Muestreo e intervalos de confianza

Estimar la varianza.
Intervalos de confianza para la media: varianza desconocida, muestras pequeñas.
Intervalo de confianza para la varianza.

Grado Biología sanitaria

M. Marvá.
Departamento de Física y Matemáticas.
UAH November 1, 2017

Sabemos estimar la media (IC) si conocemos la varianza y n>30, pero

- ¿Y si no conocemos σ?
- ¿Y si si $n < 30$?

Candidato natural: calcular la varianza de la muestra

Dada una muestra $X_1, \cdots, X_n$, calcular la media muestral $\bar{X}$ y

$$Varianza(X_1, \cdots, X_n) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\mathbf{n}}$$

Qué esperar

- $Varianza(X_1, \cdots, X_n) \approx \sigma$
- Como $Varianza(X_1, \cdots, X_n) \neq \sigma$, que **aproximadamente**
  - para la mitad de las muestras $Varianza(X_1, \cdots, X_n) > \sigma$
  - para la otra mitad de las muestras $Varianza(X_1, \cdots, X_n) < \sigma$

Ver simulaciones R

Algo de terminología:

- Una **muestra aleatoria simple (m.a.s.)** es aquella en la que todos los datos tienen la misma probabilidad de aparecer
- Un **estimador** es un estadístico (una función de la muestra) usado para estimar un parámetro desconocido de la población.
  - Es **insesgado** si su valor esperado es el valor de parámetro que estima
  - En caso contrario, se dice **sesgado**

Ejemplo:

- $\bar{X}$ es un estimador insesgado de la media poblacional
- $Var(X)$ es un estimador insesgado de la varianza poblacional
- $Varianza(X_1, \cdots, X_n)$ es un estimador sesgado de la varianza poblacional

Dada una m.a.s de la variable X, de tamaño n:

$$X_1, \ldots, X_n$$

se define la **cuasivarianza** (o **varianza muestral**) como:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\mathbf{n - 1}}.$$

Y la **cuasidesviación típica** (o **desviación típica muestral**) es la raíz cuadrada s de la cuasivarianza muestral.

$$E[s^2] = \sigma_X^2$$

Sea X una v.a. **cualquiera** de varianza desconocida. Si tomamos muestras de tamaño $\mathbf{n} > \mathbf{30}$, entonces

$$\bar{X} \sim N\left(\mu_X, \frac{\mathbf{s}}{\sqrt{n}}\right)$$

tiene una distribución aproximadamente normal. Además el intervalo de confianza al nivel $nc = (1-\alpha)$ para la media μ_X es:

$$\left(\bar{X} - z_{\alpha/2}\frac{\mathbf{s}}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\mathbf{s}}{\sqrt{n}}\right)$$

que también se escribe como

$$\mu_X = \bar{X} \pm z_{\alpha/2}\frac{\mathbf{s}}{\sqrt{n}}.$$

Se llama **error muestral** a la desviación típica de la media muestral: $s/\sqrt{n}$ (o $\sigma/\sqrt{n}$)

Ejemplo: Se ha medido la presión sanguínea (en mmHg) a 40 habitantes de Framingham. Determina el intervalo de confianza para la media con un nivel de confianza del 0.95%

197.5	122.0	160.0	121.0	116.0	126.0	144.5	123.0
107.0	130.0	126.0	121.5	100.0	106.0	124.0	143.0
159.0	168.0	101.5	131.0	132.5	129.0	146.0	117.5
163.0	127.0	145.0	102.0	181.5	118.5	122.0	112.5
119.0	150.0	123.5	118.0	150.0	114.0	120.0	168.0

Usando, por ejemplo, R: $\bar{X} = 132.15$, $s = 22.57$, $z_{0.025} = 1.96$

$$\left(132.15 - 1.96\frac{22.57}{\sqrt{40}},\ 132.15 + 1.96\frac{22.57}{\sqrt{40}}\right) = (125.2,\ 139.1)$$

La precisión de la estimación es

$$1.96\frac{22.57}{\sqrt{40}} \approx 7$$

Ejemplo (continuación): $\bar{X} = 132.15$, $s = 22.57$, $z_{0.025} = 1.96$. Determina el tamaño de la muestra para que la precisión de la estimación sea inferior a 3 unidades queremos que

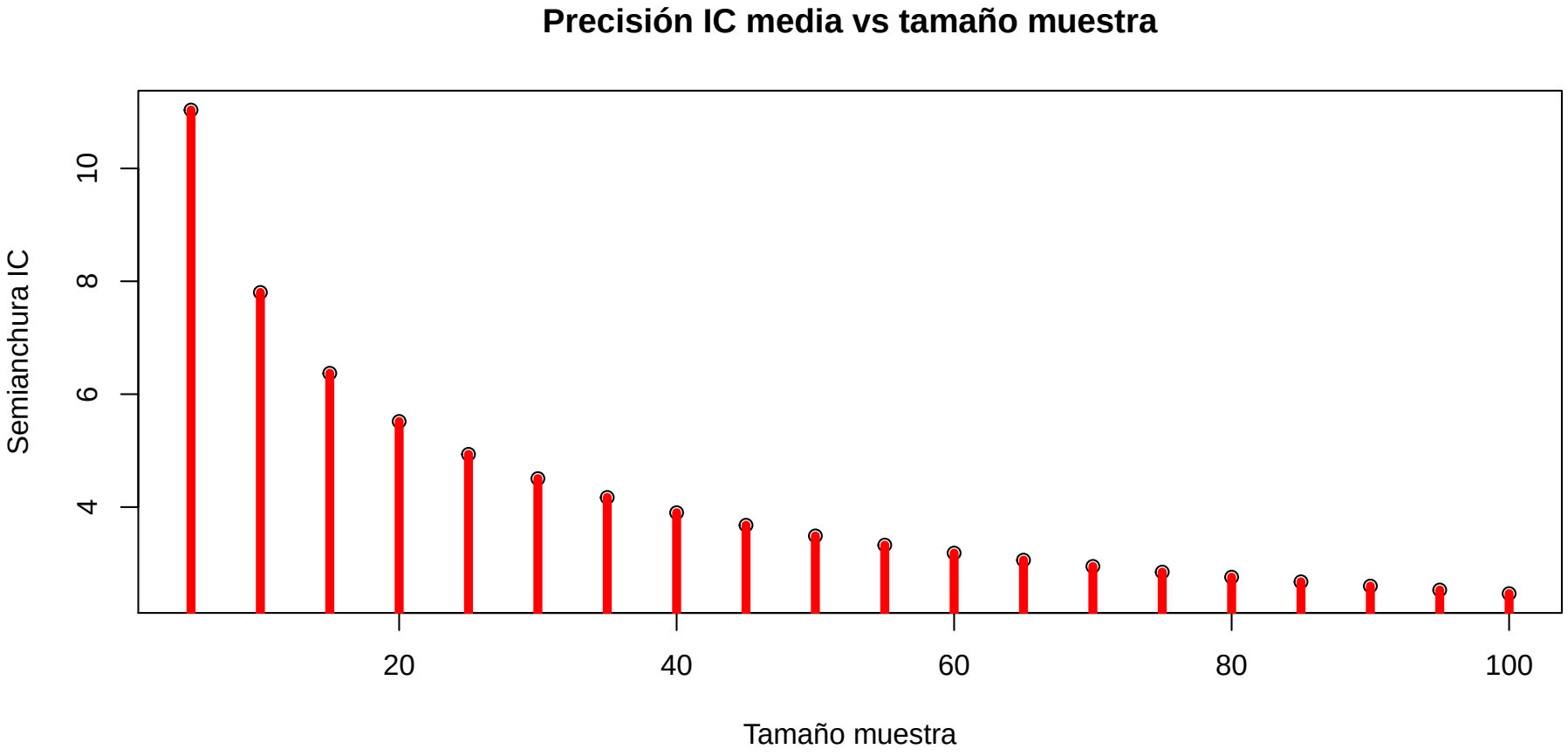
$$1.96\frac{22.57}{\sqrt{n}} < 3 \Leftrightarrow 1.96\frac{22.57}{3} < \sqrt{n}$$

es decir

$$\left(1.96\frac{22.57}{3}\right)^2 = 217.4367 < n$$

Redondear!!

Sección 6.2.4 del libro



Sección 6.2.4 del libro

Recapitulando:

- Si conocemos σ y $n > 30$

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

- Si no conocemos σ, para cualquier v.a. X y muestras de $n > 30$

$$\bar{X} \sim N\left(\mu_X, \frac{s}{\sqrt{n}}\right)$$

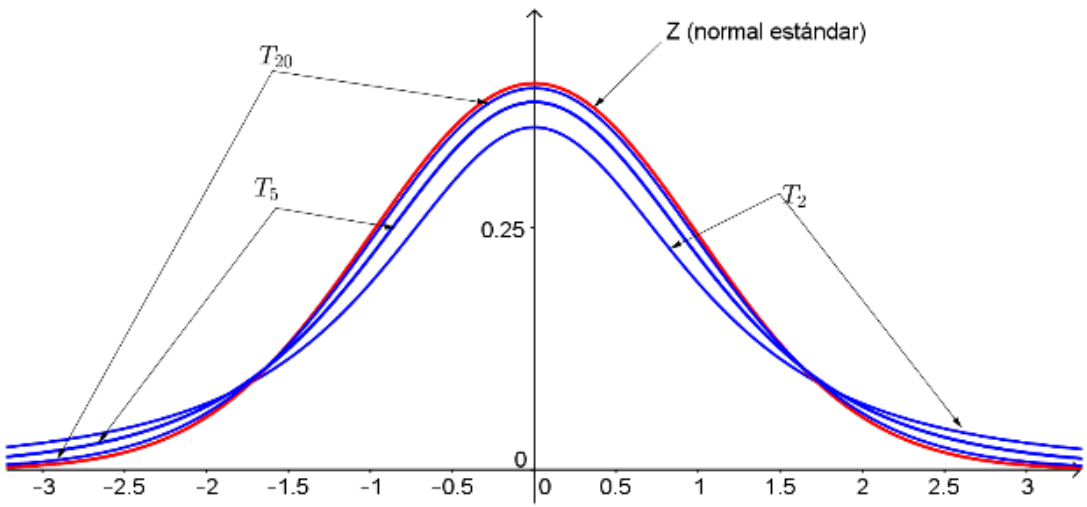
- ¿Y si las muestras son pequeñas ($n < 30$)?

Sean $X \sim N(\mu_X, \sigma_X)$ y $\bar{X}$ la media muestral de X para muestras de tamaño n. Entonces, la distribución de la variable aleatoria

$$t_{n-1} = \frac{\bar{X} - \mu_X}{s/\sqrt{n}},$$

sigue una **distribución t de Student con $n-1$ grados de libertad**.

Observa que $\lim_{n\to\infty} t_{n-1} = N(0,1)$



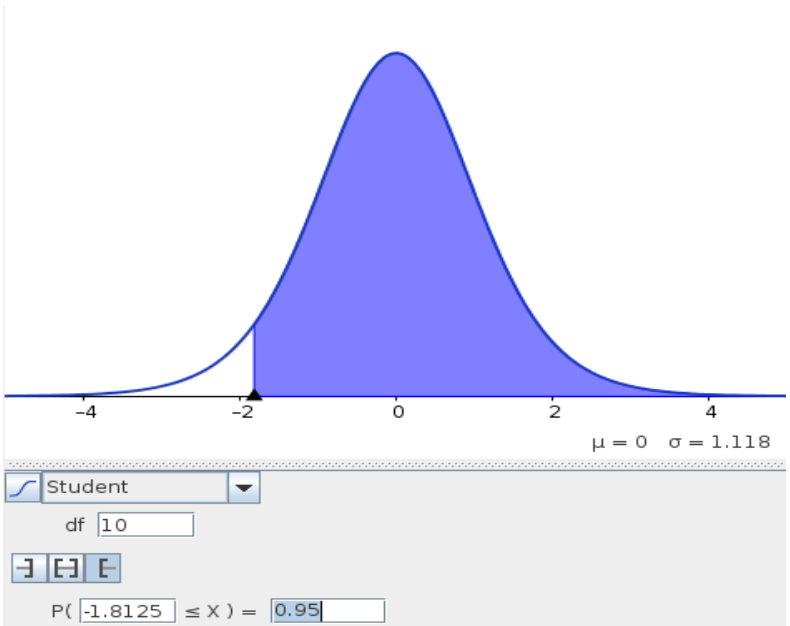
Puedes encontrar su función de distribución en la sección 6.4 del libro

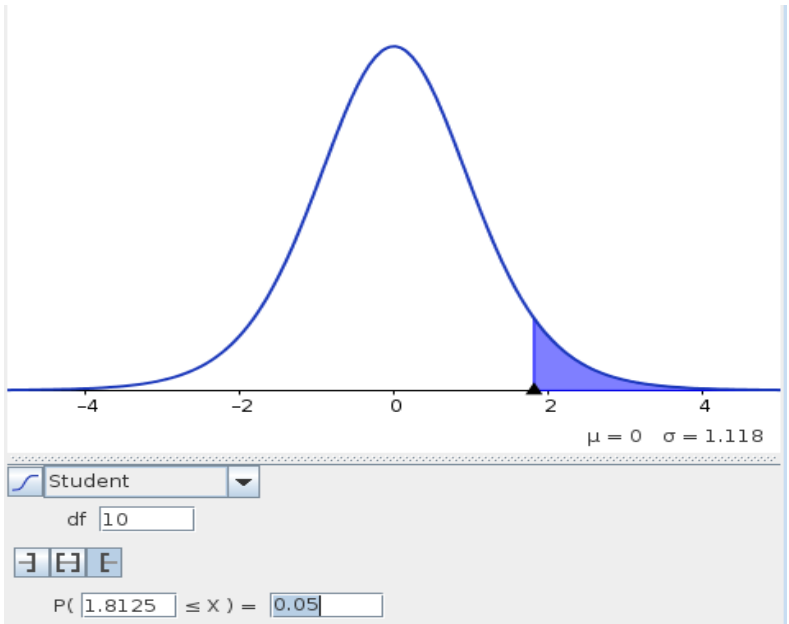
Valores críticos $t_{k;\alpha/2}$ de distribución t de Student

Sean $0 \leq p \leq 1$ y $k \in \mathbb{N}$. El **valor crítico** de la distribución t de Student con k grados de libertad asociado a p es el valor $t_{k;p}$ tal que:

$$P(t_{n-1} \leq t_{k;p}) = 1 - p.$$

$t_{k;p}$ deja una probabilidad p en su cola derecha ($1 - p$ en la cola izda)





$t_{0.95} = -1.8125$ 10 grados libertad $t_{0.05} = 1.8125$

I.C. para la media: muestras normales y σ^2 desconocida

Sea $X \sim N(\mu_X, \sigma_X)$ con varianza σ_X^2 desconocida. Para muestras de tamaño n el intervalo de confianza al nivel de confianza $nc = (1 - \alpha)$ para la media μ_X es:

$$\left(\bar{X} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right)$$

- La **semianchura** del intervalo (precisión) $t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$
- El **error estándar** de la muestra es $\frac{s}{\sqrt{n}}$

Ejemplo: Se ha medido las lipoproteinas de baja densi0dad (LDL) (en mg/dL) a 20 habitantes de Framingham. Determina el intervalo de confianza para la media con un nivel de confianza del 90% suponiendo que la concentración de LDL se distribuye de forma normal

190	157	172	130	266	193	185	170	183	152
212	97	200	162	158	132	203	111	164	244

Usando, por ejemplo, R: $\bar{X} = 174.05$, $s = 41.06$, $t_{19,0.05} = 1.73$

$$\left(174.05 - 1.73\frac{41.06}{\sqrt{20}},\ 174.05 + 1.73\frac{41.06}{\sqrt{20}}\right) = (158.1737,\ 189.9263)$$