Intervalová data a výpočet některých statistik

Milan Hladík [1] Michal Černý [2]

[1] Katedra aplikované matematiky
Matematicko-fyzikální fakulta
Univerzita Karlova

[2] Katedra ekonometrie
Fakulta informatiky a statistiky
Vysoká škola ekonomická Praha

Robust 2014

Intervalová data a příklady

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

$$[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n],$$

o nichž víme, že platí

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

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

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

Intervalová data a příklady

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

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

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

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

jsou nepozorovatelné časové okamžiky. Pak i data

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

jsou nepozorovatelná. Pozorovatelné jsou ovšem hodnoty

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

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

Některé obecné otázky

- 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 $S(x_1, \ldots, x_n)$, co o ní můžeme říci?

Hlavní otázka

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

$$[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]. \tag{1}$$

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

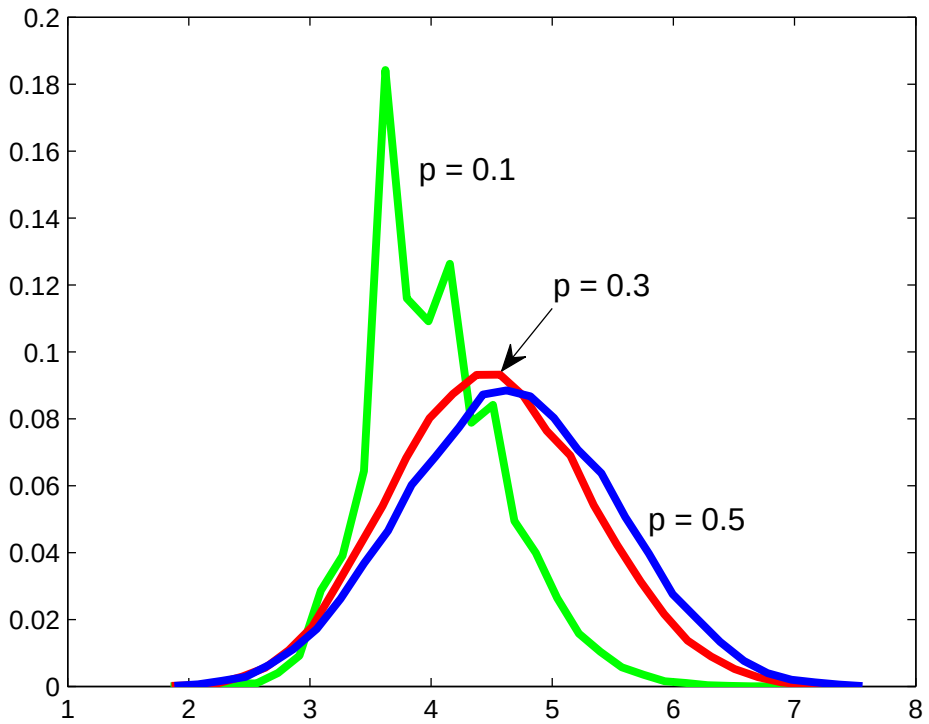
Výběrový rozptyl

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

Příklad: předpokládejme nezávislé

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

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



Výběrový rozptyl

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

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

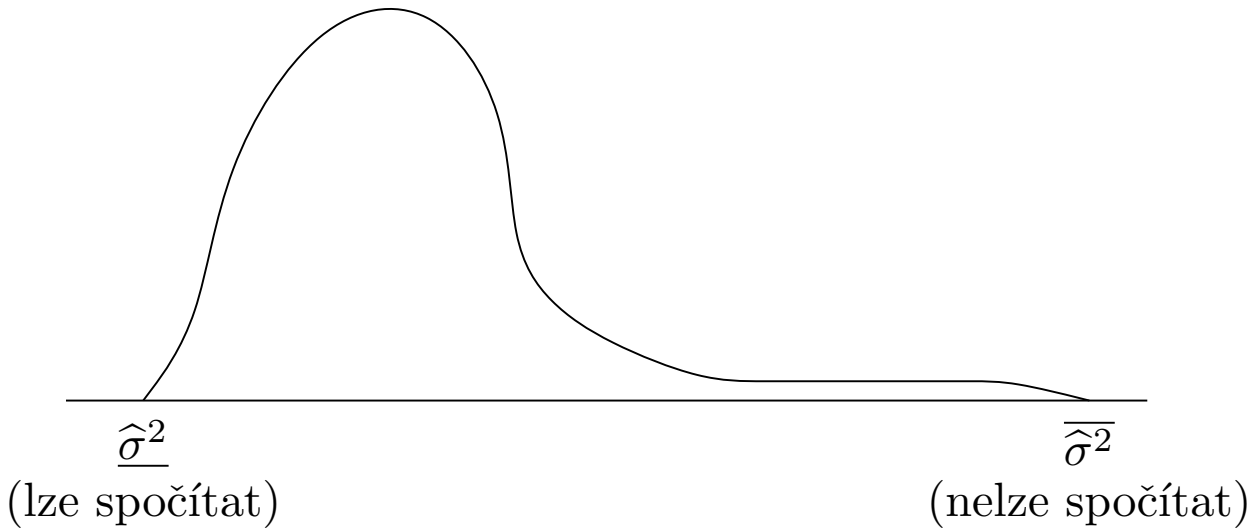
Tyto hodnoty dávají alespoň informaci $\underline{S} \leq S \leq \overline{S}$. Navíc za mírných předpokladů dokonce platí, že $[\underline{S}, \overline{S}]$ je nosičem distribuce S.

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

- Věta. Spočítat $\underline{S} = \underline{\widehat{\sigma}^2}$ lze efektivně redukcí na konvexní kvadratické programování.
- Věta. Spočítat $\overline{S} = \overline{\widehat{\sigma}^2}$ 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 $[\underline{x}_i, \overline{x}_i]$) — chceme-li se při simulaci strefit do blízkosti $\overline{\widehat{\sigma}^2}$ s rozumnou pravděpodobností, potřebujeme řádově 2^n pokusů.

Ilustrace

Řekněme, že x_i jsou nezávislé a rovnoměrně rozdělené na $[\underline{x}_i, \overline{x}_i]$. Rozdělení $\widehat{\sigma}^2$ 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 $\overline{\widehat{\sigma}^2}$, ale to nejde.)

Pozitivní výsledek

Věta. Existuje pseudopolynomiální algoritmus pro výpočet $\overline{\widehat{\sigma}^2}$.

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

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

Příklad

Data:

$$
\begin{aligned}
[\underline{x}_1, \overline{x}_1] &= [2, 3] \\
[\underline{x}_2, \overline{x}_2] &= [0, 1] \\
[\underline{x}_3, \overline{x}_3] &= [1, 3] \\
[\underline{x}_4, \overline{x}_4] &= [0, 5] \\
[\underline{x}_5, \overline{x}_5] &= [-2, 1] \\
[\underline{x}_6, \overline{x}_6] &= [-1, 0] \\
[\underline{x}_7, \overline{x}_7] &= [1, 2] \\
[\underline{x}_8, \overline{x}_8] &= [1, 6] \\
[\underline{x}_9, \overline{x}_9] &= [0, 7] \\
[\underline{x}_{10}, \overline{x}_{10}] &= [0, 2] \\
[\underline{x}_{11}, \overline{x}_{11}] &= [1, 2] \\
[\underline{x}_{12}, \overline{x}_{12}] &= [1, 3] \\
[\underline{x}_{13}, \overline{x}_{13}] &= [-1, 1] \\
[\underline{x}_{14}, \overline{x}_{14}] &= [-2, 4] \\
[\underline{x}_{15}, \overline{x}_{15}] &= [3, 4] \\
[\underline{x}_{16}, \overline{x}_{16}] &= [1, 10] \\
[\underline{x}_{17}, \overline{x}_{17}] &= [1, 2] \\
[\underline{x}_{18}, \overline{x}_{18}] &= [2, 3] \\
[\underline{x}_{19}, \overline{x}_{19}] &= [3, 4] \\
[\underline{x}_{20}, \overline{x}_{20}] &= [1, 6]
\end{aligned}
$$

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

Další výsledky o výpočtu rozptylu

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

Testové statistiky

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

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

je-li $S = S(x_1, \ldots, x_n)$ 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 $x_1, \ldots, x_n$ v intervalech $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$).
- 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“?

Další statistiky

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

t-statistika

Ve výrazu

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

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

Věta.

- Spočítat hodnotu $\overline{t}$ lze efektivně (netriviální redukcí 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.

Otázka na závěr

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.)