A sztochasztika alapjai
MBNXK262

9. előadás: Nagy számok törvényei, centrális határeloszlás-tétel; Statisztika

Kevei Péter

2023/24 tavasz

Egyenlőtlenségek

Tétel (Markov-egyenlőtlenség)

Legyen ξ *egy véletlen változó* $(\Omega, \mathcal{A}, \mathbf{P})$ *valószínűségi mezőn, melynek véges a várható értéke. Ekkor tetszőleges pozitív c konstansra*

$$\mathbf{P}(|\xi| \geq c) \leq \frac{\mathbf{E}(|\xi|)}{c}.$$

Egyenlőtlenségek

Tétel (Csebisev-egyenlőtlenség)

Legyen ξ *egy véletlen változó* $(\Omega, \mathcal{A}, \mathbf{P})$ *valószínűségi mezőn, melynek véges a szórása. Ekkor tetszőleges pozitív c konstansra*

$$\mathbf{P}(|\xi - \mathbf{E}(\xi)| \geq c) \leq \frac{\mathbf{D}^2(\xi)}{c^2}.$$

Nagy számok törvénye

Tétel (Csebisev-féle nagy számok gyenge törvénye)

Legyenek $\xi_1, \xi_2, \ldots$ páronként független, véges szórású véletlen változók, melyek közös várható értéke μ és szórásnégyzete σ^2. Ekkor tetszőleges $\varepsilon > 0$ esetén

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{\xi_1 + \ldots + \xi_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

Tétel (Bernoulli-féle nagy számok gyenge törvénye (1713))

Jelölje S_n egy p valószínűségű A esemény bekövetkezéseinek a számát egy kísérlet n független ismétlése során. Ekkor tetszőleges $\varepsilon > 0$ esetén

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) = 0.$$

CHT

Tétel (Centrális határeloszlás-tétel)

Legyenek $\xi, \xi_1, \xi_2, \ldots$ független, azonos eloszlású véletlen változók közös $\mathbf{E}(\xi) = \mu$ várható értékkel, és véges $\mathbf{D}(\xi) = \sigma$ szórással. Ekkor tetszőleges $x \in \mathbb{R}$ esetén

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\sum_{i=1}^{n}(\xi_i - \mu)}{\sqrt{n}\sigma} < x\right) = \Phi(x),$$

ahol

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} \,\mathrm{d}y$$

a standard normális eloszlás eloszlásfüggvénye.

Tétel (de Moivre–Laplace tétel)

Jelölje S_n egy p valószínűségű A esemény bekövetkezéseinek a számát egy kísérlet n független ismétlése során. Ekkor tetszőleges $x \in \mathbb{R}$ esetén

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

Példa

A kakucsretyegei polgármesterválasztáson két jelölt van: A és B. Kakucsretyege 40000 szavazója egymástól függetlenül, $1/2 - 1/2$ valószínűséggel szavaz a két jelölt egyikére. A feszült politikai helyzet miatt a szavazatok újraszámlálását rendelik el, ha a két jelöltre leadott szavazatok száma között legfeljebb 100 a különbség. Mi a valószínűsége, hogy újraszámlálásra kerül sor?

Példa

A sztochasztika alapjai kurzus 700 hallgatójának mindegyike bemegy az első előadásra. Ezt követően minden hallgató minden további előadás előtt feldob egy szabályos pénzérmét. Ha fejet kap, akkor bemegy a következő előadásra, ha írást akkor nem, és utána már egyetlen előadásra sem megy be. Véletlen Vince a 700 hallgató egyike. Mennyi a valószínűsége, hogy Vince az összes előadásra bemegy? Várhatóan hány előadáson vesz részt? Mennyi a valószínűsége, hogy a 13. (utolsó) előadáson lesz hallgató? Várhatóan hány hallgató lesz a 2., 3., utolsó előadáson? Mennyi a valószínűsége, hogy a negyedik előadáson 100-nál több hallgató vesz részt. És annak, hogy a tizenegyediken pontosan 2 hallgató vesz részt?

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

$\xi_1, \xi_2, \dots, \xi_n$ 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

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

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

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

Definíció

Az *empirikus szórásnégyzet*

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

Definíció

Az *empirikus szórásnégyzet*

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

Tétel (Steiner-tétel)

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*

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

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

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

$$\mathbf{E}(C_n(\xi, \eta) =$$

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 $\xi_1, \dots, \xi_n$ 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 $\mathbb{I}(\xi_1 < x), \dots, \mathbb{I}(\xi_n < x)$ 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: