Inferencia Estadística

Riesgo y Nociones de Optimalidad

Gabriel Martos Venturini
gmartos@utdt.edu

UTDT

UNIVERSIDAD TORCUATO DI TELLA

- Un estimador puntual es un estadístico con el que típicamente pretendemos inferir el parámetro desconocido del modelo.
- Idealmente, la distribución de probabilidad del estimador debería estar *concentrada* en torno del verdadero valor del parámetro.
- La que sigue es una discusión sobre cómo cuantificar esta característica deseable de un estimador. La noción de *riesgo* permite definir estimadores en algún sentido óptimos.
  - No es necesario circunscribir la discusión al caso de los estimadores de momentos o los estimadores máximo verosímiles; por este motivo denotaremos con W_n al estimador de interés (función de una muestra aleatoria de tamaño n) de cierto modelo estadístico de referencia.

Agenda

Riesgo de un estimador

- $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$ con $\underline{X} \in \mathcal{X}$ y $\underline{X} \sim f_{\underline{X}}(\underline{x};\theta)$.
- $W_n(\underline{X}) \equiv W_n$ es un estimador de $\theta \in \Theta$.
- Consideramos una función de pérdida $l : \mathcal{X} \times \Theta \to \mathbb{R}$.
  - Pérdida cuadrática: $l(W_n, \theta) = (W_n - \theta)^2$ (descomposición).
  - Pérdida absoluta: $l(W_n, \theta) = |W_n - \theta|$.
  - Debe reflejar las consecuencias del error en la estimación.
- El riesgo del estimador W_n se define como:

$$R(W_n, \theta) = E(l(W_n, \theta)) = \int_{\mathcal{X}} l(w_n(\underline{x}), \theta) f_{\underline{X}}(\underline{x};\theta) \mathrm{d}\underline{x}.$$

- Representa la pérdida esperada (en la métrica l) al estimar θ con W_n.
- Elementos de “Teoría de la Decisión” (CB: § 7.3.4, pp. 348).

Error Cuadrático Medio

- $\mathrm{ECM}(W_n,\theta) = E[(W_n-\theta)^2] \equiv \underbrace{[E(W_n)-\theta]^2}_{\mathrm{Sesgo}^2_\theta(W_n)} + \underbrace{E[(W_n-E(W_n))^2]}_{\mathrm{Var}_\theta(W_n)}$.

Demostración (Intuición).

$$
\begin{aligned}
\mathrm{ECM}(W_n,\theta) &= E[(W_n - E(W_n) + E(W_n) - \theta)^2] \\
&= E[(W_n - E(W_n))^2 + (E(W_n)-\theta)^2 + 2(W_n - E(W_n))(E(W_n)-\theta)] \\
&= E[(W_n - E(W_n))^2] + \underbrace{E[(E(W_n)-\theta)^2]}_{E(cte)=cte} + 2\underbrace{E[(W_n - E(W_n))(E(W_n)-\theta)]}_{(E(W_n)-E(W_n))(E(W_n)-\theta)=0}
\end{aligned}
$$

$\square$

- W_n es **insesgado** si $\mathrm{ECM}(W_n,\theta) = \mathrm{Var}_\theta(W_n)$ (para todo n y $\theta \in \Theta$).
- ECM de los estimadores máximo verosímiles cuando $X \sim N(\mu,\sigma^2)$.
  - Compara el ECM de $\widehat{\sigma}^2_n$ vs el del estimador insesgado S^2_n.

- Nota: Los estimadores insesgados no siempre son los de menor riesgo.

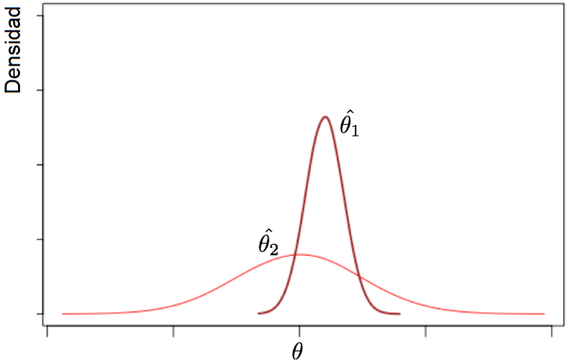


Figure: En algunos contextos específicos puede convenir asumir algo de sesgo a cambio de reducciones contundentes en la varianza del estimador.

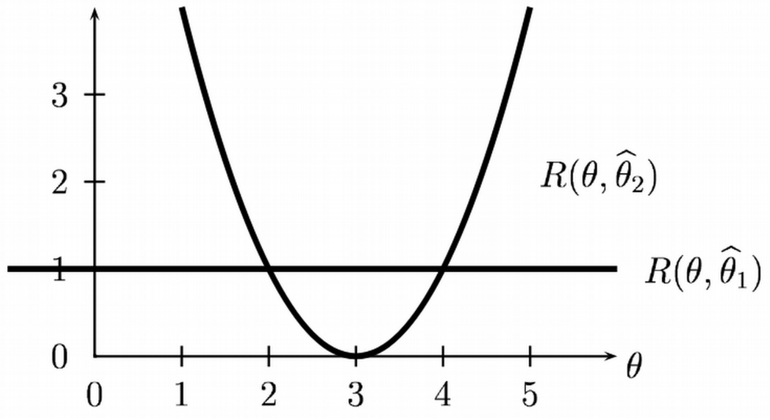
  - En regresión: Ridge y Lasso (intercambiamos sesgo por varianza).

- ¿Existe W_n : $R(W_n, \theta) \leq R(W'_n, \theta)$, para todo $\theta \in \Theta$ y $n \geq 1$?

Contraejemplo

$\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\theta, \sigma^2 = 1)$: $\widehat{\theta}_1 = \overline{X}_n$ vs. $\widehat{\theta}_2 = 3$.

$n = 1:$ $\text{ECM}(\widehat{\theta}_1, \theta) = \text{Var}(\widehat{\theta}_1) = \dfrac{\sigma^2}{n} = 1$, y $\text{ECM}(\widehat{\theta}_2, \theta) = \text{Sesgo}^2(\widehat{\theta}_2) = (\theta - 3)^2$



En general no existe un estimador *uniformemente mejor* (sobre Θ). C2

- Para poder establecer criterios de optimalidad:

1. Acotamos el conjunto de potenciales estimadores:
   - Consideramos únicamente estimadores insesgados del parámetro.
   - Dentro de este (sub)conjunto (o clase), el mejor estimador (menor ECM) será aquel que tenga la menor varianza (mayor eficiencia).
2. Identificamos la curva de riesgo con un número, de forma de poder *ordenar* a los estimadores en base a dicha cantidad.
   - Mini–max: Minimizan el máximo riesgo (criterio conservador).
3. Estimadores de mínimo riesgo Bayesiano.
   - Buscamos el estimador que minimiza la pérdida esperada aposteriori.
   - Postergamos esta discusión hasta discutir métodos Bayesianos.

Agenda

Agenda

Para $X \sim f(x;\theta)$

- Denotaremos $\mathcal{C}_\theta$ al conjunto de todos los estimadores insesgados de θ:

$$\mathcal{C}_\theta = \{W_n \,|\, E(W_n) = \theta \text{ para todo } n \geq 1 \text{ y } \theta \in \Theta\}.$$

- Si $W_n \in \mathcal{C}_\theta$, luego $\mathrm{ECM}(W_n, \theta) = \mathrm{Var}(W_n)$.

Definition (Estimador UMVUE)

$W_n^* \in \mathcal{C}_\theta$ es el **mejor estimador insesgado** de θ (UMVUE) si se cumple que $\mathrm{Var}(W_n^*) \leq \mathrm{Var}(W_n)$ para todo $n \geq 1$, $W_n \in \mathcal{C}_\theta$ y $\theta \in \Theta$.

- UMVUE: Uniform Minimum Variance Umbiased Estimator.
- No siempre existen y encontrar estos estimadores no es trivial.
  - Ejemplo: $\overline{X}_n$ y S_n^2 para λ en modelo Poisson.
- Estrategia: Hallar cota inferior para la varianza de los estimadores de θ, y si un estimador insesgado alcanza la cota entonces es UMVUE.

Condiciones de regularidad (BackUp Slide)

- Se cumplen, en general, en los modelos de la familia exponencial.

H_1) $f(x;\theta)$ es *suficientemente suave* como función de θ, de forma que:

$$\frac{\partial}{\partial\theta}\int f(x;\theta)\mathtt{d}x = \int \frac{\partial}{\partial\theta} f(x;\theta)\mathtt{d}x.$$

- Esta condición no suele cumplirse en los modelos cuyos soportes dependen de los parámetros (por ejemplo el modelo Uniforme).

H_2) Le pedimos al conjunto de estimadores de θ que estemos considerando que sus esperanzas sean funciones diferenciables respecto de θ.

$$\text{Llamemos } m_W(\theta) = E(W_n), \text{ luego pedimos que } \frac{\partial}{\partial\theta} m_W(\theta) \text{ exista.}$$

Resultado general

Teorema (Cota de Cramér–Rao)

Sea $\underline{X} = \{X_1, \ldots, X_n\}$ una muestra (no necesariamente iid) tal que $\underline{X} \sim f_{\underline{X}}(\underline{x}; \theta)$ verifica H_1 y sea W_n cualquier estimador de θ que verifica H_2:

$$V(W_n) \geq \frac{\left(\frac{\partial}{\partial\theta} E(W_n)\right)^2}{E\left(\left[\frac{\partial}{\partial\theta} \log f_{\underline{X}}(\underline{X}; \theta)\right]^2\right)}.$$

- Proof: CB § 7.3.1 pp 336 (Cauchy–Schwarz).
- Si adicionalmente asumimos que
  - $\{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x; \theta)$.
  - $W_n \in \mathcal{C}_\theta$,
- La cota tiene una expresión más simple.

- Si $W_n \in \mathcal{C}_\theta$ entonces $\frac{\partial}{\partial\theta} E(W_n) = \frac{\partial}{\partial\theta}\theta = 1.$

- Si $\{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$, luego:

$$E\Big(\Big[\frac{\partial}{\partial\theta}\log f_{\underline{X}}(\underline{X};\theta)\Big]^2\Big) \overset{i}{=} E\Big(\Big[\frac{\partial}{\partial\theta}\underbrace{\log \Pi_{i=1}^n f(X_i;\theta)}_{\ell(\theta|\mathbf{X})}\Big]^2\Big) \overset{id}{=} n\,\underbrace{E\Big(\Big[\frac{\partial}{\partial\theta}\log f(X;\theta)\Big]^2\Big)}_{i(\theta)}.$$

- Información de Fisher: $I_n(\theta) = ni(\theta)$ (no depende del estimador).

Theorem (Cramér–Rao (iid + $W_n \in \mathcal{C}_\theta$))

Sea $\{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\theta)$ y $W_n \in \mathcal{C}_\theta$ (satisfacen H_1 y H_2), luego:

$$V(W_n) \geq \frac{1}{nE\left(\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right)} = \left[I_n(\theta)\right]^{-1}.$$

- Notar que: $E\left(\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right) = \int_{\mathcal{X}} \left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]^2 f(x;\theta)\mathrm{d}x.$

- Si $W_n \in \mathcal{C}_\theta$ y $Var(W_n) = 1/I_n(\theta)$, entonces W_n es el UMVUE de θ.
    - Ejemplo: Si $X \sim \text{Pois}(\lambda)$, entonces $\widehat{\lambda}_n = \overline{X}_n$ es el UMVUE de λ.
- Notar que:

$$I_n(\theta) = E\Big(\Big[\frac{\partial}{\partial\theta}\underbrace{\log f_{\underline{X}}(\underline{X};\theta)}_{\ell(\theta)}\Big]^2\Big) \overset{iid}{=} n\,\underbrace{E\Big(\Big[\frac{\partial}{\partial\theta}\log f(X;\theta)\Big]^2\Big)}_{i(\theta)}.$$

... qué nos dice la Información de Fisher? Qué relación tiene ésta con la función de (log) versimilitud?

- $I_n(\theta) \equiv ni(\theta)$ (depende del tamaño de la muestra).
- En el ejemplo Poisson: $i(\lambda) = 1/\lambda$. Si λ es pequeño, media y varianza de X es pequeña (la población es homogénea). Cada elemento de $\underline{X}$ aporta mucha información sobre λ (mucha información = cota baja).

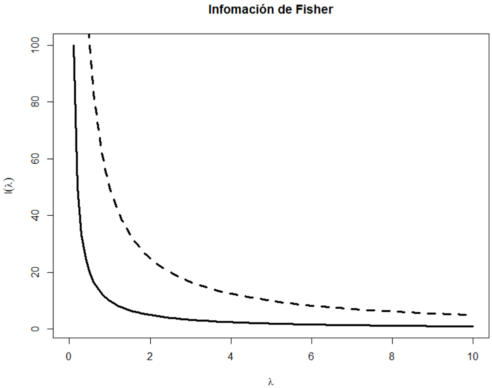


Figure: Información de Fisher para $X \sim \text{Pois}(\lambda)$: En la línea continua (—) $n = 10$ y en línea punteada (- -) $n = 50$. A medida que $\lambda \to 0$, $I_n(\lambda) \to \infty$ para todo n.

$$I_n(\theta) = \text{Var}\left(\frac{\partial}{\partial\theta}\ell(\theta;\underline{X})\right)$$

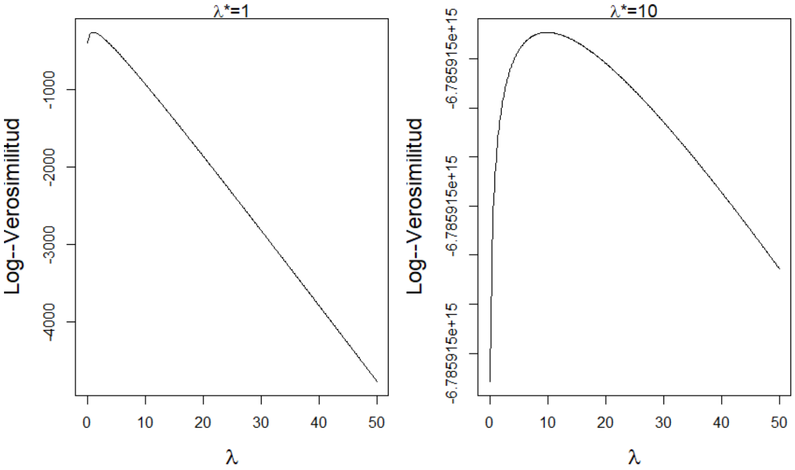


Figure: La información de Fisher cuantifica la concavidad de ℓ en torno a θ^*.

- En la slide anterior ocurre que: $I_n(\lambda = 1) > I_n(\lambda = 10)$.

Familias exponenciales

- Los modelos de la familia exponencial, verifican que:

$$E\left(\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right) = i(\theta).$$

- Luego:

$$I_n(\theta) \overset{iid}{=} -nE\left(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta|\underline{X})\right).$$

- Simplificando cálculo de la cota, en particular cuando: $\frac{\partial^2}{\partial\theta^2}\ell(\theta|\underline{X}) = c$.
- Revisitando el calculo de la cota de varianza para estimadores insesgados en el contexto del modelo Poisson / Exponencial.
- Dada $\underline{X} = \underline{x}$, llamamos información *observada* a:

$$J_n(\theta) = -\left(\frac{\partial^2}{\partial\theta^2}\ell(\theta|\underline{x})\right) = \sum_{i=1}^{n}\frac{\partial^2\log f(x_i;\theta)}{\partial\theta^2}.$$

Caso multiparámetro (familias exponenciales)

- $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x; \boldsymbol{\theta})$, con $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$.
- Consideramos estimadores $\mathbf{W}$ insesgados para $\boldsymbol{\theta}$ ($E(\mathbf{W}) = \boldsymbol{\theta}$).
- Ahora $\mathbf{I}(\boldsymbol{\theta})$ será una matriz de $d \times d$:

$$
\begin{aligned}
[\mathbf{I}(\boldsymbol{\theta})]_{ij} &= E\Big(\Big\{\frac{\partial}{\partial\theta_i}\log f_{\underline{X}}(\underline{X};\boldsymbol{\theta})\Big\}\Big\{\frac{\partial}{\partial\theta_j}\log f_{\underline{X}}(\underline{X};\boldsymbol{\theta})\Big\}\Big) \\
(\text{fam. exp.}) &= -E\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta}|\underline{X})\Big), \\
(iid) &= -nE\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X;\boldsymbol{\theta})\Big)
\end{aligned}
$$

- La cota de CR será $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$, y por tanto se verifica que:

$$\text{Var}(\mathbf{W}) \geq [\mathbf{I}(\boldsymbol{\theta})]^{-1}.$$

'Algoritmo' para encontrar un UMVUE:

1) `Check: Insesgadez y computo de varianza de tu estimador.`
2) `Construye la log-verosimilitud asociada al modelo.`
3) `Computa la derivada segunda de la log-verosimilitud.`
4) `Calcular esperanza en (3).`
5) `Utiliza (4) para calcular la información Fisher.`
6) `Compara la varianza de tu estimador contra la cota CR.`

Otro ejemplo (alumnes lo resuelven en clase):

- $X \sim \text{Bern}(\theta)$ y consideramos $\widehat{\theta} = \sum_{i=1}^{n} \frac{X_i}{n}$ (insesgado y $\text{Var}(\widehat{\theta}) = \frac{\theta(1-\theta)}{n}$).
  - Insesgado + varianza alcanza la cota de CR $\Rightarrow$ UMVUE.

Condición para alcanzar la cota

Teorema

Asumiendo que el modelo estadístico cumple H_1 y $W_n \in \mathcal{C}_\theta$, luego W_n alcanza la cota de CR si y solo si existe una función $a(\theta)$ que verifica:

$$a(\theta)[w_n - \theta] = \frac{\partial}{\partial\theta}\ell(\theta) = S(\theta) \quad \text{(proof: CB § 7.3.2).}$$

- Ejemplo: $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu_0, \sigma^2)$ con $\underline{\mu_0 \text{ conocida}}$. Existe un estimador que alcanza la cota y es $\widehat{\sigma}_n^2$ (el EMV) ya que:

$$\frac{\partial}{\partial\sigma^2}\ell(\sigma^2 \mid \mu_0) = \underbrace{\frac{n}{2\sigma^4}}_{a(\sigma^2)} \Big[\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2}_{\widehat{\sigma}_n^2} - \sigma^2 \Big] = a(\sigma^2)[\widehat{\sigma}_n^2 - \sigma^2].$$

- $\widehat{\sigma}_n^2$ es el UMVUE: Insesgado (cuando μ es conocida) + alcanza cota.

Agenda

- En nuestra discusión en torno a los UMVUE no utilizamos el concepto de suficiencia. Existen dos resultados teóricos relevantes a destacar:
  - Rao–Blackwell: Un estimador insesgado de θ se puede "mejorar" (reducir su varianza) si es condicionado por un estadístico suficiente.
    - ★ Corolario: En nuestra búsqueda de estimadores UMVUE debemos considerar solo aquellos que resultan funciones de el/los estadístico/s suficientes para el/los parámetro/s de interés.
    - ★ Los EMV en familias exponenciales son suficientes y completos :)
  - Lehmann–Scheffé: Si existe estimador insesgado de θ que sea función de un estadístico suficiente y completo, éste es el único estimador UMVUE de θ.

"Rao–Blackwellización" de estimadores

Theorem (Rao–Blackwell)

Si $W_n \in \mathcal{C}_\theta$ y T es suficiente para θ; $\phi(T) = E(W_n|T)$ es preferible a W_n:

a) *$\phi(T)$ es insesgado: $E(\phi(T)) = \theta$, (esperanza total) y*

b) *$V(\phi(T)) \leq V(W_n)$ (varianza total).*

c) *La distribución de $\phi(T)$ es independiente de θ (suficiencia).*

(proof: CB § 7: pp–342)

- Si $W_n \in \mathcal{C}_\theta$ pero no es función de un estadístico suficiente:
  - $\text{ECM}(\phi(T), \theta) \leq \text{ECM}(W_n, \theta)$ para todo $\theta \in \Theta$ y $n \geq 1$.
- **COROLARIO**: Un estimador insesgado que no es función de un estadístico suficiente será peor (mayor ECM) que otro que si lo es.
  - Ejemplo: $\widehat{\mu}_n = \overline{X}_n$ vs la Mediana cuando $X \sim N(\mu, \sigma^2)$.

Unicidad

- En el contexto de un modelo $X \sim f(x;\theta)$.

Theorem (Lehmann–Scheffé)

Sea T un estadístico suficiente y completo para θ y ϕ una función cualquiera de T; luego $\phi(T)$ es el **único** *estimador UMVUE de $E(\phi(T))$.*

- Si logramos encontrar una función ϕ del estadístico suficiente y completo T tal que $E(\phi(T)) = \theta$ (centramos correctamente a T), entonces $\phi(T)$ es el **único** estimador UMVUE de θ.
  - Ideal para modelos de la familia Exponencial ya que sabemos que el estadístico $T = \sum_{i=1}^{n} t(X_i)$ es completo y suficiente para θ.
  - Ejemplos: $\phi(T) = T/n$ resulta insesgado para los parámetros de los modelos Poisson, Exponencial y Bernoulli. Luego como T es suficiente y completo (modelos de la familia exponencial), entonces T/n es el único estimador UMVUE de dichos parámetros.

Agenda

Admisibilidad de un estimador

- Dada una función de riesgo R y un estimador W_n de $\theta \in \Theta$, decimos que W_n es **inadmisible** si existe otro estimador W_n' tal que:
  - $R(W_n', \theta) \leq R(W_n, \theta)$ para todo $\theta \in \Theta$, y
  - $R(W_n', \theta) < R(W_n, \theta)$ para al menos algún $\theta \in \Theta$.
- Un estimador es **admisible** si no es inadmisible.
  - La admisibilidad es la ausencia de una propiedad negativa.
- Los estimadores UMVUE no son necesariamente admisibles.
- ¿Existen estimadores admisibles? ¿Cómo los construimos?
  - Estimadores **minimax**.
  - Estimadores Bayesianos.

Estimadores **minimax**

- Como lo sugiere su nombre, los estimadores minimax buscan minimizar una cota máxima para el riesgo.
- W_n se dice estimador mini–max de θ si para cualquier otro estimador W_n' (en una clase $\mathcal{W}$) y una función de riesgo R ocurre que:

$$\sup_{\theta \in \Theta} R(W_n, \theta) \leq \sup_{\theta \in \Theta} R(W_n', \theta).$$

- Criterio conservador para definir optimalidad: Podríamos estar descartando estimadores W_n' que tienen poco riesgo en casi todo Θ excepto quizá en regiones pequeñas de Θ.
  - En este curso no vamos a abordar este criterio de optimalidad.
- Este tipo de aproximaciones conservadores son más simples de abordar desde el paradigma Bayesiano, dando como resultado estimadores de "regla bayesiana" que verifican también el principio de admisibilidad.

Optimalidad Bayesiana

- W_n es un estimador que sigue la regla (criterio) Bayesiana(o) si minimiza un promedio ponderado del riesgo Bayesiano $r(\pi, W_n)$.
- Dicha función se define, para una prior π de θ, como:

$$r(\pi, W_n) = E_\pi[R(W_n, \theta)] = \int_\Theta R(W_n, \theta)\pi(\theta)d\theta.$$

- W_n se dice una regla de estimación Bayesiana de θ con respecto a una prior π, si para cualquier otro estimador W_n' ocurre que:

$$r(W_n, \pi) \leq r(W_n', \pi).$$

- Bajo ciertas condiciones de regularidad sobre el modelo estadístico, podemos encontrar la expresión de W_n que minimiza $r(W_n, \pi)$ (y que en general dependerá de la elección de tu prior $\pi(\theta)$).

Agenda

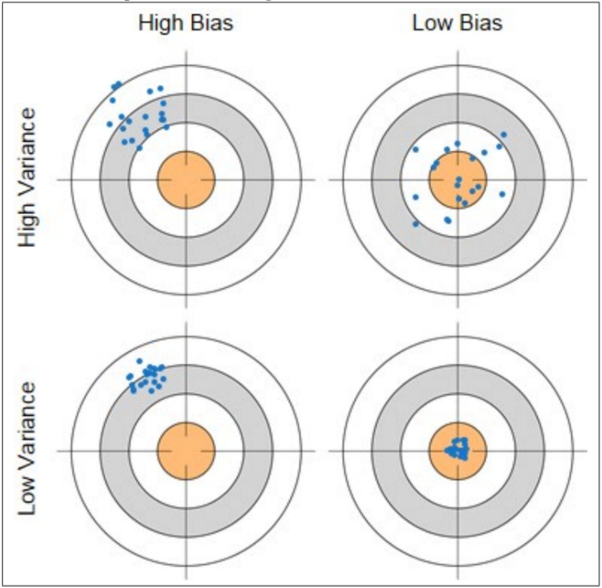


Figure: ¿Cómo será el ECM en cada caso? volver

Reforzando el mensaje:

- Si W_n tiene sesgo y varianza pequeños, entonces tiene un ECM bajo.
  - En otras palabras, la distribución de probabilidad de W_n está concentrada en torno de θ, o lo que es lo mismo
  - Con alta probabilidad el estimador W_n tomará valores cercanos al parámetro desconocido θ (garantiza que estimo razonablemente bien).
- COROLARIO: Si tengo dos estimadores W_n y W_n' y (por ejemplo) $\text{ECM}(W_n, \theta) < \text{ECM}(W_n', \theta)$ para todo posible valor de θ en la población, entonces *prefiero* W_n a W_n' como estimador de θ.
- Desafortunadamente, encontrar estimadores uniformemente mejores no es posible en general (ver otro contraejemplo en las próximas slides).

Estimar el parámetro de una Bernoulli

$\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Bern}(p)$ y comparamos dos estimadores $\widehat{p}_1$ y $\widehat{p}_2$:

$\widehat{p}_1$ la media muestral:

$$\widehat{p}_1 = \overline{X}_n \,.$$

$\widehat{p}_2$ agrega 2 fracasos y 2 éxitos artificiales a la muestra:

$$\widehat{p}_2 = \frac{\sum_{i=1}^n X_i + 2}{n+4}.$$

Calculemos primero el ECM de $\widehat{p}_1$. Sabemos que:

$$E\left(\widehat{p}_1\right) = p, \quad \text{y} \quad Var(\widehat{p}_1) = \frac{p(1-p)}{n}$$

Entonces el $\text{ECM}(\widehat{p}_1, p) = E\left[\left(\widehat{p}_1 - p\right)^2\right] = \frac{p(1-p)}{n}$.

Sesgo y varianza de $\widehat{p}_2$ (notar que $\sum_{i=1}^n X_i = T \sim Bin(n,p)$).

$$E\left(\frac{T+2}{n+4}\right) - p = \frac{E(T+2)}{n+4} - p = \frac{np+2}{n+4} - p = \frac{2(1-2p)}{n+4}.$$

El sesgo no es cero, salvo cuando $p = 0.5$. Sin embargo converge a cero cuando $n \to \infty$.

$$Var\left(\frac{T+2}{n+4}\right) = \frac{Var(T+2)}{(n+4)^2} = \frac{Var(T)}{(n+4)^2} = \frac{np(1-p)}{(n+4)^2}.$$

Entonces:

$$\text{ECM}(\widehat{p}_2, p) = E\left((\widehat{p}_2 - p)^2\right) = \left(\frac{2(1-2p)}{n+4}\right)^2 + \frac{np(1-p)}{(n+4)^2}.$$

¿Cuál es mejor?

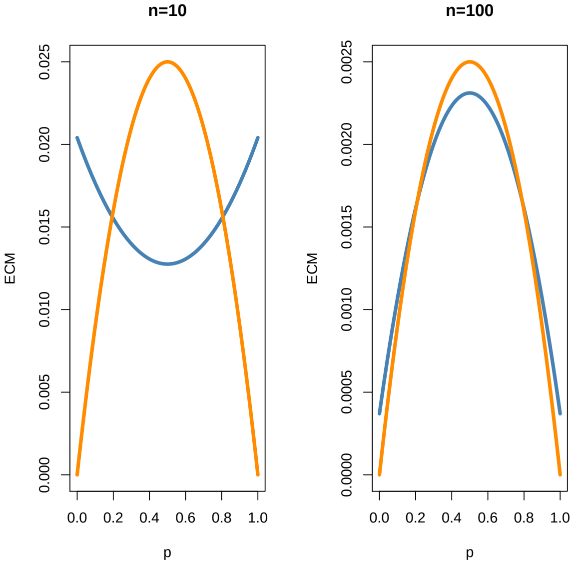


Figure: Error cuadrático medio para $\widehat{p}_1$ (naranja) y $\widehat{p}_2$ (azul). ▸ volver