

## 11. Gyakorlat

Alapstatisztikák, paraméterbecslés, MLE

### Alapstatisztikák

**Definíció (Minta és statisztika)**  $\underline{X}_n = (X_1, \dots, X_n)$  egy  $n$  elemű mintának nevezzük, ha  $X_1, \dots, X_n$  független azonos eloszlású (f.a.e.) valószínűségi változók. A minta egy realizációja  $\underline{x}_n = (x_1, \dots, x_n) \in \mathbb{R}^n$ . A  $T_n : \mathbb{R}^n \rightarrow \mathbb{R}$  statisztika a minta egy függvénye:  $T_n(\underline{X}_n)$ .

**Definíció (Mintaátlag, empirikus variancia, korrigált empirikus variancia)** Az  $\underline{X}_n = (X_1, \dots, X_n)$  minta esetén alapvető statisztikák

- A mintaátlag

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k.$$

- Az empirikus variancia és szórás

$$S_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 \quad \text{és} \quad S_n = \sqrt{S_n^2}.$$

- A korrigált empirikus variancia és szórás

$$S_n^{*2} = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2 \quad \text{és} \quad S_n^* = \sqrt{S_n^{*2}}.$$

Egy konkrét realizáción felvett értékek jelölése:

$$\bar{x}_n = \frac{1}{n} \sum_{k=1}^n x_k, \quad s_n^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x}_n)^2, \quad s_n = \sqrt{s_n^2} \quad s_n^{*2} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}_n)^2, \quad s_n^* = \sqrt{s_n^{*2}}.$$

### Paraméterbecslés

**Definíció (Torzítatlan statisztika)** Legyen  $\underline{X}_n = (X_1, \dots, X_n)$  minta egy  $\theta$  paraméterű eloszláscsaládból. Azt mondjuk, hogy a  $T_n$  statisztika torzítatlan  $f(\theta)$ -ra, ha

$$\mathbb{E}[T_n(\underline{X}_n)] = f(\theta).$$

**Állítás** Legyen  $\underline{X}_n = (X_1, \dots, X_n)$  minta  $X$  egy  $\theta$  paraméteres eloszlásából. Ekkor  $\bar{X}_n$  torzítatlan becslés  $\mathbb{E}_\theta(X)$ -re,  $S_n^{*2}$  torzítatlan becslés  $\mathbb{D}_\theta^2(X)$ -re és  $S_n^*$  torzítatlan becslés  $\mathbb{D}_\theta(X)$ -re.

**Definíció (Maximum likelihood becslés - MLE)** Legyen  $\underline{X}_n = (X_1, \dots, X_n)$  minta egy  $\theta$  paraméterű eloszláscsaládból. Ekkor  $\theta$  maximum likelihood becslése az a  $\hat{\theta}$ , amire a likelihood függvénye  $\mathcal{L}_\theta : \mathbb{R}^n \rightarrow \mathbb{R}_0^+$  maximális (amennyiben létezik) az adott  $\underline{x}_n = (x_1, \dots, x_n)$  realizáción.

- Amennyiben az eloszlás diszkrét

$$\mathcal{L}_\theta(x_1, \dots, x_n) = \prod_{k=1}^n p_\theta(x_k), \quad \text{ahol } p_\theta \text{ az eloszlás súlyfüggvénye.}$$

- Amennyiben az eloszlás folytonos

$$\mathcal{L}_\theta(x_1, \dots, x_n) = \prod_{k=1}^n f_\theta(x_k), \quad \text{ahol } f_\theta \text{ az eloszlás sűrűségfüggvénye.}$$

**Megjegyzés** Mivel  $\ln$  monoton függvény, a maximumhelyhez elég a log-likelihood függvény maximumát megkeresni. A log-likelihood függvény

$$l_\theta = \ln(\mathcal{L}_\theta).$$

**Feladatok**

1. Legyen  $X$  eloszlása az alábbi egy ismeretlen  $\theta$  paraméter esetén

$$\mathbb{P}_\theta(X=0) = \frac{2\theta}{3}, \quad \mathbb{P}_\theta(X=2) = \frac{1-\theta}{3}, \quad \mathbb{P}_\theta(X=3) = \frac{2-\theta}{3}.$$

Tekintsük az  $\underline{x}_5 = (3, 0, 2, 0, 2)$  mintarealizációt az eloszlásból.

- (a) Adjunk a realizáció alapján torzítatlan becslést  $X$  várható értékére és szórására.  
 (b) Adjunk a realizáció alapján maximum likelihood becslést  $\theta$ -ra.

**Megoldás**

- (a) Torzítatlan becslés a várható értékre a mintaátlag az adott mintarealizáción

$$\bar{x}_5 = \frac{1}{5}(3 + 0 + 2 + 0 + 2) = \frac{7}{5} = 1.4.$$

Torzítatlan becslés a szórásra a korrigált empirikus szórás az adott mintarealizáción

$$s_5^* = \sqrt{\frac{1}{5-1} ((3-1.4)^2 + (0-1.4)^2 + (2-1.4)^2 + (0-1.4)^2 + (2-1.4)^2)} \\ \approx 1.3416.$$

- (b) A likelihood függvény az adott realizáción

$$\mathcal{L}_\theta(3, 0, 2, 0, 2) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{1-\theta}{3}\right)^2 \left(\frac{2-\theta}{3}\right)^1.$$

Ekkor a log-likelihood függvény

$$l_\theta(3, 0, 2, 0, 2) = 2 \ln \left(\frac{2\theta}{3}\right) + 2 \ln \left(\frac{1-\theta}{3}\right) + \ln \left(\frac{2-\theta}{3}\right).$$

Ekkor ennek a függvénynek kell megkeresni a maximumhelyét a  $0 < \theta < 1$  intervallumon (mivel ilyen  $\theta$  értékekre lesz a feladat szövegében szereplő függvény valóban súlyfüggvény). A log-likelihood függvény első deriváltja  $\theta$  szerint

$$\frac{d}{d\theta} l_\theta(3, 0, 2, 0, 2) = 2 \cdot \frac{3}{2\theta} \cdot 2 + 2 \cdot \frac{3}{1-\theta} \cdot (-1) + \frac{3}{2-\theta} \cdot (-1) \\ = \frac{6}{\theta-1} + \frac{6}{\theta} + \frac{3}{\theta-2}.$$

Ez alapján keressük meg, hol lehet lokális szélsőértéke

$$\frac{6}{\theta-1} + \frac{6}{\theta} + \frac{3}{\theta-2} = 0 \\ 15\theta^2 - 33\theta + 12 = 0,$$

aminek a gyökei

$$\theta_1 = \frac{11 - \sqrt{41}}{10} \approx 0.4597 \quad \text{és} \quad \theta_2 = \frac{11 + \sqrt{41}}{10} \approx 1.7403.$$

Mivel  $0 < \theta < 1$ , ezért csak

$$\theta_1 = \frac{11 - \sqrt{41}}{10} \approx 0.4597$$

jöhet szóba, mint lokális szélsőérték hely. Most vizsgáljuk meg, hogy valóban maximum hely-e. A log-likelihood függvény második deriváltja  $\theta$  szerint

$$\frac{d^2}{d\theta^2} = -\frac{6}{(\theta-1)^2} - \frac{6}{\theta^2} - \frac{3}{(\theta-2)^2} < 0$$

minden  $0 < \theta < 1$  esetén, így ez lokális maximum lesz. Végül vizsgáljuk meg a határokat

$$\lim_{\theta \rightarrow 0+} l_{\theta}(3, 0, 2, 0, 2) = -\infty \quad \text{és} \quad \lim_{\theta \rightarrow 1-} l_{\theta}(3, 0, 2, 0, 2) = -\infty,$$

így ez egy globális maximum  $0 < \theta < 1$  esetén.

Azt kaptuk, hogy a MLE becslés az adott realizáció esetén  $\theta$ -ra

$$\hat{\theta} = \frac{11 - \sqrt{41}}{10} \approx 0.4597.$$

Ekkor az eloszlásra az alábbi a becslésünk

$$\mathbb{P}_{\hat{\theta}}(X = 0) \approx 0.3065, \quad \mathbb{P}_{\hat{\theta}}(X = 2) \approx 0.1801 \quad \mathbb{P}_{\hat{\theta}}(X = 3) \approx 0.5134.$$

2. Egy  $0 < \theta < 1$  valószínűséggel sikeres kísérletet az első sikerig ismétlünk egymástól függetlenül.

(a) Adjunk maximum likelihood becslést  $\theta$ -ra tetszőleges realizáció esetén.

(b) Az  $\underline{x}_{10} = (4, 3, 4, 1, 10, 6, 5, 2, 2, 3)$  10 darab megfigyelés alapján adjunk becslést  $\theta$ -ra.

### Megoldás

(a) A feladat szövege szerint  $X \sim \text{Geom}(\theta)$ -ra kell maximum likelihood becslést adni tetszőleges  $\underline{x}_n = (x_1, \dots, x_n)$  mintarealizáció esetén. Ekkor a likelihood függvény

$$\mathcal{L}_{\theta}(\underline{x}_n) = \prod_{j=1}^n \mathbb{P}(X = x_j) = \prod_{j=1}^n \theta(1 - \theta)^{x_j-1}.$$

A log-likelihood függvény

$$\begin{aligned} l_{\theta}(\underline{x}_n) &= \sum_{j=1}^n \ln(\theta) + (x_j - 1) \ln(1 - \theta) \\ &= n \ln(\theta) - n \ln(1 - \theta) + \ln(1 - \theta) \sum_{j=1}^n x_j. \end{aligned}$$

Ekkor az első deriváltja  $\theta$  szerint

$$\begin{aligned} \frac{d}{d\theta} l_{\theta}(\underline{x}_n) &= \frac{n}{\theta} + \frac{n}{1 - \theta} - \frac{1}{1 - \theta} \sum_{j=1}^n x_j \\ &= \frac{n}{\theta(1 - \theta)} - \frac{1}{1 - \theta} \sum_{j=1}^n x_j, \end{aligned}$$

ami pontosan akkor nulla, ha

$$\theta = \frac{n}{\sum_{j=1}^n x_j} = \frac{1}{\frac{1}{n} \sum_{j=1}^n x_j} = \frac{1}{\bar{x}_n},$$

vagyis amennyiben  $\theta$  a mintaátlag reciproka. Vizsgáljuk meg a második deriváltat

$$\frac{d^2}{d\theta^2} l_{\theta}(\underline{x}_n) = -\frac{n}{\theta^2} - \frac{1}{(1 - \theta)^2} \underbrace{\sum_{j=1}^n (x_j - 1)}_{\geq 0} < 0,$$

minden  $0 < \theta < 1$  esetén, így itt lokális maximuma lesz. A határookban

$$\lim_{\theta \rightarrow 0+} l_{\theta}(\underline{x}_n) = -\infty \quad \text{és} \quad \lim_{\theta \rightarrow 1-} l_{\theta}(\underline{x}_n) = -\infty,$$

így ez globális maximum  $0 < \theta < 1$  esetén, így az MLE

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

(b) Az előzőek alapján elég a mintaátlagot kiszámolni. Ekkor

$$\bar{x}_{10} = \frac{1}{10}(4 + 3 + 4 + 1 + 10 + 6 + 5 + 2 + 2 + 3) = 4.$$

Ekkor az MLE ezen a realizáción az előző feladatrészt szerint

$$\hat{\theta} = \frac{1}{\bar{x}_{10}} = \frac{1}{4}.$$

3. Legyen  $X$  eloszlása egy ismeretlen  $0 < \theta < 1$  paraméter esetén az alábbi.

$$\mathbb{P}_{\theta}(X = k) = (k-1)\theta^2(1-\theta)^{k-2}, \quad k = 2, 3, 4, \dots$$

(a) Mi az eloszlása  $X$ -nek?

(b) Adjunk maximum likelihood becslést  $\theta$ -ra tetszőleges realizáció esetén.

### Megoldás

(a) A súlyfüggvény miatt  $X \sim \text{Negbinom}(2, \theta)$ , ahol  $0 < \theta < 1$ .

(b) A likelihood függvény tetszőleges  $\underline{x}_n = (x_1, \dots, x_n)$  mintarealizáció esetén

$$\mathcal{L}_{\theta}(\underline{x}_n) = \prod_{j=1}^n \mathbb{P}(X = x_j) = \prod_{j=1}^n (x_j - 1)\theta^2(1 - \theta)^{x_j - 2}.$$

A log-likelihood függvény

$$\begin{aligned} l_{\theta}(\underline{x}_n) &= \sum_{j=1}^n [\ln(x_j - 1) + 2\ln(\theta) + (x_j - 2)\ln(1 - \theta)] \\ &= 2n\ln(\theta) - 2n\ln(1 - \theta) + \sum_{j=1}^n (x_j - 1) + \ln(1 - \theta) \sum_{j=1}^n x_j. \end{aligned}$$

Ekkor az első deriváltja  $\theta$  szerint

$$\begin{aligned} \frac{d}{d\theta} l_{\theta}(\underline{x}_n) &= \frac{2n}{\theta} + \frac{2n}{1 - \theta} - \frac{1}{1 - \theta} \sum_{j=1}^n x_j \\ &= \frac{2n}{\theta(1 - \theta)} - \frac{1}{1 - \theta} \sum_{j=1}^n x_j, \end{aligned}$$

ami pontosan akkor nulla, ha

$$\theta = \frac{2n}{\sum_{j=1}^n x_j} = \frac{2}{\frac{1}{n} \sum_{j=1}^n x_j} = \frac{2}{\bar{x}_n},$$

vagyis amennyiben  $\theta$  a mintaátlag reciproka. Vizsgáljuk meg a második deriváltat

$$\frac{d^2}{d\theta^2} l_{\theta}(\underline{x}_n) = -\frac{2n}{\theta^2} - \frac{1}{(1 - \theta)^2} \underbrace{\sum_{j=1}^n (x_j - 2)}_{\geq 0} < 0,$$

minden  $0 < \theta < 1$  esetén, így itt lokális maximuma lesz. A hátárokból

$$\lim_{\theta \rightarrow 0+} l_{\theta}(\underline{x}_n) = -\infty \quad \text{és} \quad \lim_{\theta \rightarrow 1-} l_{\theta}(\underline{x}_n) = -\infty,$$

így ez globális maximum  $0 < \theta < 1$  estén, így az MLE

$$\hat{\theta} = \frac{2}{\bar{x}_n}.$$

4. Legyen  $\underline{X}_n = (X_1, \dots, X_n)$  egy  $n$  elemű f.a.e. minta, ahol  $\theta > 0$  paraméter esetén a mintaelemek sűrűségfüggvénye.

$$f_\theta(x) = \begin{cases} \theta^2 x e^{-\theta x} & \text{ha } x > 0 \\ 0 & \text{egyébként} \end{cases}.$$

Adjunk maximum likelihood becslést  $\theta$ -ra tetszőleges realizáció esetén.

**Megoldás** A likelihood függvény tetszőleges  $\underline{x}_n = (x_1, \dots, x_n)$  realizáció esetén

$$\mathcal{L}_\theta(\underline{x}_n) = \prod_{j=1}^n \theta^2 x_j e^{-\theta x_j},$$

amennyiben minden  $j = 1, \dots, n$  esetén  $x_j > 0$ . Más esetben konstans nulla, így az MLE nem létezik. Haladjunk tovább az előző esettel. Ekkor a log-likelihood függvény

$$\begin{aligned} l_\theta(\underline{x}_n) &= \sum_{j=1}^n [2 \ln(\theta) + \ln(x_j) - \theta x_j] \\ &= 2n \ln(\theta) + \sum_{j=1}^n \ln(x_j) - \theta \sum_{j=1}^n x_j. \end{aligned}$$

Az első derivált

$$\frac{d}{d\theta} l_\theta(\underline{x}_n) = \frac{2n}{\theta} - \sum_{j=1}^n x_j,$$

ami pontosan akkor nulla, ha

$$\theta = \frac{2n}{\sum_{j=1}^n x_j} = \frac{2}{\bar{x}_n} > 0.$$

Továbbá a második derivált

$$\frac{d^2}{d\theta^2} l_\theta(\underline{x}_n) = -\frac{2n}{\theta^2} < 0,$$

minden  $\theta > 0$  esetén, így ez lokális maximum lesz. Továbbá a határokon

$$\lim_{\theta \rightarrow 0_+} l_\theta(\underline{x}_n) = -\infty,$$

így globális a maximumhely. Az MLE tehát

$$\hat{\theta} = \frac{2}{\bar{x}_n}.$$

5. Legyen  $\underline{X}_n = (X_1, \dots, X_n)$  egy  $n$  elemű f.a.e. minta egy  $\theta > 0$  paraméterű eloszlásból. Adjunk maximum likelihood becslést  $\theta$ -ra tetszőleges realizáció esetén.

- (a) Amennyiben az eloszlás  $\text{Exp}(\theta)$ .
- (b) Amennyiben az eloszlás  $\text{Uni}(0, \theta)$ .

**Megoldás**

- (a) Az  $\text{Exp}(\theta)$  eloszlás sűrűségfüggvénye

$$f_\theta(x) = \begin{cases} \theta e^{-\theta x} & \text{ha } x > 0 \\ 0 & \text{egyébként} \end{cases}.$$

A likelihood függvény tetszőleges  $\underline{x}_n = (x_1, \dots, x_n)$  realizáció esetén

$$\mathcal{L}_\theta(\underline{x}_n) = \prod_{j=1}^n \theta e^{-\theta x_j},$$

amennyiben minden  $j = 1, \dots, n$  esetén  $x_j > 0$ . Más esetben konstans nulla, így az MLE nem létezik. Haladjunk tovább az előző esettel. Ekkor a log-likelihood függvény

$$\begin{aligned} l_\theta(\underline{x}_n) &= \sum_{j=1}^n [\ln(\theta) - \theta x_j] \\ &= n \ln(\theta) - \theta \sum_{j=1}^n x_j. \end{aligned}$$

Az első derivált

$$\frac{d}{d\theta} l_\theta(\underline{x}_n) = \frac{n}{\theta} - \sum_{j=1}^n x_j,$$

ami pontosan akkor nulla, ha

$$\theta = \frac{n}{\sum_{j=1}^n x_j} = \frac{1}{\bar{x}_n} > 0.$$

Továbbá a második derivált

$$\frac{d^2}{d\theta^2} l_\theta(\underline{x}_n) = -\frac{n}{\theta^2} < 0,$$

minden  $\theta > 0$  esetén, így ez lokális maximum lesz. Továbbá a határokon

$$\lim_{\theta \rightarrow 0_+} l_\theta(\underline{x}_n) = -\infty,$$

így globális a maximumhely. Az MLE tehát

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

(b) A  $\text{Uni}(0, \theta)$  eloszlás sűrűségfüggvénye

$$f_\theta(x) = \begin{cases} \theta^{-1} & \text{ha } 0 \leq x \leq \theta \\ 0 & \text{egyébként} \end{cases}.$$

A likelihood függvény tetszőleges  $\underline{x}_n = (x_1, \dots, x_n)$  realizáció esetén

$$\mathcal{L}_\theta(\underline{x}_n) = \begin{cases} \theta^{-n} & \text{ha } 0 \leq x_j \leq \theta \\ 0 & \text{egyébként} \end{cases}.$$

Ekkor amennyiben létezik  $j = 1, \dots, n$ , hogy  $x_j < 0$ , akkor az MLE nem létezik, mivel a likelihood függvény konstans nulla. Amennyiben minden  $j = 1, \dots, n$  esetén  $0 \leq x_j$ , akkor vegyük észre, hogy  $\theta \mapsto \theta^{-n}$  szigorúan monoton csökken. Továbbá minden  $j = 1, \dots, n$  esetén  $x_j \leq \theta$ , ezért a legkisebb  $\theta$ , amire teljesül az egyenlőtlenség

$$\theta = \max \{x_1, \dots, x_n\},$$

ezért itt lesz maximális  $\mathcal{L}_\theta(\underline{x}_n)$  ebben az esetben. Vagyis ekkor az MLE

$$\hat{\theta} = \max \{x_1, \dots, x_n\}.$$