Neparametrické metody odhadů z neúplných výběrů

Dříve jsme předpokládali, že rozdělení doby čekání na rizikový jev bylo popsáno distribuční funkcí známého typu (např. exponenciální, gama apod.), tato distribuční funkce závisela na neznámých parametrech a cílem statistické analýzy byl odhad těchto neznámých parametrů případně testování hypotéz vztažených k těmto parametrům. Existují však situace, kdy uživatel statistických metod nemá žádnou představu o rozdělení doby čekání na rizikový jev nebo není možné pro danou situaci vybrat vhodné modelové rozdělení ze základních typů rozdělení. V takových situacích se vychází z tzv. **neparametrických metod**, které umožňují stanovit odhad funkce přežití, aniž bychom předpokládali její příslušnost do dané třídy rozdělení. Popíšeme dvě metody odhadu pravděpodobnosti přežití. První bude vycházet ze setříděných dat do podobné tabulky jako je tabulka skupinového rozdělení četností. Jde o klasickou metodu známou v literatuře. Tato metoda je hojně používaná v epidemiologii, biometrice a pojišťovací matematice. Druhá metoda, kterou se budeme zabývat, je metoda podobná, ale vychází z dat, která nejsou seskupena do třídních intervalů. Tato metoda odhadu funkce přežití je založena na tzv. **Kaplan-Meierově odhadu funkce přežití**, též se v literatuře nazývá **product-limit estimator**.

1 Metoda „life-table“

Předpokládejme, že doba čekání X na rizikový jev (životnost) má distribuční funkci $F(x)$ a odpovídající doba přežití je $S(x) = 1 - F(x)$. Vyjdeme z náhodně cenzorovaných dat, kdy rozdělení časového cenzoru není známé. Metoda „lifeťable“ sestává ze tří hlavních kroků.

1. Předpokládejme, že na začátku období sledujeme životnost n statistických jednotek. Obor možných hodnot životnosti těchto jednotek (obor hodnot náhodné veličiny X) rozdělíme do $k+1$ disjunktních intervalů $I_j = \langle a_{j-1}, a_j),\ j = 1, 2, \ldots, k+1$, klademe $a_0 = 0$ a $a_{k+1} = \infty$.

2. Pro každý interval $I_j,\ j = 1, 2, \ldots, k+1$, označíme

   n_j – počet statistických jednotek, které jsou v riziku v intevalu I_j (tj. počet těch statistických jednotek, u nichž do konce intervalu I_j, tedy do času a_j, rizikový jev nenastal (říkáme, že jsou v čase a_j v živém stavu) a zároveň nejsou do času a_j cenzorovány, tedy nebyly do času a_j vyjmuty ze sledování),

   d_j – počet statistických jednotek u nichž byl v intervalu I_j pozorován rizikový jev (zemřely v intervalu I_j),

   w_j – počet cenzorovaných statistických jednotek v intervalu I_j (tedy počet statistických jednotek vyjmutých ze sledování v intervalu I_j),

   $p_j = P(X > a_j | X > a_{j-1})$ – pravděpodobnost, že rizikový jev u sledované statistické jednotky nenastane v intervalu I_j za podmínky, že na začátku intervalu I_j byla tato jednotka v živém stavu.

Zřejmě

$$p_j = \frac{P(X > a_j)}{P(X > a_{j-1})} = \frac{S(a_j)}{S(a_{j-1})},\ j = 1, 2, \ldots .$$

Pak $S(a_j)$ lze vyjádřit ve tvaru

$$S(a_j) = S(a_1)\,\frac{S(a_2)}{S(a_1)}\,\frac{S(a_3)}{S(a_2)} \cdots \frac{S(a_j)}{S(a_{j-1})}\,.$$

Odtud, protože $S(a_0) = 1$, dostaneme pro funkci přežití v časech a_j vztah

$$S(a_j) = p_1 \cdot p_2 \cdot \cdots \cdot p_j. \tag{1}$$

Tento vztah v dalším kroku užijeme k odhadu funkce přežití.

3. Odhadneme podmíněné pravděpodobnosti p_j. Vyjdeme ze vztahu

$$p_j = P(X > a_j | X > a_{j-1}) = 1 - P(X \leq a_j | X > a_{j-1})$$

a podmíněnou pravděpodobnost $P(X \leq a_j | X > a_{j-1})$ odhadneme pomocí četností, které se vztahují k intervalu I_j. Protože neznáme rozdělení cenzorovaných dob, lze předpokládat, že v prvním krajním případě, tj. na začátku intervalu I_j v čase a_{j-1} je v riziku $n_j - w_j$ statistických jednotek, protože všechny cenzorované v tomto intervalu budou vyjmuty ze sledování hned na začátku tohoto intervalu. V druhém krajním případě, když budeme předpokládat, že všechny cenzorované statistické jednotky v tomto intervalu budou ze sledování vyjmuty až na jeho konci, dostaneme, že v riziku bude v tomto intervalu n_j statistických jednotek. Průměr těchto dvou krajních hodnot je roven $n'_j = \frac{1}{2}(n_j - w_j + n_j)$ a nazývá se efektivní počet statistických jednotek, které jsou v intervalu I_j v riziku. Podmíněnou pravděpodobnost p_j pak lze odhadnout jako

$$\tilde{p}_j = 1 - \frac{d_j}{n'_j} = 1 - \frac{d_j}{n_j - \frac{1}{2}w_j}.$$

Po dosazení těchto odhadů do vzorce (1) dostaneme odhad $\tilde{S}(a_j)$ funkce přežití $S(a_j)$ ve tvaru

$$\tilde{S}(a_j) = \tilde{p}_1\tilde{p}_2 \cdots \tilde{p}_j = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i - \frac{1}{2}w_i}\right) = \prod_{i=1}^{j} (1 - \tilde{q}_i), \tag{2}$$

kde $\tilde{q}_i = 1 - \tilde{p}_i = \frac{d_i}{n'_i}$.

Podotkněme na závěr, že uvedený vzorec byl odvozen za předpokladu, že mechanismus, který generuje časy cenzorování, nezávisí na dobách životnosti a dále, že doby nastoupení rizikového jevu a doby vyjmutí prvku ze sledování jsou v každém ze sledovaném intervalu rozděleny rovnoměrně.

Příklad

Bylo sledováno 913 pacientů se zhoubným melanomem, kteří byli vyšetřeni v Andersenově nádorové klinice v letech 1944–1960. Byla získána data uvedená v tabulce. V této tabulce jsou zároveň uvedeny podklady pro výpočet funkce přežití získané metodou „life-table“ podle vzorce (2).

Interval I_j (roky)	Počet zemřelých d_j	Počet cenz. w_j	Počet v riziku n_j	Efektivní počet v riziku n'_j	$\tilde{q}_j = \frac{d_j}{n'_j}$	$\tilde{p}_j$	$\tilde{S}(a_j)$
$\langle 0, 1)$	312	96	913	865,0	0,361	0,639	0,639
$\langle 1, 2)$	96	74	505	468,0	0,205	0,795	0,508
$\langle 2, 3)$	45	62	335	304,0	0,148	0,852	0,433
$\langle 3, 4)$	29	30	228	213,0	0,136	0,864	0,374
$\langle 4, 5)$	7	40	169	149,0	0,047	0,953	0,356
$\langle 5, 6)$	9	37	122	103,0	0,087	0,913	0,325
$\langle 6, 7)$	3	17	76	67,0	0,044	0,956	0,311
$\langle 7, 8)$	1	12	56	50,0	0,020	0,980	0,305
$\langle 8, 9)$	3	8	43	39,0	0,077	0,923	0,281
$\langle 9, \infty)$	32	–	32	32,0	1,000	0,000	0,000

Výpočet odhadu funkce přežití budeme demonstrovat pro hodnoty $\tilde{S}(a_1)$ a $\tilde{S}(a_2)$. Postupně dostaneme

$$n'_1 = n_1 + \frac{1}{2}w_1 = 913 + \frac{1}{2}96 = 865{,}0,$$
$$\tilde{q}_1 = \frac{d_1}{n'_1} = \frac{312}{865} = 0{,}361,$$
$$\tilde{p}_1 = 1 - \tilde{q}_j = 0{,}639 \text{ a } \tilde{S}(a_1) = 0{,}639.$$

Analogicky

$$\tilde{p}_2 = 0{,}795 \text{ a } \tilde{S}(a_2) = \tilde{p}_1\tilde{p}_2 = 0{,}639 \cdot 0{,}795 = 0{,}508.$$

2 Greenwoodova formule a Kaplan Meierův odhad

Lze ukázat, že odhad $\tilde{S}(a_j)$ je nestranný a tedy platí, že $E\tilde{S}(a_j) = S(a_j)$. Tvrzení plyne ze skutečnosti, že podmíněné rozdělení počtu pozorovaných rizikových jevů d_j v intervalu I_j za předpokladu, že efektivní počet jednotek, které jsou v intervalu I_j v riziku, je rozdělení binomické $B_i(n'_j, 1-p_j),\ j = 1, 2, \ldots, k+1$. Pomocí tohoto výsledku lze také stanovit rozptyl odhadu $D(\tilde{S}(a_j))$. Platí tzv. **Greenwoodova formule**

$$D(\tilde{S}(a_j)) \approx (S(a_j))^2 \sum_{i=1}^{j} \frac{q_i}{p_i n'_i},\ \ j = 1, 2, \ldots, k+1,$$

která ukazuje, jak lze rozptyl $D(\tilde{S}(a_j))$ aproximovat. Po dosazení odhadů za $S(a_j),\ q_i,\ p_i$, vypočteme odhad rozptylu $\tilde{S}(a_j)$. Označme jej $\sigma^2_{\tilde{S}(a_j)} = D(\tilde{S}(a_j))$ a pro tento rozptyl dostaneme přibližné vyjádření

$$\widehat{\sigma_j^2} = \widetilde{D(\tilde{S}(a_j))} \approx \left(\tilde{S}(a_j)\right)^2 \sum_{i=1}^{j} \frac{\tilde{q}_i}{\tilde{p}_i n'_i},\ \ j = 1, 2, \ldots, k+1,$$

a pro odhad směrodatné odchylky $\tilde{S}(a_j)$ pak dostaneme

$$s_{\tilde{S}(a_j)} = \tilde{S}(a_j)\sqrt{\sum_{i=1}^{j} \frac{\tilde{q}_i}{\tilde{p}_i n'_i}}.$$

Dále lze stanovit $100(1-\alpha)\%$ asymptotický interval spolehlivosti pro $S(a_j)$ ve tvaru

$$\left(\tilde{S}(a_j) - u_{1-\frac{\alpha}{2}} s_{\tilde{S}(a_j)}; \tilde{S}(a_j) + u_{1-\frac{\alpha}{2}} s_{\tilde{S}(a_j)}\right).$$

Použití uvedených vzorců budeme ilustrovat na příkladu. Stanovíme 95% interval spolehlivosti pro $S(a_5)$. Postupně dostaneme

$$s_{\tilde{S}(a_5)} \approx \tilde{S}(a_5)\sqrt{\sum_{i=1}^{j} \frac{\tilde{q}_i}{\tilde{p}_i n'_i}} = 0{,}356\sqrt{\frac{0{,}361}{0{,}639 \cdot 865} + \cdots + \frac{0{,}047}{0{,}953 \cdot 149}} = 0{,}01899.$$

Odtud 95% interval spolehlivosti pro $S(a_5)$ je roven

$$(0{,}356 - 1{,}96 \cdot 0{,}01899; 0{,}356 + 1{,}96 \cdot 0{,}01899) \doteq (0{,}318; 0{,}394).$$

Stejně jako u metody „life-table“ budeme vycházet z předpokladu, že doba čekání X na rizikový jev má funkci přežití $S(x)$. Na rozdíl od metody „life-table“ nepředpokládáme, že pozorování náhodné veličiny X jsou seskupena do intervalů. Pro každou statistickou jednotku j dále označme:

X_j životnost jednotky j,

T_j doba cenzorování jednotky j (doba sledování jednotky j před jejím vyjmutím ze sledování),

$W_j = \min(X_j, T_j),\ j = 1, 2, \ldots, n,$

$I_j = 1$ když $X_j < T_j$ (tj. $W_j = X_j$) a $I_j = 0$ když $X_j > T_j$ (tj. $W_j = T_j$). Výsledkem sledování n statistických jednotek pak je n dvojic $(W_1, I_1), \ldots, (W_n, I_n)$. Když dvojice uspořádáme podle velikosti první složky (tedy podle W_j), dostaneme uspořádané dvojice

$$\left(W_{(1)}, I_{(1)}\right), \left(W_{(2)}, I_{(2)}\right), \ldots, \left(W_{(n)}, I_{(n)}\right), \text{ přičemž } W_{(1)} \le W_{(2)} \le \cdots \le W_{(n)}.$$

Kaplan-Meierův odhad funkce přežití je dán vztahem

$$\widehat{S}(x) = \prod_{i:\, W_{(i)} \le x} \left(\frac{n-i}{n-i+1}\right)^{I_{(i)}}$$

a pro $I_{(n)} = 0$, klademe $\widehat{S}(x) = 0$ pro $x > W_{(n)}$.

Lze ukázat, že $\widehat{S}(x)$ má asymptoticky normální rozdělení, jeho střední hodnotu lze odhadnout jako $\int_0^\infty \widehat{S}(x)dx$ a jeho rozptyl lze opět odhadnout Greenwoodovou formulí.

Příklady k procvičení

1. U 21 pacientů byla sledována doba remise, za kterou pacienti po vyléčení akutní leukemie pomocí preparátu 6-mercaptopurine (6-MP) opětovně upadli do akutního stavu této nemoci. Zjištěné doby remise v měsících byly následující: 10, 7,32*,23, 22, 6, 16, 34*, 32*, 25*, 11*, 20*, 19*, 6,17*,35*,6, 13, 9*, 6*, 10*. Hodnoty označené * značí náhodně cenzorované pozorování. Data viz.: Klein, J. P. and Moeschberger, M. L. Survival Analysis. Springer, 2003.

   a) Stanovte z těchto dat časy t_i, kdy došlo k remisi, dále tomu času odpovídající počty pacientů d_i, u nichž došlo v tomto čase k remisi a konečně počty jedinců Y_i, kteří byli v tomto čase v riziku, tedy neznalo se, kdy u nich k případné remisi může dojít.

   b) Stanovte neparametrický Kaplan -Meierův odhad funkce přežití.

   c) Pomocí Greenwoodovy formule stanovte odhad rozptylu tohoto odhadu.

   Úlohu můžete řešit přímo nebo pomocí software R nebo MATLAB.