Inferencia Estadística

Estimación por intervalos

Gabriel Martos Venturini

gmartos@utdt.edu

UTDT

Outline

Outline

Motivación

- Modelo estadístico $\mathcal{F} = \{f(x;\theta), \theta \in \Theta\}$ (desconocemos θ).
- $\{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta) \rightarrow$ Estimador $\rightarrow$ Estimación puntual.
  - Los EMV tienen propiedades: La consistencia garantiza que para $n \gg 0$ luego $P(|\widehat{\theta}_n - \theta| \geq \varepsilon) \approx 0$ (eficiencia y normalidad asintóticas).
  - En general acompañamos las estimaciones puntuales con alguna cuantificación de la incertidumbre (por ejemplo el error estandard).
  - Aún así, y en general para todo $n < \infty$, $P(\widehat{\theta}_n = \theta) = 0$.
- *Estimación por intervalos*: En vez de reportar una estimación puntual, construimos un intervalo de valores plausibles para el parámetro que resultan compatibles con los datos de la muestra $\underline{x} = \{x_1, \ldots, x_n\}$.

$$[\widehat{\theta}_n - \text{algo}, \widehat{\theta}_n - \text{algo}]$$

  - La longitud del intervalo cuantifica la incertidumbre respecto de θ.

Intervalo de confianza

- Sea $\underline{X} \equiv \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$ con $\theta \in \Theta$.
- Si dos estadísticos $L_n(\underline{X})$ y $U_n(\underline{X})$ (con $P(L_n \leq U_n) = 1$) verifican[1]:

$$P_\theta\left(L_n \leq \theta \leq U_n\right) \geq 1 - \alpha \text{ para todo } \theta \in \Theta,$$

entonces decimos que $[L_n, U_n]$ es un **intervalo** de *nivel* $1 - \alpha$ para θ.

  - Notar que $[L_n(\underline{X}), U_n(\underline{X})]$ es un **intervalo aleatorio**.
  - En general L_n = estimador de θ - algo y U_n = estimador de θ + algo.

El método de inferencia tiene garantías probabilísticas ex–ante de realizarse la muestra aleatoria. Cada vez que consideramos datos concretos, la probabilidad de que $[l_n, u_n]$ contenga a θ será 0 ó 1.

[1] En general va a ocurrir que $P_\theta\left(L_n \leq \theta \leq U_n\right) = 1 - \alpha$ para todo $\theta \in \Theta$.

Interpretando correctamente los intervalos en la práctica

- Si consideramos una gran cantidad de muestras de tamaño n; aproximadamente una fracción $(1-\alpha)$ de los intervalos que construimos (basados en diferentes datos y por tanto diferentes límites l_n y u_n) contendrán al parámetro poblacional de interés.

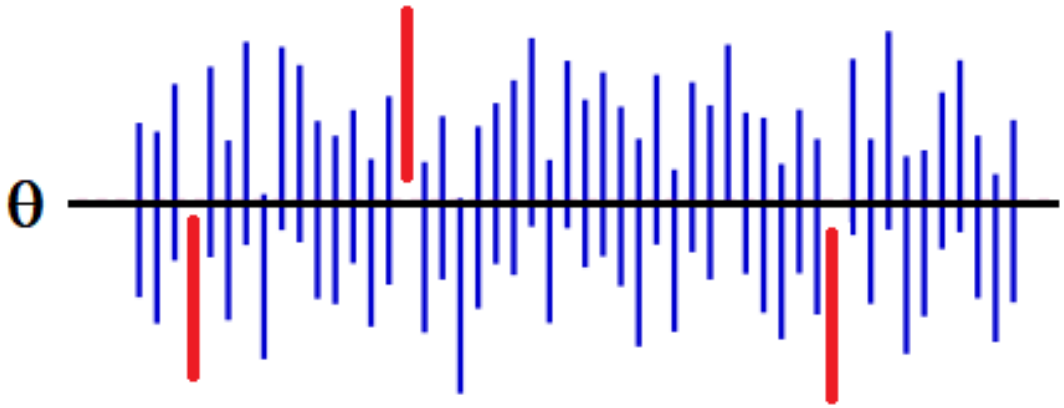


Figure: Ilustración de la interpretación frecuentista de los intervalos.

- En INCORRECTO afirmar (interpretar): "... con los datos de la muestra y el respectivo intervalo de confianza $[l_n, u_n]$ computado con los mismos, la probabilidad de que $\theta \in [l_n, u_n]$ es igual a $1-\alpha$".

Intervalos laterales (cotas de confianza)

- Así como definimos intervalos bilaterales, también podríamos estar interesados en intervalos unilaterales o cotas de confianza:
  - Intervalo de confianza $1-\alpha$ por la izquierda / cota inferior:

$$[L_n(\underline{X}), \infty) \Rightarrow P\big(L_n(\underline{X}) \leq \theta\big) \geq 1-\alpha, \ \forall \theta \in \Theta.$$

    - ★ L_n es el límite *inferior* de confianza de nivel $1-\alpha$.
  - Intervalo de confianza $1-\alpha$ por la derecha / cota superior:

$$(-\infty, U_n(\underline{X})] \Rightarrow \ P\big(\theta \leq U_n(\underline{X})\big) \geq 1-\alpha, \ \forall \theta \in \Theta.$$

    - ★ U_n es el límite *superior* de confianza de nivel $1-\alpha$.
- La interpretación en la práctica de estos intervalos laterales es similar respecto del caso bilateral discutido en la transparencia anterior.

Outline

Definición

Sea $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$, $g_n : \underline{X} \times \Theta \to \mathbb{R}$ es un **pivote** si:

1. Para cualquier $\theta \in \Theta$ fijo, $g_n(\underline{X}, \theta)$ es continua.
2. $P[g_n(\underline{X}, \theta) \leq c]$ no depende de θ.

- Generalmente los pivotes son funciones continuas respecto de algún estadístico / estimador suficiente: $g_n(\underline{X}, \theta) \equiv g(T_n(\underline{X}), \theta)$.
- Remark: Un pivote **NO es un estadístico** ya que depende de θ.
  - Para computar el valor de $g_n(\underline{X}, \theta)$ necesito darle un valor a θ; sin embargo la distribución de la v.a. $g_n(\underline{X}, \theta)$ NO depende de θ.
- Ejemplo[2]: $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu, \sigma_0^2)$.

[2] Recuerda que si $X \sim N(\mu, \sigma^2)$ entonces $\overline{X}_n \sim N(\mu, \sigma^2/n)$.

Método pivotal

- Dado $g_n(\underline{X}, \theta) \sim F_g$ y un nivel de confianza $1 - \alpha \in (0, 1)$.
- Determinamos[3] $F_g(q_{1-\alpha/2}) = \alpha/2$ y $F_g(q_{\alpha/2}) = 1 - \alpha/2$, tal que:

$$P\left(q_{1-\alpha/2} \leq g_n(\underline{X}, \theta) \leq q_{\alpha/2}\right) = 1 - \alpha.$$

- Por definición, la probabilidad anterior no depende de θ.
- El intervalo de confianza queda determinado como:

$$\text{IC}_{1-\alpha}(\theta) = \{\theta \in \Theta : \ q_{1-\alpha/2} \leq g_n(\underline{X}, \theta) \leq q_{\alpha/2}\}$$

- En la mayoría de los problemas prácticos podrás resolver el sistema de desigualdades y obtener de forma analítica el intervalo de confianza.
  - Ejemplo I: Pivote para el modelo Normal.
  - Ejemplo II: Pivote para el modelo Uniforme.

[3] Entendemos por cuantil α–th aquel valor q_α que verifica: $P(X \geq q_\alpha) = \alpha$.

Outline

- Media μ de una población con distribución normal.
  - Con varianza σ^2 conocida (y discutimos "precisión").
  - Con varianza σ^2 desconocida.
- Varianza σ^2 de una población con distribución normal.
- Diferencia de medias en poblaciones normales:
  - Con varianzas conocidas (poco útil).
  - Con varianzas desconocidas pero iguales.

... todos estos resultados surgen de utilizar pivotes (exactos), los discutimos a partir de ejemplos prácticos (algunos en R).

Intervalo para μ en población $N(\mu, \sigma_0^2)$

$$P(\overline{X}_n + z_{1-\alpha/2}\,\sigma_0/\sqrt{n} \leq \mu \leq \overline{X}_n + z_{\alpha/2}\,\sigma_0/\sqrt{n}) = 1-\alpha.$$

- Luego[4] $\text{IC}_{1-\alpha}(\mu) = [\overline{X}_n - z_{\alpha/2}\,\sigma_0/\sqrt{n}, \overline{X}_n + z_{\alpha/2}\,\sigma_0/\sqrt{n}]$.
- De una muestra de 25 ofertas de alquiler en el portal `zonaprop`, se estimó que el costo mensual promedio en alquiler por m^2 en el barrio de Nuñez asciende a 19.8 USD con un desvío estimado de 1.2 USD. **Asumiendo** normalidad y que $\sigma_0 = 1.2$, construye un intervalo con $1-\alpha = 0.95$ para el costo medio por m^2 del alquiler en Nuñez.
- **Solución**: De la tabla de la normal tenemos que $z_{\alpha/2} = z_{0.025} = 1.96$.

$$19.8 - 1.96\frac{1.2}{\sqrt{25}} \leq \mu \leq 19.8 + 1.96\frac{1.2}{\sqrt{25}},$$

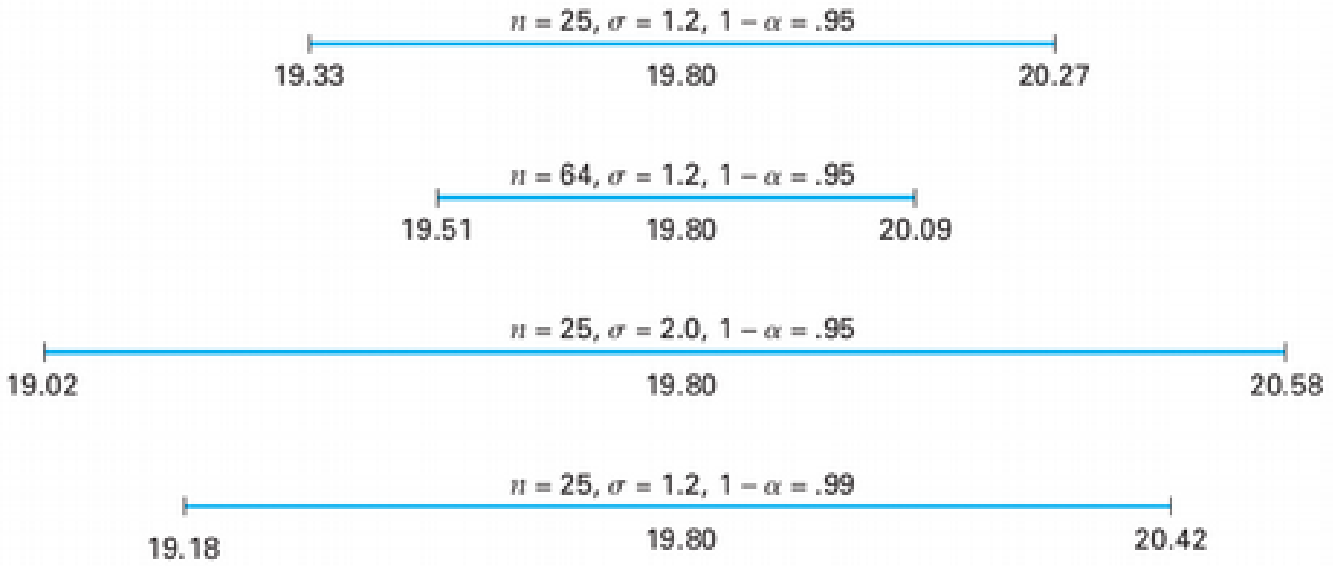
- $\text{IC}_{95\%}(\mu)$: $[19.33, 20.27]$ con un nivel de confianza del 95%.

[4] Notar que $z_{1-\alpha/2} = -z_{\alpha/2}$ (simetría de la normal).

Precisión: $U_n(\underline{X}) - L_n(\underline{X}) = 2z_{\alpha/2}\,\sigma_0/\sqrt{n}$

- (n, α, σ_0^2) determinan la longitud del intervalo (precisión):

Figure 7.6
Effects of Sample Size, Population Standard Deviation, and Confidence Level on Confidence Intervals



- La precisión (en éste caso) no depende de $\underline{X}$.
- Si el intervalo es extenso, la inferencia puede ser poco útil.
  - Deberíamos elegir n (dado α y asumiendo/estimando σ^2) antes de tomar la muestra para garantizar una longitud del intervalo específica.

- Asumiendo $\sigma = 1.2$ y $\alpha = 0.05$, determinar el valor de n necesario para garantizar una longitud de intervalo para μ de como máximo 1USD.
- **Solución**:
-

- Interesante notar que (dado n, σ y α) la longitud depende de los cuantiles. Entonces porque no elegir por ejemplo: $z_{1-\alpha/3}$ y $z_{2\alpha/3}$?
  - Notar que el intervalo aleatorio $[\overline{X}_n + z_{1-\alpha/3}\,\sigma/\sqrt{n}, \overline{X}_n + z_{2\alpha/3}\,\sigma/\sqrt{n}]$ (IC no centrado $\overline{X}_n$) también tiene una confianza de $1-\alpha$.
  - ¿Cómo elegir los cuantiles para garantizar máxima precisión?

Theorem (Maximizando la precisión del intervalo)

Siendo $f_g(t)$ la densidad unimodal del pivote $g_n(\underline{X}, \theta)$ con el que construimos el intervalo para θ. Si el intervalo $[a, b]$ satisface que:

1. $\int_a^b f_g(t)dt = 1 - \alpha$.
2. $a \leq m_o \leq b$, *donde* m_o *es la moda de* f_g.
3. $f_g(a) = f_g(b)$.

entonces $[a, b]$ es el intervalo más corto que satisface (1).

- Ej: $X \sim N(\mu, \sigma_0^2)$, luego $g(\overline{X}_n, \mu) = \frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}} \sim f_g(z) = N(0, 1)$

$$P(a \leq \sqrt{n}(\overline{X}_n - \mu)/\sigma_0 \leq b) = 1 - \alpha \Rightarrow \text{Longitud} = (b - a)\sigma_0/\sqrt{n}.$$

- Del resultado anterior $a = z_{1-\alpha/2}$ y $b = z_{\alpha/2}$ ya que se cumple:

$$z_{1-\alpha/2} \leq m_o = 0 \leq z_{\alpha/2} \text{ y } f_g(z_{1-\alpha/2}) = f_g(z_{\alpha/2}).$$

Intervalo para μ en población $N(\mu, \sigma^2)$

a. $\overline{X}_n \sim N(\mu, \sigma^2/n)$ y $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$.

b. $\overline{X}_n$ y S_n^2 son independientes (Basu).

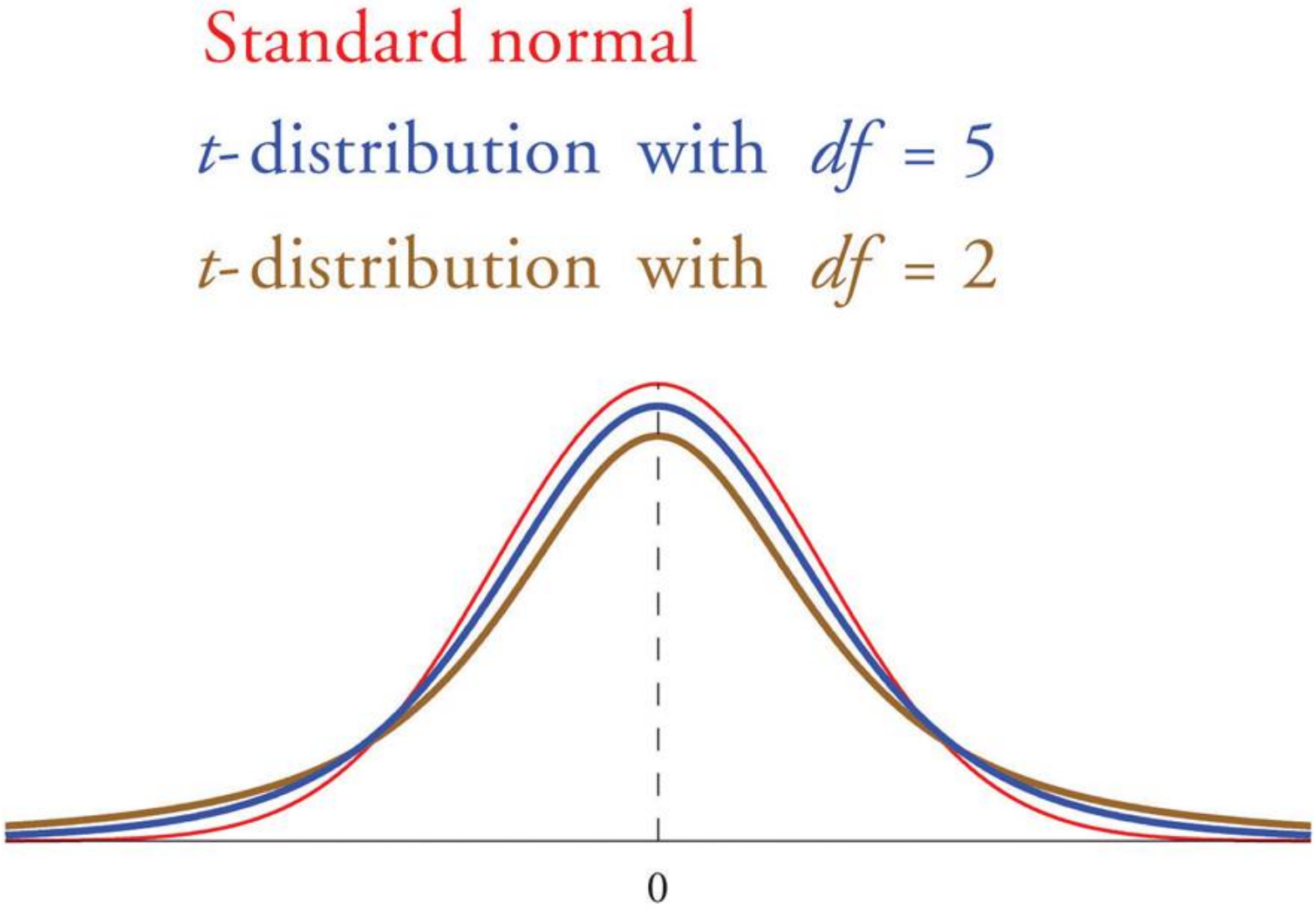
c. $Z \sim N(0,1)$ y $V \sim \chi^2_p$ son independientes, luego $T = Z/\sqrt{V/p} \sim t_p$.

$$g(\overline{X}_n, S_n^2, \mu) = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

- Luego: $P(\overline{X}_n - t_{\alpha/2}\, S_n/\sqrt{n} \leq \mu \leq \overline{X}_n + t_{\alpha/2}\, S_n/\sqrt{n}) = 1-\alpha$.
- Ejemplo: Ídem ejemplo anterior, ahora considerando σ desconocida.

$$19.8 - 2.26\frac{1.2}{\sqrt{25}} \leq \mu \leq 19.8 + 2.26\frac{1.2}{\sqrt{25}},$$

- Menos precisión ya que ahora $\text{IC}_{95\%}(\mu) = [19.26, 20.34]$.



- A medida que crece n, las diferencias con el primer caso son menores.
- Caso práctico en `R`.

Intervalo para σ^2 en población $N(\mu, \sigma^2)$

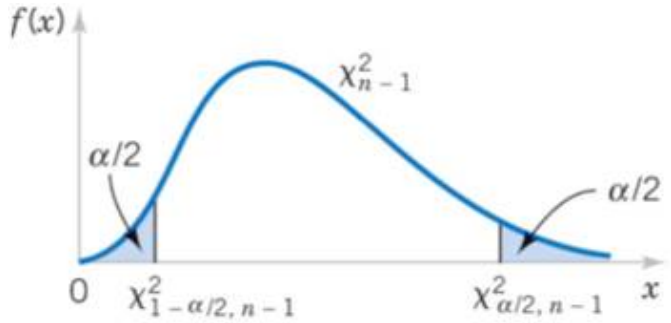
- Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu, \sigma^2)$, se tiene que

$$g(S_n^2, \sigma^2) = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

- Un intervalo aleatorio con un nivel de confianza $1 - \alpha$:

$$L_n(\underline{X}) = \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}} = U_n(\underline{X}).$$

- La distribución χ^2 es unimodal pero asimétrica:



... por lo tanto este intervalo NO maximiza la precisión.

- **Ejemplo**: En un estudio de márketing, 20 sujetos evaluaron la calidad de un producto en una escala de 0 a 100. Asumiendo normalidad[5] calcular un intervalo de confianza al 95% para la varianza de las opiniones en la población sabiendo que de los datos $s_n^2 = (7.5)^2$.

- En `R` puedes computar los cuantiles de χ^2 con: `qchisq(`p`,n-1)`.

```
> qchisq(0.025, 19)
[1] 8.906516
```

[5] Generalmente contrastamos esta hipótesis con un test de normalidad.

Diferencia de medias

- Sean $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ y $\{Y_1, \ldots, Y_m\} \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ dos muestras independientes y donde asumimos que las σ's son conocidas.
- Sea $\Delta = \mu_X - \mu_Y$ el parámetro de interés y $\widehat{\Delta} = \overline{X}_n - \overline{Y}_m$, luego:

$$g(\widehat{\Delta}, \Delta) = \frac{\widehat{\Delta} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1).$$

- Si desconocemos σ pero podemos asumir[6] que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, luego

$$P\left(\widehat{\Delta} - t_{m+n-2,\alpha/2}\, \tilde{S}\sqrt{\frac{1}{n} + \frac{1}{m}} \leq \Delta \leq \widehat{\Delta} + t_{m+n-2,\alpha/2}\, \tilde{S}\sqrt{\frac{1}{n} + \frac{1}{m}}\right) = 1 - \alpha.$$

donde $\tilde{S}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$ y $t_{m+n-2,\alpha/2}$ el cuantil $\alpha/2$ de t_{n+m-2}.

[6] Generalmente usamos un test de igualdad de varianzas.

Brecha de salarios entre mujeres y hombres

- Una consultora encuestó telefónicamente a 120 mujeres y 90 hombres, todos mayores de edad y ocupados en el sector privado.
- Con los datos de la encuesta, los salarios medios estimados ascienden a $\bar{x}_m = 35.600$ y $\bar{x}_h = 43.400$ pesos mensuales respectivamente.
- Los desvíos standard estimados fueron $s_m = 1.300$ y $s_h = 2.200$.
- Asumiendo que los salarios en la población siguen una distribución normal, computar un intervalo de confianza del 95% para $\mu_h - \mu_m$.
  - Haríamos un test para verificar si podemos rechazar $H_0 : \sigma_m^2 = \sigma_h^2$.
  - Asumamos que NO rechazamos este test.
- Como los tamaños de muestra son relativamente grandes, podríamos también utilizar un pivote aproximado (ver discusión en § 31).

Outline

Pivote aproximado

Sea $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$, $g_n : \underline{X} \times \Theta \to \mathbb{R}$ es **pivote aproximado** si:

1. Para cualquier $\theta \in \Theta$ fijo, $g_n(\underline{X}, \theta)$ es continua.
2. $g_n(\underline{X}, \theta)$ tiene una distribución aproximada G que no depende de θ.

$$P[g_n(\underline{X}, \theta) \leq x] \approx G(x).$$

- En general G es una aproximación asintótica.

$$\lim_{n\to\infty} P[g_n(\underline{X}, \theta) \leq x] = G(x).$$

- Los intervalos construidos con g_n tendrán una confianza *aproximada*.
- Más adelante discutimos varios ejemplos con pivotes aproximados.

Pivotes aproximados para EMV

- Recordemos que bajo condiciones bastante generales los EMV son asintóticamente normales, por lo tanto (usando Slutsky):

$$g(\widehat{\theta}_n, \theta) \equiv \sqrt{ni(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta) \to_F N(0,1).$$

- Luego para tamaños de muestra grandes:

$$P\left(\hat{\theta}_n - z_{\alpha/2}\frac{1}{\sqrt{ni(\widehat{\theta}_n)}} \leq \theta \leq \hat{\theta}_n + z_{\alpha/2}\frac{1}{\sqrt{ni(\widehat{\theta}_n)}}\right) \overset{n \gg 0}{\approx} 1-\alpha.$$

- $\text{IC}^{(a)}_{1-\alpha}(\theta) = [\widehat{\theta}_n - z_{\alpha/2}\widehat{\text{se}}_\theta\ ,\ \widehat{\theta}_n + z_{\alpha/2}\widehat{\text{se}}_\theta]$, con $\widehat{\text{se}}_\theta = 1/\sqrt{ni(\widehat{\theta}_n)}$.

- Ejemplo: Intervalo para λ cuando $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Exp}(\theta)$.
  - Caso de estudio II en R.

Otros resultados relevantes con pivotes aproximados

Intervalos para:

- $E(X)$ en muestras grandes (varianza finita).
  - Caso particular: Proporción en muestras grandes.
- Diferencia de medias en muestras apareadas (y grandes).
- Diferencia de medias en muestras grandes.
  - Diferencia de proporciones en dos muestras grandes.

Estos resultados surgen de utilizar el TLC para aproximar la distribución del pivote. Se asume que la muestra es "grande", bajo condiciones razonables para los datos en la población y con $n \geq 30$, las aproximaciones por medio del TCL son razonablemente buenas.

Para $E(X)$ (asumiendo $E(X^4) < \infty$)

- Por el TLC (+ Slutsky) y siendo S_n^2 consistente para Var(X).

$$g(\overline{X}_n, S_n^2, \mu) = \frac{\overline{X}_n - E(X)}{S_n/\sqrt{n}} \to_F N(0,1).$$

- Por lo que un intervalo de confianza aproximado para $E(X)$ será

$$P\left(\overline{X}_n - z_{\alpha/2}\, S_n/\sqrt{n} \leq E(X) \leq \overline{X}_n + z_{\alpha/2}\, S_n/\sqrt{n}\right) \approx 1-\alpha.$$

- De una encuesta a 100 estudiantes de secundaria sobre la cantidad de horas diarias que dedican al uso del teléfono celular se obtuvo que éstos en promedio utilizan $\bar{x} = 5$hs diarias el teléfono con $s = 2$hs.

- Elegimos $\alpha = 0.05$ (entonces $z_{\alpha/2} = 1.96$) luego:

$$\text{IC}^{(a)}_{95\%}(E(X)) = \left[5 - 1.96 \times 2/\sqrt{100}; 5 + 1.96 \times 2/\sqrt{100}\right] = \left[4.61; 5.39\right].$$

Intervalos para la proporción en muestras grandes

- Por el TLC (+ Slutsky) y consistencia del estimador de la varianza:

$$g(\widehat{p}_n, p) = \frac{\widehat{p}_n - p}{\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}} \to_F N(0,1)$$

- Intervalo de confianza de nivel $1-\alpha$ aproximado:

$$P\left(\widehat{p}_n - z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}} \leq p \leq \widehat{p}_n + z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right) \approx 1-\alpha.$$

- Muestra de $n = 200$ pacientes hospitalizados con el mismo diagnóstico y tratamiento. Se observaron complicaciones severas en 38 casos.

- **Solución**: Elegimos $\alpha = 0.01$; luego $z_{0.005} = 2.576$ y:

Diferencia de medias en muestras **apareadas**

- Sea $\{(X_1, Y_1) \ldots, (X_n, Y_n)\}$ una muestra apareada.
- Notar que las muestras NO son independientes.
- Definimos $D_i = X_i - Y_i$ y $\mu_D = E(D) = E(X) - E(Y) = \mu_X - \mu_Y$.
- Asumiendo que $V(D) = \sigma_D^2 < \infty$, luego:

$$g(\overline{D}_n; S_D^2, \mu_D) = \frac{\overline{D}_n - \mu_D}{S_D/\sqrt{n}} \to_F N(0,1).$$

- Por lo que un intervalo de confianza aproximado para μ_D será

$$P\left(\overline{D}_n - z_{\alpha/2}\, S_D/\sqrt{n} \leq \mu_D \leq \bar{D} + z_{\alpha/2}\, S_D/\sqrt{n}\right) \approx 1 - \alpha.$$

- **Nota**: Si $D \sim N(\mu_D, \sigma_D^2)$ tienes un pivote exacto y un intervalo que vale para muestras pequeñas (en ese caso el pivote sigue t_{n-1}).
  - Si la muestra es pequeña tienes que validar el supuesto $D \sim N(\mu_D, \sigma_D^2)$.

Ejemplo en clase

- Seleccionamos $n = 10$ pacientes al azar para un estudio de eficacia de un tratamiento para reducir la cantidad de azúcar en sangre.
- Medimos los niveles de glucosa pre (x_{pre}) y post tratamiento (y_{pos})[7]:

ID	1	2	3	4	5	6	7	8	9	10
$x_{i,pre}$	190	189	196	198	193	186	182	194	184	187
$y_{i,pos}$	193	172	168	181	176	175	178	177	185	169
d_i :	-3	17	28	17	17	11	4	17	-1	18

Table: Medición cantidad de glucosa en sangre (mh/dL) de c/ individuo.

- Corrimos un test de normalidad sobre las mediciones $\{d_1, \ldots, d_{10}\}$ ($d_i = x_{i,ant} - y_{i,post}$) y NO rechazamos la hipótesis de normalidad.
- De los datos: $\overline{d}_n = 12.5$ y $s_d = 9.7 mh/dL$.
- Compute un IC del 95% para el parámetro $\mu_D = E(X_{pre}) - E(Y_{pos})$.

[7] Medidas en mh/dL = miligramo por decilitro.

Diferencia de medias en muestras **independientes**

- $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$ y $\{Y_1, \ldots, Y_m\} \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$ muestras pequeñas e independientes con σ_X^2 y σ_Y^2 desconocidas y diferentes.
- Definimos $\Delta = \mu_X - \mu_Y$ y $\widehat{\Delta} = \overline{X}_n - \overline{Y}_m$, luego:

$$g(\widehat{\Delta}, \Delta) = \frac{\widehat{\Delta} - \Delta}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \to_F t_\upsilon.$$

- Donde (aproximación de Welch)

$$\upsilon = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

- Con $\min(n, m)$ grande no necesitamos asumir normalidad y vale:

$$P\left(\widehat{\Delta} - z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \leq \Delta \leq \widehat{\Delta} + z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}\right) \approx 1 - \alpha.$$

Diferencia de proporciones

- Sean $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Bern}(p_X)$ y $\{Y_1, \ldots, Y_m\} \overset{iid}{\sim} \text{Bern}(p_Y)$.
- Asumimos independencia entre las muestras.
- Nos interesa el parámetro $\Delta = p_X - p_Y$.
- Definimos $\widehat{\Delta} = \widehat{p}_X - \widehat{p}_Y$, luego los límites del intervalo aleatorio para Δ–de nivel aproximado $1 - \alpha$– estarán conformados como:

$$L(\widehat{\Delta}) = \widehat{\Delta} - z_{\alpha/2} \sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n} + \frac{\widehat{p}_Y(1-\widehat{p}_X)}{m}},$$

$$U(\widehat{\Delta}) = \widehat{\Delta} - z_{\alpha/2} \sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n} + \frac{\widehat{p}_Y(1-\widehat{p}_X)}{m}}.$$

- Tal que de esta manera se verifica que:

$$P\big(L(\widehat{\Delta}) \leq \Delta \leq U(\widehat{\Delta})\big) \approx 1 - \alpha.$$

- Para comparar la eficacia de dos tratamientos médicos distintos, se llevan a cabo estudios con pacientes en dos grupos independientes y una junta médica experta juzga si cada paciente en cada grupo fue tratado con éxito o no (si superó o no la dolencia).
- El procedimiento/tratamiento A, resulta eficaz en 63 de entre 91 pacientes tratados con A.
- El procedimiento B, en cambio, es efectivo en 42 de entre las 79 experimentaciones con éste tratamiento.
- Construya un intervalo de confianza al 90% para la diferencia de proporciones e interprete.

Outline

Definition (Región de confianza)

Sea $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x; \boldsymbol{\theta})$, donde $\boldsymbol{\theta} \in \Theta$ es un vector de parámetros desconocidos, decimos que $R_n(\underline{X}) \subset \Theta$ es una región de confianza de nivel $1 - \alpha$ si verifica que (generalmente $R_n(\underline{X})$ tiene forma de elipse):

$$P\left[\boldsymbol{\theta} \in R_n(\underline{X})\right] \geq 1 - \alpha \text{ para todo } \boldsymbol{\theta} \in \Theta,$$

es decir que $R_n(\underline{X})$ **cubre** a $\boldsymbol{\theta}$ con probabilidad (de al menos) $1 - \alpha$.

- Ejemplo 1: Pivote y regiones de confianza en un modelo normal multivariante: Ilustración caso bivariante. Extensión caso general.
  - Misma interpretación que los intervalos (ver § 6).
- Ejemplo 2: Regiones de confianza aproximadas basadas en EMV de los parámetros del modelo cuando $\boldsymbol{\theta} \in \Theta$ y $\dim(\Theta) > 1$.

Normal multivariante (ejemplo 1)

- $\underline{X} \equiv \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \overset{iid}{\sim} \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ y $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$.
- Si $\boldsymbol{\Sigma}$ fuera conocida, el EMV del parámetro de localización es:

$$\widehat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i.$$

- Luego podemos definir el pivote:

$$g(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_n) = n(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) \sim \chi^2_d.$$

- En este caso (poco habitual en la práctica) la región de confianza para el vector de parámetros $\boldsymbol{\mu}$ queda expresado como:

$$R_{1-\alpha}(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : n(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) \leq c_\alpha\},$$

donde c_α es la constante que verifica que $P(\chi^2_d \leq c_\alpha) = 1 - \alpha$.

- En la práctica reemplazamos $\mathbf{\Sigma}$ por un estimador insesgado:

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_n)(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_n)^T.$$

- Se puede demostrar que el pivote (§ KK 7.8):

$$g(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_n) = \frac{n(n-d)}{d(n-1)}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})^T \mathbf{S}_n^{-1}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) \sim F_{n-d}^{d}.$$

- Luego la región de confianza para el vector $\boldsymbol{\mu}$:

$$R_{1-\alpha}(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : \frac{n(n-d)}{d(n-1)}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})^T \mathbf{S}_n^{-1}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) \leq c_\alpha\},$$

  donde c_α es la constante que verifica que $P(F_{n-d}^{d} \leq c_\alpha) = 1 - \alpha$.

- Cuando $n \gg 0$, podemos construir elipses de confianza con el pivote aproximado (sin necesidad de asumir normalidad en los datos):

$$g(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_n) = n(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})^T \mathbf{S}_n^{-1}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) \sim_a \chi_d^2.$$

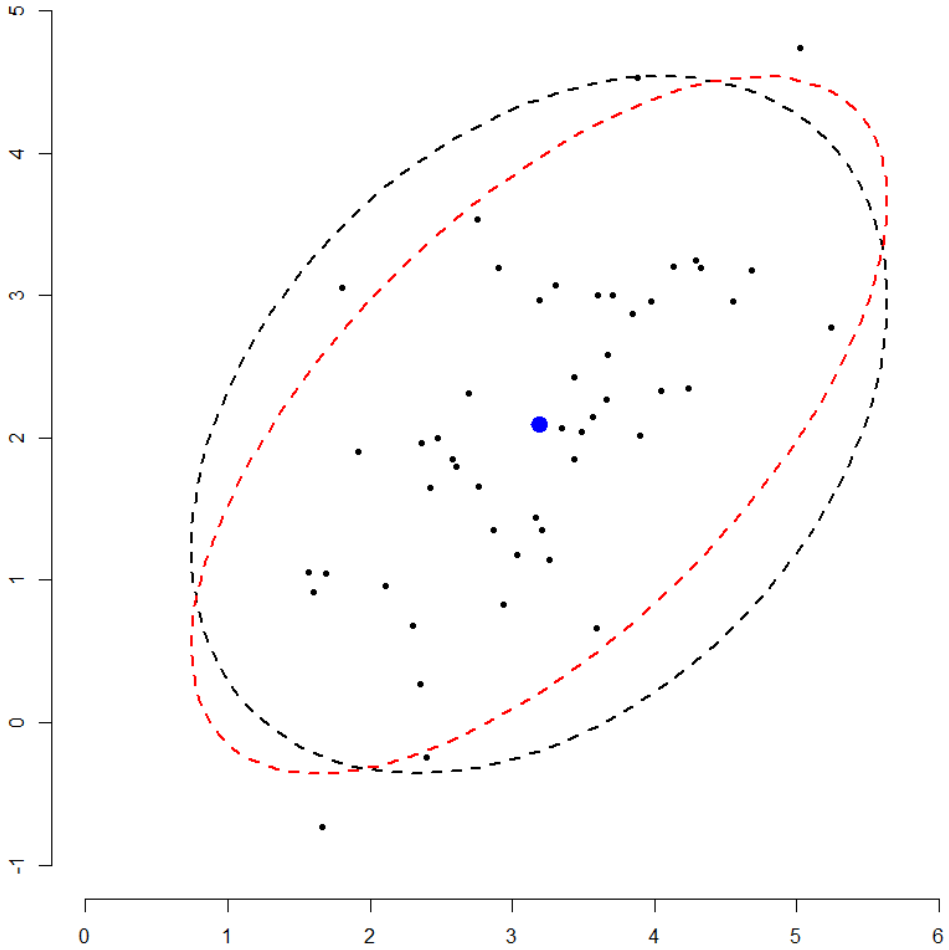


Figure: Datos simulamos de modelo Normal bivariante de parámetros $\boldsymbol{\mu} = (3, 2)$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\rho = 0.5$. El punto azul corresponde a la estimación máximo verosímil de $\widehat{\boldsymbol{\mu}}$, la elipse negra se corresponde con la región de confianza asumiendo Σ conocida y la elipse roja utilizando el estimador insesgado de Σ.

EMV con $d > 1$ (ejemplo 2)

- $2\Big(\ell_n(\widehat{\boldsymbol{\theta}}_n|\underline{X}) - \ell_n(\boldsymbol{\theta}|\underline{X})\Big) \to_F \chi^2_d$, donde $d = \dim(\Theta)$.
- Se puede demostrar la región de confianza:

$$R_{1-\alpha}(\boldsymbol{\theta}) = \{\boldsymbol{\theta} : \ell_n(\boldsymbol{\theta}|\underline{X}) \geq \ell_n(\widehat{\boldsymbol{\theta}}_n|\underline{X}) - \frac{1}{2}c_d(\alpha)\},$$

  donde $c_d(\alpha)$ es la constante que verifica $P(\chi^2_d \leq c_d(\alpha)) = 1 - \alpha$, tiene una probabilidad aproximada de $1 - \alpha$ de cubrir a $\boldsymbol{\theta}$.

- De forma equivalente (expansión de Taylor de por medio) se tiene:

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T \mathbf{I}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to_F \chi^2_d,$$

- Meeker and Escobar: *Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation* (TAS, 2010).

Si $X \sim N(\mu, \sigma^2)$ (d=2)

$$R_{1-\alpha}(\mu, \sigma^2) = \{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : \frac{n(\widehat{\mu}_n - \mu)^2}{\widehat{\sigma}_n^2} + \frac{n(\widehat{\sigma}_n^2 - \sigma^2)^2}{2\widehat{\sigma}_n^4} \leq c_\alpha\}.$$

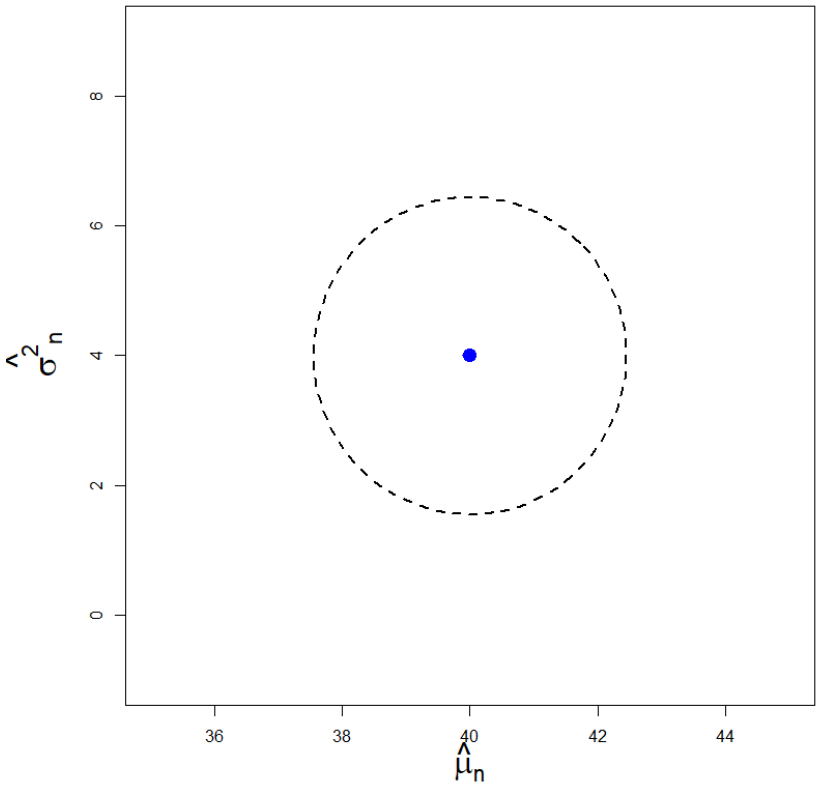


Figure: Con $n = 100$, $\widehat{\mu}_n = 40$, $\widehat{\sigma}_n^2 = 4$ y para $\alpha = 0.05$.