

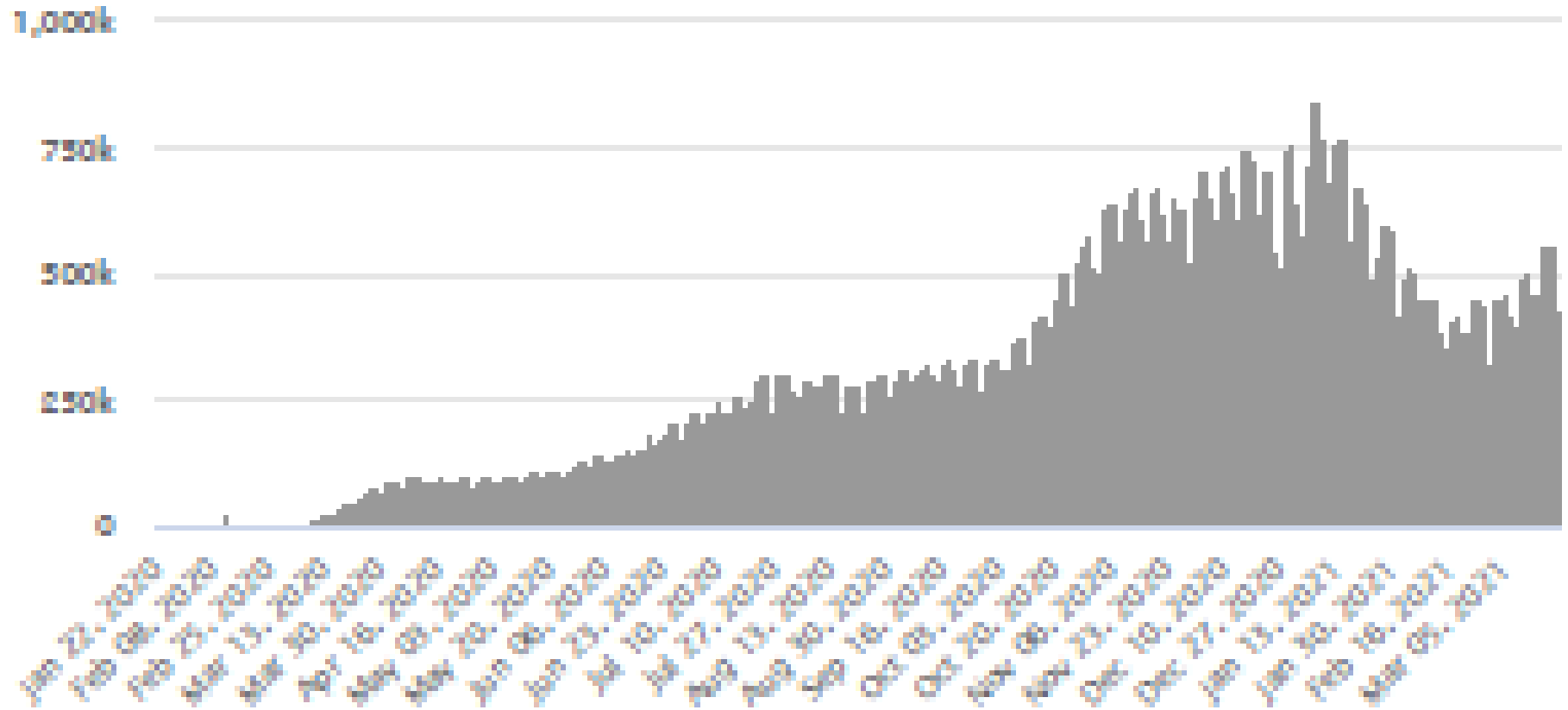
Valószínűségszámítás és Statisztika

7. előadás
2020. március 23.

Daily New Cases

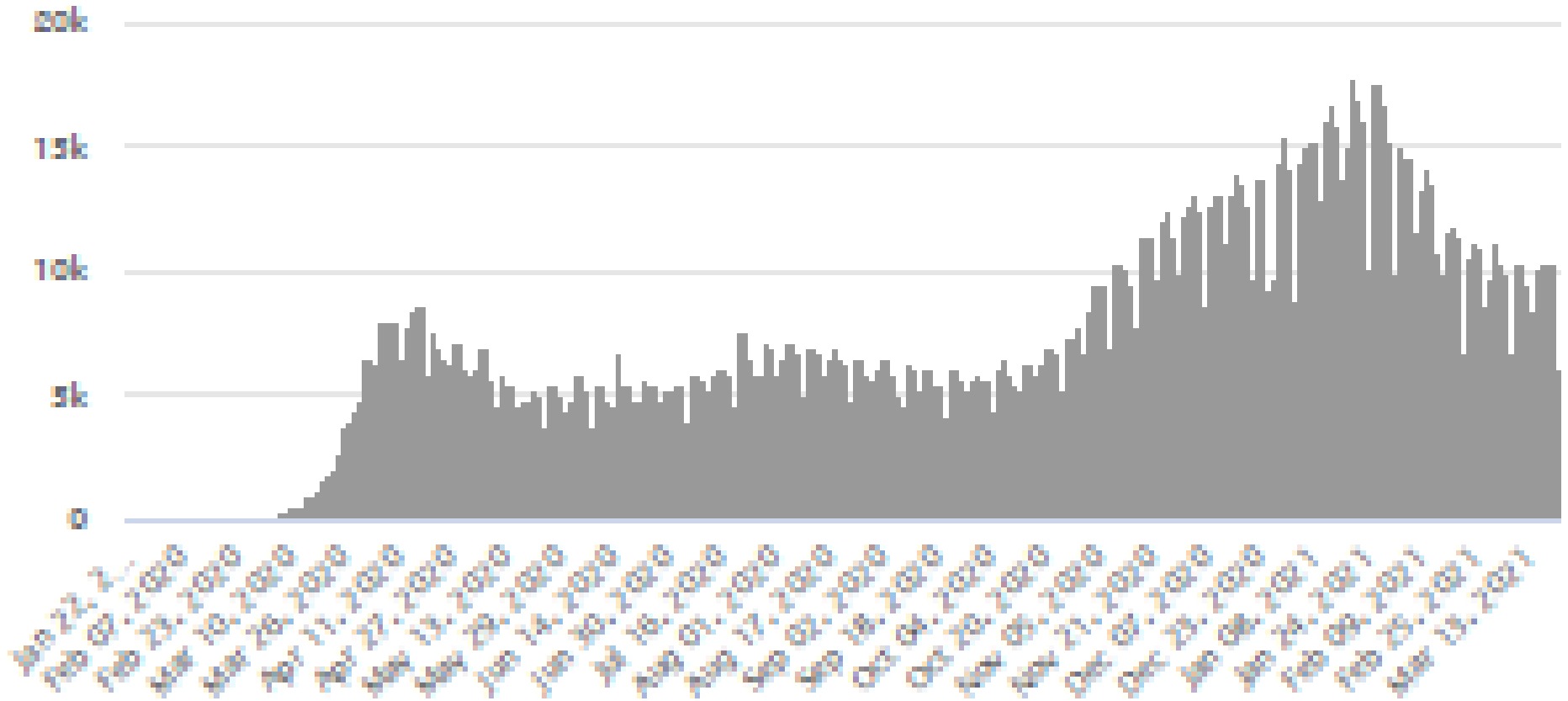
Cases per Day

Data as of 0:00 GMT+0



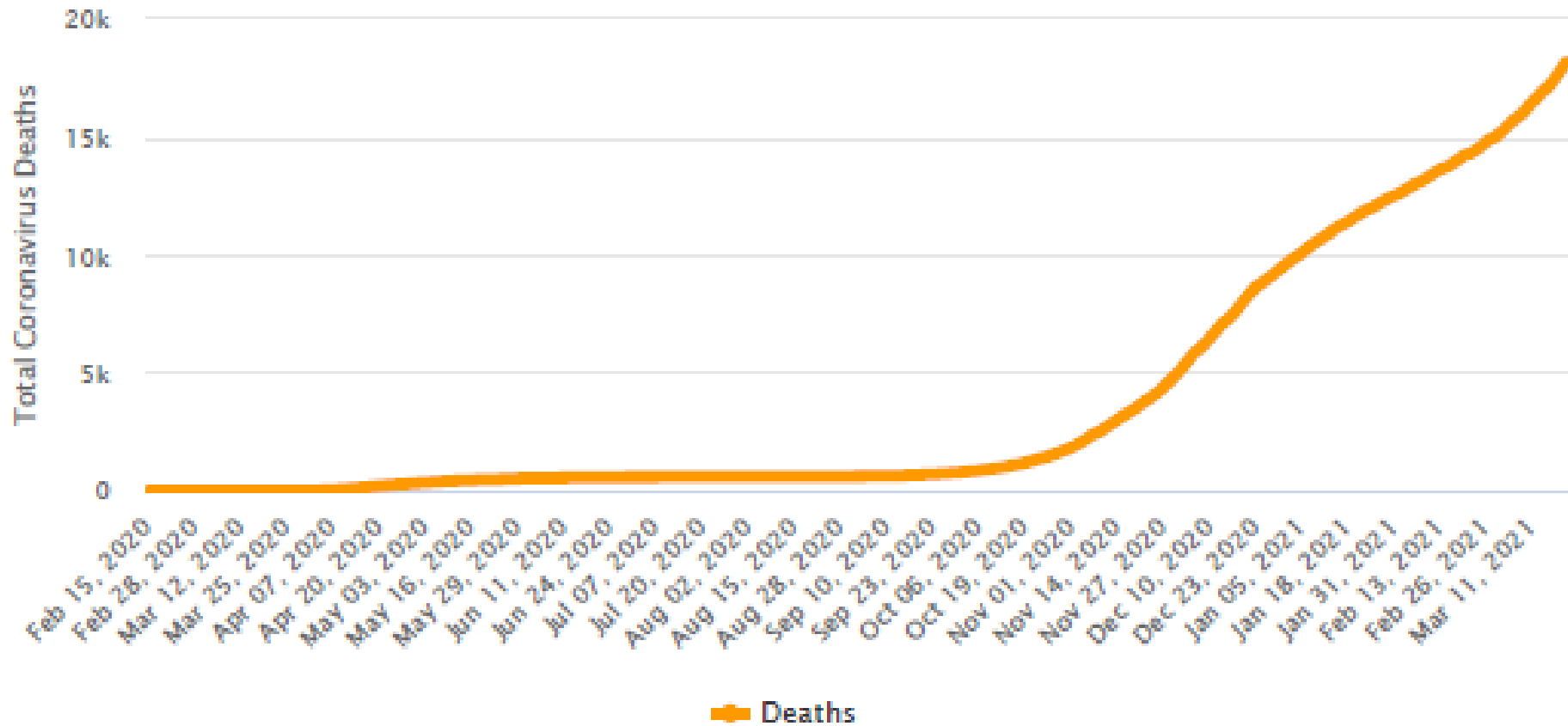
Daily Deaths

Deaths per Day
Data as of 0:00 GMT+0



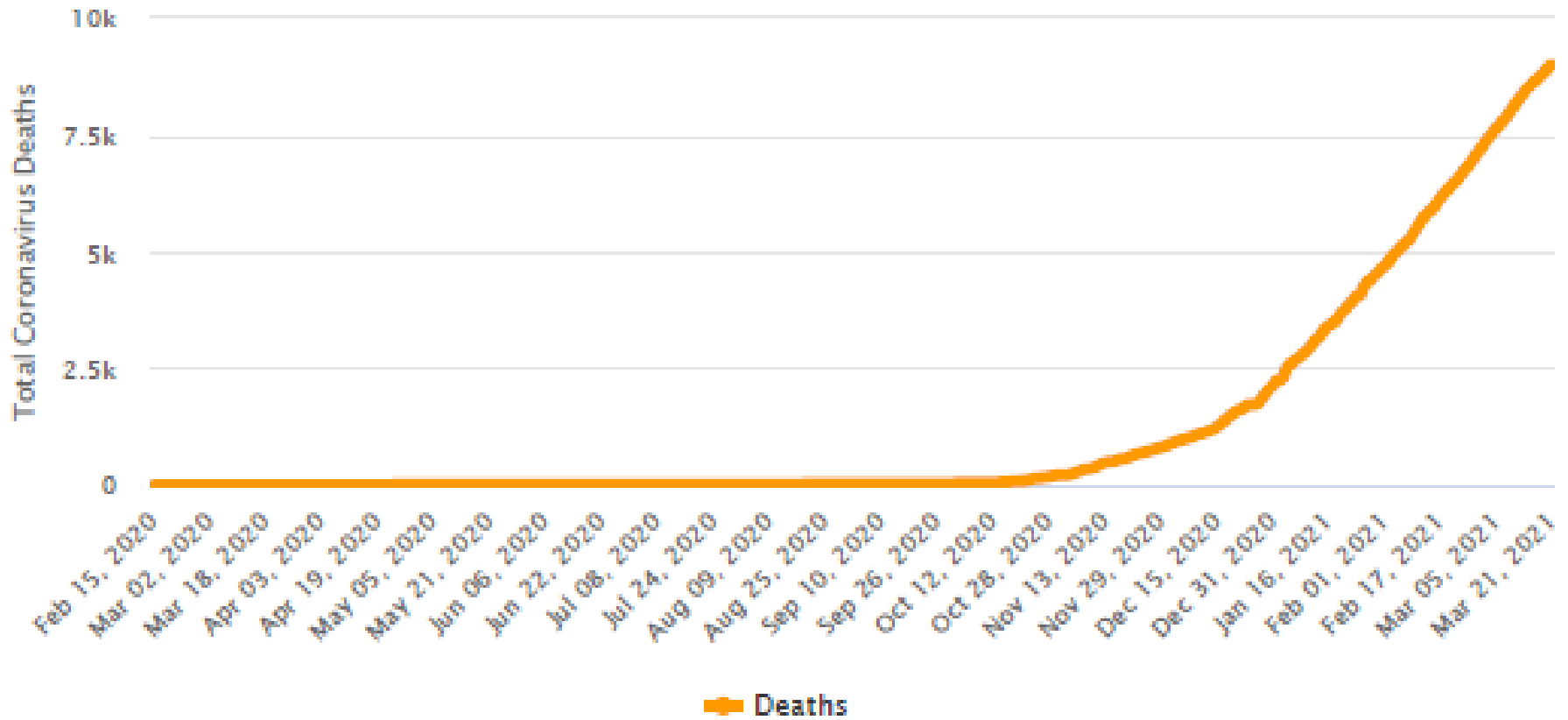
Total Deaths

(Linear Scale)



Total Deaths

(Linear Scale)



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?

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

$$P(X_i = 1) = p, n = 2.644.757, \xi = \frac{\sum_{i=1}^n X_i}{n} \Rightarrow$$

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

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

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

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

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

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

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

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

Esetünkben 0,9973 valószínűséggel $0,5132 \leq p \leq 0,5150$

Statisztikai mező

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

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

Egy érmédobás modellje

- Nem ismerjük a fejdobás valószínűségét:

$$\Omega = \{F, I\}, A = \{\emptyset; \{F\}; \{I\}; \{F, I\}\},$$

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

Minta

Def.: A $\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} : \Omega \rightarrow \mathcal{X}$

$\subseteq \mathbb{R}^n$ 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 diszkréték.
- Abszolút folytonos eloszlású minta: a mintaelemek abszolút folytonosak.

Eloszláscsaládok

$$F_g(\mathbf{s}) = P_g(\xi_1 < s_1, \dots, \xi_n < s_n)$$

Független minta esetén:

$$F_g(\mathbf{s}) = \prod_{i=1}^n P_g(\xi_i < s_i)$$

Független azonos eloszlású minta esetén:

$$F_g(\mathbf{s}) = \prod_{i=1}^n P_g(\xi_i < s_i) = \prod_{i=1}^n F_g(s_i)$$

Jelölések:

E_g : várható érték P_g esetén,

D_g : szórás P_g esetén,

f_g : sűrűségfüggvény P_g esetén (absz. folyt. minta)

$p_g(s) = P_g(\xi_i = s)$ (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.

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

Statisztikák

Def.: Statisztika: a minta függvénye.

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

Def'.: Statisztika:

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

Tapasztalati momentumok

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

mintaközép:

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

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

tapasztalati k . momentum:

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

$$T(\boldsymbol{\xi}) = \frac{\sum_{i=1}^n \xi_i^k}{n}.$$

Tapasztalati szórásnégyzet

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

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

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

Rendezett minta

- A ξ_1, \dots, ξ_n minta elemeit nagyság szerint sorbarendezeve kapjuk az $\xi_1^{(n)} \leq \xi_2^{(n)} \leq \dots \leq \xi_n^{(n)}$ – rendezett mintát.
- Ez n -dimenziós statisztika
- Mostantól: a ξ_1, \dots, ξ_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:

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

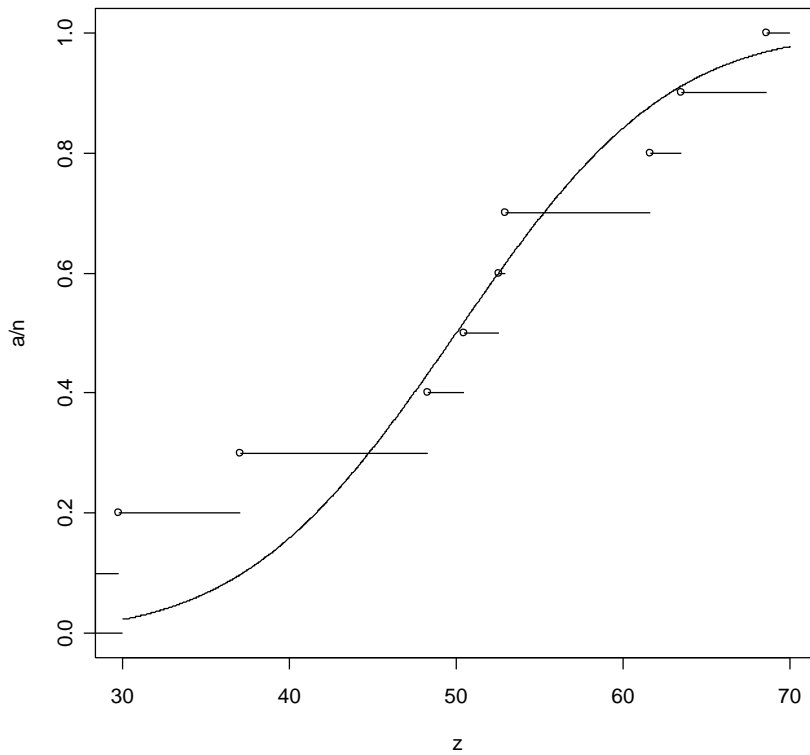
$$F_n(z) = \frac{k}{n}, \text{ ha } \xi_k^{(n)} < z \leq \xi_{k+1}^{(n)},$$

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

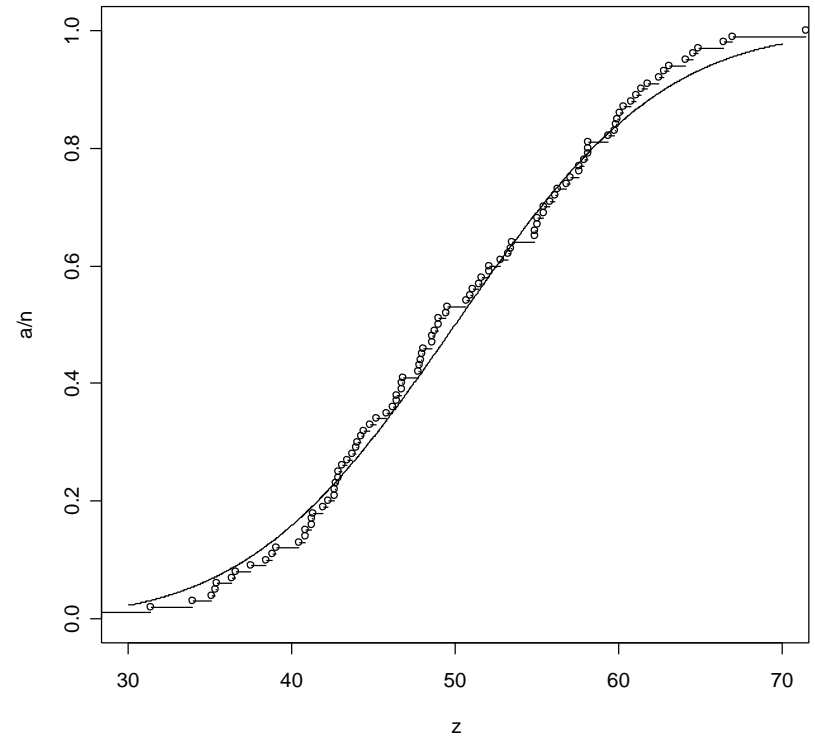
Mintaátlag éppen ennek az eloszlásnak a várható értéke.

Példa

normális eloszlás közelítése, $n=10$



normális eloszlás közelítése, $n=100$



Glivenko-Cantelli tétel ("statisztika alaptétele")

Tétel: ξ_1, \dots, ξ_n független, azonos F eloszlásfüggvényűek. Ekkor $\sup_z |F_n(z) - F(z)| \xrightarrow{n \rightarrow \infty} 0$ majdnem mindenütt (1 vszgel).

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

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

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

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

$$\begin{aligned} F_n(z) - F(z) &\leq F_n(z_{k+1}) - F(z_k) \\ &= F_n(z_{k+1}) - F(z_{k+1}) + \frac{1}{N}, \end{aligned}$$

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

Ebből következik, hogy

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

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

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

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

$$P(A_{k,N}) = 1 \text{ és } B_N = \left\{ \omega : \max_{0 \leq k \leq N} |F_n(z_k) - F(z_k)| \xrightarrow{n \rightarrow \infty} 0 \right\} = \bigcap_{k=1}^{N-1} A_{k,N}.$$

B_N – en $\limsup_{n \rightarrow \infty} |F_n(z) - F(z)| \leq \frac{1}{N}$. Ebből következik,

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

1 valószínűségű események metszete is 1 valószínűségű.

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

Becslésemélet

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

- 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 $n = 2.644.757$, $x = \{0; 1\}^n$.
- Fiúk relatív gyakorisága így 0,5141.
- Mik ennek a becslésnek a tulajdonságai?

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

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

$$n = 2.644.757,$$

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

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

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

Becslések tulajdonságai

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

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

- *Konzisztencia*: $\hat{\vartheta}(\xi) \rightarrow \vartheta$ sztochasztikusan ($n \rightarrow \infty$) 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 $E_{\vartheta}(\hat{\vartheta}_n(\xi)) \rightarrow \vartheta$
(aszimptotikus torzítatlanság)
és $D_{\vartheta}^2(\hat{\vartheta}_n(\xi)) \rightarrow 0$.

Példák

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

Becslések összehasonlítása

- Melyik a jobb becslés?

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

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

$$\hat{p}_2 = X_1, \text{ vagy}$$

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

Becslések összehasonlítása (hatásos becslések)

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

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

- 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:

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

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

$$D_\theta^2(T_1) \leq D_\theta^2\left(\frac{T_1 + T_2}{2}\right) = \frac{D_\theta^2(T_1) + 2\text{cov}(T_1, T_2) + D_\theta^2(T_2)}{4} = \frac{D_\theta^2(T_1) + \text{cov}(T_1, T_2)}{2} \Rightarrow$$

$$D_\theta^2(T_1) \leq \text{cov}(T_1, T_2) = D_\theta T_1 \square D_\theta T_2 \square R(T_1, T_2) = D_\theta^2(T_1) \square R(T_1, T_2) \leq D_\theta^2(T_1) \Rightarrow$$

$$D_\theta^2(T_1) = D_\theta^2(T_2) = \text{cov}(T_1, T_2) \Rightarrow D_\theta^2(T_1 - T_2) = D_\theta^2(T_1) - 2\text{cov}(T_1, T_2) + D_\theta^2(T_2) = 0.$$

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

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: ξ_1, \dots, ξ_n független $N(m, 1)$ minta. Ekkor

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

miközben

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

Elégséges statisztika

- Minden információt (ugyanannyit mint az eredeti minta) tartalmaz az ismeretlen paraméterre vonatkozóan.
- "Elég" az θ 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 ξ mintából képzett $T(\xi)$ statisztika elégséges θ -ra, ha a $P_{\theta}(\xi = \mathbf{x} | T(\xi) = t)$ 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 általánosítása)
- 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 -től, 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)

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

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

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

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

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.:

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

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

$\Leftarrow P_\theta(\xi = \mathbf{x} | T(\xi) = t) = 0$, 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

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

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 ξ mintából képzett $T(\xi)$ statisztika elégséges

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

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

Likelihood függvény

Def.: A ξ_1, \dots, ξ_n független, azonos eloszlású minta likelihood függvénye

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

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 ξ mintából képzett $T(\xi)$ statisztika 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 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 $\theta = (m, \sigma^2)$

$$\begin{aligned} L(\mathbf{x}, (m, \sigma^2)) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right) = \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^2 - 2x_i m + m^2)\right) = \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2nm\bar{x} + nm^2\right)\right) = \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2 - 2nm\bar{x} + nm^2\right)\right). \end{aligned}$$

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

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

Példa (egyenletes $E(0, a)$ minta)

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

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

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

$\max_{1 \leq i \leq n} x_i$ 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 megszámlolják 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 $\hat{\theta} = T(\xi) \in \Theta$, ha

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

Likelihood egyenlet

Gyakran a loglikelihood függvény maximumhelyét keresik a $\frac{\partial l(\mathbf{x}, \theta)}{\partial \theta} = 0$ egyenletet (vagy egyenletrendszer) megoldva.

Ez diszkrét minta esetén a

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

egyenletet (vagy egyenletrendszer) jelenti.

Abszolút folytonos minta esetén

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

egyenletet (vagy egyenletrendszer) 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

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