

Matematikai statisztika előadás, 3. hét, február 24.
Becslések tulajdonságai, maximumlikelihood-becslés

1. Statisztikai mező

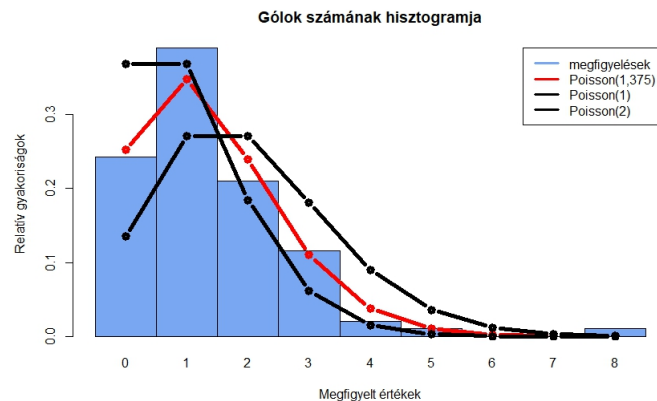
A matematikai statisztika egyik fő célja, hogy ha az X_1, X_2, \dots, X_n ismeretlen eloszlású minta, akkor erről az ismeretlen eloszlásról minél több információt nyerjen. Azt, hogy az eloszlás ismeretlen, úgy fogalmazhatjuk meg, hogy olyan valószínűségi mezőket tekintünk, ahol az eseménytér és az események halmaza ugyanaz, de a valószínűség, ami megmondja, hogy melyik esemény mennyire valószínű, eltérő.

Például annak valószínűsége, hogy egy véletlenszerűen választott ember jövedelme több 500000 forintnál, egész más lehet, ha a jövedelem (mondjuk) 300000 várható értékű és 100000 szórású normális eloszlású, vagy ha a jövedelem ezzel azonos várható értékű, de például $\alpha = 3$ rendű, vagyis végtelen szórású Pareto-eloszlással írható le.

Így juthatunk el az alábbi definícióhoz.

1.1. 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.

Ennek fontos speciális esete, amikor feltételezzük, hogy az eloszlás egy néhány paraméterrel leírható eloszláscsaládból származik, azon belül azonban nem tudjuk, hogy melyik eloszlásról van szó. Például feltételezzük, hogy a jövedelem eloszlása Pareto-eloszlás, de nem ismerjük a paramétereit, vagy feltételezzük, hogy egy betegség lappangási ideje normális eloszlású, de nem ismerjük a várható értéket és a szórásot. Az ismeretlen paraméter vagy paramétereket általánosan ϑ -val jelöljük.



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

1.2. Definíció. 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, a paramétertér egy eleme.

Például: \mathcal{P} lehet például

- a Poisson-eloszlások halmaza, $\vartheta = \lambda$ az ismeretlen paraméter, $\Theta = (0, \infty)$ a paraméter lehetséges értékeinek 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.

Az 1. ábra azt mutatja, hogy ha például a gólok számát Poisson-eloszlásúnak feltételezzük, de a paramétert ismeretlennek tekintjük, akkor néhány különböző λ érték mellett mennyire jól illeszkedik a λ -hoz tartozó eloszlás a megfigyelésekhez.

1.3. Definíció. *Statisztikai minta:* X_1, X_2, \dots, X_n valószínűségi változók az $(\Omega, \mathcal{A}, \mathcal{P})$ valószínűségi mezőn.

A minta független, ha ezek a valószínűségi változók függetlenek.

Statisztika: ha $T : \mathbb{R}^n \rightarrow \mathbb{R}^k$ egy n változós függvény, akkor a $T(X_1, \dots, X_n)$ valószínűségi változót statisztikának nevezzük.

A statisztika tehát olyan mennyiség, amit a megfigyelésekből, a mintából egy megfelelő, előre rögzített függvény alkalmazásával ki tudunk számolni.

Például $k = 1$ -re példa: $T(X_1, \dots, X_n) = \frac{X_1 + \dots + X_n}{n}$ esetén a statisztika az átlag.

Vagy $k = 2$ -re példa: $T(X_1, \dots, X_n) = (\bar{X}, s_n^*)$ az a statisztika, ami a mintából az átlagot és a korrigált tapasztalati szórást számítja ki.

2. Torzítatlanság és hatásosság

Tegyük fel, hogy a $[0, \vartheta]$ intervallumon egyenletes eloszlás ismeretlen ϑ paraméterét szeretnénk becsülni (ez analóg a német tankok problémájával: https://en.wikipedia.org/wiki/German_tank_problem, lényegében annak a folytonos változata, ahol az ismeretlen paraméter nem csak egész értékeket vehet fel).

Vegyük észre, hogy egy megfigyelés várható értéke $\vartheta/2$, így az átlag várható értéke is $\vartheta/2$, az átlag kétszeresének várható értéke éppen ϑ . Ebből kiindulva tekintsük az átlag kétszeresét, mint becslést ϑ -ra:

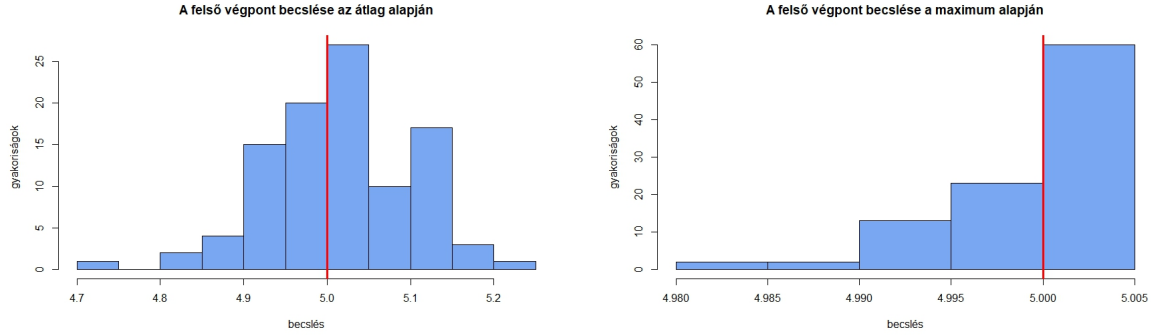
$$\begin{aligned} T_1(X_1, \dots, X_n) &= 2\bar{X}; \\ \mathbb{E}_\vartheta(T_1) &= 2\mathbb{E}_\vartheta(\bar{X}) = 2 \cdot \mathbb{E}_\vartheta(X_1) = 2 \cdot \frac{\vartheta}{2} = \vartheta; \\ D_\vartheta(T_1) &= 2D_\vartheta(\bar{X}) = \frac{2}{\sqrt{n}}D_\vartheta(X_1) = \frac{\vartheta}{\sqrt{3n}},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

felhasználva az egyenletes eloszlásról, illetve az átlag várható értékéről és szórásáról valószínűségszámításból tanultakat.

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

$$\begin{aligned} T_2(X_1, \dots, X_n) &= \frac{n+1}{n} \cdot X_n^*; & \mathbb{E}_\vartheta(T_2) &= \frac{n+1}{n} \cdot \frac{n\vartheta}{n+1} = \vartheta. \\ D_\vartheta(T_2) &= \sqrt{\frac{n \cdot \vartheta^2}{(n+2)(n+1)^2}} \leq \frac{\vartheta}{n+1}. \end{aligned}$$

A két becslés összehasonlítása látható a 2. ábrán. Itt $n = 1000$ elemű mintát használtunk, száz alkalommal kisorsolva, és elvégezve a becslést. Az ábrák a száz becslés hisztogramját mutatják,



2. ábra. A $[0, \vartheta]$ intervallumon egyenletes eloszlás paraméterének becslése $2\bar{X}$, illetve $(n+1)X_n^*/n$ alapján

a bal oldalon a $2\bar{X}$, a jobb oldalon az $(n+1)X_n^*/n$ becsléssel. Az igazi paraméter $\vartheta = 5$. Az első esetben a száz becslés átlaga: **5,015**, korrigált tapasztalati szórása: **0,086**. A második esetben a száz becslés átlaga: **4,9999**, korrigált tapasztalati szórása: **0,0049** < **0,086**.

Tehát azt látjuk, hogy bár várható érték szempontjából a két becslés hasonló, a második esetben a szórás lényegesen kisebb. Ezt az elméleti számítások is alátámasztják, a korábbiak alapján:

$$\mathbb{E}_\vartheta(T_1) = \mathbb{E}_\vartheta(T_2) = \vartheta$$

teljesül minden lehetséges $\vartheta \in \Theta$ -ra, azaz mindkét becslés **torzítatlan becslés** ϑ -ra. Ugyankor a második, a legnagyobb mintaelemet használó becslés szórása kisebb, ez **hatásosabb** a másiknál:

$$D_\vartheta(T_1) = \frac{\vartheta}{\sqrt{3n}} > \frac{\vartheta}{n+1} > D_\vartheta(T_2)$$

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

2.1. Torzítatlanság

A fenti példa alapján egy becslésre az alábbi tulajdonságokat vizsgálhatjuk. A g függvényre azért lehet szükség, mert nem mindig magát a paramétert közvetlenül szeretnénk becsülni. Például lehet, hogy a Poisson-eloszlás paramétere λ , mi azonban $\sqrt{\lambda}$ -t, vagyis az eloszlás szórását szeretnénk torzítatlanul megbecsülni, ekkor g lehet a gyökvonás. Vagy normális eloszlásnál lehet, hogy paraméternek a σ szórást tekintjük ismeretlen paraméternek, de a σ^2 szórásnégyzetet szeretnénk torzítatlanul becsülni, ekkor g a négyzetre emelés.

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

2.1. Definíció (Torzítatlanság). A T statisztika torzítatlan becslés g -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: $\mathbb{E}(2\bar{X}) = \vartheta$.

2.2. Az átlag és a szórásnégyzet torzítatlan becslése

2.1. 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 becslése a várható értéknek.

Ez az alábbi, valószínűségiszámításból ismert állításnak a következménye.

2.2. Állítás. Legyen X_1, \dots, X_n 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 $\mathbb{E}(Y) = \mathbb{E}(Z)$

□

Ebből következik, hogy a **mintaátlag** torzítatlan becslés a várható értékre.

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

A szórásra teljes általánosságban nem találhatunk torzítatlan becslést, a szórásnégyzetre azonban igen (ehhez emlékeztetőül: általában $\mathbb{E}(T)^2 \neq \mathbb{E}(T^2)$, ezért nem igaz, hogy ha T^2 torzítatlan a szórásnégyzetre, akkor T torzítatlan a szórásra, nem elég gyököt vonni).

2.3. Állítás (A szórásnégyzet torzítatlan becslése). X_1, \dots, X_n független azonos eloszlású véges szórású minta. 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 szórásnégyzetre.

Ennek bizonyításához idézzük fel az alábbi állítást.

2.4. Állítás. Legyen X_1, \dots, X_n független azonos eloszlású minta, és $D^2(X_i) < \infty$ létezik. 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 $D(Y) = D(Z)$.

□

Szintén segítségünkre lesz, ha a tapasztalati szórásnégyzetet nem definíció alapján, hanem egy másik alakban írjuk fel.

2.5. 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

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

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.

Most már kiszámíthatjuk a korigált tapasztalati szórásnégyzet várható értékét.

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

Ennek várható értékére vagyunk kíváncsiak.

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 szintén a szórásnégyzet definíciója, valamint az átlag várható értéke (2.2. állítás) és szórása (2.4. állítás) 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:

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

3. Hatásosság

Ahogy a bevezető példában is láttuk, két, a várható érték szempontjából egyformán jó becslés szórása (bizonytalansága) között lényeges különbség is lehet. Sőt, ha csak a várható értéket vennénk figyelembe, egy mintaelem, X_1 , ugyanolyan jó becslés lenne, mint 1000 mintaelem átlaga, pedig ez utóbbi sokkal informatívabb. A várható érték szempontjából egyformán jó becsléseket a szórás alapján hasonlíthatjuk össze.

3.1. Definíció (Hatásosság). Legyenek T_1, T_2 **torzítatlan** becslései a paraméter $g(\vartheta)$ 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** $g(\vartheta)$ -ra, ha $g(\vartheta)$ 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.

4. Konzisztencia

Az eddigiekben csak azt vizsgáltuk, hogy rögzített mintaelemszám esetén milyen tulajdonságai lehetnek egy becslésnek. Ahogy azonban például a Glivenko–Cantelli-tételnél, a statisztika alaptételénél is láttuk, az is fontos kérdés, hogy hogyan viselkedik egy becslésekből álló sorozat, ha a mintaelemszám végtelenhez tart. Ehhez a 3. ábrán látunk egy példát: ha X_1, X_2, \dots független, exponenciális eloszlású valószínűségi változókból álló minta, akkor $1/\bar{X}$, azaz az átlag reciprokának sorozata paraméterhez tart, legalábbis a két vizsgált paraméter esetén. Ha ez minden paraméterértékre fennáll, vagyis a becslések sorozata tart a valódi, becsülni kívánt paraméterhez, azt mondjuk, hogy a becslés konzisztens.

A példában ez teljesül, hiszen a nagy számok erős törvénye szerint \bar{X} a várható értékhez tart 1 valószínűséggel, ami exponenciális eloszlás esetén $1/\lambda$. Ebből következik, hogy $1/\bar{X} \rightarrow \lambda$ teljesül 1 valószínűséggel $n \rightarrow \infty$ esetén, és így sztochasztikusan is. Vagyis a paraméter reciproka konzisztens becslése λ -nak.

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

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

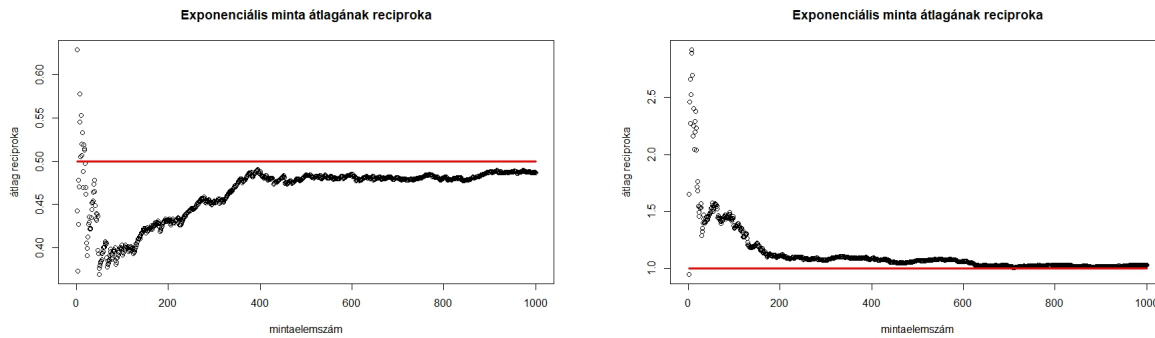
$n \rightarrow \infty$ esetén sztochasztikusan, azaz minden $\vartheta \in \Theta$ és $\varepsilon > 0$ -ra teljesül, hogy

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

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.



3. ábra. $\lambda = 0,5$ paraméterű exponenciális eloszlást generálva a mintaátlag reciproka $0,5$ -höz tart (bal oldali ábra), $\lambda = 1$ paraméter esetén ugyanez a mennyiség 1 -hez tart (jobb oldali ábra)

4.1. 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**, és torzítatlan is.

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűségekre**, és torzítatlan is.

Az s_n és s_n^* közül mindkettő konzisztens becslés a szórásra, a négyzeteik pedig konzisztens becslései a szórásnégyzetnek. Azonban torzítatlanság szempontjából csak azt állíthatjuk általánosan, hogy s_n^{*2} torzítatlan becslése a szórásnégyzetnek.

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 március 3., szerda, 9:00-ig

A Poisson-eloszlás λ paraméterét $T(X) = a\bar{X} + bs_n^{*2}$ alakú statisztikával szeretnénk becsülni.

(a) Milyen (a, b) számpárokra kapunk torzítatlan becslést?

(b) Válasszunk legalább tíz különböző a értéket (például $a = 0, 0, 1, \dots, 1$). Sorsoljunk száz darab, egymástól független, $n = 1000$ elemű Poisson-eloszlású mintát, és mind a száz minta esetén számítsuk ki a kapott $T(X)$ statisztikát. Ábrázoljuk az így kapott száz darab $T(X)$ átlagát, illetve korrigált tapasztalati szórását (lehet külön ábrán) a függvényében (a becslés nem csak a mintától, hanem a -tól is függ, minden a -ra annyi értéket kapunk, ahány minta volt, ezeknek az átlaga, illetve korrigált tapasztalati szórása kell). Milyen következtetést vonhatunk le ebből, a kipróbált a értékek közül melyik adja a „legjobb” becslést?