Intervalová data

Příklad

Výše pojistných plnění 227 klientů jedné pojišťovny:

výše plnění	počet klientů
0 - 7500	99
7500 - 17500	42
17500 - 32500	29
32500 - 67500	28
67500 - 125000	17
125000 - 300000	9
přes 300000	3

Jak reprezentovat kategorická data?
a) interval reprezentovat jedním bodem
b) nagenerovat si data z rovnoměrných rozdělení na příslušných intervalech
c) neparametrické odhady pro intervalová data
d) parametrické modely -||-

empirická distribuční funkce: $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x\} = \frac{1}{n}\{\text{počet pozorování} \le x\}$.

označme hranice intervalů $c_0 < c_1 < c_2 < \dots < c_k$ a počty pozorování v intervalu $(c_{j-1}, c_j]$ jako m_j. (c_0 a c_k může být i $\pm\infty$)

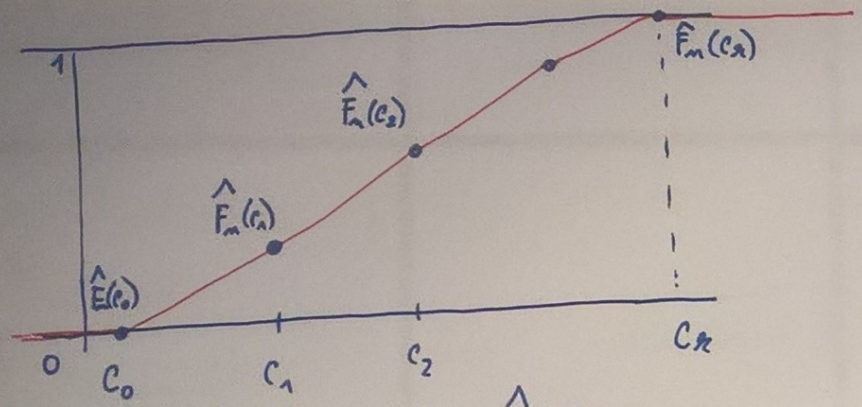
$(c_0, c_1]$ ... m_1
$(c_1, c_2]$ ... m_2
⋮
$(c_{k-1}, c_k]$ ... m_k
n

Podle definice můžeme určit hodnoty emp. distr. fce v bodech $c_0, c_1, \dots, c_k$:

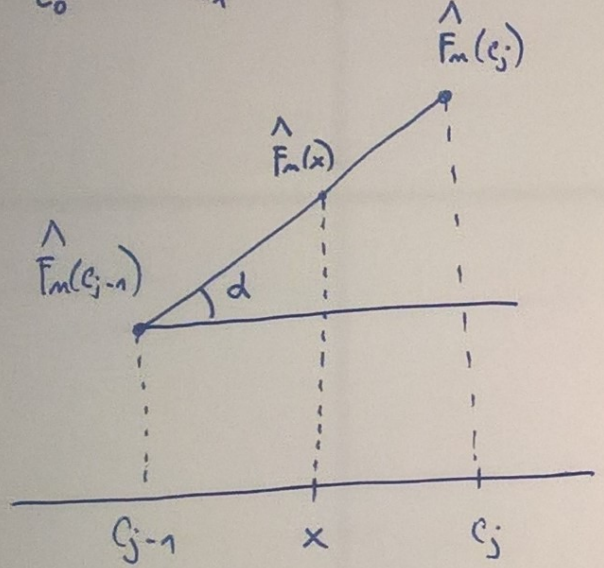
$\hat{F}_n(c_0) = 0$

$\hat{F}_n(c_j) = \frac{m_1 + m_2 + \dots + m_j}{n}$ a mezi nimi spojitě dodefinovat (ogiva)

$\hat{F}_n(c_k) = 1$



$$\hat{F}_m(x) = \begin{cases} 0 & x \le c_0 \\ \frac{c_j - x}{c_j - c_{j-1}} \cdot \hat{F}_m(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} \cdot \hat{F}_m(c_j) & , \; c_{j-1} \le x \le c_j \\ 1 & x \ge c_k \end{cases}$$



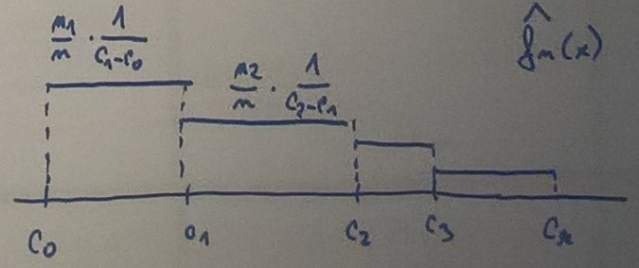
$$\operatorname{tg}\alpha = \frac{\hat{F}_m(c_j) - \hat{F}_m(c_{j-1})}{c_j - c_{j-1}}$$

$$\hat{F}_m(x) - \hat{F}_m(c_{j-1}) = (x - c_{j-1}) \cdot \operatorname{tg}\alpha$$

empirická hustota = histogram

$$\hat{f}_m(x) = \hat{F}_m'(x) \quad \text{pro } x \neq c_0, c_1, \dots, c_k$$

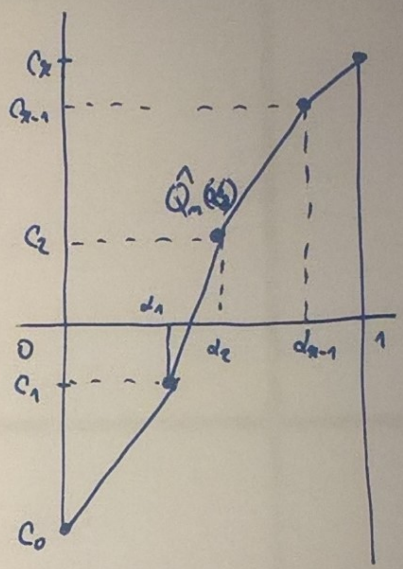
$$\hat{f}_m(x) = \frac{-\hat{F}_m(c_{j-1}) + \hat{F}_m(c_j)}{c_j - c_{j-1}} = \frac{-\frac{m_1 + \dots + m_{j-1}}{n} + \frac{m_1 + \dots + m_j}{n}}{c_j - c_{j-1}} = \frac{m_j}{n} \cdot \frac{1}{c_j - c_{j-1}} \quad \text{pro} \quad c_{j-1} \le x < c_j .$$



empirická kvantilová funkce:

určíme jako inverzní funkci k emp. distr. fci (ogive)

$$\hat{Q}_m(\alpha) = \hat{F}_m^{-1}(\alpha) \quad , 0<\alpha<1$$



řešení: $\hat{Q}_m(0) = c_0$

$\hat{Q}_m(\alpha) = c_j$ pro $\alpha = \frac{n_1+n_2+\dots+n_j}{n} =: d_j$ } a mezi nimi spojitě dodefinujeme

$\hat{Q}_m(1) = c_k$

$$\hat{Q}_m(\alpha) = \frac{d_j - \alpha}{d_j - d_{j-1}} \cdot \hat{Q}_m(d_{j-1}) + \frac{\alpha - d_{j-1}}{d_j - d_{j-1}} \cdot \hat{Q}_m(d_j) \quad , d_{j-1} \le \alpha \le d_j$$

Parametrické modely

původní data: $X_1, \dots, X_n$ náh. výběr X_i má distribuční funkci $F(x,\theta)$, resp. hustotu $f(x,\theta)$, kde θ je neznámý parametr.

metoda maximální věrohodnosti

Počítejme pravd., že daná pozorování padnou do intervalu $(c_{j-1}, c_j]$: $P(X_i \in (c_{j-1}, c_j]) = F(c_j,\theta) - F(c_{j-1},\theta) = \int_{c_{j-1}}^{c_j} f(x,\theta)\,dx$

věrohodnostní funkce $L(\theta) = \prod_{i=1}^{n} P(X_i \in (c_{j-1}, c_j]) = \left[F(c_1,\theta) - F(c_0,\theta)\right]^{n_1} \cdot \left[F(c_2,\theta) - F(c_1,\theta)\right]^{n_2} \cdot \dots \cdot \left[F(c_k,\theta) - F(c_{k-1},\theta)\right]^{n_k} =$

$$= \prod_{j=1}^{k} \left[F(c_j,\theta) - F(c_{j-1},\theta)\right]^{n_j}$$

logaritmická věrohodnostní funkce $\ell(\Theta) = \log L(\Theta) = \sum_{j=1}^{k} n_j \cdot \log\left[F(c_j, \Theta) - F(c_{j-1}, \Theta)\right]$

$\hat{\Theta}$ je odhad parametru Θ metodou max. věr. (jelikož $\hat{\Theta} = \operatorname{argmax} L(\Theta) = \operatorname{argmax} \ell(\Theta)$.

Příklad

mějme soubor pojistných plnění X_i řídí exponenciálním rozdělením s parametrem $\frac{1}{\lambda}$, $\lambda > 0$

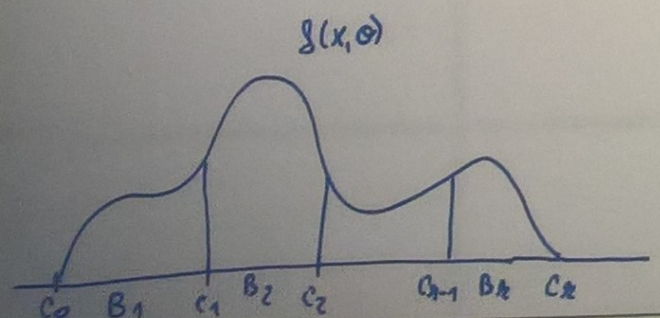
$$g(x, \lambda) = \frac{1}{\lambda} \cdot e^{-\frac{x}{\lambda}}, \quad x > 0$$

$$F(x, \lambda) = 1 - e^{-\frac{x}{\lambda}}, \quad x > 0$$

$$F(c_j, \lambda) - F(c_{j-1}, \lambda) = e^{-\frac{c_{j-1}}{\lambda}} - e^{-\frac{c_j}{\lambda}}$$

$$\ell(\lambda) = 99 \cdot \log\left(1 - e^{-\frac{7500}{\lambda}}\right) + 42 \cdot \log\left(e^{-\frac{7500}{\lambda}} - e^{-\frac{17500}{\lambda}}\right) + \dots + 3 \cdot \log\left(e^{-\frac{300\,000}{\lambda}}\right)$$ ... numerická maximalizace $\rightarrow \hat{\lambda}$ = odhad průměrné výše plnění

III. Metoda minimálního χ^2



mějme opět intervalová data

$(c_0, c_1] = B_1$ .... n_1 pozorování
$(c_1, c_2] = B_2$ ... n_2 pozorování
:
$(c_{k-1}, c_k] = B_k$ ... n_k pozorování

} celkem $n = n_1 + n_2 + \dots + n_k$ pozorování

Nejprve jako v předchozím postupujeme: $P(X_i \in B_j) = P(X_i \in (c_{j-1}, c_j]) = \int_{c_{j-1}}^{c_j} f(x,\theta)\,dx = F(c_j,\theta) - F(c_{j-1},\theta) =: p_j(\theta)$

celkem n pozorování, tedy v intervalu $B_j = (c_{j-1}, c_j]$ by mělo být $n \cdot p_j(\theta)$ pozorování (očekávané četnosti), by porovnáme s pozorovanými n_j:

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{[n_j - n \cdot p_j(\theta)]^2}{n \cdot p_j(\theta)}$$

$\tilde{\theta}$ je odhad parametru θ metodou minimálního χ^2, jestliže $\tilde{\theta} = \operatorname{argmin}_{\theta} \chi^2(\theta)$.

Poznámky

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{n_j^2 - 2 n n_j p_j(\theta) + n^2 p_j^2(\theta)}{n p_j(\theta)} = \sum_{j=1}^{k} \frac{n_j^2}{n p_j(\theta)} - 2 \cdot \underbrace{\sum_{j=1}^{k} n_j}_{=n} + n \underbrace{\sum_{j=1}^{k} p_j(\theta)}_{=1} = \sum_{j=1}^{k} \frac{n_j^2}{n p_j(\theta)} - n.$$

je-li θ_p - rozměrný parametr a $p_j(\theta)$ jsou diferencovatelné podle θ, pak pro $\tilde{\theta}$ platí:

$$-\sum_{j=1}^{k} \frac{n_j^2}{n p_j^2(\theta)} \cdot \frac{\partial p_j(\theta)}{\partial \theta_h} = 0 \quad \text{pro } h = 1, 2, \dots, p.$$

$\tilde{\theta}$ řeší soustavu: $\sum_{j=1}^{k} \frac{n_j^2}{p_j^2(\theta)} \cdot \frac{\partial p_j(\theta)}{\partial \theta_h} = 0$ pro $h = 1, 2, \dots, p$.

Poznámky

Metoda nedá použít i na „klasická" data, kdy $X_1, \dots, X_n$ tvoří náh. výběr z rozdělení s hustotou $f(x,\theta)$. Intervaly $B_1, \dots, B_k$ si vybíráme sami a dopočítáme $n_1, \dots, n_k$.

Jak volit k? Heuristická pravidla: $k \approx 2n^{\frac{2}{5}}$, nebo $k \approx 15 \cdot \left(\frac{n}{100}\right)^{\frac{2}{5}}$.

Intervaly $B_1, \dots, B_k$ se volí tak, aby všechny byly stejně pravděpodobné", tj. $p_1(\tilde{\theta}) = p_2(\tilde{\theta}) = \dots = p_k(\tilde{\theta}) = \frac{1}{k}$.

$\tilde{\theta}$ ale neznáme!

V praxi vezmeme nějaký jiný odhad θ^* parametru θ a požadujeme, aby $p_1(\theta^*) = p_2(\theta^*) = \dots = p_k(\theta^*) = \frac{1}{k}$.

Jinými slovy $c_1, c_2, \dots, c_{k-1}$ jsou $\frac{1}{k}, \frac{2}{k}, \dots, \frac{k-1}{k}$ - kvantily rozdělení s hustotou $f(x, \theta^*)$.