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

Az elégséges statisztika tulajdonságai

- 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 $D_\vartheta^2 T(\mathbf{X}) \leq D_\vartheta^2 T^*(\mathbf{X})$ 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.

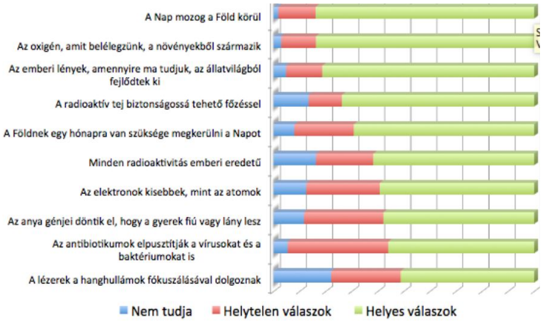
E**23.)** Adjunk meg "jó" becslést az $\{1, 2, ..., a\}$ számokon egyenletes eloszlás paraméterére a Rao-Blackwell tétel segítségével! E**24.)** Legyen $X_1, \ldots, X_n$ i.i.d. abszolút folytonos valószínűségi változók sorozata.

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!

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)

Egy feladat

E25.) Egy véletlenszám-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!
- d.) Torzítatlanok az ML-becslések? Ha nem, akkor tegyük őket torzítatlanná!
- e.) Konzisztensek az ML-becslések?

Sűrűségfüggvény becslése – magfüggvényes módszer (Parzen-Rosenblatt becslés)

$$f_n(x) = \frac{1}{n \cdot h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \text{ ahol}$$

- $K : \mathbb{R} \to \mathbb{R}$ páros sűrűségfüggvény, neve: magfüggvény
- h_n sávszélesség, pl. $h_n = n^c$, ahol $-1 < c < 0$ valós szám
- Ekkor korlátos K-ra f_n konzisztens becslés f folyt.-i pontjaiban

A leggyakoribb magfüggvények:

Magfüggvény neve	$K(x)$
Gauss	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
téglalap (rectangular)	$\frac{1}{2} I(\lvert x \rvert \le 1)$
háromszög (triangular)	$(1 - \lvert x \rvert) \cdot I(\lvert x \rvert \le 1)$
Bartlett–Epanechnikov	$\frac{3}{4}(1 - x^2) \cdot I(\lvert x \rvert \le 1)$
cosinus	$\frac{\pi}{4} \cos\left(\frac{\pi}{2} x\right) \cdot I(\lvert x \rvert \le 1)$

Mi az "optimális" h_n? $\rightsquigarrow$ amivel $f_n(x)$ "legjobban" közelíti a valódi sfv-t.

Eloszlásfüggvény/sűrűségfüggvény becslése: példák

Eloszlásfüggvény (F) becslése tapasztalati eloszlásfüggvénnyel (F_n): nemcsak torzítatlan és konzisztens, hanem egyenletesen is konvergál:

$$\sup_{x\in\mathbb{R}} |F(x) - F_n(x)| \to 0 \ \ n \to \infty$$

(Glivenko tétele)

E**26.)**

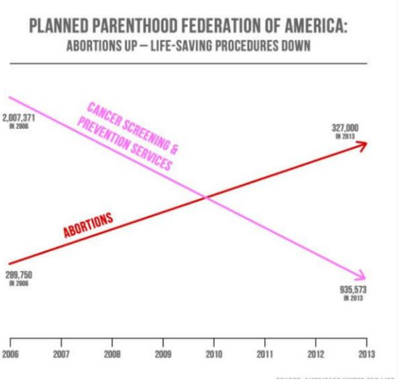
a.) Legyen a rendezett mintánk 1,2,5,6,12. Számoljuk ki a sűrűségfüggvény becslését, ha $h = 0,5$ és a téglalap-magfüggvényt alkalmazzuk!

b.) Mennyiben változik a becslés, ha a háromszög-magfüggvényt választjuk?

c.) Mennyi az R által alkalmazott default sávszélesség?

Öt perc szünet

Egy "érdekes" ábra



Furcsa COVID adatok

Horvátország	544	443	31%	▲
Németország	11294	8114	8%	▲
Dánia	720	524	9%	▲
Spanyolország	36826	-4464	-148%	▼

A Fisher-információ, becslések hatásossága

- A minta (elméleti) információja:

$$I_n(\vartheta) = E_\vartheta \left(\partial_\vartheta \ell(\vartheta, \mathbf{X})\right)^2$$

- Tulajdonságok:
  - A reguláris esetben $I_n(\vartheta) = nI_1(\vartheta)$.
  - Ha másodrendben is reguláris a függvény, akkor

$$I_n(\vartheta) = -E_\vartheta \left(\partial_\vartheta^2 \ell(\vartheta, \mathbf{X})\right)$$

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

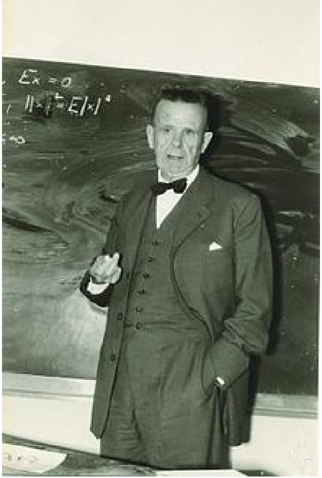
- Megjegyzés: a hatásos becslés nem mindig létezik, de ha van ilyen, akkor egyértelmű.
- Az általános (nem feltétlenül torzítatlan) esetre $E(T(\mathbf{X}) - \vartheta)^2)$ minimalitása definiálja a hatásosságot

A Fisher-információ alkalmazásai

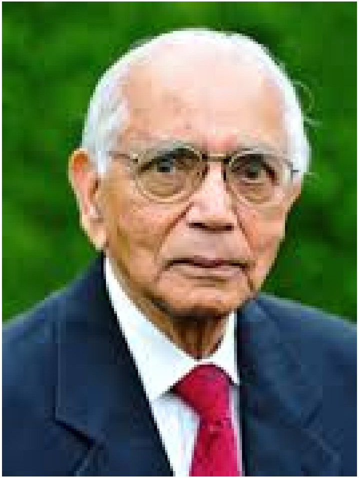
- hatásos becslés keresése az információs határ segítségével
  Cramér-Rao egyenlőtlenség: A reguláris esetben
  $$D^2_\vartheta(T(\mathbf{X})) \geq \underbrace{\frac{(g'(\vartheta))^2}{I_n(\vartheta)}}_{\text{információs határ}}$$
  Ha egy $g(\vartheta)$-ra nézve torzítatlan T statisztika esetén egyenlőség teljesül, akkor az a statisztika hatásos becslése $g(\vartheta)$-nak.
- az ML-becslés eloszlásban egy olyan normális eloszláshoz tart, aminek a szórásnégyzete a Fisher-információ inverze
- intervallumbecslés az ML-becslésre
- kísérlettervezés
- Bayes-i statisztika – Jeffrey-féle apriori eloszlás számításához
- neurális hálók, machine learning
- számítógépes agykutatás

Az információs határ felfedezői/névadói

Harald Cramér (1893 – 1985)



C. R. Rao (1920 –)



Hatásos becslés keresése az információs határral

E**27.)** Legyen $X_1, \ldots, X_n$ i.i.d. Ind(p) eloszlású.

a.) Határozzuk meg a mintában lévő Fisher-információ értékét!

b.) Mutassuk meg, hogy a relatív gyakoriság hatásos becslése a valószínűségnek!

E**28.)** Legyen $X_1, \ldots, X_n$ i.i.d. Exp(λ) eloszlású.

a.) Határozzuk meg a mintában lévő Fisher-információ értékét!

b.) Mutassuk meg, hogy a mintaátlag hatásos becslése $\frac{1}{\lambda}$-nak!

Az ML becslés aszimptotikus tulajdonságai

- Ha a likelihood függvény teljesít bizonyos regularitási feltételeket, akkor a maximum likelihood becslés $\hat{\vartheta}_n$
  - létezik
  - aszimptotikusan torzítatlan
  - aszimptotikusan hatásos:

$$\sqrt{nI_1(\vartheta)}D(\hat{\vartheta}_n) \to 1 \ (n \to \infty)$$

  - Az alábbi függvénye aszimptotikusan standard normális eloszlású:

$$\sqrt{nI_1(\vartheta)}(\hat{\vartheta}_n - \vartheta) \ (n \to \infty)$$