

## PRACTICA 7: Intervalos de confianza, una población.

Estadística, Grado en Biología Sanitaria, UAH.

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  $\text{kg/m}^2$ .
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

Guarda el contenido del fichero Pima.tr en una variable llamada pima. Se pide

1. Estima la glucosa media poblacional al nivel de confianza del 95% a partir de los datos de la muestra.
2. Estima la desviación típica poblacional de la presión sanguínea al nivel de confianza del 99% a partir de los datos de la muestra.
3. Si quisiéramos estimar el valor medio de la glucosa al nivel de confianza del 95% con una precisión de 4 mg/dl, ¿qué tamaño de muestra necesitaríamos?
4. Supón que la media muestral de la glucosa es  $\bar{X} = 124$ ,  $s = 30$  y  $n = 200$ . Si el intervalo de confianza que has obtenido para la media es  $IC_{\mu} = (120, 128)$  ¿cuál es el nivel de confianza del intervalo?

5. Estima, con un nivel de confianza del 90%, la prevalencia de la diabetes entre las mujeres de la etnia de las indias Pima.
6. Estima el índice de masa corporal media al nivel de confianza del 90% a partir de los primeros 20 datos de la muestra.

## Soluciones

Se utilizan fragmentos de las plantillas de inferencia

1. Fíjate en que dispones de los datos en bruto y de que la muestra es grande (usaremos una Z, una normal  $N(0,1)$ ):

```

library(MASS)
library(TeachingDemos)

muestra = pima$glu

# ajustar para contrastar hipotesis
mu0 = 0

# tamaño muestra
length(muestra)

[1] 200

# Elige el contraste adecuado y descomenta la línea corresp
(CHconZ = z.test(muestra, mu = mu0,
                 stdev = sd(muestra),
                 alternative = "two.sided", # "greater", "less",
                                     # "two.sided"
                 conf.level = 0.95))

One Sample z-test

data: muestra
z = 55.363, n = 200.0000, Std. Dev. = 31.6672, Std. Dev. of the sample
mean = 2.2392, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 119.5812 128.3588
sample estimates:
mean of muestra
 123.97

```

Es decir, el intervalo de confianza pedido es, redondeando a 4 cifras significativas

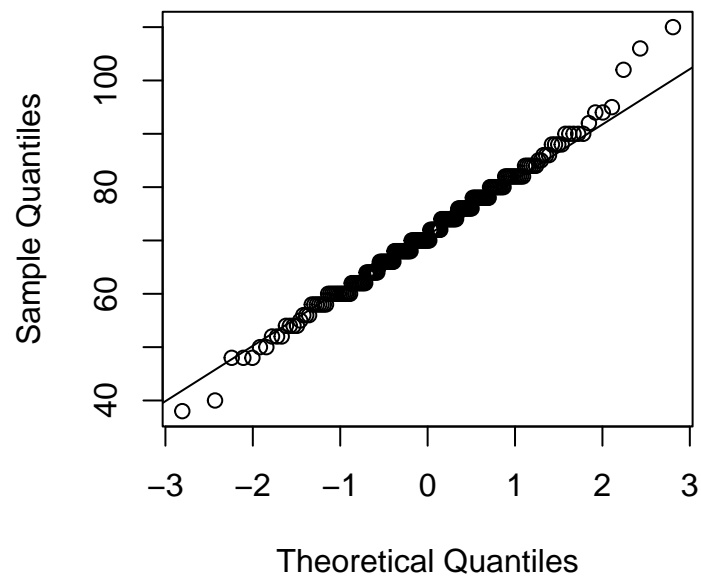
```
[1] 119.6 128.4
```

2. Estima la desviación típica poblacional de la presión sanguínea al nivel de confianza del 99% a partir de los datos de la muestra.

Lo primero es analizar la normalidad de la muestra

```
muestra = pima$bp  
qqnorm(muestra)  
qqline(muestra)
```

### Normal Q-Q Plot



Que podemos dar por buena. A partir de ahí, podemos calcular el intervalo de confianza

```
# ajustar H0, usa 1 para el IC  
sigma0 = 1  
# Opciones para H1: greater / less / two.sided
```

```
# solo para el IC, pon "two.sided"
(CHsigma2 = sigma.test(muestra,
                      sigma = sigma0,
                      alternative = "two.sided", # "greater", "less",
                                                # "two.sided"
                      conf.level = 0.99))

One sample Chi-squared test for variance

data: muestra
X-squared = 26224, df = 199, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
99 percent confidence interval:
 103.1911 173.2476
sample estimates:
var of muestra
 131.7813
```

Ten en cuenta que el resultado obtenido

```
CHsigma2$conf.int

[1] 103.1911 173.2476
attr("conf.level")
[1] 0.99
```

es el intervalo para la varianza; para obtener el intervalo de la desviación típica hay que extraer la raíz cuadrada que, con 4 cifras significativas, vale:

```
signif(sqrt(CHsigma2$conf.int[1:2]), digits = 4)

[1] 10.16 13.16
```

3. La precisión viene dada por la semianchura del intervalo de confianza

$$z_{0.025} \frac{s}{\sqrt{n}}$$

donde  $n$  es el tamaño de la muestra,

```
(z = qnorm(0.025, lower.tail = F))

[1] 1.959964
```

y  $s$  es la desviación típica muestral

```
(s = sd(pima$glu))  
## [1] 31.66723
```

Se pide determinar el valor de  $n$  tal que

$$1.959964 \frac{31.6672254}{\sqrt{n}} < 4$$

que equivale a

$$1.959964 \frac{31.6672254}{4} < \sqrt{n}$$

elevando al cuadrado ambos miembros, se tiene

$$240.7665926 < n,$$

es decir,

```
[1] 241
```

4. La semianchura del intervalo vienen dada por

$$4 = z_{\alpha/2} \frac{30}{\sqrt{200}}$$

de donde

$$4 \frac{\sqrt{200}}{30} = 1.8856181 = z_{\alpha/2}$$

esto es, buscamos  $\alpha$  tal que el valor crítico  $z_{\alpha/2}$  de una normal  $N(0,1)$  sea 1.8856181, es decir,

```
(alphamedios = pnorm(4*sqrt(200)/30, lower.tail = F))  
[1] 0.02967322
```

de donde  $\alpha$  vale

```
(alpha = 2*alphamedios)  
[1] 0.05934644
```

y el nivel de confianza es del

```
(1-alpha)*100
```

```
[1] 94.06536
```

5. Se trata de estimar por intervalos de confianza la proporción de mujeres que padece diabetes, y para ello precisamos conocer la proporción muestral y el tamaño de la muestra:

```
n = nrow(pima) #num. elementos muestra
k = sum(pima$type == "Yes") #num. exitos muestra
(pMuestral = k / n)

[1] 0.34

# ajustar para H0, usa 0.1 para el IC
p0 = 0.1

# Opciones para H1: greater / less / two.sided
# solo para el IC, pon "two.sided"
(CHp = prop.test(k, n, p = p0, correct=FALSE,
                 alternative = "two.sided", # "greater", "less",
                                     # "two.sided"
                 conf.level = 0.9))

1-sample proportions test without continuity correction

data: k out of n, null probability p0
X-squared = 128, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.1
90 percent confidence interval:
 0.2873664 0.3969047
sample estimates:
 p
0.34
```

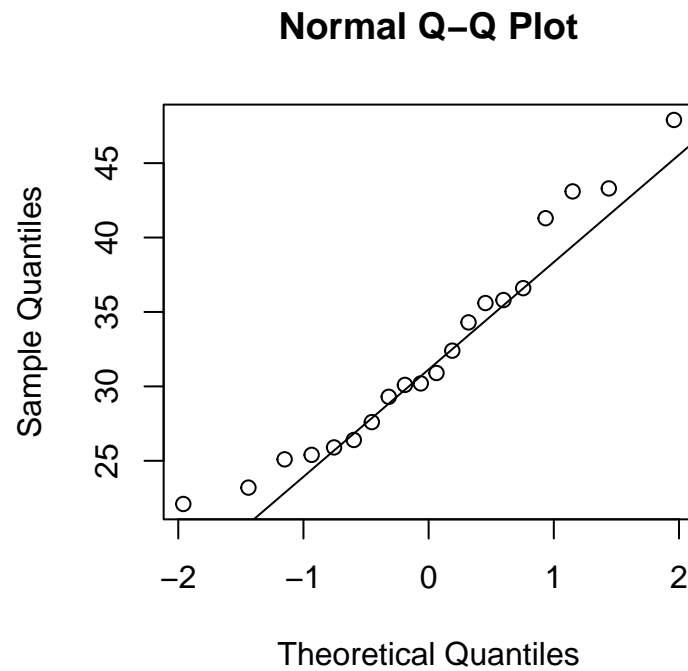
Donde el intervalo pedido, con 4 cifras significativas, es

```
signif(CHp$conf.int, digits = 4)
```

```
[1] 0.2874 0.3969
attr(,"conf.level")
[1] 0.9
```

6. En este caso se trata de una muestra pequeña, y lo primero es determinar la normalidad de los datos:

```
qqnorm(pima$bmi[1:20])  
qqline(pima$bmi[1:20])
```



No hay motivos para dudar de la normalidad de los datos, ya que están bien dispuestos sobre la recta.

Por tanto, se usa la pantalla para la media con datos en bruto y una  $t$  de Student

```
library(MASS)  
library(TeachingDemos)  
muestra = pima$bmi[1:20]  
  
# ajustar para H0, usa 0 para IC  
mu0 = 0  
  
(CHcont = t.test(muestra, mu = mu0,  
                 alternative = "two.sided", # "greater", "less",
```

```
                                # "two.sided"
conf.level = 0.9))

One Sample t-test

data: muestra
t = 19.808, df = 19, p-value = 3.795e-14
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 29.5032 35.1468
sample estimates:
mean of x
 32.325
```

Donde el intervalo pedido, con 4 cifras significativas, es

```
signif(CHcont$conf.int, digits = 4)

[1] 29.50 35.15
attr("conf.level")
[1] 0.9
```