

## Curs 18

Eroarea pătetică medie (mean square error)

Def: Fie  $x_1, x_2, \dots, x_n$  un eșantion de volum  $n$  dintr-o populație  $\Omega$  și  $\hat{\theta}_n$  un estimat pentru  $\theta$ . Definim eroarea pătetică medie:

$$MSE_{\theta}(\hat{\theta}_n) = E_{\theta}[(\hat{\theta}_n - \theta)^2]$$

Să punem că estimatul  $\hat{\theta}_1$  este mai bun decât estimatul  $\hat{\theta}_2$  (în sensul eroarei pătetică medie) dacă

$$MSE_{\theta}(\hat{\theta}_1) < MSE_{\theta}(\hat{\theta}_2)$$

P) Arătu că

$$MSE_{\theta}(\hat{\theta}) = V_{\theta}(\hat{\theta}) + b_{\theta}(\hat{\theta})^2$$

În particular, dacă estimatul  $\hat{\theta}$  este nedeplasat pt o atingere

$$MSE_{\theta}(\hat{\theta}) = V_{\theta}(\hat{\theta})$$

d): În general nu se găsește estimator pt care eroarea pătetică medie este căt mai mică

Ex:  $x_1, \dots, x_n \sim N(\mu, \sigma^2)$  și scopul este să estimăm varianta  $\sigma^2$ .

Arătu statistică  $T_n = \sum_{i=1}^n (x_i - \bar{x}_n)^2$  și către un estimat de forma  $\hat{\sigma}^2 = a T_n$ ,  $a = a(n)$  m. pozitiv.

Până acum am văzut că  $a = \frac{1}{n} \rightarrow$  varianta cuprinsă

$$V_n^2$$

$a = \frac{1}{n-1} \rightarrow$  varianta esantionului

$$S_n^2$$

Nreau să determinăm  $a$  pt care

$MSE_{\sigma^2}(\hat{\tau}_a^2)$  este minimă.

$$\hat{\tau}_a^2 = a T_n - a \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

$$MSE_{\sigma^2}(\hat{\tau}_a^2) = Var_{\sigma^2}(\hat{\tau}_a^2) + b_{\sigma^2}(\hat{\tau}_a^2)^2$$

Reamiihui: în cazul populației care vărtă că

$$\frac{(n-1) S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sigma^2} \sim \chi^2(n-1)$$

Stim:  $\mathbb{E}[\chi^2(v)] = v$

$$\begin{cases} Var(\chi^2(v)) = 2v \\ Var(\chi^2(v)) = 2v \end{cases}$$

$$\Rightarrow \frac{T_n}{\sigma^2} \sim \chi^2(n-1)$$

$$\Rightarrow E_{\sigma^2}\left(\frac{T_n}{\sigma^2}\right) = n-1 \Rightarrow E_{\sigma^2}[T_n] = (n-1)\sigma^2$$

$$Var_{\sigma^2}\left(\frac{T_n}{\sigma^2}\right) = 2(n-1) \Rightarrow Var_{\sigma^2}(T_n) = 2(n-1)\sigma^4$$

Aștept,

$$\begin{aligned} MSE_{\sigma^2}(\hat{\tau}_a^2) &= Var_{\sigma^2}(a T_n) + b_{\sigma^2}(a T_n)^2 \\ &= a^2 Var_{\sigma^2}(T_n) + [E_{\sigma^2}(a T_n) - \sigma^2]^2 \end{aligned}$$

$$= 2a^2(n-1)\sigma^4 + [a(n-1)\sigma^2 - \sigma^2]^2$$

$$= \sigma^4 [2a^2(n-1) + (a(n-1)-1)^2]$$

- 3 -

Derivând după  $\hat{\sigma}^2$  și înmulțind cu  $-2$ :

$$\frac{d}{d\hat{\sigma}^2} \text{MSE}_2(\hat{\sigma}^2) = \sigma^4 [4a(n-1) + 2(a(n-1)(n-1))]$$

$$= 2\sigma^4(n-1)[2a + a(n-1)-1]$$

$$\frac{d}{da} \text{MSE}_2(\hat{\sigma}^2) = 0 \Leftrightarrow 2a + a(n-1) - 1 = 0$$

$$\Leftrightarrow a = \frac{1}{n-1}.$$

Pentru a avea o punctie estimatoare a  $\sigma^2$  de forma

$$a \sum_{i=1}^n (x_i - \bar{x}_n)^2$$
, estimatul cu eroarea patologică medie cea mai mică este  $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ 

Est.  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$  este un estimat nediplosat

④ Deoarece  $\text{MSE}_2(\hat{\theta}_n) \rightarrow 0$  pt  $n \rightarrow \infty$  atunci  $\hat{\theta}_n$  este un estimat consistent pt  $\theta$ .

(Ex): Compararea a doi estimatori nediplosați după MSE  
Fie  $X_1, X_2, \dots, X_n \sim \text{Pois}(\theta)$ , și  $\hat{\theta}_1 = \bar{x}_n$  și  $\hat{\theta}_2 = S_n^2$  doi estimatori nediplosați pt  $\theta$ .

Vrem să determinăm care este mai bun din punctul de vedere al eroarei patologice medie:

$$\text{MSE}_2(\hat{\theta}_1) \quad (?) \quad \text{MSE}_2(\hat{\theta}_2) \quad \left| \begin{array}{l} \text{V}_{\theta_0}(S_n^2) = \\ \frac{1}{n} (\mu_4 - \frac{n-3}{n} \sigma^4) \\ \mu_4 = E[(X_i - \mu)^4] \end{array} \right.$$

$$\text{V}_{\theta_0}(\hat{\theta}_1) \quad \quad \quad \text{V}_{\theta_0}(\hat{\theta}_2)$$

## Metoda de construcție a estimării

### 1) Metoda momentelor

- sfârșitul sec XIX începutul sec XX (Karl Pearson)

Fie  $X_1, X_2, \dots, X_n$  unele egale număr de valori n dintr-o populație.

distribuție  $f_{\theta}$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$ ,  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$

Dacă  $X \sim f_{\theta}$  este un moment de ordin  $j$ ,  $1 \leq j \leq k$  sunt

$$E[X^j] = \int_{-\infty}^{+\infty} x^j f_{\theta}(x) dx, \quad j \in \{1, 2, \dots, k\}$$

În general,

$$E[X^1] = g_1(\theta_1, \dots, \theta_k)$$

$$E[X^2] = g_2(\theta_1, \dots, \theta_k)$$

$$E[X^k] = g_k(\theta_1, \dots, \theta_k)$$

Metoda momentelor presupune rezolvarea sistemului de  $k$  ecuații cu  $k$  necunoscute:

$$\left\{ \begin{array}{l} \frac{1}{n} \sum_{i=1}^n x_i = g_1(\theta_1, \dots, \theta_k) \\ \frac{1}{n} \sum_{i=1}^n x_i^2 = g_2(\theta_1, \dots, \theta_k) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n x_i^k = g_k(\theta_1, \dots, \theta_k) \end{array} \right.$$

Soluția este estimăria obținută prin metoda momentelor.

Esp:  $x_1, x_2, \dots, x_n \sim \text{Exp}(\lambda)$ ,  $\lambda = \theta \in (0, \infty)$

-5-

Cum  $\theta \in \mathbb{R} \Rightarrow \theta \in \mathbb{R}$

$$\begin{array}{ccc} \text{momentul empiric} & = & \text{momentul teoretic} \\ \text{de ordin 1} & & \text{de ordin 1} \\ \overrightarrow{x_n} = \mathbb{E}_\theta[x_1] = \frac{1}{2} & & \end{array}$$

Astfel sistemul se reduce la  $\overline{x}_n = \frac{1}{2} \Rightarrow \hat{\theta}_n = \frac{1}{\overline{x}_n}$

Esp: Trei  $x_1 = 0.42, x_2 = 0.10, x_3 = 0.65, x_4 = 0.23$   
un egratim de volum  $n=4$  din jdg.

$$f_\theta(x) = \theta x^{\theta-1}, 0 \leq x \leq 1$$

$\theta \in [0, 1] \subseteq \mathbb{R} \Rightarrow \theta \in \mathbb{R}$

$\hat{\theta}_n = \frac{1}{\overline{x}_n}$

Met. momentelor:  $\overline{x}_n = \mathbb{E}_\theta[x_1]$

$$\mathbb{E}_\theta[x_1] = \int_{-\infty}^{+\infty} x f_\theta(x) dx = \int_0^1 \theta x^\theta dx = \frac{\theta}{\theta+1}$$

$$\Rightarrow \overline{x}_n = \frac{\hat{\theta}_n}{\hat{\theta}_n + 1} \quad (\Rightarrow \hat{\theta}_n = \frac{\overline{x}_n}{1 - \overline{x}_n})$$

estimatiul  $\hat{\theta}$  obținut  
prin metoda momentelor

$$\hat{\theta}_n = \frac{0.35}{1 - 0.35} = 0.54 \quad , \quad \overline{x}_4 = 0.35$$

Expo:  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$

In acest caz  $\Theta = (\theta_1, \theta_2) \subset (\mu, \sigma^2)$

$\theta \in \Theta = \mathbb{R} \times (0, \infty) \subseteq \mathbb{R}^2$ ,  $k=2$

Met. momentelor presupune rest. unui sistem cu 2 ec.

n! = 2 neavansat!

$$\begin{cases} \bar{X}_n = E_{\theta}[X_1] \\ \frac{1}{n} \sum_{i=1}^n X_i^2 = E_{\theta}[X_1^2] \end{cases}$$

Dacă  $X_i \sim N(\mu, \sigma^2) \Rightarrow E_{\theta}[X_1] = \mu$

$$\begin{aligned} E_{\theta}[X_1^2] &= V_{\theta}(X_1) + E_{\theta}[X_1]^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Sistemul devine:

$$\begin{cases} \bar{X}_n = \mu \\ \frac{1}{n} \sum_{i=1}^n X_i^2 = \sigma^2 + \mu^2 \end{cases} \Rightarrow \begin{cases} \tilde{\mu}_n = \bar{X}_n \\ \tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \end{cases}$$

$$\Rightarrow \begin{cases} \tilde{\mu}_n = \bar{X}_n \quad - \text{ media gantimului} \\ \tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad ( \text{ varianța empirică} ) \end{cases}$$

Obs: Ce se întâmplă în cazul unei rep.  $B(k, p)$ ?

$$\begin{cases} \bar{X}_n = kp \\ \frac{1}{n} \sum_{i=1}^n X_i^2 = kp(1-p) + kp^2 \end{cases} \quad \begin{cases} p \in (0, 1) \\ k \in \mathbb{N} \end{cases}$$

2) Metoda verosimilitutii maxime (maximum likelihood)

Fie  $x_1, x_2, \dots, x_n \sim f_\theta(x)$ ,  $\theta \in \Theta \subseteq \mathbb{R}^k$

Densitatea (nura) comună a  $(x_1, x_2, \dots, x_n)$  este

$$f_\theta(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

Def: Se numește funcție de verosimilitatea asociată eguienței  $x_1, x_2, \dots, x_n$ ; funcția

$$L(\theta | x_1, x_2, \dots, x_n) = f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

este variată ca o funcție de  $\theta$

Notăm de asemenea

$$l(\theta | x_1, \dots, x_n) = \log L(\theta | x_1, \dots, x_n) - \text{logantul} \\ \text{fct de verosimilitate}$$

Obs: Vom putea întâlni doar notarea  $L(\theta | x)$  respectiv  $l(\theta)$  sau  $L(\theta | x)$  și respectiv  $l(\theta | x)$

OBS: Fct. de verosimilitatea nu este o probabilitate de probabilitate pt.  $\theta$ .

OBS: Dacă  $L(\theta_1 | x) > L(\theta_2 | x)$  (echivalent  $l(\theta_1 | x) > l(\theta_2 | x)$ ) atunci spunem că  $\theta_1$  este mai probabil decât  $\theta_2$  să fi produs obs.  $x_1, x_2, \dots, x_n$ .

Cu alte ~~cuvinte~~<sup>-8-</sup> fo<sub>1</sub> reprezintă reprezintă modelul matem. dictat fo<sub>2</sub> în ceea ce privește fitarea datelor observate.

Def: Fie  $x_1, x_2, \dots, x_n \sim f_{\theta}(x)$  și  $L(\theta|x)$ ,  $l(\theta|x)$  sunt funcții de verosimilitate resp. logaritmul lor de verosimilitate.

Numești estimare de verosimilitate maximă (MLE) punctul  $\hat{\theta}$

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta|x) = \underset{\theta \in \Theta}{\operatorname{argmax}} l(\theta|x)$$

$$\hat{\theta}_n \in \hat{\Theta}_n(x_1, x_2, \dots, x_n)$$

Ob: Dacă funcția de verosimilitate este dif. ;  $\theta = (\theta_1, \dots, \theta_k)$  atunci problemele candidatelor MLE sunt soluții ale sistemului

$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0, \quad i=1, \dots, k$$

Exp:  $x_1, x_2, \dots, x_n \sim \text{Ber}(\theta)$ ,  $\theta \in (0, 1)$

Funcția de verosimilitate:

$$L(\theta|x) = \prod_{i=1}^n f_{\theta}(x_i)$$

$$f_{\theta}(x) = \begin{cases} \theta, & x=1 \\ 1-\theta, & x=0 \end{cases} = \theta^x (1-\theta)^{1-x}$$

$$L(\theta|x) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{\sum_{i=1}^n (1-x_i)}$$

$$L(\theta|x) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} - \theta^n$$

Amen:  $\ell(\theta|x) = \log L(\theta|x) = \left( \sum_{i=1}^n x_i \right) \log(\theta) + \left( n - \sum_{i=1}^n x_i \right) \log(1-\theta)$

Dacă  $0 < \sum_{i=1}^n x_i < n$  atunci derivând

$$\frac{d}{d\theta} \ell(\theta|x) = \frac{\sum_{i=1}^n x_i}{\theta} - \left( n - \sum_{i=1}^n x_i \right) \frac{1}{1-\theta}$$

Rezolvare.

$$\frac{d}{d\theta} \ell(\theta|x) = 0 \Leftrightarrow \frac{\sum_{i=1}^n x_i}{\theta} = \frac{1}{1-\theta} \left( n - \sum_{i=1}^n x_i \right) / n$$

$$(2) \quad \frac{\bar{x}_n}{\theta} = \frac{1}{1-\theta} (1 - \bar{x}_n) \Leftrightarrow \hat{\theta}_n = \bar{x}_n$$

Dacă  $\sum_{i=1}^n x_i = 0$  sau  $\sum_{i=1}^n x_i = n$  atunci

$$\ell(\theta|x) = \begin{cases} n \log(1-\theta), & \sum_{i=1}^n x_i = 0 \\ n \log(\theta), & \sum_{i=1}^n x_i = n \end{cases}$$

în  $\ell(\theta|x)$  este monotonă în  $\theta$  și maximul se

atinge pt  $\theta = 0$  dacă  $\sum x_i = 0$       }  $\Rightarrow \hat{\theta}_n = \bar{x}_n$   
 și pt  $\theta = 1$  dacă  $\sum x_i = n$

În general,  $\boxed{\hat{\theta}_n = \bar{x}_n}$  este notul de verificare.

~~Definirea~~: Determinarea estimatorului pentru metoda momentelor.

Exp:  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$  cu  $\mu$  și  $\sigma^2$  necunoscute  
 $\theta = (\mu, \sigma^2) = (\mu, \sigma^2)$

Functia de verosimilitate:

$$L(\mu, \sigma^2 | x) = \prod_{i=1}^n f_0(x_i)$$

$$\text{unde } f_0(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{x_i-\mu}{\sigma}\right)^2} \quad (\text{distribuția } N(\mu, \sigma^2))$$

$$L(\mu, \sigma^2 | x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Logantreniuand

$$l(\mu, \sigma^2 | x) = \log L(\mu, \sigma^2 | x)$$

$$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^2}$$

trebuie extinsul

$$\begin{cases} \frac{\partial}{\partial \mu} l(\mu, \sigma^2 | x) = 0 \\ \frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2 | x) = 0 \end{cases}$$

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2 | x) = -\frac{1}{2} \sum_{i=1}^n \frac{\partial}{\partial \mu} \frac{(x_i-\mu)^2}{\sigma^2}$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2 | x) = 0 \Leftrightarrow \sum_{i=1}^n (x_i - \mu) = 0 \Leftrightarrow \hat{\mu}_n = \frac{\sum_{i=1}^n x_i}{n}$$

$$\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2 | x) = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2$$

$$\left\{ \begin{array}{l} \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2 | x) = 0 \\ \hat{\mu}_n = \bar{x}_n \end{array} \right. \quad (\Rightarrow) \quad \frac{-n}{\hat{\sigma}_n^2} + \frac{1}{\hat{\sigma}_n^4} \sum_{i=1}^n (x_i - \hat{\mu}_n)^2 = 0$$

$$(\Rightarrow) \boxed{\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Thus we have,  $\bar{x}_n$  in  $\min_{\mu} \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$  our estimation of  $\mu$  coincides with maximum pt.  $\mu$  of  $\sigma^2$  in crcl pop. uncond.