llamo p a la probabilidad de obtener cara con esta moneda, voy a realizar el contraste

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p < 0.5 \end{cases}$$

```
> binom.test(0, 5, p=0.5, alternative="less")
```

```
      Exact binomial test

data:  0 and 5
number of successes = 0, number of trials = 5, p-value = 0.03125
alternative hypothesis: true probability of success is less than
    0.5
95 percent confidence interval:
 0.0000000 0.4507197
sample estimates:
probability of success
                     0
```

El p-valor del test es 0.03125 indica que hay evidencia significativa que permite rechazar la hipótesis nula y concluir que la probabilidad de sacar cara és menor de 0.5. El intervalo de confianza que nos da este test es para el valor real de la p. En este caso, nos permite estimar que, a un nivel de confianza del 95 %, la probabilidad de sacar cara con nuestra moneda está entre 0 y 0.4507.

Cuando la muestra es grande, podemos usar también la función `prop.test`, que sirve además para contrastes de dos proporciones independientes. Su estructura es

```
prop.test(x, n, p =..., alternative=..., conf.level=...)
```

donde:

- `x` puede ser dos cosas:
  (a) Un número natural: en este caso, `R` entiende que es el número de éxitos en una muestra.
  (b) Un vector de dos números naturales: en este caso, `R` entiende que es un contraste de dos muestras y que éstos son los números de éxitos en las mismas.
- Cuando trabajamos con una sola muestra, `n` es su tamaño. Cuando estamos trabajando con dos muestras, `n` es el vector de dos entradas de sus longitudes.
- Cuando trabajamos con una sola muestra, `p` es la proporción poblacional que contrastamos. En el caso de un contraste de dos muestras, no hay que especificarlo.
- El significado de `alternative` y `conf.level`, y sus posibles valores, son los usuales.

Veamos algunos ejemplos.

Ejemplo 3.11. De 50 estudiantes de la UIB encuestados al azar, 3 han sido zurdos. Suponiendo que los estudiantes encuestados formasen una muestra aleatoria simple, ¿aporta esta encuesta evidencia de que la proporción de estudiantes zurdos en la UIB sea inferior al 10 %, el porcentaje estimado de zurdos en España? Para decidirlo, y llamando p a la

proporción de estudiantes zurdos en la UIB, vamos a realizar el contraste

$$\begin{cases} H_0 : p = 0.1 \\ H_1 : p < 0.1 \end{cases}$$

Como la muestra es grande ($n = 50$) podemos usar `prop.test`.

```
> prop.test(3, 50, p=0.1, alternative="less")

        1-sample proportions test with continuity correction

data:  3 out of 50, null probability 0.1
X-squared = 0.5, df = 1, p-value = 0.2398
alternative hypothesis: true p is less than 0.1
95 percent confidence interval:
 0.0000000 0.1539523
sample estimates:
   p
0.06
```

El p-valor obtenido en el test es 0.2398, superior a 0.05, y el intervalo de confianza del 95 % para p que hemos obtenido es $[0, 0.154[$, que contiene el valor 0.1. Por lo tanto, no podemos rechazar que un 10 % de los estudiantes de la UIB sean zurdos.

La conclusión usando el test binomial hubiera sido la misma:

```
> binom.test(3, 50, p=0.1, alternative="less")

        Exact binomial test

data:  3 and 50
number of successes = 3, number of trials = 50, p-value = 0.2503
alternative hypothesis: true probability of success is less than
    0.1
95 percent confidence interval:
 0.0000000 0.1478372
sample estimates:
probability of success
                  0.06
```

Ejemplo 3.12. Una empresa que fabrica trampas para cucarachas ha producido una nueva versión de su trampa más popular, y afirma que la nueva trampa mata más cucarachas que la vieja. Hemos llevado a cabo un experimento para comprobarlo. Hemos situado dos trampas en dos habitaciones. En cada habitación hemos soltado 60 cucarachas. La versión vieja de la trampa ha matado 40 y la nueva, 48. ¿Podemos afirmar que la nueva trampa es más efectiva que la vieja?

Digamos p_v y p_n a las proporciones de cucarachas que matan la trampa vieja y la trampa nueva, respectivamente. La hipótesis nula será que $H_0 : p_v = p_n$, y la hipótesis alternativa

$H_1 : p_v < p_n$. Los tamaños de las muestras nos permiten usar `prop.test` para realizar este contraste.

```
> prop.test(c(40,48),c(60,60),alternative="less")

        2-sample test for equality of proportions with
    continuity correction

data:  c(40, 48) out of c(60, 60)
X-squared = 2.0881, df = 1, p-value = 0.07423
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01461667
sample estimates:
   prop 1    prop 2
0.6666667 0.8000000
```

El p-valor es 0.07423, y el intervalo de confianza que nos da el test, $]-1, 0.014617[$, es para la diferencia de proporciones $p_v - p_n$ y contiene el 0, aunque por poco. En resumen, si tomáramos un nivel de significación de 0.05, no encontraríamos evidencia de que la trampa nueva sea mejor que la vieja, pero como no nos han dado un nivel de significación *a priori* y el p-valor ha caído dentro de la *zona gris* $]0.05, 0.1[$, es más seguro tomar el resultado como no concluyente: convendría llevar a cabo un nuevo experimento con más cucarachas.

La función `prop.test` sólo sirve para contrastar las proporciones de dos muestras independientes, y sólo tiene sentido para muestras grandes. Otro test que se puede usar para contrastar las proporciones de dos poblaciones independientes es el *test exacto de Fisher*, que usa una distribución hipergeométrica y que no tiene restricciones de uso.

Supongamos que evaluamos una característica dicotómica (es decir, que sólo puede tomar dos valores y por tanto define distribuciones de Bernoulli) sobre dos poblaciones independientes de individuos y resumimos los resultados en una tabla como ésta:

		Poblaciones	
		1	2
Característica	Sí	a	b
	No	c	d

Llamemos p_1 a la proporción de individuos con la característica bajo estudio dentro de la población 1 y p_2 a la proporción de individuos con la característica bajo estudio dentro de la población 2. Queremos contrastar la hipótesis nula $H_0 : p_1 = p_2$ contra alguna hipótesis alternativa. Por ejemplo, en el ejemplo de las trampas para cucarachas, las poblaciones vendrían definidas por el tipo de trampa, y la característica que tendríamos en cuenta sería si la cucaracha ha muerto o no, lo que nos daría la tabla

	Trampas	
	Viejas	Nuevas
Muertas	40	48
Vivas	20	12

Para realizar este tipo de contrastes, podemos usar el test de Fisher, que se lleva a cabo en `R` con la función `fisher.test`. Su estructura es

```
fisher.test(x, alternative=..., conf.level=...)
```

donde

- `x` es la matriz $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$.
- El significado de `alternative` y `conf.level`, y sus posibles valores, son los usuales.

Así, en el ejemplo de las trampas para cucarachas, entraríamos:

```
> Datos=rbind(c(40,48),c(20,12))
> Datos
     [,1] [,2]
[1,]   40   48
[2,]   20   12
> fisher.test(Datos, alternative="less")

        Fisher's Exact Test for Count Data

data:  Datos
p-value = 0.07392
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.084135
sample estimates:
odds ratio
 0.5029092
```

y obtenemos de nuevo un p-valor cercano a 0.74. El intervalo de confianza que nos da esta función es para el cociente p_v/p_n, y vemos que contiene el 1.

Ejemplo 3.13. Para determinar si el Síndrome de Muerte Súbita del Recién Nacido (*SIDS*, por sus siglas en inglés) tiene alguna componente genética, se estudiaron parejas de gemelos y mellizos en las que se dio algún caso de SIDS. Sean p_1 la proporción de casos con exactamente una muerte de SIDS entre las parejas de gemelos con algún caso por SIDS, y p_2 la proporción de casos con exactamente una muerte por SIDS entre las parejas de mellizos con algún caso de SIDS. La hipótesis de trabajo es que si el SIDS tiene componente genético, será más probable que un gemelo de un muerto por SIDS también lo sufra que si sólo es mellizo, y por lo tanto que en las parejas de gemelos ha de ser más raro que haya

exactamente un caso de SIDS que en las parejas de mellizos. Es decir, que $p_1 < p_2$. Así pues, queremos realizar el contraste

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 < p_2 \end{cases}$$

En un estudio de 1980 se obtuvieron los datos siguientes

		Tipo de gemelos	
		Gemelos	Mellizos
Casos de SIDS	Uno	23	35
	Dos	1	2

Vamos a realizar el contraste. Observad que damos la tabla de manera que p_1 es la proporción de parejas con un solo caso de SIDS entre aquellas de la población 1 (gemelos), y p_2 es la proporción de parejas con un solo caso de SIDS entre aquellas de la población 2 (mellizos). Por tanto hemos de aplicar `fisher.test` a esta matriz y $p_1 < p_2$ corresponderá a `alternative="less"`.

```
> Datos=rbind(c(23,35),c(1,2))
> Datos
     [,1] [,2]
[1,]   23   35
[2,]    1    2
> fisher.test(Datos, alternative="less")

        Fisher's Exact Test for Count Data

data:  Datos
p-value = 0.7841
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
  0.00000 39.73954
sample estimates:
odds ratio
  1.308589
```

El p-valor es 0.7841, muy grande, por lo que no obtenemos evidencia de componente genética en el SIDS.

Las funciones `prop.test` y `fisher.test` no pueden usarse para comparar proporciones de muestras emparejadas, en situaciones como la siguiente. Supongamos que evaluamos dos características dicotómicas sobre un mismo conjunto de n individuos y resumimos los resultados en una tabla como ésta:

		Car. 1	
		Sí	No
Car. 2	Sí	a	b
	No	c	d

donde $a + b + c + d = n$. Ahora vamos a llamar p_1 a la proporción de individuos con la característica 1, y p_2 a la proporción de individuos con la característica 2. Queremos contrastar la hipótesis nula $H_0 : p_1 = p_2$ contra alguna hipótesis alternativa.

En este caso, para realizar el contraste

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$$

cuando n es grande y el número $b + c$ de *casos discordantes* (en los que una característica da Sí y la otra da No) es razonablemente grande, pongamos $\geqslant 20$, podemos usar el *test de McNemar*, que se lleva a cabo en `R` con la instrucción `mcnemar.test`. Su sintaxis básica es

```
mcnemar.test(x)
```

donde `x` es la matriz 2×2

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

que corresponde a la tabla anterior.

Ejemplo 3.14. Para comparar la efectividad de dos tratamientos del asma, se escogieron 200 pacientes con asma severo, se los trató durante un mes con el tratamiento A, se les dejó sin tratamiento durante un mes, y a continuación se los trató durante un mes con el tratamiento B. Se anotó si durante cada periodo de tratamiento cada enfermo visitó o no el servicio de urgencias por dificultades respiratorias. Los resultados del experimento se resumen en la tabla siguiente:

		Trat. A	
		Sí	No
Trat. B	Sí	71	48
	No	30	51

Queremos determinar si hay diferencia en la efectividad de los dos tratamientos.

```
> Datos=matrix(c(71,48,30,51),nrow=2,byrow=TRUE)
> Datos
     [,1] [,2]
[1,]   71   48
[2,]   30   51
> mcnemar.test(Datos)

        McNemar's Chi-squared test with continuity correction

data:  Datos
McNemar's chi-squared = 3.7051, df = 1, p-value = 0.05425
```

El p-valor del test es 0.05425, ligeramente superior a 0.05, por lo tanto no permite concluir que haya evidencia de que la efectividad de los dos tratamientos sea diferente. Sería conveniente llevar a cabo un estudio más amplio.

Otra posibilidad para realizar este contraste, que no requiere de ninguna hipótesis sobre los tamaños de las muestras, es usar de manera adecuada el `binom.test`. Consideremos para ello la tabla siguiente, donde ahora damos las probabilidades poblacionales de las cuatro combinaciones de resultados

		Car. 1	
		Sí	No
Car. 2	Sí	p_{11}	p_{10}
	No	p_{01}	p_{00}

De esta manera $p_1 = p_{11} + p_{01}$ y $p_2 = p_{11} + p_{10}$. Entonces, $p_1 = p_2$ es equivalente a $p_{10} = p_{01}$. Si esta hipótesis nula es cierta, del total de casos discordantes, el número de casos en los que la característica A da Sí y la característica B da No debería seguir una ley binomial con $p = 0.5$. Por lo tanto,

```
binom.test(c, b+c, p=0.5)
```

permite contrastar

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$$

La ventaja de este test es que es exacto, no es necesaria ninguna hipótesis sobre los tamaños de las muestras.

Ejemplo 3.15. Usemos el test binomial para llevar a cabo el contraste del Ejemplo 3.14. Habíamos obtenido $30 + 48 = 78$ casos discordantes, de los que 48 eran casos en los que el tratamiento A había dado Sí y el tratamiento B había dado No.

```
> binom.test(48, 78, p=0.5)

      Exact binomial test

data:  30 and 78
number of successes = 30, number of trials = 78, p-value =
    0.05354
alternative hypothesis: true probability of success is not equal
     to 0.5
95 percent confidence interval:
 0.2766016 0.5016690
sample estimates:
probability of success
             0.3846154
```

Obtenemos de nuevo un p-valor en la zona gris, ligeramente superior a 0.05.

Ejemplo 3.16. Para determinar la efectividad de un test casero de VIH basado en un frotis bucal, se lo comparó con el test de VIH estándar, basado en una analítica de sangre que detecta la presencia del virus. Se tomó una muestra aleatoria de 241 individuos en situación de riesgo, y a todos se les hizo el test de VIH estándar y el nuevo test. Los resultados se resumen en la tabla siguiente

		Test estándar	
		Positivo	Negativo
Test casero	Positivo	72	10
	Negativo	2	157

Para realizar el contraste

$$\begin{cases} H_0 : p_{\text{estándar}} = p_{\text{casero}} \\ H_1 : p_{\text{estándar}} \neq p_{\text{casero}} \end{cases}$$

como el número de casos discordantes es pequeño ($10 + 2 = 12$), realizaremos el test binomial.

```
> binom.test(2, 12, p=0.5)

      Exact binomial test

data:  2 and 12
number of successes = 2, number of trials = 12, p-value =
    0.03857
alternative hypothesis: true probability of success is not equal
     to 0.5
95 percent confidence interval:
 0.02086253 0.48413775
sample estimates:
probability of success
             0.1666667
```

Obtenemos una evidencia significativa de que los tests no son igual de efectivos.

3.4. Cálculo de la potencia de un contraste

La *potencia* de un contraste de hipótesis es la probabilidad de aceptar la hipótesis alternativa si es verdadera. Es decir, si llamamos un *error de tipo II* a no rechazar la hipótesis nula cuando la alternativa es verdadera, la potencia del test es 1 menos la probabilidad de cometer un error de tipo II; usualmente, la probabilidad de cometer un error de tipo II se denota por β, y por lo tanto la potencia es $1 - \beta$. Así pues, cuánto más alta sea la potencia de un contraste, menor será la probabilidad de cometer un error de tipo II y, por lo tanto, más fiable será el contraste.

Independientemente del nivel de significación, la potencia de un contraste se puede incrementar aumentando el tamaño de la muestra, pero esto no siempre será posible en la

práctica por razones logísticas o económicas. El objetivo de un contraste será entonces maximizar la potencia manteniendo un nivel de significación y un tamaño de la muestra lo más pequeños posible. Dicho en otras palabras, se pretende maximizar la probabilidad de encontrar evidencia de la hipótesis alternativa si esta es verdadera y al mismo tiempo minimizar por un lado la probabilidad de encontrar evidencia de la hipótesis alternativa si esta es falsa y por otro los costes de realizar el contraste.

La potencia de un contraste está relacionada con la llamada *magnitud del efecto*. En un contraste, el *efecto* es la diferencia entre el valor estimado del parámetro a partir de la muestra usada y el valor que se da a dicho parámetro como hipótesis nula: por ejemplo, la diferencia entre la media muestral $\overline{x}$ y el valor contrastado μ_0 de la media poblacional. Se rechaza entonces la hipótesis nula si el efecto observado es tan grande que es muy improbable cuando la hipótesis nula es verdadera; pero, en realidad, no se tiene en cuenta si el efecto observado ha sido grande o no, sólo si es significativo. Entonces, sin entrar en detalle, digamos que la *magnitud del efecto* es una medida estadística específica de lo grande que es el efecto observado, y depende de muchos factores: desde la variabilidad de la población de la que extraemos la muestra a las características del protocolo usado en el experimento, o las diferencias entre los protocolos usados para extraer diferentes muestras. Por ejemplo, en un t-test bilateral de dos medias independientes, se define la magnitud del efecto como

$$d = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{(n_1-1)\widetilde{s}_1^2+(n_2-1)\widetilde{s}_2^2}{n_1+n_2}}}$$

donde $n_1, \overline{x}_1, \widetilde{s}_1^2$ son, respectivamente, el tamaño, la media y la varianza muestral de una muestra, y $n_1, \overline{x}_1, \widetilde{s}_2^2$ son los correspondientes valores de la otra muestra; definido de esta manera, d mide cuántas veces es mayor el efecto (en este caso, la diferencia entre medias muestrales) que el error estándard muestral.

En resumen, en un contraste de hipótesis intervienen cuatro cantidades fundamentales:

- El *tamaño* de la muestra, n: el número de observaciones que usamos en el contraste.
- El *nivel de significación*, α: la probabilidad de rechazar la hipótesis nula si es cierta.
- La *potencia*, $1-\beta$: la probabilidad de rechazar la hipótesis nula si la hipótesis alternativa es cierta.
- La *magnitud del efecto*: como hemos dicho, se trata de un valor que cuantifica la diferencia entre el parámetro estimado a partir de la muestra y su valor en la hipótesis nula. Se toma, por convenio, que, en tests sobre medias o proporciones,
  - una magnitud del efecto de valor absoluto alrededor de 0.2 es pequeña
  - una magnitud del efecto de valor absoluto alrededor de 0.5 es mediana
  - una magnitud del efecto de valor absoluto alrededor de 0.8 es grande

Estas cuatro cantidades no son independientes, sino que, a partir de tres cualesquiera de ellas, se puede calcular la cuarta. Esto tiene interés en dos contextos:

- En un contraste utilizando unas observaciones concretas, a partir de α, del tamaño de la(s) muestra(s) y de la magnitud del efecto calculada mediante la fórmula adecuada, se puede determinar la potencia.
- En un contraste para el que el investigador pueda estimar *a priori* la magnitud del efecto (a partir del diseño del experimento y de conocimientos previos sobre las poblaciones objeto de estudio), a partir de ésta y del nivel de significación se puede calcular el número mínimo de observaciones necesario para obtener la potencia deseada.

Con `R`, este tipo de cálculos lo llevan a cabo algunas funciones del paquete `pwr`. Las funciones que por ahora nos interesan son las siguientes:

- `pwr.t.test`, para utilizar en t-tests de una media, de dos medias emparejadas o de dos medias independientes usando muestras del mismo tamaño.
- `pwr.t2n.test`, para utilizar en t-tests de dos medias independientes usando muestras de distinto tamaño.
- `pwr.p.test`, para utilizar en contrastes binomiales de una proporción.
- `pwr.2p.test`, para utilizar en contrastes binomiales de dos proporciones independientes usando que usen muestras del mismo tamaño.
- `pwr.2p2n.test`, para utilizar en contrastes binomiales de dos proporciones independientes que usen muestras de distinto tamaño.

Estas funciones tienen los parámetros básicos siguientes:

- `n`: el tamaño de la muestra (o de las muestras cuando son del mismo tamaño)
- `n1` y `n2`: los tamaños de las dos muestras en `pwr.2p2n.test` y `pwr.t2n.test`
- `d` (en las dos primeras) o `h` (en las tres últimas): la magnitud del efecto
- `sig.level`: el nivel de significación α
- `power`: la potencia $1 - \beta$
- `type` (en la primera): el tipo de contraste; sus posibles valores son `"one.sample"` (para contrastes de una muestra), `"two.sample"` (para contrastes de dos muestras independientes), o `"paired"` (para contrastes de dos muestras emparejadas)
- `alternative`: el tipo de hipótesis alternativa, con sus valores usuales

Si, en una cualquiera de estas funciones, se especifican todos los parámetros `n` (o `n1` y `n2`), `d`, `sig.level` y `power` menos uno, la función da el valor del parámetro que falta.

Veamos algunos ejemplos de uso.

Ejemplo 3.17. Queremos calcular la potencia del contraste llevado a cabo en el Ejemplo 3.2. Se trataba de un contraste bilateral de una media usando un t-test, por lo que utilizaremos la función `pwr.t.test`. Los parámetros que le entraremos son:

- `n`, el tamaño de la muestra; en este ejemplo, $n = 25$.
- `d`, la magnitud del efecto. Para contrastes de una media e hipótesis nula $H_0 : \mu = \mu_0$, la magnitud del efecto se calcula con la fórmula

$$d = \frac{\overline{x} - \mu_0}{\widetilde{s}_x}.$$

  En nuestro ejemplo, $d = \frac{2.8048-2}{0.68064} \approx 1.1824$.
- `sig.level`, el nivel de significación α; en este ejemplo, $\alpha = 0.05$.

Además como es un contraste bilateral de una media, especificaremos `type="one.sample"` y `alternative="two.sided"`.

```
> #Instalamos y cargamos el paquete pwr
> x=c(2.2,2.66,2.74,3.41,2.46,2.96,3.34,2.16,2.46,2.71,2.04,
  3.74,3.24,3.92,2.38,2.82,2.2,2.42,2.82,2.84,4.22,3.64,1.77,
  3.44,1.53)
> d=(mean(x)-2)/sd(x) #Magnitud del efecto
> pwr.t.test(n=25, d=d, sig.level=0.05, type="one.sample",
    alternative="two.sided")

     One-sample t test power calculation

              n = 25
              d = 1.18241
      sig.level = 0.05
          power = 0.9998934
    alternative = two.sided
```

La potencia del contraste llevado a cabo es 0.99998934, muy alta. Si sólo hubiéramos querido obtener el valor de esta potencia, hubiera bastado añadir el sufijo `$power`.

```
> pwr.t.test(n=25, d=d, sig.level=0.05, type="one.sample",
    alternative="two.sided")$power
[1] 0.9998934
```

Ejemplo 3.18. Supongamos que, en el contraste anterior, quisiéramos calcular el tamaño mínimo de una muestra para tener un nivel de significación del 5 % y una potencia del 95 %, suponiendo que sabemos que la magnitud del efecto va a ser grande. Si no recordamos los valores pequeño, mediano y grande de la magnitud del efecto, las podemos pedir a `R` con la función

```
cohen.ES(test=..., size=...)$effect.size
```

donde el parámetro `test` indica el tipo de contraste (por ahora, `"p"` si es de proporciones y `"t"` si es de medias) y el parámetro `size` indica el «tamaño» de la magnitud del efecto (sus posibles valores son `"small"`, `"medium"` y `"large"`).

Por lo tanto, para calcular este tamaño entraríamos:

```
> pwr.t.test(d=cohen.ES(test="t", size="large")$effect.size,
   sig.level=0.05, power=0.95, alternative="two.sided",
   type="one.sample")

One-sample t test power calculation

              n = 22.32453
              d = 0.8
      sig.level = 0.05
          power = 0.95
    alternative = two.sided
```

Hubieran sido suficientes 23 observaciones.

Si sólo hubiéramos querido obtener el valor de n, hubiéramos podido usar el sufijo `$n`. Por ejemplo, ¿cuántas observaciones necesitaríamos si estimaramos que la magnitud del efecto iba a ser pequeña?

```
> pwr.t.test(d=cohen.ES(test="t", size="small")$effect.size,
   sig.level=0.05, power=0.95, alternative="two.sided",
   type="one.sample")$n
[1] 326.7952
```

En este caso necesitaríamos como mínimo 327 observaciones.

Ejemplo 3.19. Vamos a calcular la potencia del contraste

$$\begin{cases} H_0 : p_v = p_n \\ H_1 : p_v < p_n \end{cases}$$

del Ejemplo 3.12. En este caso, usamos la función `pwr.2p.test` y le entramos los parámetros siguientes:

- `n`, el tamaño de la muestra; en este ejemplo, $n = 60$.
- `h`, la magnitud del efecto. Para calcularla,[4] usamos la función `ES.h` del mismo paquete

[4] Por si a alguien le interesa, la fórmula para esta magnitud del efecto es

$$h = 2\left(\arcsin\left(\sqrt{\widehat{p}_1}\right) - \arcsin\left(\sqrt{\widehat{p}_2}\right)\right),$$

siendo $\widehat{p}_1$ y $\widehat{p}_2$ las proporciones muestrales de éxitos de las dos muestras; véase el capítulo 6 de J. Cohen, *Statistical power analysis for the behavioral sciences* (2a edición). Routledge (1988).

`pwr` y que se aplica a las proporciones muestrales de éxitos de las dos muestras: en este ejemplo, $\widehat{p}_v = 0.67$ y $\widehat{p}_n = 0.8$.

- `sig.level`, el nivel de significación, 0.05.

```
> h=ES.h(0.67,0.8) #Magnitud del efecto
> h
[1] -0.2965842
> pwr.2p.test(h=h, n=60, sig.level=0.05,
   alternative="less")$power
[1] 0.4918641
```

Hemos obtenido una potencia de, aproximadamente, un 49 %.

Para saber qué tamaño de muestra tendríamos que tomar para tener un nivel de significación del 95 % y una potencia del 90 % suponiendo que la magnitud iba a ser media, entraríamos:

```
> pwr.2p.test(h=-cohen.ES(test="p", size="medium")$effect.size,
   sig.level=0.05, power=0.9, alternative="less")$n
[1] 68.51078
```

Tendríamos que usar dos muestras de 69 cucarachas cada una. Observad que, como la hipótesis alternativa es $H_1 : p_v < p_n$, esperamos que el efecto, y por lo tanto su magnitud, sea negativo, y por esto lo hemos entrado negativo en `pwr.2p.test`.

Ejemplo 3.20. Calculemos la potencia del contraste

$$\begin{cases} H_0 : \mu_n = \mu_f \\ H_1 : \mu_n > \mu_f \end{cases}$$

llevado a cabo en el Ejemplo 3.6. Como es un contraste de dos medias independientes y los tamaños de las muestras son diferentes, usaremos la función `pwr.t2n.test` con parámetros:

- `n1`: tamaño de la primera muestra; en este ejemplo, $n_1 = 115$.
- `n2`: tamaño de la segunda muestra; en este ejemplo, $n_2 = 74$.
- `sig.level`, el nivel de significación, 0.05.
- `d`: magnitud del efecto. En este tipo de contrastes, la magnitud del efecto d se calcula con la fórmula

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)\widetilde{s}_1^2+(n_2-1)\widetilde{s}_2^2}{n_1+n_2}}}$$

```
> #Cargamos el paquete MASS, si no lo está
> bwt_nofum=birthwt[birthwt$smoke==0,]$bwt #Pesos de hijos de
    madres no fumadoras
> bwt_fum=birthwt[birthwt$smoke==1,]$bwt #Pesos de hijos de
    madres fumadoras
> n1=length(bwt_nofum)
> x1=mean(bwt_nofum)
> var1=var(bwt_nofum)
> n2=length(bwt_fum)
> x2=mean(bwt_fum)
> var2=var(bwt_fum)
> d=(x1-x2)/sqrt(((n1-1)*var1+(n2-1)*var2)/(n1+n2))
> d
[1] 0.3974624
> pwr.t2n.test(n1=n1, n2=n2, d=d, sig.level=0.05,
   alternative="greater")$power
[1] 0.8443554
```

Obtenemos una potencia de aproximadamente un 84.4 %.

Ejercicios

(1) Para satisfacer las necesidades respiratorias de los peces de agua caliente, el contenido de oxígeno disuelto debe presentar un promedio de 6.5 partes por millón (ppm), con una desviación típica no mayor de 1.2 ppm. Cuando la temperatura del agua crece, el oxígeno disuelto decrece, y esto causa la asfixia del pez.

Se realizó un estudio sobre los efectos del calor en verano en el contenido de oxígeno disuelto en un gran lago. Después de un período particularmente caluroso, se tomaron muestras de agua en 35 lugares aleatoriamente seleccionados en el lago, y se determinó el contenido de oxígeno disuelto. Los resultados (en ppm) fueron los siguientes:

```
O2=c(9.1,6.8,7.0,7.5,8.7,3.2,5.4,8.1,4.4,5.1,6.2,6.9,6.9,4.3,
    8.0,5.3,6.2,6.4,7.8,5.8,6.9,7.7,5.2,5.8,6.3,5.9,8.5,7.5,8.9,
    5.6,6.6,5.3,5.7,6.9,6.6)
```

Suponemos que estos contenidos de oxígeno siguen una distribución normal.

(a) ¿Hay evidencia de que el contenido medio de oxígeno en el lago sea inferior al nivel aceptable de 6.5 ppm?

(b) ¿Hay evidencia de que la desviación típica del contenido de oxígeno en el lago sea superior a 1.2?

(2) El *dataframe* `sleep` contiene información sobre el número de horas de sueño que añadieron dos tipos diferentes de somníferos a 10 pacientes.

```
> str(sleep)
'fecha.frame': 20 obs. of 2 variables:
$ extra: num 0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2 ...
$ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

En concreto, a 10 pacientes se les midió en una ocasión el tiempo que durmieron una noche que no tomaron ningún somnífero, en otra ocasión el tiempo que durmieron una noche que tomaron el primer somnífero, y en una tercera ocasión el tiempo que durmieron una noche que tomaron el segundo somnífero. Finalmente, para cada somnífero se anotó la diferencia "horas de sueño el día que el paciente tomó este somnífero−horas de sueño el día que el paciente no tomó ningún somnífero". La variable `extra` contiene esta diferencia, y la variable `group` es un factor que contiene el tipo de somnífero.

¿Hay evidencia de que el efecto de los dos somníferos sea diferente? En caso afirmativo, ¿cuál parece que es el más efectivo? ¿Cuál es la potencia del contraste realizado?[5]

(Supondremos que las diferencias recogidas en la variable `extra` siguen leyes normales; en caso contrario, como la muestra es pequeña, el resultado del contraste no sería fiable.)

(3) Los angiogramas son la técnica estándar para detectar un ictus, pero tienen un ligero riesgo de mortalidad (inferior al 1 %). Algunos investigadores han propuesto usar una prueba PET para diagnosticar el ictus de manera no invasiva. Sobre 64 pacientes ingresados en urgencias con síntomas de ictus se usaron ambas técnicas de diagnóstico. Los resultados obtenidos se resumen en la tabla siguiente:

		Angiograma	
		Positivo	Negativo
PET	Positivo	32	8
	Negativo	3	21

Contrastad si ambas técnicas de diagnóstico tienen la misma probabilidad de detectar un ictus.

Modelo de test

(1) Tenemos una muestra formada por los números $2, 5, 3, 5, 6, 6, 7, 2$ de una población que suponemos normal con $\sigma = 2$. Usando la función `z.test`, calculad el p-valor (redondeado a 3 cifras decimales, sin ceros innecesarios a la derecha) del contraste $H_0 : \mu = 4$ contra $H_1 : \mu \neq 4$ y decid (contestando SI, sin acento, o NO) si podemos rechazar la hipótesis nula en favor de la alternativa a un nivel de significación de 0.04. Tenéis que dar las dos respuestas en este orden, separadas por un único espacio en blanco.

[5] Hemos dado la fórmula para calcular la magnitud del efecto en contrastes bilaterales de dos medias independientes en la página 56; para contrastes bilaterales de dos medias emparejadas, la fórmula correspondiente es $d = |\overline{D}|/\widetilde{s}_D$, donde D es el vector de diferencias.

(2) Tenemos una muestra de una población normal formada por los números $2, 5, 3, 5, 6, 6, 7, 2$. Usando la función `t.test`, calculad el p-valor (redondeado a 3 cifras decimales, sin ceros innecesarios a la derecha) del contraste $H_0 : \mu = 4$ contra $H_1 : \mu \neq 4$ y decid (contestando SI, sin acento, o NO) si podemos rechazar la hipótesis nula en favor de la alternativa a un nivel de significación de 0.05. Tenéis que dar las dos respuestas en este orden, separadas por un único espacio en blanco.

(3) Tenemos dos muestras de poblaciones normales, x_1: $2, 5, 3, 5, 6, 6, 7, 2$ y x_2: $3, 2, 5, 4, 2, 2, 4, 5, 1, 6, 2$. Usando la función `t.test`, calculad el p-valor (redondeado a 3 cifras decimales, sin ceros innecesarios a la derecha) del contraste $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$ suponiendo que las varianzas son diferentes y decid (contestando SI, sin acento, o NO) si podemos rechazar la hipótesis nula en favor de la alternativa a un nivel de significación de 0.1. Tenéis que dar las dos respuestas en este orden, separadas por un único espacio en blanco.

(4) Tenemos dos muestras de poblaciones normales, x_1: $2, 5, 3, 5, 6, 6, 7, 2$ y x_2: $3, 2, 10, 9, 2, 2, 4, 5, 1, 10, 2$. Usando la función `var.test`, calculad los extremos inferior y superior de un intervalo de confianza para σ_1^2/σ_2^2 (redondeados a 3 cifras decimales, sin ceros innecesarios a la derecha) y decid (contestando SI, sin acento, o NO) si en el contraste $H_0 : \sigma_1 = \sigma_2$ contra $H_1 : \sigma_1 \neq \sigma_2$ podemos rechazar la hipótesis nula en favor de la alternativa a un nivel de significación de 0.05. Tenéis que dar las tres respuestas en este orden, separadas por un único espacio en blanco.

(5) Hemos realizado un contraste bilateral de dos medias independientes usando un t-test, con un nivel de significación de 0.1. Los tamaños de las muestras han sido 35 y 40, respectivamente, y la magnitud del efecto ha sido 0.38. ¿Cuál es la potencia del contraste realizado? Dad su valor redondeado a 3 cifras decimales, sin ceros innecesarios a la derecha.

Respuestas

(1) `0.48 NO`

(2) `0.487 NO`

(3) `0.083 SI`

(4) `0.078 1.465 NO`

(5) `0.493`