Becslések és tulajdonságaik (11. előadás)

- $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta)$ valamely Θ halmazzal (Θ a paramétertér);
- $\psi : \Theta \to \mathbb{R}$ függvény.
- Cél: olyan T statisztika keresése, amire a $T(X)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esnek egymáshoz.

Becslések és tulajdonságaik (11. előadás)

- $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta)$ valamely Θ halmazzal (Θ a paramétertér);
- $\psi : \Theta \to \mathbb{R}$ függvény.
- Cél: olyan T statisztika keresése, amire a $T(X)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esnek egymáshoz.

Definíció (Torzítatlanság)

A T statisztika torzítatlan becslés ψ-re, ha minden $\vartheta \in \Theta$-ra

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta).$$

A T statisztika torzítása a $b_T(\vartheta) = \mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) - \psi(\vartheta)$ függvény.

Példa. $X_1, X_2, \ldots, X_n$ független minta a $[0, \vartheta]$ intervallumon egyenletes eloszlásból. Ekkor $2\overline{X}$ torzítatlan becslés $\psi(\vartheta) = \vartheta$-ra.

Torzítatlan becslések

Állítás (A várható érték torzítatlan becslése)

Legyen $X_1, \ldots, X_n$ *független azonos eloszlású véges várható értékű minta. Ekkor*

$$\mathbb{E}_\vartheta(\overline{X}) = \mathbb{E}_\vartheta(X_1) \qquad \textit{minden } \vartheta \in \Theta\textit{-ra},$$

vagyis a **mintaátlag** *torzítatlan becslés* ψ*-re.*

Állítás (A szórásnégyzet torzítatlan becslése)

$X_1, \ldots, X_n$ *független azonos eloszlású véges szórású minta. Ekkor Ekkor*

$$\mathbb{E}_\vartheta(s_n^{*2}) = D_\vartheta^2(X_1) \qquad \textit{minden } \vartheta \in \Theta\textit{-ra},$$

vagyis a **korrigált tapasztalati szórásnégyzet** *torzítatlan becslés a szórásnégyzet-re.*

Bizonyításvázlat a következő oldalakon.

Az átlag várható értéke

Állítás

Legyen $X_1, \ldots, X_n$ *független azonos eloszlású minta, és* $m = \mathbb{E}(X_i) < \infty$*. Ekkor*

$$\mathbb{E}(\overline{X}) = m.$$

Az átlag várható értéke

Állítás

Legyen $X_1, \ldots, X_n$ *független azonos eloszlású minta, és* $m = \mathbb{E}(X_i) < \infty$. *Ekkor*

$$\mathbb{E}(\overline{X}) = m.$$

Bizonyítás.

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n}\mathbb{E}(X_1 + \ldots + X_n) = \frac{1}{n} \cdot nm = m.$$

Felhasználtuk a várható érték linearitását, és hogy csak eloszlástól függ:

- $\mathbb{E}(cX) = c\mathbb{E}(X)$, ha $c \in \mathbb{R}$;
- $\mathbb{E}(Y + Z) = \mathbb{E}(Y) + \mathbb{E}(Z)$;
- ha Y és Z eloszlása megegyezik, akkor $\mathbb{E}(Y) = \mathbb{E}(Z)$

Tehát a **mintaátlag** torzítatlan becslés a várható értékre.

Az átlag szórása

Állítás

Legyen $X_1, \ldots, X_n$ *független azonos eloszlású minta, és* $\mathbb{E}(X_i^4) < \infty$*. Ekkor*

$$D(\overline{X}) = \frac{D(X_1)}{\sqrt{n}}.$$

Az átlag szórása

Állítás

Legyen $X_1, \ldots, X_n$ *független azonos eloszlású minta, és* $\mathbb{E}(X_i^4) < \infty$. *Ekkor*

$$D(\overline{X}) = \frac{D(X_1)}{\sqrt{n}}.$$

Bizonyítás.

$$D(\overline{X}) = D\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{D(X_1 + \ldots + X_n)}{n} = \frac{\sqrt{n\sigma^2}}{n} = \frac{\sigma}{\sqrt{n}}.$$

Felhasználtuk a szórás alábbi tulajdonságait:

- $D(cX) = |c|D(X)$, ha $c \in \mathbb{R}$;
- $D^2(Y + Z) = D^2(Y) + D^2(Z)$, ha Y és Z függetlenek;
- ha Y és Z eloszlása megegyezik, akkor $D(Y) = D(Z)$

A tapasztalati szórásnégyzet

Állítás (A tapasztalati szórásnégyzet másik alakja)

$$s_n^2 = \frac{1}{n}\left[\sum_{k=1}^{n} X_k^2\right] - \overline{X}^2.$$

Bizonyítás. Átrendezéssel kapjuk, hogy

$$\sum_{k=1}^{n}(X_k - \overline{X})^2 = \sum_{k=1}^{n}\left[X_k^2 - 2X_k \cdot \overline{X} + \overline{X}^2\right] = \sum_{k=1}^{n} X_k^2 - 2n\overline{X} \cdot \overline{X} + n \cdot \overline{X}^2 =$$

$$= \sum_{k=1}^{n} X_k^2 - n \cdot \overline{X}^2.$$

Ebből adódik, hogy

$$s_n^2 = \frac{1}{n}\left[\sum_{k=1}^{n}(X_k - \overline{X})^2\right] = \frac{1}{n}\left[\sum_{k=1}^{n} X_k^2\right] - \overline{X}^2,$$

a tapasztalati szórásnégyzet definíciója alapján.

A korrigált tapasztalati szórásnégyzet

$$s_n^{*2} = \frac{n}{n-1} s_n^2 = \frac{n}{n-1} \left[\frac{1}{n} \left[\sum_{k=1}^{n} X_k^2 \right] - \overline{X}^2 \right] = \frac{1}{n-1} \left[\sum_{k=1}^{n} X_k^2 \right] - \frac{n}{n-1} \overline{X}^2.$$

Az első tag várható értéke a szórásnégyzet definíciója alapján:

$$\mathbb{E}_\vartheta \left(\sum_{k=1}^{n} X_k^2 \right) = \sum_{k=1}^{n} \mathbb{E}_\vartheta (X_k^2) = n \cdot \mathbb{E}_\vartheta (X_1^2) = n \cdot \left[D_\vartheta^2 (X_1) + \mathbb{E}_\vartheta (X_1)^2 \right].$$

A második tag várható értéke az átlag szórásnégyzete alapján:

$$\mathbb{E}_\vartheta (\overline{X}^2) = D_\vartheta^2 (\overline{X}^2) + \mathbb{E}_\vartheta (\overline{X})^2 = \frac{1}{n} D_\vartheta^2 (X_1) + \mathbb{E}_\vartheta (X_1)^2.$$

Vagyis valóban s_n^{*2} torzítatlan becslés a szórásnégyzetre:

$$\mathbb{E}_\vartheta (s_n^{*2}) = \frac{n}{n-1} \left[D_\vartheta^2 (X_1) + \mathbb{E}_\vartheta (X_1)^2 \right] - \frac{n}{n-1} \left[\frac{1}{n} D_\vartheta^2 (X_1) + \mathbb{E}_\vartheta (X_1)^2 \right] = D_\vartheta^2 (X_1).$$

Becslések összehasonlítása

Definíció (Hatásosság)

Legyenek T_1, T_2 **torzítatlan** *becslései a paraméter $\psi(\vartheta)$ függvényének. T_1* **hatásosabb** *T_2-nél, ha*

$$D^2_\vartheta(T_1) \leq D^2_\vartheta(T_2)$$

teljesül minden $\vartheta \in \Theta$-ra.

A T_1 becslés **hatásos** *$\psi(\vartheta)$-ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).*

Becslések összehasonlítása

Definíció (Hatásosság)

Legyenek T_1, T_2 **torzítatlan** *becslései a paraméter $\psi(\vartheta)$ függvényének. T_1* **hatásosabb** *T_2-nél, ha*

$$D_\vartheta^2(T_1) \leq D_\vartheta^2(T_2)$$

teljesül minden $\vartheta \in \Theta$-ra.

A T_1 becslés **hatásos** *$\psi(\vartheta)$-ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).*

- Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.
- A várható értékre nézve a mintaátlag hatásosabb minden $\sum_{j=1}^n c_j X_j$ alakú becslésnél (ahol $\sum_{j=1}^n c_j = 1$).

Becslések összehasonlítása

Definíció (Hatásosság)

Legyenek T_1, T_2 **torzítatlan** *becslései a paraméter $\psi(\vartheta)$ függvényének. T_1* **hatásosabb** *T_2-nél, ha*

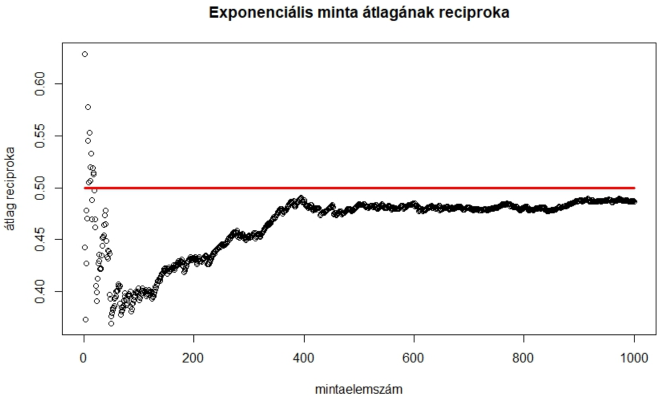
$$D_\vartheta^2(T_1) \leq D_\vartheta^2(T_2)$$

teljesül minden $\vartheta \in \Theta$-ra.

A T_1 becslés **hatásos** *$\psi(\vartheta)$-ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).*

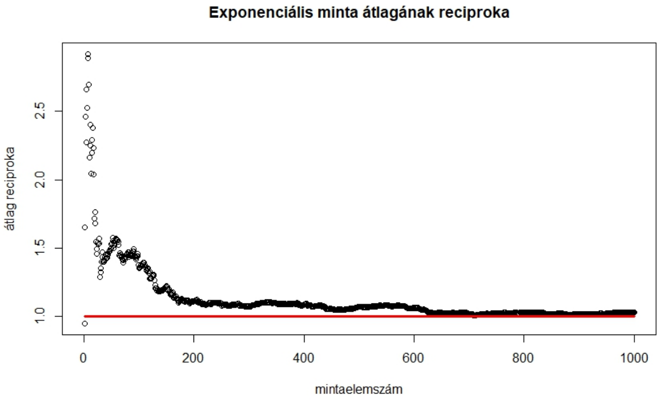
- Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.
- A várható értékre nézve a mintaátlag hatásosabb minden $\sum_{j=1}^n c_j X_j$ alakú becslésnél (ahol $\sum_{j=1}^n c_j = 1$).
- **Bizonyos feladatokban lehet a mintaátlagnál hatásosabb becslés a várható értékre**: A $[0, b]$ intervallumon egyenletes eloszlás esetén b-re $\frac{n+1}{n} \max(X_1, \ldots, X_n)$ hatásosabb a mintaátlag kétszeresénél.

Konzisztens becslés



$\lambda = 0,5$ paraméterű exponenciális eloszlást generálva a mintaátlag reciproka $0,5$-höz tart, azaz **konzisztens** becslés, hiszen ez minden λ-ra teljesül.

Konzisztens becslés



$\lambda = 1$ paraméterű exponenciális eloszlást generálva a mintaátlag reciproka 1-hez tart, azaz **konzisztens** becslés, hiszen ez minden λ-ra teljesül.

Konzisztencia

Definíció

A $T_n = T_n(X_1, \ldots, X_n)$ **konzisztens** *becsléssorozat* $\psi(\vartheta)$*-ra, ha minden* $\vartheta \in \Theta$*-ra*

$$(T_n(X_1, \ldots, X_n)) \to \psi(\vartheta)$$

$n \to \infty$ *esetén sztochasztikusan, azaz minden* $\vartheta \in \Theta$ *és* $\varepsilon > 0$*-ra teljesül, hogy*

$$\mathbb{P}_\vartheta\big(|T_n - \psi(\vartheta)| > \varepsilon\big) \to 0 \qquad (n \to \infty).$$

Konzisztencia

Definíció

A $T_n = T_n(X_1, \ldots, X_n)$ **konzisztens** *becsléssorozat* $\psi(\vartheta)$*-ra, ha minden* $\vartheta \in \Theta$*-ra*

$$(T_n(X_1, \ldots, X_n)) \to \psi(\vartheta)$$

$n \to \infty$ *esetén sztochasztikusan, azaz minden* $\vartheta \in \Theta$ *és* $\varepsilon > 0$*-ra teljesül, hogy*

$$\mathbb{P}_\vartheta\big(|T_n - \psi(\vartheta)| > \varepsilon\big) \to 0 \qquad (n \to \infty).$$

Elégséges feltétel:

$$\mathbb{E}_\vartheta(T(X)) \to \vartheta \qquad \text{és} \qquad D_\vartheta(T(X)) \to 0$$

minden $\vartheta \in \Theta$-ra.

Példák torzítatlan, konzisztens becslésekre

$X_1, X_2, \ldots$ független azonos eloszlású minta. Ekkor

$$T_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \mathbb{E}_\vartheta(X_1)$$

teljesül $n \to \infty$ esetén sztochasztikusan a nagy számok gyenge törvénye szerint, vagyis az **átlag** konzisztens becslés a **várható értékre**.

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűségre**.

Példák torzítatlan, konzisztens becslésekre

$X_1, X_2, \ldots$ független azonos eloszlású minta. Ekkor

$$T_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \mathbb{E}_\vartheta(X_1)$$

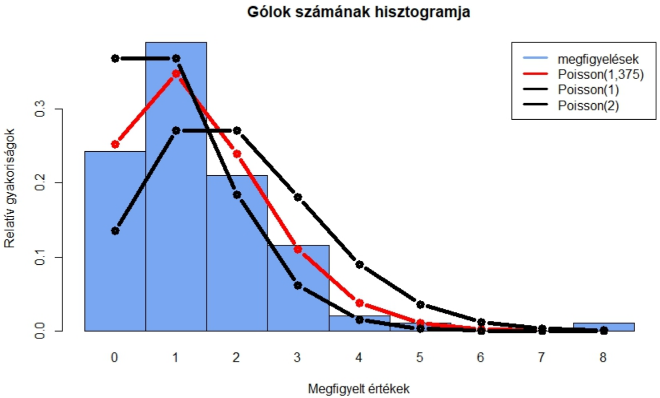
teljesül $n \to \infty$ esetén sztochasztikusan a nagy számok gyenge törvénye szerint, vagyis az **átlag** konzisztens becslés a **várható értékre**.

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűségre**.

Nevezetes eloszlások:

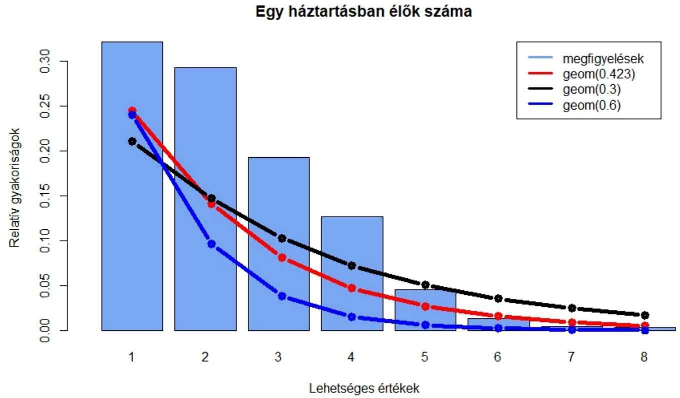
- Poisson-eloszlás λ paraméterére az átlag torzítatlan, konzisztens
- a normális eloszlás m paraméterére az átlag torzítatlan és konzisztens; a σ paraméterre a tapasztalati szórás és a korrigált tapasztalati szórás konzisztensek, de nem torzítatlanok; σ^2-re s_n^{*2} torzítatlan
- exponenciális eloszlás: $1/\overline{X}$ konzisztens λ-ra, de nem torzítatlan a paraméterre
- exponenciális eloszlás: $n \cdot \min(X_1, \ldots, X_n)$ torzítatlan, de nem konzisztens a várható értékre (vagyis $1/\lambda$-ra).

Poisson-eloszlás paraméterének becslése



A gólok számának hisztogramja $n = 95$ mérkőzésen, és különböző paraméterű Poisson-eloszlások ($\mathbb{P}_\lambda(X = k) = \lambda^k/k! \cdot e^{-\lambda}$)

Geometriai eloszlás paraméterének becslése



Egy háztartásban élők számának hisztogramja (forrás: KSH, 2011) és a geometriai eloszlás $p = 0,423$ (piros), $p = 0,3$ (fekete), és $p = 0,6$ (kék) paraméterekkel

$n = 4105698$ a háztartások száma, $1/\overline{X} = 0,423$; $\mathbb{P}_p(X = k) = (1-p)^{k-1} \cdot p$.

Maximumlikelihood-módszer

Definíció (Likelihood-függvény)

Ha az $(Y_1, \ldots, Y_n)$ *független minta diszkrét (a lehetséges értékeinek száma véges vagy megszámlálható sok), akkor a likelihood-függvénye:*

$$L_{n,\vartheta}(k_1, \ldots, k_n) = \prod_{j=1}^{n} \mathbb{P}_{j,\vartheta}(Y_j = k_j) \qquad ((k_1, \ldots, k_n) \in H).$$

Maximumlikelihood-módszer

Definíció (Likelihood-függvény)

Ha az $(Y_1, \ldots, Y_n)$ *független minta diszkrét (a lehetséges értékeinek száma véges vagy megszámlálható sok), akkor a likelihood-függvénye:*

$$L_{n,\vartheta}(k_1, \ldots, k_n) = \prod_{j=1}^{n} \mathbb{P}_{j,\vartheta}(Y_j = k_j) \qquad ((k_1, \ldots, k_n) \in H).$$

Ha az $(Y_1, \ldots, Y_n)$ *független minta abszolút folytonos, és* Y_j *sűrűségfüggvénye (a* $\mathbb{P}_\vartheta$ *valószínűség mellett)* $f_{j,\vartheta}$*, akkor a minta likelihood-függvénye:*

$$L_{n,\vartheta}(t_1, \ldots, t_n) = \prod_{j=1}^{n} f_{j,\vartheta}(t_j) \qquad (t_1, \ldots, t_n \in \mathbb{R}).$$

Maximumlikelihood-módszer

Legyen $(\Omega, \mathcal{A}, \mathbb{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$, vagyis az ismeretlen eloszlás a ϑ paraméterrel jellemezhető.

Definíció (Maximum-likelihood becslés)

A ϑ maximumlikelihood-becslése (ML-becslése) az $X_1, \ldots, X_n$ mintából $\hat{\vartheta}$, ha $\hat{\vartheta}$ maximalizálja a $\vartheta \mapsto L_{n,\vartheta}(X_1, \ldots, X_n)$ függvényt, ahol $L_{n,\vartheta}$ a minta likelihood-függvénye. Azaz, ha

$$L_{n,\hat{\vartheta}}(X_1, \ldots, X_n) \geq L_{n,\vartheta}(X_1, \ldots, X_n) \text{ minden } \vartheta \in \Theta\text{-ra.}$$

Példa. $X_1, \ldots, X_n$ függetlenek, eloszlásuk exponenciális eloszlás $\vartheta > 0$ paraméterrel. Ekkor

$$L_{n,\vartheta}(X_1, \ldots, X_n) = \prod_{j=1}^{n} f_{j,\vartheta}(X_j) = \prod_{j=1}^{n} \Big[\vartheta \exp(-\vartheta X_j) \mathbb{I}(X_j > 0)\Big],$$

amiből $\hat{\vartheta} = \frac{1}{\overline{X}}$.

Poisson-eloszlás paraméterének becslése

Tegyük fel, hogy $X_1, X_2, \ldots, X_n$ független, azonos λ paraméterű Poisson-eloszlású minta, ahol $\lambda > 0$ ismeretlen paraméter, $n = 95$, és $\overline{X} = 1,379$.

Poisson-eloszlásnál:

$$\mathbb{P}_\lambda(X_j = k) = \frac{\lambda^k}{k!} e^{-\lambda}; \qquad \mathbb{E}(X_j) = \lambda; \qquad D(X_j) = \sqrt{\lambda}.$$

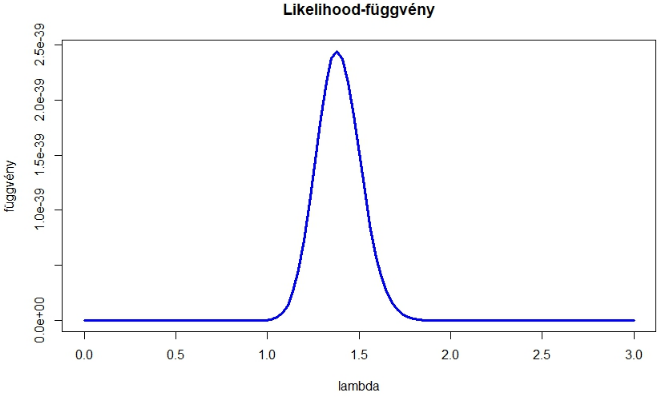
A megfigyelések az alábbiak (a gólok száma összesen $\sum_{j=1}^{n} X_j = 131$):

$$3, \quad 0, \quad 2, \quad 2, \quad 1, \quad 3, \ldots, 2.$$

Annak valószínűsége λ paraméter mellett, hogy éppen ezt a sorozatot kaptuk:

$$\begin{aligned} L_{95,\lambda}(3,0,2,\ldots,2) &= \frac{\lambda^3}{3!} e^{-\lambda} \cdot \frac{\lambda^0}{0!} e^{-\lambda} \cdot \frac{\lambda^2}{2!} e^{-\lambda} \cdot \ldots \cdot \frac{\lambda^2}{2!} e^{-\lambda} = \\ &= \frac{\lambda^{3+0+2+2\ldots+2}}{3! \cdot 0! \cdot 2! \cdot \ldots \cdot 2!} e^{-95\cdot\lambda} = \frac{\lambda^{131}}{3! \cdot 0! \cdot 2! \cdot \ldots \cdot 2!} e^{-95\cdot\lambda} \end{aligned}$$

Poisson-eloszlás paraméterének becslése



A $\frac{\lambda^{131}}{3!\cdot 0!\cdot 2!\cdot \ldots \cdot 2!} e^{-95\cdot\lambda}$ likelihoodfüggvény a $\lambda > 0$ paraméter függvényében;

mintaátlag: $\overline{X} = \frac{131}{95} = 1,379$

ML-becslés: Poisson-eloszlás

$X_1, \ldots, X_n$ függetlenek, Poisson-eloszlás $\lambda > 0$ ismeretlen paraméterrel, azaz

$$\mathbb{P}(X_j = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad (k = 0, 1, 2, \ldots).$$

Ekkor

$$L_{n,\lambda}(X_1, \ldots, X_n) = \prod_{j=1}^{n} \left(\frac{\lambda^{X_j}}{X_j!} e^{-\lambda} \right) = \frac{\lambda^{X_1}}{X_1!} e^{-\lambda} \cdot \frac{\lambda^{X_2}}{X_2!} e^{-\lambda} \cdot \ldots \cdot \frac{\lambda^{X_n}}{X_n!} e^{-\lambda}.$$

$$L_{n,\lambda}(X_1, \ldots, X_n) = \lambda^{\sum_{j=1}^{n} X_j} e^{-n\lambda} \cdot \prod_{j=1}^{n} \frac{1}{X_j!}.$$

ML-becslés: Poisson-eloszlás

$X_1, \dots, X_n$ függetlenek, Poisson-eloszlás $\lambda > 0$ ismeretlen paraméterrel, azaz

$$\mathbb{P}(X_j = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad (k = 0, 1, 2, \dots).$$

Ekkor

$$L_{n,\lambda}(X_1, \dots, X_n) = \prod_{j=1}^{n} \left(\frac{\lambda^{X_j}}{X_j!} e^{-\lambda} \right) = \frac{\lambda^{X_1}}{X_1!} e^{-\lambda} \cdot \frac{\lambda^{X_2}}{X_2!} e^{-\lambda} \cdot \dots \cdot \frac{\lambda^{X_n}}{X_n!} e^{-\lambda}.$$

$$L_{n,\lambda}(X_1, \dots, X_n) = \lambda^{\sum_{j=1}^{n} X_j} e^{-n\lambda} \cdot \prod_{j=1}^{n} \frac{1}{X_j!}.$$

$$\log L_{n,\lambda}(X_1, \dots, X_n) = \log \lambda \sum_{j=1}^{n} X_j - n\lambda + \log \prod_{j=1}^{n} \frac{1}{X_j!}$$

ML-becslés: Poisson-eloszlás

$X_1, \ldots, X_n$ függetlenek, Poisson-eloszlás $\lambda > 0$ ismeretlen paraméterrel, azaz

$$\mathbb{P}(X_j = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad (k = 0, 1, 2, \ldots).$$

Ekkor

$$L_{n,\lambda}(X_1, \ldots, X_n) = \prod_{j=1}^{n} \left(\frac{\lambda^{X_j}}{X_j!} e^{-\lambda} \right) = \frac{\lambda^{X_1}}{X_1!} e^{-\lambda} \cdot \frac{\lambda^{X_2}}{X_2!} e^{-\lambda} \cdot \ldots \cdot \frac{\lambda^{X_n}}{X_n!} e^{-\lambda}.$$

$$L_{n,\lambda}(X_1, \ldots, X_n) = \lambda^{\sum_{j=1}^{n} X_j} e^{-n\lambda} \cdot \prod_{j=1}^{n} \frac{1}{X_j!}.$$

$$\log L_{n,\lambda}(X_1, \ldots, X_n) = \log \lambda \sum_{j=1}^{n} X_j - n\lambda + \log \prod_{j=1}^{n} \frac{1}{X_j!}$$

$$\frac{\partial}{\partial \lambda} \log L_{n,\lambda}(X_1, \ldots, X_n) = \frac{\sum_{j=1}^{n} X_j}{\lambda} - n > 0 \Leftrightarrow \lambda < \overline{X}.$$

Ezért az ML-becslés: $\hat{\lambda} = \overline{X}$. A log monoton növekedését használtuk.