Leíró és matematikai statisztika

Matematika alapszak, matematikai elemző szakirány

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: `zempleni.elte.hu`

E-mail: `andras.zempleni@ttk.elte.hu`

Szoba: D 3-310

4. előadás

Fontos becslések tulajdonságai

Tétel. Legyen $X_1, \ldots, X_n$ i.i.d. minta egy ϑ paraméterű eloszláscsaládból, $h : \mathbb{R} \to \mathbb{R}$ (mérhető) függvény. Tegyük fel, hogy a táblázatban szereplő összes várható érték/szórás létezik minden ϑ esetén.

Mit becsülünk? $g(\vartheta)$	Ha mivel becsüljük? $T_n(\mathbf{X})$	Torzítatlan?	Aszimptotikusan torzítatlan?	Gyengén/ erősen konzisztens?
$E_\vartheta X_1$	$\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$	igen	igen	igen
$D^2_\vartheta X_1$	$S_n^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$	nem	igen	igen
$D^2_\vartheta X_1$	$(S_n^*)^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$	igen	igen	igen
$F_\vartheta(x)$	$F_n(x) = \frac{\sum_{i=1}^n I(X_i < x)}{n}$	igen	igen	igen
$E_\vartheta h(X_1)$	$\frac{\sum_{i=1}^n h(X_i)}{n}$	igen	igen	igen

Becslési módszerek I

Maximum likelihood módszer

(ML-módszer): Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel: $\max\limits_{\vartheta} L(\vartheta; \mathbf{x})$.

- Ez nyilván megegyezik azzal a paraméterértékkel, ahol a log-likelihood függvény veszi fel a legnagyobb értéket, azaz: $\max\limits_{\vartheta} \ell(\vartheta; \mathbf{x})$.
- Amennyiben a függvény deriválható ϑ szerint, akkor a maximumot kereshetjük a deriváltak segítségével
- Célszerűbb a likelihood függvény helyett a log-likelihood függvény maximumhelyét keresni.
- Ha ϑ 1 dimenziós, akkor $\partial_\vartheta \ell(\vartheta, \mathbf{x}) = 0$, míg ha $\vartheta = (\vartheta_1, \ldots, \vartheta_p)$ p dimenziós, akkor $\partial_{\vartheta_i} \ell(\vartheta, \mathbf{x}) = 0$ megoldásából kapjuk a becslést.
- Példa: a Poisson, exponenciális, normális eloszlás paramétereire adjunk ML becslést!

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

Feladatok

E15.) El szeretnénk dönteni egy érméről, hogy az szabályos-e, avagy cinkelt. Írjuk fel a problémát leíró statisztikai mezőt!

E16.) Legyen $X_1, \ldots, X_n$ i.i.d. Bin$(4; p)$ eloszlású valószínűségi változó, ahol $p \in (0; 1)$ ismeretlen valós paraméter.

a.) Adjuk meg a mintateret és a paraméterteret!

b.) Határozzuk meg az ismeretlen paraméter ML-becslését!

c.) Torzítatlan, illetve konzisztens az ML-becslés? Amennyiben nem torzítatlan, tegyük azzá!

d.) Adjunk torzítatlan becslést $g(p) = p^2(1-p)^2$-re!

E17.) Legyen X_1 Bin$(2; p)$ eloszlású (egyelemű) minta, ahol $p \in (0; 1)$ ismeretlen valós paraméter. Adjunk X_1 segítségével torzítatlan becslést $g(p) = \frac{1}{p}$-re!

ML becslés a normális eloszlás μ paraméterére

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

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

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

Átrendezve a $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)'_{\mu} = 0$ egyenletet, kapjuk, hogy

$$\hat{\mu} = \frac{\sum\limits_{i=1}^{n} X_i}{n} = \overline{X}.$$

Ez valóban maximum, mivel $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)''_{\mu} = -\frac{n}{\sigma^2} < 0.$

ML becslés a normális eloszlás σ paraméterére

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

$$\left(lnL(\mu,\sigma^2;\mathbf{x})\right)'_\sigma = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Átrendezve a $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)'_\sigma = 0$ egyenletet, kapjuk, hogy

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}.$$

Ez valóban maximum. Ugyanakkor μ tipikusan nem ismert, így a becslést úgy kapjuk, hogy behelyettesítjük az előzőekben μ-re kapott ML becslést:

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n}; \text{ és ebből } \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n}}$$

Becslési módszerek II

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 p ismeretlenes egyenletrendszert oldunk meg ϑ-ra: $M_1(\vartheta) = m_1, \ldots, M_p(\vartheta) = m_p$ (megjegyzés: $m_1 = \overline{x}$)
- Ha valamelyik egyenlet nem ad információt a keresett paraméterre, akkor magasabb hatványokat nézünk (míg megoldható nem lesz az egyenletrendszer)

Feladatok

E**18.)** Adjuk meg a momentum módszeres becslést a normális eloszlás paramétereire!

E**19.)** Legyen $X_1, \ldots, X_n$ i.i.d. Bin$(4; p)$ eloszlású valószínűségi változó, ahol $p \in (0; 1)$ ismeretlen valós paraméter. Határozzuk meg az ismeretlen paraméter momentum becslését!

E**20.)** Tegyük fel, hogy egy hét munkanapjain az alábbi várakozási időket mértük a 4-es 6-os villamosra (percben):

1,2 2 1,5 3 2,1

A várakozási időről tegyük fel, hogy exponenciális eloszlású.

- a.) Adjuk meg a mintateret és a paraméterteret!
- b.) Határozzuk meg az ismeretlen paraméter ML-becslését!
- c.) Határozzuk meg az ismeretlen paraméter momentum-becslését!
- d.) Torzítatlan, illetve konzisztens az ML-becslés? Amennyiben nem torzítatlan, tegyük azzá!
- e.) Mutassuk meg, hogy az $S(\mathbf{X}) = n \cdot X_1^*$ statisztika torzítatlan, de nem konzisztens becslése $g(\lambda) = \frac{1}{\lambda}$-nak!

Elégséges statisztika

- Legyen az $X_1, ..., X_n$ minta diszkrét eloszlású. $T(\mathbf{X})$ minden információt tartalmaz a ϑ paraméterre (azaz elégséges ϑ-ra), ha a
$$P_\vartheta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t)$$
feltételes valószínűség nem függ ϑ-tól.
- Faktorizáció: T pontosan akkor elégséges, ha
$$f_\vartheta(\mathbf{x}) = h(\mathbf{x}) g_\vartheta(T(\mathbf{x}))$$
(azaz a paramétertől a T-n keresztül függ a eloszlás)

E**21.)** Határozzuk meg az elégséges statisztikát a Poisson eloszlás paraméterére!

- Az abszolút folytonos esetre a faktorizáció a definíció.

E**22.)** Határozzuk meg az elégséges statisztikát az exponenciális és az egyenletes eloszlás paramétere(i)re!

Tulajdonságok

- Az elégséges statisztika nem egyértelmű: szigorúan monoton függvénybe behelyettesíthetjük
- Sőt: ha hozzáveszünk újabb statisztikákat, akkor is elégséges marad.
- Ezért a célunk minél egyszerűbb elégséges statisztika keresése.
- A maximum likelihood becslés az elégséges statisztika függvénye.

Definíció. **[** hatásosság] Egy T torzítatlan becslés hatásos, ha minden más T^* torzítatlan becslésre $D^2_\vartheta T(\mathbf{X}) \leq D^2_\vartheta T^*(\mathbf{X})$ minden $\vartheta \in \Theta$ paraméterre

- Megjegyzés: a hatásos becslés nem mindig létezik, de ha van ilyen, akkor egyértelmű.
- Az általános (nem feltétlenül torzítatlan) esetre $E_\vartheta(T(\mathbf{X}) - \vartheta)^2)$ minimalitása definiálja a hatásosságot

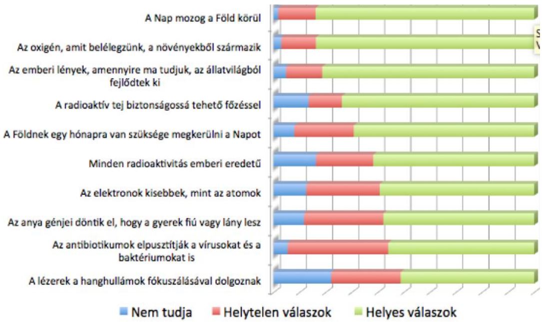
Rao-Blackwell tétel

- Rao-Blackwell tétel: Ha T torzítatlan becslés a ϑ paraméterre és S elégséges statisztika ϑ-ra, akkor megadható olyan becslés, mely
  - S függvénye
  - torzítatlan ϑ-ra
  - kisebb szórású, mint T
- $E(T|S)$ lesz a megoldás.

E**23.)** Adjunk meg "jó" becslést az $\{1, 2, ..., a\}$ számokon egyenletes eloszlás paraméterére a Rao-Blackwell tétel segítségével!

Öt perc szünet

Egy érdekes felmérés:



És az értelmezése:

- "nálunk 800 ezer ember gondolja úgy, hogy a Nap kering a Föld körül" (minden portál)
- "A megkérdezettek mintegy 40 százaléka sem azt nem tudta, hogy az atomok vagy az elektronok kisebbek-e, sem azt, hogy az antibiotikumokat milyen kórokozók ellen érdemes bevetni" (qubit)
- "...ezzel az eredménnyel Európában a 2. legjobb helyen állunk" (index)

Feladatok

E24.) Legyen $X_1, \ldots, X_n$ i.i.d. abszolút folytonos valószínűségi változók sorozata.

a.) Adjuk meg az i-edik rendezett mintaelem, azaz X_i^* eloszlás- és sűrűségfüggvényét ($1 \leq i \leq n$)!

b.) Milyen eloszlású X_i^*, amennyiben a mintaelemek $(0; 1)$-en egyenletesek? Határozzuk meg az EX_i^* mennyiséget!

E25.) Egy véletlen szám generátorral 20 véletlen számot állítunk elő egy ismeretlen (a, b) intervallumból. A kapott véletlen számok sorrendbe téve és (egyszerűség kedvéért) egészre kerekítve:

5	11	12	13	13	14	17	19	21	22
23	24	25	27	31	31	32	35	36	38

a.) Adjuk meg a mintateret és a paraméterteret!

b.) Határozzuk meg az ismeretlen paraméterek maximum likelihood becslését!

c.) Határozzuk meg a paraméterek momentum-becslését! Értékeljük a kapottakat!