

# Statisztikai becslések, konfidenciaintervallumok. Maximum likelihood- és momentumbecslés. Becslések tulajdonságai. A hipotézisvizsgálat alapfogalmai. Klasszikus próbák. Lineáris modellek, regresszió.

## Alapfogalmak

### Statisztikai minta

Az  $X_1, X_2, \dots, X_n$  közös eloszlású, egymástól független val. változók rendszerét *statisztikai mintának* nevezzük. Konkrét minta: ezeknek egy konkrét  $x_1, x_2, \dots, x_n$  előfordulása. (A két fogalom azonban gyakran keveredik, az  $X_i$  és  $x_i$  jelölést mindkettőre szokták használni.)

### Becslés

Legyen  $x_1, \dots, x_n$  egy statisztikai minta, a mintát jellemző  $a$  paramétert egy  $\hat{a}_n = T_n(x_1, \dots, x_n)$  *statisztikai függvény* vagy *statisztika* értéke alapján határozzuk meg. Az  $\hat{a}_n$ -t az  $a$  paraméter *becslésének* nevezzük. Az olyan becsléseket, amelyek megadott paramétereket becsülnek meg, paraméter vagy *pontbecsléseknek* nevezzük.

## Statisztikai függvények, összefüggések

- *Középérték* vagy *empírikus közép* vagy *átlag*

$$\bar{x}_n = \bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- *Empírikus szórásnégyzet*

$$S_n^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

vagy

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

alapján is számítható.

- *Korrigált empírikus szórásnégyzet*

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

- Ha a minta elemeit nagyság szerint rendezzük, akkor  $x_1^* \leq x_2^* \leq \dots \leq x_n^*$  *rendezett mintát* kapunk.

- Minta *középpontja*:  $\frac{x_1^* + x_n^*}{2}$

- Minta *terjedelme*:  $x_n^* - x_1^*$

- Minta *mediánja*:  $= \begin{cases} x_m^* & \text{ha } n = 2m - 1 \\ \frac{x_m^* + x_{m+1}^*}{2} & \text{ha } n = 2m \end{cases}$

- Az átlag várható értéke, ha  $E(x_i) = \mu$  minden  $i$ -re:

$$E(\bar{x}) = E\left(\frac{x_1 + \dots + x_n}{n}\right) = \frac{1}{n} E(x_1 + \dots + x_n) = \frac{1}{n} [E(x_1) + \dots + E(x_n)] = \frac{1}{n} n\mu = \mu$$

- Az átlag szórásnégyzete, ha  $D^2(x_i) = \sigma^2$  minden  $i$ -re és  $x_i$ -k függetlenek:

$$D^2(\bar{x}) = D^2\left(\frac{x_1 + \dots + x_n}{n}\right) = \frac{1}{n^2} D^2(x_1 + \dots + x_n) = \frac{1}{n^2} [D^2(x_1) + \dots + D^2(x_n)] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- Tehát röviden:  $E(\bar{x}) = \mu$  és  $D^2(\bar{x}) = \frac{\sigma^2}{n}$ . Ekkor az  $u := \sqrt{n} \frac{\bar{x} - \mu}{\sigma}$  val. változó  $N(0, 1)$  (standard normális) eloszlású.

## Paraméter becslések (pontbecslések)

Az  $x_1, \dots, x_n$  statisztikai minta  $\hat{a}_n = T_n(x_1, \dots, x_n)$  függvényét képezzük.

**Definíció 1** Az  $\hat{a}_n$  **torzítatlan becslése** az a paraméternek, ha  $E(\hat{a}_n) = a$ .

**Tétel 1** Az empirikus közép a várható érték torzítatlan becslése, azaz  $E(\bar{x}_n) = \mu$ .

A korrigált empirikus szórásnégyzet torzítatlan becslése a szórásnégyzetnek, azaz  $E(S_n^{*2}) = \sigma^2$ .

Az empirikus szórásnégyzet torzított becslése a szórásnégyzetnek:  $E(S_n^2) = \frac{n-1}{n}\sigma^2$

A relatív gyakoriság torzítatlan becslése a valószínűségnek.

**Definíció 2** Az olyan becslést, amely  $n \rightarrow \infty$  esetén torzítatlanná válik **aszimptotikusan torzítatlan becslésnek** nevezzük.

**Tétel 2** Az empirikus szórásnégyzet aszimptotikusan torzítatlan becslése a szórásnégyzetnek.

**Definíció 3** Két torzítatlan becslés közül  $\hat{a}_{n,1}$  **hatásosabb becslés**  $\hat{a}_{n,2}$ -nél, ha  $D^2(\hat{a}_{n,1}) \leq D^2(\hat{a}_{n,2})$ . (ahol  $D^2(\hat{a}_n) = E((\hat{a}_n - a)^2)$ ). Ha létezik olyan  $\hat{a}_{min}$ , melyre  $D^2(\hat{a}_{min}) \leq D^2(\hat{a}_n)$  minden  $\hat{a}_n$ -re, akkor  $\hat{a}_{min}$  a **leghatékonyabb hatásos becslés**, röviden **hatásos becslés**.

**Definíció 4** Ha az  $\hat{a}_n$  becslés sztochasztikusan konvergál az a paraméterhez, akkor a becslést **konzisztensnek** nevezzük. Sztochasztikus konvergencia:  $\lim_{n \rightarrow \infty} P(|\hat{a}_n - a| > \varepsilon) = 0$ . Az  $\hat{a}_n$  torzítatlan becslés **erősen konzisztens becslése** az a paraméternek, ha  $D^2(\hat{a}_n) \rightarrow 0, n \rightarrow \infty$ .

**Tétel 3** A relatív gyakoriság az összes torzítatlan becslés közül a legkisebb szórással rendelkező becslése a valószínűségnek. Az  $\bar{x}_n$  a  $\mu$  várható érték konzisztens becslése. Az  $S_n^{*2}, S_n^2$  a  $\sigma^2$  szórásnégyzet konzisztens becslései.

**Definíció 5** Azt a becslést, amely az adott paraméterre a legtöbb információt nyújtja **elégséges becslésnek** nevezzük.

**Tétel 4** Binomiális eloszlás esetén a relatív gyakoriság elégséges becslése a valószínűségnek.

$N(\mu, \sigma)$  normális eloszlás esetén  $(\bar{x}_n, S_n^{*2})$  elégséges becslése a  $(\mu, \sigma^2)$  paramétereknek.

## Intervallumbecslések

Kérdés: Az  $\hat{a}_n = T_n(x_1, \dots, x_n)$  torzítatlan statisztika milyen közel esik az a paraméter valódi értékéhez?

Megfogalmazva: Megadunk  $0 < \varepsilon < 1$  valószínűséget és konstruálunk  $\hat{a}_1, \hat{a}_2$  statisztikákat (ezek számok), melyekre  $P(\hat{a}_1 < a < \hat{a}_2) = 1 - \varepsilon$ .

**Definíció 6** A fenti tulajdonságú  $(\hat{a}_1, \hat{a}_2)$  intervallum a **konfidencia intervallum**, az  $\hat{a}_1, \hat{a}_2$  számok a **konfidencia (megbízhatósági) határok**.

1.  $N(\mu, \sigma)$  eloszlású statisztikai sokaság  $\mu$  várható értékére konfidenciaintervallum *ismert*  $\sigma$  szórással esetén (a-priori)

Tudjuk:  $E(\bar{x}_n) = \mu, D^2(\bar{x}_n) = \sigma^2$  esetén a  $u := \sqrt{n} \frac{\bar{x}_n - \mu}{\sigma}$  val. változó  $N(0, 1)$  eloszlású.

Keressük  $\beta$ -t, melyre  $P(|u| \leq \beta) = 1 - \varepsilon$ . Ebből:  $\Phi(\beta) = 1 - \frac{\varepsilon}{2}$ , az ennek megfelelő  $\beta$  táblázatból kikereshető. A konfidenciaintervallum:

$$[\bar{x} - \beta \frac{\sigma}{\sqrt{n}}, \bar{x} + \beta \frac{\sigma}{\sqrt{n}}]$$

2.  $N(\mu, \sigma)$  eloszlású statisztikai sokaság  $\mu$  várható értékére konfidenciaintervallum *nem ismert*  $\sigma$  szórással esetén  
Ekkor  $t = \sqrt{n} \frac{\bar{x}_n - \mu}{S_n^*}$  egy  $(n - 1)$  szabadsági fokú Student eloszlás. A konfidenciaintervallum:

$$[\bar{x} - \beta \frac{S_n^*}{\sqrt{n}}, \bar{x} + \beta \frac{S_n^*}{\sqrt{n}}]$$

3.  $N(\mu, \sigma)$  eloszlású statisztikai sokaság  $\sigma$  szórására konfidenciaintervallum

Az  $\frac{nS^2}{\sigma^2}$  val. változó  $(n-1)$  szabadsági fokú  $\chi^2$  eloszlású.  $P(\chi^2 > \chi_{\frac{\varepsilon}{2}}^2) = \frac{\varepsilon}{2}$  és  $P(\chi^2 > \chi_{1-\frac{\varepsilon}{2}}^2) = 1 - \frac{\varepsilon}{2}$ .

A konfidenciaintervallum:

$$\left( \frac{nS^2}{\chi_{\frac{\varepsilon}{2}}^2}, \frac{nS^2}{\chi_{1-\frac{\varepsilon}{2}}^2} \right)$$

## Hipotézisvizsgálat

Az  $X$  val. változó  $F(x, a)$  eloszlására vonatkozó feltevést fogalmazzunk meg.

### Csoportosításuk

- Egyik szempont szerint:  
*Egyszerű hipotézis:* Ha a feltételnek csak egyetlen val. eloszlás felel meg.  
*Összetett hipotézis:* Ha a feltételnek több val. eloszlás felel meg.
- Másik szempont szerint:  
*Paraméteres probléma:* Ha valamely val. eloszlás egy vagy több paraméterére vonatkozik a hipotézis.  
*Nem paraméteres probléma:* Legismertebbek az illeszkedésvizsgálat, homogenitás vizsgálat.

Legyen  $\mathcal{C}$  egy bizonyos típusú valószínűségi eloszlásoknak az osztálya. Ezt felbontjuk két nem üres, diszjunkt  $\mathcal{C}_1, \mathcal{C}_2$  részosztályra.

Tekintsünk egy statisztikai mintát:  $x_1, \dots, x_n$

Kérdés: Ha  $x_1, \dots, x_n$  a  $\mathcal{P} \in \mathcal{C}$  eloszlású statisztikai minta, akkor  $\mathcal{P} \in \mathcal{C}_1$ ?

*Nullhipotézis:*  $H_0 : \mathcal{P} \in \mathcal{C}_1$

*Ellenhipotézis:*  $H_1 : \mathcal{P} \in \mathcal{C}_2$

Legyen  $\hat{a} = \hat{a}_n(x_1, \dots, x_n)$  a  $\mathcal{C}_1$  osztályra vonatkozó statisztika. Ekkor léteznek  $\hat{a}_1, \hat{a}_2$  számok, hogy  $P(\hat{a} \notin (\hat{a}_1, \hat{a}_2)) = \varepsilon$ .

A  $\kappa_n(\varepsilon) = \{a : a \notin (\hat{a}_1, \hat{a}_2)\}$  halmazt *kritikus tartománynak* nevezzük. Az  $\hat{a}_n(x_1, \dots, x_n)$  a *próbat statisztika*, az  $\varepsilon$  a *próba terjedelme*, az  $(\hat{a}_1, \hat{a}_2)$  intervallum az *elfogadási tartomány*.

	$H_0$ -t elfogadjuk	$H_0$ -t elvetjük
$H_0$ áll fenn	helyes	elsőfajú hiba
$H_1$ áll fenn	másodfajú hiba	helyes

## Próbák

1. Egymintás u-próba

$N(\mu, \sigma)$  eloszlású sokaságból vett minta,  $\sigma$  ismert.

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Döntési eljárás: Kiszámítjuk  $u$ -t. (Intervallumbecslések 1. pontjánál lévő  $u$ ) Ha  $u$  beleesik az  $(1 - \varepsilon)$ -hoz számolt intervallumba, akkor a hipotézist elfogadjuk, egyébként elvetjük.

2. Egymintás t-próba

$N(\mu, \sigma)$  eloszlású sokaságból vett minta,  $\sigma$  nem ismert.

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Döntési eljárás: Kiszámítjuk  $t$ -t. (Intervallumbecslések 2. pontjánál lévő  $t$ ) Ha  $t$  beleesik az  $(1 - \varepsilon)$ -hoz számolt intervallumba, akkor a hipotézist elfogadjuk, egyébként elvetjük.

3. Kétmintás u-próba

$X \sim N(\mu_1, \sigma_1)$  eloszlású val. változó,  $X_1, \dots, X_n$  stat. minta

$Y \sim N(\mu_2, \sigma_2)$  eloszlású val. változó,  $Y_1, \dots, Y_m$  stat. minta

$\sigma_1$  és  $\sigma_2$  ismert

$$H_0 : E(X) = E(Y), \quad \text{azaz} \quad \mu_1 = \mu_2$$

$$H_1 : E(X) \neq E(Y), \quad \text{azaz} \quad \mu_1 \neq \mu_2$$

Ha  $H_0$  fennáll, akkor  $E(\bar{X} - \bar{Y}) = 0$ .

4. F-próba

$X \sim N(\mu_1, \sigma_1)$  eloszlású val. változó,  $X_1, \dots, X_n$  stat. minta

$Y \sim N(\mu_2, \sigma_2)$  eloszlású val. változó,  $Y_1, \dots, Y_m$  stat. minta

$\sigma_1$  és  $\sigma_2$  ismert

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 \neq \sigma_2$$

5.  $\chi^2$ -próbák

(a) Illeszkedésvizsgálat

$A_1, \dots, A_r$  teljes eseményrendszer,  $p_1, \dots, p_r$  pozitív számok, összegük 1.

$$H_0 : P(A_i) = p_i$$

$$H_1 : P(A_i) \neq p_i$$

(b) Homogenitásvizsgálat

$X : x_1, \dots, x_n$  és  $Y : y_1, \dots, y_n$  stat. minták.

$$H_0 : X, Y \text{ azonos eloszlásúak}$$

$$H_1 : X, Y \text{ nem azonos eloszlásúak}$$

(c) Függetlenségvizsgálat  $A_1, \dots, A_r$  és  $B_1, \dots, B_s$  teljes eseményrendszerek.

$$H_0 : P(A_i B_j) = P(A_i) P(B_j) \quad \forall i, j$$

$$H_1 : \text{az egyenlőség nem minden } i, j\text{-re teljesül}$$

## Lineáris regresszió

Legyen  $X, Y$  két val. változó, véges, nem nulla szórással. A két val. változó között

$$Y = aX + b$$

lineáris kapcsolatot feltételezünk. Olyan  $a, b$ -t keresünk, melyre

$$E((Y - aX - b)^2)$$

minimális. Ez a legkisebb négyzetek módszere.

Legyen  $(X, Y)$  párra stat. minta  $(x_1, y_1), \dots, (x_n, y_n)$ . Olyan  $a, b$ -t keresünk, melyre

$$f(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

minimális. A szélsőérték feltétele:

$$\frac{\partial f}{\partial a} = 0 \quad \text{és} \quad \frac{\partial f}{\partial b} = 0$$

A parciális deriválást elvégezve és átrendezve:

$$a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i$$
$$a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i$$

Ennek a megoldása  $\hat{a}, \hat{b}$ , az  $a, b$  becslése. A regressziós egyenes egyenlete:

$$y = \hat{a}x + \hat{b} = \hat{r} \frac{S_y}{S_x} (x - \bar{x}) + \bar{y}$$

ahol  $S_x, S_y$  az empirikus szórások,  $\hat{r}$  a korrelációs együttható becslése.