Nagy eltérés tételek - bevezetés

Legyen $X_1, X_2, \ldots, X_n$ független és azonos eloszlású,

$\mathbb{E}X_i = m$, $\operatorname{Var} X_i = \sigma^2$, $S_n = X_1 + \ldots + X_n$.

Így persze $\mathbb{E}S_n = nm$, $\operatorname{Var} S_n = n\sigma^2$, $DS_n = \sqrt{n}\sigma$.

Kérdés a $\mathbb{P}(S_n \leq K)$ vagy $\mathbb{P}(S_n \geq K)$ valószínűség, ahol $K \in \mathbb{R}$.

Persze a kérdés átfogalmazható az $\tilde{S}_n := \frac{S_n - nm}{\sqrt{n}\sigma}$ normált val. változóra: $\mathbb{E}\tilde{S}_n = 0$ és $\operatorname{Var}\tilde{S}_n = 1$,

így jobban érzékelhető, hogy az $\{S_n \leq K\}$ mennyire „extrém" vagy „meglepő":

$$\mathbb{P}(S_n \leq K) = \mathbb{P}(\tilde{S}_n \leq x), \quad \text{ahol } K = nm + \sqrt{n}\sigma x$$

$$\mathbb{P}(S_n \geq K) = \mathbb{P}(\tilde{S}_n \geq x) \quad \text{avagy } K = \mathbb{E}S_n + x\, DS_n$$

Láttuk: A centrális határeloszlás tétel arról az esetről szól, amikor $x \in \mathbb{R}$ rögzített, vagyis $\underbrace{K - \mathbb{E}S_n}_{\text{eltérés a várható értéktől}} \sim \underbrace{\sqrt{n}}_{\sqrt{n} \text{ nagyságrendű.}}$

~~Problémá~~:

<u>Nagy eltérés problémának</u> neveztük azt, amikor

$$|K - ES_n| \gg \sqrt{n}, \quad \text{pl. } K - ES_n \sim n$$

~~Az~~ Ilyenkor $P(S_n \geq k)$ becslésére a CHT <u>nem alkalmas</u>:

a keresett val.ség nagyon kicsi (vagy nagyon nagy), de a CHT becslés hibája miatt nem lehet CHT-vel megbízhatóan becsülni.

Ilyenkor hasznosak a <u>nagy eltérés tételek</u>.

Hoeffding egyenlőtlenség (Hoeffding, 1963)

(ejtsd: „höffding")

Tétel Legyenek $X_1, X_2, \dots X_n$ független valószínűségi változók (de nem feltétlen azonos eloszlásúak), és legyen $S_n = X_1 + X_2 + \dots + X_n$.

Tegyük fel, hogy mindegyik X_i korlátos: minden i-re van olyan $a_i < b_i$, hogy $a_i \le X_i \le b_i$.

Ekkor minden $t > 0$-ra

$$\mathbb{P}(S_n \ge \mathbb{E}S_n + t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$ [nagy eltérés felfelé]

és

$$\mathbb{P}(S_n \le \mathbb{E}S_n - t) \le$$ —— || —— . [nagy eltérés lefelé]

Megjegyzések:

1.) A korlátosság erős feltétel, a tétel legfőbb gyengéje.

2.) Ez egy nagy eltérés tétel: ha az X_i-k azonos eloszlásúak, akkor $(b_i - a_i)^2$ „szórásnégyzet-szerűség", így a nevezőbeli $\sum_{i=1}^{n}(b_i - a_i)^2 = n(b-a) \sim n \sim \mathrm{Var}\, S_n$, Így a jobboldal akkor lesz kicsi, ha $t \gg \sqrt{n}$.

3.) A tétel minden n-re felső becslést ad (nem csak aszimptotikusan $n \to \infty$-ben).

4.) A felső becslés általában nem éles: a tényleges érték sokkal kisebb is lehet.

5.) A tétel nagyon könnyen alkalmazható:

- kevés számolással jár,

és főleg

- keveset kell tudni az X_i-k eloszlásáról:
  - → kell az n (darabszám)
  - → kellenek a korlátok (a_i, b_i)
  - → kell az ES_n, vagyis az összeg várható értéke.

Megj: Persze $\mathbb{E}S_n = \mathbb{E}X_1 + \mathbb{E}X_2 + \dots + \mathbb{E}X_n$, de nem baj, ha külön-külön nem ismerjük őket.

Példa a Hoeffding egyenlőtlenség alkalmazására

Piripócson a kukásautó hétfőnként 1000 háztartásból viszi el a szemetet. A lakók 3 féle kukát használnak:

400 háztartásban 60 l-es kuka van
100 〃 110 〃
500 〃 ~~120~~ 120 〃.

Az egyes háztartások által a kukába tett szemét mennyisége véletlen és független, de persze legfeljebb annyi lehet, amennyi a kukába fér.

Ha nem fér bele az összes szemét a kukásautóba, akkor baj van.

Hány liter szemétnek kell elférni a kukásautóban, hogy ennek valószínűsége egészen biztosan legfeljebb 1% legyen?

A kukás cégnek persze fogalma sincs, hogy az egyes emberek külön-külön mennyi szemetet termelnek, de azt tudják, hogy az össz-mennyiség átlagosan 40000 liter szokott lenni.

Megoldás: Legyen $n=1000$ és legyen $X_1, X_2, \dots, X_n$ az egyes kukákban a szemét mennyisége, literben. Így $S_n := X_1 + X_2 + \dots + X_n$ az össz-mennyiség. Olyan K-t keresünk, amire $\mathbb{P}(S_n > K) \le 0.01$.

A $\mathbb{P}(S_n > K)$ val.ség becslésére a CHT nem alkalmas, pl. mert nincs okunk feltételezni, hogy az X_i-k azonos eloszlásúak. Ráadásul a szórásokat sem ismerjük.

A Hoeffding egyenlőtlenség viszont alkalmazható, mert az X_i-k korlátosak: $a_i \le X_i \le b_i$, ahol $a_i = 0$,

$$\text{és } b_i = \begin{cases} 60, & 400 \text{ db } i\text{-re} \quad (\text{mondjuk } i=1,\dots,400) \\ 110, & 100 \text{ db } i\text{-re} \quad (\text{mondjuk } i=401,\dots,500) \\ 120, & 500 \text{ db } i\text{-re} \quad (\text{mondjuk } i=501,\dots,1000) \end{cases}$$

Fontos a szumma!

Fontos a négyzet!

$$\text{Így } \sum_n (b_i - a_i)^2 = 400\cdot(60-0)^2 + 100\cdot(110-0)^2 + 500\cdot(120-0)^2 = 9850000$$

Vagyis a Hoeffding egyenlőtlenség szerint minden $t > 0$-ra

$$\mathbb{P}(S_n \ge ES_n + t) \le \exp\left(-\frac{2t^2}{9850000}\right).$$

A cél, hogy a jobb oldal $1\% = 0.01$ legyen:

$$\exp\left(-\frac{2t^2}{9850000}\right) = 0.01 \iff t = \sqrt{\frac{9850000}{2}\cdot(-\ln(0.01))} \approx 4762$$

Vagyis $K := \mathbb{E}S_n + t$ jó lesz $t = 4762$ választással:

ennyivel kell nagyobb kapacitás, mint az $\mathbb{E}S_n = 40000$

várható érték.

7

<u>Válasz</u>: $K = \mathbb{E}S_n + t = 44762$ literes kukásautó elég.