TESTY SHODY A HOMOGENITY PRO CENZOROVANÁ DATA

Petr Volf, ÚTIA ČSAV, Praha

Při sledování spolehlivosti či délky života se často setkáváme s cenzorovanými daty. V příspěvku uvádím několik metod na testování shody cenzorovaných dat s hypotetickým rozložením pravděpodobnosti a na zkoumání homogenity dvou souborů dat. Hledáme analogie statistik Kolmogorova či Cramera-von Misese, které jsou založeny na rozdílech mezi odhadnutou a předpokládanou distribuční funkcí.

Náhodné veličiny $X_1, X_2, \ldots, X_N$ nechť představují doby bezporuchového provozu nějakých z našeho hlediska stejných objektů. Tyto veličiny jsou nezávislé, nezáporné, stejně rozdělené s distribuční funkcí F, s funkcí spolehlivosti P=1-F. Některé z pokusů jsou však předčasně ukončeny působením jiných náhodných veličin $Y_1, Y_2, \ldots, Y_N$, také nezáporných, nezávislých vzájemně i na X_i, stejně rozdělených s distribuční funkcí G, Q=1-G. Nechť obě distr. funkce jsou spojité.

Ve skutečnosti tedy pozorujeme jen

$$Z_i = \min(X_i, Y_i) \qquad i=1,2,\ldots,N$$

a víme, zda došlo k sledované poruše či jsme se jí nedočkali, což označíme náhodným indikátorem δ_i =1 je-li $X_i \leq Y_i$
=0 jinak.

Toto je struktura náhodně (zprava) cenzorovaných dat. Přečíslujme měření tak, aby náš výběr byl uspořádaný co do hodnot Z: $Z_1 \leq Z_2 \leq \ldots \leq Z_N$. Ve zpracování cenzorovaných dat hraje pořadí důležitou roli. Uznávaným odhadem funkce spolehlivosti je Kaplanův-Meierův "Product Limit Estimate"

(1) $$P_N(t) = \prod_{i:\, Z_i \leq t} \left(\frac{N-i}{N-i+1}\right)^{\delta_i}$$

Podobným způsobem zkonstruovaný odhad pro kumulativní intenzitu poruch KIP definovanou jako $L(t) = -\ln P(t)$ je

$$L_N(t) = \sum_{i:\, Z_i \leq t} \frac{\delta_i}{N-i+1}$$

Jsou to maximálně věrohodné odhady(či alespoň jejich limity) a mají proto odpovídající vlastnosti. Vezměme T takové, že je ještě $P(T)\cdot Q(T) > 0$. Potom platí:

VĚTA (Breslow,Crowley 1974): Pro spojité P a Q náhodný proces $V_N(t) = \sqrt{N}\left(\frac{P_N(t)}{P(t)} - 1\right)$ na $\langle 0,T\rangle$ konverguje v distribuci ke gaussovskému procesu V(t) s nulovou střední hodnotou a kovarianční funkcí

(2) $$\mathrm{cov}\,(V(s),V(t)) = C(s), \quad 0 < s \leq t < T, \text{ kde } C(s) = -\int_0^s \frac{dP}{P^2 Q}.$$

Není těžké ukázat, že:

I/ $C_N(t) = \sum_{i:\, Z_i \leq t} \frac{N\delta_i}{(N-i)(N-i+1)}$ je na $\langle 0,T\rangle$ silně konzistentním odhadem C(t), plyne to ze silné konzistence odhadu P_N (např. Hall,Wellner 1980).

II/ $W_N(t) = \sqrt{N}\,(L_N(t)-L(t))$ konverguje v distribuci na $\langle 0,T\rangle$ k témuž procesu V(t) (Nair 1981, Volf 1983).

Jinými slovy,

(3) $$V_N(t) \ \left(\text{resp. } W_N(t)\right) \xrightarrow{F} \beta(C(t)) \quad \text{při } N \to \infty,$$

kde $\beta(\tau)$ je standardní Wienerův proces, proces Brownova pohybu. A to je základ, na němž se budují statistiky pro testy shody. Nebo, definujeme-li proces nazývaný Brownův most (Brownian bridge, tied Brownian process)

$$B(u) = \beta\left(\frac{u}{1-u}\right)\cdot(1-u) \qquad \text{pro } u \in \langle 0,1\rangle$$

a chceme-li vyjádřit (3) pomocí tohoto procesu, označíme $K(t)=\frac{C(t)}{1+C(t)}$ a máme

$$(4) \qquad V_N(t)/(1+C_N(t)) \sim B(K(t)) \qquad \text{pro } t \in \langle 0,T\rangle .$$

Mimochodem, můžeme se přesvědčit, že v případě bez cenzorování (t.j. Q≡1) se redukuje $P_N(t)$ na 1- E.D.Fce, C(t) na (1-P(t))/P(t). V našem případě je ovšem vliv cenzorující veličiny nezanedbatelný, proto musíme i rozptyl nahradit odhadem a odvolávat se na asymptotické vlastnosti.

2. Konfidenční oblasti a testy.

Všimneme si nyní některých vlastností rozložení trajektorií procesu Brownova pohybu (Robbins, Siegmund 1970). Pro $c,d \geq 0$

$$(5) \quad P\{\exists t\in\langle 0,T\rangle : \beta(t) \geq c+d\cdot t\} = \tilde\Phi\left(\frac{d\cdot T+c}{\sqrt{T}}\right) + e^{-2cd}\cdot\tilde\Phi\left(\frac{d\cdot T-c}{\sqrt{T}}\right),$$

kde $\tilde\Phi$ označuje 1 - distribuční funkce standardního normálního rozložení. Označíme tuto pravděpodobnost $P^+(c,d,T)$, je to totéž jako pro "Brownův most"

$$P\left\{\exists \tau\in\left(0,\frac{T}{1+T}\right\rangle : B(\tau) \geq c+(d-c)\cdot\tau\right\}$$

a dík symetrii Brownova procesu také

$$P\{\exists t\in\langle 0,T\rangle : \beta(t) \leq -c-d\cdot t\} .$$

$$(6) \quad P\{\exists t\in\langle 0,T\rangle : |\beta(t)| \geq c+dt\} = 2\tilde\Phi\left(\frac{dT+c}{\sqrt{T}}\right) - 2\sum_{l=1}^{\infty}(-1)^l e^{-2cdl^2}\cdot$$
$$\cdot\left\{\tilde\Phi\left[\frac{-dT+c(2l-1)}{\sqrt{T}}\right] - \tilde\Phi\left[\frac{dT+c(2l+1)}{\sqrt{T}}\right]\right\}.$$

Tuto pravděpodobnost budeme značit $P(c,d,T)$.

Dík (3) můžeme hned vytvořit asymptotické konfidenční oblasti pro funkci spolehlivosti(Gillespie, Fisher 1979). Jen nahradíme ještě jednou P(t) odhadem a samozřejmě musíme nahradit odhadem i kovarianční funkci C.

$$P\left\{0 \leq P(t) \leq P_N(t)\left[1+\frac{c+d\cdot C_N(t)}{\sqrt{N}}\right],\ \forall t\in\langle 0,T\rangle\right\} \sim 1-P^+(c,d,C(T))$$

Toto je jednostranná oblast spolehlivosti, pomocí níž bychom měli zachytit případy, kdy data odpovídají spíš menší funkci spolehlivosti než hypotetické P(t). Pak totiž očekáváme, že P(t) nebude v této oblasti. Opačný případ pro zachycení alternativy, že "skutečná" funkce spolehlivosti je větší než P, můžeme zkoumat pomocí oblasti $\left\{P_N(t)\left[1-\frac{c+dC_N(t)}{\sqrt{N}}\right] \leq P(t) \leq 1\right\}$.

Pravděpodobnost, že P(t) je uvnitř na celém intervalu $\langle 0,T\rangle$, je opět v limitě ($N\to\infty$) $1-P^+(c,d,C(T))$, pokud P(t) je "správná" funkce spolehlivosti.

Můžeme také vytvořit konfidenční oblast omezenou z obou stran, a pak je

$$P\left\{P_N(t)\left[1-\frac{c+dC_N(t)}{\sqrt{N}}\right] \leq P(t) \leq P_N(t)\left[1+\frac{c+dC_N(t)}{\sqrt{N}}\right]\right\} \sim 1-P(c,d,C(T)).$$

Stejně vytváříme konfidenční oblasti pro kumulativní intenzitu poruch. Jednostrannou

$$\Lambda(t) \geq \Lambda_N(t) - \frac{c+d\cdot C_N(t)}{\sqrt{N}}$$

pro zachycení alternativy, že soubor je horší než předpokládáme, pro opačnou možnost

$$\Lambda(t) \leq \Lambda_N(t) + \frac{c+d\cdot C_N(t)}{\sqrt{N}} ,$$

i oboustrannou.

Abychom zajistili (asymptoticky) požadovanou pravděpodobnost $1-\alpha$ (což znamená hladinu testu α), musíme najít odpovídající c,d jako řešení rovnice

$P(c,d,C_N(T)) = \alpha$,resp. $P^+(c,d,C_N(T)) = \alpha$ pro jednostranné oblasti. Proto je jednodušší omezit se na případ c = d (Hall, Wellner 1980), iteračním postupem je snadné najít dostatečně přesné přibližné řešení. I v případě běžných testů Kolmogorova se užívá takovýchto kritických oblastí s c=d, navíc s T: $P(T)=0$. Možnosti, které dávají $c \neq d$, případně různé horní hranice T, zkoumáme v další části práce. Při praktickém testování shody dat s modelem- funkcí spolehlivosti P, budeme také používat tuto variantu (c=d) a budeme postupovat následujícím způsobem. Zvolíme hladinu významnosti α , vezmeme T podle toho, na kterém intervalu $\langle 0,T\rangle$ nás shoda zajímá, vždy však menší než největší naměřená necenzorovaná hodnota. Odhadneme $P_N(t)$, $C_N(t)$, případně $L_N(t)$, chceme-li použít variantu pro KIP. Zjistíme

$$a = \max_{i=\overline{1,N}} \frac{|V_N(Z_i)|}{1+C_N(Z_i)} \quad , Z_i \le T$$

a spočteme odpovídající pravděpodobnost $P(a,a,C_N(T))$.

Je-li menší než α ,shodu zamítáme.

Pro test proti alternativě $P_1 > P$ zjistíme $a_1 = \max_{i=\overline{1,N}} \frac{V_N(Z_i)}{1+C_N(Z_i)}$, $Z_i \le T$

a shodu zamítneme, je-li $P^+(a_1,a_1,C_N(T)) < \alpha$.

Při testu našeho hypotetického modelu (P) proti alternativě $P_2 < P$ shodu zamítáme ve prospěch alternativy, je li $P^+(-a_2,-a_2,C_N(T)) < \alpha$,kde

$$a_2 = \min_{\substack{i=1,N \\ Z_i \le T}} \frac{V_N(Z_i)}{1+C_N(Z_i)} \quad .$$

Testujeme-li kumulativní intenzitu poruch, na místě V_N užijeme statistiku W_N .

Protože součet v (6) má členy se střídavým znaménkem, jejichž absolutní hodnota rychle klesá, stačí užít jen několika prvních členů k dosažení velké přesnosti při výpočtu $P(c,d,C)$. Při běžných $\alpha < 0{,}2$ je $c,d \sim 1$, stačí dva členy součtu k přesnosti na exp(-18).

Testy s pomocí KIP by mohly být silnější (a příklady tomu nasvědčují) v případě, že $P_N < P$, neboť pak

$$|L_N - L| \sim |\ln P_N - \ln P| = |\ln \frac{P_N}{P}| > |\frac{P_N}{P} - 1|,$$

v opačném případě je vhodnější použít variantu s funkcí spolehlivosti.

Kromě kriterií supremálního typu se užívají i kriteria součtová, jako je Cramerovo-von Misesovo. Základem je opět proces "Brownův most" a rozložení statistiky $\omega_\tau^2 = \int_0^\tau B^2(u)\,du$ .Tabulku můžeme najít např. v Pettitt,Stephens(1976). Pro nás

$$\frac{V_N(t)}{1+C_N(t)} \sim B(K(t)) \qquad \text{,kde } K(t) = \frac{C(t)}{1+C(t)},$$

takže $$\sum_{Z_i \le T} \delta_i \left[\frac{V_N(Z_i)}{1+C_N(Z_i)}\right]^2 \cdot \left[K_N(Z_i) - K_N(Z_{i-1})\right] \sim \omega^2_{K(T)} \quad (Z_0 = 0).$$

K hodnocení testu potřebujeme proto v praxi znát kvantily uříznuté Cramerovy von Misesovy statistiky s horní mezí $K_N(T)$.

3. Diskuse, testy homogenity.

Jak jsme viděli, konfidenční oblasti jsou založeny na přímkovém omezení pro trajektorie procesu Brownova pohybu. Přitom máme výběr, jakou přímku (s jakým sklonem) použijeme. Po převodu na proces "Brownův most" to vypadá tak, že tento proces, který je na $\langle 0,1\rangle$ symetrický co do rozložení kolem $\frac{1}{2}$,roven 0 v krajních bodech, s maximálním rozptylem v $\frac{1}{2}$, se snažíme omezovat přímkou. Zdálo by se, že omezení jinou křivkou by monlo být těsnější a test pak silnější. Blíže se tímto problémem zabývali Borovkov, Syčeva (1970), optimální se v tomto smyslu zdá $\sqrt{u(1-u)}$ uvnitř intervalu $\langle 0,1\rangle$, zatímco na krajích je těsné ohraničení dáno zákonem dvojitého logaritmu. Navíc, potíž by byla ve výpočtu odpovídajících pravděpodobností. Proto zůstáváme u přímkových hranic. Sílu testu pak můžeme zvětšit tím, že bereme v potaz shodu na různých intervalech v $\langle 0,T\rangle$ a volíme různé parametry přímky c,d.

Pokud jde o rozsah výběru, Hall,Wellner (1980) docházejí k závěru, že největší nepřesnost vzniká nikoli aplikací limitního rozložení, ale odhadem $C_N(t)$, který rychle roste pro vysoká pořadí pozorování. To znamená, že zbytečná šířka konfidenční oblasti, zpočátku srovnatelná s šířkou asymptotické kolmogorovské oblasti, se u větších pozorování zvětšuje.Samozřejmě, zvětšuje se s rostoucím počtem cenzorovaných hodnot.

Stejně jako v případě bez cenzorování není problém přejít od testů shody k testům homogenity dvou výběrů, například udělat analogii testu Kolmogorova-Smirnova.

Představme si, že máme dva cenzorované náhodné výběry rozsahů M a N, jako realizace vzájemně nezávislých náhodných veličin. Předpokládejme, že v obou skupinách mají cenzorované veličiny tutéž funkci spolehlivosti P. Udělejme z prvního výběru odhady $P_M(t)$, $C_M(t)$, $L_M(t)$, z druhého $P_N(t)$,$C_N(t)$,$L_N(t)$.Pro asymptotické závěry předpokládejme ještě, že existuje kladná a konečná limita $\frac{M}{N}$, rostou-li neomezeně rozsahy výběrů. Opět nás zajímají hodnoty v nějakém intervalu $\langle 0,T\rangle$.

Označíme

$$V_{M,N}(t)=\sqrt{\frac{MN}{M+N}}\left(\frac{P_M(t)}{P_N(t)}-1\right) , \qquad W_{M,N}(t)=\sqrt{\frac{MN}{M+N}}\left(L_M(t)-L_N(t)\right),$$

$$A_{M,N}(t)=\frac{1}{M+N}\left[N\cdot C_M(t)+M\cdot C_N(t)\right],$$

$$K_{M,N}(t)=\frac{A_{M,N}(t)}{1+A_{M,N}(t)} , \qquad K(t)=\lim_{M,N\to\infty} K_{M,N}(t) .$$

Pak podle Breslow, Crowley (1974) a důsledků znovu dostaneme statistiky s asymptotickým rozložením procesu "Brownův most"

$$V_{M,N}(t)\left(1-K_{M,N}(t)\right) \sim W_{M,N}(t)\left(1-K_{M,N}(t)\right) \sim B(K(t)) .$$

S těmito statistikami již můžeme testovat homogenitu dvou výběrů, když kritické oblasti získáme z (5), (6), nejlépe opět pro c=d.

Poznámka 1. Je také vypracováno několik variant testů homogenity, které jsou zobecněním testů Wilcoxona (Lee,Desu, Gehan 1975),některé jsou již i v souboru programů BMDP.

Poznámka 2. Pro tříděná data je zpracována obdoba běžných χ^2 testů shody založených na porovnání výběrových a teoretických četností ve třídách, s logickým rozšířením na testy homogenity (Volf 1983).

Tabulka 1. Kvantily (κ) rozložení statistiky ω_τ^2, podle Pettitt,Stephens(1976)

$P\{\omega_\tau^2 < \kappa\}$	τ: 0.5	0.6	0.7	0.8	0.9
0.50	0.0536	0.0707	0.0877	0.1030	0.1147
0.90	0.1890	0.2407	0.2861	0.3205	0.3412
0.95	0.2579	0.3269	0.3861	0.4298	0.4548
0.975	0.3295	0.4167	0.4906	0.5439	0.5733
0.99	0.4271	0.5393	0.6330	0.6997	0.7352

Tabulka 2. Některé parametry přímky omezující proces B(u) na $\langle 0,\tilde{\tau}\rangle$s pravděpodobností $1-\alpha$:

a) jednostranně, t.j. řešení rovnice $P^+(c,d,\frac{\tau}{1-\tau}) = \alpha$

τ :	0.5	0.6	0.7	0.8	0.8	0.8
α =0.05 c=d :	1.132	1.182	1.208	1.221	c: 1.521 d: 0.971	0.722 2.072
α =0.1 c=d :	0.975	1.023	1.053	1.069		

b) oboustranně, jako řešení rovnice $P(c,d,\frac{\tau}{1-\tau}) = \alpha$

α =0.05 c=d :	1.273	1.320	1.347	1.356	c: 1.655 d: 1.105	0.855 2.155
α =0.10 c=d :	1.133	1.182	1.209	1.222		

Literatura:

Borovkov A.A.,Syčeva N.M.(1968),O asymptotičeski optimalnych neparametričeskich kriterijach, Teor. ver. i prim. 13,3.

Breslow N.,Crowley J.(1974),A large sample study of the life table and product limit estimates under random censorship,AS 2,p.437.

Gillespie M.J.,Fisher L.(1979),Confidence bands for the Kaplan-Meier survival curve estimate,AS 7,1.

Hall W.J.,Wellner J.A.(1980),Confidence bands for a survival curve from censored data, Biometrika 67,1.

Lee E.T.,Desu M.M.,Gehan E.A.(1975),A Monte Carlo study of the power of some two-sample tests, Biometrika 62,2.

Nair V.N.(1981),Plots and tests for goodness of fit with randomly censored data, Biometrika 68,1.

Pettitt A.N.,Stephens M.A.(1976),Modified Cramer-von Mises statistics for censored data,Biometrika 63,2.

Robbins H.,Siegmund D.(1970),Boundary crossing probabilities for the Wiener process and sample sums, AMS 41,5.

Volf P.(1983),Estimates of distribution function and tests of fit from censored and grouped data, Trans. 9-th Prague Conf. on Inf. Th..., Academia Praha.