Valószínűségszámítás és Statisztika

8. előadás

2017. november 12.

Példa

- Milyen valószínűséggel születik fiúgyermek?
- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Fiúk relatív gyakorisága így 0,5141.
- Igaz-e, hogy a valószínűség 0,5? És 0,1?

$$X_i = \begin{cases} 1, & i.\text{fiú} \\ 0, & i.\text{lány} \end{cases} \Rightarrow$$

$$P(X_i = 1) = p,\ n = 2.644.757,\ \xi = \sum_{i=1}^{n} X_i \Big/ n \quad \Rightarrow$$

$$EX_i = p, D^2 X_i = p(1-p),\ P\left(\frac{\sum_{i=1}^{n} X_i - nEX_1}{DX_1\sqrt{n}} < x \right) \sim \Phi(x) \Rightarrow$$

$$P\left(-u < \sqrt{\frac{n}{p(1-p)}}(\xi - p) < u \right) \sim 2\Phi(u) - 1$$

$$p = 0.5 \Rightarrow \sqrt{\frac{n}{p(1-p)}}(\xi - p) = 37$$

$$u = 4 \Rightarrow 2\Phi(u) - 1 = 0,999936$$

$$p(1-p) \le \frac{1}{4} \Rightarrow$$

$$2\Phi(u) - 1 \sim P\left(-u < \sqrt{\frac{n}{p(1-p)}} (\xi - p) < u \right) \le$$

$$\le P\left(-u < 2\sqrt{n} (\xi - p) < u \right) =$$

$$= P\left(\frac{-u}{2\sqrt{n}} < (\xi - p) < \frac{u}{2\sqrt{n}} \right) = P\left(\xi - \frac{u}{2\sqrt{n}} < p < \xi + \frac{u}{2\sqrt{n}} \right)$$

Esetünkben 0,9973 valószínűséggel $0{,}5132 \le p \le 0{,}5150$

Statisztikai mező

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

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

Egy érmedobás modellje

- Nem ismerjük a fejdobás valószínűségét:

$$\Omega = \{\mathrm{F}, \mathrm{I}\}, A = \{\varnothing; \{\mathrm{F}\}; \{\mathrm{I}\}; \{\mathrm{F}, \mathrm{I}\}\},$$

$$P_p(\{\mathrm{F}\}) = p, P_p(\{\mathrm{I}\}) = 1 - p, p \in [0,1].$$

Minta

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.

Mintatér

- Def: ✠ mintatér: a minta lehetséges értékeinek halmaza. Elemei a mintaértékek.
- *n*-elemű valós minta esetén: $✠=☼^n$
- *n*-elemű pozitív egész értékű minta esetén: $✠=☠^n$
- Példa: egy biztosítónál 10 napon keresztül figyelték a bejelentett károk számát, ekkor $✠=☪_0^n$

Egy benzinkútnál tankoló autók száma 5 napon keresztül

- Megfigyelések: 78, 89, 167, 90, 85
- Minta realizációja:

$(78, 89, 167, 90, 85)^T$

- Mintanagyság: 5

A minták típusai

- Független minta: a mintaelemek függetlenek.
- Független azonos eloszlású minta: a mintaelemek függetlenek és azonos eloszlásúak.
- Diszkrét minta: a mintaelemek diszkrétek.
- Abszolút folytonos eloszlású minta: a mintaelemek abszolút folytonosak.

Eloszláscsaládok

$$F_\vartheta(\mathbf{s}) = P_\vartheta\left(\xi_1 < s_1, \ldots, \xi_n < s_n\right)$$

Független minta esetén:

$$F_\vartheta(\mathbf{s}) = \prod_{i=1}^{n} P_\vartheta\left(\xi_i < s_i\right)$$

Független azonos eloszlású minta esetén:

$$F_\vartheta(\mathbf{s}) = \prod_{i=1}^{n} P_\vartheta\left(\xi_i < s_i\right) = \prod_{i=1}^{n} F_\vartheta\left(s_i\right)$$

Jelölések:

E_ϑ: várható érték P_ϑ esetén,

D_ϑ: szórás P_ϑ esetén,

f_ϑ: sűrűségfüggvény P_ϑ esetén (absz. folyt. minta)

$p_\vartheta(s) = P_\vartheta\left(\xi_i = s\right)$ (diszkrét minta)

Példák

- Egy érmedobás. Fej esetén 1-et írunk, írás esetén 0-át.

$$p_p(k) = P_p(\xi_1 = k) = \begin{cases} p & k = 1 \\ 1-p & k = 0 \end{cases} = p^k (1-p)^{1-k}$$

Benzinkutas példa. Azt feltételezzük, hogy megfigyeléseink független, azonos eloszlású Poissonok.

$$p_\lambda(k) = P_\lambda(\xi_i = k) = \lambda^k e^{-\lambda} / k!,\ k = 0,1,2,\ldots$$

Statisztikák

Def.: Statisztika: a minta függvénye.

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

Def'.: Statisztika:

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

Példák

Tapasztalati momentumok:

$X = R^n,$

mintaközép: $T(\mathbf{x}) = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad T(\boldsymbol{\xi}) = \overline{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n},$

tapasztalati k. momentum: $T(\mathbf{x}) = \frac{\sum_{i=1}^{n} x_i^k}{n}, \quad T(\boldsymbol{\xi}) = \frac{\sum_{i=1}^{n} \xi_i^k}{n}.$

Tapasztalati szórásnégyzet

$$X = R^n,$$

$$T(\mathbf{x}) = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n}, \quad T(\boldsymbol{\xi}) = s^2 = \frac{\sum_{i=1}^{n} \left(\xi_i - \bar{\xi} \right)^2}{n}$$

Rendezett minta

- A $\xi_1,\ldots,\xi_n$ minta elemeit nagyság szerint sorbarendezve kapjuk az $\xi_1^{(n)} \le \xi_2^{(n)} \le \ldots \le \xi_n^{(n)}$ rendezett mintát.
- Ez n-dimenziós statisztika
- Mostantól: a $\xi_1,\ldots,\xi_n$ minta elemei független, azonos eloszlásúak.
- Ha feltesszük, hogy a közös eloszlásuk abszolút folytonos, akkor felírható a rendezett minta k-adik elemének, $\xi_k^{(n)}$ -nek a sűrűségfüggvénye. (gyakorlat)
- Spec.: minimum, maximum.
- Def.: minta terjedelme: $\xi_n^{(n)} - \xi_1^{(n)}$

Tapasztalati eloszlásfüggvény

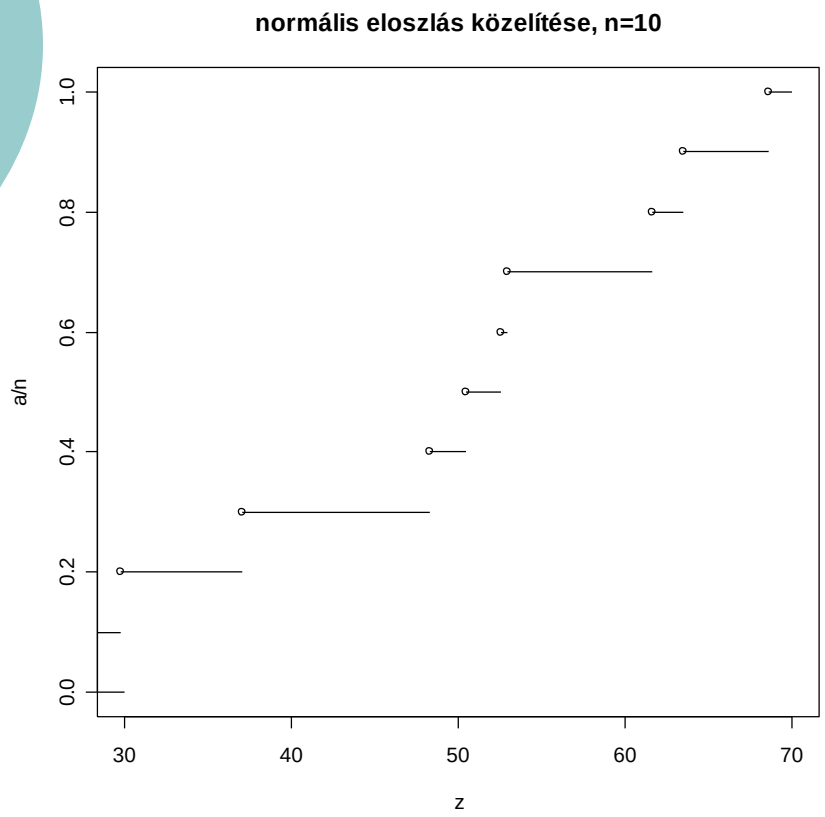
- Előző előadáson már szerepelt.
- Tapasztalati eloszlás eloszlásfüggvénye: tapasztalati eloszlásfüggvény:

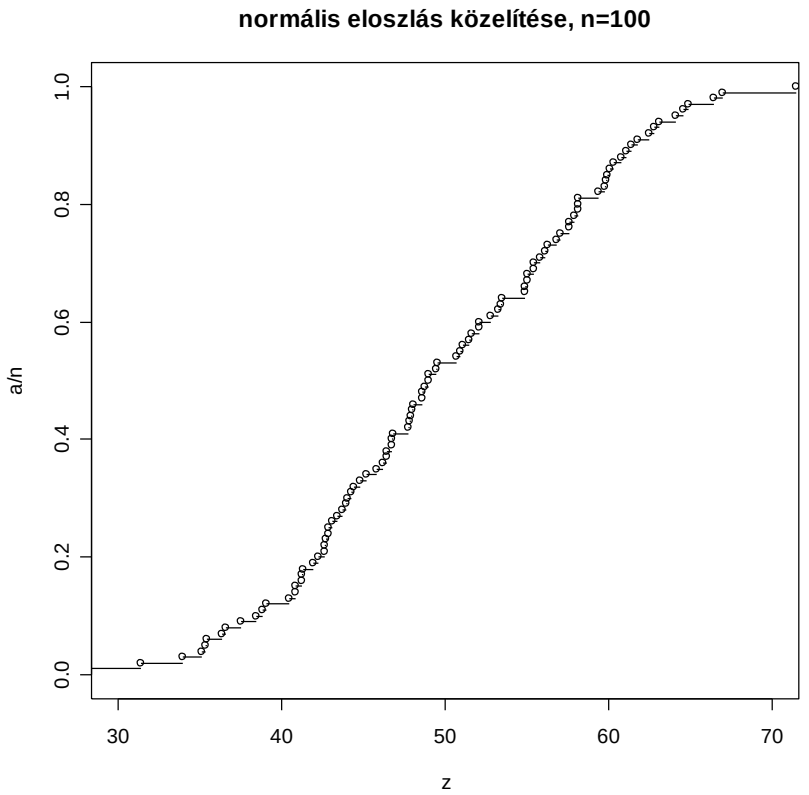
$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i < z\}$$

$$F_n(z) = \frac{k}{n}, \text{ ha } \xi_k^{(n)} < z \le \xi_{k+1}^{(n)},\ \xi_0^{(n)} = -\infty,\ \xi_{n+1}^{(n)} = \infty$$

Mintaátlag éppen ennek az eloszlásnak a várható értéke.

Példa


normális eloszlás közelítése, n=10
a/n
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
z


normális eloszlás közelítése, n=100
a/n
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
z

Glivenko-Cantelli tétel (”statisztika alaptétele”)

Tétel: $\xi_1, ..., \xi_n$ független, azonos F eloszlásfüggvényűek. Ekkor $\sup_z \left|F_n(z) - F(z)\right| \xrightarrow[n\to\infty]{} 0$ majdnem mindenütt (1 vszgel).

Biz.: Csak folytonos F eloszlásfüggvényekre látjuk be. Ebből következik, hogy tetszőleges pozitív egész N-hez léteznek olyan valós $z_1, ..., z_N$ számok, hogy

$$F(z_0) = 0, F(z_1) = \frac{1}{N}, ..., F(z_i) = \frac{i}{N}, ..., F(z_{N-1}) = \frac{N-1}{N}, F(z_N) = 1,$$

$z_0 = -\infty, z_N = \infty.$

Becsléselmélet

- 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.

Példa (Milyen valószínűséggel születik fiúgyermek?)

- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Ekkor n= 2.644.757, ✠={0;1}^{n}.
- Fiúk relatív gyakorisága így 0,5141.
- Mik ennek a becslésnek a tulajdonságai?

$$X_i = \begin{cases} 1, & i.\text{fiú} \\ 0, & i.\text{lány} \end{cases} \Rightarrow$$

$$P_p\left(X_i = 1\right) = p,\ n = 2.644.757,$$

$$\hat{p} = \hat{p}(\mathbf{X}) = \sum_{i=1}^{n} X_i \Big/ n \Rightarrow$$

$$E_p\,\hat{p} = p,\ \hat{p} \xrightarrow[n\to\infty]{} p \ \text{ mm.}$$

Becslések tulajdonságai

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

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

- *Konzisztencia:* $\hat{\vartheta}(\xi) \quad \vartheta \rightarrow$ 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

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

Példák

- Poisson eloszlás paraméterére: mintaátlag
- Exponenciális eloszlás paraméterére:
  - mintaátlag reciproka: aszimptotikusan torzítatlan, konzisztens
  - $n \cdot \min(X_1,\ldots,X_n)$ torzítatlan, de nem konzisztens
- Szórásnégyzetre

Becslések összehasonlítása

- Melyik a jobb becslés?

$$X_i = \begin{cases} 1, & i.\text{fiú} \\ 0, & i.\text{lány} \end{cases}, \; P_p\left(X_i = 1\right) = p,$$

$$\hat{p}_1 = \sum_{i=1}^{n} X_i \Big/ n, \quad \hat{p}_2 = X_1, \text{ vagy}$$

$$\hat{p}_3 = \sum_{i=1}^{[n/2]} X_i \Big/ [n/2] \; ?$$

Becslések összehasonlítása (hatásos becslések)

- Torzítatlan becslésekre: T_1 hatásosabb becslése $h(\theta)$-nak a T_2-nél, ha

$$D_\theta^2\left(T_1(\underline{X})\right) \leq D_\theta^2\left(T_2(\underline{X})\right)$$

teljesül minden θ paraméterértékre.

Példa: a mintaátlag hatásosabb becslés a várható értékre minden

$$\sum_{i=1}^{n} c_i X_i$$

alakú becslésnél.

Hatásos becslés

- Def.: A *T* torzítatlan becslés hatásos, ha minden más torzítatlan becslésnél hatásosabb.
- Miért a torzítatlanokra? Furcsa példa: azonosan 0-val becsüljük az ismeretlen paramétert.
- Ezért érdemes a hatásos becsléseket csak a torzítatlan becslések között keresni.
- Átlagos négyzetes eltérés:

$$E_\theta\left(T(\underline{X})-\theta\right)^2$$

Hatásos becslés egyértelműsége

- **Áll.:** Amennyiben T_1 és T_2 hatásos becslései $h(\theta)$-nak, akkor 1 valószínűséggel megegyeznek minden lehetséges paraméter esetén.

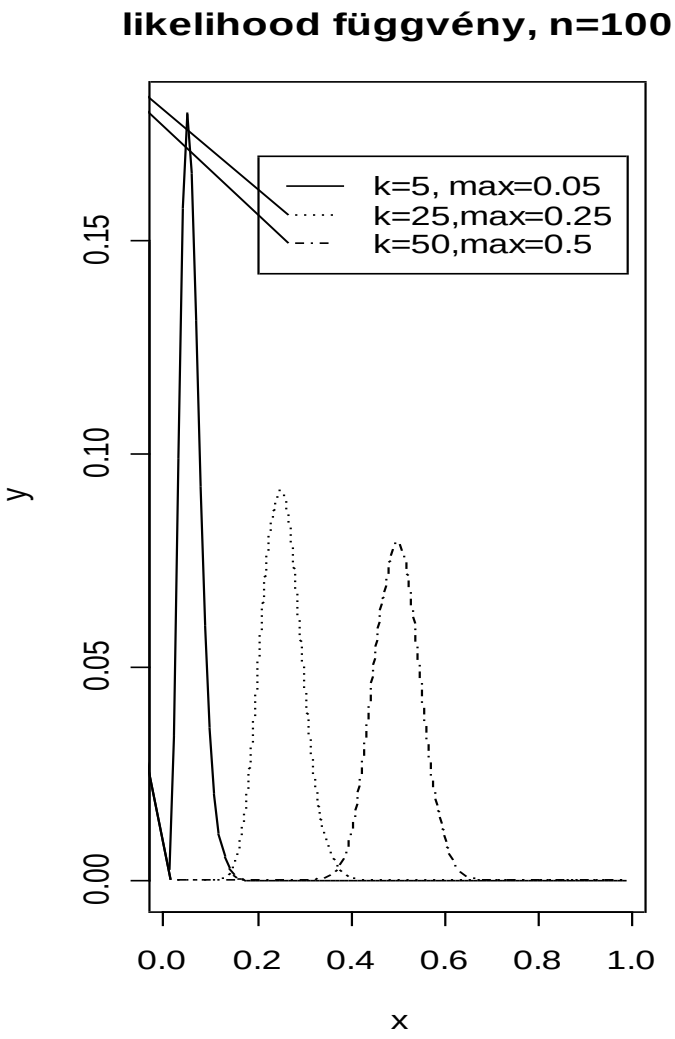
Becslési módszerek

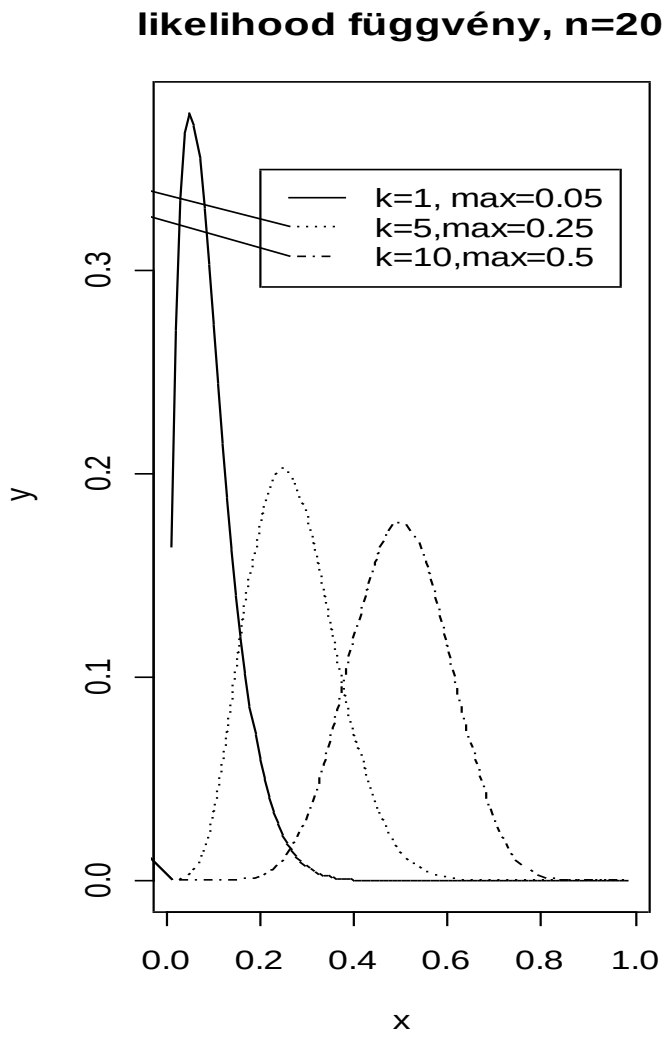
- Eddig: „ad hoc” módszerek
- Általános eljárás kellene
  - Példa: valószínűség becslése, n kísérletből. Jelölje k a sikeresek számát (X_i i=1,...,n indikátorminta)

$$P\left(\sum_{i=1}^{n} X_i = k\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

Most p függvényében nézzük, k rögzített (elnevezés: likelihood függvény).

A likelihood függvény maximumhelye logikus választás a valószínűség becslésének





A módszer általánosan

$$L(\theta;\underline{x}) = f_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i)$$

(a likelihood függvény) maximumhelye lesz a θ paraméter maximum likelihood becslése.

Ha a függvény deriválható, a loglikelihood függvény

$$l(\theta;\underline{x}) = \ln f_\theta(\underline{x}) = \sum_{i=1}^{n} \ln f_\theta(x_i)$$

maximumhelye deriválással

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

megoldásaként megtalálható

Példák

- valószínűségre: relatív gyakoriság
- Poisson eloszlás paraméterére: $\bar{x}$
- Exponenciális eloszlás paraméterére: $1/\bar{x}$
- Normális eloszlás várható értékére: $\bar{x}$
- A módszer többdimenzós paraméter becslésére is használható: N(m,σ) esetén $\left(\bar{x}, \sum \left(x_i - \bar{x}\right)^2 / n\right)$ a maximum likelihood becslés.