

Seminar 14

Estimatori. Teorema de limită centrală

14.1 Probleme rezolvate

1. Cererea de memorie pentru o aplicație, ca proporție 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 și apoi să se estimeze θ în funcție de media de selecție \bar{x} a unei selecții aleatoare de volum n .

b) Să se determine un estimator al parametrului θ din selecția 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 aplicației cu diferite date de intrare.

Rezolvare:

a) Mai întâi calculăm media teoretică:

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

Dacă $m = M(X)$ și \bar{x} este media de selecție a unui eșantion de valori x_1, x_2, \dots, x_n , atunci din egalitatea impusă $\hat{m} = \bar{x}$ se determină un estimator al parametrului θ :

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

b) Pentru valorile înregistrate avem $\bar{x} = 0.6$, deci un estimator pentru parametrul θ este $\hat{\theta} = \frac{2\bar{x} - 1}{1 - \bar{x}} = 0.5$.

2. Pentru a estima rata sosirii λ a cererilor de acces la o bază de date s-au monitorizat intervalele de timp dintre 10 cereri consecutive și s-au înregistrat valorile:

$$0.2, 0.1, 0.1, 0.05, 0.05, 0.2, 0.8, 0.5, 0.2, 0.8.$$

Care este estimatorul ratei sosirilor, $\hat{\lambda}$?

Rezolvare:

Sosirile cererilor la o bază de date este un proces Poisson (N_t) , $t \geq 0$, de rată $\lambda > 0$. Intervalul inter-sosirilor are distribuția exponențială, $X \sim \text{Exp}(\theta = 1/\lambda)$. Dar parametrul θ pentru distribuția exponențială este valoarea medie a variabilei X .

Prin urmare din datele înregistrate putem calcula un estimator al lui θ , adică a mediei lungimii intervalelor inter-sosirilor: $\hat{\theta} = 0.3$. Din relația $\theta = 1/\lambda$, adică $\lambda = 1/\theta$, obținem un estimator al ratei sosirilor ca fiind $\hat{\lambda} = 1/\hat{\theta} = 1/0.3 = 3.33$.

3. Fie populația \mathcal{P} formată dintr-un tip de circuite. Caracteristica ce dorim să o investigăm prin sondaj statistic este durata de viață a acestor circuite, știind că aceasta este exponențial distribuită, cu parametrul θ necunoscut. Măsurând timpul de viață (în ani) a 10 circuite, se obțin valorile:

0.8830, 1.96511, 1.9189, 4.8448, 0.9208 3.4377, 1.7162, 4.2327, 5.9435, 8.3128.

Să determinăm estimatorul de verosimilitate maximă pentru θ (adică pentru media duratei de viață a acestui tip de circuite).

Rezolvare:

Densitatea de probabilitate a distribuției exponențiale este

$$f_{\theta}(x) = \begin{cases} 0, & \text{dacă } x < 0, \\ \frac{1}{\theta} e^{-x/\theta}, & \text{dacă } x \geq 0. \end{cases} \quad (14.1)$$

Astfel, funcția de verosimilitate este

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^n x_i/\theta}. \quad (14.2)$$

Funcția logaritmică cu bază mai mare ca 1 are derivata întâi pozitivă. Notând cu $h = \ln$ și cu $\ell(\theta) = h(L(\theta))$, avem că funcția $\ell'(\theta) = h(L(\theta))L'(\theta)$ are același semn ca derivata lui L , deci ℓ și L au aceleași puncte de extrem și de aceeași natură. Pentru simplitatea calculelor, determinăm punctul de maxim absolut (dacă acesta există) pentru ℓ și acesta va fi punct de maxim absolut și pentru L :

$$\ell(\theta) = \ln L(\theta; x_1, x_2, \dots, x_n) = -n \ln \theta - \frac{\sum_{i=1}^n x_i}{\theta}. \quad (14.3)$$

Avem

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2}. \quad (14.4)$$

Rezolvând ecuația $\ell'(\theta) = 0$ în raport cu θ , obținem punctul $\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$, care este maxim absolut pentru ℓ , deci și pentru L . Reamintim condiția suficientă de maxim: $\ell^{(n)}(x_0) < 0$, n -par, unde x_0 este punct critic, i.e. $\ell'(x_0) = 0$. Aici, $\ell''(\bar{x}) = -\frac{n}{\bar{x}^2} < 0$.

În concluzie,

$$\operatorname{argmax} L(\theta; x_1, x_2, \dots, x_n) = \bar{x},$$

estimatorul de verosimilitate maximă a parametrului θ a distribuției exponențiale este media de selecție. În cazul exemplului dat, estimatorul verosimilității maxime a mediei de viață a circuitelor este media selecției:

$$\hat{\theta} = \frac{x_1 + x_2 + \dots + x_{10}}{10} = 3.525.$$

Observație: Valorile de selecție au fost generate simulând o variabilă $X \sim \operatorname{Exp}(\theta = 3.5)$, deci estimatorul verosimilității maxime $\hat{\theta} = 3.525$ este "destul de bun".

4. Un simulator al distribuției Bernoulli

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

de parametru p necunoscut, generează stringul de biți:

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

Să se determine estimatorul verosimilității maxime al parametrului p pe baza eșantionului de biți.

Rezolvare: Notând parametrul p necunoscut cu θ , distribuția de probabilitate a variabilei X este $p_X(\theta; b) = P(X = b)$, unde b este bitul 1 sau 0.

$$p_X(\theta; b) = \begin{cases} \theta, & \text{dacă } b = 1, \\ 1 - \theta, & \text{dacă } b = 0. \end{cases}$$

Funcția de verosimilitate asociată eșantionului de biți generați este

$$L(\theta; b_1, b_2, \dots, b_{26}) = p_X(\theta; b_1) p_X(\theta; b_2) \cdots p_X(\theta; b_{26}) = \theta^{12} (1 - \theta)^{14},$$

unde 12 este numărul de biți 1 din string, iar 14 numărul de biți 0. Se logaritmează și se determină punctul de maxim absolut al funcției $\ell(\theta) = \ln(L(\theta))$. Cum

$$\ell(\theta) = 12 \ln(\theta) + 14 \ln(1 - \theta),$$

rezultă

$$\ell'(\theta) = \frac{12}{\theta} - \frac{14}{1 - \theta} = \frac{12(1 - \theta) - 14\theta}{\theta(1 - \theta)}.$$

Ecuția $\ell'(\theta) = 0$ are soluția $\hat{\theta} = \frac{12}{26}$. Se verifică că acesta este un punct de maxim pentru ℓ , deci estimatorul verosimilității maxime pentru parametrul p al distribuției Bernoulli: într-adevăr, $\ell''(\frac{12}{26}) < 0$. Se observă că $\hat{\theta}$ este egal cu numărul biților 1 din string supra numărul total de biți. Acest estimator al lui p este de fapt probabilitatea intuitivă de a obține bitul 1: numărul cazurilor favorabile din string supra numărul cazurilor posibile.

5. Variabila aleatoare X , care dă numărul de zone defecte ale unui CD de un anumit tip, are următoarea distribuție de probabilitate: $p(x) = P(X = x)$, unde

x	$p(x)$
0	0.75
1	0.15
2	0.10

- a) Să se calculeze media și abaterea standard a numărului de zone defecte ale CD-ului.
b) Ce distribuție de probabilitate are media de selecție a unui eșantion de 400 de CD-uri din tipul investigat? Care este media și dispersia acestei distribuții?
c) Care este probabilitatea ca media numărului de zone defecte/CD într-un lot de 400 de CD-uri să fie mai mică decât 0.3?

Rezolvare:

- a) $m = M(X) = 0.15 + 0.2 = 0.35$, iar

$$\sigma^2(X) = (0 - 0.35)^2 0.75 + (1 - 0.35)^2 0.15 + (2 - 0.35)^2 0.1 = 0.4275$$

și deci abaterea standard este $\sigma = \sigma(X) = \sqrt{\sigma^2(X)} = 0.6538$.

- b) Volumul eșantionului fiind mare, conform teoremei limită centrală

$$\overline{X}_{400} \sim \text{ApN}(m, \sigma^2/400),$$

unde $m = 0.35$, $D^2 = \sigma^2/400 = 0.00106$, iar $D = 0.032$.

- c) Avem

$$\begin{aligned} P(\overline{X}_{400} < 0.3) &= F_{\overline{X}}(0.3) = \Phi\left(\frac{0.3 - m}{D}\right) = \Phi\left(\frac{0.3 - 0.35}{0.032}\right) \\ &= \Phi(-1.5625) = 1 - \Phi(1.5625) = 1 - 0.94 = 0.06. \end{aligned}$$

14.2 Probleme propuse

6. Fie densitatea de probabilitate

$$f(x) = \begin{cases} 2\theta x^{2\theta-1} & \text{daca } 0 \leq x \leq 1 \\ 0 & \text{altfel} \end{cases},$$

unde parametru $\theta > 0$ este necunoscut.

- a). Sa se determine media teoretica si apoi estimatorul parametrului θ .
b). Sa se determine estimatorul verosimilitatii maxime al parametrului θ pe baza unui esantion oarecare de volum n.

7. Un simulator al distributiei geometrice de parametru $p \in (0, 1)$ genereaza sirul de numere: 2, 3, 1, 14, 1, 3, 4, 1, 11, 10.

Reamintim ca daca X este o variabila distribuita geometric, atunci $f(x) = P(X = x) = pq^{x-1}$, $x = 1, 2, \dots$, unde p este probabilitatea succesului, iar $q = 1 - p$.

a). Sa se determine estimatorul verosimilitatii maxime al lui p pe baza unui esantion oarecare x_1, \dots, x_n de observatii ale variabilei distribuite geometric X .

b). Sa se determine numeric un estimator nedeplasat al parametrului p pe baza esantionului din problema.

8. Se consideră o buclă **for**:

```
for(i=1;i<=n;i++) // n>30
{
    executa blocul B;
}
```

Știind că timpul de execuție al blocului B este o variabilă aleatoare de distribuție de probabilitate necunoscută, având media $m = 60ms$ și abaterea standard de $\sigma = 8ms$, iar execuțiile succesive ale blocului sunt independente, să se determine distribuția de probabilitate a timpului de execuție a buclei **for**. Care este probabilitatea ca timpul de execuție al buclei în cazul $n = 50$ să fie cuprins între $0.75s$ sec și 1 sec?

Indicație Notând cu T_i timpul celei de-a i -a execuții a blocului B, timpul total de execuție

$$T = T_1 + T_2 + \dots + T_{50}$$

este aproximativ normal distribuit.

9. In primul an de operare Dropbox Romania va accepta un milion de clienti. Se estimeaza ca cererea de memorie de stocare X_i , de catre un user, i , $i = 1 \dots 10^6$, are media $m = 1.5Gb$ si abaterea standard de $\sigma = 0.5Gb$.

Ce volum, x , de Gb trebuie asigurat, daca cu o probabilitate de $p = 0.9$, cererea totala, C , va fi de cel putin x Gb? Se va folosi $z_{0.1} = -1.28$.

10. Fie X_1, X_2, \dots, X_{50} variabile aleatoare discrete i.i.d. având distributia de probabilitate Poisson de parametrul $\lambda = 2.5$. Care este distributia de probabilitate a variabilei aleatoare medie aritmetica $\bar{X}_{50} = (X_1 + X_2 + \dots + X_{50})/50$?