

## Konfidenciaintervallumok (6. előadás)

Legyen  $(\Omega, \mathcal{A}, \mathcal{P})$  statisztikai mező,  $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$  és  $\underline{X} = (X_1, \dots, X_n)$  független azonos eloszlású minta. Tegyük fel, hogy  $\vartheta$  valós paraméter, vagyis  $\Theta \subseteq \mathbb{R}$ .

### Definíció

Azt mondjuk, hogy a  $(T_1(\underline{X}), T_2(\underline{X}))$  intervallum legalább  $1 - \alpha$  megbízhatósági szintű konfidenciaintervallum  $\vartheta$ -ra, ha minden  $\vartheta \in \mathbb{R}$  esetén teljesül, hogy

$$\mathbb{P}_\vartheta(T_1(\underline{X}) < \vartheta < T_2(\underline{X})) \geq 1 - \alpha.$$

A konfidenciaintervallum megbízhatósági szintje:  $\inf_{\vartheta \in \Theta} \{\mathbb{P}_\vartheta(\vartheta \in (T_1, T_2))\}$ .

## Konfidenciaintervallum a várható értékre

Legyenek  $Z_0, Z_1, \dots, Z_n$  független  $N(0, 1)$  eloszlásúak, és  $t_{f, \alpha}$  az  $f$  szabadsági fokú  $\alpha$  terjedelmű kétoldali  $t$ -próba kritikus értéke, azaz az  $f$  szabadsági fokú  $t$ -eloszlás  $1 - \alpha/2$ -kvantilise:

$$1 - \alpha/2 = \mathbb{P}(Y \leq t_{f, \alpha}) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \dots + Z_f^2}} \leq t_{f, \alpha}\right).$$

Az  $Y = \frac{Z_0}{\sqrt{Z_1^2 + \dots + Z_f^2}}$  valószínűségi változó eloszlása  $f$  szabadsági fokú  **$t$ -eloszlás**.

### Állítás (Konfidenciaintervallum a várható értékre, ismeretlen szórás)

*Tegyük fel, hogy  $X_1, \dots, X_n$  független  $N(m, \sigma^2)$  normális eloszlású valószínűségi változók ( $m, \sigma$  ismeretlenek). Ekkor a*

$$(T_1, T_2) = \left( \bar{X} - t_{n-1, \alpha} \cdot \frac{S_n^*}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha} \cdot \frac{S_n^*}{\sqrt{n}} \right)$$

*intervallum  $1 - \alpha$  megbízhatósági szintű kétoldali konfidenciaintervallum az eloszlás várható értékére.*

# Konfidenciaintervallum a szórásra

Fisher–Bartlett-tétel:

$$\frac{(n-1)s_n^{*2}}{\sigma^2} = \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2,$$

azaz a hányados eloszlása  $n - 1$  szabadsági fokú  $\chi^2$  (ami megegyezik  $n - 1$  darab független standard normális eloszlás négyzetösszegének eloszlásával).

## Állítás

Legyen  $X_1, X_2, \dots, X_n$  független normális eloszlású minta. Ekkor az eloszlás  $\sigma^2$  szórásnégyzetére  $1 - \alpha$  megbízhatósági szintű konfidenciaintervallum az alábbi:

$$(T_1, T_2) = \left( \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{c_{n-1, 1-\alpha/2}}; \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{c_{n-1, \alpha/2}} \right),$$

ahol  $c_{f,q}$  az  $f$ -szabadsági fokú  $q$  terjedelmű  $\chi^2$ -próba kritikus értéke, azaz az  $f$  szabadsági fokú  $\chi^2$ -eloszlás  $q$ -kvantilise.

Ezzel a választással nem a legrövidebb intervallumot kapjuk.



## Konfidenciaintervallum a valószínűsége

Az  $A$  esemény valószínűsége  $p \in [0, 1]$  ismeretlen paraméter. Ezt szeretnénk megbecsülni. Ha  $n$  kísérletből az  $A$  esemény  $X$ -szer következett be:

$$\mathbb{E}_p(X) = np; \quad D_p(X) = \sqrt{np(1-p)}; \quad \mathbb{E}_p\left(\frac{X}{n}\right) = p; \quad D_p\left(\frac{X}{n}\right) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

$X \sim \text{Bin}(n, p)$  binomiális eloszlású, viszont a relatív gyakoriságot,  $X/n$ -t nem tudjuk egyetlen eloszlással közelíteni.

Ezért  $X/n$  eloszlását  $\hat{p}$  várható értékű,  $\frac{\hat{p}(1-\hat{p})}{\sqrt{n}}$  szórású normális eloszlással közelítjük, ahol  $\hat{p}$  az esemény relatív gyakorisága a mintában.

$1 - \alpha$  megbízhatósági szintű konfidenciaintervallum  $p$ -re:

$$\left( \hat{p} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}; \hat{p} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right).$$

## Konfidenciaintervallum az ML-becslés alapján

Ha likelihoodfüggvény teljesít bizonyos regularitási feltételeket, akkor a  $\vartheta$  paraméternek az  $X_1, X_2, \dots, X_n$  mintából számolt  $\hat{\vartheta}_n$  maximumlikelihood-becslése

- létezik;
- aszimptotikusan torzítatlan:  $\lim_{n \rightarrow \infty} \mathbb{E}_{\vartheta}(\hat{\vartheta}_n) = \vartheta$  minden  $\vartheta \in \Theta$ -ra;
- aszimptotikusan hatásos:  $\lim_{n \rightarrow \infty} \sqrt{nl_1(\vartheta)} D_{\vartheta}(\hat{\vartheta}_n) = 1$  minden  $\vartheta \in \Theta$ -ra;
- aszimptotikusan normális eloszlású:  $\sqrt{nl_1(\vartheta)}(\hat{\vartheta}_n - \vartheta)$  eloszlásban tart a standard normális eloszláshoz minden  $\vartheta \in \Theta$ -ra  $n \rightarrow \infty$  esetén.

## Konfidenciaintervallum az ML-becslés alapján

Ha likelihoodfüggvény teljesít bizonyos regularitási feltételeket, akkor a  $\vartheta$  paraméternek az  $X_1, X_2, \dots, X_n$  mintából számolt  $\hat{\vartheta}_n$  maximumlikelihood-becslése

- létezik;
- aszimptotikusan torzítatlan:  $\lim_{n \rightarrow \infty} \mathbb{E}_{\vartheta}(\hat{\vartheta}_n) = \vartheta$  minden  $\vartheta \in \Theta$ -ra;
- aszimptotikusan hatásos:  $\lim_{n \rightarrow \infty} \sqrt{nI_1(\vartheta)} D_{\vartheta}(\hat{\vartheta}_n) = 1$  minden  $\vartheta \in \Theta$ -ra;
- aszimptotikusan normális eloszlású:  $\sqrt{nI_1(\vartheta)}(\hat{\vartheta}_n - \vartheta)$  eloszlásban tart a standard normális eloszláshoz minden  $\vartheta \in \Theta$ -ra  $n \rightarrow \infty$  esetén.

Ez alapján **aszimptotikus** konfidenciaintervallum, ami  $n \rightarrow \infty$  esetén  $1 - \alpha$ -hoz tartó valószínűséggel tartalmazza  $\vartheta$ -t:

$$\left( \hat{\vartheta}_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{\sqrt{\hat{I}_n(\vartheta)}}; \hat{\vartheta}_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{\sqrt{\hat{I}_n(\vartheta)}} \right),$$

ahol  $\hat{I}_n(\vartheta) = n \cdot I_1(\hat{\vartheta}_n)$ , vagyis a Fisher-információ kifejezésébe a maximumlikelihood-becslést írjuk be.

## Házi feladat március 20., 9:00-ig

Legyen  $X_1, X_2, \dots, X_n$  független,  $n = 10000$  elemű,  $N(m, \sigma)$  eloszlású minta, ahol  $m \neq 0$  és  $\sigma \neq 1$  tetszőlegesen választott értékek.

Csoportosítsuk a mintaelemeket úgy, hogy minden csoportba tíz megfigyelés essen (első tíz, második tíz, stb.) Legyen  $a_j$  a  $j$ . csoportba eső megfigyelések átlaga,  $s_j^*$  pedig a  $j$ . csoportba eső megfigyelések korrigált tapasztalati szórása ( $j = 1, 2, \dots, 1000$ ). Legyen továbbá

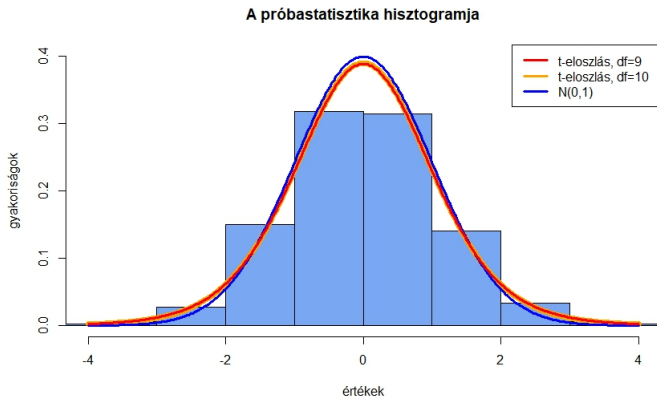
$$Z_j = \frac{a_j - m}{s_j^*} \cdot \sqrt{10} \quad (j = 1, 2, \dots, 1000).$$

Készítsünk a  $Z_j$  értékekből hisztogramot, és ábrázoljuk ezt együtt az  $f = 9$  szabadsági fokú  $t$ -eloszlás, az  $f = 10$  szabadsági fokú  $t$ -eloszlás, illetve a standard normális eloszlás sűrűségfüggvényével (lehet egy ábrán, de három külön ábrán is). Melyik sűrűségfüggvény illeszkedik a legjobban? Melyik a  $Z_j$  valódi sűrűségfüggvénye?

## Házi feladat március 20-ig, megoldás

```
minta<-rnorm(10000, m=12, sd=4)
a=1:1000; for(j in 1:1000){a[j]=mean(minta[((j-1)*10+1):(j*10)])}
s=1:1000; for(j in 1:1000){s[j]=sd(minta[((j-1)*10+1):(j*10)])}
z=(a-12)/s*sqrt(10)
hist(z, col="#79a7f2", main="A próbastatisztika hisztogramja", xlab="értékek",
ylab="gyakoriságok", freq=F, ylim=c(0,0.4), xlim=c(-4,4))
x=seq(from=-4, to=4, by=0.02); y=dt(x, df=9); v=dt(x, df=10); u=dnorm(x)
lines(v x, lwd="7", col="orange")
lines(u x, lwd="3", col="blue")
lines(y x, lwd="3", col="red")
legend("topright", c("t-eloszlás, df=9", "t-eloszlás, df=10", "N(0,1)"), col=c("red",
"orange", "blue"), lwd="3")
```

# Házi feladat március 20-ig, megoldás



A  $t = \frac{\bar{X} - m}{s_n^*} \sqrt{n}$  próbastatisztikából készült hisztogram, a  $t$ -eloszlás sűrűségfüggvénye 9 és 10 szabadsági fokokkal, valamint a standard normális eloszlás sűrűségfüggvénye

# Bayes-becslések

Eddig: **frekventista hozzáállás**, a  $\Theta$  paraméterter rögzített halmaz saját struktúra nélkül, a tulajdonságoknak (például konzisztencia, torzítatlanság) minden  $\vartheta \in \Theta$ -ra teljesülniük kell.

Hátrány például: egy szabályosnak tűnő érmével  $n$  dobásból mindegyik fej lett. Legyen  $p$  az írás valószínűsége. Ezt relatív gyakorisággal becsüljük:  $\hat{p} = 0$  (ez torzítatlan, konzisztens), az  $n$ -től függetlenül.

# Bayes-becslések

Eddig: **frekventista hozzáállás**, a  $\Theta$  paraméterter rögzített halmaz saját struktúra nélkül, a tulajdonságoknak (például konzisztencia, torzítatlanság) minden  $\vartheta \in \Theta$ -ra teljesülniük kell.

Hátrány például: egy szabályosnak tűnő érmével  $n$  dobásból mindegyik fej lett. Legyen  $p$  az írás valószínűsége. Ezt relatív gyakorisággal becsüljük:  $\hat{p} = 0$  (ez torzítatlan, konzisztens), az  $n$ -től függetlenül.

**Bayes-i hozzáállás**: a paramétert magát is valószínűségi változónak tekintjük, ebbe beépítve valamilyen előzetes információt. Például a pénzérménél, mivel szabályosnak tűnik, azt tesszük fel, hogy nagy valószínűséggel az  $1/2$ -hez közel van az értéke. Például feltesszük, hogy az eloszlása beta-eloszlású a  $[0, 1]$  intervallumon.

# Beta-eloszlás

## Definíció

Az  $X$  valószínűségi változó beta-eloszlású  $a, b$  paraméterekkel, ha sűrűségfüggvénye:

$$f(x) = \frac{(a+b-1)!}{(a-1)!(b-1)!} x^{a-1}(1-x)^{b-1} dx, \quad \text{ha } x \in [0, 1],$$

és nulla különben.

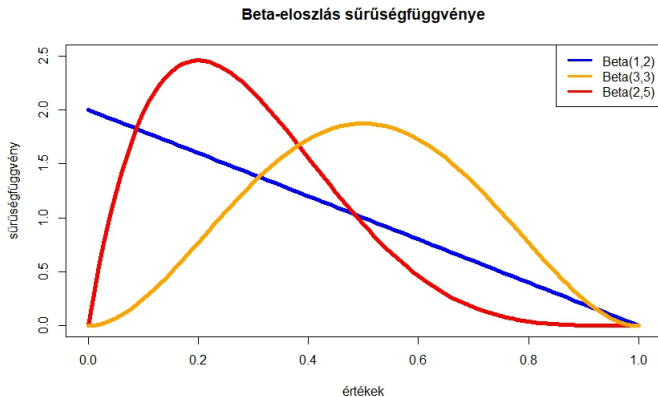
Az  $(a, b)$  paraméterű beta-eloszlás várható értéke és szórása:

$$\mathbb{E}(X) = \frac{a}{a+b}; \quad D(X) = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}.$$

Legyen  $X_1, X_2, \dots, X_n$  független minta a  $[0, 1]$  intervallumon egyenletes eloszlásból. Ekkor  $X_k^*$ -nak, vagyis a nagyság szerint  $k$ . legnagyobb mintaelemnek az eloszlása beta-eloszlás  $a = k$  és  $b = n - k + 1$  paraméterekkel. Következmény:

$$\mathbb{E}(X_k^*) = \frac{k}{n+1}; \quad D(X_k^*) = \sqrt{\frac{k(n-k+1)}{(n+1)^2(n+2)}}.$$

# Beta-eloszlás



Beta-eloszlás sűrűségfüggvénye különböző paraméterpárok mellett

## A priori és a posteriori eloszlás

- **a priori eloszlás:** eloszlás a  $\Theta$  paraméterterén, sűrűségfüggvénye legyen  $\pi$ . Ez tartalmazza a paraméterről az előzetes információt, feltevést (az előző példában  $\pi$  lehet beta(3,3) eloszlás sűrűségfüggvénye)
- **prediktív eloszlás:** a minta eloszlása az a priori eloszlás alapján, feltétel nélkül. Sűrűségfüggvénye:

$$f_{\pi}(x) = \int_{\Theta} f_{\vartheta}(x)\pi(\vartheta)d\vartheta,$$

ahol  $f_{\vartheta}(x)$  a minta sűrűségfüggvénye a  $\vartheta$  paraméter mellett.

## A priori és a posteriori eloszlás

- **a priori eloszlás:** eloszlás a  $\Theta$  paraméterterén, sűrűségfüggvénye legyen  $\pi$ . Ez tartalmazza a paraméterről az előzetes információt, feltevést (az előző példában  $\pi$  lehet  $\text{beta}(3,3)$  eloszlás sűrűségfüggvénye)
- **prediktív eloszlás:** a minta eloszlása az a priori eloszlás alapján, feltétel nélkül. Sűrűségfüggvénye:

$$f_{\pi}(x) = \int_{\Theta} f_{\vartheta}(x)\pi(\vartheta)d\vartheta,$$

ahol  $f_{\vartheta}(x)$  a minta sűrűségfüggvénye a  $\vartheta$  paraméter mellett.

- **a posteriori eloszlás:** a minta megfigyelt értékeire feltételesen mi lesz a  $\vartheta$  paraméter eloszlása a  $\Theta$  paraméterterén. Sűrűségfüggvénye:

$$\pi^*(\vartheta|\underline{X} = \underline{x}) = \frac{L_{\vartheta}(X_1, \dots, X_n)\pi(\vartheta)}{L_{\pi}(X_1, \dots, X_n)},$$

ahol  $L_{\vartheta}$  a likelihood-függvény  $\vartheta$  mellett,  $L_{\pi}$  pedig az  $f_{\pi}$ -ből számolt likelihood-függvény:  $L_{\pi}(X_1, \dots, X_n) = \prod_{j=1}^n f_{\pi}(X_j)$ .

## A priori és a posteriori eloszlás: példa

Egy érmével dobva  $n$  dobásból  $k$  írás lett. Az írások száma legyen  $Y$ . Az írás (I) valószínűsége  $\vartheta$ . Az **a priori eloszlás** legyen  $\text{beta}(a,b)$  a  $\Theta = [0, 1]$  paraméterterén. Sűrűségfüggvénye  $\pi$ .

**Prediktív eloszlás:** annak valószínűsége, hogy írást dobunk, vagyis (diszkrét eloszlásnál sűrűségfüggvény helyett a valószínűséget használhatjuk):

$$\mathbb{P}_\pi(I) = \int_{\Theta} \mathbb{P}_{\vartheta}(I) \pi(\vartheta) d\vartheta = \int_0^1 \vartheta \cdot \pi(\vartheta) d\vartheta = \frac{a}{a+b},$$

hiszen éppen a  $\text{beta}(a,b)$  eloszlás várható értéke jelent meg.

**A posteriori eloszlás:** feltéve, hogy  $n$  dobásból  $k$  írás lett, mi a  $\vartheta$  paraméter eloszlása a  $[0, 1]$  intervallumon. Ez  $\text{beta}$ -eloszlás  $k+a$  és  $n-k+b$  paraméterekkel:

$$\begin{aligned} \pi^*(\vartheta | Y = k) &= \frac{L_{\vartheta}(X_1, \dots, X_n) \pi(\vartheta)}{L_{\pi}(X_1, \dots, X_n)} = \frac{\mathbb{P}_{\vartheta}(Y = k) \pi(\vartheta)}{\mathbb{P}_{\pi}(Y = k)} = \\ &= \frac{\binom{n}{k} \vartheta^k (1 - \vartheta)^{n-k} \cdot \frac{(a+b-1)!}{(a-1)!(b-1)!} \vartheta^{a-1} (1 - \vartheta)^{b-1}}{\binom{n}{k} \left(\frac{a}{a+b}\right)^k \left(1 - \frac{a}{a+b}\right)^{n-k}} = \\ &= C_{a,b} \vartheta^{k+a-1} (1 - \vartheta)^{n-k+b-1}. \end{aligned}$$

# Bayes-becslés

**Veszteségfüggvény:** a veszteség, ha az igazi  $\vartheta$  paraméter helyett annak  $\hat{\vartheta}$  becslését használjuk, ez  $W(\vartheta, \hat{\vartheta})$ . Ez nemnegatív,  $\vartheta - \hat{\vartheta}$  függvénye, például  $(\vartheta - \hat{\vartheta})^2$  vagy  $|\vartheta - \hat{\vartheta}|$ .

## Definíció

A  $\vartheta \in \Theta$  paraméter Bayes-becslése az a  $\hat{\vartheta} = T(X_1, \dots, X_n)$  becslés, melyre az

$$R_Q(T) = \mathbb{E}_\vartheta(\mathbb{E}(W(T(X_1, \dots, X_n), \vartheta)))$$

a priori bayesi rizikót.

# Bayes-becslés

**Veszteségfüggvény:** a veszteség, ha az igazi  $\vartheta$  paraméter helyett annak  $\hat{\vartheta}$  becslését használjuk, ez  $W(\vartheta, \hat{\vartheta})$ . Ez nemnegatív,  $\vartheta - \hat{\vartheta}$  függvénye, például  $(\vartheta - \hat{\vartheta})^2$  vagy  $|\vartheta - \hat{\vartheta}|$ .

## Definíció

A  $\vartheta \in \Theta$  paraméter Bayes-becslése az a  $\hat{\vartheta} = T(X_1, \dots, X_n)$  becslés, melyre az

$$R_Q(T) = \mathbb{E}_{\vartheta}(\mathbb{E}(W(T(X_1, \dots, X_n), \vartheta)))$$

a priori bayesi rizikót.

## Tétel

Ha a  $W(x, y) = (x - y)^2$  négyzetes veszteségfüggvényt használjuk, akkor a paraméter  $g(\vartheta)$  függvényének Bayes-becslése az a  $g$  várható értéke az a posteriori eloszlás szerint:

$$\widehat{g(\vartheta)} = \int_{\Theta} g(\vartheta) \pi^*(\vartheta | \underline{X} = \underline{x}) d\vartheta.$$

## A priori és a posteriori eloszlás: példa

Egy érmével dobva  $n$  dobásból  $k$  írás lett. Az írások száma legyen  $Y$ . Az írás (l) valószínűsége  $\vartheta$ . Az **a priori eloszlás** legyen  $\text{beta}(a,b)$  a  $\Theta = [0, 1]$  paraméterterén. Sűrűségfüggvénye  $\pi$ . Az **a posteriori eloszlás**  $\text{beta}$ -eloszlás  $k + a$  és  $n - k + b$  paraméterekkel.

$W(x, y) = (x - y)^2$  négyzetes veszteségfüggvény esetén a  $\vartheta$  paraméter becslése:

$$\hat{\vartheta} = \int_{\Theta} \vartheta \cdot \pi^*(\vartheta | \underline{X} = \underline{x}) d\vartheta = \int_0^1 \vartheta \cdot \pi^*(\vartheta | \underline{X} = \underline{x}) d\vartheta = \frac{k + a}{n - k + b},$$

hiszen éppen a  $k + a$  és  $n - k + b$  paraméterű  $\text{beta}$ -eloszlás várható értéke jelent meg.

Tehát például ha  $n = 10$  dobásból  $k = 0$  írás van, és  $\text{beta}(2,2)$  az a priori eloszlás, akkor

$$\hat{\vartheta} = \frac{2}{12} = \frac{1}{6} = 16,67\%.$$

Tehát például ha  $n = 100$  dobásból  $k = 0$  írás van, és  $\text{beta}(2,2)$  az a priori eloszlás, akkor

$$\hat{\vartheta} = \frac{2}{102} = 1,96\%.$$

## Házi feladat március 27., 9:00-ig

Tegyük fel, hogy az emberek testmagassága normális eloszlású, várható értéke  $m$  ismeretlen paraméter, szórása  $s = 10$ . Az  $m$  paraméterről tegyük fel, hogy az apriori eloszlása normális, várható értéke  $\mu$ , szórása  $\sigma$ .

- Határozzuk meg az  $m$  paraméter Bayes-bebecslését a húszelemű mintából (ez  $\mu$ -nek és  $\sigma$ -nak egy függvénye).
- A kapott kifejezést ábrázoljuk  $\mu$  függvényében, egy tetszőlegesen rögzített, de 15-nél nem nagyobb  $\sigma$  mellett.
- A kapott kifejezést ábrázoljuk  $\sigma$  függvényében, egy tetszőlegesen rögzített, 170 és 180 közötti  $\mu$  érték mellett.