

# 1. (6-7 hét) Leíró statisztikák, statisztikai alapfogalmak, becslések (maximum likelihood, momentum)

## Elmélet

**Definíció (Minta).**  $X_1, \dots, X_n$  valószínűségi változó sorozat. A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak. Realizációja:  $x_1, \dots, x_n$

**Definíció (Statisztika).** A minta valamely függvénye, pl.:

Mintaátlag v. átlag:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Tapasztalati szórás:  $S_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$

Korrigált tapasztalati szórás:  $S_n^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

Szórási együttható (vagy relatív szórás):  $V = \frac{S_n}{\bar{X}} = \frac{S_n}{\bar{X}} 100\%$  (az átlagtól való átlagos eltérés százalékban)  
/megjegyzés: lehet a korrigált tapasztalati szórással is számolni/

k-adik tapasztalati momentum ( $k \geq 1, k \in \mathbb{Z}$ ):  $m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$

Tapasztalati módusz: a legtöbbször előforduló érték

Rendezett minta:  $X_1^* \leq \dots \leq X_n^*$  a mintaelemek nem csökkenő sorrendben

Tapasztalati medián:  $X_{\frac{n+1}{2}}^*$ , ha  $n$  páratlan és  $\frac{X_{\frac{n}{2}}^* + X_{\frac{n}{2}+1}^*}{2}$ , ha  $n$  páros

Terjedelem:  $R = X_n^* - X_1^*$  (legnagyobb – legkisebb mintaelem)

z-kvantilis:  $q_z = \inf\{x : F(x) \geq z\}$ . Ha  $F$  invertálható, akkor  $q_z = F^{-1}(z)$ .

Tapasztalati z-kvantilis:  $q_z$  értelmezése: a mintaelemek  $z$ -ed része legfeljebb a  $q_z$ ,  $(1-z)$ -ed része pedig legalább a  $q_z$  értéket veszi fel ( $0 < z < 1$ ); sokféleképpen számolható, pl. interpolációs módszerrel: először megállapítjuk a sorszámot:  $(n+1)z = e + t$  ( $e$ : egészrész,  $t$ : törtrész), majd kiszámoljuk a  $z$ -kvantilist:  $q_z = X_e^* + t(X_{e+1}^* - X_e^*)$ .

Kvartilisek: Speciális kvartilisek, alsó (vagy első) kvantilis:  $Q_1 = q_{\frac{1}{4}}$ ,  
medián:  $Q_2 = q_{\frac{1}{2}}$ ,  
felső (vagy harmadik) kvantilis:  $Q_3 = q_{\frac{3}{4}}$

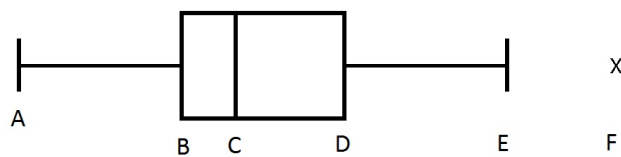
Interkvantilis terjedelem:  $IQR = q_{\frac{3}{4}} - q_{\frac{1}{4}} = Q_3 - Q_1$

Tapasztalati eloszlásfüggvény:  $F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i < x)$

ahol  $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \geq x \end{cases}$  indikátor függvény

**Tétel (Glivenko-Cantelli).** Az  $F_n(x)$  tapasztalati eloszlásfüggvény és az  $F(x)$  elméleti eloszlásfüggvény közötti eltérés maximuma 1 valószínűséggel 0-hoz konvergál, ami azt jelenti, hogy elég nagy minta esetén  $F_n(x)$  értéke minden  $x$ -re tetszőlegesen közel van  $F(x)$  értékéhez és  $n$ -et növelve mindenütt annak közelében marad.

**Definíció (Boxplot).**



$A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$ ,  $B = Q_1$ ,  $C = Q_2$ ,  $D = Q_3$ ,  $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$   
F: kieső értékek, azokat rünetjük fel pontokként, amik A-n vagy E-n kívülre esnek

Legyenek  $X_1, X_2, \dots, X_n$  független, azonos eloszlású valószínűségi változók (minta) egy  $\vartheta$  paraméterrel és legyen  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ . A becslés a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.

**Definíció** (Torzítatlan becslés). A  $\vartheta$  valós paraméter  $T(\mathbf{X})$  becslése torzítatlan, ha  $E(T(\mathbf{X})) = \vartheta$  minden  $\vartheta$  paraméterértékre.

**Definíció** (Likelihood függvény).  $L(\vartheta; \mathbf{x}) = f_{\vartheta}(\mathbf{x}) = \prod_{i=1}^n f_{\vartheta}(x_i)$ , ha az eloszlás abszolút folytonos

$$L(\vartheta; \mathbf{x}) = P_{\vartheta}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n P_{\vartheta}(X_i = x_i), \text{ ha az eloszlás diszkrét}$$

**Definíció** (Log-likelihood függvény).  $l(\vartheta; \mathbf{x}) = \ln(L(\vartheta; \mathbf{x}))$

### Paraméterbecslési módszerek:

#### Maximum likelihood módszer (ML-módszer):

Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel (azaz diszkrét esetben az ismeretlen paraméter azon értéket keressük, amely mellett a bekövetkezett eredmény maximális valószínűségű):  $\max_{\vartheta} L(\vartheta; \mathbf{x})$ . Ez nyilván megegyezik azzal a paraméterértékkel, ahol a log-likelihood függvény veszi fel a legnagyobb értéket, azaz:  $\max_{\vartheta} l(\vartheta; \mathbf{x})$ .

Amennyiben a függvény deriválható  $\vartheta$  szerint, akkor a maximumot kereshetjük a szokásos módon, a deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan  $n$ -szeres szorzatot kellene deriválni, amelyeknek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.

Ha  $\vartheta$  1 dimenziós, akkor  $\partial_{\vartheta} l(\vartheta, \mathbf{x}) = 0$ , míg ha  $\vartheta = (\vartheta_1, \dots, \vartheta_p)$   $p$  dimenziós, akkor  $\partial_{\vartheta_i} l(\vartheta, \mathbf{x}) = 0$  megoldásából kapjuk a becslést. (A második deriváltak segítségével ellenőrizzük, hogy valóban maximum.)

**Tétel** (ML-becslés invariáns tulajdonsága). Ha  $\vartheta$  ML-becslése  $\hat{\vartheta}$ , akkor tetszőleges  $g$  függvény esetén  $g(\vartheta)$  ML-becslése  $g(\hat{\vartheta})$ .

#### Momentum módszer:

A mintából számítható tapasztalati momentumokat ( $m_i := \frac{1}{n} \sum_j x_j^i$ ) egyenlővé tesszük az elméleti momentumokkal ( $M_i(\vartheta) := E_{\vartheta} X^i$ ), mégpedig annyit, amennyiből a paramétereket meg tudjuk határozni.  $p$  darab ismeretlen paraméter esetén tipikusan  $p$  ismeretlenes egyenletrendszert oldunk meg  $\vartheta$ -ra:  $M_1(\vartheta) = m_1, \dots, M_p(\vartheta) = m_p$  (megjegyzés:  $m_1 = \bar{x}$ )

### Feladatok

**1.1. Feladat.** Legyen  $X_1, \dots, X_n$  független, azonos eloszlású valószínűségi változók  $m$  várható értékkel. Célunk az ismeretlen  $m$  paraméter becslése. Tekintsük az alábbi statisztikákat és állapítsuk meg, hogy melyek torzítatlanok! Amelyik nem torzítatlan, hogyan tudnánk torzítatlanná tenni?

$$T_1(\mathbf{X}) = X_8, \quad T_2(\mathbf{X}) = \frac{X_9 + X_{19}}{9}, \quad T_3(\mathbf{X}) = \bar{X}$$

#### Megoldás

$E(T_1(\mathbf{X})) = E(X_8) = m$ , így  $T_1$  torzítatlan

$E(T_2(\mathbf{X})) = E\left(\frac{X_9 + X_{19}}{9}\right) = \frac{E(X_9) + E(X_{19})}{9} = \frac{2m}{9}$ , így  $T_2$  nem torzítatlan, viszont  $\frac{9}{2}T_2$  már igen

$E(T_3(\mathbf{X})) = E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n E(X_i)\right) = m$ , így  $T_3$  torzítatlan

**1.2. Feladat.** Adjon torzítatlan becslést a független, azonos  $E[0, \vartheta]$  eloszlású  $X_1, \dots, X_n$  minta  $\vartheta$  paraméterére a mintaátlag segítségével!

#### Megoldás

Mivel  $X_1, \dots, X_n \sim E[0, \vartheta]$ , így  $E(X_i) = \frac{\vartheta}{2}$ . Ekkor  $E(\bar{X}) = \frac{1}{n} E(\sum_{i=1}^n X_i) = \frac{n}{n} E(X_1) = \frac{\vartheta}{2}$ , tehát  $E(2\bar{X}) = \vartheta$ , vagyis  $2\bar{X}$  torzítatlan becslése  $\vartheta$ -nak.

**1.3. Feladat.** Legyen az alábbi gyakorisági tábla egy 20 elemű minta, a következő diszkrét eloszlásból:  
 $P(X_i = -1) = c, P(X_i = 1) = 3c, P(X_i = 2) = 1 - 4c$  ( $i = 1, \dots, 20$  és  $c$  az ismeretlen paraméter,  $0 < c < \frac{1}{4}$ ).

érték	-1	1	2
gyakoriság	4	10	6

Határozza meg  $c$  ML-becslését és  $c$  becslését a momentum módszerrel!

**Megoldás**

$c$  ML-becslése:

$$L(c, \mathbf{x}) = P(X_1 = x_1, \dots, X_{20} = x_{20}) = c^4(3c)^{10}(1 - 4c)^6$$

$$\ln L(c, \mathbf{x}) = 4 \ln(c) + 10 \ln(3c) + 6 \ln(1 - 4c)$$

$$(\ln L(c, \mathbf{x}))'_c = \frac{4}{c} + \frac{10}{c} - \frac{6 \cdot 4}{1 - 4c}$$

Átrendezve a  $(\ln L(c, \mathbf{x}))'_c = 0$  egyenletet, kapjuk, hogy

$$\begin{aligned} \frac{14}{c} - \frac{24}{1 - 4c} &= 0 \\ 14(1 - 4c) - 24c &= 0 \\ 14 &= 80c \end{aligned}$$

így  $\hat{c} = \frac{7}{40} = \frac{21}{120}$ . Ez valóban maximum, mivel  $(\ln L(c, \mathbf{x}))''_c$ -t kiértékelve a  $\hat{c}$  helyen  $(\ln L(c, \mathbf{x}))''_c = -\frac{14}{c^2} - \frac{96}{(1 - 4c)^2} < 0$ .

$c$  becslése momentum-módszerrel:

$$M_1(c) = EX = -1 \cdot c + 1 \cdot 3c + 2 \cdot (1 - 4c) = 2 - 6c, \quad m_1 = \frac{1}{20}(-1 \cdot 4 + 1 \cdot 10 + 2 \cdot 6) = 0,9$$

így az  $M_1(c) = m_1$  egyenletet  $c$ -re megoldva kapjuk, hogy  $\hat{c} = \frac{2 - 0,9}{6} = \frac{11}{60} = \frac{22}{120}$

Megjegyzés: Két különböző módszerrel különböző becslést kaptunk  $c$ -re, ami nem hiba, hanem a statisztikai módszerek sajátossága.

**1.4. Feladat.** Legyenek  $X_1, X_2, \dots, X_n$  független azonos eloszlású valószínűségi változók az alábbi eloszlásokból. Számolja ki az ismeretlen paraméter ML-becslését!

- a)  $Bin(m, p)$  binomiális eloszlás, ahol  $m \in \mathbb{N}$  adott és  $p$  a paraméter  
 b)  $Exp(\lambda)$  exponenciális eloszlás  
 c)  $N(\mu, \sigma^2)$  normális eloszlás, ahol  $\sigma \in \mathbb{N}$  adott és  $\mu$  a paraméter

**Megoldás**

a)

$$L(m, p; \mathbf{x}) = \prod_{k=1}^n \binom{m}{x_k} p^{x_k} (1 - p)^{m - x_k} \quad (x_k = 0, 1, \dots, m)$$

$$\ln L(m, p; \mathbf{x}) = \sum_{k=1}^n \ln \binom{m}{x_k} + \ln p \sum_{k=1}^n x_k + \ln(1 - p) \sum_{k=1}^n (m - x_k)$$

$$(\ln L(m, p; \mathbf{x}))'_p = \frac{1}{p} \sum_{k=1}^n x_k + \frac{-1}{1 - p} \sum_{k=1}^n (m - x_k) = \frac{1}{p} \sum_{k=1}^n x_k + \frac{-1}{1 - p} \left( nm - \sum_{k=1}^n x_k \right) = \frac{1}{p} n\bar{x} + \frac{-1}{1 - p} (nm - n\bar{x})$$

Átrendezve a  $(\ln L(m, p; \mathbf{x}))'_p = 0$  egyenletet, kapjuk, hogy

$$\begin{aligned} \frac{1}{p} n\bar{x} + \frac{-1}{1 - p} (nm - n\bar{x}) &= 0 \\ \frac{\bar{x}}{p} - \frac{m - \bar{x}}{1 - p} &= 0 \\ \bar{x} - p\bar{x} - pm + p\bar{x} &= 0 \end{aligned}$$

így  $\hat{p} = \frac{\bar{X}}{m}$ . Ez valóban maximum, mivel  $(\ln L(m, p))''_p$ -t kiértékelve a  $\hat{p}$  helyen  $(\ln L(m, p; \mathbf{x}))''_p = \frac{-n\bar{x}}{p^2} + \frac{-n(m-\bar{x})}{(1-p)^2} = -n \left( \frac{\bar{x}}{p^2} + \frac{m-\bar{x}}{(1-p)^2} \right) < 0$ .

b)

$$L(\lambda; \mathbf{x}) = \prod_{k=1}^n \lambda e^{-\lambda x_k} \quad (x_k > 0)$$

$$\ln L(\lambda; \mathbf{x}) = \sum_{k=1}^n \ln \lambda e^{-\lambda x_k} = \sum_{k=1}^n \ln \lambda + \sum_{k=1}^n \ln e^{-\lambda x_k} = n \ln \lambda - \lambda \sum_{k=1}^n x_k = n \ln \lambda - \lambda n \bar{x}$$

$$(\ln L(\lambda; \mathbf{x}))'_\lambda = \frac{n}{\lambda} - \sum_{k=1}^n x_k = \frac{n}{\lambda} - n\bar{x}$$

Átrendezve a  $(\ln L(\lambda; \mathbf{x}))'_\lambda = 0$  egyenletet, kapjuk, hogy  $\hat{\lambda} = \frac{1}{X}$ . Ez valóban maximum, mivel  $(\ln L(\lambda))''_\lambda = -\frac{n}{\lambda^2} < 0$ .

c)

$$L(\mu, \sigma^2; \mathbf{x}) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_i-\mu)^2}$$

$$\ln L(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{k=1}^n (x_i - \mu)^2$$

$$(\ln L(\mu, \sigma^2; \mathbf{x}))'_\mu = -\frac{1}{2\sigma^2} (-2) \sum_{k=1}^n (x_i - \mu)$$

Átrendezve a  $(\ln L(\mu, \sigma^2; \mathbf{x}))'_\mu = 0$  egyenletet, kapjuk, hogy  $\hat{\mu} = \frac{\sum_{k=1}^n X_i}{n} = \bar{X}$ . Ez valóban maximum, mivel  $(\ln L(\mu, \sigma^2; \mathbf{x}))''_\mu = -\frac{n}{\sigma^2} < 0$ .

**1.5. Feladat.** Határozza meg az ismeretlen paraméter ML-becslését, ha a minta  $E[a, 1]$  eloszlású!

**Megoldás**

A paraméter függvényében nem deriválható a likelihood függvény (ugrik):

$$\begin{aligned} L(a; \mathbf{x}) &= \prod_{i=1}^n \frac{1}{1-a} I(a \leq x_i \leq 1) = \frac{1}{(1-a)^n} I(a \leq x_1, x_2, \dots, x_n \leq 1) = \\ &= \frac{1}{(1-a)^n} I(a \leq x_1^* \leq \dots \leq x_n^* \leq 1) = \frac{1}{(1-a)^n} I(a \leq x_1^*) I(x_n^* \leq 1) \end{aligned}$$

Az  $I(a \leq x_1^*) I(x_n^* \leq 1)$  rész 0 vagy 1 lehet, tehát úgy kell megválasztani a paramétereiket, hogy 1 legyen:  $a \leq x_1^*$  és  $x_n^* \leq 1$  teljesüljön. Mivel a  $(-\infty, x_1^*]$  intervallumon az  $\frac{1}{(1-a)^n}$  függvény maximuma az  $a = x_1^*$  pontban van, így  $\hat{a} = X_1^*$ .

**1.6. Feladat.** Legyenek  $X_1, X_2, \dots, X_n$  független azonos  $E[a, b]$  eloszlású valószínűségi változók. Számolja ki az ismeretlen paraméterek becslését a momentum módszerrel!

**Megoldás**

$$M_1(a, b) = E(X) = \frac{a+b}{2}, \quad m_1 = \bar{x}$$

$$M_2(a, b) = E(X^2) = D^2(X) + E(X)^2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2, \quad m_2 = \frac{1}{n} \sum_{k=1}^n x_k^2$$

Így  $M_1(a, b) = m_1$  és  $M_2(a, b) = m_2$ -ből kapjuk, hogy

$$\frac{a+b}{2} = m_1$$

$$\frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2 = m_2$$

és ezt oldjuk meg  $a, b$ -re, először  $m_1$  és  $m_2$ -vel kifejezve. Átrendezve a fenti adja, hogy  $\frac{(b-a)^2}{12} = m_2 - m_1^2$ , így

$$b - a = \sqrt{12(m_2 - m_1^2)}$$

$$b + a = 2m_1.$$

Ezeket összeadva kapjuk, hogy  $b = m_1 + \sqrt{3(m_2 - m_1^2)}$  és  $a = m_1 - \sqrt{3(m_2 - m_1^2)}$ . Azaz a paraméterek becslése a momentum módszerrel:

$$\hat{a} = \bar{X} - \sqrt{3 \left( \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}^2 \right)} = \bar{X} - \sqrt{3} S_n \quad \text{és} \quad \hat{b} = \bar{X} + \sqrt{3 \left( \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}^2 \right)} = \bar{X} + \sqrt{3} S_n$$