4. (7-8 hét) Leíró statisztikák, statisztikai alapfogalmak: becslések (maximum likelihood, momentum)

Elmélet

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

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

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

Tapasztalati szórás: $S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ *(az átlagtól való átlagos abszolút eltérés)*

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

Szórási együttható (vagy relatív szórás): $V = \frac{S_n}{\overline{X}} = \frac{S_n}{\overline{X}}100\%$ *(az átlagtól való átlagos eltérés százalékban)*
/megjegyzés: lehet a korrigált tapasztalati szórással 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: $X_1^* \leq ... \leq X_n^*$ *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*

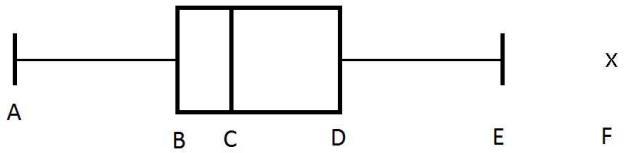
Terjedelem: $R = X_n^* - X_1^*$ *(legnagyobb − legkisebb mintaelem)*
z-kvantilis: $q_z = \inf\{x : F(x) \geq z\}$*. Ha* F *invertálható, akkor* $q_z = F^{-1}(z)$*.*
Tapasztalati z-kvantilis: q_z *értelmezése: a mintaelemek* z*-ed része legfeljebb a* q_z*, (*$1-z$*)-ed része pedig legalább a* q_z *értéket 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 kvantilisek, alsó (vagy első) kvartilis: $Q_1 = q_{\frac{1}{4}}$*,*
medián: $Q_2 = q_{\frac{1}{2}}$*,*
felső (vagy harmadik) kvartilis: $Q_3 = q_{\frac{3}{4}}$
Interkvartilis 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 valószínűséggelen 0-hoz konvergál, ami azt jelenti, hogy elég nagy minta esetén* $F_n(x)$ *éréke minden* x*-re tetszőleges 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 tüntetjük fel pontokként, amik A-n vagy E-n kívülre esnek

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

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

$L(\vartheta; \mathbf{x}) = P_\vartheta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$, 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 ϑ szerint, akkor a maximumot kereshetjük a szokásos módon, a deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan n-szeres szorzatot kellene deriválni, amelyiknek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.

Ha ϑ 1 dimenziós, akkor $\partial_\vartheta l(\vartheta, \mathbf{x}) = 0$, míg ha $\vartheta = (\vartheta_1, ..., \vartheta_p)$ 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 ϑ 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 ϑ-ra: $M_1(\vartheta) = m_1, \ldots M_p(\vartheta) = m_p$ (megjegyzés: $m_1 = \overline{x}$)

Feladatok

4.1. Feladat. Legyen $X_1, \ldots, X_n$ 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?

$T_1(\mathbf{X}) = X_8, \qquad T_2(\mathbf{X}) = \dfrac{X_9 + X_{19}}{9}, \qquad T_3(\mathbf{X}) = \overline{X}$

Megoldás

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

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

$E(T_3(\mathbf{X})) = E(\overline{X}) = E\left(\dfrac{1}{n}\sum_{1=1}^{n} E(X_i)\right) = m$, *így* T_3 *torzítatlan*

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

Megoldás

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

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

érték	-1	1	2
gyakoriság	4	10	6

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

Megoldás

c ML-becslése:

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

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

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

Átrendezve a $(lnL(c, \mathbf{x}))'_c = 0$ *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* $(lnL(c, \mathbf{x}))''_c$*-t kiértékelve a* $\hat{c}$ *helyen* $(lnL(c; \mathbf{x}))''_c = -\frac{14}{c^2} - \frac{96}{(1 - 4c)^2} < 0$.

c becslése momentum-módszerrel:

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

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

4.4. 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 μ a paraméter

Megoldás

(Továbbá lehet, hogy érdemes megjegyezni, hogy az $\overline{x} = m$ *eset külön meggondolást igényel.)*

a)

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

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

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

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

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

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

b)

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

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

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

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

c)

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

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

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

Átrendezve a $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)'_\mu = 0$ *egyenletet, kapjuk, hogy* $\hat{\mu} = \dfrac{\sum\limits_{k=1}^{n} X_i}{n} = \overline{X}$. *Ez valóban maximum, mivel* $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)''_\mu = -\dfrac{n}{\sigma^2} < 0.$

4.5. 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):

$$L(a;\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{1-a} I(a \le x_i \le 1) = \frac{1}{(1-a)^n} I(a \le x_1, x_2, ..., x_n \le 1) =$$

$$= \frac{1}{(1-a)^n} I(a \le x_1^* \le ... \le x_n^* \le 1) = \frac{1}{(1-a)^n} I(a \le x_1^*) I(x_n^* \le 1)$$

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

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

Megoldás

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

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

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:*

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