

Intervalová data a výpočet některých statistik

Milan Hladík¹ Michal Černý²

¹ Katedra aplikované matematiky
Matematicko-fyzikální fakulta
Univerzita Karlova

² Katedra ekonometrie
Fakulta informatiky a statistiky
Vysoká škola ekonomická Praha

Robust 2014

Intervalová data. Nechť data $x_1, \dots, x_n$ jsou nepozorovatelná. Pozorovatelné jsou jen intervaly

$$[\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_n, \bar{x}_n],$$

o nichž víme, že platí

$$\underline{x}_i \leq x_i \leq \bar{x}_i, \quad i = 1, \dots, n.$$

Příklad 1. Namísto dat $x_1, \dots, x_n$ pozorujeme pouze „zaokrouhlené“ hodnoty

$$\underline{x}_i = \lfloor x_i \rfloor, \quad \bar{x}_i = \lceil x_i \rceil, \quad i = 1, \dots, n.$$

Příklad 2. Namísto dat $x_1, \dots, x_n$ pozorujeme pouze „zašuměné“ hodnoty

$$\underline{x}_i = x_i - \gamma_i, \quad \bar{x}_i = x_i + \delta_i, \quad i = 1, \dots, n,$$

kde γ_i a δ_i jsou nezáporné náhodné veličiny.

Příklad 3. Nechť $X(t)$ je pozorovatelný náhodný proces s časem $t \geq 0$.
Nechť

$$\tau_1 \in [0, 1], \quad \tau_2 \in [1, 2], \quad \dots, \quad \tau_n \in [n-1, n]$$

jsou nepozorovatelné časové okamžiky. Pak i data

$$x_1 = X(\tau_1), \quad \dots, \quad x_n = X(\tau_n)$$

jsou nepozorovatelná. Pozorovatelné jsou ovšem hodnoty

$$\underline{x}_i = \min_{t \in [i-1, i]} X(t), \quad \bar{x}_i = \max_{t \in [i-1, i]} X(t), \quad i = 1, \dots, n,$$

které jistě splňují $\underline{x}_i \leq x_i \leq \bar{x}_i$.

- Jaké jsou další zajímavé mechanismy generující intervalová data (vedle zaokrouhlování, diskretizace, klasifikace do tříd, ...)?
- Jak se chovat v lineárním regresním modelu

$$y = X\beta + \varepsilon,$$

jestliže namísto dat (X, y) pozorujeme jen intervaly $([\underline{X}, \overline{X}], [\underline{y}, \overline{y}])$, o nichž víme, že platí $\underline{X} \leq X \leq \overline{X}$ a $\underline{y} \leq y \leq \overline{y}$?

- Co když nevíme nic více? A co když naopak víme něco dalšího, například známe rozdělení X na $[\underline{X}, \overline{X}]$ a/nebo rozdělení y na $[\underline{y}, \overline{y}]$?
- Je-li dána statistika $S(x_1, \dots, x_n)$, co o ní můžeme říci?

- Uvažme například, že $x_1, \dots, x_n$ je výběr z $N(\mu, \sigma^2)$. Pozorujeme ale jen naše intervaly

$$[\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_n, \bar{x}_n]. \quad (1)$$

- **Hlavní otázka.** Je dána statistika (= funkce dat) $S(x_1, \dots, x_n)$, např. $\hat{\mu}$, $\hat{\sigma}^2$, t -ratio apod. Co o ní můžeme říci, známe-li jen intervaly (1)?
- Na $x_1, \dots, x_n$ můžeme nahlížet jako na náhodné veličiny na intervalech (1) s jistým rozdělením. Pak i hodnota $S = S(x_1, \dots, x_n)$ je náhodná veličina. Můžeme něco říci o jejím rozdělení?

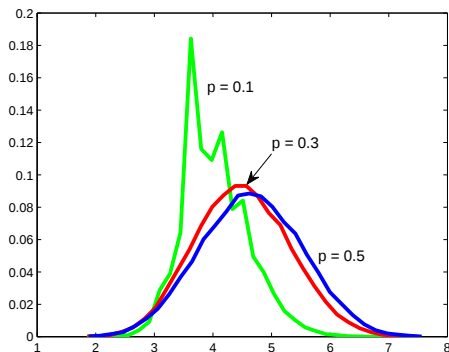
Výběrový rozptyl

Za statistiku S vezměme $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j \right)^2$.

Příklad: předpokládejme nezávislé

$$x_i = \begin{cases} \underline{x}_i & \text{s pravděpodobností } p, \\ \bar{x}_i & \text{s pravděpodobností } 1 - p. \end{cases}$$

Pak rozdělení $\hat{\sigma}^2$ může vypadat například:



Položme si tuto otázku: je-li dána statistika S , dokážeme spočítat alespoň

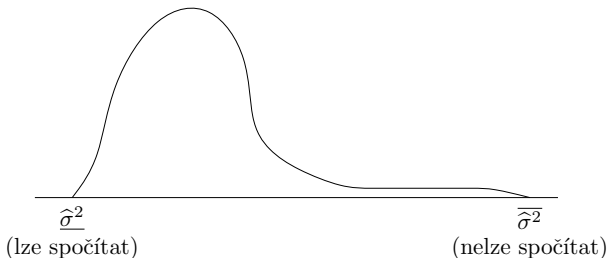
$$\begin{aligned}\bar{S} &= \sup\{S(x_1, \dots, x_n) : (\forall i) x_i \in [\underline{x}_i, \bar{x}_i]\}, \\ \underline{S} &= \inf\{S(x_1, \dots, x_n) : (\forall i) x_i \in [\underline{x}_i, \bar{x}_i]\}?\end{aligned}$$

Tyto hodnoty dávají alespoň informaci $\underline{S} \leq S \leq \bar{S}$. Navíc za mírných předpokladů dokonce platí, že $[\underline{S}, \bar{S}]$ je nosičem distribuce S .

Za statistiku S opět vezměme $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j\right)^2$.

- **Věta.** Spočítat $\underline{S} = \underline{\hat{\sigma}^2}$ lze efektivně redukcí na konvexní kvadratické programování.
- **Věta.** Spočítat $\bar{S} = \overline{\hat{\sigma}^2}$ je NP-těžký problém („neexistuje obecná metoda pracující v lepším čase než 2^n “).
- Dokonce ani (přímočará) simulace příliš nepomáhá (např. při rovnoměrném rozdělení x_i na $[\underline{x}_i, \bar{x}_i]$) — chceme-li se při simulaci strefit do blízkosti $\hat{\sigma}^2$ s rozumnou pravděpodobností, potřebujeme řádově 2^n pokusů.

Řekněme, že x_i jsou nezávislé a rovnoměrně rozdělené na $[\underline{x}_i, \overline{x}_i]$. Rozdělení $\hat{\sigma}^2$ si pak lze představovat např. podle obrázku:



Důsledek. Neexistuje ani metoda, která by dokázala efektivně vyčíslit hodnoty funkce hustoty, distribuční funkce, kvantilové funkce apod. (Kdyby taková metoda existovala, pak bychom dokázali pomocí půlení intervalu efektivně aproximovat hodnotu $\overline{\sigma}^2$, ale to nejde.)

Věta. Existuje **pseudo**polynomiální algoritmus pro výpočet $\overline{\sigma^2}$.

To zhruba znamená: Jsou-li kraje intervalů $\underline{x}_i$, $\overline{x}_i$ **celá čísla**, která nejsou příliš velká, pak dokážeme spočítat $\overline{\sigma^2}$ i při velkém n .

- **Polynomiální algoritmus** pracuje v polynomiálním čase vzhledem k **binárnímu kódování** celých čísel, tj. v čase

$$\text{polynom}(\log |\underline{x}_1| + \log |\overline{x}_1| + \cdots + \log |\underline{x}_n| + \log |\overline{x}_n|).$$

- **Pseudopolynomiální algoritmus** pracuje v polynomiálním čase vzhledem k **unárnímu kódování** celých čísel, tj. v čase

$$\text{polynom}(|\underline{x}_1| + |\overline{x}_1| + \cdots + |\underline{x}_n| + |\overline{x}_n|).$$

Data:

$[x_1, \bar{x}_1]$	$= [2, 3]$
$[x_2, \bar{x}_2]$	$= [0, 1]$
$[x_3, \bar{x}_3]$	$= [1, 3]$
$[x_4, \bar{x}_4]$	$= [0, 5]$
$[x_5, \bar{x}_5]$	$= [-2, 1]$
$[x_6, \bar{x}_6]$	$= [-1, 0]$
$[x_7, \bar{x}_7]$	$= [1, 2]$
$[x_8, \bar{x}_8]$	$= [1, 6]$
$[x_9, \bar{x}_9]$	$= [0, 7]$
$[x_{10}, \bar{x}_{10}]$	$= [0, 2]$
$[x_{11}, \bar{x}_{11}]$	$= [1, 2]$
$[x_{12}, \bar{x}_{12}]$	$= [1, 3]$
$[x_{13}, \bar{x}_{13}]$	$= [-1, 1]$
$[x_{14}, \bar{x}_{14}]$	$= [-2, 4]$
$[x_{15}, \bar{x}_{15}]$	$= [3, 4]$
$[x_{16}, \bar{x}_{16}]$	$= [1, 10]$
$[x_{17}, \bar{x}_{17}]$	$= [1, 2]$
$[x_{18}, \bar{x}_{18}]$	$= [2, 3]$
$[x_{19}, \bar{x}_{19}]$	$= [3, 4]$
$[x_{20}, \bar{x}_{20}]$	$= [1, 6]$

Počet kroků pseudopoly algoritmu:

n	pseudopol.	„brute-force“ metoda (2^n)
2	6	4
3	14	8
4	30	16
5	62	32
6	126	64
7	230	128
8	428	256
9	784	512
10	1 272	1 024
11	1 814	2 048
12	2 442	4 096
13	3 220	8 192
14	4 314	16 384
15	5 513	32 768
16	7 715	65 536
17	10 004	131 072
18	12 380	262 144
19	14 880	524 288
20	17 812	1 048 576

- **Věta.** Nejen přesný, ale dokonce i **přibližný** výpočet hodnoty $\overline{\sigma^2}$ s **libovolnou absolutní chybou** je NP-těžký.
- **Problém.** Jak je to s výpočtem $\overline{\sigma^2}$ s **relativní chybou**?
- Víme jen:
 - existuje polynomiální algoritmus na $\overline{\sigma^2}$ s relativní chybou $\varrho = 1$;
 - aproximace $\overline{\sigma^2}$ s relativní chybou $\varrho \leq 2^{-2n-1}$ je NP-těžká.

K čemu může být dobrý interval $[\underline{S}, \overline{S}]$, kde

$$\overline{S} = \sup\{S(x_1, \dots, x_n) : (\forall i) x_i \in [\underline{x}_i, \overline{x}_i]\},$$

$$\underline{S} = \inf\{S(x_1, \dots, x_n) : (\forall i) x_i \in [\underline{x}_i, \overline{x}_i]\},$$

je-li $S = S(x_1, \dots, x_n)$ testová statistika pro nějaký test?

Je-li C kritický obor (na pevně zvolené hladině významnosti), pak můžeme činit alespoň dílčí závěry, máme-li štěstí:

- Je-li $[\underline{S}, \overline{S}] \subseteq C$, pak víme, že test zamítá nulovou hypotézu (bez ohledu na to, kde konkrétně leží data $x_1, \dots, x_n$ v intervalech $[\underline{x}_1, \overline{x}_1], \dots, [\underline{x}_n, \overline{x}_n]$).
- Analogicky, je-li $[\underline{S}, \overline{S}] \cap C = \emptyset$, pak víme, že test nulovou hypotézu nezamítne.
- **Problém.** A jak se zachovat ve třetím případě? Co když je např. průnik $[\underline{S}, \overline{S}] \cap C$ neprázdný, ale „malý“?

- Spočítat $\underline{S}$ a $\overline{S}$ je snadné, je-li S lineární funkcí proměnných $x_1, \dots, x_n$, například $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$.
- Obecněji: Spočítat $\underline{S}$ a $\overline{S}$ je snadné, lze-li předpis pro S napsat jako vzorec, v němž se každá z proměnných $x_1, \dots, x_n$ vyskytuje *nanejvýš jednou*, například

$$\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

- Někdy je to ale těžké: například není těžké ukázat, že pro F -statistiku je výpočet $\underline{F}$ i $\overline{F}$ NP-těžký.
- A co slibovaná t -statistika?

Ve výrazu

$$t = \sqrt{n} \frac{|\hat{\mu} - \mu_0|}{\hat{\sigma}} = \sqrt{n} \frac{\left| \frac{1}{n} \sum_{j=1}^n x_j - \mu_0 \right|}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{j=1}^n x_j)^2}},$$

kde μ_0 je libovolná pevná konstanta, se normuje rozptylem. Není proto překvapivé, že platí

Věta.

- Spočítat hodnotu $\bar{t}$ lze efektivně (netriviální redukci na konvexní optimalizaci),
- spočítat hodnotu $\underline{t}$ je NP-těžké,
- spočítat hodnotu $\underline{t}$ je dokonce NP-těžké i s **libovolnou** absolutní chybou (tj. např. i s chybou 1000!),
- spočítat hodnotu $\underline{t}$ lze v pseudopolynomiálním čase.

Zajímá nás ekonometrická regrese

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 \pi_t + \varepsilon_t,$$

kde

- t indexuje čas,
- C_t = spotřební výdaje,
- Y_t = příjem,
- π_t = inflace predikovaná v období t pro období $t + 1$.

Otázka. Problém je, že hodnoty π_t nejsou pozorovatelné. Máme k dispozici jen intervaly $[\underline{\pi}_t, \overline{\pi}_t]$, např. interval predikcí „expertů“ či intervalovou predikci jiného modelu. Co pak můžeme dělat? Co můžeme například říci o běžných estimátorech regresních parametrů?

Děkujeme za pozornost.

(Některé z prezentovaných výsledků vyjdou v CSDA.)