Matematika A4 - Valószínűségszámítás 10. gyakorlat
ML-becslés, konfidenciaintervallum

Paraméterbecslés

Legyen $\mathcal{X}_n = (X_1, X_2, \ldots, X_n)$ egy n elemű minta! Ez alapján akarjuk az X valószínűségi változó θ paraméterét becsülni.

- Átlag: $\overline{\mathcal{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ torzítatlan becslés $\mathbb{E}(X)$-re
- Tapasztalati szórás: $\hat{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathcal{X}}_n)^2}$ torzított becslés $\mathbb{D}(X)$-re
- Torzítatlan becslés szórásra: $\hat{s}^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{\mathcal{X}}_n)^2}$ (1)

Maximum Likelihood becslés

Azt a θ értéket akarjuk meghatározni, amire a látott $\mathcal{X}_n$ realizáció a legvalószínűbb. Ehhez maximalizálni kell θ-ban az $l(\mathcal{X}_n, \theta)$ **likelihood függvény**t:

- Diszkrét eset: $l(\mathcal{X}_n, \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i \,|\, \theta)$, ahol $\mathbb{P}(X_i \,|\, \theta)$ annak a valószínűsége, hogy ha az igazi paraméter θ, akkor X_i-t látunk
- Folytonos eset: $l(\mathcal{X}_n, \theta) = \prod_{i=1}^{n} f(X_i \,|\, \theta)$, ahol $f(X_i \,|\, \theta)$ az X változó sűrűségfüggvénye X_i-ben, ha az igazi paraméter θ

Ehhez elég az $L(\mathcal{X}_n, \theta) = \ln(l(\mathcal{X}_n, \theta))$ **log likelihood** függvényt maximalizálni. Vagyis keressük azt a $\hat{\theta}$ paramétert, amire $\frac{\partial}{\partial\theta} L\left(\mathcal{X}_n, \hat{\theta}\right) = 0$. Azaz

- Diszkrét: $\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\left(\mathbb{P}\left(X_i \,|\, \hat{\theta}\right)\right) = 0$
- Folytonos: $\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\left(f\left(X_i \,|\, \hat{\theta}\right)\right) = 0$

1. feladat Tegyük fel, hogy X súlyfüggvénye az ismeretlen θ paraméter függvényében az alábbi:

$$\mathbb{P}(X = 0 \,|\, \theta) = \frac{5\theta}{4}, \quad \mathbb{P}(X = 2 \,|\, \theta) = \frac{2 - 2\theta}{4}, \quad \mathbb{P}(X = 4 \,|\, \theta) = \frac{1 - 2\theta}{4}, \quad \mathbb{P}(X = 5 \,|\, \theta) = \frac{1 - \theta}{4}.$$

Az alábbi mintát látjuk: $\mathcal{X}_8 = (0, 2, 4, 4, 5, 2, 5, 5)$.

(a) Mennyi a minta átlaga, tapasztalati szórása? Adjunk a szórásra torzítatlan becslést!

(b) Adjunk ML becslést θ paraméterre!

Megoldásvázlat

(a) Átlag: $\overline{\mathcal{X}}_8 = \frac{\sum_{i=1}^{8} X_i}{8} = \frac{27}{8}$.

Tapasztalati szórás: $\hat{s} = \sqrt{\frac{1}{8}\sum_{i=1}^{8}\left(X_i - \frac{27}{8}\right)^2} = \frac{\sqrt{191}}{8}$.

Torzítatlan becslés szórásra: $\hat{s}^* = \sqrt{\frac{1}{7}\sum_{i=1}^{8}\left(X_i - \frac{27}{8}\right)^2} = \sqrt{\frac{n}{n-1}} \cdot \hat{s} = \sqrt{\frac{191}{56}}$.

(b) Először vegyük észre, hogy a valószínűségek nemnegativitása miatt $0 \leq \theta \leq \frac{1}{2}$. A log likelihood függvény:

$$L(\mathcal{X}_8, \theta) = \ln\left(\frac{5\theta}{4}\right) + 2\ln\left(\frac{2-2\theta}{4}\right) + 2\ln\left(\frac{1-2\theta}{4}\right) + 3\ln\left(\frac{1-\theta}{4}\right)$$

A $\frac{\partial}{\partial\theta}L(\mathcal{X}_8, \theta) = 0$ feltételből azt kapjuk, hogy

$$\frac{1}{\theta} - \frac{2}{1-\theta} - \frac{4}{1-2\theta} - \frac{3}{1-\theta} = 0.$$

Közös nevezőre hozás és a nevezővel szorzás után azt kapjuk, hogy $16\theta^2 - 12\theta + 1 = 0$. Ennek gyökei $\frac{3\pm\sqrt{5}}{8}$, amiből a nagyobb több, mint $\frac{1}{2}$, ami nem lehet, hiszen ekkor $\mathbb{P}(X = 4|\theta) < 0$ lenne. Vagyis $\hat{\theta} = \frac{3-\sqrt{5}}{8}$.

2. feladat Legyen $f(x) = a^2 x\mathrm{e}^{-ax}$, ha $x \geq 0$. Adjunk ML becslést az a paraméterre tetszőleges $\mathcal{X}_n = (X_1, \ldots, X_n)$ minta esetén!

Megoldásvázlat Ekkor $\ln f(x|a) = 2\ln a + \ln x - ax$, vagyis $L(\mathcal{X}_n, a) = 2n\ln a + \sum_{i=1}^{n} \ln x_i - a\sum_{i=1}^{n} x_i$.

A $\frac{\partial}{\partial a}L(\mathcal{X}_n, a) = 0$ feltételből $\frac{2n}{a} - \sum_{i=1}^{n} x_i = 0$. Vagyis $\hat{a} = \frac{2n}{\sum_{i=1}^{n} x_i} = \frac{2}{\overline{\mathcal{X}}_n}$.

3. feladat Egy berendezés élettartamára a következő mintánk van (években):

$$\mathcal{X}_{16} = (2,\ 3,\ 4,\ 2.5,\ 3,\ 3,\ 3.5,\ 4,\ 2,\ 4,\ 3,\ 3,\ 3.5,\ 2,\ 3,\ 4).$$

(a) Mennyi a minta átlaga?

(b) Adjunk ML becslést λ-ra, ha feltesszük, hogy a berendezés élettartamának eloszlása $\mathrm{EXP}(\lambda)$!

(c) Adjunk ML becslést μ-re, ha feltesszük, hogy a berendezés élettartamának eloszlása $\mathcal{N}(\mu, 1)$!

Megoldásvázlat

(a) $\overline{\mathcal{X}}_{16} = \frac{49.5}{16} = \frac{99}{32}$

(b) Ekkor $\ln f(x|\lambda) = \ln\lambda - \lambda x$, így $\frac{\partial}{\partial\lambda}(\ln f(x\,|\,\lambda)) = \frac{1}{\lambda} - x$. Tehát

$$\frac{\partial}{\partial\lambda}L(\mathcal{X}_{16}, \lambda) = \sum_{i=1}^{16}\left(\frac{1}{\lambda} - x_i\right) = \frac{16}{\lambda} - 49.5.$$

A $\frac{\partial}{\partial\lambda}L(\mathcal{X}_{16}, \lambda) = 0$ feltételből $\hat{\lambda} = \frac{16}{49.5} = \frac{1}{\overline{\mathcal{X}}_{16}}$.

(c) Ekkor $\ln f(x\,|\,\mu) = \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2}$, így $\frac{\partial}{\partial\mu}(\ln f(x|\mu)) = x - \mu$. Tehát

$$\frac{\partial}{\partial\mu}L(\mathcal{X}_{16}, \mu) = \sum_{i=1}^{16}(x_i - \mu) = 49.5 - 16\mu.$$

A $\frac{\partial}{\partial\mu}L(\mathcal{X}_{16}, \mu) = 0$ feltételből $\hat{\mu} = \frac{49.5}{16} = \overline{\mathcal{X}}_{16}$.

4. feladat Egy cinkelt érmét dobálunk, ahol a fej valószínűsége p.

(a) A következő adatsorunk van az első fej megjelenésére:

$$\mathcal{X}_{16} = (2,\ 3,\ 1,\ 2,\ 3,\ 3,\ 2,\ 1,\ 4,\ 4,\ 4,\ 2,\ 3,\ 2,\ 1,\ 4).$$

Adjunk ML becslést p-re!

(b) Mi az ML becslés p-re tetszőleges $\mathcal{X}_n = (a_1, a_2, \ldots, a_n)$ minta esetén?

Megoldásvázlat Legyen X az első fej megjelenéséig a dobások száma. Ekkor $X \sim \mathrm{GEO}(p)$, azaz $\mathbb{P}(X = k \,|\, p) = (1-p)^{k-1} \cdot p$. Így $\ln(\mathbb{P}(X = k \,|\, p)) = (k-1)\ln(1-p) + \ln p$. Ezt deriválva

$$\frac{\partial}{\partial p} \ln\left(\mathbb{P}(X = k \,|\, p)\right) = -\frac{k-1}{1-p} + \frac{1}{p} = \frac{1-kp}{p(1-p)}.$$

(a) Tehát $\frac{\partial}{\partial p} L(\mathcal{X}_{16}, p) = \sum\limits_{i=1}^{16} \frac{1-k_i p}{p(1-p)} = \frac{16}{p(1-p)} - \frac{41}{1-p}$.

A $\frac{\partial}{\partial p} L(\mathcal{X}_{16}, p) = 0$ feltételből $\hat{p} = \frac{16}{41}$.

(b) A fenti számolás alapján $\frac{\partial}{\partial p} L(\mathcal{X}_n, p) = \sum\limits_{i=1}^{n} \frac{1-a_i p}{p(1-p)} = \frac{n}{p(1-p)} - \frac{\sum_{i=1}^{n} a_i}{1-p}$.

A $\frac{\partial}{\partial p} L(\mathcal{X}_n, p) = 0$ feltételből $\hat{p} = \frac{n}{\sum_{i=1}^{n} a_i} = \frac{1}{\overline{\mathcal{X}}_n}$.

Konfidenciaintervallum

Ismeretlen θ paraméter lehetséges értékére egy intervallumot adunk $\mathcal{X}_n$ minta alapján, amiben az elvárt valószínűséggel beleesik az igazi értéke a paraméternek:

$$\mathbb{P}(l(\mathcal{X}_n) \le \theta \le u(\mathcal{X}_n)) = 1 - \alpha.$$

$[l(\mathcal{X}_n),\ u(\mathcal{X}_n)]$ a **konfidenciaintervallum**, $1-\alpha$ a **konfidenciaszint**.

Kétoldali Z-próba

$X \sim \mathcal{N}(\mu,\ \sigma^2)$, ahol σ ismert, μ ismeretlen. Ekkor $\mathcal{X}_n$ n elemű minta alapján μ-re az $1-\alpha$ szinthez tartozó konfidenciaintervallum:

$$\left(\overline{\mathcal{X}}_n - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}},\ \overline{\mathcal{X}}_n + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right),$$

ahol $z_{\alpha/2}$ az az érték, aminél csak $\alpha/2$ valószínűséggel nagyobb egy standard normális eloszlású változó, azaz $\Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$.

Kétoldali T-próba

Ugyanaz a feladat, mint kétoldali Z-próbánál, de ismeretlen σ-val. Ekkor μ-re az $1-\alpha$ szinthez tartozó konfidenciaintervallum:

$$\left(\overline{\mathcal{X}}_n - \frac{t_{\alpha/2,n-1} \cdot \hat{s}^*}{\sqrt{n}},\ \overline{\mathcal{X}}_n + \frac{t_{\alpha/2,n-1} \cdot \hat{s}^*}{\sqrt{n}}\right),$$

ahol $\hat{s}^*$ torzítatlan becslés a szórásra (lásd fent (1) képlet), míg $t_{\alpha/2,n-1}$ az az érték, aminél csak $\alpha/2$ valószínűséggel nagyobb egy $n-1$ szabadságfokú T-eloszlású változó. Azaz $T_{n-1}\left(t_{\alpha/2,n-1}\right) = 1 - \frac{\alpha}{2}$, ahol T_{n-1} jelöli az $(n-1)$ szabadságfokú T-eloszlás eloszlásfüggvényét. Ennek az értékei is táblázatból olvashatóak ki.

Egyoldali Z-próba

Hasonló a feladat, mint kétoldali Z-próbánál, de most egy $(-\infty,\, a)$ (vagy $(a,\, \infty)$) alakú intervallumot akarunk meghatározni, amibe $1-\alpha$ valószínűséggel beleesik μ. Ekkor a megfelelő konfidenciaintervallum:

$$\left(-\infty\,,\, \overline{\mathcal{X}}_n + \frac{z_\alpha \cdot \sigma}{\sqrt{n}}\right) \qquad \left(\text{vagy } (a,\infty) \text{ esetén: } \left(\overline{\mathcal{X}}_n - \frac{z_\alpha \cdot \sigma}{\sqrt{n}}\,,\, \infty\right)\right),$$

ahol z_α az az érték, aminél csak α valószínűséggel nagyobb egy standard normális eloszlású változó, azaz $\Phi\left(z_\alpha\right) = 1-\alpha$.

5. feladat Egy X normális eloszlású véletlen változóval generáltuk az alábbi mintát:

$$\mathcal{X}_9 = (5,\, 6,\, 7,\, 4,\, 5.5,\, 6,\, 8,\, 7,\, 10).$$

(a) Adjunk X várható értékére 95%-os konfidenciaintervallumot!

(b) Hogyan változik az intervallum, ha megtudjuk, hogy az eloszlás szórása 2?

(c) Adjuk meg mindkét esetben a 99%-os konfidenciaintervallumot is!

(d) Mekkora az az érték, aminél 90%-os valószínűséggel kisebb a várható érték? (A szórásról továbbra is tudjuk, hogy 2.)

Megoldásvázlat

(a) Kétoldali T-próbát használunk, mert a szórás ismeretlen.

Az átlag 6.5, $\hat{s}^* \approx 1.77$ és $t_{0.025,8} = 2.306$. Így a keresett intervallum:

$$\left(6.5 - \frac{2.306 \cdot 1.77}{\sqrt{9}},\; 6.5 + \frac{2.306 \cdot 1.77}{\sqrt{9}}\right).$$

(b) Ha a szórás ismert, akkor kétoldali Z-próbát használhatunk. $z_{0.025} = 1.96$. Vagyis az intervallum:

$$\left(6.5 - \frac{1.96 \cdot 2}{\sqrt{9}},\; 6.5 + \frac{1.96 \cdot 2}{\sqrt{9}}\right).$$

(c) Most $t_{0.005,8} = 3.355$, míg $z_{0.005} = 2.58$, vagyis ismeretlen szórás esetén az intervallum

$$\left(6.5 - \frac{3.355 \cdot 1.77}{\sqrt{9}},\; 6.5 + \frac{3.355 \cdot 1.77}{\sqrt{9}}\right),$$

míg $\sigma = 2$ esetén

$$\left(6.5 - \frac{2.58 \cdot 2}{\sqrt{9}},\; 6.5 + \frac{2.58 \cdot 2}{\sqrt{9}}\right).$$

(d) Az egyoldali Z-próbánál látott képlettel: $\overline{\mathcal{X}}_n + \frac{z_\alpha \sigma}{\sqrt{n}} = 6.5 + \frac{1.28 \cdot 2}{\sqrt{9}}$.