A sztochasztika alapjai
MBNXK262

10. előadás: CHT; statisztika

Kevei Péter

2023/24 tavasz

Ismétlés

Tétel (Centrális határeloszlás-tétel)

$\xi, \xi_1, \xi_2, \ldots$ *független, azonos eloszlású vv,* $\mathbf{E}(\xi) = \mu$, $\mathbf{D}(\xi) = \sigma$

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\sum_{i=1}^{n}(\xi_i - \mu)}{\sqrt{n}\sigma} < x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,\mathrm{d}y.$$

Tétel (de Moivre–Laplace tétel)

S_n: egy p valószínűségű A esemény bekövetkezéseinek a száma n kísérlet során

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} < x\right) = \Phi(x).$$

Példa

Budapesten meg akarják állapítani a dohányosok p arányát. Ehhez kiválasztanak n egyént úgy, hogy minden választásnál mindenki ugyanakkora valószínűséggel kerül kiválasztásra, és csak ezek közt nézik meg a dohányosok k számát. Legalább mekkora legyen az n, hogy a kapott $p' = k/n$ arány legalább 0,95 valószínűséggel legfeljebb 0,005 hibával közelítse a valódi p arányt, akármi is $p \in (0, 1)$?

Statisztikai alapfogalmak

Független, azonos eloszlású véletlen változók egy $\xi, \xi_1, \xi_2, \ldots$ sorozatát *statisztikai mintának* nevezzük. Közös (ismeretlen) eloszlást *háttéreloszlás*. A minta egy adott realizációját $x_1, \ldots, x_n$ jelöli. A minta egy T függvényét *statisztikának* nevezzük.

Alapstatisztikák

$\xi_1, \xi_2, \ldots, \xi_n$ egy n-elemű minta.

Definíció

Az

$$\overline{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

a *mintaátlag*.

Torzítatlanság, konzisztencia

$\overline{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i$

$\mathbf{E}\left(\overline{\xi}_n\right) =$

Torzítatlanság, konzisztencia

$\overline{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i$

$\mathbf{E}\left(\overline{\xi}_n\right) =$

$$\mathbf{P}\left(|\overline{\xi}_n - \mu| > \varepsilon\right) \to 0$$

Az *empirikus szórásnégyzet*

$$V_n(\xi) = V_n = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2.$$

Az *empirikus szórásnégyzet*

$$V_n(\xi) = V_n = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2.$$

Tétel (Steiner-formula)

Tetszőleges $x_1, \ldots, x_n$ értékekre és c valós számra

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + (\overline{x}_n - c)^2.$$

A $c = 0$ választással

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i^2 - (\overline{\xi}_n)^2.$$

A $c = 0$ választással

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i^2 - (\overline{\xi}_n)^2.$$

$$\mathbf{E}(V_n) =$$

A $c = 0$ választással

$$V_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i^2 - (\overline{\xi}_n)^2.$$

$$\mathbf{E}(V_n) =$$

Korrigált empirikus szórásnégyzet

$$V_n^*(\xi) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2 = \frac{n}{n-1}V_n(\xi)$$

torzítatlan becslése a szórásnégyzetnek.

Tétel

Ha $\mathbf{E}(\xi^4) < \infty$*, akkor mind az empirikus szórásnégyzet, mind a korrigált empirikus szórásnégyzet gyengén konzisztens becslése a szórásnégyzetnek.*

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

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

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 a*

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

empirikus kovariancia gyengén konzisztens becslése a kovarianciának.

Jelölés

A konkrét realizációból számolt értékeket a megfelelő kisbetűkkel jelöljük, így konkrét $x_1, \ldots, x_n$ realizációhoz tartozó mintaátlag és empirikus szórásnégyzet

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad v_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_i)^2.$$

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.

Állítás

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

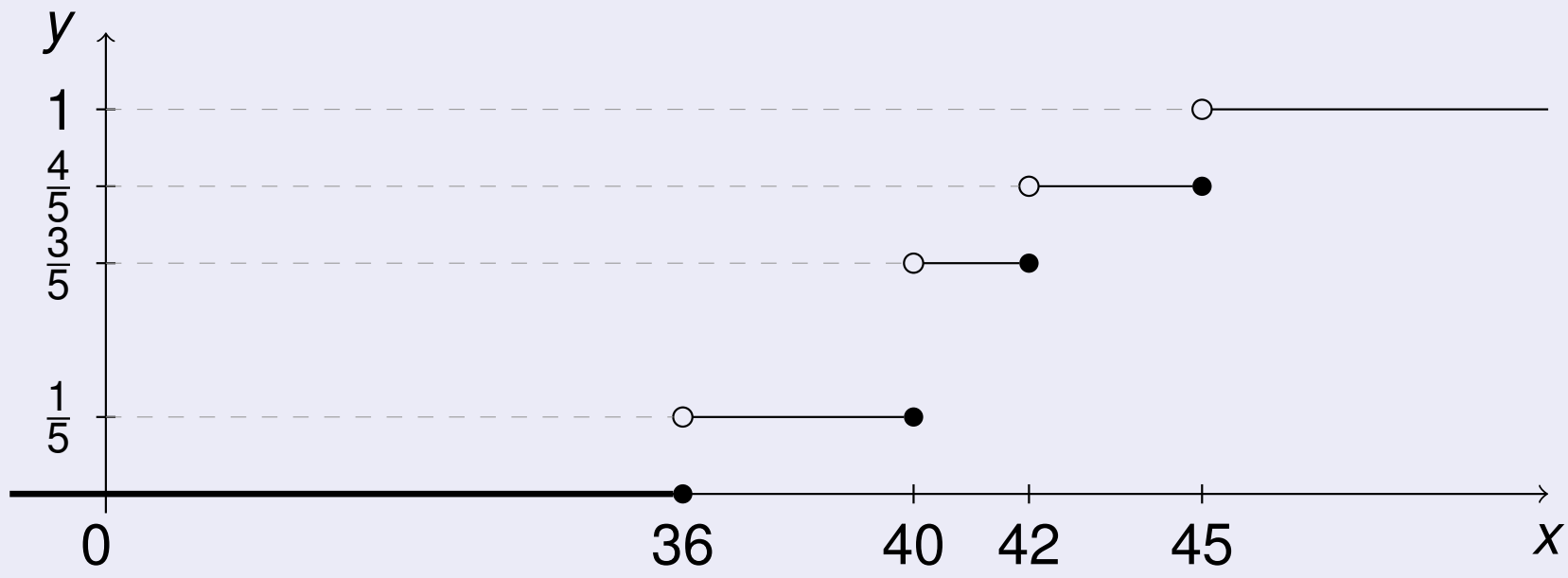
$$\mathbf{E}(F_n(x)) = F(x), \ \mathbf{D}^2(F_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

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

$\xi_1, \ldots, \xi_n$ független Bernoulli(θ)-eloszlású.

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

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.

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

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