Statisztika alapok

A statisztika alapfeladata: $X \in \mathbb{R}$ valószínűségi változó adott, de ismeretlen. $X_1, X_2, X_3 \ldots$ független és X-szel azonos eloszlású.

Ezek egyetlen realizációja $x_1, x_2, \ldots, x_n$ ← VÉGES hosszú sorozat, ezt megfigyeljük: $\boxed{\text{MINTÁT VESZÜNK } X\text{-ből}}$

[Azaz: rendelkezésünkre áll n db szám.]

Kérdés: Mi X eloszlása?

1. Válasz: Nyilván nem lehet tudni.

Legegyszerűbb eset: Egy hamis érmén a fej valsége p: ez adott, de nem ismerjük.

Feldobtuk $n = 100$-szor, és azt kaptuk, hogy

$$\underbrace{F\,I\,F\,I\,F\,F\,F\,F\,I\,I\ldots\ldots F\,F}_{60 \text{ db } F,\ 40 \text{ db } I}.$$

Mennyi p értéke?

Legelső természetes tipp: $\frac{60}{100} = 0.6$.

Igen ám, de

- Lehet, hogy $p = 0.5$? Válasz: Hát persze, lehet.
- Lehet, hogy $p = 0.01$? Válasz: Igen, az is lehet.
- Na jó, de azt gondoljuk, hogy $p = 0.6$?

Válasz: Nem, nem gondoljuk.

Miért is lenne ilyen szép kerek szám?

Simán lehet, hogy $p = 0.601032$, csak hát $n = 100$ kísérletből ilyet lehetetlen kimérni.

- Azért mégis valószínűbb, hogy $p = 0.6$, mint hogy $p = 0.01$, nem?

Válasz: De, igen ....?...

Második, kicsit értelmesebb állítás:

Nem tudjuk, hogy p mennyi, de nagy valószínűséggel ~~a~~ 0.5 és 0.7 közöt van.

[Nem fogom megmutatni, de igazából jelen esetben 95%]

MICSODA?!?! Hol itt a valószínűség?

Válasz: Sehol :(ugyanis

→ a p szám egy objektív létező
(bár nem ismerjük).

→ Mint ilyen, vagy 0,5 és 0,7 között van,
vagy sem.

→ Valószínűségről akkor lenne értelmes beszélni,
ha sok „kísérlet"-et végezhetnénk, ezek
egy részében $p \in [0,5 ; 0.7]$ lehetne
más részében meg nem.
De itt nincsenek kísérletek.

Terv a statisztika mikro-kurzusra:

- Megértjük, hogy milyen értelemben „legjobb tipp" a $p = 0.6$
- Értelmessé tesszük a „95% val.séggel $p \in [0.5, 0.7]$ kijelentést.
- Ha az érme gyártója azt állítja, hogy igazából $p = 0.4$, eldöntjük, hogy higgyünk-e neki vagy sem.

Lényeg röviden: Ha valaki bízik abban, hogy $p \in [0.5 ; 0.7]$,
az bízik a szerencséjében, de
Nem abban bízik, hogy szerencséje lesz,
hanem abban, hogy szerencséje volt.
(amikor a kísérletet végezte).

Megj.1 másik véglet / legnehezebb eset:

Titokzatos jelek érkeznek a világűrből – mindegyik egy szám. Forrásukról, jelentésükről fogalmunk sincs.

Megfigyeltünk 10-et, és az jött ki, hogy

-3 17 5 108 1210.3

1 10 100 -7 -70

Mi lehet X eloszlása?

Ez így végképp reménytelen:

- lehet, hogy $X \sim Uni[-100; 1300]$
- de az is lehet, hogy X diszkrét, és csak ezt a 10 értéket veheti fel.

Megj.2: Más lenne a helyzet, ha vehetnénk egy $X_1, X_2, X_3, \ldots$ $\boxed{\text{végtelen}}$ mintát.

Ebből 1 valószínűséggel pontosan rekonstruálható lenne bármilyen eloszlás.

Kár, hogy a statisztikában minden minta véges :(

Momentum-becslések

Legyen X val.-változó, $m = \mathbb{E}X$, $\sigma^2 = \mathrm{Var}\,X = D^2X$.

Mintát veszünk X-ből: $x_1, x_2, \ldots, x_n$

Vajon mennyi lehet m és σ?

Def: átlag: $\bar{x} := \dfrac{x_1 + x_2 + \ldots + x_n}{n}$

Fő hír: $\mathbb{E}\bar{x} = m$, vagyis az átlag torzítatlan becslése a várható értéknek.

Def: torzítatlan becslése: $\mathbb{E}(\text{becslés}) = \text{tényleges érték}$

Persze: értsd jól: Ha $X_1, X_2, \ldots, X_n \sim X$ függetlenek és $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$, akkor $\bar{X}$ val.-változó,

és $\mathbb{E}\bar{X} = \frac{1}{n}\left[\mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_n\right] = \frac{1}{n}\left[\underbrace{m + m + \ldots + m}_{n \text{ db}}\right]$

$= m.$

Persze, miután a kísérletet elvégeztük és kijön mondjuk, hogy $\bar{x} = 0.6$, ez már csak egy szám,

$\mathbb{E}\bar{x} = \mathbb{E}\,0.6 = 0.6$ nem túl értelmes, és főleg általában $\neq m$.

Gonosz megjegyzés:

Ezért kár volt ennyit számolni.

Legyen a becslésünk simán csak X_1: az elsőt felírom,
a többit meg se nézem.

Ez is torzítatlan becslése m-nek, hiszen $\mathbb{E}X_1 = m$.

Jó hír 2: Ha $n \to \infty$, akkor

$$\underset{\bar{X}(n)}{\overset{\|}{\bar{X}}} = \frac{X_1 + X_2 + \dots + X_n}{n} \longrightarrow m$$ 1 valószínűséggel,

ez a nagy számok törvénye.

Avagy: $\bar{x}$ Def: konzisztens becslés m-~~re~~re:

becslés $\xrightarrow{n \to \infty}$ tényleges érték

Ebben az értelemben, ha n nagy, akkor $\bar{x}$ jobb becslés, mint X_1.

Játsszuk el most ugyanezt σ-ra, hiszen

$\sigma^2 = \mathrm{Var}\,X = \mathbb{E}\left((X-m)^2\right)$ maga is egy várható érték:

$\boxed{K:}$ Miért ne legyen a becslése

$$\frac{(X_1-m)^2 + (X_2-m)^2 + \dots + (X_n-m)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2 \; ?$$

Válasz: Ez rendben lenne, ha m-et ismernénk ....

[Megj: És tényleg: ha m-et ismerjük, akkor

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2$$ torzítatlan és konzisztens becslése σ^2-nek.]

De általában m-et nem ismerjük.

Ötlet: jobb híján írjuk be m helyére a becslését, $\bar{X}$-ot!

Def: Az $X_1, X_2 \dots X_n$ mintából számolt tapasztalati szórásnégyzet

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Tétel (könnyű számolás):

$$S_n^2 = \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right]}_{\substack{\text{négyzetek}\\ \text{átlaga}}} - \underbrace{\bar{X}^2}_{\substack{\text{átlag}\\ \text{négyzete}}}$$

Köv: S_n^2 konzisztens becslése σ^2-nek.

Biz: $n \to \infty$-re

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \xrightarrow[\text{törvénye}]{\text{Nagy számok}} \mathbb{E}(X^2)\ ;\quad \bar{X} \xrightarrow{\text{N. SZ. T.}} m$$

$$\Rightarrow S_n^2 \longrightarrow \mathbb{E}(X^2)-m^2=\sigma^2 \qquad \square$$

~~Köv: S_n~~

<u>Def</u>: $S_n := \sqrt{S_n^2}$ a <u>tapasztalati szórás</u>.

<u>Köv</u>: S_n konzisztens becslése σ-nak.

Biz: Hát persze: $S_n^2 \to \sigma^2 \Longrightarrow S_n = \sqrt{S_n^2} \to \sqrt{\sigma^2} = \sigma$ $\square$

<u>Kérdés</u>: No, és S_n^2 <u>torzítatlan</u> becslése σ^2-nek?

<u>Válasz</u>: Közép-szörnyű számolás, de ÉRDEMES végiggondolni:

$$\mathbb{E} S_n^2 = \underbrace{\mathbb{E}\, \frac{1}{n} \sum_{i=1}^{n} X_i^2}_{\text{könnyű}} - \underbrace{\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)^2\right]}_{\text{ezt nézzük meg jobban}}$$

$$\| \quad \text{könnyű}$$

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{E}(X_i^2)}_{\mathbb{E}(X^2)} = \frac{1}{n} n \mathbb{E}(X^2) = \mathbb{E}(X^2)$$

2. tag: $\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)^2\right] = \mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) \frac{1}{n}\left(\sum_{j=1}^{n} X_j\right)\right] =$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}(X_i X_j) \leftarrow n^2 \text{ db tag}$$

Ebből
- ha $i=j$, akkor $X_j = X_i \Rightarrow \mathbb{E}(X_i X_j) = \mathbb{E}(X_i^2) = \mathbb{E}(X^2)$, ilyen tag van n darab
- ha $i \neq j$, akkor X_i és X_j <u>független</u> $\Rightarrow \mathbb{E}(X_i X_j) = \mathbb{E}X_i \mathbb{E}X_j = (\mathbb{E}X)^2$; ilyen tag van (n^2-n) darab

Vagyis

$$2.\ \text{tag} \ \underline{\underline{\text{felt.}}}\ \frac{1}{n^2}\Big[n\,\mathbb{E}(X^2) + \overbrace{(n^2-n)}^{n(n-1)}(\mathbb{E}X)^2\Big]$$

$$= \frac{1}{n}\mathbb{E}(X^2) - \frac{n-1}{n}(\mathbb{E}X)^2$$

Egybeírva:

$$\mathbb{E}\,S_n^2 = \underbrace{\mathbb{E}(X^2) - \frac{1}{n}\mathbb{E}(X^2)}_{\underbrace{\left(1-\frac{1}{n}\right)}_{\frac{n-1}{n}}\mathbb{E}(X^2)} + \frac{n-1}{n}(\mathbb{E}X)^2 = \frac{n-1}{n}\Big[\mathbb{E}(X^2) - (\mathbb{E}X)^2\Big]$$

$$= \frac{n-1}{n}\,\mathrm{Var}X = \frac{n-1}{n}\sigma^2$$

<u>Tétel:</u> $\mathbb{E}\,S_n^2 = \frac{n-1}{n}\sigma^2$.

Vagyis S_n^2 $\boxed{\text{NEM}}$ torzítatlan becslése σ^2-nek, csak

<u>aszimptotikusan torzítatlan</u>: (Def.)

$$\mathbb{E}(\text{becslés}) \xrightarrow{n\to\infty} \text{tényleges érték}$$

Ezért hát

<u>Def:</u> Az $X_1, X_2, \dots X_n$ mintából számolt <u>korrigált</u> <u>tapasztalati szórásnégyzet</u> legyen $S_n^{*2} := \frac{n}{n-1}S_n^2$.

A <u>korrigált tapasztalati szórás</u> legyen $S_n^* := \sqrt{S_n^{*2}}$.

<u>Köv.:</u> $\mathbb{E}(S_n^{*2}) = \sigma^2$

Tétel: S_n^{*2} torzítatlan és konzisztens becslése σ^2-nek.

Megj. 1: ~~Kova~~ $S_n^* = \sqrt{S_n^{*2}}$ viszont NEM torzítatlan becslése σ-nak: $\mathbb{E}\sqrt{S_n^{*2}} \neq \sqrt{E(S_n^{*2})}$.

A szórásnégyzet hasznosabb / kellemesebb fogalom, mint a szórás.

Megj. 2: $n=1$ esetén ~~S_n~~ $S_n^2 = 0$, viszont $S_n^{*2} = \frac{1}{1-1} \cdot 0$ nincs értelmezve.

Ez így van jól: $n=1$ elemű mintából

- az hagyján, hogy valaki várható értéket becsül
- de szórást már csak ne akarjon, pláne ne nullának.

Megj. 3: Mögöttes jelenség: Az n db adatunkból

- első lépésben megbecsültük a várható értéket
- így 2. lépésben σ^2 becsléséhez „már csak $n-1$ $\underbrace{\text{szabadsági fokunk}}_{\text{„független adat}}$ maradt" – ezért kell „n helyett $(n-1)$-gyel osztani"

<u>Megj:</u> Móricka rengeteg számot diktál 1-esével Pistikének azzal, hogy számolja ki az átlagukat és a (korrigált) tapasztalati szórás(négyzet)ukat.
(üket)

Pistike $\boxed{\text{nem jegyzi meg}}$ a számokat, csak mindig az aktuális

mert kevés a memóriája

- darabszámot: ehhez minden lépésben hozzáad 1-et
- összeget: ehhez — // — ja a legújabb számot
- négyzetösszeget: ehhez — // — legújabb szám négyzetét.

Az így eltárolt <u>3 darab</u> számból:

$$n\text{-ből},\quad \sum_{i=1}^{n} x_i\text{-ből és}\quad \sum_{i=1}^{n} x_i^2\text{-ből}$$

bármikor $\boxed{\text{könnyen}}$ számolható, ami kell:

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

sokkal könnyebben, mint ha az összes számból számolna $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$ alapján

ez csak a legvégén derül ki,

~~Így működik a klasszikus zseb~~

Igy müködik a klassikus zsebszámológépek statisztika üzemmódja.