УВОД У СТАТИСТИКУ час 8

19. април ’17.

Метод максималне веродостојности

Пример 1

Претпостави се да је дат фаличан новчић, при чему се зна да је просечан (очекиван) удео броја палих писама једна од три вредности $p \in \{0.2, 0.3, 0.8\}$. Изводи се експеримент који се састоји у бацању овог новчића двапут и регистровању броја палих писама. Илустровати идеју ММВ – за оцену непознатог параметра p бира се вредност за коју је вероватноћа реализације датог реализованог узорка највећа.

Решење:

Ова ситуација математички се може моделирати простим сл. узорком (X_1, X_2) обима 2 за обележје X са Бернулијевом $B(1, p)$ расподелом, при чему је параметарски скуп (скуп допустивих вредности за p) скуп $\{0.2, 0.3, 0.8\}$.

Приметити да оцењивање методом момената у овом примеру не даје прихватљиву оцену непознатог параметра p . Наиме:

$$p = EX = \bar{X}_n,$$

па би једине могуће оцењене вредности биле: $\hat{p} = 0$ или $\hat{p} = 0.5$ или $\hat{p} = 1$, а ниједна од њих се не налази у параметарском скупу.

Зато се разматра вероватноћа из заједничког закона расподеле простог сл. узорка:

$$L(x_1, x_2; p) = p^{x_1+x_2}(1-p)^{2-(x_1+x_2)},$$

$x_j \in \{0, 1\}$, $j = 1, 2$. Вредности ове вероватноће дате су у следећој таблици за различите парове (x_1, x_2) и допустиве вредности p:

$p/(x_1, x_2)$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
0.20	0.64	0.16	0.16	0.04
0.30	0.49	0.21	0.21	0.09
0.80	0.04	0.16	0.16	0.64

Стога је оцена $\hat{p}$ која максимизује вероватноћу реализованог пара (x_1, x_2) дата са:

$$\hat{\boldsymbol{p}} = \begin{cases} \boldsymbol{0.20}, & \text{ако је } (\boldsymbol{x_1}, \boldsymbol{x_2}) = (\boldsymbol{0}, \boldsymbol{0}) \\ \boldsymbol{0.30}, & \text{ако је } (\boldsymbol{x_1}, \boldsymbol{x_2}) \in \{(\boldsymbol{0}, \boldsymbol{1}), (\boldsymbol{1}, \boldsymbol{0})\} \\ \boldsymbol{0.80}, & \text{ако је } (\boldsymbol{x_1}, \boldsymbol{x_2}) = (\boldsymbol{1}, \boldsymbol{1}) \end{cases}.$$

Пример 2

Нека обележје X има $P(\lambda)$ расподелу, где је $\lambda > 0$.
Наћи оцену непознатог параметра λ на основу простог случајног узорка $(X_1, X_2, \dots, X_n)$ обима n из обележја X.

Решење:

Функција веродостојности дата је са:

$$L(x_1, x_2, \dots, x_n; \lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{j=1}^{n} x_j}}{\prod_{j=1}^{n} x_j!},$$

$x_j \in \mathbb{N}_0$, а њен логаритам са:

$$\ln L(x_1, x_2, \dots, x_n; \lambda) = -n\lambda + \sum_{j=1}^{n} x_j \ln \lambda - \ln\left(\prod_{j=1}^{n} x_j!\right).$$

Једначина максимума веродостојности је:

$$\frac{d}{d\lambda} \ln L(x_1, x_2, \dots, x_n; \lambda) = -n + \sum_{j=1}^{n} \frac{x_j}{\lambda} = 0,$$

чије је решење

$$\hat{\lambda} = \sum_{j=1}^{n} \frac{x_j}{n} = \bar{x}_n.$$

Могуће је проверити да је ово решење тачка у којој се заиста достиже максимум, рачунањем другог извода:

$$\frac{d^2}{d\lambda^2}\ln L(x_1, x_2, \dots, x_n; \lambda) = -\sum_{j=1}^{n}\frac{x_j}{\lambda^2}$$

и евалуирањем у тачки $\hat{\lambda}$, када се добија $-\frac{n}{\bar{x}_n} < 0$.

Дакле, оцена непознатог параметра λ, добијена ММВ, је статистика $\hat{\lambda} = \bar{X}_n$.

Лако се показује да је ова оцена непристрасна. △

Приметити да су за $X \in P(\lambda)$ оцене непознатог параметра λ методом момената:

$$\hat{\lambda} = \bar{X}_n,$$

и

$$\hat{\lambda} = \bar{S}_n^2$$

(јер је $\lambda = EX = DX$).

Пример 2 (примена)

Обележје и узорак: подаци о броју саобраћајних незгода на територији општине са око 102000 становника за 15 случајно одабраних дана без падавина, током 2016. г.

Реализовани узорак: $(0, 2, 8, 4, 2, 2, 3, 5, 1, 2, 1, 0, 4, 3, 4)$.

На основу ових података оценити вероватноћу $P\{X = 0\}$, где је X посматрано обележје.

Како је број возача (релативно) велики, а за сваког од њих је (релативно) мала вероватноћа да учествује у саобраћајној незгоди одређеног дана, чини се оправданим претпоставити да је број незгода X у току једног дана случајна величина са Пуасоновом $\mathcal{P}(\lambda)$ расподелом где је $\lambda > 0$ непознат параметар.

Треба оценити: $\theta = \theta(\lambda) = P\{X = 0\} = e^{-\lambda}$.

Модел описан у Примеру 2 сада се репараметризује користећи: $\lambda = -\ln\theta$, након чега се добија нова функција веродостојности $L^*(x_1, x_2, \dots, x_n; \theta)$. Њу би требало максимизовати по θ како би се добила жељена оцена.

Даље је:

$$\ln L^*(x_1, x_2, \dots, x_n; \theta) = n\ln\theta + \sum_{j=1}^{n} x_j \ln(-\ln\theta) - \ln\left(\prod_{j=1}^{n} x_j!\right),$$

$$\frac{d}{d\theta} L^*(x_1, x_2, \dots, x_n; \theta) = \frac{n}{\theta} + \sum_{j=1}^{n} \frac{x_j}{\theta \ln\theta} = 0,$$

па је: $\hat{\theta}\left(= \hat{\theta}(\lambda) = \theta(\hat{\lambda})\right) = e^{-\bar{x}_n} = e^{-\hat{\lambda}}$.

Пример 2 (примена – наставак)

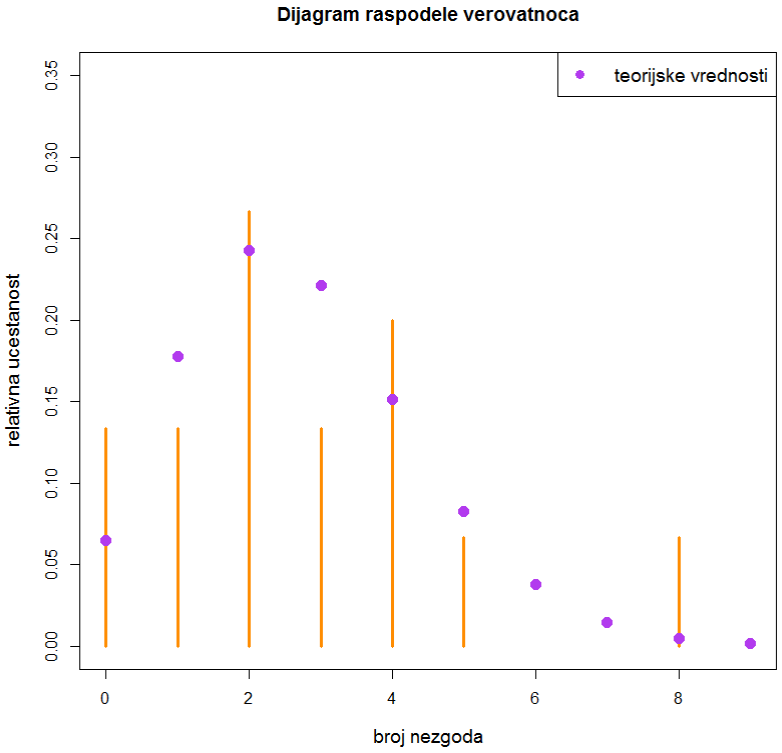
На основу реализованог узорка добија се следећа вредност узорачке средине:

$$(\hat{\lambda} =)\bar{x}_n = \frac{41}{15} \approx 2.7333,$$

па је оцењена вредност $\hat{\theta}$ ММВ тражене вероватноће: $\boldsymbol{\hat{\theta} \approx 0.065}$.

При одређеним генералним претпоставкама $100 \cdot \hat{\theta}$ је уједно и проценат дана у 2016. г. без падавина током којих се није догодила ниједна саобраћајна незгода.

Визуелно слагање података са моделом – слагање узорачког дијаграма расподеле вероватноћа и дијаграма расподеле вероватноћа $\mathcal{P}(\hat{\lambda})$ расподеле



Пример 3

Нека обележје X има $\gamma(\alpha,\beta)$, где су $\alpha,\beta \in (0,+\infty)$.

Наћи оцене непознатих параметара α и β на основу простог случајног узорка $(X_1, X_2, \dots, X_n)$ обима n из обележја X.

Решење:

Функција веродостојности дата је са:

$$L(x_1, x_2, \dots, x_n; \alpha, \beta) = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\left(\prod_{j=1}^{n} x_j\right)^{\alpha-1} e^{-\beta \sum_{j=1}^{n} x_j},$$

$x_j > 0$, а њен логаритам са:

$$\ln L(x_1, x_2, \dots, x_n; \alpha, \beta) = n\alpha \ln \beta - n \ln \Gamma(\alpha) + (\alpha - 1) \ln\left(\prod_{j=1}^{n} x_j\right) - \beta \sum_{j=1}^{n} x_j .$$

Пример 3 (наставак)

Парцијални изводи су:

$$\frac{\partial}{\partial\alpha}\ln L(x_1,x_2,\dots,x_n;\alpha,\beta)=n\ln\beta-\frac{n\Gamma'(\alpha)}{\Gamma(\alpha)}+\ln\left(\prod_{j=1}^{n}x_j\right),$$

$$\frac{\partial}{\partial\beta}\ln L(x_1,x_2,\dots,x_n;\alpha,\beta)=\frac{n\alpha}{\beta}-\sum_{j=1}^{n}x_j\,.$$

Означи се: $\tilde{x}_n := \left(\prod_{j=1}^{n}x_j\right)^{1/n}$ (узорачка геометријска средина) и $\Psi(\alpha) := \Gamma'(\alpha)/\,\Gamma(\alpha)$. Изједначавањем горњих парцијалних извода са нулом добијају се једначине:

$$\ln\hat{\alpha}-\Psi(\hat{\alpha})-\ln\frac{\bar{x}_n}{\tilde{x}_n}=0,$$

$$\hat{\beta}=\frac{\hat{\alpha}}{\bar{x}_n}.$$

Ово је пример када се једначине максимума веродостојности не могу, у општем случају, експлицитно решити. △

Пример 3 (специјалан случај: $\alpha = 1$)

Обележје X има $\varepsilon(\beta)$, $\beta > 0$.

Ту је:

$$L(x_1, x_2, \dots, x_n; \beta) = \beta^n e^{-\beta \sum_{j=1}^{n} x_j}, \; x_j > 0,$$

$$\ln L(x_1, x_2, \dots, x_n; \beta) = n \ln \beta - \beta \sum_{j=1}^{n} x_j .$$

Једначина максимума веродостојности је:

$$\frac{d}{d\beta} \ln L(x_1, x_2, \dots, x_n; \beta) = \frac{n}{\beta} - \sum_{j=1}^{n} x_j = 0,$$

па је

$$\hat{\beta} = \frac{1}{\bar{x}_n}.$$

Приметити да се методом момената добија иста оцена непознатог параметра експоненцијалне расподеле.

Пример 4 (тешкоће у примени ММВ)

Нека обележје X има $U(0,\theta)$ расподелу, где је $\theta \in (0,+\infty)$.

Наћи оцену непознатог параметра θ на основу простог случајног узорка $(X_1, X_2, \dots, X_n)$ обима n из обележја X.

Решење:

Функција веродостојности дата је са:

$$L(x_1, x_2, \dots, x_n; \theta) = \frac{1}{\theta^n},$$

$x_j \in (0,\theta)$. Ово је опадајућа функција (по θ) на целом $\mathbb{R}^+$, па се она стога максимизира најмањом могућом вредношћу θ. Међутим, сви елементи узорка морају бити мањи од θ. Стога тражена најмања могућа вредност за θ мора бити $\max\limits_{1\le j\le n} x_j$, па је највећа статистика поретка оцена параметра θ ММВ, тј.

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{X_{(n)}}.$$

△

Слична ситуација појављује се кад год је носач расподеле обележја од интереса функција параметра који се оцењује.

Пример 4 (додатак – пример пристрасне али постојане оцене)

Посматра се оцена $\hat{\theta} = X_{(n)}$ на основу простог сл. узорка обима n из $U(0,\theta)$ расподеле.

Важи:

$$E\hat{\theta} = \frac{n}{n+1}\theta,$$

па ова оцена није непристрасна (иако јесте асимптотски непристрасна, при $n \to +\infty$).

$\max_{1\le j\le n} X_j \in (0,\theta)$

Са друге стране, за $\epsilon \in (0,\theta)$,

$$P\{|\hat{\theta}-\theta| \le \epsilon\} = P\{\theta-\epsilon \le \hat{\theta} \le \theta\} = \int_{\theta-\epsilon}^{\theta} f_{\hat{\theta}}(x)dx = \int_{\theta-\epsilon}^{\theta} \frac{nx^{n-1}}{\theta^n}dx$$

$$= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n.$$

За остале $\epsilon > 0$ ова вероватноћа једнака је 1. Стога је за $\forall \epsilon > 0$:

$$\lim_{n\to+\infty} P\{|\hat{\theta}-\theta| > \epsilon\} = 0,$$

што показује да је ова оцена постојана.

Пример 4 (додатак – ефикасност оцене)

Посматрају се две оцене: $\hat{\theta}_1 = \frac{n+1}{n} X_{(n)}$ и $\hat{\theta}_2 = 2\bar{X}_n$, на основу простог сл. узорка обима n из $U(0,\theta)$ расподеле.

Обе оцене су непристрасне. Која је боља?!

Да би се добио одговор на ово питање требало би упоредити њихове дисперзије:

$$D\hat{\theta}_1 = \frac{1}{n(n+2)}\theta^2,$$

$$D\hat{\theta}_2 = \frac{1}{3n}\theta^2.$$

За сваки природан број $n \in \mathbb{N}$: $D\hat{\theta}_1 \leq D\hat{\theta}_2$, па је оцена $\hat{\theta}_1$ боља (у смислу, ефикаснија) од оцене $\hat{\theta}_2$.