

Becslések és tulajdonságaik (3. előadás)

- $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamely Θ halmazzal (Θ a paraméterter);
- $\psi : \Theta \rightarrow \mathbb{R}$ 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 (3. előadás)

- $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamely Θ halmazzal (Θ a paraméterter);
- $\psi : \Theta \rightarrow \mathbb{R}$ 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

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

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

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

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.

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.

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

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 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:

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

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

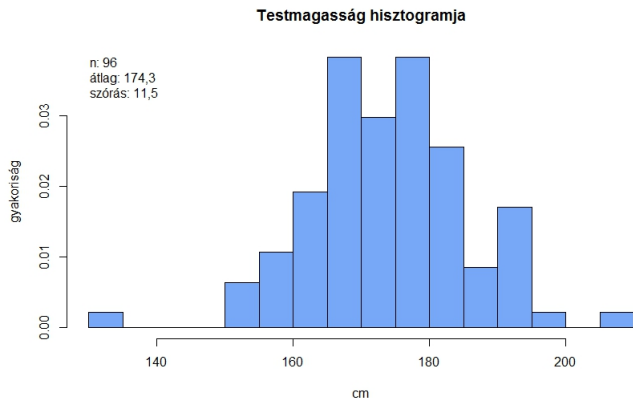
$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

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 másikonál.
- 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.

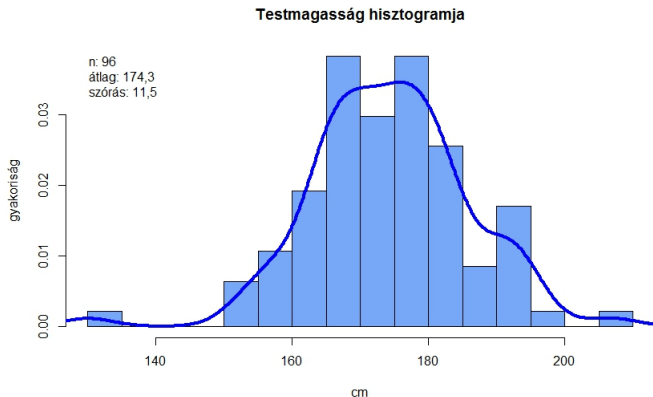
A sűrűségfüggvény becslése



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?

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:

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

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:

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

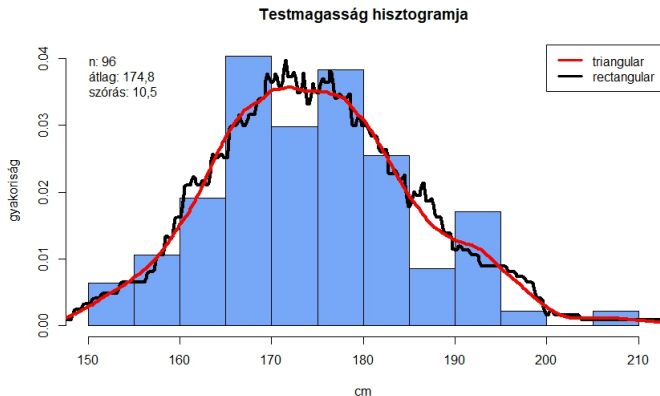
Szokásos sáv szélesség-választások (normális eloszlás és Gauss-magfüggvény esetén az első optimális):

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

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).

Házi feladat március 6., 9:00-ig

Tekintsük a testmagasságokról kapott húszelemű adatsort, ábrázoljuk a hisztogramot (úgy, hogy az összterület egy legyen), és illesszük rá a sűrűségfüggvény becslését

- háromszöges magfüggvénnyel úgy, hogy a sávszélesség az alapértelmezett harmada;
- téglalapos úgy, hogy a sávszélesség az alapértelmezettnek a háromszorosa;
- Epanechnikov-magfüggvénnyel úgy, hogy a sávszélesség az alapértelmezett

(Alapértelmezett: ez tetszőleges beállítás szerinti alapértelmezett lehet, viszont a három megoldásnál ugyanaz legyen.)

Házi feladat február 27., 9:00-ig

Generáljunk $n = 100000$ elemű exponenciális eloszlású mintákat $\lambda = 1/2$, $\lambda = 1$ és $\lambda = 5$ paraméterekkel.

- Mindhárom különböző paraméterre ábrázoljuk k függvényében (de csak a $k = 1, 2, \dots, 1000$ értékekre) az első k mintaelem átlagának reciprokát, és ezt hasonlítsuk össze a mintához tartozó paraméterrel.
- Csoportosítsuk a mintaelemeket ezresével, és legyen Y_i az i . csoport (vagyis az $(i-1) \cdot 1000 + 1, (i-1) \cdot 1000 + 2, \dots, i \cdot 1000$ indexű mintaelemek) átlagának reciproka. Számítsuk ki az Y_1, Y_2, \dots, Y_{100} átlagát, és ezt hasonlítsuk össze a λ paraméterrel.
- Értelmezzük az eredményeket: mennyire látható, hogy exponenciális eloszlásnál az átlag reciproka konzisztens, de nem torzítatlan becslése a paraméternek?

Házi feladat február 27., 9:00-ig, megoldás

Például $\lambda = 0,5$ -re a mintaátlag reciprokának változása:

```
minta1<-rexp(100000, rate=0.5)
```

```
atlagrep1<-1:1000
```

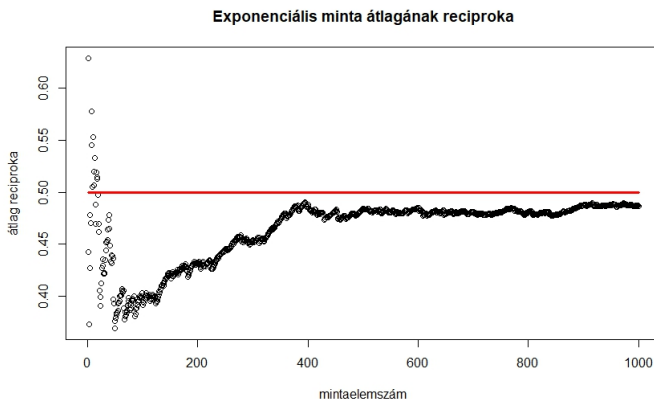
```
for(j in 1:1000)atlagrep1[j]=1/mean(minta1[1:j])
```

```
plot(atlagrep1, main="Exponenciális minta átlagának reciproka", xlab="mintaelemsz",  
ylab="átlag reciproka")
```

```
x=rep(0.5, 1000)
```

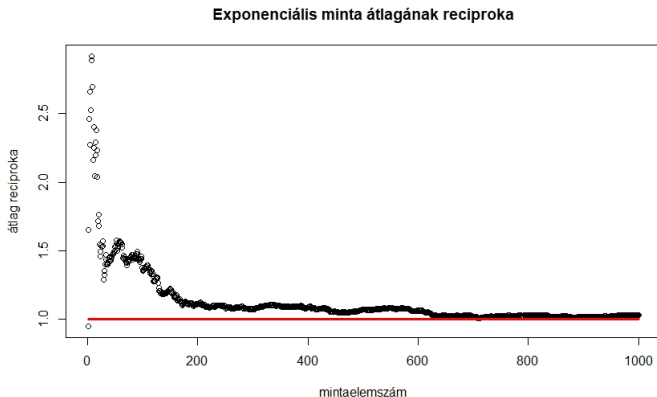
```
lines(x, lwd=2, col="red")
```

Házi feladat február 27-ig, megoldás



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

Házi feladat február 27-ig, megoldás



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

Házi feladat február 27-ig, megoldás

$\lambda = 0,5$ paraméterű exponenciális eloszlásnál:

```
> minta1<-rexp(100000, rate=0.5)
```

```
> atlagrep1<-1:100
```

```
> for(j in 1:100)atlagrep1[j]=1/mean(minta1[((j-1)*1000+1):(j*1000)])
```

```
> mean(atlagrep1)
```

```
[1] 0.5016493
```

Hasonlóképpen $\lambda = 1$ -re:

```
> mean(atlagrep2)
```

```
[1] 1.002733
```

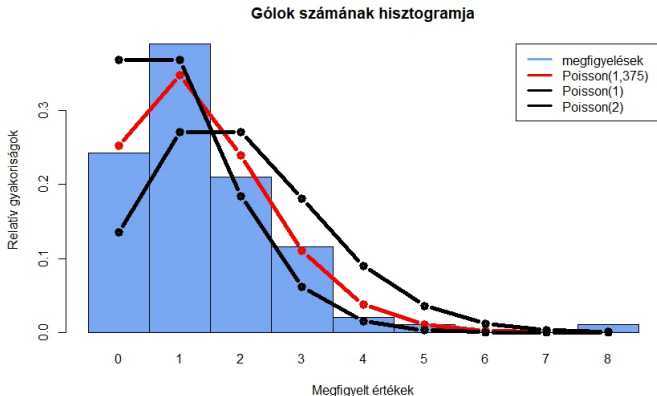
Hasonlóképpen $\lambda = 5$ -re:

```
> mean(atlagrep3)
```

```
[1] 4.981833
```

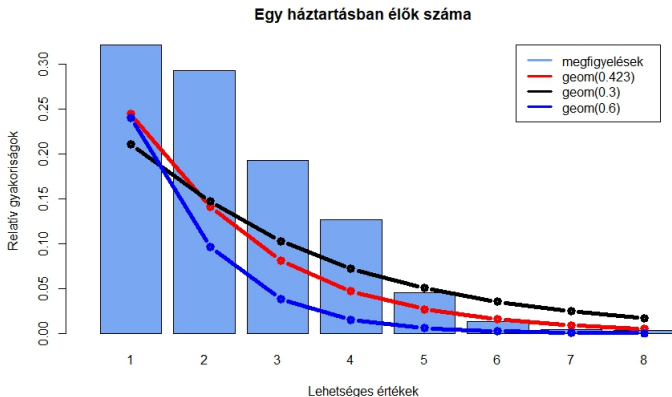
Ugyan az átlag reciproka nem torzítatlan becslés a paraméterre, a megfigyelt értékek nem különböznek nagyon a valódi értékektől.

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, $1/\bar{X} = 0,423$; $\mathbb{P}_p(X = k) = (1 - p)^{k-1} \cdot p$.

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 maximalizálja a $\vartheta \mapsto L_{n,\vartheta}(X_1, \dots, X_n)$ függvényt, ahol $L_{n,\vartheta}$ a minta likelihood-függvénye. Azaz, ha

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