Valószínűségszámítás és Statisztika

9. előadás

2017. november 19.

Statisztikai mező (ismétlés)

$$\left(\Omega, A, P_\vartheta\right),\ \vartheta \in \Theta$$

statisztikai mező, ha Θ paraméterhalmaz és $\left(\Omega, A, P_\vartheta\right)$ minden paraméter esetén valószínűségi mező.

Minta (ismétlés)

Def.: A $\boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} : \Omega \to X \subseteq R^n$ valószínűségi vektorváltozót

mintának nevezzük.

n : mintanagyság

ξ_i : i. mintaelem

Def.: minta realizációja: $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ a konkrét megfigyelt számsorozat.

Statisztikák (ismétlés)

Def.: Statisztika: a minta függvénye.

$$T : X \rightarrow R^k$$

Def'.: Statisztika:

$$T(\xi), \text{ ha } T : X \rightarrow R^k \text{ függvény.}$$

Becsléselmélet (ismétlés)

- A minta eloszlásának ismeretlen paraméterét közelítjük a minta függvényével
- Def.: becslőfüggvény: $\hat{\vartheta} : X \rightarrow \Theta$
- Def.: becslés: $\hat{\vartheta}(\xi)$

- A becslések maguk is statisztikák. Szubjektíven: olyan statisztikák, amik jól közelítik az ismeretlen paramétert.

Becslések tulajdonságai (ismétlés)

- Def.: *Torzítatlanság:* A paraméter $\hat{\vartheta}(\xi)$ becslése torzítatlan, ha

$$E_{\vartheta}\left(\hat{\vartheta}(\xi)\right) = \vartheta, \ \forall \vartheta \in \Theta.$$

- *Konzisztencia*: $\hat{\vartheta}(\xi) \rightarrow \vartheta$ sztochasztikusan ($n \rightarrow \infty$) minden paraméterértékre.
- Példák:
  - Valószínűség becslése relatív gyakorisággal.
  - Glivenko tétele: a tapasztalati eloszlásfüggvény egyenletesen is konvergál az elméleti eloszlásfüggvényhez.
  - Várható érték becslése mintaátlaggal

Konzisztencia

- Elégséges feltétel $E_{\vartheta}\left(\hat{\vartheta}_n(\xi)\right) \to \vartheta$ (aszimptotikus torzítatlanság) és

$$D_{\vartheta}^{2}\left(\hat{\vartheta}_n(\xi)\right) \to 0$$

Mit kell tudni a mintáról?

- Benzinkutas példa. Megfigyelések: 78, 89, 167, 90, 85.
- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.

Kár-szám	0	1	2	3	4	5	6	7	> 7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006

Mennyi információt hordoz a statisztika?

Példa: $\xi_1, ..., \xi_n$ független $N(m,1)$ minta. Ekkor

$$\overline{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n} \sim N(m, \frac{1}{n})$$ eloszlású (függ m-től!),

miközben

$$s^2 = \frac{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}{n}$$ eloszlása nem függ m-től!

Elégséges statisztika

- Minden információt (ugyanannyit mint az eredeti minta) tartalmaz az ismeretlen paraméterre vonatkozóan.
- ”Elég” az ő értékét ismerni.
- Ismeretében már ”nincs bizonytalanság” a mintában (úgy értve, hogy egyértelmű a minta eloszlása, már nem függ az ismeretlen paramétertől).

Elégséges statisztika diszkrét minta esetén

Def.: A diszkrét $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges Θ-ra, ha a $P_\theta(\boldsymbol{\xi} = \mathbf{x} \,|\, T(\boldsymbol{\xi}) = t)$ feltételes valószínűség nem függ θ-tól

Feltételes várható érték

Legyenek X és Y diszkrét val. változók. $E(X|Y)$ az a val. változó, ami az $Y=y_k$ eseményen az $E(X|Y=y_k)$ értéket veszi fel.

Tulajdonságok:

- Ha $X \geq 0$, akkor $E(X|Y) \geq 0$
- $E(E(X|Y))=EX$ (a teljes várható érték tételének általánosítasa)
- Ha X_1, X_2 várható értéke véges, akkor $E(c_1X_1+c_2X_2|Y)=c_1E(X_1|Y)+c_2E(X_2|Y)$
- Ha X független Y-tól, akkor $E(X|Y)=E(X)$
- Ha X és $h(Y)$ várható értéke véges, akkor $E(h(Y)X|Y)=h(Y)E(X|Y)$
- Teljes szórásnégyzet tétele:

$$D^2(X) = D^2(E(X|Y)) + E(D^2(X|Y))$$

Tétel (Neyman-féle faktorizációs):

A diszkrét $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika pontosan akkor elégséges Θ-ra, ha $\exists\, g_\theta(t)$ és $h(\mathbf{x})$ úgy, hogy $\forall \theta \in \Theta$ és $\mathbf{x} \in \mathcal{X}$-ra

$$P_\theta(\boldsymbol{\xi} = \mathbf{x}) = h(\mathbf{x}) g_\theta(T(\mathbf{x})).$$

Biz.:

$$\Rightarrow T(\boldsymbol{\xi}) \text{ elégséges, ekkor } P_\theta(\boldsymbol{\xi} = \mathbf{x}) = P_\theta(T(\boldsymbol{\xi}) = T(\mathbf{x})) \frac{P_\theta(\boldsymbol{\xi} = \mathbf{x}, T(\boldsymbol{\xi}) = T(\mathbf{x}))}{P_\theta(T(\boldsymbol{\xi}) = T(\mathbf{x}))}$$

$$= P_\theta(T(\boldsymbol{\xi}) = T(\mathbf{x})) P_\theta(\boldsymbol{\xi} = \mathbf{x} \,|\, T(\boldsymbol{\xi}) = T(\mathbf{x})) = g_\theta(T(\mathbf{x})) h(\mathbf{x}).$$

$\Leftarrow P_\theta(\boldsymbol{\xi} = \mathbf{x} \,|\, T(\boldsymbol{\xi}) = t) = 0$, ha $t \neq T(\mathbf{x})$. Amennyiben ez teljesül:

$$P_\theta(\boldsymbol{\xi} = \mathbf{x} \,|\, T(\boldsymbol{\xi}) = t) = \frac{P_\theta(\boldsymbol{\xi} = \mathbf{x}, T(\boldsymbol{\xi}) = t)}{P_\theta(T(\boldsymbol{\xi}) = t)} = \frac{P_\theta(\boldsymbol{\xi} = \mathbf{x}, T(\boldsymbol{\xi}) = t)}{P_\theta(T(\boldsymbol{\xi}) = t)} = \frac{P_\theta(\boldsymbol{\xi} = \mathbf{x})}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} P_\theta(\boldsymbol{\xi} = \mathbf{y})}$$

$$= \frac{h(\mathbf{x}) g_\theta(T(\mathbf{x}))}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y}) g_\theta(T(\mathbf{y}))} = \frac{h(\mathbf{x}) g_\theta(t)}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y}) g_\theta(t)} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y})}.$$

Ez nem függ θ-tól!

Példa (Poisson minta)

η_i – k független λ Poissonok. Ekkor

$$P_\lambda\left(\eta_1 = k_1, \ldots, \eta_n = k_n\right) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda} =$$

$$= h(\mathbf{k}) g_\lambda\left(\sum_{i=1}^{n} k_i\right),$$

ahol

$$h(\mathbf{k}) = \prod_{i=1}^{n} \frac{1}{k_i!}, \quad g_\lambda(t) = \lambda^t e^{-n\lambda}.$$

Elégséges statisztika általában

Def.: A $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges Θ-ra, ha minden $\mathbf{x} \in \mathbf{R}^n$-re a $P_\theta(\boldsymbol{\xi} < \mathbf{x} \,|\, T(\boldsymbol{\xi}) = t) = P_\theta(\xi_1 < x_1, ..., \xi_n < x_n \,|\, T(\boldsymbol{\xi}) = t)$ feltételes eloszlásfüggvény nem függ θ-tól.

Probléma: A feltételes valószínűség és várható érték fogalmát nem tanultuk általánosan!

Likelihood függvény

Def.: A $\xi_1, ..., \xi_n$ független, azonos eloszlású minta likelihood függvénye

$$L(\mathbf{x}, \theta) = \begin{cases} P_\theta(\boldsymbol{\xi} = \mathbf{x}) = \prod_{i=1}^{n} P_\theta(\xi_i = x_i) & \text{diszkrét minta esetén} \\ f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) & \text{abszolút folytonos minta esetén} \end{cases}$$

ahol f_θ ξ_i sűrűségfüggvénye.

$l(\mathbf{x}, \theta) = \ln L(\mathbf{x}, \theta)$ a loglikelihood függvény.

Abszolút folytonos eset

- Definíció a faktorizációval

Def.:

Az abszolút folytonos $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges Θ-ra, ha $\exists\, g_\theta(t)$ és $h(\mathbf{x})$ úgy, hogy $\forall\, \theta \in \Theta$ és $\mathbf{x} \in X$ -ra a likelihood függvény felírható a következő alakban:

$$L(\mathbf{x}, \theta) = h(\mathbf{x}) g_\theta(T(\mathbf{x})).$$

Becslési módszerek

- Példa: Egy tóban *N* hal van, számukat nem ismerjük. Első héten kihalásznak 1000 halat és megjelölik őket. A következő héten kihalásznak 5000-et és megszámolják a megjelölteket. 50-et találnak. Becsüljük meg *N*-et!

Természetes eljárás

Jelölje ξ a másodjára kihúzott halak számát.

Tudjuk, hogy ez hipergeometrikus eloszlású, így

$$L(50, N) = P_N(\xi = 50) = \frac{\binom{1000}{50}\binom{N-1000}{4950}}{\binom{N}{5000}}.$$

Becslés

$$\hat{N} : L(50, \hat{N}) = \max_N L(50, N) \Rightarrow \hat{N} = 100000$$

Maximum likelihood becslés

- Definíció heurisztikusan: azt a paraméterértéket keressük, amelyre az adott minta bekövetkezési valószínűsége maximális.

Def.: θ maximum likelihood becslése $\hat{\theta} = T(\boldsymbol{\xi}) \in \Theta$, ha

$$L(\boldsymbol{\xi}, \hat{\theta}) = \max_{\theta \in \Theta} L(\boldsymbol{\xi}, \theta)$$

Likelihood egyenlet

Gyakran a loglikelihood függvény maximumhelyét keresik a

$\frac{\partial l(\mathbf{x},\theta)}{\partial \theta} = 0$ egyenletet (vagy egyenletrendszert) megoldva.

Ez diszkrét minta esetén a

$$\sum_{i=1}^{n} \frac{\partial \ln P_\theta(\xi_i = x_i)}{\partial \theta} = 0$$

egyenletet (vagy egyenletrendszert) jelenti.

Abszolút folytonos minta esetén

$$\sum_{i=1}^{n} \frac{\partial \ln f_\theta(x_i)}{\partial \theta} = 0$$

egyenletet (vagy egyenletrendszert) oldjuk meg.

Adottak sorszámozott gömbök (lottóhúzás) 1-től *N*-ig.
Visszatevéses húzás esetén becsüljük meg *N*-t!

$$\mathrm{P}_N(\xi_i = k) = \frac{1}{N}\chi\{k \le N\}$$

$$\mathrm{L}(\underline{k}, \lambda) = \mathrm{P}_N(\xi_1 = k_1, \ldots, \xi_n = k_n) = \frac{1}{N^n}\chi\{\max_i k_i \le N\}$$

$$\widehat{N} = \max_i \xi_i$$

Momentum módszer

- Ha az eloszlást k db paraméter határozza meg, akkor k db egyenletből kaphatunk rájuk becslést. Az egyenletek a tapasztalati és az elméleti momentumok egybevetéséből adódnak:

$$m_i(\underline{\theta}) = E_{\underline{\theta}}(X^i)$$

$$m_i(\underline{\theta}) = \frac{\sum_{j=1}^{n} \left(\xi_j\right)^i}{n}$$

- Példa: egyenletes eloszlás az [a,b] intervallumon.