KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Analýza migrujících domácností v šetření Životní podmínky

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2013

Vypracovala:
Bc. Markéta Švecová
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně pod vedením RNDr. Karla Hrona, Ph.D., a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 19. 3. 2013

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. a konzultantům na Českém statistickém úřadě RNDr. Jaromíru Kalmusovi a Mgr. Mirkovi Otáhalovi, Ph.D. za obětavou spolupráci, připomínky a rady. Také bych ráda poděkovala své rodině, která mě po celou dobu studia podporovala.

Obsah

Úvod

Každý rok Český statistický úřad (ČSÚ) provádí šetření Životní podmínky u vybraných domácností. Ty musí být sledovány po dobu 4 let, a to i v případě, že se celá domácnost nebo její část přestěhuje. Dohledávání těchto subjektů znamená pro Český statistický úřad zvýšení nákladů.

Úkolem této diplomové práce je na základě různých statistických nástrojů ukázat, jaké subjekty migrují a zda jejich vynechání způsobí změnu výsledků šetření.

V první kapitole je čtenář seznámen s metodikou šetření Životní podmínky a s postupem výběru domácností. Následující kapitoly jsou již věnovány analýze dat, přičemž druhá kapitola nejprve obsahuje přehled typů proměnných, kterých může statistický znak nabývat. Ve třetí kapitole jsou pak uvedeny metody analýzy jednorozměrných dat - číselné charakteristiky a grafické nástroje, jako je například histogram, boxplot, kvantilový graf, Q-Q graf a další. Následující kapitola je věnována analýze vícerozměrných dat. Zde se zaměřuji zejména na biplot, který je konstruován na základě analýzy hlavních komponent. Další částí této kapitoly jsou i bagplot a jeho grafická interpretace, testy dobré shody a korelační analýza.

V poslední části této práce se věnuji již samotné analýze dat pomocí zmíněných statistických metod. I když některé používané nástroje statistické analýzy jsou jednoduché, pro řešení našeho problému jsou plně dostačující a používání složitějších metod by nemělo smysl. Data byla k dispozici na Českém statistickém úřadě, protože nemohou být volně šířena. Ze stejného důvodu není v práci přiložena ani ukázka těchto dat.

1 Šetření Životní podmínky

Následující kapitola byla zpracována pomocí internetových stránek [3], na kterých můžeme najít podrobnější informace a výsledky šetření.

Šetření Životní podmínky se stalo pro Českou republiku závazné po vstupu do Evropské unie (2004) a provádí se od roku 2005. Šetření je národní modifikací celoevropského šetření European Union - Statistics on Income and Living Conditions (EU-SILC). Pomocí šetření můžeme srovnávat data nejen v rámci České republiky, ale i porovnat jednotlivé zúčastněné země mezi sebou díky jednotné metodice šetření.

V rámci šetření získáváme údaje o úrovni příjmů v jednotlivých typech domácností, o způsobu bydlení, o vybavení domácností a o pracovních a zdravotních podmínkách dospělých osob v dané domácnosti.

V šetření Životní podmínky ČSÚ využívá tzv. rotační panel, kdy jsou vybrané domácnosti navštěvovány vždy po roce během 4 let. Díky dlouhodobějšímu pozorování domácností můžeme sledovat vývoj jejich ekonomické a sociální situace. Každý rok se část domácností, které již byly navštěvovány 4 roky, obmění.

Výběrovou jednotkou šetření je byt, který je zvolen dvoustupňovým náhodným výběrem. Nejprve se náhodně vybere sčítací obvod a následně se z vybraného obvodu vybere 10 bytů. Jednotkami zjišťování jsou tzv. hospodařící domácnosti tvořené osobami, které se ve vybraném bytě společně podílí na nákladech na své potřeby. Rozhodující tedy není trvalé bydliště. Osoby z domácnosti, se kterými tazatel provede šetření v prvním roce, tvoří tzv. panelovou složku šetření.

Šetření se v domácnostech provádí pomocí těchto dotazníků: dotazník za byt, dotazník za domácnost, dotazník za osobu a modulu, který se mění podle aktuálního zaměření šetření EU-SILC a rozšiřuje tak některou z šetřených oblastí.

Při šetření je kladen důraz na anonymitu. Získaná data jsou chráněna zákonem o státní statistické službě č. 89/1995 Sb. a zákonem o ochraně osobních údajů č. 101/2000 Sb. Tedy všechny osoby, které nějakým způsobem pracují s daty, jsou vázány mlčenlivostí.

1.1 Problém migrace obyvatelstva

Pokud se celá domácnost nebo některá osoba tvořící tzv. panelovou složku šetření z dané domácnosti odstěhuje, podléhá šetření na nové adrese, což je dáno dle nařízení EU. Tazatel tedy musí takové osoby dohledat a provést šetření. Dohledávání těchto osob je velmi náročné a nákladné. Tazatel nemá mnoho možností, jak zjistit informace o jejich nové adrese. Pokud se z bytu odstěhovala jen část panelových osob, může tazatel získat informace od ostatních osob, které v dané domácnosti zůstaly. Ale pokud se odstěhuje celá domácnost, je v podstatě odkázán pouze na komunikaci se sousedy a většinou novou adresu nezíská, nebo ji získá špatnou či neúplnou. Bohužel nezbývá nic jiného, než na získanou adresu tazatele vyslat a zjistit, zda tam daná osoba opravdu žije, což pro Český statistický úřad znamená náklady navíc. Mým úkolem v této diplomové práci je zjistit, zda dohledávání takových osob je opravdu nutné. Jinak řečeno, budeme analyzovat, zda chybějící hodnoty migrujících domácností ovlivní celkovou strukturu souboru dat či nikoliv.

2 Typy proměnných

Tato kapitola byla vytvořena na základě zdrojů [6] a [21]. Data, která chceme analyzovat, se získávají pomocí zjišťování hodnot námi definovaných proměnných (statistických znaků). Proměnné můžeme chápat jako charakteristiky daných objektů a u každé proměnné známe postup ke zjištění její hodnoty. Každá proměnná má své rozdělení pravděpodobností (teoretické), resp. rozdělení četností (výběrové). To je dáno hodnotami, kterých proměnná nabývá, a údaji, jak často se dané hodnoty v datech vyskytují. Proměnné můžeme rozdělit do tříd podle různých hledisek.

Prvním pohledem je dělení dle vztahu mezi proměnnými, kdy rozlišujeme *závisle proměnné (cílové)* a *nezávisle proměnné (prediktory)*. Změna závisle proměnné je způsobena změnou nezávisle proměnné (např. změna BMI je vyvolána změnou váhy).

Někdy do tohoto vztahu vstupuje tzv. *rušivá proměnná*, jejíž působení zkresluje rozhodování o vztahu mezi závisle a nezávisle proměnnou. Rušivá proměnná může být známá i neznámá, měřitelná nebo neměřitelná. Důležité je si uvědomit, zda taková proměnná při analýze našich dat existuje (např. počet úrazů páteře při pádu na lyžích je závislý na použití ochranných pomůcek páteře, přičemž rušivá proměnná může být rychlost jízdy na lyžích před pádem).

Nezávislé proměnné (prediktory) můžeme dále dělit dle možnosti zásahu. *Přirozený prediktor* nelze nijak měnit, je charakteristikou (vlastností) daného objektu (např. výška, váha, věk). Oproti tomu *manipulativní prediktor* lze ovlivňovat (např. teplota prostředí, vlhkost vzduchu).

Další kategorizace je na kolektivní a individuální proměnné. Individuální proměnné charakterizují daný objekt (např. výška a rozloha budovy), kolektivní proměnné popisují podskupinu, do které objekt patří (např. daň z nemovitosti v dané lokalitě).

Nejznámější dělení je podle použitého měřítka. Za *nominální měřítko* považujeme slovní popis nebo kód, kdy o hodnotách proměnné můžeme pouze říci, zda jsou stejné nebo různé (např. místo bydliště, pohlaví, rodinný stav, skupina

řidičského průkazu). U těchto proměnných lze pouze zjišťovat rozdělení četností. V případě *ordinálního měřítka* můžeme kromě rozdělení tříd určit pořadí hodnot proměnné (např. školní klasifikace, stupeň vzdělání). Nominální a ordinální proměnné označujeme jako *kvalitativní*.

Vlastnosti ordinálního měřítka má i *měřítko intervalové*, které navíc umožňuje zjistit vzdálenosti jednotlivých údajů. Hodnoty jsou tedy čísla a má smysl počítat jejich rozdíly (např. teplota měřená ve °C). Nejvíce informací nese *poměrové měřítko*, kde k vlastnostem intervalového měřítka přibývá definování absolutní nuly. Máme tedy možnost říci, kolikrát je jedna hodnota lepší než druhá (např. teplota měřená ve °F). Intervalové a poměrové proměnné nazýváme *kvantitativní*.

Kvantitativní proměnné můžeme dále rozdělit podle hodnot, kterých nabývají. *Spojitá proměnná* může nabývat libovolných hodnot z daného intervalu (např. váha, výška jednotlivce). Oproti tomu *diskrétní proměnná* nabývá pouze konečného (případně spočetného) počtu hodnot (např. počet lidí v domácnosti, počet bytových jednotek v rodinném domě).

3 Analýza jednorozměrných dat

Při analýze statistického souboru chceme zjistit jeho strukturu, odhalit zvláštnosti v datech a ověřit předpoklady pro statistické zpracování. V případě jednorozměrných dat jsou sledovány hodnoty pouze jednoho statistického znaku.

3.1 Číselné charakteristiky jednorozměrných dat

Při vytvoření této kapitoly byly použity zejména zdroje [6], [7], [9], [10], [16] a [17].

Získaný datový soubor můžeme v případě kvantitativního znaku znázornit pomocí číselných charakteristik. Ukazatele, které poskytují informaci o rozložení (koncentraci) hodnot získaného souboru, označujeme jako míry polohy. Ukazatele, které určují rozptýlení neboli variabilitu kolem charakteristiky polohy, nazýváme míry variability. K získání charakteristik se využívá tzv. (výběrových) momentů.

3.1.1 Momenty

Definice 3.1. *Nechť* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *je soubor hodnot statistického znaku* X *naměřený na* n *statistických jednotkách. Obecný moment* k*-tého řádu kolem bodu* a *je definován vztahem*

$$m_k(a) = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^k. \tag{1}$$

Pro hodnotu $a = 0$ dostáváme vztah

$$v_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k, \tag{2}$$

který nazýváme *počáteční obecný moment* k*-tého řádu*. Hodnotu n nazýváme rozsah souboru.

Jestliže $a = \bar{x}$, kde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ označuje aritmetický průměr, dostáváme *centrální obecný moment* k*-tého řádu* definovaný vztahem

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k. \tag{3}$$

3.1.2 Míry polohy

Pro určení hodnoty, kolem které se data soustřeďují, používáme již zmíněný aritmetický průměr, dále modus a kvantily.

Aritmetický průměr si můžeme představit jako pomyslné těžiště dat a je jedním z výběrových momentů (počáteční obecný moment 1. řádu). Bohužel tato hodnota je silně ovlivněna odlehlými pozorováními.

Hodnotu, která se v datech vyskytuje s nejvyšší četností, nazýváme *modus* a značíme ji $\hat{x}$. Modus neboli modální hodnota je tedy nejčastější hodnotou, které nabývá daná proměnná (statistický znak). V praxi ji nejčastěji zjišťujeme pomocí histogramu (viz dále) jako průměr z krajních hodnot toho intervalu, který dosahuje na základě dat nejvyšší četnosti. Na rozdíl od aritmetického průměru může být modů více, a to v situaci, kdy máme několik intervalů s nejvyšší četností.

Kvantil je hodnota, která dělí soubor dat na dvě části. Přitom první část tvoří hodnoty, které jsou menší nebo rovny než daný kvantil a druhá část je tvořena hodnotami většími něž daný kvantil nebo jemu rovnými.

Definice 3.2. *Nechť $\alpha \in (0,1)$. Pak $\alpha-$kvantil statistického znaku X je takové reálné číslo x_α, které splňuje, že minimálně $100\alpha\%$ hodnot statistického souboru je $\leq x_\alpha$ a $100(1-\alpha)\%$ hodnot je $\geq x_\alpha$.*

Speciálním případem kvantilu je *medián* $(0,5-kvantil)$, který rozdělí uspořádanou množinu hodnot znaku na dvě poloviny. Oproti aritmetickému průměru je medián málo citlivý k odlehlým hodnotám v datovém souboru.

Pro určení hodnoty mediánu uspořádáme pozorování do neklesající posloupnosti $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Při lichém počtu pozorování je medián jednoznačně určen hodnotou, která leží uprostřed této posloupnosti. Tedy

$$x_{0,5} = x_{(\frac{n+1}{2})}. \tag{4}$$

Při sudém počtu pozorování je medián určen aritmetickým průměrem prostředních dvou hodnot,

$$x_{0,5} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}. \tag{5}$$

Obecně α-kvantil obdržíme pomocí vztahu

$$x_\alpha = \begin{cases} x_{([np]+1)}, & np \neq [np], \\ \frac{x_{(np)}+x_{(np+1)}}{2}, & np = [np], \end{cases} \tag{6}$$

kde symbol $[\cdot]$ označuje funkci „celá část“.

Kvantily, které mají hodnotu α menší než $0,5$, nazveme *dolními kvantily*. Kvantily s hodnotou α vetší než $0,5$ označujeme jako *kvantily horní*. Speciálně pro $\alpha = 0,25$ nazveme číslo $x_{0,25}$ *dolním kvartilem* a pro $\alpha = 0,75$ je číslo $x_{0,75}$ *horním kvartilem*. Pro α rovno přirozeným násobkům jedné desetiny obdržíme *decily*.

3.1.3 Míry variability

K posouzení *míry variability* používáme rozptyl, směrodatnou odchylku, variační koeficient, interkvartilové rozpětí a medián absolutních odchylek.

Definice 3.3. *Rozptyl je dán přepisem*

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{7}$$

a směrodatná odchylka pomocí vztahu

$$s = \sqrt{s_n^2}. \tag{8}$$

Rozptyl je tedy definován jako průměrná kvadratická odchylka naměřených hodnot od jejich aritmetického průměru. Má však jiný rozměr než je tomu u původních dat, proto se zavádí směrodatná odchylka, která pomocí odmocniny získává zpět měřítko původních dat. Stejně jako aritmetický průměr je i hodnota směrodatné odchylky silně ovlivněna odlehlými pozorováními.

Definice 3.4. *Variační koeficient V je definován podílem*

$$V = \frac{s}{\bar{x}}. \tag{9}$$

Variační koeficient je bezrozměrná veličina. V praxi ji používáme pro porovnání variability mezi statistickými soubory, resp. mezi proměnnými o různých jednotkách. Při vynásobení 100 vyjadřuje variabilitu v procentech.

Definice 3.5. *Interkvartilové rozpětí je dáno vztahem*

$$IQR = x_{0,75} - x_{0,25}, \tag{10}$$

kde $x_{0,25}$ je dolní kvartil a $x_{0,75}$ je horní kvartil.

Tato charakteristika se standardně používá i k popisu tvaru rozdělení datového souboru. Oproti směrodatné odchylce není interkvartilové rozpětí tak citlivé vůči odlehlým pozorováním.

Definice 3.6. *Medián absolutních odchylek (MAD=Median Absolute Deviation) je dán přepisem*

$$MAD = x_{0,5}\{|x_i - x_{0,5}|\}. \tag{11}$$

MAD, který je kvantilovou alternativou ke směrodatné odchylce, je tedy definován jako medián z absolutních hodnot odchylek jednotlivých pozorování od mediánu těchto pozorování.

3.1.4 Koeficienty šikmosti a špičatosti

Tyto charakteristiky popisují tvar rozdělení dat, odhalí jeho asymetrii a strmost. Pomocí nich také zjišťujeme, jak se rozdělení dat podobá (teoretické) hustotě normálního rozdělení (tzv. Gaussově křivce), což je funkce daná vztahem

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{12}$$

kde parametry $\mu \in \mathbf{R}$ a $\sigma^2 > 0$. K výpočtu šikmosti a špičatosti se nejčastěji využívá centrálních momentů.

Koeficient šikmosti (někdy nazýván jako koeficient asymetrie) je číslo charakterizující míru sešikmení (nesymetrie) kolem aritmetického průměru a výpočet provádíme pomocí vzorce

$$S_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}^3} = \frac{m_3}{m_2^{\frac{3}{2}}}. \tag{13}$$

Používáme tedy druhý a třetí centrální moment. Pokud míra šikmosti je rovna nule, jedná se o symetrické rozdělení. $S_1 > 0$ značí rozdělení s prodlouženým pravým koncem, $S_1 < 0$ identifikuje rozdělení s prodlouženým levým koncem.

Koeficient špičatosti (někdy označován jako koeficient excesu) vyjadřuje koncentraci hodnot kolem aritmetického průměru. Charakterizuje tedy špičatost (resp. plochost) tvaru rozdělení dat a je dán vztahem

$$S_2 = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} = \frac{m_4}{m_2^2}. \tag{14}$$

Pro hodnotu 3 odpovídá koncentrace hodnot normálnímu rozdělení. V praxi se používá modifikovaný koeficient špičatosti, kde od koeficientu špičatosti odečtěme 3, tj.

$$S_2 = \frac{m_4}{m_2^2} - 3. \tag{15}$$

Takto vypočtený koeficient špičatosti má pro normální rozdělení hodnotu 0. Kladná hodnota značí špičatější křivku, záporná hodnota plošší. Čím je větší špičatost, tím jsou data více soustředěna kolem středu rozdělení, reprezentovaného aritmetickým průměrem.

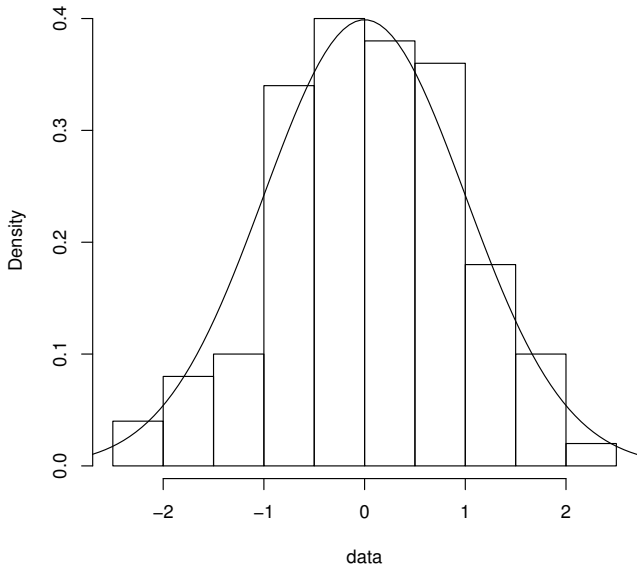
3.2 Grafické nástroje

V následujícím textu jsou uvedeny některé grafické nástroje, které se používají k charakterizaci dat. Další grafické metody a informace můžeme najít například v literatuře [6], [14] a internetových zdrojích [12], [13], [16], [18] a [19], pomocí nichž byl vytvořen tento text.

Důležitými prostředky pro průzkumovou analýzu dat jsou grafické nástroje. Díky těmto metodám můžeme nejen zjednodušeně popsat data, ale i získat informace o rozdělení dat a posoudit jejich statistické vlastnosti.

3.2.1 Histogram

Nejčastěji používaným nástrojem pro zobrazení hodnot jedné proměnné je histogram. Tento graf tvoří soustava obdélníků, jejichž výška znázorňuje četnosti (absolutní nebo relativní), s nimiž statistický znak nabývá dané třídy hodnot (tvoří šířku obdélníku). Pro dobré zobrazení je nutné zvolit optimální počet tříd pokrývajících celou škálu hodnot znaku. Se zmenšujícím se počtem dat se zmenšuje i počet těchto tříd.



Obrázek 1: Histogram výběru z normovaného normálního rozdělení

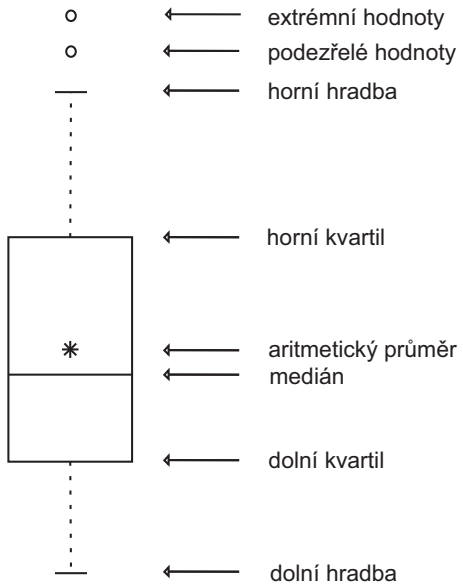
Důležitý je tvar histogramu. Pomocí něj můžeme určit základní tvar rozdělení, z něhož data pochází. Nejčastěji se tvar histogramu porovnává s ideální křivkou (hustotou normálního rozdělení, zmíněnou Gaussovou křivkou - symetrická funkce zvonovitého tvaru). Pokud tvar histogramu připomíná tuto křivku, můžeme usoudit, že data pochází z normálního rozdělení.

Z grafu můžeme také vyčíst nejčetnější hodnoty statistického znaku (modus $\hat{x}$) pomocí nejvyšších sloupců (třídy s nejvyšší četností). Dále lze identifikovat oblasti

bez hodnot díky mezerám v grafu a určit odlehlá pozorování. Ukázka histogramu je uvedena na obrázku 1.

3.2.2 Boxplot (Krabicový graf)

Boxplot neboli krabicový graf s anténami (někdy označovány jako vousy) slouží pro posouzení centrální tendence dat a jejich rozptýlenosti. Pomocí tohoto grafu můžeme také posoudit sešikmení rozdělení datového souboru a odhalit přítomnost odlehlých pozorování.



Obrázek 2: Boxplot

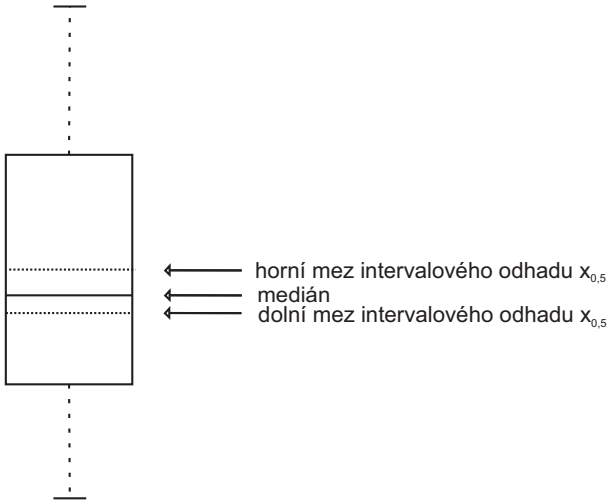
Jak vidíme na obrázku 2, krabicový graf se skládá z obdélníku a antén. Obdélník obsahuje přibližně polovinu hodnot souboru dat. Spodní a horní hrana obdélníku je určena dolním a horním kvartilem. Obdélník je rozdělen na dvě části pomocí mediánu. Umístění mediánu v obdélníku charakterizuje sešikmení. Pokud medián dělí obdélník na dvě poloviny, je rozdělení dat symetrické. Pokud se medián přibližuje k jedné z hran obdélníku, rozdělení dat je sešikmené v opačném směru. Uvnitř obdélníku může být zakreslen i aritmetický průměr. Úsečky, které vychází z obdélníku, jsou nazývané jako antény nebo vousy a jsou zakončeny tzv. vnitřními hradbami. Dolní hradba je dána hodnotou $x_{0,25} - 1,5 \cdot IQR$, horní

hradba hodnotou $x_{0,75} + 1,5 \cdot IQR$. Hodnoty, které neleží uvnitř intervalu daného hradbami, jsou považovány za podezřelé hodnoty a jsou znázorněny v grafu kroužky. Někdy se do grafu znázorňují i extrémní hodnoty, které jsou menší než $x_{0,25} - 3 \cdot IQR$ nebo větší než $x_{0,75} + 3 \cdot IQR$.

3.2.3 Vrubový boxplot

Obdobou krabicového grafu je vrubový krabicový graf, který je ukázán na obrázku 3. Tento graf navíc poskytuje informaci o variabilitě mediánu. Ta je v grafu znázorněna pomocí pruhu, který značí intervalový odhad mediánu ($I_D \leq x_{0,5} \leq I_H$). Meze intervalu jsou vyjádřeny vztahy

$$I_D = x_{0,5} - \frac{1,57 \cdot IQR}{\sqrt{n}}, \qquad I_H = x_{0,5} + \frac{1,57 \cdot IQR}{\sqrt{n}}. \tag{16}$$

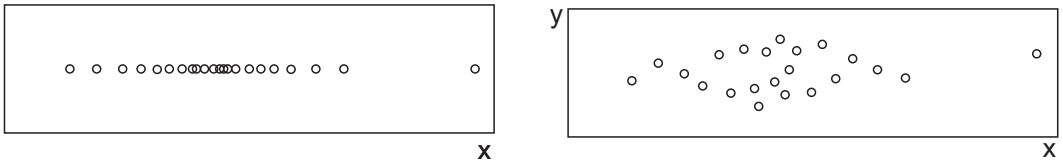


Obrázek 3: Vrubový boxplot

3.2.4 Diagram rozptýlení, rozmítnutý diagram rozptýlení

Graf rozptýlení je prosté znázornění dat na osu x. I když tento způsob je velmi jednoduchý, má značnou vypovídající schopnost o datech. Můžeme odhalit jak odlehlá pozorování, tak i lokální koncentraci dat. Pokud je v grafu více

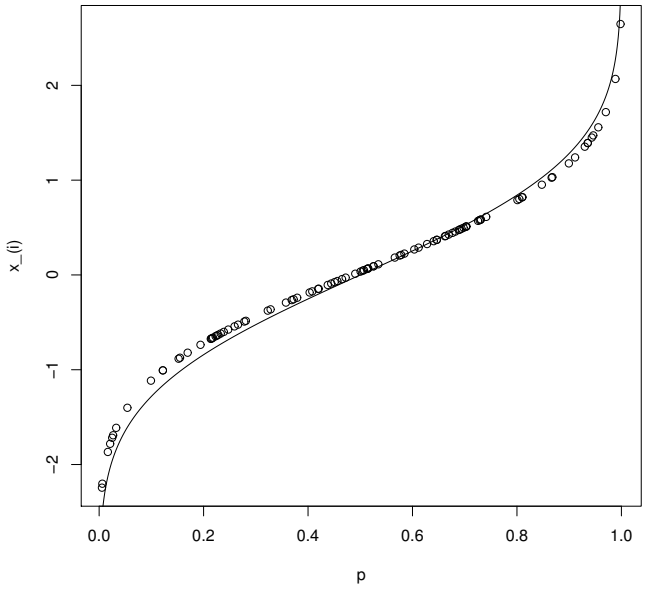
koncentrací (shluků) dat, svědčí to o nehomogenitě výběrového souboru.



Obrázek 4: Diagram rozptýlení (vlevo) a rozmítnutý diagram rozptýlení (vpravo)

Jestliže se v grafu nacházejí místa s velkou koncentrací hodnot, je vhodné pro lepší přehlednost zvolit rozmítnutý diagram rozptýlení. Tento graf má body vhodně rozptýleny ve směru osy y a nedochází tak k jejich překrývání. Oba grafy jsou znázorněny na obrázku 4.

3.2.5 Kvantilový graf



Obrázek 5: Kvantilový graf výběru z normovaného normálního rozdělení

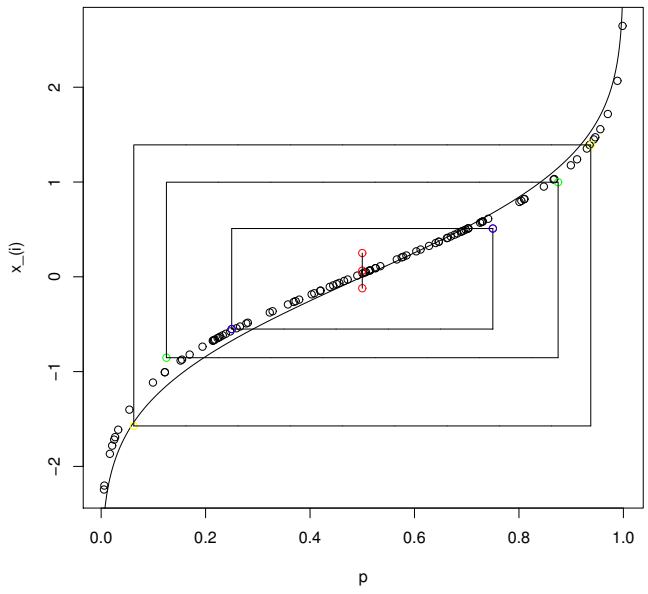
Z tohoto grafu můžeme snadno získat informace o rozdělení a odhalit asymetrii v datech, shluky dat a odlehlá pozorování. Graf je tvořen body, jejichž

souřadnice jsou pořadové pravděpodobnosti (na ose x) a hodnoty uspořádaného datového souboru $x_{(i)}$ (na ose y). Tvar křivky, do níž jsou uspořádány body v rovině, dává informaci o sešikmení datového souboru. Pokud křivka má sigmoidální tvar, značí symetrické rozdělení dat. Pro rozdělení sešikmená je křivka konvexně rostoucí (sešikmení k vyšším hodnotám), nebo konkávně rostoucí (sešikmění k nižším hodnotám). Pro snadnější porovnání s jednotlivými typy rozdělení se do tohoto grafu zakreslují i kvantilové funkce příslušných rozdělení, jak je uvedeno na obrázku 5.

3.2.6 Graf rozptýlení s kvantily

Body v tomto zobrazení jsou shodné s body v kvantilovém grafu a i zde se pro porovnání mohou zakreslit kvantilové funkce příslušných rozdělení. Do grafu rozptýlení s kvantily jsou pro usnadnění interpretace navíc zakresleny obdélníky - kvartilový, oktilový a sedecilový. Nejmenší obdélník - kvartilový je dán souřadnicemi: na ose x pořadové pravděpodobnosti $2^{-2} = 0,25$ a $1 - 2^{-2} = 0,75$, na ose y kvartily $x_{0,25}$ a $x_{0,75}$. Prostřední obdélník - oktilový má souřadnice: na ose x pořadové pravděpodobnosti $2^{-3} = 0,125$ a $1 - 2^{-3} = 0.875$, na ose y kvantily $x_{0.125}$ a $x_{0.875}$. Největší obdélník - sedecilový je dán souřadnicemi: na ose x pořadové pravděpodobnosti $2^{-4} = 0,0625$ a $1 - 2^{-4} = 0,937$, na ose y kvantily $x_{0,0625}$ a $x_{0,0937}$. Dále se do grafu zakresluje i intervalový odhad mediánu ($x_{0,5} \pm \frac{1,57 \cdot IQR}{\sqrt{n}}$) jako horizontální úsečka se středem v bodě $[0,5, x_{0,5}]$. Graf rozptýlení s kvantily máme znázorněn na obrázku 6.

Na základě obdélníků můžeme zjistit informace o rozdělení výběru. Symetrické rozdělení je charakterizováno tím, že jednotlivé obdélníky jsou symetricky uvnitř sebe. Nesymetrická rozdělení se sešikmením k vyšším hodnotám mají vzdálenosti mezi dolními hradbami obdélníků výrazně menší než jsou tyto vzdálenosti mezi horními hradbami, pro sešikmení k nižším hodnotám je tomu naopak. Dále můžeme identifikovat odlehlá pozorování, kdy mimo kvartilový obdélník křivka náhle vzroste (hodnota směrnice roste nade všechny meze). Odhalit můžeme i vícemodální rozdělení tak, že v kvartilovém obdélníku je několik úseků, kdy je



Obrázek 6: Graf rozptýlení s kvantily výběru z normovaného normálního rozdělení

křivka téměř rovnoběžná s osou x (má téměř nulové směrnice).

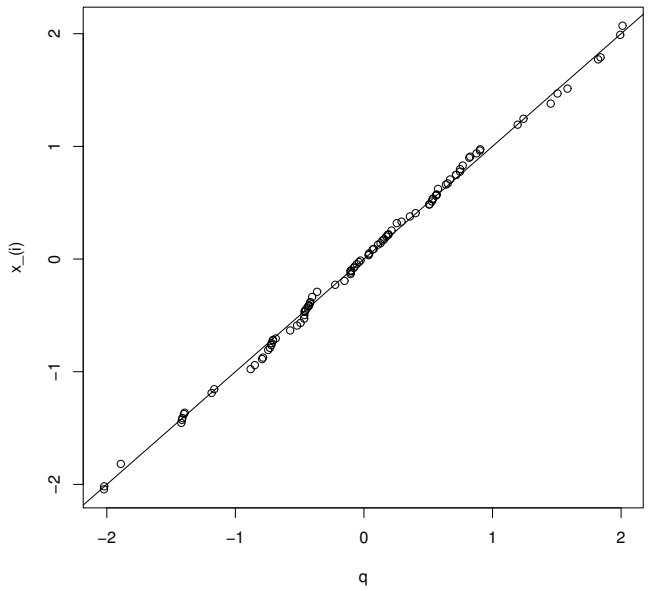
3.2.7 Q-Q graf

Q-Q neboli kvantil-kvatilový graf je jedním z nejpoužívanějších nástrojů k posouzení rozdělení dat. V tomto grafu porovnáváme odhad kvantilové funkce výběrového rozdělení, který je dán uspořádanými hodnotami $x_{(i)}$, s kvantilovou funkcí teoretického rozdělení $Q_T(P)$. Pokud platí přibližná rovnost těchto kvantilů, tj.

$$x_{(i)} \approx Q_T(p_i), \tag{17}$$

je výběrové rozdělení shodné se zvoleným teoretickým rozdělením (p_i je pořadová pravděpodobnost) a závislost uspořádaných hodnot $x_{(i)}$ na kvantilech teoretického rozdělení $Q_T(p_i)$ je lineární, tedy Q-Q grafem bude přímka.

3.2.8 Rankitový graf



Obrázek 7: Rankitový graf - Q-Q graf s normovaným normálním rozdělení

Obdobou Q-Q grafu je graf rankitový, kdy výběrové rozdělení porovnáváme s rozdělením normálním. Ukázka grafu je uvedena v obrázku 7. Průběh grafu umožňuje zařadit rozdělení dat do skupin dle šikmosti, špičatosti a délky konců. Pokud je graf konvexní, popř. konkávní, rozdělení je sešikmené. Esovitý tvar křivky zase indikuje odlišnou špičatost, než má normální rozdělení.

4 Analýza vícerozměrných dat

Následující kapitola byla zpracována pomocí literatury [2], [4] [5], [6], [7], [8], [11], [14] a [20].

Podkladem pro analýzu vícerozměrných dat je tzv. *datová matice* $\mathbf{X}$ (někdy nazývána zdrojovou maticí). Tato matice je tvořena hodnotami zjišťovaných proměnných. Řádky vyjadřují jednotlivé objekty a sloupce vyjadřují hodnoty daného pozorování příslušných objektů. Celkový počet objektů označíme n a počet proměnných zjišťovaných u každého objektu označíme p. Tedy

$$\mathbf{X} = \{x_{ij}\}_{i=1,\dots,n}^{j=1,\dots,p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}, \tag{18}$$

kde x_{ij} je hodnota j-té proměnné zjištěna u i-tého objektu a vektor $\boldsymbol{x}_i^T = (x_{i1}, \dots, x_{ip})$ značí i-tý objekt, který je charakterizován p hodnotami příslušných proměnných.

V mnoha případech se jednotlivá měření provádí v různých jednotkách, proto se zdrojová matice před samotnou analýzou upravuje centrováním nebo standardizací. V případě centrování se od prvků ve sloupci odečte jejich sloupcový aritmetický průměr. Standardizaci, neboli normování, provedeme tak, že centrované hodnoty vydělíme příslušnou sloupcovou směrodatnou odchylkou.

4.1 Charakteristiky vícerozměrných dat

Obdobně jako jednorozměrná data lze popsat pomocí číselných charakteristik, můžeme i vícerozměrná data charakterizovat pomocí průměru a (výběrové) varianční matice.

Definice 4.1. *Průměr matice $\boldsymbol{X}$ je definován vztahem*

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i. \tag{19}$$

Vektor $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)$ je tedy složen z aritmetických průměrů jednotlivých sloupců datové matice.

Definice 4.2. *Varianční matice* $\boldsymbol{S}$ *je dána jako,*

$$\boldsymbol{S} = (s_{ij}) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T. \tag{20}$$

Tato symetrická pozitivně definitní matice řádu p má na hlavní diagonále rozptyly jednotlivých proměnných, mimo diagonálu pak kovariance jednotlivých dvojic proměnných, které jsou dány vztahem $s_{ij} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$, kde $i, j = 1, \ldots, p$, $i \neq j$.

Z varianční matice můžeme pomocí vhodného normování získat korelační matici.

Definice 4.3. *Matici* $\boldsymbol{R} = (r_{jk})$, $j, k = 1, \ldots, p$, *jejíž prvky jsou korelační koeficienty, tj.*

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} \cdot s_{kk}}}, \tag{21}$$

kde s_{jk} *značí prvky matice* $\boldsymbol{S}$*, nazveme korelační maticí.*

4.2 Korelace

Korelace označuje míru těsnosti vztahu mezi proměnnými. Tento vztah nastává, pokud určité hodnoty dané proměnné mají tendenci se vyskytovat společně s určitými hodnotami jiné proměnné. K určení intenzity vztahů mezi proměnnými se používají korelační koeficienty. Jednotlivé typy korelačních koeficientů se liší podle typů proměnných, pro které byly odvozeny. Zde popíšeme podrobněji jen párový (Pearsonův) korelační koeficient (tvořící prvky výše uvedené korelační matice), který budeme používat.

Pearsonův korelační koeficient je nejdůležitější mírou síly lineárního vztahu

dvou proměnných X a Y a je dán vztahem

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{22}$$

Korelační koeficient nabývá hodnot $-1 \leq r_{xy} \leq 1$. Pokud $|r_{xy}| = 1$, leží všechny body na jedné přímce a y-ovou souřadnici bodu můžeme vypočítat z x-ové souřadnice pomocí lineárního vztahu. Čím více se hodnota koeficientu blíží ke krajním bodům intervalu $\langle -1, 1\rangle$, tím více jsou dané proměnné korelované (tj. jejich lineární vztah je silnější). Jestliže $r_{xy} = 0$, označíme X a Y za nekorelované.

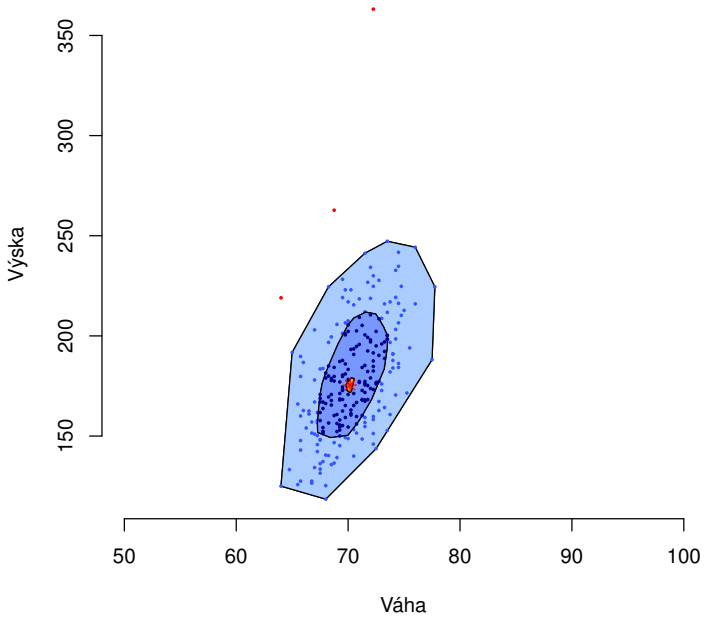
Ze vzorce korelačního koeficientu je zřejmé, že jeho hodnota se nezmění, pokud změníme jednotky měření jednotlivých proměnných.

Nevýhodou Personova korelačního koeficientu je, že je ovlivněn odlehlými pozorováními a nerozlišuje závisle a nezávisle proměnné. Také nedává úplný popis dat i při velmi silném lineárním vztahu (vhodnější je rovnice přímky, která udává konkrétní vztah). Korelace sama o sobě navíc ještě neprokazuje, že změna proměnné X skutečně způsobuje změnu v proměnné Y.

4.3 Bagplot

Bagplot je zobecněním boxplotu ve dvourozměrném prostoru. Využívá se v situacích, kdy chceme graficky znázornit výsledky měření dvou proměnných.

Bagplot tvoří několik hlavních částí. První částí je tzv. bag, který obsahuje 50% dat mezi horním a dolním kvartilem. Tato část odpovídá obdelníku v boxplotu. V grafu je tato oblast zakreslena nejtmavší barvou. Stejně jako v krabicovém grafu je i zde zaznačen medián, tentokrát pomocí bodu uvnitř bagu. Další částí je loop (tzv. smyčka), která reprezentuje vousy boxplotu. V grafu tuto oblast tvoří světlejší barva. Hranice světlejší oblasti odděluje odlehlá pozorování a je nazývána fence (tzv. oplocení, hradba). Za touhle hranicí jsou odlehlé body vyznačeny pomocí křížků.



Obrázek 8: Bagplot - váha, výška

Stejně jako boxplot i bagplot znázorňuje jednotlivé charakteristiky dat. Informaci o poloze nám dává umístění hloubkového (dvourozměrného) mediánu. Rozpětí je dáno velikostí bagu a orientace bagu nám dává informaci o korelaci mezi proměnnými. Pokud je graf „rostoucí“, jedná se o kladnou korelaci. V opačném případě je korelace záporná. Dále můžeme získat přehled i o sešikmení v datech, a to podle tvaru bagu a smyčky a obdobně jako u boxplotu podle umístění mediánu. Ukázka bagplotu je uvedena na obrázku 8, který zobrazuje data pro výšku a váhu dospělých mužů dostupná ze zdroje [1].

4.4 Analýza hlavních komponent

Analýza hlavních komponent (PCA= Principal Component Analysis) je jednou z nejpoužívanějších metod vícerozměrné analýzy dat. Cílem této metody je redukce počtu proměnných pomocí tzv. *hlavních komponent*, které lze popsat

jako lineární kombinace původních proměnných. Pro redukci proměnných musí být ovšem splněn předpoklad, že původní proměnné jsou mezi sebou (pokud možno) silně korelované. Čím je korelace mezi původními proměnnými větší, tím méně nových proměnných bude zapotřebí. Oproti regresní analýze nejsou proměnné děleny na závislé a nezávislé. Chceme tedy původních p proměnných x_i nahradit menším počtem nových nekorelovaných proměnných z_j (tzv. hlavních komponent), které obsahují co nejvíce informací (reprezentovaných rozptyly původních proměnných) o výchozím datovém souboru. Díky nekorelovanosti každá z nových proměnných vysvětluje jinou vlastnost dat. Základními charakteristikami hlavních komponent jsou tedy jejich rozptyly. Pomocí nich můžeme seřadit tyto nové proměnné dle jejich důležitosti, tj. $var(z_1) > var(z_2) > \ldots > var(z_p)$. Při analýze hlavních komponent předpokládáme, že pouze několik z nich má nezanedbatelný rozptyl, ostatní komponenty s malým nebo nulovým rozptylem pak můžeme zanedbat. Takto získáme pro popis původních p proměnných menší počet nových proměnných z_j. Největší část informace o variabilitě původních dat je tedy obsažena v první hlavní komponentě.

Například v literatuře [14] je ukázáno, že rozptyl j-té hlavní komponenty je roven j-tému vlastnímu číslu varianční matice $\mathbf{S}$, tj. $var(z_j) = \lambda_j$ a rozptyl všech hlavních komponent je roven stopě matice $\mathbf{S}$, $tr(\mathbf{S}) = \sum_{j=1}^{p} \lambda_j$.

Jelikož platí, že součet rozptylů hlavních komponent musí dát součet rozptylů původních proměnných, můžeme pomocí podílu

$$P_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{23}$$

určit, jakou část variability daná hlavní komponenta obsahuje. Pokud součet prvních m podílů je blízký 1 (tedy vyjadřuje téměř 100% variability, obvykle stačí 70-80%), těchto m hlavních komponent dostatečně vysvětluje variabilitu původních proměnných, ostatní zanedbáme. I pokud je původních proměnných velký počet, může být počet m ponechaných hlavních komponent velmi malý (často 2-5). Těchto m komponent tvoří vlastní model hlavních komponent PCA. Ztrátu

informace vzniklou při redukci proměnných nazveme *chybou modelu PCA* nebo *špatnou mírou těsností proložení* modelu PCA. V praxi jsou výsledky metody hlavních komponent zobrazeny pomocí biplotu, jehož konstrukci popíšeme v následujícím textu.

4.4.1 Biplot

Biplot je dvourozměrné znázornění proměnných a objektů do jednoho grafu. Dvoudimenzionální zobrazení dat je přehledné a výhodné pro manipulaci. V tomto textu se budeme zabývat metodou konstrukce biplotu, která je založena na analýze hlavních komponent.

Projekce dat ve dvou dimenzích předpokládá, že hodnost datové matice je 2. Pro datové matice vyšších hodností použijeme pouze první dvě hlavní komponenty. Předpokládáme tedy, že dvě dimenze reprezentují většinu variability původních dat.

Při konstrukci biplotu vycházíme z toho, že matici $\mathbf{X}$ vyjádříme pomocí vztahu

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{24}$$

kde $\mathbf{U}$ a $\mathbf{V}$ jsou ortogonální matice rozměrů $n \times n$ a $p \times p$ a $\mathbf{D}$ je matice typu $n \times p$, kde $d_{ii} \geq 0$ pro $i = 1, \ldots, min(n, p)$ a zbývající prvky jsou nulové. Kladné hodnoty d_{ii} budeme nazývat cenné hodnoty matice $\mathbf{X}$. Matice $\mathbf{UD}$ přitom obsahuje (výběrové) hodnoty hlavních komponent. Pokud je hodnost $h(\mathbf{X}) = k < min(n, p)$, můžeme $\mathbf{X}$ vyjádřit pomocí vztahu

$$\mathbf{X} = \sum_{i=1}^{k} d_{ii}\mathbf{u}_i\mathbf{v}_i^T, \tag{25}$$

kde $\mathbf{u}_i$ je i-tým sloupcem matice $\mathbf{U}$, $\mathbf{v}_i$ je i-tým sloupcem matice $\mathbf{V}$. Sloupce $\mathbf{u}_i$ matice $\mathbf{U}$ jsou vlastní vektory matice $\mathbf{X}\mathbf{X}^T$ příslušné vlastnímu číslu d_{ii}^2 a sloupce $\mathbf{v}_i$ matice $\mathbf{V}$ jsou vlastní vektory matice $\mathbf{X}^T\mathbf{X}$ příslušné vlastnímu číslu d_{ii}^2, což vychází ze vztahů

$$\mathbf{X}\mathbf{X}^T\mathbf{u}_i = d_{ii}^2\mathbf{u}_i, \tag{26}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{v}_i = d_{ii}^2\mathbf{v}_i. \tag{27}$$

Mějme tedy datovou matici $\mathbf{X}$ o rozměrech $n \times p$ s hodností k. Princip konstrukce biplotu je založen na aproximaci matice $\mathbf{X}$ pomocí matice $\mathbf{X}_{(2)}$ s hodností 2. Tato aproximace je nejlepší z hlediska minimalizace součtu čtverců odchylek prvků matice $\mathbf{X}_{(2)}$ od příslušných prvků matice $\mathbf{X}$ a je dělána podle výše uvedeného singulárního rozkladu. Jelikož matice $\mathbf{X}_{(2)}$ má pouze dva sloupce, pracujeme dále pouze s prvními dvěma sloupci matic $\mathbf{U}$ a $\mathbf{V}$, nové matice označme $\mathbf{U}_2$ a $\mathbf{V}_2$ s rozměry $n \times 2$ a $p \times 2$, a s příslušnou maticí $\mathbf{D}_2$. Tvar aproximace je tedy

$$\mathbf{X} \approx \mathbf{X}_{(2)} = \mathbf{U}_2\mathbf{D}_2\mathbf{V}_2^T = (\mathbf{u}_1, \mathbf{u}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix}. \tag{28}$$

Matice $\mathbf{X}_{(2)}$ může být také rozdělena následovně,

$$\mathbf{X}_{(2)} = \mathbf{G}\mathbf{H}^T, \tag{29}$$

kde matice $\mathbf{G}$ a $\mathbf{H}$ jsou dány takto

$$\mathbf{G} = (\mathbf{u}_1, \mathbf{u}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c}, \tag{30}$$

$$\mathbf{H} = (\mathbf{v}_1, \mathbf{v}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{31}$$

pro $0 \leq c \leq 1$. Po volbě c budou rozděleny první dvě cenné hodnoty mezi maticemi $\mathbf{G}$ a $\mathbf{H}$. Biplot potom tvoří řádky matic $\mathbf{G}$ a $\mathbf{H}$, tedy $n+p$ dvourozměrných vektorů. Pro $c = 1$ jsou matice $\mathbf{G}$ a $\mathbf{H}$ dány následovně.

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_{1.}^T \\ \vdots \\ \mathbf{g}_{n.}^T \end{pmatrix} = \sqrt{n-1}\,(\mathbf{u}_1, \mathbf{u}_2), \tag{32}$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_{1.}^T \\ \vdots \\ \mathbf{h}_{p.}^T \end{pmatrix} = \frac{1}{\sqrt{n-1}}\,(\mathbf{v}_1, \mathbf{v}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}. \tag{33}$$

Dále dostaneme následující vlastnosti, kde pro zjednodušení předpokládáme, že pracujeme již s centrovanými daty:

- Součiny řádků matic $\mathbf{G}$ a $\mathbf{H}$ nám dávají aproximaci matice $\mathbf{X}$,

$$\mathbf{g}_{i.}^T\mathbf{h}_{j.} = \sqrt{n-1}\mathbf{u}_{i.}^T\frac{1}{\sqrt{n-1}}(\mathbf{v}_{j.}^T\mathbf{D}_2)^T = \mathbf{u}_{i.}^T\mathbf{D}_2\mathbf{v}_{j.} \approx x_{ij}. \tag{34}$$

- Součin mezi řádky matice $\mathbf{H}$ nám dává aproximaci kovarianční matici $\mathbf{S}$,

$$\mathbf{H}\mathbf{H}^T = (\frac{1}{\sqrt{n-1}}\mathbf{V}_2\mathbf{D}_2)(\frac{1}{\sqrt{n-1}}\mathbf{D}_2\mathbf{V}_2^T) = \frac{1}{n-1}\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2^T =$$

$$= \frac{1}{n-1}(\mathbf{V}_2\mathbf{D}_2\mathbf{U}_2^T)(\mathbf{U}_2\mathbf{D}_2\mathbf{V}_2^T) = \frac{1}{n-1}\mathbf{X}_{(2)}^T\mathbf{X}_{(2)} \approx \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \mathbf{S}. \tag{35}$$

Zde jsme využili toho, že matice $\mathbf{D}$ má na hlavní diagonále nezáporné hodnoty (tedy $\mathbf{D}\mathbf{D}^T = \mathbf{D}^2$), a že platí $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Z předešlého vyplývá, že diagonální prvky matice $\mathbf{H}\mathbf{H}^T$ reprezentují hodnoty čtverců euklidovské normy, tj. $||\mathbf{h}_{j.}||^2 = \mathbf{h}_{j.}^\mathbf{T}\mathbf{h}_{j.}$, což nám dává aproximaci rozptylu.

- Platí, že kosinus mezi řádky matice $\mathbf{H}$ nám dává aproximaci korelací mezi proměnnými,

$$\cos(\mathbf{h}_{i.}, \mathbf{h}_{j.}) = \frac{\mathbf{h}_{i.}^T\mathbf{h}_{j.}}{||\mathbf{h}_{j.}||\,||\mathbf{h}_{j.}||} \approx r_{ij}. \tag{36}$$

- Další vlastností je, že euklidovská vzdálenost mezi řádky matice $\mathbf{G}$ aproximuje Mahalanobisovu vzdálenost mezi pozorováními.

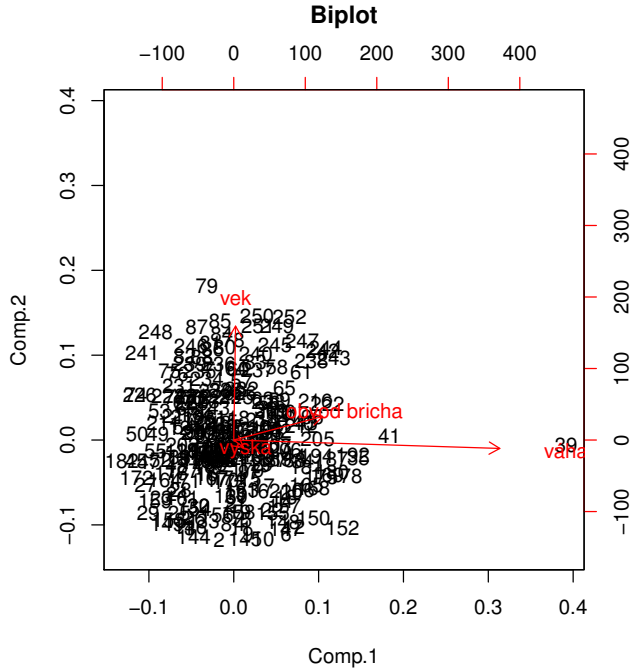
$$\left|\left|\mathbf{g}_{i.} - \mathbf{g}_{j.}\right|\right|^2 = (\mathbf{g}_{i.} - \mathbf{g}_{j.})^T(\mathbf{g}_{i.} - \mathbf{g}_{j.}) = (n-1)(\mathbf{u}_{i.} - \mathbf{u}_{j.})^T(\mathbf{u}_{i.} - \mathbf{u}_{j.}) \approx$$

$$\approx (\mathbf{x}_{i.} - \mathbf{x}_{j.})^T\mathbf{S}^{-1}(\mathbf{x}_{i.} - \mathbf{x}_{j.}), \tag{37}$$

neboť

$$\mathbf{x}_{i.}^T\mathbf{S}^{-1}\mathbf{x}_{j.} \approx (\mathbf{u}_{i.}^T\mathbf{D}\mathbf{V}^T)(n-1)(\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T)(\mathbf{V}\mathbf{D}\mathbf{u}_{j.}) = (n-1)\mathbf{u}_{i.}^T\mathbf{u}_{j.} \tag{38}$$

pro $i, j = 1, \ldots, n$ ($\mathbf{x}_{i.}$ značí i-tý řádek matice $\mathbf{X}$).



Obrázek 9: Biplot

V rámci grafické interpretace biplotu nám řádky matice **G** udávají body a řádky matice **H** vrcholy šipek vycházejících z bodu, který je dán dvojicí aritmetických průměrů sloupců matice **G**. Pokud se jedná o centrovaná data, jedná se o bod [0,0]. Jednotlivé body v grafu vyjadřují objekty a šipky jednotlivé statistické znaky. Délky šipek udávají směrodatné odchylky příslušných statistických znaků. Tedy velikost šipky nám říká, jak je daný statistický znak významný (čím větší rozptyl, tím má znak větší vliv na uspořádaní dat). Kosinus úhlu mezi jednotlivými šipkami znázorňuje korelaci mezi příslušnými statistickými znaky. Čím je velikost úhlu bližší 0° (resp. 180°), tím je kladný (resp. záporný) vztah mezi příslušnými statistickými znaky těsnější a naopak. Umístění jednotlivých bodů vzhledem k šipkám nám udává, jak si daný objekt stojí v porovnání s ostatními objekty. Graf biplotu je uveden na obrázku 9, který je sestaven na základě dat ze zdroje [1] pro proměnné věk, výška (ve stopách), váha (v librách) a obvod

břicha (v centimetrech). Z grafu vidíme, že největší korelaci mají proměnné váha a obvod břicha. Naopak proměnné váha a věk u dospělých mužů spolu dle biplotu téměř nesouvisí, protože úhel mezi příslušnými šipkami je okolo 90 stupňů. Pokud se zaměříme na velikosti šipek, vidíme, že největší směrodatnou odchylku má proměnná váha, oproti tomu výška má tuto charakteristiku velmi malou. V grafu vidíme také odlehlé pozorování s č. 39. Tento objekt je nejblíže k šipce charakterizující váhu. To značí, že svou hmotností převažuje nad ostatními.

4.5 Testy dobré shody

Testy dobré shody jsou nástrojem k testování shody relativní struktury tříd dané proměnné s nějakým teoretickým modelem. V této kapitole uvedeme test dobré shody, který porovnává empirické četnosti $n_1, \ldots, n_k$ (skutečně napozorované četnosti tříd statistického souboru), kde k je počet tříd, s očekávanými (teoretickými) četnostmi $np_1^0, \ldots, np_k^0$, kde $p_1^0, \ldots, p_k^0$ jsou pravděpodobnosti jednotlivých tříd v předpokládaném multinomickém pravděpodobnostním rozdělení. Přitom n značí rozsah souboru, dále $n_1 + \ldots + n_k = n$ a $p_1^0 + \ldots + p_k^0 = 1$. Testujeme nulovou hypotézu, že pravděpodobnosti jednotlivých tříd jsou rovny teoretickým pravděpodobnostem, tj.

$$H_0 : p_1 = p_1^0, \ldots, p_k = p_k^0 \quad \text{oproti} \quad H_A : p_i \neq p_i^0 \quad \text{pro alespoň jedno } i.$$

Danou hypotézu ověřujeme pomocí testovací statistiky ve tvaru

$$Z = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i},$$

která má za platnosti nulové hypotézy asymptoticky χ^2 rozdělení o $k-1$ stupních volnosti. Nulovou hypotézu zamítneme, pokud realizace testovací statistiky $Z \in W = \langle \chi^2_{k-1}(1-\alpha), \infty \rangle$, kde $\alpha \in (0,1)$ a $\chi^2_{k-1}(1-\alpha)$ značí $(1-\alpha)$-kvantil příslušného rozdělení. Pokud hodnota testovací statistiky leží v kritickém oboru W, značí to špatnou shodu souboru dat s teoretickým rozdělením pravděpodobnosti, nulovou hypotézu tedy zamítáme.

5 Zpracování dat z šetření Životní podmínky

Na Českém statistickém úřadě máme k dispozici dva soubory s daty z šetření Životní podmínky 2010. První soubor obsahuje data pro celé domácnosti. Šetření bylo prováděno celkem u 9098 domácností, u kterých byla zjišťována a zaznamenávána data v následujících oblastech:

- DOHLED - zda daná domácnost byla v době šetření na původní adrese, nebo se přestěhovala.

Hodnota	Význam hodnoty
0	domácnost byla na původní adrese
1	domácnost byla na jiné adrese

- OSOB - počet členů dané domácnosti.

- EA - počet ekonomicky aktivních členů dané domácnosti.

- NAKLADY - měsíční náklady na bydlení dané domácnosti.

- NAKL_ZATEZ - jak velkou zátěží jsou náklady na bydlení pro danou domácnost.

Hodnota	Význam hodnoty
1	velká zátěž
2	určitá zátěž
3	žádná zátěž

- CP_PRIJ - čisté příjmy domácnosti za rok 2009.

- VEL - velikost obce, ve které daná domácnost žije.

Hodnota	Význam hodnoty
1	do 199 obyvatel
2	200 až 499 obyvatel
3	500 až 999 obyvatel
4	1 000 až 1 999 obyvatel
5	2 000 až 4 999 obyvatel
6	5 000 až 9 999 obyvatel
7	10 000 až 49 999 obyvatel
8	50 000 až 99 999 obyvatel
9	100 000 a více obyvatel

Druhý soubor obsahuje data zjišťovaná celkem u 21 379 jednotlivců v oblastech:

- DOHLED - zda jednotlivec byl v době šetření na původní adrese, nebo se přestěhoval.

Hodnota	Význam hodnoty
0	jednotlivec byl na původní adrese
1	jednotlivec byl na jiné adrese

- POHL - pohlaví.

Hodnota	Význam hodnoty
1	muž
2	žena

- VZD - nejvyšší dokončené vzdělání šetřeného jedince.

Hodnota	Význam hodnoty
0	předškolní děti, neukončený 1. stupeň ZŠ
1	první stupeň ZŠ
2	druhý stupeň ZŠ
3	vyučení, nižší střední vzdělání
4	úplné střední vzdělání s maturitou
5	nástavbové studium, pomatur. kurzy, absolvování 2 či více SŠ
6	vyšší odborné studium (DiS.)
7	vysokoškolské bakalářské studium
8	vysokoškolské magisterské či inženýrské
9	doktorské studium (Ph.D., CSc., DrSc.)

- VEK - věk dosažený k 31.12.2009.

- CPPRIJ - čistý příjem jednotlivce za rok 2009.

- EA - převažující ekonomická aktivita jedince.

Hodnota	Význam hodnoty
1	zaměstnanec - plný úvazek
2	zaměstnanec - částečný úvazek
3	samostatně činný - plný úvazek
4	samostatně činný - částečný úvazek
5	nezaměstnaný
6	student
7	ve starobním důchodu
8	v invalidním důchodu
9	v domácnosti, péče o děti nebo jiné osoby blízké
0	ostatní ekonomicky neaktivní

Při analýze dat se nejprve zaměříme na domácnosti a poté na jednotlivce. Výpočty a tvorbu grafů budeme provádět ve statistickém softwaru R za pomocí nápovědy [15]. Excelovský soubor s daty nejprve převedeme na formát csv. Pro načtení dat použijeme funkci read.csv2().

5.1 Četnostní tabulky pro domácnosti

Pro lepší představu o přestěhovaných domácnostech uvedeme četnostní tabulky pro kategoriální proměnné, které získáme pomocí příkazu table(). Jak již bylo řečeno, celkový počet domácností je 9098 a z nich se přestěhovalo 260, což je 2,868%. Z tabulek četností můžeme zjistit, zda stěhování není typické pro nějakou skupinu domácností a při jejich vynechání bychom mohli ztratit právě charakteristické rysy této skupiny.

V tabulce 1 vidíme, že největší procento odstěhovaných je u domácností se 7 členy. Nicméně počet těchto domácností je v souboru malý a odstěhování jedné domácnosti může být náhodné. Obdobně to můžeme říci i o ostatních domácnostech, které jsou v souboru zastoupeny v malém rozsahu. Další největší procenta přestěhovaných mají čtyřčlenné domácnosti a tříčlenné domácnosti.

OSOB/DOHLED	0	1	% přestěhovaných
1	2499	62	2,421
2	3081	70	2,222
3	1500	59	3,784
4	1364	59	4,146
5	302	9	2,862
6	66	0	0
7	20	1	4,762
8	5	0	0
9	1	0	0

Tabulka 1: Stěhování domácností podle počtu osob

Z tabulky 2 je zřejmé, že procenta u domácností s jedním a dvěma ekonomicky aktivními členy jsou více jak dvojnásobná, než je tomu u ostatních. Při vynechání migrujících subjektů můžeme tedy ztratit část informace právě u tohoto druhu domácností.

EA/DOHLED	0	1	% přestěhovaných
0	3385	61	1,770
1	2706	107	3,804
2	2259	90	3,831
3	397	1	0,251
4	86	1	1,149
5	5	0	0

Tabulka 2: Stěhování domácností podle počtu ekonomicky aktivních osob

NAKL_ZATEZ/DOHLED	0	1	% přestěhovaných
1	2180	97	4,260
2	5843	151	2,519
3	815	12	1,451

Tabulka 3: Stěhování domácností podle nákladové zátěže

VEL/DOHLED	0	1	% přestěhovaných
1	231	10	4,149
2	584	11	1,849
3	869	18	2,029
4	830	29	3,376
5	1041	36	3,343
6	737	25	3,281
7	1925	64	3,218
8	1032	34	3,189
9	1589	33	2,035

Tabulka 4: Stěhování domácností podle velikosti měst

Nejvyšší procentuální hodnota v tabulce 3 je v prvním řádku. Proto musíme také vzít v úvahu, zda při vynechání migrujících domácností neztratíme i část informace o subjektech s velkou nákladovou zátěží.

V tabulce 4 je největší procento u nejmenšího města, oproti tomu nejmenší procento je u druhého nejmenšího města. Zde by nemusela být ztráta informace významná, protože rozdílnost domácností žijících ve dvou nejmenších městech nemusí být tak velká.

5.2 Číselné charakteristiky pro soubor domácnosti

V následujících tabulkách jsou uvedeny vypočtené číselné charakteristiky pro jednotlivé statistické znaky. Postupně pro migrující domácnosti, poté pro nemigrující domácnosti a pro celý soubor. K výpočtu pomocí softwaru R použijeme funkce mean(), median(), quantile(), var(), sd(), IQR(), MAD(), sd()/mean(), max()-min(), mean(scale()^ 3), mean(scale()^ 4)−3.

Jak je vidět z tabulky 5 a 6 charakteristiky polohy migrujících domácností se liší od nemigrujících, což značí, že migrující domácnosti mají trochu jiné rozložení hodnot proměnných. Nejmenší odlišnost je u velikosti města, tedy velikost města nemá na migraci vliv. Odlišné hodnoty můžeme vidět i u šikmosti a špičatosti. Například u příjmů a nákladů jsou hodnoty migrujících domácností méně sešikmené k vyšším hodnotám a jsou méně soustředěny kolem středu.

Dále se zaměříme na porovnání tabulek 6 a 7. Tyto charakteristiky se ve většině případů liší jen velmi málo, což je dáno již tím, že proporcionální zastoupení migrujících domácností v celkovém datovém souboru je relativně malé. Největší odlišnost je u dolního kvartilu nákladové zátěže. I u koeficientů šikmosti a špičatosti jsou velmi malé odlišnosti.

	Průměr	Medián	D. kvartil	H. kvartil	Šikmost	Špičatost
OSOB	2,5654	2	2	4	0,3154	-0,6144
EA	1,1308	1	1	2	0,0066	-0,6576
NAKLADY	4901,946	4499,5	3311,0	5964,5	1,3171	3,3313
NAKL_ZATEZ	1,6731	2	1	2	0,0818	-0,6984
CP_PRIJ	342505,2	300546,5	178220,0	412498,2	2,5825	9,2677
VEL	5,9577	7	4	8	-0,4882	-0,6262

Tabulka 5: Charakteristiky polohy, šikmost a špičatost pro migrující domácnosti

Z tabulek variabilit 8, 9, 10 opět vidíme větší rozdíly pro migrující domácnosti. Pokud porovnáme charakteristiky variability pro všechny domácnosti a pro nemigrující, vidíme, že rozdíly ve variabilitě jsou již menší. Při zaměření na charakteristiku IQR zjistíme na první pohled změnu u nákladové zátěže, dále je změna i u příjmů. U ostatních proměnných nejsou rozdíly, nebo v případě nákladů velmi malé.

	Průměr	Medián	D. kvartil	H. kvartil	Šikmost	Špičatost
OSOB	2,3435	2	1	3	0,8035	0,2873
EA	0,9939	1	0	2	0,6257	-0,2485
NAKLADY	4727,705	4434,5	3383,25	5655,75	3,2247	35,8404
NAKL_ZATEZ	1,8456	2	2	2	-0,0271	-0,0759
CP_PRIJ	332889,2	280110	186000,0	421733,2	4,7420	55,7354
VEL	5,9951	7	4	8	-0,3941	-0,9578

Tabulka 6: Charakteristiky polohy, šikmost a špičatost pro nemigrující domácnosti

	Průměr	Medián	D. kvartil	H. kvartil	Šikmost	Špičatost
OSOB	2,3499	2	1	3	0,7886	0,2517
EA	0,9978	1	0	2	0,6122	-0,2560
NAKLADY	4732,684	4438,5	3382	5660	3,1580	34,5957
NAKL_ZATEZ	1,8406	2	1	2	-0,0245	-0,0960
CP_PRIJ	333164	280416	185760,8	421681,5	4,6641	53,8931
VEL	5,9941	7	4	8	-0,3963	-0,9491

Tabulka 7: Charakteristiky polohy, šikmost a špičatost pro všechny domácnosti

	Rozptyl	Směr. odch.	IQR	MAD	V
OSOB	1,4668	1,2111	2	1,4826	0,4721
EA	0,6160	0,7849	1	1,4826	0,6941
NAKLADY	5603066	2367,08	2653,5	1968,151	0,4829
NAKL_ZATEZ	0,3136	0,5600	1	0	0,3347
CP_PRIJ	65512648865	255954,4	234278,2	177244,1	0,7473
VEL	4,7743	2,1850	4	2,9652	0,3668

Tabulka 8: Charakteristiky variability pro migrující domácnosti

	Rozptyl	Směr. odch.	IQR	MAD	V
OSOB	1.4615	1.2089	2	1.4826	0.5159
EA	0.9150	0.9565	2	1.4826	0.9624
NAKLADY	4788997	2188.378	2272.5	1674.597	0.4629
NAKL_ZATEZ	0.3151	0.5613	0	0	0.3041
CP_PRIJ	54212956625	232836.8	235733.2	172552.4	0.6994
VEL	5.3926	2.3222	4	2.9652	0.3873

Tabulka 9: Charakteristiky variability pro nemigrující domácnosti

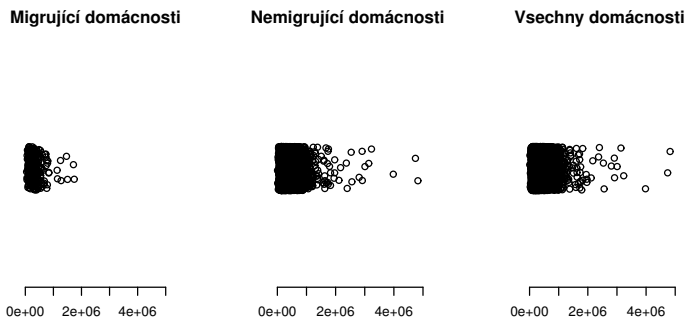
	Rozptyl	Směr. odch.	IQR	MAD	V
OSOB	1,4628	1,2095	2	1,4826	0,5147
EA	0,9069	0,9523	2	1,4826	0,9544
NAKLADY	4812490	2193,739	2278	1680,527	0,4635
NAKL_ZATEZ	0,3158	0,5620	1	0	0,3053
CP_PRIJ	54531277141	233519,3	235920,8	172741,4	0,7009
VEL	5,3745	2,3183	4	2,9652	0,3868

Tabulka 10: Charakteristiky variability pro všechny domácnosti

5.3 Grafy pro soubor domácností

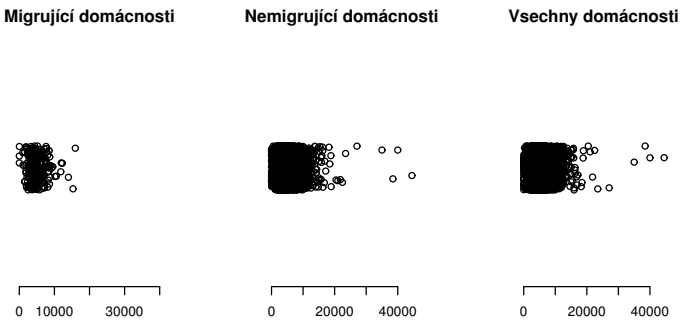
V následující sekci se zaměříme na porovnání souborů pomocí grafických nástrojů. Některé grafy je vhodné použít jen pro určitý druh proměnných. A proto vybrané grafy uvedeme jen pro některé statistické znaky, pro které to bude mít nějakou vypovídající schopnost.

5.3.1 Diagram rozptýlení a rozmítnutý diagram rozptýlení



Obrázek 10: Diagram rozptýlení a rozmítnutý diagram rozptýlení pro příjmy

Z obrázku 10 vyplývá, že přestěhované domácnosti jsou převážně s příjmem přibližně do 700 000 Kč. To může být způsobeno tím, že domácnosti vydělávající vysoké částky nemají potřebu se stěhovat právě z důvodu, že v místě bydliště mají dobré zaměstnání. U grafu pro nemigrující a pro všechny domácnosti není viditelný rozdíl, což může být z důvodu velkého množství překrývajících se dat. Tento graf jsem získali v softwaru R pomocí příkazu stripchart().



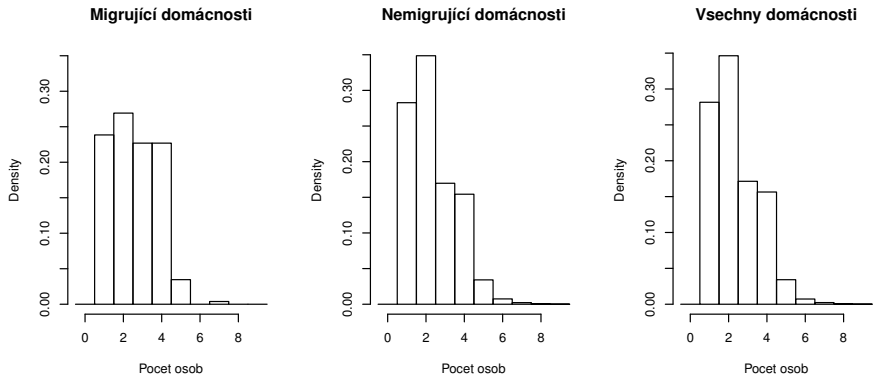
Obrázek 11: Diagram rozptýlení a rozmítnutý diagram rozptýlení pro náklady

V grafu 11 vidíme, že přestěhované domácnosti jsou ty, které mají nízké náklady na bydlení (do 10 000 Kč). Výjimkou jsou domácnosti s velmi nízkými náklady (do 2 000 Kč), které dle grafu převážně podléhají šetření na původní adrese. To může být z toho důvodu, že domácnost bydlí ve vlastním bytě nebo domě, který je již splacený a náklady na bydlení tvoří pouze energie. Grafy pro nemigrované domácnosti a všechny domácnosti nejsou na pohled dle rozložení hodnot výrazně odlišné, ale data se nám opět překrývají.

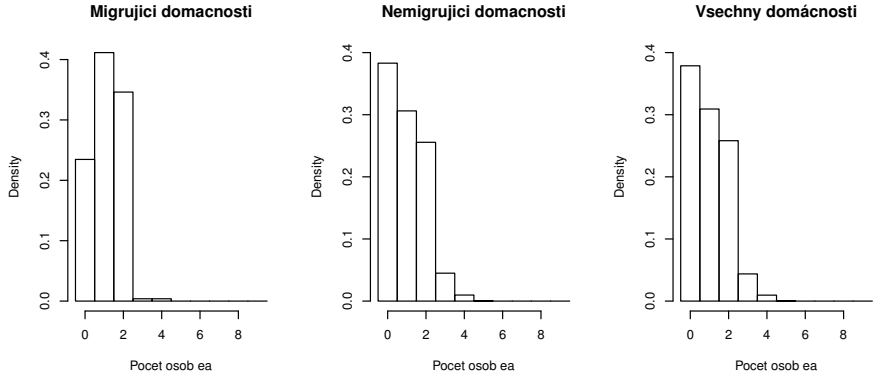
5.3.2 Histogram

Pro přehled o rozložení spojitých proměnných uvedeme histogram. V softwaru R použijeme funkce hist(). Grafy pro četnosti uvedeme i pro diskrétní a kategoriální proměnné, abychom mohli porovnat i rozdíly relativních četností těchto proměnných pro migrující, nemigrující a všechny domácnosti.

Z uvedených grafů pro kategoriální proměnné vidíme, že relativní četnosti se při vynechání migrujících domácností viditelně neliší od relativních četností všech domácností. Oproti tomu u migrujících domácností jsou tyto rozdíly na první pohled zřetelné. Největší rozdíl v relativních četnostech je u počtu ekonomicky aktivních osob. Zde je vidět, že daleko menší zastoupení u přestěhovaných mají domácnosti bez ekonomicky aktivních členů. Větší rozdíl můžeme pozorovat i u ekonomické zátěže, kde je větší zastoupení migrujících domácností s velkou ná-



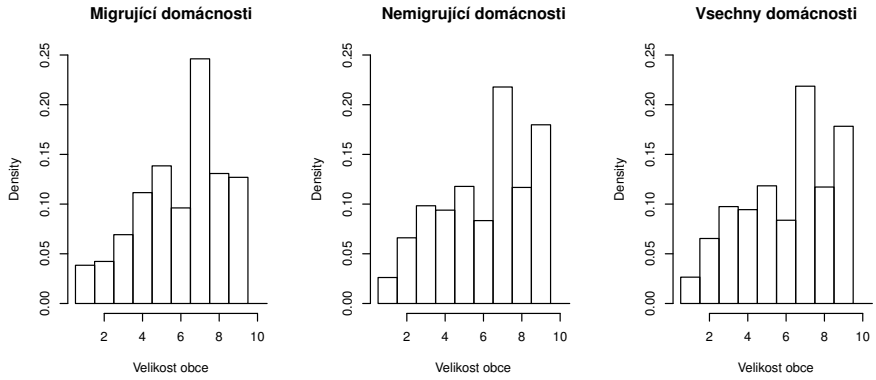
Obrázek 12: Relativní četnosti pro počet osob



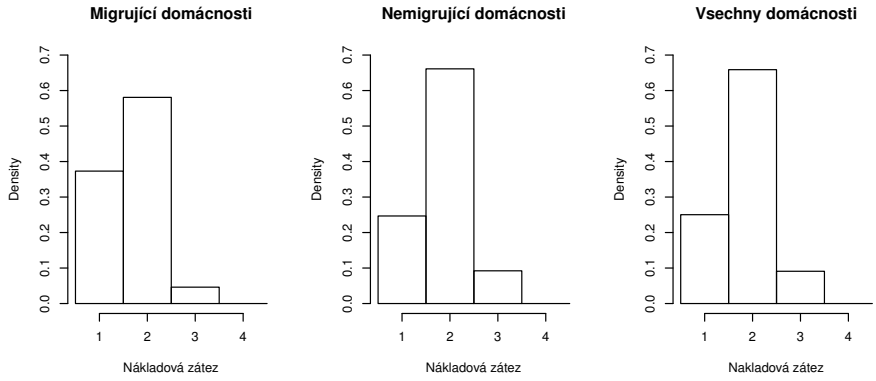
Obrázek 13: Relativní četnosti pro počet osob ea

kladovou zátěží. Nejmenší rozdíl můžeme vidět u velikosti města, tedy struktura souboru nemigrujících domácností je hodně podobná struktuře všech domácností.

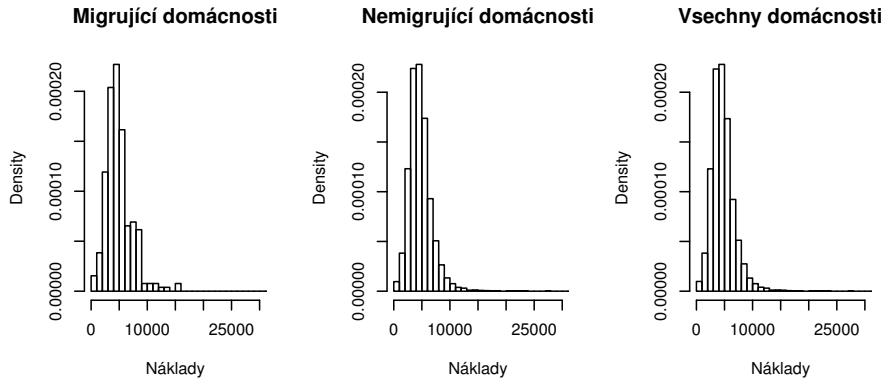
Pokud se podíváme na histogram nákladů (obrázek 16) a příjmů (obrázek 17), zjistíme, že rozdělení těchto spojitých proměnných se také nijak neliší při vypuštění hodnot přestěhovaných domácností oproti souboru pro všechny domácnosti. Dle tvaru histogramu vidíme, že náklady a příjmy mají sešikmení k vyšším hodnotám, což je více vidět u příjmů.



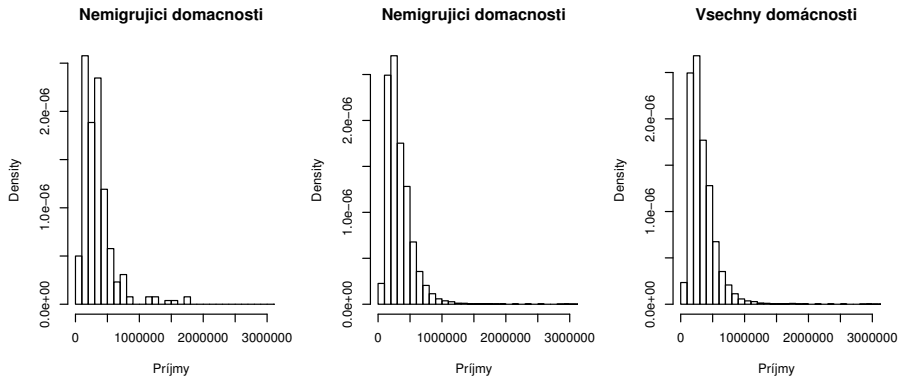
Obrázek 14: Relativní četnosti pro velikost obce



Obrázek 15: Relativní četnosti pro nákladovou zátěž



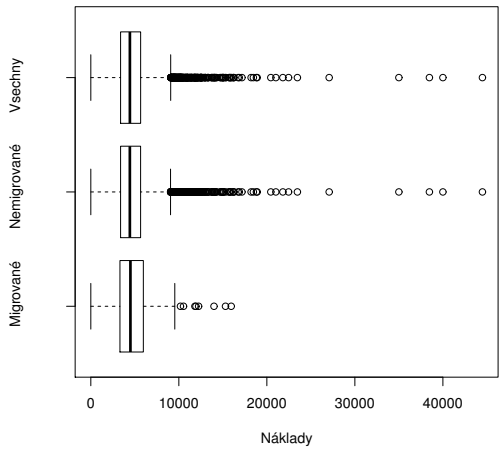
Obrázek 16: Histogram pro náklady



Obrázek 17: Histogram pro příjmy

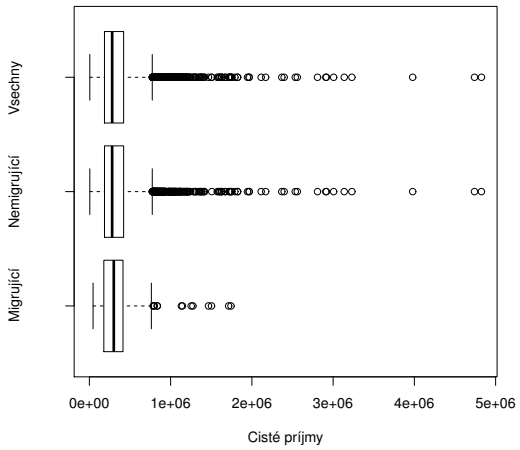
5.3.3 Boxplot

Pro grafické znázornění mediánů, kvartilů a odlehlých hodnot použijeme boxplot. K získání tohoto grafu v softwaru R slouží funkce boxplot(). V jednotlivých grafech jsou uvedeny boxploty spojitých proměnných vždy pro migrované domácnosti, nemigrované domácnosti a pro všechny domácnosti.



Obrázek 18: Boxplot pro náklady

I z obrázku 18 vidíme, že tvar boxplotu se při vynechání migrujících domácností neliší od boxplotu pro všechny domácnosti. Migrující domácnosti mají



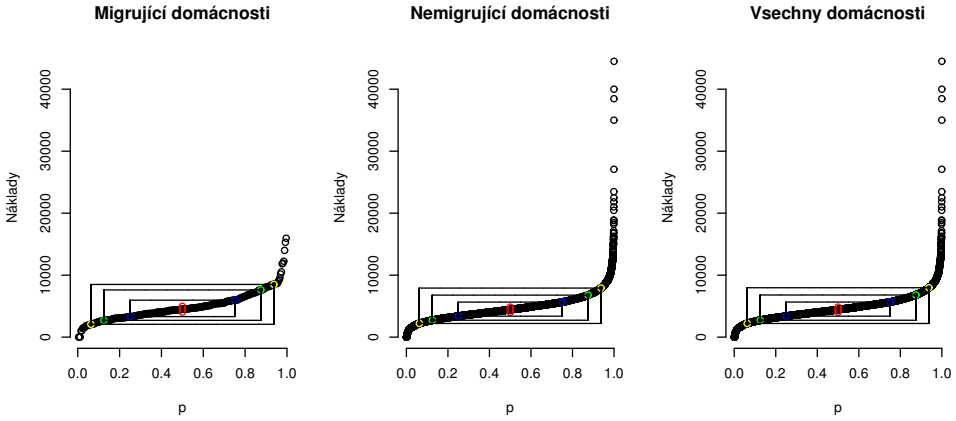
Obrázek 19: Boxplot pro příjmy

větší interkvartilové rozpětí a nemají tolik odlehlých hodnot. To vypovídá opět o tom, že se nestěhují domácnosti s velkými náklady na bydlení. To můžou být právě ty domácnosti, které mají velký příjem a pokryjí i velké náklady na bydlení bez problémů.

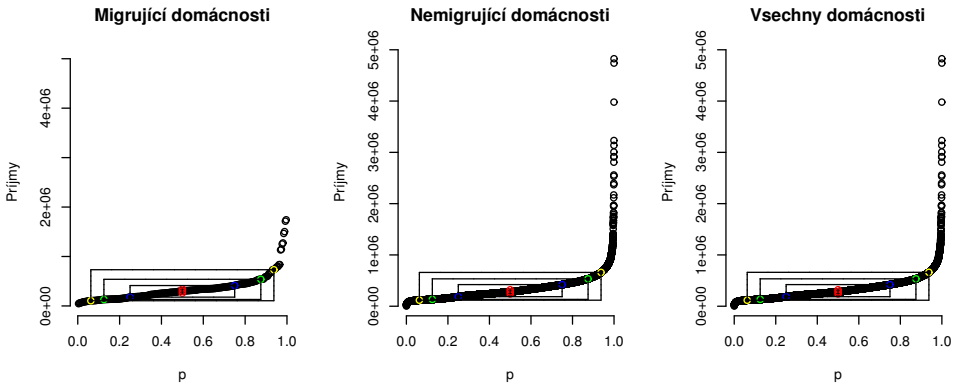
Větší sešikmení k vyšším hodnotám je u boxplotu příjmů 19. Zde vidíme, že tvar boxplotu pro nemigrované domácnosti a všechny domácností se také viditelně nezmění. U migrujících domácností je znatelný posun mediánu k vyšším hodnotám, a že se nestěhují domácnosti s extrémně velkým příjmem.

5.3.4 Graf rozptýlení s kvantily

Pro získání grafu rozptýlení s kvantily jsme použili klasickou funkci plot(). Pořadové pravděpodobnosti jsme určili pomocí vzorce $p_i = \frac{i}{n+1}$, kde i je pořadí hodnoty uspořádaného datového souboru a n je celkový počet hodnot. Grafy 20 a 21 nám potvrzují sešikmení rozdělení nákladů i příjmů k vyšším hodnotám, protože vzdálenosti mezi dolními hradbami obdélníků jsou menší než vzdálenosti mezi horními hradbami. Pokud srovnáme graf pro všechny domácnosti a pro nemigrující, nevidíme rozdíl ve tvaru křivky, do níž jsou data seřazena. Velmi podobný tvar u obou proměnný má i křivka pro migrující domácnosti s tím roz-



Obrázek 20: Graf rozptýlení s kvantily pro náklady

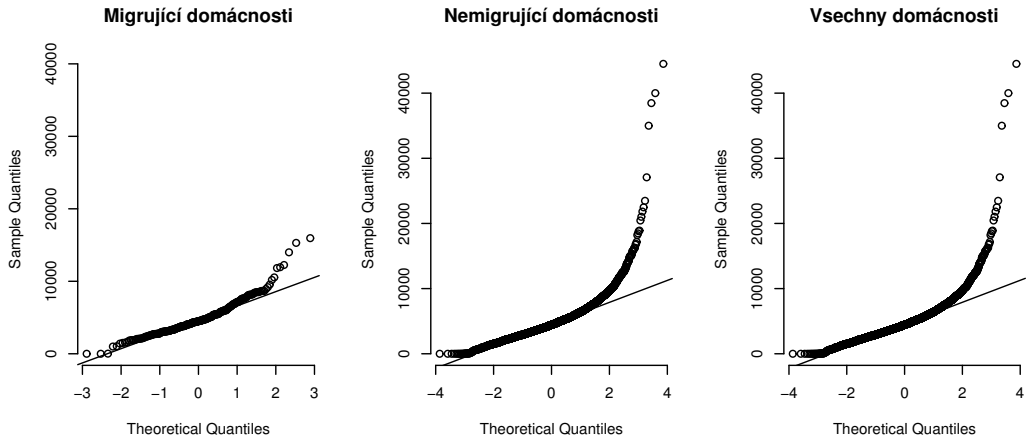


Obrázek 21: Graf rozptýlení s kvantily pro příjmy

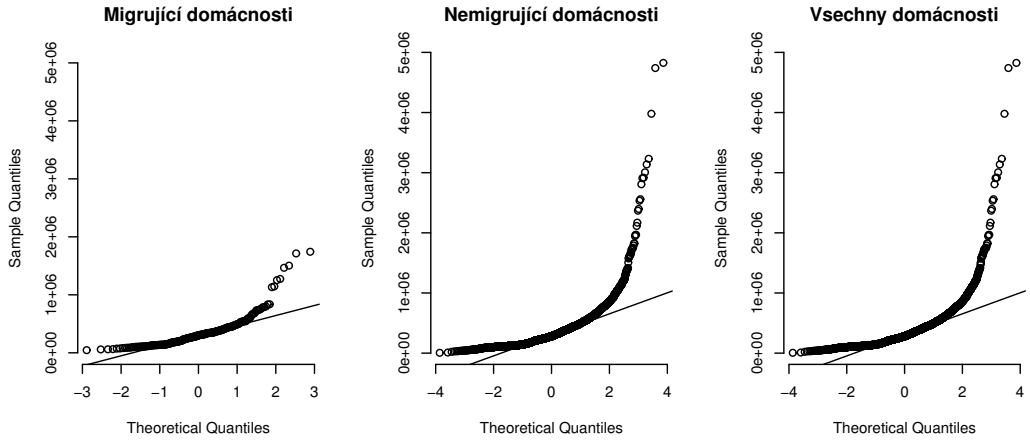
dílem, že křivka končí dříve, což opět vypovídá o nestěhování domácností s velmi vysokými náklady a příjmy.

5.3.5 Q-Q graf

V této části budeme srovnávat rozdělení nákladů a příjmů s normálním rozdělením. V softwaru R použijeme funkce qqnorm() a qqline(). V obrázku 22 a 23 opět vidíme, že náklady a příjmy mají sešikmení k vyšším hodnotám. Toto sešikmení je daleko menší u migrujících domácností, kde opět můžeme vidět, že se nestěhují domácnosti s vysokými příjmy a náklady.



Obrázek 22: Q-Q graf pro náklady

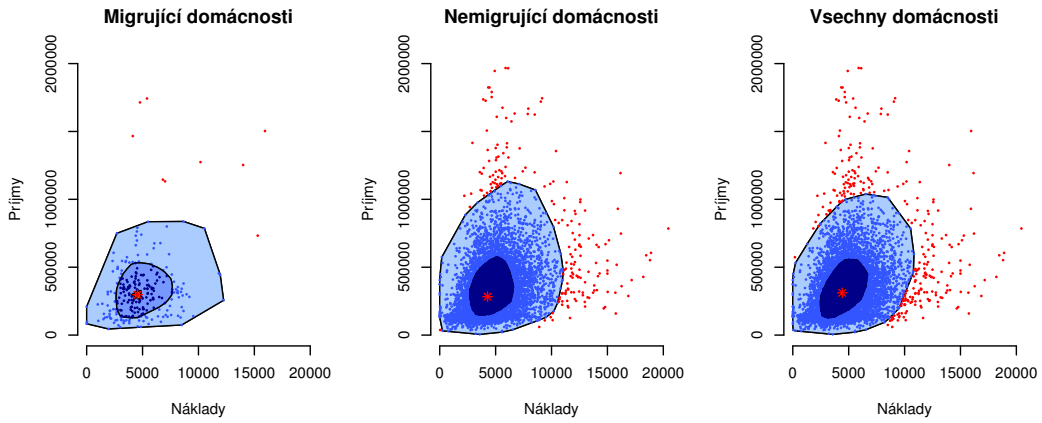


Obrázek 23: Q-Q graf pro příjmy

5.3.6 Bagplot

Pro vykreslení bagplotu v softwaru R je zapotřebí balíček aplpack, ve kterém použijeme funkci bagplot(). V souboru domácnosti vytvoříme tento graf pro spojité proměnné náklady a příjmy.

Dle orientace grafu bagplotu 24 vidíme kladnou korelaci mezi proměnnými náklady a příjmy, ta je zřetelnější v grafu pro všechny domácnosti. Dále můžeme opět pozorovat asymetrii rozdělení dat z vejčitého tvaru bagu a umístění mediá-



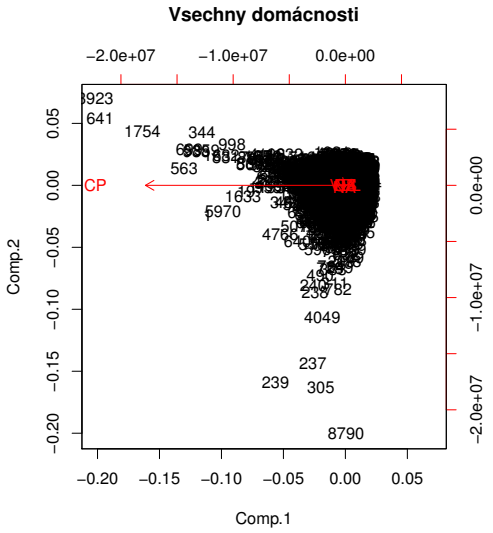
Obrázek 24: Bagplot pro náklady a příjmy

nu, který se nenachází uprostřed bagu, což je více vidět v prostředním grafu, který znázorňuje nemigrované domácnosti. Tedy při srovnání tvaru bagplotu pro všechny domácnosti a pro nemigrované domácnosti vidíme rozdíl ve tvaru bagu a smyčky i v umístění v mediánu. Migrované domácnosti mají už výraznější rozdíl, což je způsobeno i velmi rozdílným počtem hodnot.

5.3.7 Biplot

Pro získání informace o uspořádání objektů a o korelaci mezi jednotlivými statistickými znaky použijeme biplot, který v softwaru R získáme pomocí funkce biplot(princomp(x)). K vypsání informací o jednotlivých hlavních komponentách použijeme funkci summary(princomp(x)). Názvy proměnných jsme označili pomocí zkratek: OS (počet osob), EA (počet ekonomicky aktivních osob), NA (náklady), NZ (nákladová zátěž), CP (příjmy), VEL (velikost města).

Biplot na obrázku 25, který je sestaven z dat pro všechny domácnosti, není vůbec přehledný. Důvodem je, že proměnná příjmy má největší vliv (její hodnoty jsou daleko vyšší, než jsou hodnoty ostatních statistických znaků). Uspořádání objektů v grafu zprava doleva nám dává informaci o velikosti příjmů. Objekty, které jsou nejvíce vlevo, mají největší příjmy. Bohužel jinou informaci z biplotu nemůžeme získat. Nevhodnost tohoto biplotu můžeme vidět i z tabulky 11, kde



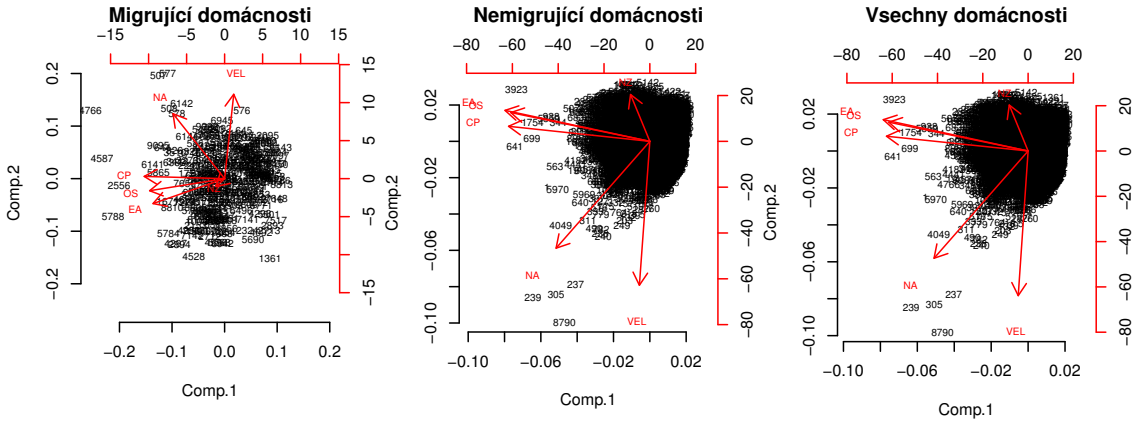
Obrázek 25: Biplot pro všechny domácnosti

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	2,335075e+05	2,086689e+03	2,232581e+00	1,075305e+00
Proportion of Variance	9,999201e-01	7,985075e-05	9,140676e-11	2,120444e-11
Cumulative Proportion	9,999201e-01	1,000000e+00	1,000000e+00	1,000000e+00
	Comp.5	Comp.6		
Standard deviation	6,523775e-01	5,328649e-01		
Proportion of Variance	7,804798e-12	5,207124e-12		
Cumulative Proportion	1,000000e+00	1,000000e+00		

Tabulka 11: Charakteristiky hlavních komponent

ve druhém řádku vidíme, že první hlavní komponenta vysvětluje 99,9% celkové variability souboru. Pro migrující a nemigrující domácnosti by nastala stejná situace, proto příslušné grafy biplotů nebudeme uvádět.

K řešení tohoto problému slouží škálování. Po přeškálování hodnot sestavíme nový biplot, kde již nebudou vysoké nebo nízké hodnoty statistických znaků ovlivňovat interpretaci biplotu. V softwaru R provedeme škálování pomocí funkce scale(x, center = TRUE, scale = TRUE) a z těchto nových hodnot získáme biploty v obrázku 26 a uvedeme charakteristiky hlavních komponent (tabulka 12 pro migrující domácnosti, tabulka 13 pro nemigrující domácnosti, tabulka 14 pro všechny domácnosti). Ve třetím řádku můžeme vyčíst, že první dvě hlavní



Obrázek 26: Biplot pro domácnosti po škálování

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1,4432839	1,1247323	1,0630860	0,7940234	0,67945442
Proportion of Variance	0,3485185	0,2116512	0,1890859	0,1054846	0,07724013
Cumulative Proportion	0,3485185	0,5601697	0,7492556	0,8547402	0,93198032
	Comp.6				
Standard deviation	0,63761147				
Proportion of Variance	0,06801968				
Cumulative Proportion	1,00000000				

Tabulka 12: Charakteristiky hlavních komponent pro migrující domácnosti

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1,5304622	1,1060684	1,0322776	0,74963975	0.66165047
Proportion of Variance	0,3904299	0,2039209	0,1776196	0,09367056	0.07297181
Cumulative Proportion	0,3904299	0,5943509	0,7719705	0,86564103	0.93861284
	Comp.6				
Standard deviation	0,60686182				
Proportion of Variance	0,06138716				
Cumulative Proportion	1,00000000				

Tabulka 13: Charakteristiky hlavních komponent pro nemigrující domácnosti

komponenty u tabulek pro všechny a nemigrující domácnosti vysvětlují necelých 60% celkové variabilikty, u migrujících domácností je to 56%.

Z obrázku 26 vidíme, že grafy pro nemigrující a všechny domácnosti se viditelně neliší. Pokud se zaměříme na úhly mezi šipkami, zjistíme, že největší

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1,5280408	1,1063169	1,0334972	0,74946487	0.6620635
Proportion of Variance	0,3891942	0,2040119	0,1780390	0,09362656	0.0730627
Cumulative Proportion	0,3891942	0,5932062	0,7712451	0,86487170	0.9379344
	Comp.6				
Standard deviation	0,6102070				
Proportion of Variance	0,0620656				
Cumulative Proportion	1,0000000				

Tabulka 14: Charakteristiky hlavních komponent pro všechny domácnosti

korelaci budou mezi sebou mít znaky počet osob a počet ekonomicky aktivních osob. Vysokou korelaci budou mít i dvojice proměnných počet osob a příjmy, počet ekonomicky aktivních osob a příjmy. Tyto korelace se sníží v případě migrujících domácností. Dále můžeme sledovat, že velikost příjmů obecně prakticky nesouvisí s velikostí obce. Přitom s rostoucí velikostí obce rostou náklady. Pokud se zaměříme na uspořádání objektů v grafu, opět pozorujeme, že subjekty, které mají hodně vysoké příjmy i náklady, jsou nemigrující.

5.4 Korelační matice

Pro zjištění vztahů mezi jednotlivými statistickými znaky použijeme kromě biplotu korelační matici, jejíž prvky jsou párové korelační koeficienty jednotlivých proměnných.

Korelační matice pro migrují domácnosti:

	OSOB	EA	NAKLADY	NAKL_ZATEZ	CP_PRIJ	VEL
OSOB	1,0000	0,3972	0,3170	-0,0794	0,4463	-0,1864
EA	0,3972	1,0000	0,1321	0,0625	0,4946	-0,1318
NAKLADY	0,3197	0,1321	1,0000	-0,1105	0,3272	0,2734
NAKL_ZATEZ	-0,0794	0,0625	-0,1105	1,0000	0,2139	0,0612
CP_PRIJ	0,4463	0,4946	0,3272	0,2139	1,0000	-0,0147
VEL	-0,1864	-0,1318	0,2734	0,0612	-0,0147	1,0000

Korelační matice pro nemigrují domácnosti:

	OSOB	EA	NAKLADY	NAKL_ZATEZ	CP_PRIJ	VEL
OSOB	1,0000	0,6031	0,3016	-0,0315	0,5143	-0,0953
EA	0,6031	1,0000	0,2717	0,1013	0,5912	-0,0344
NAKLADY	0,3016	0,2717	1,0000	-0,0979	0,3077	0,2850
NAKL_ZATEZ	-0,0315	0,1013	-0,0979	1,0000	0,1728	0,0335
CP_PRIJ	0,5143	0,5912	0,3077	0,1728	1,0000	0,0346
VEL	-0,0953	-0,0344	0,2850	0,0335	0,0346	1,0000

Korelační matice pro všechny domácnosti:

	OSOB	EA	NAKLADY	NAKL_ZATEZ	CP_PRIJ	VEL
OSOB	1,0000	0,5983	0,3024	-0,0344	0,5120	-0,0978
EA	0,5983	1,0000	0,2682	0,0990	0,5881	-0,0366
NAKLADY	0,3024	0,2682	1,0000	-0,0988	0,3084	0,2846
NAKL_ZATEZ	-0,0344	0,0990	-0,0988	1,0000	0,1735	0,0344
CP_PRIJ	0,5120	0,5881	0,3084	0,1735	1,0000	0,0331
VEL	-0,0978	-0,0366	0,2846	0,0344	0,0331	1,0000

Stejně jako jsme viděli u biplotu i zde má nejvyšší korelační koeficient dvojice počet osob a počet ekonomicky aktivních osob. Oproti tomu nejmenší korelaci vidíme u dvojic počet osob a nákladová zátěž, počet ekonomicky aktivních osob a velikost města. V případě srovnání korelačních koeficientů pro soubor se všemi domácnostmi a pro nemigrující domácnosti jsou vidět rozdíly v hodnotách jen malé. Větší rozdíly jsou v tabulce pro migrující domácnosti, kde má tentokrát největší korelaci dvojice příjem a počet ekonomicky aktivních osob.

5.5 Testy dobré shody

V rámci testů dobré shody chceme zjistit, zda se struktury hodnot jednotlivých kategoriálních proměnných migrujících domácností a nemigrujících domácností liší od struktury všech domácností. Budeme tedy testovat nulovou hypotézu, zda $p_i = p_i^0$, $i = 1, \ldots, k$, kde p_i^0 jsou pravděpodobnosti daných kategorií v souboru pro všechny domácnosti a k je počet tříd dané proměnné. Pokud nulovou hypotézu zamítneme, značí to, že struktury souborů jsou odlišné. V souboru domácností máme dvě kategoriální proměnné - nákladová zátěž a velikost města. Nejprve si tedy vypočítáme teoretické pravděpodobnosti pro nákladovou zátěž ze souboru všech domácností, kde m je celkový počet domácností, m_i je počet domácností v dané kategorii a $k = 3$:

NAKL_ZATEZ	m_i	$p_i^0 = \frac{m_i}{m}$
1	2277	0,2503
2	5994	0,6588
3	827	0,0909
m	9098	

Nyní tyto pravděpodobnosti použijeme pro výpočet testovací statistiky $Z = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}$ pro soubor nemigrujících domácností:

NAKL_ZATEZ	n_i	p_i^0	np_i	$\frac{(n_i - np_i^0)^2}{np_i^0}$
1	2180	0,2503	2212,15	0,4672
2	5843	0,6588	5822,44	0,0726
3	815	0,0909	803,37	0,1684
n	8838		Z	0,7082

Kde n je celkový počet nemigrujících domácností a n_i počet nemigrujících domácností i-té třídy. Hodnotu testovacího kritéria $Z = 0{,}7082$ porovnáváme s kritickým oborem $W = \langle \chi_2^2(0,95), \infty) = \langle 5,991; \infty)$. Hodnota Z do kritického oboru nepatří, proto nulovou hypotézu o rovnosti pravděpodobností nelze zamítnout. Struktura četností v případě nemigrujících domácností se tedy (na hladině testu $\alpha = 0,05$) neliší od struktury všech vyšetřovaných domácností.

Stejným postupem vypočteme hodnotu testovací statistiky pro migrující domácnosti, kde tentokrát n je počet migrujících domácností a n_i počet migrujících domácností v i-té kategorii:

NAKL_ZATEZ	n_i	p_i^0	np_i	$\frac{(n_i - np_i^0)^2}{np_i^0}$
1	97	0,2503	65,08	15,656
2	151	0,6588	171,29	2,403
3	12	0,0909	23,63	5,724
n	260		Z	23,78

Kritický obor je opět $W = \langle \chi_2^2(0{,}95), \infty) = \langle 5{,}991; \infty)$. Hodnota testovací statistiky leží uvnitř intervalu W, proto nulovou hypotézu zamítáme a může říci, že struktura nákladové zátěže u migrujících domácností není stejná jako struktura u všech domácností.

Dále vypočítáme teoretické pravděpodobnosti ze souboru všech domácností pro kategoriální proměnnou velikost města, kde $k = 9$:

VEL	m_i	$p_i^0 = \frac{m_i}{m}$
1	241	0,0265
2	595	0,0654
3	887	0,0975
4	859	0,0944
5	1077	0,1184
6	762	0,0837
7	1989	0,2186
8	1066	0,1172
9	1622	0,1783
m	9098	

A určíme hodnotu testovacího kritéria pro nemigrující domácnosti:

VEL	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
1	231	0,0265	234,113	0,041
2	584	0,0654	577,996	0,062
3	869	0,0975	861,652	0,063
4	830	0,0944	834,452	0,024
5	1041	0,1184	1046,222	0,026
6	737	0,0837	740,224	0,014
7	1925	0,2186	1932,159	0,0267
8	1032	0,1172	1035,536	0,012
9	1589	0,1783	1575,647	0,113
n	8838		Z	0,382

Kritický obor je v tomto případě $W = \langle \chi_8^2(0{,}95), \infty) = \langle 15{,}51; \infty)$ a s hodnotou $Z = 0{,}382$ nulovou hypotézu nelze zamítnout.

Nyní otestujeme shodu s migrujícími domácnostmi:

VEL	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
1	10	0,0265	6,8872	1,4069
2	11	0,0654	17,0037	2,1198
3	18	0,0975	25,3484	2,1303
4	29	0,0944	24,5483	0,8073
5	36	0,1184	30,7782	0,8859
6	25	0,0837	21,7762	0,4773
7	64	0,2186	56,8411	0,9016
8	34	0,1172	30,4638	0,4105
9	33	0,1783	46,3530	3,8466
n	260		Z	13,00

S hodnotou testovací statistiky $Z = 13{,}00$ nemůžeme nulovou hypotézu zamítnout, protože kritický obor je opět $W = \langle \chi^2_8(0{,}95), \infty) = \langle 15{,}51; \infty)$ a tato hodnota do něj nepatří.

5.6 Četnostní tabulky pro jednotlivce

Pro každou kategoriální proměnnou uvedeme nejprve opět tabulky četností spolu s procentem přestěhovaných osob v dané kategorii. Celkem máme v souboru 21 379 jednotlivců, z nichž se přestěhovalo 667 lidí, tedy 3,120%.

POHL/DOHLED	0	1	% přestěhovaných
1	9941	309	3,015
2	10771	358	3,217

Tabulka 15: Stěhování jednotlivců podle pohlaví

Z tabulky 15 vidíme, že procento přestěhovaných mužů není moc odlišné od procenta přestěhovaných žen. Proto můžeme říci, že pohlaví nemá vliv na stěhování.

Pokud se zaměříme na stěhování osob podle vzdělání (tabulka 16), zjistíme, že procento u skupin 0 (předškolní, neukončený 1.stupeň ZŠ), 1 (1.stupeň ZŠ) a 6 (DiS.) je znatelně vyšší, než je tomu u ostatních skupin. Proto při vynechání přestěhovaných osob může dojít ke zkreslení informace.

Nejvetší rozdíly procent vidíme u tabulky 17, kde se nejvíce stěhují osoby v domácnosti a ostatní ekonomicky neaktivní osoby, dále pak samostatně činní na částečný úvazek. Naopak nejméně se stěhují osoby v důchodu.

VZD/DOHLED	0	1	% přestěhovaných
0	2160	143	6,209
1	790	41	4,934
2	3093	79	2,491
3	6546	163	2,430
4	5717	174	2,954
5	205	4	1,914
6	139	6	4,138
7	341	10	2,849
8	1643	45	2,666
9	78	2	2,500

Tabulka 16: Stěhování jednotlivců podle vzdělání

EA/DOHLED	0	1	% přestěhovaných
0	3022	185	5,769
1	7391	245	3,208
2	230	4	1,709
3	1119	43	3,701
4	44	2	4,348
5	709	24	3,274
6	1506	43	2,776
7	5259	50	0,942
8	759	17	2,191
9	673	54	7,428

Tabulka 17: Stěhování jednotlivců podle ekonomické aktivity

Stěhování osob v domácnosti a osob předškolního a školního věku může být způsobeno rozvodovostí, kdy se jeden z manželů po rozvodu odstěhuje spolu s potomkem.

5.7 Charakteristiky pro soubor jednotlivců

V následujících tabulkách uvedeme charakteristiky pro soubor jednotlivci. Opět vždy pro migrující část respondentů, nemigrující část a pro všechny osoby dohromady.

Pokud porovnáme tabulky 18, 19 a 20, vidíme velký rozdíl u migrujících domácností, kdy se stěhují zejména mladší respondenti. Rozdíl je i u dalších proměnných kromě pohlaví, ale charakteristiky pro proměnnou pouze se dvěma kategoriemi nejsou moc vypovídající. Pokud vyřadíme všechny přestěhované osoby, charakteristiky polohy se oproti původnímu souboru více změní u příjmů.

	Průměr	Medián	D. kvartil	H. kvartil	Šikmost	Špičatost
POHL	1,5367	2	1	2	-0,1470	-1,9814
VZD	2,8306	3	1	4	0,6504	0,3479
VEK	28,9805	28	13	39	0,5763	-0,0487
CPRIJ	133510,3	110400	0	193354	3,5957	21,7783
EA	2,6087	1	0	5	0,9871	-0,5830

Tabulka 18: Charakteristiky polohy pro migrující jednotlivce

Obdobně je to i u charakteristik variability (tabulky 21, 22, 23), kde vidíme velké rozdíly u migrujících domácností. Při porovnání nemigrujících domácností a všech domácností jsou už rozdíly menší.

	Průměr	Medián	D. kvartil	H. kvartil	Šikmost	Špičatost
POHL	1,5200	2	1	2	-0,0802	-1,9937
VZD	3,2626	3	2	4	0,7619	0,9494
VEK	42,8762	44	24	61	-0,1084	-0,9878
CPRIJ	142046,9	130327	60800	190803	6,4629	115,4178
EA	3,5200	1	1	7	0,2918	-1,6366

Tabulka 19: Charakteristiky polohy pro nemigrující jednotlivce

	Průměr	Medián	D. kvartil	H. kvartil	Šikmost	Špičatost
POHL	1,5206	2	1	2	-0,0823	-1,9933
VZD	3,2491	3	2	4	0,7522	0,9246
VEK	42,4426	43	24	61	-0,0853	-0,9934
CPRIJ	141780,6	129960	55500	190985	6,3284	110,5124
EA	3,4916	1	1	7	0,3115	-1,6201

Tabulka 20: Charakteristiky polohy pro všechny jednotlivce

	Rozptyl	Směr. odch.	IQR	MAD	V
POHL	0,2490	0,4990	1	0,3247	0
VZD	4,6725	2,1616	3	1,4826	0,7637
VEK	381,2203	19,5249	26	19,2738	0,6737
CPRIJMY	28628523435	169199,7	193354	163679	1,2673
EA	9,3106	3,0513	5	1,4826	1,1697

Tabulka 21: Charakteristiky variability pro migrující jednotlivce

	Rozptyl	Směr. odch.	IQR	MAD	V
POHL	0,2496	0,4996	1	0	0,3286
VZD	3,9585	1,9896	2	1,4826	0,6124
VEK	512,842	22,6460	37	26,6868	0,5336
CPRIJMY	21074340495	145170	135485	94923,46	1,0239
EA	9,4447	3,0732	6	1,4826	0,8802

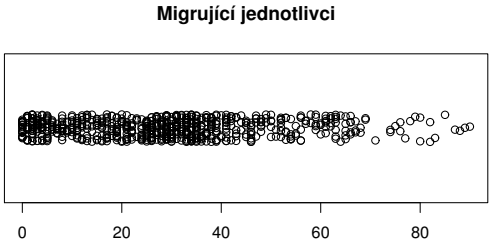
Tabulka 22: Charakteristiky variability pro všechny jednotlivce

	Rozptyl	Směr. odch.	IQR	MAD	V
POHL	0,2496	0,4996	1	0	0,3287
VZD	3,9299	1,9824	2	1,4826	0,6076
VEK	511,0749	22,6070	37	26,6868	0,5273
CPRIJMY	20830165819	144326,6	130003	92325,95	1,0160
EA	9,4236	3,0698	6	1,4826	0,8721

Tabulka 23: Charakteristiky variability pro nemigrující jednotlivce

5.8 Grafy pro soubor jednotlivci

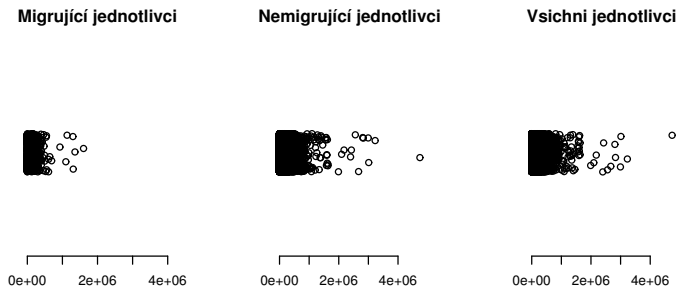
5.8.1 Diagram rozptýlení a rozmítnutý diagram rozptýlení



Obrázek 27: Rozmítnutý diagram rozptýlení migrujících domácností pro věk

Zde uvedeme grafy pro statistické znaky věk a příjmy jednotlivců. V obrázku 27 je uveden pouze diagram rozptýlení pro migrující domácnosti. Je to z toho důvodu, že ostatní diagramy pro proměnnou věk byly značně nepřehledné. V grafu opět vidíme, že se daleko méně stěhují starší osoby, což nám potvrzuje informaci i z číselných charakteristik.

Graf 28 nám dává informaci o rozložení příjmů. Ta nám ukazuje, že stěhování proběhlo u osob, které v šetření udávají příjmy do 500 000 Kč. Tedy jak bylo již řečeno, lidé s vysokými příjmy se nestěhují. Pokud ale srovnáme diagramy pro nemigrující jednotlivce a pro všechny jednotlivce, znatelný rozdíl nevidíme. To může být způsobeno velkým množstvím překrytých hodnot.

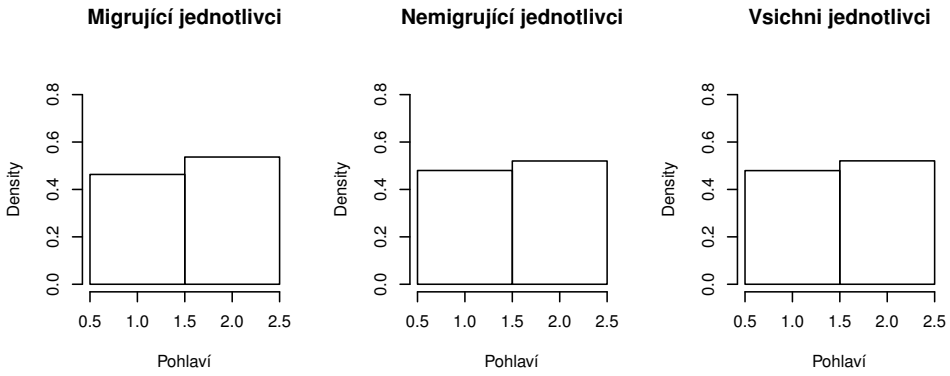


Obrázek 28: Rozmítnutý diagram rozptýlení pro příjmy

5.8.2 Histogram

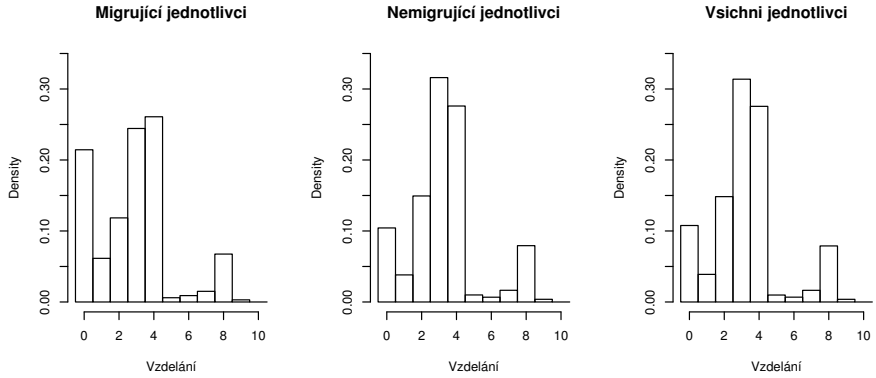
V této části uvedeme kromě histogramů pro věk a příjmy i grafy relativních četností pro kategoriální proměnné.

V jednotlivých grafech vidíme, že relativní četnosti nemigrujících osob se výrazně neliší od relativních četností souboru všech jednotlivců. Rozdíly jsou daleko patrnější u grafů pro migrující jednotlivce. Z obrázku 29 je zřejmé, že relativní četnosti pro pohlaví se liší velmi málo. To znamená, že struktura mužů a žen je ve všech třech souborech stejná a pohlaví nemá vliv na stěhování. Pokud se zaměříme na vzdělání (obrázek 30), vidíme, že větší zastoupení u migrujících osob má skupina 0 (předškolní děti, neukončený 1. stupeň ZŠ), oproti tomu menší zastoupení má skupina 3 (vyučení). Na obrázku 31 opět pozorujeme, že soubor přestěhovaných má menší zastoupení starších osob, protože relativní četnosti pro osoby ve starobním důchodu jsou zde menší než v souboru pro všechny jednotlivce. Oproti tomu stoupla relativní četnost ostatních ekonomicky neaktivních osob.

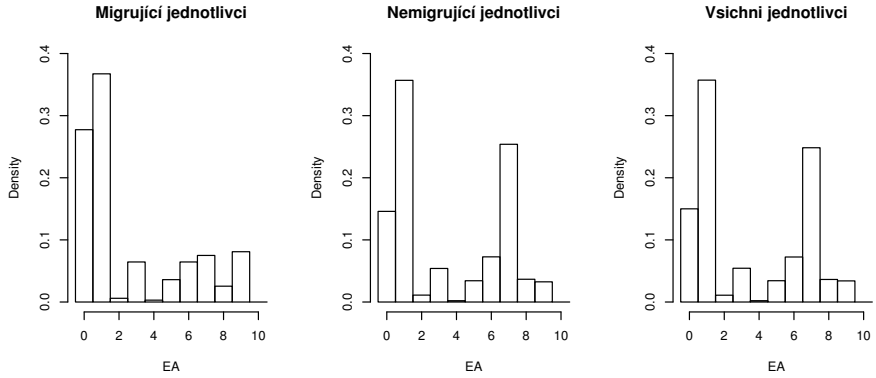


Obrázek 29: Relativní četnosti pro pohlaví

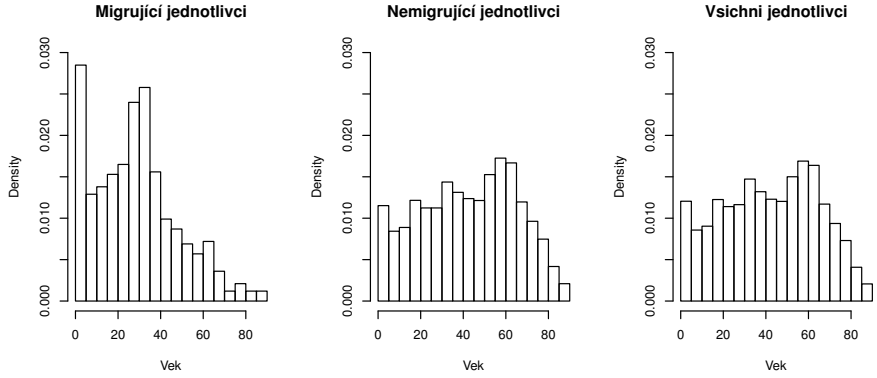
Histogram 32 nám opět potvrzuje, že se méně stěhují starší osoby. Výraznější navýšení relativních četností u migrujících mají osoby do 5 let, a poté osoby s věkem mezi 25-35 lety. V histogramu pro příjmy (obrázek 33) vidíme u migrujících, že se zvýšil podíl osob s příjmem do 50 000 Kč a oproti tomu se snížil podíl jednotlivců s příjmy mezi 100 000 - 150 000 Kč. Při srovnání všech osob a



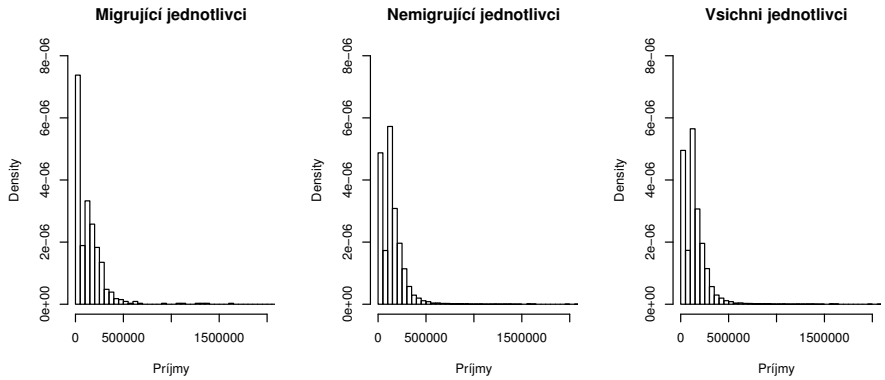
Obrázek 30: Relativní četnosti pro vzdělání



Obrázek 31: Relativní četnosti pro ekonomickou aktivitu



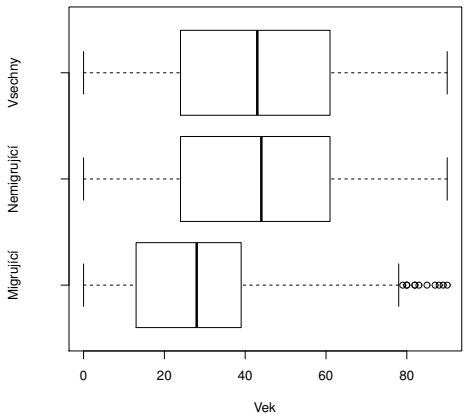
Obrázek 32: Histogram pro věk



Obrázek 33: Histogram pro příjmy

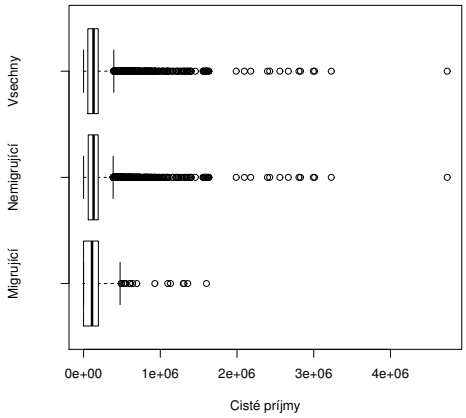
nemigrujících osob opět nevidíme rozdíl ani u jedné proměnné.

5.8.3 Boxplot



Obrázek 34: Boxplot pro věk

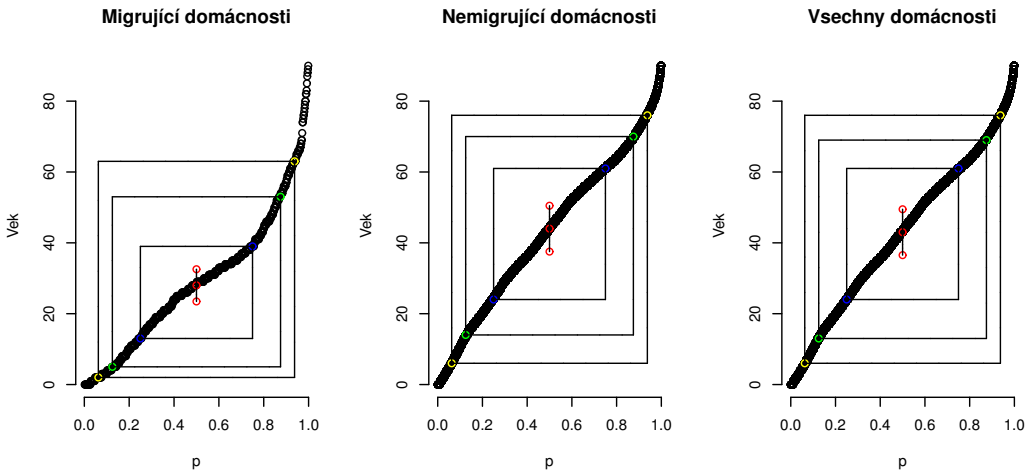
Grafy boxplotu uvedeme pro proměnnou věk a příjmy jednotlivců. Obrázek 34 opět potvrzuje, že se nestěhují starší lidé. Vidíme velký posun mediánu i celého krabicového grafu k menším hodnotám. Užší je i mezikvartilové rozpětí. Pokud se zaměříme na porovnání všech a nemigrujících osob, zaznamenáme mírný posun mediánu u druhého boxplotu k vyšším hodnotám. U boxplotu příjmů (obrázek



Obrázek 35: Boxplot pro příjmy

35) vidíme také vetší rozdíl u migrujících domácností, kde se nestěhují osoby s extrémně velkými příjmy. Dojde také k rozšíření mezikvartilového rozpětí. Při porovnání prvních dvou boxplotů na obrázku 35 rozdíl nevidíme.

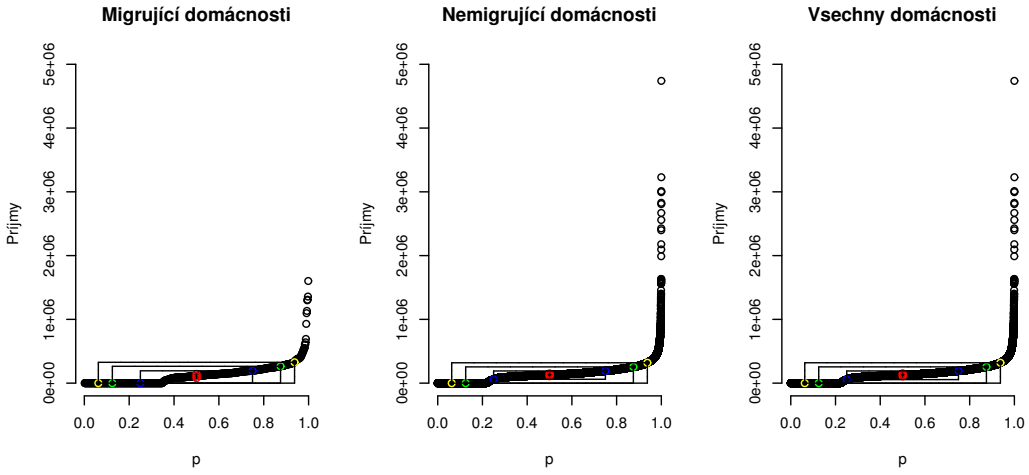
5.8.4 Graf rozptýlení s kvantily



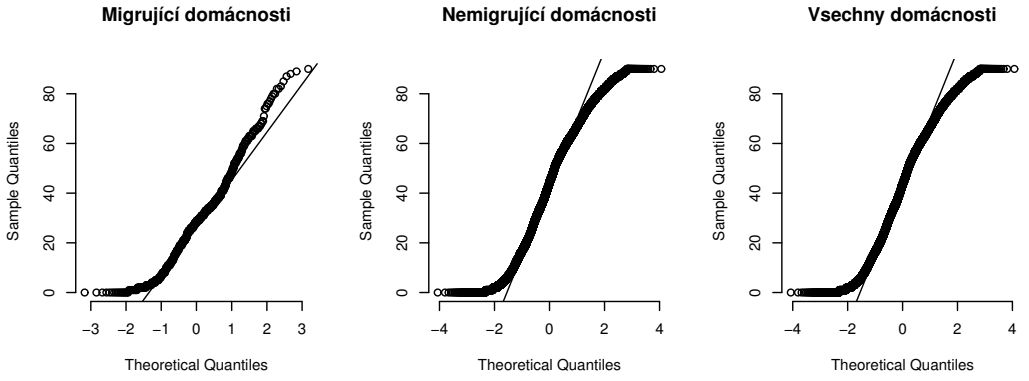
Obrázek 36: Graf rozptýlení s kvantily pro věk

V grafu rozptýlení s kvantily pro věk a příjmy vyplývají stejné závěry, jako

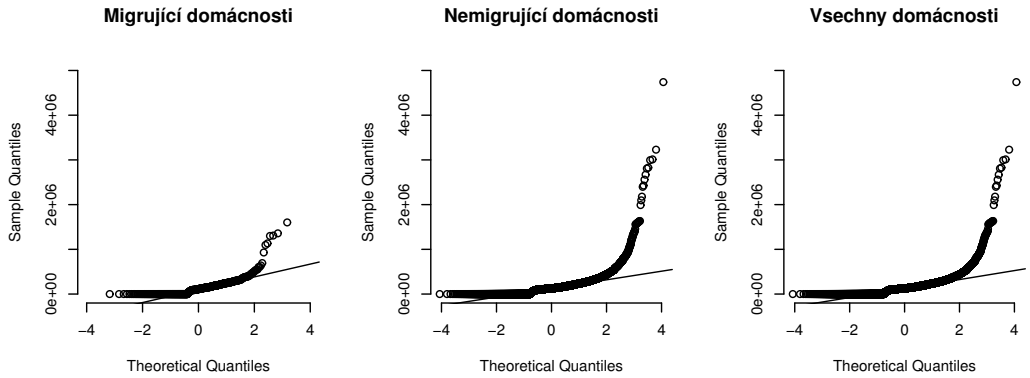
v předchozím textu. V grafu 36 opět vidíme stěhování převážně mladších respondentů. Na obrázku 37 vidíme sešikmení výsledné křivky k vyšším hodnotám, a že se nestěhují jednotlivci s velmi vysokými příjmy.



Obrázek 37: Graf rozptýlení s kvantily pro příjmy



Obrázek 38: Q-Q graf pro věk



Obrázek 39: Q-Q graf pro příjmy

5.8.5 Q-Q graf

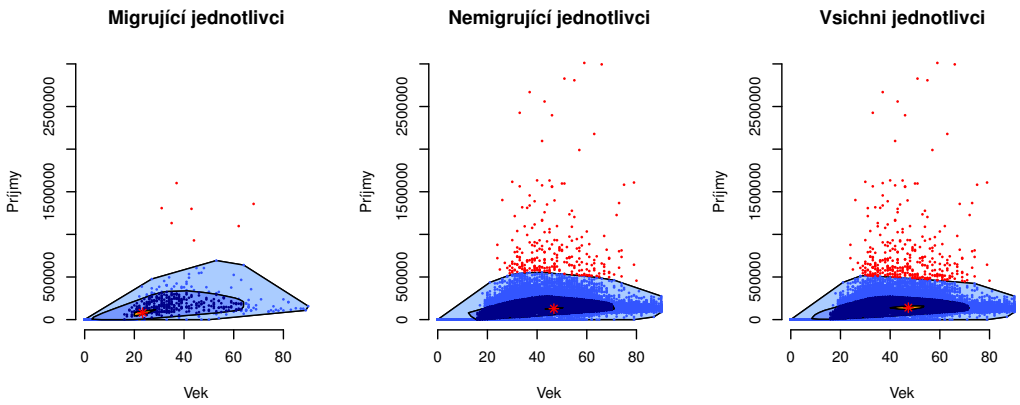
Esovitý tvar křivky pro proměnnou věk (graf 38) nám dává informaci, že rozdělení má jinou špičatost, než normální rozdělení. Toto je více patrné u souboru nemigrujících a všech jednotlivců, jejichž grafy nejsou viditelně rozdílné. Rozdíl u těchto dvou skupin nevidíme ani u grafu pro příjmy (obrázek 39).

5.8.6 Bagplot

Bagplot jsme tentokrát sestavili pro dvojici proměnných věk a příjmy (obrázek 40). Zde vidíme, že migrující jsou převážně mladší lidé, protože bag a medián je posunutý více doleva. Vynechání těchto osob ovšem v prostředním bagu nevyvolá větší změnu oproti bagu vpravo, protože těchto osob je malý počet.

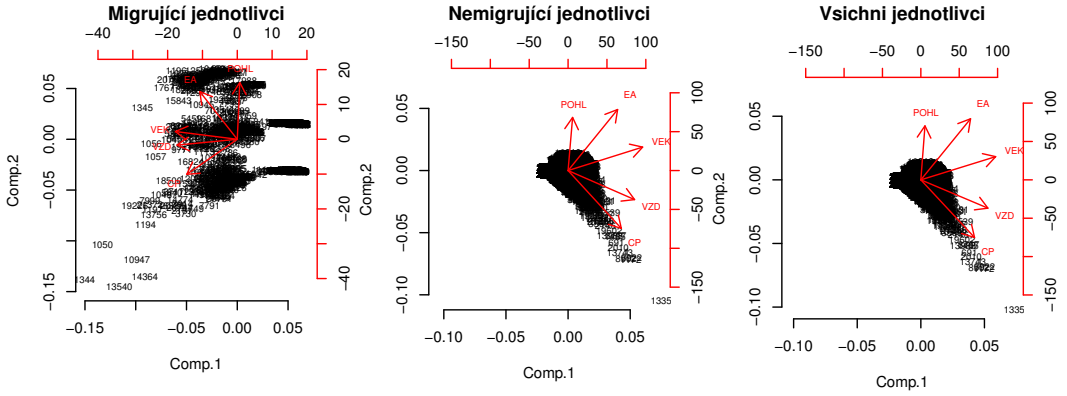
5.8.7 Biplot

Jelikož mezi statistickými znaky je opět proměnná, která svými hodnotami převažuje nad ostatními (v našem případě příjmy), přejdeme ke škálování. Po přeškálování dostaneme biploty v obrázku 41 a charakteristiky příslušných hlavních komponent v tabulkách 24, 25 a 26, kde první dvě hlavní komponenty u obou souborů pro všechny a nemigrující jednotlivce charakterizují přes 66%



Obrázek 40: Bagplot pro věk a příjmy

variability, u migrujících je to přes 70%. Z grafu vidíme, že prostřední a pravý biplot se od sebe neliší. V případě migrujících domácností pozorujeme, že u dvojice proměnných věk a vzdělání se zvýší jejich kladná korelace, protože příslušné šipky svírají menší úhel. Současně si můžeme všimnout výrazně odlišné struktury zobrazených objektů.



Obrázek 41: Biplot pro jednotlivce

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1,4854624	1,1479121	0,8475963	0,6414096	0,58170201
Proportion of Variance	0,4419823	0,2639362	0,1438997	0,0824048	0,06777706
Cumulative Proportion	0,4419823	0,7059185	0,8498181	0,9322229	1,00000000

Tabulka 24: Charakteristiky hlavních komponent pro migrující domácnosti

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1,3816437	1,1869340	0,9339554	0,7116677	0,55069368
Proportion of Variance	0,3818063	0,2817761	0,1744630	0,1012991	0,06065563
Cumulative Proportion	0,3818063	0,6635823	0,8380453	0,9393444	1,00000000

Tabulka 25: Charakteristiky hlavních komponent pro nemigrující domácnosti

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1,385810	1,1845181	0,9320667	0,7091190	0,55191902
Proportion of Variance	0,384112	0,2806298	0,1737578	0,1005747	0,06092577
Cumulative Proportion	0,384112	0,6647418	0,8384996	0,9390742	1,00000000

Tabulka 26: Charakteristiky hlavních komponent pro všechny domácnosti

5.9 Korelační matice

I v souboru pro jednotlivce srovnáme korelace proměnných u migrujících, nemigrujících a všech osob.

Korelační matice pro migrující jednotlivce:

	POHL	VZD	VEK	CPRIJMY	EA
POHL	1,0000	-0,0228	-0,0028	-0,2007	0,2308
VZD	-0,0228	1,0000	0,5652	0,5348	0,2893
VEK	-0,0028	0,5652	1,0000	0,4628	0,4651
CPRIJMY	-0,2007	0,5348	0,4628	1,0000	-0,0060
EA	0,2308	0,2893	0,4651	-0,0060	1,0000

Korelační matice pro nemigrující jednotlivce:

	POHL	VZD	VEK	CPRIJMY	EA
POHL	1,0000	-0,0123	0,0709	-0,1631	0,1680
VZD	-0,0123	1,0000	0,3612	0,4945	0,1463
VEK	0,0709	0,3612	1,0000	0,3015	0,5533
CPRIJMY	-0,1631	0,4945	0,3015	1,0000	-0,1262
EA	0,1680	0,1463	0,5533	-0,1262	1,0000

Korelační matice pro všechny jednotlivce:

	POHL	VZD	VEK	CPRIJMY	EA
POHL	1,0000	-0,0128	0,0679	-0,1645	0,1695
VZD	-0,0128	1,0000	0,3686	0,4961	0,1527
VEK	0,0679	0,3686	1,0000	0,3055	0,5524
CPRIJMY	-0,1645	0,4961	0,3055	1,0000	-0,1211
EA	0,1695	0,1527	0,5524	-0,1211	1,0000

Největší korelaci mají dvojice věk a ekonomická aktivita, vzdělání a příjmy. Naopak téměř nekorelované jsou dvojice pohlaví a věk, pohlaví a vzdělání. Při srovnání tabulek pro všechny a pro nemigrující jednotlivce vidíme opět jen drobné rozdíly v jednotlivých korelačních koeficientech. Největší rozdíl má korelační koeficient u dvojic věk a vzdělání, ekonomická aktivita a vzdělání. Jak jsme mohli pozorovat i u biplotu, korelační koeficient migrujících jednotlivců se výrazně zvýší u dvojice proměnných věk a vzdělání, i u ostatních dvojic proměnných můžeme pozorovat dílčí změny.

5.10 Testy dobré shody

Zde budeme opět porovnávat soubor se všemi jednotlivci se soubory pro nemigrující a migrující osoby. Kategoriální proměnné jsou pohlaví, vzdělání a ekonomická aktivita. Nejprve si vypočítáme teoretické pravděpodobnosti ze souboru pro všechny osoby pro kategoriální proměnnou pohlaví:

POHL	m_i	$p_i^0 = \frac{m_i}{m}$
1	10250	0,479
2	11129	0,521
m	21379	

Následně vypočítáme hodnotu testovací statistiky pro nemigrující osoby:

POHL	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
1	9941	0,479	9921,05	0,040
2	10771	0,521	10790,95	0,037
n	20712		Z	0,077

Kritický obor $W = \langle \chi_1^2(0{,}95), \infty) = \langle 3{,}84; \infty)$ neobsahuje hodnotu $Z = 0{,}077$, proto nulovou hypotézu nelze zamítnout.

POHL	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
1	309	0,479	309,48	0,007
2	358	0,521	347,51	0,317
n	667		Z	0,324

Kritický obor zůstává stejný, $W = \langle 3{,}84; \infty)$, proto nulovou hypotézu nemůžeme ani v případě migrujících jednotlivců zamítnout.

Dále vypočítáme teoretické pravděpodobnosti pro vzdělání:

VZD	m_i	$p_i^0 = \frac{m_i}{m}$
0	20303	0,1077
1	831	0,0389
2	3172	0,1484
3	6709	0,3138
4	5891	0,2756
5	209	0,0098
6	145	0,0068
7	351	0,0164
8	1688	0,0789
9	80	0,0037
n	39379	

Tyto pravděpodobnosti použijeme opět pro výpočet testovací statistiky pro nemigrující osoby:

VZD	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
0	2160	0,1077	2230,6824	2,2397
1	790	0,0389	805,6968	0,3058
2	3093	0,1484	3073,6608	0,1217
3	6546	0,3138	6499,4256	0,3337
4	5717	0,2756	5708,2272	0,0135
5	205	0,0098	202,9776	0,0202
6	139	0,0068	140,8416	0,0241
7	341	0,0164	339,6768	0,0052
8	1643	0,0789	1634,1768	0,0476
9	78	0,0037	76,6344	0,0243
n	20712		Z	3,115

Kritický obor je $W = \langle \chi^2_9(0{,}95), \infty) = \langle 16{,}92; \infty)$ a nulou hypotézu nelze zamítnout, protože hodnota testovací statistiky nenáleží do tohoto intervalu. Dále otestujeme migrující domácnosti:

VZD	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
0	143	0,1077	71,8359	70,4986
1	41	0,0389	25,9463	8,7340
2	79	0,1484	98,9828	4,0342
3	163	0,3138	209,3046	10,2440
4	174	0,2756	183,8252	0,5251
5	4	0,0098	6,5366	0,9844
6	6	0,0068	4,5356	0,4728
7	10	0,0164	10,9388	0,0806
8	45	0,0789	52,6263	1,1052
9	2	0,0037	2,4679	0,0887
n	667		Z	96,7676

Hodnota $Z = 96{,}7676$ leží uvnitř oboru $W = \langle 16{,}92; \infty)$, a proto nulovou hypotézu zamítáme. Tedy struktura vzdělání pro migrující osoby není stejná, jako struktura pro všechny osoby.

Nyní určíme pravděpodobnosti poslední kategoriální proměnné, ekonomické aktivity:

EA	m_i	$p_i^0 = \frac{m_i}{m}$
0	3207	0,1500
1	7636	0,3572
2	234	0,0109
3	1162	0,0543
4	46	0,0022
5	733	0,0343
6	1549	0,0725
7	5309	0,2483
8	776	0,0363
9	727	0,0340
n	21379	

Vypočteme testovací kritérium pro nemigrující jednotlivce:

EA	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
0	3022	0,1500	3106,8	2,3146
1	7391	0,3572	7398,3264	0,0073
2	230	0,0109	225,7608	0,0796
3	1119	0,0543	1124,6616	0,0285
4	44	0,0022	45,5664	0,0538
5	709	0,0343	710,4216	0,0028
6	1506	0,0725	1501,62	0,0128
7	5259	0,2483	5142,7896	2,6260
8	759	0,0363	751,8456	0,0681
9	673	0,0340	704,208	1,3830
n	20712		Z	6,5765

Jelikož máme opět stejný počet tříd, bude kritický obor totožný s předešlým. Hodnota $Z = 6{,}5765$ nepatří do intervalu $W = \langle 16{,}92; \infty)$, tedy nulovou hypotézu nelze zamítnout.

Jako poslední vypočteme testovací statistiku pro migrující jednotlivce:

EA	n_i	p_i^0	np_i^0	$\frac{(n_i - np_i^0)^2}{np_i^0}$
0	185	0,1500	100,05	72,1290
1	245	0,3572	238,2524	0,1911
2	4	0,0109	7,2703	1,4710
3	43	0,0543	36,2181	1,2699
4	2	0,0022	1,4674	0,1933
5	24	0,0343	22,8781	0,0550
6	43	0,0725	48,3575	0,5936
7	50	0,2483	165,6161	80,7113
8	17	0,0363	24,2121	2,1483
9	54	0,034	22,678	43,2608
n	667		Z	202,0256

Ani struktura vzdělání migrujících jednotlivců není totožná se strukturou u všech jednotlivců, protože hodnota testovací statistiky patří do kritického oboru $W = \langle 16{,}92; \infty)$ a nulovou hypotézu na hladině testu $\alpha = 0,05$ zamítáme.

Závěr

Čtenář se v této práci seznámil s různými metodami analýzy jak jednorozměrných, tak i mnohorozměrných dat. Následně byly tyto metody použity na reálný problém, analýzu struktury migrujících domácností v šetření Životní podmínky.

Psaní této práce bylo pro mne přínosné, protože jsem se dozvěděla zajímavé informace o práci Českého statistického úřadu a konkrétně o zmíněném šetření Životní podmínky.

V uvedené práci jsem se přesvědčila, že vypuštění hodnot migrujících subjektů výrazně nezměnilo výsledky šetření i přesto, že struktura migrujících domácností a jednotlivců není totožná se strukturou původního souboru pro všechny subjekty. Je to z toho důvodu, že procento migrace je u obou skupin malé. Problém by mohl nastat, pokud by se v České republice zvýšila migrace obyvatelstva při zachování relativní struktury migrujících domácností a jednotlivců.

Z výsledků statistické analýzy vyplývá, že se procentuálně nejvíce stěhovaly domácnosti se 3-4 členy, s 1-2 ekonomicky aktivním osobami a s velkou nákladovou zátěží. U jednotlivců jsme zaznamenali zvýšení relativního počtu přestěhovaných u osob v domácnosti, osob mezi 25-35 lety, dětí předškolního věku a s neukončeným 1. stupněm ZŠ. Naopak nejméně se stěhovaly osoby v důchodu. Z grafů bylo také zřejmé, že se nestěhují osoby s velmi vysokými příjmy a náklady na bydlení. Přestože může být podrobné studium všech uvedených výsledků časově náročné, věřím, že může pomoci odhalit případné další zajímavé informace týkající se uvedené problematiky.

Otázkou pak zůstává, zda se v následujících letech zvýší migrace obyvatelstva natolik, aby vynechání těchto subjektů následně ovlivnilo výsledky šetření. To by mohlo být předmětem jiné analýzy, kdy by se zkoumala tendence růstu migrace obyvatelstva na základě dosavadního vývoje v České republice s přihlédnutím na vývoj migrace v podobných zemích Evropy. Doufám, že k ní bude má práce inspirací.

Seznam obrázků

Seznam tabulek

Literatura

[1] Body fat data [online], dostupné z: http://lib.stat.cmu.edu/datasets/bodyfat, [citováno 7. 3. 2013].

[2] Filzmoser, P., Multivariate Statistik. Wien: TU, 2007.

[3] Internetové stránky českého statistického úřadu [online], dostupné z: http://www.czso.cz, [citováno 4. 10. 2012].

[4] Hebák, P. a kol., Vícerozměrné statistické metody (1), 1. vydání. Praha: Informatorium, 2004.

[5] Hebák, P. a kol., Vícerozměrné statistické metody (3), 1. vydání. Praha: Informatorium, 2005.

[6] Hendl, J., Přehled statistických metod, 3. vydání. Praha: Portál, 2009.

[7] Hron, K., Kunderová, P., Základy počtu pravděpodobnosti a metod matematické statistiky. Olomouc: VUP, 2013.

[8] Kalivodová, A., Biplot a jeho aplikace, UPOL, 2010 (Bakalářská práce) [online], dostupné z: http://theses.cz/id/warxzj/72260-112175808.pdf, [citováno 7. 2. 2013].

[9] Kubánová, J., Statistické metody pro ekonomickou a technickou praxi. Bratislava: Statis, 2004.

[10] Kunderová, P., Úvod do teorie pravděpodobnosti a matematické statistiky, 2. vydání. Olomouc: VUP, 2004.

[11] Meloun, M., Militký, J., Kompendium statistického zpracování dat, 1. vydání. Praha: Academia, 2002.

[12] Meloun, M., Militký, J., Nejlepší odhady polohy a rozptýlení chemických dat [online], dostupné z: http://meloun.upce.cz/docs/publication/136.pdf, [citováno 25. 10. 2012].

[13] Meloun, M.,Průzkumová analýza jednorozměrných dat [online], dostupné z: http://meloun.upce.cz/docs/books/2.pdf, [citováno 5. 10. 2012].

[14] Meloun, M., Militký, J., Statistická analýza experimentálních dat, 2. vydání. Praha: Academia, 2004.

[15] Nápověda k softwaru R [online], dostupné z: http://www.r-project.org/, [citováno 1. 2. 2013].

[16] Otyepka, M., Banáš, P., Otyepková, E., Základy zpracování dat [online], dostupné z: http://fch.upol.cz/skripta/zzd/chemo/chemo.pdf, [citováno 11. 10. 2012].

[17] Pavlík, J. a kol., Aplikovaná statistika. Praha: VŠCHT, 2005.

[18] Pokorný, M., Matematické metody vyhodnocování experimentu. Olomouc: Moravská vysoká škola Olomouc, o. p. s., 2010 [online], dostupné z: http://www.mvso.cz/Files/WEB/APSYS/31Matematicke_metody_vyhodnocovani_experimentu.pdf,

[citováno 11. 10. 2012].

[19] Průzkumová analýza dat [online], dostupné z: http://user.mendelu.cz/drapela/Statisticke_metody/teorie%20text%20II.pdf, [citováno 5. 10. 2012].

[20] Rousseeuw P. J., Ruts I., Tukey J. W., The agplot: A bivariate boxplot, The American Statistician 53/4, 382-387, (1999) [online], dostupné z: http://venus.unive.it/romanaz/ada2/bagplot.pdf, [citováno 15. 2. 2013].

[21] Typy proměnných [online], dostupné z:

http://iastat.vse.cz/typy_promennych.html, [citováno 9. 1. 2013].