

Statisztikai mező (3. előadás)

Definíció

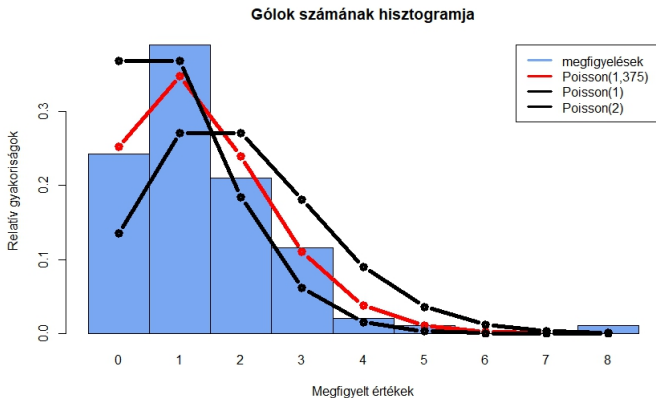
Az $(\Omega, \mathcal{A}, \mathcal{P})$ hármast **statisztikai mezőnek** nevezzük, ha minden $\mathbb{P} \in \mathcal{P}$ -re $(\Omega, \mathcal{A}, \mathbb{P})$ Kolmogorov-féle valószínűségi mező.

Paraméteres statisztika mező: $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$. Ekkor ϑ az ismeretlen paraméter, mely egy $\Theta \subseteq \mathbb{R}^q$ ismert halmaz eleme.

Például: \mathcal{P} lehet például

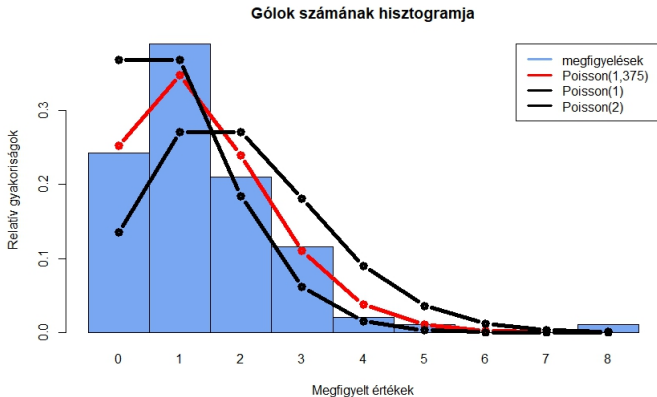
- a λ paraméterű Poisson-eloszlások halmaza;
- a normális eloszlások halmaza (ekkor $\vartheta = (m, \sigma)$ az ismeretlen paraméter);
- az $[a, b]$ intervallumon egyenletes eloszlások halmaza (ekkor $\vartheta = (a, b)$ az ismeretlen paraméter).

Poisson-eloszlás paraméterének becslése



A gólok számának hisztogramja $n = 95$ mérkőzésen, és különböző paraméterű Poisson-eloszlások. Ekkor

Poisson-eloszlás paraméterének becslése

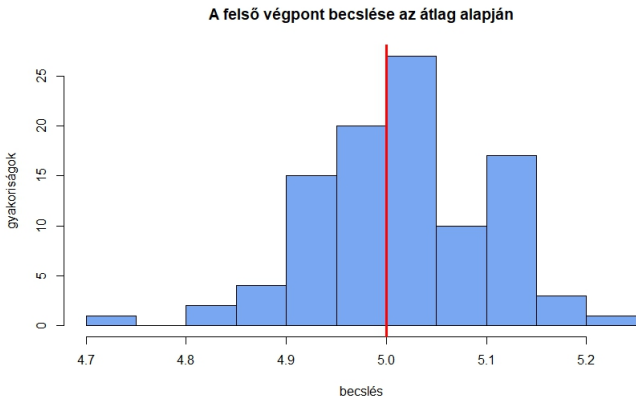


A gólok számának histogramja $n = 95$ mérkőzésen, és különböző paraméterű Poisson-eloszlások. Ekkor

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X_1) = \lambda \quad \text{minden } \lambda\text{-ra.}$$

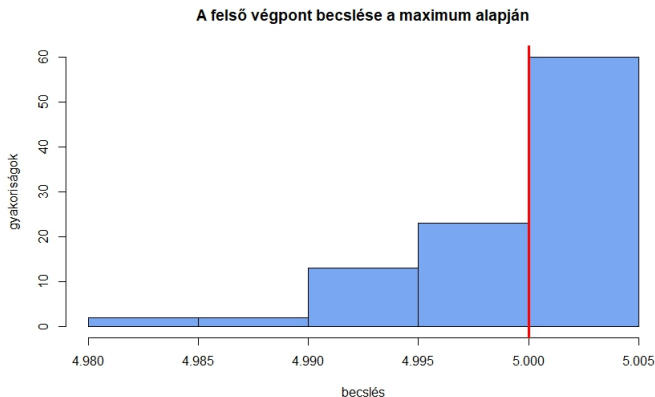
Ha az átlaggal becsüljük a paramétert, akkor a becslésünk várható értéke λ .

Egyenletes eloszlás paraméterének becslése



A $[0, \vartheta]$ intervallumon egyenletes eloszlás paraméterének becslése $T(X) = 2\bar{X}$ -szel, $n = 1000$ elemű mintából. Az igazi paraméter $\vartheta = 5$. Száz ismétlésből a becslések átlaga: **5,015**, korigált tapasztalati szórása: **0,086**.

Egyenletes eloszlás paraméterének becslése



A $[0, \vartheta]$ intervallumon egyenletes eloszlás paraméterének becslése a maximum alapján $T(X) = \frac{n+1}{n} \max(X_1, \dots, X_n)$ -nel, $n = 1000$ elemű mintából. Az igazi paraméter $\vartheta = 5$. Száz ismétlésből a becslések átlaga: **4,9999**, korigált tapasztalati szórása: **0,0049** < **0,086**.

Egyenletes eloszlás paraméterének becslése

X_1, X_2, \dots, X_n független azonos eloszlású, a $[0, \vartheta]$ intervallumon egyenletes eloszlású valószínűségi változók. Itt $\vartheta > 0$ ismeretlen paraméter. Tekintsük a ϑ két különböző becslését:

$$T_1(X_1, \dots, X_n) = 2\bar{X}; \quad \mathbb{E}_{\vartheta}(T_1) = 2 \cdot \frac{\vartheta}{2} = \vartheta; \quad D_{\vartheta}(T_1) = \frac{\vartheta}{\sqrt{3n}}.$$

Másrészt, felhasználva, hogy $X_n^* = \max(X_1, \dots, X_n)$ sűrűségfüggvénye: $f_{\vartheta}(t) = (nt^{n-1}/\vartheta^n)\mathbb{I}(0 \leq t \leq \vartheta)$, egy másik becslést is találhatunk:

$$T_2(X_1, \dots, X_n) = \frac{n+1}{n} \cdot X_n^*; \quad \mathbb{E}_{\vartheta}(T_2) = \frac{n+1}{n} \cdot \frac{n\vartheta}{n+1} = \vartheta.$$

$$D_{\vartheta}(T_2) = \sqrt{\frac{n \cdot \vartheta^2}{(n+2)(n+1)^2}} \leq \frac{\vartheta}{n+1} < \frac{\vartheta}{\sqrt{3n}}.$$

Egyenletes eloszlás paraméterének becslése

X_1, X_2, \dots, X_n független azonos eloszlású, a $[0, \vartheta]$ intervallumon egyenletes eloszlású valószínűségi változók. Itt $\vartheta > 0$ ismeretlen paraméter. Tekintsük a ϑ két különböző becslését:

$$T_1(X_1, \dots, X_n) = 2\bar{X}; \quad \mathbb{E}_{\vartheta}(T_1) = 2 \cdot \frac{\vartheta}{2} = \vartheta; \quad D_{\vartheta}(T_1) = \frac{\vartheta}{\sqrt{3n}}.$$

Másrészt, felhasználva, hogy $X_n^* = \max(X_1, \dots, X_n)$ sűrűségfüggvénye: $f_{\vartheta}(t) = (nt^{n-1}/\vartheta^n)\mathbb{I}(0 \leq t \leq \vartheta)$, egy másik becslést is találhatunk:

$$T_2(X_1, \dots, X_n) = \frac{n+1}{n} \cdot X_n^*; \quad \mathbb{E}_{\vartheta}(T_2) = \frac{n+1}{n} \cdot \frac{n\vartheta}{n+1} = \vartheta.$$

$$D_{\vartheta}(T_2) = \sqrt{\frac{n \cdot \vartheta^2}{(n+2)(n+1)^2}} \leq \frac{\vartheta}{n+1} < \frac{\vartheta}{\sqrt{3n}}.$$

Mindkét becslés **torzítatlan** ϑ -ra, és a második **hatásosabb**.

Becslések és tulajdonságaik

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

Becslések és tulajdonságaik

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

Definíció (Torzítatlanság)

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

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

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

Példa. X_1, X_2, \dots, X_n független minta a $[0, \vartheta]$ intervallumon egyenletes eloszlásból.

Becslések és tulajdonságaik

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

Definíció (Torzítatlanság)

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

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

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

Példa. X_1, X_2, \dots, X_n független minta a $[0, \vartheta]$ intervallumon egyenletes eloszlásból. Ekkor $2\bar{X}$ torzítatlan becslés $g(\vartheta) = \vartheta$ -ra: $\mathbb{E}(2\bar{X}) = \vartheta$.

Torzítatlan becslések

Állítás (A várható érték torzítatlan becslése)

Legyen X_1, \dots, X_n független azonos eloszlású véges várható értékű minta. Ekkor

$$\mathbb{E}_\vartheta(\bar{X}) = \mathbb{E}_\vartheta(X_1) \quad \text{minden } \vartheta \in \Theta\text{-ra,}$$

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

Állítás (A szórásnégyzet torzítatlan becslése)

X_1, \dots, X_n független azonos eloszlású véges szórású minta. Ekkor Ekkor

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

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

Az átlag várható értéke

Állítás

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

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

Az átlag várható értéke

Állítás

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

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

Bizonyítás.

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

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.

Speciálisan: a **relatív gyakoriság** torzítatlan becslés egy esemény valószínűségére.

Az átlag szórása

Állítás

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

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

Az átlag szórása

Állítás

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

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

Bizonyítás.

$$D(\bar{X}) = D\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{D(X_1 + \dots + X_n)}{n} = \frac{\sqrt{nD^2(X_1)}}{n} = \frac{D(X_1)}{\sqrt{n}}.$$

Felhasználtuk a szórás alábbi tulajdonságait:

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

Tapasztalati szórásnégyzet

Állítás (A tapasztalati szórásnégyzet másik alakja)

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

Bizonyítás. Átrendezéssel kapjuk, hogy

$$\begin{aligned} \sum_{k=1}^n (X_k - \bar{X})^2 &= \sum_{k=1}^n [X_k^2 - 2X_k \cdot \bar{X} + \bar{X}^2] = \sum_{k=1}^n X_k^2 - 2n\bar{X} \cdot \bar{X} + n \cdot \bar{X}^2 = \\ &= \sum_{k=1}^n X_k^2 - n \cdot \bar{X}^2. \end{aligned}$$

Ebből adódik, hogy

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

a tapasztalati szórásnégyzet definíciója alapján.

Korrigált tapasztalati szórásnégyzet

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

Az első tag várható értéke a szórásnégyzet definíciója alapján:

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

A második tag várható értéke az átlag szórásnégyzete alapján:

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

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

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

Becslések összehasonlítása

Definíció (Hatásosság)

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

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

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).

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

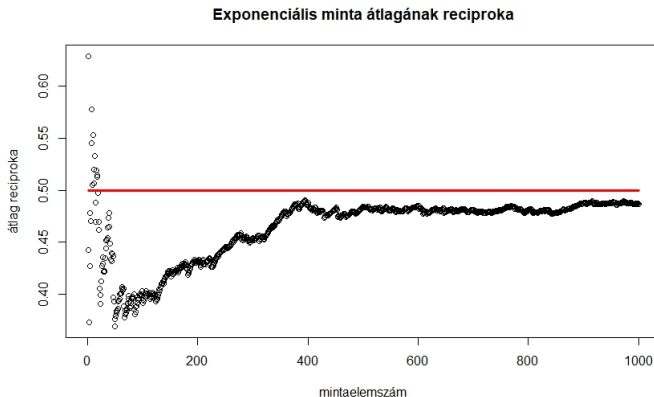
$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

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).

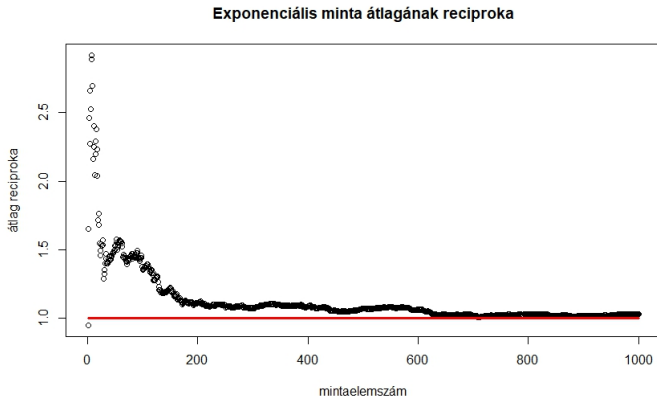
- Nem mindig létezik hatásos becslés, és lehetséges, hogy T_1 és T_2 közül egyik sem hatásosabb a másiknál.
- A várható értékre nézve a mintaátlag hatásosabb minden $\sum_{j=1}^n c_j X_j$ alakú becslésnél (ahol $\sum_{j=1}^n c_j = 1$).
- **Bizonyos feladatokban lehet a mintaátlagnál hatásosabb becslés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

Exponenciális eloszlás



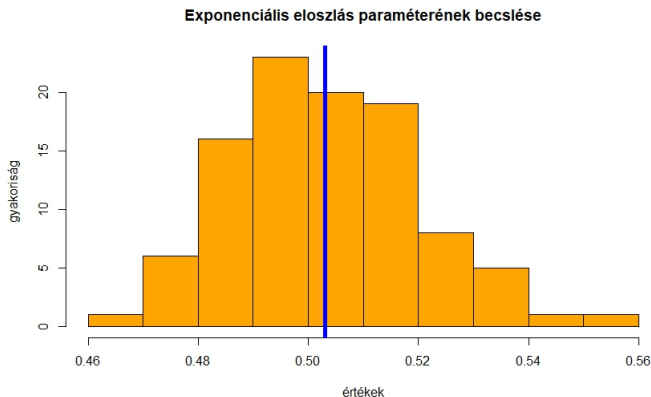
$\lambda = 0,5$ paraméterű exponenciális eloszlást generálva a mintaátlag reciproka 0,5-höz tart, azaz **konzisztens** becslés, hiszen ez minden λ -ra teljesül.

Exponenciális eloszlás



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

Exponenciális eloszlás



$\lambda = 0,5$ paraméterű exponenciális eloszlásból 100000 elemű mintát generáltunk, és ezresével csoportosítottuk a megfigyeléseket. Az ezres csoportokhoz hozzárendeltük az átlag reciprokát. Ennek hisztogramja látható. Az átlagok átlaga: **0,5031**. Az átlagok szórása: **0,0168**.

Konzisztencia

Definíció

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

$$(T_n(X_1, \dots, X_n)) \rightarrow \psi(\vartheta)$$

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

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

Konzisztencia

Definíció

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

$$(T_n(X_1, \dots, X_n)) \rightarrow \psi(\vartheta)$$

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

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

Elégséges feltétel:

$$\mathbb{E}_\vartheta(T_n(X)) \rightarrow \vartheta \quad \text{és} \quad D_\vartheta(T_n(X)) \rightarrow 0$$

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

Példák torzítatlan, konzisztens becslésekre

X_1, X_2, \dots független azonos eloszlású minta. Ekkor

$$T_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}_\theta(X_1)$$

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

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűsége**re.

Példák torzítatlan, konzisztens becslésekre

X_1, X_2, \dots független azonos eloszlású minta. Ekkor

$$T_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}_\theta(X_1)$$

teljesül $n \rightarrow \infty$ 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: $1/\bar{X}$ konzisztens λ -ra, de nem torzítatlan a paraméterre
- exponenciális eloszlás: $(n+1) \cdot \min(X_1, \dots, X_n)$ torzítatlan, de nem konzisztens a várható értékre (vagyis $1/\lambda$ -ra).

Házi feladat február 26., 8:15-ig: megoldás

Tekintsük az utazással töltött időkből gyűjtött 25 elemű mintát. Készítsük el a sűrűségfüggvény becslését úgy, hogy

- csak az első 5 megfigyelést használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- csak a nők adatait használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- csak a férfiak adatait használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- az összes megfigyelést használjuk, és a magfüggvény az Epanechnikov-magfüggvény;
- az összes megfigyelést használjuk, és a magfüggvény a Gauss-magfüggvény.

Nem kell mindegyikhez külön ábra, az összehasonlítás kedvéért lehet egy ábrán több görbe is.

Házi feladat február 26., 8:15-ig: megoldás

```
plot(density(no_utazas, kernel='epanechnikov'), col='red', lwd="3", main="Utazási  
idő sűrűségfüggvényének becslése", xlab="idő (perc)", ylab="f(x)")
```

```
lines(density(ferfi_utazas, kernel='epanechnikov'), col='blue', lwd="3")
```

```
lines(density(utazas, kernel='epanechnikov'), col='black', lwd="3")
```

```
legend("topright", c("nők (17)", "férfiak (11)", "összesen (28)"), col=c("red",  
"blue", "black"), lwd="3")
```

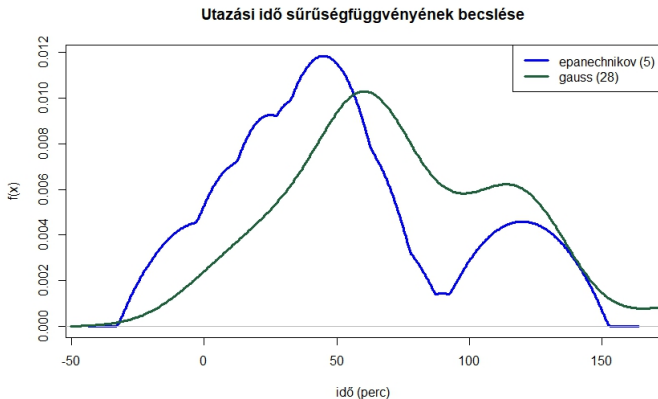
```
plot(density(ot, kernel='epanechnikov'), col='blue', lwd="3", main="Utazási idő  
sűrűségfüggvényének becslése", xlab="idő (perc)", ylab="f(x)")
```

```
lines(density(utazas, kernel='gauss'), col="#1f613d", lwd="3")
```

```
legend("topright", c("epanechnikov (5)", "gauss (28)"), col=c("blue", "#1f613d"),  
lwd="3")
```

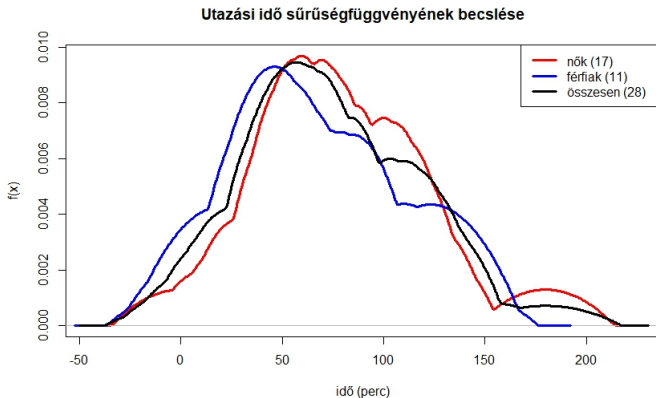
excel: descriptive statistics → kernel density estimation (magfüggvény: options)

Házi feladat február 26., 8:15-ig: megoldás



Az utazási idő sűrűségfüggvényének becslése az első öt mintaelem alapján Epanechnikov-magfüggvénnyel és az összes mintaelem alapján Gauss-magfüggvénnyel

Házi feladat február 26., 8:15-ig: megoldás



Az utazási idő sűrűségfüggvényének becslése Epanechnikov-magfüggvénnyel külön a nők és a férfiak esetében, illetve összesen

Maximumlikelihood-módszer

Van két telefontöltőnk, melyek ránézésre megkülönböztethetlenné. A jó minden kipróbálásnál a többletől függetlenül 90%, a rossz 20% valószínűséggel működik.

Az egyik kábelt kiválasztottuk, 10 kipróbálásból pontosan 8-szor működött. Ez melyik kábel lehetett?

Maximumlikelihood-módszer

Van két telefontöltőnk, melyek ránézésre megkülönböztethetetlenek. A jó minden kipróbálásnál a többletől függetlenül 90%, a rossz 20% valószínűséggel működik.

Az egyik kábelt kiválasztottuk, 10 kipróbálásból pontosan 8-szor működött. Ez melyik kábel lehetett?

$$\mathbb{P}(\text{a jó kábel 8-szor működik}) = \binom{10}{8} 0,9^8 0,1^2 = \mathbf{19,8\%}.$$

$$\mathbb{P}(\text{a rossz kábel 8-szor működik}) = \binom{10}{8} 0,2^8 0,8^2 = \mathbf{0,00007\%}.$$

Inkább a jó kábel lehetett, amit kiválasztottunk.

Maximumlikelihood-módszer

Van egy telefontöltőnk, mely minden kipróbálásnál a többitől függetlenül p valószínűséggel működik.

Itt az ismeretlen paraméter: $p \in [0, 1]$.

A kábel 10 kipróbálásból pontosan 8-szor működött. Mennyi lehet p értéke?

Maximumlikelihood-módszer

Van egy telefontöltőnk, mely minden kipróbálásnál a többletől függetlenül p valószínűséggel működik.

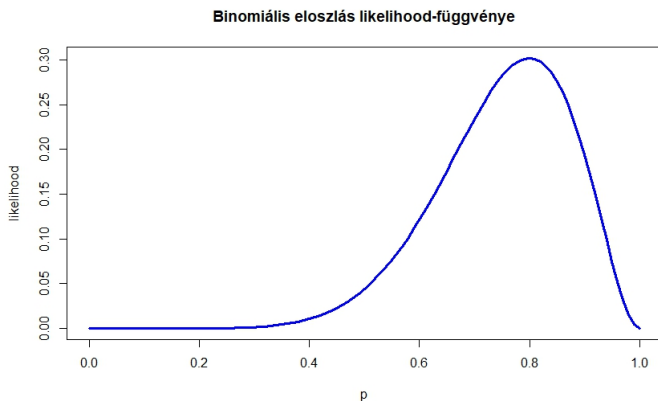
Itt az ismeretlen paraméter: $p \in [0, 1]$.

A kábel 10 kipróbálásból pontosan 8-szor működött. Mennyi lehet p értéke?

$$\mathbb{P}_p(\text{a kábel 8-szor működik}) = \binom{10}{8} p^8 (1-p)^2.$$

Kérdés: ez milyen $p \in [0, 1]$ -re lesz a legnagyobb?

Maximumlikelihood-módszer



A likelihood-függvény: $\mathbb{P}_p(\text{a kábel 8-szor működik}) = \binom{10}{8} p^8 (1-p)^2$ a p ismeretlen paraméter függvényében. A maximumhelye: $\hat{p} = 0,8$.

Maximumlikelihood-módszer

Definíció (Likelihood-függvény)

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

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

Maximumlikelihood-módszer

Definíció (Likelihood-függvény)

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

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

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

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

Maximumlikelihood-módszer

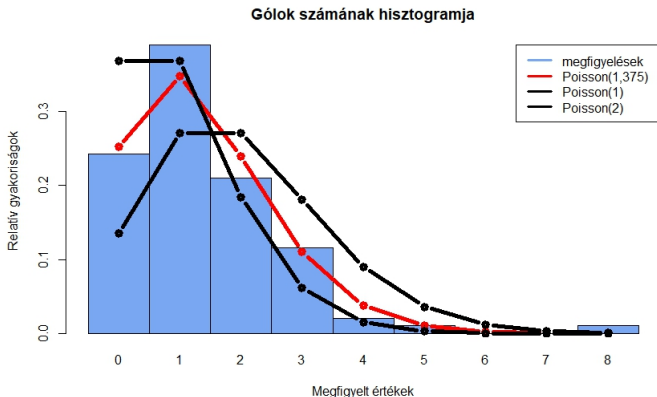
Legyen $(\Omega, \mathcal{A}, \mathbb{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$, vagyis az ismeretlen eloszlás a ϑ paraméterrel jellemezhető.

Definíció (Maximum-likelihood becslés)

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

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

Poisson-eloszlás paraméterének becslése



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

Poisson-eloszlás paraméterének becslése

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

Poisson-eloszlásnál:

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

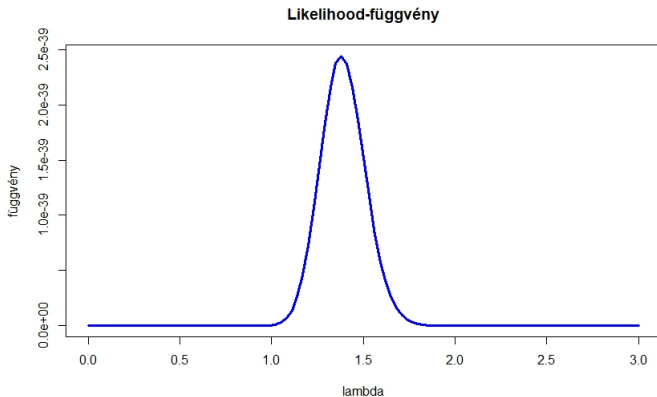
A megfigyelések az alábbiak (a gólok száma összesen $\sum_{j=1}^n X_j = 131$):

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

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

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

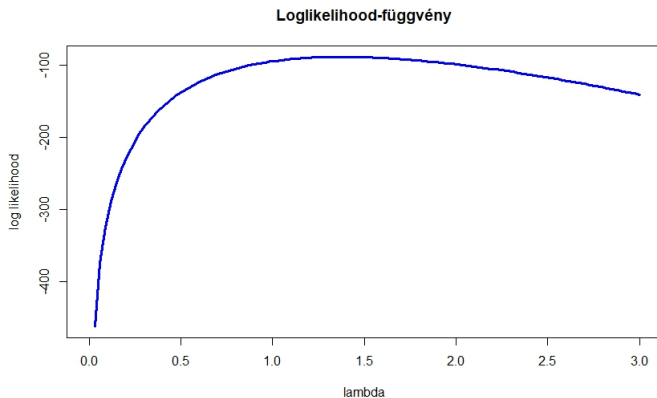
Poisson-eloszlás paraméterének becslése



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

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

Poisson-eloszlás paraméterének becslése



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

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

ML-becslés: Poisson-eloszlás

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

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

Ekkor

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

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

ML-becslés: Poisson-eloszlás

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

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

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

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

Ekkor

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

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

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

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

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

Házi feladat március 4., 8:15-ig

Tegyük fel, hogy egy véletlenszerűen kiválasztott ismerős által az utóbbi egy hónapban nézett sorozatok száma **geometriai eloszlású**, a paraméter, $p > 0$ ismeretlen.

A rendelkezésre álló 25 elemű mintára vonatkozóan írjuk fel a likelihood-függvény képletét, és ábrázoljuk is ezt a függvényt.

Határozzuk meg a p paraméter maximumlikelihood-becslését (a konkrét adatokból), azaz azt a \hat{p} -t, amire a likelihoodfüggvény a legnagyobb (elég az ábra alapján).