Valószínűségszámítás és Statisztika

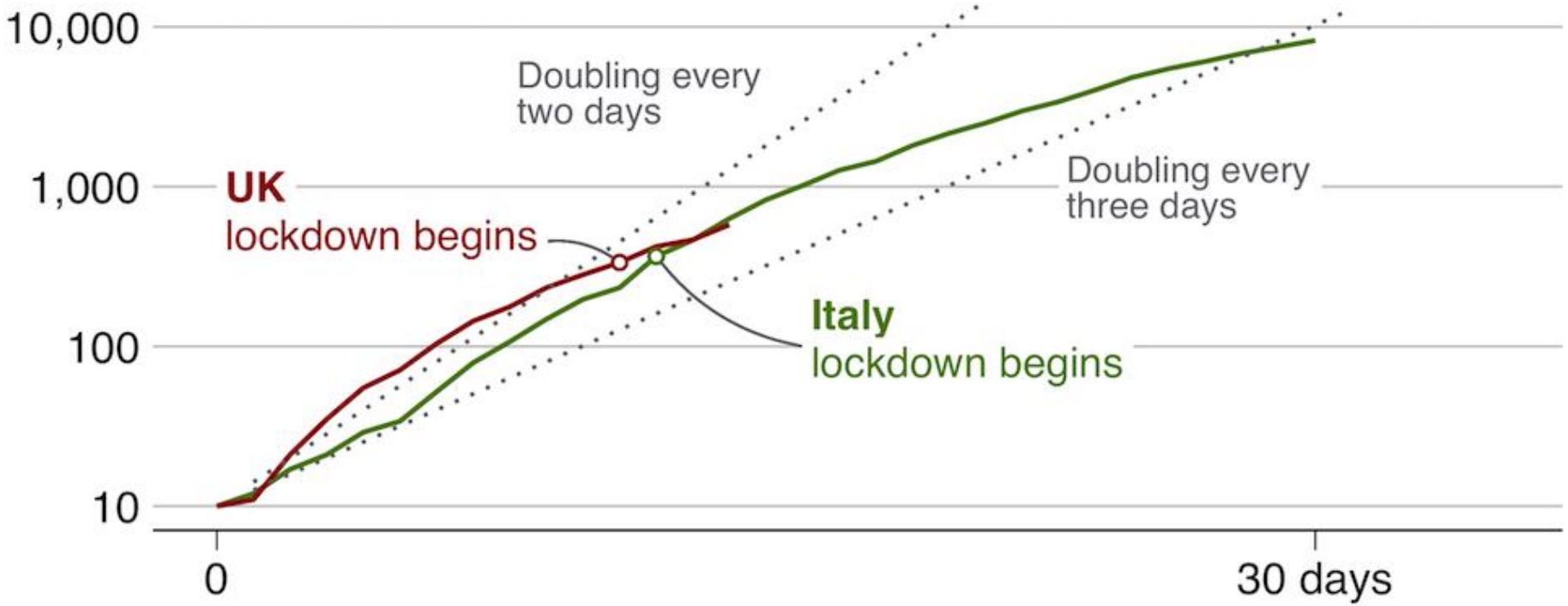
7. előadás

2020. március 31.

XI26VK

UK deaths initially rose faster than Italy's but both countries have since slowed

Compared from the day on which the 10th death was announced in each country



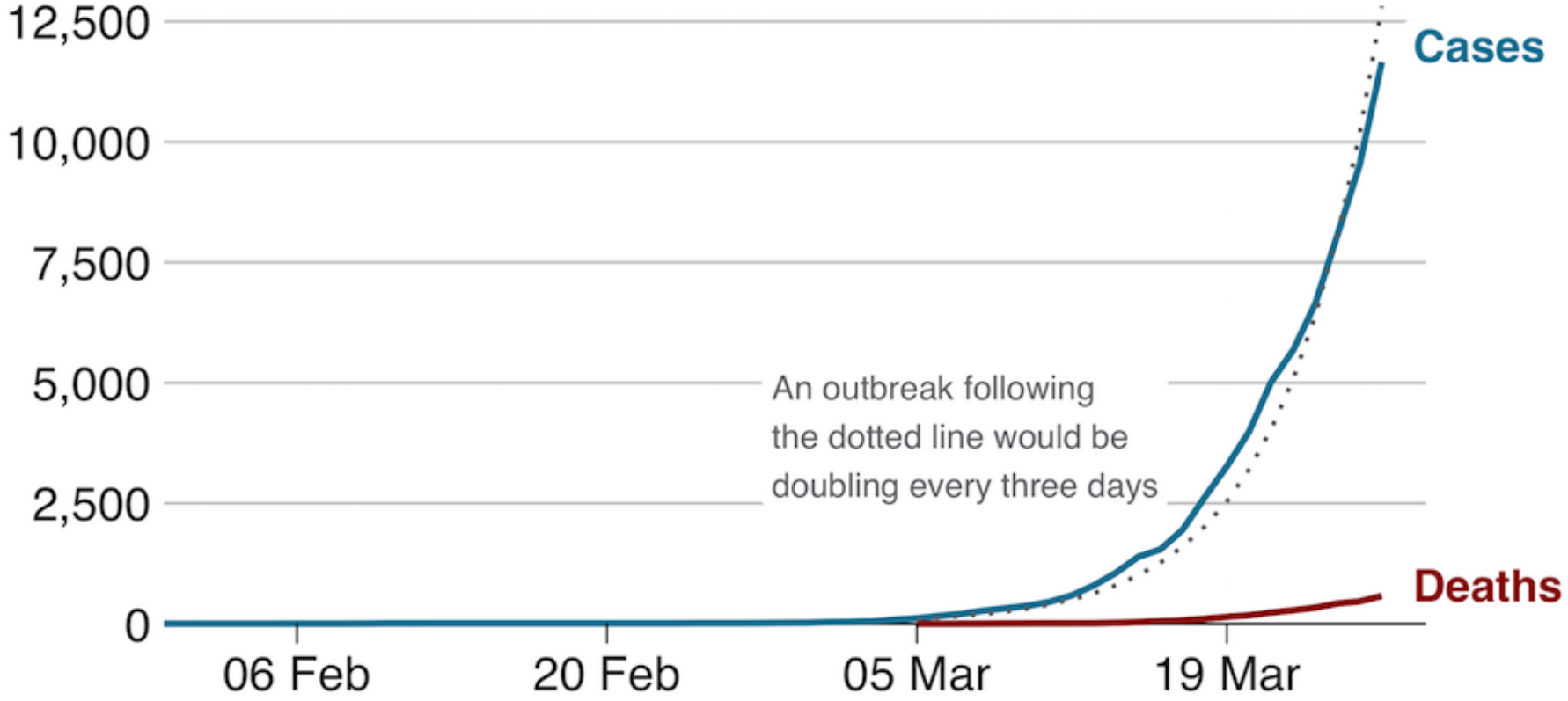
Log scale

Source: Johns Hopkins University. As of 26 Mar 22:00 GMT

BBC

The UK outbreak has so far followed the expected pattern for an epidemic

Confirmed cases are doubling about every three days

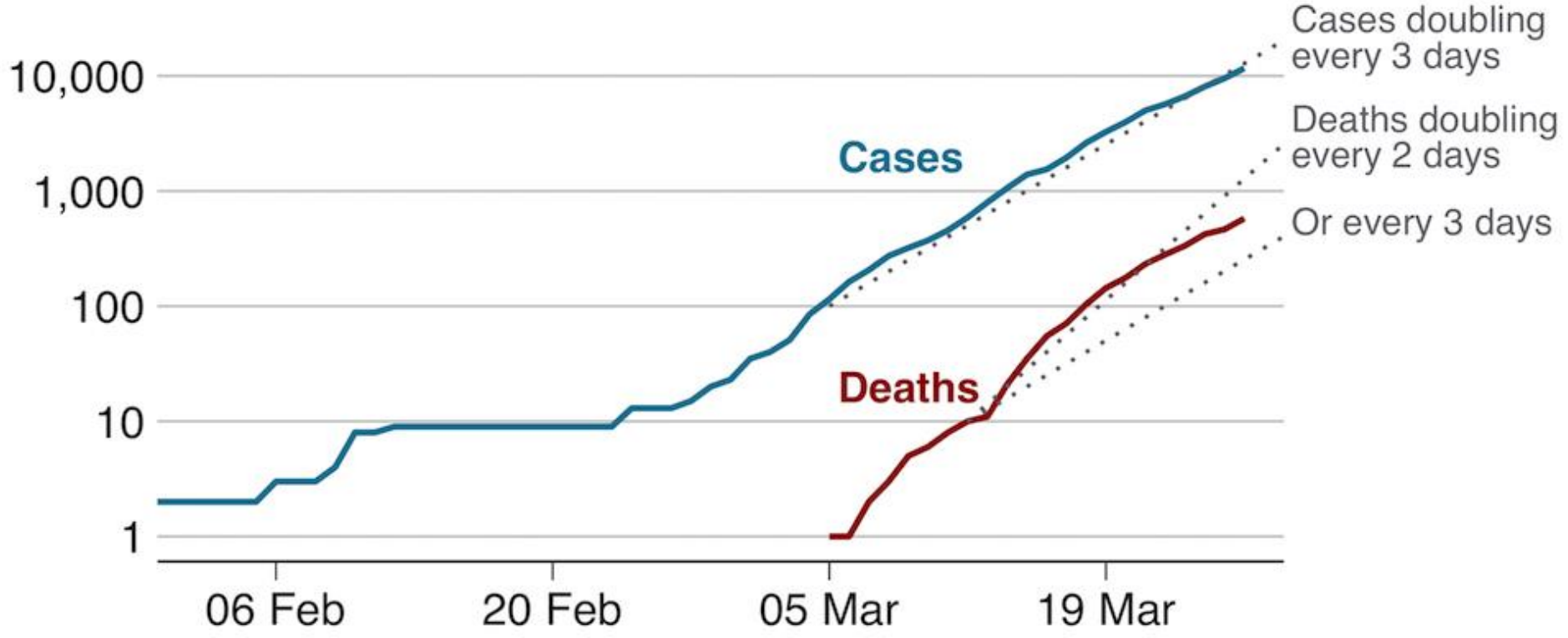


Source: Johns Hopkins University. As of 26 Mar 22:00 GMT

BBC

The UK outbreak has so far followed the expected pattern for an epidemic

This is the same information as the chart above but shown in a different way



Log scale

Source: Johns Hopkins University. As of 26 Mar 22:00 GMT

BBC

Példa

- Milyen valószínűséggel születik fiúgyermek?
- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Fiúk relatív gyakorisága így 0,5141.
- Igaz-e, hogy a valószínűség 0,5? És 0,1?

$$X_i = \begin{cases} 1, & i.\text{fiú} \\ 0, & i.\text{lány} \end{cases} \Rightarrow$$

$$P\left(X_i = 1\right) = p,\ n = 2.644.757,\ \xi = \sum_{i=1}^{n} X_i \Big/ n \quad \Rightarrow$$

$$EX_i = p, D^2 X_i = p(1-p),\ P\left(\frac{\sum_{i=1}^{n} X_i - nEX_1}{DX_1\sqrt{n}} < x \right) \sim \Phi(x) \Rightarrow$$

$$P\left(-u < \sqrt{\frac{n}{p(1-p)}}\left(\xi - p\right) < u \right) \sim 2\Phi(u) - 1$$

$$p = 0.5 \Rightarrow \sqrt{\frac{n}{p(1-p)}}\left(\xi - p\right) = 37$$

$$u = 4 \Rightarrow 2\Phi(u) - 1 = 0,999936$$

$$p(1-p) \le \frac{1}{4} \Rightarrow$$

$$2\Phi(u) - 1 \sim P\left(-u < \sqrt{\frac{n}{p(1-p)}} \left(\xi - p \right) < u \right) \le$$

$$\le P\left(-u < 2\sqrt{n} \left(\xi - p \right) < u \right) =$$

$$= P\left(\frac{-u}{2\sqrt{n}} < \left(\xi - p \right) < \frac{u}{2\sqrt{n}} \right) = P\left(\xi - \frac{u}{2\sqrt{n}} < p < \xi + \frac{u}{2\sqrt{n}} \right)$$

Esetünkben 0,9973 valószínűséggel $0{,}5132 \le p \le 0{,}5150$

Statisztikai mező

$$\left(\Omega, A\ , P_{\vartheta}\right),\ \vartheta \in \Theta$$

statisztikai mező, ha Θ paraméterhalmaz és $\left(\Omega, A\ , P_{\vartheta}\right)$ minden paraméter esetén valószínűségi mező.

Egy érmedobás modellje

- Nem ismerjük a fejdobás valószínűségét:

$$\Omega = \{\mathrm{F}, \mathrm{I}\}, A = \{\varnothing; \{\mathrm{F}\}; \{\mathrm{I}\}; \{\mathrm{F}, \mathrm{I}\}\},$$

$$P_p(\{\mathrm{F}\}) = p, P_p(\{\mathrm{I}\}) = 1 - p, p \in [0,1].$$

Minta

Def.: A $\boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} : \Omega \to \mathcal{X} \subseteq \mathbb{R}^n$ valószínűségi vektorváltozót mintának nevezzük.

n: mintanagyság

ξ_i: i. mintaelem

Def.: minta realizációja:

$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ a konkrét megfigyelt számsorozat.

Mintatér

- Def: $\mathcal{X}$ mintatér: a minta lehetséges értékeinek halmaza. Elemei a mintaértékek.
- *n*-elemű valós minta esetén: $\mathcal{X} = \mathbb{R}^n$
- *n*-elemű pozitív egész értékű minta esetén: $\mathcal{X} = \mathbb{N}^n$
- Példa: egy biztosítónál 10 napon keresztül figyelték a bejelentett károk számát, ekkor $\mathcal{X} = \mathbb{Z}_0^{10}$

Az elmúlt 5 napban elhunyt koronavírusos betegek száma

- Megfigyelések: 0, 1, 2, 2, 1
- Minta és realizációja:

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{pmatrix} \text{ és } \begin{pmatrix} 0 \\ 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}$$

Mintanagyság: 5

A minták típusai

- Független minta: a mintaelemek függetlenek.
- Független azonos eloszlású minta: a mintaelemek függetlenek és azonos eloszlásúak.
- Diszkrét minta: a mintaelemek diszkrétek.
- Abszolút folytonos eloszlású minta: a mintaelemek abszolút folytonosak.

Eloszláscsaládok

$$F_\vartheta(\mathbf{s}) = P_\vartheta\left(\xi_1 < s_1, \ldots, \xi_n < s_n\right)$$

Független minta esetén:

$$F_\vartheta(\mathbf{s}) = \prod_{i=1}^{n} P_\vartheta\left(\xi_i < s_i\right)$$

Független azonos eloszlású minta esetén:

$$F_\vartheta(\mathbf{s}) = \prod_{i=1}^{n} P_\vartheta\left(\xi_i < s_i\right) = \prod_{i=1}^{n} F_\vartheta\left(s_i\right)$$

Jelölések:

E_ϑ: várható érték P_ϑ esetén,

D_ϑ: szórás P_ϑ esetén,

f_ϑ: sűrűségfüggvény P_ϑ esetén (absz. folyt. minta)

$p_\vartheta(s) = P_\vartheta\left(\xi_i = s\right)$ (diszkrét minta)

Példák

- Egy érmedobás. Fej esetén 1-et írunk, írás esetén 0-át.

$$p_p(k) = P_p(\xi_1 = k) = \begin{cases} p & k = 1 \\ 1-p & k = 0 \end{cases} = p^k (1-p)^{1-k}$$

Koronavírusos példa. Azt feltételezzük, hogy megfigyeléseink független, azonos eloszlású Poissonok.

$$p_\lambda(k) = P_\lambda(\xi_i = k) = \lambda^k e^{-\lambda} / k!,\ k = 0,1,2,\ldots$$

Statisztikák

Def.: Statisztika: a minta függvénye.

$$T: \mathcal{X} \to \mathbb{R}^k$$

Def'.: Statisztika:

$$T(\boldsymbol{\xi}), \text{ ha } T: \mathcal{X} \to \mathbb{R}^k \text{függvény.}$$

Tapasztalati momentumok

$\mathcal{X} = \mathbb{R}^n$

mintaközép:

$$T(\mathbf{x}) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad T(\boldsymbol{\xi}) = \bar{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n},$$

tapasztalati k. momentum:

$$T(\mathbf{x}) = \frac{\sum_{i=1}^{n} x_i^k}{n}, \qquad T(\boldsymbol{\xi}) = \frac{\sum_{i=1}^{n} \xi_i^k}{n}.$$

Tapasztalati szórásnégyzet

$$\mathcal{X} = \mathbb{R}^n,$$

$$T(\mathbf{x}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n},$$

$$T(\boldsymbol{\xi}) = s^2 = \frac{\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2}{n}$$

Rendezett minta

- A $\xi_1, \dots, \xi_n$ minta elemeit nagyság szerint sorbarendezve kapjuk az $\xi_1^{(n)} \leq \xi_2^{(n)} \leq \cdots \leq \xi_n^{(n)}$ – rendezett mintát.
- Ez n-dimenziós statisztika
- Mostantól: a $\xi_1, \dots, \xi_n$ minta elemei független, azonos eloszlásúak.
- Ha feltesszük, hogy a közös eloszlásuk abszolút folytonos, akkor felírható a rendezett minta k-adik elemének, $\xi_k^{(n)}$ -nek a sűrűségfüggvénye. (gyakorlat)
- Spec.: minimum, maximum.
- Def.: minta terjedelme: $\xi_n^{(n)} - \xi_1^{(n)}$

Tapasztalati eloszlásfüggvény

- Tapasztalati eloszlás eloszlásfüggvénye: tapasztalati eloszlásfüggvény:

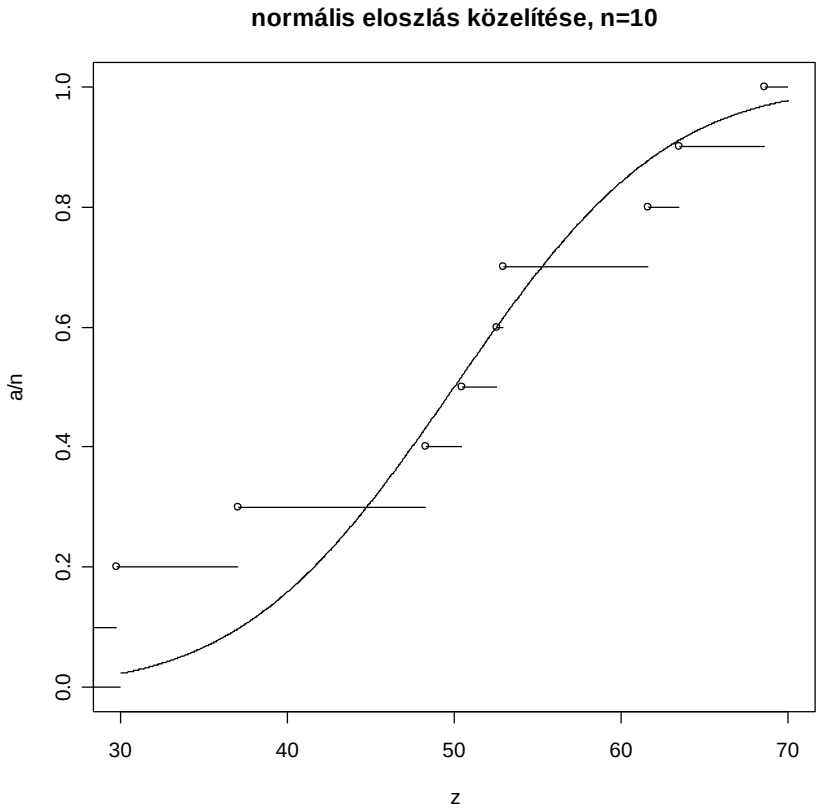
$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i < z\}$$

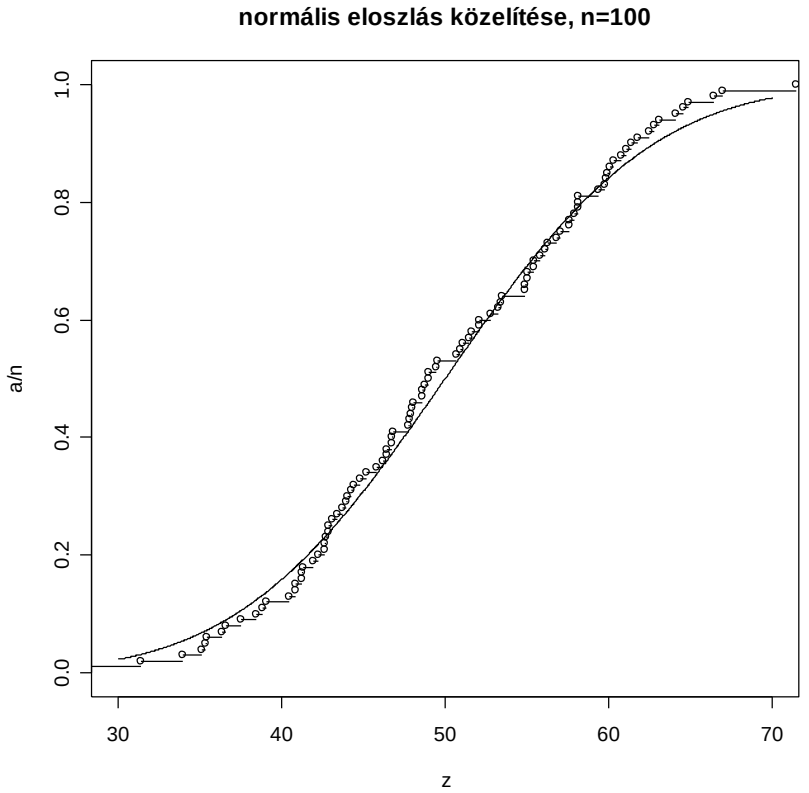
$$F_n(z) = \frac{k}{n}, \text{ha } \xi_k^{(n)} < z \leq \xi_{k+1}^{(n)},$$

$$\xi_0^{(n)} = -\infty, \xi_{n+1}^{(n)} = \infty$$

Mintaátlag éppen ennek az eloszlásnak a várható értéke.

Példa


normális eloszlás közelítése, n=10
a/n
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
z


normális eloszlás közelítése, n=100
a/n
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
z

Glivenko-Cantelli tétel (”statisztika alaptétele”)

Tétel: $\xi_1, \ldots, \xi_n$ független, azonos F eloszlásfüggvényűek. Ekkor $\sup_z |F_n(z) - F(z)| \xrightarrow[n\to\infty]{} 0$ majdnem mindenütt (1 vszgel).

Biz.: Csak folytonos F eloszlásfüggvényekre látjuk be. Ebből következik, hogy tetszőleges pozitív egész N−hez léteznek olyan valós $z_1, \ldots, z_N$ számok, hogy

$$F(z_0) = 0, F(z_1) = \frac{1}{N}, \ldots, F(z_i) = \frac{i}{N}, \ldots, F(z_{N-1}) = \frac{N-1}{N},$$

$$F(z_N) = 1,$$

$$z_0 = -\infty, z_N = \infty.$$

Ekkor, ha $z \in [z_k, z_{k+1})$, akkor

$$F_n(z) - F(z) \le F_n(z_{k+1}) - F(z_k)$$

$$= F_n(z_{k+1}) - F(z_{k+1}) + \frac{1}{N},$$

$$F_n(z) - F(z) \ge F_n(z_k) - F(z_{k+1}) = F_n(z_k) - F(z_k) - \frac{1}{N}.$$

Ebből következik, hogy

$$\sup_z |F_n(z) - F(z)| \le \max_{0 \le k \le N} |F_n(z_k) - F(z_k)| + \frac{1}{N}.$$

Tudjuk, hogy rögzített x − re

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi\{\xi_i < x\},$$

ahol $\chi\{\xi_i < x\}$ független, azonos eloszlású indikátor valószínűségi változók, melyek várható értéke

$$= E(\chi\{\xi_i < x\}) = P(\xi_i < x) = F(x).$$

Így a nagy számok erős törvénye szerint

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i < x\} \xrightarrow[n\to\infty]{} E\chi\{\xi_i < x\} = F(x) \text{ mm.}$$

Legyen $A_{k,N} = \left\{\omega: \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i(\omega) < z_k\} \xrightarrow[n\to\infty]{} F(z_k)\right\}$, ekkor

$P(A_{k,N}) = 1$ és $B_N = \{\omega: \max_{0\le k\le N} |F_n(z_k) - F(z_k)| \xrightarrow[n\to\infty]{} 0\} = \bigcap_{k=1}^{N-1} A_{k,N}$.

B_N − en $\limsup_{n\to\infty} |F_n(z) - F(z)| \le \frac{1}{N}$. Ebből következik,

hogy $\bigcap_{N=1}^{\infty} B_N$ −en $\limsup_{n\to\infty} |F_n(z) - F(z)| = 0$.

1 valószínűségű események metszete is 1 valószínűségű.

Így $\bigcap_{N=1}^{\infty} B_N = \bigcap_{N=1}^{\infty} \bigcap_{k=1}^{N-1} A_{k,N}$ is 1 valószínűségű.

Becsléselmélet

- A minta eloszlásának ismeretlen paraméterét közelítjük a minta függvényével
- Def.: becslőfüggvény: $\hat{\vartheta}: \mathcal{X} \rightarrow \Theta$
- Def.: becslés: $\hat{\vartheta}(\boldsymbol{\xi})$

- A becslések maguk is statisztikák. Szubjektíven: olyan statisztikák, amik jól közelítik az ismeretlen paramétert.

Példa (Milyen valószínűséggel születik fiúgyermek?)

- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Ekkor $n = 2.644.757$, $\mathcal{X} = \{0; 1\}^n$.
- Fiúk relatív gyakorisága így 0,5141.
- Mik ennek a becslésnek a tulajdonságai?

$$X_i = \begin{cases} 1, & i.\text{fiú} \\ 0, & i.\text{lány} \end{cases} \Rightarrow$$

$$P_p(X_i = 1) = p,$$

$$n = 2.644.757,$$

$$\hat{p} = \hat{p}(\mathbf{X}) = \frac{\sum_{i=1}^{n} X_i}{n} \Rightarrow$$

$$E_p\hat{p} = p,$$

$$\hat{p} \xrightarrow[n\to\infty]{} p \text{ mm.}$$

Becslések tulajdonságai

- Def.: *Torzítatlanság:* A paraméter $\hat{\vartheta}(\boldsymbol{\xi})$ becslése torzítatlan, ha

$$E_{\vartheta}\left(\hat{\vartheta}(\boldsymbol{\xi})\right) = \vartheta, \forall \vartheta \in \Theta.$$

- *Konzisztencia*: $\hat{\vartheta}(\boldsymbol{\xi}) \to \vartheta$ sztochasztikusan $(n \to \infty)$ minden paraméterértékre.
- Példák:
  - Valószínűség becslése relatív gyakorisággal.
  - Glivenko tétele: a tapasztalati eloszlásfüggvény egyenletesen is konvergál az elméleti eloszlásfüggvényhez.
  - Várható érték becslése mintaátlaggal

Konzisztencia

- Elégséges feltétel $E_\vartheta\left(\hat{\vartheta}_n(\boldsymbol{\xi})\right) \to \vartheta$ (aszimptotikus torzítatlanság) és $D_\vartheta{}^2\left(\hat{\vartheta}_n(\boldsymbol{\xi})\right) \to 0$.

Példák

- Poisson eloszlás paraméterére: mintaátlag
- Exponenciális eloszlás paraméterére:
  - mintaátlag reciproka: aszimptotikusan torzítatlan, konzisztens
  - $n \cdot \min(X_1, \ldots, X_n)$ torzítatlan, de nem konzisztens
- Szórásnégyzetre

Becslések összehasonlítása

- Melyik a jobb becslés?

$$X_i = \begin{cases} 1, & i.\,\text{fiú} \\ 0, & i.\,\text{lány} \end{cases}, P_p(X_i = 1) = p,$$

$$\hat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{n},$$
$$\hat{p}_2 = X_1, \text{vagy}$$

$$\hat{p}_3 = \frac{\sum_{i=1}^{[n/2]} X_i}{[n/2]}?$$

Becslések összehasonlítása (hatásos becslések)

- Torzítatlan becslésekre: T_1 hatásosabb becslése $h(\Theta)$-nak a T_2-nél, ha

$$D_\theta^2\left(T_1(\underline{X})\right) \leq D_\theta^2\left(T_2(\underline{X})\right)$$

- teljesül minden Θ paraméterértékre.

Példa: a mintaátlag hatásosabb becslés a várható értékre minden

$$\sum_{i=1}^{n} c_i X_i$$

alakú becslésnél.

Hatásos becslés

- Def.: A T torzítatlan becslés hatásos, ha minden más torzítatlan becslésnél hatásosabb.
- Miért a torzítatlanokra? Furcsa példa: azonosan 0-val becsüljük az ismeretlen paramétert.
- Ezért érdemes a hatásos becsléseket csak a torzítatlan becslések között keresni.
- Átlagos négyzetes eltérés:

$$E_\theta\left(T(\underline{X}) - \theta\right)^2$$

Hatásos becslés egyértelműsége

Áll.: Amennyiben T_1 és T_2 hatásos becslései $h(\theta)$-nak, akkor 1 valószínűséggel megegyeznek minden lehetséges paraméter esetén.

$E_\theta T_1 = E_\theta T_2 = h(\theta)$, továbbá $D_\theta T_1 = D_\theta T_2$. Ebből

$$\mathrm{D}_\theta^2(T_1) \le \mathrm{D}_\theta^2\left(\frac{T_1+T_2}{2}\right) = \frac{\mathrm{D}_\theta^2(T_1)+2\,\mathrm{cov}(T_1,T_2)+\mathrm{D}_\theta^2(T_2)}{4} = \frac{\mathrm{D}_\theta^2(T_1)+\mathrm{cov}(T_1,T_2)}{2} \Rightarrow$$

$$\mathrm{D}_\theta^2(T_1) \le \mathrm{cov}(T_1,T_2) = D_\theta T_1 \cdot D_\theta T_2 \cdot R(T_1,T_2) = \mathrm{D}_\theta^2(T_1) \cdot R(T_1,T_2) \le \mathrm{D}_\theta^2(T_1) \Rightarrow$$

$$\mathrm{D}_\theta^2(T_1) = \mathrm{D}_\theta^2(T_2) = \mathrm{cov}(T_1,T_2) \Rightarrow \mathrm{D}_\theta^2(T_1-T_2) = \mathrm{D}_\theta^2(T_1) - 2\,\mathrm{cov}(T_1,T_2) + \mathrm{D}_\theta^2(T_2) = 0.$$

Így $E_\theta(T_1 = T_2) = 1 \;\forall \theta \in \Theta$.

Mit kell tudni a mintáról?

- Benzinkutas példa. Megfigyelések: 78, 89, 167, 90, 85.
- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.

Kár-szám	0	1	2	3	4	5	6	7	> 7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006

Mennyi információt hordoz a statisztika?

Példa: $\xi_1, \ldots, \xi_n$ független $N(m,1)$ minta. Ekkor

$$\bar{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n} \sim N(m, \frac{1}{n})$$ eloszlású (függ m-től!),

miközben

$$s^2 = \frac{\sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}{n}$$ eloszlása nem függ m-től!

Elégséges statisztika

- Minden információt (ugyanannyit mint az eredeti minta) tartalmaz az ismeretlen paraméterre vonatkozóan.
- ”Elég” az ő értékét ismerni.
- Ismeretében már ”nincs bizonytalanság” a mintában (úgy értve, hogy egyértelmű a minta eloszlása, már nem függ az ismeretlen paramétertől).

Elégséges statisztika diszkrét minta esetén

Def.: A diszkrét $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges θ−ra, ha a $P_\theta(\boldsymbol{\xi} = \mathbf{x} | T(\boldsymbol{\xi}) = t)$ feltételes valószínűség nem függ θ−tól

Feltételes várható érték

Legyenek X és Y diszkrét val. változók.

$E(X|Y)$ az a val. változó, ami az $Y = y_k$ eseményen az $E(X|Y = y_k)$ értéket veszi fel.

Tulajdonságok:

- Ha $X \geq 0$, akkor $E(X|Y) \geq 0$
- $E(E(X|Y)) = EX$ (a teljes várható érték tételének általánosítasa)
- Ha X_1, X_2 várható értéke véges, akkor $E(c_1X_1 + c_2X_2|Y) = c_1E(X_1|Y) + c_2E(X_2|Y)$
- Ha X független Y –tól, akkor $E(X|Y) = E(X)$
- Ha X és $h(Y)$ várható értéke véges, akkor $E(h(Y)X|Y) = h(Y)E(X|Y)$
- Teljes szórásnégyzet tétele:

$$D^2(X) = D^2(E(X|Y)) + E(D^2(X|Y))$$

Példa (indikátor minta)

$$X_i = \begin{cases} 1, & p \text{ valószínűséggel} \\ 0, & 1-p \text{ valószínűséggel} \end{cases} \Rightarrow P_p\left(X_i = x\right) = p^x (1-p)^{1-x}, x = 0 \text{ és} 1.$$

$$P_p\left(\mathbf{X} = \mathbf{x} \middle| \sum_{i=1}^n X_i = t\right) = P_p\left(X_1 = x_1, ..., X_n = x_n \middle| \sum_{i=1}^n X_i = t\right) =$$

$$\frac{P_p\left(X_1 = x_1, ..., X_n = x_n, \sum_{i=1}^n X_i = t\right)}{P_p\left(\sum_{i=1}^n X_i = t\right)} = \begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \dfrac{P_p\left(X_1 = x_1, ..., X_n = x_n\right)}{P_p\left(\sum_{i=1}^n X_i = t\right)} & \sum_{i=1}^n x_i = t \end{cases} =$$

$$\begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \dfrac{p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} & \sum_{i=1}^n x_i = t \end{cases} = \begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \dfrac{1}{\binom{n}{t}} & \sum_{i=1}^n x_i = t \end{cases}$$

Tétel (Neyman-féle faktorizációs):

A diszkrét ξ mintából képzett $T(\xi)$ statisztika pontosan akkor elégséges Θ-ra, ha $\exists\, g_\theta(t)$ és $h(\mathbf{x})$ úgy, hogy $\forall \theta \in \Theta$ és $\mathbf{x} \in \mathcal{X}$-ra

$P_\theta(\xi = \mathbf{x}) = h(\mathbf{x}) g_\theta(T(\mathbf{x}))$.

Biz.:

$\Rightarrow T(\xi)$ elégséges, ekkor $P_\theta(\xi = \mathbf{x}) = P_\theta(T(\xi) = T(\mathbf{x})) \dfrac{P_\theta(\xi = \mathbf{x}, T(\xi) = T(\mathbf{x}))}{P_\theta(T(\xi) = T(\mathbf{x}))}$

$= P_\theta(T(\xi) = T(\mathbf{x})) P_\theta(\xi = \mathbf{x} \,|\, T(\xi) = T(\mathbf{x})) = g_\theta(T(\mathbf{x})) h(\mathbf{x})$.

$\Leftarrow P_\theta(\xi = \mathbf{x} \,|\, T(\xi) = t) = 0$, ha $t \neq T(\mathbf{x})$. Amennyiben ez teljesül:

$$P_\theta(\xi = \mathbf{x} \,|\, T(\xi) = t) = \frac{P_\theta(\xi = \mathbf{x}, T(\xi) = t)}{P_\theta(T(\xi) = t)} = \frac{P_\theta(\xi = \mathbf{x}, T(\xi) = t)}{P_\theta(T(\xi) = t)} = \frac{P_\theta(\xi = \mathbf{x})}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} P_\theta(\xi = \mathbf{y})}$$

$$= \frac{h(\mathbf{x}) g_\theta(T(\mathbf{x}))}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y}) g_\theta(T(\mathbf{y}))} = \frac{h(\mathbf{x}) g_\theta(t)}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y}) g_\theta(t)} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y})}.$$

Ez nem függ θ-tól!

Példa (Poisson minta)

η_i – k független λ Poissonok. Ekkor

$$P_\lambda\left(\eta_1 = k_1, \ldots, \eta_n = k_n\right) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda} =$$

$$= h(\mathbf{k}) g_\lambda\left(\sum_{i=1}^{n} k_i\right),$$

ahol

$$h(\mathbf{k}) = \prod_{i=1}^{n} \frac{1}{k_i!}, \quad g_\lambda(t) = \lambda^t e^{-n\lambda}.$$

Elégséges statisztika általában

Def.: A $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges

Θ-ra, ha minden $\mathbf{x} \in \mathbf{R}^n$-re a $P_\theta(\boldsymbol{\xi} < \mathbf{x} \,|\, T(\boldsymbol{\xi}) = t) = P_\theta(\xi_1 < x_1, \ldots, \xi_n < x_n \,|\, T(\boldsymbol{\xi}) = t)$ feltételes eloszlásfüggvény nem függ θ-tól.

Probléma: A feltételes valószínűség és várható érték fogalmát nem tanultuk általánosan!

Likelihood függvény

Def.: A $\xi_1, ..., \xi_n$ független, azonos eloszlású minta likelihood függvénye

$$L(\mathbf{x}, \theta) = \begin{cases} P_\theta(\boldsymbol{\xi} = \mathbf{x}) = \prod_{i=1}^{n} P_\theta(\xi_i = x_i) & \text{diszkrét minta esetén} \\ f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) & \text{abszolút folytonos minta esetén} \end{cases}$$

ahol f_θ ξ_i sűrűségfüggvénye.

$l(\mathbf{x}, \theta) = \ln L(\mathbf{x}, \theta)$ a loglikelihood függvény.

Abszolút folytonos eset

- Definíció a faktorizációval

Def.:

Az abszolút folytonos $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges $\Theta-$ra, ha $\exists\ g_\theta(t)$ és $h(\mathbf{x})$ úgy, hogy $\forall\theta \in \Theta$ és $\mathbf{x} \in \mathcal{X}-$ra a likelihood függvény felírható a következő alakban:
$L(\mathbf{x}, \theta) = h(\mathbf{x}) g_\theta(T(\mathbf{x}))$.

Példa (normális $N(m,\sigma^2)$ minta)

ξ_i – k független, $N(m,\sigma^2)$ eloszlásúak. Ekkor $\theta=(m,\sigma^2)$

$$L(\mathbf{x},(m,\sigma^2))=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-m)^2}{2\sigma^2}\right)=$$

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i^2-2x_im+m^2\right)\right)=$$

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2-2nm\bar{x}+nm^2\right)\right)=$$

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\bar{x}^2-2nm\bar{x}+nm^2\right)\right).$$

Ebből következik, hogy $\left(\sum_{i=1}^{n}x_i^2,\bar{x}\right)$ elégséges statisztika.

Hasonlóan $\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2/n,\bar{x}\right)$ is.

Példa (egyenletes $E(0,a)$ minta)

ξ_i – k független, $E(0,a)$ eloszlásúak. Sűrűségfüggvényük

$$f_a(x)=\begin{cases}1/a & 0\le x\le a\\ 0 & \text{különben}\end{cases}$$

$$L(\mathbf{x},a)=\prod_{i=1}^{n}\frac{1}{a}\chi\{x_i\le a\}=\frac{1}{a^n}\chi\left\{\max_{1\le i\le n}x_i\le a\right\}\Rightarrow$$

$\max\limits_{1\le i\le n}x_i$ elégséges!

Becslési módszerek

- Példa: Egy tóban N hal van, számukat nem ismerjük. Első héten kihalásznak 1000 halat és megjelölik őket. A következő héten kihalásznak 5000-et és megszámolják a megjelölteket. 50-et találnak. Becsüljük meg N-et!

Természetes eljárás

Jelölje ξ a másodjára kihúzott halak számát.

Tudjuk, hogy ez hipergeometrikus eloszlású, így

$$L(50, N) = P_N(\xi = 50) = \frac{\binom{1000}{50}\binom{N-1000}{4950}}{\binom{N}{5000}}.$$

Becslés

$$\hat{N} : L(50, \hat{N}) = \max_N L(50, N) \Rightarrow \hat{N} = 100000$$

Maximum likelihood becslés

- Definíció heurisztikusan: azt a paraméterértéket keressük, amelyre az adott minta bekövetkezési valószínűsége maximális.

Def.: θ maximum likelihood becslése $\hat{\theta} = T(\boldsymbol{\xi}) \in \Theta$, ha

$$L(\boldsymbol{\xi}, \hat{\theta}) = \max_{\theta \in \Theta} L(\boldsymbol{\xi}, \theta)$$

Likelihood egyenlet

Gyakran a loglikelihood függvény maximumhelyét keresik a

$\frac{\partial l(\mathbf{x},\theta)}{\partial \theta} = 0$ egyenletet (vagy egyenletrendszert) megoldva.

Ez diszkrét minta esetén a

$$\sum_{i=1}^{n} \frac{\partial \ln P_\theta(\xi_i = x_i)}{\partial \theta} = 0$$

egyenletet (vagy egyenletrendszert) jelenti.

Abszolút folytonos minta esetén

$$\sum_{i=1}^{n} \frac{\partial \ln f_\theta(x_i)}{\partial \theta} = 0$$

egyenletet (vagy egyenletrendszert) oldjuk meg.

Példa (indikátor)

$$L(\mathbf{x}, p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i},$$

$$l(\mathbf{x}, p) = \ln L(\mathbf{x}, p) = \left(\sum_{i=1}^{n} x_i \right) \ln p + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p).$$

Likelihood egyenlet

$$\frac{\partial l(\mathbf{x}, p)}{\partial p} = \left(\sum_{i=1}^{n} x_i \right) \frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i \right) \frac{1}{1-p} = 0$$

Ennek megoldása

$$p = \frac{\sum_{i=1}^{n} x_i}{n}.$$

És ez valóban maximumhely!

Így a ML becslés

$$\hat{p} = \frac{\sum_{i=1}^{n} \xi_i}{n}.$$