Matematikai statisztika
Témakör 2: Becslések kismintás tulajdonságai

Elek Péter

1. Sokaság és minta

Statisztikai következtetéselmélet

- *Statisztikai következtetéselmélet* célja: következtetni a sokaság jellemzőire a sokaságból vett minta alapján
- pl. pontbecslés, intervallumbecslés, hipotézisvizsgálat
- Szoros kapcsolat a valószínűségszámítással
  - de gyakorlatban: teljesülnek-e az adott valószínűségi modell feltételei?

Sokaság

- *Sokaság (populáció)*: a statisztikai vizsgálat tárgyát képező egységek összessége
  - ezek valamilyen tulajdonságának (pl. tömeg, életkor) jellemzői érdekelnek minket (pl. várható érték, eloszlás)
- Lényeg: a sokasági tulajdonság ismeretlen eloszlása
- Lehet véges vagy végtelen, folytonos vagy diszkrét stb.
- Példák:
  - magyarországi felnőtt férfiak testmagassága
  - egy évfolyam diákjainak statisztika pontszáma
  - kockadobás

Minta

- A sokasági eloszlás jellemzőire szeretnénk következtetni a minta alapján
- *Minta*: a sokaságból vett megfigyelések halmaza
- *Független, azonos eloszlású (f.a.e.) minta* (independent identically distributed, *i.i.d.*): $X_1, X_2, \ldots, X_n$ mintaelemek függetlenek és az $f(x)$ sokasági sűrűségfüggvénnyel (eloszlással) jellemezhetők (n a mintaelemszám)
  - Ekkor a minta együttes sűrűségfüggvénye $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.
  - Példák: f.a.e. mintavétel végtelen sokaságból, *visszatevéses* mintavétel véges sokaságból
- Példa: 100-szor feldobunk egy kockát, és a kocka jellemzőire következtetünk

Eltérések a f.a.e. mintától

- *Visszatevés nélküli* minta véges sokaságból
  - Ekkor a mintaelemek azonos eloszlásúak (belátható), de természetesen nem függetlenek!
  - Elnevezés: egyszerű véletlen minta
  - Ha a mintaelemszám sokkal kisebb a sokaság méreténél, akkor nincs nagy jelentősége
- Egyéb eljárások (pl. *rétegzett minta*)
- (Nem valószínűségi mintavétel: pl. hólabda-módszer)
- Néhány gyakori eltérés a közgazdasági gyakorlatban:
  - idősorok és paneladatbázisok
  - térbeli összefüggőség

Paraméteres és nemparaméteres eljárások

- *Paraméteres* eljárások: feltételezik, hogy a minta egy – ismeretlen paraméterű – paraméteres eloszláscsaládból származik
  - Példák:
    - Bernoulli(p) (pl. demokrata vagy republikánus)
    - Normális μ és σ^2 paraméterekkel (pl. testmagasság)
  - Jelölés: $f(x, \theta)$ vagy $f_\theta(x)$ az ismeretlen sűrűségfüggvény (ahol θ az esetleg többdimenziós paraméter).
- *Nemparaméteres* (eloszlásfüggetlen) eljárások: ha a lehetséges eloszlások halmaza "bő" (pl. összes folytonos eloszlás)

2. Paraméterek, statisztikák, mintavételi eloszlások

Paraméter

- Keressük a sokasági eloszlás ismeretlen paraméterét a minta alapján
- Példák: p (Bernoulli), μ, σ (normális), λ (Poisson), θ (exponenciális)

(Minta)statisztika

- *(Minta)statisztika*: az $x_1, x_2, \ldots, x_n$ mintaelemek függvénye
  - (tehát a mintaelemekből számolt numerikus érték)
- Példa 1: mintaátlag: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Példa 2: tapasztalati medián: $m = Med(x_1, x_2, \ldots, x_n)$
- Példa 3: tapasztalati variancia: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$
- Megjegyzés:
  - a paraméter tehát az ismeretlen sokasági eloszlás jellemzője,
  - míg a mintastatisztikát a mintából számítjuk

A mintastatisztika maga is egy valószínűségi változó!

- Jelölés: $W = T(X_1, X_2, \ldots, X_n)$
  - egy konkrét mintarealizáció esetén pedig $w = T(x_1, x_2, \ldots, x_n)$
- (a paraméterek viszont a klasszikus (nem Bayes-i) megközelítésben nem valószínűségi változók)

Becslőfüggvények és becslések (pontbecslés)

- *(Pont)becslőfüggvény* (point estimator) olyan mintastatisztika, amelyet egy paraméter becslésére használunk: $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$
- A *pontbecslés* (point estimate) ebből konkrét számok behelyettesítésével adódik : $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$.
- Nem fogjuk mindig a becslőfüggvényt és becslést így megkülönböztetni.
- Példák:
  - Mintaátlag a várható érték becslőfüggvénye
  - Tapasztalati medián a várható érték (és a sokasági medián) egy másik becslőfüggvénye
  - Tapasztalati variancia a sokasági variancia becslőfüggvénye

Becslőfüggvények tulajdonságai

- Milyen a "jó" becslőfüggvény?
- Példa: a mintaátlag vagy a tapasztalati medián a várható érték "jobb" becslőfüggvénye?
- Becslőfüggvények fontosabb lehetséges tulajdonságai:
  - *Kismintás (véges mintás)* tulajdonságok: rögzített n elemű minta
  - *Nagymintás (aszimptotikus)* tulajdonságok: $n \to \infty$ eset

Mintavételi eloszlás és meghatározása

- *Mintavételi eloszlás*: a mintastatisztika (vagy becslőfüggvény) mint valószínűségi változó eloszlása
- Egyes esetekben analitikusan, máskor szimulációval határozhatjuk meg.
- Szimuláció:
  - Mintavétel az eloszlásból
  - Mintastatisztika (becslés) kiszámítása
  - Ennek ismétlése nagyon sokszor (pl. 10000-szer)
  - A becslés 10000 realizációja tekinthető a becslőfüggvény mintavételi eloszlásának.
- Példa: kockadobás $n = 10$-szer, mintaátlag kiszámítása, majd az eljárás ismétlése pl. 10000-szer

Példa

- A sokaság három elemből áll: $\{0, 3, 12\}$ és mindegyik bekövetkezési valószínűsége $p = \frac{1}{3}$.
- Visszatevéssel $n = 3$ elemű mintát veszünk a sokaságból.
- Mi a tényleges sokasági eloszlás?
- Soroljuk fel a lehetséges mintákat és azok bekövetkezési valószínűségeit, és minden esetben számoljuk ki a mintaátlagot és a tapasztalati mediánt!
- Határozzuk meg a mintaátlag és a tapasztalati medián mintavételi eloszlását!

3. Torzítatlanság, relatív hatásosság és MSE-kritérium

Torzítatlanság

- Becsülje a $\widehat{\theta}$ becslőfüggvény a θ sokasági paramétert.
  - $\widehat{\theta}$ tehát valószínűségi változó, és eloszlása függ θ-tól.
- $\widehat{\theta}$ a θ paraméter *torzítatlan becslőfüggvénye*, ha minden θ esetén $E(\widehat{\theta}) = \theta$.
- Ha $E(\widehat{\theta}) \neq \theta$ legalább egy θ érték esetén, akkor $\widehat{\theta}$ a θ torzított becslőfüggvénye.
- Torzítás: $Bias_{\widehat{\theta}}(\theta) = E\left(\widehat{\theta}\right) - \theta$
  - Ez szintén a θ függvénye.

Példa

- Példa (korábbról): $\overline{X}$ és m mintavételi eloszlása az adott sokasági eloszlás mellett:

	0	1	2	3	4	5	6	8	9	12
$\overline{x}$	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$
m	$\frac{7}{27}$	0	0	$\frac{13}{27}$	0	0	0	0	0	$\frac{7}{27}$

- Melyik a várható érték torzítatlan becslőfüggvénye?
  - $\overline{X}$ mindig torzítatlanul becsüli $E(X)$-et (ebben az esetben is)
  - m az $E(X)$-et és a sokasági mediánt sem becsüli torzítatlanul itt
  - (Azonban pl. normális eloszlás esetén igen, mert a normális eloszlás szimmetrikus.)

Relatív hatásosság

- Legyen $\widehat{\theta}_1$ és $\widehat{\theta}_2$ a θ paraméter két *torzítatlan* becslőfüggvénye.
- $\widehat{\theta}_1$ *legalább olyan hatásos*, mint $\widehat{\theta}_2$, ha $Var(\widehat{\theta}_1) \leq Var(\widehat{\theta}_2)$ minden θ esetén.
  - Azaz a mintavételi eloszlás szórása sehol sem nagyobb.
  - Torzítatlan becslőfüggvényeket hasonlítunk össze!

Hatásosság

- A $\widehat{\theta}$ torzítatlan becslés *hatásos* (máshogy: minimális varianciájú torzítatlan becslőfüggvény – MVUE), ha legalább olyan hatásos, mint bármelyik más torzítatlan becslőfüggvény.
  - Azaz bármely más $\widehat{\theta}_2$ torzítatlan becslőfüggvényre $Var(\widehat{\theta}) \leq Var(\widehat{\theta}_2)$ minden θ esetén.
- Megjegyzés: a hatásosságot néha máshogy definiálják (eléri a Cramer-Rao-határt, lásd később).
- Mivel a mintavételi variancia függhet θ-tól, hatásos becslőfüggvény nem mindig létezik.

Átlagos négyzetes eltérés (MSE)-kritérium

- A varianciák összehasonlítása csak torzítatlan becslőfüggvények esetén értelmes.
- De nem mindig érdemes ragaszkodni a torítatlansághoz (lásd később és a feladatok között).
- Általánosabban az *átlagos négyzetes eltérés* (mean squared error – MSE) használható a becslőfüggvények összehasonlítására.
- $MSE\left(\widehat{\theta}\right) = E\left[\left(\widehat{\theta} - \theta\right)^2\right]$.
- *MSE kritérium*: azt a becslőfüggvényt választjuk, amelynek legkisebb az MSE értéke.
- Állítás: $MSE\left(\widehat{\theta}\right) = Var\left(\widehat{\theta}\right) + \left[Bias\left(\widehat{\theta}\right)^2\right]$

Bizonyítás: $MSE\left(\widehat{\theta}\right) = Var\left(\widehat{\theta}\right) + \left[Bias\left(\widehat{\theta}\right)^2\right]$

- Azonos átalakításokkal:

$$\begin{aligned} MSE\left(\widehat{\theta}\right) = E\left[\left(\widehat{\theta} - \theta\right)^2\right] &= E\left[\left((\widehat{\theta} - E(\widehat{\theta})) - (\theta - E(\widehat{\theta}))\right)^2\right] \\ &= E\left[\left(\widehat{\theta} - E(\widehat{\theta})\right)^2\right] - 2E\left[\left(\widehat{\theta} - E(\widehat{\theta})\right)\left(\theta - E(\widehat{\theta})\right)\right] \\ &\quad + E\left[\left(\theta - E(\widehat{\theta})\right)^2\right] \\ &= Var\left(\widehat{\theta}\right) - 0 + Bias^2\left(\widehat{\theta}\right). \end{aligned}$$

Torzítás és szórás közötti átváltás

- *Torzítás-szórás közötti átváltás* (bias-variance tradeoff): gyakran érdemes egy kis torzítást megengedni a variancia csökkentése érdekében
- "Shrinkage" becslőfüggvények stb., lásd a feladatok között
- Vagy általánosabban: különböző statisztikai eljárások (lineáris regressziótól a rugalmasabb eljárásokig)

Példa (folyt.)

- Melyik korábbi becslőfüggvény (mintaátlag illetve tapasztalati medián) rendelkezik kisebb MSE-ral az adott paraméternél (azaz amikor mindhárom kimenet egyformán valószínű)?
- $Var(\overline{X}) = \frac{1}{27}(0-5)^2 + \frac{3}{27}(1-5)^2 + \ldots + \frac{1}{27}(12-5)^2 = 8.6667$ és $MSE(\overline{X}) = Var(\overline{X})$ a torzítatlanság miatt.
- $Var(m) = \frac{7}{27}(0-4.5556)^2 + \frac{13}{27}(3-4.5556)^2 + \frac{7}{27}(12-4.5556)^2 = 20.9136$ és $MSE(m) > Var(m)$ a torzítás miatt.
- Így $\overline{X}$ jobb, mint m az MSE-kritérium szerint. (Mint láttuk, az első torzítatlan is, míg a második nem.)

4. Mintaátlag és tapasztalati variancia kismintás tulajdonságai

4.1. Mintaátlag tulajdonságai

Mintaátlag tulajdonságai f.a.e. minta esetén

- Legyen $X_1, X_2, \ldots, X_n$ f.a.e. minta egy μ várható értékű és σ^2 varianciájú eloszlásból.
- Ekkor $\overline{X}$ torzítatlanul becsli μ-t.
  - $E(\overline{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\sum_{i=1}^n \mu = \frac{1}{n}(n\mu) = \mu$.
- A mintavételi szórás pedig arányos $1/\sqrt{n}$-nel.
  - $Var(\overline{X}) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}$
  - $sd(\overline{X}) = \sigma/\sqrt{n}$

Legjobb lineáris torzítatlan becslőfüggvény (BLUE)

- *Legjobb lineáris torzítatlan becslőfüggvény* (best linear unbiased estimator – BLUE): egy becslőfüggvény BLUE, ha a legkisebb varianciával rendelkezik minden olyan torzítatlan becslőfüggvény között, amely a mintaelemek lineáris kombinációja.
- Megjegyzés:
  - a mintaátlag nem mindig becsüli hatásosan μ-t (vannak "furcsa" ellenpéldák),
  - de az igaz, hogy legjobb *lineáris* torzítatlan becslőfüggvény μ-re (f.a.e. minta esetén).

Bizonyítás: mintaátlag BLUE f.a.e. minta esetén

- Legyen $\widehat{\theta}$ egy tetszőleges lineáris torzítatlan becslőfüggvény: $\widehat{\theta} = \sum_{i=1}^{n} a_i X_i$, ahol $E\left(\widehat{\theta}\right) = \mu$.
- Ekkor $E\left(\widehat{\theta}\right) = E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E\left(X_i\right) = \mu$, tehát $\sum_{i=1}^{n} a_i = 1$ teljesül.
- Továbbá $Var\left(\widehat{\theta}\right) = Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var\left(X_i\right) = \sigma^2 \sum_{i=1}^{n} a_i^2$.
- Mivel $Var\left(\overline{x}\right) = \frac{\sigma^2}{n}$, csak azt kell igazolnunk, hogy $\sum_{i=1}^{n} a_i^2 \geq \frac{1}{n}$, ha $\sum_{i=1}^{n} a_i = 1$.
- Ez viszont a számtani és négyzetes közép közötti egyenlőtlenségből adódik.

Mintaátlag tulajdonságai normális eloszlású minta esetén

- Legyen $X_1, X_2, \ldots, X_n$ f.a.e. és $N(\mu, \sigma^2)$ eloszlású minta ismeretlen μ és σ paraméterekkel.
- Ekkor μ hatásos becslés is, nem csak BLUE.
- Továbbá $\overline{X}$ is normális eloszlású: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Feladat 1

- Egy gép a beállítása szerint átlagosan 500 gramm gabonapelyhet tölt minden dobozba, az ettől való eltérés nem szisztematikus (azaz zérus várható értékű) és 20 gramm szórású normális eloszlású változó.

- A minőségbiztosítás egy $n = 25$ elemű véletlen mintát vesz a dobozokból, és megméri a tömegüket. Akkor állítják le a folyamatot, ha az átlagos tömeg 510 grammnál vagyobb vagy 490 grammnál kisebb.
- Mennyi a megállás valószínűsége?

4.2. Tapasztalati variancia tulajdonságai

Korrigálatlan tapasztalati variancia

- *Korrigálatlan* tapasztalati variancia: $s^2 = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}$
- $E(s^2) = \frac{n-1}{n}\sigma^2$, tehát ez σ^2 torzított becslőfüggvénye.
- Bizonyítás:
  - $s^2 = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n} = \frac{\sum_{i=1}^{n}\left[(X_i-\mu)-\left(\overline{X}-\mu\right)\right]^2}{n} = \ldots = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n} - \left(\overline{X}-\mu\right)^2$.
  - Várható értékeket véve:
    - $E\left[(X_i-\mu)^2\right] = Var(X_i) = \sigma^2$, és
    - $E\left[\left(\overline{X}-\mu\right)^2\right] = E\left[\left(\overline{X}-E(\overline{X})\right)^2\right] = Var(\overline{X}) = \frac{\sigma^2}{n}$.
  - Tehát $E(s^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$.

Korrigált tapasztalati variancia

- *Korrigált* tapasztalati variancia: $s^{*2} = \frac{n}{n-1}\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}$
- Ez tehát σ^2 torzítatlan becslőfüggvénye.
  - Egy szabadsági fokot "elvesztünk", mert μ-t $\overline{X}$-gal becsültük a képletben.

Tapasztalati variancia további tulajdonságai normális eloszlású minta esetén

- Legyen $X_1, \ldots, X_n$ normális eloszlású f.a.e. minta μ és σ^2 paraméterekkel.
- Ekkor könnyen láthatóan $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$ eloszlású.
- Ha μ-t $\overline{X}$-gal becsüljük, akkor:

$$(n-1)\,s^{*2} = \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - n\left(\overline{X}-\mu\right)^2 \sim \sigma^2\chi^2_{n-1}.$$

- Ez utóbbit nem bizonyítjuk (teljes indukcióval történhet).

Normális eloszlású minta (folyt.)

- Mivel $(n-1)s^{*2} = ns^2$, ezért $\frac{ns^2}{\sigma^2} = \frac{(n-1)s^{*2}}{\sigma^2}$ is χ^2_{n-1}-eloszlású.
- Néhány további tulajdonság (a teljesség kedvéért):
  - $Var\left(\frac{(n-1)s^{*2}}{\sigma^2}\right) = 2(n-1)$ (miért?)
  - $Var\left(s^{*2}\right) = \frac{\sigma^4}{(n-1)^2} Var\left(\frac{(n-1)s^{*2}}{\sigma^2}\right) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1}$
  - $Var(s^2) = \frac{2(n-1)\sigma^4}{n^2}$

Feladat 2

- Mennyi a valószínűsége, hogy az előző példában az $n = 25$ elemű mintából számított korrigált tapasztalati szórás meghaladja a 24,5 grammot?

Standardizált mintaátlag normális eloszlású mintában

- Standardizált mintaátlag: $\frac{\overline{X}-\mu}{s^*/\sqrt{n}}$
- Legyen $X_1, X_2, \ldots, X_n$ f.a.e. normális eloszlású minta.
- Tudjuk, hogy
  - $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ és
  - $\sqrt{\frac{(n-1)s^{*2}/\sigma^2}{n-1}} = \frac{s^*}{\sigma} \sim \sqrt{\frac{\chi^2_{n-1}}{n-1}}$.
- Továbbá a két változó egymástól független (nem bizonyítjuk).
- Ezért $\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{s^*/\sigma} = \frac{\overline{X}-\mu}{s^*/\sqrt{n}} \sim t_{n-1}$.

Tananyag

- slide-ok és feladatok
- Wooldridge Appendix C.1-C.2
- Amemiya 1, 7.1-7.2 (kivéve 7.1.3, 7.2.2, 7.2.4, 7.2.6)