INTERVALOVÝ STATISTICKÝ SOUBOR A JEHO CHARAKTERISTIKY

INTERVAL STATISTICAL SAMPLE DATA AND ITS CHARACTERISTICS

Zdeněk Karpíšek, Marianna Dražanová, Veronika Lacinová, Jakub Šácha

Abstrakt: V článku je prezentováno netradiční pojetí a popis číselných charakteristik statistického souboru, které je založeno na pojmech a metodách intervalové analýzy, při modelování číselného statistického souboru, jehož pozorované hodnoty jsou nepřesné. Tyto nepřesnosti jsou vyjádřeny pomocí intervalů, které v aplikacích můžeme stanovit expertně.

Klíčová slova: intervalová aritmetika, intervalová funkce, intervalový statistický soubor, intervalové charakteristiky, metoda Monte Carlo

Abstract: The article presents a non-traditional conception and description of the numerical characteristics of statistical sample data, which is based on the notions and methods of interval analysis, at modeling of the numeric sample data whereof observed values they are inaccurate. These inaccuracies are expressed by the help of an intervals which in applications we can determine expert.

Keywords: interval arithmetic, interval function, interval statistical sample data, interval characteristics, Monte Carlo method

JEL klasifikace: C53, C46, E01

1 ÚVOD

Motivací k tomuto příspěvku bylo získat intervalové odhady číselných charakteristik statistického souboru z expertních nebo statistických intervalových odhadů hodnot pozorované ekonomické nebo finanční veličiny (znaku, ukazatele, indikátoru), a to pomocí intervalové analýzy. Důvodem k takovému modelování je skutečnost, že v praxi se často nepřesné hodnoty

pozorovaných veličin považují za zcela přesné, tedy nikoli např. ve formě intervalů. Závěry vyvozené klasickými statistickými metodami z těchto nepřesných hodnot pak ale nemusí odpovídat skutečnosti. Abychom získali seriozní a rigorózní matematický model pro praktické aplikace, je proto definován intervalový statistický soubor včetně jeho intervalových číselných charakteristik.

2 INTERVALOVÁ ANALÝZA A METODA MONTE CARLO

Intervalovým číslem rozumíme [2], [3] uzavřený reálný interval $[a, b]$, $a \le b$, kde a, b jsou reálná čísla. *Aritmetické operace s intervalovými čísly* definujeme vztahy:

$$\begin{aligned}
&[a,b]+[c,d]=[a+c,b+d],\\
&[a,b]-[c,d]=[a-d,b-c],\\
&[a,b]\cdot[c,d]=[\min\{ac,ad,bc,bd\},\max\{ac,ad,bc,bd\}],\\
&[a,b]/[c,d]=[a,b]\cdot[1/d,1/c] \text{ pro } 0\notin[c,d].
\end{aligned} \tag{1}$$

Pro $\forall a \in (-\infty,\infty)$ klademe $a=[a,a]$. Jestliže $a > 0$, pak píšeme $[a,b]>0$ atd. Zřejmě je

$$\lambda[a,b]=\begin{cases}[\lambda a,\lambda b] \text{ pro } \lambda>0,\\ 0 \quad \text{pro } \lambda=0,\\ [\lambda b,\lambda a] \text{ pro } \lambda<0,\end{cases} \tag{2}$$

kde $\lambda \in (-\infty,\infty)$.V aplikacích dle potřeby symbol operace násobení $\cdot$ vynecháváme a symbol operace dělení / nahrazujeme zlomkem.

Jestliže *J*, *K*, *L*, *M* jsou intervalová čísla, pak platí:

$$\begin{aligned}
&J+K=K+J,\\
&J+(K+L)=(J+K)+L,\\
&J\cdot K=K\cdot J,\\
&J\cdot(K\cdot L)=(J\cdot K)\cdot L,\\
&0+J=J,\\
&1\cdot J=J,\\
&J\cdot(K+L)\subset(J\cdot K)+(J\cdot L).
\end{aligned} \tag{3}$$

Speciálně pro $K\cdot L>0$ je $J\cdot(K+L)=J\cdot K+J\cdot L$.

Jestliže $J \subset L$ a $K \subset M$, pak

$$\begin{aligned} &J + K \subset L + M, \\ &J - K \subset L - M, \\ &J \cdot K \subset L \cdot M, \\ &J / K \subset L / M, \ (0 \notin M). \end{aligned} \tag{4}$$

Jestliže $J = [a,b]$ a $K = [c,d]$, pak pro $a \geq 0,\ c \geq 0$ je $J \cdot K = [ac, bd]$, pro $b \leq 0,\ d \leq 0$ je $J \cdot K = [bd, ac]$ a pro $a > 0, c > 0$ je $J / K = [a/d, b/c]$.

Jestliže $y = f(x_1,...,x_n)$ je reálná funkce a $I_1,...,I_n$ jsou intervalová čísla, pak ***intervalovou hodnotou*** této funkce rozumíme intervalové číslo (pokud existuje)

$$[\min f(x_1,...,x_n), \max f(x_1,...,x_n)], \tag{5}$$

kde $(x_1,...,x_n) \in I_1 \times \cdots \times I_n$, a hovoříme o ***intervalové funkci*** $f(I_1,...,I_n)$.

Jestliže spojitá funkce $y = f(x_1,...,x_n)$ na množině $I_1 \times \cdots \times I_n$ je rostoucí ve všech nezávisle proměnných na množině $I_1 \times \cdots \times I_n$, pak

$$\begin{aligned} &[\min f(x_1,...,x_n), \max f(x_1,...,x_n)] = \\ &[f(\min I_1,...,\min I_n), f(\max I_1,...,\max I_n)]. \end{aligned}$$

Analogicky pro spojitou funkci $y = f(x_1,...,x_n)$ klesající ve všech nezávisle proměnných na množině $I_1 \times \cdots \times I_n$ je

$$\begin{aligned} &[\min f(x_1,...,x_n), \max f(x_1,...,x_n)] = \\ &[f(\max I_1,...,\max I_n), f(\min I_1,...,\min I_n)]. \end{aligned}$$

Jestliže funkce $y = f(x_1,...,x_n)$ není rostoucí ani klesající ve všech svých proměnných, určíme její intervalovou hodnotu určením jejího absolutního minima a absolutního maxima na množině $I_1 \times \cdots \times I_n$ pomocí obvyklých analytických postupů nebo některé nelineární optimalizační metody na PC výpočtem vázaných extrémů na množině $I_1 \times \cdots \times I_n$, případně ji odhadneme simulační metodou ***Monte Carlo*** [1]. Tato metoda spočívá v realizaci dostatečně velkého počtu náhodných pokusů, kdy z intervalových čísel $I_1,...,I_n$ vybereme náhodná čísla $x_{ij} \in I_i$, $i = 1,...,n$ a $j = 1,...,N$. Počet vybraných *n*-tic $(x_{1j},...,x_{nj})$ obvykle volíme aspoň *N* = 10000 a čísla x_{ij} jsou hodnoty vzájemně nezávislých náhodných veličin X_{ij} s rovnoměrnými

rozděleními pravděpodobnosti na intervalech $I_1,...,I_n$. Pro všechny *n*-tice $(x_{1j},...,x_{nj})$ vypočteme hodnoty funkce $y = f(x_1,...,x_n)$ a získáme tak statistický soubor funkčních hodnot $(y_1,...,y_N)$. Intervalovou hodnotu $[\min f(x_1,...,x_n), \max f(x_1,...,x_n)]$ této funkce v intervalových číslech $I_1,...,I_n$ pak aproximujeme intervalovým číslem $[\min y_j, \max y_j]$, $j = 1,...,N$.

3 INTERVALOVÝ STATISTICKÝ SOUBOR A JEHO INTERVALOVÉ ČÍSELNÉ CHARAKTERISTIKY

V tomto oddílu zapisujeme intervalová čísla *I* jako intervaly $[\min x, \max x]$ apod.

Jestliže místo hodnot x_i kvantitativního statistického souboru $(x_1,...,x_n)$ [1] uvažujeme intervalová čísla $[\min x_i, \max x_i]$, tj. intervaly obsahující x_i, $i = 1,...,n$, dostaneme ***intervalový statistický soubor***

$$\left([\min x_1, \max x_1],...,[\min x_n, \max x_n]\right). \quad (6)$$

Analogicky modelujeme intervalově i dvourozměrný kvantitativní statistický soubor $\left((x_1, y_1),...,(x_n, y_n)\right)$ [1] ***dvourozměrným intervalovým statistickým souborem***

$$\left([\min x_1, \max x_1]\times[\min y_1, \max y_1],...,[\min x_n, \max x_n]\times[\min y_n, \max y_n]\right) \quad (7)$$

Pro popis intervalového statistického souboru používáme následující intervalové číselné charakteristiky, které vycházejí z vlastností intervalových čísel a způsobu určení intervalové hodnoty intervalové funkce (5) v oddílu 2.

Základní ***intervalové charakteristiky polohy*** intervalového statistického souboru jsou:

1. ***Intervalový aritmetický průměr*** je intervalová hodnota $[\min \bar{x}, \max \bar{x}]$ funkce $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ na intervalovém statistickém souboru (6). Platí, že

$$[\min \bar{x}, \max \bar{x}] = \left[\frac{1}{n}\sum_{i=1}^{n} \min x_i, \frac{1}{n}\sum_{i=1}^{n} \max x_i\right]. \tag{8}$$

Vlastnosti intervalového aritmetického průměru jsou:

a) $y = ax + b \Rightarrow [\min \bar{y}, \max \bar{y}] = a[\min \bar{x}, \max \bar{x}] + b$
pro reálné konstanty *a*, *b*,

b) $[\min(\overline{x+y}), \max(\overline{x+y})] = [\min \bar{x} + \min \bar{y}, \max \bar{x} + \max \bar{y}]$
pro soubory se stejným rozsahem.

Analogicky se definuje ***intervalový vážený aritmetický průměr***

$$[\min \bar{x}, \max \bar{x}] = \left[\frac{\sum_{i=1}^{n} w_i \min x_i}{\sum_{i=1}^{n} w_i}, \frac{\sum_{i=1}^{n} w_i \max x_i}{\sum_{i=1}^{n} w_i}\right], \tag{9}$$

kde $w_i \geq 0$ jsou neintervalové ***váhy***.

2. ***Intervalový medián*** je intervalová hodnota $[\min \tilde{x}, \max \tilde{x}]$ funkce

$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$ pro lichá n, resp. $\frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right]$ pro sudá n, na intervalovém statistickém souboru (6). Pro výpočet intervalového mediánu je vhodná jeho intervalová aproximace metodou Monte Carlo z oddílu 2. Vlastnosti intervalového mediánu jsou:

a) $y = ax + b \Rightarrow [\min \tilde{y}, \max \tilde{y}] = a[\min \tilde{x}, \max \tilde{x}] + b$
pro reálné konstanty *a*, *b*,

b) $\left[\min(\widetilde{x+y}), \max(\widetilde{x+y})\right] = [\min \tilde{x} + \min \tilde{y}, \max \tilde{x} + \max \tilde{y}]$
pro uspořádané soubory se stejným rozsahem.

3. ***Intervalový modus*** $[\min \hat{x}, \max \hat{x}]$ dostaneme (pokud existuje) aproximací podobně jako intervalový medián metodou Monte Carlo z oddílu 2.

Základní ***intervalové charakteristiky proměnlivosti*** (***variability***) intervalového statistického souboru jsou:

1. ***Intervalový rozptyl*** (***disperze***, ***variance***) je intervalová hodnota $[\min s^2, \max s^2]$ funkce $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \overline{x}^2$ na intervalovém statistickém souboru (6), kde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$. Pro výpočet intervalového rozptylu je vhodná jeho intervalová aproximace metodou Monte Carlo z oddílu 2. Vlastnosti intervalového rozptylu jsou:

a) $[\min s^2, \max s^2] \geq 0$,

b) $y = ax + b \Rightarrow [\min s^2(y), \max s^2(y)] = \left[a^2 \min s^2(x), a^2 \max s^2(x)\right]$ pro reálné konstanty *a*, *b*.

2. ***Intervalová směrodatná odchylka*** $[\min s, \max s] = [\sqrt{\min s^2}, \sqrt{\max s^2}]$.

Vlastnosti směrodatné odchylky jsou:

a) $[\min s, \max s] \geq 0$,

b) $y = ax + b \Rightarrow [\min s(y), \max s(y)] = \left[|a| \min s(x), |a| \max s(x)\right]$ pro reálné konstanty *a*, *b*.

3. ***Intervalový variační koeficient*** je intervalová hodnota $[\min v, \max v]$ funkce $v = \frac{s}{\overline{x}}$ na intervalovém statistickém souboru (6) pro $s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ a $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$. Platí, že $[\min v(ax), \max v(ax)] = \left[\min v(x), \max v(x)\right]$ pro reálnou konstantu $a \geq 0$ a analogicky pro $a < 0$ (záměna min a max). Tato relativní míra variability se uvádí také v %.

4. ***Intervalové rozpětí*** $[\min R, \max R]$ dostaneme aproximací podobně jako intervalový medián metodou Monte Carlo z oddílu 2 pro $R = \max x_i - \min x_i$. Intervalové rozpětí má stejné vlastnosti jako intervalová směrodatná odchylka.

Základní ***intervalovou charakteristikou souměrnosti*** intervalového statistického souboru je ***intervalový koeficient šikmosti*** (***asymetrie***) $[\min A, \max A]$, což je intervalová hodnota funkce $A = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$, na intervalovém statistickém

souboru (6), kde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a $s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Pro výpočet intervalového koeficientu šikmosti je vhodná jeho intervalová aproximace metodou Monte Carlo z oddílu 2. Vlastnosti intervalového koeficientu šikmosti jsou:

a) $[\min A, \max A] > 0 \Leftrightarrow$ většina intervalových hodnot $[\min x_i, \max x_i]$ je „menší“ než $[\min \overline{x}, \max \overline{x}]$,
b) $[\min A, \max A] = 0 \Leftrightarrow$ intervalové0 hodnoty $[\min x_i, \max x_i]$ jsou rozloženy víceméně souměrně vzhledem k $[\min \overline{x}, \max \overline{x}]$,
c) $[\min A, \max A] < 0 \Leftrightarrow$ většina intervalových hodnot $[\min x_i, \max x_i]$ je „větší“ než $[\min \overline{x}, \max \overline{x}]$,
d) $y = ax + b \Rightarrow [\min A(y), \max A(y)] = [\min A(x), \max A(x)]$
pro reálnou konstantu $a \geq 0$ a analogicky pro $a < 0$ (záměna min a max) a libovolnou reálnou konstantu *b.*

Příklad. Statistickým šetřením byly získány ceny Kč/l benzínu Natural 95 v červnu 2021 v deseti čerpacích stanicích v různých regionech ČR. Tyto hodnoty považujeme za nepřesné reprezentanty skutečných cen v ČR z důvodu jejich neznámých změn během daného období a způsobu výběru čerpacích stanic. V Tabulce 1 jsou uvedeny původní pozorované hodnoty ceny a jejich expertně stanovené intervalové hodnoty.

Tabulka 1: Pozorované ceny a jejich intervalové odhady

i	x_i	min x_i	max x_i
1	31,2	30,2	32,1
2	30,9	29,9	31,8
3	32,1	31,1	33,0
4	31,5	30,5	32,4
5	32,4	31,4	33,3
6	33,3	32,3	34,2
7	32,9	31,9	33,8
8	32,9	31,9	33,8
9	32,2	31,2	33,1
10	31,9	30,9	32,8

Zdroj: vlastní

V Tabulce 2 jsou prezentovány vypočtené základní číselné charakteristik pro původní „neintervalový“ statistický soubor a také intervalové číselné charakteristiky pro intervalový statistický soubor, které byly získány výpočtem pomocí vzorců z tohoto oddílu a aplikací pomocí metody Monte Carlo pro N = 10 000 experimentů. Výpočty byly provedeny v Excelu.

Tabulka 2: Vypočtené číselné charakteristiky

Číselné charakteristiky	Neintervalové hodnoty pro původní soubor	Intervalové hodnoty pro intervalový soubor	
		min	max
Aritmetický průměr	32,13	31,54735	32,48954
Medián	32,15	31,64035	32,64674
Rozptyl	0,5461	0,275236	1,488481
Směrodatná odchylka	0,738986	0,52463	1,220033
Variační koeficient (%)	2,299987	1,632173	3,795472
Rozpětí	2,4	1,61037	3,831346
Koeficient šikmosti	-0,09732	-1,34586	1,177304

Zdroj: vlastní

Z výsledků v Tabulce 2 vidíme, že intervalový přístup přináší významně více informací o variabilitě číselných charakteristik původního statistického souboru nežli klasické bodové popisné charakteristiky.

Nabízí se řada dalších intervalových číselných charakteristik intervalového statistického souboru. Např. pro intervalové poměrové statistické znaky (cenové a množstevní indexy, úrokové míry apod.) je nutno místo intervalového aritmetického průměru použít ***intervalový geometrický průměr***

$$\left[\min \bar{x}_g, \max \bar{x}_g\right] = \left[\sqrt[n]{\prod_{i=1}^{n} \min x_i}, \sqrt[n]{\prod_{i=1}^{n} \max x_i}\right] \tag{10}$$

a ve speciálních případech (např. pro statistické znaky vyjadřující rychlost nějakého děje) používáme ***intervalový harmonický průměr***

$$\left[\min \bar{x}_h, \max \bar{x}_h\right] = \left[\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\min x_i}\right)^{-1}, \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\max x_i}\right)^{-1}\right]. \tag{11}$$

Podle definice koeficientu korelace [1] a definice hodnoty intervalové funkce (5) je ***intervalový koeficient korelace*** $\left[\min r, \max r\right]$ dvourozměrného intervalového statistického souboru (7) intervalová hodnota funkce

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{12}$$

na množině všech dvojic $(x_i, y_i) \in [\min x_i, \max x_i] \times [\min y_i, \max y_i]$, kde $i = 1,...,n$ a $\overline{x}$, $\overline{y}$ jsou aritmetické průměry statistických souborů $(x_1,...,x_n)$, $(y_1,...,y_n)$. Pro výpočet intervalového koeficientu korelace je vhodná jeho intervalová aproximace metodou Monte Carlo z oddílu 2. Aplikace intervalového koeficientu korelace je možno najít v článku [3], kde jde o vyjádření korelace kvartálních hodnot hrubého domácího produktu HDP České republiky a denních kurzů CZK vzhledem k EUR a USD v letech 1999 až 2018.

ZÁVĚR

Matematické modelování neurčitých veličin je od počátku 20. století založeno zejména na jejich stochastickém pojetí, intervalové analýze a teorii fuzzy množin. Stochastické pojetí umožňuje aplikaci matematicko-statistických metod a má víceméně objektivní charakter, avšak je často omezeno neznalostí pozorovaných pravděpodobnostních rozdělení a složitostí výpočtů. Intervalové a fuzzy pojetí má sice subjektivní charakter, ale umožňuje naopak respektovat expertní pohled na neurčitost pozorovaných dat. Výsledky naznačené v tomto článku dokládají vhodnost intervalového přístupu pomocí intervalové analýzy a navíc uvedená metodika výpočtu intervalových číselných charakteristik nevyžaduje složité výpočty na PC. Intervalový přístup a jeho přínos při hledání lineárního trendu časové řady je popsán ve [4] a [5], kde jsou také prezentovány dosti překvapivé výsledky.

AFILACE

Příspěvek je součástí řešení výzkumných grantových projektů IGA_AS_03_01/2 a IGA AS_04 AKADEMIE STING v Brně.

POUŽITÉ ZDROJE

[1] MONTGOMERY, D. C. a RUNGER, G. *Applied Statistics and Probability for Engineers.* 5th ed. New York: John Wiley, 2010. 784 s. ISBN 978-0-470-05304-1.

[2] MOOR, R. E., KEARFOTT, R. B. a CLOUD, M. J. *Introduction to Interval Analysis*. Philadelphia: SIAM 2009, 459 s. ISBN 978-0-898716-69-6.

[3] KARPÍŠEK, Z., SLÁDKOVÁ, J. a DRAŽANOVÁ, M. Intervalová korelace ekonomic-kých indikátorů. *ACTA STING*, 2/2020, Volume 9, s. 22-28, ISSN 1805-6873.

[4] KARPISEK, Z., LACINOVA, V., SADOVSKY, Z. a SCHNEIDER, A. Is the Increasing Trend Always Really Increasing? MENDEL 2016 – 2^{2th} International Conference on Soft Computing. Brno, 2016. *Mendel Series*, Volume 2016, p. 229-234. ISSN 1803-3814, ISBN 978-80-214-5365-4.

[5] KARPÍŠEK, Z., DRAŽANOVÁ, M. a LACINOVÁ, V. Lineární regresní model intervalové časové řady. *ACTA STING*, 1/2019, Volume 8, s. 22-28, ISSN 1805-6873.

AUTOŘI

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

doc. Ing. Marianna Dražanová, CSc., Katedra ekonomiky a řízení, AKADEMIE STING, o.p.s., Stromovka 1, 637 00 Brno, e-mail: drazanova@post.sting.cz.

Ing. Veronika Lacinová, Ph.D., Katedra kvantitativních metod, Fakulta vojenského leadershipu, Univerzita obrany, Kounicova 156/65, 662 10 Brno, e-mail: veronika.lacinova@unob.cz.

Ing. Jakub Šácha, Ph.D., Ústav statistiky a operační analýzy, Provozně ekonomická fakulta, Mendelova univerzita v Brně, Zemědělská 1, 613 00 Brno, e-mail: jakub.sacha@mendelu.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. Marianna Dražanová, CSc., Department of Economics and Management, STING ACADEMY, Stromovka 1, 637 00 Brno, Czech Republic, e-mail: drazanova@post.sting.cz.

Ing. Veronika Lacinová, Ph.D., Department of Quantitative Methods, Faculty of Military Leadership, University of Defence in Brno, Kounicova 156/65, 662 10 Brno, Czech Republic, e-mail: veronika.lacinova@unob.cz.

Ing. Jakub Šácha, Ph.D., Department of Statistics and Operation Analysis, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: jakub.sacha@mendelu.cz.