

Matematikai statisztika (2. előadás)

Statisztikai minta: (X_1, X_2, \dots, X_n) valószínűségi változók (azaz: valószínűségi vektorváltozó).

Mintaelemszám: n

A minta **független**, ha az (X_1, X_2, \dots, X_n) valószínűségi változók függetlenek (például ha a megkérdezetteket függetlenül választottuk, vagy ha a mérések nem befolyásolják egymást), azaz

$$\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \dots, X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \cdot \dots \cdot \mathbb{P}(X_n \leq t_n)$$

teljesül tetszőleges t_1, t_2, \dots, t_n valós számok esetén.

Az (X_1, X_2, \dots, X_n) valószínűségi változók azonos eloszlásúak, de **eloszlásuk nem ismert**: nem tudjuk, hogy mennyi $\mathbb{P}(X_1 \leq t)$, mennyi X_1 várható értéke, szórása, vagy hogy két mennyiség között milyen erős a korreláció. A cél az adatok alapján

- a valószínűségi változók eloszlásának minél jobb megismerése
- a várható érték, szórás stb. becslése
- az eloszlásra vonatkozó hipotézisek eldöntése
- több valószínűségi változó együttes viselkedésének leírása

Középértékek: medián

Minta: (X_1, X_2, \dots, X_n) , mintaelemszám: n .

Definíció (medián)

Ha n páratlan: a rendezett minta középső, $(n+1)/2$. elemét, azaz $X_{(n+1)/2}^*$ -t a minta **mediánjának** nevezzük.

Ha n páros: a rendezett minta $n/2$. és $n/2 + 1$. elemének átlagát, azaz a

$$\frac{X_{n/2}^* + X_{n/2+1}^*}{2}$$

mennyiséget a minta mediánjának nevezzük.

Megjegyzés: páros n esetén a teljes $[X_{n/2}^*, X_{n/2+1}^*]$ intervallumot (vagy annak bármely elemét) is a minta mediánjának lehet hívni.

Példa: a Duna vízállásáról kapott húszelemű minta mediánja:

$$\frac{1}{2}(X_{10}^* + X_{11}^*) = \frac{1}{2}(135 + 148) = 141,5.$$

Középértékek: az átlag és a medián összehasonlítása

Normális eloszlás

500 elemű független minta: X_1, X_2, \dots, X_{500} függetlenek, eloszlásuk normális eloszlás $m = 1$ várható értékkel és $\sigma = 1$ szórással

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.9840	0.2847	0.9842	0.9863	1.6930	3.6110

Exponenciális eloszlás

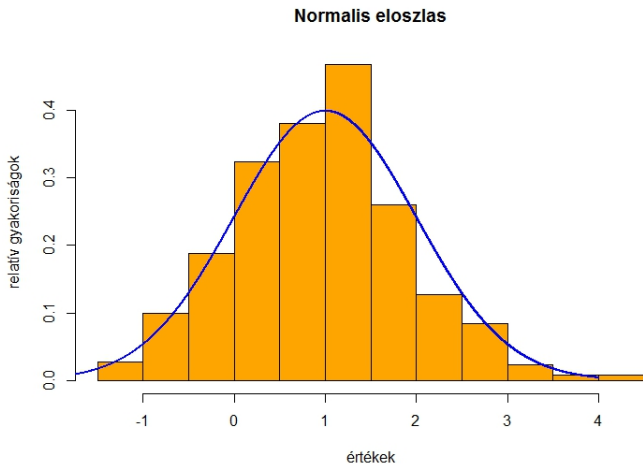
500 elemű független minta: Y_1, Y_2, \dots, Y_{500} függetlenek, eloszlásuk exponenciális eloszlás $\lambda = 1$ paraméterrel. $\mathbb{E}(Y_k) = 1$ és $D(Y_k) = 1$ minden $k = 1, 2, \dots, 500$ -ra.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001326	0.282700	0.637300	0.984900	1.349000	5.895000

```
exp=rexp(n=500, rate=1)
```

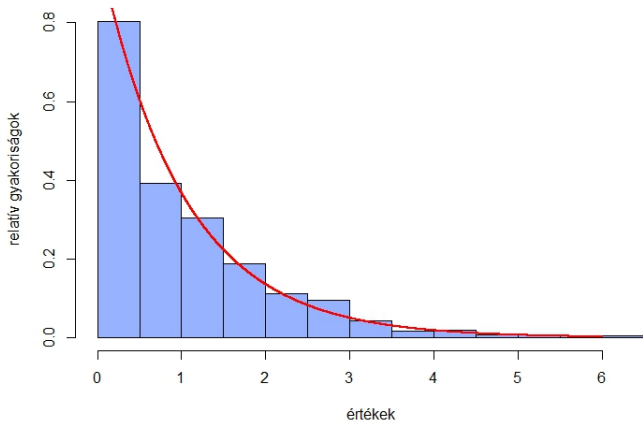
```
summary(exp)
```

A normális eloszlású minta hisztogramja



Az exponenciális eloszlású minta hisztogramja

Exponenciális eloszlás



Az átlag és a medián összehasonlítása

Az átlag

- "több információt használ"
- érzékenyebb a kiugró adatokra, azaz egy hibás mérés is könnyen megváltoztathatja
- nem szimmetrikus esetben eltérhet a leggyakrabban megfigyelt értékektől

A mediánt is érdemes használni, ha

- vannak kiugró (esetleg hibás) adatok;
- ha az eloszlás nem szimmetrikus, és az átlag és a medián jelentősen különbözik (mint a fenti példában az exponenciális eloszlás esetén).

Középértékek közelítése osztályközös gyakoriságokkal

Tegyük fel, hogy az adatokat nem ismerjük pontosan, csak a hisztogramot, vagyis hogy az egyes osztályokba, intervallumokba hány megfigyelés esik. Legyen x_j a j . osztályközép (az alsó és felső határ átlaga), és f_j a j . osztályba eső megfigyelések száma, továbbá $n = f_1 + f_2 + \dots + f_k$ az összes megfigyelés száma. Ekkor

- az átlag közelítése:

$$\frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{n};$$

Középértékek közelítése osztályközös gyakoriságokkal

Tegyük fel, hogy az adatokat nem ismerjük pontosan, csak a hisztogramot, vagyis hogy az egyes osztályokba, intervallumokba hány megfigyelés esik. Legyen x_j a j . osztályközép (az alsó és felső határ átlaga), és f_j a j . osztályba eső megfigyelések száma, továbbá $n = f_1 + f_2 + \dots + f_k$ az összes megfigyelés száma. Ekkor

- az átlag közelítése:

$$\frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{n};$$

- a medián közelítése:

$$t_{\text{me}} + \frac{n/2 - F_{\text{me}-1}}{f_{\text{me}}} \cdot h_{\text{me}},$$

ahol t_{me} a mediánt tartalmazó osztály alsó határa, $F_{\text{me}-1}$ a mediánt tartalmazó osztályt megelőző osztályok gyakoriságainak összege, f_{me} a mediánt tartalmazó osztály gyakorisága, h_{me} a mediánt tartalmazó osztály szélessége.

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

Statisztikai minta: (X_1, X_2, \dots, X_n) valószínűségi változók (azaz: valószínűségi vektorváltozó).

Mintaelemszám: n

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

Statisztikai minta: (X_1, X_2, \dots, X_n) valószínűségi változók (azaz: valószínűségi vektorváltozó).

Mintaelemszám: n

A minta **független**, ha az (X_1, X_2, \dots, X_n) valószínűségi változók függetlenek (például ha a megkérdezetteket függetlenül választottuk, vagy ha a mérések nem befolyásolják egymást), azaz

$$\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \dots, X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \cdot \dots \cdot \mathbb{P}(X_n \leq t_n)$$

teljesül tetszőleges t_1, t_2, \dots, t_n valós számok esetén.

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

Statisztikai minta: (X_1, X_2, \dots, X_n) valószínűségi változók (azaz: valószínűségi vektorváltozó).

Mintaelemszám: n

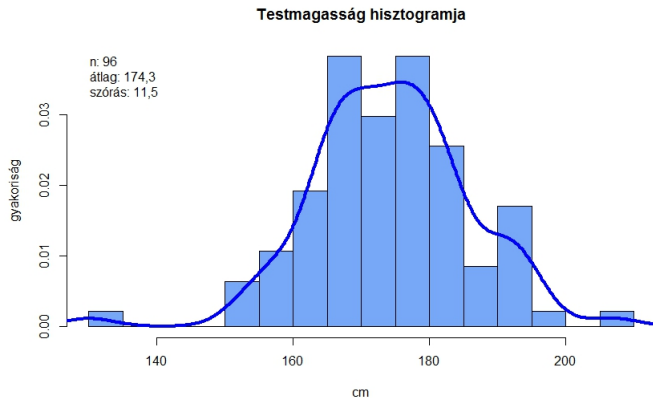
A minta **független**, ha az (X_1, X_2, \dots, X_n) valószínűségi változók függetlenek (például ha a megkérdezetteket függetlenül választottuk, vagy ha a mérések nem befolyásolják egymást), azaz

$$\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \dots, X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \cdot \dots \cdot \mathbb{P}(X_n \leq t_n)$$

teljesül tetszőleges t_1, t_2, \dots, t_n valós számok esetén.

Az (X_1, X_2, \dots, X_n) valószínűségi változók **eloszlása nem ismert**: nem tudjuk, hogy mennyi $\mathbb{P}(X_1 \leq t)$, vagy mennyi X_1 várható értéke, szórása. A cél a valószínűségi változók eloszlásának a becslése, rá vonatkozó hipotézisek eldöntése a megfigyelések, vagyis az adatok alapján.

Hisztogram és sűrűségfüggvény becslése



A testmagasság hisztogramja $n = 96$ elemű mintából (valós adatokból), a sűrűségfüggvény becslése Gauss-magfüggvénnyel.

A sűrűségfüggvény becslése

X_1, X_2, \dots, X_n független azonos eloszlású abszolút folytonos minta. A sűrűségfüggvény f , azaz

$$\mathbb{P}(a \leq X_1 \leq b) = \int_a^b f(t) dt \quad \text{minden } a < b\text{-re.}$$

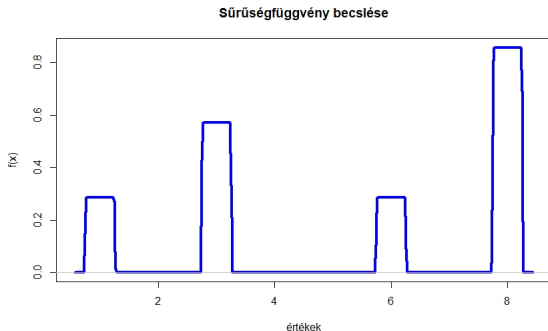
Az f függvény ismeretlen. Hogyan tudjuk $f(t)$ értékét becsülni az X_1, \dots, X_n megfigyelések segítségével?

Hisztogram:

$$\mathbb{P}(a \leq X_1 \leq b) = \int_a^b f(t) dt \approx \frac{1}{n} \sum_{j=1}^n \mathbb{I}(a < X_j \leq b),$$

azaz a becslés a és b közé eső mintaelemek aránya.

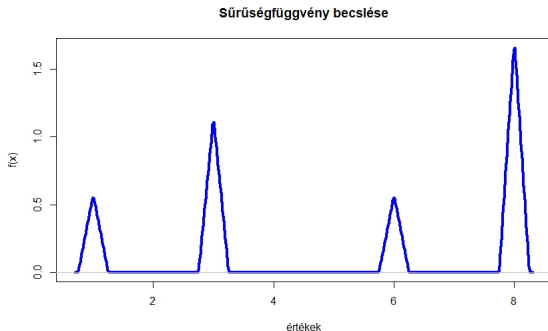
A sűrűségfüggvény becslése téglalapos magfüggvénnyel



Minta (X_1, \dots, X_7) : 1, 3, 3, 6, 8, 8, 8. **Téglalap magfüggvény**: $k(y) = 1/2$, ha $-1 \leq y \leq 1$, nulla különben, azaz $k(y) = \frac{1}{2}\mathbb{I}(|y| \leq 1)$ és h az **ablakszélesség**.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right) = \frac{1}{nh} \sum_{j=1}^n \frac{1}{2} \mathbb{I}(|t - X_j| < h).$$

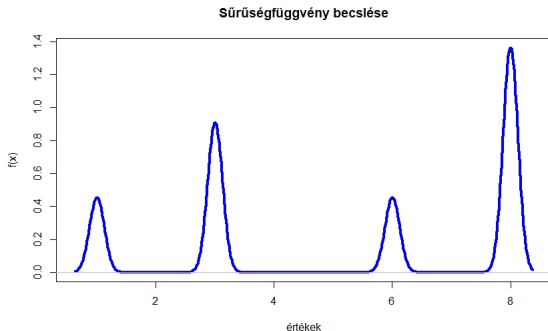
A sűrűségfüggvény becslése háromszöges magfüggvénnyel



Minta (X_1, \dots, X_7) : 1, 3, 3, 6, 8, 8, 8. **Háromszöges magfüggvény:** $k(y) = \max(1 - |y|, 0)$ és $h = 1/2$ az ablakszélesség.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right).$$

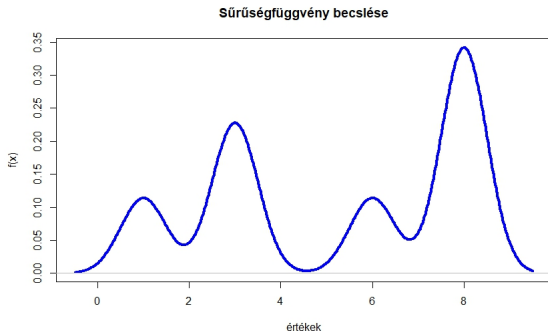
A sűrűségfüggvény becslése Gauss-magfüggvénnyel



Minta (X): 1, 3, 3, 6, 8, 8, 8. **Gauss-magfüggvény:** $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ és $h = 1/2$ az ablakszélesség.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right) = \frac{1}{n \cdot h \cdot \sqrt{2\pi}} \sum_{j=1}^n \exp\left(-\frac{(y - X_j)^2}{2h^2}\right).$$

A sűrűségfüggvény becslése Gauss-magfüggvénnyel



Minta (X): 1, 3, 3, 6, 8, 8, 8. Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ és $h = 2$ **az ablakszélesség.**

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right) = \frac{1}{n \cdot h \cdot \sqrt{2\pi}} \sum_{j=1}^n \exp\left(-\frac{(y - X_j)^2}{2 \cdot 2^2}\right).$$

Parzen–Rosenblatt-becslés

Legyen $k : \mathbb{R} \rightarrow \mathbb{R}_+$ olyan függvény, mely korlátos, $\lim_{y \rightarrow \infty} yk(y) = 0$, továbbá h_n olyan számsorozat, melyre $\lim_{n \rightarrow \infty} h_n = 0$ és $\lim_{n \rightarrow \infty} nh_n = \infty$. A sűrűségfüggvény becslése a t pontban a Parzen–Rosenblatt-módszerrel a k magfüggvénnyel és h_n sávszélességgel az X_1, \dots, X_n független minta alapján:

$$\hat{f}_n(t) = \frac{1}{n \cdot h_n} \sum_{j=1}^n k\left(\frac{t - X_j}{h_n}\right).$$

Parzen–Rosenblatt-becslés

Legyen $k : \mathbb{R} \rightarrow \mathbb{R}_+$ olyan függvény, mely korlátos, $\lim_{y \rightarrow \infty} yk(y) = 0$, továbbá h_n olyan számsorozat, melyre $\lim_{n \rightarrow \infty} h_n = 0$ és $\lim_{n \rightarrow \infty} nh_n = \infty$. A sűrűségfüggvény becslése a t pontban a Parzen–Rosenblatt-módszerrel a k magfüggvénnyel és h_n sávszélességgel az X_1, \dots, X_n független minta alapján:

$$\hat{f}_n(t) = \frac{1}{n \cdot h_n} \sum_{j=1}^n k\left(\frac{t - X_j}{h_n}\right).$$

Megfelelő feltételek mellett $\hat{f}_n(t) \rightarrow f(t)$ minden t -re, ha $n \rightarrow \infty$. Szokásos magfüggvények például:

- Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$.
- Háromszög magfüggvény: $k(y) = (1 - |y|)$, ha ez nemnegatív, nulla különben.
- Epanechnikov-magfüggvény: $k(y) = \frac{3}{4}(1 - y^2)$, ha ez nemnegatív, nulla különben.
- Téglalap magfüggvény: $k(y) = 1/2$, ha $-1 \leq y \leq 1$, nulla különben.

Parzen–Rosenblatt-becslés

A sűrűségfüggvény becslése a t pontban a Parzen–Rosenblatt-módszerrel a k magfüggvénnyel és h_n sáv szélességgel az X_1, \dots, X_n független minta alapján:

$$\hat{f}_n(t) = \frac{1}{n \cdot h_n} \sum_{j=1}^n k\left(\frac{t - X_j}{h_n}\right).$$

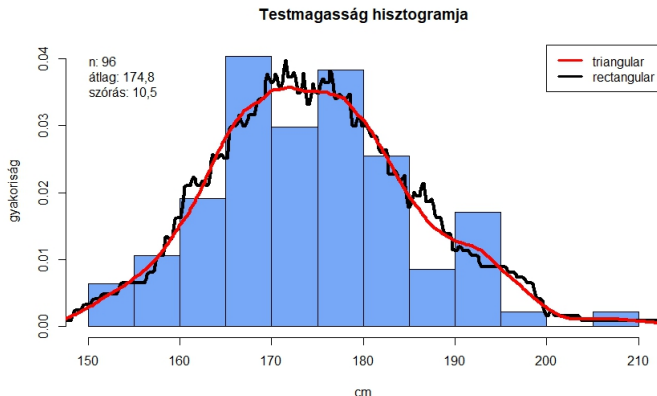
Szokásos sáv szélesség-választások (normális eloszlás és Gauss-magfüggvény esetén az első optimális), ezekre $h_n \rightarrow 0$, de $nh_n \rightarrow \infty$:

$$h_n = 0,7 \cdot \frac{s_n^*}{n^{1/5}}; \quad h_n = 0,7 \cdot \frac{\min(s_n^*, q)}{n^{1/5}},$$

ahol s_n^* a korrigált tapasztalati szórás, q a harmadik és első kvartilis távolsága.

Ugyanúgy, mint a hisztogramnál, a túl nagy sáv szélesség túl kevésbé részletes ábrához, a túl kicsi sáv szélesség túl részletes ábrához vezet.

Hisztogram és sűrűségfüggvény becslése

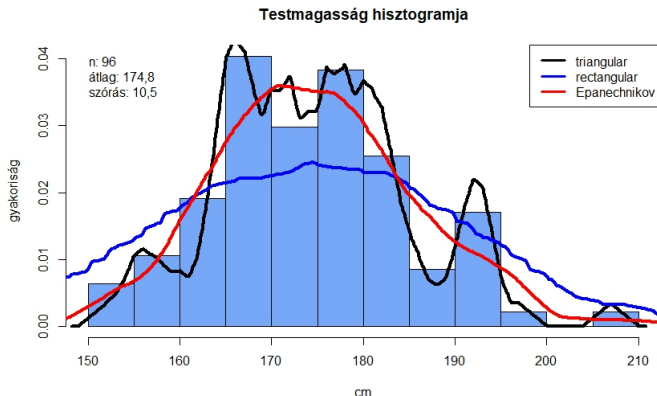


A testmagasság hisztogramja $n = 96$ elemű mintából (valós adatokból), háromszöges (piros) és téglalapos (fekete) magfüggvénnyel (ez utóbbihoz túl kicsi a sávszélesség).

Testmagasság sűrűségfüggvényének becslése

```
> testmagassag <- c(186, 180, 197, 191, 181, 178, 193, 177, 167, 163, 164, 170, 178)
> hist(testmagassag, col="#76a8f7", xlab="cm", ylab="gyakoriság",
main="Testmagasság hisztogramja", breaks=12, prob=TRUE)
> legend("topleft", c("n: 96", "átlag: 174,8", "szórás: 10,5"),
bty="n")
> lines(density(testmagassag, kernel="triangular", adjust=1/3),
lwd="4", col="black")
> lines(density(testmagassag, kernel="rectangular", adjust=3),
lwd="4", col="blue")
> lines(density(testmagassag, kernel="epanechnikov"),
lwd="4", col="red")
> legend("topright", c("triangular", "rectangular", "Epanechnikov"),
col=c("black", "blue", "red"), lwd="3", bty="o")
```

Testmagasság sűrűségfüggvényének beclése



Testmagasság sűrűségfüggvényének beclése: háromszöges magfüggvény 1/3-szoros sávszélességgel (fekete), téglalapos magfüggvény 3-szoros sávszélességgel (kék), Epanechnikov-magfüggvény alapértelmezett sávszélességgel (piros)

Indexek számítása

Egy időben változó mennyiség egy időszakban (tárgyidőszakban) mért értékeit szeretnénk egy korábbi, hasonló időszakban (bázisidőszakban) mért értékekkel összehasonlítani, hogy az átlagos változást leírhassuk. Például tekinthetjük a fogyasztói árindexet (például: https://www.ksh.hu/stadat_files/ara/hu/ara0040.html, vagy http://www.ksh.hu/interaktiv/fogyar_radar/index.html).

Tegyük fel, hogy az árindexbe az $1, 2, \dots, n$ termékek forgalmát építik be. Legyen

- $q_{0,j}$ a j . termékből eladott mennyiség a bázisidőszakban;
- $q_{1,j}$ a j . termékből eladott mennyiség a tárgyidőszakban;
- $p_{0,j}$ a j . termék egységára a bázisidőszakban;
- $p_{1,j}$ a j . termék egységára a tárgyidőszakban.

Árindexek számítása

Tegyük fel, hogy az árindexbe az $1, 2, \dots, n$ termékek forgalmát építik be. Legyen

- $q_{0,j}$ a j . termékből eladott mennyiség a bázisidőszakban;
- $q_{1,j}$ a j . termékből eladott mennyiség a tárgyidőszakban;
- $p_{0,j}$ a j . termék egységára a bázisidőszakban;
- $p_{1,j}$ a j . termék egységára a tárgyidőszakban.

Bázisidőszaki súlyozású vagy Laspeyres-féle árindex: annak hányadosa, hogy az új árakkal, de a bázisidőszak fogyasztásával mennyivel nőtt az összes kiadás a régebbi időszakhoz képest, azaz

$$\frac{\sum_{j=1}^n q_{0,j} p_{1,j}}{\sum_{j=1}^n q_{0,j} p_{0,j}}$$

Tárgyidőszaki súlyozású vagy Paasche-féle árindex: annak hányadosa, hogy az új árakkal és az új fogyasztással mennyivel nőtt az összes kiadás, azaz

$$\frac{\sum_{j=1}^n q_{1,j} p_{1,j}}{\sum_{j=1}^n q_{1,j} p_{0,j}}$$

Volumenindexek számítása

Tegyük fel, hogy az árindexbe az $1, 2, \dots, n$ termékek forgalmát építik be. Legyen

- $q_{0,j}$ a j . termékből eladott mennyiség a bázisidőszakban;
- $q_{1,j}$ a j . termékből eladott mennyiség a tárgyidőszakban;
- $p_{0,j}$ a j . termék egységára a bázisidőszakban;
- $p_{1,j}$ a j . termék egységára a tárgyidőszakban.

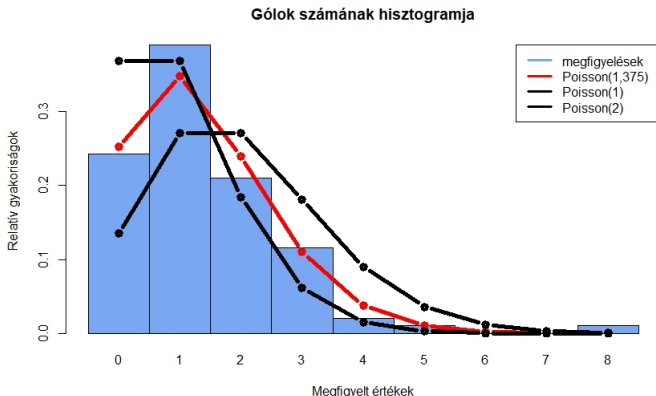
Bázisidőszaki súlyozású vagy Laspeyres-féle volumenindex: a régi árakkal számolva hányszorosára nőtt az összes kiadás, azaz

$$\frac{\sum_{j=1}^n q_{1,j} p_{0,j}}{\sum_{j=1}^n q_{0,j} p_{0,j}}$$

Tárgyidőszaki súlyozású vagy Paasche-féle volumenindex: az új árakkal számolva hányszorosára nőtt az összes kiadás, azaz

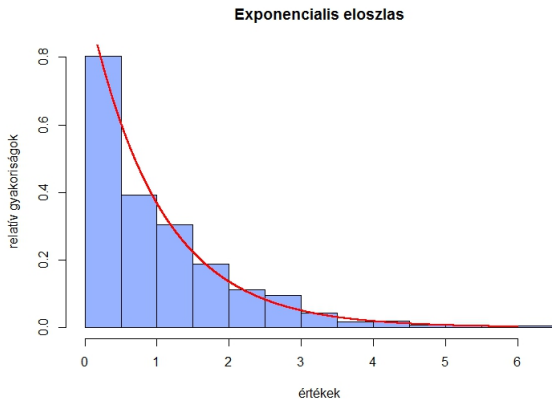
$$\frac{\sum_{j=1}^n q_{1,j} p_{1,j}}{\sum_{j=1}^n q_{0,j} p_{1,j}}$$

Poisson-eloszlás paraméterének becslése



A gólok számának hisztogramja $n = 95$ mérkőzésen, és különböző paraméterű Poisson-eloszlások ($\mathbb{P}_\lambda(X = k) = \lambda^k / k! \cdot e^{-\lambda}$). $\lambda = 1,375$ a gólok átlagos száma.

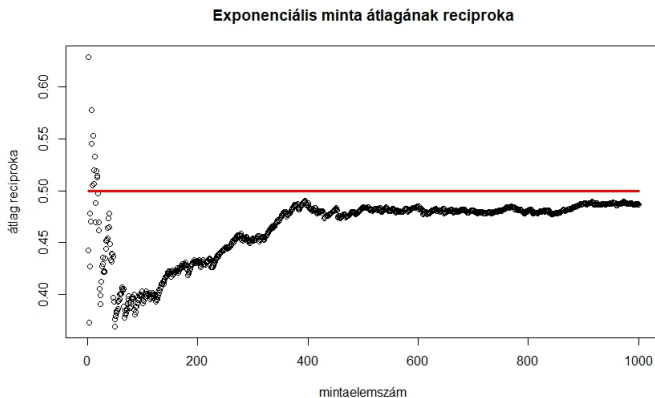
Az exponenciális eloszlású minta hisztogramja



Exponenciális eloszlású minta hisztogramja és sűrűségfüggvénye:

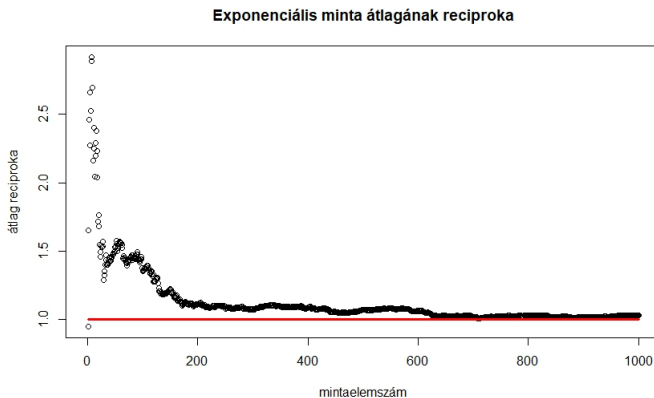
$$f(x) = \lambda \exp(-\lambda x) \mathbb{I}(x > 0); \quad \mathbb{E}(X) = D(X) = \frac{1}{\lambda}.$$

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.

Statisztikai mező

Definíció

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

Paraméteres statisztikai 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).

Minta és statisztika

Definíció (Minta)

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező. Egy

$$\underline{X} = (X_1, X_2, \dots, X_n) : \Omega \rightarrow B \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót (n elemű) **mintának** nevezünk. Itt B a mintatér, n a minta elemszáma vagy nagysága. A minta független, ha az X_1, X_2, \dots, X_n valószínűségi változók függetlenek.

Minta és statisztika

Definíció (Minta)

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező. Egy

$$\underline{X} = (X_1, X_2, \dots, X_n) : \Omega \rightarrow B \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót (n elemű) **mintának** nevezünk. Itt B a mintatér, n a minta elemszáma vagy nagysága. A minta független, ha az X_1, X_2, \dots, X_n valószínűségi változók függetlenek.

Definíció (Statisztika)

Legyen $T : B \rightarrow \mathbb{R}^k$ függvény. Ekkor a $T(X_1, X_2, \dots, X_n)$ valószínűségi változót statisztikának nevezzük.

Például: $T(X_1, \dots, X_n) = \bar{X}$ a mintaátlag, vagy $T(X_1, \dots, X_n) = s_n^*$ a korrigált tapasztalati szórásnégyzet.