

**EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY**

Evidenčné číslo: 17200/D/2011/0005203382

**ANALÝZA VÝBRANÝCH SOCIÁLNO-EKONOMICKÝCH
UKAZOVATEĽOV S OHĽADOM NA ODLAHLÉ ÚDAJE**

Dizertačná práca

2011

Ing. He Nguyen Dinh

**EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY**

**ANALÝZA VÝBRANÝCH SOCIÁLNO-EKONOMICKÝCH
UKAZOVATEĽOV S OHĽADOM NA ODLAHLÉ ÚDAJE**

Dizertačná práca

Študijný program: Ekonometria a operačný výskum

Študijný odbor: 3.3.25 Ekonometria a operačný výskum

Školiace pracovisko: Katedra operačného výskumu a ekonometrie

Školiteľ: Prof. Ing. Milan Terek, PhD.

Bratislava 2011

Ing. He Nguyen Dinh

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a že som uviedol všetku použitú literatúru.

Dátum:

.....

(podpis študenta)

Pod'akovanie

Ďakujem môjmu školiťovi, Prof. Ing. Milan Terek, PhD. za cenné rady, čas, pripomienky a odborný dohľad počas celého obdobia vypracovávania dizertačnej práce.

ABSTRAKT

NGUYEN DINH, He: *Analýza vybraných sociálno-ekonomických ukazovateľov ohľadom na odlahlé údaje*. – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra operačného výskumu a ekonometrie. – prof. Ing. Milan Terek, PhD. – Bratislava: FHI EU, 2011, počet strán 113. Práca je rozdelená do 6 kapitol, obsahuje 12 obrázkov, 21 tabuliek a 3 príloh.

Vyhľadávanie odlahlých hodnôt (Outliers) sa v súčasnosti stáva jedným z dôležitých procesov pri analýze dát. Stretávame sa s ním v mnohých úlohách, ktoré sa týkajú analýzy dát veľkého počtu premenných. Detekcia odlahlých hodnôt je jedným z prvých krokov na získanie konzistentnej analýzy. Hoci sú odlahlé hodnoty často považované za chyby alebo šum, môžu niesť dôležité informácie. Na druhej strane by mohli nepriaznivo ovplyvniť parametre modelu ich skreslením, znížením sily testu, čo vedie k nesprávnym výsledkom analýz.

Cieľom dizertačnej práce je priniesť všeobecný prehľad o odlahlých údajoch a aktualizovať nové prístupy, trendy a metódy na ich detekciu. Nové a aktuálne prístupy a metódy budeme aplikovať pri analýze miezd. Ich charakteristiky sú veľmi dôležité sociálno-ekonomické ukazovatele.

Tradične sa napríklad sleduje ukazovateľ Priemerná hrubá nominálna mesačná mzda, Priemerná hrubá hodinová mzda a iné. Niektoré ukazovatele sa formulujú s ohľadom na odlahlé údaje. Nazveme ich ukazovatele s ohľadom na odlahlé údaje. Z hľadiska štatistiky ide o netradičné charakteristiky polohy rozdelenia. Konkrétne budeme analyzovať možnosti zastreihnutých priemerov, winsorizovaného priemeru a M-estimátorov.

Kľúčové slová: Odlahlé údaje, detekcia odlahlých hodnôt, bootstrapová metóda, analýza miezd, hrubá mesačná mzda.

ABSTRACT

NGUYEN DINH, He: *Modeling the import of goods and services of the Slovak Republic*. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Operations Research and Econometrics. – prof. Ing. Milan Terek, PhD. . – Bratislava: FEI UE, 2011, no. of pages 113. Thesis is divided into 6 chapters, includes 12 figures, 21 tables and 3 annexes.

In the statistical analysis we always have to face with outliers in the observed data. Outlier detection is a primary step in many data-mining applications. In presence of outliers, special attention should be taken to assure the robustness of the used estimators. Outlier detection for data mining is often based on distance measures, clustering and spatial methods.

This doctoral thesis aims to introduce how to detect outliers in observed data and the ways to deal with outliers. We present several methods for outlier detection, while distinguishing parametric vs. nonparametric procedures. We try to apply new approaches and actual methods in the wage's analysis, which has very important characteristics in the social-economics aspect.

We focus to observe average nominal monthly brutto wages. Some indicators are designed with the regards to outliers. In statistic these local characteristics of distribution are non-traditional. Computing of one-step M-Estimator and modified one-step M-Estimator of location are described. A comparison of the trimmed mean and M-estimators of location is given. Finally, the application of some non-traditional measures of location in analysis of wages is given.

Key words: outliers, trimmed mean, detection, M-Estimator, analysis of wages, bootstrap, monthly brutto wages.

OBSAH

ÚVOD.....	10
1 Súčasný stav riešenej problematiky doma a v zahraničí.....	13
1.1 Odlahlé údaje.....	13
1.2 Tradičný prístup k analýze mzdy na Slovensku.....	16
2 Ciele práce	20
3 Detekcia a eliminácia odlahlých hodnôt.....	222
3.1 Detekcia odlahlých údajov	22
3.1.1 Box-plot	22
3.1.2 Dixonov test na detekciu extrémov.....	22
3.1.3 Používanie Z hodnoty na detekciu odlahlých hodnôt.....	24
3.1.4 Modifikovaná metóda Z hodnoty.....	25
3.1.5 Grubbsov test na detekciu odlahlých hodnôt.....	29
3.2 Odlahlé údaje a spôsob ich eliminácie	30
3.2.1 Eliminácia odlahlých hodnôt metódou <i>zastrihnutého priemeru (Trimmed mean)</i>	31
3.2.2 Eliminácia odlahlých hodnôt metódou <i>jednostranného zastrihnutého priemeru</i>	31
3.2.3 Eliminácia odlahlých hodnôt metódou Winsorizovaného priemeru (Winsorized mean)	33
3.2.4 Eliminácia odlahlých hodnôt metódou jednostranného Winsorizovaného priemeru (One side Winsorized mean).....	33
4 Nové netradičné charakteristiky a metódy induktívnej štatistiky	34
4.1 Charakteristiky polohy	34
4.1.1 M-estimátory.....	34
4.1.2 Výpočet M-estimátora polohy	36

4.1.3	Modifikovaný jednokrokový M-estimátor a jeho výpočet.....	38
4.1.4	Výberový priemer, medián, zastrihnutý priemer a jednokrokové M-estimátory	39
4.2	Intervaly spoľahlivosti	40
4.2.1	Intervaly spoľahlivosti pre zastrihnutú strednú hodnotu.....	40
4.2.2	Intervaly spoľahlivosti pre zastrihnutý rozptyl σ^2 normálneho rozdelenia	41
4.3	Testovanie štatistických hypotéz	42
4.3.1	Testovanie štatistických hypotéz o zastrihnutej strednej hodnote	42
4.3.2	Testovanie štatistických hypotéz o zastrihnutom rozptyle.....	44
4.4	Bootstrapová metóda	45
4.4.1	Základné pojmy a charakteristiky	45
4.4.2	Percentilová metóda.....	51
4.4.3	Bootstrapové odhadovanie rozptylu estimátorov percentilovou metódou	53
4.4.4	Bootstrap-t interval	53
4.4.5	Symetrické intervaly spoľahlivosti	56
4.4.6	Percentilové bootstrapové intervaly spoľahlivosti pre výberovú strednú hodnotu	57
4.4.7	Percentilové bootstrapové intervaly spoľahlivosti založené na M-estimátoroch.....	59
5	Výsledky práce	62
5.1	Charakteristika sledovaného ukazovateľa.....	62
5.1.1	Podstata a formy mzdy.....	62
5.1.2	Charakteristika výberovej vzorky ISPZ	63
5.1.3	Opisné charakteristiky pre hmes-mzda	65
5.1.4	Detekcia odľahlých údajov pre hmes-mzda	68
5.1.5	Netradičné charakteristiky polohy v analýze miezd	72
5.2	Analýza miezd vo veľkých firmách.....	78
5.2.1	Opisné charakteristiky pre hmes-mzda vo veľkých firmách.....	78

5.2.2	Detekcia odľahlých údajov pre hmes-mzda vo veľkých firmách	81
5.2.3	Netradičné charakteristiky polohy v analýze miezd	83
5.2.4	Bootstrapový interval spoľahlivosti pre hmes-mzda.....	89
5.2.5	Aplikácia bootstrapovej metódy na výpočet intervalu spoľahlivosti pre zastrihnutú strednú hodnotu a charakteristiku polohy, založenú na M-estimátore.....	90
6	Prínos dizertačnej práce.....	97
	ZÁVER.....	97
	LITERATÚRA.....	105
	PRÍLOHY	111

ÚVOD

V posledných rokoch tohto storočia sa svetová ekonomika dostala do novej etapy, v ktorej je rýchlosť regulácie a prispôsobenia sa na zmeny trhu prvotnou úlohou. Vďaka revolúcii v technológii môžeme neustále zvyšovať kvalitu života, ale musíme žiť s vyššou rýchlosťou a s vyšším tlakom konkurencie. Informácia a kvalita informácie sa stávajú nevyhnutným faktorom úspešného života.

Informáciu potrebujeme pri každej analýze, pri každom rozhodnutí, pri každej činnosti a jej kvalita je kľúčom k úspechu. V dnešnom období môžeme mať hneď k dispozícii dôležitú informáciu vďaka internetu a rozhodujúcim faktorom je prístup k informácii a spôsob analýzy.

Pod každým vysokým vrcholom je pevný základ. Proces spracovania vzorky a charakterizovania základných vlastností je základom štatistickej analýzy. V takom prípade sa využívajú indukzívne štatistické metódy. Jedným veľkým problémom v tomto procese sú extrémne hodnoty (*môže ísť o odľahlé hodnoty*), ktoré sú všeobecne definované ako pozorované hodnoty, ktoré sú neobvykle oddelené od ostatných hodnôt v súbore. Odľahlé hodnoty môžu meniť hodnoty vypočítanej štatistiky, dávať skreslené informácie, napríklad zväčšený odhad rozptylu, prípadne spôsobujú, že priemer nie je dobrou charakteristikou stredu rozdelenia.

Vyhľadávanie odľahlých hodnôt (Outliers) sa v súčasnosti stáva jedným z dôležitých procesov pri analýze dát. Stretávame sa s ním v mnohých úlohách, ktoré sa týkajú analýzy dát veľkého počtu premenných. Detekcia odľahlých hodnôt je jedným z prvých krokov na získanie konzistentnej analýzy.

Hoci sú odľahlé hodnoty často považované za chyby alebo šum, môžu niesť dôležité informácie. Na druhej strane by mohli nepriaznivo ovplyvniť parametre modelu ich skreslením, znížením sily testu, čo vedie k nesprávnym výsledkom analýz. Pred modelovaním a vlastnou analýzou je preto dôležitá ich identifikácia.

Detekcia a skúmanie odľahlých údajov majú veľké množstvo aplikácií, ako sú podvody s kreditnými kartami, zistenia daňových podvodov, klinické testy, nezrovnalosti vo volebných analýzach, čistenie dát, narušenia siete, predpovede počasia, geografické informačné systémy, analýzy športových výsledkov a ďalšie úlohy pre hĺbkovú analýzu dát, v analýze časových radov, v regresnej analýze atď.

Ďalší postup, po detekcii odľahlých hodnôt je hľadanie spôsobu ich odstránenia, aplikovania, alebo vytvorenia špeciálneho priestoru na ich skúmanie.

Cieľom dizertačnej práce je priniesť všeobecný prehľad o odľahlých údajoch a aktualizovať nové prístupy, trendy a metódy na ich detekciu. Nové a aktuálne prístupy a metódy budeme aplikovať pri analýze miezd. Ich charakteristiky sú veľmi dôležité sociálno-ekonomické ukazovatele.

Tradične sa napríklad sleduje ukazovateľ Priemerná hrubá nominálna mesačná mzda, Priemerná hrubá hodinová mzda a iné. Niektoré ukazovatele sa formulujú s ohľadom na odľahlé údaje. Nazveme ich ukazovatele s ohľadom na odľahlé údaje. Z hľadiska štatistiky ide o netradičné charakteristiky polohy rozdelenia. Konkrétne budeme analyzovať možnosti zastrihnutých priemerov, winsorizovaného priemeru a M-estimátorov.

Predkladaná dizertačná práca je rozdelená do piatich kapitol. V prvej kapitole je stručne a všeobecne popísaný prehľad o súčasnom stave a trende k odľahlým údajom v zahraničí a na Slovensku.

V druhej kapitole je uvedený hlavný cieľ a čiastkové ciele dizertačnej práce.

Obsahom tretej kapitoly je teoretické spracovanie aktuálnych metód a prístupov na detekciu odľahlých údajov od jednoduchých, založených na grafe *box-plot* alebo kvartilovom rozpätí, Z-hodnoty až po moderné parametrické a neparametrické prístupy ako *Dixonov test* alebo *Grubbsov test*. Vysvetlené sú pojmy a teoretické základy, ktoré je nutné poznať a ovládať pre správne použitie jednotlivých metód, ako napr. pojem *mediánová absolútna odchýlka* (*median absolute deviation*) – MAD, *zastrihnutý priemer*, *winsorizovaný priemer* a iné.

Štvrtá kapitola obsahuje nové metódy a prístupy na výpočet netradičných charakteristík polohy a indukčnej štatistiky. V tejto kapitole sa venujeme výpočtom M-estimátora a aplikácií bootstrapovej metódy v indukčnej štatistike pre základný súbor a aj pre zastrehaný súbor.

Posledná kapitola je piata kapitola, v ktorej aplikujeme teoretické poznatky z predchádzajúcich kapitol na analýzu mesačnej mzdy na Slovensku. Pri tejto analýze použijeme štvrťročné údaje o cene práce "ISCP (MPSVR SR) 1-04 zaradené do Programu štátnych štatistických zisťovaní na roky 2009-2011 vyhláškou Štatistického úradu SR č. 416/2008 Z. z., čiastka 152, str. 3487. Analýzu mzdy uskutočníme pomocou softwaru Statgraphics Centurion XV, Microsoft Office Excel a S-plus 8.

1 Súčasný stav riešenej problematiky doma a v zahraničí

1.1 Odlahlé údaje

Odlahlé hodnoty sú všeobecne definované ako pozorované hodnoty, ktoré sú neobvykle oddelené od ostatných hodnôt v súbore¹.

Odlahlé hodnoty môžu meniť hodnoty vypočítanej štatistiky, napríklad zvýšenie odhadu rozptylu, prípadne spôsobujú, že priemer nie je dobrou charakteristikou stredu rozdelenia.

Testy odlahlých údajov odpovedajú na základe niekoľkých jednoduchých predpokladov na otázku, či je pozorovaná hodnota odlahlou hodnotou.

Na začiatku boli odlahlé hodnoty spojené s terminológiami ako *nereprezentatívne* (*unrepresentative*), *nekontrolované* (*rogue*) a *odlahlé* (*outlying*) pozorovania v súbore. Tieto hodnoty sú často spomínané ako kontaminovanie dát, ktoré znížili a skreslili informáciu zdroja údajov. Je to prirodzené, keď hľadáme význam interpretácie a zoskupenia odlahlých hodnôt alebo metódy na ich manipuláciu, ktoré nám umožnia odlahlé hodnoty odmietat', aby sme mohli obnoviť vlastnosti údajov².

Treba však poznamenať, že odlahlé údaje (*outliers*) nemusia byť vždy mätúce, zlé alebo chybné. Naopak, existujú situácie, v ktorých sú odlahlé 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 odlahlých údajov³.

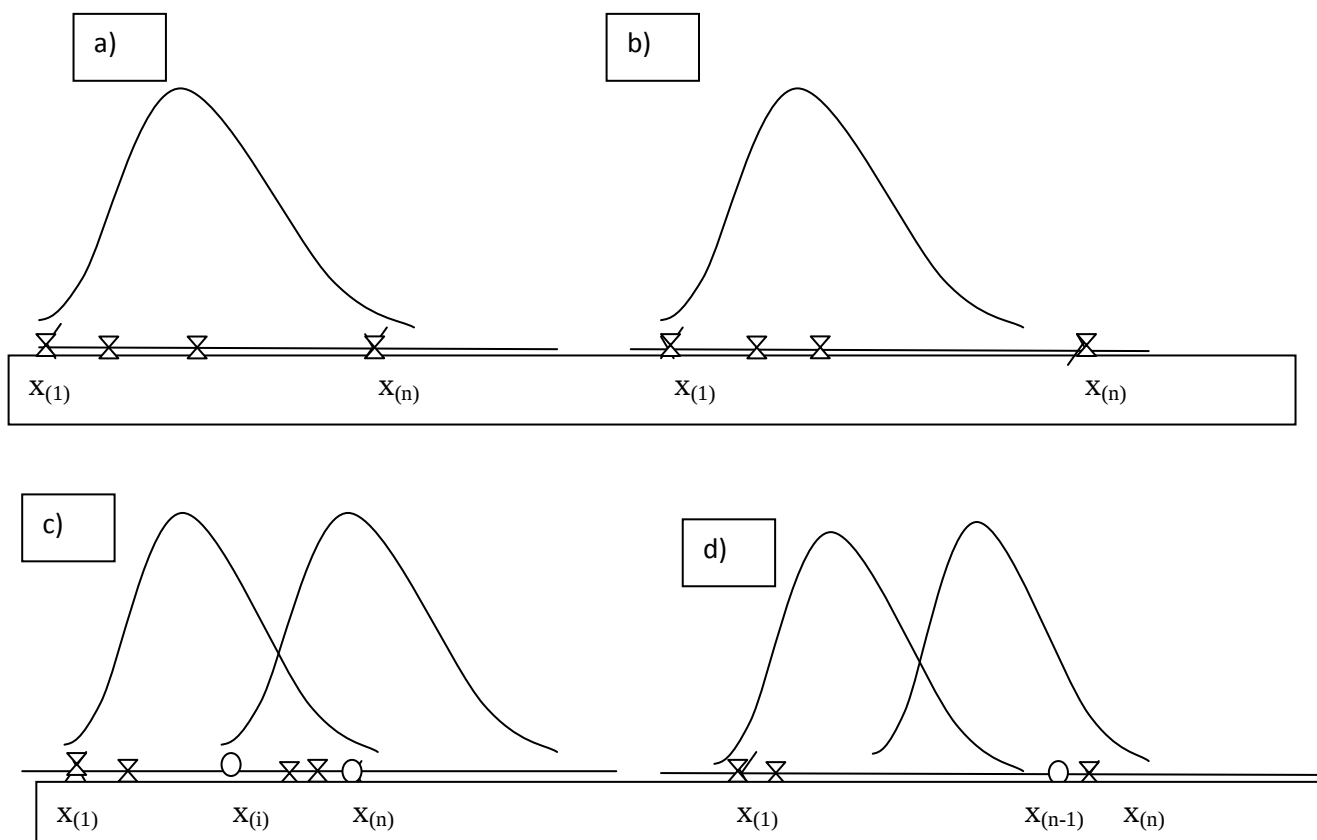
Pre správne pochopenie pojmu odlahlých údajov potrebujeme rozlišovať extrémne hodnoty, odlahlé hodnoty a kontaminanty.

¹ <http://www.charuaggarwal.net/outl.pdf>

² BARNETT, V. – LEWIS, T.: *Outliers in Statistical data*. UK: John Wiley & Sons, 1998, s. 3

³ Terek, M. – Nguyen Dinh He: Možnosti využitia niektorých netradičných charakteristík v analýze miezd. *Ekonomické rozhlady* 1/2011.

Nech $x_1, x_2, \dots, x_n$ sú hodnoty náhodného výberu s rozsahom n , ktoré majú distribučnú funkciu F . Usporiadame ich od najmenšej po najväčšiu. Dostaneme $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, kde $x_{(i)}$ je poriadková hodnota a ich zobrazenia v rôznych prípadoch nasledovné:



Obr.1.1 Definovanie extrémnych hodnôt, odľahlých hodnôt a kontaminantov

(Zdroj: vlastné spracovanie)

Na obr.1.1(a) hodnoty $x_{(1)}$ a $x_{(n)}$ sú definované ako extrémne hodnoty, ale nie sú odľahlé hodnoty. Na ďalšom obr.1.1(b) $x_{(n)}$ je horná odľahlá hodnota a $x_{(1)}$ môžeme považovať za dolnú odľahlú hodnotu. Takže, extrémne hodnoty môžu byť alebo nemusia byť odľahlé hodnoty, ale odľahlé hodnoty sú vždy extrémne hodnoty⁴. Obr.1.1(c) ukazuje hodnoty, ktoré sú kontaminanty a zároveň sú extrémne hodnoty, ale nie sú odľahlé. Naopak ako na obr.1.1(d) vidíme hodnoty, ktoré nie sú extrémne kontaminanty, ale sú odľahlé hodnoty.

⁴ BARNETT, V. – LEWIS, T.: *Outliers in Statistical data*. UK: John Wiley & Sons, 1998, s. 9

Prvý významný autor, ktorý svojou prácou položil prvý aspekt štatistickej metódy pri analýze odľahlých údajov bol Pierce (1852).⁵

Neskôr Calvin (1949) vo svojej učebnici chémie odporúča čitateľom používať Chauvenetovu metódu na zamietnutie odľahlých údajov, ale nebolo to efektívne.

Niektoré základné charakteristiky odľahlých údajov boli ako ilustračné príklady v anglickom zákone o manželstve, ktoré sú spojené s prácou Halduma v roku 1949. Haldum v tejto práci skúmal žiadosti o rozvod, ktoré sú podané hneď po svadbe.⁶

Presná definícia odľahlých hodnôt často závisí od predpokladov, týkajúcich sa dátovej štruktúry a použitej detekčnej metódy. Napriek tomu niektoré definície sú považované za všeobecné. Hawkins (Hawkins, 1980) definoval odľahlé hodnoty ako hodnoty, ktoré sa odkláňajú od iných hodnôt s cieľom vzbudiť podozrenie, že to bolo generované iným mechanizmom.

Barnet a Lewis (Barnett a Lewis, 1994) uvádzajú, že odľahlá hodnota je tá, ktorá sa výrazne odlišuje od ostatných hodnôt súboru, v ktorom sa nachádza. Podobne, Johnson (Johnson, 1992) definoval odľahlé hodnoty v dátovom súbore ako hodnoty, ktoré sa zdajú byť nekonzistentné so zvyškom hodnôt tohto súboru.

Členenie metód detekcie odľahlých údajov môže byť rozdelené v závislosti od toho, či sa aplikujú v jednorozmerných alebo viacrozmerných štatistických metódach. Základné detekčné metódy sa delia na parametrické (štatistické) metódy a neparametrické metódy.

Štatistické parametrické metódy sú založené na známom rozdelení pravdepodobnosti základného súboru (napr. Hawkins, 1980; Rousseeuw a Leora, 1987; Barnett a Lewis, 1994), alebo aspoň sú založené na štatistických odhadoch parametrov neznámeho rozdelenia (Hadi, 1992; Caussinus a Roiz, 1990).

Tieto metódy považujú odľahlé hodnoty za hodnoty, ktoré sa odchyľujú od daných predpokladov. Také metódy sú často nevhodné pre vysoko-dimenziálne dáta a pre ľubovoľné dáta bez predchádzajúcich znalostí o základných dátových rozdeleniach (Papadimitriou et al., 2002).

Neparametrické metódy detekcie odľahlých hodnôt sú založené na vzdialenostiach. Tieto metódy sú zvyčajne schopné manipulácie rozsiahlych databáz (Knorr a Ng, 1997; Knorr a Ng,

⁵ Tamže

⁶ BARNETT, V. – LEWIS, T.: *Outliers in Statistical data*. UK: John Wiley & Sons, 1998, s.4

1998; Fawcett a Provost, 1997, Williams a Huang, 1997; Mouchel a Schönlau, 1998; Knorr et al., 2000; Knorr et al., 2001; Jin et al., 2001; Breunig et al., 2000; Williams et al., 2002; Hawkins et al., 2002; Bay a Schwabacher, 2003). Iná trieda metód detekcie odlahlých hodnôt je založená na zoskupovacích technikách, kde môže byť zhuk malých rozmerov považovaný za zhuk odlahlých hodnôt (Kaufman a Rousseeuw, 1990; a Ng Han, 1994; Ramaswamy et al., 2000; Barbara a Chen, 2000; Shekhar a Chawla, 2002; Shekhar a Lu, 2001; Shekhara Lu, 2002; Acuna a Rodriguez, 2004).

Hu a Sung (Hu a Sung, 2003) navrhli metódu na identifikáciu odlahlých hodnôt s vysokou i nízkou hustotou súboru.

V dnešnom období, skúmanie odlahlých údajov začína byť populárne. Problém odlahlých údajov je veľmi aktuálna téma v každej oblasti. V súčasnosti sa sledujú a skúmajú odlahlé hodnoty v ekonomike, chémii, technike a medicíne. Problém odlahlých údajov nie je len pre jednorozmerné údaje, ale skúma sa aj pre viacrozmerné údaje a štruktúrované dáta.

Na Slovensku zatiaľ skúmanie odlahlých údajov nemá širšie uplatnenie, o čom svedčí aj minimum publikovanej literatúry v slovenskom jazyku, vrátane publikácií v českom jazyku. Komplexné a štruktúrované spracovanie problematiky ešte nebolo publikované. Na Slovensku sú zatiaľ len publikácie, v ktorých sa odlahlým údajom venujú len v určitej kapitole.

Slovenskí štatistickí, ktorí sa hlbšie venujú problematike odlahlých údajov sú Terek⁷, ktorý sa zaoberá novými prístupmi a metódami zisťovania odlahlých údajov a ich aplikáciou pri analýze ekonomických ukazovateľov, ďalej Horníková⁸ a Labudová⁹, ktoré sa zaoberajú problematikou hĺbkovej analýzy údajov.

1.2 Tradičný prístup k analýze mzdy na Slovensku

⁷ Pozri http://www.fhi.sk/sk/katedry/zamestnanci/prof.-ing.-milan-terek_-phd..html

⁸ Pozri http://www.fhi.sk/sk/katedry-fakulty/zamestnanci/ing.-mgr.-adriana-hornikova_-phd..html

⁹ Pozri <http://www.fhi.sk/sk/katedry/zamestnanci/rndr.-viera-labudova.html>

V tejto dizertačnej práci aplikujeme teórie odľahlých údajov a ich prístupov pri analýze hrubej mesačnej mzdy na Slovensku. *TREXIMA* je firma, ktorá sa zaoberá analýzou mzdy na Slovensku podľa projektu ISCP¹⁰.

Spoločnosť *TREXIMA* je špecializovanou výskumno-štatistickou a poradensko-konzultačnou organizáciou, ktorá disponuje skúsenými profesionálmi v oblastiach výberových štatistických zisťovaní pracovných pozícií, trhu práce, ceny a nákladov práce, klasifikácie zamestnaní, riadenia ľudských zdrojov, analytického hodnotenia pracovných miest, štandardizácie pracovných činností, racionalizácie práce, v oblasti hĺbkových retrospektívnych analýz, prognózovania vývoja miezd, zamestnanosti a produktivity. Trexima od roku 1992 sleduje osvedčené metódy a poznatky v oblasti štatistiky miezd v jednotlivých štátoch Európy a z celého sveta. Systematicky využíva pozitívne domáce i zahraničné skúsenosti odborníkov z oblasti odmeňovania, štatistiky a ekonomiky hodnotenia práce¹¹.

Odborníci z Treximy sú v kontakte s 4000 spravodajskými jednotkami z celej SR. Systematická spolupráca a spätná väzba z týchto partnerských organizácií významne zvyšuje kvalitu i odbornosť výsledkov a výstupov z jednotlivých celoštátnych štatistických zisťovaní a je kľúčom úspešnosti jednotlivých informačných systémov.

Cieľom spoločnosti je poskytovať svojim zákazníkom kvalitné a komplexné služby na vysokej profesionálnej úrovni. Trexima je držiteľom Certifikátu systému manažmentu kvality, ktorý bol schválený spoločnosťou Lloyd's Register Quality Assurance podľa noriem ISO 9001:2000 a je zameraný na výskum miezd, sledovanie nákladov práce a pracovných podmienok vo vybraných organizáciách SR, analýzu štatistických údajov a poskytovanie výsledných produktov štátnym orgánom, ako aj na výskumnú a poradenskú činnosť v oblasti manažmentu ľudských zdrojov.

Od roku 1999 je Trexima zaregistrovaná ako výskumno-poradenská špecializovaná organizácia v centrálnom registri Európskej komisie pod evidenčným číslom 100 436.

¹⁰ Informačný systém o cene práce

¹¹ Pozri <http://www.trexima.sk/det.asp?id=1>

Informačný systém o cene práce (ISCP), ktorého gestorom je od roku 1992 Ministerstvo práce, sociálnych vecí a rodiny SR, zahŕňa aj pravidelné longitudinálne štvrťročné zisťovanie priemerných hodinových a mesačných zárobkov zamestnancov podľa jednotlivých zamestnaní.

Dôležitou prednosťou ISCP je minimalizácia prácnosti na strane respondenta (zamestnávateľskej organizácie), získavanie súboru vstupných dát modernou elektronickou formou, čo je zárukou kvalitných a dôveryhodných výsledkov¹².

Priemerná mesačná mzda sa vypočíta spriemerovaním celkovej mesačnej mzdy účtovanej na mesačnej báze a jej vydelením priemernou dennou úrovňou zamestnanosti v štvrťroku.

Okrem údajov o priemerných hrubých nominálnych mesačných mzdách sa publikujú aj indexy reálnej mzdy oproti rovnakému štvrťroku predchádzajúceho roka (reálne mzdy sú nominálne mzdy deflované indexom spotrebiteľských cien za rovnaké obdobie).

Dosiahnutá priemerná hrubá mesačná mzda vo 4. štvrťroku 2010 bola 821,07 EUR a jej medziročný nárast bol 3,09 %. Po odpočítaní poistného na zdravotné, sociálne poistenie a preddavkov na daň z príjmov fyzických osôb z priemernej hrubej mesačnej mzdy, predstavovala priemerná čistá mesačná mzda 650,31 EUR a jej podiel z hrubej mzdy bol 79,20 %.

Tabuľka 1.1 Výstup opisnej štatistiky hrubej mesačnej mzdy

Ukazovateľ	Euro	
	4. Q. 2009	4. Q. 2010
Priemerná hrubá mesačná mzda v hospodárstve	796,40	821,07
1. Decíl	333,33	343,42
1. Kvartil	444,74	456,56
Medián	625,35	644,01
3. Kvartil	861,94	892,7
9. Decíl	1241,22	1289,68

(zdroj: výstup ISCP 104 –Trexima)

Použité opisné štatistiky a pojmy pri tejto analýze:

¹² Výstup ISCP 104 – Trexima, Bratislava, 2010

- Prvý decil je taká hodnota, že 10% hodnôt sledovaného znaku (hodinového zárobku alebo priemernej mesačnej mzdy) je menších alebo sa rovná prvému decilu.
- Deviaty decil je taká hodnota, že 90% hodnôt sledovaného znaku (hodinového zárobku alebo priemernej mesačnej mzdy) je menších alebo sa rovná deviatemu decilu.
- Medián je hodnota uprostred vzostupne usporiadanej rady hodnôt sledovaného znaku. Polovica hodnôt je menších alebo sa rovná mediánu.
- Priemer je súčet nameraných hodnôt delený ich počtom (aritmetický priemer)
- Prvý kvartil je taká hodnota, že 25% hodnôt sledovaného znaku (hodinového zárobku alebo priemernej mesačnej mzdy) je menších alebo sa rovná prvému kvartilu.
- Tretí kvartil je taká hodnota, že 75% hodnôt sledovaného znaku (hodinového zárobku alebo priemernej mesačnej mzdy) je menších alebo sa rovná tretiemu kvartilu.

Rozpätie medzi najvyššími a najnižšími zárobkami, ktoré je možné vyčísliť podielom 9. a 1. decilu. je v sledovanom období 3,919. Podiel mediánu a priemeru je 0,750, čo znamená, že polovica všetkých zamestnancov poberá mzdu nižšiu ako 75% priemernej mzdy v podnikateľskej sfére.

2 Ciele práce

Hlavným cieľom dizertačnej práce je priniesť poznatky o odlahlých údajoch, nové prístupy na ich zisťovanie a aplikácia týchto prístupov. Ciele dizertačnej práce je možné zhrnúť do nasledujúcich bodov:

➤ **Vytvorenie uceleného systém poznatkov detekcie odlahlých údajov ako prvý základ na analýzu odlahlých údajov**

Na Slovensku sa zatiaľ venuje veľmi malá pozornosť skúmaniu odlahlých údajov a v slovenskej literatúre absentuje publikácia, ktorá by sa komplexne a systematicky venovala tejto problematike. Existuje len zopár publikácií, ktoré sa problematike odlahlých údajov venujú, ale to len čiastočne a okrajovo, lebo odlahlé údaje sú väčšinou iba stručne spomenuté. Predkladaná dizertačná práca sa pokúsi čiastočne zaplniť tento prázdny priestor v publikáciách.

➤ **Definovanie pojmov a vytvorenie štruktúry symboliky**

Skoro všetky predchádzajúce publikácie boli napísané len čiastočne s rôznym zameraním. Problematika odlahlých údajov bola vždy spomenutá ako vedľajšia vec, ktorá vznikla kvôli súvislostiam k hlavnej úlohe. Je veľmi málo publikácií zameraných na štatistické prístupy, v ktorých je teoretická časť doplnená praktickou aplikáciou alebo ilustračným príkladom. Vzhľadom na túto situáciu, najmä, že autori pristupujú k danej problematike z rôznych oblastí a z rôznych hľadísk, potom vznikli rôzne definície a odlišné symboliky. Táto dizertačná práca má za cieľ do určitej miery vytvoriť systémovo teoretické poznatky, definície a zjednotenie symboliky.

➤ **Prínos nových poznatkov, prístupov a metód na spracovanie odľahlých údajov a popísanie metód s využitím výpočtovej techniky**

Vo väčšine publikácií je problematika odľahlých údajov len stručne popísaná a málo z autorov ukazuje, ako sa dajú aplikovať spomenuté poznatky na riešenie problému odľahlých údajov. V tejto práci uvidíme detailne a podrobne nové prístupy a metódy na analýzu odľahlých údajov a ich aplikáciu v danej problematike. V 3. a 4. kapitole sú podrobne vysvetlené pojmy, vzorce, definície a ilustračné príklady. V praktickej časti dizertačnej práce bol využívaný štatistický balík *Statgraphics Centurion*. Jednoduché mechanické výpočty boli realizované v tabuľkovom procesore *Microsoft Office Excel* a špeciálne výpočty boli spracované pomocou štatistického programu *S-Plus 8*.

➤ **Aplikácia poznatkov a metód pri analýze hrubej mesačnej mzdy**

V dizertačnej práci budeme analyzovať hrubú mesačnú mzdu podľa klasickej metódy, ktorá nám umožňuje získať základné informácie opisnej štatistiky. Ďalej budeme aplikovať nové poznatky na detekciu odľahlých údajov pre daný súbor, aby sme zistili prítomnosť odľahlých hodnôt. Pomocou testu normality overíme predpoklad o normalite rozdelenia pozorovaného znaku.

Skúmame údaje, ktoré sú získané len z veľkých firiem, aby sme mohli aplikovať nové metódy a prístupy na skúmanie rozličných charakteristík polohy budeme ilustrovať na analýze mesačných miezd 718 666 zamestnancov všetkých veľkých firiem v SR v 2. polroku 2009. Ide teda o výsledky vyčerpávajúceho zisťovania. Pomocou nových metód a prístupov zistíme odľahlé hodnoty v pozorovanom súbore cez procedúru *Outlier Identification*.

Ďalej vypočítame netradičné charakteristiky polohy pre mzdu, ako sú zastrihnutý priemer, M-estimátor a modifikovaný M-estimátor, pomocou *Statgraphics Centurion* a *Microsoft Office Excel* a porovnáme získané výsledky.

Použijeme bootstrap metódy z indukčnej štatistiky. Aplikujeme bootstrap metódu na zistenie intervalu spoľahlivosti pre strednú hodnotu, zastrihnutú strednú hodnotu a intervalu spoľahlivosti charakteristiky polohy, založenej na M-estimátore a intervalu spoľahlivosti pre rozptyl.

3 Detekcia a eliminácia odľahlých hodnôt

3.1 Detekcia odľahlých údajov

3.1.1 Box-plot

Box-plot je grafické znázornenie rozdelenia dát. Graf zobrazuje dolný kvartil (Q_1) a horný kvartil (Q_3) spolu s mediánom. Medián je 50. percentil. Dolný kvartil je 25. percentil, a horný kvartil je 75. percentil. Horné a dolné hranice (lower and upper fences) sa zvyčajne nastavujú na pevnú vzdialenosť od kvartilového rozpätia $R_Q^4 = (Q_3 - Q_1)$.

Tukey (1997) navrhoval metódu zisťovania odľahlých údajov pomocou „hranice“. „Hranice“ nastavíme nasledovne:

- Dolná vnútorná hranica: $Q_1 - 1,5R_Q^4$
- Horná vnútorná hranica: $Q_3 + 1,5R_Q^4$
- Dolná vonkajšia hranica: $Q_1 - 3R_Q^4$
- Horná vonkajšia hranica: $Q_3 + 3R_Q^4$

Akékoľvek pozorovanie za vnútornými hranicami sa pokladá za potenciálny extrém (outsider). Bod za vonkajšími hranicami sa pokladá za jasný extrém. Aj keď údaje nemajú normálne rozdelenie, môžeme použiť box-plot¹³.

3.1.2 Dixonov test na detekciu extrémov.

Dixonov test na odľahlé hodnoty, je veľmi jednoduchý, ľahký na výpočet a vysvetlenie. Dixonov test je založený na pomere rozdielov hodnôt. Vďaka tomu je Dixonov test efektívny na testovanie každej konkrétnej pozorovanej hodnoty.

¹³ LAROSE, D. T.: *Discovering Knowledge in Data. An Introduction to Data Mining*. USA: J. Wiley and Sons 2005, s. 38 - 39.

Dixonov test dáva dobré výsledky aj pre malé súbory. Nevyžaduje splnenie predpokladu normality dát. V závislosti od počtu podozrivých odľahlých hodnôt, sú na identifikáciu potenciálnych odľahlých hodnôt použité rôzne pomery.

V prvej skupine, kde je potenciálna odľahlá hodnota najväčšie a najmenšie pozorovanie, sa používa testovacia štatistika R_{10} pre Dixonov test nasledovne:

Testovanie najväčšej pozorovanej hodnoty

$$R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1} \quad (3.1)$$

Testovanie najmenšej pozorovanej hodnoty:

$$R_{10} = \frac{x_2 - x_1}{x_n - x_1} \quad (3.2)$$

V druhej skupine, kde je potenciálna odľahlá hodnota druhé najväčšie a druhé najmenšie pozorovanie, sa používa testovacia štatistika R_{11} pre Dixonov test nasledovne:

Testovanie najväčšej pozorovanej hodnoty ako odľahlá hodnota s vylúčením najmenšej hodnoty:

$$R_{11} = \frac{x_n - x_{n-1}}{x_n - x_2} \quad (3.3)$$

Testovanie najmenej pozorovanej hodnoty ako odľahlá hodnota m s vylúčením najväčšej hodnoty:

$$R_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1} \quad (3.4)$$

3.1.3 Používanie Z hodnoty na detekciu odľahlých hodnôt

Používanie Z hodnoty na detekcie odľahlých údajov je veľmi jednoduchá metóda identifikácie odľahlých údajov. Táto metóda je založená na výberovom priemere a výberovom rozptyle. Z hodnota je podiel absolútnej hodnoty rozdielu medzi pozorovanou hodnotou a priemerom smerodajnej odchýlky¹⁴:

$$Z = \frac{|x_i - \bar{x}|}{s} \quad (3.5)$$

kde

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

V roku 1987 Douzenis a Rakow navrhovali 2,70 ako hranicu pre potenciálny extrém. Keď bude hodnota Z väčšia ako 2,70 potom pozorovaná hodnota x je považovaná za potenciálny odľahlý údaj. Táto metóda má rovnaký základ ako box-plot metóda, lebo ak je Z väčšie ako 2,70 znamená to, že x_i je väčšia ako 1,5 násobok kvartilového rozpätia. Pokračovanie tejto myšlienky –skutočný extrém sa dá zistiť keď jeho Z hodnota je väčšia ako 4,72.

Metóda Z – hodnoty je dobrá, jednoduchá ale nevhodná pre súbory, ktoré majú malý počet údajov. V roku 1993 Iglewicz a Hoaglin zmenili spôsob vypočtu Z – hodnoty, aby mohli odstrániť nepresnosti pri testovaní odľahlých hodnôt v obmedzenom počte dát. Namiesto smerodajnej odchýlky v menovateli používali odhad smerodajnej odchýlky:

$$\tilde{s} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

¹⁴ <http://itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>

3.1.4 Modifikovaná metóda Z hodnoty

Za predpokladu normality považujeme za odľahlé hodnoty tie, ktoré majú vzdialenosť od strednej hodnoty väčšiu ako dvojnásobok smerodajnej odchýlky:

$$|x_i - \mu| > 2\sigma$$

Na základe tejto myšlienky v roku 1990 navrhli Rousseuw a Zomeren ďalší prístup na zistenie odľahlých hodnôt :

$$\frac{|x_i - \mu|}{\sigma} > 2,24 \quad (3.6)$$

S miernou modifikáciou Rousseuw a Zomeren považujú za odľahlé hodnoty tie, ktorých vzdialenosť strednej hodnoty je väčšia ako 2,24 násobok smerodajnej odchýlky¹⁵. Tento vzorec platí pre základný súbor, keď poznáme strednú hodnotu (μ) a rozptyl (σ^2) základného súboru.

V skutočnosti málokedy poznáme μ a σ . Vtomto prípade budeme testovať extrémny podľa nasledovného vzťahu:

$$\frac{|x_i - \bar{x}|}{s} > 2,24 \quad (3.7)$$

Príklad

Máme údaje: 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 6, 6, 6, 6, 6, 1500

Výberový priemer je 96,875 a výberová smerodajná odchýlka je 374,127. Testujeme či 1500 je odľahlá hodnota.

¹⁵ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 77

$$\frac{|x_i - \bar{x}|}{s} = \frac{|1500 - 96,875|}{374,127} = 3,75$$

3,75 je väčšie ako 2,24, čo znamená, že hodnotu 1500 považujeme za odľahlú hodnotu. Skutočne v tomto prípade 1500 je veľmi neobvykle veľká hodnota takže test bol dobrý.

Problém toho testu sa prejaví, ak pridáme do súboru hodnotu 15000. Je zrejmé, že 1500 a 15000 sú neobvykle veľké v porovnaní s ostatnými hodnotami. Podobne, ako v predchádzajúcom príklade, zistíme výberový priemer 973,529 a výberovú smerodajnú odchýlku 3632,65. Testujeme, či je hodnota 1500 odľahlá hodnota:

$$\frac{|x - \bar{x}|}{s} = \frac{|1500 - 973,529|}{3632,65} = 0,145$$

Podľa výsledku testu hodnota 1500 nie je odľahlá hodnota, hoci v skutočnosti ňou je. Tento príklad ukazuje slabú stránku tejto metódy a zároveň je ilustráciou známeho problému *maskovania* (*masking*). Extrémy ovplyvňujú výpočet výberového priemeru aj výberovej smerodajnej odchýlky. Okrem toho odľahlé hodnoty sa ovplyvňujú samé medzi sebou a dochádza k *maskovaniu* ich prítomnosti, keď na ich detekcie používame Rousseuw – Zomerenov vzťah.

Na odstránenie tohto problému potrebujeme novú metódu na detekciu odľahlých hodnôt, ktorá nie je ovplyvnená extrémami. Typická charakteristika odľahlých hodnôt sa prejavuje v ich veľkostiach. Cez veľkosť hodnoty ovplyvňujú výpočet výberového priemeru a výberovej smerodajnej odchýlky. Znamená to, že potrebujeme hľadať spôsob detekcie, ktorý nie je ovplyvnený veľkosťou hodnoty. Z toho hľadiska vidíme, že medián spĺňa naše kritérium. Najprv sa oboznámime s novým ukazovateľom MAD – *Mediánová absolútna odchýlka* (*median*

absolute deviation). MAD je robustná miera variability jednorozmerných číselných súborov. To môže poukazovať na výberový parameter, ktorý sa dá odhadnúť podľa nasledovného vzorca:

$$\text{MAD} = \text{median}_i(|x_i - M|) \quad (3.8)$$

kde

$$M = \text{median}_i(x_i) \quad (3.9)$$

MAD sa počíta ako medián zo všetkých absolútnych odchýlok x_i od ich mediánu¹⁶.

Za predpokladu normality sa medián rovná strednej hodnote, to znamená, že:

$$|x_i - \mu| = |x_i - M|$$

MAD je medián súboru absolútnych odchýlok X_i a z toho vyplýva:

$$\frac{1}{2} = P(|x_i - \mu| \leq \text{MAD}) = P\left(\frac{|x_i - \mu|}{\sigma} \leq \frac{\text{MAD}}{\sigma}\right) \quad (3.10)$$

Podľa teórie pravdepodobnosti vieme, že:

$$\frac{|x_i - \mu|}{\sigma} = Z$$

Dosadíme do vzorca 3.10 a dostaneme nový tvar:

$$\frac{1}{2} = P(|x_i - \mu| \leq \text{MAD}) = P\left(\frac{|x_i - \mu|}{\sigma} \leq \frac{\text{MAD}}{\sigma}\right) = P\left(|Z| \leq \frac{\text{MAD}}{\sigma}\right)$$

¹⁶ http://en.wikipedia.org/wiki/Median_absolute_deviation

$$P\left(|Z| \leq \frac{\text{MAD}}{\sigma}\right) = \frac{1}{2}. \text{ Znamená } P\left(Z \leq \frac{\text{MAD}}{\sigma}\right) = \frac{3}{4}$$

Z má normované normálne rozdelenie a z toho vychádza vzťah medzi rozptylom a MAD:

$$P\left(Z \leq \frac{\text{MAD}}{\sigma}\right) = \frac{3}{4} \quad \text{alebo} \quad \frac{\text{MAD}}{\sigma} = \phi^{-1}\left(\frac{3}{4}\right) \approx 0,6745$$
$$\Rightarrow \frac{\text{MAD}}{\sigma} \approx 0,6745 \quad \text{alebo} \quad \sigma = 1,4826 \text{ MAD}$$

Keď máme náhodný výber z normálneho rozdelenia, potom:

$$\frac{\text{MAD}}{0,6745} = \text{MADN} \quad (3.11)$$

je dobrým bodovým odhadom smerodajnej odchýlky σ základného súboru.

Takže už sme našli dva ukazovatele M a MAD, ktoré nie sú ovplyvnené veľkosťou hodnoty. Použitím týchto ukazovateľov získame nový prístup na detekciu odľahlých hodnôt.

Teraz x bude odľahlá hodnota, keď:

$$\frac{|x_i - M|}{\text{MADN}} > 2,24 \quad (3.12)$$

Vráťme sa k predchádzajúcemu príkladu s nasledovnými údajmi:

1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 6, 6, 6, 6, 6, 1500, 15000

$M = 3$, potom dostaneme absolútne odchýlky 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 1497, 14997.

Z týchto absolútnych odchýlok zistíme veľmi jednoducho medián odchýlok, alebo $\text{MAD} = 2$,

potom $\text{MADN} = 2,965$. Testujeme hodnotu 1500, ktorá nebola definovaná ako odľahlá hodnota

v prechádzajúcom príklade:

$$\frac{|x - M|}{\text{MADN}} = \frac{1500 - 2}{2,965} = 505,277 > 2,24$$

Podľa toho testu je hodnota 1500 odľahlou hodnotou.

3.1.5 Grubbsov test na detekciu odľahlých hodnôt.

Grubbsov test (Grubbs 1969 Štefanský 1972) sa používa na detekciu odľahlých údajov v jednorozmerných dátach. Je založený na predpoklade normality. To znamená, že by sme mali najprv overiť, či naše dáta možno aproximovať normálnym rozdelením pred použitím Grubbsovho testu¹⁷.

Grubbsov test zistí jednu odľahlú hodnotu naraz. Táto odľahlá hodnota sa vynechá z dátového súboru a test sa opakuje, kým nie sú zistené odľahlé hodnoty. Avšak, viac iterácií zmení pravdepodobnosti detekcie, a test by nemal byť používaný pri súboroch s menším počtom prvkov ako sedem.

Grubbsov test je tiež známy ako maximálny normovaný reziduálny test. Nulová a alternatívna hypotéza má tvar:

H_0 : Neexistuje žiadna odľahlá hodnota v súbore.

H_1 : Existuje aspoň jedna odľahlá hodnota v súbore.

Testovacia štatistika je vyjadrená vzťahom:

$$G = \frac{\max |Y_i - \bar{Y}|}{s} \quad (3.13)$$

¹⁷ Podrobný prehľad o metódach detekcie odľahlých údajov je napríklad v BARNETT, V. – LEWIS, T.: *Outliers in Statistical Data*. New York: Wiley and Sons, 1994. ISBN 978-0471930945.

kde, $\bar{Y}$ a s označujú výberový priemer a smerodajnú odchýlku. Grubbsova testovacia štatistika je najväčšia absolútna odchýlka od priemeru, meraná v smerodajných odchýlkach.

Ide o bilaterálnu verziu testu. Grubbsov test môže tiež byť definovaný ako jeden z týchto jednostranných testov

Test či minimálna hodnota súboru je odľahlá používa testovaciu štatistiku:

$$G_1 = \frac{\bar{Y} - Y_{\min}}{s} \quad (3.15)$$

kde $Y_{\min}$ označuje minimálnu hodnotu súboru.

Test, či maximálna hodnota súboru je odľahlá používa testovaciu štatistiku:

$$G_2 = \frac{Y_{\max} - \bar{Y}}{s} \quad (3.16)$$

kde $Y_{\max}$ označuje maximálnu hodnotu v súboru.

Keď α je hladina významnosti, N je rozsah súboru, potom kritická oblasť testu v obojstrannom teste je:

$$G > \frac{(N-1)}{N} \sqrt{\frac{t_{(\alpha/(2N), N-2)}^2}{N-2 + t_{(\alpha/2N, N-2)}^2}} \quad (3.17)$$

$t_{(\alpha/2N, N-2)}^2$ je kritická hodnota pre t -rozdelenia so stupňami voľnosti $v = N - 2$ a hladinou významnosti $\alpha/2N$ ¹⁸.

3.2 Odľahlé údaje a spôsob ich eliminácie

¹⁸ http://en.wikipedia.org/wiki/Grubbs%27_test_for_outliers

3.2.1 Eliminácia odľahlých hodnôt metódou *zastrihnutého priemeru (Trimmed mean)*

Táto metóda je založená na veľmi jednoduchom princípe. Odľahlé hodnoty považujeme za hodnoty, ktoré sú neobvykle malé alebo veľké hodnoty v súbore. Keď chceme aby v skúmanom súbore neexistovali odľahlé hodnoty, jednoducho z pôvodného súboru vyberieme určitý podiel najväčších a najmenších hodnôt.

Po vynechaní veľkých a malých hodnôt dostaneme nový súbor, ktorý označujeme ako *zastrihnutý súbor (trimmed set)*. *Zastrihnutý priemer* ($\bar{X}_t$) je výberový priemer, ktorý je vypočítaný zo *zastrihnutého súboru*.

Všeobecne nech je γ nejaká konštanta, pre ktorú platí $0 < \gamma < 0,5$. Označme $\eta = 100\gamma$ a vytvorme množinu $g = [\gamma n]$, kde $[\gamma n]$ znamená γn zaokrúhlime nadol na najbližšie celé číslo¹⁹. Potom γ – *zastrihnutý priemer* je priemer vypočítaný z hodnôt, z ktorých bolo g najmenších a najväčších hodnôt pozorovaní odstránených²⁰:

$$\bar{X}_{t,\eta} = \frac{1}{n-2g} (X_{g+1} + X_{g+2} + \dots + X_{n-g}) \quad (3.18)$$

Podstatnou otázkou je, v koľkých percentách by sme mali „odstrihnúť“. *Zastrihnutý priemer* predpokladá „odstrihnúť“ na oboch stranách súboru. Z toho vyplýva, že niekedy stratíme údaje, keď má súbor len maximálne alebo minimálne hodnoty ako odľahlé. Máme podobnú situáciu aj pre M-estimátor. Odľahlé hodnoty sú len na jednej strane umele posúvajú medián k tejto strane a to zvyšuje mediánovú absolútnu odchýlku. Odpoveďou na tento problém je *jednostranný zastrihnutý priemer (one side trimmed mean)*.

3.2.2 Eliminácia odľahlých hodnôt metódou *jednostranného zastrihnutého priemeru*

¹⁹ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 63

²⁰ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 63

V druhej časti sme skúmali problém na detekciu odľahlých hodnôt. Pomocou testu môžeme zistiť, či súbor má odľahlé hodnoty na oboch stranách alebo má odľahlé hodnoty na jednej strane. Najvhodnejšie testy sú Dixonov a Grubbsov test.

Keď zistíme, že súbor má odľahlé hodnoty na ľavej strane potom „odstrihneme“ zo súboru najmenšie hodnoty na ľavej strane. Vypočítame výberový priemer z nového súboru dostaneme $\bar{X}_l$, ktorý označuje ako *ľavo-zastrihnutý priemer*.

Podobne ako *zastrihnutý priemer* nech je γ nejaká konštanta, pre ktorú platí $0 < \gamma < 0,5$. Označme $\eta = 100\gamma$ a vytvorme množinu $g = [\gamma n]$, kde $[\gamma n]$ znamená γn zaokrúhlime nadol na najbližšie celé číslo. Potom γ – *zastrihnutý priemer* je priemer vypočítaný z hodnôt, z ktorých bolo g najmenších hodnôt pozorovaní odstránených:

$$\bar{X}_{l,\eta} = \frac{1}{n-g} (x_{g+1} + x_{g+2} + \dots + x_n) \quad (3.19)$$

Rovnaký prístup ako na *ľavo-zastrihnutý priemer* možno uplatniť pri pravo *zastrihnutom* priemere.

Nech je γ nejaká konštanta, pre ktorú platí $0 < \gamma < 0,5$. Označme $\eta = 100\gamma$ a vytvorme množinu $g = [\gamma n]$, kde $[\gamma n]$ znamená γn zaokrúhlime nadol na najbližšie celé číslo. Potom γ – *zastrihnutý priemer* je priemer vypočítaný z hodnôt, z ktorých bolo g najväčších hodnôt pozorovaní odstránených:

$$\bar{X}_{r,\eta} = \frac{1}{n-g} (x_{g+1} + x_{g+2} + \dots + x_n) \quad (3.20)$$

3.2.3 Eliminácia odľahlých hodnôt metódou Winsorizovaného priemeru (Winsorized mean)

Winsorizovaný priemer je podobný ako *zastrihnutý priemer*, ale „odstrihnuté“ údaje sa nahradia inými údajmi. Inak povedané, neodstrihneme najväčšie a najmenšie pozorované

hodnoty, ale každá z nich sa nahradí najväčšou hodnotou, ktorá by už nebola „odstrihnutá“ pri výpočte zastrihnutého priemeru²¹. Potom γ -winsorizovaný priemer sa vypočíta podľa vzťahu:

$$\bar{x}_{w,\gamma} = \frac{1}{n} \{ (g+1)x_{g+1} + x_{g+2} + \dots + x_{n-g-1} + (g+1)x_{n-g} \} \quad (3.21)$$

3.2.4 Eliminácia odľahlých hodnôt metódou jednostranného Winsorizovaného priemeru (One side Winsorized mean)

Ako jednostranný zastrihnutý priemer u winsorizovaného priemeru dokážeme rozšíriť winsorizovaný priemer na *ľavo-winsorizovaný priemer* a *pravo-winsorizovaný priemer*.

Ľavo-winsorizovaný priemer vypočítame ako *ľavo-zastrihnutý priemer*, ale neodstrihneme najmenšie pozorované hodnoty, každá z nich sa nahradí najväčšou hodnotou, ktorá by už nebola „odstrihnutá“ pri výpočte ľavo-zastrihnutého priemeru. Potom γ -ľavo-winsorizovaný priemer sa vypočíta podľa vzťahu:

$$\bar{x}_{lw,\gamma} = \frac{1}{n} \{ (g+1)x_{g+1} + x_{g+2} + \dots + x_{n-1} + x_n \} \quad (3.22)$$

Podobne γ -pravo-winsorizovaný priemer sa vypočíta podľa vzťahu:

$$\bar{x}_{rw,\gamma} = \frac{1}{n} \{ x_1 + x_2 + \dots + x_{n-g-1} + (g+1)x_{n-g} \} \quad (3.23)$$

²¹ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 64

4 Nové netradičné charakteristiky a metódy indukčnej štatistiky

4.1 Charakteristiky polohy

4.1.1 M-estimátory²²

M-estimátory tvoria inú triedu charakteristík polohy, ktoré majú značný praktický význam.

Keď pre nejakých n pozorovaní $x_1, x_2, \dots, x_n$ chceme nájsť také c , ktoré minimalizuje sumu štvorcov odchýlok

$$\sum_{i=1}^n (x_i - c)^2 \quad (4.1)$$

musí byť $\sum_{i=1}^n (x_i - c)^2 = 0$. Z toho $c = \bar{x}$. Keď teda hľadáme charakteristiku polohy, založenú na minimalizácii štvorcov chýb daných vzťahom (4.1), vedie to k použitiu výberového priemeru.

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

$$\sum_{i=1}^n |x_i - c| \quad (4.2)$$

²² TEREK. M.: *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 2010. CD ROM. ISBN 978-80-225-3076-7. s. 371 – 384.

hodnotu výrazu (2) minimalizuje výberový medián M ²³.

Všeobecne existuje nekonečne veľa spôsobov merania vzdialenosti, ktoré vedú k príslušným charakteristikám polohy. Keby sme napríklad vzdialenosť merali pomocou

$\sum_{i=1}^n |x_i - c|^a$, 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 symetrická²⁴) tak, že nájdeme také c , ktoré vyhovuje rovnici

$$\Psi(x_1 - c) + \Psi(x_2 - c) + \dots + \Psi(x_n - c) = 0 \quad (4.3)$$

Charakteristiky polohy, založené na rovnici (4.3) sa nazývajú M-estimátory.

Napríklad pre $\Psi(x) = x$ je $\Psi(x - c) = x - c$, čo vedie k priemeru, $\Psi(x) = \text{sign}(x)$ vedie k mediánu ($\text{sign}(x)$ sa rovná -1 , 0 alebo 1 v závislosti od toho, či je číslo x menšie, rovné alebo väčšie ako nula). Rozličné miery vzdialenosti vedú k rozličným charakteristikám polohy. Treba riešiť dilemu, akú mieru vzdialenosti je vhodné použiť.

Praktický význam má napríklad Huberova funkcia Ψ , kde $\Psi(x) = x$ za predpokladu, že $|x| < K$, kde K je nejaká konštanta, ktorú treba určiť. Pre $x < -K$: $\Psi(x) = -K$, pre $x > K$:

$\Psi(x) = K$. Iná známa funkcia sa nazýva dvoj váhová (biweight): $\Psi(x) = x(1 - x)^2$ pre $|x| < 1$, inak $\Psi(x) = 0$. Teraz si stručne všimneme problematiku detekcie odľahlých údajov.

²³ Tamže, s. 67.

²⁴ Pre niektoré funkcie Ψ vyžaduje nájdenie vhodných charakteristík polohy špeciálny prístup, keď ide o zošikmené rozdelenia.

4.1.2 Výpočet M-estimátora polohy

Niektoré charakteristiky polohy majú takú vlastnosť, že keď pre násobíme všetky hodnoty pozorovaní konštantou, môžeme rovnakou konštantou pre násobiť aj uvažovanú charakteristiku polohy.

Keď použijeme vo vzťahu (4.3), Huberovu funkciu Ψ spolu s mierou stupnice, ktorá má vysoký bod prelomu, výsledný M-estimátor bude mať rovnaký bod prelomu ako miera stupnice²⁵. Keď použijeme MAD ktorá má bod prelomu 0,5, vo vzťahu (3) ako mieru stupnice, výsledný M-estimátor, bude mať pri použití Huberovej funkcie Ψ bod prelomu tiež rovný 0,5²⁶. Rovnica (4.3) bude mať tvar

$$\Psi\left(\frac{x_1 - c}{MAD}\right) + \Psi\left(\frac{x_2 - c}{MAD}\right) + \dots + \Psi\left(\frac{x_n - c}{MAD}\right) = 0 \quad (4.4)$$

kde c je ako predtým, charakteristika polohy. Pri hľadaní c možno využiť obyčajnú iteratívnu metódu alebo možno použiť len jeden jej krok. Táto druhá stratégia vedie k nájdeniu *jednokrokového M-estimátora (one-step M-estimator)*. Nech n_1 je počet pozorovaní X_i , pre ktoré:

$$\frac{(x_i - M)}{MADN} < -K \quad (4.5)$$

a n_2 je počet pozorovaní, pre ktoré

²⁵ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 82.

²⁶ Pozri WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 83

$$\frac{(x_i - M)}{MADN} > K$$

Najčastejšie sa používa: $K = 1,28$. Potom jednokrokový M-estimátor polohy (založený na Huberovej funkcii Ψ) je

$$\hat{\mu}_{os} = \frac{K(MADN)(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} \quad (4.6)$$

kde $x_{(i)}$ je i -ta poriadková štatistika, ktorá je definovaná podľa usporiadaných hodnôt $x_1 \leq x_2 \leq \dots \leq x_n$.

Pri výpočte hodnoty tohto estimátora sa postupuje tak, že sa metódou založenou na MADN nájdú odľahlé hodnoty, s tým rozdielom, že namiesto vzťahu (4.5) sa pri detekcii odľahlých hodnôt použije vzťah (4.7)

$$\frac{|x - M|}{MADN} \quad (4.7)$$

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 mediánom²⁷.

²⁷ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 82.

4.1.3 Modifikovaný jednokrokový M-estimátor a jeho výpočet²⁸

Niekedy sa používa jednoduchá modifikácia jednokrokového M-estimátora:

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} x_{(i)}}{n - n_1 - n_2} \quad (4.8)$$

V tomto estimátore sa na determináciu n_1 a n_2 používa $K = 2,24$. Bod prelomu konečného výberu tohto estimátora je $0,5^{29}$.

Jednokrokový M-estimátor aj modifikovaný jednokrokový M-estimátor majú oproti zastreihnutému priemeru niektoré výhody. Tieto estimátory napríklad nemusia pri výpočte vyradiť žiadne hodnoty. Keď je rozdelenie napríklad zošikmené doprava, môže byť želateľné vyradiť viac hodnôt z pravého ako z ľavého konca rozdelenia. Tieto estimátory takúto možnosť poskytujú.

Pokúsime sa o stručné porovnanie zastreihnutých priemerov, M-estimátorov a modifikovaných M-estimátorov. Zastreihnuté 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. M-estimátory sú založené na tom, ako meriame celkovú vzdialenosť medzi nejakou charakteristikou polohy a pozorovaniami. Huberova miera vzdialenosti vedie k jednokrokovému M-estimátoru, ktorý na rozdiel od modifikovaného M-estimátora realizuje úpravu odhadu na

²⁸ TEREK. M.: *Analýza odlahlých 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 2010. CD ROM. ISBN 978-80-225-3076-7. s. 371 – 384.

²⁹ Tamže. s. 83.

základe miery stupnice, keď sa počet vyradených hodnôt z pravého konca rozdelenia líši od počtu vyradených hodnôt z ľavého konca rozdelenia³⁰.

4.1.4 Výberový priemer, medián, zastrihnutý priemer a jednokrokové M-estimátory

Pri výberoch z normálneho rozdelenia poskytuje najpresnejšie hodnoty odhadu výberový priemer. Presnosť výberového mediánu je výrazne menšia, presnosť výberového 20% zastrihnutého priemeru a jednokrovového M-estimátora je len o málo menšia. Presnosť modifikovaného jednokrovového M-estimátora je o málo menšia ako presnosť jednokrovového M-estimátora. Empirický výskum ukázal, že pri veľmi malých odchýlkach od normálneho rozdelenia základného súboru, môžu výberový medián, 20% zastrihnutý priemer a jednokrokové M-estimátory poskytovať oveľa presnejšie hodnoty odhadu stredu symetrického rozdelenia ako výberový priemer.

Pri odhadovaní strednej hodnoty základného súboru s normálnym rozdelením je najlepším odhadom výberový priemer. Čím viac sa rozdelenie základného súboru vzdáľuje od normálneho k rozdeleniu s mohutnejšími koncami, stáva sa výberový priemer oveľa menej presnejším odhadom strednej hodnoty základného súboru ako výberový 20% zastrihnutý priemer. Smerodajná chyba 20% zastrihnutého priemeru však môže byť výrazne väčšia ako smerodajná chyba jednokrovových M-estimátorov³¹.

Všeobecne, keď je rozdelenie základného súboru zošikmené, jeho stredná hodnota môže poskytovať veľmi skreslený obraz o typickej hodnote, ktorú hľadáme. Okrem toho už malé odchýlky od normálneho rozdelenia môžu značne rozšíriť intervaly spoľahlivosti pre strednú

³⁰ TEREK, M.: *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 2010, CD ROM, ISBN 978-80-225-3076-7, s. 371 – 384.

³¹ TEREK, M.: *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 2010, CD ROM, ISBN 978-80-225-3076-7, s. 371 – 384.

hodnotu. V takýchto prípadoch sa zdá vhodné prejsť k intervalu spoľahlivosti Tukeyho-McLaughlina pre zastrihnutú strednú hodnotu $\mu_{t(\gamma)}$.

4.2 Intervaly spoľahlivosti

4.2.1 Intervaly spoľahlivosti pre zastrihnutú strednú hodnotu

Intervalový odhad pre zastrihnutú strednú hodnotu je založený na podobnom prístupe ako pre obyčajnú strednú hodnotu. Znamená to, že používame výberový priemer a výberovú smerodajnú odchýlku na odhadovanie parametra v základnom súbore cez Studentovo rozdelenie.

Prvá úloha intervalového odhadu je hľadanie odhadu smerodajnej odchýlky pre γ – zastrihnutý súbor. Keď poznáme s_w^2 je γ -winsorizovaný rozptyl, n je rozsah súboru. Potom γ – zastrihnutý rozptyl sa vypočíta podľa vzťahu³²:

$$s_t^2 = \frac{s_w^2}{(1 - 2\gamma)^2 n} \quad (4.9)$$

Podobne máme vzorec na výpočet γ - jednostranného zastrihnutého rozptylu:

$$s_t^2 = \frac{s_w^2}{(1 - \gamma)^2 n} \quad (4.10)$$

Spôsob výpočtu $(1 - \alpha)\%$ intervalu spoľahlivosti pre zastrihnutú strednú hodnotu ponúkli Tukey a McLaughlin (1963).

³² Terek, M.(2010): *Netradičné charakteristiky polohy 2*. Materiál pre firmu Trexima Bratislava, v rámci riešenia projektu Európskeho sociálneho fondu, NP XIV-2 Zamestnanosť a sociálna inklúzia.

Nech h je počet neodstrihnutých pozorovaní. $t_{1-\frac{\alpha}{2}}(h-1)$ je $\left(1-\frac{\alpha}{2}\right)100\%$ kvantil Studentovho rozdelenia so stupňami voľnosti $v = n - 2g - 1 = h - 1$. $\bar{X}_t$ je hodnota γ – výberového zastreihnutého priemeru a S_w je hodnota winsorizovanej smerodajnej odchýlky (ide o hodnotu výberovej smerodajnej odchýlky winsorizovaných údajov) náhodného výberu rozsahu n ³³. Potom $(1 - \alpha)100\%$ interval spoľahlivosti pre γ – zastreihnutú strednú hodnotu μ_t je

$$\bar{X}_t - t_{1-\alpha/2} \frac{S_w}{(1-2\gamma)\sqrt{n}} < \mu_t < \bar{X}_t + t_{1-\alpha/2} \frac{S_w}{(1-2\gamma)\sqrt{n}} \quad (4.11)$$

kde $t_{1-\alpha/2}$ je kvantil Studentovo rozdelenia so stupňami voľnosti $v = n - 2g - 1 = h - 1$

Rovnaký spôsob môžeme aplikovať na výpočet intervalu spoľahlivosti pre M-estimátor a modifikovaný jednokrokový M-estimátor keď n je dostatočne veľké číslo. V prípade ak $n < 100$, v súbore sa vyžaduje symetrické rozdelenie. V praxi je veľmi ťažko nastaviť konkrétne n aby bolo výberovým súborom s dostatočne veľkým rozsahom.

S využitím bootstrapových metód možno vypočítať presné intervaly spoľahlivosti aj pre M-estimátory.

4.2.2 Intervaly spoľahlivosti pre zastreihnutý rozptyl σ^2 normálneho rozdelenia

Ak n je rozsah výberového súboru zo základného súboru s normálnym rozdelením, je možné vypočítať $(1 - \alpha)100\%$ -ný interval spoľahlivosti pre rozptyl σ^2 použitím výberovej

³³ Terek. M.(2010): *Netradičné charakteristiky polohy 2*. Materiál pre firmu Trexima Bratislava, v rámci riešenia projektu Európskeho sociálneho fondu. NP XIV-2 Zamestnanosť a sociálna inklúzia.

charakteristiky $W_1 = \frac{(n-1)S^2}{\sigma^2}$, ktorá je náhodnou premennou s χ^2 –rozdelením pravdepodobnosti s $\nu = n - 1$ stupňami. Dostaneme interval spoľahlivosti pre rozptyl σ^2 v tvare³⁴

$$\frac{(h-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(h-1)S^2}{\chi^2_{\frac{\alpha}{2}}}$$

Nech s_w^2 je γ -winsorizovaný rozptyl, h je počet neodstrihnutých pozorovaní. Potom γ -obojsstranný zastrihnutý rozptyl sa vypočíta podľa vzťahu:

$$s_t^2 = \frac{s_w^2}{(1-2\gamma)^2 n} \quad \text{a} \quad (4.12)$$

Teda pri zvolenej spoľahlivosti $(1 - \alpha)$ a $\nu = n - 2g - 1$ stupňov voľnosti máme hľadaný interval spoľahlivosti pre parameter σ^2 v tvare

$$\frac{(h-1)S_t^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(h-1)S_t^2}{\chi^2_{\frac{\alpha}{2}}} \quad (4.13)$$

4.3 Testovanie štatistických hypotéz

4.3.1 Testovanie štatistických hypotéz o zastrihnutej strednej hodnote.

Spôsob testovania zastrihnutej strednej hodnoty vychádza zo spôsobu testovania hypotéz pre obyčajnú strednú hodnotu. Z toho hľadiska máme nulovú hypotézu pre *zastrihnutú strednú hodnotu*.

³⁴ PACÁKOVÁ, V. a kol.: *Štatistika pre ekonómov*. Bratislava: Vydavateľstvo EKONÓM, 2003. ISBN 80-89047-74-2.

$$H_0 : \mu_t = \mu_0$$

Nech h je počet neodstrihnutých pozorovaní, $t(h-1)$ Studentovho rozdelenia so stupňami voľnosti $(h-1)$, $\bar{X}_t$ je hodnota γ – výberového zastreihnutého priemeru a S_w je hodnota winsorizovanej smerodajnej odchýlky (ide o hodnotu výberovej smerodajnej odchýlky winsorizovaných údajov) náhodného výberu rozsahu n ³⁵.

Potom podľa prechádzajúcej časti o intervale spoľahlivosti máme vhodné testovacie kritérium pre γ – zastreihnutú strednú hodnotu μ_t :

$$T = \frac{(1-2\gamma)(\bar{X}_t - \mu_0)}{S_w / \sqrt{n}} \quad (4.14)$$

Hodnotu testovacej charakteristiky porovnáme s kritickými hodnotami, ktorými sú kvantily rozdelenia $t(h-1)$ v závislosti od hladiny významnosti α a typu alternatívnej hypotézy H_1 . Kritické oblasti pre zvolené α pri rôznych alternatívach obsahuje tabuľka 4.1

Tabuľka 4.1

Kritické oblasti pre rôzne alternatívy H_1

Alternatívna hypotéza	Kritické hodnoty	Kritická oblasť
$H_1 : \mu_t \neq \mu_0$	$-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$	$\nu_\alpha = \left(-\infty, -t_{1-\frac{\alpha}{2}}\right) \cup \left(t_{1-\frac{\alpha}{2}}, +\infty\right)$
$H_1 : \mu_t > \mu_0$	$t_{1-\alpha}$	$\nu_\alpha = (t_{1-\alpha}, +\infty)$
$H_1 : \mu_t < \mu_0$	$-t_{1-\alpha}$	$\nu_\alpha = (-\infty, -t_{1-\alpha})$

(zdroj: vlastné spracovanie)

³⁵ WILCOX, R. R.: *Applying Contemporary Statistical Techniques*. USA: Academic Press, 2003, s. 167-170

4.3.2 Testovanie štatistických hypotéz o zastrihnutom rozptyle

Na základe určitých informácií, ktoré nepochádzajú z výberového súboru, môžeme niekedy predpokladať, že parameter σ_t^2 sa rovná hodnote σ_0^2

Rovnakým spôsobom, ako testovanie obyčajného rozptylu, môžeme sformulovať nulovú hypotézu.

$$H_0 : \sigma_t^2 = \sigma_0^2$$

ktorú chceme overiť na základe výberových údajov v súbore oproti niektorej z alternatív

$$H_1 : \sigma_t^2 \neq \sigma_0^2$$

$$H_1 : \sigma_t^2 > \sigma_0^2$$

$$H_1 : \sigma_t^2 < \sigma_0^2$$

Pritom máme vhodné testovacie kritérium pre zastrihnutý rozptyl:

$$W = \frac{(h-1)S^2}{\sigma_0^2} \quad (4.15)$$

Nech s_w^2 je γ -winsorizovaný rozptyl, h je počet neodstrihnutých pozorovaní. Potom γ -obojsstranný zastrihnutý rozptyl sa vypočíta podľa vzťahu:

$$s_t^2 = \frac{s_w^2}{(1-2\gamma)^2 n}$$

Potom dostaneme testovaciu charakteristiku pre zastrihnutý rozptyl v tvare:

$$W_1 = \frac{(h-1)S_w^2}{(1-2\gamma)n\sigma_0^2} \quad (4.16)$$

kde s_w^2 je γ -winsorizovaný rozptyl, h je počet neodstrihnutých pozorovaní. Táto testovacia charakteristika má rozdelenie $\chi^2(h-1)$. Kritické hodnoty a kritické oblasti možných alternatívnych hypotéz sú na hladine významnosti α v tabuľke 4.2.

Tabuľka 4.2

Kritické oblasti pre rôzne alternatívy H_1

Alternatívna hypotéza	Kritické hodnoty	Kritická oblasť
$H_1 : \sigma_t^2 \neq \sigma_0^2$	$\chi_{\frac{\alpha}{2}}^2(h-1), \chi_{1-\frac{\alpha}{2}}^2(h-1)$	$\nu_\alpha = \left(0, \chi_{\frac{\alpha}{2}}^2(h-1)\right) \cup \left(\chi_{1-\frac{\alpha}{2}}^2(h-1), +\infty\right)$
$H_1 : \sigma_t^2 > \sigma_0^2$	$\chi_{1-\alpha}^2(h-1)$	$\nu_\alpha = (\chi_{1-\alpha}^2(h-1), +\infty)$
$H_1 : \sigma_t^2 < \sigma_0^2$	$\chi_\alpha^2(h-1)$	$\nu_\alpha = (0, \chi_\alpha^2(h-1))$

(zdroj: vlastné spracovanie)

4.4 Bootstrapová metóda

4.4.1 Základné pojmy a charakteristiky

Metóda bootstrap patrí medzi počítačovo intenzívne metódy pre štatistickú analýzu dát, ktorá sa využíva najmä na výpočet intervalov spoľahlivosti a testovanie hypotéz. *Bootstrap* využíva podobnosť medzi výberom a základným súborom. Metóda je založená na myšlienke rekonštrukcie rozdelenia základného súboru mnohonásobným náhodným vyberaním s opakovaním z pozorovaného výberu. Aplikácia metódy vychádza z predpokladu, že pozorovaný výber je v skutočnosti základný súbor. Výbery možno získavať z fiktívneho základného súboru – z náhodného výberu, ktorý máme k dispozícii.

Názov metódy vychádza z tzv. Münchhausenovej metódy, ktorá vo všeobecnosti označuje princíp, že nejaký systém sa sám naštartuje. Označenie sa vzťahuje na legendu o barónovi Münchhauseni, ktorý sa z bažiny dostal tak, že sa sám vytiahol za vlastné vlasy. V americkej verzii na to použil remene na čižmách (bootstraps), preto sa táto metóda nazýva bootstrap alebo bootstrapping³⁶.

Bootstrap je prevzorkovaný postup, lebo z pôvodného súboru dát urobíme prevzorkovanie. Niektoré postupy prevzorkovania podobné metóde bootstrap už boli predtým. Využitie počítačov na simulácie prebiehalo už koncom štyridsiatych rokov. Až v roku 1979 bola *bootstrap* akceptovaná ako silný štatistický nástroj na odhadovanie štandardnej odchýlky vďaka Efronovi³⁷, ktorý zjednotil myšlienky a spájal neparametrickú bootstrap “prevzorkovanie dát s opakovaním - *resamples the data with replacement*” s predchádzajúcimi štatistickými nástrojmi ako *jackknife* a *delta* metóda s hlavným účelom poskytnutia informácie čitateľom, ktorí nemajú hlboké poznatky v matematickej oblasti. Ďalšími autormi v tejto oblasti sú Efron a Tibshirani (1993), Chernick (1999), Davison a Hinkley (1997), Hall a Hall (1995), Lunneborg (2000), Mooney a Duval (1993), Shao a Tu (1995).

Uvažujme nezávisle identické rozdelenie (iid) s náhodnými veličinami $X_1, \dots, X_n$ a s distribučnou funkciou F . Nech $\theta = \theta(F)$ je nejaká charakteristika rozdelenia; je to pre nás neznámy parameter, ktorý musíme odhadnúť na základe realizácie náhodného výberu.

Nech $T_n = T_n(X_1, \dots, X_n)$ je štatistika pre odhad parametra θ , nech $R_n = R_n(X_1, \dots, X_n)$ je jej vhodná štandardizovaná verzia, napr. $R_n = \sqrt{n}(T_n - \theta)$, alebo nejaká iná funkcia. Nech $H_n(x) = P[R_n(X_1, \dots, X_n, F) \leq x]$ je distribučná funkcia štatistiky R_n .

Explicitné odvodenie rozdelenia H_n aj výpočet číselných charakteristík môže byť v jednotlivých prípadoch veľmi náročne, alebo niekedy analyticky nevykonateľné a to aj vtedy, keď poznáme distribučnú funkciu F . V takom prípade (keď distribučná funkcia F je známa) možno postupovať metódou *Monte Carlo*, ktorá generuje dlhú sériu nezávislých náhodných výberov z rozdelenia s danou distribučnou funkciou (tzv. viackrát umele opakovat' pokus), pri

³⁶ Terek, M.(2010): *Netradičné charakteristiky polohy 2*. Materiál pre firmu Trexima Bratislava, v rámci riešenia projektu Európskeho sociálneho fondu, NP XIV-2 Zamestnanosť a sociálna inklúzia.

³⁷ [http://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)](http://en.wikipedia.org/wiki/Bootstrapping_(statistics))

každom opakovaní spočítať hodnotu príslušnej charakteristiky a jej skutočné rozdelenie aproximovať empirickým rozdelením, ktoré získame zo série takto umelo získaných hodnôt³⁸.

Ak distribučná funkcia F je neznáma, čo je oveľa častejší prípad, je možné aproximovať H_n asymptotickým rozdelením, ktoré možno odvodiť na základe centrálnych limitných viet z teórie pravdepodobnosti. Presnosť takej aproximácie je však ovplyvnená a obmedzená počtom pozorovaní, ktoré sú skutočne k dispozícii.

Metóda *bootstrap* ponúka riešenie, ktoré kombinuje tzv. *substitučný princíp* a metódu *Monte Carlo*.

Začneme najprv *substitučný princíp*. Nech $F_n(x)$ je nejaký odhad distribučnej funkcie. Najčastejšie sa uvažuje empirická distribučná funkcia založená na náhodnom výbere $X_1, \dots, X_n$ v nasledovnom tvare:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I[X_i \leq x]$$

kde $I[A]$ sa označuje ako indikátor množiny A . Pri daných hodnotách $x_1, \dots, x_n$ je $F_n(x)$ známa funkcia.

Nech $x_1^*, \dots, x_n^*$ je nezávislý náhodný výber z F_n , potom pri daných pozorovaných $x_1, \dots, x_n$ sú $x_1^*, \dots, x_n^*$ podmienené nezávislé rovnako rozdelené náhodné veličiny, z ktorých každá nadobúda hodnoty $x_1, \dots, x_n$ s pravdepodobnosťou $\frac{1}{n}$. Súbor $x_1, \dots, x_n$ sa nazýva *bootstrapový výber*. V ďalších úvahách pôvodný výber nahradíme bootstrapovým výberom a neznámu distribučnú funkciu F známou distribučnou funkciou F_n . Dostaneme parameter $\theta^* = \theta(F)$ a štatistiku $T_n^* = T_n(x_1^*, \dots, x_n^*)$ a $R_n^* = R_n(x_1^*, \dots, x_n^*, F_n)$.

Teraz môžeme definovať charakteristiku ako:

$$E(T_n^*) = \int T_n(x_1, \dots, x_n) d(F_n(x_1), \dots, F_n(x_n)),$$

³⁸ <http://www.statspol.cz/robust/robust2004/praskova.pdf>

$$D(T_n^*) = \int [T_n(x_1, \dots, x_n) - E(T_n^*)]^2 d(F_n(x_1), \dots, F_n(x_n)),$$

a distribučná funkcia

$$H_n^*(x) = P^*(R_n(x_1^*, \dots, x_n^*, F_n) \leq x)$$

$$H_n^*(x) = P(R_n(x_1^*, \dots, x_n^*, F_n) \leq x | X_1, \dots, X_n);$$

Máme tzv. *teoretické charakteristiky* a *teoretickú distribučnú funkciu*, ktoré sú získané metódou *bootstrap*.

Povedali sme že, $\hat{H}_n$ je konzistentný odhad H_n , ak

$$\rho(\hat{H}_n, H_n) \rightarrow 0 \quad \text{pri} \quad n \rightarrow \infty$$

Pri slabšej konzistencii alebo silnej konzistencii, kde ρ je nejaká metrika na priestore distribučných funkcií. Najčastejšie metriky v tomto prípade sú *supremálne metriky*

$$\rho_\infty(G, H) = \sup_{x \in R} |G(x) - H(x)|$$

Alebo tzv. *Mallowsova vzdialenosť*, ktorá je pre distribučné funkcie G a H zo skupiny distribučných funkcií s konečnými r -tými momentmi definovaná predpisom

$$\tilde{\rho}_r(G, H) = \inf_{T_{X,Y}} (E|X - Y|^r)^{1/r}$$

kde $T_{X,Y}$ je množina všetkých možných združených rozdelení vektorov (X,Y) , ktorých marginálne rozdelenia sú G a H . Podľa vlastností *Mallowsvej vzdialenosti* a súvislosti s konvergenciou v distribúcii náhodných veličín, môžeme definovať konzistenciu bootstrapových charakteristík prirodzeným spôsobom³⁹.

Nech $\hat{D}(T_n)$ je konzistentný odhad rozptylu T_n , potom

$$\hat{D}(T_n)/D(T_n) \rightarrow 1 \quad \text{pri } n \rightarrow \infty$$

Pre praktické použitie sú teoretické bootstrapové charakteristiky vhodné len v prípade, že sú explicitnými funkciami pozorovaní $x_1, \dots, x_n$. Presné stanovenie bootstrapového rozdelenia by sa vyžadovalo vykonanie všetkých n^n možných výberov s opakovaním zo základného súboru pozorovaných hodnôt $x_1, \dots, x_n$. To je však uskutočniteľné len pre výbery s malým rozsahom. Aj keby sme sa obmedzili len na vzájomne rôzne výbery, máme ich stále ešte $\binom{2n-1}{n}$, čo pre $n = 10$ máme 92378 výberov.

Kvôli tomu sa na bootstrapový výber $x_1^*, \dots, x_n^*$ so známou distribučnou funkciou F_n aplikuje metóda Monte Carlo, keď sa mnohokrát (B -krát) generuje nezávislý náhodný výber z rozdelenia F_n , pri každom opakovaní sa spočítajú hodnoty T_n^*, R_n^* a z nich sa stanoví aritmetický priemer. Dostaneme bootstrapový odhad pôvodného rozdelenia a pôvodných charakteristík.

Napr. bootstrapový odhad rozptylu T_n dostaneme tak, že opakujeme nezávislý náhodný výber z rozdelenia F_n B -krát a spočítame vždy hodnotu štatistiky T_n^* . Dostávame také hodnoty $T_{n,1}^*, \dots, T_{n,B}^*$, z ktorých vypočítame:

$$D(T_n^*) = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{k=1}^B T_{n,k}^* \right)^2$$

³⁹ <http://www.statapol.cz/robust/robust2004/praskova.pdf>

Podobne odhadneme distribučnú funkciu štatistiky R_n ako

$$\widehat{H}_n^*(x) = \frac{1}{B} \sum_{b=1}^B I\{R_n(X_{1,b}^*, \dots, X_{n,b}^*, F_n) \leq x\}$$

kde $\{x_{1,b}^*, \dots, x_{n,b}^*\}, b = 1, \dots, B$, sú nezávislé výbery z F_n .

Ďalej aplikujeme bootstrap metódu pre strednú hodnotu.

Nech $x_1, \dots, x_n$ sú hodnoty nezávislého náhodného výberu pozorovaní a nech sú $x_1^*, \dots, x_n^*$ hodnoty nezávislého náhodného bootstrapového výberu veľkosti n , ktoré sú získané náhodným výberom n hodnôt s opakovaním z $x_1, \dots, x_n$. Potom hodnota výberového priemeru bootstrapového výberu je

$$\bar{X}^* = \frac{1}{n} \sum X_i^*.$$

Predpokladajme, že opakujeme proces generovania bootstrapového výberu B -krát. Označme príslušné získané hodnoty výberových priemerov $\bar{x}_1^*, \dots, \bar{x}_B^*$.

Potom približný $(1-\alpha)100\%$ bootstrapový interval spoľahlivosti pre μ je

$$(\bar{x}_{(\ell+1)}^*, \bar{x}_{(u)}^*),$$

kde $\bar{x}_1^* \leq \dots \leq \bar{x}_B^*$ je B hodnôt výberových priemerov z bootstrapových výberov, ktoré sú zoradené vzostupne, $\ell = \alpha B / 2$ zaokrúhlené na najbližšie celé číslo a $u = B - \ell$.

Vo všeobecnosti, $\bar{x}_{(\ell+1)}^*$ a $\bar{x}_{(u)}^*$ obsahujú $(1-\alpha)$ percentný stred B bootstrapových výberových priemerov.

Normálny výberový rozptyl S^2 vieme určiť podľa nasledovného vzorca

$$est\sigma^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

kde x_i označuje pozorované hodnoty nezávislého výberu, $\bar{x}$ je výberový priemer, n je počet pozorovaní. *Bootstrapová smerodajná chyba* (bootstrap standard error) S_b je smerodajná odchýlka bootstrapového rozdelenia štatistiky. *Bootstrap* nám tiež umožňuje počítať smerodajnú chybu pre štatistiku jednoducho pomocou *bootstrapovej percentilovej metódy*

4.4.2 Percentilová metóda

Podrobnejšie si všimneme *percentilovú bootstrapovú metódu*. Začneme príkladom.

Bootstrapový výber rozsahu n sa získa náhodným vyberaním s opakovaním z pozorovaných hodnôt. Keďže ide o náhodné vyberanie s opakovaním, jednotlivé náhodné výbery sa môžu líšiť. Potom sa môžu líšiť aj príslušné hodnoty výberového priemeru.

Predstavme si, že takto získame B bootstrapových výberov rozsahu n a pre každý vypočítame hodnotu výberového priemeru. Keď je B dostatočne veľké, dostaneme množinu hodnôt výberového priemeru, ktorá poskytuje aproximáciu výberového rozdelenia výberového priemeru. Potom hranice $(1-\alpha)100\%$ stredu výberového rozdelenia výberového priemeru sú hranice *percentilového bootstrapového intervalu spoľahlivosti pre strednú hodnotu μ* .⁴⁰

Nech $X_1, \dots, X_n$ sú hodnoty pozorovaní a $X_1^*, \dots, X_n^*$ sú hodnoty pozorovaní bootstrapového výberu rozsahu n , ktorý sa získa náhodným vyberaním s opakovaním. Hodnota výberového priemeru bootstrapového výberu je

⁴⁰ Terek, M.(2010): *Netradičné charakteristiky polohy* 2. Materiál pre firmu Trexima Bratislava, v rámci riešenia projektu Európskeho sociálneho fondu, NP XIV-2 Zamestnanosť a sociálna inklúzia

$$\bar{X}^* = \frac{1}{n} \sum_{i=1}^n \bar{X}_i^*$$

Predpokladajme, že B krát opakujeme proces generovania výberového priemeru bootstrapového výberu. Označme príslušné hodnoty výberových priemerov $\bar{x}_1^*, \dots, \bar{x}_B^*$. Potom približný $(1 - \alpha)100\%$ percentilový bootstrapový interval spoľahlivosti pre strednú hodnotu μ je

$$(\bar{x}_{(l+1)}^*, \bar{x}_{(u)}^*)$$

kde $\bar{x}_{(1)}^* \leq \dots \leq \bar{x}_{(B)}^*$ je B hodnôt výberových priemerov z bootstrapových výberov (hodnôt poriadkových štatistík $\bar{x}_{(1)}^*, \dots, \bar{x}_{(B)}^*$), $l = \frac{\alpha}{2} B$, zaokrúhlené na najbližšie prirodzené číslo a

$$u = B - l.$$

Poznamenajme, že poriadková štatistika je štatistika určená svojim poradím v neklesajúcom usporiadaní náhodných premenných. Zoberme napríklad pozorované hodnoty vo výbere z príkladu:

3, 14, 11, 6, 23, 24, 32, 6, 10, 36, 6, 32

Potom pozorované hodnoty poriadkových štatistík sú

3, 6, 6, 6, 10, 11, 14, 23, 24, 32, 32, 36

Tieto hodnoty predstavujú realizácie poriadkových štatistík $x_{(1)}, x_{(2)}, \dots, x_{(10)}$.

4.4.3 Bootstrapové odhadovanie rozptylu estimátorov percentilovou metódou

Nech $\hat{\mu}$ je nejaká charakteristika polohy (napríklad M-estimátor alebo modifikovaný M-estimátor) a nech $\hat{\mu}^*$ je jej hodnota založená na bootstrapovom výbere. Nech $\hat{\mu}_b^*$

($b = 1, \dots, B$) je B hodnôt bootstrapových odhadov charakteristiky polohy. Potom hodnota odhadu rozptylu charakteristiky $\hat{\mu}$ je⁴¹

$$S^2 = \frac{1}{B-1} \sum_{b=1}^B (\hat{\mu}_b^* - \bar{\mu}^*)^2$$

$$\text{Kde } \bar{\mu}^* = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b^*$$

4.4.4 Bootstrap-t interval

Bootstrap-t (alebo percentil-t) metóda je založená na základnej myšlienke, keď poznáme distribučnú funkciu rozdelenia, ktoré má nasledujúci tvar.

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}},$$

⁴¹ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 211-216

Môžeme vypočítať interval spoľahlivosti pre strednú hodnotu základného súboru.

Prvé prístupy *Bootstrap-t metódy* pre odhad T rozdelenia začíname rovnako ako pri percentilovej metóde: Získame rozsah bootstrapovej vzorky n . Nech $x_1, \dots, x_n$ sú hodnoty nezávislého náhodného výberu pozorovaní a nech $X_1^*, \dots, X_n^*$ hodnoty nezávislého náhodného bootstrapového výberu rozsahu n , ktoré sú získané náhodným výberom n hodnôt s opakovaním z $x_1, \dots, x_n$. Nech $\bar{x}^*$ je priemer a S^* smerodajná odchýlka založené na tejto bootstrapovej vzorke⁴².

Keď je

$$\bar{X}^* = \frac{1}{n} \sum X_i^*$$

a

$$S^* = \sqrt{\frac{1}{n-1} \sum (X_i^* - \bar{X}^*)^2},$$

Potom dostaneme

$$T^* = \frac{\bar{X}^* - \bar{X}}{S^* / \sqrt{n}}.$$

⁴² WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 211-216

Vieme, že ak získame bootstrapovú vzorku, potom dostaneme priemer $\bar{x}$ rozdelenia, z ktorého bola bootstrapová vzorka získaná. V bootstrapovej metóde $\bar{x}$ reprezentuje μ a $\bar{x}^*$ reprezentuje $\bar{x}$.

Ak B -krát opakujeme uvedený proces, získame približnú hodnotu výberového T rozdelenia a predovšetkým máme odhad jeho 0,025 a 0,975 kvantilov. Odhad týchto kvantilov je založený na 95%-nom strede hodnôt T^* .

Nech $T_{(1)}^* \leq T_{(2)}^* \leq \dots \leq T_{(B)}^*$ je B bootstrapových T^* hodnôt, ktoré sú zoradené vzostupne a nech $\ell = 0,025B$ zaokrúhlené na najbližšie celé číslo a $u = B - \ell$, dostaneme odhad 0,025 a 0,975 kvantilov T rozdelenia je $T_{(\ell+1)}^*$ a $T_{(u)}^*$. Výsledný 95%-ný interval spoľahlivosti pre μ je

$$\left(\bar{X} - T_{(u)}^* \frac{S}{\sqrt{n}}, \bar{X} - T_{(\ell+1)}^* \frac{S}{\sqrt{n}} \right)$$

a nazýva sa *interval spoľahlivosti s rovnakými koncami*.

Testovanie hypotéz pomocou bootstrap-t metódy

Na testovanie $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

vypočítame

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

H_0 zamietneme, ak

$$T \leq T_{(\ell+1)}^*$$

alebo ak

$$T \geq T_{(u)}^*.$$

4.4.5 Symetrické intervaly spoľahlivosti

Aplikácia bootstrap-t metódy pre symetrické intervaly spoľahlivosti môže byť spomenutá len pri testovaní obojstranných hypotéz. Lepšie je používať T^* , ktorá je definovaná podľa bootstrap-t⁴³

$$T^* = \frac{|\bar{X}^* - \bar{X}|}{S^* / \sqrt{n}}$$

a zamietneme $H_0: \mu = \mu_0$, ak $|T| \geq T_{(c)}^*$, kde $c = (1 - \alpha)B$ je zaokrúhlené na najbližšie celé číslo a $T_{(1)}^* \leq \dots \leq T_{(B)}^*$ je B bootstrap T^* hodnôt zoradených vzostupne. Približný $(1 - \alpha)$ interval spoľahlivosti pre μ je daný ako

$$\bar{X} \pm T_{(c)}^* \frac{S}{\sqrt{n}}$$

⁴³ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 211-216

a nazveme ho *symetrický obojstranný interval spoľahlivosti*. Pojem *symetrický obojstranný interval spoľahlivosti* znamená, že $(T_{(c)}^* S / \sqrt{n})$ je pripočítané a odpočítané od priemeru z každej strany pri výpočte intervalu spoľahlivosti.

S veľkým rozsahom súboru má *symetrický obojstranný interval spoľahlivosti* niekoľko teoretických výhod oproti intervalu spoľahlivosti s rovnakými koncami. Hlavným problémom je to, že ak je rozsah súboru malý, potom pravdepodobnosť chyby prvého druhu môže byť neuspokojivá. V niektorých prípadoch sa aktuálna pravdepodobnosť pokrytia týchto dvoch metód líši veľmi málo⁴⁴.

4.4.6 Percentilové bootstrapové intervaly spoľahlivosti pre zastreihnutú strednú hodnotu

Používanie percentilovej metódy

Pri tejto metóde sa používajú podobné postupy na odhadovanie intervalu spoľahlivosti zastreihnutej strednej hodnoty ako pri odhadovaní *bootstrapového intervalu* spoľahlivosti pre μ . Jednoducho $\bar{x}$ nahradíme $\bar{x}_t$. Znamená to, že budeme generovať bootstrapovú vzorku s rozsahom n , vypočítame zastreihnutý priemer $\bar{x}_t^*$. Tento proces budeme opakovať B -krát, potom dostaneme príslušné hodnoty výberových priemerov $\bar{x}_{t1}^*, \dots, \bar{x}_{tB}^*$. Potom približný $(1 - \alpha)100\%$ percentilový bootstrapový interval spoľahlivosti pre zastreihnutú strednú hodnotu μ_t je

$$(\bar{X}_{t(l+1)}^*, \bar{X}_{t(u)}^*)$$

⁴⁴ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 216

kde $\bar{x}_{t(1)}^* \leq \dots \leq \bar{x}_{t(B)}^*$ je B hodnôt výberových priemerov z bootstrapových výberov (hodnôt poriadkových štatistík $\bar{x}_{t(1)}^*, \dots, \bar{x}_{t(B)}^*$), $l = \frac{\alpha}{2} B$, zaokrúhlené na najbližšie prirodzené číslo a

$$u = B - l^{45}.$$

Aplikovanie bootstrap-t metódy

Bootstrap-t metóda môže byť aplikovaná pre odhadovanie intervalu spoľahlivosti zastrihnutého priemeru nasledovným spôsobom. Generujeme bootstrapovú vzorku s rozsahom n a vypočítame zastrihnutý priemer a winsozirozanú štandardnú odchýlku, ktoré označíme $\bar{x}_t^*$ a S_w^* . Ďalší krok vypočítame t-hodnotu

$$T_t^* = \frac{(1 - 2\gamma)(\bar{x}_t^* - \bar{x}_t)}{S_w^* / \sqrt{n}}$$

kde γ je hodnota odstrihnutia. Opakujeme tento proces B -krát, potom dostaneme B -krát príslušných hodnôt T_t^* .

Nech $T_{t(1)}^* \leq T_{t(2)}^* \leq \dots \leq T_{t(B)}^*$ je B bootstrapových T_t^* hodnôt, ktoré sú zoradené vzostupne a nech $\ell = 0,025B$ zaokrúhlené na najbližšie celé číslo a $u = B - \ell$, dostaneme odhad 0,025 a 0,975 kvantilov T_t rozdelenia je $T_{t(\ell+1)}^*$ a $T_{t(u)}^*$. Výsledný 95%-ný interval spoľahlivosti pre μ je

⁴⁵ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 221-226

$$\bar{x}_t - T_{t(u)}^* \frac{S_w}{(1-2\gamma)\sqrt{n}}, \bar{x}_t + T_{t(l+1)}^* \frac{S_w}{(1-2\gamma)\sqrt{n}}$$

a nazýva sa *interval spoľahlivosti s rovnakými koncami*.

Testovanie hypotéz pomocou bootstrap-t metódy

Na testovanie $H_0: \mu_t = \mu_{t_0}$, vypočítame

$$T = \frac{\bar{x}_t - \mu_{t_0}}{s_t / \sqrt{n}}$$

H_0 zamietneme, ak

$$T \leq T_{(\ell+1)}^*$$

alebo ak

$$T \geq T_{(u)}^*.$$

4.4.7 Percentilové bootstrapové intervaly spoľahlivosti založené na M-estimátoroch

Keď je rozsah výberu aspoň 20, percentilová bootstrapová metóda dáva dobré výsledky s M-estimátorom⁴⁶. Budeme hľadať $(1 - \alpha)100\%$ percentilový bootstrapový interval pre μ_{os} –

⁴⁶ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 228.

charakteristiku polohy základného súboru, odhadovanú pomocou $\hat{\mu}_{os}$ (jednokrokový M-estimátor definovaný rovnicou (4.6). Treba generovať B bootstrapových výberov a na ich základe získať B hodnôt estimátora $\hat{\mu}_{os}$. Tieto hodnoty treba usporiadať vzostupne⁴⁷:

$$\hat{\mu}_{os(1)}^* \leq \dots \leq \hat{\mu}_{os(B)}^*$$

Potom $(1 - \alpha)100\%$ percentilový bootstrapový interval spoľahlivosti pre μ_{os} je

$$(\hat{\mu}_{os(l+1)}^*, \hat{\mu}_{os(u)}^*)$$

kde $l = \frac{\alpha}{2}B$, zaokrúhlené na najbližšie prirodzené číslo a $u = B - l$.

Pre $\alpha = 0,05$ plne postačuje⁴⁸ voliť $B = 399$ ⁴⁹.

Ked' chceme vypočítať $(1 - \alpha)100\%$ percentilový bootstrapový interval pre μ_{mom} – charakteristiku polohy základného súboru, odhadovanú pomocou $\hat{\mu}_{mom}$, potom sa $\hat{\mu}_{os}$ jednoducho nahradí $\hat{\mu}_{mom}$, ktorý sa vypočíta podľa vzťahu

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

⁴⁷ Terek, M.(2010): *Netradičné charakteristiky polohy* 2. Materiál pre firmu Trexima Bratislava, v rámci riešenia projektu Európskeho sociálneho fondu, NP XIV-2 Zamestnanosť a sociálna inklúzia

⁴⁸ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 228

⁴⁹ Na výpočte interval spoľahlivosti pre M-estimátor na istotu zvolíme minimálne $n = 20$ a $B = 399$ (Wilcox, s. 228)

kde $x_{(i)}$ je i -ta poriadková štatistika, n_1 je počet pozorovaní x_i , pre ktoré:

$$\frac{(x_i - M)}{MADN} < -K$$

a n_2 je počet pozorovaní, pre ktoré

$$\frac{(x_i - M)}{MADN} > K$$

Používa sa $K = 2,24$. Intervaly spoľahlivosti pre μ_{mom} poskytujú dobré výsledky už pre

$B = 500$ ⁵⁰.

Pre oba intervaly možno očakávať dobré výsledky⁵¹ už pre $n = 20$.

⁵⁰ WILCOX, R. R. (2003): *Applying Contemporary Statistical Techniques*. USA: Academic Press. ISBN 0-12-751541-0, s. 228

⁵¹ Tamže.

5 Výsledky práce

5.1 Analýza miezd

5.1.1 Podstata a formy mzdy

Mzda je cena za prácu. Je výsledkom fungovania trhu. Vyplýva z dopytu po práci a ponuky práce. Kategória „mzda“ je všeobecný pojem pre mzdy, platy a ostatné formy odmien a kompenzácií za prácu.

Mzda vystupuje v dvoch základných formách

- časová mzda,
- úkolová mzda.

Prvou formou mzdy, historicky a logicky, bola časová mzda. Z nej sa postupne odvodila úkolová mzda, ktorá je vlastne premenenou formou časovej mzdy.

Časová mzda je forma mzdy, ktorá vyjadruje odmenu za určitý čas (za hodinu, deň, týždeň, mesiac), a podľa toho rozlišujeme hodinovú, dennú, týždennú, mesačnú mzdu. Pri časovej mzde vymedzujeme celkovú sumu mzdy a cenu práce za jednu časovú jednotku (hodinu).

Úkolová mzda je premenenou formou, modifikáciou časovej mzdy. Pri jej určení sa vychádza z priemerného počtu výrobkov, ktoré pracovník vyrobí za určitý čas.

Úkolová mzda sa stáva dôležitým faktorom zvyšovania intenzity práce. Pri danej úkolovej sadzbe sa pracovník usiluje pracovať intenzívnejšie, aby si tak zvýšil mzdu. Pri úkolovej mzde má podnikateľ možnosť meniť sadzbu za výrobok, súčiastku alebo pracovnú operáciu.

Nominálna mzda predstavuje sumu peňazí, ktorú pracovník dostáva. Tá môže byť v úkolovej forme alebo časovej forme.

Reálna mzda predstavuje sumu tovarov a služieb, ktoré si pracovník za svoju nominálnu mzdu môže kúpiť. Reálna mzda je ukazovateľ spotreby pracovníka, je rozhodujúca pre jeho životnú úroveň. Keď porovnávame mzdy v jednotlivých krajinách, porovnávame reálne mzdy.

Kompenzačná mzda - cieľom je kompenzovať (vynahradiť) určité znevýhodnenia, ktoré sú spojené s danou prácou, súvisí s rizikom, ľudia pracujúci v rizikových pracoviskách majú vďaka rizikovej práci vyššiu mzdu (napr. práce v baniach, pri peciach...), alebo môže byť tiež spojená s prácou v sťažených podmienkach (napr. na Aljaške, kde je chladnejšie klíma).

Dynamická mzda - vznikne, ak v určitej lokalite enormne stúpne dopyt po určitom druhu práce. Napríklad z dôvodov zvýšenia výstavby budú mzdy v stavebníctve vyššie ako inde v krajine. Tento jav je len dočasný.

5.1.2 Charakteristika výberovej vzorky ISPZ

Horizonty globálneho myslenia v hospodárskej politike Európskej únie (EÚ) vedú k identifikácii vzájomnej podmienenosti produktivity, výšky disponibilného príjmu a pracovnej výkonnosti ekonomicky aktívneho obyvateľstva. Stanovenie mzdy zachováva národné i regionálne špecifická a vyvoláva potrebu adekvátneho poznania technologického vývoja a jeho vplyvu na obsah a organizáciu práce. K poznaniu mzdového vývoja zohľadňujúceho globálnu, socioekonomickú situáciu sú nevyhnutné kvalitné, medzinárodne plne kompatibilné štatistické údaje.

Štatistické údaje o mzdovom vývoji patria k dôležitým strategickým nástrojom pri riešení aktuálnych otázok trhu práce na národnej úrovni. Takéto informácie sú nevyhnutné pri napĺňaní cieľov Vlády SR v oblasti zamestnanosti a sociálnej inklúzie, ako aj v oblasti zlepšovania podnikateľského prostredia, ktoré sú vytýčené v Národnom programe reforiem SR 2010, schválenom 17. novembra 2010 Vládou SR, vychádzajúc z Európskej stratégie 2020.

Pravidelné sledovanie a vyhodnocovanie úrovne zárobkov a štruktúry miezd v detailnom členení podľa konkrétnych zamestnaní vyplýva z Nariadenia ES č. 530/1999 na zabezpečenie štrukturálnej štatistiky príjmov, Nariadenia Európskej komisie č. 72/2002 o kvalite zisťovania štruktúry miezd zamestnancov, z dohovoru Medzinárodnej organizácie práce (MOP) o štatistikách práce č. 160/1985 a Odporúčania MOP č. 170/1985 k štatistikám práce, ktorá zaviazala sociálnych partnerov vybudovať tento informačný systém.

Informačný systém o cene práce (ISCP), ktorého gestorom je od roku 1992 Ministerstvo práce, sociálnych vecí a rodiny SR, zahŕňa aj pravidelné longitudinálne⁵² štvrťročné zisťovanie priemerných hodinových a mesačných zárobkov zamestnancov podľa jednotlivých zamestnaní.

Zisťovanie je pod názvom „Štvrťročný výkaz o cene práce“ ISCP (MPSVR SR) 1-04 zaradené do Programu štátnych štatistických zisťovaní na roky 2009-2011 vyhláškou Štatistického úradu SR č. 416/2008 Z. z., čiastka 152, str. 3487.

Dôležitou prednosťou ISCP je minimalizácia prácnosti na strane respondenta (zamestnávateľskej organizácie), získavanie súboru vstupných dát modernou elektronickou formou, čo je zárukou kvalitných a dôveryhodných výsledkov.

Hlavným cieľom zisťovania je zabezpečiť objektívne a aktuálne informácie na účely:

- Poskytovania aktuálnych údajov výborom NR SR, jednotlivým ministerstvám pre zabezpečenie vlastných úloh, sociálnym partnerom, jednotlivým stupňom a typom škôl, vedecko-výskumným inštitúciám, ústredným orgánom, orgánom štátnej správy a samosprávy, všetkým účastníkom trhu práce atď.,
- Riešenia súdnych sporov,
- Ďalšieho administratívneho spracovávanía údajov v Sociálnej poisťovni,
- Porovnávanía údajov na národnej a nadnárodnej úrovni v zmysle požiadaviek Európskeho štatistického úradu, Medzinárodnej organizácii práce a OECD,
- Získavania zahraničných investícií (prostredníctvom zastupiteľských úradov, agentúry SARIO a priamo),
- Kolektívneho vyjednávania na všetkých stupňoch,
- Vyhodnocovania rovnosti príležitostí žien a mužov,
- Informovania masovo-komunikačných prostriedkov, odbornej verejnosti a občanov SR.

Zverejňujú sa údaje o priemernej hrubej mesačnej mzde zamestnanca (bez manažérskych a podnikateľských príjmov). Priemerná mesačná mzda za štvrťrok zahŕňa hrubú mzdu, ktorá pozostáva zo základného platu, odmien, príplatkov za stratu času a nadčasy a iných príplatkov.

⁵² Logitudinálne(*longitudinally*) je podĺžne

Údaje sa získavajú prostredníctvom štvrťročných zisťovaní v organizáciách a kvalifikovaným odhadom za zamestnancov u živnostníkov. Štvrťročné zisťovanie zahŕňa všetky odvetvia ekonomickej činnosti na celom území Slovenskej republiky.

Priemerná mesačná mzda sa vypočíta spriemerovaním celkovej mesačnej mzdy účtovanej na mesačnej báze a jej vydelením priemernou dennou úrovňou zamestnanosti v štvrťroku.

Okrem údajov o priemerných hrubých nominálnych mesačných mzdách sa publikujú aj indexy reálnej mzdy oproti rovnakému štvrťroku predchádzajúceho roka (reálne mzdy sú nominálne mzdy deflované indexom spotrebiteľských cien za rovnaké obdobie).

5.1.3 Opisné charakteristiky pre hmes-mzda

Na základe údajov z výberového zisťovania „Štvrťročný výkaz o cene práce“ sme sledovali ukazovateľ *hrubej mesačnej mzdy*, o ktorom sme predpokladali, že nepochádza z normálneho rozdelenia.

V súbore sa nachádzalo spolu 956844 pracovníkov. Opisné charakteristiky sú uvedené vo výstupe z programu *Statgraphics Centurion* v tabuľke 5.1.

Tabuľka 5.1 Opisné charakteristiky ukazovateľa „*hrubá mesačná mzda*“ v 2. polroku 2009

Count	956844
Average	812,106
Median	659,333
Mode	295,5
Variance	642891
Standard deviation	801,805
Coeff. of variation	98,73%
Standard error	0,819687
MAD	202,198
Minimum	6,21167
Maximum	99144,7
Range	99108,48
Lower quartile	482,283
Upper quartile	911,056
Skewness	21590,8

(zdroj: Vlastné výpočty)

Podľa výsledkov v tabuľke môžeme povedať, že 50% pracovníkov má hrubú mesačnú mzdu menšiu alebo rovnú 659,33 Eur. Priemerná mesačná mzda na Slovensku v danom štvrtroku je 812,106 Eur, ktorá sa výrazne líši od mediánu. Najmenšia hrubá mesačná mzda na Slovensku v danom štvrtroku je 6,21 Eur. Ide zrejme o mesačnú mzdu pracovníka, ktorý pracoval len na čiastočný úväzok. Najväčšia hrubá mesačná mzda na Slovensku v danom štvrtroku je 99114,7 Eur. Môže ísť napríklad o netypický príjem top manažéra. Pri výpočte hodnoty výberového priemeru sa však tieto a podobné hodnoty brali do úvahy. Z toho je odvodená aj „obrovská“ hodnota rozpätia miezd⁵³, ktorá je 99108,48 EUR. Prvý kvartil v danom súbore je 482,28, čo znamená, že na Slovensku v danom štvrtroku malo 25% pracovníkov mesačnú mzdu menšiu alebo rovnú 482,28 Eur. Tretí kvartil v danom súbore je 911,06 Eur, čo znamená, že na Slovensku v danom štvrtroku malo 75% pracovníkov mesačnú mzdu menšiu alebo rovnú 911,06 Eur. Smerodajná odchýlka je 801,805 EUR, variačný koeficient je 98,73%. Hodnota koeficienta šikmosti je 21590,8, ktorá hovorí, že hrubá mesačná mzda nemá symetrické rozdelenie a jej rozdelenie je výrazne pravostranné zošikmenie.

Zdá sa, že aritmetický priemer nie je v tomto prípade najlepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy.

Podľa tabuľky 5.2 vidíme vysokú koncentráciu v rozdelení mzdy a v sledovanom pozorovaní existujú prvky, ktoré majú neobvyklé veľké hodnoty a zatiaľ ich môžeme považovať za potenciálne odľahlé hodnoty. Na objasnenie toho problému sa pozrieme na histogram rozdelenia mzdy.

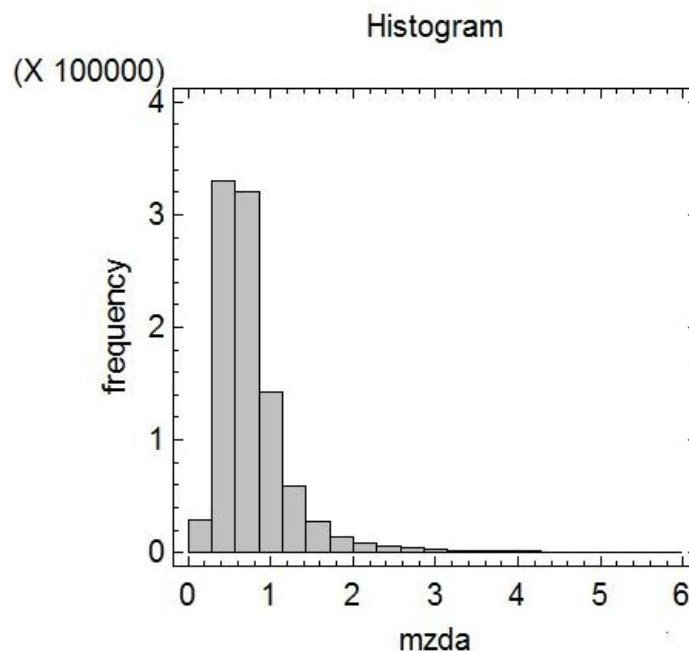
⁵³ Ide o rozdiel medzi najväčšou a najmenšou mzdou.

Tabuľka 5.2 Rozdelenie „hrubá mesačná mzda“

	Lower	Upper			Relative	Cumulative	Cum, Rel,
Class	Limit	Limit	Midpoint	Frequency	Frequency	Frequency	Frequency
	at or below	0		0	0	0	0
1	0	285,714	142,857	13214	0,0184	13214	0,0184
2	285,714	571,429	428,571	228568	0,318	241782	0,3364
3	571,429	857,143	714,286	247552	0,3445	489334	0,6809
4	857,143	1142,86	1000	117876	0,164	607210	0,8449
5	1142,86	1428,57	1285,71	50594	0,0704	657804	0,9153
6	1428,57	1714,29	1571,43	22278	0,031	680082	0,9463
7	1714,29	2000	1857,14	11541	0,0161	691623	0,9624
8	2000	2285,71	2142,86	7146	0,0099	698769	0,9723
9	2285,71	2571,43	2428,57	4719	0,0066	703488	0,9789
10	2571,43	2857,14	2714,29	3341	0,0046	706829	0,9835
11	2857,14	3142,86	3000	2321	0,0032	709150	0,9868
12	3142,86	3428,57	3285,71	1679	0,0023	710829	0,9891
13	3428,57	3714,29	3571,43	1236	0,0017	712065	0,9908
14	3714,29	4000	3857,14	1009	0,0014	713074	0,9922
15	4000	4285,71	4142,86	795	0,0011	713869	0,9933
16	4285,71	4571,43	4428,57	644	0,0009	714513	0,9942
17	4571,43	4857,14	4714,29	504	0,0007	715017	0,9949
18	4857,14	5142,86	5000	450	0,0006	715467	0,9955
19	5142,86	5428,57	5285,71	366	0,0005	715833	0,9961
20	5428,57	5714,29	5571,43	276	0,0004	716109	0,9964
21	5714,29	6000	5857,14	248	0,0003	716357	0,9968
	above	6000		2309	0,0032	718666	1

(zdroj: vlastné výpočty)

Podľa tvaru histogramu na obrázku 5.1 môžeme predpokladať, že výber nepochádza z normálneho rozdelenia. Histogram nám ukazuje, že hrubá mesačná mzda má výrazne pravostranné zošikmenie a neobvyklé veľké hodnoty, ktoré považujeme za potenciálne odľahlé hodnoty.



Obr. 5.1 Histogram mzdy
(zdroj: vlastné výpočty)

Viac informácií o tomto predpoklade uvidíme v nasledovných tabuľkách výstupu procedúry *Distribution fitting* z *Statgraphics Centurion* (*Analyze/variable data/Distribution Fitting*).

Statgraphics Centurion umožňuje navrhnúť normálne rozdelenie s ľubovoľnými parametrami a testuje, či je to vhodné rozdelenie.

Tabuľka 5.3 Výstup z procedúry *Distribution Fitting*

Fitted Distributions	
Normal	
mean = 850,715	
standard deviation = 906,076	

(zdroj: výstup z *Statgraphics*)

5.1.4 Detekcia odľahlých údajov pre hmes-mzda

Detekciu odľahlých údajov môžeme uskutočniť cez program *Statgraphics Centurion* v ponuke *Analyze/ Variable data /Outlier Identification*. Prvým výstupom v *Statgraphics* sú

netradičné charakteristiky polohy daného znaku s 5% -ným zastrihnutím. Ďalší výstup nám ukazuje testy na zisťovanie odľahlých údajov.

<u>Outlier Identification - hmes_mzda</u>	
Data variable: hmes_mzda	
956 844 values ranging from 6,211 to 99114,7	
Number of values currently excluded: 0	
Location estimates Trimming: 5.0%	
Sample mean	812,106
Sample median	659,333
Trimmed mean	722,425
Winsorized mean	751,725
Scale estimates	
Sample std. deviation	801,805
MAD/0,6745	299,775
Sbi	316,077
Winsorized sigma	406,723

Obr. 5.2 Výstup z procedúry Outlier Identification
(zdroj: výstup z Statgraphics)

Podľa obrázku 5.2 vidíme, že zastrihnutý priemer mzdy je 722,425 Eur s 5% zastrihnutím a táto hodnota je menšia ako priemerná mzda v pôvodnej forme. Winsorizovaný priemer je 751,725, čo je viacej ako zastrihnutý priemer, ale stále menej ako priemerná mzda v pôvodnej forme.

Metóda založená na Dixonovom teste a MADN

Opačný ako Grubbsov test, ktorý testuje súbor ako celok sú Dixonov test a test založený na MADN (*meadian absolute deviation*), ktoré testujú každú hodnotu v súbore a zisťujú, či je odľahlá alebo nie je.

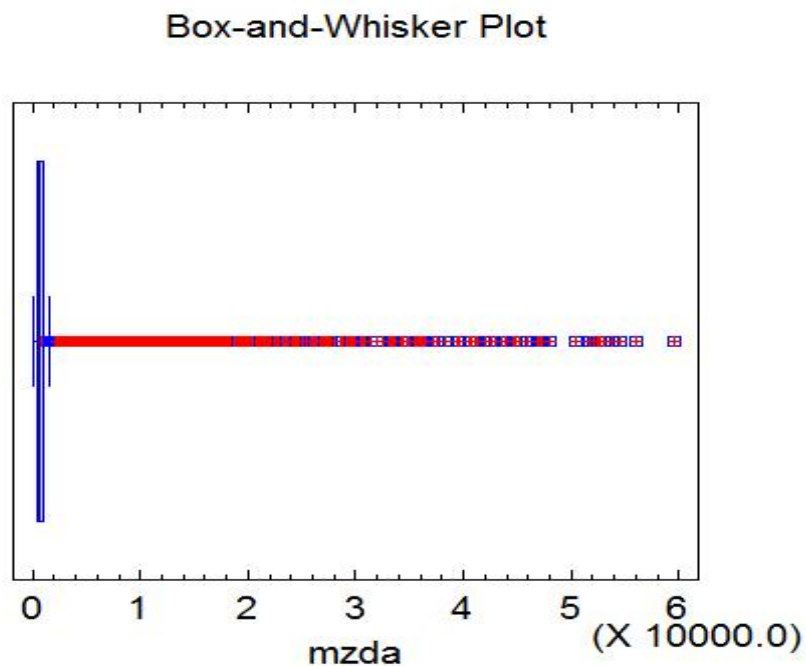
Podľa tabuľky 5.4 vidíme, že v tomto teste sú negatívne výsledky pre daný pozorovaný súbor. To znamená, že v našom sledovanom súbore existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu našich induktívnych odhadov a iných analýz.

Tabuľka 5.4 Testy z procedúry Outlier Identification

		<i>Studentized Values</i>	<i>Studentized Values</i>	<i>Modified</i>
<i>Row</i>	<i>Value</i>	<i>Without Deletion</i>	<i>With Deletion</i>	<i>MAD Z-Score</i>
1	6,21167	-1,0051	-1,0051	-2,17871
2	13,7058	-0,995753	-0,995754	-2,15371
3	13,8308	-0,995598	-0,995599	-2,15329
4	15,0362	-0,994094	-0,994095	-2,14927
5	16,6616	-0,992067	-0,992068	-2,14385
...				
956840	53108,9	65,2238	65,3693	174,963
956841	53864,1	66,1657	66,3177	177,483
956842	54546,2	67,0164	67,1742	179,758
956843	55941,4	68,7565	68,927	184,412
956844	59648,3	73,3797	73,5871	196,778

(zdroj: výstup z Statgraphics)

Metóda založená na grafe *Box-plot*



Obr. 5.3 Box - plot hrubej mesačnej mzdy

(zdroj: výstup z Statgraphics)

Box-plot je grafické znázornenie rozdelenia dát. Graf zobrazuje dolný kvartil (Q_1) a horný kvartil (Q_3) spolu s mediánom. Medián je 50. percentil. Dolný kvartil je 25. percentil, a horný kvartil je 75. percentil. Horné a dolné ploty (lower and upper fences) sa zvyčajne nastavujú na pevnú vzdialenosť od kvartilového rozpätia $R_Q^4 = (Q_3 - Q_1)$.

Podľa grafu vidíme veľa potenciálnych odľahlých hodnôt, ktoré ležia za vnútornými hranicami a jasných odľahlých hodnôt.

Metóda založená na kvartilovom rozpätí R_Q

Metóda využíva pri formulácii robustného pravidla na detekciu odľahlých údajov kvartilové rozpätie:

$$R_Q = x_{3/4} - x_{1/4} = 911,05 - 482,28 = 428,77$$

kde $x_{\frac{3}{4}}$ je horný kvartil.

$x_{\frac{1}{4}}$ – dolný kvartil.

Hodnota je odľahlá, keď:

- je väčšia alebo rovná $(x_{\frac{3}{4}} + 1,5 R_Q)$, alebo
- je menšia alebo rovná $(x_{\frac{1}{4}} - 1,5 R_Q)$.
-

Podľa tabuľky 5.1 poznáme, že maximálna hodnota v súbore je 99114,7, čo je oveľa väčšia než $1,5R_Q = 659,76$. Týmto výsledkom môžeme potvrdiť existenciu odľahlých hodnôt v pozorovanom súbore mzdy.

Zo všetkých uvedených testov na detekcie odľahlých údajov môžeme potvrdiť, že v sledovanom súbore mzdy existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu odhadu a sú príčinou na vytvorenie skreslených informácií. Tento výsledok je v súlade so skutočnosťou, keď

počet pracovníkov s extrémne vysokými mzdami je veľmi malý a ich vysoké mzdy sú len dočasné a nestabilné.

5.1.5 Netradičné charakteristiky polohy v analýze miezd

Netradičné charakteristiky polohy vypočítané pomocou Statgraphics Centurion

Pomocou procedúru *Analyze / Variable Data / One-Variable Analysis* v programe *Statgraphics Centurion* môžeme vypočítať hodnoty niektorých netradičných ukazovateľov pre analýzu hrubej mesačnej mzdy, ako *zastrihnutý priemer*, *Winsorizovaný priemer*, *mediánovú absolútnu odchýlku* a *Winsorizovanú sigma*. Zvolili sme 5% zastrihnutie a dostaneme:

Tabuľka 5.5 Výstup zo Statgraphics pre netradičné charakteristiky s 5% zastrihnutím

Count	861159
Average	722,426
Median	659,333
Mode	440,681
5% Trimmed mean	722,425
5% Winsorized mean	751,725
Standard deviation	292,91
Coeff. of variation	40,55%
Standard error	0,31564
5% Winsorized sigma	385,851
MAD	179,726
Minimum	315,055
Maximum	1715,78
Range	1400,72
Lower quartile	499,985
Upper quartile	875,39
Skewness	-526,107

(zdroj: výstup z Statgraphics)

Podľa tabuľky 5.5 dostaneme 5% zastrihnutý priemer rovný 722,426, čo znamená, že keby sme odstrihli z daného súboru 5%, potom priemerná mesačná hrubá mzda na Slovensku v danom štvrtroku je 722,426 Eur. Hodnota 5%-ného zastrihnutého priemeru je 722,426EUR, teda skoro o 90 EUR menšia ako predtým. Výrazne menšia je aj smerodajná odchýlka (292,91 EUR) a variačný koeficient (40,55%), rozpätie miezd je výrazne menšie (len 1400,72 EUR oproti 99 108,488 EUR predtým), medián sa nezmenil, kvartilové rozpätie sa zmenilo len málo. Rozdelenie miezd má oveľa 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%-ný zastrihnutý priemer je v tomto prípade lepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy ako aritmetický priemer. Winsorizovaná priemerná mesačná hrubá mzda na Slovensku v danom štvrtroku je 751,725 a je to viac ako zastrihnutý priemer lebo pri Winsorizovanom priemere sme nevyradili odľahlé hodnoty, ale sme ich len nahradili menšími hodnotami.

Tabuľka 5.6 Výstup zo Statgraphics pre netradičné charakteristiky s 10% zastrihnutím

Count	765476
Average	699,23
Median	659,333
Mode	440,681
10% Trimmed mean	699,229
10% Winsorized mean	726,373
Standard deviation	225,883
Coeff. of variation	32,30%
Standard error	0,258177
10% Winsorized sigma	330,571
MAD	157,848
Minimum	367,748
Maximum	1302,16
Range	934,411
Lower quartile	517,664
Upper quartile	843,789
Skewness	207,681

(zdroj: výstup z Statgraphics)

Teraz preskúmame 10%-ný zastrihnutý priemer. Výsledky analýzy súboru miezd, v ktorom bolo 10% najväčších a 10% najmenších miezd odstrihnutých sú v tabuľke 5.6. Zmenila sa šikmosť, z negatívnej na pozitívnu a je už len 207,681. Rozpätie miezd klesá na 934,411 EUR, kvartilové rozpätie sa zmenilo minimálne, medián sa prakticky nezmenil.

Hodnota 10%-ného zastrihnutého priemeru je 699,23 EUR, smerodajná odchýlka sa znížila na 225,883 EUR a variačný koeficient je 32,30%.

Nakoniec preskúmame charakteristiky polohy s 20%-ným zastrihnutím. V tabuľke 5.7 sú výsledky analýzy súboru miezd, v ktorom bolo 20% najväčších a 20% najmenších miezd „odstrihnutých“. Hodnota 20%-ného zastrihnutého priemeru znovu klesá na 676,719 EUR a len málo sa líši od mediánu, ktorý sa nezmenil, je stále 659,333 EUR. Modus sa zvýšil od 400 EUR do 500 EUR. Rozpätie miezd pokleslo na 551,043 EUR, výrazne pokleslo aj kvartilové rozpätie a hlavne šikmosť, ktorá sa zmenila, z pozitívnej na negatívnu a má hodnotu len -103,8. Smerodajná odchýlka poklesla na 147,306 EUR a variačný koeficient na 21,77%.

Tabuľka 5.7 Výstup zo Statgraphics pre netradičné charakteristiky s 20% zastrihnutím

Count	574107
Average	676,719
Median	659,333
Mode	500
20% Trimmed mean	676,719
20% Winsorized mean	694,79
Standard deviation	147,306
Coeff. of variation	21,77%
Standard error	0,194413
20% Winsorized sigma	270,414
MAD	115,468
Minimum	446,377
Maximum	997,42
Range	551,043
Lower quartile	553,023
Upper quartile	788,068
Skewness	-103,8

(zdroj: výstup z Statgraphics)

Zásadnou otázkou je, aký podiel údajov by sme mali odstihnúť. Pri riešení mnohých praktických problémov je vhodné použiť 20%-ný výberový zastrihnutý priemer, ktorý často poskytuje lepší výsledok ako výberový priemer a medián (Wilcox, s. 63).

Podľa výsledkov našej analýzy práve odstihnutie 20% údajov z každej strany rozdelenia poskytuje najkompaktnejší opis rozdelenia miezd. Podobne, 20%-ný zastrihnutý priemer – 676,719 EUR, dobre charakterizuje „typickú mesačnú mzdu“ zamestnanca v sledovanom období.

Iné netradičné charakteristiky polohy

Na analýzu hrubej mesačnej mzdy aplikujeme viacero prístupov a metód a pri niektorých výpočtoch musíme priamo využiť Microsoft Office Excel.

MADN a výpočet M-estimator

$$\frac{MAD}{0,6745} = MADN = \frac{202,198}{0,6745} = 299,745$$

Výpočet M-estimátora

Nech n_1 je počet pozorovaní X_i , pre ktoré:

$$\frac{(X_i - M)}{MADN} < -K \quad \text{a} \quad n_2 \text{ je počet pozorovaní, pre ktoré } \frac{(X_i - M)}{MADN} > K$$

M je medián výberového súboru .

Najčastejšie sa používa: $K = 1,28$. Potom jednokrokový M-estimátor polohy (založený na Huberovej funkcii Ψ)

Používame Microsoft Office Excel. V prvom kroku vypočítame hodnotu $\frac{(X_i - M)}{MADN}$ pre každý údaj v pozorovanom výbere. V druhom kroku budeme porovnávať $K = 1,28$ so všetkými

vypočítanými hodnotami aby sme mohli určiť počet n_1 a n_2 . Pri výpočte M-estimátora sa použili hodnoty v tabuľke 5.8. Z pravej strany rozdelenia sa odstránilo 170475 a z ľavej strany rozdelenia 24504 odľahlých hodnôt.

Tabuľka 5.8 Výpočet M-estimátora

MAD	202,198
Koeficient	0,6745
MADN	299,774
Median	659,333
n_1	24504
n_2	170475
K	1,28
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	472189611,9

(zdroj: vlastné výpočty)

Dosadíme do vzorca na výpočet M-estimátora dostaneme:

$$\hat{\mu}_{os} = \frac{K(\text{MADN})(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = 693,299$$

Dostali sme ďalšiu charakteristiku, pomocou ktorej možno dobre charakterizovať typickú mesačnú mzdu zamestnanca na Slovensku v sledovanom období. Hodnota 693,299 EUR sa len málo líši od 20%-ného zastrihnutého priemeru 676,719 EUR, ale výrazne od aritmetického priemeru, ktorý je až 812,106 EUR. Zdá sa že hodnotu aritmetického priemeru výrazne ovplyvnil menší počet netypicky vysokých miezd, čím sa výrazne zmenšila jeho výpovedná schopnosť ako typickej hodnoty.

Výpočet modifikovaného M estimátora

Podobný prístup ako pri výpočte M-estimátor používame Microsoft Office Excel na všetky výpočty, aby sme mohli nájsť chýbajúce prvky, ktoré potrebujeme na výpočet M-estimátora.

V prvom kroku vypočítame hodnotu $\frac{(X_i - M)}{MADN}$ pre každý údaj v pozorovanom výbere.

V druhom kroku budeme porovnávať $K = 2,24$ pre modifikovaný M-estimátor so všetkými vypočítanými hodnotami, aby sme mohli určiť počet n_1 a n_2

Dosadíme do vzorca na výpočet modifikovaného M-estimátora. Dostaneme:

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = 658,355$$

Tabuľka 5.9 Výpočet modifikovaného M-estimátora

MAD	202,198
Koeficient	0,6745
MADN	299,775
Median	659,333
n_1	0
n_2	90379
K	2,24
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	570441925

(zdroj: vlastné výpočty)

Hodnota modifikovaného M-estimátora hrubej mesačnej mzdy na Slovensku v danom štvrtroku je 658,355 Euro. Hodnota 658,355 EUR sa len málo líši od 20%-ného zastrihnutého priemeru 676,719 EUR, ale výrazne od aritmetického priemeru, ktorý je až 812,106 EUR. Dostali sme ešte jednu charakteristiku, pomocou ktorej možno dobre charakterizovať typickú mesačnú mzdu zamestnanca na Slovensku v sledovanom období.

Budeme skúmať údaje, ktoré sú získané len z veľkých firiem, aby sme mohli porovnať presnosť induktívnych úsudkov s presnými výsledkami z vyčerpávajúceho zisťovania. Nové metódy a prístupy na skúmanie rozličných charakteristík polohy budeme ilustrovať na analýze mesačných miezd 718 666 zamestnancov všetkých veľkých firiem⁵⁴ v SR v 2. polroku 2009. Ide teda o výsledky vyčerpávajúceho zisťovania⁵⁵.

5.2 Analýza miezd vo veľkých firmách

5.2.1 Opisné charakteristiky pre hmes-mzda vo veľkých firmách

Tabuľka 5.10 Hodnoty charakteristík pre všetky údaje vo veľkých firmách.

Count	718666
Average	850,715
Median	684,995
Mode	307,7
Standard deviation	906,076
Coeff. of variation	106,51%
Minimum	29,01
Maximum	99114,7
Range	99085,7
Lower quartile	510,4
Upper quartile	950,24
Skewness	11452,2

(zdroj: výstup z Statgraphics. vlastné výpočty)

⁵⁴ Za veľké sa považujú firmy, ktoré majú 250 a viac zamestnancov.

⁵⁵ Údaje (anonymizované), poskytla firma Trexima Bratislava.

V súbore sa nachádzalo spolu 718 666 pracovníkov. Opisné charakteristiky sú uvedené vo výstupe z programu *Statgraphics Centurion* v tabuľke 5.10.

Podľa výsledkov v tabuľke môžeme povedať, že 50% pracovníkov má mesačnú mzdu menšiu alebo rovnú 684,995 Eur. Priemerná mesačná mzda na Slovensku v danom štvrtroku je 850,715 Eur, ktorá sa výrazne líši od mediánu. Minimálna mesačná mzda na Slovensku v danom štvrtroku je 29,01 Eur. Ide zrejme o mesačnú mzdu pracovníka, ktorý pracoval len na čiastočný úväzok.

Maximálna mesačná mzda na Slovensku v danom štvrtroku je 99 114,7 Eur. Môže ísť napríklad o netypický príjem top manažéra. Pri výpočte hodnoty výberového priemeru sa však tieto a podobné hodnoty brali do úvahy. Z toho je odvodená aj „obrovská“ hodnota rozpätia miezd⁵⁶, ktorá je 99 085,6 EUR. Prvý kvartil v danom súbore je 510,4, čo znamená, že na Slovensku v danom štvrtroku malo 25% pracovníkov mesačnú mzdu menšiu alebo rovnú 510,4 Eur.

Tretí kvartil v danom súbore je 950,24 Eur, čo znamená, že na Slovensku v danom štvrtroku malo 75% pracovníkov mesačnú mzdu menšiu alebo rovnú 950,24 Eur. Smerodajná odchýlka je 906,076 EUR, variačný koeficient je 106,51%. Hodnota koeficienta šikmosti je 11 452,2, ktorá hovorí, že hrubá mesačná mzda nemá symetrické rozdelenie a jej rozdelenie je výrazne pravostranné zošikmenie.

Zdá sa, že aj v tomto súbore aritmetický priemer nie je najlepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy.

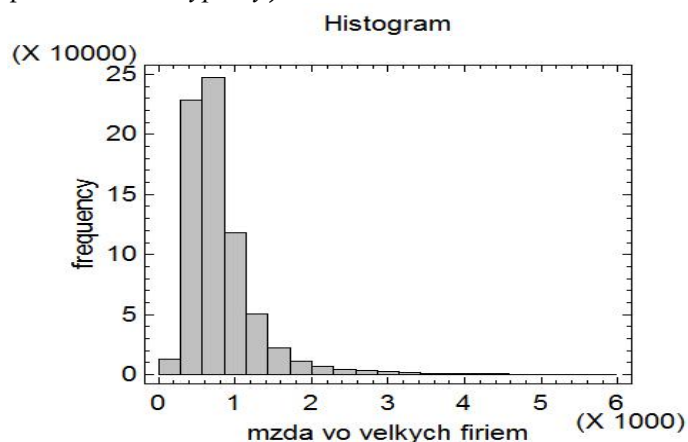
Podľa tabuľky 5.11 vidíme vysokú koncentráciu v rozdelení mzdy a v sledovanom pozorovaní existujú prvky, ktoré majú neobvyklé veľké hodnoty a zatiaľ ich môžeme považovať za potenciálne odlahlé hodnoty. Na overenie použijeme ich histogram a testy normality cez *Statgraphics* s ponukou *Distribution Fitting*.

⁵⁶ Ide o rozdiel medzi najväčšou a najmenšou mzdou.

Tabuľka 5.11 Rozdelenie „hrubá mesačná mzda“ vo veľkých firmách

	Lower	Upper			Relative	Cumulative	Cum, Rel,
Class	Limit	Limit	Midpoint	Frequency	Frequency	Frequency	Frequency
	at or below	0		0	0	0	0
1	0	57,1429	28,5714	122	0,0002	122	0,0002
2	57,1429	114,286	85,7143	1113	0,0015	1235	0,0017
3	114,286	171,429	142,857	2265	0,0032	3500	0,0049
4	171,429	228,571	200	3840	0,0053	7340	0,0102
5	228,571	285,714	257,143	5874	0,0082	13214	0,0184
6	285,714	342,857	314,286	22669	0,0315	35883	0,0499
7	342,857	400	371,429	41765	0,0581	77648	0,108
8	400	457,143	428,571	49340	0,0687	126988	0,1767
9	457,143	514,286	485,714	56457	0,0786	183445	0,2553
10	514,286	571,429	542,857	58337	0,0812	241782	0,3364
11	571,429	628,571	600	61056	0,085	302838	0,4214
12	628,571	685,714	657,143	57200	0,0796	360038	0,501
13	685,714	742,857	714,286	50735	0,0706	410773	0,5716
14	742,857	800	771,429	41967	0,0584	452740	0,63
15	800	857,143	828,571	36594	0,0509	489334	0,6809
16	857,143	914,286	885,714	32510	0,0452	521844	0,7261
17	914,286	971,429	942,857	26638	0,0371	548482	0,7632
18	971,429	1028,57	1000	23149	0,0322	571631	0,7954
19	1028,57	1085,71	1057,14	19100	0,0266	590731	0,822
20	1085,71	1142,86	1114,29	16479	0,0229	607210	0,8449
21	1142,86	1200	1171,43	13869	0,0193	621079	0,8642
	above	1200		97587	0,1358	718666	1

(zdroj: výstup z Statgraphics. vlastné výpočty)



Obr. 5.4 Histogram mzdy vo veľkých firmách

(zdroj: výstup z Statgraphics)

Podľa tvaru histogramu na obrázku 5.4 môžeme predpokladať, že výber nepochádza z normálneho rozdelenia. Viac informácií o tomto predpoklade uvidíme v nasledovných tabuľkách výstupu procedúry *Distribution fitting* zo *Statgraphics Centurion* (*Analyze/variable data/Distribution Fitting*).

5.2.2 Detekcia odľahlých údajov pre hmes-mzda vo veľkých firmách

Detekciu odľahlých údajov môžeme uskutočniť cez program *Statgraphics Centurion* v ponuke *Analyze/ Variable data /Outlier Identification*. Prvým výstupom z programu *Statgraphics* sú netradičné charakteristiky polohy daného znaku s 15%-ným zastrihnutím. Ďalší výstup nám ukazuje testy na zisťovanie odľahlých údajov. Podobnými prístupmi ako v predchádzajúcej časti ukážeme, že aj v tomto novom výbere existujú odľahlé hodnoty.

Data variable: mzda vo veľkych firmám 718666 values ranging from 29,01 to 99114,7 Number of values currently excluded: 0	
Location estimates	
Sample mean	850,715
Sample median	684,995
Trimmed mean	716,478
Winsorized mean	740,634
Scale estimates	
Sample std. deviation	906,076
MAD/0,6745	303,892
Sbi	320,701
Winsorized sigma	363,404

Obr. 5.5 výstup z procedúry Outlier Identification
(zdroj: výstup z *Statgraphics*)

Metóda založená na kvartilovom rozpätí R_Q

Metóda využíva pri formulácii robustného pravidla na detekciu odľahlých údajov kvartilové rozpätie⁵⁷:

$$R_Q = x_{3/4} - x_{1/4} = 950,24 - 510,4 = 439,84$$

kde $x_{\frac{3}{4}}$ je horný kvartil,

$x_{\frac{1}{4}}$ je dolný kvartil.

Hodnota je odľahlá, keď:

- je väčšia alebo rovná $(x_{\frac{3}{4}} + 1,5 R_Q)$ alebo
- je menšia alebo rovná $(x_{\frac{1}{4}} - 1,5 R_Q)$.

Znamená to, že hodnota v danom súbore je odľahlá, keď je väčšia ako 1610 alebo menšia ako -149,36. Podľa tabuľky *summary statistic* máme maximálnu hodnotu v pozorovanom súbore 99 114,7 > 1610. Na základe tejto metódy môžeme povedať, že daný súbor má odľahlé hodnoty.

Metóda založená na Dixonov test a MADN

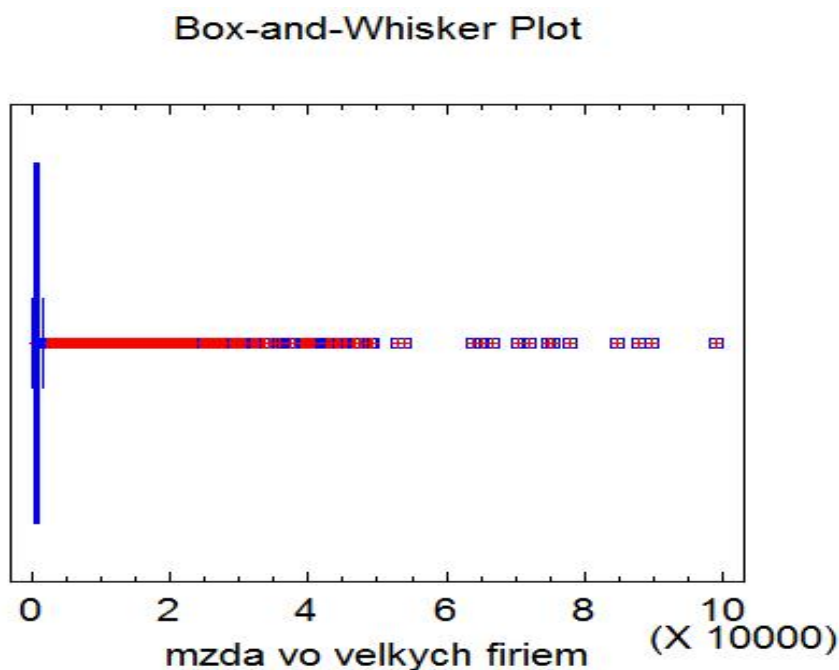
Tabuľka 5.12 Testy z procedúry Outlier Identification

		Studentized Values		Modified
Row	Value	Without Deletion	With Deletion	MAD Z-Score
1	29,01	-0,906883	-0,906885	-2,15861
3	30,36	-0,905394	-0,905395	-2,15417
2	30,36	-0,905394	-0,905395	-2,15417
4	31,96	-0,903628	-0,903629	-2,14891
5	33,2	-0,902259	-0,90226	-2,14483
...				
718662	77771,0	84,8939	85,3229	253,663
718663	84597,9	92,4285	92,9828	276,128
718664	87800,7	95,9633	96,5842	286,667
718665	89666,7	98,0227	98,6847	292,807
718666	99114,7	108,45	109,349	323,897

(zdroj: výstup z Statgraphics)

Podľa Tabuľky 5.12 vidíme, že v tomto teste sú negatívne výsledky pre daný pozorovaný súbor. To znamená, že v našom sledovanom súbore existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu našich induktívnych odhadov a iných analýz.

Metóda založená na *box-plot*



Obr. 5.6 Box - plot hrubej mesačnej mzdy
(zdroj: výstup z Statgraphics)

Zo všetkých uvedených testov na detekciu odľahlých údajov môžeme potvrdiť, že v sledovanom súbore mzdy existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu odhadu a sú príčinou na vytvorenie skreslených informácií. Tento výsledok je v súlade so skutočnosťou, keď počet pracovníkov s extrémne vysokými mzdami je veľmi malý a ich vysoké mzdy sú len dočasné a nestabilné. A je to dôvod, aby sme používali nové prístupy a metódy pri analýze tohto súboru.

5.2.3 Netradičné charakteristiky polohy v analýze miezd vo veľkých firmách

Netradičné charakteristiky polohy pomocou Statgraphics Centurion

Pomocou procedúry *Analyze / Variable Data / One-Variable Analysis* v programe *Statgraphics Centurion* môžeme vypočítať hodnoty niektorých netradičných ukazovateľov pre analýzu hrubej mesačnej mzdy, ako *zastrihnutý priemer*, *Winsorizovaný priemer*, *mediánovú absolútnu odchýlku* a *Winsorizovanú sigma*. Zvolili sme 5% zastrihnutie a dostaneme:

Tabuľka 5.13 Výstup zo Statgraphics pre netradičné charakteristiky s 5% zastrihnutím

Count	646800
Average	754,076
Median	684,995
Mode	370
5% Trimmed mean	754,075
5% Winsorized mean	784,164
Variance	89169
Standard deviation	298,612
Coeff. of variation	39,60%
5% Winsorized sigma	393,125
MAD	182,085
Minimum	342,97
Maximum	1766,95
Range	1423,98
Lower quartile	528,44
Upper quartile	912,7
Skewness	246,303

(zdroj: výstup z Statgraphics)

Podľa tabuľky 5.13 dostaneme 5% zastrihnutý priemer rovný 754,075, čo znamená, že keby sme odstrihli z daného súboru 5%, potom priemerná mesačná hrubá mzda vo veľkých firmách na Slovensku v danom štvrtroku je 754,075Eur. Hodnota 5%-ného 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 (39,60%), rozpätie miezd je výrazne menšie (len 1 423,98 EUR oproti 99 085,7 EUR predtým), medián sa nezmenil, kvartilové rozpätie sa zmenilo len málo. Rozdelenie miezd má oveľa menšiu variabilitu a je menej zošíkmené. Zdá sa, že toto rozdelenie lepšie charakterizuje celkovú situáciu v oblasti miezd

zamestnancov veľkých firiem v skúmanom období. Podobne sa zdá, že 5%-ný zastrihnutý priemer je v tomto prípade lepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy ako aritmetický priemer. Winsorizovaná priemerná mesačná hrubá mzda na Slovensku v danom štvrtroku je 784,164 EUR a je to viac ako zastrihnutý priemer lebo pri Winsorizovanom priemere sme nevyradili odľahlé hodnoty, ale sme ich len nahradili menšími hodnotami.

Tabuľka 5.14 Výstup zo Statgraphics pre netradičné charakteristiky s 10% zastrihnutím

Count	574934
Average	730,456
Median	685
Mode	1039,5
10% Trimmed mean	730,454
10% Winsorized mean	757,86
Standard deviation	53336,5
Coeff. of variation	230,947
Standard error	31,62%
10% Winsorized sigma	336,786
MAD	159,815
Minimum	392,79
Maximum	1342,18
Range	949,39
Lower quartile	546,08
Upper quartile	880,01
Skewness	-1335,38

(zdroj: výstup z Statgraphics)

Teraz preskúmame 10%-ný zastrihnutý priemer. Výsledky analýzy súboru miezd, v ktorom bolo 10% najväčších a 10% najmenších miezd odstrihnutých sú v tabuľke 5.14. Zmenila sa šikmosť, z pozitívnej na negatívnu. Rozpätie miezd klesá na 949,39 EUR, kvartilové rozpätie sa zmenilo minimálne, medián sa prakticky nezmenil.

Hodnota 10%-ného zastrihnutého priemeru je 730,454 EUR, smerodajná odchýlka znižuje na 230,947 EUR a variačný koeficient je 31,62%.

Nakoniec preskúmame charakteristiky polohy s 20%-ným zastrihnutím. V tabuľke 5.15 sú výsledky analýzy súboru miezd, v ktorom bolo 20% najväčších a 20% najmenších miezd

„odstrihnutých“. Hodnota 20%-ného zastrihnutého priemeru znovu klesá na 706,747 EUR a len málo sa líši od mediánu, ktorý sa nezmenil, je stále 706,747 EUR. Rozpätie miezd pokleslo na 563,33 EUR, výrazne pokleslo aj kvartilové rozpätie a hlavne šikmosť, ktorá sa zmenila opäť na pozitívnu a má hodnotu len -58,6193. Smerodajná odchýlka poklesla na 151,162 EUR a variačný koeficient na 21,39%.

Tabuľka 5.15 Výstup zo Statgraphics pre netradičné charakteristiky s 20% zastrihnutím

Count	431202
Average	706,75
Median	685
Mode	520,5
20% Trimmed mean	706,747
20% Winsorized mean	726,414
Standard deviation	22850,1
Coeff. of variation	151,162
Standard error	21,39%
20% Winsorized sigma	276,96
MAD	116,82
Minimum	474,25
Maximum	1037,58
Range	563,33
Lower quartile	580,5
Upper quartile	821,87
Skewness	58,6193

(zdroj: výstup z Statgraphics)

Podľa výsledkov nasej analýze práve odstrihnutie 20% údajov z každej strany rozdelenia poskytuje najkompaktnejší opis rozdelenia miezd. Podobne, 20%-ný zastrihnutý priemer – 706,747 EUR, dobre charakterizuje „typickú mesačnú mzdu“ zamestnanca v sledovanom období.

Iné netradičné charakteristiky polohy

Na analýzu hrubej mesačnej práce aplikujeme viacero prístupov a metód a pri niektorých výpočtoch musíme priamo využiť Microsoft Office Excel.

MADN a výpočet M-estimátora

$$\frac{\text{MAD}}{0,6745} = \text{MADN} = \frac{202,198}{0,6745} = 299,745$$

Výpočet M-estimátora

Nech n_1 je počet pozorovaní X_i , pre ktoré:

$$\frac{(X_i - M)}{\text{MADN}} < -K \quad \text{a} \quad n_2 \text{ je počet pozorovaní, pre ktoré } \frac{(X_i - M)}{\text{MADN}} > K$$

M je medián výberového súboru .

Najčastejšie sa používa: $K = 1,28$. Potom jednokrokový M-estimátor polohy (založený na Huberovej funkcii Ψ) je

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

Používame Microsoft Office Excel. V prvom kroku vypočítame hodnotu $\frac{(X_i - M)}{\text{MADN}}$ pre každý údaj v pozorovanom výbere. V druhom kroku budeme porovnávať $K = 1,28$ so všetkými vypočítanými hodnotami aby sme mohli určiť počet n_1 a n_2 .

Tabuľka 5.16 Výpočet M-estimátora

MAD	204,973
Koeficient	0,6745
MADN	303,888
Median	684,995
n_1	15495
n_2	131472
K	1,28
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	368842265,65

(zdroj: vlastné výpočty)

Pri výpočte M-estimátora sa použili hodnoty v tabuľke 5.16. Z pravej strany rozdelenia sa odstránilo 170 475 a z ľavej strany rozdelenia 24 504 odľahlých hodnôt.

Dosadíme do vzorca na vypočítanie M-estimátora dostaneme:

$$\hat{\mu}_{os} = \frac{K(\text{MADN})(n_2 - n_1) + \sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = 724,08$$

Dostali sme ďalšiu charakteristiku, pomocou ktorej možno dobre charakterizovať typickú mesačnú mzdu zamestnanca na Slovensku v sledovanom období. Hodnota 724,08 EUR sa len málo líši od 20%-ného zastrihnutého priemeru 676,719 EUR, 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.

Výpočet Modifikovaného M estimátora

Podobný prístup ako pri výpočte M-estimátor používame Microsoft Office Excel na všetky výpočty, aby sme mohli nájsť chýbajúce prvky, ktoré potrebujeme na výpočet M-estimátora. V prvom kroku vypočítame hodnotu $\frac{(X_i - M)}{\text{MADN}}$ pre každý údaj v pozorovanom výbere. V druhom kroku budeme porovnávať $K = 2,24$ pre Modifikovaný M-estimátor so všetkými vypočítanými hodnotami aby sme mohli určiť n_1 a n_2

Dosadíme do vzorca na vypočítanie modifikovaného M-estimátora. Dostaneme:

$$\hat{\mu}_{mom} = \frac{\sum_{i=n_1+1}^{n-n_2} X_{(i)}}{n - n_1 - n_2} = 688,874$$

Tabuľka 5.17 Výpočet Modifikovaného M-estimátora

MAD	204,973
Koeficient	0,6745
MADN	303,888
Median	684,995
n_1	0
n_2	68372
K	2,24
$\sum_{i=n_1+1}^{n-n_2} X_{(i)}$	447970516,51

(zdroj: vlastné výpočty)

Hodnota modifikovaného M-estimátora hrubej mesačnej mzdy na Slovensku v danom štvrtroku je 688,874 Euro. Hodnota 688,874 EUR sa len málo líši od 20%-ného zastrihnutého priemeru, ale výrazne od aritmetického priemeru, ktorý je až 850,715 EUR. Dostali sme ešte jednu charakteristiku, pomocou ktorej možno dobre charakterizovať typickú mesačnú mzdu zamestnanca na Slovensku v sledovanom období.

5.2.4 Bootstrapový interval spoľahlivosti pre hmes-mzda

Bootstrapové intervaly percentilovej metódy

Používame bootstrap metódu na odhad 95% intervalu spoľahlivosti pre strednú hodnotu a smerodajnú odchýlku *hrubej mesačnej mzdy* pomocou procedúry *Analyze / Variable Data / One-Variable Analysis / Confidence Intervals* v programe *Statgraphics Centurion*. Pre nastavenie percent intervalu spoľahlivosti a počtu prevzorkovania pri využití bootstrapovej metódy, sme prostredníctvom *Pane options* v dialógom okne označili voľbu *Include Bootstrap Number of Subsamples*, kde sme mohli nastaviť požiadavky podľa potreby.

V tomto príklade sme zvolili 95% interval spoľahlivosti a 1000^{58} krát prevzorkovanie s rozsahom výberu $n = 20$. Dostaneme:

Tabuľka 5.18 Bootstrapový interval spoľahlivosti

	Confidence Interval
Mean	(836,689;866,283)
Standard deviation	(786,728;1025,946)

(zdroj: výstup z Statgraphics)

Z tabuľky 5.18 vidíme, že s 95% spoľahlivosťou priemerná hrubá mesačná mzda na Slovensku v danom štvrtroku leží v intervale (836,689;866,283). Interval spoľahlivosti priemernej mesačnej mzdy málo pohybuje od skutočnej priemernej hrubej mzdy 850,715 EUR získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem). Podobne ako pri priemere, dostaneme smerodajnú odchýlku hrubej mesačnej mzdy na Slovensku v danom štvrtroku, ktorá leží v intervale (786,728;1025,946). A tento interval tiež veľmi dobre reprezentuje skutočnú smerodajnú odchýlku 906,076 získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem).

5.2.5 Aplikácia bootstrapovej metódy na výpočet intervalu spoľahlivosti pre zastreihnutú strednú hodnotu a charakteristiku polohy, založenú na M-estimátore

Aplikácia bootstrapovej metódy na výpočet intervalu spoľahlivosti pre zastreihnutú strednú hodnotu

V predchádzajúcej časti sme zistili, že mzda nespĺňa predpoklad o normalite, ktorý je veľmi dôležitý na odhadovanie intervalu spoľahlivosti. Avšak vďaka metóde bootstrap dokážeme odhadnúť interval spoľahlivosti pre dané pozorované hodnoty aj keď nemajú normálne rozdelenie pravdepodobnosti.

⁵⁸ $\beta = 1000$ je stanovený na odhadovaní strednej hodnoty a štandardnej odchýlky (Wilcox, s. 210, 212)

Použijeme program S-plus 8 na výpočet intervalu spoľahlivosti pre 5%-nú zastrihnutú strednú hodnotu, v ktorom sme zvolili prevzorkovaný súbor s rozsahom $n = 20$ a zároveň sme nastavili počet prevzorkovaní $B = 2000$ ⁵⁹ dostaneme:

```
Trimmed 5%
*** Bootstrap Results ***
Call:
bootstrap(data = SDF1, statistic = trimmed mean, B = 2000, seed = 20, trace = F, assign.frame1
= F, save.indices = F)

Number of Replications: 2000

Summary Statistics:

```

	Observed	Bias	Mean	SE
mean	754,1	0,00845	754,1	0,369

```
BCa Percentiles:

```

	2.5%	5%	95%	97.5%
mean	753,402	753,499	754,735	754,846

Obr. 5.7 Výstup z S-plus interval spoľahlivosti mesačnej mzdy s 5% zastrihnutím
(zdroj: výstup z S-plus)

S 95% spoľahlivosťou 5% zastrihnutá stredná hodnota miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 753,402 EUR do 754,846 EUR. Tento interval veľmi dobre reprezentuje skutočný 5%-ný zastrihnutý priemer 754,076 získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem s 5% zastrihnutím).

⁵⁹ Pri výpočte intervalu spoľahlivosti pre zastrihnutú strednú hodnotu zvolili $n = 20$ a $B = 2000$ (Wilcox, s. 224)

10% trimmed

*** Bootstrap Results ***

Call:

bootstrap(data = SDF1, statistic = trimmed mean, B = 2000, seed = 20, trace = F,
assign.frame1 = F, save.indices = F)

Number of Replications: 2000

Summary Statistics:

	Observed	Bias	Mean	SE
mean	730,5	0.00767	730,5	0,307

BCa Percentiles:

	2.5%	5%	95%	97.5%
Mean	729,852	729,963	730,960	731,069

Obr. 5.8 Výstup z S-plus interval spoľahlivosti mesačnej mzdy s 10% zastrihnutím
(zdroj: výstup z S-plus)

Podobne vypočítame interval spoľahlivosti pre 10%-nú zastrihnutú strednú hodnotu miezd vo veľkých firmách. S 95% spoľahlivosťou 10% zastrihnutá stredná hodnota miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 729,852 EUR do 731,069 EUR. Hodnota 10%-ného zastrihnutého priemeru získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem s 10% zastrihnutím) je 730,454 EUR, ktorá je z tohto intervalu.

Nakoniec preskúmame interval spoľahlivosti pre 20%-nú zastrihnutú strednú hodnotu miezd vo veľkých firmách, ktorú sme podľa predchádzajúcej analýzy považovali za veľmi dobrú hodnotu na charakterizovaní typickej mesačnej mzdy pracovníka veľkej firmy v sledovanom období

S 95% spoľahlivosťou 20% zastrihnutá stredná hodnota miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 706,319 EUR do 707,234 EUR. Hodnota 20%-ného zastrihnutého priemeru získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem s 20% zastrihnutím) je 706,747 EUR, ktorá tiež je zo získaného intervalu.

20% trimmed

*** Bootstrap Results ***

Call:

bootstrap(data = SDF1, statistic = trimmed mean, B = 2000, seed = 20, trace = F, assign.frame1 = F, save.indices = F)

Number of Replications: 2000

Summary Statistics:

	Observed	Bias	Mean	SE
mean	706,8	0.011	706,8	0,231

BCa Percentiles:

	2.5%	5%	95%	97.5%
Mean	706,319	706,395	707,141	707,234

Obr. 5.9 Výstup z S-plus interval spoľahlivosti mesačnej mzdy s 20% zastrihnutím
(zdroj: výstup z S-plus)

Aplikácia bootstrapovej metódy na výpočet intervalu spoľahlivosti pre charakteristiku polohy, μ_{os} , založenú na M-estimátore

Použijeme program S-plus 8 na výpočet intervalu spoľahlivosti pre μ_{os} , v ktorom sme zvolili prevzorkovaný súbor s rozsahom $n = 20$ a zároveň sme nastavili počet prevzorkovaní $B = 399$ dostaneme:

S 95% spoľahlivosťou $\hat{\mu}_{os}$ miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 723,432 EUR do 724,862 EUR. Porovnáme s presným výsledkom výpočtu M-estimátora 724,08 získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem). Vidíme, že interval veľmi dobre reprezentuje skutočnú hodnotu μ_{os}

*** Bootstrap Results ***				
Call: bootstrap(data = SDF1, statistic = mest, B = 399, seed = 20, trace = F, assign.frame1 = F, save.indices = F) Number of Replications: 399				
Summary Statistics:				
	Observed	Bias	Mean	SE
mest	724,1	0.00867	724,1	0,217
BCa Percentiles:				
	2.5%	5%	95%	97.5%
mest	723,432	723,528	724,696	724,862

Obr. 5.10 Výstup z S-plus interval spoľahlivosti M-estimátora mesačnej mzdy
(zdroj: výstup z S-plus)

Aplikácia bootstrapovej metódy na výpočet intervalu spoľahlivosti pre charakteristiku polohy μ_{mom} založenú na modifikovanom M-estimátore

Použijeme program S-plus 8 na výpočet intervalu spoľahlivosti pre μ_{mom} , v ktorom sme zvolili prevzorkovaný súbor s rozsahom $n = 20$ a zároveň sme nastavili počet prevzorkovaní $B = 500$ ⁶⁰ dostaneme:

S 95% spoľahlivosťou μ_{mom} vo veľkých firmách na SR v danom sledovanom období je z intervalu od 688,336 EUR do 689,646 EUR. Tento interval tiež veľmi dobre vyjadruje skutočný modifikovaný M-estimátor 688,874, ktorý je získaný zo súboru všetkých 718 666 zamestnancov veľkých firiem

⁶⁰ Na výpočte interval spoľahlivosti pre M-estimátor na istotu zvolíme minimálne $n = 20$ a $B = 500$ (Wilcox, s. 228)

*** Bootstrap Results ***				
Call: bootstrap(data = SDF1, statistic = mom, B = 500, seed = 20, trace = F, assign.frame1 = F, save.indices = F) Number of Replications: 500				
Summary Statistics:				
	Observed	Bias	Mean	SE
mom	688,9	0.00268	688,9	0,116
BCa Percentiles:				
	2.5%	5%	95%	97.5%
mom	688,336	688,446	689,568	689,646

Obr. 5.11 Výstup z S-plus interval spoľahlivosti M-estimátora mesačnej mzdy
(zdroj: výstup z S-plus)

Hodnota 20%-ného zastrihnutého priemeru je 706,747 EUR, hodnota M-estimátora je 724,08 EUR, hodnota modifikovaného M-estimátora je 688,874 EUR a hodnota mediánu⁶¹ 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 miezd⁶².

Pomocou bootstrapovej metódy môžeme určiť interval spoľahlivosti pre tieto nové netradičné charakteristiky polohy.

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

⁶¹ Keď uvažujeme o všetkých údajoch.

⁶² Terek, M. – Nguyen Dinh He: Možnosti využitia niektorých netradičných charakteristík v analýze miezd. *Ekonomické rozhlady* 1/2011.

samozrejme užitočné počítať aj charakteristiky variability, šikmosti, prípadne iné charakteristiky, ktoré poskytnú iné pohľady na celkovú situáciu⁶³.

⁶³ Terek, M. – Nguyen Dinh He: Možnosti využitia niektorých netradičných charakteristík v analýze miezd. *Ekonomické rozhlady* 1/2011

6 Prínos dizertačnej práce

Opíšeme najprv výsledky opisnej analýzy. Z pôvodného súboru cez tradičný prístup a opisné štatistiky sme dostali základné informácie o mzde: medián je 659,33, čo znamená, že 50% pracovníkov má mesačnú mzdu menšiu alebo rovnú 659,33 Eur. Priemerná mesačná mzda na Slovensku v danom štvrťroku je 812,106 Eur, ktorá sa výrazne líši od mediánu. Minimálna mesačná mzda na Slovensku v danom štvrťroku je 6,21 Eur. Ide zrejme o mesačnú mzdu pracovníka, ktorý pracoval len na čiastočný úväzok. Maximálna mesačná mzda na Slovensku v danom štvrťroku je 99114,7 Eur. Môže ísť napríklad o netypický príjem top manažéra, Pri výpočte hodnoty výberového priemeru sa však tieto a podobné hodnoty brali do úvahy. Z toho je odvodená aj „obrovská“ hodnota rozpätia miezd⁶⁴, ktorá je 99108,48 EUR. Prvý kvartil v danom súbore je 482,28, čo znamená, že na Slovensku v danom štvrťroku malo 25% pracovníkov mesačnú mzdu menšiu alebo rovnú 482,28 Eur. Tretí kvartil v danom súbore je 911,06 Eur, čo znamená, že na Slovensku v danom štvrťroku malo 75% pracovníkov mesačnú mzdu menšiu alebo rovnú 911,06 Eur. Smerodajná odchýlka je 801,805 EUR, variačný koeficient je 98,73%. Hodnota koeficienta šikmosti je 21590,8, ktorá hovorí, že hrubá mesačná mzda nemá symetrické rozdelenie a jej rozdelenie je výrazne pravostranné zošikmenie.

Histogram ukazuje, že hrubá mesačná mzda v tomto súbore má výrazne pravostranné zošikmenie a neobvyklé veľké hodnoty, ktoré považujeme za potenciálne odľahlé hodnoty.

Zo všetkých testov na detekcie odľahlých údajov vyplýva, že v sledovanom súbore mzdy existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu odhadu a sú príčinou na vytvorenie skreslených informácií. Tento výsledok je v súlade so skutočnosťou, keď počet pracovníkov s extrémne vysokými mzdami je veľmi malý a ich vysoké mzdy sú len dočasné a nestabilné.

Potom sme skúmali súbor s 5% zastrihnutím a dostali sme iné výsledky. Dostali sme 5% zastrihnutý priemer rovný 722,426, čo znamená, že keby sme odstihli z daného súboru 5%, a charakterizovali typickú mzdu pomocou zastrihnutého priemeru, potom typická mesačná hrubá mzda na Slovensku v danom štvrťroku je 722,426 Eur, skoro o 90 EUR menšia ako keď ju

⁶⁴ Ide o rozdiel medzi najväčšou a najmenšou mzdou.

charakterizujeme pomocou aritmetického priemeru. Výrazne menšia je aj smerodajná odchýlka (292,91 EUR) a variačný koeficient (40,55%), rozpätie miezd je výrazne menšie (len 1400,72 EUR oproti 99 108,488 EUR predtým), medián sa nezmenil, kvartilové rozpätie sa zmenilo len málo. Rozdelenie miezd má oveľa 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%-ný zastrihnutý priemer je v tomto prípade lepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy ako aritmetický priemer. Winsorizovaná priemerná mesačná hrubá mzda na Slovensku v danom štvrtroku je 751,725a je to viac ako zastrihnutý priemer lebo pri Winsorizovanom priemere sme nevyradili odľahlé hodnoty, ale sme ich len nahradili menšími hodnotami.

Nakoniec preskúmame charakteristiky polohy s 20%-ným zastrihnutím. V tabuľke 5.8 sú výsledky analýzy súboru miezd, v ktorom bolo 20% najväčších a 20% najmenších miezd „odstrihnutých“. Hodnota 20%-ného zastrihnutého priemeru znovu klesá na 676,719 EUR a len málo sa líši od mediánu, ktorý sa nezmenil, je stále 659,333 EUR. Rozpätie miezd pokleslo na 551,043 EUR, výrazne pokleslo aj kvartilové rozpätie a hlavne šikmosť, ktorá sa zmenila, z pozitívnej na negatívnu a má hodnotu len -103,8. Smerodajná odchýlka poklesla na 147,306 EUR a variačný koeficient na 21,77%.

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

Podľa výsledkov našej analýze práve odstrihnutie 20% údajov z každej strany rozdelenia poskytuje najkompaktnejší opis rozdelenia miezd. Podobne, 20%-ný zastrihnutý priemer – 676,719 EUR, dobre charakterizuje „typickú mesačnú mzdu“ zamestnanca v sledovanom období.

Dostali sme ďalšiu charakteristiku, pomocou ktorej možno dobre charakterizovať typickú mesačnú mzdu zamestnanca na Slovensku v sledovanom období. Hodnota 693,299 EUR sa len málo líši od 20%-ného zastrihnutého priemeru 676,719 EUR, ale výrazne od aritmetického priemeru, ktorý je až 812,106 EUR. Zdá sa že hodnotu aritmetického priemeru výrazne ovplyvnil

menší počet netypicky vysokých miezd, čím sa výrazne zmenšila jeho výpovedná schopnosť ako typickej hodnoty.

Hodnota modifikovaného M-estimátora hrubej mesačnej mzdy na Slovensku v danom štvrtroku je 658,355 Euro. Hodnota 658,355 EUR sa len málo líši od 20%-ného zastrihnutého priemeru 676,719 EUR, ale výrazne od aritmetického priemeru, ktorý je až 812,106 EUR. Dostali sme ešte jednu charakteristiku, pomocou ktorej možno dobre charakterizovať typickú mesačnú mzdu zamestnanca na Slovensku v sledovanom období.

V ďalšej časti sme preskúmali nový súbor, v ktorom údaje boli zistené len z veľkých firiem, aby sme mohli posúdiť presnosť výsledkov ktoré poskytuje aplikácia induktívnych metód. Charakteristiky polohy sme najprv vypočítali v základnom súbore pracovníkov veľkých firiem rozsahu 718 666 zamestnancov, získali sme teda výsledky vyčerpávajúceho zisťovania. Potom sme uplatnili metódy induktívnej štatistiky. Výsledky sme porovnali.

V tomto súbore existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu odhadu a sú príčinou na vytvorenie skreslených informácií. Tento výsledok ukazuje, že v tomto prípade aritmetický priemer nie je najlepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy. Súčasne 20%-ný zastrihnutý priemer – 706,747 EUR, ktorý odporúčal Wilcox veľmi dobre charakterizuje „typickú mesačnú mzdu“ zamestnanca veľkej firmy v sledovanom období.

Podobne ako v predchádzajúcej analýze môžeme potvrdiť, že v sledovanom súbore mzdy existujú odľahlé hodnoty, ktoré ovplyvňujú kvalitu odhadu a sú príčinou na vytvorenie skreslených informácií. Tento výsledok je v súlade so skutočnosťou, keď počet pracovníkov s extrémne vysokými mzdami je veľmi malý a ich vysoké mzdy sú len dočasné a nestabilné. A je to dôvod, aby sme používali nové prístupy a metódy pri analýze tohto súboru.

5% zastrihnutý priemer rovný 754,075, teda skoro o 100 EUR menšia ako hodnota aritmetického priemeru. Výrazne menšia je aj smerodajná odchýlka (298,612 EUR) a variačný koeficient (39,60%), rozpätie miezd je výrazne menšie (len 1423,98 EUR oproti 99 085,7 EUR predtým), medián sa nezmenil, kvartilové rozpätie sa zmenilo len málo. Rozdelenie miezd má oveľa 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%-ný zastrihnutý priemer je v tomto prípade lepšou charakteristikou „typickej“ mesačnej mzdy pracovníka veľkej firmy ako aritmetický priemer. Winsorizovaná priemerná mesačná hrubá mzda na Slovensku v danom štvrtroku je 784,164 EUR a je to viac ako zastrihnutý priemer lebo pri Winsorizovanom priemere sme nevyradili odľahlé hodnoty, ale sme ich len nahradili menšími hodnotami.

Podľa výsledkov vykonanej analýzy práve odstrihnutie 20% údajov z každej strany rozdelenia poskytuje najkompaktnejší opis rozdelenia miezd. Podobne, 20%-ný zastrihnutý priemer – 706,747 EUR, dobre charakterizuje „typickú mesačnú mzdu“ zamestnanca v sledovanom období.

Hodnota M-estimátora 724,08 EUR sa len málo líši od 20%-ného zastrihnutého priemeru 676,719 EUR, ale výrazne od aritmetického priemeru, ktorý je až 850,715 EUR.

Hodnota modifikovaného M-estimátora hrubej mesačnej mzdy na Slovensku v danom štvrtroku je 688,874 Euro. Hodnota 688,874 EUR sa len málo líši od 20%-ného zastrihnutého priemeru, ale výrazne od aritmetického priemeru, ktorý je až 850,715 EUR. Dostali sme ešte jednu charakteristiku, pomocou ktorej možno dobre tiež charakterizovať typickú mesačnú mzdu zamestnanca veľkej firmy na Slovensku v sledovanom období.

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, kvantily, 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 variability, šikmosti, prípadne iné charakteristiky, ktoré poskytnú iné pohľady na celkovú situáciu⁶⁵.

V 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.

⁶⁵ TEREK, M. – NGUYEN DINH HE: Možnosti využitia niektorých netradičných charakteristík v analýze miezd. *Ekonomické rozhľady* 1/2011.

Teraz si všimneme výsledky aplikácie niektorých induktívnych metód.

Aplikáciou bootstrapovej metódy na odhadovanie intervalu spoľahlivosti tohto súboru s rozsahom výberu $n = 20$ a počtom prevzorkovania $B = 1000$ sme zistili, že s 95% spoľahlivosťou priemerná hrubá mesačná mzda na Slovensku v danom štvrtroku leží v intervale (836,689;866,283). Interval spoľahlivosti pre očakávanú mesačnú mzdu pokrýva skutočnú priemernú hrubú mzdu 850,715 EUR získanú vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem). Podobne ako pri priemere, dostaneme štandardnú odchýlku hrubej mesačnej mzdy na Slovensku v danom štvrtroku, ktorá leží v intervale (786,728;1025,946).

Na výpočet intervalu spoľahlivosti sme použili program S-plus 8, v ktorom sme zvolili prevzorkovaný súbor s rozsahom $n = 20$ a zároveň sme nastavili počet prevzorkovaní $B = 2000$ pre zastrihnutú strednú hodnotu. S 95% spoľahlivosťou 5% zastrihnutá stredná hodnota miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 753,402 EUR do 754,846 EUR. Podobne vypočítame interval spoľahlivosti pre 10%-nú zastrihnutú strednú hodnotu miezd vo veľkých firmách. S 95% spoľahlivosťou 10% zastrihnutá stredná hodnota miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 729,852 EUR do 731,069 EUR. Hodnota 10%-ného zastrihnutého priemeru získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem s 10% zastrihnutím) je 730,454 EUR, a je z vypočítaného intervalu.

Nakoniec preskúmame interval spoľahlivosti pre 20%-nú zastrihnutú strednú hodnotu miezd vo veľkých firmách, ktorá podľa predchádzajúcej analýzy sme považovali za veľmi dobrú hodnotu na charakterizovaní typickej mesačnej mzdy pracovníka veľkej firmy v sledovanom období. S 95% spoľahlivosťou 20% zastrihnutá stredná hodnota miezd vo veľkých firmách na SR v danom sledovanom období je z intervalu od 706,319 EUR do 707,234 EUR. Hodnota 20%-ného zastrihnutého priemeru získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem s 20% zastrihnutím) je 706,747 EUR, teda je z vypočítaného intervalu.

Pokračovali sme v tomto procese aj pre M-estimátor. Použili sme program S-plus 8 na výpočet intervalu spoľahlivosti pre M-estimátor, v ktorom sme zvolili prevzorkovaný súbor s rozsahom $n = 20$ a zároveň sme nastavili počet prevzorkovaní $B = 399$ dostaneme: S 95% spoľahlivosťou μ_{os} je z intervalu od 723,432 EUR do 724,862 EUR. Porovnáme tento výsledok s presným výsledkom výpočtu M-estimátora 724,08 získaným vyčerpávajúcim skúmaním (na súbore všetkých 718 666 zamestnancov veľkých firiem). Vidíme, že interval pokrýva skutočnú hodnotu.

S 95% spoľahlivosťou μ_{mom} je z intervalu od 688,336 EUR do 689,646 EUR s $n = 20$ a $B = 500$.

Tento interval tiež pokrýva skutočnú hodnotu μ_{mom} 688,874, ktorý je získaný zo súboru všetkých 718 666 zamestnancov veľkých firiem.

Vidno, že bootstrapové percentilové intervaly spoľahlivosti poskytujú veľmi presné výsledky. Všetky 95%-né intervaly pokrývajú skutočné hodnoty charakteristík. Pritom rozsahy realizovaných výberov boli len $n = 20$. V niektorých situáciách možno náhodné vyčerpávajúce zisťovanie nahradiť výberovým skúmaním pri ktorom stačí preskúmať veľmi malú vzorku. Podobné postupy možno odporúčať napríklad pri niektorých marketingových prieskumoch.

ZÁVER

Cieľom dizertačnej práce bolo poskytnúť ucelený obraz o odľahlých údajoch a poskytnúť základné poznatky pre skúmanie odľahlých údajov na Slovensku a aplikovať opísané metódy na konkrétnych príkladoch z ekonomickej praxe.

Dizertačná práca je členená na teoretickú a aplikačnú časť. Teoretická časť sa skladá zo štyroch kapitol, ktoré obsahujú základné poznatky o odľahlých údajoch a prístupy k ich aplikácii. Praktická časť obsahuje aplikácie opísaných metód pri analýze hrubej mesačnej mzdy na Slovensku, ktoré sú uvedené v piatej kapitole.

Prvá kapitola prináša všeobecný prehľad o odľahlých údajoch a ich rozvojový trend v zahraničí. V tejto kapitole sú tiež uvedené klasické prístupy pri analýze mzdy na Slovensku.

V druhej kapitole sú formulované ciele dizertačnej práce.

Tretia kapitola rozoberá dva základné poznatky o odľahlých údajoch: detekcia odľahlých hodnôt a základné prístupy na ich odstránenie.

V prvej časti je opísaná konštrukcia metód detekcie odľahlých hodnôt. Tu sme uviedli jednoduché metódy na detekciu odľahlých hodnôt, ktoré sú založené na základe kvartilového rozpätia alebo grafu Box-plot. Pomocou Grubbsovho testu môžeme testovať odľahlé hodnoty v súbore ako celku. V tejto časti opisujeme aj ďalšie empirické metódy, ktoré sú založené na MAD alebo Z-hodnote a Dixonov princíp.

Druhá časť tejto kapitoly sa zaoberá odstránením odľahlých hodnôt v štatistickej analýze, kde je uvedený zastrehaný priemer, jednostranný zastrehaný priemer, Winsorizovaný priemer, jednostranný Winsorizovaný priemer.

V štvrtej kapitole sú uvedené nové prístupy a metódy indukčnej štatistiky s ohľadom na odľahlé údaje. Tieto metódy umožňujú vypočítať hodnoty opisných štatistík a odhadovať parametre základných súborov s ohľadom na odľahlé údaje. Kombinácia bootstrap metódy s

novými prístupmi pri analýze odľahlých hodnôt, nám umožnila skúmať odľahlé údaje na novej, vyššej úrovni.

Teoretickú časť dizertačnej práce uzatvárajú uvedené dve kapitoly venované najčastejšie odhadovaným parametrom – strednej hodnote a rozptylu. Pre obidva parametre sú podrobne opísané prístupy, teoretické základy a ilustračné príklady.

V aplikačnej časti práce sú použité údaje z výberového zisťovania ISCP 2009-2011 pod názvom „Štvrťročný výkaz o cene práce“. Väčšina výpočtov a testy boli realizované pomocou štatistických programov *Statgraphics Centurion* a *S-plus*, pomocné výpočty a triedenia boli urobené v tabuľkovom procesore Microsoft Office Excel. Realizovala sa opisná analýza a boli aplikované niektoré indukčnej štatistiky založené na bootstrape.

LITERATÚRA

- [1] ACUNA E. – RODRIGUEZ C. A.: *Meta analysis study of outlier detection methods in classification*. Department of Mathematics, University of Puerto Rico at Mayaguez: 2004.
- [2] BAKYTOVÁ, H. – URGON, M –KONTŠEKOVÁ, O: *Základy štatistiky*. Bratislava: Vydavateľstvo, Alfa, 1979.
- [3] BARBARA, D., CHEN, P.: *Using the fractal dimension to cluster datasets*. In: Proc. ACM KDD 2000, 260-264, 2000.
- [4] BARNETT, V. – LEWIS, T.: *Outliers in Statistical Data*. John Wiley& Sons, 1994. ISBN 10: 0471930946
- [5] BAY S. D. – SCHWABACHER, M.: *Mining distance-based outliers in near linear time with randomization and a simple pruning rule*. In Proc. of the ninth ACM SIGKDD Conference on Knowledge Discovery and Data Mining. USA: Washington, DC, 2003.
- [6] BOX G. E. P. – JENKINS G. M.: *Times Series Analysis, Forecasting and Control*. In Oakland, CA: Holden Day, 1976.
- [7] BREUNIG M.M. – KRIEGEL H.P. – NG R.T., SANDER J.: *Lof: Identifying densitybased local outliers*. In Proc. ACM SIGMOD Conf. 2000. ISBN 93–104, 2000.
- [8] CHERNICK, M. R.: *Bootstrap methods A Guide for Practitioners and Researchers*. Canada: John Wiley& Sons, 1994. ISBN 10: 0471930946
- [9] DAVISON, A., C., HINKLEY, D., V.: *Bootstrap methods and their application*. Cambridge University Press, 1997. ISBN: 9780521574716

- [10] DAVID H. A.: *Robust estimation in the presence of outliers*. In Robustness in Statistics, eds. R. L. Launer and G.N. Wilkinson, New York: Academic Press, ISBN 61-74, 1979.
- [11] DAVIES, L. – GATHER, U : *The identification of multiple outliers*. In Journal of the American Statistical Association, 1993. 88(423), ISSN 782-792.
- [12] DUMOUCHEL, W. – SCHONLAU, M.: *A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities*. In Proceedings of the 4th International Conference on Knowledge Discovery and Data-mining (KDD98), ISBN 1998. 189–193,
- [13] EFRON, B.: *Bootstrap methods: another look at the jackknife*. Ann. Statist, 1979. 7, 1 – 26.
- [14] FAWCETT, T. – PROVOST, F.: *Adaptive fraud detection*. In Data-mining and Knowledge Discovery, 1(3), ISBN 1997. 291–316.
- [15] FERGUSON, T. S.: *On the Rejection of outliers*. In Proceedings of the Fourth.
- [16] GRUBBS, F. E.: *Procedures for detecting outlying observations in Samples*. In Technometrics, ISBN 11, 1-21, 1969.
- [17] HADI, A. S.: *Identifying multiple outliers in multivariate data*. In Journal of the Royal Statistical Society. Series B, ISSN 1992. 54, 761-771.
- [18] HADI, A. S.: *A modification of a method for the detection of outliers in multivariate samples*. In Journal of the Royal Statistical Society, Series B, 56(2), 1994.
- [19] HALL, P.: *The bootstrap and the Edgeworth expansion*. Springer-Verlag, New York, 1992.

- [20] HAWKINS, D.: *Identification of Outliers*. In Chapman and Hall, 1980.
- [21] HAWKINS, S. – HE, H. X. – WILLIAMS, G. J. – BAXTER, R. A.: *Outlier detection using replicator neural networks*. In Proceedings of the Fifth International Conference and Data Warehousing and Knowledge Discovery (DaWaK02), Aixen Provence, France, 2002.
- [22] HUŤKA, V. – HORAKOVÁ, G.: *Teória pravdepodobnosti 1*, Bratislava : Vydavateľstvo EKONÓM, 2002. ISBN 80-225-1511-6.
- [23] HU, T. – SUNG, S. Y.: *Detecting pattern-based outliers*. In: *Pattern Recognition Letters*, 24, 3059-3068.
- [24] IGLEWICS, B. – MARTINEZ, J.: *Outlier Detection using robust measures of scale*. In *Journal of Statistical Computation and Simulation*, 15, ISBN 1982. 285-293.
- [25] JIN, W. – TUNG, A. – HAN, J.: *Mining top-n local outliers in large databases*. In Proceedings of the 7th International Conference on Knowledge Discovery and Data-mining (KDD01), San Francisco, CA, 2001.
- [26] KAUFMAN, L. – ROUSSEEUW, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York: 1990.
- [27] KOLLIOS, G. – GUNOPULOS, D. – KOUDAS, N. – BERCHTOLD, S.: *Efficient biased sampling for approximate clustering and outlier detection in large data sets*. In: *IEEE Transactions on Knowledge and Data Engineering*, 2003. 15 (5), ISBN 1170-1187.
- [28] KNORR E. – NG, R.: *A unified approach for mining outliers*. In Proceedings Knowledge Discovery KDD, 1997. ISBN 219-222.
- [29] KNORR, E. – NG, R. – TUCAKOV, V.: *Distance-based outliers*. In *Algorithms and applications*, VLDB Journal: Very Large Data Bases, 2000. ISSN 237-253.

[30] KNORR, E., NG., R., T., TUCAKOV, V.: *Distance-based outliers*. In: Algorithms and applications,”

VLDB Journal: Very Large Data Bases, 8(3-4):237-253, 2000.

[31] KNORR, E., NG., R., T., ZAMAR, R., H.: *Robust space transformations for distance based Operations*. In: Proceedings of the 7th International Conference on Knowledge Discovery and Data-mining (KDD01), 126-135, San Francisco, CA, 2001.

[32] KNORR, E., NG. R. : *Algorithms for mining distance-based outliers in large Datasets*. In: Proc. 24th Int. Conf. Very Large Data Bases (VLDB), 392-403, 24-27, 1998.

[33] LUNNEBORG, C., E.: *Data analysis by resampling: Concept and Applications*. Duxbury, 2000. ISBN 0-534-221-106

[34] LIU, H. – SHAH, S. – JIANG, W.: *On-line outlier detection and data cleaning*. In Computers and Chemical Engineering, 2004. 28, ISBN 1635–1647.

[35] LU, C. – CHEN, D. – KOU, Y.: *Algorithms for spatial outlier detection*. In Proceedings of the 3rd IEEE International Conference on Data-mining (ICDM'03), Melbourne, FL, 2003.

[36] MOONEY, C., Z., DUVAL, R.: *Bootstrapping: A nonparametric approach to statistical inference*. California: Sage publications, 1993. ISBN 080-395-381-X.

[37] NG., R.T., HAN, J.: *Efficient and Effective Clustering Methods for Spatial Data Mining*. In: Proceedings of Very Large Data Bases Conference, 144-155, 1994.

[38] PACÁKOVÁ, V. a kol.: *Štatistika pre ekonómov*. Bratislava: Vydavateľstvo EKONÓM, 2003. ISBN 80-89047-74-2.

- [39] PAPADIMITRIOU, S., KITAWAGA, H., GIBBONS, P.G., FALOUTSOS, C.: *LOCI: Fast Outlier Detection Using the Local Correlation Integral*. In: Intel research Laboratory Technical report no. IRP-TR-02-09, 2002.
- [40] PENNY, K. I. – JOLLIFFE, I. T.: *A comparison of multivariate outlier detection methods for clinical laboratory safety data*. In The Statistician 50(3), 2001. ISBN 295-308.
- [41] PERARSON, R. K.: *Outliers in process modeling and identification*. In: IEEE Transactions on Control Systems Technology, 10, 55-63, 2002.
- [42] RAMASWAMY, S., RASTOGI, R., SHIM, K.: *Efficient algorithms for mining outliers from large data sets*. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Dalas, TX, 2000.
- [43] ROSNER, B.: *On the detection of many outliers*. In Technometrics, 17, 1975, ISBN 221-227.
- [44] R. J. A. – LITTLE, & D, B. RUBIN: *Statistical Analysis With Missing Data*. John Wiley & Sons (1987), ISBN 0-471-80243-9.
- [45] ROUSSEEUW, P. – LEORY, A.: *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics, 1987.
- [46] SHAO J., TU, D.: *The jackknife and bootstrap*. Springer-Verlag, New York, 1995.
- [47] SHEKHAR, S. – LU, C. T. – ZHANG, P.: *Algorithms and Applications (A Summary of Results)*. In Proc. of the Seventh ACM-SIGKDD Conference on Knowledge Discovery and Data Mining, SF, CA, 2001.

[48] SHEKHAR, S. – LU, C. T. – ZHANG, P.: *Detecting Graph-Based Spatial Outlier*. In *Intelligent Data Analysis, an International Journal*, 6(5), ISSN 451–468, 2002.

[49] SHEKHAR, S., CHAWLA, S.: *A Tour of Spatial Databases*, Prentice Hall, 2002.

[50] TEREK, M., Horníková, A., Labudová, V.: *Hĺbková analýza údajov*. Bratislava: Iura Edition 2010, 265 s., ISBN 978-80-8078-336-5.

[51] TEREK, M.: *Analýza odlahlých 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 2010, CD ROM, ISBN 978-80-225-3076-7, s. 371 – 384.

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

[53] TEREK, M.(2010): *Netradičné charakteristiky polohy 2*. In: Materiál pre firmu Trexima Bratislava, v rámci riešenia projektu Európskeho sociálneho fondu, NP XIV-2 Zamestnanosť a sociálna inklúzia.

[54] TEREK, M. – NGUYEN DINH HE: Možnosti využitia niektorých netradičných charakteristík v analýze miezd. *Ekonomické rozhlady* 1/2011.

[55] TUKEY, J.W.: *Exploratory Data Analysis*. Addison-Wesley, 1977.

[56] WILLIAMS G. J., Huang Z.: *Mining the knowledge mine: The hot spots methodology for mining large real world databases*. In: Abdul Sattar, editor, *Advanced Topics in Artificial Intelligence*, volume 1342 of *Lecture Notes in Artificial Intelligence*, 340–348, Springer, 1997

PRÍLOHY

P. 1 Kritické hodnoty pre Dixonov test

Testovacia štatistika	N	Hladina významnosti α					
		0.30	0.20	0.10	0.05	0.01	0.005
$R_{10} = \frac{x_2 - x_1}{x_n - x_1}$	3	0.684	0.781	0.886	0.941	0.988	0.994
	4	0.471	0.56	0.679	0.765	0.889	0.926
	5	0.373	0.451	0.557	0.642	0.78	0.821
	6	0.318	0.386	0.482	0.56	0.698	0.74
	7	0.281	0.344	0.434	0.507	0.637	0.68
$R_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}$	8	0.318	0.385	0.479	0.554	0.683	0.725
	9	0.288	0.352	0.441	0.512	0.635	0.677
	10	0.265	0.325	0.409	0.477	0.597	0.639
$R_{21} = \frac{x_3 - x_1}{x_{n-1} - x_1}$	11	0.391	0.442	0.517	0.576	0.679	0.713
	12	0.37	0.419	0.49	0.546	0.642	0.675
	13	0.351	0.399	0.467	0.521	0.615	0.649
$R_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1}$	14	0.37	0.421	0.492	0.546	0.641	0.674
	15	0.353	0.402	0.472	0.525	0.616	0.647
	16	0.338	0.386	0.454	0.507	0.595	0.624
	17	0.325	0.373	0.438	0.49	0.577	0.605
	18	0.314	0.361	0.424	0.475	0.561	0.589
	19	0.304	0.35	0.412	0.462	0.547	0.575
	20	0.295	0.34	0.401	0.45	0.535	0.562
	21	0.287	0.331	0.391	0.44	0.524	0.551
	22	0.28	0.323	0.382	0.43	0.514	0.541
	23	0.274	0.316	0.374	0.421	0.505	0.532
	24	0.268	0.31	0.367	0.413	0.497	0.524
	25	0.262	0.304	0.36	0.406	0.489	0.516

(zdroj: <http://www.watpon.com/table/dixontest.pdf>)

P. 2 Kritické hodnoty G pre Grubbsov obojstranný test

N	G hodnota $\alpha=0.05$	G hodnota $\alpha=0.01$	N	G hodnota $\alpha=0.05$	G hodnota $\alpha=0.01$	N	G hodnota $\alpha=0.05$	G hodnota $\alpha=0.01$
3	1.1543	1.1547	15	2.5483	2.8061	80	3.3061	3.6729
4	1.4812	1.4962	16	2.5857	2.8521	90	3.3477	3.7163
5	1.715	1.7637	17	2.62	2.894	100	3.3841	3.754
6	1.8871	1.9728	18	2.6516	2.9325	120	3.4451	3.8167
7	2.02	2.1391	19	2.6809	2.968	140	3.4951	3.8673
8	2.1266	2.2744	20	2.7082	3.0008	160	3.5373	3.9097
9	2.215	2.3868	25	2.8217	3.1353	180	3.5736	3.946
10	2.29	2.4821	30	2.9085	3.2361	200	3.6055	3.9777
11	2.3547	2.5641	40	3.0361	3.3807	300	3.7236	4.0935
12	2.4116	2.6357	50	3.1282	3.4825	400	3.8032	4.1707
13	2.462	2.699	60	3.1997	3.5599	500	3.8631	4.2283
14	2.5073	2.7554	70	3.2576	3.6217	600	3.9109	4.274

(Zdroj: <http://www.sediment.uni-goettingen.de/staff/dunkl/software/pep-grubbs.pdf>)

P. 3 Kritické hodnoty G pre Grubbsov jednostranný test

N	G hodnota $\alpha=0.05$	G hodnota $\alpha=0.01$	N	G hodnota $\alpha=0.05$	G hodnota $\alpha=0.01$	N	G hodnota $\alpha=0.05$	G hodnota $\alpha=0.01$
3	1.1531	1.1546	15	2.409	2.7049	80	3.1319	3.5208
4	1.4625	1.4925	16	2.4433	2.747	90	3.1733	3.5632
5	1.6714	1.7489	17	2.4748	2.7854	100	3.2095	3.6002
6	1.8221	1.9442	18	2.504	2.8208	120	3.2706	3.6619
7	1.9381	2.0973	19	2.5312	2.8535	140	3.3208	3.7121
8	2.0317	2.2208	20	2.5566	2.8838	160	3.3633	3.7542
9	2.1096	2.3231	25	2.6629	3.0086	180	3.4001	3.7904
10	2.1761	2.4097	30	2.7451	3.1029	200	3.4324	3.822
11	2.2339	2.4843	40	2.8675	3.2395	300	3.5525	3.9385
12	2.285	2.5494	50	2.957	3.3366	400	3.6339	4.0166
13	2.3305	2.607	60	3.0269	3.4111	500	3.6952	4.0749
14	2.3717	2.6585	70	3.0839	3.471	600	3.7442	4.1214

Zdroj: <http://www.lcgceurope.com/lcgceurope/data/articlestandard/lcgceurope/502001/4509/article.pdf>.