2. előadás - Statisztikai alapismeretek

Sztochasztikus rendszerek matematikája

2016. szeptember 12.

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 $\Psi(\hat{\theta})$.

Amivel most nem foglalkozunk:

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

Információs határ

Definíció

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

Alapfeladat

Tegyük fel, hogy adott egy $\xi_1, \ldots, \xi_n$ független minta - mérési eredmények, megfigyelések. Ezek alapján dönteni akarunk különböző kérdésekben:

- A minta egy adott eloszlás(család)ból származik-e? (Pl. villamosok közötti idő)
- A közös várható érték megegyezik-e egy előírt mennyiséggel? (Pl. tablettában lévő hatóanyag mennyisége)
- Szignifikánsan eltér-e a közös várható érték az előírtnál, és ha igen, akkor milyen irányban? Kevesebb avagy több?
- Adott genetikai minta származhat-e egy bizonyos személytől?

Alapfeladat

A statisztikai hipotézisek vizsgálata abból indul ki, hogy adott $P_\theta, \theta \in \Theta$ eloszlásra, vagy annak valamely paraméterére egy megadott állítás érvényes-e vagy sem.

Ezt a feltételezést nullhipotézisnek nevezzük:

$$\mathrm{H}_0 : \theta \in \Theta_0$$

Pl. a gyógyszeres példában valóban az előírt mennyiség stimmel $\mathrm{E}\xi = \mu$.

Az ellenhipotézis a nullhipotézis (valamilyen értelemben vett)tagadása, azaz

$$\mathrm{H}_1 : \theta \in \Theta_1,$$

ahol $(\Theta_0 \cup \Theta_1 = \Theta)$. Pl. a fenti példában $\mathrm{E}\xi \neq \mu_0$.

A statisztikai próba és a döntési eljárás

Azt az eljárást, ami alapján döntünk, statisztikai próbának nevezzük.

Jelölje tehát ξ a $(\xi_1, \ldots, \xi_n)$ minta értékkészletét, és legyen $f(\xi_1, \ldots, \xi_n)$ a próbastatisztikánk. Ésszerűen választunk két halmazt:

- E az elfogadási tartomány
- K a kritikus tartomány

Világos, hogy $E \cup K = \mathbb{R}$ és $E \cap K = \emptyset$.

- $f(\xi_1, \ldots, \xi_n) \in E$ esetén elfogadjuk H_0-t,
- $f(\xi_1, \ldots, \xi_n) \in K$ esetén elutasítjuk H_0-t (azaz H_1-et fogadjuk el)

A statisztikai próba és a döntési eljárás

A próba alapján kétféle módon követhetünk el hibát:

	H_0-t elfogadjuk	*H_0-t elutasítjuk*
H_0 fennáll	helyes a döntés	elsőfajú hiba
H_0 nem áll fenn	másodfajú hiba	helyes a döntés

Cél a gyakorlatban: az elsőfajú hiba "kordában" tartása, és közben a másodfajú hiba minimalizálása, amennyire csak lehet.

Azaz lerögzítjük az elsőfajú hiba nagyságát, és olyan statisztikai próbát keresünk, melynél az adott elsőfajú hibanagyság mellett a másodfajú hiba a lehető legkisebb. Azaz

$$P(\mathrm{H}_0 \text{ igaz, de } f(\xi_1, \ldots, \xi_n) \in K) = \alpha,$$

ahol α az adott elsőfajú hiba nagysága.

Paraméteres és nemparaméteres próbák

- **Paraméteres próba:** a vizsgált változók eloszlásfüggvényeit véges sok paraméter egyértelműen meghatározza (pl. normális eloszlású)
- **Nemparaméteres próba:** az egyes eloszlásfüggvények nem azonosíthatók egy, vagy több szám együttesével (pl. ha csak annyit tudunk, hogy az eloszlás folytonos)

u-próbák családja (4 eset)

Egymintás kétoldali u-próba: cél egy normális eloszlású v.v ismeretlen μ várható értékére vonatkozó hipotézis tesztelése ismert σ_0 szórás mellett. Azaz

$$\mathrm{H}_0 : \mu = \mu_0, \quad \mathrm{H}_1 : \mu \neq \mu_0$$

A próbastatisztika

$$u := f(\xi_1, \ldots, \xi_n) = \frac{\bar{\xi}_n - \mu_0}{\sigma_0/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Adott α elsőfajú hiba mellett az elfogadási tartomány

$$E = [-a(\alpha), a(\alpha)]$$

alakú, ahol a-t úgy választjuk meg, hogy

$$P(|u| > a) = 2(1 - \phi(a)) = \alpha$$

teljesüljön. (Lásd standard normális eloszlás táblázata.)

u-próbák családja (4 eset)

Egymintás egyoldali u-próba: cél egy normális eloszlású v.v ismeretlen μ várható értékére vonatkozó hipotézis tesztelése ismert σ_0 szórás mellett, de most a hipotézis egyoldali, azaz

$$\mathrm{H}_0 : \mu = \mu_0, \quad \mathrm{H}_1 : \mu < \mu_0$$

A próbastatisztika

$$u := f(\xi_1, \ldots, \xi_n) = \frac{\bar{\xi}_n - \mu_0}{\sigma_0/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Adott α elsőfajú hiba mellett az elfogadási tartomány

$$E = [-a(\alpha), \infty]$$

alakú, ahol a-t úgy választjuk meg, hogy

$$P(u < a) = \phi(a) = \alpha$$

teljesüljön. (Lásd standard normális eloszlás táblázata.)

u-próbák családja (4 eset)

Kétmintás kétoldali u-próba: cél két normális eloszlású v.v ismeretlen μ és ν várható értékeikre vonatkozó hipotézis tesztelése ismert σ_0, σ_1 szórások mellett:

$$\mathrm{H}_0 : \mu = \nu, \quad \mathrm{H}_1 : \mu \neq \nu$$

A próbastatisztika

$$u := f(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_m) = \frac{\bar{\xi}_n - \bar{\eta}_m}{\sqrt{\frac{\sigma_0^2}{n} + \frac{\sigma_1^2}{m}}} \sim \mathrm{N}(0,1)$$

Adott α elsőfajú hiba mellett az elfogadási tartomány $E = [-a(\alpha), a(\alpha)]$ alakú, ahol a-t úgy választjuk meg, hogy

$$P(|u| > a) = 2(1 - \phi(a)) = \alpha$$

teljesüljön. (Lásd standard normális eloszlás táblázata.)

u-próbák családja (4 eset)

Kétmintás egyoldali u-próba: cél két normális eloszlású v.v ismeretlen μ és ν várható értékeikre vonatkozó hipotézis tesztelése ismert σ_0, σ_1 szórások mellett, de most a hipotézis egyoldali, azaz

$$\mathrm{H}_0 : \mu = \nu, \quad \mathrm{H}_1 : \mu < \nu$$

A próbastatisztika

$$u := f(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_m) = \frac{\bar{\xi}_n - \bar{\eta}_m}{\sqrt{\frac{\sigma_0^2}{n} + \frac{\sigma_1^2}{m}}} \sim \mathrm{N}(0,1)$$

Adott α elsőfajú hiba mellett az elfogadási tartomány $E = [-a(\alpha), \infty]$ alakú, ahol a-t úgy választjuk meg, hogy

$$P(u < a) = \phi(a) = \alpha$$

teljesüljön. (Lásd standard normális eloszlás táblázata.)

t-próbák családja

Egymintás kétoldali t-próba: cél egy normális eloszlású v.v ismeretlen μ várható értékére vonatkozó hipotézis tesztelése ismeretlen σ_0 szórás mellett:

$$\mathrm{H}_0 : \mu = \mu_0, \quad \mathrm{H}_1 : \mu \neq \mu_0$$

A próbastatisztika

$$t := f(\xi_1, \ldots, \xi_n) = \frac{\bar{\xi}_n - \mu_0}{s_n^*/\sqrt{n}} \sim t_{n-1}$$

Adott α elsőfajú hiba mellett az elfogadási tartomány

$$E = [-a(\alpha), a(\alpha)]$$

alakú, ahol a-t úgy választjuk meg, hogy

$$P(|t| > a) = \alpha$$

teljesüljön. (Lásd t-eloszlás táblázata.)

t-próbák családja

Kétmintás kétoldali t-próba: cél két normális eloszlású v.v ismeretlen μ és ν várható értékeikre vonatkozó hipotézis tesztelése ismeretlen, de közös σ_0 szórás mellett:

$$\mathrm{H}_0 : \mu = \nu, \quad \mathrm{H}_1 : \mu \neq \nu$$

A próbastatisztika

$$t := \frac{\bar{\xi}_n - \bar{\eta}_m}{\sqrt{(n-1)s_n^2 + (m-1)r_m^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t_{n+m-2}$$

Adott α elsőfajú hiba mellett az elfogadási tartomány $E = [-a(\alpha), a(\alpha)]$ alakú, ahol a-t úgy választjuk meg, hogy

$$P(|t| > a) = \alpha$$

teljesüljön. (Lásd t-eloszlás táblázata.)

t-próbák családja

Egy- és kétmintás egyoldali t-próba: Az eljárás ugyanaz, mint a kétoldali esetekben, a hipotézis is a szokásos: egymintás esetben

$$\mathrm{H}_0 : \mu = \mu_0, \quad \mathrm{H}_1 : \mu > \mu_0,$$

míg kétmintás esetben

$$\mathrm{H}_0 : \mu = \nu, \quad \mathrm{H}_1 : \mu > \nu.$$

A Student eloszlás szimmetriája miatt ekkor az elfogadási tartományt definiáló $a(\alpha)$ meghatározására a

$$P(t > a) = P(|t| > a)/2 = \alpha$$

egyenletet kell megoldanunk a táblázat segítségével.

Fontos: a CHT miatt nagy n esetén akkor is alkalmazhatóak az u és t próbák, ha a v.v.-k nem normális eloszlásúak.

F-próba

Láttuk, hogy a kétmintás t-próba esetén ellenőriznünk kell azt, hogy a minták szórása megegyezik-e. Erre szolgál az F-próba. Legyenek

$$\mathrm{H}_0 : \sigma_0^2 = \sigma_1^2, \quad \mathrm{H}_1 : \sigma_0^2 \neq \sigma_1^2.$$

A próbastatisztika

$$F = F(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_m) = \frac{s^2}{r^2} \sim F(n-1, m-1)$$

ahol s^2 és r^2 a korrigált tapasztalati szórásnégyzetek. Mindig feltesszük, hogy $s^2 > r^2$. Adott α elsőfajú hiba mellett az elfogadási tartomány $E = [1, a(\alpha)]$ alakú, ahol a-t úgy választjuk meg, hogy

$$P(F > a) = \alpha/2$$

teljesüljön. (Lásd F-eloszlás táblázata.)

Alapfeladat

A vizsgálandó kérdések típusai:

- Szabályos-e egy dobókocka? - illeszkedés-vizsgálat
- Két minta azonos eloszlású-e? - homogenitás-vizsgálat
- Független-e egymástól két ismérv, pl. képesség, avagy vásárlási szokások, stb... - függetlenség-vizsgálat

A fenti feladatokra mind használható úgynevezett χ^2-próba. Az elnevezés a próbastatisztikák határeloszlására utal.

Illeszkedésvizsgálat

Adott $\xi_1, \ldots, \xi_n$ i.i.d. mintáról szeretnénk eldönteni, hogy egy adott, $x_1, \ldots, x_k$ értékű eloszlásból származik-e. Azaz

$$\mathrm{H}_0 : P(\xi_1 = x_j) = p_j,\ 1 \le j \le k, \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ nem teljesül.}$$

A próbastatisztika

$$T := \sum_{j=1}^{k} \frac{(np_j - \nu_j)^2}{np_j} \sim \chi^2_{k-1}$$

ahol $\nu_j = \sharp\{i : \xi_i = x_j\}$, a mintában előforduló x_j értékek száma. Adott α mellett a kritikus értéket a

$$P(T > a) = \alpha$$

összefüggés megoldása adja.

Homogenitásvizsgálat

Adott $\xi_1, \ldots, \xi_m$ és $\eta_1, \ldots, \eta_n$ független minták esetén azt szeretnénk eldönteni, hogy a két minta eloszlása megegyezik-e.

Legyen $(x_1, \ldots, x_k)$ az értékhalmaz, ν_j az x_j érték gyakorisága a ξ minta esetén, (μ_j) az η minta esetén, (p_j) és (r_j) pedig az eloszlások, $j = 1, \ldots, k$. Ekkor

$$\mathrm{H}_0 : p_j = r_j,\ 1 \le j \le k, \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ nem teljesül.}$$

A próbastatisztika

$$T := \frac{1}{mn} \sum_{j=1}^{k} \frac{(n\mu_j - m\nu_j)^2}{\nu_j + \mu_j} \sim \chi^2_{k-1}$$

Adott α mellett a kritikus értéket a $P(T > a) = \alpha$ összefüggés megoldása adja.

Függetlenségvizsgálat

Adott két szempont és n megfigyelés, az első szempont szerint k, a második szerint pedig l osztály. Független lesz-e a két szempont egymástól? Legyen

$$A_i = \{\text{az első szempont szerint az } i.\text{ kategóriába esik a megfigyelés}\},$$

$$B_j = \{\text{a második szempont szerint a } j.\text{ kategóriába esik a megfigyelés}\}.$$

Ekkor

$$\mathrm{H}_0 : P(A_i \cap B_j) = P(A_i)P(B_j)\,\forall i, j, \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ nem teljesül.}$$

A próbastatisztika itt is χ^2 eloszlású $(k-1)(l-1)$ szabadságfokkal. Adott α mellett a kritikus értéket a $P(T > a) = \alpha$ összefüggés megoldása adja.