

Becslések

Olyan esetekben, amikor valamiért nem tudjuk vagy nem akarjuk a teljes sokaságot megvizsgálni, hogy meghatározzuk a fontosabb statisztikai mutatóit, becslést alkalmazunk. A becslés lényege, hogy egy minta alapján próbálunk ezekre a mutatókra következtetni.

[Megnézem a kapcsolódó epizódot](#)

A megbízhatósági szintet konfidencia szintnek nevezzük. A konfidencia szint szokásos jelölése $1 - \alpha$.

[Megnézem a kapcsolódó epizódot](#)

Az $1 - \alpha$ megbízhatósági szinthez, vagy másként konfidencia szinthez tartozó konfidencia intervallumok azok az intervallumok, amik a sokasági átlagot $1 - \alpha$ valószínűséggel tartalmazzák.

A konfidencia intervallum végpontjai:

$$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ ahol}$$

$\bar{x}$ = a minta átlaga

n = a minta elemszáma

σ = a teljes [sokaság](#) szórása

$Z_{1-\frac{\alpha}{2}}$ pedig a [standard normális eloszlás](#) $1 - \frac{\alpha}{2}$ -höz tartozó Z értéke.

[Megnézem a kapcsolódó epizódot](#)

$$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ ahol}$$

$\bar{x}$ = a minta átlaga

n = a minta elemszáma

σ = a teljes [sokaság](#) szórása

$Z_{1-\frac{\alpha}{2}}$ pedig a [standard normális eloszlás](#) $1 - \frac{\alpha}{2}$ -höz tartozó Z értéke.

[Megnézem a kapcsolódó epizódot](#)

A FAE minta azt jelenti, hogy a [mintavétel](#) során bármely mintaelemet azonos eséllyel választunk ki. Ilyen a visszatevéses [mintavétel](#), vagy pedig abban az esetben ha az alap [sokaság](#) elemszáma nagyon nagy, akkor a visszatevés nélküli [mintavétel](#) is.

[Megnézem a kapcsolódó epizódot](#)

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \text{ ahol}$$

$1 - \alpha =$ konfidencia szint

$\bar{x} =$ a minta átlaga

$n =$ a minta elemszáma

$s =$ a teljes [sokaság](#) szórása, a sokasági szórás nem ismert

$t_{1-\frac{\alpha}{2}}$ pedig a [t-eloszlás](#) $1 - \frac{\alpha}{2}$ -höz tartozó értéke.

[Megnézem a kapcsolódó epizódot](#)

$$\bar{p} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \text{ ahol}$$

$1 - \alpha =$ konfidencia szint

$\bar{p} =$ a minta alapján kapott valószínűség

$n =$ a minta elemszáma

$Z_{1-\frac{\alpha}{2}}$ pedig a [standard normális eloszlás](#) $1 - \frac{\alpha}{2}$ -höz tartozó Z értéke.

[Megnézem a kapcsolódó epizódot](#)

$$\frac{(n-1) \cdot s^2}{\chi^2_{1-\frac{\alpha}{2}}(v)} < \sigma^2 < \frac{(n-1) \cdot s^2}{\chi^2_{\frac{\alpha}{2}}(v)} \text{ ahol}$$

$1 - \alpha =$ konfidencia szint

$\bar{p} =$ a minta alapján kapott valószínűség

$n =$ a minta elemszáma

$s =$ a minta szórása, a sokasági szórás nem ismert

$\chi^2(v)$ pedig a khi-négyzet [eloszlás](#) megfelelő értéke

[Megnézem a kapcsolódó epizódot](#)

Az EV-minta abban különbözik a FAE-mintától, hogy a kiválasztott mintaelemek nem függetlenek egymástól.

Ez olyankor fordulhat elő, ha a teljes [sokaság](#) mérete viszonylag kicsi a minta elemszámához képest. EV-minták esetén tehát a minta fontos jellemzőjévé válik, hogy mekkora a teljes [sokaság](#), amelynek elemszámát N jelöli.

[Megnézem a kapcsolódó epizódot](#)

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \text{ ahol}$$

$1 - \alpha$ = konfidencia szint

$\bar{x}$ = a minta átlaga

n = a minta elemszáma

N = a teljes [sokaság](#) elemszáma

s = a minta szórása

$t_{1-\frac{\alpha}{2}}$ pedig a [t-eloszlás](#) $1 - \frac{\alpha}{2}$ -höz tartozó értéke.

[Megnézem a kapcsolódó epizódot](#)

$$\bar{p} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \cdot \sqrt{1 - \frac{n}{N}} \text{ ahol}$$

$1 - \alpha$ = konfidencia szint

$\bar{p}$ = a minta alapján kapott valószínűség

n = a minta elemszáma

N = a teljes [sokaság](#) elemszáma

$Z_{1-\frac{\alpha}{2}}$ pedig a [standard normális eloszlás](#) $1 - \frac{\alpha}{2}$ -höz tartozó Z értéke.

[Megnézem a kapcsolódó epizódot](#)

Ha a teljes sokaságot felosztjuk viszonylag homogén rétegekre, és a mintát is ezen a rétegek szerint vizsgáljuk, a variancia csökkenthető.

$$\hat{\bar{x}}_R \pm Z_{1-\frac{\alpha}{2}} \cdot s_{\hat{\bar{X}}_R}$$

$1 - \alpha$ = konfidencia szint

$\bar{x}$ = a minta átlaga

n = a minta elemszáma

n_j = a minta j -edik rétegének elemszáma

N = a teljes [sokaság](#) elemszáma

N_j = a teljes [sokaság](#) j -edik rétegének elemszáma

W_j = a teljes [sokaság](#) j -edik rétegének a teljes sokasághoz viszonyított aránya

s_j = a minta j -edik rétegének szórása

$$\hat{\bar{X}}_R = \sum_{j=1}^M W_j \bar{x}_j$$

$$s_{\hat{\bar{X}}_R}^2 = \sum_{j=1}^M W_j^2 \frac{s_j^2}{n_j} \left(1 - \frac{n_j}{N_j}\right)$$

[Megnézem a kapcsolódó epizódot](#)

A kétmintás becslésekre akkor van szükség, amikor két [sokaság](#) valamilyen paraméterét, leginkább az átlagát szeretnénk összehasonlítani.

A kétmintás [becslések](#) lehetnek független mintás [becslések](#) vagy páros mintás [becslések](#).

[Megnézem a kapcsolódó epizódot](#)

Ha mindkét [sokaság](#) közel normális eloszlású, akkor az [átlagok](#) különbségének becslésére ez a formula van forgalomban.

$$d \pm t_{1-\frac{\alpha}{2}} \cdot s_d \text{ ahol } d = \bar{x} - \bar{y}$$

$$s_d = s_c \cdot \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}} \text{ itt } s_c^2 = \frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}$$

$1 - \alpha$ = konfidencia szint

$\bar{x}$ = az egyik minta átlaga

$\bar{Y}$ = a másik minta átlaga

n_X = az egyik minta elemszáma

n_Y = a másik minta elemszáma

A szabadságfok $v = n_X + n_Y - 2$

[Megnézem a kapcsolódó epizódot](#)

Egy becslést torzítatlannak nevezünk, ha az egyes mintákból kapott [becslések](#) várható értéke megegyezik a becslni kívánt mennyiséggel.

Ez a tulajdonság azt jelenti, hogy a becslés során kapott értékek a becslni kívánt érték körül ingadoznak, és ez az ingadozás szimmetrikus. A torzítatlan becsléseket mindig előnyben részesítjük a torzítottakkal szemben.

[Megnézem a kapcsolódó epizódot](#)

A kérdés az, hogy ha egy sokasági jellemzőre több becslés jöhet szóba, hogyan válasszunk közülük, vagyis mikor tekintünk egy becslést jónak, kettő közül melyiket tekintjük jobbnak és kijelenthetjük-e valamelyikről, hogy a legjobb?

Két alapvető szempont alapján szoktuk a becsléseket versenyeztetni. Az egyik, a már jól ismert torzítatlanság, vagyis a becslésnek az a tulajdonsága, hogy az összes lehetséges mintán vett [becslések](#) átlaga megegyezik a becslni kívánt sokasági jellemzővel. A másik az úgynevezett minimális variancia kritérium.

A minimális variancia kritérium azt jelenti, hogy ha van két torzítatlan becslésünk, akkor a kettő közül azt tekintjük jobbnak, aminek az összes mintán vett értékeinek varianciája kisebb.

[Megnézem a kapcsolódó epizódot](#)

$$MSE(\hat{\theta}) = var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$$

Az első tag a varianciát, a második tag a várható értéktől való eltérést, vagyis a torzítottságot méri. Ha a becslés torzítatlan, $E(\hat{\theta}) = \theta$ így ez a második tag nulla. Két becslés közül azt részesítjük előnyben, amelyre MSE kisebb.

Az $E(\hat{\theta}) - \theta$ különbségre, vagyis a torzítás mértékére az angol bias szó alapján a Bs jelölés van forgalomban. Használatos tehát az

$$MSE(\hat{\theta}) = var(\hat{\theta}) + Bs^2(\hat{\theta})$$

Képlet is.

[Megnézem a kapcsolódó epizódot](#)

Eddigi vizsgálódásaink egyik legfontosabb eredménye a mintaátlagot eloszlásának jellemzése. Ha a teljes [sokaság](#) átlaga μ és szórása pedig σ , akkor az ebből vett n elemű minták átlagai olyan eloszlással helyezkednek el, aminek átlaga szintén μ , a szórása pedig $\frac{\sigma}{\sqrt{n}}$.

Ezt az utóbbit a minta standard hibájának szokás nevezni. A standard hiba tehát azt mondja meg, hogy a minta [átlagok](#) mekkora szórással ingadoznak a tényleges sokasági átlag körül.

[Megnézem a kapcsolódó epizódot](#)

Mintavételi hibának azokat a hibákat nevezzük, amik kimondottan azért fordulnak elő, mert nem tudjuk, vagy nem akarjuk a teljes sokaságot vizsgálni. A mintavételi hiba tehát a [sokaság](#) eloszlásán és a mintavételi eljárásen kívül főleg a minta elemszáma határozza meg. Mivel pedig ezeket általában már a mintavételt megelőzően ismerjük, a mintavételi hibának megvan az a kellemes tulajdonsága, hogy legtöbbször előre megállapítható. Vagyis még el sem végeztük a mintavételt, de már tudjuk, hogy mekkora lesz a [mintavétel](#) során elkövetett hiba. Ez a kellemes tulajdonság lesz a kiindulópont a [becslések](#) és később a hipotézisvizsgálatok elméletének kiépítésében.

[Megnézem a kapcsolódó epizódot](#)
