Neparametrické metody

- Přestože parametrické metody zaujímají klíčovou úlohu ve statistické analýze dat, je možné některé problémy řešit i při neparametrickém přístupu.

- V této přednášce uvedeme neparametrické odhady funkce spolehlivosti (doplněk distribuční funkce do jedničky), pomocí kterých lze odhadnout momenty doby do poruchy a některé jiné charakteristiky.

- Nevýhodou všech neparametrických odhadů je však nemožnost získat informaci o chování chvostů rozdělení. Na druhou stranu výhodou odhadů, které zde odvodíme, je možnost jejich sestrojení z poměrně „divokých“ dat. Tak můžeme vyřešit nejednu situaci, se kterou se v praxi setkáme.

- Uvedené odhady se často používají nejen v teorii spolehlivosti, ale i v **klinickém výzkumu a pojišťovnictví, při analýze tabulek úmrtnosti.**

KAPLAN-MEIERův odhad (angl. též product limit estimate)

Nechť X označuje náhodnou veličinu „dobu do poruchy“, $R(x)$ odpovídající funkci spolehlivosti, $0 = x_0 < x_1 < \ldots < x_k$ „význačné“ časové okamžiky a $J_i = (x_{i-1}, x_i]$, $i = 1, \ldots, k$. Předpokládejme, že experiment je uspořádán tak, že v okamžiku $x_0 = 0$ začneme pozorovat n identických prvků, a že údaje o průběhu experimentu můžeme zjišťovat pouze v časových okamžicích $x_1, \ldots, x_k$. Získáme údaje jsou:

n_i … počet prvků neporouchaných a sledovaných do okamžiku x_{i-1} (okamžik x_{i-1} v to nepočítaje),

d_i … počet poruch v intervalu J_i,

v_i … počet prvků, které se ztratily ze sledování v intervalu J_i,

w_i … počet prvků, které byly záměrně vyjmuty ze sledování $i = 1, \ldots, k$.

Položme $n_0 = n$, $d_0 = v_0 = w_0 = 0$.

Pro $t \in J_i$ můžeme hodnotu $R(t)$ odhadnout pomocí

$$est\ R(t) = 1 - \frac{d_{i-1}}{n_{i-1}}$$

KAPLAN-MEIERův odhad

Tento odhad (někdy nazývaný odhadem založeným na redukovaném výběru) však ignoruje informaci obsaženou ve v_i a w_i. Myšlenka zahrnut i tuto informaci je založena na následující pravděpodobnostní úvaze. Označme $E_i = \{X \geq x_i\}$, $p_i = P(E_i \mid E_{i-1})$, $i = 1, \ldots, k$. Potom zřejmě platí

$$P\left(E_k\right) = P\left(E_k \middle| E_{k-1}\right) P\left(E_{k-1}\right) = \ldots = P\left(E_k \middle| E_{k-1}\right) P\left(E_{k-1} \middle| E_{k-2}\right) \ldots P\left(E_1\right)$$

$$= \prod_{i=1}^{k} p_i$$

(1)

KAPLAN-MEIERův odhad

V případě, že v intervalu J_i nedošlo ke „ztrátám“, a že žádné prvky nebyly vyjmuty ze sledování, můžeme p_i odhadnout pomocí veličiny $1 - d_i / n_i$. V případě, že v intervalu J_i došlo ke „ztrátám“ nebo záměrnému vyjmutí ze sledování, se předpokládá, že ztracené a vyjmuté prvky byly sledovány polovinu příslušného intervalu. Potom počítáme s tzv. efektivním počtem prvků sledovaných v intervalu J_i

$$n_i = n_i - \frac{1}{2}\left(v_i + w_i\right) \qquad (2)$$

Za odhad podmíněných pravděpodobností pi potom vezmeme

$$\tilde{p}_i = 1 - \frac{d_i}{n_i} \qquad (3)$$

Což spolu se (1) vede k odhadu funkce spolehlivosti

$$\tilde{R}(t) = \prod_{i=1}^{j} \tilde{p}_i \qquad x_j < t \le x_{j+1} \quad j = 1,\ldots,k-1 \qquad (4) \qquad \tilde{R}(t) = 1, \qquad t < x_1$$

KAPLAN-MEIERův odhad

Vychází z vyjádření (1), ale za „význačné“ okamžiky bere přímo okamžiky, kdy se prvek porouchal nebo byl vyjmut ze sledování. Podobně jako (4), ani Kaplan-Meierův odhad nerozlišuje prvky, které se ztratily, a prvky, které byly vyjmuty ze sledování. Můžeme proto předpokládat, že data jsou náhodně cenzorována a výsledkem experimentu je n dvojic

$$(W_1, I_1) \ldots (W_n, I_n)$$

kde W_j je okamžik poruchy resp. vyjmutí j-tého prvku ze sledování a $I_j = 1$ resp. $I_j = 0$ podle toho, zda dříve došlo k poruše resp. vyjmutí.

Předpokládejme, že ve výběru $W_1, \ldots, W_n$ nedošlo ke shodám, a utvořme uspořádaný náhodný výběr $W_{(1)} < \ldots < W_{(n)}$. Nechť $I_{(j)}$ je indikátor odpovídající $W_{(j)}$, $j = 1, \ldots, n$. (Pozor, $I_{(1)}, \ldots, I_{(n)}$ nejsou uspořádána!). Za význačné okamžiky vezmeme $W_{(1)} < \ldots < W_{(n)}$ .Označme nyní

n_i … počet prvků neporouchaných do okamžiku $W_{(i)}$ (okamžik $W_{(i)}$ v to nepočítejte),

d_i … počet poruch v okamžiku $W_{(i)}$.

Potom za odhady podmíněných pravděpodobností p_i můžeme vzít

$$\hat{p}_i = 1 - \frac{d_i}{n_i} \qquad i = 1,...n \tag{5}$$

Poznamenejme, že

$\hat{p}_i = 1 - \frac{1}{n_i}$ jestliže $I_{(i)} = 1$

$\hat{p}_i = 1$, jestliže $I_{(i)} = 0$

Kaplan-Meierův odhad funkce spolehlivosti je potom

$$\hat{R}(t) = \prod_{i:W_{(i)}<t} \hat{p}_i \qquad t \le W_{(n)},$$

$$\hat{R}(t) = 0 \qquad t > W_{(n)}, \tag{6}$$

A prázdný součin definujeme jako rovný jedné, tj.

$$\hat{R}(t) = 1 \qquad t \le W_{(1)}$$

KAPLAN-MEIERův odhad

Alternativní tvar Kaplan-Meierova odhadu je

$$\hat{R}(t) = \prod_{i:W_{(i)}<t} \left(1 - \frac{1}{n_i}\right)^{I_{(i)}} = \prod_{i:W_{(i)}<t} \left(\frac{n-i}{n-i+1}\right)^{I_{(i)}} \qquad t \le W_{(n)},$$

$$\hat{R}(t) = 0 \qquad t > W_{(n)}, \tag{7}$$

Reálná data mohou obsahovat shody. V takovém případě modifikujeme Kaplan-Meierův odhad následujícím způsobem. Nechť R_i označuje pořadí dvojic (W_i, 1 – I_i) v lexikografickém uspořádání posloupnosti

Potom modifikovaný Kaplan-Meierův odhad je

$$\hat{R}(t) = \prod_{i:W_{(i)}<t} \left(1 - \frac{1}{n_i}\right)^{I_{(i)}} = \prod_{i:W_{(i)}<t} \left(\frac{n-R_i}{n-R_i+1}\right)^{I_{(i)}} \qquad t \le W_{(n)},$$

$$\hat{R}(t) = 0 \qquad t > W_{(n)}, \tag{8}$$

KAPLAN-MEIERův odhad - příklad

Uvažujme $n = 11$ pozorování

9, 13, 13 + ,18, 23, 28 + , 31, 34, 45 + , 48, 161 + .

(Symbolem + označujeme podle úmluvy okamžiky, ve kterých došlo k cenzorování.)

Odpovídají pořadí $R_1, \ldots, R_{11}$ jsou zřejmě 1, 2, 3, 4, …, 11.

Dále $I_3 = I_6 = I_9 = I_{11} = 0$, ostatní I_i jsou rovna jedné. Kaplan-Meierův odhad je funkce schodovitá zleva spojitá, jejíž hodnoty se mění pouze v bodech $W_{(i)}$ s $I_{(i)} = 1$ a v bodě $W_{(n)}$.

V našem případě máme 11 pozorování: 9, 13, 13 + ,18, 23, 28 + , 31, 34, 45 + , 48, 161 + .

$R(9) = 1,$

$$R(13) = \left(\frac{n - R_1}{n - R_1 + 1}\right)^1 = \left(\frac{11 - 1}{11}\right)^1 = 0.91$$

$$R(18) = R(13)\left(\frac{n - R_2}{n - R_2 + 1}\right)^{I_2}\left(\frac{n - R_3}{n - R_3 + 1}\right)^{I_3} = 0.91\left(\frac{9}{10}\right)^1\left(\frac{8}{9}\right)^0 = 0.82$$

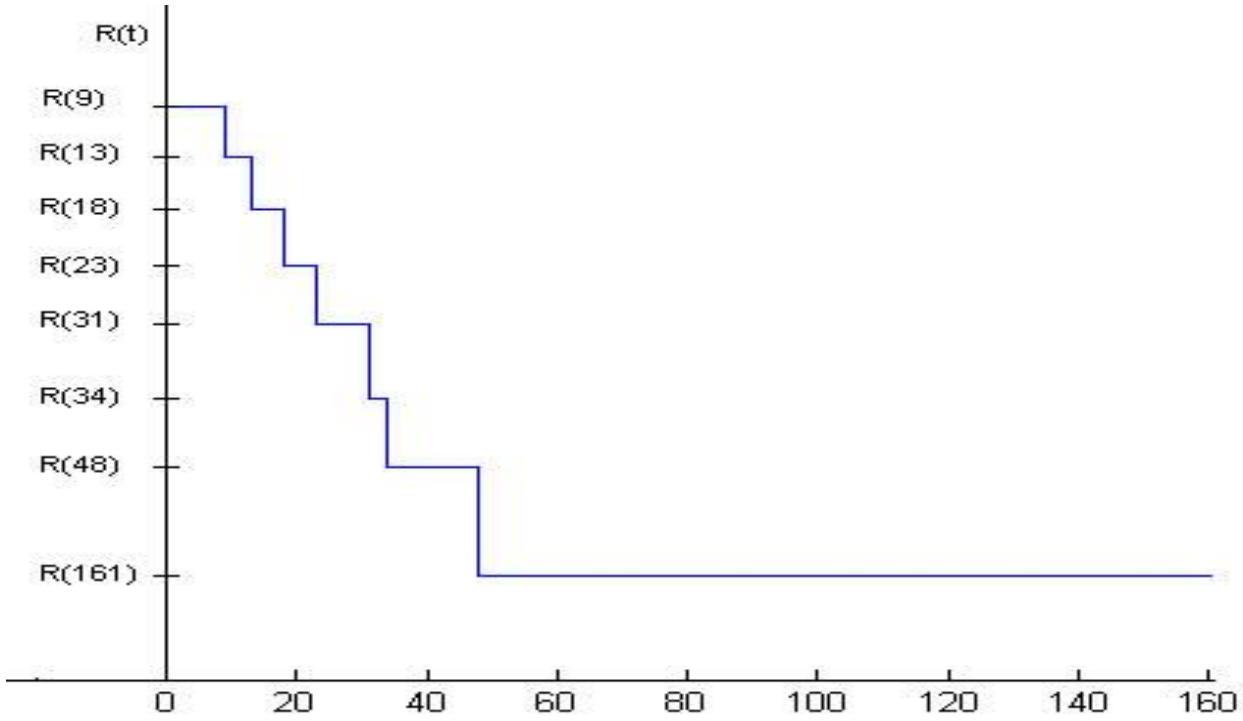
$$R(23) = R(18)\left(\frac{n - R_4}{n - R_4 + 1}\right)^{I_4} = 0.82\left(\frac{7}{8}\right)^1 = 0.72$$

$$R(31) = R(23)\left(\frac{n - R_5}{n - R_5 + 1}\right)^{I_5}\left(\frac{n - R_6}{n - R_6 + 1}\right)^{I_6} = 0.72\left(\frac{6}{7}\right)^1\left(\frac{5}{6}\right)^0 = 0.62$$

a podobně

$$R(34) = .49, \qquad R(48) = .37, \qquad R(161) = .18$$

Jestliže tedy $9 < t \leq 13$, je $R(t)=R(13) = 0.91$, jestliže $48 < t \leq 161$, pak $R(t)=R(161)=0.18$, atd.



Poznámka:

- Pro úplný výběr je Kaplan-Meierův odhad totožný s empirickou funkcí spolehlivosti.
- Asymptotické vlastnosti Kaplan-Meierův odhadu v případě náhodného cenzorování jsou uvedeny v následující větě.

Věta 1. Asymptotické rozdělení $\hat{R}(t)$

Nechť distribuční funkce F doby do poruchy X a distribuční funkce G časového cenzoru jsou spojité. Nechť $t > 0$ je takové, že $R(t) = 1 - F(t) > 0$.

Potom

$$\sqrt{n}\left(\hat{R}(t)-R(t)\right)\xrightarrow{D}N\left(0,R^2(t)\int_0^t\left(\left(1-F(x)\right)\left(1-G(x)\right)\right)^{-2}dP\left(X<x,I=1\right)\right) \tag{9}$$

Bez důkazu.

Rozptyl aproximujeme nejčastěji pomocí

$$Var\,\hat{R}(t)=\hat{R}^2(t)\sum_{i:W_{(i)}<t}\frac{I_{(i)}}{(n-i)(n-i+1)}$$

což je tzv. **GREENWOODova formule.**

GREENWOODova formule

V praxi je třeba nahradit rozptyl asymptotického rozdělení ve větě 1. nějakým odhadem. Jeden z možných postupů je tento.

Předně je patrné, že pravděpodobnost $P(X < x, I = 1)$ je možné odhadnout pomocí relativní četnosti jako

$$\hat{P}(X < x, I = 1) = \frac{1}{n} \sum_i I_i,$$

Takže $\hat{P}$ jakožto funkce x má skoky velikosti $1/n$ v bodech W_i s $I_i = 1$.

Dále $H(x) = 1 - (1 - F(x))(1 - G(x))$ je distribuční funkce náhodné veličiny W, takže H můžeme odhadnout pomocí obyčejné empirické distribuční funkce založené na výběru $W_1, \ldots, Wn$:

$$\hat{H}(x) = \frac{1}{n} \sum_{i\ 1} I(W_i < x).$$

Vzhledem k chování $\hat{P}$ potřebujeme znát odhady $\hat{H}$ pouze v bodech $W_{(i)}$.

Místo $(1 - \hat{H}(x))^2$ v (9) použijeme „symetrizovaný“ odhad

$$\left(1 - \hat{H}(x)\right)\left(1 - \hat{H}(x+)\right).$$

Vzhledem k tomu, že

$$\hat{H}\left(W_{(i)}\right) = \frac{i-1}{n}, H\left(W_{(i)}+\right) = \frac{i}{n},$$

je možné odhadnout rozptyl v (9) pomocí

$$\hat{R}(t) \sum_{i:W_{(i)}<t} \left(\left(1 - \hat{H}\left(W_{(i)}\right)\right)\left(1 - \hat{H}\left(W_{(i)}+\right)\right)\right) \frac{I_{(i)}}{n},$$

z čehož

$$Var\,\hat{R}(t) = \hat{R}^2(t) \sum_{i:W_{(i)}<t} \frac{I_{(i)}}{(n-i)(n-i+1)} \qquad (10)$$

Poslední vzorec je v literatuře znám jako **GREENWOODova formule.**