PRACTICA 9: inferencia sobre dos poblaciones. Soluciones.

Objetivos

- Hacer inferencia (estimación por intervalos de confianza y contraste de hipótesis) sobre dos poblaciones para la media, la varianza y la proporción.
- Comprobar las condiciones que permiten hacer inferencia sobre los parámetros antes mencionados.
- Interpretar los resultados obtenidos. En particular,
  - Combinar la información que proporcionan los intervalos y el p-valor asociados al contraste de una hipótesis estadística.
  - Distinguir entre diferencia estadísticamente significativa y científicamente significativa.

El fichero Pima.tr se refiere a mujeres de la etnia Pima de al menos 21 años de edad que viven cerca de Phoenix, Arizona. Se hizo un test a cada una de ellas de diabetes de acuerdo con los criterios de la Organización Mundial de la SAlud (WHO); los datos fueron recogidos por el Instituto Nacional para la Diabetes y enfermedades Digestivas y de riñón de Estados Unidos. Las variables medidas fueron

1. npreg: número de embarazos.
2. glu: concentración de glucosa en sangre tras un test oral de tolerancia a la glucosa (mg/dm).
3. bp: presión diastólica (mm Hg).
4. skin: grosor del pliege de la piel en el triceps (mm).
5. bmi: índice de masa corporal, en kg/m^2.
6. ped: diabetes pedigree function. Función que asigna la probabilidad de padecer diabetes a partir de la historia familiar.
7. age: edad, en años.
8. type: Yes or No, para diabetes, de acuerdo con los criterios de la WHO.

Enunciados y soluciones

Empieza por leer el fichero `Practica07-pima.csv` y guardar su contenido en la variable `pima`. Puedes usar la función de R `read.table()` (no olvides fijar correctamente la carpeta de trabajo)

```
pima = read.table(file = "Practica07-pima-tr-BS.csv", sep = ";", header = TRUE, dec = ".")
```

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

Visualiza las primeras filas de la tabla

```
head(pima, 4)

##   npreg glu bp skin  bmi   ped age type
## 1     5  86 68   28 30.2 0.364  24   No
## 2     7 195 70   33 25.1 0.163  55  Yes
## 3     5  77 82   41 35.8 0.156  35   No
## 4     0 165 76   43 47.9 0.259  26   No
```

1. **Se quiere comparar si la disparsión en la presión arterial es la misma entre las indias a la que se les diagnosticó diabetes y las que no**:

   (a) **¿Se puede afirmar, con un nivel de confianza del 95%, que la dispersión (varianza) es la misma?**

   Se pide contrastar la hipótesis

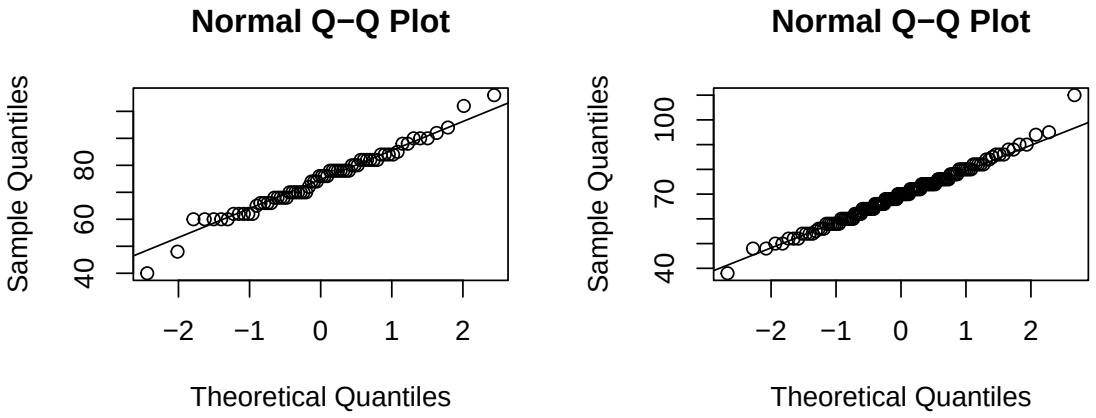
   $$H_0 : \sigma^2_{Yes} = \sigma^2_{No} \qquad H_1 : \sigma^2_{Yes} \neq \sigma^2_{No}$$

   Es necesario que las muestras provengan de poblaciones normales, lo que supondremos cierto

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

qqnorm(pima$bp[pima$type == "Yes"])
qqline(pima$bp[pima$type == "Yes"])

qqnorm(pima$bp[pima$type == "No"])
qqline(pima$bp[pima$type == "No"])
```




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

La mayor parte de los puntos quedan bien alineados, por lo que no hay suficiente motivo para dudar de su normalidad. Además, es psible hacer la comprobación mediante un contraste de hipótesis que debe complementar el análisis gráfico. El test de Shapiro plantea H_0: "la muestra viene de una población normal" frente a H_1: "la muestra NO proviene de un apoblación normal"

```
shapiro.test(pima$bp[pima$type == "Yes"])


Shapiro-Wilk normality test

data:  pima$bp[pima$type == "Yes"]
W = 0.98025, p-value = 0.3545
```

```
shapiro.test(pima$bp[pima$type == "No"])


Shapiro-Wilk normality test

data:  pima$bp[pima$type == "No"]
W = 0.98976, p-value = 0.4409
```

de forma análoga el test de Kolmogorov-Smirnov de libreria **nortset** (no olvides instalarla si no la tienes) es otro contraste de normalidad comunmente utilizado

```
library(nortest)
lillie.test(pima$bp[pima$type == "Yes"])


Lilliefors (Kolmogorov-Smirnov) normality test

data:  pima$bp[pima$type == "Yes"]
D = 0.080451, p-value = 0.3381
```

```
lillie.test(pima$bp[pima$type == "No"])


Lilliefors (Kolmogorov-Smirnov) normality test

data:  pima$bp[pima$type == "No"]
D = 0.052892, p-value = 0.4872
```

En todos los casos el pvalor es suficientemente grande como para no rechazar H_0 (es decir, no rechazar que las muestras prienen de poblaciones normales).
Usando la plantilla correspondiente (`CocienteVar_2pob_enBruto.R`)

```
## LECTURA DESDE FICHERO
muestra1 = pima$bp[pima$type == "Yes"]
muestra2 = pima$bp[pima$type == "No"]

# 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 = 1.0923, num df = 67, denom df = 131, p-value = 0.6603
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7286571 1.6851708
sample estimates:
ratio of variances
          1.092318
```

El p-valor es grande, y el 1 (el valor del cociente cuando las varianzas son iguales) está en el intervalode confianza para el cociente de medias, por lo que no parece haber motivo para rechazar la hipótesis nula (a falta de cuantificar cuán diferentes son las varianzas con el intervalo de confianza).

(b) **¿Qué relación (numérica) hay entre las varianzas?**

El intervalo de confianza indica que

$$0.7287 < \frac{\sigma^2_{Yes}}{\sigma^2_{No}} < 1.685$$

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

$$0.7287 * \sigma^2_{No} < \sigma^2_{Yes} < 1.685 * \sigma^2_{No}$$

o, lo que es lo mismo, la varianza σ^2_{Yes} es entre 0.7287 y 1.685 veces la varianza σ^2_{No}. Por tanto, σ^2_{Yes} es entre un 72.87% y 168.5% del valor de σ^2_{No} con un nivel de confianza del 95%.

2. **Determina si se puede considerar que el valor medio del grosor pliegue en el triceps (al nivel de significación del 10%) es mayor entre los individuos a los que se les diagnosticó diabetes y a los que no**.

Para poder afirmar que el grosor medio del triceps es mayor en los individuos con hipertensión necesito una evidencia muestral muy fuerte a favor de esa afirmación, por eso se contrastan las hipótesis

$$H_0 : \mu_{Yes} \leq \mu_{No} \qquad H_1 : \mu_{Yes} > \mu_{No}$$

Se definen las variables (se puede trabajar con las columnas de la tabla directamente)

```
hiper   = pima$skin[pima$type == "Yes"]
nohiper = pima$skin[pima$type== "No"]
```

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

```
length(hiper)

## [1] 68

length(nohiper)

## [1] 132
```

Como las muestras son grandes, se usa la normal

```
library(BSDA)

## Loading required package:  lattice
##
## Attaching package:  'BSDA'
## The following object is masked from 'package:datasets':
##
##     Orange

muestra1 = hiper
muestra2 = nohiper

s1 = sd(muestra1)
s2 = sd(muestra2)

# Opciones para H1: greater / less / two.sided
# solo para el IC, pon "two.sided"
(CHz = z.test(x = muestra1, y = muestra2, sigma.x = s1, sigma.y = s2,
       alternative = "greater", # "greater", "less", "two.sided"
       conf.level = 0.9))
```

```
##
##  Two-sample z-Test
##
## data:  muestra1 and muestra2
## z = 3.3421, p-value = 0.0004157
## alternative hypothesis: true difference in means is greater than 0
## 90 percent confidence interval:
##  3.645687       NA
## sample estimates:
## mean of x mean of y
##  33.11765  27.20455

CHz$conf.int

## [1] 3.645687       NA
## attr(,"conf.level")
## [1] 0.9
```

es decir, el p-valor 4.157233×10^{-4} sugiere rechazar H_0 al nivel de significación $\alpha = 0.1$.

Además, el intervalo de confianza (este es unilateral) indica que la diferencia entre las medias es de al menos 3.646 (aproximadamente). Habría que determinar si es o no una diferencia relevante desde el punto de vista científico recurriendo al conocimiento de un experto (3mm ronda el 10% de los valores medios de cada subpoblación, lo que sugiere que no es una cantidad despreciable).

3. **Supón (por suponer) que el índice de masa corporal de cada individuo se ha reducido un 7% tras cinco años. Se trata de una disminución significativa? Trabaja al 5% de significación. Pista, se trata de comparar `pima$bmi` con `pima$bmi` $* 0.93$.**

   Se trata de muestras pareadas, porque a cada individu se le ha medido el `bmi` dos veces con 5 años de diferencia. Por tanto, hay que crear una nueva variable

```
D = pima$bmi - pima$bmi*0.93
```

como el tamaño de la muestra es grande, es posible trabajar con una normal. Ojo, porque lo que queremos comprobar es si, efectivamente, hay una reducción, es decir, si $\mu_{antes} > \mu_{despues}$. Para afirmarlo, tendremos que observar una evidencia muestral potente. Por ello se elige

$$H_0 : \mu_{antes} \leq \mu_{despues} \qquad H_1 : \mu_{antes} > \mu_{despues}$$

```
library(TeachingDemos)


Attaching package: 'TeachingDemos'
The following object is masked from 'package:BSDA':
   z.test

(CHz = z.test(D, sd = sd(D), alternative = "greater", conf.level = 0.95))


One Sample z-test

data:  D
z = 74.538, n = 200.000000, Std. Dev. = 0.429115, Std. Dev. of the
sample mean = 0.030343, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
```

```
 2.21179     Inf
sample estimates:
mean of D
   2.2617
```

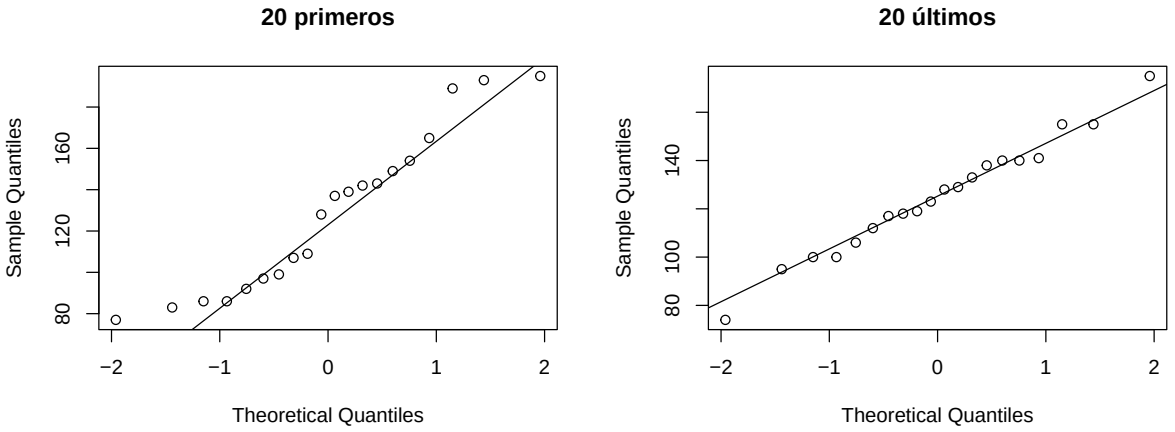
El p-valor es muy pequeño. Y el intervalo de confianza indica que, como poco, el índice de masa corporal era (aproximadamente) $2.21 kg/m^2$ mayor antes que después de 5 años. Podemos afirmar que hubo una reducción con un nvel de confianza del 95%.

4. **Un estudioso afirma que el nivel medio de glucosa de los primeros 20 individuos de la tabla menor que el de los 20 últimos. Contrasta dicha afirmación.**

   Se trata de comparar las medias de muestras pequeñas, por lo que hay que analizar su normalidad:

```
par(mfrow = c(1,2))
qqnorm(pima$glu[1:20], main = "20 primeros")
qqline(pima$glu[1:20])

qqnorm(pima$glu[181:200], main = "20 últimos")
qqline(pima$glu[181:200])
```




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

shapiro.test(pima$glu[1:20])


Shapiro-Wilk normality test

data:  pima$glu[1:20]
W = 0.92541, p-value = 0.126

shapiro.test(pima$glu[181:200])


Shapiro-Wilk normality test

data:  pima$glu[181:200]
W = 0.98947, p-value = 0.9975
```

en ambos casos tanto los qqplots como los contrastes sugieren normalidad, por lo que podemos hacer el contraste de medias con la t de Student.

En este caso se contrasta

$$H_0 : \mu_{primeros} \leq \mu_{ultimos} \qquad H_1 : \mu_{primeros} > \mu_{ultimos}$$

lo que equivale a

$$H_0 : \mu_{primeros} - \mu_{ultimos} \leq 0 \qquad H_1 : \mu_{primeros} - \mu_{ultimos} > 0$$

con el resultado de

```
(CHt = t.test(pima$glu[1:20], pima$glu[181:200], mu = 0,
              alternative = "greater", conf.level = 0.95))


Welch Two Sample t-test

data:  pima$glu[1:20] and pima$glu[181:200]
t = 0.35863, df = 31.807, p-value = 0.3611
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -13.40665       Inf
sample estimates:
mean of x mean of y
    128.5     124.9

mean(pima$glu[1:20])

[1] 128.5
```

El p-valor es grande, lo que sugiere no rechazar H_0. Además, el intervalo de confianza indica que el nivel de glucosa de los 20 últimos puede de, como mucho, hasta 13.4 unidades mayor que el de las 20 primeras, pero que el de las primeras también podría ser mayor que el de las 20 últimas (y no se especifica cuánto). La muestra no sugiere rechazar H_0.

5. **Se quiere analizar la (posible) asociación entre las variables tener diabetes diagnosticada (`type`) y el índice de masa corporal (`bmi`). Para ello analiza la diferencia de proporciones de indias con diabetes entre los individuos por debajo del primer cuartil y por encima del tercer cuartil respecto de la variable `bmi`. En concreto, considera los siguientes individuos.**

```
# por debajo del primer cuartil
table(pima$type[pima$bmi <= quantile(x = pima$bmi, 0.25)])


 No Yes
 46   4

# por encima del tercer cuartil
table(pima$type[pima$bmi > quantile(x = pima$bmi, 0.75)])


 No Yes
 26  23
```

Comenta los resultados.

Las variables estarían relacionadas si la proporción de indias con diabetes diagnosticada fuera diferente en cada uno de los grupos. Es decir, si estar por debajo de Q_1 o por encima de Q_3 en la variable `bmi` influye en la proporción de indias diagnosticadas. Por contra, si la proporción de indias diagnosticadas en cada grupo es similar (no diferente) podría arguirse que no hay relación entre ambas variables. Trabaja con un nivel de confianza del 90%

El contraste que se plantea es

$$H_0 : p_1 = p_2, \qquad H_2 : p_1 \neq p_2$$

donde p_1 se refiere a la proporción de indias con diabetes diagnosticada en el primer cuartil y p_2 opr encima del tercer cuartil.

```
n1 = 46 + 4 # tamanno muestra 1
n2 = 26 + 23 # tamanno muestra 2
k1 = 4 # num de exitos 1
k2 = 23   # num de exitos 2

# Si tienes proporciones muestrales p1, p2, entonces
# comenta las dos lineas anteriores. Después, 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
(CHp = prop.test(c(k1, k2), n = c(n1, n2), correct = FALSE,
                 alternative = "two.sided",
                 conf.level = 0.9))


2-sample test for equality of proportions without continuity
correction

data:  c(k1, k2) out of c(n1, n2)
X-squared = 18.918, df = 1, p-value = 1.365e-05
alternative hypothesis: two.sided
90 percent confidence interval:
 -0.5225591 -0.2562164
sample estimates:
   prop 1    prop 2
0.0800000 0.4693878
```

El p-valor es 1.3647655×10^{-5} muy pequeño, lo que sugiere rechazar H_0. De hecho, el intervalo de confianza $(-0.5226, -0.2562)$ indica que cabe esperar, al 90 de confianza, que la incidencia de la diabetes sea entre 0.2562 y 0.5226 puntos porcentuales superior en en el cuartil con mayor `bmi` que en el cuartil más pequeño.

Por tanto, con un nivel de confianza del 90% podemos afirmar que las variable están relacionadas, y que la presencia de diabetes es mayor entre los individuos con mayor índice de masa corporal. Ojo, que esto no quiere decir que haya relación de causa-efecto.