PRACTICA 7: contraste de hipótesis paramétricos. Soluciones.

Grado en Ciencias Ambientales, UAH, 2021/22.

Objetivos

- Determinar las condiciones de aplicabilidad de los contrastes de hipótesis paramétricos y el tipo de contraste a utilizar en cada situación.
- Realizar contrastes de hipótesis paramétricos para la media, la varianza y la proporción con una y dos poblaciones.
- Complementar la información que proporciona el p-valor con la del intervalo de confianza.
- Interpretar los resultados obtenidos.

Los datos que aparecen en el fichero `p1-robles.csv` se refieren a un estudio realizado sobre un robledal cercano a una planta industrial, parte de los cuales se visualizan en la tabla. Se han seleccionado robles de dos variedades (A y B) y ubicados en cuatro zonas distintas. Además la mitad de ellos han sido sometidos a cierto tratamiento (codificados con 1), los no tratados (codificados con 0). Sobre cada árbol se han medido las concentraciones (mg/kg) de ocho elementos químicos en sus hojas: hierro, manganeso y zinc (metales pesados); calcio y magnesio (metales alcalinot?rreos); potasio (metal alcalino); y fósforo y nitrógeno (no metales).

```
head(datos, 4)
```

```
  Num Hierro Manganeso   Zinc Calcio Magnesio Potasio Fosforo Nitrogeno Zona
1   1  0.058    0.0303 0.0089  2.365    0.400   2.632   0.145     2.776    1
2   2  0.060    0.0294 0.0109  2.745    0.432   2.495   0.161     2.918    1
3   3  0.058    0.0289 0.0090  2.513    0.349   2.396   0.169     4.826    1
4   4  0.059    0.0275 0.0090  2.361    0.349   1.979   0.155     4.893    1
  Variedad Tratamiento
1        A           0
2        A           0
3        A           0
4        A           0
```

Responde de forma concisa y razonadamente a las siguientes preguntas:

1. **Contrasta la hipótesis de que la media estimada para la población para la variable Magnesio es distinta de 0.41, al nivel de significación del 5%. Usa R para calcular el p-valor y el intervalo de confianza. Calcula, además, la región de rechazo. Comenta los resultados**.

   La media de cualquier muestra nunca valdrá exáctamente 0.41, de modo que lo que puedes medir es si la media de tu muestra es demasiado diferente de 0.41, por lo que se trata de un contraste bilateral sobre la media. En concreto, contrastar

   $$H_0 : \mu_{Mg} = 0.41 \qquad H_1 : \mu_{Mg} \neq 0.41$$

   Como la población es grande ($n > 30$) se usa una normal, en concreto, la plantilla `media_1pob_T_o_Z_enBruto.R`.

   ```
   library(MASS)
   library(TeachingDemos)

   # Ajusta los parametros de read.table.
   datos = read.table(file = "p1-robles.csv",
                      sep = ";", header = TRUE, dec = ".")
   muestra = datos$Magnesio
   ```

```r
# ajustar para contrastar hipotesis
mu0 = 0.41

# Opciones para alternativa: greater / less / two.sided
# Elige el contraste adecuado y descomenta la linea correspondiente
# (CHconT = t.test(muestra, mu = mu0, alternative = "two.sided", conf.level = ))
(CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                 alternative = "two.sided", conf.level = 0.95))
```

```
One Sample z-test

data:  muestra
z = -1.0402, n = 38.000000, Std. Dev. = 0.075792, Std. Dev. of the
sample mean = 0.012295, p-value = 0.2982
alternative hypothesis: true mean is not equal to 0.41
95 percent confidence interval:
 0.3731125 0.4213086
sample estimates:
mean of muestra
      0.3972105
```

Por otro lado, se pide la región de rechazo. Por la teoría sabemos que, en caso de ser cierta H_0, se tiene que (busca la desviación típica muestral y el tamaño de la muestra en la salida de R):

$$\bar{X}_{Mg} \sim N(\mu_0,\, s/\sqrt{n}) = N(0.41,\, 0.076/\sqrt{38})$$

por tanto, los extremos de la regic'on de rechazo los marcan los valores de $\bar{X}_{Mg}$ que dejan por debajo y por encima de sí una probabilidad de 0.025 ($\alpha/2$), es decir

```r
qnorm(c(0.025, 0.975), mean = 0.41, sd = 0.076/sqrt(38))
```

```
[1] 0.3858359 0.4341641
```

Un enfoque alternativo (pero totalmente equivalente) es usar la N(0,1) estandarizando la media muestral. La expresión, en este caso, es

$$\mu_{Mg} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

es decir, la región de no rechazo está entre

```r
(a  = mu0 - qnorm(0.025, lower.tail = F)*sd(muestra)/sqrt(length(muestra)))
```

```
[1] 0.3859019
```

y

```r
(b = mu0 + qnorm(0.025, lower.tail = F)*sd(muestra)/sqrt(length(muestra)))
```

```
[1] 0.4340981
```

Observa que la semi-amplitud de la región de no rechazo es $z_{\alpha/2}\frac{s}{\sqrt{n}}$, que es la misma que la del intervalo de confianza. La diferencia es que el intervalo de confianza está centrado en la media muestral y la región de no rechazo en la media poblacional. Así, es equivalente decir que

(a) No hay evidencia muestral para rechazar H_0 porque 0.41 (la media hipotetizada) está en el intervalo de confianza (0.3731, 0.4213), que es donde esperamos que esté con probabilidad 0.95.

(b) No hay evidencia muestral para rechazar H_0 porque la media observada 0.3972105 (la media muestral) está en la región de no rechazo (0.3859, 0.4341), que contiene el 95% de las medias muestrales más probables si no se rechaza H_0.

(c) El p-valor vale $0.2982451 > \alpha = 0.05$, por lo que no se rechaza H_0.

Muy importante: H_0 **nunca** se acepta, en todo caso, no se rechaza.

2. **¿Puede admitirse, al nivel de significación del 5%, que la contaminación media por Potasio está por encima de 1.5? Calcula, además, la región de rechazo. Comenta los resultados.**

Recuerda que H_0 nunca se acepta; como mucho, diremos que no hay evidencias para rechazarla (que no es lo mismo). Por eso, para admitir que la concentración media es mayor que 1.5, hay que suponer lo contrario (H_0, que es menor o igual que 1.5) y, si la muestra contradice dicha hipótesis, al rechazarla nos quedamos con la alternativa. Es decir, podremos admitir que la concentración media de Potasio es superior a 1.5.

Se trata de un contraste de hipótesis unilateral sobre la media:

$$H_0 : \mu_K \leq 1.5 \qquad H_1 : \mu_K > 1.5$$

Como la población es grande ($n > 30$) se usa una normal, en concreto, la plantilla `media_1pob_T_o_Z_enBruto.R`.

```
library(MASS)
library(TeachingDemos)

# Ajusta los parametros de read.table.
datos = read.table(file = "p1-robles.csv",
                   sep = ";", header = TRUE, dec = ".")
muestra = datos$Potasio

# ajustar para contrastar hipotesis
mu0 = 1.5

# Opciones para alternativa: greater / less / two.sided
# Elige el contraste adecuado y descomenta la linea correspondiente
# (CHconT = t.test(muestra, mu = mu0, alternative = "two.sided", conf.level = ))
(CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                 alternative = "greater",
                 conf.level = 0.95))


One Sample z-test

data:  muestra
z = 5.1474, n = 38.000000, Std. Dev. = 0.514993, Std. Dev. of the
sample mean = 0.083543, p-value = 1.321e-07
alternative hypothesis: true mean is greater than 1.5
95 percent confidence interval:
 1.79261     Inf
sample estimates:
mean of muestra
       1.930026
```

Observa que, en este caso, el intervalo de confianza

$$(1.793, \infty)$$

tiene un aspecto raruno, porque su extremo superior es $+\infty$. Este tipo de intervalos no suelen explicarse a nivel de primer curso de grado, pero se usan en *la realidad*. Esencialmente, nos dice que con una probablidad del 95%, a la vista de la muestra, la concentración de Potasio será de, al menos, 1.793mg/kg.

En este caso, de nuevo, la teoría dice que, en caso de ser cierta H_0, se tiene que (busca la desviación típica muestral y el tamaño de la muestra en la salida de R):

$$\bar{X}_K \sim N(1.5,\, 0.51/\sqrt{38})$$

Ahora la región de rechazo está formada por el 5% de los valores de $\bar{X}_K$ más contradictorios con H_0, es decir, el 5% de los que más se alejan de 1.5 por su derecha. Por tanto, la región de rechazo la forman los valores de la media muestral mayores que

```
(h  = qnorm(0.05, 1.5, sd(muestra)/sqrt(length(muestra)), lower.tail = F))

[1] 1.637416
```

Observa que las regiones de rechazo/no rechazo, el intervalo de confianza y el p-valor son coherentes entre sí:

(a) La muestra sugiere rechazar H_0 porque 1.5 (la media hipotetizada) no está en el intervalo de confianza $(1.793, \infty)$, que es donde esperamos que esté con probabilidad 0.95.

(b) Hay evidencia muestral para rechazar H_0 porque la media observada 1.93 (la media muestral) está en la regi'on de rechazo (1.637, $+\infty$), que contiene el 5% de las medias muestrales más contradictorias con H_0.

(c) El p-valor vale $1.3208149 \times 10^{-7} < \alpha = 0.05$, por lo que se rechaza H_0.

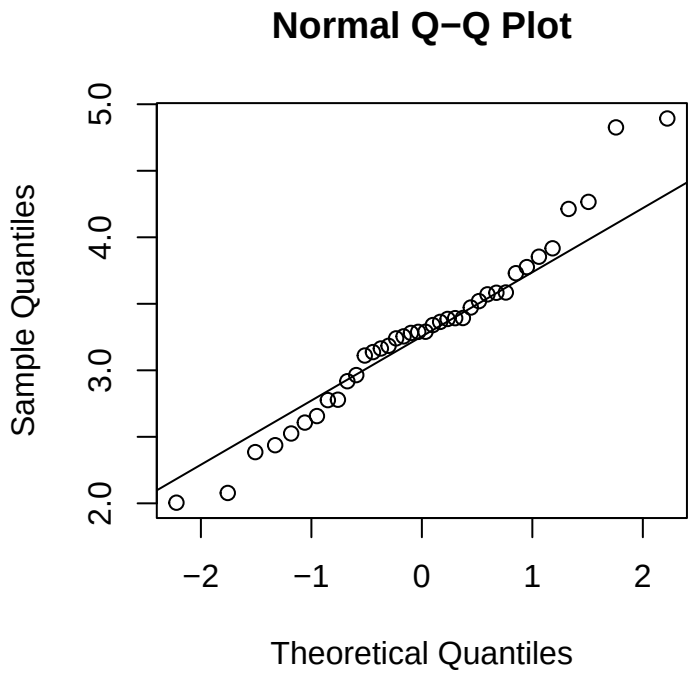
3. **¿Puede admitirse, al nivel de significación del 1%, que la varianza del Nitrógeno es inferior a 0.3?**

Se pide un contraste de hipótesis unilateral sobre la varianza:

$$H_0 : \sigma_N^2 \geq 0.3 \qquad H_1 : \sigma_N^2 < 0.3$$

Como se trata de la varianza, hay que comprobar la normalidad de los datos:

```
qqnorm(datos$Nitrogeno)
qqline(datos$Nitrogeno)
```



excepto por unos pocos datos en los extremos los puntos están bastante alineados, por lo que no rechazamos la normalidad de los datos.

También se puede usar el contraste de Shapiro (H_0 es que los datos provienen de una distribución normal)

```
shapiro.test(datos$Nitrogeno)


Shapiro-Wilk normality test

data:  datos$Nitrogeno
W = 0.96759, p-value = 0.3311
```

que arroja un p-valor compatible con asumir la normalidad de los datos.

Ahora se usa la plantilla `varianza_1pob_enBruto.R`

```
library(MASS)
library(TeachingDemos)

datos = read.table(file = "p1-robles.csv",
                   sep = ";", header = TRUE, dec = ".")
muestra = datos$Nitrogeno

# ajustar para contrastar hipotesis
sigma0 = sqrt(0.3)
# Opciones para alternativa: greater / less / two.sided
(CH = sigma.test(muestra, sigma = sigma0,
                 alternative = "less", conf.level = 0.99))


One sample Chi-squared test for variance
```

```
data:  muestra
X-squared = 50.756, df = 37, p-value = 0.9346
alternative hypothesis: true variance is less than 0.3
99 percent confidence interval:
 0.000000 0.762856
sample estimates:
var of muestra
     0.4115347
```

por lo que no se rechaza la hipótesis nula; observa que la varianza muestral es mayor que 0.3.

Como 0.3 está en el intervalo de confianza para la varianza (y no está cerca de sus extremos) no hay duda: la muestra no apoya el rechazar H_0. Para que el intervalo de confianza sugiriera rechazar H_0 tendríamos que haber obtenido algo como $(0, 0.2)$ (y un p-valor pequeño, claro). Si la varianza estuviera con probabilidad 0.99 entre $(0, 0.2)$, rechazaríamos la posibilidad de que valiera 0.3 o más.

4. **En relación a la variable Nitrógeno, ¿tienes motivos para dudar de que la mitad de los robles presenta una concentración superior a 3, y la otra mitad inferior a 3, con un nivel de significación del 10%?**

Si llamamos p a la proporción de robles con una concentración de Nitrógeno superior a 3, se trata de contrastar la hipótesis:

$$H_0 : p = 1/2 \qquad H_1 : p \neq 1/2$$

Se usa la plantilla `proporcion_1pob.R` en la que

```
datos = read.table(file = "p1-robles.csv",
                   sep = ";", header = TRUE, dec = ".")
muestra = datos$Nitrogeno
n = length(muestra)  #num. elementos muestra
k = sum(muestra>3)  #num. exitos muestra
pMuestral = k / n

# ajustar para contrastar hipotesis
p0 = 0.5

# Opciones para alternativa: greater / less / two.sided
prop.test(k, n, conf.level =0.9 , p = p0, correct=FALSE, alternative = "two.sided")


1-sample proportions test without continuity correction

data:  k out of n, null probability p0
X-squared = 6.7368, df = 1, p-value = 0.009444
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.5787774 0.8142895
sample estimates:
        p
0.7105263
```

el p valor es menor que el nivel de significiación $\alpha = 0.1$, por lo que se rechaza H_0.

Ahora el intervalo de confianza no contiene el valor de la proporción muestra estipulado en H_0, lo que corrobora la decisión tomada. Además, $1/2$ está alejado el extremo inferior del intervalo, por lo que estamos seguros de la decisión tomada en el contrate.

5. **Sobre la variable Calcio, supón que dispones de 100 datos que arrojan una media muestral $\bar{X} = 2.97$ y una desviación típica muestal de $s = 0.1$. ¿Se puede afirmar que el valor medio de la concentración de Calcio es distinta de 3, con un nivel de significación del 5%? ¿Te parece relevante la diferencia?**

   Se trata de contrastar la hipótesis:

$$H_0 : \mu_{Ca} = 3 \qquad H_1 : \mu_{Ca} \neq 3$$

   Se usa la plantilla `media_1pob_T_o_Z_estadisticos.R` en la que

```
datos = read.table(file = "p1-robles.csv",
                   sep = ";", header = TRUE, dec = ".")
muestra = datos$Calcio
library(MASS)
library(TeachingDemos)

n = 100
xbar = 2.97
# Asegurate de usar s y no s^2
s = 0.1

muestra = mvrnorm(n = n, mu = xbar, Sigma = s^2, empirical = TRUE)

# ajustar para contrastar hipotesis
mu0 = 3

# Opciones para alternativa: greater / less / two.sided
# Elige el contraste adecuado y descomenta la linea correspondiente
# (CHconT = t.test(muestra, mu = mu0, alternative = "two.sided", conf.level = ))
(CHconZ = z.test(muestra, mu = mu0, stdev = s,
                 alternative = "two.sided", conf.level = 0.95))



One Sample z-test

data:  muestra
z = -3, n = 1e+02, Std. Dev. = 1e-01, Std. Dev. of the sample mean =
1e-02, p-value = 0.0027
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.9504 2.9896
sample estimates:
mean of muestra
           2.97
```

   el p valor es menor que el nivel de significiación $\alpha = 0.05$, por lo que desde un punto de vista estadístico, se rechaza H_0. Sin embargo, la distancia entre el extremo superior del intervalo de confianza y el valor de la media poblacional establecido en H_0 es muy pequeña

```
[1] 0.01040036
```

   por lo que cabría preguntarse si esa diferencia es científicamente significativa (y haría falta recurrir alguien con conocimento experto en este problema concreto).

6. **Para la variable Manganeso, se afirma que la concentración media es menor que** 0.01**. Se toma la decisión de rechazar esta hipótsis si se observa una concentración muestral**

mayor o igual que 0.013 **¿Qué nivel de significación está asociado con esa regla de decisión?**

El contraste de hipótesis subyacente es $H_0 : \mu_{Mg} \leq 0.01$ frente a $H_1 : \mu_{Mg} > 0.01$, y el valor 0.013 marca el inicio de la región de rechazo. Así, la probabilidad

```
pnorm(0.013, mean = 0.01,
      sd = sd(datos$Manganeso)/sqrt(length(datos$Manganeso)),
      lower.tail = FALSE)

[1] 0.0553651
```

nos da el nivel de significación pedido.

7. **En relación a la dispersión de la concentración de Potasio, ¿es distinta para cada variadad de roble? Usa un nivel de significación del 10%.**
La pregunta es sobre el cociente de varianzas, en concreto, como las variables de roble son

```
unique(datos$Variedad)

[1] "A" "B"
```

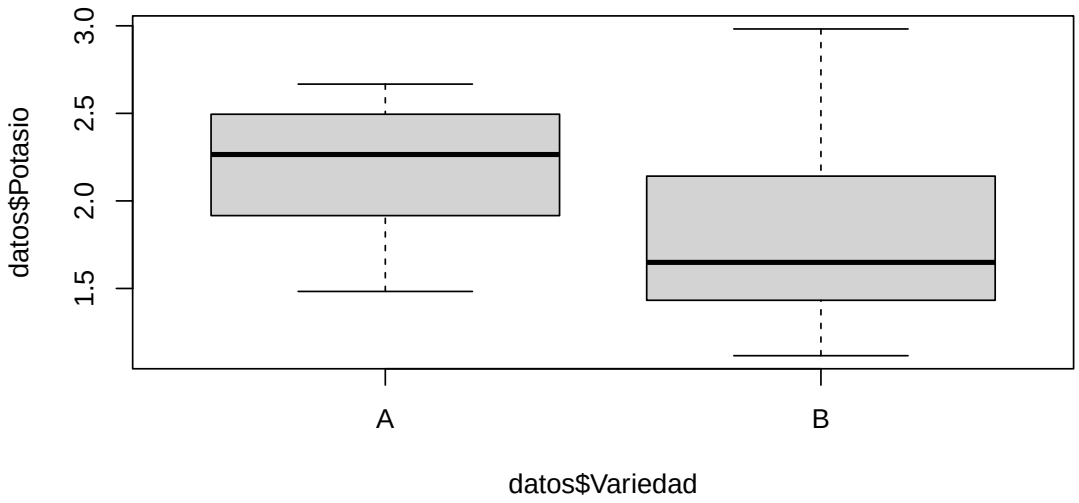
contrastar, para el Potasio,

$$H_0 : \sigma_A^2 = \sigma_B^2 \qquad H_1 : \sigma_A^2 \neq \sigma_B^2$$

para lo que se necesita la plantilla de cociente de varianzas, datos en bruto, dos poblaciones.

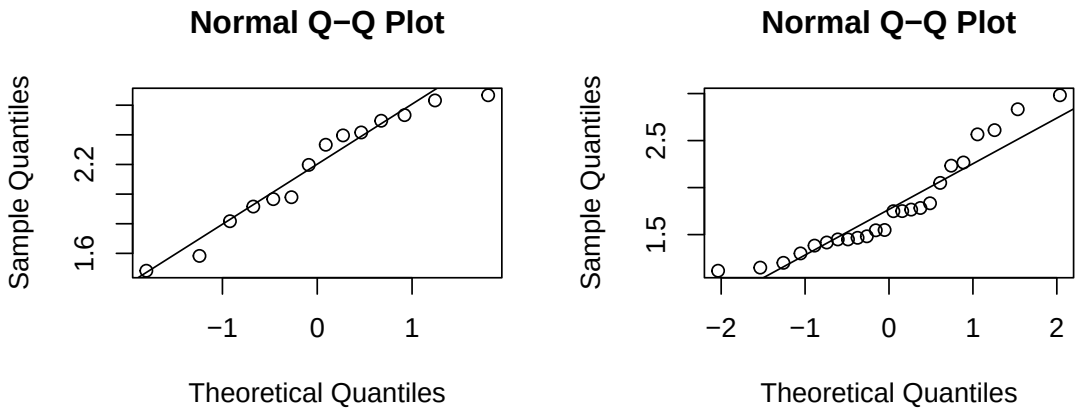
Una primera aproximación consiste en visualizar los boxplots para comparar, *grosso modo*, la dispersión de los datos

```
boxplot(datos$Potasio ~ datos$Variedad)
```



El boxplot parece indicar cierta diferencia, pero no si esa diferencia es estadísticamente significativa. Para decidir hay que hacer el contraste de varianzas y lo primero es evaluar la normalidad de los datos

```r
## LECTURA DESDE FICHERO
tabla = read.table(file = "p1-robles.csv", sep = ";", header = T)
muestra1 = tabla$Potasio[tabla$Variedad == "A"]
muestra2 = tabla$Potasio[tabla$Variedad == "B"]
par(mfrow = c(1,2))
qqnorm(muestra1)
qqline(muestra1)
qqnorm(muestra2)
qqline(muestra2)
```




```r
par(mfrow = c(1,1))
```

Ambas nubes de puntos están bastante alineadas, por no que no dudamos de que los datos provienen de una población normal. De nuevo el contraste de Shapiro (H_0 es que los datos provienen de una distribución normal)

```r
shapiro.test(muestra1)
```

```
Shapiro-Wilk normality test

data:  muestra1
W = 0.93455, p-value = 0.353
```

```r
shapiro.test(muestra2)
```

```
Shapiro-Wilk normality test

data:  muestra2
W = 0.90167, p-value = 0.02335
```

el primero arroja un p-valor compatible con asumir la normalidad de la muestra, no así el segundo (al nivel de confianza que plantea el ejercicio). Seguimos adelante (porque estamos en clase ;))

Así, contrastamos la diferencia de varianzas

```r
var.test(x = muestra1, y = muestra2,
         alternative = "two.sided",
         conf.level = 0.9)
```

```
F test to compare two variances

data:  muestra1 and muestra2
F = 0.51551, num df = 13, denom df = 23, p-value = 0.2164
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
 0.2369972 1.2520124
sample estimates:
ratio of variances
         0.5155068
```

y el p-valor indica que la muestra no contradice H_0, por lo que no se rechaza. Observa que el intervalo de confianza contiene al 1, lo que apoya esta decisión.

8. **¿Podemos decir que la contaminación por Potasio es distinta en ambas variedades? (indica la respuesta para diferentes niveles de significación: 0.1, 0.05, 0.01).**
El contraste en este caso es

$$\mu_A = \mu_B \qquad \mu_A \neq \mu_B$$

En este caso hay que comparar las medias. Las muestras tienen tamaños

```
# para escribir menos crear esta variables
length(muestra1)

[1] 14
```

y

```
# para escribir menos crear esta variables
length(muestra2)

[1] 24
```

es decir, son muestras pequeñas. Del ejercicio anterior sabes que no hay motivo para dudar de que las muestras provengan de poblaciones normales, de modo que podemos usar la t de Student para comparar las medias. Por el ejercicio anterior sabes también que se puede considerar que las varianzas son iguales. Con la plantilla correspondiente se obtiene, para $\alpha = 0.1$

```
t.test(x = muestra1, y = muestra2, var.equal = TRUE, mu = 0,
       alternative = "two.sided", conf.level = .9)


Two Sample t-test

data:  muestra1 and muestra2
t = 2.3464, df = 36, p-value = 0.02458
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 0.1076207 0.6597603
sample estimates:
mean of x mean of y
 2.172357  1.788667
```

para $\alpha = 0.05$

```r
t.test(x = muestra1, y = muestra2, var.equal = TRUE, mu = 0,
       alternative = "two.sided", conf.level = .95)
```

```
Two Sample t-test

data:  muestra1 and muestra2
t = 2.3464, df = 36, p-value = 0.02458
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0520573 0.7153237
sample estimates:
mean of x mean of y
 2.172357  1.788667
```

para $\alpha = 0.01$

```r
t.test(x = muestra1, y = muestra2, var.equal = TRUE, mu = 0,
       alternative = "two.sided", conf.level = .99)
```

```
Two Sample t-test

data:  muestra1 and muestra2
t = 2.3464, df = 36, p-value = 0.02458
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.06099864  0.82837959
sample estimates:
mean of x mean of y
 2.172357  1.788667
```

es decir, para $\alpha = 0.1, 0.05$ se rechaza H_0 mientras que para $\alpha = 0.01$ no hay evidencia muestral para rechazarla. Observa, y es importante, que uno de los extremos de los intervalos de confianza el muy próximo a 0, y que en el caso de para $\alpha = 0.05$ y para $\alpha = 0.01$ la diferencia es nimia, probablemente no sea significativa (relevante) desde el punto de vista científico.

Los datos que contiene el fichero `Practica06-ciudades.csv` (el que usamos en la práctica anterior) se refieren (a excepción de la última columna) a 41 ciudades de Estados Unidos y fueron extraídos de distintas revistas publicadas por el gobierno de este país durante los años 1969 y 1970. Las variables que nos interesan son:

- CIUDAD (nombre de la ciudad).
- SO2_*tr* (la variable SO2 transformada adecuadamente).
- En la última columna se añaden datos ficticios correspondientes a una posible evolución de la variable SO2, en el año 1975; el nombre que recibe esta variable, en el fichero, es SO2tr_75.

```
datos = read.table(file = "Practica06-ciudades.csv", sep = ";", header = TRUE, dec = ".")
```

9. **Cinco años después, en 1975, se vuelve a medir la variable SO2_ tr, y se codifica como SO2_ tr75. Se pide**

   (a) **¿Se puede decir, al 10% de nivel de significación, que el nivel de SO2tr se incrementó entre 1970 y 1975?**
   En este caso los datos están pareados. Por tanto, hay que crear una nueva variable, que llamaremos `incremento`:

   ```
   incremento = datos$SO2_tr75 - datos$SO2_tr
   ```

   de nuevo habría que comprobar la normalidad de la nueva variable, pero se asume que lo es. A partir de la próxima práctica, que dispondrás de las herramientas adecuadas, habrá que hacer la comprobación, claro).
   Hay que contrastar las hipótesis

   $$H_0 : \mu_{incremento} \leq 0 \qquad H_1 : \mu_{incremento} > 0$$

   La muestra es grande porque

   ```
   length(incremento)
   ```

   ```
   [1] 41
   ```

   por lo que podemos usar un normal:

   ```
   library(TeachingDemos)
   muestra = incremento
   mu0 = 0
   (CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                    alternative = "greater", conf.level = 0.9))


   One Sample z-test

   data:  muestra
   z = 1.7591, n = 41.00000, Std. Dev. = 2.13879, Std. Dev. of the sample
   mean = 0.33402, p-value = 0.03928
   alternative hypothesis: true mean is greater than 0
   90 percent confidence interval:
    0.1594967       Inf
   sample estimates:
   mean of muestra
         0.5875637
   ```

   y concluir que sí hubo un incremento estadísticamente significativo (aunque este hecho no indica de por sí que fuera un incremento científicamente significativo).

(b) **Cuantifica la evolución de la concentración de SO2 entre los años 1970 y 1975**. El intervalo de confianza lo obtuvimos como subproducto del contraste en el apartado anterior: obsrva que el hecho de que el intervalo no contenga al cero corrobora que una de ellas supera a la otra y, como es positivo, efectivamente ha habido un incremento en la concentración de SO2. De hecho, este incremento es de entre 0.1595 y ∞.

10. **Hemos comprobado que, sobre esta muestra, el 56% de las ciudades con más de 1 millón de habitantes (POP>100) tienen un plan de sostenibilidad medioambiental. Este mismo porcentaje en Canadá, medido sobre una muestra de 35 ciudades, es del 72%. ¿Podemos admitir, al 99% de confianza, que hay diferencias significativas a este respecto entre ambos países?**

De nuevo hay que contrastar hipótesis. En este caso sobre proporciones: las hipótesis

$$H_0 : p_{USA} = p_{Canada} \qquad H_1 : p_{USA} \neq p_{Canada}$$

Veamos cuántas ciudades de la muestras supera el millón de habitantes

```
sum(datos$POP > 100)

[1] 39
```

esto es importante para poder consignar el tamaño de la muestra 1

```
n1 =  sum(datos$POP > 100) # tamanno muestra 1
n2 =  35 # tamanno muestra 2
# k1 =   # num de exitos 1
# k2 =   # num de exitos 2

# Si tienes proporciones muestrales p1, p2, entonces
# comenta las dos lineas anteriores. Despus, descomenta
# y usa estas cuatro lineas:
p1 = 0.56
p2 = 0.72
k1 = n1 * p1
k2 = n2 * p2
# Opciones para H1: greater / less / two.sided
(CH = prop.test(c(k1, k2), n = c(n1, n2), correct = FALSE,
                alternative = "two.sided",
                conf.level = 0.99))


2-sample test for equality of proportions without continuity
correction

data:  c(k1, k2) out of c(n1, n2)
X-squared = 2.039, df = 1, p-value = 0.1533
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.4430829  0.1230829
sample estimates:
prop 1 prop 2
  0.56   0.72
```

El p-valor indica que las proporciones se pueden considerar iguales.

Si se quisiera cuantificar entre qué valores está la diferencia, basta con fijarse en el intervalo de confianza para la diferencia de proporciones:

```
[1] -0.4431  0.1231
```

es decir, $p_{USA} - p_{Canada} \in (-0.4431, 0.1231)$ lo que equivale a

$$p_{Canada} - 0.4431 < p_{USA} < p_{Canada} + 0.1231$$