SROVNÁNÍ STATISTICKÝCH A GNOSTICKÝCH ODHADŮ PARAMETRU POLOHY NA REÁLNÝCH DATECH

P.Kovanic a J.Novovičová, ÚTIA ČSAV, Praha

1.Úvod

Stigler (6) ve své studii porovnával účinnost 11 odhadů parametru polohy na několika souborech reálných fyzikálních dat z 18. a 19. století, pro která nynější "správná" hodnota je známá. Jedenáct odhadů parametru polohy zahrnovalo průměr, medián, výběr M- a L- odhadů a adaptivních odhadů.

Datové soubory byly vybrány z měření sloužících k určení vzdálenosti země od slunce (James Short 1761), z experimentů k určení rychlosti světla (Michelsen a Newcomb 1874, 1882) a z měření sloužících k stanovení průměrné hustoty země (Cavendish 1798). Závěry této studie jsou založeny na 24 datových souborech (o rozsazích 17-100 měření). Část souborů (8) je závislá na ostatních.

Za účelem ilustrace, jak pracují gnostické odhady v porovnání s některými statistickými, jsme rozšířili výše zmíněnou studii Stiglera o gnostický odhad,který bude spolu s ostatními použitými statistickými odhady popsán v odstavci 2.

2. Odhady použité pro srovnání

Předpokládejme,že je dán výběr $\{x_1,x_2,\dots,x_n\}$ a že data jsou uspořádána tak,že $x_1 \leq x_2 \leq \dots \leq x_n$.Do této studie bylo zahrnuto následujících 12 odhadů:

2.1. a 2.2. Průměr $\bar{x}$ a medián $\tilde{x}$.

2.3. až 2.5. 10%, 15% a 25% useknuté průměry (trimmed means). 100α% useknutý průměr je definován vztahem

$$\bar{x}_\alpha = \frac{1}{n(1-2\alpha)}\left[\sum_{i=g+1}^{n-g} x_i + p(x_g + x_{n-g+1})\right]$$

kde $g=[\alpha n]+1$ a $p=g-\alpha n$.

2.6. až 2.8. M- odhady (Huber P15,Andrews ANT a Tukey Biweight). Odhady tohoto typu jsou definovány jako řešení rovnice

$$\sum_{i=1}^{n} \psi\left(\frac{x_i-T}{\hat{\sigma}}\right) = 0$$

kde $\hat{\sigma}$ je odhad měřítka (zde je to medián absolutních reziduí kolem mediánu nebo kolem dřívější hodnoty odhadu, je-li tato rovnice řešena iteračním postupem) a ψ je funkce,kterou je třeba znát.

Huber P15 ((1),str.13) je jednokrokový M- odhad,kde

$$\psi(u) = u \qquad |u| \leq k$$
$$= k.\mathrm{sign}\,u \qquad |u| > k,$$

k=1.5 a $\hat{\sigma}$ =medián $|x_i-M|$, kde M je medián.

Andrews ANT ((1),str 15) je M- odhad,kde

$$\psi(u) = \sin(u/2.1) \qquad |u| < 2.1\pi$$
$$= 0 \qquad \text{jinde}$$

a $\hat{\sigma}$ je mediám absolutních odchylek od předchozí hodnoty T.

Tukey Biweight je M- odhad,kde

$$\psi(u) = u.w(u),$$
$$w(u) = (1-u^2)^2 \qquad |u| \leq 1$$
$$= 0 \qquad |u| > 1.$$

Tento odhad je počítán iteračním postupem:

$$T_{i+1} = \frac{\sum_j w((x_j-T_i)/c\,\hat{\sigma}_i).x_j}{\sum_j w((x_j-T_i)/c\,\hat{\sigma}_i)}$$

kde c=6.0, $\hat{\sigma}_i$ = medián $|x_j-T_i|$.

Poslední dva odhady doporučili k zařazení do Stiglerovy studie D.F.Andrews a J.W.Tukey.Huberův odhad byl zařazen na základě své dobré kvality zjištěné ve studii (1).

2.9. Edgeworthův odhad je vážený průměr dolního kvartilu,mediánu a horního kvartilu s vahami v poměru 5:6:5.

2.10. Outmean je vpodstatě průměr měření vyřazených při výpočtu $x_{0.25}$. Je určen vztahem

$$\bar{x}^{c}_{0.25} = 2\bar{x} - \bar{x}_{0.25}.$$

2.11. Hoggův odhad T_1 je adaptivní odhad,který navrhl R.V.Hogg a doporučil ho zařadit do Stiglerovy studie. Je určen vztahem

$$\begin{aligned} T_1 &= \bar{x}^{c}_{0.25} && \text{jestliže} && Q < 2.0 \\ &= \bar{x} && \text{jestliže} && 2.0 \leq Q \leq 2.6 \\ &= \bar{x}_{3/16} && \text{jestliže} && 2.6 \leq Q \leq 3.2 \\ &= \bar{x}_{3/8} && \text{jestliže} && 3.2 < Q, \end{aligned}$$

kde Q je míra "váhy na koncích" výběru daná vztahem

$$Q = \frac{U(0.05)-L(0.05)}{U(0.5)-L(0.5)}$$

kde $L(\alpha)$ je průměr horních 100α % výběru x_i a $U(\alpha)$ je průměr horních 100α% výběru x_i.

2.12. Gnostický odhad se počítá ve čtyřech krocích:

1.krok - exponencializace dat:

$$z_i = \exp((2x_i - x_1 - x_n)/(x_n - x_1))$$

2.krok - odhad parametru měřítka s v gnostickém modelu dat

$$z_i = z_0 \exp(s\,\Omega_i) \tag{2.1}$$

kde z_0 je parametr polohy datového souboru $z_1,\ldots,z_n$, s je parametr měřítka, Ω_i je charakteristika vlivu neurčitosti na z_i.Odhad se najde řešením úlohy

$$\tilde{s} = \arg\min_{s} \max_{j} \max\left[(!p_{cj} - (j-1)/n!),(!p_{cj} - j/n!)\right] \tag{2.2}$$

kde $j=1,\ldots,n$

$$p_{cj} = (1 + h_{cj})/2, \quad h_{cj} = w_j^{-1}\sum_{i}^{n} h_{ij}$$

$$h_{ij} = (q_{ij}^{-2} - q_{ij}^{2})/(q_{ij}^{2} + q_{ij}^{-2})$$

$$w_j = \left[\left(\sum_{i}^{n} f_{ij}\right)^2 + \left(\sum_{i}^{n} h_{ij}\right)^2\right]^{1/2}, f_{ij} = 2/(q_{ij}^{2} + q_{ij}^{-2})$$

$$q_{ij} = (z_i/z_j)^{1/2} \tag{2.3}$$

3.krok - odhad parametru polohy datového souboru $z_1,\ldots z_n$:

Interpretujme z_j v (2.3) jako nezávisle proměnnou. Pak řešme úlohu

$$\tilde{z}_0 = \arg\max_{z_j}\left(\frac{dp_{cj}}{dz_j}\right)$$

4.krok - spětná transformace výsledku

Odhad $\tilde{x}_G$ parametru polohy souboru $x_1,\ldots,x_n$ se získá transformací inverzní k (2.1):

$$\tilde{x}_G = (\log(\tilde{z}_0)(x_n - x_1) + x_n + x_1)/2$$

Gnostické algoritmy odhadování parametru polohy a měřítka nejsou obecně známé,proto zaslouhují podrobnější komentář:

Ke kroku 1: Zpracovávaná data byla upravována aditivními operacemi. Všechny ověřované estimátory jsou ekvivariantní vůči posunutí. Exponencializací "aditivně rušených" dat získáme "multiplikativně rušená" data z .Gnostický odhad parametru polohy je invariantní k násobení dat konstantou, proto bude jeho logaritmus ekvivariantní k posunutí dat x. Exponencializace rovněž zajišťuje splnění prvního gnostického axiomu.
Ke kroku 2: Veličina h_{ij} je estimační irrelevance datové položky z_i vzhledem k položce z_j. Veličina h_{oj} je celková estimační irrelevance dat $z_1,\dots,z_n$ vzhledem k položce z_j. Veličina p_{oj} je pak gnostickým odhadem hodnoty distribuční funkce dat v bodě z_j. Parametr měřítka se tedy hledá z požadavku minimalizace maximální odchylky gnostické distribuční funkce dat od empirické distribuční funkce.
Ke kroku 3: Odhad parametru polohy se tedy hledá jako poloha maxima gnostické hustoty dat.Tato úloha má řešení vždy, avšak v případě odlehlých dat může mít více než jedno řešení,např.$\tilde{z}_{01},\dots,\tilde{z}_{0m}$. Pak se přijme takové řešení $\tilde{z}_0$,pro které platí

$$p_o(\tilde{z}_0) \geq p_o(\tilde{z}_{0k}) \quad \text{pro každé } k.$$

Z 24 datových souborů zkoumaných v Stiglerově studii mělo jediný gnostický parametr polohy 23 souborů. V jediném případě souboru o rozsahu 17 se od položek rozložených v intervalu 7.71 až 9.71 "separovala" svým vlastním parametrem polohy položka 5.76 a položka 9.67. Parametr polohy 15 "řádných" dat byl 8.42. Záměna položky 5.76 za libovolnou hodnotu z intervalu $(0,\infty)$ vedla ke změnám parametru polohy celého souboru pouze v intervalu (8.418, 8.501).Je zajímavé,že nejnižší hodnota dat by měla "svůj vlastní" parametr polohy a byla by proto označena za odlehlou pro jakoukoliv hodnotu menší než 6.9.

Závěrem komentáře ke gnostickému odhadu je třeba poznamenat, že všechny použité veličiny se odvozují ze dvou axiomů gnostické teorie a z modelu dat ((2)-(4)).

3. Srovnání odhadů parametru polohy.

3.1. Metodika Stiglera a jeho závěry.

Stigler zavedl ve studii (6) označení θ_j pro správnou hodnotu parametru polohy j-tého souboru dat. Podstatné je,že za tuto "správnou" hodnotu přijímá dnešní hodnotu veličiny,která měla být výsledkem měření provedených před mnoha desítkami let. Dále značí $\hat{\theta}_{ij}$ i-tý odhad veličiny θ_j, t.j.výsledek aplikace i-té metody odhadování na j-tý datový soubor. Průměrnou absolutní odchylku odhadů pro j-tý datový soubor

$$s_j = \frac{1}{k}\sum_{i=1}^{k} |\hat{\theta}_{ij} - \theta_j| \qquad (3.1)$$

(kde k je počet testovaných odhadů)

používá k zavedení relativní odchylky

$$e_{ij} = |\hat{\theta}_{ij} - \theta_j|/s_j \qquad (3.2)$$

a indexu relativní odchylky

$$RE(i) = \frac{1}{m}\sum_{j=1}^{m} e_{ij} \qquad (3.3)$$

(kde m je počet souborů dat),

který spolu se svým rozptylem

$$(SE(i))^2 = \frac{1}{m-1}\sum_{j=1}^{m} (e_{ij} - RE(i))^2 \qquad (3.4)$$

může být použit pro hodnocení jednotlivých metod odhadování. Kromě těchto ukazatelů používá Stigler i index relativního pořadí odhadu stanovený jako průměr umístění dané metody odhadování v uspořádání podle vzdálenosti od správného výsledku.

Stigler má zato,že metoda odhadování je tím lepší,čím bližší jsou jí získané odhady k dnes známým hodnotám měřených fyzikálních konstant,tj. "odhad je tím lepší,čím menší je hodnota indexu RE(i)". Test realizuje na skupině 20 datových souborů "malých", tj. o rozsahu od 17 do 29 dat a na skupině 4 "velkých" souborů, získaných z malých souborů jejich

skládáním. Je třeba poznamenat, že 4 z malých souborů jsou navzájem závislé. Přijatý princip hodnocení aplikovaný na malé i velké soubory s využitím jak indexu relativní odchylky tak i indexu relativního pořadí odhadů přivádí Stiglera k závěru, že 10% useknutý průměr spolu s průměrem dávají nejlepší výsledky a že "moderní" odhady nemají cenu ani času potřebného k jejich výpočtu,neboť daly v tomto srovnání výsledky podstatně horší ((6) str. 1064).

3.2. Srovnání dle rozptylu odhadů.

Studie S.M.Stiglera vyvolala velkou pozornost. Spolu s touto studií byla opublikována i řada diskusních příspěvků, z nichž vyplývá, že nelze zanedbat podstatné vychýlení historických dat oproti dnes známým hodnotám fyzikálních konstant. Aby zcela vyloučili vliv volby "správné" hodnoty na srovnání metod odhadování, použili autoři později opublikovaného článku (5), navazujícího na Stiglerovu studii hodnocení podle rozptylu výsledků, dosažených jednou metodou odhadování na různých datových souborech. Z takového srovnání však vyplynuly zcela opačné závěry, než ze Stiglerova srovnání. O tom se ostatně lze přesvědčit i v jeho studii (6), kde jsou uvedeny i hodnoty odmocnin z výběrových rozptylů. Použijeme-li k hodnocení tyto veličiny pro 20 malých souborů, dostaneme jako nejlepší ze statistických odhadů Hoggův,Edgeworthův a 25% useknutý průměr, zatímco 10% useknutý průměr spolu s průměrem se přesunou až na 8. a 9. místo. Důležité je však i to, že zatímco nejlepší tři robustní odhady mají střední kvadratickou chybu 7 až 8%, dosahuje chyba desetiprocentního useknutého průměru a průměru 19%. Použitím robustních metod odhadování lze tedy podstatně zvýšit kvalitu odhadů.

3.3. Další hlediska pro srovnání.

Každý z testovaných odhadů může být považován za experta, veličina e_{ij} je pak výrokem i-tého experta o parametru polohy j-tého datového souboru. Všichni experti jsou kvalifikovaní a mají za sebou četné úspěchy v řadě praktických případů. Navíc je o nich známo z teorie, že jsou nestranní. Jsou i objektivní, protože své soudy zakládají jen na datech. Každý z nich se může v dílčích případech mýlit, ale jejich souhrnný názor by měl platit jako to nejlepší, co lze z dat vytěžit, vždyť jde o výběr z nejlepších expertů oboru. Přijmeme-li toto stanovisko a vezmeme-li v úvahu, že 84.5% hodnot e_{ij} padá do intervalu (0.8, 1.2) a že hustota odhadů e_{ij} má ostré maximum v jedničce, máme dobrý důvod přijmout za "správný" výsledek odhadování vyjádřený relativní proměnnou e_{ij} jedničku. Pak je třeba považovat odhad za tím lepší, čím je příslušné e_{ij} bližší k jedničce. Kromě rozptylu a rozpětí odhadů pořízených toutéž metodou (které na "správné" hodnotě nezávisejí) můžeme proto použít pro hodnocení odhadu i průměrnou odchylku odhadu od jedničky. Užijeme-li tato kriteria, dostaneme uspořádání odhadů,které se jen nepodstatně liší od uspořádání dle rozptylů. Závěry o výhodnosti robustních metod odhadování lze tedy podpořit hodnocením provedeným podle tří kriterií,dávajících stejné výsledky.

4. Srovnání gnostických odhadů se statistickými

Typ odhadu	$\lvert RE(i)-1\rvert$	SE(i)	$\max_j e_{ij} - \min_j e_{ij}$
Gnostický	0.001	0.046	0.15
Edgeworth	0.011	0.063	0.29
Hogg T_1	0.017	0.065	0.28
Andrews ANT	0.026	0.153	0.66
25% usek.průměr	0.029	0.076	0.28
15% usek.průměr	0.032	0.106	0.44
Tukey Biweight	0.044	0.138	0.64
Průměr	0.078	0.217	1.04
Huber P15	0.084	0.218	0.86
Outmean	0.087	0.634	2.53
10% usek.průměr	0.097	0.219	0.82
Medián	0.124	0.290	0.96

Pro hodnocení praktické použitelnosti výsledků gnostické teorie dat byl výběr testovaných odhadů rozšířen o gnostický odhad, aplikovaný na stejná data. Pro hodnocení byly využity stejné veličiny (3.1)-(3.4), kde k=12. Ukázalo se, že gnostický odhad má nejmenší rozptyl ze všech testovaných odhadů nezávisle na tom, uvažují-li se pouze malé soubory, pouze velké soubory, všech 24 souborů společně nebo jen 16 malých nezávislých souborů. Výsledky dosažené pro poslední z uvažovaných případů jsou shrnuty do tabulky na předcházející straně.

Z tabulky je patrné, že gnostický odhad získává příznivé hodnocení nejen podle nejmenšího rozptylu, ale i z ostatních hledisek. Je třeba zdůraznit, že podle Stiglera je nejzávažnějším úkolem robustního odhadování "zabránit nejhoršímu". Z tohoto hlediska nabývá na důležitosti zejména srovnání podle rozpětí.

Literatura:

(1) Andrews D.T., Bickel P.J., Hampel F.R., Huber P.J., Rogers W.H. and Tukey J.W. (1972). Robust Estimates of Location: Survey and Advances. Princeton Univer. Press.

(2) Kovanic P. (1984). Gnostical Theory of Individual Data, Problems of Control and Information Theory (PCIT),Vol.13, No.4, 259-274.

(3) Kovanic P. (1984). Gnostical Theory of Small Samples of Real Data, PCIT, Vol.13, No.5, 304-319.

(4) Kovanic P. (1984). On Relation Between Information and Physics, PCIT, Vol.13,No 6, 383-399.

(5) Rocke D.M., Downs G.W., Rocke A.J. (1982). Are Robust Estimators Really Necessary? Technometrics 24, No 2, 95-101.

(6) Stigler S.M. (1977). Do Robust Estimators Work with Real Data? Annals of Stat., Vol.5, No.6, 1055-1098.