Valószínűség, konfidenciaintervallum

Valószínűség a mindennapokban

Köznyelvi jelentés: tapasztalat alapú becslés (n megfigyelt esetből hányszor történt meg egy adott esemény). Pl. „valószínűleg mindjárt elered az eső” (mert ha ilyen borús az ég, gyakran esik), „valószínűleg idén sem lesz fizetésemelés” (mert tíz éve nem volt).

A valószínűség soha nem jelent biztos tudást! Néha mégsem esik, ha borús az ég, és néha mégis van fizetésemelés.

Intuitív becslésnek kevés fokozata van: *nem túl valószínű, elég valószínű, nagyon valószínű, több mint valószínű*.

Valószínűség a szerencsejátékban

Fej vagy írás egy érme feldobásakor?

Megfigyelés: 10 dobás, 20, 30...

Valószínűség a szerencsejátékban

Fej vagy írás egy érme feldobásakor?

Megfigyelés: 10 dobás, 20, 30...

Fejek száma egyre jobban közelíti a 0,5-ös értéket.

Empirikus valószínűség P definíciója:

P = fej/összes dobás

ahol a dobások száma a végtelenhez közelít.

$\Rightarrow$ valószínűség értéke mindig 0 (egyáltalán nem valószínű) és 1 (teljesen biztos) között mozog.

Példák

1. Adott szám dobása kockával.
2. Ász húzása egy 32 lapos kártyapakliból.
3. Kétszer egymás után fej dobása.
4. Mekkora a valószínűsége annak, hogy egy véletlenszerűen kiválasztott magyar állampolgár katolikus, ha az összes megkérdezett közötti arány katolikus 51%, református 16%, evangélikus 3%, nem vallásos 14,5%.
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%.

Példák

1. Adott szám dobása kockával: adott szám/összes szám $= 1/6$ $= 0{,}1667$.

Példák

1. Adott szám dobása kockával: adott szám/összes szám $= 1/6 = 0{,}1667$.
2. Ász húzása egy 32 lapos kártyapakliból: ászok száma/összes kártya $= 4/32 = 0{,}125$.

Példák

1. Adott szám dobása kockával: adott szám/összes szám $= 1/6 = 0{,}1667$.
2. Ász húzása egy 32 lapos kártyapakliból: ászok száma/összes kártya $= 4/32 = 0{,}125$.
3. kétszer egymás után fej dobása:
   (fej+fej)+(fej+írás)+(írás+fej)+(írás+írás) $= 1/4 = 0{,}25$.

Példák

1. Adott szám dobása kockával: adott szám/összes szám $= 1/6 = 0{,}1667$.
2. Ász húzása egy 32 lapos kártyapakliból: ászok száma/összes kártya $= 4/32 = 0{,}125$.
3. kétszer egymás után fej dobása: **(fej+fej)**+(fej+írás)+(írás+fej)+(írás+írás) $= 1/4 = 0{,}25$.
4. Mekkora a valószínűsége annak, hogy egy véletlenszerűen kiválasztott magyar állampolgár katolikus, ha az összes megkérdezett közötti arány katolikus 51%, református 16%, evangélikus 3%, nem vallásos 14,5%: 51% $= 0{,}51$.

Példák

1. Adott szám dobása kockával: adott szám/összes szám = 1/6 = 0,1667.
2. Ász húzása egy 32 lapos kártyapakliból: ászok száma/összes kártya = 4/32 = 0,125.
3. kétszer egymás után fej dobása: **(fej+fej)**+(fej+írás)+(írás+fej)+(írás+írás) = 1/4 = 0,25.
4. Mekkora a valószínűsége annak, hogy egy véletlenszerűen kiválasztott magyar állampolgár katolikus, ha az összes megkérdezett közötti arány katolikus 51%, református 16%, evangélikus 3%, nem vallásos 14,5%: 51% = 0,51.
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%: 0,224*0,5 = 0,112.

Becslés

Inferenciális statisztika: oksági vagy relációs alapú statisztika. A minta értékei alapján következtet a populációra.

DE: a minta alapján a populációra csak **becsléseket** tehetünk.

1. probléma: különböző minták különböző átlagokat eredményeznek, még véletlenszerű kiválasztás esetén is.

2. probléma: véges az időnk, csak véges mintával tudunk dolgozni.

Konfidenciaintervallum

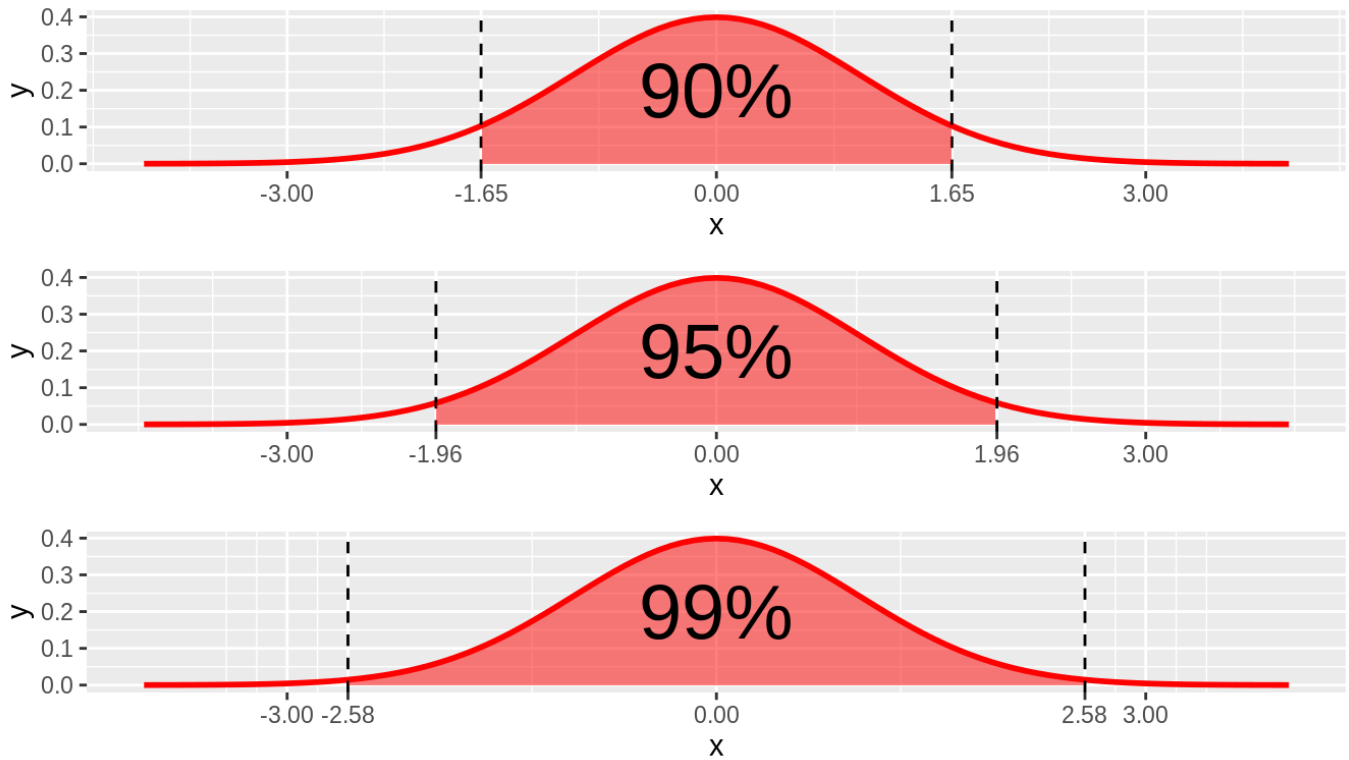
Kérdés: igaz-e, hogy a véletlen mintánk átlaga beleesik az ismeretlen populációátlag körül szóródó mintaátlagokba?

Nehézség: μ-t, vagyis a populáció átlagát nem ismerjük, csak $\bar{x}$-et, azaz a mintaátlagot.
$\Rightarrow$ Döntés nem lehetséges, csak egy adott valószínűségi határon, azaz **konfidenciaintervallumon** belüli valószínűség megállapítása.

A konfidenciaszintet mi határozzuk meg önkényesen. A 95%-os szint azt jelzi, hogy a nullhipotézisről hozott döntésünk 95%-ban lesz megbízható. Jelölése: $p = 0,95$.

A konfidenciaintervallum jelentősége

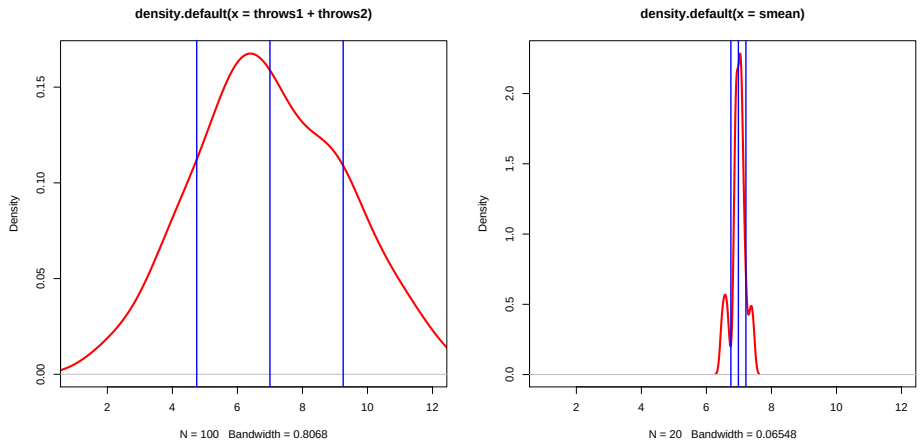
A valószínűségeket a standard normális eloszlás függvénye alapján állapítjuk meg.



$p = 0.95$ esetén a konfidenciaintervallum az átlagtól 1,96 szórásnyira eső két érték közé esik.

A mintától a populációig

Két kockával dobunk 100-szor.
Bal: egy alkalommal dobtunk 100-szor, egy mintánk van.
Jobb: 20-szor dobtunk 100-szor, húsz minta átlagai, amelyek egyenként 100 dobásból állnak.



Egy minta átlaga = 7, 20 minta átlagának átlaga = 6,98.
Egy minta szórása = 2,25, húsz minta átlagainak szórása = 0,23.

```
mean(rep(sample(1:6,100,replace=T)+
+sample(1:6,100,replace=T),20))
```

A normális eloszlás jelentősége

A normális eloszlás nem csak egy adott mintára lehet jellemző, hanem egy adott populációból vett több mintára is.

Feltételezés: n minta $\bar{x}$ átlagai normális eloszlást mutatnak a populáció μ átlaga körül, ha a populáció szórása σ.

A mintaátlagok μ körüli szórása egyenlő a standard hibával (*standard error*), azaz $se = \frac{s}{\sqrt{n}}$ -vel.

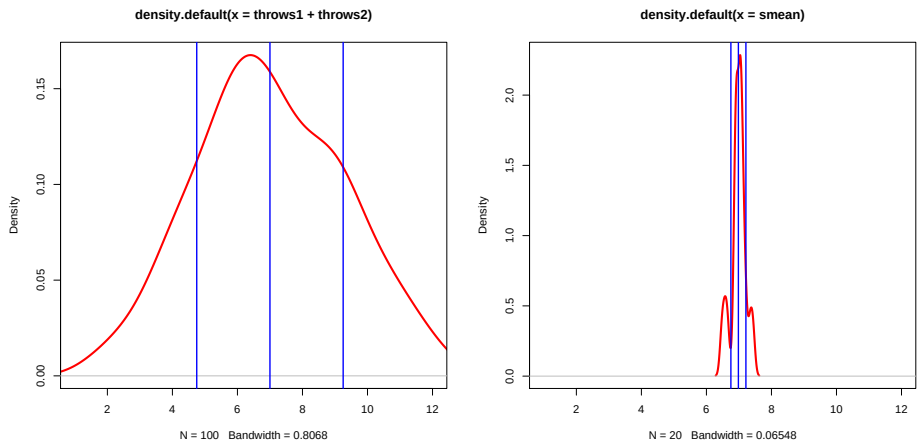
A szórás az egyes adatpontok mintaátlagtól való távolságát fejezi ki.

A standard hiba a mintaátlagok populációátlagtól való távolságát fejezi ki.

Előny: egyetlen minta átlaga és szórása alapján következtethetünk a populáció ismeretlen értékeire.

A minta és a populáció szórása

Két kockával dobunk 100-szor.
Bal: egyetlen, 100 dobásból álló minta összegei. Jobb: 20 minta átlagai, amelyek egyenként 100 dobásból állnak.



Egyetlen minta szórása $= 2{,}25$, 20 minta átlagának szórása $= 0{,}23$.

Standard hiba EGYETLEN mintából számolva:
$2,25/\sqrt{100} = 0,225$
$\rightarrow$ a 20 mintából számolt szórás jó közelítése akkor is, ha csak egyetlen mintánk van.

Pontbecslés

Véletlen minta átlaga függ a véletlentől, azaz egy **becsült pont**.

Mennyire megbízható a becslés egy véletlen minta alapján? A populáció és a minta viszonyát itt egy nagy mintával és az ebből vett részmintákkal demonstráljuk.

Példa: A magyar 18–20 éves férfiak átlagos testmagasságát akarjuk megtudni. A teljes populációt nem tudjuk vizsgálni, ezért veszünk egy mintát: a Névtelen Egyetemre beiratkozó fiúkat mérőléc mellé állítjuk szeptemberben. A minta 300 főből áll. Ennek ismert az átlaga, szórása és a mérete.

$\bar{x} = 178{,}0$ cm
$s = 6{,}3$ cm

Mi történik, ha ebből a mintából kisebb részmintákat veszünk? Mi a részmintákból számolt átlagok szórása a 300 fős minta átlaga körül?

A részmintákból számolt átlagok szórása kiszámolható a teljes minta szórásából és a részminták elemszámából. Standard hiba: $se = \frac{s}{\sqrt{n}}$.

A véletlenül kiválasztott tíz fős minták szórása a teljes minta átlaga körül:
$se = 6,3/\sqrt{10} = 1,99$ cm.

Véletlen ötven fős minták szórása a teljes minta átlaga körül
$se = 6,3/\sqrt{50} = 0,89$ cm.

$\Rightarrow$ Minél nagyobb az elemszám, annál kisebb a standard hiba, vagyis a feltételezett átlagok szórása, Vagyis nagyobb elemszám mellett az egyes mintaátlagok jobban közelítik a nagy minta (általában véve a populáció) átlagát.

Feladat

Fájl letöltése:

http://phon.nytud.hu/mady/courses/statistics/materials/testmagassag.txt

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.
átlag: `mean()`
szórás: `sd()`
gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?

Feladat

Fájl letöltése:

http://phon.nytud.hu/mady/courses/statistics/materials/testmagassag.txt

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.
átlag: `mean()`
szórás: `sd()`
gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?
`mean(testmagassag$height[1:10])`

Teljes minta standard hibája?

Feladat

Fájl letöltése:

http://phon.nytud.hu/mady/courses/statistics/materials/testmagassag.txt

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.
átlag: `mean()`
szórás: `sd()`
gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?
`mean(testmagassag$height[1:10])`

Teljes minta standard hibája?
`sd(testmagassag$height)/sqrt(300)`
0,36

Kiindulás

- A populációból (itt: 300 elemű nagy mintából) véletlenszerűen vett kisebb minták átlagai normális eloszlásúak.
- Normális eloszlás esetén az átlagok 95%-a a populáció átlagától $\pm$ 1,96 egységnyi szórást mutat. Az átlagok szórását a standard hibával jellemezzük, vagyis a $s/\sqrt{n}$ képlettel számoljuk ki.
- Keresssük μ-t, a populáció eloszlásának középpontját.

Tehát:

$p(-1,96 * se + \mu < \bar{x} < \mu + 1,96 * se) = 0,95$

Cél: a 95%-os konfidenciaintervallumon belüli határértékek meghatározása negatív és pozitív irányban.

Konfidenciaintervallum $\bar{x}$ alapján

$$p(-1,96 * se + \mu < \bar{x} < \mu + 1,96 * se) = 0,95$$

$$-\mu$$

$$p(-1,96 * se < \bar{x} - \mu < 1,96 * se) = 0,95$$

$$* -1$$

$$p(1,96 * se > \mu - \bar{x} > -1,96 * se) = 0,95$$

$$+\bar{x}$$

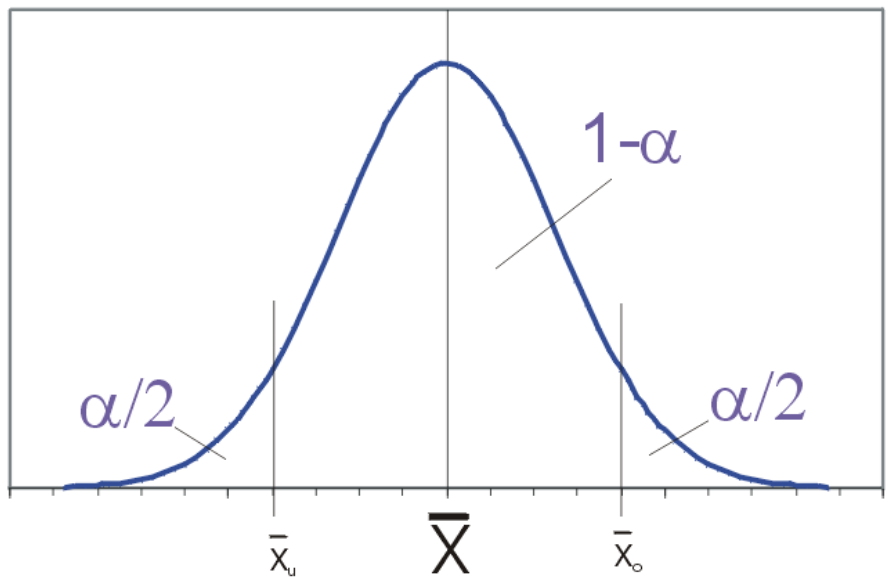
$$p(1,96 * se + \bar{x} > \mu > \bar{x} - 1,96 * se) = 0,95$$

$$p(-1,96 * se + \bar{x} < \mu < \bar{x} + 1,96 * se) = 0,95$$

Konfidenciaszint

Konfidenciaintervallum: értéktartomány, amely a becsülendő paramétert előre rögzített valószínűséggel tartalmazza.

Konfidenciaintervallumon kívüli tartomány: $\alpha = 1 - p$.



Ha $\bar{x}$ nem esik a 95%-os konfidenciaintervallumba, akkor is tartozhat az adott populációhoz! Tévedés valószínűsége 5%, ez az ún. alfa-hiba.

Kiindulási hipotézis tesztelése

Hipotézis állítása falszifikáción keresztül, azaz az állításunk **ellenhipotézisét** teszteljük.

Az empirikus vizsgálatokban általában abban vagyunk érdekeltek, hogy a vizsgált érték $1 - p$, azaz α tartományba essen, hiszen többnyire azt akarjuk megmutatni, hogy a mintánk különbözik egy másik populációtól.
$\Rightarrow$ A szignifikanciaszintet α értékével szokás megadni, azaz 0,05 vagy 5%.

Ha azt akarjuk bizonyítani, hogy egy adott minta NEM tartozik az adott p konfidenciaintervallumba, akkor a mintának negatív és pozitív irányban az $\alpha/2$ tartományba kell tartoznia. Tehát egy szimmetrikus, azaz kétoldalas tesztnél az azonosság elutasítása 2,5%-ra teljesül.

A mintaméret jelentősége a hipotézistesztelés szempontjából

A $p = 0,95$-es konfidenciahatár kritikus értékei:

alsó kritikus érték: $-1.96 * se + \bar{x}$

felső kritikus érték: $+1.96 * se + \bar{x}$

A standard hiba kiszámítása:

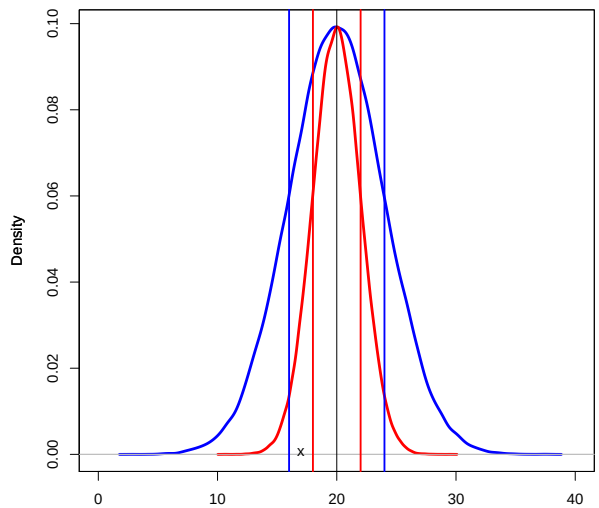
$$se = \frac{s}{\sqrt{n}}$$

Ha n alacsony, a standard hiba nagyobb, mert a nevezőben $\sqrt{n}$ szerepel $\Rightarrow$ nagyobb standard hiba esetén a kritikus érték nagyobb lesz, ezért H_0 elutasításának lehetősége kisebb.

Kisebb és nagyobb standard hiba és hipotézistesztelés

Két populáció, amelyek átlaga $\mu = 20$. 1. populáció szórása $se = 1$, 2. populációé $se = 2$.

Az első populáció alsó kritikus értéke: 20–1*1.96 = 18.04, a másodiké: 20–2*1.96 = 16,08.
Egy minta, amely átlaga $\bar{x} = 17$, ehhez a populációhoz tartozik?



1. populáció: $17 < 18.04 \Rightarrow H_0$-t elutasíthatjuk.
2. populáció: $17 > 16.08 \Rightarrow H_0$-t nem utasíthatjuk el.

Feladat

Számoljuk ki a `testmagassag` R-objektum első tíz elemének átlagát. Beleesik a teljes, 300 elemű minta 95%-os konfidenciaintervallumába?

Első tíz elem átlagának kiszámítása:
`mean(testmagassag$height[1:10])`
175.9037

`gmodels` csomagban `ci()` függvény. Teljes, 300 elemszámú populáció 95%-os konfidenciaintervallumának határai:
`ci(testmagassag$height,0.95) Estimate CI lower CI upper Std. Error 178.0349657 177.3166967 178.7532346 0.3649871`
A teljes populációból vett 10 fős minta átlaga kívül esik a 95%-os konfidenciahatárokon. Ha ismert a populáció átlaga, elutasítanánk azt a hipotézist, hogy az első tíz érték ehhez a populációhoz tartozik.

Feladat

A Fogbarát Cukorgyár ellen panasz érkezik. A márka vásárlói azt állítják, hogy a csomagoláson szereplő 1000 g helyett a csomagban lévő kristálycukor valójában mindig kevesebb. A cukorgyár szerint az értékek átlaga pontosan 1000 g, és valóban előfordul ingadozás, ennek törvényben rögzített szórása 10 g.
A fogyasztóvédelem 50 mintát vesz az ország különböző helyein található üzletekből.
Az eredményt a `cukor` vektorban rögzítik. Letölthető innen:
http://phon.nytud.hu/mady/courses/statistics/materials/cukor.RData

Igaz-e, hogy a fogyasztók átlagosan 1000 g cukrot visznek haza, és hogy a gyártó nem károsítja meg őket a saját javára?

Nullhipotézis: a minta átlaga beleesik az összes cukorcsomagnál megállapított 10 g-os szórás által megszabott konfidenciaintervallumba.

Konfidenciahatárok megállapítása:
`library(gmodels) ci(cukor)`

Igazat mond a cukorgyár?

A kritikus fogyasztók maguk kezdenek mérőkampányba az interneten, és a `cukor.ism` objektumban található méréseket folytatják a külön e célból beszerzett patikamérlegen.

Hasonlít-e az átlag a fogyasztóvédelem által megállapítotthoz? Igaz-e rá, hogy beleesik a törvény által megengedett ingadozásba?

És az igaz, hogy a fogyasztók ugyanolyan arányban visznek haza kicsivel több cukrot, mint kevesebbet?

Bízz bennem! Én is bízom magamban.