O STATISTICKÉ INFERENCI VE WEIBULLOVĚ ROZDĚLENÍ

Jan Hurt
Matematicko-fyzikální fakulta Univesity Karlovy
Sokolovská 86, 186 00 Praha 8

1. ÚVOD

Weibullovo rozdělení jistým způsobem zobecňuje exponenciální rozdělení. Weibullovo rozdělení s parametry $\theta > 0, \beta > 0$, které budeme označovat $W(\theta, \beta)$ má distribuční funkci

$$F(x) = 1 - \exp(-(x/\theta)^{\beta}), \quad x > 0 \qquad (1.1)$$
$$= 0 \text{ jinak}$$

a hustotu

$$f(x) = (\beta/\theta)(x/\theta)^{\beta-1}\exp(-(x/\theta)^{\beta}), \quad x > 0 \qquad (1.2)$$
$$= 0 \text{ jinak.}$$

Pro $\beta = 1$ je tedy Weibullovo rozdělení totožné s exponenciálním rozdělením s parametrem θ. Weibullovo rozdělení $W(\theta, \beta)$ má střední hodnotu

$$\mu = \theta\Gamma(1+1/\beta) \qquad (1.3)$$

a rozptyl

$$\sigma^2 = \theta^2(\Gamma(1+2/\beta) - \Gamma^2(1+1/\beta)), \qquad (1.4)$$

takže variační koeficient

$$V = \left(\frac{\Gamma(1+2/\beta)}{\Gamma^2(1+1/\beta)} - 1\right)^{1/2} \qquad (1.5)$$

je pouze funkcí β. Tohoto faktu se využívá pro konstrukci odhadu parametru β momentovou metodou.

Snad nejčastěji se Weibullovo rozdělení používá v teorii spolehlivosti. Odpovídající intenzita poruch

$$r(x) = \frac{f(x)}{1 - F(x)} = (\beta/\theta)(x/\theta)^{\beta-1}, \quad x > 0, \qquad (1.6)$$

je totiž v závislosti na parametru β klesající ($\beta < 1$), konstantní ($\beta = 1$) nebo rostoucí ($\beta > 1$). Přitom pro $\beta < 2$ je $r(x)$ konkávní a pro $\beta > 2$ konvexní. Ukazuje se, že díky těmto skutečnostem lze pomocí Weibullova rozdělení dobře aproximovat doby do poruchy (doby života) prvků (systémů) s monotónní intenzitou poruch. Na základě teoretických úvah je též možné dojít k Weibullově rozdělení jakožto limitnímu rozdělení (teorie extrémních hodnot).

Poznamenejme, že někdy je užitečná následující reparametrizace (1.1): položíme-li $\eta = \ln\theta$, $\xi = 1/\beta$ a $Y = \ln X$, kde X má rozdělení $W(\theta, \beta)$, má Y rozdělení s distribuční funkcí

$$G(y) = 1 - \exp(-\exp(\frac{y - \eta}{\xi})), \quad -\infty < y < \infty. \qquad (1.7)$$

Rozdělení s touto distribuční funkcí je asymptotické rozdělení extrémních hodnot typu I. Parametry η a ξ zde mají význam parametrů polohy a měřítka, tj., distribuční funkci $G(y)$ lze vyjádřit ve tvaru $G(y) = H((y - \eta)/\xi)$, kde H je distribuční funkce nezávislá na parametrech, v našem případě $H(z) = 1 - \exp(-\exp(z))$. Tohoto faktu se využívá při odhadu neznámých parametrů polohy a měřítka.

2. CENZOROVÁNÍ

V aplikacích je často potřebné provádět statistické závěry na základě neúplných výběrů. Zmíníme se proto o prakticky nejdůležitějších uspořádáních experimentů poskytujících tzv. cenzorované výběry. Předpokládejme, že náhodná veličina X je doba do poruchy sledovaného prvku. V čase $t = 0$ začneme pozorovat n prvků stejného typu. Sledované prvky se časem porouchávají. V případě, že pozorování provádíme až do doby, kdy se všechny prvky porouchají, získáme úplný výběr $X_1,\ldots,X_n$ dob do poruchy. Z časových či ekonomických důvodů však může být nereálné konat pozorování až do okamžiku selhání posledního sledovaného prvku a proto se experiment ukončí podle předem daného pravidla dříve.

Je-li předepsán časový okamžik T, ve kterém je experiment ukončen (bez ohledu na to, ke kolika poruchám skutečně došlo), jsou takto získaná pozorování časově cenzorována, přičemž toto cenzorování se nazývá cenzorování typu I. Druhý možný způsob cenzorování zvaný cenzorování typu II spočívá v tom, že experiment je ukončen v okamžiku, kdy se porouchá r-tý sledovaný prvek, kde r je předem dané číslo. Data jsou tzv. cenzorována poruchou.

Zatímco při cenzorování typu I je počet skutečně pozorovaných dob do poruchy náhodná veličina a doba trvání experimentu předem dána, je při cenzorování typu II počet pozorovaných poruch předem dán a doba trvání experimentu je dána dobou do poruchy r-tého porouchaného prvku a je tudíž náhodnou veličinou.

Při sledování složitých systémů se můžeme setkat s případem, kdy se časové cenzory systém od systému liší, a kdy ukončení pozorování u každého systému je více méně náhodné. Skončí-li pozorování systémů $1,\ldots,n$ v okamžicích $T_1,\ldots,T_n$, je ekonomicky neúnosné považovat získaná pozorování jakožto cenzorovaná v okamžiku $T = \min(T_1,\ldots,T_n)$ a aplikovat metody cenzorování typu I. Místo toho je vhodné považovat časový cenzor též za náhodnou veličinu. V takovém případě mluvíme o náhodném cenzorování.

Mimo uvedené typy uspořádání experimentu se vyskytují i další, často vnějšími podmínkami vynucená uspořádání.

3. METODY ODHADU

Omezíme se pouze na prakticky nejdůležitější metody, a to metodu maximální věrohodnosti a metodu nejmenších čtverců. Předpokládejme, že $X_1,\ldots,X_n$ je náhodný výběr z rozdělení s distribuční funkcí (1.1). Logaritmus sdružené hustoty výběru (věrohodnostní funkce) je

$$(3.1)\quad n\ln\beta - n\beta\ln\theta + (\beta-1)\Sigma_1^n \ln X_i - \Sigma_1^n (X_i/\theta)^\beta \quad .$$

Po derivování podle β a θ obdržíme soustavu věrohodnostních rovnic

$$(3.2)\quad n/\beta + \Sigma \ln X_i + \ln\theta\,(\theta^{-\beta}\Sigma X_i^\beta - n) - \theta^{-\beta}\Sigma X_i^\beta \ln X_i = 0,$$

$$(3.3)\quad \theta = (n^{-1}\Sigma X_i^\beta)^{1/\beta} \quad .$$

Dosazením (3.3) do (3.2) dostaneme

$$(3.4)\qquad \frac{\Sigma X_i^{\beta} \ln X_i}{\Sigma X_i^{\beta}} - \frac{1}{\beta} - \frac{1}{n}\Sigma \ln X_i = 0.$$

Řešení věrohodnostních rovnic pak spočívá v tom, že iteračním postupem získáme řešení $\hat{\beta}$ rovnice (3.4) a to dosadíme do (3.3). Velmi účinnou metodou pro řešení rovnice (3.4) se ukazuje Newton-Raphsonova metoda. Numerické experimenty ukazují, že konvergence metody je velmi ryhlá, a že příliš nezávisí na zvoleném počátečním řešení. Je proto možné zvolit počáteční řešení rozumným odhadem. Dříve se doporučovalo užít za počáteční řešení odhad získaný momentovou metodou na základě vyjádření variačního koeficientu (1.5). Numerický proces je však náročný na přesnost - doporučuje se proto provádět výpočty ve dvojnásobné aritmetice.

Jednoduchá je adaptace věrohodnostních rovnic v případě cenzorovaných pozorování. Výsledkem experimentu při cenzorování typu I je prvních r hodnot pořádkových statistik $X_{(1)} \leq \ldots \leq X_{(r)} \leq T$ a informace $X_{(r+1)} > T$. Je-li F distribuční funkce základního souboru a f odpovídající hustota, je věrohodnostní funkce

$$(3.5)\qquad \frac{n!}{(n-r)!} \prod_1^r f(X_{(i)})(1 - F(T))^{n-r}.$$

Konkrétně po zlogaritmování (3.5) a po úpravě dostaneme pro Weibullovo rozdělení soustavu věrohodnostních rovnic

$$(3.6)\qquad \frac{\Sigma_1^r X_{(i)}^{\beta} \ln X_{(i)} + (n-r)T^{\beta} \ln T}{\Sigma_1^r X_{(i)}^{\beta} + (n-r)T^{\beta}} - \frac{1}{\beta} - \frac{1}{r}\Sigma_1^r \ln X_{(i)} = 0$$

$$(3.7)\qquad \theta = \left(r^{-1}\left(\Sigma_1^r X_{(i)}^{\beta} + (n-r)T^{\beta}\right)\right)^{1/\beta},$$

které se řeší analogicky jako rovnice pro úplný výběr. Při cenzorování typu II je výsledkem experimentu prvních r hodnot pořádkových statistik $X_{(1)} \leq \ldots \leq X_{(r)}$. Odpovídající věrohodnostní funkce je

$$(3.8)\qquad \frac{n!}{(n-r)!} \prod_1^r f(X_{(i)})(1 - F(X_{(r)}))^{n-r}.$$

Věrohodnostní rovnice pro cenzorování typu II jsou tedy zcela analogické rovnicím pro cenzorování typu I s tím, že místo T píšeme $X_{(r)}$.

V případě úplných výběrů je možno provádět statistické závěry na základě známých asymptotických vlastností maximálně věrohodných odhadů. Asymptotické rozptyly a kovariance maximálně věrohodných odhadů parametrů Weibullova rozdělení v případě cenzorování typu II odvodili HARTER a MOORE. Odvodit vlastnosti (byť i asymptotické) maximálně věrohodných odhadů v případě cenzorování typu I tradičními metodami zdá se doposud být neschůdné. Často se můžeme setkat s tvrzením, že výsledky platné pro cenzorování typu II je možné použít i pro cenzorování typu I. Z intuitivního hlediska se však toto tvrzení zdá být nesprávné proto, že ignoruje informaci $X_{(r+1)} > T$. Numerická

Časový cenzor T	1	2
Počet simulací, pro něž r=21	340	375
MC průměr $\bar{\hat{\theta}}$	0.705	1.333
Výběrový rozptyl $\hat{\theta}$	0.0061	0.0242
MC průměr $\bar{\hat{\beta}}$	0.5452	0.5044
Výběrový rozptyl $\hat{\beta}$	0.0127	0.0084

Tabulka

ilustrace ukazuje, že tomu tak skutečně je. V jedné studii metodami Monte Carlo (MC) byly studovány odhady parametrů Weibullova rozdělení s parametry $\theta = 1$, $\beta = 0.5$, rozsah základního souboru n=30 pro cenzorování typu I s časovými cenzory T=1 a T=2. Pro každý cenzor bylo provedeno N=3000 simulací náhodných výběrů, řešeny věrohodnostní rovnice a výsledky registrovány. V tabulce jsou uvedeny výsledky pro marginální rozdělení s r=21. Je patrné, že statistická indukce by v obou případech byla velice rozdílná.

Metodu nejmenších čtverců pro odhad parametrů polohy a měřítka je možné aplikovat na Weibullovo rozdělení po reparametrizaci (1.7), tj. převedením na model extrémních hodnot typu I. Je-li $X_{(1)},\ldots,X_{(n)}$ uspořádaný náhodný výběr z rozdělení s distribuční funkcí $G(x) = H((x-\eta)/\xi)$, nezávisí rozdělení

(3.9) $$Y_{(i)} = (X_{(i)} - \eta)/\xi \quad , \; i=1,\ldots,n$$

na neznámých parametrech. Tato skutečnost inspiruje k sestrojení regresního modelu

(3.10) $$X_{(i)} = \eta + \xi EY_{(i)} + u_{(i)} \; , \quad i=1,\ldots,n,$$

kde $u_{(i)}$ jsou poruchy s nulovými středními hodnotami a kovariancemi

(3.11) $$\mathrm{cov}(u_{(i)},u_{(j)}) = \xi^2 \mathrm{cov}(Y_{(i)},Y_{(j)}).$$

Kovariance na pravé straně (3.11) nezávisí na neznámých parametrech. Výhodou této formulace je, že z n rovnic (3.10) můžeme využít k odhadu jen některé. Parametry η a ξ nyní můžeme odhadnout jednak obyčejnou metodou nejmenších čtverců, tj. za předpokladu nekorelovanosti poruch, jednak zobecněnou metodou nejmenších čtverců s využitím znalosti kovariancí $\mathrm{cov}(Y_{(i)},Y_{(j)})$ pro konkrétní rozdělení. V obou případech jsou odhady ve tvaru lineárních kombinací pořádkových statistik, přičemž koeficienty těchto kombinací jsou pro nejužívanější rozdělení tabelovány. Zajímavou a pravděpodobně dosud neřešenou otázkou je robustnost odhadů obyčejnou metodou nejmenších čtverců.

Odvodit jiné než asymptotické vlastnosti odhadů získaných výše uvedenými metodami je prakticky neproveditelné. Proto i konstrukce konfidenčních intervalů a testování hypotéz se provádí na základě asymptotických vlastností.

Existují i další, obvykle méně efektivní metody odhadu (zjednodušenné lineární, kvantilové aj.).

4. LITERATURA

BURY,K.V.: Statistical models in applied science, Wiley, New York, 1975.

JOHNSON,N.L.-KOTZ,S.: Continuous univariate distributions-1, Houghton Mifflin, Boston, 1970.

MANN,N.R.-SCHAFER,R.E.-SINGPURWALLA,N.D.: Methods for statistical analysis of reliability and life data, Wiley, New York, 1974.