PRACTICA 7: intervalos de confianza. Soluciones.

Estadística, Grado en Biología, UAH, 2022/23.

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 `Practica07-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 sobre la variable SO2_*tr*:

1. 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

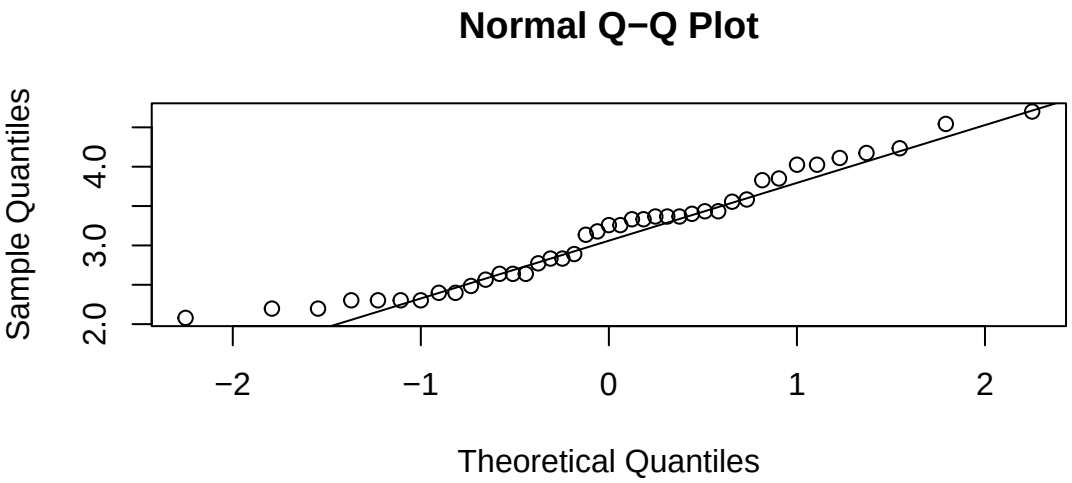
```
(inter1 = signif(unname(CHconZ$conf.int[1:2]), digits = 4))

## [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.

2. 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
```

3. 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 estimación, 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:

(a) Puedes (y debes) pensar sobre la precisión que quieres para tu estimación ANTES de ir a por los datos; eso ahorra sorpresas desagradables.

(b) Para conseguir el doble de precisión haría falta muchos más que el doble de datos.

4. 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`

5. 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`

```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

# 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.

6. Por un lado, en el ejercicio **??** se tiene que con un nivel de confianza del 95% el valor medio poblacional de $SO2_{tr}$ supera el umbral de 2.9. Por otro lado, del ejercicio **??** se desprende que entre el 40% y el 70% de las poblaciones superan dicho umbral con el mismo nivel de confianza. ¿Resultan contradictorios ambos resultados?

   No hay contradicción entre ambos resultados. Ten en cuenta que cualquier media es el resumen de medidas individuales. El que la media esté por encima de 2.9 no implica que todos los valores a partir de los que se calcula estén por encima de dicho valor. Algunos de ellos podrían estar por debajo. El ejercicio 5 nos dice como mínimo el 40% de ellas supera el umbral, y como mucho lo hace el 70%. Por tanto, las que NO supera el umbral suponen entre un 30% y un 60 % del total de ciudades.

7. 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`

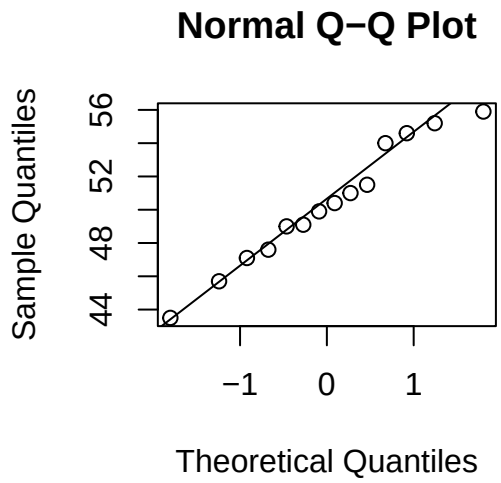
```r
ciudades = read.table(file = "Practica06-ciudades.csv", sep = ";", header = TRUE, dec = ".")
continhum = ciudades$TEMP[ciudades$CLIMA == "continental-humedo"]
```

   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`.

```
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 esperar que se alcance esta otra temperatura.

8. En el primer ejercicio hemos averiguado que la estimación puntual de la media de SO2_tr es

```
[1] 3.153004
```

Del intervalo de confianza calculado se deduce que la precisión de la estimación (la semi anchura del intervalo) al nivel de confianza del 95% es

```
[1] 0.21497
```

En el ejercicio 3 se pregunta cuánto hay que aumentar el tamaño de la muestra para mejorar la precisión hasta 0.1 unidades. Esa estrategia implica aumentar la muestra desde 200 hasta 190 individuos, lo que implica consumir muchos recursos.

Otra opción es la siguiente: reducir el nivel de confianza reducie el tamanño del intervalo, lo que aumenta la precisión (ojo, a costa de reducir en nivel de certidumbre).

Para explorar esa vía (redondeando valores para que salgan números más limpios) supón que la media muestral de la variable SO2_tr es $\bar{X} = 3.3$, $s = 0.7$ y $n = 41$. Considera el intervalo de confianza

$IC_\mu = (3.2, 3.4)$ (es decir, tiene precisión 0.1, pero has mantanido la muestra en 41 elementos) ¿cuál es el nivel de confianza de este intervalo?

La semianchura del intervalo vienen dada por

$$0.1 = z_{\alpha/2}\frac{30}{\sqrt{200}}$$

de donde

$$0.1\frac{\sqrt{41}}{0.7} = 0.914732 = z_{\alpha/2}$$

esto es, buscamos α tal que el valor crítico $z_{\alpha/2}$ de una normal N(0,1) sea 0.914732, es decir,

```
(alphamedios = pnorm(0.1*sqrt(41)/0.7, lower.tail = F))

[1] 0.1801662
```

de donde α vale

```
(alpha = 2*alphamedios)

[1] 0.3603323
```

y el nivel de confianza es del

```
(1-alpha)*100

[1] 63.96677
```

Como conclusión, la reducción en el nivel de confianza es muy grande, y se desaconsejaría esta estrategia.