

Matematikai statisztika (2. előadás)

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 (2. előadás)

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 (2. előadás)

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

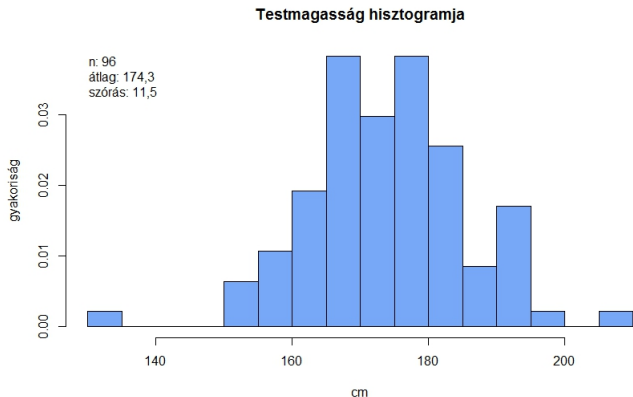
A minta **független**, ha az (X_1, X_2, \dots, X_n) valószínűségi változók függetlenek (például ha a megkérdezetteket függetlenül választottuk, vagy ha a mérések nem befolyásolják egymást), azaz

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

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.

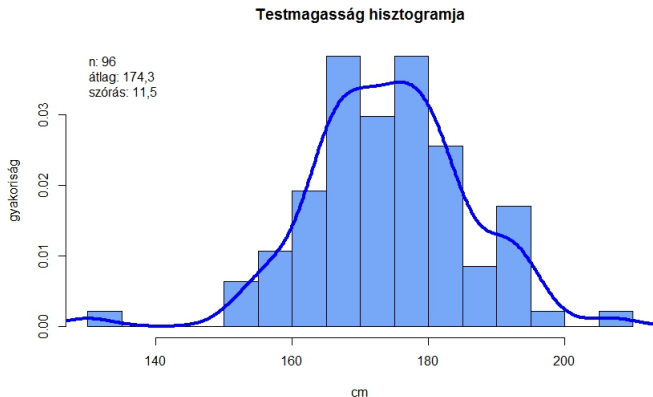
A sűrűségfüggvény becslése



A testmagasság hisztogramja $n = 96$ elemű mintából (valós adatokból).

Hogyan becsülhető **a testmagasság sűrűségfüggvénye**? A hisztogram közelíti a sűrűségfüggvényt, de nem világos, hogy milyen intervallumhosszal érdemes számolni.

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

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

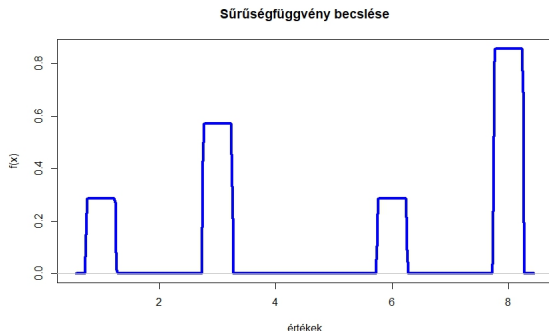
Az f függvény ismeretlen. Hogyan tudjuk $f(t)$ értékét becsülni az X_1, \dots, X_n megfigyelések segítségével?

Hisztogram:

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

azaz a becslés a és b közé eső mintaelemek aránya.

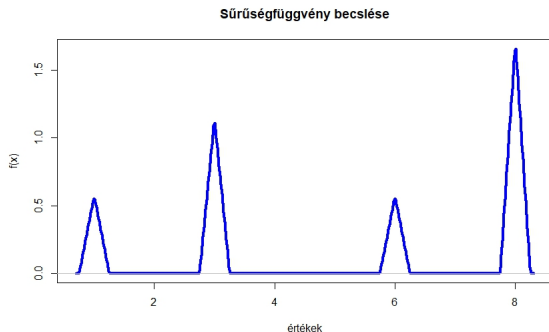
A sűrűségfüggvény becslése téglalapos magfüggvénnyel



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**.

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

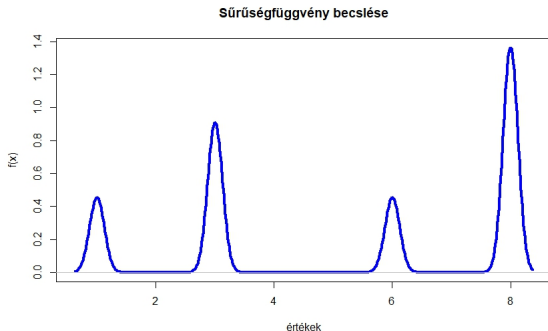
A sűrűségfüggvény becslése háromszöges magfüggvénnyel



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.

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

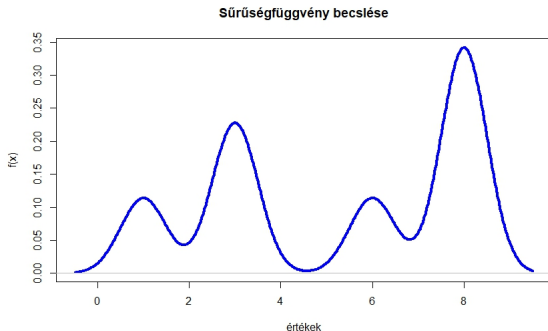
A sűrűségfüggvény becslése Gauss-magfüggvénnyel



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:

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

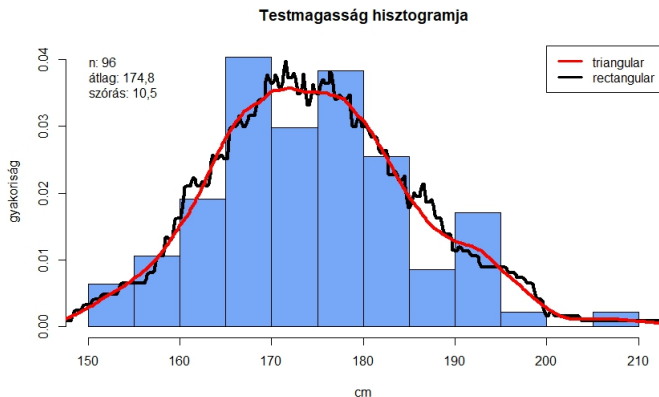
Szokásos sáv szélesség-választások (normális eloszlás és Gauss-magfüggvény esetén az első optimális), ezekre $h_n \rightarrow 0$, de $nh_n \rightarrow \infty$:

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

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

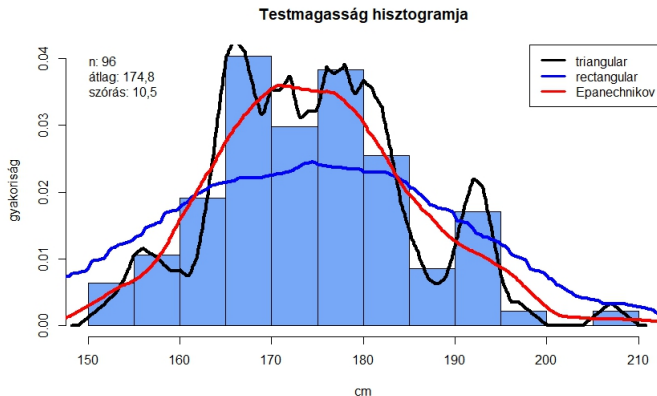
Ugyanúgy, mint a hisztogramnál, a túl nagy sáv szélesség túl kevésbé részletes ábrához, a túl kicsi sáv szélesség túl részletes ábrához vezet.

Hisztogram és sűrűségfüggvény becslése



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

Testmagasság sűrűségfüggvényének becslése



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

Házi feladat február 26., 8:15-ig

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 19-ig: megoldás

Kérdezzünk meg legalább huszonöt felnőtt ismerőst a

- a nemükről;
- arról, hogy hány filmsorozatot (vagy tévésorozatot) néztek rendszeresen az elmúlt egy hónapban (az nem kell, hogy melyik sorozatokat)
- arról, hogy egy hétköznapon átlagosan hány percet töltenek közlekedéssel.

Az adatokra az egész félév során szükség lesz a házi feladatoknál, azzal együtt, hogy melyik válaszok tartoznak össze.

Tekintsük a közlekedéssel töltött időket. Készítsünk (a) hisztogramot; (b) boxplotot; (c) ábrázoljuk a tapasztalati eloszlásfüggvényt külön a férfiak és a nők esetében, illetve összesítve. Milyen következtetéseket vonhatunk le a kapott ábrákról?

Házi feladat február 19-ig: megoldás

```
hist(no_utazas, col="#79a7f2", xlab="Utazással töltött idő (perc)",  
ylab="Relatív gyakoriságok", main="Utazással töltött idő  
a nők esetében")
```

```
hist(ferfi_utazas, col="#79a7f2", xlab="Utazással töltött idő (perc)",  
ylab="Relatív gyakoriságok", main="Utazással töltött idő  
a férfiak esetében")
```

```
utazas<-c(no_utazas, ferfi_utazas)
```

```
hist(utazas, col="#79a7f2", xlab="Utazással töltött idő (perc)",  
ylab="Relatív gyakoriságok", main="Utazással töltött idő  
hisztogramja")
```

Házi feladat február 19-ig: megoldás

```
boxplot(ferfi_utazas, no_utazas, utazas, col="yellow",
names=c("férfiak", "nők", "összesen"),
ylab="utazással töltött idő (perc)", main="Utazással töltött idő
boxplotja")
```

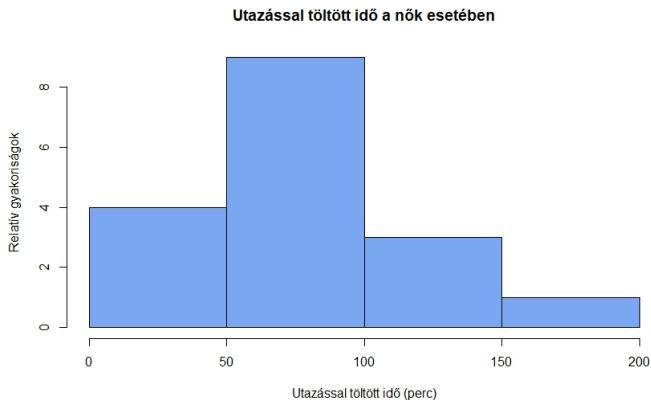
```
plot(ecdf(no_utazas), lwd="5", col="red", main="Utazással töltött
idő tapasztalati eloszlásfüggvénye", xlab="idő (perc)",
ylab="tapasztalati eloszlásfüggvény")
```

```
lines(ecdf(ferfi_utazas), lwd="5", col="blue")
```

```
legend("topleft", c("nők", "férfiak"), col=c("red", "blue"),
lwd="3")
```

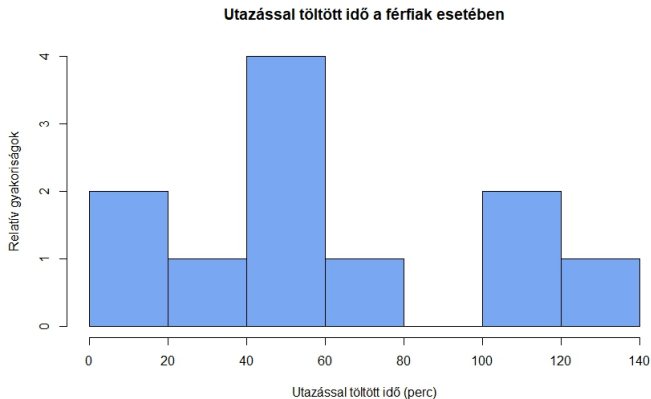
```
plot(ecdf(utazas), lwd="5", col="red", main="Utazással töltött idő
tapasztalati eloszlásfüggvénye",
xlab="idő (perc)", ylab="tapasztalati eloszlásfüggvény")
```

Utazással töltött idő hisztogramja



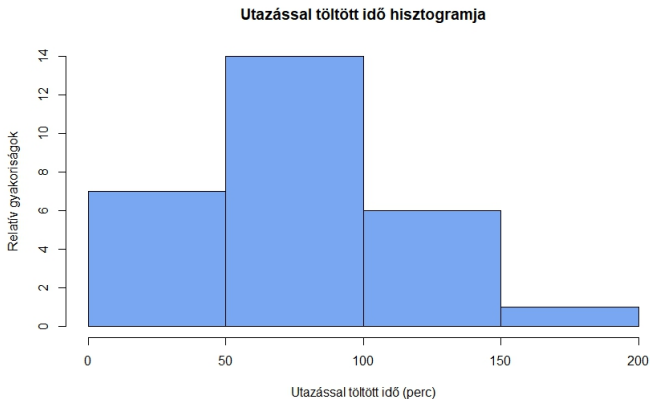
Utazással töltött idő hisztogramja a nők esetében ($n_1 = 17$)

Utazással töltött idő hisztogramja



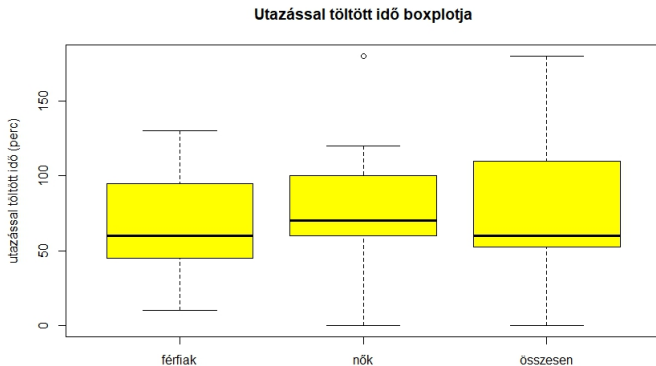
Utazással töltött idő hisztogramja a férfiak esetében ($n_2 = 11$)

Utazással töltött idő hisztogramja



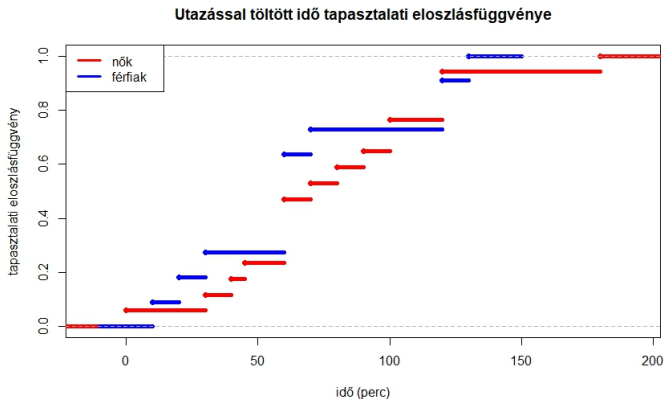
Utazással töltött idő hisztogramja a teljes minta esetében ($n = 28$)

Utazással töltött idő boxplotja



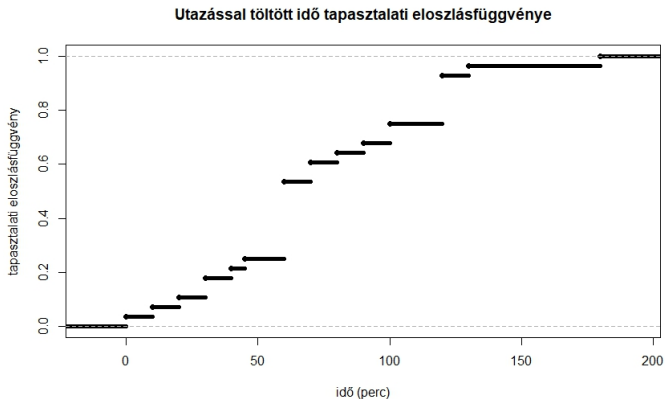
Utazással töltött idő boxplotja ($n_1 = 17$ nő, $n_2 = 11$ férfi, $n = 28$ összes megfigyelés)

Utazással töltött idő tapasztalati eloszlásfüggvénye



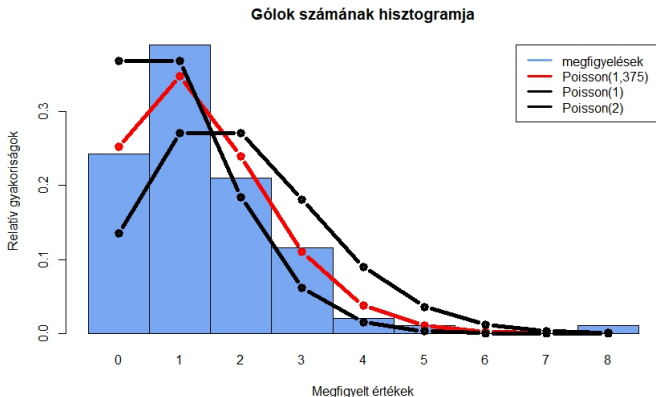
Utazással töltött idő tapasztalati eloszlásfüggvénye ($n_1 = 17$ nő, $n_2 = 11$ férfi)

Utazással töltött idő tapasztalati eloszlásfüggvénye



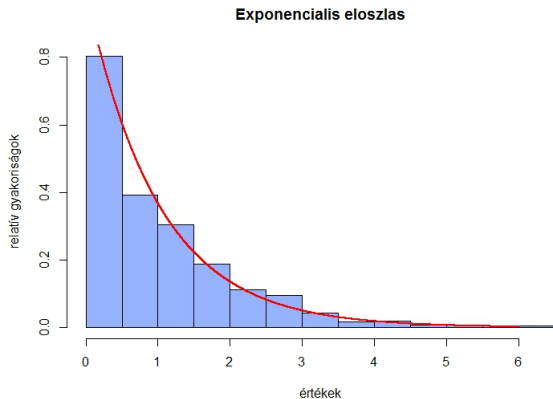
Utazással töltött idő tapasztalati eloszlásfüggvénye ($n = 28$ megkérdezett)

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.

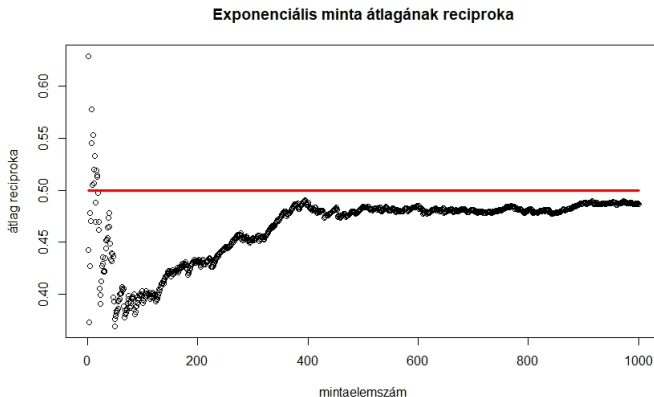
Az exponenciális eloszlású minta hisztogramja



Exponenciális eloszlású minta hisztogramja és sűrűségfüggvénye:

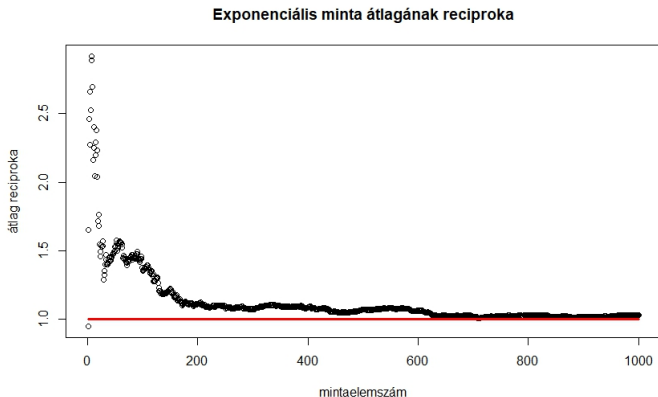
$$f(x) = \lambda \exp(-\lambda x) \mathbb{I}(x > 0); \quad \mathbb{E}(X) = D(X) = \frac{1}{\lambda}.$$

Exponenciális eloszlás



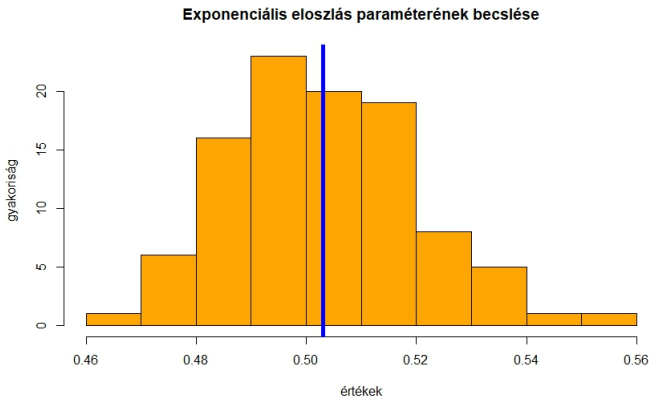
$\lambda = 0,5$ paraméterű exponenciális eloszlást generálva a mintaátlag reciproka 0,5-höz tart, azaz **konzisztens** becslés, hiszen ez minden λ -ra teljesül.

Exponenciális eloszlás



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

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**.

Statisztikai mező

Definíció

Az $(\Omega, \mathcal{A}, \mathcal{P})$ hármast **statisztikai mezőnek** nevezük, 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).

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.

Becslések és tulajdonságaik

- $(\Omega, \mathcal{A}, \mathcal{P})$ 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.

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.