evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERZITA OBRANY
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Parametrické metody odhadů z neúplných výběrů 1

Dříve popsané metody pro hledání bodových odhadů vycházely z předpokladu, že je dán náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení o distribuční funkci $F(x, \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ je neznámý parametr, který je potřeba odhadnout. Označíme-li X zkoumanou náhodnou veličinu, která má distribuční funkci $F(x, \boldsymbol{\theta})$, pak náhodný výběr $X_1, X_2, \ldots, X_n$ lze považovat za n nezávislých kopií zkoumané náhodné veličiny X na n statistických jednotkách. Když X značí dobu čekání na rizikový jev (např. poruchu nějaké součásti, dobu života jedince dané populace a pod.), je často obtížné získat úplný náhodný výběr $X_1, X_2, \ldots, X_n$.

Například při zkouškách životnosti složitých systémů, v pojišťovnictví, při konstrukci tabulek úmrtnosti a pod. je třeba analyzovat dobu čekání na rizikový jev ještě předtím, než dojde k jeho realizaci u všech n sledovaných prvků. V teorii spolehlivosti mohou být některé prvky vyjmuty ze sledování ještě před ukončením experimentu. V klinických pokusech dochází k úmrtí z příčin, které nejsou předmětem zkoumání nebo se pacienti prostě odstěhují a opět není možné získat úplný výběr $X_1, X_2, \ldots, X_n$. Mnohdy je ekonomicky neúnosné provádět experiment až do okamžiku porouchání všech sledovaných prvků. Často je to nemožné i z věcného hlediska, např. sledovaná doba čekání na rizikový jev překračuje dobu po níž má experimentátor reálnou možnost tento rizikový jev sledovat. V situacích, kdy sledujeme dobu čekání na rizikový jev u n statistických jednotek po nějakou dobu a během sledované doby nedojde k rizikovému jevu u všech n sledovaných jednotek, mluvíme o **neúplných náhodných výběrech** nebo o **cenzorovaných náhodných výběrech**.

1 Metoda maximální věrohodnosti pro cenzorované výběry

Předpokládejme, že sledujeme n statistických jednotek a rozdělíme je do dvou skupin. Do první skupiny dáme všechny statistické jednotky, během jejichž sledování došlo k pozorování rizikového jevu (tedy u každé jednotky v této skupině známe dobu do poruchy). Tuto množinu statistických jednotek označíme J_1, zřejmě $J_1 \subset \{1, 2, \ldots, n\}$. Do druhé skupiny dáme všechny statistické jednotky, u nichž během sledované doby nedošlo k pozorování rizikového jevu (daná jednotka se během sledované doby neporouchala). Tuto množinu označíme J_0. Zřejmě $J_0 \cup J_1$ je množina všech sledovaných statistických jednotek $\{1, 2, \ldots, n\}$.

Statistické jednotky z množiny J_1 nazýváme necenzorované a označíme X_i dobu, kdy došlo k pozorování rizikového jevu u jednotky i, $i \in J_1$. Oproti tomu statistické jednotky z množiny J_0 nazýváme cenzorované a označíme t_i pro $i \in J_0$ dobu, po kterou byla statistická jednotka i sledována a k poruše nedošlo (říkáme, že t_i je doba cenzorování jednotky i, $i \in J_0$). Výsledkem statistického šetření jsou potom doby pozorování rizikového jevu X_i, $i \in J_1$, a doby cenzorování t_i, $i \in J_0$. Předpokládáme, že náhodné veličiny X_i, $i \in J_1$ jsou nezávislé, každá má hustotu $f(x, \boldsymbol{\theta})$. Pokud jde o cenzorované statistické jednotky, máme k dispozici pouze informaci, že k poruše i-té cenzorované jednotky došlo až po čase t_i, $i \in J_0$, a pravděpodobnost tohoto jevu pro i-tou jednotku, $i \in J_0$, je $P(X_i > t_i) = 1 - F(t_j, \boldsymbol{\theta}) = S(t_j, \boldsymbol{\theta})$, kde S je funkce přežití závislá na parametru $\boldsymbol{\theta}$. Můžeme tedy zapsat věrohodnostní funkci popisující výsledek pozorování ve tvaru

$$L(\boldsymbol{\theta}) = \prod_{i \in J_1} f(x_i, \boldsymbol{\theta}) \prod_{i \in J_0} S(t_i, \boldsymbol{\theta}).$$

Jejím logaritmováním dostaneme logaritmickou věrohodnostní funkci $l(\boldsymbol{\theta})$ ve tvaru

$$l(\boldsymbol{\theta}) = \sum_{i \in J_1} \ln f(x_i, \boldsymbol{\theta}) + \sum_{i \in J_0} \ln S(t_i, \boldsymbol{\theta}). \tag{1}$$

Operační program Vzdělávání pro konkurenceschopnost
Název projektu: Inovace magisterského studijního programu Fakulty ekonomiky a managementu
Registrační číslo projektu: CZ.1.07/2.2.00/28.0326

PROJEKT JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY.

Když nyní využijeme vztahu $S(t,\boldsymbol{\theta}) = e^{-\int_0^t s(x,\boldsymbol{\theta})dx}$ mezi funkcí přežití S a rizikovou funkcí $s(x,\boldsymbol{\theta})$ a vztahu $f(x,\boldsymbol{\theta}) = s(x,\boldsymbol{\theta})S(x,\boldsymbol{\theta})$, můžeme logaritmickou věrohodnostní funkci přepsat do tvaru

$$l(\boldsymbol{\theta}) = \sum_{i\in J_1} \ln s(x_i,\boldsymbol{\theta}) - \sum_{i\in J_0}\int_0^{x_i} s(x,\boldsymbol{\theta})dx - \sum_{i\in J_0}\int_0^{t_i} s(x,\boldsymbol{\theta})dx$$

a odtud, když položíme $w_i = x_i,\ i \in J_1$ a $w_i = t_i,\ i \in J_0$, dostaneme logaritmickou věrohodnostní funkci ve tvaru

$$l(\boldsymbol{\theta}) = \sum_{i\in J_1} \ln s(x_i,\boldsymbol{\theta}) - \sum_{i=1}^{n}\int_0^{w_i} s(x,\boldsymbol{\theta})dx.$$

Maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ lze získat maximalizací logaritmické věrohodnostní funkce $l(\boldsymbol{\theta})$. Tedy $\widehat{\boldsymbol{\theta}}$ je maximálně věrohodný odhad $\boldsymbol{\theta}$, když platí

$$l(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}).$$

Maximálně věrohodný odhad obvykle hledáme řešením věrohodnostních rovnic

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} = 0,\ \ j = 1,2,\ldots,r.$$

V některých situacích se vychází z uspořádaného náhodného výběru $X_{(1)},\ldots,X_{(n)}$ místo náhodného výběru $X_1,\ldots,X_n$, pro něž byly v tomto odstavci odvozeny věrohodnostní rovnice. Princip metody je stejný, jako v případě neuspořádaného výběru, jen místo hustoty $f(x)$ se vyjde z hustoty uspořádaného náhodného výběru.

Budeme zabývat třemi základními modelovými situacemi:

1. cenzorování časem (cenzorování typu I),
2. cenzorování poruchou (cenzorování typu II),
3. náhodné cenzorování.

$X_{(1)}, X_{(2)},\ldots,X_{(n)}$ bude značit uspořádaný náhodný výběr $X_1, X_2,\ldots,X_n$. Tedy $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ a $x_{(1)}, x_{(2)},\ldots,x_{(n)}$ bude značit realizaci uspořádaného náhodného výběru. Dále $S(x) = S(x,\boldsymbol{\theta}) = 1 - F(x,\boldsymbol{\theta})$ bude značit funkci přežití zkoumané veličiny X a $f(x) = f(x,\boldsymbol{\theta})$ bude značit hustotu příslušnou k distribuční funkci $F(x,\boldsymbol{\theta})$.

1.1 Cenzorování časem

S cenzorováním časem se setkáváme při experimentálním uspořádání, kdy v okamžiku $t = 0$ zahájíme sledování n statistických jednotek a sledujeme je po čas $T > 0$. Čas T je předem pevně daný a nazývá se **časový cenzor**. Experiment ukončíme v okamžiku T bez ohledu na to, u kolika statistických jednotek byl rizikový jev (porucha) pozorován. Výsledkem takto uspořádaného experimentu je náhodná veličina m udávající počet statistických jednotek, u nichž byl do času T pozorován rizikový jev (porucha) a dále doby poruch $X_{(1)}, X_{(2)},\ldots,X_{(m)}$ pozorované u m statistických jednotek do doby T. Dále pak z výsledku experimentu vyplývá, že doba poruchy $(m+1)$-ní statistické jednotky $X_{(m+1)} > T$. Zřejmě náhodná veličina m má obor hodnot $\{0,1,2,\ldots,n\}$ a doba trvání experimentu je pevná a rovná se T. Sdružené rozdělení výsledku experimentu lze popsat sdruženým rozdělením pravděpodobností uspořádaného náhodného výběru $X_{(1)}, X_{(2)},\ldots,X_{(m)}$ a náhodné veličiny m. Když budeme pravděpodobnostní funkci náhodné veličiny m chápat jako speciální případ hustoty – diskrétní

hustota (formálně se pravděpodobnostní funkce zavádí jako hustota vzhledem k tzv. čítací míře), lze zapsat sdruženou hustotu náhodných veličin $X_{(1)}, \ldots, X_{(m)}$, m při cenzorování časem ve tvaru

$$f(x_{(1)}, x_{(2)}, \ldots, x_{(m)}, m) = \frac{n!}{(n-m)!} \left(\prod_{i=1}^{m} f\left(x_{(i)}\right) \right) S^{n-m}(T) \tag{2}$$

$$\text{pro } 0 < x_{(1)} < x_{(2)} < \cdots < x_{(m)} < T, \;\; m = 0, 1, \ldots, n.$$

Příklad

Předpokládejme, že doba čekání na rizikový jev je náhodná veličina X a $X \sim Ex(\lambda)$. Stanovte maximální věrohodný odhad parametru λ v případě, že v rámci experimentu bylo sledováno n statistických jednotek a pozorování byla cenzorována časem s časovým cenzorem T. Dále stanovte maximálně věrohodný odhad parametrické funkce $\tau(\lambda) = \frac{1}{\lambda}$, která udává střední dobu čekání EX na rizikový jev.

Řešení: Protože $X \sim Ex(\lambda)$ je

$$f(x) = \lambda e^{-\lambda x} \text{ pro } x > 0,$$
$$f(x) = 0 \qquad \text{pro } x < 0.$$

Odtud $S(x) = 1 - F(x) = 1 - \int_0^x f(y)dy = e^{-\lambda x}$ pro $x > 0$. Pak dosazením do (2) dostaneme pro věrohodnostní funkci výsledku experimentu při cenzorování časem vyjádření

$$\begin{aligned}
&\mathcal{L}(\lambda; x_{(1)}, \ldots, x_{(m)}, m) = \\
&\quad = f(x_{(1)}, \ldots, x_{(m)}, m; \lambda) = \frac{n!}{(n-m)!} \left(\prod_{i=1}^{m} \lambda e^{-\lambda x_{(i)}} \right) \left(e^{-\lambda T} \right)^{n-m} = \\
&\quad = \frac{n!}{(n-m)!} \lambda^m e^{-\lambda \sum_{i=1}^{m} x_{(i)}} e^{-\lambda T(n-m)}
\end{aligned}$$

pro $m \in \{0, 1, \ldots, n\}$ a $0 < x_{(1)} < \cdots < x_{(m)} < T$. Odtud vypočteme logaritmickou věrohodnostní funkci

$$\ell(\lambda) = \ln \mathcal{L}(\lambda) = \ln \left(\frac{n!}{(n-m)!} \right) + m \ln \lambda - \lambda \sum_{i=1}^{m} x_{(i)} - \lambda T(n-m).$$

Potom dostaneme věrohodnostní rovnici $\frac{\partial \ell}{\partial \lambda} = 0$ ve tvaru

$$\frac{m}{\lambda} - \sum_{i=1}^{m} x_{(i)} - T(n-m) = 0.$$

Odtud

$$\frac{1}{\lambda} = \frac{1}{m} \left(\sum_{i=1}^{m} x_{(i)} + T(n-m) \right)$$

$$\widehat{\lambda} = \left[\frac{1}{m} \left(\sum_{i=1}^{m} X_{(i)} + T(n-m) \right) \right]^{-1}$$

$$\widehat{\tau}(\lambda) = \tau(\widehat{\lambda}) = \frac{1}{\widehat{\lambda}} = \frac{1}{m} \sum_{i=1}^{m} X_{(i)} + \frac{n-m}{m} T$$

je maximálně věrohodným odhadem parametrické funkce $\tau(\lambda) = \frac{1}{\lambda}$. Tedy $\widehat{\tau}$ je maximálně věrohodný odhad střední doby čekání na rizikový jev.

1.2 Cenzorování poruchou

Při cenzorování poruchou sledujeme v čase $t = 0$ celkem n experimentálních jednotek a pozorování ukončíme přesně v okamžiku, kdy sledovaný rizikový jev byl pozorován právě u m experimentálních jednotek, přičemž m je nyní pevné dané předem zvolené přirozené číslo, $m = \{1, 2, \ldots, n\}$. Výsledkem experimentu je potom m hodnot $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(m)}$, které představují hodnoty prvních m dob čekání na rizikový jev u n statistických jednotek. V tomto případě je doba trvání experimentu $X_{(m)}$.

Při cenzorování poruchou je výsledkem experimentu pozorování náhodných veličin $X_{(1)}, \ldots, X_{(m)}$ a informace, že u $n-m$ statistických jednotek byla doba čekání na rizikový jev větší než $X_{(m)}$. Proto lze sdruženou hustotu výsledku experimentu zapsat ve tvaru

$$f(x_{(1)}, \ldots, x_{(r)}) = \frac{n!}{(n-m)!} \left(\prod_{i=1}^{m} f\left(x_{(i)}\right) \right) S^{n-m}(x_{(m)}) \tag{3}$$

pro $0 < x_{(1)} < x_{(2)} < \cdots < x_{(m)} < \infty$.

Příklad
Předpokládejme, že doba čekání na rizikový jev je náhodná veličina X a $X \sim Ex(\lambda)$. Stanovte maximální věrohodný odhad parametru λ v případě, že v rámci experimentu bylo sledováno n statistických jednotek a pozorování byla cenzorována poruchou m prvků, $m \leq n$. Dále stanovte maximálně věrohodný odhad parametrické funkce $\tau(\lambda) = \frac{1}{\lambda}$.

Řešení: Protože $X \sim Ex(\lambda)$, dostaneme po dosazení za f a S do (3) vyjádření věrohodnostní funkce ve tvaru

$$\begin{aligned}
\mathcal{L}(\lambda; x_{(1)}, \ldots, x_{(m)}) &= f(x_{(1)}, \ldots, x_{(m)}; \lambda) = \\
&= \frac{n!}{(n-m)!} \left(\prod_{i=1}^{m} \lambda e^{-\lambda x_{(i)}} \right) \left(e^{-\lambda x_{(m)}} \right)^{n-m} = \\
&= \frac{n!}{(n-m)!} \lambda^m e^{-\lambda \sum_{i=1}^{m} x_{(i)}} e^{-\lambda(n-m)x_{(m)}}; \;\; \lambda > 0.
\end{aligned}$$

Odtud dostaneme logaritmickou věrohodnostní funkci ve tvaru

$$\ell(\lambda) = \ln \mathcal{L}(\lambda) = \ln \frac{n!}{(n-m)!} + m \ln \lambda - \lambda \sum_{i=1}^{m} x_{(i)} - \lambda(n-m)x_{(m)}; \;\; \lambda > 0.$$

Po dosazení do věrohodnostní rovnice $\frac{\partial \ell}{\partial \lambda} = 0$ dostaneme

$$\frac{m}{\lambda} - \sum_{i=1}^{m} x_{(i)} - (n-m)x_{(m)} = 0.$$

Řešením věrohodnostní rovnice dostaneme maximální věrohodný odhad parametru λ ve tvaru

$$\widehat{\lambda} = \left(\frac{1}{m} \sum_{i=1}^{m} X_{(i)} + \frac{n-m}{m} X_{(m)} \right)^{-1}$$

a maximálně věrohodný odhad střední doby čekání na rizikový jev (tj. odhad parametrické funkce $\tau(\lambda)$) ve tvaru

$$\widehat{\tau}(\lambda) = \tau(\widehat{\lambda}) = \frac{1}{m}\sum_{i=1}^{m} X_{(i)} + \frac{n-m}{m} X_{(m)}.$$

Příklady k procvičení

1. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z exponenciální rozdělení $Ex(\lambda)$. Metodou maximální věrohodnosti stanovte odhad parametru λ.

2. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z normálního rozdělení $N(\mu, \sigma^2)$. Metodou maximální věrohodnosti stanovte odhad parametrů μ σ.

3. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z Poissonova rozdělení $Po(\lambda)$. Metodou maximální věrohodnosti stanovte odhad parametrů λ.

4. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z logaritmicko-normálního rozdělení $LN(\mu, \sigma^2)$. Metodou maximální věrohodnosti stanovte odhad parametrů μ σ.

5. Je dán náhodný výběr $X_1, X_2, \ldots, X_n$ rozsahu n z binomického rozdělení $Bi(n, \theta)$ n je pevně dané. Metodou maximální věrohodnosti stanovte odhad parametrů θ.

6. Napište věrohodnostní rovnice pro odhady neznámých parametrů z příkladů 1 – 5 za předpokladu, že výběry jsou zprava cenzorované časem T, tedy cenzorování je typu I. Které z těchto věrohodnostních rovnic dokážete analyticky řešit?

7. Napište věrohodnostní rovnice pro odhady neznámých parametrů z příkladů 1 – 5 za předpokladu, že výběry jsou zprava cenzorované časem poruchou, tedy počtem experimentálních jednotek m tedy cenzorování je typu II. Které z těchto věrohodnostních rovnic dokážete analyticky řešit?

8. Napište věrohodnostní rovnice pro odhady neznámých parametrů z příkladů 1 – 5 za předpokladu, že výběry jsou zleva cenzorované časem T, tedy cenzorování je typu I. Které z těchto věrohodnostních rovnic dokážete analyticky řešit?