UŽITÍ METODY BOOTSTRAP V INDEXOVÉ ANALÝZE

USE OF BOOTSTRAP METHOD IN INDEX ANALYSIS

Zdeněk Karpíšek Alena Kocmanová,
Zdeněk Sadovský, Veronika Lacinová

Abstrakt: *Článek je zaměřen na možnosti využití metody bootstrap pro výpočet intervalových odhadů středních hodnot individuálních jednoduchých indexů z pozorovaných hodnot množstevních a cenových znaků. Užití metody bootstrap v dané oblasti je vhodné i v případech, kdy neznáme rozdělení pravděpodobnosti náhodných veličin modelujících uvedené statistické znaky.*

Klíčová slova: *bootstrap, bootstrapový intervalový odhad, index, intervalový index*

Abstract: *The article is oriented on possibilities of the bootstrap method usage for calculation of the interval estimations of individual single indexes means from the observed values of quantity and price. Use of the bootstrap method in this sphere is suitable also when we don't know probability distributions of random variables modeling the listed quantities.*

Keywords: *bootstrap, bootstrap interval estimation, index, interval index*

JEL Classification: C02, C43, C65

1 ÚVOD

Metoda bootstrapových intervalových odhadů je užitečná, potřebujeme-li určit intervalové odhady parametrů pozorované náhodné veličiny, resp. náhodného vektoru, nebo testovat statistické hypotézy o těchto parametrech, ale:

(1) neznáme nebo neodhadneme rozdělení pravděpodobnosti této náhodné veličiny, resp. náhodného vektoru,

(2) rozsah výběru není dostatečně velký, abychom mohli aplikovat asymptotické odhady.

Od vydání prvního článku (Efron 1979) se metoda bootstrap velmi rozvinula (Efron, Tibshirani 1993), (Davison, Hinkley 2006) a našla uplatnění v mnoha oblastech aplikací matematické statistiky. Dostala se proto také alespoň částečně do některých profesionálních statistických softwarových produktů.

2 PRINCIP METODY BOOTSTRAP

Předpokládejme, že pozorujeme nějakou náhodnou veličinu *X*. Ze statistického souboru pozorovaných hodnot $(x_1,\ldots,x_n)$ s rozsahem *n* této náhodné veličiny vytvoříme nový statistický soubor $(x_1^*,\ldots,x_n^*)$ s rozsahem *n* náhodným výběrem hodnot x_i s opakováním (s vracením) z původního souboru $(x_1,\ldots,x_n)$. Takto získaný náhodný výběr se nazývá ***bootstrapový výběr***, resp. ***bootstrapový soubor***. Bootstrapový výběr pak *B*-krát opakujeme. Počet všech různých bootstrapových výběrů je

$$\binom{n+B-1}{B}.$$

Obecný postup aplikace metody bootstrap popisuje algoritmus:

1. Získání původního statistického souboru.
2. Výpočet statistik pro původní statistický soubor.
3. Vytvoření bootstrapových výběrů.
4. Výpočet bootstrapových statistik.
5. Výpočet bootstrapových intervalových odhadů, resp. testování hypotéz.

3 BOOTSTRAPOVÉ ODHADY

Předpokládejme, že chceme odhadnout parametr θ pozorované náhodné veličiny (vektoru) X. Bootstrapový odhad parametru θ je založen na těchto krocích (Efron, Tibshirani 1993), (Davison, Hinkley 2006):

1. Z pozorovaných hodnot $(x_1,\ldots,x_n)$ náhodného výběru $(X_1,\ldots,X_n)$ vypočítáme odhad $\hat{\theta}$ parametru θ.

2. Realizujeme B náhodných bootstrapových výběrů $(x_1^*, \ldots, x_n^*)$ o rozsahu n z pozorovaných hodnot $(x_1, \ldots, x_n)$. Obvykle volíme $B >> n$.
3. Z každého bootstrapového výběru vypočítáme odhad $\hat{\theta}_{b,j}$ parametru θ, $j = 1, \ldots, B$.

Odtud získáme ***bootstrapový odhad rozptylu*** $D(\hat{\theta})$

$$\hat{D}(\hat{\theta})_b = \frac{1}{B-1} \sum_{j=1}^{B} \left(\hat{\theta}_{b,j} - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_{b,j} \right)^2$$

a ***bootstrapový odhad směrodatné odchylky*** $\sigma(\hat{\theta})$

$$\hat{\sigma}(\hat{\theta})_b = \sqrt{\hat{D}(\hat{\theta})_b} \,.$$

Odhad $\hat{\theta}$ vypočtený z původního statistického souboru $(x_1, \ldots, x_n)$ je bodovým odhadem parametru θ, ale můžeme jej dle potřeby také nahradit aritmetickým průměrem

$$\frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_{b,j} \,.$$

Pomocí bootstrapových výběrů získáme ***bootstrapové intervalové odhady se spolehlivostí*** $1-\alpha$ střední hodnoty, rozptylu a směrodatné odchylky náhodné veličiny X. Nechť $\bar{x}$ je aritmetický průměr a s^2 je rozptyl původního statistického souboru $(x_1, \ldots, x_n)$, $\bar{x}_{b,j}$ je aritmetický průměr a $s_{b,j}^2$ je rozptyl statistického souboru z j–tého bootstrapového výběru, $j = 1, \ldots, B$. Potom:

1. ***Bootstrapový intervalový odhad střední hodnoty*** $E(X)$ ***se spolehlivostí*** $1-\alpha$ je

$$\left\langle \bar{x} - t_{b,1-\alpha/2} \frac{s}{\sqrt{n-1}} ; \bar{x} - t_{b,\alpha/2} \frac{s}{\sqrt{n-1}} \right\rangle,$$

kde $t_{b,P}$ je P–kvantil statistického souboru $(t_{b,1}, \ldots t_{b,B})$ a $t_{b,j} = \frac{\bar{x}_{b,j} - \bar{x}}{s_{b,j}} \sqrt{n-1}$ pro $j = 1, \ldots, B$.

2. ***Bootstrapový intervalový odhad rozptylu*** $D(X)$ ***se spolehlivostí*** $1-\alpha$ je

$$\left\langle \frac{ns^2}{\chi^2_{b,1-\alpha/2}}; \frac{ns^2}{\chi^2_{b,\alpha/2}} \right\rangle,$$

kde $\chi^2_{b,P}$ je P–kvantil statistického souboru $(\chi^2_{b,1},\ldots,\chi^2_{b,B})$ a $\chi^2_{b,j} = \dfrac{ns^2_{b,j}}{s^2}$ pro $j = 1,\ldots,B$.

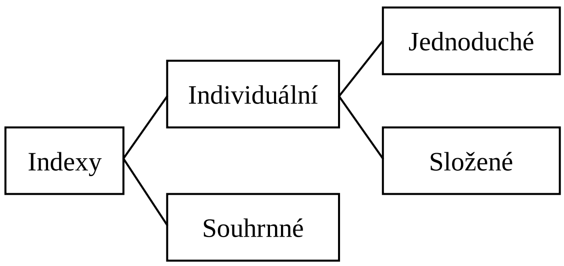
3. ***Bootstrapový intervalový odhad směrodatné odchylky*** $\sigma(X)$ ***se spolehlivostí*** $1-\alpha$ obdržíme z bootstrapového intervalového odhadu $D(X)$ pomocí odmocniny.

Existuje řada dalších způsobů stanovení intervalových odhadů uvedených i jiných parametrů, založených na bootstrapových výběrech (Efron, Tibshirani 1993), (Davison, Hinkley 2006). Mimo uvedených oboustranných intervalových odhadů se používají dle potřeby také analogicky získané jednostranné bootstrapové intervalové odhady.

4 BOOTSTRAPOVÉ INTERVALOVÉ ODHADY INDEXŮ

Indexy patří mezi poměrné kvantitativní statistické znaky a vyjadřují změnu sledovaného kvantitativního znaku nebo souboru znaků u jedné nebo více statistických jednotek během nějakého časového intervalu nebo vlivem nějakého faktoru. Indexy se obvykle konstruují ve tvaru zlomku, kde v čitateli je hodnota znaku ve srovnávaném, tzv. ***běžném období***, a ve jmenovateli hodnota tohoto znaku v tzv. základním období (Seger, Hindls 1995), (Karpíšek, Sadovský 2005). *Pro* jednotlivé znaky se používají ***individuální jednoduché indexy***, pro skupinu homogenních znaků ***individuální složené indexy*** a pro heterogenní skupinu znaků pak ***souhrnné*** (***agregátní***) ***indexy*** - viz Obr. 1.

Obr. 1: Druhy indexů



Zdroj: vlastní zpracování

Porovnávané znaky (veličiny) dělíme na ***intenzitní***, vyjadřující cenu, intenzitu apod., které značíme písmenem p a ***extenzitní***, vyjadřující množství, objem, produkci apod., které značíme písmenem q. V ekonomických aplikacích se nejčastěji užívají:

- pro intenzitní znaky ***cenové*** indexy,
- pro extenzitní znaky ***množstevní indexy***,
- pro spojení extenzitních a intenzitních znaků ***hodnotové indexy***.

Individuální jednoduché indexy jsou:

1. ***Individuální jednoduchý cenový index*** $I_p = \dfrac{p_1}{p_0}$.
2. ***Individuální jednoduchý množstevní index*** $I_q = \dfrac{q_1}{q_0}$.
3. ***Individuální jednoduchý hodnotový index*** $I_h = \dfrac{q_1 p_1}{q_0 p_0} = I_q I_p$.

Bootstrapové intervalové odhady středních hodnot, rozptylů a směrodatných odchylek cen, množství a individuálních jednoduchých indexů se spolehlivostí $1-\alpha$ určíme postupem:

1. Pozorováním dvojic cen a množství stejné komodity u n statistických jednotek ve dvou různých časových obdobích nebo regionech apod. získáme čtyřrozměrný statistický soubor
$$\begin{pmatrix} p_{01},\ldots,p_{0n} \\ q_{01},\ldots,q_{0n} \\ p_{11},\ldots,p_{1n} \\ q_{11},\ldots,q_{1n} \end{pmatrix}.$$
2. Řádky čtyřrozměrného souboru z kroku 1 jsou jednorozměrné statistické soubory cen $(p_{01},\ldots,p_{0n})$ a $(p_{11},\ldots,p_{1n})$, množství $(q_{01},\ldots,q_{0n})$ a $(q_{11},\ldots,q_{1n})$, a výpočtem individuálních jednoduchých indexů obdržíme jednorozměrné statistické soubory $(I_{p1},\ldots,I_{pn})$, $(I_{q1},\ldots,I_{qn})$ a $(I_{h1},\ldots,I_{hn})$.
3. Ze čtyřrozměrného souboru z kroku 1 provedeme B bootstrapových výběrů podle sloupců.

4. Ze všech bootstrapových výběrů pak dostaneme analogickým způsobem jako v kroku 2 jednorozměrné bootstrapové výběry cen, množství a individuálních jednoduchých indexů.
5. Bootstrapové výběry pak zpracujeme způsobem popsaným v oddílu 3.

Příklad. Statistickým šetřením u náhodně vybraných 20 řidičů z České republiky byly zjištěny měsíční nákupy a ceny benzínu Natural 95 v květnu a září 2012. Získané hodnoty jsou v levé části Tab. 1 (jde o expertně simulovaný soubor) a v pravé části této tabulky jsou vypočtené individuální jednoduché indexy.

Tab. 1: Ceny (Kč/litr), množství (litr/měsíc) a individuální jednoduché indexy.

i	p_0	q_0	p_1	q_1	I_p	I_q	I_h
1	35,2	400	35,9	380	1,0199	0,9500	0,9689
2	35,2	350	36,2	350	1,0284	1,0000	1,0284
3	35,6	350	35,6	310	1,0000	0,8857	0,8857
4	35,8	320	36,5	290	1,0196	0,9063	0,9240
5	35,9	410	36,4	350	1,0139	0,8537	0,8655
6	35,9	420	36,9	380	1,0279	0,9048	0,9300
7	36,5	450	36,9	450	1,0110	1,0000	1,0110
8	36,6	280	37,1	310	1,0137	1,1071	1,1223
9	36,9	290	37,8	250	1,0244	0,8621	0,8831
10	37,1	240	37,5	190	1,0108	0,7917	0,8002
11	37,2	190	37,9	200	1,0188	1,0526	1,0724
12	37,5	200	38,3	180	1,0213	0,9000	0,9192
13	37,9	280	37,5	250	0,9894	0,8929	0,8834
14	38,3	150	38,6	200	1,0078	1,3333	1,3438
15	38,3	180	38,9	220	1,0157	1,2222	1,2414
16	38,5	180	39,2	190	1,0182	1,0556	1,0747
17	38,9	210	39,4	250	1,0129	1,1905	1,2058
18	39,1	160	39,9	170	1,0205	1,0625	1,0842
19	39,2	130	39,5	120	1,0077	0,9231	0,9301
20	39,5	210	40,2	200	1,0177	0,9524	0,9693

Zdroj: vlastní zpracování

Následující vypočtené hodnoty byly získány pomocí software Statgraphics Centurion XV. Základní číselné charakteristiky statistických souborů z Tab. 1 jsou v Tab. 2 a Tab. 3 obsahuje konfidenční a bootstrapové intervalové odhady se spolehlivostí 0,95. Počet bootstrapových výběrů byl $B = 500$.

Tab. 2: Základní číselné charakteristiky

	p_0	q_0	p_1	q_1	I_p	I_q	I_h
Average	37,255	270,0	37,81	262,0	1,01498	0,992325	1,00717
Median	37,15	260,0	37,65	250,0	1,0167	0,9512	0,9691
Stnd. deviation	1,40055	99,7365	1,38864	86,9119	0,00920	0,13835	0,14029
Var. coeff. (%)	3,759	36,939	3,673	33,172	0,906	13,942	13,929
Minimum	35,2	130,0	35,6	120,0	0,9894	0,7917	0,8002
Maximum	39,5	450,0	40,2	450,0	1,0284	1,3333	1,3438
Range	4,3	320,0	4,6	330,0	0,039	0,5416	0,5436
Stnd. skewness	0,12414	0,72448	0,29142	0,94371	-2,03531	1,78192	1,61728
Stnd. kurtosis	-1,19977	-1,04824	-1,02689	-0,46534	1,89232	0,51270	0,30505

Zdroj: vlastní zpracování

Tab. 3: Konfidenční a bootstrapové intervalové odhady

		p_0		q_0	
Confidence Intervals	Mean	36,5995	37,9105	223,322	316,678
	Stnd. deviation	1,06511	2,04561	75,8486	145,672
Bootstrap Intervals	Mean	36,67	37,87	228,0	315,0
	Stnd. deviation	1,05903	1,63239	1,05903	1,63239
	Median	36,2	38,3	195,0	335,0
		p_1		q_1	
Confidence Intervals	Mean	37,1601	38,4599	221,324	302,676
	Stnd. deviation	1,05605	2,02821	66,0957	126,941
Bootstrap Intervals	Mean	37,215	38,45	231,5	298,0
	Stnd. deviation	1,00315	1,6048	62,7359	108,487
	Median	36,9	38,75	200,0	310,0
		I_p		I_q	
Confidence Intervals	Mean	1,01067	1,01929	0,927575	1,05708
	Stnd. deviation	0,006997	0,013438	0,105214	0,202070
Bootstrap Intervals	Mean	1,0105	1,01856	0,9382	1,05134
	Stnd. deviation	0,005255	0,012392	0,086399	0,178028
	Median	1,01195	1,01975	0,9216	1,0747

		I_h	
Confidence Intervals	Mean	0,941513	1,07283
	Stnd. deviation	0,106687	0,204900
Bootstrap Intervals	Mean	0,95181	1,07761
	Stnd. deviation	0,086399	0,178028
	Median	0,9216	1,0747

Zdroj: vlastní zpracování

Z Tab. 3 je zřejmé, že intervalové odhady (ostatně jako vždy) oproti bodovým odhadům realizovaným číselnými charakteristikami z Tab. 2 vypovídají daleko více o středních hodnotách, směrodatných odchylkách a mediánech cen množství i indexů. Navíc vypočtené bootstrapové intervaly středních hodnot a směrodatných odchylek jsou povětšinou menší než vypočtené konfidenční intervaly týchž parametrů, takže jsou při stejné spolehlivosti 0,95 přesnější. Nezanedbatelná je také skutečnost, že bootstrapové intervaly nezávisejí na rozděleních pravděpodobnosti pozorovaných cen a množství, takže není nutno verifikovat obvyklý požadavek, že ceny a množství mají normální rozdělení. Navíc předpoklad normálního rozdělení cen asi neobstojí, protože ceny stanovují prodejci. Měsíční nákupy pak určují zákazníci, takže ani množství nemají obvykle normální rozdělení.

5 ZÁVĚR

Použití bootstrapových odhadů v indexové analýze je netradiční. Bohužel ani konfidenční intervalové odhady se v ekonomických průzkumech moc nepoužívají anebo jejich použití nebývá seriózní. Naopak bootstrapové intervalové odhady umožňují vyrovnat se zejména s problémem neznámého pozorovaného rozdělení pravděpodobnosti, s extrémně odchýlenými hodnotami a také s nepříliš velkými rozsahy pozorovaných statistických souborů. Podobně jako popsané bootstrapové odhady individuálních jednoduchých indexů můžeme také získat bootstrapové odhady individuálních složených indexů a souhrnných indexů.

Metoda bootstrap nachází uplatnění i v jiných oblastech matematické statistiky, např. při fitování rozdělení pravděpodobnosti kategoriální veličiny (Karpíšek, Lacinová 2010). Jiný nežli v tomto článku prezentovaný způsob stanovení intervalových hodnot ekonomických indexů, který využívá tzv. ***intervalovou aritmetiku***, je popsán v článku (Karpíšek et al. 2012).

POUŽITÉ ZDROJE

[1] DAVISON, A. C. and HINKLEY, D. V. Bootstrap Methods and their Applications. Cambridge: Cambridge University Press, 2006. ISBN 0-521-57471-4.

[2] DUDEK, P. Řízení zpravodajského systému ve skupině Unipetrol. In: INSOURCE 2008: konference o profesionálních informačních zdrojích, Praha 5. - 6. února 2008: sborník příspěvků konference [online]. Praha: Albertina icome Praha, 2008 [cit. 27.4.2014]. Dostupné z: http://www.insource.cz/pdf/2008/dudek-petr2.pdf. ISBN 978-80-7224-565-9

[3] EFRON, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7(1), 1979, pp. 1 - 26.

[4] EFRON, B. and TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. New York: Chapmann & Hall, 1993. ISBN 0-412-04231-2.

[5] KARPÍŠEK, Z. a SADOVSKÝ, Z. *Matematické metody 2*. Elektronický učební text. Brno: AKADEMIE STING, 2007.

[6] KARPÍŠEK, Z., a LACINOVÁ, V. *Odhady diskrétního rozdělení pravděpodobnosti s použitím kvazinorem a bootstrapu.* Analýza dat 2010/II – Statistické metody pro technologii a výzkum. Pardubice: TriloByte, CQR, 2010, pp. 131-145. ISBN 978-80-904053-3-2.

[7] KARPÍŠEK, Z., KOCMANOVÁ, A., KRÁL, D., a LACINOVÁ, V. Aplikace intervalové aritmetiky v indexové analýze. *ACTA STING*, Brno, 2012, č. 1, s. 13-24, ISSN 1805-1391 (Print), ISSN 1805-6873 (Online).

[8] SEGER, J. a HINDLS, R. *Statistické metody v tržním hospodářství*. Praha: Victoria Publishing, 1995, ISBN 80-7187-058-7.

PODĚKOVÁNÍ:

Příspěvek je součástí řešení výzkumného projektu AKADEMIE STING IGA_AS_03 „Podpora řízení podniků“, grantového projektu GAČR reg. č. P403/11/2085 „Konstrukce metod pro vícefaktorové měření komplexní podnikové výkonnosti ve vybraném odvětví“ a projektu TAČR TA02021449 "Systém inteligentních alarmů v energetickém provozu jaderných elektráren".

AUTOŘI:

doc. RNDr. Zdeněk Karpíšek, CSc., Katedra aplikovaných disciplín, AKADEMIE STING, o.p.s., Stromovka 1, 637 00 Brno, Česká republika, e-mail: karpisek@sting.cz

doc. Ing. Alena Kocmanová, Ph.D., Fakulta podnikatelská, Vysoké učení technické v Brně, Technická 2, Brno, Česká republika, e-mail: kocmanova@fbm.vutbr.cz

doc. Ing. Zdeněk Sadovský, CSc., Katedra účetnictví a daní, AKADEMIE STING, o.p.s., Stromovka 1, 637 00 Brno, Česká republika, e-mail: sadovsky@sting.cz

Ing. Veronika Lacinová, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně, Technická 2, 616 69 Brno, Česká republika, e-mail: v.neradova@email.cz

AUTHORS:

doc. RNDr. Zdeněk Karpíšek, CSc., Department of Applied Disciplines, STING ACADEMY, Stromovka 1, 637 00 Brno, Czech Republic, e-mail: karpisek@sting.cz

doc. Ing. Alena Kocmanová, Ph.D., Faculty of Business and Management, Brno University of Technology, Technická 2, Brno, Czech Republic, e-mail: kocmanova@fbm.vutbr.cz

doc. Ing. Zdeněk Sadovský, CSc., Department of Accounting and Taxes, STING ACADEMY, Stromovka 1, 637 00 Brno, Czech Republic, e-mail: sadovsky@sting.cz

Ing. Veronika Lacinová, Department of Mathematics, Faculty of Civil Engineering, Brno University of Technology, Technická 2, Brno, Czech Republic, e-mail: v.neradova@email.cz