1. előadás - Valószínűségszámítási alapismeretek

Sztochasztikus rendszerek matematikája

2016. szeptember 5.

Tartalom

1. előadás:

- Technikai kérdések a tárggyal kapcsolatban.
- A félév szakmai anyaga címszavakban, tematika.
- Valószínűségszámítási alapismeretek ismétlése.
- Statisztikai alapok ismétlése - becsléselmélet.

Követelmények

- A tárgyból írásbeli vizsga lesz. Félév közben 3 **kötelező házi feladat** lesz (megfelelt/nem felelt meg), amelyek az aláírás megszerzéséhez kellenek, a vizsgajegy kialakításában nem játszanak szerepet. Aláírást az a hallgató kaphat, akinek legalább két házi feladata megfelelt minősítésű.
- Osztályozás a szokásos módon.
- A tárgykövetelmények külön fájlban elérhetőek a honlapon.

Adminisztratív ügyek

- www.math.bme.hu/∼ftamas
- Az előadás slidejai képezik a tananyag törzsét, ezek fent lesznek a honlapon.
- Ajánlott szakirodalom:
  - S. Karlin, H. Taylor: A first course in stochastic processes, Academic Press, 2003 (Van magyar nyelvű fordítás a könyvtárban.)
  - Bognár Jánosné, Göndőcs F., Kászonyi L., Kováts A., Michaletzky Gy., Somogyi Á., Székely J. G.:Matematikai Statisztika, Tankönyvkiadó, 1986. (egyetemi jegyzet)
  - Ramu Ramanathan: Bevezetés az ökonometriába alkalmazásokkal, Panem Kiadó, 2003.
  - L. Györfi, L. Győri, M. Pintér: Tömegkiszolgálás informatikai rendszerekben, Műegyetemi Kiadó, 2005.

Gyakorlatok tudnivalói

- Minden páros héten lesznek gyakorlatok alkalmanként két órában, a páratlan heteken csak előadás lesz.
- Gyakorlatok helyszíne a H épület 207-es laborja.
- Gyakorlatok között nincs átjárás, mivel a termek befogadóképessége korlátos.

Az előadások tematikája

- Valószínűségszámítási és statisztikai alapismeretek ismétlése
- Többváltozós statisztika
- Lineáris regresszió
- Idősorelemzés
- Markov láncok, Poisson folyamat
- Tömegkiszolgálási modellek

Valószínűségszámítási alapok

Kolmogorov-féle valószínűségi mező: $(\Omega, \mathcal{A}, P)$, ahol

- Ω: valószínűségi tér (mintatér), $\omega \in \Omega$ elemekkel (elemi események)
- $\mathcal{A}$: σ-algebra
- P: valószínűségi mérték $\mathcal{A}$-n

Valószínűségi változó (v.v.): olyan $X : \Omega \to \mathbb{R}$ függvény, melyre minden B Borel-halmaz esetén $\{\omega : X(\omega) \in B\} \in \mathcal{A}$.

Diszkrét valószínűségi változó: X értékkészlete megszámlálható halmaz, és ekkor az eloszlás olyan (x_i, p_i) párok sorozata, ahol x_i az X értékkészletének egy felsorolása, $p_i = P(X = x_i)$.

Valószínűségi vektorváltozó (v.v.v.): p darab v.v. vektora egy p-dimenziós valószínűségi vektorváltozó.

Valószínűségszámítási alapok

Eloszlásfüggvény: $\underline{X}$ p-dimenziós v.v.v. esetén

$$F_{\underline{X}}(\underline{x}) = F_{X_1,\dots,X_p}(x_1,\dots,x_p) = P(\omega | X_1(\omega) < x_1, \dots, X_p(\omega) < x_p).$$

Sűrűségfüggvény: ha $F(x)$ abszolút folytonos a Lebesgue-mértékre nézve, akkor létezik, és

$$f(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, \quad f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1.$$

Várható érték: diszkrét v.v. esetén

$$\mathrm{E}X = \sum_i x_i p_i,$$

ha a sor abszolút konvergens, míg abszolút folytonos esetben

$$\mathrm{E}X = \int_\Omega X(\omega)\,\mathrm{d}P(\omega) = \int_{\mathbb{R}} x\ \mathrm{d}F(x) = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x.$$

Valószínűségszámítási alapok

Momentumok: $\mathrm{E}(X^k)$ a k-dik momentum, $\mu_k = \mathrm{E}(X - \mathrm{E}X)^k$, $k \in \mathbb{N}$, a k-adik centrális momentum. Speciálisan

- $\mu_2 = \sigma^2 = Var(X) = D^2(X)$ a szórásnégyzet
- μ_3/σ^3 az aszimmetria mértéke (skewness)
- μ_4/σ^4 a csúcsosság és farok-eloszlás mértéke (kurtosis)

Tétel

A magasabb momentumok létezéséből következik az alacsonyabb momentumok létezése, a megfordítás viszont nem igaz.

Függetlenség: X és Y függetlenek, ha $\forall A, B$ Borel-halmazra

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Speciálisan

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Valószínűségszámítási alapok

Ekkor $\mathrm{E}(XY) = \mathrm{E}X \cdot \mathrm{E}Y$, és $Var(X+Y) = Var(X) + Var(Y)$.

Feltételes valószínűség:

$$P(A|B) = \frac{P(AB)}{P(B)}, \quad \text{ha} \quad P(B) \neq 0.$$

Kovariancia, korreláció:

$$Cov(X,Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)].$$

Speciálisan $Cov(X,X) = Var(X)$. A két változó közötti kapcsolat erősségét méri, de számszerű értéke függ a változók értékeinek nagyságrendjétől. A kovariancia szimmetrikus, bilineáris függvény.

Valószínűségszámítási alapok

A nagyságrendtől függést küszöböli ki a **korrelációs együttható:**

$$R(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$|R(X,Y)| \leq 1$, és a **két változó közötti lineáris kapcsolat erősségét méri:**

- ha $R \approx \pm 1$, akkor majdnem lineáris a kapcsolat a változók közt.
- ha $R \approx 0$, akkor nemlineáris a kapcsolat a változók közt.

Tétel

Független v.v-k korrelálatlanok (azaz $R(X,Y) = 0$), de ez visszafelé nem igaz.

Nevezetes eloszlások - diszkrét

(n, p) **paraméterű binomiális eloszlás:** lehetséges értékei a $k = 1, 2, \ldots, n$ számok, eloszlása pedig

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

λ **paraméterű Poisson eloszlás:** lehetséges értékei a természetes számok, eloszlása pedig

$$p_k = P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Nevezetes eloszlások - abszolút folytonos

λ paraméterű exponenciális eloszlás: sűrűségfüggvénye

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

Standard normális eloszlás: sűrűségfüggvénye

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

(m, σ) paraméterű normális eloszlás: sűrűségfüggvénye

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

NSZT és CHT

Tétel (Nagy számok Bernoulli-féle gyenge törvénye)

Ha ξ_i, $i \in \mathbb{N}$ *független, azonos eloszlású v.v-k olyan sorozata, melyre* $D^2\xi = \sigma^2 < \infty$, *akkor* $\forall \varepsilon \geq 0$ *esetén*

$$P\left(\left|\frac{\sum_{i=1}^n \xi_i}{n} - \mathrm{E}\xi_1\right| \geq \varepsilon\right) \to 0, \quad n \to \infty.$$

Tétel

Az előző tétel feltételei mellett $\forall x \in \mathbb{R}$ *esetén*

$$P\left(\frac{\sum_{i=1}^n \xi_i - nm}{\sqrt{n}\sigma} < x\right) \to P(\eta < x), \quad n \to \infty,$$

ahol $m = \mathrm{E}\xi_1$, $\eta \sim N(0,1)$. *Azaz* $\sum \xi_i \approx N(nm, \sqrt{n}\sigma)$.

Néhány további nevezetes eloszlás

n-edrendű, λ paraméterű Gamma-eloszlás: n db független $exp(\lambda)$ eloszlású v.v. összegének eloszlása.

n szabadságfokú χ^2 eloszlás: n darab független standard normális v.v. négyzeteinek összege.

n szabadságfokú t-eloszlás (Student eloszlás):

$$t = \frac{\sqrt{n} \cdot \eta}{\sqrt{\xi_1^2 + \ldots + \xi_n^2}},$$

ahol $\eta, \xi_1, \ldots, \xi_n \sim N(0,1)$ függetlenek.

n, m szabadságfokú F-eloszlás:

$$t = \frac{\frac{1}{n}(\xi_1^2 + \ldots + \xi_n^2)}{\frac{1}{m}(\eta_1^2 + \ldots + \eta_m^2)},$$

$\eta_1, \ldots, \eta_m \sim N(0,1)$ függetlenek, $\xi_1, \ldots, \xi_n \sim N(0,1)$ függetlenek.

Mi a statisztika?

- "A statisztika a matematika azon ága, melynek alapfeladata az, hogy a politikus kezébe olyan eszközt adjon, mellyel tetszőleges állítás és annak ellentéte is tudományos alapon igazolható." (ismeretlen forrás)
- A statisztika a világ számszerűsíthető tényeinek szisztematikus összegyűjtésével és elemzésével foglalkozó tudományos módszer és gyakorlat.

Feladat, cél: a tapasztalati adatokból az információk kinyerése, statisztikai törvényszerűségek feltárása, következtetések levonása és felhasználása. Modellépítés, paraméterbecslés, következtetések, hipotézisek vizsgálata.

Alapfogalmak

Kiindulópont: $(\Omega, \mathcal{F}, \mathcal{P})$ statisztikai mező, ahol a $\mathcal{P}$ mértékcsalád olyan P eloszlásokból áll, melyekkel $(\Omega, \mathcal{F}, P)$ valószínűségi mező. A probléma éppen a megfelelő eloszlás kiválasztása valamilyen $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ paraméteres családból.

Vizsgálódás tárgya: X v.v., pl. testmagasság, melynek pontos P eloszlását nem ismerjük, csak annyit tudunk, hogy $P \in \mathcal{P}$.

Cél: paraméterbecslés, hipotézis-vizsgálat

Eszköz: a statisztikai minta, mely valamilyen véletlen mennyiségre vonatkozó véges számú független megfigyelés (lehetséges) eredménye, azaz véges sok független, azonos eloszlású v.v.

Realizáció: egy konkrét kimenetel, $T(X) = T(X_1, \ldots, X_n)$ a statisztika, ahol T mérhető függvény.

Alapfogalmak

Alapstatisztikák: kiinduló tájékozódás az $X_1, \ldots, X_n$ mintáról

- Mintaátlag: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Tapasztalati szórásnégyzet: $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- Korrigált tapasztalati szórásnégyzet: $S_n^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- Mintaátlag standardizált hibája: $\frac{\overline{X} \cdot \sqrt{n}}{S_n^*}$
- k. tapasztalati centrális momentum: $M_k^c = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^k$

Alapfogalmak

- Ferdeség (szimmetria): $\dfrac{M_3^c}{(M_2^c)^{3/2}}$
- Lapultság: $\dfrac{M_4^c}{(M_2^c)^2} - 3$
- Tapasztalati kovariancia: (X_i, Y_i), $i = 1, \ldots, n$, 2-dim i.i.d. (független és azonos eloszlású) minta,

$$C = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X} \cdot \overline{Y}$$

- Tapasztalati korrelációs együttható:

$$R = \frac{C}{S_X S_Y},$$

ahol S_X és S_Y a komponensek tapasztalati szórásai.

Becsléselmélet

Feladat: a θ paramétert vagy annak valamilyen $\Psi(\theta)$ függvényét szeretnénk becsülni az $X = (X_1, \ldots, X_n)$ i.i.d. minta alapján konstruált $\hat{\theta}_n = T_n(X)$ (avagy $\widehat{\Psi(\theta_n)} = T_n(X)$) statisztika segítségével. Ő lesz majd a paraméter, avagy megfelelő függvényének becslése.

Az olyan becsléseket, amelyek megadott paraméterek becslését adják, paraméter- vagy pontbecsléseknek nevezzük.

A becslés jóságának mérése: a valódi paraméter körüli ingadozás és a sztochasztikus konvergencia segítségével történik.

Alapfogalmak röviden

Torzítatlanság: $T_n(X)$ torzítatlan becslés $\Psi(\theta)$-ra, ha

$$\mathrm{E}_\theta(T_n(X)) = \Psi(\theta), \quad \forall \theta \in \Theta.$$

Pl.: az átlag mindig torzítatlan becslése a várható értéknek, ha ez véges.

Aszimpotitkus torzítatlanság: a $T_n(X)$ statisztika-sorozat aszimptotikusan torzítatlan becslése $\Psi(\theta)$-nak, ha

$$\lim_{n\to\infty} \mathrm{E}_\theta(T_n(X)) = \Psi(\theta), \quad \forall \theta \in \Theta.$$

Pl.: szórásnégyzetre a tapasztalati szórásnégyzet. (A korrigált változat torzítatlan is!)

Alapfogalmak röviden

Konzisztencia: $T_n(X)$ konzisztens becslés $\Psi(\theta)$-ra, ha $T_n(X) \to \Psi(\theta)$, $n \to \infty$ sztochasztikusan, azaz bármely $\varepsilon > 0$ esetén

$$P(|T_n(X) - \Psi(\theta)| > \varepsilon) \to 0, \quad n \to \infty, \quad \forall \theta \in \Theta.$$

Erős konzisztencia: $T_n(X)$ erősen konzisztens becslés $\Psi(\theta)$-ra, ha konzisztens és szórása nullához tart $n \to \infty$ esetén.

Pl.: $\overline{X}$ konzisztens és erősen is konzisztens becslése a várható értéknek, és ez persze nem más, mint a nagy számok erős és gyenge törvénye.

Hatásosság: Ha T_1 és T_2 torzítatlan becslések $\Psi(\theta)$-ra, akkor T_1 hatásosabb, mint T_2, ha

$$D^2_\theta(T_1) \leq D^2_\theta(T_2), \quad \forall \theta \in \Theta.$$

Egy becslés hatásos, ha minden más becslésnél hatásosabb.

Becslési elvek

Maximum-likelihood becslés: az ismeretlen $\Psi(\theta)$ paraméter becsléseként azt a $\Psi(\hat{\theta}_n)$ értéket vesszük, amely mellett az $x_1, \ldots, x_n$ minta valószínűsége a legnagyobb.

- Diszkrét esetben az

$$L_\theta(x_1, \ldots, x_n) = p_\theta(x_1) \cdot \ldots \cdot p_\theta(x_n) = \prod_{i=1}^{n} P(X = x_i)$$

likelihood-függvény maximumát keressük.

- Folytonos esetben az

$$L_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \cdot \ldots \cdot f_\theta(x_n) = \prod_{i=1}^{n} f(x_i)$$

likelihood-függvény maximumát keressük, ahol f a megfelelő sűrűségfüggvény.

Becslési elvek

Tétel

Ha $\hat{\theta}$ a θ paraméter maximum-likelihood becslése, akkor a $\Psi(\theta)$ maximum-likelihood becslése éppen $\widehat{\Psi(\theta_n)}$.

Amivel most nem foglalkozunk:

- momentumok módszere - a paraméter kiszámítása elméleti momentumokkal, majd becslése a tapasztalati momentumokkal.
- legkisebb négyzetek módszere - ezt majd a regressziónál tárgyaljuk részletesen

Információs határ

Legyen paraméterünk egydimenziós, és $L(x, \theta)$ a likelihood-függvény!

Definíció

Az $X_1, \ldots, X_n$ statisztikai minta Fisher-féle információjának az

$$I_n(\theta) = \mathrm{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log L(X, \theta) \right)^2 \right] \geq 0$$

mennyiséget hívjuk, ha derivált létezik és a várható érték véges.

Tétel

Ha a fenti becslés torzítatlan és $D^2_\theta(T) < \infty$ *minden* $\theta \in \Theta$ *esetén, akkor*

$$D^2_\theta(T) \geq \frac{(\Psi'(\theta))^2}{I_n(\theta)}.$$

Konfidencia-intervallumok

A pontbecslések helyett térjünk most át az intervallum-becslésekre.

Legyen $X_1, \ldots, X_n$ statisztikai minta a P_θ eloszláscsaládból. A θ paraméterhez olyan (T_1, T_2) véletlen hosszúságú intervallumot keresünk, melyre

$$P(T_1 \leq \theta < T_2) \geq 1 - \varepsilon,$$

ahol $\varepsilon > 0$ kicsi. Itt T_1 és T_2 maguk is v.v.-k, a minta valamilyen függvényei. Ezt az intervallumot a θ paraméterre vonatkozó legalább $1 - \varepsilon$ megbízhatósági szintű konfidencia-intervallumnak nevezzük.