A sztochasztika alapjai
MBNXK262

11. előadás: Becslések

2022/23 tavasz

Az $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ minta *empirikus kovarianciája*

$$C_n(\xi, \eta) = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \overline{\xi}_n)(\eta_i - \overline{\eta}_n) = \frac{1}{n} \sum_{i=1}^{n} \xi_i \eta_i - \overline{\xi}_n \overline{\eta}_n.$$

Az $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ minta *empirikus kovarianciája*

$$C_n(\xi, \eta) = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \overline{\xi}_n)(\eta_i - \overline{\eta}_n) = \frac{1}{n} \sum_{i=1}^{n} \xi_i \eta_i - \overline{\xi}_n \overline{\eta}_n.$$

$$\mathbf{E}(C_n(\xi, \eta)) = \frac{n-1}{n} \mathbf{Cov}(\xi, \eta)$$

Tétel

Legyen $(\xi, \eta), (\xi_1, \eta_1), (\xi_2, \eta_2), \ldots$ *olyan statisztikai minta, melyre* $\mathbf{E}_\theta(\xi^4) < \infty$, $\mathbf{E}_\theta(\eta^4) < \infty$, *minden* $\theta \in \Theta$ *esetén. Ekkor* $C_n(\xi, \eta)$ *empirikus kovariancia gyengén konzisztens becslése a kovarianciának.*

Definíció

Az $\xi_1, \ldots, \xi_n$ minta *empirikus eloszlásfüggvénye*

$$F_n(x) = \frac{1}{n} \left| \{ i : \xi_i < x \} \right| .$$

Indikátorváltozókkal:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\xi_i < x).$$

Definíció

Az $\xi_1, \ldots, \xi_n$ minta *empirikus eloszlásfüggvénye*

$$F_n(x) = \frac{1}{n} \left| \{ i : \xi_i < x \} \right| .$$

Indikátorváltozókkal:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\xi_i < x).$$

A függetlenség miatt $\mathbb{I}(\xi_1 < x), \ldots, \mathbb{I}(\xi_n < x)$ független Bernoulli-eloszlású véletlen változók $p = F(x)$ paraméterrel.

Korábban beláttuk, hogy független, azonos paraméterű Bernoullik összege binomiális. Innen adódik a következő:

Állítás

Legyen $\xi_1, \xi_2, \ldots$ az F háttéreloszlásból származó minta, és tekintsük ennek az F_n empirikus eloszlásfüggvényét. Ekkor

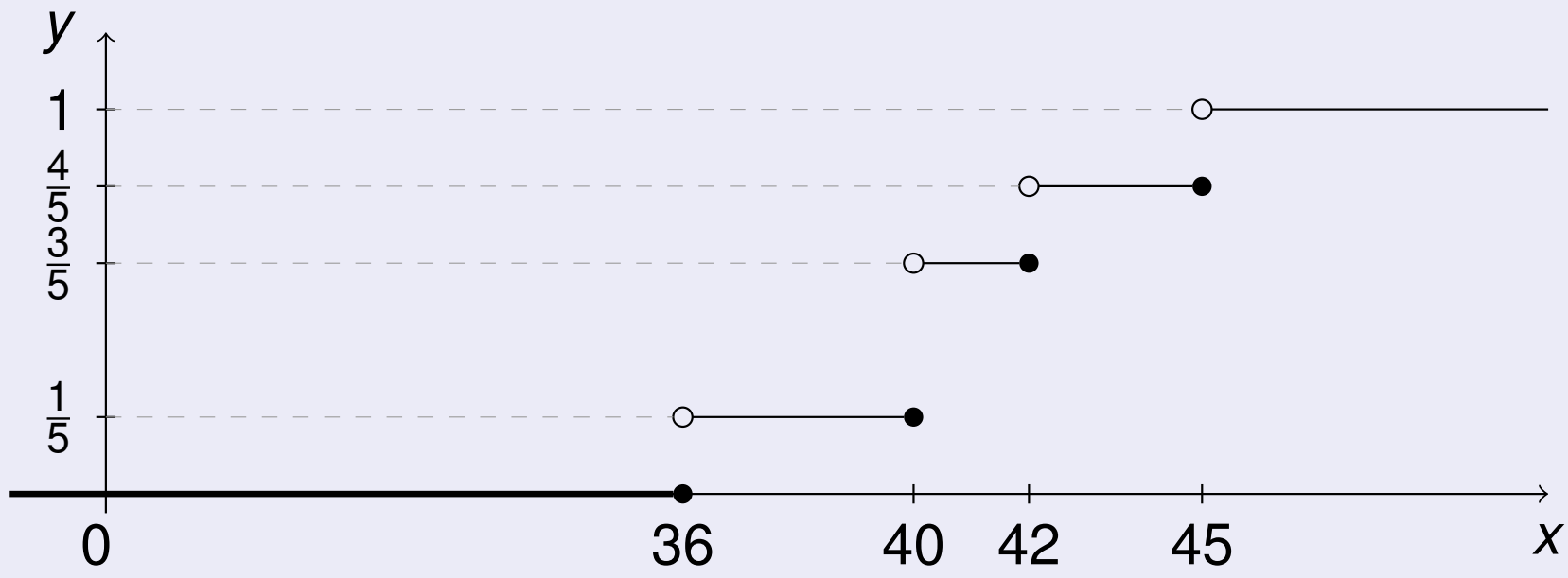
$$\mathbf{E}(F_n(x)) = F(x),\ \mathbf{D}^2(F_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

Továbbá, tetszőleges $x \in \mathbb{R}$ esetén, tetszőleges $\varepsilon > 0$ számra

$$\lim_{n\to\infty} \mathbf{P}(|F_n(x) - F(x)| > \varepsilon) = 0.$$

Példa

Az $x_1 = 40$, $x_2 = 45$, $x_3 = 40$, $x_4 = 42$, $x_5 = 36$ minta empirikus eloszlásfüggvénye:



Maximum likelihood módszer

Példa

Egy dobozban két pénzérme van. Az egyik szabályos, a másik cinkelt, 0,7 valószínűséggel ad fejet. Az egyik érmével 4-szer dobunk. Az eredmény 3 fej és 1 írás. Vajon melyik érmével dobtunk?

Maximum likelihood módszer

Tekintsünk egy $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mezőt, ahol $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$, ahol $\Theta \subset \mathbb{R}^k$ (általában egy-, néha kétdimenziós). Egy adott $\mathbf{x} = (x_1, \ldots, x_n)$ realizáció esetén azt a θ paramétert fogadjuk el, mely mellett a legnagyobb a valószínűsége az adott realizációnak.

Példa

Bernoulli-eloszlásból veszünk mintát, ahol a paraméter $\theta \in [0,1]$ ismeretlen, ezt akarjuk becsülni. $\xi_1, \ldots, \xi_n$ független Bernoulli(θ)-eloszlású. A függetlenség miatt

$$\mathbf{P}_\theta\left((\xi_1, \ldots, \xi_n) = (x_1, \ldots, x_n)\right) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Innen látjuk, hogy csak az számít a mintából, hogy hányszor következett be a vizsgált esemény. Tehát az $(x_1, \ldots, x_n)$ realizációhoz tartozó valószínűség, ha θ a valódi paraméter

$$L_\theta(x_1, \ldots, x_n) = \theta^{s_n}(1-\theta)^{n-s_n},$$

ahol $s_n = \sum_{i=1}^n x_i$.

$$L_\theta(x_1, \ldots, x_n) = \theta^{s_n}(1-\theta)^{n-s_n},$$

ahol $s_n = \sum_{i=1}^{n} x_i$.
Azt a θ értéket gondoljuk az igazi paraméternek, mely esetén az adott kimenetel a legvalószínűbb. Azaz a θ paraméter becslésére azt a $\widehat{\theta}$ értéket választjuk, melyre

$$L_{\hat{\theta}}(\mathbf{x}) = \sup_{\theta \in [0,1]} L_\theta(\mathbf{x}).$$

Adott s, n értékek mellett keressük a

$$L_\theta = \theta^s (1-\theta)^{n-s}$$

függvény maximumát a $\theta \in [0,1]$ intervallumon. Lederiválva

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L_\theta = \theta^{s-1}(1-\theta)^{n-s-1}(s - n\theta).$$

Látjuk, hogy a derivált pozitív a $[0, s/n)$ intervallumon, s/n helyen 0, és negatív az $(s/n, 1]$ intervallumon. Ezek szerint a

$$\widehat{\theta}(\mathbf{x}) = \frac{s_n}{n}$$

becslést kapjuk, ami éppen a relatív gyakoriság, vagy empirikus várható érték.

ML általában

Definíció

Ha a háttéreloszlás diszkrét, akkor a *likelihood-függvény*

$$L_\theta(\mathbf{x}) = L_\theta(x_1, \ldots, x_n) = \mathbf{P}_\theta(\xi = \mathbf{x}) = \prod_{j=1}^{n} \mathbf{P}_\theta(\xi = x_j),$$

ha folytonos, akkor a *likelihood-függvény*

$$L_\theta(\mathbf{x}) = L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_\theta(x_i).$$

A θ paraméter *maximum likelihood becslése* az $\mathbf{x}$ minta alapján

$$\widehat{\theta}(\mathbf{x}) = \operatorname{argmax}_{\theta \in \Theta} L_\theta(\mathbf{x}).$$

Log-likelihood

A szorzatalak miatt általában kényelmesebb a likelihood függvény logaritmusával számolni, amit log-likelihood függvénynek nevezünk, azaz

$$\ell_\theta(\mathbf{x}) = \log L_\theta(\mathbf{x}).$$

Mivel a logaritmus függvény szigorúan monoton nő, ezért L és ℓ maximumhelye megegyezik.

Hipergeometrikus eloszlás

Egy területen meg akarjuk becsülni az ott élő madarak ismeretlen N számát. Meggyűrűzünk M madarat, majd befogunk $n \leq M$ madarat, melyek közül m van meggyűrűzve. Tegyük föl, hogy $m \geq 1$, különben fogjunk még madarat. Megadjuk N ML becslését.

Hipergeometrikus eloszlás

Egy területen meg akarjuk becsülni az ott élő madarak ismeretlen N számát. Meggyűrűzünk M madarat, majd befogunk $n \leq M$ madarat, melyek közül m van meggyűrűzve. Tegyük föl, hogy $m \geq 1$, különben fogjunk még madarat. Megadjuk N ML becslését.
A likelihood függvény

$$L_N(m) = \mathbf{P}_N(\xi = m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}},$$

ahol ξ jelöli a másodszorra befogottak közül meggyűrűzöttek számát. A feladat meghatározni, hogy ez milyen N esetén lesz maximális. Ez az N lesz a ML becslés.

Egyszerű számolással

$$\frac{L_{N+1}(m)}{L_N(m)} = \frac{(N+1-m)(N+1-n)}{(N+1)(N+1-M-n+m)} > 1.$$

pontosan akkor teljesül, ha

$$N < \frac{nM}{m} - 1.$$

Ezek szerint az $(L_N(m))_{N \geq M}$ sorozat monoton nő $[nM/m]$-ig, ahol $[\cdot]$ az egészrészfüggvény. Ezek szerint N ML becslése

$$\widehat{N} = \left[\frac{nM}{m}\right].$$

Exponenciális eloszlás

Legyenek $\xi_1, \ldots, \xi_n$ független, Exp(λ) eloszlású véletlen változók. Adott $\mathbf{x} = (x_1, \ldots, x_n)$ realizáció esetén a log-likelihood függvény

$$\ell_\lambda(\mathbf{x}) = \log \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Innen a

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

likelihood egyenlet adódik, aminek megoldása

$$\widehat{\lambda} = \frac{1}{\overline{x}_n}.$$

Ez valóban maximumhely.

Egyenletes eloszlás

Legyenek $\xi_1, \ldots, \xi_n$ független (a, b)-n egyenletes eloszlású véletlen változók. Adott $\mathbf{x}$ realizáció esetén jelölje $x_{\min}$ és $x_{\max}$ a legkisebb és legnagyobb mintaelemet. Ekkor

$$L_{(a,b)}(\mathbf{x}) = \left(\frac{1}{b-a}\right)^n \mathbb{I}(a \leq x_{\min}, x_{\max} \leq b).$$

Ez akkor maximális (ne deriváljunk ész nélkül!!), amikor az indikátor 1, és $b - a$ a lehető legkisebb. Tehát a

$$(\widehat{a}, \widehat{b}) = (x_{\min}, x_{\max})$$

lesz az (a, b) ML becslése. Könnyen látható, hogy ez aszimptotikusan torzítatlan, konzisztens becslés.

Momentumok módszere

Tegyük fel, hogy $\theta = (\theta_1, \ldots, \theta_k)$, azaz $k \geq 1$ paraméterünk van. Általában $k = 1$, néha $k = 2$ (normális, egyenletes). Vezessük be a

$$m_j = \mathbf{E}_\theta(\xi^j) = g_j(\theta_1, \ldots, \theta_k), \quad j = 1, 2, \ldots, k,$$

jelölést. Tfh, az első k momentum egyértelműen meghatározza a paramétereket. Ekkor vannak olyan $h_1, \ldots, h_k$ függvények (m inverze), hogy

$$h_i(m_1, \ldots, m_k) = \theta_i, \quad i = 1, 2, \ldots, k.$$

Definíció

A $\theta = (\theta_1, \ldots, \theta_k)$ *momentum becslése* a

$$\widehat{\theta}_i = h_i(\widehat{m}_1, \ldots, \widehat{m}_k), \quad i = 1, 2, \ldots, k,$$

statisztika, ahol $\widehat{m}_i$ az empirikus i-edik momentum, azaz

$$\widehat{m}_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i, \quad i = 1, 2, \ldots, k.$$

A nagy számok gyenge törvénye szerint $\lim_{n\to\infty} \widehat{m}_i = m_i$. Innen pedig egyszerűen adódik, hogy $\widehat{\theta}_i$ konzisztens becslése θ_i-nek minden i-re.

Poisson-eloszlás

A Poisson-eloszlás paramétere megegyezik a várható értékkel, azaz $m_1 = \mathbf{E}_\lambda(\xi) = \lambda$, azaz $h_1(x) = x$, vagyis a λ paraméter momentum becslése

$$\widehat{\lambda} = \widehat{m}_1 = \overline{x}_n,$$

éppen a mintaátlag.

Exponenciális eloszlás

Az exponenciális eloszlás esetén azaz $m_1 = \mathbf{E}_\lambda(\xi) = \frac{1}{\lambda}$, azaz $h_1(x) = x^{-1}$, vagyis a λ paraméter momentum becslése

$$\widehat{\lambda} = \frac{1}{\widehat{m}_1} = \frac{1}{\overline{x}_n},$$

éppen a ML becslés.

Normális

Ha $\xi \sim \mathrm{N}(\mu, \sigma^2)$, akkor $m_1 = \mathbf{E}_{(\mu,\sigma^2)}(\xi) = \mu$ és $m_2 = \mathbf{E}_{(\mu,\sigma^2)}(\xi^2) = \sigma^2 + \mu^2$. Tehát

$$h_1(m_1, m_2) = m_1, \quad h_2(m_1, m_2) = m_2 - m_1^2.$$

Az empirikus momentumokat behelyettesítve

$$\widehat{\mu} = \widehat{m}_1 = \overline{x}_n$$
$$\widehat{\sigma^2} = \widehat{m}_2 - (\overline{x}_n)^2 = v_n.$$

Azaz a mintaátlag és az empirikus szórásnégyzet a megfelelő becslések.
A mintaátlag torzítatlan, az empirikus szórásnégyzet aszimptotikusan torzítatlan, mindkettő konzisztens.

Lineáris regresszió – motiváció

A következő modellt vizsgáljuk. Egy x ismert bemenethez tartozik egy y ismert kimenet. A két változó kapcsolatát $y = f(x) + \text{hiba}$ formula adja meg, ahol a hiba valamilyen értelemben kicsi, szimmetrikus. Feltesszük, hogy ez véletlen, várható értéke 0. A hiba eredhet a mérés pontatlanságából (mérési hiba), hozzáadott zajból.
Nem ismerjük az f függvényt, ezt akarjuk meghatározni. Ez a mesterséges intelligencia egyik alapfeladata, a *tanulás*.

Lineáris regresszió

A legegyszerűbb esetet vizsgáljuk, amikor $f(x) = ax + b$ lineáris függvény, ahol $a, b \in \mathbb{R}$ nem ismertek. Ekkor az a, b értékek meghatározása, becslése a feladat.
Adott az (x_i, y_i), $i = 1, 2, \ldots, n$ minta. Nem várunk pontos lineáris illeszkedést, nem lesz olyan a, b, hogy $y_i = ax_i + b$ minden $i = 1, 2, \ldots, n$ esetén teljesüljön. Négyzetes hibára minimalizálunk, azaz keressük azt az $(\hat{a}, \hat{b})$ párt, melyre

$$h(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

négyzetes hiba minimális.

Azaz egy kétváltozós függvény minimumhelyét keressük. Ez most egyszerű struktúrájú, hiszen a-ban, b-ben másodfokú. A $z_i = y_i - ax_i$ jelöléssel

$$\begin{aligned} h(a,b) &= \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \\ &= \sum (z_i - b)^2 = \sum z_i^2 - 2b \sum z_i + b^2 n \\ &= n(b - \overline{z}_n)^2 + \sum z_i^2 - n(\overline{z}_n)^2 \end{aligned}$$

adódik. Az első tag akkor minimális, ha $b = \overline{z}_n$.

Mivel $\overline{z}_n = \overline{y}_n - a\overline{x}_n$ és

$$\frac{1}{n}\sum x_i^2 - (\overline{x}_n)^2 = v_n,$$
$$\frac{1}{n}\sum x_i y_i - \overline{x}_n\overline{y}_n = c_n$$

így, a maradék ($z_i = y_i - ax_i$)

$$\begin{aligned}
&\sum z_i^2 - n(\overline{z}_n)^2 \\
&= a^2\left(\sum x_i^2 - n(\overline{x}_n)^2\right) - 2a\left(\sum x_i y_i - n\overline{x}_n\overline{y}_n\right) + \sum y_i^2 - n(\overline{y}_n)^2 \\
&= nv_n\left(a - \frac{c_n}{v_n}\right)^2 - n\frac{c_n^2}{v_n} + \sum y_i^2 - n(\overline{y}_n)^2.
\end{aligned}$$

Tehát $h(a, b)$ kifejezést négyzetek összegére bontottuk

$$h(a, b) = n(b - \overline{z}_n)^2 + nv_n \left(a - \frac{c_n}{v_n} \right)^2 + \text{ konstans},$$

ahol a konstans nem függ a, b értékétől.

Tehát $h(a,b)$ kifejezést négyzetek összegére bontottuk

$$h(a,b) = n(b - \overline{z}_n)^2 + nv_n \left(a - \frac{c_n}{v_n} \right)^2 + \text{ konstans},$$

ahol a konstans nem függ a, b értékétől.
Tehát $h(a,b)$ pontosan akkor minimális, ha a négyzetek 0-k, azaz

$$b = \overline{z}_n = \overline{y}_n - a\overline{x}_n.$$

és

$$a = \frac{c_n}{v_n} = \frac{\sum_{i=1}^n (y_i - \overline{y}_n)(x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}.$$

Ez a *legkisebb négyzetek módszere*.
Vegyük észre, hogy pontosan a korábban megkapott elméleti értékek empirikus változatait kaptuk.

Konfidenciaintervallumok

Definíció

A $(T_1(\xi), T_2(\xi))$ statisztikapárral definiált intervallum *legalább* $1-\varepsilon$ *megbízhatósági szintű konfidenciaintervallum a* $\psi(\theta)$ *paraméterre*, ha

$$\mathbf{P}_\theta(T_1(\xi) < \psi(\theta) < T_2(\xi)) \geq 1-\varepsilon, \quad \forall \theta \in \Theta,$$

ahol $\varepsilon > 0$ előre adott kicsi szám.
Amennyiben a fönti $\geq$ helyett $=$ szerepel, akkor $(T_1(\xi), T_2(\xi))$ *pontosan* $1-\varepsilon$ *megbízhatósági szintű konfidenciaintervallum*.

Az $1-\varepsilon$ érték a megbízhatósági szint, általában 95% vagy 99%.

Konfidenciaintervallum normális eloszlás várható értékére ismert szórás esetén

Legyenek $\xi_1, \ldots, \xi_n \sim \mathrm{N}(\mu, \sigma_0^2)$ független véletlen változók, ahol μ az ismeretlen paraméter, σ_0^2 ismert. Megadunk μ-re pontosan $1 - \varepsilon$ megbízhatósági szintű konfidenciaintervallumot.
Mivel $\overline{\xi}$ torzítatlan, erősen konzisztens becslés μ-re, ezért az intervallumot $(\overline{\xi} - r_\varepsilon, \overline{\xi} + r_\varepsilon)$ alakban keressük.
Szükségünk lesz az alábbi állításra, melyet nem bizonyítunk.

Tétel

Ha ξ, η független normális eloszlású véletlen változók akkor $\xi + \eta$ is normális eloszlású.

Ebből indukcióval az is következik, hogy független normálisok összege normális, és persze a várható értékek összeadódnak (azok mindig), és a szórásnégyzetek is összeadódnak (hiszen függetlenek az összeadandók). Ezért $\overline{\xi} \sim \mathrm{N}(\mu, \sigma_0^2/n)$. Tehát

$$\begin{aligned}
\mathbf{P}_\mu(\overline{\xi} - r_\varepsilon < \mu < \overline{\xi} + r_\varepsilon) &= \mathbf{P}_\mu\left(-\frac{r_\varepsilon}{\sigma_0/\sqrt{n}} < \frac{\overline{\xi}-\mu}{\sigma_0/\sqrt{n}} < \frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) \\
&= \Phi\left(\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) - \Phi\left(-\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) \\
&= 2\Phi\left(\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) - 1 = 1 - \varepsilon,
\end{aligned}$$

azaz

$$\Phi\left(\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) = 1 - \frac{\varepsilon}{2}.$$

Tehát az

$$u_{\varepsilon/2} = \Phi^{-1}(1 - \varepsilon/2)$$

jelöléssel, ahol Φ^{-1} a Φ függvény inverze, a keresett $1 - \varepsilon$ megbízhatósági szintű konfidenciaintervallum

$$\left(\overline{\xi} - \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}, \overline{\xi} + \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}} \right).$$

Vegyük észre, hogy a megbízhatósági szint növelésével, azaz ε csökkenésével, az intervallum hossza nő, a mintaelemszám növelésével pedig csökken.