Odhad funkce přežití

Jana Kahounová*

Předmětem zájmu metod analýzy přežití je sledování určitého jevu na daných objektech. Tomuto jevu budeme říkat „selhání“, i když se nemusí ve všech případech jednat o negativní událost. Metody analýzy přežití se zabývají daty – hodnotami náhodné veličiny X, která představuje dobu do výskytu určitého jevu – dobu od počátku výzkumu do „selhání“. Tyto metody se zpočátku využívaly zejména v medicíně a pro náhodnou veličinu X se vžil termín doba života nebo doba přežití. Využití těchto metod je však daleko širší, např. ve strojírenství můžeme sledovat dobu do poruchy nějakého zařízení. Odtud také pochází alternativní název – teorie spolehlivosti. V tomto textu budeme nezápornou náhodnou veličinu X nazývat *doba do poruchy*.

Předpokládejme, že začneme pozorovat n objektů stejného typu, přičemž experiment je organizován tak, že pozorování začne u všech objektů ve stejném okamžiku $t = 0$ a pozorování budeme provádět tak dlouho, dokud všechny objekty nebudou mít poruchu. Takto získáme *úplný náhodný výběr*, který tvoří nezávislé náhodné veličiny X_j, $j = 1, 2, ..., n$ představující skutečné doby do poruchy u jednotlivých objektů. V mnoha případech se jedná o dlouhý časový horizont a z technických či ekonomických důvodů nemůžeme provést experiment až do konce. Musíme tedy vhodně rozhodnout o ukončení pozorování. V tomto případě dostaneme *neúplný – cenzorovaný náhodný výběr*, kdy máme úplnou informaci jen u $r < n$ pozorování.

Podle toho jaká pravidla si stanovíme pro ukončení pozorování, pracujeme s různými modely cenzorování.

Cenzorování typu I

Při cenzorování typu I je pozorování ukončeno v předem určeném okamžiku $T > 0$. Číslo T je tzv. časový cenzor. Jedná se zřejmě o cenzorování časem. Tímto postupem získáme prvních r pozorování – pořádkových statistik

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(r)} \le T, \quad X_{(r)} \le T \le X_{(r+1)}.$$

O zbývajících $n - r$ objektech víme pouze to, že u nich je skutečná doba do poruchy větší než T, tzn. $X_{(j)}$, $j = r + 1, r + 2, ..., n$ jsou cenzorovaná pozorování.

* Doc. Ing. Jana Kahounová, CSc.; Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky, VŠE v Praze, kahoun@vse.cz.

Cenzorování typu II.

Tak jako v předešlém případě všech n objektů vstupuje do pozorování současně. Cenzorování typu II je cenzorování poruchou. Je předepsán počet poruch r, $1 \le r \le n$. Doba trvání experimentu je doba do poruchy r-tého objektu $X_{(r)}$. Opět zde máme úplnou informaci u prvních r pozorování

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(r)},$$

o ostatních víme, že jsou větší než $X_{(r)}$.

Náhodné cenzorování

V tomto případě sledujeme dobu do poruchy X a časový cenzor T. Označíme jako náhodnou veličinu $W_j = \min(X_j, T_j)$, $j = 1, 2, ..., n$. U tohoto typu cenzorování nepředpokládáme, že všech n objektů vstupuje do experimentu současně. Náhodná veličina T_j, $j = 1, 2, ..., n$ pak představuje dobu do cenzorování u j-tého objektu. Předpokládáme, že náhodné veličiny X_j a T_j jsou nezávislé. Dále zavedeme veličinu C_j, tzv. cenzorovací index, který nám poskytuje informaci o tom, zda se jedná o skutečnou dobu do poruchy nebo dobu do cenzorování. V případě, že pozorování je v čase T_j necenzorované, bude $C_j = 1$ a $W_j = X_j$. U cenzorovaných pozorování bude $C_j = 0$ a $W_j = T_j$. Výsledkem experimentu bude nyní dvourozměrný náhodný výběr s prvky (W_j, T_j), j = 1, 2, ..., n.

Funkce přežití a další formy popisu rozdělení doby do poruchy

Uvažujme nezápornou náhodnou veličinu X spojitého typu představující dobu do poruchy s distribuční funkcí

$$F(x) = P(X \le x) \tag{1}$$

a s hustotou pravděpodobnosti $f(x)$, přičemž ve všech bodech, kde existuje derivace distribuční funkce je

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x). \tag{2}$$

Funkce $F(x)$ udává pravděpodobnost, že v intervalu $(0, x\rangle$ dojde k poruše.

Funkce přežití, resp. funkce spolehlivosti $S(x)$ vyjadřuje pravděpodobnost, že v intervalu $(0, x\rangle$ k poruše nedojde

$$S(x) = P(X > x) = 1 - F(x) = \int_x^\infty f(t)\mathrm{d}t. \tag{3}$$

Tyto i další funkce budeme uvažovat jen na intervalech $(0, \infty)$, protože jejich hodnoty na intervalu $(-\infty, 0)$ jsou vzhledem k nezápornosti náhodné veličiny X zřejmé. Např. pro $x < 0$ je $F(x) = 0$, $S(x) = 1$.

Uvedené funkce jsou dvě ekvivalentní vyjádření rozdělení nezáporné spojité náhodné veličiny. Je možno použít i další funkce jako např. funkci rizikovou. *Riziková funkce r*(*x*) je definována vztahem

$$r(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(x \le X < x + \Delta \mid X \ge x),\ x > 0, \Delta > 0. \tag{4}$$

Z definice podmíněné pravděpodobnosti dostaneme

$$r(x) = \frac{f(x)}{S(x)},\ \ S(x) > 0. \tag{5}$$

Protože

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[1 - S(x)\right] = -\frac{\mathrm{d}}{\mathrm{d}x}\left[S(x)\right],$$

bude

$$r(x) = \left[-\frac{\mathrm{d}}{\mathrm{d}x} S(x) \right] \Big/ S(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \ln S(x). \tag{6}$$

Výraz $\Delta r(x)$ udává přibližně pravděpodobnost poruchy ve velmi krátkém časovém intervalu $(x, x + \Delta)$ podmíněnou bezporuchovým provozem po dobu x. Funkci $r(x)$ můžeme interpretovat jako riziko poruchy za jednotku času (nebo také jako intenzitu poruch) během procesu stárnutí.

Odhad funkce přežití

Je-li známo, že pravděpodobnostní rozdělení doby do poruchy je popsáno funkcí $F(x, \boldsymbol{\Theta})$ kde $\boldsymbol{\Theta} = \left[\Theta_1, \Theta_2, \ldots, \Theta_P\right]'$ je vektor neznámých parametrů, na nichž toto rozdělení závisí, pak odhad funkce přežití stanovíme obvykle na základě odhadu parametru $\boldsymbol{\Theta}$.

Předpokládejme, že jako model rozdělení doby do poruchy použijeme jednoparametrické exponenciální rozdělení s hustotou pravděpodobnosti

$$\begin{aligned} f(x;\lambda) &= \lambda \mathrm{e}^{-\lambda x}, x > 0, \lambda > 0, \\ &= 0 \text{ jinak.} \end{aligned} \tag{7}$$

Funkce přežití v bodě x, tj. pravděpodobnost, že zařízení bude pracovat bez poruchy po dobu alespoň rovnou x

$$S(x) = \mathrm{e}^{-\lambda x}. \tag{8}$$

K odhadu parametrické funkce $S(x)$ je možno také použít Raovu-Blackwellovu větu (viz např. [1]). Vyjdeme z jednoduchého nestranného odhadu

$$U = 1 \quad \text{pro} \quad X_1 \geq x,$$
$$= 0 \quad \text{pro} \quad X_1 < x$$

a stanovíme úplnou postačující statistiku pro λ

$$V = \sum_{j=1}^{n} X_j,$$

která má rozdělení Gama s parametry n a λ. Nejlepším nestranným odhadem funkce $S(x)$ je pak statistika $\varphi(V)$, pro kterou platí

$$\varphi(V) = E(U \mid v) = P\left(X_1 \geq x \middle| \sum_{j=1}^{n} X_j = v \right), \quad \text{tj.}$$

$$\varphi(V) = \left(1 - \frac{x}{\sum_{j=1}^{n} X_j} \right)^{n-1} \quad \text{pro} \quad \sum_{j=1}^{n} X_j \geq x,$$

$$= 0 \qquad\qquad \text{pro} \quad \sum_{j=1}^{n} X_j < x.$$

U výběrů většího rozsahu lze použít maximálně věrohodné odhady. Pro maximálně věrohodný odhad parametrické funkce $\gamma(\boldsymbol{\Theta}) = \gamma(\Theta_1, \Theta_2, \ldots \Theta_P)$ platí

$$\hat{\gamma}(\Theta_1, \Theta_2, \ldots \Theta_P) = \gamma(\hat{\Theta}_1, \hat{\Theta}_2, \ldots \hat{\Theta}_P). \tag{9}$$

Z rozdělení $f(x; \boldsymbol{\Theta})$ pořídíme úplný náhodný výběr $\mathbf{X} = \left[X_1, X_2, \ldots, X_n \right]'$. Věrohodnostní funkce při daném $\mathbf{x} = \left[x_1, x_2, \ldots, x_n \right]'$ má tvar

$$L\left(\mathbf{x}; \boldsymbol{\Theta}\right) = \prod_{j=1}^{n} f(x_j; \boldsymbol{\Theta}). \tag{10}$$

Maximálně věrohodný odhad $\hat{\boldsymbol{\Theta}} = \left[\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_P \right]$ dostaneme řešením soustavy věrohodnostních rovnic

$$\frac{\partial \ln L(\mathbf{x}; \boldsymbol{\Theta})}{\partial \Theta_k} = 0, \quad k = 1, 2, \ldots P. \tag{11}$$

Nyní budeme uvažovat neúplný náhodný výběr, který tvoří náhodně cenzorovaná pozorování. Předpokládáme, že X a T jsou nezávislé spojité náhodné veličiny. Rozdělení náhodné veličiny X závisí na vektorovém parametru $\boldsymbol{\Theta}_1$ a je popsáno distribuční funkcí F, resp. hustotou pravděpodobnosti f a rozdělení náhodné veličiny T závisí na vektorovém parametru $\boldsymbol{\Theta}_2$ a je popsáno distribuční funkcí G, resp. hustotou pravděpodobnosti g. Je možno dokázat, že rozdělení náhodného vektoru (**W**, **C**) má hustotu pravděpodobnosti

$$h(w,c) = \left\{ f(w)\left[1-G(w)\right]\right\}^{c} \left\{ g(w)\left[1-F(w)\right]\right\}^{1-c}, \quad w>0, \quad c=0,1. \tag{12}$$

Pak věrohodnostní funkce při daném výsledku experimentu bude

$$L(\mathbf{w},\mathbf{c},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) = L(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) = \prod_{j=1}^{n} h(w_j,c_j;\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) = \prod_{j=1}^{n} \left\{ f(w_j)\left[1-G(w_j)\right]\right\}^{c_j} \left\{ g(w_j)\left[1-F(w_j)\right]\right\}^{1-c_j}.$$

Označíme množinu indexů necenzorovaných pozorování jako R a množinu cenzorovaných pozorování jako Z. Jak bylo již výše řečeno, jestliže $j \in R$, pak $c_j = 1$ a jestliže $j \in Z$, pak $c_j = 0$. Po dalších úpravách věrohodnostní funkce bude

$$L(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) = \prod_{j\in R} f(w_j) \prod_{j\in Z}\left[1-F(w_j)\right] \prod_{j\in Z} g(w_j) \prod_{j\in R}\left[1-G(w_j)\right].$$

Označíme

$$\begin{aligned} L_1(\boldsymbol{\Theta}_1) &= \prod_{j\in R} f(w_j) \prod_{j\in Z}\left[1-F(w_j)\right], \\ L_2(\boldsymbol{\Theta}_2) &= \prod_{j\in Z} g(w_j) \prod_{j\in R}\left[1-G(w_j)\right]. \end{aligned} \tag{13}$$

Jestliže rozdělení uvažovaných veličin nezávisí na společných parametrech a mezi $\boldsymbol{\Theta}_1$ a $\boldsymbol{\Theta}_2$ neexistuje funkční závislost, dostaneme maximálně věrohodné odhady řešením soustavy rovnic

$$\frac{\partial \ln L_1(\boldsymbol{\Theta}_1)}{\partial \Theta_{1k}} = 0, \quad k = 1,2,\ldots P. \tag{14}$$

Celý postup ilustrujeme opět na jednoduchém příkladě jednoparametrického exponenciálního rozdělení s hustotou pravděpodobnosti (7). V případě úplného náhodného výběru bude věrohodnostní funkce (10)

$$L(\mathbf{x};\lambda) = \lambda^n \exp\left(-\lambda \sum_{j=1}^{n} x_j \right)$$

a řešením věrohodnostní rovnice (11)

$$\frac{n}{\hat{\lambda}} - \sum_{j=1}^{n} x_j = 0$$

je

$$\hat{\lambda} = \frac{n}{\sum_{j=1}^{n} x_j} = \frac{1}{\overline{x}}, \overline{x} > 0.$$

Podle vztahu (8) bude maximálně věrohodný odhad funkce přežití

$$\hat{S}(x) = \exp\left(-\frac{x}{\overline{x}} \right).$$

Při náhodně cenzorovaném výběru využijeme (13) a

$$L_1(\lambda) = \prod_{j \in R} \lambda \mathrm{e}^{-\lambda w_j} \prod_{j \in Z} \mathrm{e}^{-\lambda w_j}.$$

Označíme počet prvků množiny R jako $|R|$, takže

$$L_1(\lambda) = \lambda^{|R|} \prod_{j \in R} \mathrm{e}^{-\lambda w_j} \prod_{j \in Z} \mathrm{e}^{-\lambda w_j}.$$

Pak rovnice (14)

$$\frac{|R|}{\hat{\lambda}} - \sum_{j=1}^{n} w_j = 0$$

dává výsledek

$$\hat{\lambda} = \frac{|R|}{\sum_{j=1}^{n} w_j}$$

a odhad funkce přežití

$$\hat{S}(x) = \exp\left(-\frac{|R| x}{\sum_{j=1}^{n} w_j} \right).$$

Uvažujme nyní náhodný výběr $X_1, X_2, ..., X_n$ z rozdělení, jehož tvar neznáme. Na základě realizace tohoto náhodného výběru chceme odhadnout distribuční funkci $F(x)$. Dobrým odhadem je *empirická distribuční funkce*, kterou budeme značit $\hat{F}_n(x)$. Empirickou distribuční funkci definujeme jako funkci, která každému x přiřazuje relativní četnost pozorování menších nebo rovných x. Označíme-li M_x počet veličin X_j, které splňují nerovnosti $X_j \le x$, pak

$$\hat{F}_n(x) = \frac{M_x}{n}. \tag{15}$$

Zřejmě platí, že

$$\begin{aligned} \hat{F}_n(x) &= 0, \quad x < X_{(1)} \\ &= \frac{j}{n}, \quad X_{(j)} \le x < X_{(j+1)}, \quad j = 1, \ldots, n-1, \\ &= 1, \quad x \ge X_{(n)}. \end{aligned} \tag{16}$$

Empirická distribuční funkce je schodovitá funkce, je rovna 0 pro všechna $x < X_{(1)}$, je rovna $1/n$ pro všechna $X_{(1)} \le x < X_{(2)}$, atd. To znamená, že je konstantní v intervalech $\langle X_{(j)}, X_{(j+1)})$ se skokem $1/n$ v každém bodě $X_{(j)}$, v němž funkce $\hat{F}_n(x)$ má diskontinuitu.

Veličina $M_x = n\hat{F}_n(x)$ má binomické rozdělení s parametry n a $\pi = F(x)$. Pro střední hodnotu, resp. rozptyl náhodné veličiny, řekněme X, použijeme obvyklé symboly $E(X)$, resp. $D(X)$. Pak zřejmě platí

$$E(M_x) = E\left[n\hat{F}_n(x)\right] = nF(x),$$

$$D(M_x) = D\left[n\hat{F}_n(x)\right] = nF(x)\left[1-F(x)\right],$$

takže

$$E\left[\hat{F}_n(x)\right] = F(x), \tag{17}$$

$$D\left[\hat{F}_n(x)\right] = \frac{F(x)\left[1-F(x)\right]}{n}. \tag{18}$$

Funkce $\hat{F}_n(x)$ je tedy nestranným odhadem distribuční funkce $F(x)$ a protože

$$\lim_{n\to\infty} \frac{F(x)\left[1-F(x)\right]}{n} = 0, \tag{19}$$

je funkce $\hat{F}_n(x)$ také odhadem konzistentním.

Z Moivreovy-Laplaceovy věty plyne, že pro $n \to \infty$ má náhodná veličina

$$\frac{\left[\hat{F}_n(x) - F(x)\right]\sqrt{n}}{\sqrt{F(x)\left[1-F(x)\right]}} \tag{20}$$

asymptotické normální rozdělení $N(0;1)$.

Obdobně můžeme postupovat i pro náhodně cenzorovaná data. Nechť doby do poruchy $X_1, X_2, ..., X_n$ tvoří náhodný výběr ze spojitého rozdělení s distribuční funkcí F a doby do cenzorování $T_1, T_2, ..., T_n$ představují náhodný výběr ze spojitého rozdělení s distribuční funkcí G. Náhodné veličiny X_j, T_j jsou nezávislé. V případě náhodného cenzorování pozorujeme hodnoty veličin $W_j = \min(X_j, T_j)$ a hodnoty cenzorovaného indexu C_j, $j = 1, 2, ..., n$. Z těchto hodnot pak provedeme odhad distribuční funkce F, resp. funkce přežití S náhodných veličin $X_1, X_2, ..., X_n$. Jedná se o určitou modifikaci empirické distribuční funkce $\hat{F}_n(x)$, resp. *empirické funkce přežití* $\hat{S}_n(x) = 1 - \hat{F}_n(x)$.

Nechť $W_{(1)}, W_{(2)}, ..., W_{(n)}$ je uspořádaný náhodný výběr dob do poruchy nebo cenzorování. Vzhledem k tomu, že F, G jsou funkce spojité, nastane s pravděpodobností jedna

$$W_{(1)} < W_{(2)} < ... < W_{(n)}.$$

Označíme jako

$C_{(i)} = 1$, $W_{(i)}$ je necenzorováno,
$= 0$, $W_{(i)}$ je cenzorováno.

Pak odhad funkce přežití má tvar

$$\hat{S}(x) = \prod_{i:\, W_{(i)} \le x} \left(1 - \frac{C_{(i)}}{Y_i}\right), \quad x \le W_{(n)}, \qquad (21)$$
$$= 0, \qquad x > W_{(n)},$$

kde Y_i je počet objektů, které pozorujeme těsně před událostí v čase $W_{(i)}$, tedy

$$Y_i = n - i + 1, \quad i = 1, 2, \ldots, n.$$

Vztah (21) je tzv. Kaplanův-Meierův odhad funkce přežití. V literatuře se můžeme setkat s různými alternativními tvary tohoto odhadu. V případě, že při experimentu nenastane žádné cenzorování, tj.

$$C_{(1)} = C_{(2)} = \ldots = C_{(n)} = 1,$$

pak dosazením do (21) dostáváme

$$\hat{S}(x) = \prod_{i:\, W_{(i)} \le x} \left(1 - \frac{1}{n - i + 1}\right) \qquad (22)$$

a Kaplanův-Meierův odhad se zjednoduší na empirickou funkci přežití $\hat{S}_n(x)$.

Závěr

Analýza přežití představuje skupinu metod, které původně sloužily pro analyzování dat získaných při laboratorních výzkumech živočichů nebo při klinických studiích lidí. Využití je však daleko širší. Kromě strojírenství, kde se používá alternativní název – analýza spolehlivosti, dochází ke značnému rozšíření těchto metod i do takových oborů jako je sociologie, marketing, pojišťovnictví apod.

Jednou z forem popisu pravděpodobnostního rozdělení náhodné veličiny – doby do výskytu určité události je funkce přežití. Obsahem článku jsou možnosti odhadu této funkce v případě úplného i neúplného náhodného výběru. Jsou zde uvedeny možnosti odhadu za předpokladu určitého pravděpodobnostního modelu rozdělení sledované náhodné veličiny i jedna z možností neparametrického odhadu funkce přežití. Některé postupy jsou ilustrovány na jednoduchém případě, kdy předpokládáme exponenciální model pravděpodobnostního rozdělení uvažované náhodné veličiny.

Literatura

[1] ANDĚL, J., 1978: *Matematická statistika.* Praha, SNTL/Alfa, 1978.

[2] FLEMING, T. R. – HARRINGTON, D. P., 1991: *Counting Processes and Survival Analysis.* New York, John Wiley and Sons, 1991.

[3] HURT, J., 1984: *Teorie spolehlivosti.* Praha, SPN, 1984.

[4] LEE, E. T., 1992: *Statistical Methods for Survival Data Analysis.* New York, John Wiley, 1992.

[5] LIKEŠ, J. – MACHEK, J., 1981: *Počet pravděpodobnosti.* Praha, SNTL, 1981.

[6] LIKEŠ, J. – MACHEK, J., 1983: *Matematické statistika.* Praha, SNTL, 1983.

[7] MILLER, Jr., R. G., 1981: *Survival Analysis*. New York, John Wiley, 1981.

Odhad funkce přežití

Jana Kahounová

Abstrakt

Aplikace analýzy dat o „přežití“ (survival data) zaznamenala velký rozvoj a metody analýzy přežití se rozšířily z oblasti biomedicíny a techniky do řady dalších oborů. Problému odhadu pravděpodobnostního rozdělení doby do výskytu určitého jevu v případě, kdy jsou k dispozici cenzorovaná data, je věnována značná pozornost. V tomto článku se zabýváme možnostmi odhadu funkce přežití za různých situací. Jestliže nejsme ochotni činit parametrické předpoklady o přesné formě pravděpodobnostního rozdělení doby do „selhání“ a doby do cenzorování, ale jsme ochotni předpokládat jejich nezávislost, pak je vhodné využít Kaplanův-Meierův odhad, který, kromě jiných vhodných vlastností, je konzistentní.

Klíčová slova: analýza přežití; funkce přežití; cenzorovaná pozorování.

Estimation of Survival Function

Abstract

In the past decade applications of the statistical methods for survival data analysis have been extended beyond biomedical and reliability research to other fields. The term survival data has been used in a broad sense for data involving time to a certain event such a failure, response, death and so on. Survival times are subjected random variations and like any random variable, they form a distribution. The ability to estimate a survival distribution in the presence of censoring is important and has been studied extensively. This paper is concerned with estimators of survival function. If one is not willing to make parametric assumptions about the exact form of the underlying survival and censoring distributions but is willing to assume independence between survival and censoring variables, Kaplan and Meier provided an estimator which is consistent, among other desirable properties.

Key words: survival analysis; survival function; censored observations.