ESTADÍSTICA I
Tema 4:
Estimación por intervalos de confianza

- El concepto de intervalo de confianza (IC)
- IC aproximados basados en el TCL: intervalos para una proporción
- Determinación del mínimo tamaño muestral
- Construcción de IC: el método de la cantidad pivotal
- Las distribuciones t de Student y χ^2
- Intervalos de confianza en poblaciones normales

El concepto de intervalo de confianza

Sea una muestra $X_1, \ldots, X_n$ de una v.a. con distribución de probabilidad dependiente de un parámetro desconocido $\theta \in \Theta \subset \mathbb{R}$.

Sean dos estadísticos $T_n^{(1)}(X_1, \ldots, X_n)$ y $T_n^{(2)}(X_1, \ldots, X_n)$ con $T_n^{(1)} < T_n^{(2)}$ y un valor $\alpha \in (0, 1)$. Supongamos que se verifica

$$\mathbb{P}_\theta\{T^{(1)}(X_1, \ldots, X_n) < \theta < T^{(2)}(X_1, \ldots, X_n)\} = 1 - \alpha, \ \forall \theta.$$

Entonces para una realización concreta de la muestra, $x_1, \ldots, x_n$, se dice que $(T^{(1)}(x_1, \ldots, x_n), T^{(2)}(x_1, \ldots, x_n))$ es un intervalo de confianza para θ con nivel de confianza $1 - \alpha$ y lo denotaremos $\mathrm{IC}_{1-\alpha}(\theta)$.

El método de la "cantidad pivotal"

Una metodología general para obtener un intervalo de confianza para θ consiste en encontrar una función $Q(\theta; X_1, \ldots, X_n)$ (llamada "cantidad pivotal") cuya distribución no dependa de θ y sea conocida (al menos de modo aproximado). A partir de esta distribución, fijado un valor $\alpha \in (0, 1)$ se obtienen dos valores $q_1(\alpha)$ y $q_2(\alpha)$ tales que

$$\mathbb{P}_\theta\{q_1(\alpha) < Q(\theta; X_1, \ldots, X_n) < q_2(\alpha)\} = 1 - \alpha.$$

Despejando θ se obtiene una expresión del tipo

$$\mathbb{P}_\theta\{T_n^{(1)}(X_1, \ldots, X_n) < \theta < T_n^{(2)}(X_1, \ldots, X_n)\} = 1 - \alpha,$$

que ya proporciona directamente el intervalo de confianza.

Un ejemplo: intervalo de confianza para la media de una normal con varianza conocida

Supongamos que $X_1, \ldots, X_n$ son v.a.i.i.d. $N(\mu, \sigma)$, donde μ es un parámetro desconocido y σ es conocida. Se sabe que

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \text{ y, tipificando, } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Por tanto, si para cualquier $\alpha \in (0,1)$, z_α denota el cuantil $1-\alpha$ en la normal estándar (e.d., $\Phi(z_\alpha) = 1-\alpha$, siendo Φ la función de distribución de la $N(0,1)$) tenemos

$$\mathbb{P}_\mu\left\{-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\} = 1-\alpha$$

y, despejando,

$$\mathbb{P}_\mu \left\{ \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha.$$

Se concluye que

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right)$$

es un intervalo de confianza de nivel $1 - \alpha$ para μ.

Interpretación intuitiva en términos "frecuentistas":
Si, por ejemplo, $1 - \alpha = 0.95$ y extraemos muchas muestras de una $N(0, 1)$ aproximadamente en el 95% de los casos el intervalo de confianza contiene al verdadero valor $\mu = 0$ del parámetro.

Cuando aceptamos que el modelo que generó los datos de una muestra es normal, lo habitual es suponer que la media μ y la desviación típica σ son desconocidas y hay que estimarlas a partir de los datos. Por ello, R no tiene una orden para calcular intervalos de confianza para la media μ de una normal con varianza σ^2 *conocida*. Sin embargo, podemos programarlo nosotros mismos:

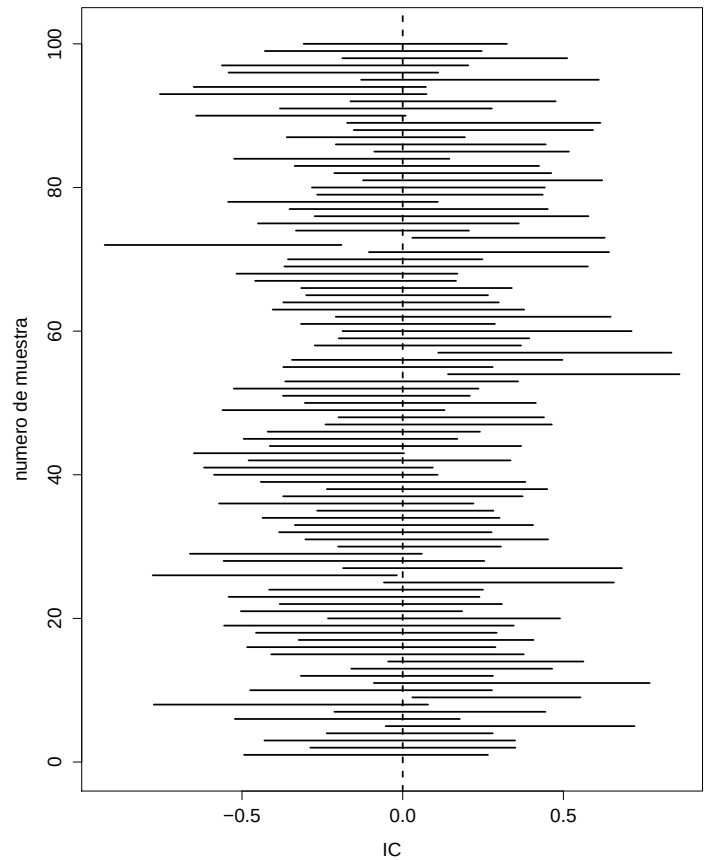
```
norm.interval = function(datos, varianza = var(datos),
  nivel.conf = 0.95)
{
  z = qnorm((1 - nivel.conf)/2, lower.tail = FALSE)
  m = mean(datos)
  dt = sqrt(varianza/length(datos))
  c(m - z * dt, m + z * dt)
}

source("norm.interval.R")
X = rnorm(50,0,1)
norm.interval(X)
[1] -0.2566292  0.4148183
norm.interval(X,1)
[1] -0.1980862  0.3562753
```

Podemos muestrear 100 intervalos de confianza y dibujarlos:

```
nMC = 100 ; n = 30
mu = 0 ; sigma = 1
muestras = matrix(rnorm(nMC * n,mu,sigma),n)
int.conf = apply(muestras,2,norm.interval,varianza=1)
sum(int.conf[1,] <= mu & int.conf[2,] >= mu)
[1] 94
```

```
plot(range(int.conf), c(0, 1+nMC),
  type = "n", xlab = "IC",
  ylab = "numero de muestra")
for (i in 1:nMC) {
      lines(int.conf[, i], rep(i,2),
    lwd=2)
      }
abline(v = 0, lwd = 2, lty = 2)
```



Intervalos de confianza "asintóticos" basados en el TCL

El intervalo de confianza para la media de una normal (con σ conocida)

$$\text{IC}_{0.95}(\mu) = \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

se deducía inmediatamente de la propiedad

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{1}$$

Por el TCL, el resultado (1) es cierto aproximadamente (cuando n es "grande") cualquiera que sea la distribución de las X_i, siempre que $\mathbb{V}(X) < \infty$. Por tanto se tiene, para n suficientemente grande,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{aprox.}{\sim} N(0,1). \tag{2}$$

Sustituyendo σ por un estimador consistente $\hat{\sigma}$ se tiene una nueva aproximación

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \overset{aprox.}{\sim} N(0,1), \tag{3}$$

de la que se obtiene el siguiente intervalo de confianza para $\mu = \mathbb{E}(X)$ con nivel aproximado $1 - \alpha$

$$\left(\bar{x} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

Este intervalo es (aproximadamente) válido, para cualquier distribución, siempre que n sea lo bastante grande.

Una aplicación importante: Intervalo de confianza (aproximado) para una proporción p

Sean $X_1, \ldots, X_n$ iid Bernoulli(p). Por el TCL

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{aprox.}{\sim} N(0,1)$$

y reemplazando p por su estimador natural $\hat{p} = \bar{X}$, obtenemos que el intervalo de confianza aproximado para p es,

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right). \qquad (4)$$

Ejemplo: Se estima la proporción p de piezas defectuosas en la producción de una fábrica con una muestra de 200 piezas de las cuales 8 resultan ser defectuosas. Obtener un intervalo de confianza de nivel 0.95 para p.

Sustituyendo en (4) obtenemos

$$\begin{aligned} IC_{0.95}(p) = \left(\frac{8}{200} \pm 1.96 \sqrt{\frac{0.04 \cdot 0.96}{200}} \right) &= (0.04 \pm 0.02716) \\ &= (0.01284, 0.06716). \end{aligned}$$

Supongamos que este "error de estimación" (la mitad de la longitud del IC) se considera insatisfactorio y se desea obtener un intervalo con un error de, como mucho, 0.01. ¿Qué tamaño muestral habría que elegir?

Debemos tener

$$1.96 \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq 0.01$$

Ejemplo (cont.): Como valor de $\bar{x}$ podemos tomar (a modo de aproximación) el obtenido en la muestra anterior. Entonces

$$1.96\sqrt{\frac{0.04 \cdot 0.96}{n}} \leq 0.01$$

Despejando, obtenemos $n = 1.96^2 \left(\frac{0.04 \cdot 0.96}{0.01^2} \right) = 1475.17$. Por tanto, habría que tomar $n \geq 1476$.

Cuando se quiere determinar el tamaño muestral necesario para obtener un error ϵ y no se tiene ninguna información previa sobre el valor de p se puede actuar "poniéndose en el caso peor" (es decir, en el que da un intervalo de confianza más amplio) que es $p = 1/2$. En el ejemplo anterior se tendría

$$n = 1.96^2 \left(\frac{0.5 \cdot 0.5}{0.01^2} \right) = 9604.$$

La distribución χ^2

Estamos interesados en obtener intervalos de confianza exactos, válidos para cualquier n, para σ^2 en una normal.

Para ello presentamos una distribución auxiliar que tiene una especial importancia en estadística, la distribución χ^2.

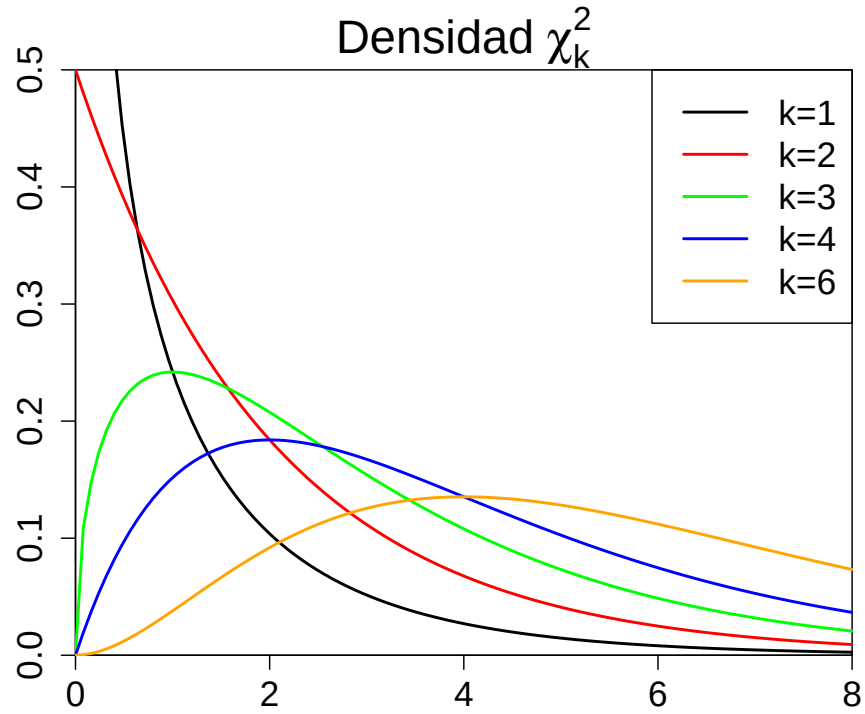
En realidad la distribución χ^2_k (distribución ji-cuadrado con k grados de libertad) es la distribución $\gamma(1/2, k/2)$.

La densidad de una v.a. Y con distribución de probabilidad χ^2_k es

$$g(y;k) = \frac{1}{2^{k/2}\Gamma(k/2)} e^{-\frac{y}{2}} y^{\frac{k}{2}-1},$$

donde $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$.

$$\mathbb{E}(\chi^2_k) = k \qquad \mathbb{V}(\chi^2_k) = 2k$$



La función característica de la χ^2_k es

$$\phi(t) = \mathbb{E}(e^{itY}) = \int_{\mathbb{R}} e^{ity} g(y;k)dy = (1-2it)^{-k/2}.$$

Se puede probar que, si $Z_1, \ldots, Z_n$ son vaiid con distribución $N(0,1)$, entonces

$$Z_1^2 + \ldots + Z_n^2 \sim \chi^2_n.$$

La distribución de S^2 en una $N(\mu, \sigma)$: intervalo de confianza para σ^2

Se puede demostrar que, si $X_1, \ldots, X_n$ son vaiid $N(\mu, \sigma)$ y

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1},$$

entonces

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Este resultado proporciona directamente una cantidad pivotal y, en consecuencia, un intervalo de confianza de nivel $1-\alpha$ para σ^2:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} \right),$$

donde $\chi^2_{k;\beta}$ denota el valor que "deja a la derecha" una probabilidad β en la distribución χ^2_k.

Ejemplo: Se tomaron las tensiones sanguíneas de una muestra aleatoria de 10 pacientes hipotensos, obteniéndose las mediciones:

10 10.5 11 10.7 10.8 12 11.5 9.1 11.3 9.9 .

Suponiendo una distribución normal de las tensiones en la población de hipotensos observada, hallar un intervalo de confianza al nivel del 90% para la varianza σ^2 de esta población.

```
var.interval = function(datos, nivel.conf = 0.95) {
gl = length(datos) - 1
chiinf = qchisq((1 - nivel.conf)/2, gl)
chisup = qchisq((1 - nivel.conf)/2, gl, lower.tail=FALSE)
v = var(datos)
c(gl * v/chisup, gl * v/chiinf)
}
```

```
X = c(10 , 10.5 , 11 , 10.7 , 10.8 , 12 , 11.5 , 9.1 ,
  11.3 , 9.9)
source("var.interval.R")
var.interval(X,nivel.conf=0.9)
[1] 0.3851297 1.9596327
```

La distribución t de Student

Sea $Z \sim N(0,1)$ y $W \sim \chi^2_k$. Supongamos que Z y W son independientes. La distribución de la v.a.

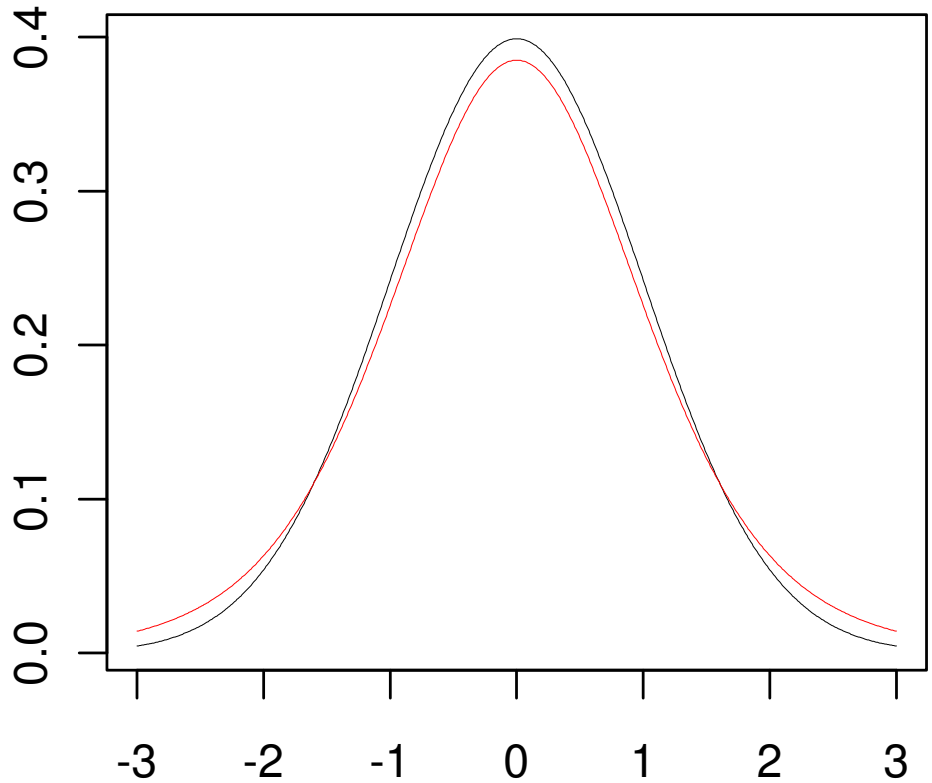
$$T = \frac{Z}{\sqrt{W/k}}$$

se denomina t de Student con k grados de libertad, t_k. La función de densidad de esta distribución es

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}\Gamma(k/2)} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$$

Demostración de este resultado en p. 223 de Casella y Berger.

La gráfica tiene una forma similar a la de la $N(0,1)$ pero con las colas "más pesadas". Para valores grandes de k ($k \geq 50$) ambas distribuciones son casi idénticas.



La figura muestra la densidad de la t_7 (en rojo) y la de la $N(0,1)$ (en negro).

LEMA DE FISHER-COCHRAN.- *Si $X_1, \ldots, X_n$ son v.a.i.i.d. con distribución $N(\mu, \sigma)$ entonces $\bar{X}$ y S^2 son estadísticos independientes*.

La demostración se puede encontrar en la p. 218 del libro de Casella y Berger. Se basa en el hecho de que $\bar{X}$ y el vector aleatorio $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$ son independientes (lo cual se demuestra a su vez calculando la función característica del vector aleatorio $(\bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$.)

Una consecuencia importante: intervalo de confianza exacto para μ en $N(\mu, \sigma)$ cuando σ es desconocida

Sea $X_1, \ldots, X_n$ una muestra de una distribución $N(\mu, \sigma)$ con σ desconocida. En virtud del Lema de Fisher-Cochran, se tiene

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Tenemos, por tanto, una cantidad pivotal para la media μ que lleva de inmediato al siguiente intervalo de confianza de nivel $1 - \alpha$:

$$IC_{1-\alpha}(\mu) = \left(\bar{x} - t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} \right),$$

donde $t_{n-1;\alpha/2}$ representa el valor que "deja a la derecha" una probabilidad de $\alpha/2$ en la distribución t de Student con $n - 1$ grados de libertad.

Ejemplo: El fichero `tortugas.txt` contiene medidas del caparazón de tortugas pintadas (*Chrysemys picta marginata*), 24 hembras y 24 machos. Los datos (de Jolicoeur y Mosimann 1960) son una tabla con 48 observaciones de las variables:

Longitud (en mm.) del caparazón
Anchura (en mm.) del caparazón
Altura (en mm.) del caparazón
Género (hembra $= 0$, macho $= 1$)

Suponiendo normalidad de la variable "Altura" en las hembras, obtener un intervalo de confianza de nivel 0.95 para estimar la esperanza de esta variable. Obtener también un intervalo de confianza de nivel 0.90 para la varianza.

```
Datos = read.table("tortugas.txt",header=T)
Hembras = (Datos$Sexo==0)
Altura = Datos$Altura
AlturaH = Altura[Hembras]
```

```
mean(AlturaH)
[1] 52.04167
var(AlturaH)
[1] 64.73732

var.interval(AlturaH,nivel.conf=0.9)
[1]  42.33307 113.74330

t.test(AlturaH,conf.level=0.95)

            One Sample t-test

data:  AlturaH
t = 31.687, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
48.64416 55.43917
sample estimates:
mean of x
52.04167
```

Referencias

Casella, G., Berger, R.L. (2002). *Statistical Inference*. Second Edition. Duxbury. Thomson Learning. Capítulo 9.

Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall/CRC. Capítulo 7.