Parametrické metody odhadů z neúplných výběrů 2

1 Metoda maximální věrohodnosti pro cenzorované výběry

1.1 Náhodné cenzorování

Při sledování složitých reálných systémů často nemáme možnost uspořádat experiment ideálně a nelze přitom využít cenzorování časem nebo poruchou, ale je potřeba vyjít z tzv. provozních dat. Podobně je tomu při sledování životnosti jedinců dané populace. Protože jedinci často migrují, mohou se dostat mimo naši kontrolu dříve, než zjistíme, zda ke sledovanému rizikovému jevu došlo či nikoliv. Např. sledujeme-li soubor pacientů, kteří jsou po závažném onemocnění v septickém stavu a sledování provádíme s ohledem na jejich přežití, nemusíme přesně znát dobu trvání septického stavu. Pacient může být během septického stavu převezen do jiného zařízení, kde již není pod naší kontrolou, případně dlouhodobé sledování pacienta pro potřeby dané studie musíme z časových důvodů předčasně ukončit apod. V takových případech uvažujeme n statistických jednotek a u každé z nich pozorujeme buď náhodnou veličinu X, která udává dobu čekání na rizikový jev nebo náhodnou veličinu T, která udává dobu sledování pacienta. Náhodná veličina T se nazývá časový cenzor. Lze tedy každé statistické jednotce přiřadit hodnotu X nebo T podle toho, která z těchto hodnot je menší. Formálně zapsáno může být výsledek sledování n dvojic $(W_1, I_1), (W_2, I_2), \ldots, (W_n, I_n)$, kde

$W_j = \min(X_j, T_j)$,

$I_j = 1$, jestliže $W_j = X_j$, a tedy, j-té pozorování X je necenzorované a rizikový jev u j-té jednotky byl pozorován v čase X_j,

$I_j = 0$, jestliže $W_j = T_j$, to znamená, že j-té pozorování X je cenzorováno v čase $T = T_j$, tedy statistická jednotka j byla vyjmuta ze sledování dříve, než došlo k nastoupení rizikového jevu, čas vyjmutí ze sledování byl T_j, $T_j < X_j$ a T_j je náhodná veličina.

Při náhodném cenzorování předpokládáme, že doba čekání na rizikový jev X a časový cenzor T jsou nezávislé náhodné veličiny. Rozdělení X popíšeme stejně jako dříve hustotou $f(x)$ nebo distribuční funkcí $F(x)$ a rozdělení časového cenzoru T popíšeme hustotou $g(t)$ nebo distribuční funkcí $G(t)$, přičemž obě tato rozdělení mohou záviset na neznámých parametrech, tedy $F(x) = F(x, \boldsymbol{\theta}_1)$ a $G(x) = G(x, \boldsymbol{\theta}_2)$. Vektor $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ je pak vektor všech neznámých parametrů a pro jednoduchost budeme předpokládat, že vektory $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ neobsahují společné parametry. Za uvedených předpokladů je potom výsledkem pozorování n nezávislých dvojic $(W_j, I_j),\ j = 1, 2, \ldots, n$. Dříve než uvedeme věrohodnostní funkci, která odpovídá náhodnému cenzorování, zavedeme funkci $H(w, i)$ pro $w > 0$ a $i \in \{0, 1\}$ vztahem $H(w, i) = P(W \leq w, I = i)$. Z uvedených předpokladů pak plyne, že

$$H(w,1) = P(W_j < w, I_j = 1) = F(w) - \int_0^w f(x)G(x)dx$$
$$\text{a}H(w,0) = P(W_j \leq w, I_j = 0) = G(w) - \int_0^w F(x)g(x)dx.$$

Odtud derivováním dostaneme funkci

$$\begin{aligned} h(w,1) &= \frac{dH(w,1)}{dw} = f(w) - f(w)G(w) = f(w)\big(1 - G(w)\big),\ w > 0 \\ h(w,0) &= \frac{dH(w,0)}{dw} = g(w) - F(w)g(w) = g(w)\big(1 - F(w)\big),\ w > 0. \end{aligned} \tag{1}$$

Zřejmě funkce $h(w,i)$ odpovídá sdružené hustotě náhodných veličin W a I. Odtud plyne, že věrohodnostní funkce výsledku experimentu $(W_1,I_1),\dots,(W_n,I_n)$ při náhodném cenzorování je rovna

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{j=1}^{n} h(W_j, I_j). \tag{2}$$

Poznamenejme ještě pro úplnost, že pro stručnost zápisu jsme v hustotách f, g, h a distribučních funkcích F, G, H explicitně nevyjádřili závislosti na parametrech $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}$. Odhad parametru $\boldsymbol{\theta}$ potom získáme maximalizací věrohodnostní funkce (2) nebo jednodušeji maximalizací logaritmické věrohodnostní funkce

$$\ell(\boldsymbol{\theta}) = ln\mathcal{L}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln h(W_j, I_j).$$

Protože na pořadí sčítanců ve vyjádření logaritmické věrohodnostní funkce $\ell(\boldsymbol{\theta})$ nezáleží, lze $\ell(\boldsymbol{\theta})$ zapsat ve tvaru

$$\ell(\boldsymbol{\theta}) = \sum_{j\in J_1} \ln h(W_j, 1) + \sum_{j\in J_0} \ln h(W_j, 0), \tag{3}$$

kde $J_1 \subseteq \{1,2,\dots,n\}$ je množina statistických jednotek j, pro které je $I_j = 1$, tj. $J_1 = \{j\colon I_j = 1\}$ a $J_0 \subseteq \{1,2,\dots,n\}$ je množina statistických jednotek $\{j\colon I_j = 0\}$. Tedy J_1 odpovídá statistickým jednotkám, u nichž byl rizikový jev pozorován a J_0 odpovídá statistickým jednotkám, které byly cenzorovány a rizikový jev nebyl pozorován. Když dosadíme za $h(W_j,0)$ a $h(W_j,1)$ ve vzorci (3) ze vzorce (1), dostaneme logaritmickou věrohodnostní funkci $\ell(\boldsymbol{\theta})$ ve tvaru

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{j\in J_1} \ln\Big(f(W_j)\big(1-G(W_j)\big)\Big) + \sum_{j\in J_0} \ln\Big(g(W_j)\big(1-F(W_j)\big)\Big) = \\
&= \sum_{j\in J_1} \ln\Big(f(X_{(j)})\big(1-G(X_{(j)})\big)\Big) + \sum_{j\in J_0} \ln\Big(g(T_j)\big(1-F(T_j)\big)\Big) = \\
&= \sum_{j\in J_1} \ln f(X_{(j)}) + \sum_{j\in J_0} \ln\big(1-F(T_{(j)})\big) + \sum_{j\in J_0} \ln g(T_j) + \sum_{j\in J_1} \ln\big(1-G(X_{(j)})\big).
\end{aligned}$$

Dále vzhledem k předpokladu, že f a F závisí pouze na parametru $\boldsymbol{\theta}_1 = (\theta_{11},\dots,\theta_{1r_1})$ a g a G pouze na parametru $\boldsymbol{\theta}_2 = (\theta_{21},\dots,\theta_{2r_2})$, dostaneme logaritmicko věrohodnostní funkci $\ell(\boldsymbol{\theta})$ jako součet

$$\ell(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2),$$

kde

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}_1) &= \sum_{j\in J_1} \ln f(X_{(j)}) + \sum_{j\in J_0} \ln\big(1-F(T_j)\big), \\
\ell_2(\boldsymbol{\theta}_2) &= \sum_{j\in J_0} \ln g(T_j) + \sum_{j\in J_1} \ln\big(1-G(X_{(j)})\big).
\end{aligned}$$

Věrohodnostní rovnice pro odhad parametru $\boldsymbol{\theta}_1$ jsou

$$\frac{\partial \ell_1}{\partial \theta_{1i}} = 0,\ i = 1,\dots,r_1$$

a věrohodnostní rovnice pro odhad θ_2 jsou

$$\frac{\partial \ell_2}{\partial \theta_{2i}} = 0,\ i = 1,\dots,r_2.$$

Jejich řešením dostaneme maximálně věrohodné odhady parametrů $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ při náhodném cenzorování.

1.2 Testovací statistiky v cenzorovaných výběrech

Budeme vycházet z cenzorovaného náhodného výběru, který byl pořízen z pozorování n statistických jednotek. Označíme J_0 množinu cenzorovaných jednotek a množinu J_1 množinu necenzorovaných statistických jednotek. Ve speciálním případě, kdy pozorování nejsou cenzorována, je množina J_0 prázdná a jedná se o speciální případ modelu s cenzorováním. Hustotu doby čekání na rizikový jev opět označíme $f(x, \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1 \ldots, \theta_r)$ je r-rozměrný parametr. V aplikacích je častá situace, kdy se zajímáme pouze o některé složky vektoru $\boldsymbol{\theta}$, budeme předpokládat, že jsou to parametry $\theta_1, \ldots, \theta_k,\ k \leq r$, a budeme je nazývat **cílové parametry**. Ostatní parametry $\theta_{k+1}, \ldots, \theta_r$ nejsou předmětem našeho zájmu, netýká se jich testovaná hypotéza a nazývají se **rušivé parametry**. Například když pracujeme s normálním rozdělením $N(\mu, \sigma^2)$, je $r = 2,\ \boldsymbol{\theta} = (\mu, \sigma)$ a testovaná hypotéza se často týká jenom parametru $\theta_1 = \mu$ $(k = 1)$ a druhý parametr $\theta_2 = \sigma$ je rušivým parametrem.

Dále označíme $\boldsymbol{\theta}_C = (\theta_1, \ldots, \theta_k)$ vektor cílových parametrů a $\boldsymbol{\theta}_R = (\theta_{k+1}, \ldots, \theta_r)$ vektor rušivých parametrů (rušivý vektorový parametr). Budeme uvažovat nulovou hypotézu $H_0 \colon \boldsymbol{\theta}_C = \boldsymbol{\theta}_0$, kde $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0k})$ je daný známý vektor. V tomto označení lze psát $\boldsymbol{\theta} = (\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ a při platnosti nulové hypotézy je $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_R)$. Logaritmickou věrohodnostní funkci $\ell(\boldsymbol{\theta})$ pak můžeme po dosazení za $\boldsymbol{\theta}$ psát ve tvaru $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ a za platnosti H_0 je $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_0, \boldsymbol{\theta}_R)$.

(LR) Věrohodnostní poměr (z anglického likelihood ratio)

$$LR = 2\left(l(\widehat{\boldsymbol{\theta}}_C, \widehat{\boldsymbol{\theta}}_R) - \ell(\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\theta}}_R)\right),$$

kde $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_C, \widehat{\boldsymbol{\theta}}_R)$ je maximálně věrohodný odhad parametru $\boldsymbol{\theta} = (\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ a $\widetilde{\boldsymbol{\theta}}_R$ je maximálně věrohodný odhad parametru $\boldsymbol{\theta}_R$ za platnosti nulové hypotézy $\boldsymbol{\theta}_C = \boldsymbol{\theta}_0$. Tedy platí pro něj, že $\ell(\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\theta}}_R) = \max\limits_{\boldsymbol{\theta}_R} \ell(\boldsymbol{\theta}_0, \boldsymbol{\theta}_R)$ a lze jej získat řešením věrohodnostních rovnic

$$\frac{\partial \ell(\boldsymbol{\theta}_0, \boldsymbol{\theta}_R)}{\partial \theta_j} = 0,\ \ j = k+1, \ldots, r.$$

(W) Waldova statistika

$$W = (\widehat{\boldsymbol{\theta}}_C - \boldsymbol{\theta}_0)' \mathbf{J}_{11\cdot 2}(\widehat{\boldsymbol{\theta}}_C, \widehat{\boldsymbol{\theta}}_R)(\widehat{\boldsymbol{\theta}}_C - \boldsymbol{\theta}_0),$$

kde matice $\mathbf{J}_{11\cdot 2}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ závislá na parametru $\boldsymbol{\theta} = (\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ se vypočte podle vzorce $\mathbf{J}_{11\cdot 2}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$, přičemž matice $\mathbf{J}_{11},\ \mathbf{J}_{12},\ \mathbf{J}_{21}$ jsou bloky tzv. **Fisherovy informační matice** $\mathbf{J} = \mathbf{J}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ definované pomocí logaritmické věrohodnostní funkce $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ vztahem

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) = \left(-E\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right)_{\substack{i=1,\ldots,r \\ j=1,\ldots,r}} = \begin{pmatrix} \mathbf{J}_{11}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)\ \mathbf{J}_{12}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) \\ \mathbf{J}_{21}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)\ \mathbf{J}_{22}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) \end{pmatrix}.$$

V uvedeném vztahu značí E střední hodnotu (logaritmická věrohodnostní funkce $\ell(\boldsymbol{\theta})$ závisí na náhodném výběru) a bloky $\mathbf{J}_{11},\ \mathbf{J}_{12},\ \mathbf{J}_{21}$ a $\mathbf{J}_{22}$ matice $\mathbf{J}$ jsou postupně typu $k \times k,\ k \times (r-k),\ (r-k) \times k$ a $(r-k) \times (r-k)$.

(LM) **Skórová funkce** (odpovídající test je založený na Lagrangeových multiplikátorech, nazývá se též Raův test)

$$LM = \left[\mathbf{U}_C(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}_R)\right]' \left[\mathbf{J}_{11\cdot 2}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}_R)\right]^{-1} \mathbf{U}_C(\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}_R),$$

kde $\mathbf{U}_C(\boldsymbol{\theta}) = \mathbf{U}_C(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R)$ je prvních k složek tzv. **skórového vektoru**

$$\mathbf{U} = \mathbf{U}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_r} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_C(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) \\ \mathbf{U}_R(\boldsymbol{\theta}_C, \boldsymbol{\theta}_R) \end{pmatrix}.$$

Tedy vektor $\mathbf{U}_C(\boldsymbol{\theta}_C, \widehat{\boldsymbol{\theta}}_R)$ je k-rozměrný a vektor $\mathbf{U}_R(\boldsymbol{\theta}_C, \widehat{\boldsymbol{\theta}}_R)$ je $(r-k)$-rozměrný.

Když jsou splněny podmínky regularity, mají všechny tři uvedené statistiky za platnosti nulové hypotézy H_0 asymptoticky Pearsonovo χ^2 rozdělení o k stupních volnosti. H_0 pak zamítáme na hladině významnosti, když daná statistika překročí $(1-\alpha)$-kvantil Pearsonovo rozdělení $\chi^2(k)$.

Příklady k procvičení

1. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z exponenciální rozdělení $Ex(\lambda)$. Stanovte testovací statistiky:

   a) věrohodnostní poměr LR,

   b) Waldovu statistiku W,

   c) skórovou statistiku LM.

   Úlohu můžete řešit numericky pro vybraná simulovaná data s použitím vhodného software R, nebo MATLAB.

2. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z exponenciální rozdělení $Ex(\lambda)$. Za předpokladu, že tento náhodný výběr je cenzorovaný časem s časovým cenzorem T stanovte testovací statistiky:

   a) věrohodnostní poměr LR,

   b) Waldovu statistiku W,

   c) skórovou statistiku LM.

   Úlohu můžete řešit numericky pro vybraná simulovaná data s použitím vhodného software R, nebo MATLAB.

3. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z exponenciální rozdělení $Ex(\lambda)$. Za předpokladu, že tento náhodný výběr je cenzorovaný poruchou počtem cenzorovaných jednotek $m < n$ stanovte testovací statistiky:

   a) věrohodnostní poměr LR,

   b) Waldovu statistiku W,

   c) skórovou statistiku LM.

   Úlohu můžete řešit numericky pro vybraná simulovaná data s použitím vhodného software R, nebo MATLAB.

4. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z exponenciální rozdělení $Ex(\lambda)$. Za předpokladu, že tento náhodný výběr je náhodně cenzorovaný a časový cenzor T má exponenciální rozdělení $Ex(\delta)$, stanovte testovací statistiky:

   a) věrohodnostní poměr LR,

   b) Waldovu statistiku W,

   c) skórovou statistiku LM.

Úlohu můžete řešit numericky pro vybraná simulovaná data s použitím vhodného software R, nebo MATLAB.

5. Řešte úlohy z příkladů 1 – 4, za předpokladu, že výběry jsou cenzorované zleva.

   Úlohu můžete řešit numericky pro vybraná simulovaná data s použitím vhodného software R, nebo MATLAB.