# PRACTICA 6: intervalos de confianza. Soluciones.

## Estadística, Grado en en Ciencias Ambientales, UAH.

### Objetivos

- Determinar las condiciones de aplicabilidad de los intervalos de confianza y el tipo de intervalo a utilizar en cada situación.
- Construir los intervalos de confianza para la media, la varianza y la proporción.
- Interpretar los resultados obtenidos.

Los datos que contiene el fichero `Practica06-ciudades.csv` 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 aparecen son:

- CIUDAD (nombre de la ciudad).
- SO2 (contenido de SO2 en el aire medido en microgramos por metro cúbico).
- TEMP (promedio de la temperatura anual en grados Fahrenheit).
- MAN (número de empresas de manufacturación con 20 o más empleados).
- POP (tamaño demográfico en decenas de miles según el censo de 1970).
- VIENTO (promedio de la velocidad del viento anual en millas por hora)
- LLUVIA (promedio de la precipitación anual en pulgadas).
- DIASLLUVIA (promedio del número de días que llueve al año).
- CLIMA (variable categórica que especifica el tipo de clima).
- 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.

Empieza por leer el fichero `Practica06-ciudades.csv` y guardar su contenido en la variable `ciudades`. Puedes usar la función de R `read.table()` (no olvides fijar correctamente la carpeta de trabajo)

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

o bien usar el botón `Import dataset` de RStudio.

Visualiza las primeras filas de la tabla

```
head(ciudades, 4)

##          CIUDAD SO2 TEMP MAN POP VIENTO LLUVIA DIASLLUVIA       CLIMA   SO2_tr
## 1       Phoenix  10 70.3 213 582    6.0   7.05         36 continental 2.302585
## 2    LittleRock  13 61.0  91 132    8.2  48.52        100 subtropical 2.564949
## 3 San Francisco  12 56.7 453 716    8.7  20.66         67 subtropical 2.484907
## 4        Denver  17 51.9 454 515    9.0  12.95         86 continental 2.833213
##   SO2_tr75
## 1  4.68068
## 2  4.15766
## 3  6.47724
## 4  2.03278
```

**ADVERTENCIA**: el contenido de las plantillas de inferencia que aparecen en las soluciones es ligeramente diferente de la versión actual. Pero no debería haber problema para que localices que la información esencial y, por supuesto, las soluciones son independientes de la versión de la platilla.

Responde de forma concisa y razonadamente a las siguientes preguntas:

1. Para la variable SO2_*tr*:

   (a) Intervalo de confianza al 95% para la media. Como disponemos de una muestra (datos en bruto) que tiene 41 datos la media muestras se distribuye como una normal (si fuera una muestra pequeña de una población normal usaríamos una t de Student). Se usa la plantilla de la normal con datos en bruto `media_1pob_T_o_Z_enBruto.R`
   que se rellena como sigue

```
library(MASS)
library(TeachingDemos)

# Ajusta los parametros de read.table.
ciudades = read.table(file = "Practica06-ciudades.csv", sep = ";",
                    header = TRUE, dec = ".")
muestra = ciudades$SO2_tr

# ajustar para contrastar hipotesis
mu0 = 0

# Opciones para alternativa: greater / less / two.sided
# Elige el contraste adecuado y descomenta la l??nea correspondiente
# (CHconT = t.test(muestra, mu = mu0, alternative = "", 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 = 28.747, n = 41.00000, Std. Dev. = 0.70230, Std. Dev. of the sample
## mean = 0.10968, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.938034 3.367974
## sample estimates:
## mean of muestra
##        3.153004
```

   De donde se tiene

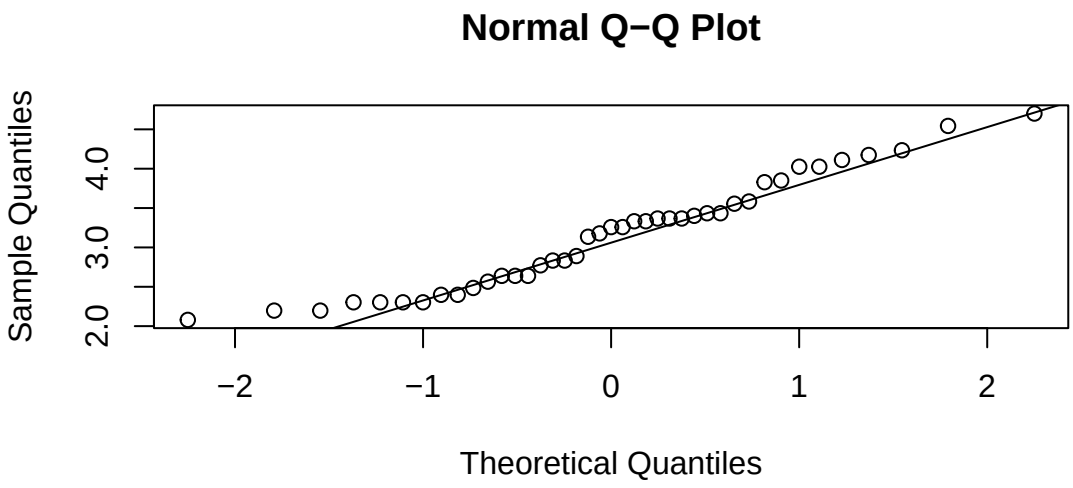
```
(inter1 = signif(unname(CHconZ$conf.int[1:2]), digits = 4))

## [1] 2.938 3.368
```

   es decir, el nivel de SO2tr está entre $2.938 mg/m^3$ y $3.368 mg/m^3$ con probabilidad 0.95.

   (b) Intervalo de confianza al 95% para la desviación típica.

   Para hacer inferencia (calcular un intervalo de confianza o hacer un contraste de hipótesis) sobre la varianza es necesario que la variable provenga de una muestra normal. Es preferible usar un qqplot que un histograma o un boxplot. Compara las tres opciones

```
qqnorm(ciudades$SO2_tr)
qqline(ciudades$SO2_tr)
```




El qqplot sugiere no rechazar la normalidad de los datos, porque no hay fuertes desviaciones respecto de la recta.

Ahora hay que usar la plantilla `varianza_1pob_enBruto.R` como sigue

```
library(MASS)
library(TeachingDemos)

ciudades = read.table(file = "Practica05-ciudades.csv", header = TRUE,
                    sep = ";", dec = ".")
muestra = ciudades$SO2_tr

# ajustar para contrastar hipotesis
sigma0 = 1
# Opciones para alternativa: greater / less / two.sided
(CH = sigma.test(muestra,
                 sigma = sigma0,
                 alternative = "two.sided",
                 conf.level = 0.95))
signif(CH$conf.int, digits = 4)
```

El resultado es

```
##
##  One sample Chi-squared test for variance
##
## data:  muestra
## X-squared = 19.729, df = 40, p-value = 0.005957
## alternative hypothesis: true variance is not equal to 1
## 95 percent confidence interval:
##  0.3324631 0.8074693
## sample estimates:
## var of muestra
##      0.4932232
```

de donde nos interesa

```
## [1] 0.3324631 0.8074693
```

(c) Si quisiéramos estimar el valor medio de $SO2_tr$ al nivel de confianza del 95% con una precisión de 0.1, ¿qué tamaño de muestra necesitaríamos?
Siempre que se toma una medida hay que asignarle una precisión (el error de medida). Por ejemplo, si se mide la altura de un individuos con un metro que mide en centímetros, un resultado serón

$179 \pm 1cm$. De la misma manera, el intervalo de confianza para la media poblacional es una medida de la media junto con un error que incorpora el hecho de que sólo accedemos a parte de los datos (la muestra).
Pero la forma que tiene es la misma que el caso de la altura:

$$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Lo que nos piden es el tamaño de la muestra para que el error de la estimacin, que vale La amplitud del intervalo es

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

sea menor que 0.1. Con los datos del problema, esto equivale a

$$1.96\frac{0.7023}{\sqrt{n}} < 0.1$$

basta con despejar

$$1.96\frac{0.7023}{0.1} < \sqrt{n}$$

para obtener que n debe ser el número natural mayor que

```
alfa = 0.05
(qnorm(1-alfa/2)*sd(ciudades$SO2_tr)/0.1)^2

## [1] 189.4697
```

es decir

```
## [1] 190
```

Hay al menos dos comentarios sobre este ejercicio:

i. Puedes (y debes) pensar sobre la precisión que quieres para tu estimación ANTES de ir a por los datos; eso ahorra sorpresas desagradables.
ii. Para conseguir el doble de precisión haría falta muchos más que el doble de datos.

(d) Se considera que valores de $SO2_tr$ superiores a 2.90 entrañan un riesgo grave para la salud. ¿Dirías, a un 95% de confianza, que es el caso de las ciudades americanas? ¿Se mantiene la misma conclusión al 99% de confianza?

Se pide el intervalo de confianza al 95% para la media de $SO2_{tr}$. Se calcula con la plantilla `media_1pob_T_o_Z_enBruto.R`

(e) Con un nivel de confianza del 95%, ¿qué porcentaje de ciudades verifica SO2_tr > 2.90?

Para calcular la estimación puntual hay que contar casos los favorables y dividir entre el total:

```
sum(ciudades$SO2_tr>2.9)/length(ciudades$SO2_tr)

## [1] 0.5609756
```

Recuerda que el conjunto de ciudades seleccionadas es sólo una muestra, por lo que la mejor respuesta es calcular el intervalo de confianza para la proporción a partir de la muestra. Se calcula con la plantilla `proporcion_1pob.R`

```
n = length(ciudades$SO2_tr)     #num. elementos muestra
k = sum(ciudades$SO2_tr>2.9)    #num. exitos muestra
(pMuestral = k / n)

## [1] 0.5609756
```

```r
# ajustar para contrastar hipotesis
p0 = 0.5
# Opciones para alternativa: greater / less / two.sided
CH = prop.test(k, n, conf.level = 0.95, p = p0,
          correct=FALSE, alternative = "two.sided")

signif(CH$conf.int, digits = 4)

## [1] 0.4104 0.7011
## attr(,"conf.level")
## [1] 0.95
```

Se concluye que, redondeando, entre el 40% (mejores estimaciones) y el 70% (peores estimaciones) de las ciudades podían tener dicho problema.

2. Calcula un intervalo de confianza del 92% para la temperatura media de las ciudades con clima continental húmedo. ¿Cabía esperar, un año cualquiera, que la temperatura media rondara los 50 grados Farenheit? ¿Y los 55?

Lo primero es seleccionar los datos con los que calcular el intervalo. Se guardan en la variable `continhum`

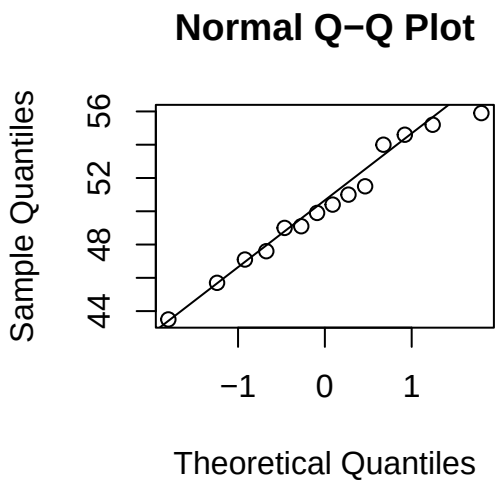
```r
ciudades = read.table(file = "Practica06-ciudades.csv", sep = ";", header = TRUE, dec = ".")
continhum = ciudades$TEMP[ciudades$CLIMA == "continental-humedo"]
```

observa que hay

```r
length(continhum)

## [1] 14
```

datos, es decir, la muestra es pequeña. Por eso tenemos que recurrir a una t de Student, lo que implica que es necesario que los datos provengan de una distribución normal. A la vista del qqplot

```r
qqnorm(continhum)
qqline(continhum)
```




no hay motivos para rechazar la normalidad de los datos. El intervalo de confianza se calcula con la plantilla `media_1pob_T_o_Z_enBruto.R`.

```r
library(MASS)
library(TeachingDemos)

muestra = continhum

# ajustar para contrastar hipotesis
mu0 = 0
# 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 = 0.92))
# (CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra), alternative = "", conf.level = ))
```

y el intervalo es

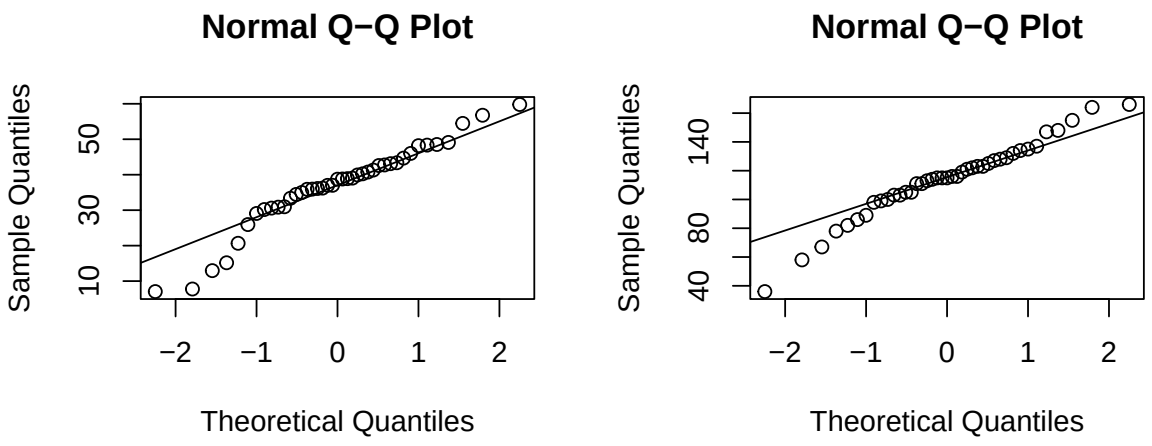
```
## [1] 48.45 52.20
```

Puesto que 50 está dentro del intervalo, sí cabría esperar que se alcance esa temperatura. En cambio, 55 está fuera y en consecuencia no cabía esperarla.

3. Compara la variabilidad (a través de la varianza) de las variables lluvia y días de lluvia con un nivel de confianza del 95% en las ciudades de Estados Unidos

   Se pide hacer un intervalo de confianza para el cociente de varianzas. Es necesario que las muestras provengan de poblaciones normales, y lo comprobamos con los correspondientes qqplots:

```r
par(mfrow = c(1,2))
qqnorm(ciudades$LLUVIA)
qqline(ciudades$LLUVIA)

qqnorm(ciudades$DIASLLUVIA)
qqline(ciudades$DIASLLUVIA)
```




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

La mayoría de los puntos quedan convenientemente alineados, por lo que no hay suficiente motivo para dudar de la normalidad de esos datos (aunque lo más apropiado es hacer un contraste de hipótesis para asegurarse).

Usando la plantilla correspondiente (`CocienteVar_2pob_enBruto.R`)

```r
## LECTURA DESDE FICHERO
ciudades = read.table(file = "Practica06-ciudades.csv", sep = ";", header = T, dec = ".")
muestra1 = ciudades$LLUVIA
muestra2 = ciudades$DIASLLUVIA

# Opciones para alternativa: greater / less / two.sided
(varCH = var.test(muestra1, muestra2, alternative = "two.sided"))

##
##  F test to compare two variances
##
## data:  muestra1 and muestra2
## F = 0.19723, num df = 40, denom df = 40, p-value = 1.184e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1051764 0.3698385
## sample estimates:
## ratio of variances
##          0.1972265
```

El intervalo de confianza indica que

$$0.1052 < \frac{\sigma^2_{lluvia}}{\sigma^2_{dias-lluvia}} < 0.3698$$

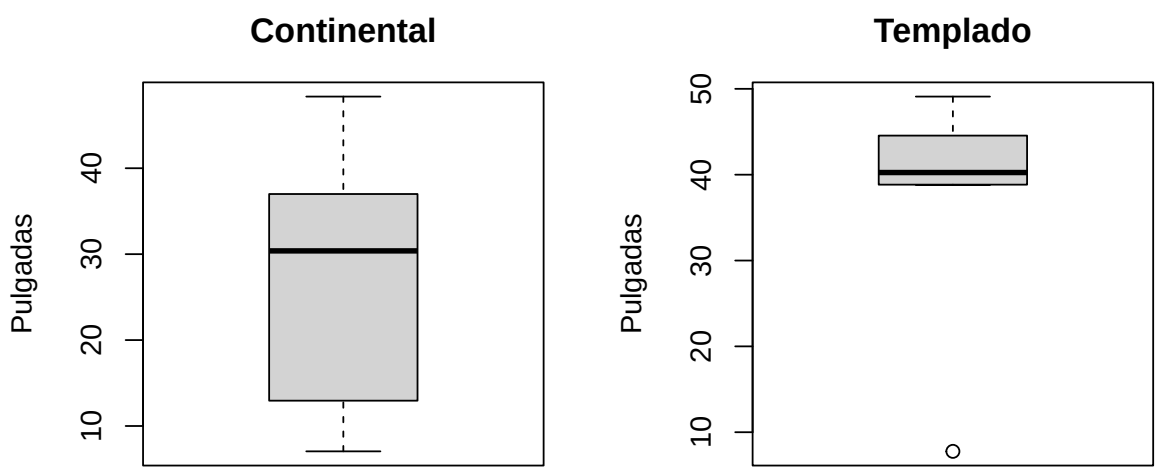
con un nivel de confianza del 95%. En concreto, el intervalo indica que

$$0.1052 * \sigma^2_{dias-lluvia} < \sigma^2_{lluvia} < 0.3698 * \sigma^2_{dias-lluvia}$$

o, lo que es lo mismo, la varianza $\sigma^2_{lluvia}$ es entre 0.1052 y 0.3698 veces la varianza $\sigma^2_{dias-lluvia}$. Por tanto, $\sigma^2_{lluvia}$ es menor que $\sigma^2_{dias-lluvia}$ con un nivel de confianza del 95%.

4. Compara la cantidad media de lluvia caída en ciudades de clima templado y de clima continental al nivel de confianza del 90%. En el trabajo real, es siempre importante hacer una exploración previa de los datos:

```r
par(mfrow = c(1,2), mar = c(1,5,3,1))
boxplot(ciudades$LLUVIA[ciudades$CLIMA == "continental"], main = "Continental",
        ylab = "Pulgadas")
boxplot(ciudades$LLUVIA[ciudades$CLIMA == "templado"], main = "Templado",
        ylab = "Pulgadas")
```



```
par(mfrow = c(1,2))
```

Hay un dato cláramente atípico, que en este caso vamos a eliminar. Te guiamos en el proceso:

- ¿De qué ciudad se trata? Es una ciudad de CLIMA templado y LLUVIA inferior a 10 (ver boxplot)

  ```
  ciudades[(ciudades$CLIMA == "templado") & (ciudades$LLUVIA < 10), ]

  ##         CIUDAD SO2 TEMP MAN POP VIENTO LLUVIA DIASLLUVIA    CLIMA   SO2_tr
  ## 23 Albuquerque  11 56.8  46 244    8.9   7.77         58 templado 2.397895
  ##    SO2_tr75
  ## 23  4.92607
  ```

  Se trata de Albuquerque.

- ¿cantidad de lluvia ha caído en Albuquerque? Puedes hacer

  ```
  ciudades$LLUVIA[ciudades$CIUDAD == "Albuquerque"]

  ## [1] 7.77
  ```

- ¿Piensas que este dato se corresponde con el resto de ciudades de clima templado? En caso negativo, descarta provisionalmente este dato. Para eso copia, pega y ejecuta el siguiente código:

  ```
  SinAlbqq = ciudades[ciudades$CIUDAD != "Albuquerque" , ]
  ```

  y termina el ejercicio con la tabla SinAlbqq.

Te recordamos la pregunta: ¿observas diferencias significativas, al 90% de confianza, entre la precipitación media en las ciudades de clima contnental y htemplado? Haz la comparación asumiendo y no asumiendo varianzas iguales. Calcula en cada caso un intervalo de confianza para la diferencia de medias, e interprétalo.

Se pide comparar los valores medios de precipitación en dos poblaciones diferentes. Concretamente, la precipitación media caída en ciudades de clima templado y de clima continental. Las variables no son pareadas, y llamaremos a cada grupo `lluviaTemp` y `lluviaCont`, respectivamente:

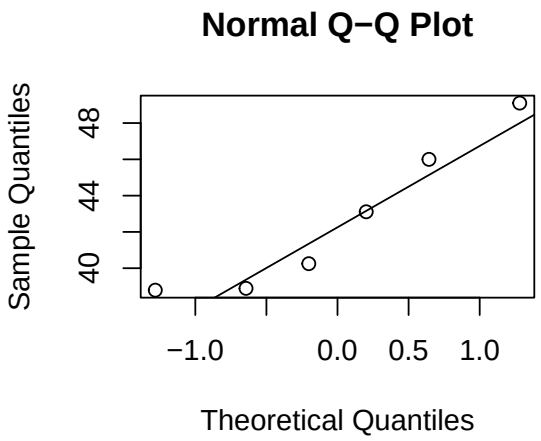
```
lluviaTemp = SinAlbqq$LLUVIA[SinAlbqq$CLIMA == "templado"]
lluviaCont = SinAlbqq$LLUVIA[SinAlbqq$CLIMA == "continental"]
```

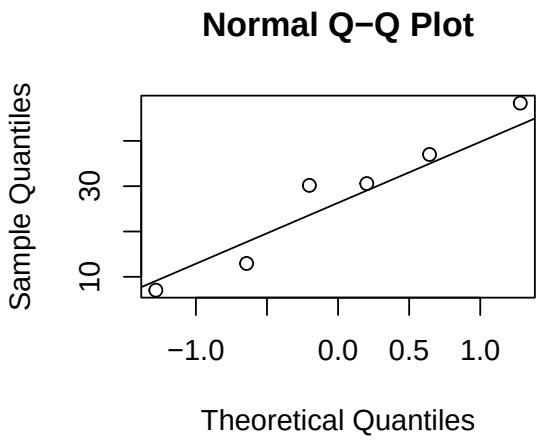
Veamos el tamaño de cada muestra para decidir si se usa una normal o la t de Student

```
length(lluviaTemp)

## [1] 6

length(lluviaCont)

## [1] 6
```

Como las muestras son pequeñas, hay que usar la t de Student, y para ello es preciso comprobar la normalidad de los datos

```
par(mfrow = c(1,2))
qqnorm(lluviaTemp)
qqline(lluviaTemp)

qqnorm(lluviaCont)
qqline(lluviaCont)
```




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

No hay fuertes desviaciones respecto de la recta, por lo que no hay motivos para rechazar la normalidad de los datos.

Ahora hay que decidir si se puede suponar o no que las varianzas sean iguales; para eso usar la plantilla correspondientes (intervalo confianza para codiente de varianzas datos en bruto) se obtiene

```
muestra1 = lluviaTemp
muestra2 = lluviaCont

# Opciones para alternativa: greater / less / two.sided
(CH = var.test(x = muestra1, y = muestra2, alternative = "two.sided", conf.level = 0.9))

##
##  F test to compare two variances
##
## data:  muestra1 and muestra2
## F = 0.075004, num df = 5, denom df = 5, p-value = 0.01294
## alternative hypothesis: true ratio of variances is not equal to 1
## 90 percent confidence interval:
##  0.0148513 0.3787946
## sample estimates:
## ratio of variances
##         0.07500395
```

Como el intervalo de confianza no contiene al 1, las varianzas deben ser consideradas distintas. Veamos esto en detalle: lo que dice el intervalo de confianza es

$$0.015 < \frac{\sigma_1^2}{\sigma_2^2} < 0.38$$

Como el cociente es más pequeño que 1, entonces el numerador es más pequeño que el denominador (con un nivel de confianza del 90%), es decir, que esas varianzas no pueden ser iguales (con un nivel de confianza del 90%). Si queremos afinar más, podemos decir que $\sigma_1^2$ es entre 0.015 y 0.38 veces $\sigma_2^2$.

La conclusión sobre la igualdad de varianzas habría sido análoga si el intervalo tuviera los dos extremos mayores que 1 (pero en ese caso $\sigma_2^2$ sería menor que $\sigma_1^2$). Razonando de forma análoga, se consideraría que las varianzas no son diferentes si el 1 estuviera en el intervalo de confianza para el cociente de varianzas.

A partir de ahí, usar la plantilla diferencia de medias usando t varianza distintas. El intervalo calculado es para `lluviaTemp - lluviaCont`:

```r
muestra1 = lluviaTemp
muestra2 = lluviaCont

(CHcont = t.test(x = muestra1, y = muestra2, var.equal = FALSE,
       alternative = "two.sided", conf.level = 0.9))

##
##  Welch Two Sample t-test
##
## data:  muestra1 and muestra2
## t = 2.3167, df = 5.7458, p-value = 0.06161
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   2.319241 27.694092
## sample estimates:
## mean of x mean of y
##  42.69000  27.68333
```

como el intervalo de confianza tiene ambos extremos positivos, la media de lluviaTemp es mayor que la de lluviaCont. Podemos afinar más y decir que la precipitación en las ciudades de clima templado es entre c y 27.69 $l/m^2$ superior.

Si las varianzas fueran diferentes, el intervalo obtenido habría sido

```r
(CHcont1 = t.test(x = muestra1, y = muestra2, var.equal = TRUE,
       alternative = "two.sided", conf.level = 0.9))

##
##  Two Sample t-test
##
## data:  muestra1 and muestra2
## t = 2.3167, df = 10, p-value = 0.04301
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   3.266395 26.746938
## sample estimates:
## mean of x mean of y
##  42.69000  27.68333
```

Fíjate en que este intervalo es menor que el obtenido para variazas diferentes. En concreto la semi-anchura del intervalo es 1 $l/m^2$ menor. Es decir, la media de lluviaTemp sigue siendo mayor que la de lluviaCont, pero en este caso la diferencia se ve recortada: como mínimo de 1 $l/m^2$ más (de 2.319 a 3.266 ) y, como máximo, de 1 $l/m^2$ menos (de 27.69 a 26.75 ).

5. Admitiendo como ciertos, para el valor de SO2_ tr en 1975 los datos de la variable SO2tr_ 75, estima al 90% de confianza, la variación del nivel de SO2_ tr entre 1970 y 1975.
Se trata de averiguar si la concentración media de $SO2_{tr}$ a aumentado o no (si la diferencia de medias es positiva o no). En este caso los datos están pareados. Por tanto, hay que crear una nueva variable, que llamaremos `incremento`:

```r
muestra = ciudades$SO2_tr - ciudades$SO2_tr75
```

La muestra es grande porque

```r
length(muestra)

## [1] 41
```

por lo que podemos usar un normal:

```r
mu0 = 0
(CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                alternative = "two.sided", 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 = 0.07857
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  -1.13698169 -0.03814565
## sample estimates:
## mean of muestra
##      -0.5875637
```

El hecho de los dos extremos del intervalo de confianza sean negativos implica que la diferencia de medias de 1970 y 1975 es negativa, es decir, hab??a menos SO2 en 1970 que en 1975, y que esa reducción es de entre 0.038 y 1.1 unidades.

6. 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%. A partir de los valores muestrales, usa un intervalo de confianza al 99% para estimar a nivel poblacional la diferencia entre ambas proporciones.
Se trata de calcular el intervalo de confianza para la diferencia de proporciones:

```r
n1 =  41 # 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 = 2.0823, df = 1, p-value = 0.149
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.4394479  0.1194479
## sample estimates:
## prop 1 prop 2
##   0.56   0.72
```

El intervalo indica que

$$-0.4394 < p_1 - p_2 < 0.1194$$

lo que equivale a

$$p_2 - 0.4394 < p_1 < p2 + 0.1194$$

es decir, la proporción $p_1$ de ciudades en Estados unidos es entre un 43.94% menor y 11.94% mayor que en Canadá.