Milan Terek
Nguyen Dinh He

MOŽNOSTI VYUŽITIA NIEKTORÝCH NETRADIČNÝCH CHARAKTERISTÍK V ANALÝZE MIEZD[1]

Abstract: *The paper deals with an analysis of how to use certain measures of location, based on analysis of outliers. Contaminants, extreme values and outliers are characterised. Trimming mean and M-estimators are described. Then the possibilities of detecting outliers are analysed. Computing of one-step M-Estimator and modified one-step M-Estimator of location is described. Further, the trimmed mean and M-estimators of location are compared. Finally, some non-traditional measures of location in the analysis of salaries are described.*

Keywords: *outliers, trimmed mean, M-Estimator, analysis of salaries.*

JEL: C 13

Úvod

Všimneme si niektoré netradičné charakteristiky polohy – zastrihnutú strednú hodnotu a charakteristiku polohy odhadovanú pomocou M-estimátora. Pri ich formulácii sa berú do úvahy odľahlé údaje. Preto sa problematike chápania a detekcie odľahlých údajov budeme venovať pomerne podrobne.
Najprv si však všimnime dve „tradičné“ charakteristiky polohy – strednú hodnotu μ a medián základného súboru. Tieto charakteristiky majú poskytovať typickú hodnotu premennej, hodnotu, ktorá dobre reprezentuje všetky jej hodnoty.

Niekedy však leží stredná hodnota rozdelenia v niektorom z jeho koncov a poskytuje veľmi skreslený obraz o polohe rozdelenia, resp. o úrovni hodnôt.

Všimneme si najprv najznámejšie výberové charakteristiky polohy – *výberový priemer* $\overline{X}$ a *výberový medián* $X_{1/2}$.

Aj hodnota výberového priemeru môže ležať v niektorom z koncov rozdelenia početností. Uvedieme príklad takejto situácie.

[1]Článok vznikol s podporou grantovej agentúry VEGA v rámci projektu č.1/0440/10.

Príklad č. 1. Náhodne sa vybralo 100 mladých mužov. Odpovedali na otázku, koľko rozličných partneriek by chceli v živote mať. Výsledky triedenia odpovedí sú v tabuľke č. 1.

Tab. č. 1

Želaný počet rozličných partneriek

Počet Partneriek x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	17	40	60	100	8000
Počet odpovedí n_i	5	49	6	1	2	6	1	3	1	1	11	1	4	1	2	2	1	1	1	1

Hodnota výberového priemeru je 85,90. Táto hodnota nemá prakticky žiadnu výpovednú schopnosť, pretože 98 % hodnôt pozorovaní je menších ako táto hodnota. Čo spôsobilo takýto výsledok? Neobvykle veľké hodnoty premennej u niekoľkých respondentov.

1 Odľahlé údaje, extrémne údaje a kontaminanty

Prakticky vo všetkých súboroch pozorovaní sa nachádzajú údaje, ktoré sa natoľko líšia od ostatných, že naznačujú existenciu nejakého zvláštneho zdroja chýb, o ktorom sme v teoretických predpokladoch neuvažovali a ktorého zahrnutie do úvah môže spôsobiť iba skomplikovanie a nesprávne nasmerovanie analýzy ([1], s. 3). Ide o *odľahlé údaje.*

Odľahlé údaje v množine údajov možno definovať ako pozorovania, ktoré sa zdajú byť nekonzistentné s ostatnými údajmi v množine údajov ([1], s. 7). Niekedy sa odľahlé údaje definujú jednoducho ako neobvykle malé alebo veľké hodnoty v množine údajov.

Treba však poznamenať, že *odľahlé údaje* (*outliers*) nemusia byť vždy mätúce, zlé alebo chybné. Naopak, existujú situácie, v ktorých sú odľahlé údaje vítané, pretože niečo naznačujú, napríklad nejaký neobyčajne úspešný priemyselný postup. Často je ich existencia veľmi užitočná pri odhaľovaní podvodov. V týchto prípadoch tiež nemusí byť nevyhnutné prijať niektorý z extrémov: vyradiť ich (s rizikom straty užitočnej informácie) alebo ich zahrnúť do analýzy (s rizikom kontaminácie údajov). Niekedy môže byť užitočné použiť robustné metódy štatistickej indukcie, ktoré využívajú všetky údaje, ale minimalizujú účinok odľahlých údajov.

Treba tiež poznamenať, že problému odľahlých údajov sa nemožno jednoducho vyhnúť tak, že ho nebudeme brať do úvahy. Všetci, ktorí pracujú s údajmi, jednoducho musia prijímať rozhodnutia, ktoré sa týkajú odľahlých údajov – či ich zahrnúť do analýzy alebo nie, či ich použitie nejako ohraničiť a pod.

Všimnime si teraz aj pojmy *extrémne údaje* (*extreme data*) a *kontaminanty* (*contaminants*).

Uvedieme príklad ([1], s. 4 – 6). Ide o súdne rozhodnutie o rozvode manželstva z roku 1949. Navrhovateľ zdôvodňuje svoj návrh tým, že dieťa sa v ich manželstve narodilo 349 dní po jeho odchode od manželky na služobný pobyt a teda nemôže byť jeho, pretože nemohlo byť počaté v čase jeho prítomnosti. Podľa odborných analýz, ktoré mal súd k dispozícii, je priemerná dĺžka materstva 280 dní. Doba 349 dní je prekvapujúco dlhá. Navrhovateľ ju považoval za odľahlú hodnotu. Samozrejme si neželal, aby sa ignorovala. Naopak, chcel, aby sa brala do úvahy a viedla k príslušnému dôsledku (rozvodu). Všeobecne, rozdielne ciele môžu viesť k rozdielnym postojom k odľahlým údajom. Navrhovateľ nebol úspešný. Sudca súhlasil s tým, že doba materstva je ohraničená, ale že 349 dní, aj keď je to málo pravdepodobná dĺžka, nie je mimo rozpätia možnej dĺžky materstva. Iný súd v podobnom prípade rozhodol, že 340 dní je vo svetle modernej gynekológie nemožná dĺžka materstva. V inom podobnom prípade z roku 1951 súd uvažoval o hranici 360 dní.

Pokúsme sa teraz na základe tohto príkladu charakterizovať pojmy odľahlé údaje, extrémne údaje a kontaminanty. Doba 349 dní je extrémny údaj, ktorý označíme odľahlý údaj. Toto ale neurčuje status pozorovania 349. Môže ísť o správne pozorovanie z rozdelenia dĺžky materstva, ktoré je prekvapivo veľké. Takto o ňom rozhodol aj súd. Môže ale ísť aj o kontaminant – pozorovanie z iného rozdelenia (v príklade – z rozdelenia s neskorším začiatkom počatia). Presne takýto názor mal v príklade navrhovateľ. Všeobecne ale napríklad aj údaj 280 dní môže byť kontaminant. Skutočná dĺžka materstva mohla byť v tomto prípade napríklad 260 dní a dieťa bolo počaté o 20 dní neskôr, než sa predpokladalo.

Extrémne hodnoty môžu, ale nemusia byť odľahlé hodnoty. Odľahlé hodnoty sú vždy extrémne hodnoty. Odľahlé hodnoty môžu, ale nemusia byť kontaminanty, kontaminanty môžu, ale nemusia byť odľahlé hodnoty. Nepoznáme spôsob na určenie, či nejaké pozorovanie je alebo nie je kontaminant. Môžeme sústrediť pozornosť len na odľahlé hodnoty ako na možný prejav kontaminácie.

Pri určovaní odľahlých údajov ide o *ohodnotenie integrity množiny údajov*. Údaje z výberu možno analyzovať, aby sa ohodnotila platnosť nejakého uvažovaného modelu, alebo aby sa odhadli jeho parametre, alebo sa o nich testovali hypotézy. Potrebujeme metódy na ohodnotenie, vyradenie alebo minimalizáciu vplyvu odľahlých údajov (metódy robustné voči odľahlým údajom). Odľahlú hodnotu charakterizuje hlavne jej účinok na analytika (ide v nejakom zmysle o prekvapujúco extrémnu hodnotu).

Všimnime si teraz procedúru riešenia problému odľahlých údajov. Skúmame nejakú množinu údajov. Predpokladajme, že sa rozhodneme, že odľahlé údaje sú v množine prítomné. Potom si musíme položiť otázku: "Ako by sme mali reagovať na prítomnosť odľahlých údajov a aké princípy a metódy by sme mali použiť na ich vylúčenie, modifikáciu ich hodnôt alebo účinku predtým, ako začneme analyzovať množinu zostávajúcich údajov?" Odpoveď závisí od formy základného súboru, použité techniky závisia hlavne od prijatého modelu základného súboru. Metódy na spracovanie odľahlých údajov majú teda relatívnu formu: relatívnu vzhľadom na základný model.

V príklade 1 vidno, že výberový priemer je výrazne ovplyvnený jedinou odľahlou hodnotou. Všimnime si teraz výberový medián $X_{1/2}$. V príklade hodnota výberového mediánu[2] je 1 a určite lepšie charakterizuje typické želania vybratých mužov.

2 Výberový zastrihnutý priemer

Všimnime si teraz *výberový zastrihnutý priemer* (*sampl*e *trimmed mean*). Hodnota výberového zastrihnutého priemeru sa vypočíta z údajov, z ktorých bol vynechaný určitý podiel najväčších a najmenších hodnôt. Špeciálnym prípadom výberového zastrihnutého priemeru je výberový priemer, v ktorom neboli pri výpočte vynechané žiadne údaje z pôvodného výberového súboru. Napríklad 10 % výberový zastrihnutý priemer znamená, že pri výpočte jeho hodnoty pre konkrétny výber sa 10 % najmenších a 10 % najväčších hodnôt pozorovaní vynechá.

Príklad č. 1 – pokračovanie 1

Vypočítame hodnotu 10 % a 20 % zastrihnutého priemeru z údajov v príklade č. 1. Výsledky sú takéto.

10 % výberový zastrihnutý priemer $\overline{X}_{t(0,1)}$ je

$$\overline{X}_{t(0,1)} = \frac{1}{80}\left(x_{11} + x_{12} + ... + x_{90}\right) = 3{,}6$$

20 % výberový zastrihnutý priemer $\overline{X}_{t(0,2)}$ je

$$\overline{X}_{t(0,2)} \approx 3{,}02$$

Príklad ilustruje výrazne lepšiu výpovednú schopnosť výberového zastrihnutého priemeru v porovnaní s výberovým priemerom.

Zásadnou otázkou je, aký podiel údajov by sme mali „odstrihnúť“. Pri riešení mnohých praktických problémov je vhodné použiť 20 % výberový zastrihnutý priemer, ktorý často poskytuje lepší výsledok ako výberový priemer a medián ([5], s. 63).

3 M-estimátory

M-estimátory tvoria inú triedu charakteristík polohy, ktoré majú značný praktický význam. Keď pre nejakých *n* pozorovaní X_1, X_2, X_n... chceme nájsť také číslo *c*, ktoré minimalizuje sumu štvorcov odchýlok

[2]Hodnotu, pre ktorú platí, že polovica jednotiek vo výbere má hodnoty premennej menšie alebo rovné ako táto hodnota, nezveme hodnotou výberového mediánu.

$$\sum_{i=1}^{n}\left(X_i - c\right)^2 \qquad (1)$$

musí byť $\sum_{i=1}^{n}\left(X_i - c\right)^2 = 0$. Z toho c = $\overline{X}$. Keď teda hľadáme charakteristiku polohy, založenú na minimalizácii štvorcov chýb daných vzťahom (1), vedie to k použitiu výberového priemeru ([5], s. 66).

Keď na meranie vzdialeností X_i od *c* použijeme sumu absolútnych hodnôt rozdielov

$$\sum_{i=1}^{n}\left|X_i - c\right| \qquad (2)$$

hodnotu výrazu (2) minimalizuje výberový medián $X_{1/2}$ ([5], s. 67).

Všeobecne existuje nekonečne veľa spôsobov merania vzdialenosti, ktoré vedú k príslušným charakteristikám polohy. Keby sme napríklad vzdialenosť merali pomocou $\sum_{i=1}^{n}\left|X_i - c\right|^a$, potom $a = 1$ vedie k výberovému mediánu, $a = 2$ vedie k výberovému priemeru.

Majme nejakú funkciu Ψ takú, že: $\Psi(-x) = -\Psi(x)$. Vhodnú charakteristiku polohy získame (za predpokladu, že funkcia hustoty je symetrická[3]) tak, že nájdeme také *c*, ktoré vyhovuje rovnici

$$\Psi(X_1 - c) + \Psi(X_2 - c) + ... + \Psi(X_n - c) = 0 \qquad (3)$$

Charakteristiky polohy, založené na rovnici (3), sa nazývajú M-estimátory. Pri výpočte ich hodnoty je nevyhnutná detekcia odľahlých údajov, preto si najskôr všimneme možnosti tejto detekcie.

3.1 Detekcia odľahlých údajov

Prvá stratégia detekcie odľahlých údajov, ktorú si všimneme, je založená na výberovom priemere a výberovom rozptyle.

Keď prijmeme predpoklad o normálnom rozdelení základného súboru, je obvyklé považovať za odľahlú hodnotu, ktorá je vzdialená od strednej hodnoty viac ako o 2,24 smerodajných odchýlok. Inak, hodnota *x* je považovaná za odľahlú, keď

[3]Pre niektoré funkcie Ψ vyžaduje nájdenie vhodných charakteristík polohy špeciálny prístup, keď ide o zošikmené rozdelenia.

$$\frac{|x-\mu|}{\sigma} > 2{,}24$$

Keď predpokladáme normálne rozdelenie základného súboru so strednou hodnotou μ a smerodajnou odchýlkou σ, pravdepodobnosť, že hodnota bude považovaná za odľahlú, je 0,025. Strednú hodnotu a smerodajnú odchýlku σ väčšinou nepoznáme a odhadujeme ich najčastejšie pomocou výberového priemeru a výberovej smerodajnej odchýlky. Potom možno formulovať takéto rozhodovacie pravidlo:

Hodnota *x* je považovaná za odľahlú, keď

$$\frac{|x-\bar{x}|}{s} > 2{,}24 \qquad (4)$$

kde $s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$ je smerodajná odchýlka

Príklad č. 2

Uvažujme o takýchto hodnotách:

1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 500

Hodnota výberového priemeru je $\bar{x} = 31{,}76$, hodnota výberovej smerodajnej odchýlky je *s* = 120,67. Odhadneme vzdialenosť hodnoty 500 od strednej hodnoty, meranej v smerodajných odchýlkach:

$$\frac{|500 - 31{,}76|}{120{,}67} \approx 3{,}88$$

V tomto prípade, podľa rozhodovacieho pravidla (4), hodnota 500 je odľahlá hodnota.

Príklad č. 3

Pridajme k údajom z príkladu č. 2 ešte jednu hodnotu: 1 000. Potom je súbor údajov takýto

1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 500, 1 000

Hodnota výberového priemeru je $\bar{x} = 85{,}56$, hodnota výberovej smerodajnej od-

chýlky je s = 256,49. Odhadneme vzdialenosť hodnoty 500 od strednej hodnoty, meranej v smerodajných odchýlkach:

$$\frac{|500 - 85{,}56|}{256{,}49} \approx 1{,}62$$

Podľa rozhodovacieho pravidla (4), hodnota 500 nie je odľahlá, aj keď z jednoduchého porovnania hodnôt je zrejmé, že ide o odľahlú hodnotu.

Posledný príklad je ilustráciou problému, ktorý je známy ako *maskovanie* (*masking*). Odľahlé hodnoty ovplyvňujú výpočet výberového priemeru aj výberovej smerodajnej odchýlky, čo môže na druhej strane maskovať ich prítomnosť, keď na ich detekciu využívame vzťah (4) ([5], s. 77).

Potrebujeme také rozhodovacie pravidlo na detekciu odľahlých hodnôt, ktoré nie je ovplyvnené odľahlými hodnotami. Uvedieme jednu robustnú metódu na detekciu odľahlých hodnôt.

3.1.1 Metóda založená na MADN

Najprv si všimneme jednu charakteristiku variability, ktorá sa nazýva *mediánová absolútna odchýlka* (*median absolute deviation*) – MAD. Na výpočet hodnoty MAD je nevyhnutné najprv vypočítať hodnotu $x_{1/2}$ výberového mediánu $X_{1/2}$, potom vypočítať absolútne hodnoty odchýlok hodnôt od mediánu:

$$|x_i - x_{1/2}| \quad \text{pre } i = 1, 2, \ldots n$$

MAD je medián absolútnych hodnôt odchýlok hodnôt od mediánu. Možno ukázať, že keď máme náhodný výber z normálneho rozdelenia, potom

$$\text{MADN} = \frac{\text{MAD}}{0{,}6745}$$

je dobrým bodovým odhadom smerodajnej odchýlky σ základného súboru (Wilcox, s. 73).

Vráťme sa teraz k hľadaniu rozhodovacieho pravidla na detekciu odľahlých hodnôt, ktoré samotné nie je ovplyvnené odľahlými hodnotami. Jednou z možností je takéto rozhodovacie pravidlo.

Hodnota x je odľahlá, keď

$$\frac{|x - x_{1/2}|}{\text{MADN}} > 2{,}24 \qquad (5)$$

Príklad č. 3 – pokračovanie 1

Posúďme odľahlosť hodnoty 500 pomocou rozhodovacieho pravidla (5).

Hodnota výberového mediánu $x_{1/2} = 3$. Absolútne hodnoty odchýlok hodnôt od mediánu, usporiadané podľa veľkosti vzostupne, sú:

0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 497, 997

Potom MAD = 1 a MADN ≈ 1,48.

Dosaďme za x, $x_{1/2}$ a MADN do vzťahu (5). Dostaneme:

$$\frac{|500-3|}{1{,}48} \approx 335{,}81 > 2{,}24$$

Podľa rozhodovacieho pravidla (5), hodnota 500 je odľahlá.

3.2 Jednokrokový M-estimátor polohy

Nech n_1 je počet pozorovaní X_i, pre ktoré:

$$\frac{(X_i - X_{1/2})}{\text{MADN}} < -K$$

a n_2 je počet pozorovaní, pre ktoré

$$\frac{(X_i - X_{1/2})}{\text{MADN}} > K$$

Najčastejšie sa používa: $K = 1{,}28$. Potom jednokrokový M-estimátor polohy je

$$\hat{\mu}_{os} = \frac{K(\text{MADN})(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} \qquad (6)$$

kde $X_{(i)}$ je i-ta poriadková štatistika.[4]

[4]Poriadková štatistika je určená svojím poradím v neklesajúcom usporiadaní náhodných premenných.

Pri výpočte hodnoty tohto estimátora sa postupuje tak, že sa metódou založenou na MADN nájdu odľahlé hodnoty, s tým rozdielom, že namiesto vzťahu (5) sa pri detekcii odľahlých hodnôt použije vzťah (7)

$$\frac{\left|x - x_{1/2}\right|}{\text{MADN}} > K \qquad (7)$$

Hodnoty určené ako odľahlé sa vylúčia a z ostatných sa vypočíta priemer. Z technických dôvodov je výpočet hodnoty estimátora založený na MADN a počte odľahlých hodnôt nad a pod mediánom ([5], s. 82).

Príklad č. 1 – pokračovanie 2

Vypočítajme pre údaje z príkladu 1 hodnotu jednokrokového M-estimátora s $K = 1{,}28$.

Hodnota výberového mediánu je $x_{1/2} = 1$. V tabuľke č. 2 sú v prvom riadku absolútne hodnoty odchýlok hodnôt premennej od mediánu.

Tab. č. 2

Odchýlky od mediánu

$\left\|x_i - x_{\frac{1}{2}}\right\|$	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	16	39	59	99	7 999
n_i	5	49	6	1	2	6	1	3	1	1	11	1	4	1	2	2	1	1	1	1

Tab. č. 3

Odchýlky od mediánu usporiadané vzostupne

$\left\|x_i - x_{\frac{1}{2}}\right\|$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	16	39	59	99	7 999
n_i	49	11	1	2	6	1	3	1	1	11	1	4	1	2	2	1	1	1	1

V tabuľke č. 3 nájdeme medián odchýlok MAD = 1. Potom vypočítame MADN:

$$\text{MADN} = \frac{\text{MAD}}{0{,}6745} = \frac{1}{0{,}6745} \approx 1{,}4826$$

V tabuľke č. 4 sú odchýlky od mediánu delené MADN.

Tab. č. 4

Odchýlky od mediánu delené MADN

x_i	0	1	2	3	4	5	6	7	8
$\left(x_i - x_{\frac{1}{2}}\right)$/MADN	–0,67	0	0,67	1,35	2,02	2,69	3,37	4,05	4,72
n_i	5	49	6	1	2	6	1	3	1

Tab. č. 5

Odchýlky od mediánu delené MADN – pokračovanie

x_i	9	10	11	12	13	14	17	40	60	100	8 000
$\left(x_i - x_{\frac{1}{2}}\right)$/MADN	5,4	6,07	6,74	7,42	8,09	8,77	10,79	26,31	39,79	66,77	5 395,25
n_i	1	11	1	4	1	2	2	1	1	1	1

Odľahlé hodnoty určíme podľa vzťahu (7), pričom položíme $K = 1{,}28$. Vidíme, že žiadne hodnoty menšie ako medián nie sú odľahlé. Potom $n_1 = 0$. Ďalej je v tabuľkách č. 4 a č. 5 vidno, že z hodnôt väčších ako medián považujeme 40 hodnôt, podľa (7), za odľahlé. Potom $n_2 = 40$. Po dosadení do (6) dostaneme

$$\hat{\mu}_{os} = \frac{1{,}28(1{,}4826)(40-0)+1\cdot 49+2\cdot 6}{100-0-40} \approx 2{,}28$$

Podľa tohto estimátora, „typický“ želaný počet rozličných partneriek vo výbere je približne 2,28.

3.3 Modifikovaný jednokrokový M-estimátor

Niekedy sa používa jednoduchá modifikácia jednokrokového M-estimátora:

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n-n_1-n_2} \qquad (8)$$

V tomto estimátore sa na determináciu n_1 a n_2 používa $K = 2{,}24$.

Príklad č. 1 – pokračovanie 3

Vypočítajme pre údaje z príkladu č. 1 hodnotu modifikovaného jednokrokového M-estimátora. Zrejme $n_1 = 0$ a $n_2 = 37$. Potom po dosadení do (8) dostaneme

$$\hat{\mu}_{mom} = \frac{1 \cdot 49 + 2 \cdot 6 + 3 \cdot 1 + 4 \cdot 2}{100 - 0 - 37} = 1{,}14$$

Je zrejmé, že oba M-estimátory majú v príklade č. 1 oveľa lepšiu výpovednú schopnosť ako výberový priemer. Pri výpočte $\hat{\mu}_{os}$ sa berie do úvahy pomer odľahlých údajov, ktoré sú väčšie ako medián a menšie ako medián. V príklade č. 1 máme len odľahlé hodnoty väčšie ako medián, čo spôsobilo zväčšenie hodnoty $\hat{\mu}_{os}$. Pri výpočte $\hat{\mu}_{mom}$ sa vzťah medzi odľahlými údajmi, ktoré ležia vpravo a vľavo od mediánu, neberie do úvahy.

4 Porovnanie zastrihnutého priemeru a M-estimátorov

Jednokrokový M-estimátor aj modifikovaný jednokrokový M-estimátor majú oproti zastrihnutému priemeru niektoré výhody. Zastrihnuté priemery vyradia pri výpočte fixovaný podiel najmenších a najväčších hodnôt, M-estimátory a modifikované M-estimátory empiricky determinujú počet hodnôt, ktoré sa majú pri výpočte vyradiť. Okrem toho umožňujú vyradiť rozličný počet hodnôt z pravého a ľavého konca rozdelenia, prípadne nevyradiť žiadne hodnoty.

Na základe uvedených charakteristík je samozrejme možné realizovať aj induktívne úsudky o príslušných parametroch základného súboru. Možno napríklad počítať intervaly spoľahlivosti pre zastrihnutú strednú hodnotu, percentilové bootstrapové intervaly spoľahlivosti[5] založené na M-estimátoroch a pod.

5 Charakteristiky polohy v analýze miezd

Jedným z ukazovateľov hodnotenia životnej úrovne obyvateľstva a disparít medzi jednotlivými sektormi, resp. odvetviami národného hospodárstva, je priemerná mesačná mzda. Výpovedná schopnosť tohto ukazovateľa však býva často malá v dôsledku viacerých faktorov. Ide hlavne o časté výrazné zošikmenie rozdelenia a o prítomnosť odľahlých údajov.

V ([2], s. 44) sa uvádza, že aplikácia aritmetického priemeru v oblasti analýzy miezd nie je vhodná napríklad vtedy, keď rozdelenie je výrazne zošikmené (priemer je posunutý v smere zošikmenia a nie je dobrou charakteristikou centra rozdelenia), a ďalej, že aplikácia zastrihnutého priemeru a M-estimátorov je v tejto oblasti vhodná napríklad vtedy, keď súčasne chceme využiť čo najviac údajov a eliminovať vplyv odľahlých údajov (s. 46).

[5]Podrobnejšie o bootstrapových intervaloch spoľahlivosti pozri napr. v [5].

Možnosti aplikácie rozličných charakteristík polohy budeme ilustrovať na analýze mesačných miezd 718 666 zamestnancov všetkých veľkých firiem[6] v SR v 2. polroku 2009. Ide teda o výsledky vyčerpávajúceho zisťovania.[7]

V tabuľke č. 6 sú hodnoty charakteristík, vypočítané pomocou programového systému Statgraphics Centurion.[8] Priemerná mesačná mzda je 850,715 eur. Výrazne sa líši od mediánu, ktorý je 684,995 eur. Možno si všimnúť aj iné zaujímavé hodnoty charakteristík. Najnižšia mzda (Minimum) je napríklad 29,01 eur. Ide zrejme o mesačnú mzdu pracovníka, ktorý pracoval len na čiastočný úväzok. Zaujímavá je aj najvyššia mesačná mzda (Maximum) jedného pracovníka,[9] ktorá je 99 114,7 eur. Môže ísť napríklad o netypický príjem top manažéra, prípadne ide o zadanie chybného údaja spravodajskou jednotkou, prípadne chybu pri predbežnom spracovaní. Pri výpočte hodnoty výberového priemeru sa však tieto a podobné „neprimerane malé" a „neprimerane veľké" hodnoty brali do úvahy. Z toho je odvodená aj „obrovská" hodnota rozpätia miezd,[10] ktorá je 99 085,6 eur. Horný kvartil je 950,239, dolný kvartil je 510,4 eur, kvartilové rozpätie je 439,839 eur. Smerodajná odchýlka je 906,076 eur, variačný koeficient je 106,508 %. Rozdelenie miezd je výrazne zošikmené doprava – hodnota koeficienta šikmosti je 1 1452,2. Zdá sa, že aritmetický priemer nie je v tomto prípadne najlepšou charakteristikou „typickej" mesačnej mzdy pracovníka veľkej firmy.

Všimnime si teraz 5 % zastrihnutý priemer. Výsledky analýzy súboru miezd, v ktorom bolo 5 % najväčších a 5 % najmenších miezd „odstrihnutých" sú v tabuľke č. 7. Hodnota 5 % zastrihnutého priemeru je 754,075 eur, teda skoro o 100 eur menšia ako predtým. Výrazne menšia je aj smerodajná odchýlka (298,612 eur) a variačný koeficient (39,5997 %), rozpätie miezd je výrazne menšie (len 1 423,98 eur oproti 99 085,6 eur predtým), medián sa nezmenil, kvartilové rozpätie sa zmenilo len málo. Rozdelenie miezd má menšiu variabilitu a je menej zošikmené. Zdá sa, že toto rozdelenie lepšie charakterizuje celkovú situáciu v oblasti miezd zamestnancov veľkých firiem v skúmanom období. Podobne sa zdá, že 5 % zastrihnutý priemer je v tomto prípadne lepšou charakteristikou „typickej" mesačnej mzdy pracovníka veľkej firmy ako aritmetický priemer.

[6]Za veľké sa považujú firmy, ktoré majú 250 a viac zamestnancov.
[7]Údaje (anonymizované) poskytla firma Trexima Bratislava.
[8]Aj všetky ostatné výpočty, ktorých výsledky sú v tabuľkách č. 7 až 11 sme realizovali pomocou programového systému Statgraphics Centurion.
[9]Ide o priemernú mesačnú mzdu v 2. polroku 2009.
[10]Ide o rozdiel medzi najvyššou a najnižšou mzdou.

Tab. č. 6

Hodnoty charakteristík pre všetky údaje

Počet údajov	718 666
Aritmetický priemer	850,715
Medián	684,995
Modus	307,7
Smerodajná odchýlka	906,076
Variačný koeficient	106,508 %
Minimum	29,01
Maximum	99 114,7
Rozpätie	99 085,6
Dolný kvartil	510,4
Horný kvartil	950,239
Kvartilové rozpätie	439,839
Šikmosť	11 452,2

Tab. č. 7

Hodnoty charakteristík pre údaje bez 5 % najväčších a 5 % najmenších hodnôt

Počet údajov	646 800
Zastrihnutý priemer	754,075
Medián	684,995
Modus	370,0
Smerodajná odchýlka	298,612
Variačný koeficient	39,5997 %
Minimum	342,97
Maximum	1 766,95
Rozpätie	1 423,98
Dolný kvartil	528,44
Horný kvartil	912,7
Kvartilové rozpätie	384,26
Šikmosť	246,303

Teraz preskúmajme 10 % zastrihnutý priemer. Výsledky analýzy súboru miezd, v ktorom bolo 10 % najvyšších a 10 % najnižších miezd „odstrihnutých“ sú v tabuľke č. 8. Hodnota 10 % zastrihnutého priemeru je 730,454 eur, smerodajná odchýlka je už len 230,947 eur a variačný koeficient je 31,6168 %. Rozpätie miezd je len 949,39 eur, kvartilové rozpätie sa zmenilo minimálne, medián sa prakticky nezmenil. Zmenila sa šikmosť – z pozitívnej na negatívnu.

Tab. č. 8

Hodnoty charakteristík pre údaje bez 10 % najväčších a 10 % najmenších hodnôt

Počet údajov	57 4934
Zastrihnutý priemer	730,454
Medián	685,0
Modus	1 039,5
Smerodajná odchýlka	230,947
Variačný koeficient	31,6168%
Minimum	392,79
Maximum	1 342,18
Rozpätie	949,39
Dolný kvartil	546,08
Horný kvartil	880,01
Kvartilové rozpätie	333,93
Šikmosť	-1 335,38

Nakoniec si všimneme 20 % zastrihnutý priemer. Výsledky analýzy súboru miezd, v ktorom bolo 20 % najvyšších a 20 % najnižších miezd „odstrihnutých“ sú v tabuľke č. 9. Hodnota 20 % zastrihnutého priemeru je 706,747 eur a len málo sa líši od mediánu, ktorý sa nezmenil, je stále 685 eur. Smerodajná odchýlka poklesla na 151,162 eur a variačný koeficient na 21,3884 %. Rozpätie miezd pokleslo na 563,33 eur, výrazne pokleslo aj kvartilové rozpätie a hlavne šikmosť, ktorá je opäť pozitívna a má hodnotu len 58,6193.

Tab. č. 9

Hodnoty charakteristík pre údaje bez 20 % najväčších a 20 % najmenších hodnôt

Počet údajov	431 202
Zastrihnutý priemer	706,747
Medián	685,0
Modus	520,5
Smerodajná odchýlka	151,162
Variačný koeficient	21,3884 %
Minimum	474,25
Maximum	1 037,58
Rozpätie	563,33
Dolný kvartil	580,5
Horný kvartil	821,87
Kvartilové rozpätie	241,37
Šikmosť	58,6193

Zásadnou otázkou je, aký podiel údajov by sme mali „odstrihnúť“. Pri riešení mnohých praktických problémov je vhodné použiť 20 % výberový zastrihnutý priemer, ktorý často poskytuje lepší výsledok ako výberový priemer a medián ([5], s. 63).

Zdá sa, že aj v „našej“ analýze práve odstrihnutie 20 % údajov z každej strany rozdelenia poskytuje najkompaktnejší opis rozdelenia miezd. Podobne 20 % zastrihnutý priemer – 706,747 eur dobre charakterizuje „typickú mesačnú mzdu“ zamestnanca veľkej firmy v sledovanom období.

Všimnime si teraz M-estimátory. Pri výpočte M-estimátora sa použili hodnoty v tabuľke č. 10. Z ľavej strany rozdelenia sa odstránilo 15 495 a z pravej strany rozdelenia 131 472 odľahlých hodnôt.

Tab. č. 10

Hodnoty na výpočet M-estimátora

Medián	684,995
MAD	204,973
MADN	303,888
K	1,28
n_1	15 495
n_2	131 472
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	368842265,65

Po dosadení do vzťahu (6) dostaneme:

$$\hat{\mu}_{os} = \frac{1{,}28 \cdot 303{,}888 \cdot (131472 - 15495) + 368842265{,}65}{718666 - 15495 - 131472} = 724{,}08$$

Dostali sme ďalšiu charakteristiku, pomocou ktorej možno dobre charakterizovať „typickú“ mesačnú mzdu zamestnanca veľkej firmy v sledovanom období. Hodnota 724,08 eur sa len málo líši od 20 % zastrihnutého priemeru, ale výrazne od aritmetického priemeru, ktorý je až 850,715 eur. Zdá sa že hodnotu aritmetického priemeru výrazne ovplyvnil menší počet netypicky vysokých miezd, čím sa výrazne zmenšila jeho výpovedná schopnosť ako „typickej“ hodnoty.

Vypočítame ešte hodnotu modifikovaného M-estimátora. Pri jeho výpočte boli použité hodnoty v tabuľke č. 11.

Tab. č. 11

Hodnoty na výpočet modifikovaného M-estimátora
Po dosadení do vzťahu (8) dostaneme:

Medián	684,995
MAD	204,973
MADN	303,888
K	2,24
n_1	0
n_2	68 372
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	447970516,51

$$\hat{\mu}_{mom} = \frac{447970516,51}{718666 - 0 - 68372} = 688,874$$

Dostali sme ešte jednu charakteristiku, pomocou ktorej možno dobre charakterizovať „typickú“ mesačnú mzdu zamestnanca veľkej firmy v sledovanom období. Hodnota 688,874 eur sa tiež len málo líši od 20 % zastrihnutého priemeru a od M-estimátora, ale výrazne od aritmetického priemeru, ktorý sa rovná 850,715 eur.

Záver

Hodnota 20 % zastrihnutého priemeru je 706,747 eur, hodnota M-estimátora je 724,08 eur, hodnota modifikovaného M-estimátora je 688,874 eur a hodnota mediánu[11] je 684,995 eur. Hodnoty charakteristík sa len málo líšia a oscilujú okolo hodnoty 700 eur. Každá z týchto charakteristík určite lepšie charakterizuje typickú mesačnú mzdu pracovníka veľkej firmy v sledovanom období ako aritmetický priemer, ktorý sa rovná 850,715 eur a je evidentne výrazne ovplyvnený malým množstvom neobvykle vysokých miezd.

Nazdávame sa, že v oblasti analýz miezd podobne ako v mnohých iných oblastiach je užitočné počítať a interpretovať okrem základných charakteristík polohy, ako sú aritmetický priemer, medián, modus, kvartily, prípadne niektoré iné kvantily aj niektoré netradičné charakteristiky polohy, ktoré môžu značne obohatiť celkovú predstavu o situácii. Okrem toho je, samozrejme, užitočné počítať aj charakteristiky

[11]Keď uvažujeme o všetkých údajoch.

variability, šikmosti, prípadne iné charakteristiky, ktoré poskytnú iné pohľady na celkovú situáciu.

V „našej“ analýze zastrihnuté priemery a M-estimátory určite pomohli získať reálnejší a pravdivejší obraz o typickej mesačnej mzde pracovníka veľkej firmy v SR v prvom polroku 2009.

Literatúra

[1] BARNETT, V. – LEWIS, T. (1994): *Outliers in Statistical Data*. New York: Wiley and Sons. ISBN 978-0471930945.

[2] HALLEY, R. M. (2004): Measures of Central Tendency, Location, and Dispersion in Salary Survey Research. In: *Compensation and Benefits,* 2004/36, 39, s. 39 – 52.

[3] TEREK, M. (2008): Analýza odľahlých údajov. In: *Forum Statisticum Slovacum* 6/2008, s. 152 – 157. ISSN 1336-7420.

[4] TEREK, M. (2010): Analýza odľahlých údajov a niektoré charakteristiky polohy. In: *Zborník príspevkov z Medzinárodnej vedeckej konferencie Znalostná ekonomika a jej odraz v ekonomickej teórii a hospodárskej praxi.* Bratislava: EU, CD ROM, ISBN 978-80-225-3076-7, s. 371 – 384.

[5] WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0.

[6] STN ISO 3534-1. *Štatistika. Slovník a značky. Časť 1: Všeobecné štatistické termíny a termíny používané v teórii pravdepodobnosti.* Bratislava: Slovenský ústav technickej normalizácie 2008.