NAHODNÉ CENZOROVÁNÍ A JEHO APLIKACE

Jan HURT, MFF UK Praha

1. Úvod

Předpokládejme, že sledujeme určitý prvek (výrobek, systém), který se během doby svého využití může porouchat. Označme X dobu do poruchy (tj. doby bezporuchového chodu) tohoto prvku. Z povahy věci vyplývá, že X je nezáporná náhodná veličina. Označme $F(x;\theta)$ její distribuční funkci, která závisí na neznámém, obecně vektorovém parametru θ. Předpokládejme dále, že $f(x;\theta)$ je odpovídající hustota. V teorii spolehlivosti se používá též funkce spolehlivosti $R(x;\theta)=1-F(x;\theta)$ znamenající pravděpodobnost bezporuchového chodu na časovém intervalu $[0,x]$ a intenzita poruch $r(x;\theta) = f(x;\theta)/R(x;\theta)$ vystihující dobře chování chvostů rozdělení. Statistická analýza se pak soustředí na 1) odhad funkcí r, R, 2) odhad parametrů, 3) odhad charakteristik spolehlivosti, tj. parametrických funkcí (střední doba do poruchy, 100γ%-ní kvantil doby do poruchy, tj. tzv. $100(1-\gamma)$%-ní život apod.). V teorii spolehlivosti se zabýváme většinou chováním chvostů rozdělení, z čehož vyplývá, že parametrické metody převažují.

2. Pravděpodobnostní modely

Podobnou úlohu jako hraje v klasické parametrické analýze normální rozdělení, hraje v teorii spolehlivosti exponenciální rozdělení $Exp(\theta)$ s funkcí spolehlivosti $R(x;\theta) = e^{-x/\theta}$. Jeho charakteristickou vlastností je, že má konstantní intenzitu poruch $r(x;\theta) = 1/\theta$, což odráží skutečnost, že exponenciální rozdělení "nemá paměť". Velice oblíbené a užívané je Weibullovo rozdělení $W(\theta,\beta)$ s funkcí spolehlivosti $R(x;\theta) = \exp(-(x/\theta)^{\beta})$. Lze jím dobře vystihnout rozdělení doby do poruchy s monotonní intenzitou poruch. V závislosti na parametru β je intenzita poruch klesající ($\beta<1$), konstantní ($\beta=1$), rostoucí konkávní ($1<\beta<2$), rostoucí konvexní ($\beta\geq 2$). Dále jsou užívána rozdělení normální, logaritmicko-normální a rozdělení gamma.

3. Výběrové plány

V čase $t=0$ začneme pozorovat n identických prvků. Označme jejich doby do poruchy $X_1,\dots,X_n$. Některá možná uspořádání experimentu jsou tato:

3.1. Úplný výběr. Experiment se provádí tak dlouho, dokud se všechny prvky neporouchají. Výsledkem experimentu je náhodný výběr $X_1,\dots,X_n$.

3.2. Cenzorování typu I (cenzorování časem). Předepíše se $T>0$ zvané časový cenzor. Experiment se ukončí v okamžiku T bez ohledu na to, kolik prvků se porouchalo. Označme $X_{(1)} \leq \dots \leq X_{(n)}$ uspořádaný náhodný výběr z $X_1,\dots,X_n$. Výsledkem experimentu je potom prvních r hodnot pořádkových statistik $X_{(1)} \leq \dots \leq X_{(r)} \leq T$ a informace $X_{(r+1)} \geq T$. Zde je r náhodná veličina.

3.3. Cenzorování typu II (cenzorování poruchou). Předepíše se přirozené číslo $r\leq n$. Experiment se ukončí v okamžiku poruchy r-tého prvku. Výsledkem experimentu je prvních r hodnot pořádkových statistik $X_{(1)}\leq\dots\leq X_{(r)}$. Doba trvání experimentu je náhodná veličina.

3.4. Náhodné cenzorování. Vzniklo při studiu spolehlivosti složitých sytémů a v lékařském výzkumu. Objekty $1,\dots,n$ jsou pozorovány do okamžiků $T_1,\dots,T_n$, které jsou svou povahou náhodné. Jednoduchý přístup spočívá v převedení na cenzorování typu I s časovým cenzorem $T = \min(T_1,\dots,T_n)$. To však není příliš ekonomické. Alternativní přístup spočívá v tom, že spolu s dobou do poruchy X uvažujeme náhodnou veličinu T, tzv. časový cenzor. Položme $I_i = 1$, je-li $X_i < T_i$ a $I_i = 0$, je-li $X_i \geq T_i$. Hodnota $I_i = 1$ odpovídá tomu, že i-té pozorování není cenzorované (hodnota $I_i=0$ tomu, že i-té pozorování je cenzorované.) Položme $W_i = \min(X_i,T_i)$. Výsledkem experimentu je nyní

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

tj. dvourozměrný (úplný) náhodný výběr.

4. Metody odhadu

Metody odhadu jsou v zásadě parametrické a neparametrické, přičemž parametrické metody z výše zmíněných důvodů převládají. Největší význam z parametrických metod má metoda maximální věrohodnosti. Uveďme věrohodnostní funkce pro výběrové plány z odstavce 3. Pro cenzorování typu I je věrohodnostní funkce

$$f(X_{(1)},\dots,X_{(r)},r) = \frac{n!}{(n-r)!}\prod_{i=1}^{r} f(X_{(i)})R^{n-r}(T),$$
$$0 \leq X_{(1)} \leq \dots \leq X_{(r)} \leq T,$$
$$r = 0,1,\dots,n.$$

(Pro jednoduchost zápisu zde i v dalším nevyznačujeme závislost na parametru θ). Pro cenzorování typu II je věrohodnostní funkce

$$f(X_{(1)},\dots,X_{(r)}) = \frac{n!}{(n-r)!}\prod_{i=1}^{r} f(X_{(i)})R^{n-r}(X_{(r)}), \qquad 0 \leq X_{(1)} \leq \dots \leq X_{(r)} .$$

Předpokládejme nyní, že časový cenzor T má rozdělení s distribuční funkcí $G(t;\theta)$ a hustotou $g(t;\theta)$ (θ je obecně vektorový parametr!), že X a T jsou nezávislé. Nalezněme nyní rozdělení vektoru (W_i,I_i) :

$$P(W_i<w,\ I_i=1) = P(X_i<w,\ X_i<T_i) = \iint_{x<w,\,x<t} dF(x)\,dG(t) = F(w) - \int_0^w f(x)G(x)\,dx,$$

analogicky

$$P(W_i< w,\ I_i=0) = G(w) - \int_0^w F(x)g(x)\,dx,$$

z čehož

$$\frac{\partial}{\partial w} P(W_i< w,\ I_i = 1) = f(w)(1-G(w))$$

$$\frac{\partial}{\partial w} P(W_i< w,\ I_i = 0) = g(w)(1-F(w)).$$

Definujme funkci h takto:

$$h(u,v) = \begin{cases} 0 & u \leq 0 \\ g(u)(1-F(u)) & u>0,\ v=0 \\ f(u)(1-G(u)) & u>0,\ v=1. \end{cases}$$

Distribuční funkce náhodného vektoru (W_i,I_i) je potom

$$H(w,y) = \int_0^w \int_0^y h(u,v)\,d\mu(u),\ d\lambda(v),$$

kde λ je čítací míra přiřazující bodům 0 a 1 míru 1 a μ je Lebesguerova míra na přímce. Funkce h je zřejmě hustota vzhledem k součinové míře $\mu \times \lambda$. Věrohodnostní funkce je potom

$$f((W_1,I_1),\dots,(W_n,I_n)) = \prod_{i=1}^{n} h(W_i,I_i).$$

Je to tedy věrohodnostní funkce dvourozměrného (nezávislého) náhodného výběru a je možno použít standardních metod.

Příklad 4.1. Předpokládejme, že X má rozdělení $Exp(1/\alpha)$ a T rozdělení $Exp(1/\beta)$, tj. rozdělení s hustotami $f(x) = \alpha e^{-\alpha x}$ a $g(t) = \beta e^{-\beta t}$. Bez újmy na obecnosti předpokládejme, že $W_1 = X_1,\dots,W_r = X_r$, $W_{r+1} = T_{r+1},\dots,W_n = T_n$. Řešením věrohodnostních rovnic je

$$\hat{\alpha} = \frac{r}{\sum_1^r X_i + \sum_{r+1}^n T_i}, \qquad \hat{\beta} = \frac{n-r}{\sum_1^r X_i + \sum_{r+1}^n T_i}.$$

Fisherova informační matice je

$$I(\alpha,\beta) = \begin{pmatrix} \frac{1}{\alpha(\alpha+\beta)} & 0 \\ 0 & \frac{1}{\beta(\alpha+\beta)} \end{pmatrix}.$$

V tomto případě je možné ověřit, že jsou splněny podmínky regularity a tudíž je možné použít výsledků o asymptotické normalitě maximálně věrohodných odhadů.

Příklad 4.2. Předpokládejme, že X má rozdělení $W(a^{-1/\beta},\beta)$ a T rozdělení $Exp(1/b)$. Soustava věrohodnostních rovnic pak zní

$$\hat{b} = \frac{n-r}{\sum_1^r X_i + \sum_{r+1}^n T_i} \tag{4.1}$$

$$\hat{a} = \frac{r}{\sum_1^r X_i^{\beta} + \sum_{r+1}^n T_i^{\beta}} \tag{4.2}$$

$$\hat{\beta} = \frac{r}{\hat{a}\sum_1^r X_i^{\beta}\ln X_i + \hat{a}\sum_{r+1}^n T_i^{\beta}\ln T_i - \sum_1^r X_i}. \tag{4.3}$$

Soustava se řeší tak, že se za $\hat{a}$ v (4.3) dosadí ze (4.2) čímž se (4.3) stane nelineární rovnicí pro jednu neznámou $\hat{\beta}$. Numerická zkušenost ukazuje, že velmi efektivní metodou pro řešení této rovnice je metoda Newton-Raphsonova. Za počáteční řešení se doporučuje vzít $\beta_0=1.5$. Přesnosti řádu 10^{-4} se obvykle dosáhne již po několika iteracích. Získaná hodnota $\hat{\beta}$ se dosadí do (4.2). O vlastnostech takto získaných odhadů není podobně jako v obecném případě mnoho známo.

Dalšími užívanými metodami jsou metoda nejmenších čtverců a metoda kvantilových odhadů.

Na závěr uveďme ještě neparametrický Kaplan-Meierův odhad funkce spolehlivosti $R(x)$. Nechť $W_{(1)} < \dots < W_{(n)}$ je uspořádaný náhodný výběr z $W_1,\dots,W_n$. Pak

$$\tilde{R}(x) = \prod_{j=1}^{k}\left[\frac{n-j}{n-j+1}\right]^{I_j}, \qquad W_{(k-1)} \le x < W_{(k)}$$
$$= 0, \qquad x \ge W_{(n)}.$$

Ukazuje se, že $\tilde{R}(x)$ je silně konzistentní a asymptoticky normální, a že $\tilde{R}(x)$ je maximálně věrohodný odhad ve třídě všech distribucí. Výhodou tohoto odhadu je, že je neparametrický, poměrně jednoduše se počítá a umožňuje výpočet některých charakteristik spolehlivosti (např. střední hodnoty doby do poruchy). Nevýhodou je nemožnost provádět statistické závěry pro hodnoty větší než $W_{(n)}$ a jistá ztráta eficience ve srovnání s odhady parametrickými.

5. Literature

[1] Breslow,N. - Crowley,J.: A large sample study of the life table and product limit estimates under random censorship. Annals of Statistics 2 (1974), 437-453.

[2] Gill,R.D. : Testing with replacement and the product limit estimator. Annals of Statistics 9 (1974), 853-860.

[3] Roubík,F. : Spolehlivost systémů. Diplomová práce MFF UK Praha, 1980.