CLASE 2 - MÉTODOS DE ESTIMACIÓN - Matías Carrasco

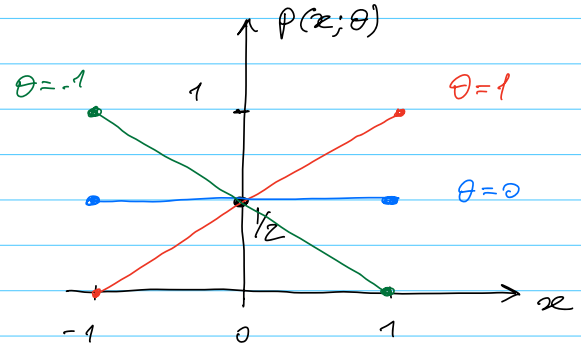
REPASO:

DEF: Un **modelo estadístico** es un par $(\mathcal{P}, \mathcal{Z})$ con:

- $\mathcal{Z}$ un espacio de **observaciones**
- $\mathcal{P}$ una **familia** de **distribuciones** de probabilidad en $\mathcal{Z}$

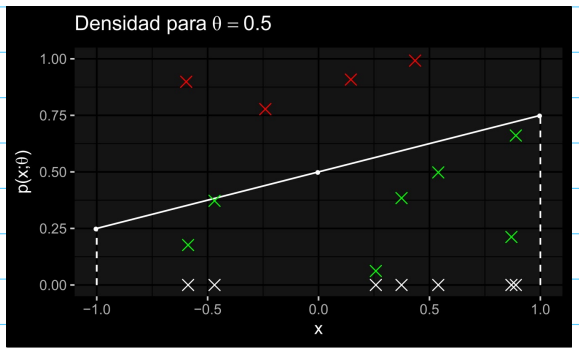
EJEMPLO: Consideremos $\mathcal{Z} = [-1, 1]$ $\mathcal{P} = \{p(x;\theta) : \theta \in [-1,1]\}$

$$p(x;\theta) = \frac{1}{2}(1 + \theta x) \quad x \in [-1,1]$$



En la clase pasada obtuvimos una muestra con el método de aceptación-rechazo:

$$-0.47,\ 0.89,\ 0.26,\ -0.59,\ 0.37,\ 0.54,\ 0.87$$



DEF: Sea $\mathcal{M} = (\mathcal{P}, \mathcal{Z})$ modelo estadístico.
Siempre podemos escribir $\mathcal{P}$ de la forma $\mathcal{P} = \{p(z;\theta) : \theta \in \Theta\}$
Es decir, siempre podemos "parametrizar" $\mathcal{P}$ con algún espacio de parámetros Θ.

Dependiendo de la dimensión de Θ decimos que $\mathcal{M}$ es **paramétrico** o **no paramétrico**:

Si $\dim \Theta$ $\begin{cases} \text{No depende del tamaño de la muestra} \Rightarrow \mathcal{M} \text{ es paramétrico} \\ \text{Depende del tamaño de la muestra} \Rightarrow \mathcal{M} \text{ es no paramétrico.} \end{cases}$

El ejemplo anterior es un ejemplo de modelo paramétrico.

En breve veremos un ejemplo de modelo no paramétrico.

Un poco de terminología

La muestra anterior consiste de una muestra i.i.d de $p(x;\theta)$

Esto quiere decir que tenemos $X_1, X_2, \dots, X_n$ v.a. independientes todas ellas con distribución $p(x;\theta)$.

La muestra es una realización concreta de las variables:

$$(X_1)_{obs} = x_1, \quad (X_2)_{obs} = x_2, \quad \dots, \quad (X_n)_{obs} = x_n$$

Siempre usamos $\begin{cases} X_i & \text{para la variable aleatoria} \\ x_i & \text{para la realización } (X_i)_{obs} \end{cases}$

En el ejemplo: $\begin{cases} x_1 = -0.47, \ x_2 = 0.89, \ \dots, \ x_7 = 0.87 \\ n = 7 \end{cases}$

El problema de estimación:

El problema de estimación consiste en **a partir** de la muestra observada $\{(X_i)_{obs}\} = \{x_i\}_{i=1}^n$ **estimar** el **parámetro** θ.

DEF: Un **estimador** es una **función** que a partir de una muestra genera una estimación para θ:

$$\hat{\theta}_n : \mathbb{R}^n \longrightarrow \Theta$$
$$(x_1 \dots x_n) \mapsto \hat{\theta}_n(x_1 \dots x_n)$$

Podemos considerar $\hat{\theta}_n$ como una **variable aleatoria** si escribimos

$$\boxed{\hat{\theta}_n(X_1, \dots, X_n)}$$

Con este punto de vista un estimador tiene una **distribución asociada**!

Para nosotros será **clave** entender algunas **características** de la distribución de un estimador, por ejemplo su **sesgo** y su **varianza**.

Máxima Verosimilitud

Es el método más popular para **definir estimadores**.

$M = (\mathcal{P}, \mathcal{Z})$ modelo estadístico $\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$

$Z_1, \dots, Z_n$ muestra iid de la distribución real θ_0.

DEF: La **verosimilitud** de un parámetro θ es la "**probabilidad**" de la muestra

$$V_n(\theta) = p(z_1;\theta)\, p(z_2;\theta) \cdots p(z_n;\theta)$$
$$= \prod_{i=1}^{n} p(z_i;\theta)$$

El negativo del logaritmo de la verosimilitud es:

$$\ell_n(\theta) = -\ln V_n(\theta) = -\sum_{i=1}^{n} \ln\left(p(z_i;\theta)\right)$$

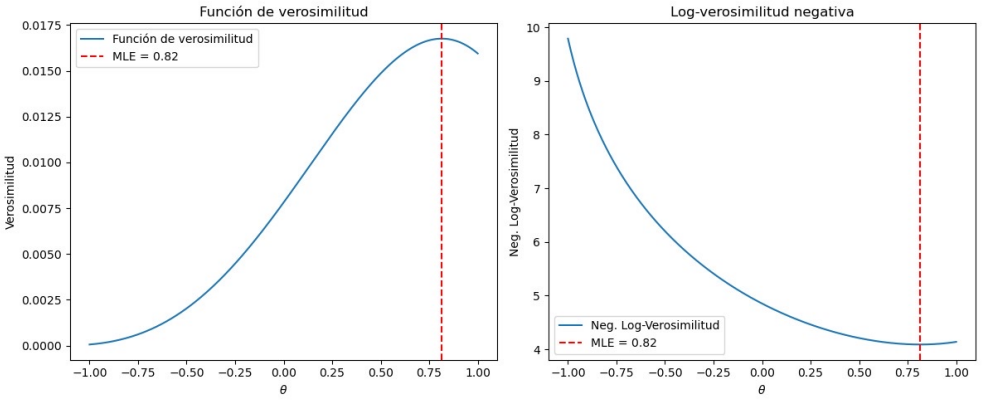
El estimador de máxima verosimilitud (MLE) es aquel valor de θ que maximiza $V_n(\theta)$

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} V_n(\theta)$$

Como el logaritmo es monótono, esto es equivalente a

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \ell_n(\theta)$$

EJEMPLO: en el ejemplo anterior



$$\ell(\theta) = -\sum_{i=1}^{n} \ln\left(\frac{1}{2}(1+\theta X_i)\right) = n\ln 2 - \sum_{i=1}^{n} \ln(1+\theta X_i)$$

Para la muestra dada $\hat{\theta}_{obs} = 0.82$

EJEMPLO: $\mathcal{M} = (\mathcal{P}, \mathcal{Z})$ $\quad \mathcal{Z} = \{0,1\}$

$\mathcal{P} = \{ p(z;\theta) : \theta \in [0,1] \}$

Bernoulli $\quad p(z;\theta) = \begin{cases} \theta & \text{si } z=1 \\ 1-\theta & \text{si } z=0 \end{cases} = \theta^z (1-\theta)^{1-z}$

Si tenemos una muestra $Z_1 \dots Z_n$:

$$V_n(\theta) = \prod_{i=1}^{n} p(Z_i;\theta) = \theta^{\sum z_i} (1-\theta)^{n-\sum z_i}$$

$$\ell_n(\theta) = -\left(\sum z_i\right) \ln\theta - \left(n - \sum z_i\right) \ln(1-\theta)$$

$$\begin{cases} \ell_n'(\theta) = -\dfrac{\sum z_i}{\theta} + \dfrac{(n-\sum z_i)}{1-\theta} \\ \ell_n''(\theta) = \dfrac{\sum z_i}{\theta^2} + \dfrac{(n-\sum z_i)}{(1-\theta)^2} > 0 \end{cases}$$

$$\ell_n'(\theta) = 0 \iff \frac{\sum z_i}{\theta} = \frac{(n-\sum z_i)}{1-\theta}$$

$$\iff (1-\theta)\sum z_i = (n-\sum z_i)\theta$$

$$\iff \sum z_i - \theta \sum z_i = (n-\sum z_i)\theta$$

$$\iff \sum z_i = n\theta$$

$$\iff \hat{\theta} = \sum z_i / n = \boxed{\overline{Z}_n} \rightarrow \text{promedio}$$

EJEMPLO: $\mathcal{M} = (\mathcal{P}, \mathcal{Z})$ $\quad \mathcal{Z} = \mathbb{R}$

$\mathcal{P} = \{ N(x;\mu,\sigma^2) : \theta = (\mu,\sigma^2) \}$

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$ **Normal**

$$V_n(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z_i-\mu)^2}$$

$$\ell_n(\mu,\sigma^2) = \frac{1}{2} n \ln 2\pi + \frac{1}{2} n \ln(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - \mu)^2$$

$$\begin{cases} \dfrac{\partial \ell_n}{\partial \mu} = -\dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (z_i - \mu) = 0 \\ \dfrac{\partial \ell_n}{\partial \sigma^2} = \dfrac{n}{2\sigma^2} - \dfrac{1}{2\sigma^4} \displaystyle\sum_{i=1}^{n} (z_i - \mu)^2 = 0 \end{cases}$$

$$\begin{cases} \hat{\mu} = \overline{z}_n \\ \hat{\sigma}^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (z_i - \overline{z})^2 \end{cases}$$

Sesgo y Varianza de un estimador

Un estimador $\hat{\theta}_n$ depende de la muestra $X_1, \dots, X_n \Rightarrow \hat{\theta}_n$ es una v.a.

$\theta_0 \in \Theta$ es el valor real que queremos estimar

DEF: 1) El error cuadrático medio (MSE) de $\hat{\theta}_n$ es:

$$MSE(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \theta_0)^2\right]$$

La esperanza es como un promedio en todas las muestras posibles

2) El sesgo de $\hat{\theta}_n$ es

$$\text{Sesgo}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta_0$$

3) La varianza de $\hat{\theta}_n$ es

$$\text{Var}(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2\right]$$

Por ejemplo:

$\mathbb{E}[\hat{\theta}_n]$ = Promedio$[\hat{\theta}_n]$ sobre $x_1 \dots x_n$ posibles

Estas tres cantidades están relacionadas de la siguiente manera:

PROPOSICIÓN: $\boxed{MSE(\hat{\theta}_n) = Sesgo(\hat{\theta}_n)^2 + Var(\hat{\theta}_n)}$

DEMOSTRACIÓN:

$$MSE(\hat{\theta}_n) \overset{\text{DEF}}{=} \mathbb{E}\left[(\hat{\theta}_n - \theta_0)\right]^2 \overset{\text{DISTRIBUTIVA}}{=} \mathbb{E}\left[\hat{\theta}_n^2 - 2\hat{\theta}_n\theta_0 + \theta_0^2\right]$$

$$\overset{\text{LINEALIDAD}}{=} \mathbb{E}\left[\hat{\theta}_n^2\right] - 2\theta_0\,\mathbb{E}\left[\hat{\theta}_n\right] + \theta_0^2$$

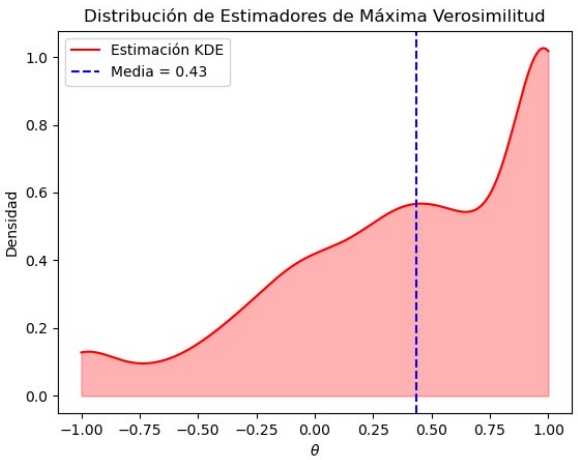
$$= Var(\hat{\theta}_n) + \mathbb{E}[\hat{\theta}_n]^2 - 2\theta_0\,\mathbb{E}[\hat{\theta}_n] + \theta_0^2$$

$$= Var(\hat{\theta}_n) + \underbrace{\left[\mathbb{E}[\hat{\theta}_n] - \theta_0\right]^2}_{Sesgo(\hat{\theta}_n)} = Var(\hat{\theta}_n) + Sesgo(\hat{\theta}_n)^2$$

EJEMPLO: Consideremos nuevamente el ejemplo $\mathcal{X} = [-1, 1]$

y $\mathcal{P} = \{p(x;\theta) : \theta \in [-1,1]\}$ $\quad p(x;\theta) = \frac{1}{2}(1+\theta x)$, $-1 \le x \le 1$

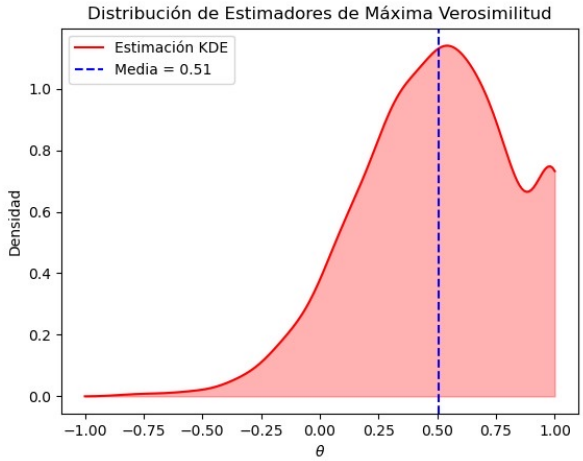
Consideremos ahora muchas (2000) muestras de tamaño $n=7$ y grafiquemos la distribución de $\hat{\theta}$ el estimador MLE.



la varianza de $\hat{\theta}_7$ en este caso es

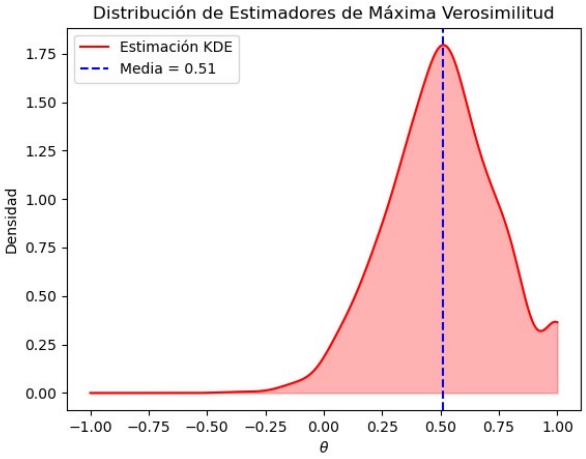
$Var(\hat{\theta}_7) = 0.32$

Observemos ahora el efecto que tiene el tamaño de la muestra en el MLE:



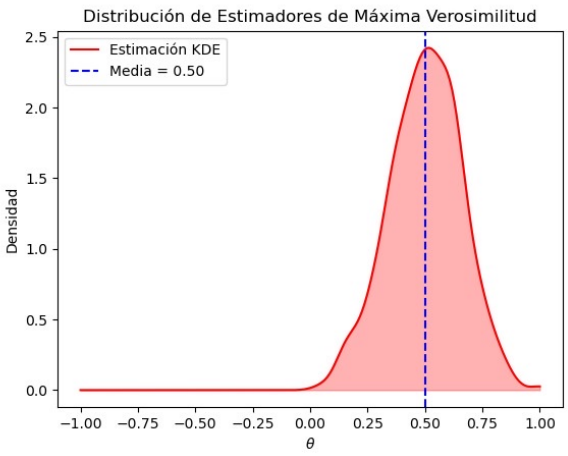
$n = 25$

$Var(\hat{\theta}_{25}) = 0.10$



$n = 50$

$Var(\hat{\theta}_{50}) = 0.06$



$n = 100$

$Var(\hat{\theta}_{100}) = 0.02$

Vemos que al aumentar n:

1) Sesgo $(\hat{\theta}_n) \to 0$ (el MLE es **asintóticamente insesgado**)
2) Var $(\hat{\theta}_n) \to 0$

1) y 2) juntos implican $MSE(\hat{\theta}_n) \to 0$: $\hat{\theta}_n$ es **consistente**

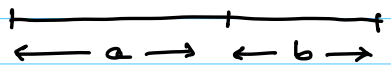
Además vemos que la distribución de $\hat{\theta}_n$ es cada vez más parecida a la distribución normal. El MLE es **asintóticamente normal**.

Cómo se calcula el MLE en la práctica: métodos de optimización

Veremos solo un ejemplo para ilustrar el concepto:

El Algoritmo Golden Search

La proporción áurea:



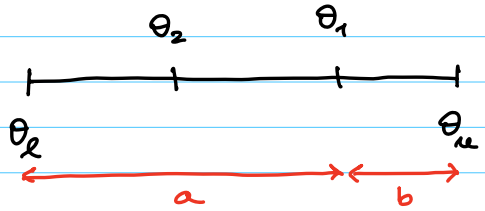
$$\frac{\text{grande}}{\text{chico}} = \frac{\text{todo}}{\text{grande}}$$

$$\frac{a}{b} = \frac{a+b}{a} = 1 + \left(\frac{a}{b}\right)^{-1}$$

$$\phi = 1 + \frac{1}{\phi}, \quad \phi^2 - \phi - 1 = 0$$

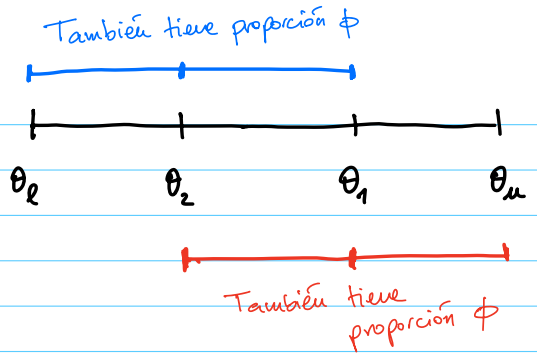
$$\boxed{\phi = \frac{1+\sqrt{5}}{2} = 1.618\ldots}$$

Dado un intervalo $[\theta_\ell, \theta_u]$ podemos subdividirlo de acuerdo a la proporción áurea desde ambos extremos:



$$\frac{a}{b} = \frac{a+b}{a} \qquad \frac{\theta_1 - \theta_\ell}{\theta_u - \theta_1} = \frac{\theta_u - \theta_\ell}{\theta_1 - \theta_\ell} = \phi$$

Hacemos lo mismo con θ_2



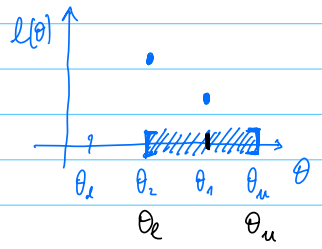
El Algoritmo:

1 - Comenzar con dos valores iniciales θ_l y θ_u que encierran al mínimo de la función

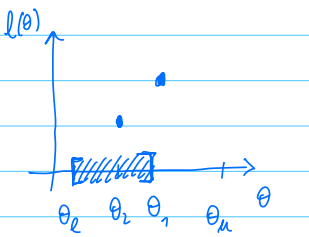
2 - Subdividir:
$$\begin{cases} \theta_1 = \theta_l + \dfrac{\theta_u - \theta_l}{\phi} \\ \theta_2 = \theta_u - \dfrac{\theta_u - \theta_l}{\phi} \end{cases}$$

3 - Evaluar $\ell(\theta_1)$ y $\ell(\theta_2)$

4 - Si $\ell(\theta_1) < \ell(\theta_2)$:
$$\hat{\theta} \leftarrow \theta_1$$
$$\theta_l \leftarrow \theta_2$$
$$\theta_u \leftarrow \theta_u$$
$$\theta_2 \leftarrow \theta_1$$
$$\theta_1 \leftarrow \theta_l + \frac{\theta_u - \theta_l}{\phi}$$



Si $\ell(\theta_1) > \ell(\theta_2)$:
$$\hat{\theta} \leftarrow \theta_2$$
$$\theta_u \leftarrow \theta_1$$
$$\theta_l \leftarrow \theta_l$$
$$\theta_1 \leftarrow \theta_2$$
$$\theta_2 \leftarrow \theta_u - \frac{\theta_u - \theta_l}{\phi}$$



5 - Repetir hasta alcanzar $|\theta_u - \theta_l| < \underbrace{\text{tolerancia}}_{\tau}$

EJERCICIO: Implementarlo en el ejemplo

Un modelo no paramétrico: el histograma

$\mathcal{Z} = [0,1]$ y queremos estimar una densidad $p(x)$

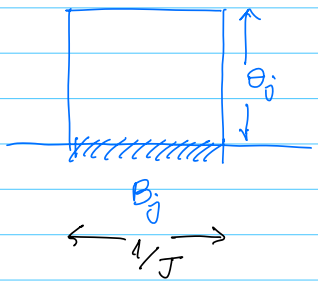
Recordar que para hacer un histograma:

- Dividimos $[0,1]$ en intervalos (bins) $\{B_j\}_{j=1}^{J}$ de ancho Δ
- Contamos cuantos elementos de la muestra $X_1 \cdots X_n$ caen en cada bin B_j.

Supongamos dado el ancho Δ de los bins.

Definimos $\mathcal{P}$ la familia de distribuciones de la siguiente manera:

$$p(x;\theta) = \sum_{j=1}^{J} \theta_j \mathbb{1}_{\{x \in B_j\}}$$



El parámetro es $\theta = (\theta_1, \dots, \theta_J)$ que tiene dimensión J. $(= 1/\Delta)$

Notar que $\frac{1}{J} \sum_{j=1}^{J} \theta_j = 1$ pues queremos área $= 1$.

(Observar también que $1/J = \Delta$)

O sea que $\Theta = \left\{ \theta = (\theta_1, \dots, \theta_J) : \frac{1}{J} \sum_{j=1}^{J} \theta_j = 1 \right\}$

Se puede ver que el MLE $\hat{\theta}$ es:

$$\hat{\theta}_j = J \times \frac{\#\{i : X_i \in B_j\}}{n}$$

Es decir: el histograma clásico es el MLE.
Para obtener consistencia debemos hacer $\Delta \to 0$ a medida que $n \to \infty$, por lo que el modelo es NO PARAMÉTRICO.