Valószínűségszámítás és Matematikai Statisztika

Miskolc, 2025.

Dr. Glavosits Tamás

10. előadás

Statisztika I.,
Pontbecslések

1. Szükséges matematikai előismeretek

A logaritmus azonosságai

Legyenek $x, y > 0$, $c \in \mathbb{R}$.

1. $\ln(xy) = \ln(x) + \ln(y)$;
2. $\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y)$;
3. $\ln(x^c) = c\ln(x)$;
4. $\ln(e^x) = x$.

Az 1. azonosság véges sok tényezőre is igaz, azaz ha $x_1, x_2, \ldots, x_n$ pozitív valós számok, akkor

$$\ln\left(\prod_{i=1}^{n} = x_i\right) = \sum_{i=1}^{n} \ln(x_i).$$

A szummázás azonosságai

Legyenek $x_1, x_2, \ldots, x_n$ és $y_1, y_2, \ldots, y_n$ tetszőleges valós számsorozatok, $\lambda \in \mathrm{R}$. Ekkor

1. $\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$;
2. $\sum_{i=1}^{n} \lambda x_i = \lambda \sum_{i=1}^{n} x_i$;
3. $\sum_{i=1}^{n} 1 = n$.

A deriválás azonosságai

Legyenek $f_1, f_2, \ldots, f_n$ deriváltható függvények, $\lambda \in \mathbb{R}$. Ekkor

1. $\frac{\partial}{\partial v} \sum_{i=1}^{n} f_i(v) = \sum_{i=1}^{n} \frac{\partial}{\partial v} f_i(v)$;
2. $\frac{\partial}{\partial v} \lambda f(v) = \lambda \frac{\partial}{\partial v} f(v)$.

Az 1. azonosság a deriválás additív, a 2. azonosság a homogén tulajdonságát fejezi ki.

2. A statisztika részterületei

A statisztika részterületei

- **Mintavételezés**
- **Becsléselmélet**
  - pontbecslés
  - intervallumbecslés
- **Hipotéziselmélet**

Minta és mintarealizáció

- **Minta:** $\xi_1, \xi_2, \ldots, \xi_n$ azonos eloszlású valószínűségi változók. Ha ezek függetlenek, akkor független mintáról beszélünk.
- **Mintarealizáció:** $\xi_1, \xi_2, \ldots, \xi_n$ valós számok vagy vektorok.
- **Mintaelemszám:** n.
  $n \leqslant 50$ kis minta
  $50 \leqslant n \leqslant 500$ közepes minta
  $n \geqslant 500$ nagy minta.

Mintavételezés

- **Reprezentatív minta:** a populáció minden tagjának egyforma esélye van a mintába kerülésre.
- **Rétegzett mintavétel:** rétegképző ismérveket használunk. Ezek az ismérvek kapcsolatban vannak azzal a paraméterrel, amit vizsgálunk. Ezek az ismérvek segítenek abban, hogy a minta emlékeztessen a sokaság összetételére.

Becslések

A statisztika célja (többek között) az eloszlás ismeretlen paraméterének (vagy paramétereinek) a meghatározása.

- **Pontbecslés:** megmondjuk, hogy mi az ismeretlen paraméter.
- **Intervallumbecslés:** megadjuk azt az intervallumot, amelybe az ismeretlen paraméter (előre adott) nagy valószínűséggel beletartozik.

Statisztika, torzítatlan becslés

Definíció

Egy $g_n : \mathbb{R}^n \to \mathbb{R}$ vagy $(\mathbb{R}^n \to \mathbb{R}^k)$ függvénysorozatot **statisztiká**nak nevezünk.

Definíció

Egy (g_n) becsléssorozatot a $\vartheta \in \Theta$ ismeretlen paraméter

- **Torzítatlan becslés**nek nevezzük, ha $\mathbb{E}(g_n(\xi_1, \xi_2, \ldots, \xi_n)) = \vartheta$.
- **Asszimptotikusan torzítatlan becslés**nek nevezzük, ha $\lim_{n\to\infty} \mathbb{E}(g_n(\xi_1, \xi_2, \ldots, \xi_n)) = \vartheta$.

Konzisztens becslés, hatásosabb statisztika

Definíció

A (g_n) becsléssorozatot az ismeretlen ϑ paraméter **konzisztens becslés**nek nevezzük, ha

$$\lim_{n\to\infty} \mathbb{P}(|g_n(\xi_1, \xi_2, \ldots, \xi_n) - \vartheta| < \varepsilon) = 0$$

tetszőleges $\varepsilon > 0$ esetén, (más szavakkal a $g_n(\xi_1, \xi_2, \ldots, \xi_n)$ valószínűségi változó sztochasztikusan konvergál a ϑ-hoz).

Definíció

Legyen a g_n és a $g_n^{'}$ két torzítatlan becslése a ϑ paraméternek (azaz $\mathbb{E}(g_n(\xi_1, \xi_2, \ldots, \xi_n) = \vartheta$ és $\mathbb{E}(g_n^{'}(\xi_1, \xi_2, \ldots, \xi_n) = \vartheta)$. Azt mondjuk, hogy a g_n **becslés hatásosabb a** $g_n^{'}$ **becslésnél**, ha kisebb a szórása, azaz

$$\mathbb{D}^2(g_n(\xi_1, \xi_2, \ldots, \xi_n) \leqslant \mathbb{D}^2(g_n^{'}(\xi_1, \xi_2, \ldots, \xi_n).$$

3. Alapstatisztikák

Alapstatisztikák

A következő alapstatisztikákkal fogunk megismerkedni

- átlag ($\overline{\xi}$);
- empirikus szórásnégyzet (s_n^2);
- korrigált empirikus szórásnégyzet (s_n^{*2});
- empirikus medián (med);
- medián abszolút eltérés (MAD);
- tapasztalati eloszlásfüggvény ($\mathbb{F}_n^*$);
- hisztogramok.

Átlag

Definíció

A $\xi_1, \xi_2, \ldots, \xi_n$ minta $\overline{x}$ módon jelölt **átlag**a:

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

Tétel (Átlag tulajdonsága)

Az átlag az elméleti várható érték (m) *torzítatlan becslése.*

Proof.

A bizonyítás a várható érték additivitása és homogenitása alapján nyilvánvaló. □

Tapasztalati szórásnégyzet

Definíció

Tapasztalati (vagy empirikus) szórásnégyzet:

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

Az s_n^2 tulajdonságai:

1. **Kiszámítása:**

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

Az itt szereplő $\frac{1}{n}\sum_{i=1}^{n}(\xi_i)^2 := m_2$ mennyiséget tapasztalati második momentumnak nevezzük.
Így $s_n^2 = m_2 - (\overline{\xi})^2$.

2. **Steiner formula:** tetszőleges $a \in \mathbb{R}$ esetén:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\overline{x} - a)^2.$$

3. Ha $\xi_1, \xi_2, \ldots, \xi_n$ páronként korrelálatlanok, akkor

$$\mathbb{E}(s_n^2) = \frac{n-1}{n}\sigma^2.$$

A Steiner formula bizonyítása

Legyen $a \in \mathbb{R}$ tetszőleges. Ekkor a teve szabályt kell alkalmazni.

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - a) + (a - \overline{x}))^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((x_i - a)^2 + 2(x_i - a)(a - \overline{x}) + (a - \overline{x})^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 + \underbrace{2(a - \overline{x})\frac{1}{n}\sum_{i=1}^{n}(x_i - a)}_{-2(\overline{x}-a)^2} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(a - \overline{x})^2}_{(\overline{x}-a)^2} =$$

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

Megjegyzés: A Steiner-formulából $a = 0$ választással kijön az 1. állítás

Az $\mathbb{E}(s_n^2) = \frac{n-1}{n}\sigma^2$ azonosság bizonyítása.

Legyen $\xi_1, \xi_2, \ldots, \xi_n$ páronként korrelálatlan minta, m az elméleti várható értékkel és σ^2 elméleti szórásnégyzettel.
A bizonyításhoz a Steiner formulát kell alkalmazni $a = m$ esetben.
A 2. állítás és a várható érték additivitása és homogenitása alapján kapjuk, hogy

$$\mathbb{E}(s_n^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi_i - m)^2 - \mathbb{E}(\overline{\xi} - m)^2.$$

Nyilvánvaló, hogy

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi - m)^2 = \frac{1}{n}n\sigma^2 = \sigma^2.$$

Megmutatjuk, hogy $\mathbb{E}(\overline{\xi} - m)^2 = \frac{1}{n}\sigma^2$, ami egy kicsit számolásigényesebb. A bizonyítás során felhasználjuk az n-tagú összeg négyzetére vonatkozó azonosságot, azaz

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n} a_i^2 + 2\sum_{i<j} a_i a_j.$$

A páronkénti korrelálatlanságot is felhasználva kapjuk, hogy

$$\mathbb{E}(\overline{\xi}-m)^2=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i-\frac{1}{n}\sum_{i=1}^{n}m\right)^2=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i-m)\right)^2=$$

$$=\frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}(\xi_i-m)\right)^2=$$

$$=\frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}(\xi_i-m)^2+2\sum_{i<j}(\xi_i-m)(\xi_j-m)\right)=$$

$$=\frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}(\xi_i-m)^2+2\sum_{i<j}\underbrace{\mathbb{E}(\xi_i-m)(\xi_j-m)}_{0}\right)=$$

$$=\frac{1}{n^2}n\sigma^2=\frac{1}{n}\sigma^2.$$

Az előzőek alapján kapjuk, hogy

$$\mathbb{E}(s_n^2)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi_i-m)^2-\mathbb{E}(\overline{\xi}-m)^2=\sigma^2-\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2.$$

Tétel

A tapasztalati szórásnégyzet az elméleti szórásnégyzet asszimptótikusan torzítatlan becslése.

Definíció

Korrigált tapasztalati szórásnégyzet:

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

azonban

$$s_n^{*2} = \frac{n}{n-1} s_n^2$$

módon is számolható.

Tétel

Az s_n^{*2} az elméleti szórásnégyzet torzítatlan becslése.

Proof.

Evidens a $\mathbb{E}(s_n^*) = \frac{n-1}{n}\sigma^2$ miatt. □

Definíció

Ha ξ egy abszolút folytonos valószínűségi változó, akkor a ξ **medián**jának azt a $\nu \in \mathbb{R}$ számot nevezzük, amelyre $\mathbb{F}(\nu) = \frac{1}{2}$.

Definíció

Legyen a rendezett minta: $\xi_1^* \leqslant \xi_2^* \leqslant \cdots \leqslant \xi_n^*$.
Tapasztalati medián:

$$\operatorname{med}(\xi_i) := \begin{cases} \dfrac{\xi_k^* + \xi_{k+1}^*}{2}, & \text{ha } n = 2k, \\ \xi_k^*, & \text{ha } n = 2k-1. \end{cases}$$

Definíció

Jelölje med a $\xi_1, \xi_2, \ldots, \xi_n$ minta tapasztalati mediánját. Ekkor a

$$\operatorname{MAD}(\xi_1, \ldots, \xi_n) = \operatorname{med}(|\xi_1 - \operatorname{med}|, \ldots, |\xi_n - \operatorname{med}|)$$

módon definiáljuk.

Definíció

Tapasztalati eloszlásfüggvény Legyen $\xi_1^* < \xi_2^* < \cdots < \xi_n^*$ egy rendezett minta. Ekkor az $\mathbb{F}_n^* : \mathbb{R} \to \mathbb{R}$

$$\mathbb{F}_n^*(x) := \begin{cases} 0, \text{ ha } x \leqslant \xi_1^*, \\ \dfrac{k}{n}, \text{ ha } \xi_k^* < x \leqslant \xi_{k+1}^*, \\ 1, \text{ ha } x > \xi_n^*. \end{cases}$$

módon definiált függvényt **tapasztalati eloszlásfüggvény**nek nevezzük.

Tehát $\mathbb{F}_n^*$ egy olyan monoton növekvő balról folytonos lépcsős függvény, amely a rendezett minta minden elemén $\frac{1}{n}$-et ugrik.
A tapasztalati eloszlásfüggvény más módon is bevezethető:

$$\mathbb{F}_n^*(x) := \frac{1}{n} \sum_{i=1}^{n} I_{]\xi_i,+\infty[}(x),$$

ahol az $I_H : \mathbb{R} \to \mathbb{R}$ függvény

$$I_H(x) := \begin{cases} 1, \text{ ha } x \in H; \\ 0, \text{ egyébként.} \end{cases}$$

módon van definiálva tetszőleges $H \subseteq \mathbb{R}$ esetén. Az I_H függvényt a H halmaz **indikátorfüggvény**ének nevezzük.
Ez utóbbi definíció jobb, mint az előző, mivel nem igényli a mintaelemek páronkénti különbözőségét.

Glivenko-Cantelli tétel

Tétel

Glivenko-Cantelli tétel vagy ***a matematikai statisztika alaptétele:***

$$\mathbb{P}\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|\mathbb{F}_n^*(x)-\mathbb{F}_n(x)|=0\right)=1$$

ahol $\mathbb{F}$ az elméleti eloszlásfüggvény, ami azt jelenti, hogy a tapasztalati eloszlásfüggvény visszaadja annak az eloszlásnak az eloszlásfüggvényét, amelyből a minta származik.

4. Becslési módszerek

Maximum likelihood becslés

Tegyük fel, hogy $\xi_1, \ldots, \xi_n$ egy független minta ϑ ismeretlen paraméterrel.

$$L(\xi_1, \ldots, \xi_n, \vartheta) = \begin{cases} \prod_{i=1}^{n} p(\xi_i, \vartheta) & \text{diszkrét minta esetén,} \\ \prod_{i=1}^{n} f(\xi_i, \vartheta) & \text{abszolút folytonos minta esetén} \end{cases}$$

ahol $p(\cdot, \vartheta)$ jelöli azt az eloszlást, amelyből a diszkrét minta származik, illetve abszolút folytonos esetben $f(\cdot, \vartheta)$ jelöli azt az ismeretlen eloszlás sűrűségfüggvényét az L függvényt **likelihood függvény**nek nevezzük.
Jelölje $\widehat{\vartheta}$ az ismeretlen $\vartheta \in \Theta$ paraméter maximum likelihood becslését.

A $\widehat{\vartheta}$ meghatározása:

A $\widehat{\vartheta}$ az a hely, ahol az $L(\cdot)$ függvény a maximumát felveszi, azaz

$$L(\widehat{\vartheta}) = \max_{\vartheta \in \Theta} L(\vartheta).$$

A technikai kivitelezéshez felhasználjuk azt az analízisből ismert tételt, hogyha egy deriválható függvénynek egy pontban lokális maximuma, vagy minimuma van, ott az első derivált eltűnik. Azokat a pontokat, ahol egy deriválható függvény első deriváltja eltűnik a függvény stacionárius pontjainak nevezzük. Tehát meg kell keresnünk az $\vartheta \to L(\vartheta)$ függvény stacionárius pontjait.

Mivel a $\vartheta \to L(\vartheta)$ függvény szorzat alakú, és a szorzat alakú függvényt nem könnyű deriválni, így vesszük a $\vartheta \to L(\vartheta)$ függvény logaritmusát, a

$$\vartheta \to l(\vartheta) := \ln(L(\vartheta))$$

függvényt, amelyet log-likelihood függvénynek nevezünk és $l(\vartheta)$ módon jelölünk.
A log-likelihood függvénynek több jó tulajdonsága is van:

- mivel az $\ln$ függvény szigorúan növekvő, így a $\vartheta \to l(\vartheta)$ függvénynek ugyanazok a maximumhelyei, mint a $\vartheta \to L(\vartheta)$ függvénynek;
- a logaritmus szorzatot összegbe visz, így a deriválás könnyebben kivitelezhető.

Összegezve: Meg kell oldanunk a $\frac{\partial}{\partial \vartheta} l(\vartheta) = 0$ egyenletet, majd meg kell vizsgálnunk, hogy a kapott $\widehat{\vartheta}$ valóban maximum helye $\vartheta \to l(\vartheta)$ függvénynek. Ez utóbbi vizsgálattól rendszerint eltekintünk.

A maximum Likelihood módszer tulajdonságai:

- Nem mindig ad torzítatlan becslést;
- Mindig konzistens becslést ad;
- A leghatásosabb becslést adja (amennyiben van ilyen).

k-adik elméleti momentum, k-adik empirikus momentum

Ha egy ξ egy valószínűségi változó, akkor a ξ k-**adik elméleti momentum**át

$$\nu_k := \mathbb{E}(\xi^k)$$

módon definiáljuk (amennyiben létezik).
Ha $\xi_1, \xi_2, \ldots, \xi_n$ egy független minta úgy, hogy az eloszlásnak, amelyből származik létezik a k-adik momentuma, akkor ezt a k-**adik empirikus momentum**mal közelítjük, ami

$$m_k := \frac{1}{n} \sum_{i=1}^{n} \xi_i^k$$

módon van definiálva.

A momentumok módszere

Adott egy $\xi_1, \xi_2, \ldots, \xi_n$ független minta ismeretlen paraméterekkel, a célunk az ismeretlen paraméterek meghatározása.
Az ismeretlen paraméterek kifejezhetők elméleti momentumokkal, bár ehhez gyakran nemlineáris egyenletrendszert kell megoldanunk.
Ekkor az ismeretlen paraméterek becslésekor az elméleti momentumok helyett a tapasztalati momentumokat használjuk.

Például:

- A független minta a $\mathrm{Poiss}(\lambda)$ eloszlásból származik, és az ismeretlen λ paraméter becslése a cél. Mivel ekkor $\mathbb{E}(\xi) = \lambda$, így a λ ismeretlen paraméter becslése $\widehat{\lambda} = \overline{\xi}$. Ebben az esetben, mivel $\mathbb{D}^2(\xi) = \lambda$, így az ismeretlen paraméter akár $\widehat{\lambda} = m_2 - (\overline{\xi})^2$ módon is becsülhető lenne, de ebben az esetben a $\widehat{\lambda} = \overline{\xi}$ becslést kell választani, mivel alacsonyabb fokú.
- Ha a független minta $\mathrm{Exp}(\lambda)$ eloszlásból származik, akkor $\mathbb{E}(\xi) = \frac{1}{\lambda}$, így az ismeretlen λ paraméter $\widehat{\lambda} = \frac{1}{\overline{\xi}}$ módon becsülhető.
- Ha az ismeretlen minta $\mathcal{N}(m, \sigma^2)$ eloszlásból származik, akkor az m ismeretlen paraméter $\widehat{m} = \overline{\xi}$, a σ^2 ismeretlen paraméter $\widehat{\sigma^2} = m_2 - (\overline{\xi})^2$ módon becsülhető.

Vége a 10. előadásnak