Estimatori punctuali. Teorema limită centrală

Fie (X, f_θ) un **model statistic** asociat unei populaţii, adică X este o variabilă aleatoare ce cuantifică o caracteristică comună a indivizilor populaţiei respective şi a cărei distribuţie depinde de un parametru necunoscut $\theta \in \Theta \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$, iar f_θ este densitatea de probabilitate a variabilei aleatoare X în cazul în care X este o variabilă aleatoare continuă, respectiv $f_\theta(x) = P(X = x)$, $x \in \mathbb{R}$, dacă X este variabilă aleatoare discretă.

Vectorul θ poate să aibă o singură componentă, caz în care distribuţia de probabilitate depinde de un singur parametru, de exemplu distribuţia Bernoulli depinde de parametrul $p \in (0, 1)$, dar poate să fie şi un vector cu mai multe componente, caz în care distribuţia de probabilitate depinde de mai mulţi parametrii, de exemplu distribuţia normală depinde de $m \in \mathbb{R}$ şi $\sigma^2 > 0$, adică $\theta = (m, \sigma^2) \in \mathbb{R} \times (0, \infty)$.

Dacă $X_1, X_2, \ldots, X_n$ sunt variabile aleatoare independente identic distribuite (i.i.d.) ca şi X, atunci vectorul $(X_1, X_2, \ldots, X_n)$ se numeşte **selecţie aleatoare** de volum n asociată modelului statistic (X, f_θ), iar dacă $(x_1, x_2, \ldots, x_n)$ este o realizare a selecţiei aleatoare $(X_1, X_2, \ldots, X_n)$, caz în care se spune că $x_1, x_2, \ldots, x_n$ sunt valori de selecţie ale variabilei aleatoare X, atunci orice funcţie

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n),$$

de variabile $x_1, x_2, \ldots, x_n$, se numeşte **estimator punctual** al parametrului θ. Cu alte cuvinte, un estimator punctual al parametrului θ este o realizare a variabilei aleatoare $\hat{\theta}(X_1, X_2, \ldots, X_n)$.

Tipuri de estimatori:

Un estimator $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ al parametrului θ se numeşte:

- **estimator consistent** dacă

$$\lim_{n\to\infty} P\left(|\hat{\theta}(X_1, X_2, \ldots, X_n) - \theta| > \varepsilon\right) = 0, \text{ pentru orice } \varepsilon > 0,$$

  adică distribuţia variabilei $\hat{\theta}(X_1, X_2, \ldots, X_n)$ este din ce în ce mai concentrată în jurul parametrului θ pe măsură ce volumul selecţiei aleatoare creşte;

- **estimator nedeplasat (centrat)** dacă

$$M\left(\hat{\theta}(X_1, X_2, \ldots, X_n)\right) = \theta.$$

Fie $\widehat{\theta}_1$, $\widehat{\theta}_2$ doi estimatori nedeplasaţi ai parametrului θ. Spunem că $\widehat{\theta}_1$ este **mai eficient** decât $\widehat{\theta}_2$ dacă are loc inegalitatea:

$$\sigma^2\left(\widehat{\theta}_1(X_1, X_2, \ldots, X_n)\right) < \sigma^2\left(\widehat{\theta}_2(X_1, X_2, \ldots, X_n)\right),$$

ceea ce este echivalent cu

$$\frac{\sigma^2\left(\widehat{\theta}_1(X_1, X_2, \ldots, X_n)\right)}{\sigma^2\left(\widehat{\theta}_2(X_1, X_2, \ldots, X_n)\right)} < 1.$$

Cu alte cuvinte, cu cât varianţa (dispersia) variabilei $\hat{\theta}(X_1, X_2, \ldots, X_n)$ este mai mică, cu atât estimatorul $\hat{\theta}$ este mai eficient.

- Determinarea mediei şi dispersiei pentru variabila medie aritmetică $\overline{X}$.

Dacă $(X_1, X_2, \ldots, X_n)$ este o selecţie aleatoare de volum n asociată unui model statistic (X, f_θ), cu $M(X) = m$ şi $\sigma^2(X) = \sigma^2$, atunci variabila medie aritmetică

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

are media m şi dispersia $D^2 = \frac{\sigma^2}{n}$.

În plus, dacă variabilele aleatoare i.i.d. $X_1, X_2, \ldots, X_n$ au o **distribuţie normală**,

$$X_i \sim N(m, \sigma^2),\ i = \overline{1, n},$$

atunci media lor aritmetică $\overline{X}$ are de asemenea o **distribuţie normală** de medie m şi dispersie $D^2 = \dfrac{\sigma^2}{n}$, adică

$$\overline{X} \sim N(m, D^2 = \sigma^2/n). \tag{1}$$

- **Teorema limită centrală** (enunţ, consecinţe)

Dacă (X_n) este un şir de variabile aleatoare independente şi identic distribuite având media comună m şi abaterea standard σ, iar

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

este şirul variabilelor medie aritmetică, atunci pentru n suficient de mare ($n \geq 30$) distribuţia de probabilitate a variabilelor $\overline{X}_n$ este **aproximativ normală** de medie m şi dispersie $D^2 = \sigma^2/n$. Notăm

$$\overline{X}_n \sim ApN(m, D^2 = \sigma^2/n).$$

În plus, pentru n suficient de mare, avem

$$S_n = X_1 + X_2 + \cdots + X_n \sim ApN(n\,m, D^2 = n\sigma^2). \tag{2}$$

- **Estimarea mediei**: Media de selecţie, definită prin

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

este un estimator nedeplasat al mediei $m = M(X)$, adică

$$\widehat{m} = \overline{x}$$

este un estimator nedeplast al parametrului m.

- **Estimarea varianţei (dispersiei)**: Dispersia de selecţie, definită prin

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

este un estimator nedeplasat al dispersiei $\sigma^2 = \sigma^2(X)$.

Estimatorul verosimilităţii maxime (maximum likelihood estimator):

Fie (X, f_θ) un model statistic şi $x_1, x_2, \ldots, x_n$ valori de selecţie ale lui X (se mai spune că $x_1, x_2, \ldots, x_n$ este un eşantion de volum n). **Funcţia de verosimilitate** (likelihood function) asociată eşantionului $x_1, x_2, \ldots, x_n$ este $L : \Theta \subset \mathbb{R}^d \to \mathbb{R}$,

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) = f_\theta(x_1) \cdot f_\theta(x_2) \cdots f_\theta(x_n).$$

Punctul de maxim global al funcţiei de verosimilitate (dacă există) se numeşte **estimatorul verosimilităţii maxime** al parametrului

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \Theta.$$

Cu alte cuvinte, estimatorul verosimilităţii maxime al lui θ (dacă acesta există) este

$$\hat{\theta} = \operatorname{argmax} L(\theta),$$

unde prin $\operatorname{argmax} L(\theta)$ se înţelege argumentul care maximizează funcţia L.

Probleme rezolvate

Exemplul 1. Cererea de memorie pentru o aplicaţie, ca proporţie din memoria ce poate fi alocată de un utilizator, este o variabilă aleatoare X ce are densitatea de probabilitate

$$f(x) = \begin{cases} (\theta+1)x^{\theta}, & 0 < x < 1, \\ 0, & \text{în rest.} \end{cases}$$

a) Să se determine media teoretică $M(X)$ a variabilei aleatoare X şi apoi un estimator nedeplasat al parametrului θ în funcţie de media de selecţie $\overline{x}$ a unei selecţii aleatoare de volum n.

b) Să se determine un estimator nedeplasat al parametrului θ din selecţia următoare:

$$0.2, 0.4, 0.5, 0.7, 0.8, 0.9, 0.9, 0.6, 0.6, 0.4,$$

rezultată în urma rulării aplicaţiei cu diferite date de intrare.

Rezolvare: a) Mai întâi calculăm media teoretică:

$$M(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_0^1 (\theta+1)x^{\theta+1}dx = (\theta+1)\frac{x^{\theta+2}}{\theta+2}\Big|_0^1 = \frac{\theta+1}{\theta+2}.$$

Dacă $m = M(X)$ şi $\overline{x}$ este media de selecţie a unui eşantion $x_1, x_2, \ldots, x_n$, atunci din egalitatea impusă

$$\hat{m} = \overline{x}$$

se determină un estimator nedeplasat al parametrului θ:

$$\frac{\hat{\theta}+1}{\hat{\theta}+2} = \overline{x} \quad \Leftrightarrow \quad \hat{\theta} = \frac{2\overline{x}-1}{1-\overline{x}}.$$

b) Pentru valorile înregistrate deducem că $\overline{x} = 0.6$, deci un estimator nedeplasat al parametrului θ este $\hat{\theta} = \dfrac{2\overline{x}-1}{1-\overline{x}} = 0.5$.

□

Exemplul 2. Fie $\mathcal{P}$ populaţia formată dintr-un tip de circuite. Caracteristica ce dorim să o investigăm este durata de viaţă a acestor circuite, ştiind că aceasta este exponenţial distribuită, cu parametrul $\theta > 0$ necunoscut.

Măsurând timpul de viaţă (în ani) a 10 circuite, se obţin valorile:

0.8830, 1.9651, 1.9189, 4.8448, 0.9208 3.4377, 1.7162, 4.2327, 5.9435, 8.3128.

Să se determine estimatorul verosimilităţii maxime al parametrului θ (adică pentru media duratei de viaţă a acestui tip de circuite).

Soluţie. Densitatea de probabilitate a distribuţiei exponenţiale este

$$f_\theta(x) = \begin{cases} 0, & \text{dacă } x < 0, \\ \frac{1}{\theta} e^{-\frac{x}{\theta}}, & \text{dacă } x \geq 0. \end{cases}$$

Astfel, funcţia de verosimilitate este

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \left(\frac{1}{\theta} e^{-\frac{x_i}{\theta}} \right) = \frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}}.$$

Notăm cu

$$\ell(\theta) = \ln(L(\theta)).$$

Funcţiile ℓ şi L au aceleaşi puncte de extrem şi de aceeaşi natură, deoarece funcţia $x \mapsto \ln x$ este strict crescătoare. Pentru simplificarea calculelor, vom determina punctul de maxim global (dacă acesta există) pentru ℓ, iar acesta va fi punct de maxim global şi pentru L. Avem

$$\ell(\theta) = \ln L(\theta) = -n \ln \theta - \frac{\sum_{i=1}^{n} x_i}{\theta},$$

deci

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}.$$

Rezolvând ecuaţia

$$\ell'(\theta) = 0,$$

în raport cu θ, obţinem punctul

$$\theta = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x},$$

care este punct de maxim global pentru ℓ, deci şi pentru L. În concluzie, estimatorul verosimilităţii maxime al parametrului θ al distribuţiei exponenţiale este media de selecţie, $\hat{\theta} = \overline{x}$. În cazul exemplului dat, estimatorul verosimilităţii maxime al mediei de viaţă a circuitelor este:

$$\hat{\theta} = \overline{x} = \frac{x_1 + x_2 + \cdots + x_{10}}{10} = 3.4175.$$

Observaţie: Valorile de selecţie alese mai sus au fost generate simulând o variabilă aleatoare $X \sim Exp(\theta = 3.5)$,

$$> \text{ rexp}(10, 1/3.5) \quad \# \text{ în limbajul de programare R, rate}=1/\theta,$$

deci estimatorul verosimilităţii maxime $\hat{\theta} = 3.4175$ este destul de bun.

□

Exemplul 3. Un simulator al distribuţiei Bernoulli

$$X : \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix},$$

de parametru $p \in (0, 1)$ necunoscut, generează stringul de biţi:

$$1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0.$$

Să se determine estimatorul verosimilităţii maxime al parametrului p pe baza eşantionului de biţi generaţi.

Soluţie: În cazul de faţă parametrul necunoscut este $\theta = p$. Cum X este o variabilă aleatoare discretă, avem

$$f_p(x) = P(X = x) = \begin{cases} p, & \text{dacă } x = 1, \\ 1-p, & \text{dacă } x = 0. \end{cases}$$

Funcţia de verosimilitate asociată eşantionului de biţi generaţi este

$$L(p) = L(p; x_1, x_2, \ldots, x_{25}) = f_p(x_1) \cdot f_p(x_2) \cdots f_p(x_{25}) = p^{12}(1-p)^{13},$$

căci stringul are 12 biţi de 1 şi 13 biţi de 0. Se logaritmează şi se determină punctul de maxim global al funcţiei

$$\ell(p) = \ln L(p) = 12 \ln(p) + 13 \ln(1-p).$$

Cum

$$\ell'(p) = \frac{12}{p} - \frac{13}{1-p} = \frac{12(1-p) - 13p}{p(1-p)} = \frac{12 - 25p}{p(1-p)},$$

ecuaţia

$$\ell'(p) = 0$$

are soluţia $p = \frac{12}{25}$. Se arată că $p = \frac{12}{25}$ este un punct de maxim al funcţiei ℓ, deci estimatorul verosimilităţii maxime pentru parametrul p al distribuţiei Bernoulli, dedus din stringul de biţi generaţi mai sus, este

$$\hat{p} = \frac{12}{25} = 0.48,$$

adică este egal cu numărul biţilor de 1 din string supra numărul total de biţi. Acest estimator al lui p este de fapt probabilitatea intuitivă de a obţine bitul 1: numărul cazurilor favorabile supra numărul cazurilor posibile din string.

□

Probleme propuse

1. Numărul de accesări ale `http://cs.upt.ro` într-o oră este o variabilă aleatoare Poisson X de parametru $\lambda > 0$ necunoscut. În 5 ore consecutive numărul de accesări înregistrate a fost

$$10, 12, 18, 15, 15.$$

Să se estimeze parametrul λ din aceste date. Să se determine estimatorul verosimilităţii maxime al lui λ.

2. Există ipoteza că timpul de aşteptare T, pentru a primi acces la un server internet, are densitatea de probabilitate

$$f(x) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{în rest,} \end{cases}$$

unde θ este un parametru pozitiv necunoscut. Pentru a estima parametrul θ s-au făcut 5 observaţii care au condus la următorii timpi de aşteptare (în secunde):

$$0.4, 0.7, 0.2, 0.6, 0.1.$$

Să se calculeze media teoretică a timpului de aşteptare şi apoi un estimator nedeplasat al parametrului θ. Să se determine estimatorul verosimilităţii maxime pentru parametrul θ pe baza datelor din eşantion.

3. Deduceţi estimatorul verosimilităţii maxime al parametrului p al distribuţiei $\text{Bin}(n, p)$, unde n este un număr natural nenul **cunoscut**.

4. Şirul de valori de mai jos

$$3.9, 2.4, 2.1, 3.6, 3.0$$

a fost generat apelând de 5 ori consecutiv generatorul $2 + (\texttt{b} - 2) * \texttt{urand}()$; Să se estimeze parametrul b.

Indicaţie: Numerele date sunt valori de observaţie asupra unei variabile aleatoare $X \sim Unif\,[2, b)$. Media unei variabile aleatoare $X \sim Unif\,[a, b)$ este $M(X) = (a + b)/2$. Parametrul b se poate calcula determinând un estimator nedeplasat al mediei $m = M(X)$, $\hat{m} = \overline{x}$.

5. Să se determine estimatorul verosimilităţii maxime pentru parametrul θ al distribuţiei uniforme

$$f_\theta(x) = \begin{cases} \dfrac{1}{\theta}, & x \in [0, \theta], \\ 0, & \text{în rest,} \end{cases}$$

pe baza valorilor de observaţie $x_1, x_2, \ldots, x_n$.

Indicaţie: Funcţia de verosimilitate este

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \frac{1}{\theta^n}, \text{ dacă } \theta \geq \max\{x_1, x_2, \ldots, x_n\}.$$

Funcţia $L : [\max\{x_i \, : \, i = \overline{1,n}\}, \infty) \to \mathbb{R}$ are derivata $L'(\theta) = -n/\theta^{n+1} < 0$, deci este strict descrescătoare şi maximul global îl atinge în $\max\{x_i \, : \, i = \overline{1,n}\}$. Deci estimatorul verosimilităţii maxime este $\hat{\theta} = \max\{x_i \, : \, i = \overline{1,n}\}$.

6. Fie X_i, $i = \overline{1,100}$, variabile aleatoare i.i.d. de distribuţie Bernoulli($p = 0.2$). Să se precizeze distribuţia de probabilitate a variabilei medie aritmetică

$$\overline{X} = \frac{1}{100} \sum_{i=1}^{100} X_i$$

şi să se estimeze probabilitatea $P(\overline{X} > 0.3)$.

7. Un DVD conţine o zonă neînregistrată de 350 MB. Să se estimeze probabilitatea ca această zonă de memorie să fie suficientă pentru a copia 300 de fotografii ce au dimensiuni independente, de medie 1 MB şi abatere standard 0.5 MB.