ODHADY HUSTOTY PRAVDĚPODOBNOSTI USEKNUTÉHO ROZDĚLENÍ

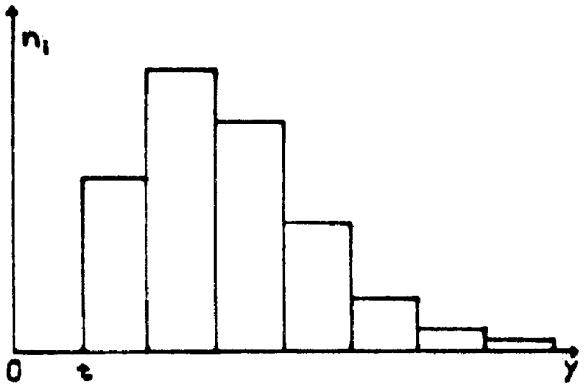
Viktor Beneš

1. Úvod

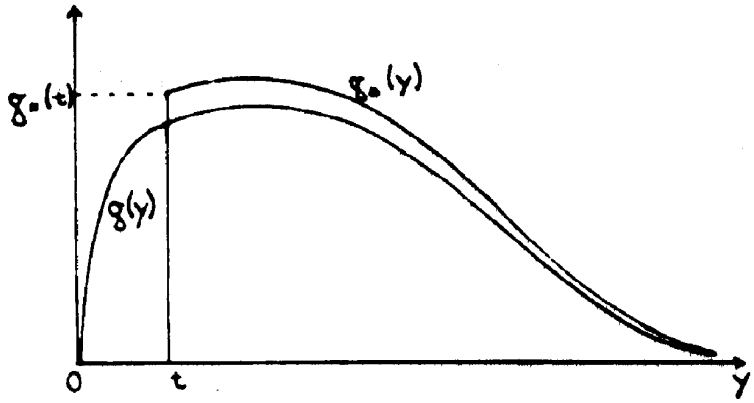
V řadě aplikací získáváme data ve formě useknutých výběrů z neznámé hustoty pravděpodobnosti. Na obr.1 je například histogram četností takového zleva useknutého výběru

$$0 < t \leq y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)} \qquad (1)$$

který vzniká třeba vlivem kladné rozlišovací schopnosti přístrojů. Pozorování menší než bod useknutí t nejsou do výběru zahrnuta a ani není znám jejich počet ne rozdíl od výběrů cenzorovaných.



Obr.1: Histogram useknutého výběru



Obr.2: Zleva useknutá hustota pravděpodobnosti $g_*(y)$

Příspěvek se zabývá metodami odhadu hustoty pravděpodobnosti g(y) neznámého rozdělení na základě výběru (1), tj. výběru ze zleva useknutého rozdělení s hustotou $g_*(y)$ /viz obr.2/, která je s g vázána vztahem

$$g_*(y) = \frac{g(y)}{1-G(t)} \quad , \text{ pro } t \leq y \qquad (2)$$

$$= 0 \quad , \text{ pro } y < t,$$

kde G je distribuční funkce příslušná k g. Zřejmě půjde vždy o metody extrapolačního typu, založené na jistých předpokladech regularity či apriorní znalosti chování sledované náhodné veličiny na intervalu $\langle 0,t)$, neboť ztracenou informaci nelze nahradit pomocí statistické inference. Další výklad je rozdělen do tří kapitol podle typu použitých metod: parametrické, neparametrické a kombinované.

2. Parametrické metody

Odhady parametrů předepsaných hustot pravděpodobnosti na základě useknutého výběru lze provádět běžnými postupy pomocí metody momentů nebo metody maximální věrohodnosti, pouze řešení příslušných rovnic je obtížnější než v neuseknutém případě. Těmito postupy se zabývala řada autorů již v padesátých a šedesátých letech, zde uvádíme jediný příklad: použití metody maximální věrohodnosti na odhady parametrů useknutého normálního rozdělení podle Cohena (1961). Pro výběr (1) buď

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_{(i)} \, , \qquad s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_{(i)}^2 - n\bar{y}^2 \right) \qquad (3)$$

Potom odhady parametrů μ a σ^2 normálního rozdělení mají tvar

$$\hat{\mu} = \bar{y} - \hat{\theta}(\bar{y} - t), \qquad \hat{\sigma}^2 = s^2 + \hat{\theta}(\bar{y} - t)^2 \tag{4}$$

kde hodnota $\hat{\theta}$ v závislosti na $\hat{\gamma} = s^2/(\bar{y}-t)^2$ je v citovaném článku tabelována, rovněž jsou tam uvedeny výrazy pro asymptotické rozptyly a kovariance odhadů.

V některých aplikacích se může stát, že bod useknutí t je neznámý. V práci Tate(1959) byl odvozen nestranný odhad bodu useknutí ve tvaru

$$\hat{t} = y_{(1)} - [1 - G(y_{(1)})][ng(y_{(1)})]^{-1} \tag{5}$$

závislý na hustotě pravděpodobnosti g, která je ve většině případů neznámá. Proto se v praxi nejčastěji používá přirozený odhad $\hat{t}=y_{(1)}$, který je pouze asymptoticky nestranný.

3. Neparametrické metody

Postupy v tomto odstavci jsou založeny na neparametrických odhadech hustoty $g_n(y)$ /viz obr.2/ na intervalu $\langle t,\infty)$ a extrapolaci parametrickou funkcí $g_e(y)$ na intervalu $\langle 0,t)$, přičemž v bodě t se požaduje spojitost, případně i spojitost derivací. Analyticky lze celou extrapolační proceduru vyjádřit v následujících krocích:

a/ globální neparametrický odhad hustoty pravděpodobnosti $g_n(y)$, $y\in(t,\infty)$ např. pomocí algoritmu Silvermana (1982),

b/ lokální neparametrický odhad $g_n^{(i)}(t)$, i=0,1,...,k-1 je stupeň derivace,

c/ odhad parametrů zvolené extrapolační funkce $g_e(y,\psi_1,\psi_2,\ldots,\psi_k)$, $y\in\langle 0,t)$, řešením systému rovnic $g_e^{(i)}(t,\hat{\psi}_1,\ldots,\hat{\psi}_k)=\hat{g}_n^{(i)}(t)$, i=0,1,...,k-1 /na levé resp. pravé straně rovnic jde o derivace zleva resp. zprava/

d/ výsledné odhady hustoty g mají tvar

$$\begin{aligned}\hat{g}(y) &= \theta\, g_e(y,\hat{\psi}_1,\ldots,\hat{\psi}_k), \quad y\in\langle 0,t)\\ &= \theta\, g_n(y), \quad y\geq t\end{aligned} \tag{6}$$

$$\theta = \Big[1 + \int_0^t g_e(y,\hat{\psi}_1,\ldots,\hat{\psi}_k)dy\Big]^{-1}.$$

Dále se budeme zabývat krokem b/ uvedené procedury, jehož provedení je nejméně střejné. V praxi často vystačíme s volbou k=1, kdy např. pro nezáporná data s předpokládanou hodnotou g(0)=C se volí g_e jako přímka procházející počátkem. Pro i=0 v b/ lze použít dva v literatuře uvedené postupy:

První z nich budeme nazývat W-odhad, byl odvozen v práci Swanepoel,Wyk(1981).

$$\hat{g}_n(t) = (r_n-1)[(n+1)(y_{(r_n)}-y_{(1)})]^{-1} \tag{7}$$

Platí, že pro $r_n/n\downarrow 0$, $r_n^{-1}\log n=o(1)$ je $\hat{g}_n(t)\to g_n(t)$ skoro jistě při $n\to\infty$. Dále pro $r_n=o(n^{2/5})$ je $\sqrt{r_n}(g_n(t)/\hat{g}_n(t)-1)\xrightarrow{D} N(0,1)$ /konvergence v distribuci k normálnímu rozdělení/ Při praktickém použití odhadu (7) se volí obyčejně r_n rovno celé části čísla $n^{2/3}$. Ukazuje se, že u jednovrcholových hustot vychýlení odhadu klesá s bodem useknutí blížícím se zleva k modu.

Druhý odhad je jádrový, odvozený Falkem(1984). Nechť $0<h_n\to 0$ pro $n\to\infty$ a jádro K splňuje $\int_{-\infty}^{\infty}K(x)dx=1$. Tvar jádrového odhadu je standardní:

$$\hat{g}_n(t) = \frac{1}{nh_n}\sum_{i=1}^{n} K\Big(\frac{y_{(i)}-y_{(1)}}{h_n}\Big) \tag{8}$$

V citované práci je dokázána asymptotická normalita odhadu (8) pro volbu jádra $K(x)=e^{-x}I_{(0,\infty)}(x)$, při splnění podmínky $nh_n^3\to 0$, ve tvaru

$$\sqrt{nh_n}\,[\hat{g}_s(t)-g_s(t)] \xrightarrow{D} N(0, f(t)/2) .$$

Před diskusí možností praktického použití odhadu (8) uvádíme výsledky simulační studie z práce Beneše (1987), ve které jsou posouzeny statistické vlastnosti obou odhadů (7), (8) ještě s dalšími metodami včetně parametrických. Simulovaná data v konkrétní aplikaci vycházejí z dané hustoty pravděpodobnosti fyzikálně odůvodnitelné a umožňují porovnání s teoretickými výsledky.

Příklad 1: Uvažme trojrozměrný neprůhledný materiálový vzorek, ve kterém jsou v matrici umístěny částice kulového tvaru s náhodným průměrem, jehož první dva obecné momenty α_1, α_2 chceme odhadnout. Předpokládá se, že náhodný bodový proces středů částic je stacionární a nezávislý na průměrech částic. V rovině řezu vzorku pozorujeme kruhové řezy částic, jejichž průměry mají hustotu pravděpodobnosti $g(y)$ a momenty β_1. Předpokládejme, že precipitace částic se řídí Ardell-Lifschitz-Slyozovým zákonem s hustotou pravděpodobnosti průměrů $f(x)=24\mu^3 x(2\mu-x)^{-5}\exp[-3x(2\mu-x)^{-1}]$, $0\le x\le 2\mu$, odsud se pomocí rovnice $g(y)=y\alpha_1^{-1}\int_y^\infty f(x)(x^2-y^2)^{-1/2}dx$ stanoví $g(y)$. Předpokládejme vliv kladné rozlišovací schopnosti při měření, tj. řezy s průměrem menším než t nelze pozorovat. Pro pevné $\mu=1$ byly z $g(y)$ opakovaně simulovány useknuté výběry tvaru (1) s cílem porovnat následující metody odhadu α_1, α_2:

1/ Při volbě lineární extrapolace $g_s(y,\psi)=\psi y$ lze z (6) a z výše uvedené integrální rovnice odvodit odhady $\hat{\alpha}_1=\pi c(tg_s(t)/2+1)/2$, $\hat{\alpha}_2=2c(t^2 g_s(t)/3+\beta_1^*)$, kde platí $c=(g_s(t)+\beta_{-1}^*)^{-1}$. Hodnota $g_s(t)$ byla odhadnuta pomocí (8) s volbou h_n aproximativně optimálního vzhledem ke střední kvadratické chybě: $h_n=\sqrt[3]{g(t)[g^{(4)}(t)]^2 n]^{-1}}$

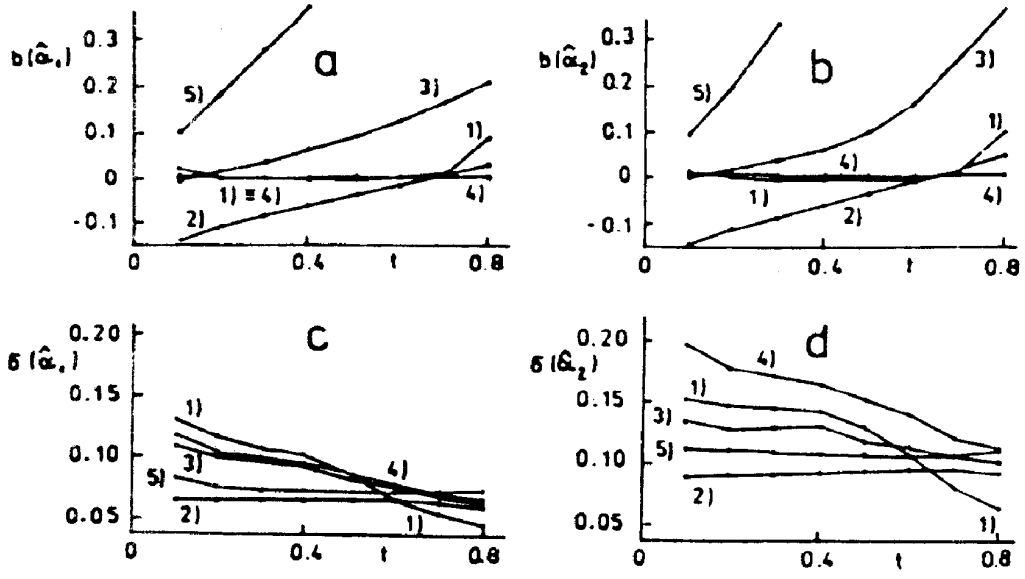
2/ Stejně jako v 1/, pouze s užitím (7) místo (8).

3/ Odhad tvaru $\hat{\alpha}_r=\hat{\alpha}_r^t/\hat{\alpha}_0^t$, $r=1,2$, kde useknuté momenty $\alpha_r^t=\int_t^\infty x^r f(x)dx$ jsou odhadnuty podle Jakemana a Andersena (1975).

4/ $\hat{\alpha}_1=8\hat{\mu}/9$, $\hat{\alpha}_2=8\hat{\mu}^2/9$, kde $\hat{\mu}$ je odhad parametru μ získaný momentovou metodou srovnáním teoretických a odhadnutých hodnot podílu α_1^t/α_0^t.

5/ Zanedbání useknutí, řešení pomocí odhadu $\hat{\beta}_r=n^{-1}\sum y_{(i)}^r$.

Na základě $m=500$ pseudonáhodných výběrů rozsahu $n=100$ na každé z úrovní $t=0.1k$, $k=1,2,\ldots,8$, tj. mezi nulou a modem hustoty g, jsou na obr.3 znázorněny grafy vychýlení $b(\hat{\alpha}_r)$ a směrodatných odchylek $\sigma(\hat{\alpha}_r)$, $r=1,2$, odhadů momentů v závislosti na t. Malé vychýlení metody 1/ a 4/ plyne ze znalosti rozdělení g, v případě jeho neznalosti nastávají potíže s volbou h_n. Vlastnosti odhadu 2/ jsou též



Obr.3: a/ vychýlení $\hat{\alpha}_1$, b/ vychýlení $\hat{\alpha}_2$, c/ směrodatná odchylka $\hat{\alpha}_1$, d/ směrodatná odchylka $\hat{\alpha}_2$.

uspokojivě a domnívá se, že neparametrické metody jsou schopny nahradit metody parametrické v dané situaci.

Simulační studie potvrdila teoretické vlastnosti odhadů (7) a (8), především fakt, že vychýlení v rostoucím bodem useknutí neroste. Pro velmi malá t lze naopak vliv useknutí zanedbat. Řád konvergence obou odhadů je $n^{-1/3}$.

Zobecnění vztahu (7) pro i-tou derivaci v bodě t tvaru

$$\hat{g}_n^{(i)}(t) = \frac{(-1)^i}{nb_n^{i+1}} \sum_{j=1}^{n} K^{(i)}\left(\frac{y_{(j)}-t}{b_n}\right) \qquad (9)$$

je asymptoticky nestranným odhadem pro $\int x^j K^{(i)}(x)dx=0$ pro $j<i$, $=(-1)^i$ pro $j=i$ a $<\infty$ pro $j>i$. Konsistence je zaručena při $\lim nb_n^{2i+1}=\infty$ a okno

$$b_n = \left\{\frac{(2i+1)g(t)\int[K^{(i)}(y)]^2dy}{2n[g_n^{(i+1)}(t)\int yK(y)dy]^2}\right\}^{\frac{1}{2i+3}} \qquad (10)$$

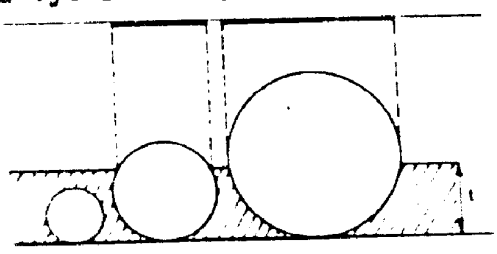
je asymptoticky optimální vzhledem ke střední kvadratické chybě (9). Pro praktické použití odhadu (9) by bylo možné k volbě okna použít analogii Sheatherovy (1986) metody, která byla odvozena pro neuseknuté rozdělení. Pro $i>0$ to ovšem představuje náročný výpočet a kvality odhadu nejsou dobré. Proto lze závěrem doporučit užívání jednoparametrické extrapolační funkce s odhadu (7).

4. Kombinované metody

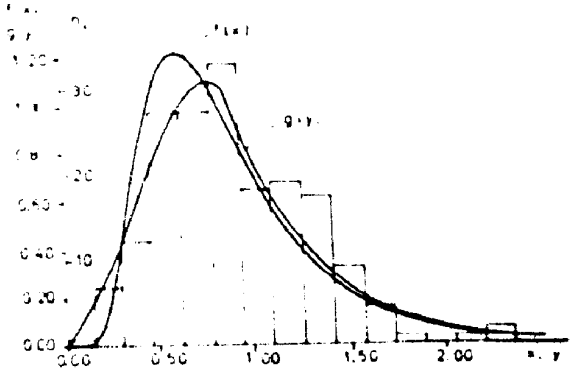
V této kapitole se omezíme na jednu ukázku z aplikace, kde jsou kombinovány parametrické i neparametrické metody odhadu. Bližší podrobnosti lze najít v práci Horálek, Beneš (1988). Při impaktních metodách výzkumu sušení rozprašováním se částice ze spreje zachycují na podložce s tenkou vrstvou kapaliny tloušťky t. V případě ponoření kapiček s neprůhlednosti kapaliny vzniká model z obr.4, kdy na fotografii podložky pozorujeme kruhové objekty, drobné kapičky nejsou pozorovány. Transformace mezi průměry částic resp.kruhů s hustotami pravděpodobnosti f,g je

$$g(y) = y[2t[1-F(t)]]^{-1}f[(y^2+4t^2)/4t], 0<y\leq 2t, \quad = f(y)/[1-F(t)], \ 2t<y. \qquad (11)$$

Předpokládejme, že rozdělení f je logaritmicko-normální, cílem je na základě naměřených průměrů kruhů $\{y_i\}$, i=1,...,n, odhadnout jeho parametry μ, σ^2. Je-li t známé, je úloha po transformaci $x_i=(y_i^2+4t^2)/4t$, $0<y_i\leq 2t$, $x_i=y_i$, $y_i>2t$, řešitelná Cohenovou metodou z kapitoly 2. Pro t neznámé bylo užito kombinace momentové metody a K-odhadu (7), kterým byl vyloučen parametr σ^2. Odhad zbylých neznámých t, μ byl získán numerickým řešením soustavy momentových rovnic. Na obr.5 je výsledek zpracování výběru rozsahu n=180 /histogram/, odhad g a výsledný tvar f.



Obr.4: Model pro impaktní experiment



Obr.5: Numerický příklad

Literatura:
Beneš V(1987)Acta Stereol.6/III,185-190
Cohen AC(1961)Technometrics 3,535-541
Falk M(1984)S.Afr.Statist.J.18,91-96
Horálek V,Beneš V(1988)Metrika 35,63-76
Jakeman AJ,Anderssen RS(1975)J.Microscopy 105,121-133
Sheather SJ(1986)Comp.Statist&Data Anal.4,61-65
Silverman J(1982)Appl.Statist. AS176
Swanepoel JWH,Wyk JWJ(1981)S.Afr.Statist.J. 15,167-172
Tate RF(1959)Ann.Math.Stat.30,341-366