Štefan Lyócsa – Eduard Baumöhl – Tomáš Výrost

Kvantitatívne metódy v ekonómii II.

ELFA 2013

Štefan Lyócsa
Eduard Baumöhl
Tomáš Výrost

Kvantitatívne metódy v ekonómii II.

Košice, 2013

Recenzenti:

Dr. h. c. prof. RNDr. Michal Tkáč, CSc.

Katedra hospodárskej informatiky a matematiky, Podnikovohospodárska fakulta so sídlom v Košiciach, Ekonomická univerzita v Bratislave

Ing. Silvia Megyesiová, PhD.

Katedra hospodárskej informatiky a matematiky, Podnikovohospodárska fakulta so sídlom v Košiciach, Ekonomická univerzita v Bratislave

Mgr. Svatopluk Svoboda

Accenture Services s.r.o, Praha, Česká Republika

Publikácia neprešla jazykovou korektúrou. Za odbornú stránku a jazykovú úpravu textu zodpovedajú autori.

Umiestnenie: http://www.econometrics.sk

Dostupné od: 07 / 2013

Vydanie prvé

Rozsah: 22.4 AH

ISBN 978-80-8086-210-7

O B S A H

Úvod

Predkladané skriptá nadväzujú na našu publikáciu Kvantitatívne metódy v ekonómii I. (KMvE II.) a rovnomenný predmet vyučovaný na Podnikovohospodárskej fakulte so sídlom v Košiciach, Ekonomickej univerzity v Bratislave (EUBA-PHF). K napísaniu tejto publikácie sme boli motivovaní potrebou vytvoriť pomocný učebný text, teda skriptá, pre študentov, ktorí absolvovali predmety KMvE I., KMvE II., a Finančná ekonometria na EUBA-PHF v rokoch 2009 až 2012. Na týchto predmetoch je našim cieľom naučiť študentov aplikovať kvantitatívne nástroje na analýzu a riešenie ekonomických problémov. Predkladané skriptá KMvE II., rovnako ako aj skriptá KMvE I., slúžili len ako doplňujúci materiál pre študentov, ktorý absolvovali rovnomenné predmety. Druhým, nemenej dôležitým motívom k napísaniu tejto publikácie bola skutočnosť, že naši diplomanti a bakalári potrebovali určitý manuál pre prácu v programe R. Z týchto dôvodov sme sa pri písaní orientovali najmä na aplikácie a prácu v tomto programe. Cieľom bolo ponúknuť čitateľovi spracovanie niektorých základných kvantitatívnych štatistických metód induktívnej štatistiky tak, aby ich bol čitateľ schopný v konkrétnom príklade prakticky použiť.

Séria publikácií Kvantitatívne metódy v ekonómii sa nevenuje matematickej štatistike. Aj keď bolo našou snahou ponúknuť čo najpresnejší výklad, intuitívny náhľad do princípov induktívnej štatistiky a použitých metód bol pri písaní pre nás prioritou. Týmto spôsobom nutne došlo k určitému kompromisu medzi jednoduchosťou a presnosťou písania, kde sme sa neraz rozhodli prikloniť práve k jednoduchosti. Ak má čitateľ záujem o exaktný opis matematických ideí v pozadí štatistických metód, musí siahnuť po inej publikácii.

Táto publikácia taktiež nie je o programovaní. V texte je napísané pomerne veľké množstvo kódov z programu R, ale programovanie tvorí podstatne viac ako len súbor jednoduchých príkazov, cyklov a zopár podmienok. Našim cieľom bolo napísať tie príkazy čo najjednoduchšie (nie čo najúspornejšie), aby aj študent, ktorý si knihu otvorí na posledných stranách knihy, vedel zreprodukovať riešené príklady a obrázky. Preto čitateľovi, ktorý chce vedieť o programovaní v R viac, odporúčame znova inú ako túto publikáciu.

Predkladaná kniha je rozdelená do siedmych kapitol a príkladov. V prvej kapitole sa venujeme vymedzeniu induktívnej štatistiky. Našou snahou bolo prezentovať taký pohľad na induktívnu štatistiku, aby čitateľ vedel rozpoznať, kedy je vhodné jej použitie. V druhej, pomerne stručnej kapitole, sme sa rozhodli venovať metódam zberu údajov, medzi ktorými sa

pri prieskumoch a výskumoch najčastejšie rozhodujeme. Tretia kapitola má za cieľ čitateľovi priblížiť problematiku odhadov. Tradične sa táto téma začína vysvetľovať na bodových odhadoch a potom sa prechádza na intervalové odhady čoho sme sa držali aj v tejto publikácii. Testovanie štatistických hypotéz je zrejme základným rozhodovacím nástrojom, ktorý štatistika pre ekonómov ponúka. Ide o rozsahovo najobšírnejšiu kapitolu, ktorú sme rozdelili na: (i) základné testy o parametroch populácie (stredné hodnoty, podiely, rozptyly), (ii) testy dobrej zhody (najmä overovanie normality), (iii) testy overujúce prítomnosť extrémnych hodnôt, (iv) niektoré ďalšie (spravidla neparametrické) testy nielen o základných parametroch populácie (napr. test náhodnosti a nezávislosti), (v) princípy metódy ANOVA. Piata kapitola mohla byť súčasťou aj predošlej kapitoly. Rozhodli sme sa ju však oddeliť, keďže meranie a testovanie významnosti závislosti medzi dvoma premennými predstavuje určitý prechodový mostík medzi predkladanou publikáciou a ďalšou plánovanou publikáciou venujúcou sa ekonometrii. Šiesta kapitola predstavuje úvod do analýzy viacrozmerných dát.

1 Induktívna analýza empirických údajov

Pri deskriptívnej (opisnej) štatistike sú predmetom nášho záujmu vlastnosti štatistického súboru, ktorý sa snažíme opísať pomocou niekoľkých jednoduchých charakteristík. Spravidla nás zaujíma charakteristika polohy štatistického súboru (aritmetický priemer, medián, prípadne modus) a variabilita hodnôt v štatistickom súbore (najčastejšie rozptyl a smerodajná odchýlka). K opisnej štatistike môžeme priradiť aj vizualizáciu dát. Prostredníctvom vhodne zvoleného grafu vieme často povedať viac, ako počítaním komplikovaných vzorcov. Pri deskriptívnej štatistike však platí, že predmetom nášho záujmu sú len tie štatistické jednotky, ktoré máme v štatistickom súbore. Naproti tomu pri induktívnej štatistike je našim zámerom na základe údajov zo štatistického súboru (reprezentatívnej vzorky) závery zovšeobecniť na tzv. populáciu.

Kým deskriptívna štatistika sa používala v rôznych podobách už niekoľko tisícročí, základy induktívnej štatistiky tak, ako ju poznáme a používame dnes, vznikali začiatkom minulého storočia. V roku 1908 W. S. Gosset publikoval článok „*The probable error of the mean*“ v časopise Biometrika, pod pseudonymom „STUDENT“. W. S. Gosset pracoval ako chemik v známom pivovare Arthur Guiness Son’s and Co., kde zamestnancom nebolo umožnené verejne publikovať výsledky svojej práce. Gosset potreboval štatistické metódy, pomocou ktorých by na základe málo početných vzoriek vedel urobiť racionálne rozhodnutia o celej populácii. Výsledkom jeho snaženia bolo tzv. *t*-rozdelenie pravdepodobnosti, z ktorého sa odvodil dnes už známy Studentov *t*-test (Lehman, 1999). Známy štatistik sir R. A. Fisher rozpoznal potenciál a význam *t*-testu a v rámci štatistiky výrazne prispel k rozvoju disciplíny, ktorej v súčasnosti hovoríme induktívna štatistika.

Aplikácia štatistických metód môže pomôcť ekonómom, marketérom a manažérom pri rozhodovaní sa. Pridaná hodnota využívania štatistických metód pritom spravidla rastie so závažnosťou prijímaného rozhodnutia. Marketér môže použiť štatistiku na opis kupujúcich a potenciálnych zákazníkov, alebo pri hľadaní faktorov, ktoré z potenciálnych zákazníkov robia kupujúcich. Manažérovi ľudských zdrojov pomôže zistiť, aké faktory súvisia so spokojnosťou zamestnancov, napr. ktoré faktory sú tzv. hygienické, a ktoré predstavujú motivátory (v rámci Herzbergovej dvojfaktorovej teórie; Herzberg, 1964). Prípadne mu štatistická analýza môže pomôcť zistiť, aký typ zamestnanca má väčšiu tendenciu k zmene pracovného miesta a tým prispieť k zníženiu fluktuácie zamestnancov. Ekonómovi môže štatistika pomôcť napríklad pri identifikácii informácií, ktoré môžu vplývať na vývoj kurzu

alebo volatility akcií a mien, či v konkrétnych podmienkach tej ktorej krajiny existuje vzťah medzi nezamestnanosťou a mierou inflácie; a samozrejme pri množstve iných situácií.

V horšom prípade sa štatistika neraz stotožňuje s evidenciou a v lepšom sa chápe pod štatistikou iba deskriptívna analýza. Ozajstným prínosom štatistiky pre ekonómov je schopnosť pomôcť charakterizovať javy, ktoré z objektívnych príčin nemôžeme merať vždy a v každej situácii a sme tak odkázaní iba na určitý obmedzený počet pozorovaní, t. j. na určitú vzorku. Toto je realita, s ktorou sa v hospodárskej praxi môžeme stretnúť najčastejšie.

Pokiaľ množina štatistických jednotiek, ktoré sú predmetom zovšeobecnenia štatistického výskumu, má spoločný jeden alebo viac štatistických znakov, nazývame túto množinu **populáciou**. Ľubovoľnú vlastnú podmnožinu populácie nazývame **vzorka** danej populácie. Pomocou metód induktívnej štatistiky využívame údaje zo vzorky k tvorbe záverov a k prijatiu racionálnych rozhodnutí týkajúcich sa populácie. Populáciu vieme spravidla vymedziť z troch hľadísk: predmet (čo sa analyzuje), miesto (kde sa štatistické jednotky nachádzajú) a čas (kedy sa údaje získali, prípadne na aké obdobie sa robí zovšeobecnenie). Situácie, keď nemôžeme získať údaje od všetkých štatistických jednotiek vznikajú, ak:

- populácia má nekonečný rozsah,
- meranie štatistického znaku môže viesť k deštrukcii alebo poškodeniu štatistickej jednotky,
- získanie hodnôt štatistického znaku je veľmi nákladné,
- máme k dispozícii iba stredné hodnoty jedného (rovnakého) procesu generujúceho pozorovania v časových radoch.

Populácia teplôt (predmet) na rôznych miestach bazéna (miesta) v určitom čase (čas) je príkladom populácie s nekonečným rozsahom populácie, keďže v priestore bazéna môžeme vykonať meranie teploty prakticky na ľubovoľnom mieste. Naproti tomu počet študentov na ekonomickej univerzite je príkladom populácie s konečným počtom. Neraz je predmetom štatistického znaku určitá vlastnosť, ktorá sa prejavuje zničením alebo deštrukciou štatistického znaku. Príkladom môžu byť rôzne akceleračné testy produktov s ohľadom na ich životnosť, bezpečnosť, spoľahlivosť a pod. Konkrétnym príkladom môže byť stanovenie chute a akosti jablka. Aby sme posúdili chuť jabĺk, nemusíme ochutnať všetky jablká stromu. Racionálnou alternatívou je ochutnať iba vybrané jablká (vo vybranom počte) a na základe nich rozhodnúť o akosti všetkých jabĺk stromu. Získavaním údajov týmto „výberovým“ spôsobom sa môžeme častejšie stretnúť v praxi, ktorý sa prejavuje získavaním údajov

dopytovaním. Ak máme záujem získať informácie o zákazníkoch, konkurentoch, produktoch či zamestnancoch, je spravidla neefektívne (väčšinou aj nerealizovateľné) dopytovať sa každej štatistickej jednotky. Existujú tu teda určité protichodné snahy. Na jednej strane je našim záujmom dozvedieť sa niečo o populácii, na druhej strane môžeme k tomu využiť iba údaje získané na základe určitého výberu z populácie, obmedzeného súboru z tzv. vzorky, resp. výberovej vzorky.

Kľúčová otázka pre pochopenie významu induktívnej štatistiky by sa dala formulovať tak, či závery vypočítané z údajov zo vzorky a použitím deskriptívnej štatistiky môžeme zovšeobecniť a tvrdiť, že platia aj pre populáciu. Odpoveď je, že nie vždy. Práve snaha zovšeobecniť závery získané zo vzorky je **predmetom induktívnej štatistiky**[1] a často je našou snahou pri využívaní štatistických metód v praxi. Uveďme si len niekoľko príkladov, kedy je potrebné „*siahnuť*" po induktívnej štatistike:

- Sú zákazníci spokojnejší s novými službami?
- Aké sú preferencie politických strán?
- Je nový liek účinný?
- Chce si objednať náš produkt viacej mužov alebo žien?
- Majú zákazníci radšej farbu žltú alebo čiernu?
- Sú naši zamestnanci oproti predošlému roku spokojnejší?
- Znížila sa doba obratu pohľadávok oproti poslednému mesiacu, alebo sú rozdiely náhodné?

Príklad 1.1

Prečo nie je vždy možné zovšeobecniť závery zo vzorky použitím deskriptívnej štatistiky na celú populáciu? Podnikateľ si plánuje založiť predajňu v meste v novom regióne. V roku 2009 v meste, priľahlých dedinách a osadách, býva spolu 21 000 pracujúcich, dospelých obyvateľov, ktorí predstavujú jeho potenciálnych zákazníkov. Základné demografické údaje k určitému dátumu vie získať od štatistického úradu a sú verejne dostupné. Má k dispozícii niektoré štatistické znaky všetkých štatistických jednotiek. V jeho prípade teda ide o populáciu. Nanešťastie, kľúčovou informáciou, pomocou ktorej bude vedieť odhadnúť kúpny potenciál trhu, je objem peňazí, ktoré minie jeden dospelý

[1] Nie vždy sa induktívna štatistika používa v tomto význame. Pomerne často (stretávame sa s tým najmä v psychológii) sa vykonáva induktívna štatistika na skupine ľudí, ale výsledky sa nezovšeobecňujú na širšiu skupinu (vzorka nie je v žiadnej etape vybratá náhodne), ale iba na tú istú skupinu. Merania na tejto skupine ľudí sa totižto považujú za náhodné. Iným, podobným prípadom, je analýza údajov na úrovni krajín. Odlišným prípadom (od toho nášho) je analýza časových radov. Bližšie sa tejto problematike nebudeme venovať a budeme sa držať nami prezentovaného pohľadu na induktívnu štatistiku.

človek na telekomunikačné služby v regióne v priebehu jedného roka. Aby tieto údaje získal, uskutoční štatistický prieskum, pričom pre jednoduchosť predpokladajme, že vyberie povedzme iba 6 obyvateľov, teda $n = 6$. Získané výsledky v EUR sú:

520, 600, 400, 240, 300, 280

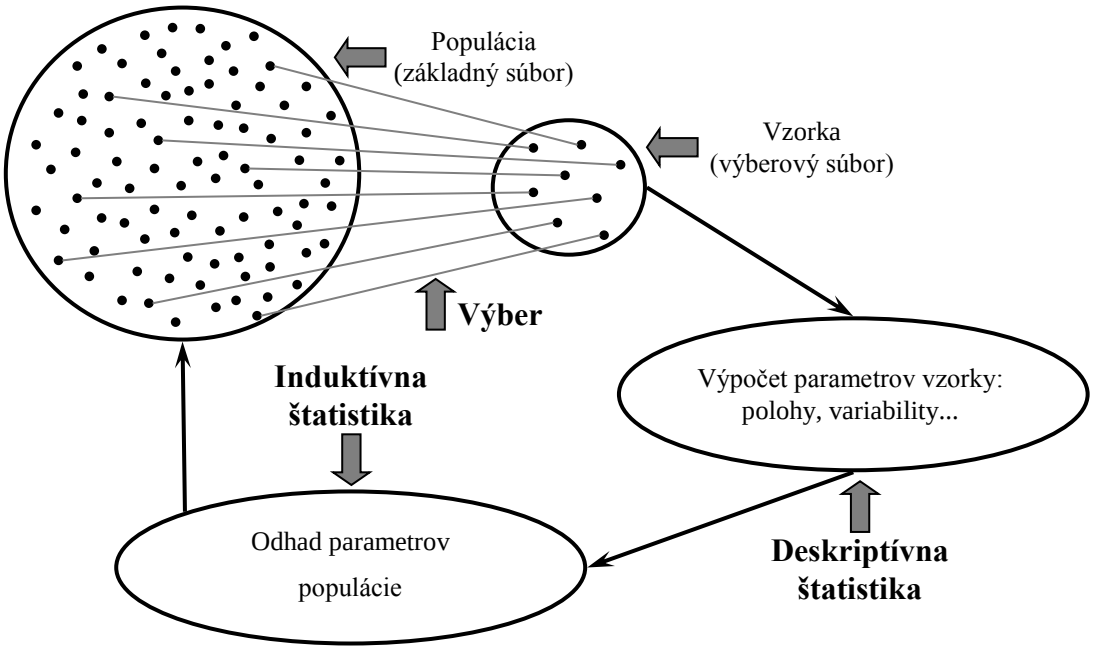
Koľko zvyknú obyvatelia regiónu minúť na telekomunikačné služby? To, čo nás zaujíma je stredná hodnota minutých peňazí na telekomunikačné služby v celej populácii. Strednú hodnotu by sme mohli **odhadnúť** na základe vzorky. Vypočítaním aritmetického priemeru dôjdeme k hodnote 390. Je aj stredná hodnota populácie 390,- EUR? V skutočnosti je to veľmi málo pravdepodobné[2]. Ak by podnikateľ do vzorky zobral ďalšieho respondenta, musel by minúť na telekomunikačné služby presne 390,- EUR, aby náš jednočíselný odhad strednej hodnoty ostal 390,- EUR. To isté platí pre každého ďalšieho respondenta (do vzorky by sme postupne pridávali po jednom respondentovi, nie naraz viac respondentov). Z toho sa zdá byť zrejmé, že je viac pravdepodobné, že stredná hodnota výdavkov na telekomunikačné služby v regióne bude v skutočnosti odlišná.

Všimnime si, že v uvedenom príklade je hodnota aritmetického priemeru závislá na tom, koho sme zobrali do vzorky. Ďalej si všimnime, že úplne istí o výške priemerných výdavkov v celej populácii by sme si boli iba vtedy, ak by sme mali k dispozícii údaje o všetkých štatistických jednotkách populácie. V opačnom prípade sme si istí iba s určitou „*konfidenciou*“ (určitou istotou), resp. s príslušnou pravdepodobnosťou. Môžeme vidieť, že aj v tejto na prvý pohľad jednoduchej úlohe sa otvorilo hneď niekoľko problémov. Ako vybrať tých „*správnych*“ respondentov, aby vzorka nebola skreslená? Nakoľko si môžeme byť istí našim odhadom strednej hodnoty? Odpovede na tieto otázky sú predmetom metód výberu štatistických jednotiek do vzorky, bodových a intervalových odhadov, ktorým sa budeme stručne venovať v ďalších kapitolách.

Štatistické metódy a postupy, pomocou ktorých môžeme zovšeobecniť výsledky z údajov získaných prostredníctvom výberovej vzorky, sú hlavnou témou nasledujúcich kapitol. V predošlom príklade bolo našim cieľom okrem iného upriamiť pozornosť čitateľa na dôležitosť výberu vzorky (pozri Obrázok 1.1), ako aj na využívanie konceptu pravdepodobnosti v induktívnej štatistike. Predmet tohto textu je podstatne širší ako štatistický výber vzoriek, a preto sa prvej z uvedených problematík budeme venovať len okrajovo.

---

[2] Za predpokladu, že ide o odhad spojitého parametra, je pravdepodobnosť neúspešného odhadu parametra takmer isto rovná 1.



Obrázok 1.1: Princíp induktívnej štatistiky

Zdroj: vlastné spracovanie

2 Metódy výberu vzoriek (sampling)

Spôsob, akým vyberáme štatistické jednotky z populácie do výberového súboru, budeme nazývať **sampling** (aj keď v slovenskej literatúre sa môžeme stretnúť aj s pojmom „*vzorkovanie*“, rozhodli sme sa používať anglický pojem). Dôvod, prečo existuje pomerne rozsiahla teória o samplingu, si môžeme vysvetliť existenciou dvoch, niekedy antagonických cieľov: tým prvým je snaha získať reprezentatívnu vzorku a tým druhým je snaha získať tzv. náhodnú vzorku. Niekedy pod náhodnou vzorkou budeme rozumieť vzorku údajov s nezávislými náhodnými premennými.[3] Nezávislosť pritom môžeme chápať tak, že napríklad výber jednej štatistickej jednotky do výberovej vzorky neovplyvní výber ostatných. Nech jav A zodpovedá tomu, že do vzorky bude vybraná jedna štatistická jednotka a $P(A)$ je pravdepodobnosťou nastania tohto javu. Podobne nech jav B zodpovedá tomu, že do vzorky bude vybraná druhá štatistická jednotka a $P(B)$ je príslušná pravdepodobnosť. Potom ak platí: $P(A \cap B) = P(A)P(B)$, hovoríme, že výber týchto dvoch štatistických jednotiek je nezávislý.

V praxi to vieme uskutočniť len zriedkakedy, najmä ak ide o analýzy vykonávané v spoločenských vedách, k akým radíme aj ekonómiu. Ak vyberieme z populácie jedného zákazníka A, zmeriame jeho štatistické znaky a zákazníka do populácie „*nevrátime*“, tak výber ďalšieho zákazníka B je do určitej miery podmienený tým, že sme pred tým vybrali už jedného zákazníka. Ak sa počet zákazníkov v populácii medzi tým nezmení, pravdepodobnosť výberu zákazníka B bude väčšia po výbere zákazníka A ako pred výberom zákazníka A. V tomto prípade striktne povedané nejde o nezávislý výber. Na tomto mieste je dôležité pripomenúť, že väčšina štatisticko-matematického aparátu induktívnej štatistiky sa opiera práve o *iid* predpoklady. Na druhej strane treba pripomenúť, že dodržať podmienky *iid* je v praxi väčšinou nemožné a „miera nedodržania“ *iid* predpokladov je do istej miery subjektívna a jej dopady ťažko predpovedateľné. V popísanom prípade v tomto odseku väčšinou platí, že ak je veľkosť populácie a vzorky pomerne veľká, dopady takéhoto výberu (bez vrátenia už vybraného respondenta späť do vzorky), ovplyvňujú výsledky štatistickej analýzy v minimálnej miere.

Z kvalitatívneho hľadiska je nemenej dôležité získať reprezentatívnu vzorku. Reprezentatívnou vzorkou rozumieme takú vzorku, ktorá odráža dôležité charakteristiky populácie, z ktorej sme túto vzorku získali. Problém spočíva v tom, že ani najlepšia metóda

---

[3] Dôležitým špecifickým prípadom je tzv. ***iid*** vzorka, z angl. **i**ndependent **i**dentically **d**istributed, čiže vzorka nielen s nezávislými náhodnými premennými, ale zároveň s náhodnými premennými s rovnakým rozdelením pravdepodobnosti.

výberu nemusí zabezpečiť reprezentatívnosť vzorky a to napriek tomu, že ide o *iid* vzorku. Otázkou, na ktorú je vhodné odpovedať je, načo je dobrá *iid* vzorka, pokiaľ nie je reprezentatívna, a teda nemôžeme na základe nej zovšeobecniť získané závery? O reprezentatívnosti vzorky môžeme hovoriť vtedy, pokiaľ poznáme určité (z hľadiska štatistickej analýzy dôležité) charakteristiky populácie, ktoré by mali byť v približne rovnakom pomere zastúpené aj vo vzorke. Môže sa stať, že na základe štatistického výberu (napríklad náhodného výberu), ktorý bude spĺňať podmienky *iid*, vyberieme z populácie o rovnakom podiele mužov a žien do vzorky iba samé ženy. Ak je pohlavie dôležitým parametrom pre štatistický výskum, takáto vzorka bude z praktického hľadiska (ako aj z hľadiska induktívnej štatistiky) nepoužiteľná, a to aj napriek tomu, že pôjde o *iid* vzorku. Ak na druhej strane sa budeme držať toho, že chceme reprezentatívnosť vzorky dodržať kontrolovaním počtu mužov a žien, takáto vzorka nebude (ako celok) *iid*.

Neraz sa môže stať, že sa sampling použije v situácii, v ktorej to nie je potrebné. Napríklad biológovi stačí skúmať určitý typ bunky a sledovať procesy v nej, keďže všetky bunky toho druhu a za tých podmienok budú z fyziologického hľadiska rovnaké, a teda aj procesy, ktoré sú predmetom skúmania, môže biológ zovšeobecniť na všetky bunky daného druhu. Ide o identickosť štatistických jednotiek, ktorá v ekonómii, v sociológii a v psychológii nie je veľmi bežná. Môžeme sa napríklad stretnúť s tzv. experimentálnou ekonómiou, ktorej mnohé závery sa týkajú experimentov vykonávaných na študentoch, pričom rozhodovacie procesy študentov nemusia mať nutne deterministický charakter a už vôbec nie charakter zovšeobecniteľný pre všetkých ľudí.

Získať vzorku, ktorá bude do istej miery spĺňať jednak potrebu reprezentatívnosti a jednak podmienky *iid* je predmetom teórie výberu štatistických jednotiek do vzorky – samplingu. Existujú mnohé metódy, ktoré sa v závislosti od okolností v praxi využívajú. Nie všetky sú však z pohľadu induktívnej štatistiky vhodné. Rozoznávame dve základné kategórie metód samplingu:

- Pravdepodobnostné metódy – pravdepodobnosť výberu štatistickej jednotky do vzorky je známa a výber štatistickej jednotky je náhodný, resp. závislý na náhode.
- Nepravdepodobnostné metódy – pravdepodobnosť výberu štatistickej jednotky do vzorky nemusí byť známa, prípadne výber štatistickej jednotky nie je náhodný.

Na záver tohto krátkeho prehľadu o výbere štatistických jednotiek spomenieme ešte dva javy, s ktorými sa čitateľ v praxi môže stretnúť. Prvým je situácia, keď zo štatistických jednotiek (oslovených respondentov, účastníkov prieskumu) niektoré odmietnu poskytnúť údaje. Ide o tzv. **non respondentov**. Ide o jav, ktorý je z pohľadu štatistickej analýzy

väčšinou negatívny. Non respondenti môžu byť nositeľmi dôležitých informácií, ktoré však z určitého dôvodu odmietnu uviesť a prieskumu sa nezúčastnia. To môže naše závery skresliť ťažko predvídateľným spôsobom.

Druhým javom je **pod-reprezentatívnosť** vzorky. Tá je charakteristická nevyváženým zastúpením kľúčových charakteristík populácie tak, ako sme to uviedli v predchádzajúcom prípade (príklad vysokého podielu žien vo vzorke). V prípade ak máme vo vzorke vysoký podiel neodpovedajúcich (hodnota je subjektívna, avšak spravidla sa za vysoký podiel považuje viac ako 30 % opýtaných) alebo je vzorka pod-reprezentovaná, je zovšeobecniteľnosť týchto výsledkov otázna.

2.1 Pravdepodobnostné metódy

Na prvý pohľad sa môže zdať zvláštne používanie pravdepodobnostných metód, založených na náhodnom výbere. V skutočnosti však, práve „*náhodnosť*“ výberu zaručuje dobré *iid* vlastnosti vzorky. Stručne si charakterizujeme najčastejšie pravdepodobnostné metódy.

2.1.1 Jednoduchý náhodný výber

Každá štatistická jednotka je vybraná na základe náhodného procesu. Môžeme ho uskutočniť s tzv. vrátením štatistickej jednotky späť do populácie alebo bez vrátenia. Za predpokladu, že populácia je dostatočne veľká, z pohľadu *iid* ako aj reprezentatívnosti je výsledok týchto dvoch obmien porovnateľný. Ide o spôsob výberu, ktorý je vo výskume najviac preferovaný.

Príklad 2.1
Spoločnosť zamestnávajúca 700 zamestnancov chce uskutočniť hĺbkový prieskum spokojnosti zamestnancov. Keďže podrobné dopytovanie sa všetkých zamestnancov nie je praktické, rozhodli sa odvodiť názory na základe náhodne vybranej vzorky o rozsahu 120 zamestnancov. Vybral sa zoznam všetkých zamestnancov a každému sa priradilo jedno zo 700 prirodzených čísel: 1, 2, 3, ..., 700. Pomocou generátora náhodných čísel sa náhodne vygenerovalo 120 čísel. Ak sa nejaké číslo už raz opakovalo, zobralo sa ďalšie číslo v poradí. Následne sa zistilo, o ktorých zamestnancov sa jedná a tí boli pozvaní na prieskum.

V programe R je ku generovaniu náhodných čísel možné použiť funkciu `sample()`, ktorá dokáže zabezpečiť, aby sa zvolené čísla neopakovali:

```
> a <- sample(1:700, 120, replace = F)
```

Vektor `"a"` predstavuje 120 náhodne vybraných čísel bez vrátenia (bližšie pozri `?sample`).

2.1.2 Systematický náhodný výber

Prvá štatistická jednotka sa z určitého zoznamu vyberie náhodne a potom každá ďalšia jednotka v štatistickom súbore je *n*-tá nasledujúca od predchádzajúcej vybranej štatistickej jednotky. Ak sa dôjde na koniec zoznamu, počíta sa plynule od začiatku. Interval výberu I_V je podielom rozsahu populácie N a požadovaného rozsahu vzorky n. Ak I_V nie je celé číslo, tak I_V sa pre jednotlivé intervaly zaokrúhľuje (nadol) na celé číslo.

Príklad 2.2
Developerskú spoločnosť zaujíma, ako sa obyvatelia mestskej časti postavia k vybudovaniu nového obchodného domu. Kým názory komunálnych politikov sú už zmapované, názory priamo od občanov nie. Developera pritom zaujíma, či sa v prípade potreby vie oprieť o podporu rezidentov. V zozname obyvateľov s trvalým pobytom v mestskej časti figuruje 25 280 obyvateľov. Developer sa rozhodol osloviť 500 rezidentov. Z abecedného zoznamu sa náhodne vybral jeden respondent a potom každý ďalší $\lfloor 25280/500 \rfloor = 50$-ty rezident, až pokiaľ ich nebolo požadovaných 500. Ak sa rezident odmietol vyjadriť, tak v jednom prípade sa to vyhodnocovalo ako negatívne stanovisko, v druhom sa postupovalo k ďalšiemu respondentovi v poradí.

```
> a <- c(1:25280, 1:25280)
> b <- a[sample(1:25280, size = 1)+floor(25280/500)*(0:499)]
```

Najprv sme si vytvorili vektor `a`, v ktorom sa nachádzajú ako keby dva rovnaké číselné zoznamy respondentov. Druhý vektor `b` predstavuje konkrétne poradia zo zoznamu. Tieto poradia sa vyberajú nasledovne. Najprv sme vybrali prvú hodnotu pomocou funkcie `sample(1:25280, size = 1)`. Toto je prvý rezident, ktorého sme zo zoznamu náhodne vybrali. Potom sme k nemu pripočítali hodnotu `floor(25280/500)`, a to bol druhý rezident zo zoznamu (funkcia `floor()` zaokrúhľuje čísla k najbližšiemu

najnižšiemu celému číslu). Postupne sme pripočítali násobky hodnôt `floor(25280/500)` k prvej hodnote. Potom, čo sme došli na koniec zoznamu, sa pokračuje plynule od začiatku zoznamu (preto vektor `a` predstavuje dva zoznamy).

2.1.3 Stratifikovaný náhodný výber

Pri stratifikovanom náhodnom výbere sa využívajú apriórne informácie o populácii. Vyberú sa kľúčové charakteristiky populácie (napr. pohlavie muž/žena a vzdelanie vyššie/stredné/nižšie), ktoré sú pre ďalší výskum kľúčové. Jednotlivé charakteristiky by sme mohli chápať aj ako sub-populácie. V rámci nich sa náhodne vyberajú štatistické jednotky do výberovej vzorky. Podiel týchto štatistických jednotiek vo výberovej vzorke pritom môže byť proporcionálny, t. j. s rovnakým podielom ako v populácii, alebo disproporcionálny[4].

Ak je našim cieľom zistiť, či je určitý postup alebo liek účinný na ľuďoch, testuje sa na niekoľkých vzorkách. Jedným z najčastejších stratifikátorov je pohlavie. Ide o charakteristiku ľudskej populácie, ktorá je v týchto typoch výskumov spravidla dosť dôležitá. Môže sa totižto stať, že liek bude mať iný efekt na mužov ako na ženy. Aby sa tieto odlišnosti mohli zachytiť, je želateľné, aby boli vo výskume v primeranom pomere zastúpené ženy aj muži, ak to má z povahy výskumu zmysel. Rovnaký príklad môžeme nájsť aj v obchode, kde pri predaji parfumov sú dôvody nákupu u žien a mužov spravidla odlišné. Muži kupujú ženám, ženy sebe alebo pre mužov. Pohlavie sa tu zdá byť znovu vhodným stratifikátorom.

2.1.4 Náhodný zhlukový výber

Ak je populácia príliš veľká alebo je jej rozsah nekonečný, je neraz vhodnou alternatívou k jednoduchému náhodnému výberu náhodný zhlukový výber. Ten z prirodzených zhlukov (pod zhlukom v tomto prípade budeme rozumieť množinu štatistických jednotiek s rovnakou hodnotou určitej charakteristiky, napr. miesto trvalého bydliska, odvetvie práce, typ auta a pod.) náhodne vyberie určitý počet zhlukov a potom sa z každého zhluku náhodne vyberú štatistické jednotky. Náhoda sa teda prejaví dvakrát – jednak pri výbere zhlukov, a jednak pri výbere štatistických jednotiek v rámci nich.

[4] V takom prípade sa využívajú techniky váženia odpovedí respondentov.

Príklad 2.3

Existuje odvetvová klasifikácia, ktorá vytvára vyše 100 rôznych odvetví. Zoberme si, že predmetom nášho záujmu je zistiť, či príslušnosť podnikov k nejakému odvetviu môže mať vplyv na priemernú finančnú výkonnosť týchto podnikov, meranú pomocou ukazovateľa ROA (z angl. ***R**eturn **O**n **A**ssets*, rentabilita aktív). Ide o tradičnú úlohu v odvetvovej analýze. Môže mať príslušnosť podnikov k odvetviu súvis so ziskovosťou podnikov? Ak je nepraktické získavať údaje od všetkých odvetví, jednou z alternatív je uskutočniť náhodný zhlukový výber. Zhluky reprezentujú jednotlivé odvetvia. Z týchto odvetví sa náhodne vyberie určitý počet – povedzme 10 a v rámci týchto odvetví sa náhodne vyberie požadovaný počet spoločností, ktoré do nich patria.

2.2 Nepravdepodobnostné metódy

V určitých situáciách nie je dosť dobre možné použiť pravdepodobnostné metódy a výskumník (resp. manažér) je obmedzený na nepravdepodobnostné výbery, neraz malého rozsahu. Na jednej strane je dôvera v zovšeobecniteľnosť záverov získaných z nepravdepodobnostných metód obmedzená, na strane druhej pri menšej vzorke neraz existuje priestor podrobnejšie skúmať štatistické jednotky, teda merať väčší počet štatistických znakov. V každom prípade, pokiaľ to je možné, odporúča sa používať pravdepodobnostné metódy vytvárania vzorky.

2.2.1 Pohodlný sampling

Do výberovej vzorky sa dostanú tie štatistické jednotky, od ktorých je jednoduché získať požadované údaje. Môže napríklad ísť o údaje, ku ktorým sa môžeme dostať čo najrýchlejšie alebo najlacnejšie, avšak nebudú reprezentovať populáciu a zrejme nebudú ani *iid* (tzn. *Independent and identically distributed*).

Jedným z prípadov je oslovovanie ľudí na ulici alebo dotazníky na internete. Poslednú dobu sú veľmi populárne ankety umiestňované na sociálnej sieti. Neraz sa môžeme stretnúť s tým, že zberatelia dát nevyberajú respondentov náhodne, ale odchytávajú ich na ulici v blízkosti miesta dopytovania. V prípade dotazníkov na internete je ďalším problémom skutočnosť, že na dotazník reagujú spravidla iba tí respondenti, ktorí majú extrémne vyhranený názor na témy, ktoré sa ich prieskumom dopytujeme. Napríklad extrémne pozitívny alebo extrémne negatívny a cítia potrebu tento svoj názor vyjadriť.

Týmto spôsobom sa do vzorky dostanú pomerne vyhranené odpovede a zovšeobecňovanie je otázne. Ak na histograme pozorujeme multi-modálnosť rozdelenia početností, niekedy je to indikátorom práve vyššie spomínanej situácie. Inou formou pohodlného samplingu je príjem tovaru na sklad, kde zamestnanec zodpovedný za kontrolu vyberie stále iba prvú paletu – lebo je to pohodlné. Ak si to dodávateľ všimne, ľahko sa môže stať, že to bude zneužívať – skúmaná vzorka potom zjavne nie je reprezentatívna.

2.2.2 Kvótny sampling

Ďalšou alternatívou samplingu je tzv. kvótny sampling. Najprv dôjde k zavedeniu kvót pre kľúčové charakteristiky populácie (napr. pohlavie, vzdelanie a iné). Ak získame potrebný počet respondentov s uvedenými kvótami, potom sa výsledky ďalších respondentov, ktorí prekročili svojou účasťou na prieskume stanovenú kvótu, do výsledkov nebudú započítavať.

2.2.3 Zámerný sampling

Ide o metódu, s ktorou sa často môžeme stretnúť v manažérskom výskume. Vzorka sa vyberá cielene na základe určitých a priori stanovených charakteristík a získavanie sa vedie osobným rozhovorom. Populáciu môžu napríklad reprezentovať vyššie postavení manažéri v hutníckom odvetví, pričom do vzorky sa vyberú vybraní manažéri, u ktorých sa predpokladá znalosť, skúsenosti a pod.

2.2.4 Snowball sampling

Využíva existujúce (zväčša sociálne) väzby medzi štatistickými jednotkami (osobami) na získavanie potrebných údajov, a to najmä v situáciách, kde nie je ľahké dostať sa k štatistickým jednotkám. Ide teda o populácie ťažko dosiahnuteľné, akými sú drogoví dealeri, drogovo závislí, rôzne undergroundové sociálne skupiny, skupiny organizovaného zločinu a pod.

Uvedené môže fungovať nasledovne. Na získavanie údajov sa použije osoba, ktorá má kontakt a dôveru v cieľovej populácii. Táto osoba následne cez svoje kontakty alebo priamo s respondentmi vykonáva rozhovor, pri ktorom získava potrebné štatistické údaje.

Aby sme mohli zovšeobecniť závery z údajov získaných z výberového súboru, musíme použiť metódy induktívnej štatistiky. Základný aparát, ktorý induktívna štatistika používa, sú tzv. bodové a intervalové odhady vybraných parametrov (polohy, variability a pod.) a testovanie štatistických hypotéz, teda tvrdení o populácii, ktoré sa overujú na základe údajov zo vzorky. Tejto oblasti sa budeme venovať v nasledujúcich kapitolách.

3 Odhady

3.1 Bodové odhady

Pri práci s výberovým súborom je jedným z našich prvých cieľov charakterizovať základné parametre populácie. Pod parametrami základného súboru pritom rozumieme charakteristiky, ktoré súhrnne opisujú štatistické znaky základného súboru. Príkladom je stredná hodnota (hodnoty parametrov základného súboru budeme označovať gréckymi písmenami), ktorú môžeme voľne interpretovať ako hodnotu, okolo ktorej sa „sústreďujú" ostatné hodnoty v populácii.[5] Označme si hodnotu parametra, ktorú chceme odhadnúť ako θ. Akúkoľvek funkciu hodnôt vzorky nazývame **štatistika**. Za odhad, tzv. estimátor, parametra θ môžeme považovať hodnotu určenú akoukoľvek štatistikou. Parameter θ je deterministická (spravidla však neznáma) konštanta. Z predchádzajúcej kapitoly je zrejmé, že ak tento parameter odhadneme z údajov náhodnej vzorky ako $\hat{\theta}$ (odhad budeme označovať strieškou), potom aj $\hat{\theta}$ bude náhodnou premennou.

Ako by sme mohli postupovať, keby sme chceli na základe údajov zo vzorky odhadnúť skutočnú strednú hodnotu populácie? Z predchádzajúceho odseku vyplýva, že ho môžeme odhadnúť použitím údajov zo vzorky a ľubovoľnej štatistiky: napríklad modálnou hodnotou, mediánom, aritmetickým priemerom alebo dokonca aj náhodne vybraným číslom z empirického štatistického súboru. Štatistika totiž nemá využívať parametre, ale namerané hodnoty. Ktorú štatistiku použiť? Je zrejmé, že niektoré odhady sú „lepšie" ako iné. Intuitívne sa zdá, že odhadnúť strednú hodnotu náhodným číslom by asi nebolo tak dobré, ako ju odhadnúť cez modálnu hodnotu vzorky. Ďalej odhadnúť strednú hodnotu modálnou hodnotou nemusí byť tak dobré, ako ju odhadnúť priemerom. Ako ale rozlíšime, čo je dobrý odhad? Existuje určitá skupina vlastností, ktoré od odhadov vyžadujeme. Tieto vlastnosti si v ďalšom texte stručne a pomerne intuitívne popíšeme.

3.1.1 Neskreslenosť (unbiasedness)

Ak stredná hodnota estimátora $\hat{\theta}$ parametra θ je rovná θ, potom hovoríme, že odhad parametra θ je neskreslený:

$$E[\hat{\theta}] = \theta \qquad (3.1)$$

[5] Pre formálnejšiu definíciu pozri publikáciu Kvantitatívne metódy v ekonómii I. (Lyócsa et al., 2013).

Intuitívne si to môžeme predstaviť nasledujúcim spôsobom. Vyberme náhodnú vzorku z populácie a vypočítajme odhad parametra θ pomocou štatistiky $\hat{\theta}$. Uvedený postup opakujeme ľubovoľný počet krát, pokiaľ nebudeme s určitou istotou schopní odhadnúť rozdelenie pravdepodobnosti parametra $\hat{\theta}$. Ak je estimátor $\hat{\theta}$ neskresleným, potom stredná hodnota tohto rozdelenia pravdepodobnosti by mala byť rovná skutočnej hodnote parametra θ. Rozdiel $E[\hat{\theta}] - \theta$ nazývame skreslením.

Ak je $\hat{\theta}(n)$ odhad parametra vyjadrený ako funkcia veľkosti rozsahu výberového súboru, potom menej striktnou alternatívou k predošlej definícii je nasledujúca, ktorú si nazveme asymptotická neskreslenosť:

$$\lim_{n\to\infty} E[\hat{\theta}(n)] = \theta \qquad (3.2)$$

Táto vlastnosť tvrdí, že stredná hodnota nášho odhadu sa bude rovnať skutočnej hodnote odhadovaného parametra θ v limitnom prípade (nekonečne veľkého výberového súboru).[6]

Môže sa stať, že nájdeme niekoľko štatistík, ktoré sú neskresleným odhadom parametra θ. Aby sme si medzi nimi mohli vybrať, je vhodné uvažovať aj o iných kritériách na hodnotenie estimátorov.

3.1.2 Efektívnosť (efficiency)

Uvažujme o dvoch neskreslených estimátoroch. Zrejme estimátor, ktorého variabilita bude menšia, budeme považovať za „lepší“, ako estimátor s vyššou variabilitou. Môžeme to chápať aj tak, že v prípade jedného výberu bude pravdepodobnosť, že náš odhad bude bližšie k skutočnej hodnote odhadovaného parametra väčšia pri estimátore s menšou variabilitou, ako pri estimátore s väčšou variabilitou.

Variabilitu estimátora si v niektorých prípadoch môžeme analyticky odvodiť, prípadne je alternatívnou možnosťou použiť simulačné metódy. Ak za odhad strednej hodnoty použijeme aritmetický priemer (estimátor), jeho variabilitu vieme odhadnúť ako σ^2/n. Keďže spravidla smerodajná odchýlka σ sledovaného štatistického znaku nie je známa (je to neznámy parameter populácie), nahradíme ju výberovou smerodajnou odchýlkou s (pre vzťah na výpočet pozri Kapitolu 3.1.4). Najmenšiu hodnotu variability, ktorú môže nadobúdať

[6] Neraz nás zaujíma, ako rýchlo bude konvergovať stredná hodnota nášho odhadu k skutočnej hodnote (v závislosti od veľkosti vzorky).

estimátor deterministického parametra populácie je možné vyjadriť pomocou tzv. Cramér-Raovej nerovnosti (pozri napr. Tkáč, 2001).

Druhým univerzálnejším prístupom je použiť tzv. **bootstrapingovú metódu**. Táto metóda sa stala v posledných rokoch pomerne populárna. Jej výhoda je, že nepotrebujeme vedieť, akým rozdelením pravdepodobnosti sa riadi odhadovaný parameter. Stručne si princíp môžeme popísať nasledujúcim spôsobom (tejto technike sa ešte budeme venovať neskôr). Nech $X_1, X_2,...., X_n$ je náhodný výber z určitej populácie a parameter populácie θ odhadujeme pomocou parametra $\hat{\theta}$. Získané hodnoty náhodných premenných v skúmanej vzorke označme $x_1, x_2, ... x_n$. Nech je medzi týmito n hodnotami s rôznych čísel ($s \leq n$). Potom vieme vytvoriť variačný rad $x_{(1)}, x_{(2)}, ... x_{(s)}$. Pomocou generátora pseudo-náhodných čísel získame tzv. boostrapové vzorky z rovnakého empirického rozdelenia pravdepodobnosti $f_n(x) = P_n(X = x)$ (ide o výber n prvkov s vrátením) a pre každú z týchto vzoriek odhadneme hľadaný parameter θ.

$$\hat{f}_n(x) = \begin{cases} 0, & ak \quad x < x_{(1)} \\ \dfrac{n_j}{n}, & ak \quad x = x_{(j)} \\ 0, & ak \quad x > x_{(s)} \end{cases} \qquad (3.3)$$

kde n_j je absolútna početnosť j-tej hodnoty vo výberovom súbore ($j = 1, 2, ..., s$). Počet náhodných výberov nie je presne stanovený, keďže veľa závisí od komplexnosti výpočtu odhadovaného parametra. Za dolnú hranicu môžeme považovať okolo 1000 iterácií. Na druhej strane by mal byť počet dostatočne veľký k tomu, aby zabezpečil určitú numerickú stabilitu výsledku. Získame tak súbor odhadovaných parametrov $\hat{\theta}$, z ktorých si môžeme vypočítať smerodajnú odchýlku. K bootstrapingovej metóde sa podrobnejšie vrátime v ďalšej časti.

Príklad 3.1

Finančná inštitúcia pripravovala špecifický produkt pre rodiny s deťmi. Z databázy sa dostala k nasledujúcim 38 údajom, kde náhodnú premennú reprezentuje vek dieťaťa v rodine.

```
vek <- c(4, 4, 4, 5, 5, 5, 5, 5, 7, 7, 7, 7, 8, 9, 10, 10, 10,
  10, 10, 12, 12, 12, 12, 12, 12, 13, 13, 13, 13, 13, 13, 13,
  15, 15, 15, 15, 18, 18)
```

Cieľom je odhadnúť strednú hodnotu veku detí. Aritmetický priemer na základe vzorky je `mean(vek) = 10.28947`. Variabilitu aritmetického priemeru môžeme odhadnúť ako

`sqrt(var(vek)/length(vek))` = `0.6454876`. Rozhodli sme sa taktiež odhadnúť smerodajnú odchýlku priemeru pomocou boostrapingu. Z empirického rozdelenia pravdepodobnosti sme vygenerovali 1000 vzoriek o rozsahu 38 pozorovaní pomocou generátora pseudo-náhodných čísel. Pre každú vzorku sme vypočítali aritmetický priemer a na záver sme vypočítali aj príslušnú výberovú smerodajnú odchýlku vzorky priemerov. Výsledky sú veľmi podobné ako tie, ku ktorým dospejeme pomocou vzťahu s^2/n.

```
> means <- c()
> for (i in 1:1000) {
+ a <- sample(1:length(vek), size = length(vek), replace = T)
+ b <- vek[a]
+ means[i] <- mean(b)
+ }
> sd(means)
[1] 0.6568982
```

Za predpokladu, že na odhad parametra populácie použijeme štatistiku, ktorá je skreslená, „výhodnosť“ použitia takej štatistiky môžeme posúdiť pomocou priemernej sumy štvorcov (*MSE*) estimátora $\hat{\theta}$. Pripomeňme, že pre skreslený estimátor platí $E[\hat{\theta}] \neq \theta$. Potom očakávaný štvorec rozdielu medzi estimátorom $\hat{\theta}$ a skutočnou hodnotou parametra θ je ***MSE***.

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = \int_{-\infty}^{\infty} \left(\hat{\theta} - \theta\right)^2 f(\hat{\theta}) d\hat{\theta} \qquad (3.4)$$

Myšlienka použitia MSE spočíva okrem iného v tom, že na jednej strane môžeme mať estimátor, ktorý je skreslený, ale pokiaľ má zároveň „výrazne“ nižšiu variabilitu ako neskreslený estimátor, môže byť jeho použitie výhodnejšie. Všimnime si, že rozdiel v zátvorke nie je rozdiel odhadu od strednej hodnoty odhadu (pripomeňme, že pokiaľ $\hat{\theta}$ je skresleným odhadom, neplatí $E(\hat{\theta}) = \theta$).

Ak by sme v predošlom vzťahu zapísali:

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] = D[\hat{\theta}] \qquad (3.5)$$

Existuje vzťah medzi:

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = D[\hat{\theta}] + \left(E[\hat{\theta}] - \theta\right)^2 \qquad (3.6)$$

$$
\begin{aligned}
MSE(\hat{\theta}) &= E\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= E\left[\left[\left(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta\right)\right]^2\right] \\
&= E\left[\left[\left(\hat{\theta}-E(\hat{\theta})\right)+\left(E(\hat{\theta})-\theta\right)\right]^2\right] \\
&= E\left[\left(\hat{\theta}-E(\hat{\theta})\right)^2\right]+2E\left[\left(\hat{\theta}-E(\hat{\theta})\right)\left(E(\hat{\theta})-\theta\right)\right]+E\left[\left(E(\hat{\theta})-\theta\right)^2\right] \\
&= D\left[\hat{\theta}\right]+2\left[E\left(\hat{\theta}\right)-E\left(\hat{\theta}\right)\right]\left[E(\hat{\theta})-\theta\right]+E\left[\left(E(\hat{\theta})-\theta\right)^2\right] \\
&= D\left[\hat{\theta}\right]+E\left[\left(E(\hat{\theta})-\theta\right)^2\right]
\end{aligned}
$$

čo je vlastne:

$$
MSE(\hat{\theta}) = D\left[\hat{\theta}\right]+bias^2 \tag{3.7}
$$

3.1.3 Konzistentnosť (consistency)

Pokiaľ použijeme estimátor $\hat{\theta}$ na odhad parametra θ, pričom použijeme údaje zo vzorky o rozsahu *n*, tak potom zrejme je lepší taký typ estimátora $\hat{\theta}$, pre ktorý platí, že so zvyšujúcim sa rozsahom súboru bude väčšia aj naša istota, že hodnota estimátora $\hat{\theta}$ bude bližšie k skutočnej hodnote parametra θ. Pri zvyšovaní rozsahu súboru nárast „našej istoty" neznamená nutne monotónne približovanie sa odhadu k skutočnej hodnote. Intuitívne uvedené kritérium naznačuje, že ak je veľkosť výberovej vzorky totožná s veľkosťou populácie, tak by malo platiť $\hat{\theta}=\theta$. Formálne vieme konzistentnosť vyjadriť nasledujúcim vzťahom ($\forall\varepsilon>0$):

$$
\lim_{n\to\infty} P\left(\left|\hat{\theta}(n)-\theta\right|<\varepsilon\right)=1 \tag{3.8}
$$

alebo:

$$
\lim_{n\to\infty} P\left(\left|\hat{\theta}(n)-\theta\right|\geq\varepsilon\right)=0 \tag{3.9}
$$

kde $\hat{\theta}(n)$ je odhad parametra θ, ktorý je vyjadrený ako funkcia rozsahu štatistického súboru. Prvý výraz si môžeme interpretovať nasledovne[7]: ak si zvolíme ľubovoľné malé kladné číslo $\varepsilon > 0$, tak chyba, ktorej sa pri odhade dopustíme s $n \to \infty$ bude takmer isto menšia ako toto číslo ε. Ide o pomerne dôležitú vlastnosť estimátora vo väčších vzorkách.

[7] Presnejšia interpretácia by si vyžadovala definovanie konvergencie rozdelení pravdepodobnosti, prípadne konvergenciu v pravdepodobnosti.

Samozrejme, ak náš estimátor spĺňa všetky zo spomínaných vlastností, je to preferovaná situácia.[8]

Našim cieľom bolo načrtnúť vybrané kritériá odhadov a hlbšia analýza je mimo rámec tohto textu. To, čo je pre hospodársku prax a empirický výskum dôležité, je vedieť aké vlastnosti majú nami použité estimátory. Pre väčšinu aplikácii si vystačíme s určitým zoznamom najvhodnejších bodových odhadov.

3.1.4 Najpoužívanejšie bodové odhady

V tejto časti zavedieme značenie, ktoré je odlišné od značenia, ktoré sme použili v opisnej štatistike.[9] V prípade, ak hovoríme o výberových charakteristikách, označujeme ich písmenami latinskej (rímskej) abecedy, kým v prípade charakteristík populácie budeme používať grécke písmená. Zároveň si nadefinujeme také odhady, ktoré patria k najlepším odhadom (podľa vlastností popísaných vyššie).

Výberový priemer

Ide o bodový odhad strednej hodnoty, ktorý je neskresleným, konzistentným a efektívnym odhadom.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3.10)$$

Výberový medián

Za predpokladu, že hodnoty výberového súboru o rozsahu *n* sú zoradené vo variačnom rade a *n* je nepárne číslo, potom výberový medián je hodnota na $\frac{(n+1)}{2}$ mieste variačného radu. Za predpokladu, že rozsah súboru *n* je párne číslo, potom výberový medián je v intervale od hodnoty na $\frac{n}{2}$ po $\frac{n+2}{2}$ mieste, resp. je zvykom vypočítať aritmetický priemer z hodnôt, ktoré patria na uvedenú pozíciu variačného radu.

[8] V ekonometrii sa stretávame s preferenciou mať najmä neskreslené a konzistentné odhady.
[9] Pozri Kvantitatívne metódy v ekonómii I. (Lyócsa et al., 2013)

Výberový modus

Je najpočetnejšou hodnotou výberového súboru. V prípade, ak neexistuje jedna najpočetnejšia hodnota, potom existujú iné metódy (ktoré si tu nebudeme prezentovať) odhadu modálnej hodnoty.

Výberový rozptyl a výberová smerodajná odchýlka

Na rozdiel od rozptylu je výberový rozptyl mierne upravený tak, aby spĺňal podmienky pre bodové odhady rozoberané vyššie. Ide teda o neskreslený, konzistentný a efektívny odhad rozptylu.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (3.11)$$

$$s = \sqrt{s^2} \qquad (3.12)$$

Na rozdiel od výberového rozptylu, výberová smerodajná odchýlka je skresleným (podhodnoteným) odhadom populačnej smerodajnej odchýlky:

$$E[s] = E[\sqrt{s^2}] \leq \sqrt{E[s^2]} = \sigma \qquad (3.13)$$

Výberový koeficient šikmosti

V literatúre sa často môžeme stretnúť s tým, že koeficient šikmosti a výberový koeficient šikmosti sú matematicky totožné. Koeficient šikmosti ako odhad skutočného koeficientu šikmosti populácie je skreslený (v angl. *biased*). Jeho neskreslený tvar je nasledujúci (Bacon, 2008):

$$\gamma_3 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 \qquad (3.14)$$

Výberový koeficient špicatosti

Podobne ako pri koeficiente šikmosti, aj pri koeficiente špicatosti (máme na mysli tzv. excess kurtosis) sa v literatúre môžeme stretnúť s tým, že výberový koeficient špicatosti a koeficient špicatosti sú totožné. V skutočnosti by však znova išlo o skreslený odhad. Jeho neskreslený tvar je nasledujúci (ide o tzv. *sample excess kurtosis*, to znamená, že hodnoty väčšie ako 0 znamenajú hodnoty „*špicatejšie*“ ako v prípade normálneho rozloženia hodnôt; Bacon, 2008):

$$\gamma_4 = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right) - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (3.15)$$

Výberový podiel

V praxi sa často stretávame s potrebou odhadnúť podiel (časť z celku, percento) štatistických jednotiek s určitou hodnotou štatistického znaku, napr. podiel nezhodných výrobkov, podiel neuspokojených zákazníkov, podiel nesplatených záväzkov a pod. Na základe výberového súboru odhadujeme podiel tzv. výberovým podielom, pre ktorý platí:

$$p_v = \frac{M_n(A)}{n} \qquad (3.16)$$

kde $M_n(A)$ je počet štatistických prvkov s požadovanou vlastnosťou A vo výberovom súbore o rozsahu n.

3.1.5 Simulácia vlastností bodových odhadov

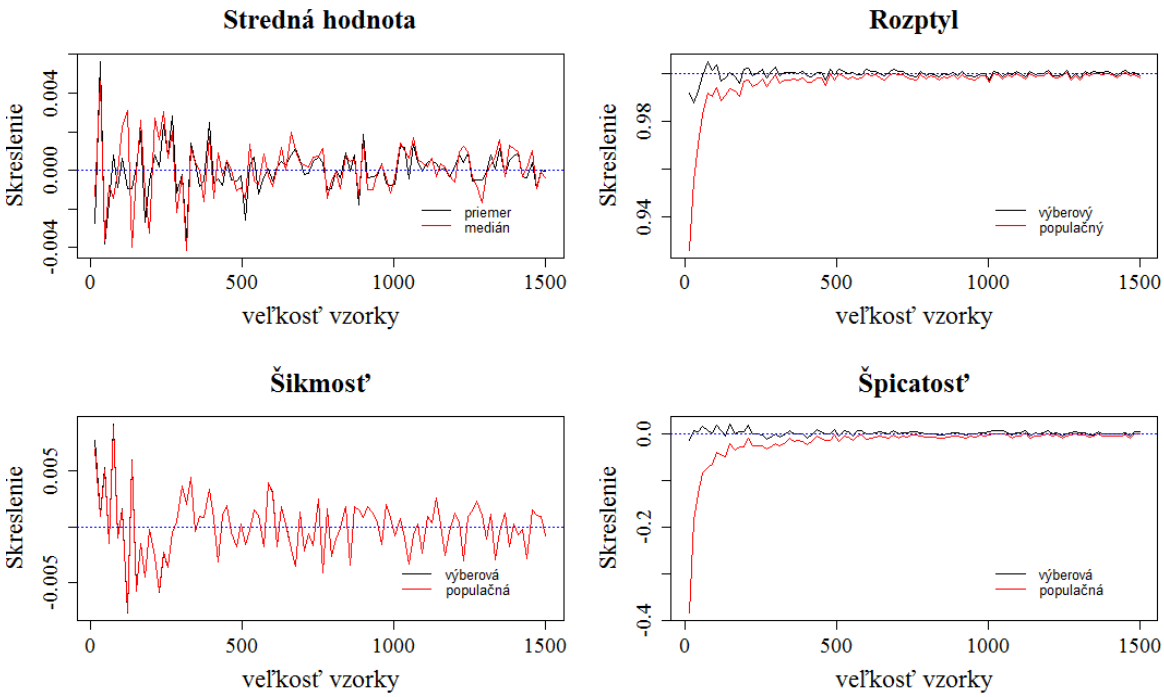
Uskutočnili sme jednoduchý pokus pomocou simulácií v programe R. Cieľom simulácie bolo vypočítať skreslenie a *MSE* odhadu štyroch parametrov: strednej hodnoty μ, rozptylu σ^2, šikmosti γ_3 a špicatosti γ_4. Plán experimentu je nasledujúci: z normálneho rozdelenia pravdepodobnosti so strednou hodnotou $\mu = 0$ a s rozptylom $\sigma^2 = 1$ sa náhodne vyberie n pozorovaní. Skutočné hodnoty odhadovaných parametrov poznáme: $\mu = 0$, $\sigma^2 = 1$, $\gamma_3 = 0$ a $\gamma_4 = 0$. Hľadané parametre populácie sa odhadnú nasledovne: stredná hodnota sa odhadne pomocou aritmetického priemeru v jednom a mediánu v druhom prípade, rozptyl σ^2 sa odhadne výberovým rozptylom v jednom a rozptylom v druhom prípade. Šikmosť aj špicatosť sa bude odhadovať v jednom prípade s populačných vzorcov a v druhom pomocou výberových charakteristík. V simulácii budeme zvyšovať rozsah súboru n = 15, 30, ..., 1500. Pre každý z týchto rozsahov uskutočníme 2000 pokusov. Zaujímať nás bude ako sa s rastúcou vzorkou mení skreslenie a *MSE* pre každý jeden z odhadov. Nasledujúci kód predstavuje tvorbu funkcií na výpočet výberovej šikmosti a špicatosti. Argumentom týchto funkcií je vektor údajov, ktorého výberové charakteristiky nás zaujímajú.

```
> sample_skew <- function(data) {
+ L <- length(data)
+ sample_skew <- (L/((L - 1)*(L - 2))) * sum(((data -
  mean(data))/sd(data))^3)
+ return(sample_skew)
+ }
------------------------------------------------------------
```

```
> sample_kurt <- function(data) {
+ L <- length(data)
+ sample_kurt <- ((L*(L + 1))/((L - 1)*(L - 2)*(L - 3))) *
  sum(((data - mean(data))/sd(data))^4) - (3*(L - 1)^2)/((L -
  2)*(L - 3))
+ return(sample_kurt)
+ }
```

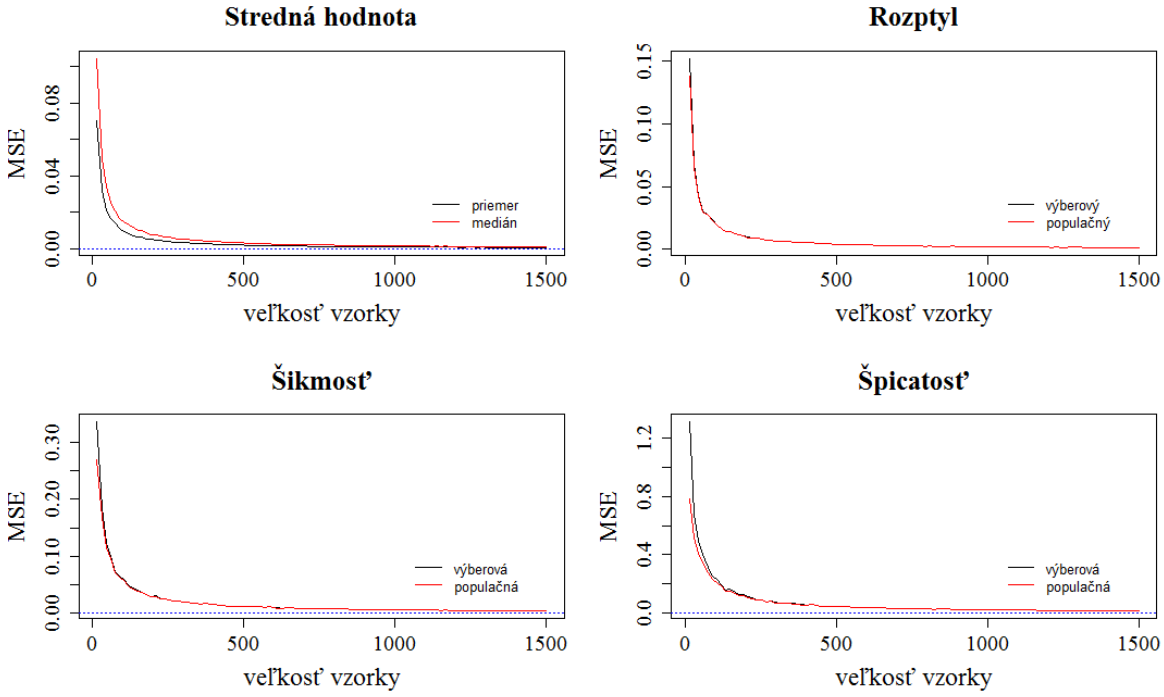
Výsledky sme zobrazili na nasledujúcich dvoch obrázkoch (pozri Obrázok 3.1 a Obrázok 3.2), kde výberové a populačné odhady uvádzané v obrázkoch sú nesreslené a skreslené odhady. Pri odhadoch strednej hodnoty a šikmosti nebadať veľké rozdiely medzi priemerom a mediánom na jednej strane a šikmosťou a výberovou šikmosťou na strane druhej. Oproti tomu, aspoň vizuálne sa zdajú byť rozdiely (najmä pre menšie vzorky) významné pri porovnaní rozptylu a výberového rozptylu na jednej a špicatosti a výberovej špicatosti na druhej strane. Výberové charakteristiky mali menšie skreslenie. Pri MSE sme namerali zaujímavé výsledky pri strednej hodnote, kde sa zdá byť efektívnosť aritmetického priemeru vyššia (menšie *MSE*) ako mediánu. Na druhej strane, pri menších vzorkách mala výberová špicatosť mierne menšiu efektívnosť. Tieto výsledky sú samozrejme len orientačné a do značnej miery závisia od rozdelenia, z ktorého sme simulovali vzorky, ako aj od parametrov rozdelenia.

```
> library(moments)
> sample <- seq(from = 15, to = 1500, by = 15)
> MVM <- matrix(ncol = 10, nrow = length(sample)) # Matica
  výsledkov - priemery z odhadov
> MVV <- matrix(ncol = 10, nrow = length(sample)) # Matica
  výsledkov - MSE z odhadov
> j = 1 #n = 100
> for (n in sample) {
+ MPV <- matrix(ncol = 10, nrow = 2000) # Matica predbežných
  výsledkov
+ for (i in 1:2000) {
+ a = rnorm(n)
+ MPV[i,] <- c(mean(a), median(a), var(a)*((n-1)/n), var(a),
  sqrt(var(a)*((n-1)/n)), sqrt(var(a)), skewness(a),
  sample_skew(a), kurtosis(a) - 3, sample_kurt(a))
+ }
+ MVM[j,] <- c(apply(MPV, 2, FUN = mean))
+ MVV[j,] <- c(mean((MPV[,1] - 0)^2), mean((MPV[,2] - 0)^2),
  mean((MPV[,3] - 1)^2), mean((MPV[,4] - 1)^2), mean((MPV[,5] -
  1)^2), mean((MPV[,6] - 1)^2), mean((MPV[,7] - 0)^2),
  mean((MPV[,8] - 0)^2), mean((MPV[,9] - 0)^2), mean((MPV[,10] -
  0)^2))
+ j = j + 1
+ }
```



Obrázok 3.1: Skreslenie odhadov

Zdroj: vlastné spracovanie, výstup zo softvéru R



Obrázok 3.2: *MSE* odhadov

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(2,2))
> plot(seq(from = 15, to = 1500, by = 15), MVM[,1], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "Skreslenie", xlab = "veľkosť vzorky", main = "Stredná
  hodnota", cex.main = 1.9, ylim = c(range(MVM[,1], MVM[,2])))
> lines(seq(from = 15, to = 1500, by = 15), MVM[,2], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 0, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("priemer", "medián"), lty =
  1, col = c("black", "red"), inset = 0.08, bty = "n")
> plot(seq(from = 15, to = 1500, by = 15), MVM[,4], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "Skreslenie", xlab = "veľkosť vzorky", main =
  "Rozptyl", cex.main = 1.9, ylim = c(range(MVM[,3], MVM[,4])))
> lines(seq(from = 15, to = 1500, by = 15), MVM[,3], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 1, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("výberový", "populačný"), lty
  = 1, col = c("black", "red"), inset = 0.08, bty = "n")
> plot(seq(from = 15, to = 1500, by = 15), MVM[,8], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "Skreslenie", xlab = "veľkosť vzorky", main =
  "Šikmosť", cex.main = 1.9, ylim = c(range(MVM[,7], MVM[,8])))
> lines(seq(from = 15, to = 1500, by = 15), MVM[,7], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 0, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("výberová", "populačná"), lty
  = 1, col = c("black", "red"), inset = 0.08, bty = "n")
> plot(seq(from = 15, to = 1500, by = 15), MVM[,10], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "Skreslenie", xlab = "veľkosť vzorky", main =
  "Špicatosť", cex.main = 1.9, ylim = c(range(MVM[,9],
  MVM[,10])))
> lines(seq(from = 15, to = 1500, by = 15), MVM[,9], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 0, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("výberová", "populačná"), lty
  = 1, col = c("black", "red"), inset = 0.08, bty = "n")
-------------------------------------------------------------
> par(mfrow = c(2,2))
> plot(seq(from = 15, to = 1500, by = 15), MVV[,1], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "MSE", xlab = "veľkosť vzorky", main = "Stredná
  hodnota", cex.main = 1.9, ylim = c(range(MVV[,1], MVV[,2])))
> lines(seq(from = 15, to = 1500, by = 15), MVV[,2], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 0, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("priemer", "medián"), lty =
  1, col = c("black", "red"), inset = 0.08, bty = "n")
> plot(seq(from = 15, to = 1500, by = 15), MVV[,4], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "MSE", xlab = "veľkosť vzorky", main = "Rozptyl",
  cex.main = 1.9, ylim = c(range(MVV[,3], MVV[,4])))
> lines(seq(from = 15, to = 1500, by = 15), MVV[,3], type = "l",
  lwd = 1.5, col = "red")
```

```
> abline(h = 1, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("výberový", "populačný"), lty
  = 1, col = c("black", "red"), inset = 0.08, bty = "n")
> plot(seq(from = 15, to = 1500, by = 15), MVV[,8], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "MSE", xlab = "veľkosť vzorky", main = "Šikmosť",
  cex.main = 1.9, ylim = c(range(MVV[,7], MVV[,8])))
> lines(seq(from = 15, to = 1500, by = 15), MVV[,7], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 0, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("výberová", "populačná"), lty
  = 1, col = c("black", "red"), inset = 0.08, bty = "n")
> plot(seq(from = 15, to = 1500, by = 15), MVV[,10], type = "l",
  lwd = 1.5, family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "MSE", xlab = "veľkosť vzorky", main = "Špicatosť",
  cex.main = 1.9, ylim = c(range(MVV[,9], MVV[,10])))
> lines(seq(from = 15, to = 1500, by = 15), MVV[,9], type = "l",
  lwd = 1.5, col = "red")
> abline(h = 0, col = "blue", lwd = 1.5, lty = 3)
> legend("bottomright", legend = c("výberová", "populačná"), lty
  = 1, col = c("black", "red"), inset = 0.08, bty = "n")
```

3.2 Intervalový odhad

Je prirodzenou snahou štatistickej analýzy získať čo najpresnejší bodový odhad hodnoty skutočného parametra. Za predpokladu, že hodnota parametra je tzv. spojitá premenná (napr. teplota môže nadobúdať ľubovoľné hodnoty v určitom intervale), potom pravdepodobnosť, že bodový odhad bude odlišný od skutočnej hodnoty parametra populácie je takmer isto rovná 1 (inak povedané, skoro určite neodhadneme skutočnú teplotu). Na druhej strane existuje potreba definovať určitú mieru precíznosti, resp. konfidencie. Samotný bodový odhad nehovorí nič o tom, nakoľko sme „blízko“ ku skutočnej hodnote parametra populácie. Našim cieľom je získať metódu, pomocou ktorej by sme mohli vyhlásiť, že skutočná hodnota parametra sa s určitou pravdepodobnosťou nachádza v nejakom vymedzenom intervale. V tejto podkapitole sa budeme venovať práve tvorbe takýchto intervalov pre najčastejšie odhadované parametre populácie. Na záver si tiež ukážeme, ako použiť bootstrapingovú metódu pre výpočet takýchto intervalov.

Pri odhadoch sa zrejme najčastejšie môžeme stretnúť s troma druhmi intervalov: intervaly konfidenčné (týmto bude venovaná táto kapitola), tolerančné intervaly a predikčné intervaly. Účelom **konfidenčných intervalov** je s určitou pravdepodobnosťou $\gamma = 1 - \alpha$ (konfidenčnou pravdepodobnosťou) stanoviť interval, v ktorom sa nachádza skutočný parameter θ populácie. Parameter α je tzv. **hladina významnosti** a predstavuje pravdepodobnosť, že skutočný parameter θ sa v konfidenčnom intervale **nenachádza.** Pod

pojmom **tolerančný interval** sa spravidla rozumejú intervaly, v ktorých sa s určitou pravdepodobnosťou budú nachádzať konkrétne hodnoty populácie, nie parametre populácie. Účelom predikčných intervalov je s určitou pravdepodobnosťou odhadnúť interval, v ktorom sa vyskytne budúca konkrétna hodnota (nové meranie, nová realizácia procesu) populácie alebo parameter populácie.

Ak k tvorbe konfidenčných intervalov pristupujeme analyticky, vychádzame zo známeho rozdelenia pravdepodobnosti odhadovaného parametra θ. Našou snahou je na základe výberového súboru odhadnúť parametre tohto rozdelenia pravdepodobnosti a následne určiť také hranice, pre ktoré bude platiť nasledujúca rovnosť:

$$P(d \leq \theta \leq h) = 1 - \alpha \qquad (3.17)$$

kde d je dolnou a h je hornou hranicou konfidenčného intervalu. Obe tieto hranice sú pritom **náhodné premenné**, keďže závisia od hodnôt výberového súboru, ktorý (predpokladáme) sme získali na základe **náhodného výberu**. Postup tvorby konfidenčného intervalu si ukážeme na tvorbe konfidenčných intervalov pre rozptyl.

Za predpokladu, že populácia má normálne rozdelenie pravdepodobnosti, so strednou hodnotou μ a rozptylom σ^2, je známe, že náhodná premenná $(n-1)s^2/\sigma^2$ má χ^2 rozdelenie s $n-1$ stupňami voľnosti. Keďže χ^2 je známe rozdelenie, vieme určiť hodnotu dolného, resp. horného kvantilu (d_χ - dolný kvantil, h_χ - horný kvantil) medzi ktorými sa s $1-\alpha=\gamma$ pravdepodobnosťou (niekedy sa po prenásobení 100% označuje aj ako konfidencia) bude nachádzať skutočná hodnota hľadaného populačného parametra. Označme si $\chi^2_{\alpha,\ k}$ ako hodnotu χ^2 rozdelenia s k stupňami voľnosti, od ktorej je náhodne vybraná hodnota z χ^2_k rozdelenia s k stupňami voľnosti menšia s pravdepodobnosťou α. Inak povedané $P(\chi^2_k < \chi^2_{\alpha,k}) = \alpha$. Potom konfidenčný interval pre rozptyl si vieme určiť z nasledujúceho vzťahu:

$$P\left(\chi^2_{(\alpha/2),k} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{(1-\alpha/2),k}\right) = 1-\alpha \qquad (3.18)$$

Po jednoduchých úpravách si vieme odvodiť interval pre rozptyl σ^2:

$$P\left(\frac{\chi^2_{(\alpha/2),k}}{(n-1)s^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{(1-\alpha/2),k}}{(n-1)s^2}\right) = 1-\alpha \qquad (3.19)$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2),k}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(\alpha/2),k}}\right) = 1-\alpha \qquad (3.20)$$

Hodnota $\chi^2_{1-\alpha/2,\ k=(n-1)}$ a $\chi^2_{\alpha/2,\ k=(n-1)}$ býva uvedená v štatistických tabuľkách, pričom je možné ju odčítať na základe známeho počtu stupňov voľnosti k. V programe R k tomu budeme používať funkciu `qchisq()`. Keďže ide o kvantil χ^2 rozdelenia, tak pre označenie

hodnoty dolného kvantilu používame označenie $\chi^2_{\alpha/2,\,(n-1)}$ a na označenie horného kvantilu používame $\chi^2_{1-\alpha/2,\,(n-1)}$. V literatúre sa môžeme stretnúť aj s opačným označením dolnej a hornej hranice vo výraze *(3.18)*, čo je dôsledok inak zadefinovaného výrazu $\chi^2_{\alpha,\,k}$, ktorý sa chápe ako $P(\chi^2_k > \chi^2_{\alpha,\,k}) = \alpha$.

Pozornému čitateľovi určite neušlo, že pre stanovenie kvantilov χ^2 rozdelenia sme použili označenie $1 - \alpha/2$ a nie $1 - \alpha$. Dôvod je ten, že ak nás zaujíma 95 % konfidencia, tak potom od horného kvantilu hodnoty väčšie a od dolného kvantilu hodnoty menšie musia predstavovať spolu 5 % všetkých hodnôt. Z toho vyplýva[10], že ak nás zaujíma dolný kvantil, tak nás v danom prípade zaujíma $\chi^2_{\alpha/2,\,(n-1)}$ v našom prípade $\chi^2_{0.025,\,(n-1)}$ a v prípade horného kvantilu $\chi^2_{1-\alpha/2,\,(n-1)}$, čomu zodpovedá $\chi^2_{0.975,\,(n-1)}$.

Doteraz sme uvažovali o obojstranných intervaloch spoľahlivosti. Bežne sa však môžeme stretnúť s potrebou určiť **jednostranný interval**.

Príklad 3.2

Príklad na tvorbu jednostranného intervalu si môžeme opísať nasledovne. Rýchlosť, s akou zákaznícke centrum vybaví požiadavku zákazníka, je parameter, ktorý môžeme merať časom. Je našim záujmom, aby priemerný čas vybavenia zákazníka nepresiahol určitú hodnotu *t*. Na základe náhodných meraní môžeme odhadnúť pravostranný konfidenčný interval pre priemernú hodnotu vybavenia požiadavky zákazníka. Pravostranný v zmysle, že bude mať iba horné číselné ohraničenie a dolný kvantil nahradíme najmenšou zmysluplnou hodnotou, pri čase zrejme 0.

3.2.1 Niektoré vlastnosti intervalových odhadov

Na veľkosť konfidenčných intervalov môžeme vplývať nasledujúcimi spôsobmi:

- spôsobom akým vyberáme hodnoty do výberového súboru (uprednostňuje sa náhodný výber),
- rozsahom výberového súboru,
- voľbou hladiny významnosti α,
- metódou výpočtu konfidenčného intervalu.

V tejto časti sa budeme venovať vplyvu dvoch parametrov na veľkosť konfidenčných intervalov: rozsahu štatistického súboru *n* a voľbe hladiny významnosti α.

[10] Všimnite si rozdiel pri jednostranných konfidenčných intervaloch.

Rozsah štatistického súboru

Hranice konfidenčných intervalov sú náhodné premenné, ktoré závisia od hodnôt, ktoré máme k dispozícii vo výberovom súbore. Preto nie je možné jednoznačne povedať, aký bude vplyv zväčšenia rozsahu štatistického súboru na šírku intervalu. Vo všeobecnosti je však snahou mať čo najväčší rozsah štatistického súboru. Ak by sme zvýšili rozsah štatistického súboru a ostatné parametre by sa nezmenili, potom spravidla šírka intervalu je pri konštantnej konfidencii menšia.[11]

Hladina významnosti

Je zrejme racionálnym cieľom získať taký konfidenčný interval, v ktorom s čo najväčšou pravdepodobnosťou budeme očakávať skutočnú hodnotu parametra θ. Keď je konfidencia voliteľný parameter, prečo si nevybrať 100 % konfidenciu? Prečo si nevybrať hladinu významnosti $\alpha = 0.00$?

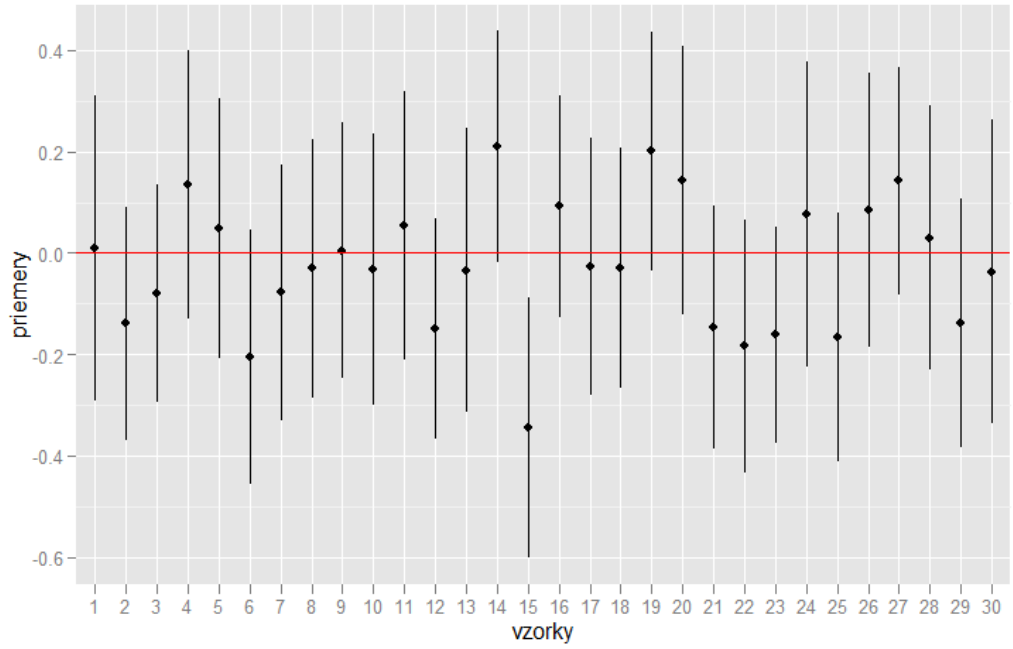
Predstavme si, že by sme mali odhadnúť vonkajšiu teplotu vzduchu, pričom máme určiť interval, v ktorom sa táto teplota bude s určitou konfidenciou nachádzať. Ak by bolo kritérium „určite sa nepomýliť“, tak by sme mohli vytvoriť interval od 0 K až[12] po $+\infty$ K a naša konfidencia by bola 100 %. Avšak praktická hodnota takto vytvoreného intervalu je minimálna.

Z predchádzajúceho príkladu môžeme intuitívne tušiť, že za inak nezmenených podmienok, s klesajúcou konfidenciou by sa šírka intervalu nemala zväčšovať. Na nasledujúcom obrázku (pozri Obrázok 3.3), môžeme vidieť na osi x nasimulované vzorky a na osi y príslušné konfidenčné intervaly (odhadovali sme strednú hodnotu). Skutočnú hodnotu parametra sme zvýraznili horizontálnou čiarou (červená čiara s hodnotou $y = 0$). Vytvorili sme 30 konfidenčných intervalov s 95 % konfidenciou, z ktorých v jednom prípade sa skutočná hodnota nenachádza v nami definovanom intervale.

---

[11] V predchádzajúcom príklade tvorby konfidenčného intervalu pre rozptyl sa rozsah vzorky nachádza nie len v čitateli, ale aj v menovateli, keďže sa používa pre výber vhodného kvantilu (počet stupňov voľnosti $k = n - 1$). S rastom veľkosti vzorky tak rastie nie len hodnota čitateľa ale aj menovateľa. Na akom definičnom obore ktorý výraz (či menovateľ alebo čitateľ) rastie rýchlejšie v závislosti od veľkosti vzorky n, to necháme na čitateľa.

[12] Pri meraní teploty v stupňoch Kelvina je 0 tzv. teplota absolútnej nuly, čiže najnižšia možná teplota, ktorá je fyzikálne definovaná (odhliadnuc od kvantových javov).



Obrázok 3.3: Simulované intervaly spoľahlivosti pre strednú hodnotu s 95 % konfidenciou

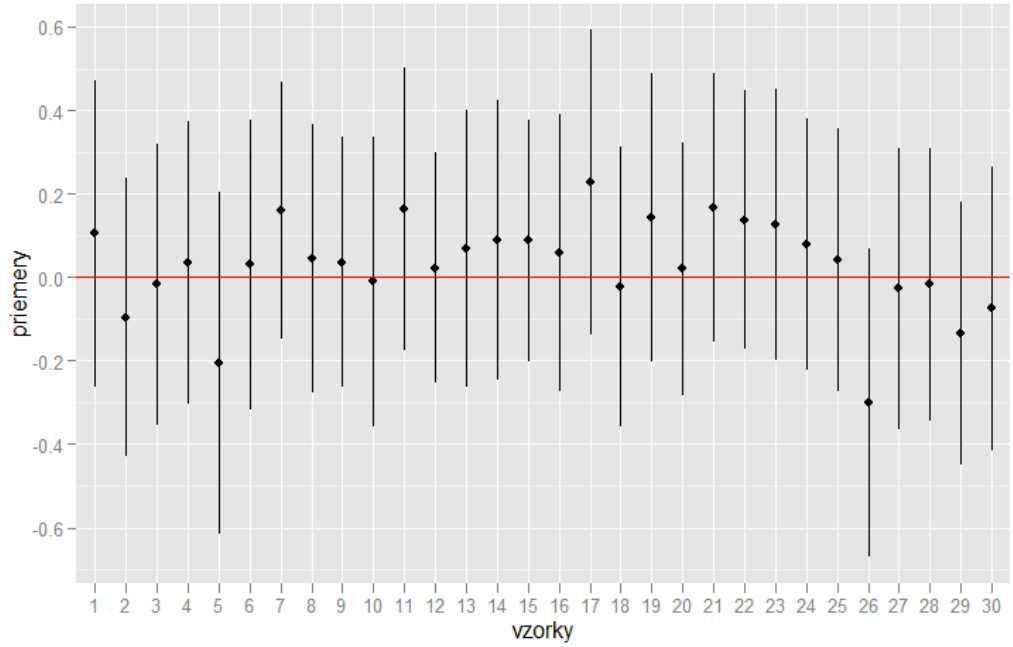
Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> library(ggplot2)
> polomer <- c()
> priemery <- c()
> kvantil <- qnorm(0.975, lower.tail = T)
> for (i in 1:30) {
+ a <- rnorm(60)
+ priemery <- c(priemery , mean(a))
+ polomer <- c(polomer, kvantil * (var(a)/60)^0.5)
+ }
> data <- data.frame(priemery, polomer, vzorky = factor(1:30))
> limits <- aes(ymax = priemery + polomer, ymin = priemery -
  polomer)
> p <- ggplot(data, aes(x = vzorky, y = priemery, fill =
  vzorky))
> p + geom_pointrange(limits)+ geom_hline(yintercept = 0, col =
  "red") + theme(legend.position = "none")
```

Z predchádzajúceho obrázku (Obrázok 3.3) je teda možné vidieť, koľkokrát by sme sa na základe konfidenčného intervalu „trafili“ a interval by obsahoval aj skutočnú hodnotu parametra. Takáto situácia v praxi nenastane, keďže skúmame zväčša iba jeden pokus. Máme iba jednu možnosť uskutočniť náhodný výber a z neho potom vypočítať interval spoľahlivosti pre hľadaný parameter.

Na ďalšom obrázku (pozri Obrázok 3.4) môžeme vidieť rovnakú situáciu, pričom sme uvažovali o 99 % konfidencii. V tomto prípade, už všetky intervaly zahŕňali aj skutočnú hodnotu parametra populácie. S rastúcou konfidenciou väčší počet intervalov zahŕňa skutočnú

hodnotu parametra. Tento jav by bol lepšie pozorovateľný pri väčšom počte intervalov. V kódoch programu R stačí zmeniť rozsah iterácie parametra „*i*“ v cykle `for()`, napríklad z 30 na 100 alebo 1000. S rastúcou konfidenciou sa šírka intervalu spravidla (za inak nezmenených podmienok) rozširuje. Preto je dôležité nájsť optimálnu rovnováhu medzi hladinou významnosti a požadovaným rozpätím konfidenčného intervalu. V praxi sa volí štandardne hladina významnosti $\alpha = 0.05$, avšak v určitých, bezpečnosť alebo život ohrozujúcich prípadoch môžeme uvažovať o hladine významnosti ešte menšej, napríklad $\alpha = 0.01$ alebo dokonca až $\alpha = 0.001$.



Obrázok 3.4: Simulované intervaly spoľahlivosti pre strednú hodnotu s 99 % konfidenciou

Zdroj: vlastné spracovanie, výstup zo softvéru R

3.3 Konfidenčné intervaly pre odhad vybraných parametrov

Po osvojení si základných vlastnostní konfidenčných intervalov si v nasledujúcich vzťahoch ukážeme výpočty obojstranných (ako aj jednostranných) konfidenčných intervalov pre najčastejšie odhadované parametre základných súborov. V texte budeme uvádzať postup, pri ktorom vždy začneme s podmienkami, ktoré sa k použitiu daného postupu viažu. Tu je potrebné pripomenúť, že podmienky je nutné overiť. Overenie štatistických podmienok (zväčša tzv. podmienku normality) si ukážeme v niektorých nasledujúcich kapitolách tejto publikácie. V poslednej časti si ukážeme, akým spôsobom urobiť odhad konfidenčných intervalov použitím najjednoduchšej verzie bootstrapingovej metódy.

3.3.1 Konfidenčný interval pre μ ak poznáme σ²

Ide o pomerne zriedkavú situáciu. Aj keď sa to tak môže zdať, znalosť rozptylu populácie σ^2 a priori nepredpokladá aj znalosť μ. Stredná hodnota sa môže meniť a rozptyl môže zostať konštantný. Takáto situácia sa môže vyskytnúť v prípade, ak považujeme rozptyl σ^2 za konštantný, známy z historických alebo iných údajov a mení sa iba charakteristika polohy, ktorú potrebujeme odhadnúť. Napríklad variabilita času vybavenia zákazníkov v banke môže byť konštantná, ale ak je v dôsledku údržby preťažená sieť, softvér reaguje na príkazy pomalšie a mení sa stredná hodnota vybavenia (a zákazníci čakajú dlhšie).

Tabuľka 1: Konfidenčné intervaly pre μ, ak poznáme σ^2

Predpoklady

- Hodnoty populácie majú normálne rozdelenie pravdepodobnosti,
- μ nie je známe, σ^2 je známe,
- μ odhadujeme na základe náhodnej vzorky o veľkosti n.

Obojstranný interval

$$P\left(\bar{x}-\left|z_{\frac{\alpha}{2}}\right|\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{x}+\left|z_{\frac{\alpha}{2}}\right|\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Pravostranný interval

$$P\left(-\infty\le\mu\le\bar{x}+\left|z_{\alpha}\right|\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Ľavostranný interval

$$P\left(\bar{x}-\left|z_{\alpha}\right|\frac{\sigma}{\sqrt{n}}\le\mu\le\infty\right)=1-\alpha$$

$\bar{x}$ – výberový aritmetický priemer, $z_{\alpha/2}$ a z_{α} – je kvantil normovaného normálneho rozdelenia, n – je rozsah výberového súboru, σ – smerodajná odchýlka populácie.

Príklad 3.3

V obchodnom reťazci zaviedli nové pravidlá správania sa predajcov k zákazníkom. Jeden z cieľov, ktoré si vedenie od zmeny sľubovalo, bol kratší čas vybavenia zákazníka, pričom by nemalo dôjsť k zníženiu jeho spokojnosti. Týmto spôsobom sa zvýši flexibilita predajcov, ktorí tak budú mať viac času venovať sa obchodu. Na základe náhodného výberu sa zaznamenali nasledujúce pozorovania (v minútach):

```
> cas <- c(2.1, 3.4, 2.3, 2.1, 1.9, 1.8, 1.6, 1.5, 1.3, 2.2,
  1.3, 1.4, 1.5, 1.8, 1.9, 2.1, 2.6, 1.4, 1.7, 1.5, 2.5, 2, 2,
  1.4)
```

Z minulých meraní vedenie vie, že čas strávený so zákazníkom sa riadi normálnym rozdelením s rozptylom $\sigma^2 = 0.45$. Pred zmenou si bolo s 95 % konfidenciou vedenie isté,

že stredná hodnota času tráveného so zákazníkom bola v intervale od 1.930 do 2.467. Došlo podľa Vás k pozitívnej zmene? Je vhodnejšie použiť jednostranné intervaly?

Pre obojstranný interval je výpočet nasledujúci:

$$P\left(\bar{x}-\left|z_{\frac{\alpha}{2}}\right|\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{x}+\left|z_{\frac{\alpha}{2}}\right|\frac{\sigma}{\sqrt{n}}\right)=P\left(1.88-1.96\frac{0.671}{\sqrt{24}}\le\mu\le1.88+1.96\frac{0.671}{\sqrt{24}}\right)$$

$$P(1.619\le\mu\le2.155)=0.95$$

V programe R si ho môžeme realizovať (manuálne) nasledovne:

```
> spodny <- mean(cas)-abs(qnorm(0.025)) * (0.45/length(cas))^0.5
> spodny
[1] 1.619121
> horny <- mean(cas) + abs(qnorm(0.025))*(0.45/length(cas))^0.5
> horny
[1] 2.155879
```

Hodnotu kvantilu *z* sme si vypočítali pomocou funkcie `qnorm()` a `abs()`. Keďže normované normálne rozdelenie je symetrické rozdelenie okolo strednej hodnoty 0, platí $|z_{(\alpha/2)}| = |z_{(1-\alpha/2)}|$. Porovnaním konfidenčných intervalov vidíme, že zrejme došlo k pozitívnej zmene. Či táto nami nameraná zmena je len dôsledkom náhody, alebo je to nenáhodná (systematická) zmena, môžeme rigorózne overiť pomocou metód štatistického testovania hypotéz, ktoré budú predmetom ďalších častí tejto publikácie.

Viac ako obojstranný interval by vedenie zrejme mal zaujímať pravostranný konfidenčný interval. Ak by predajcovia trávili so zákazníkmi menej času, bolo by to lepšie – v ideálnom prípade žiaden čas pri rovnakej spokojnosti zákazníkov a rovnakom predaji.

$$P\left(-\infty\le\mu\le\bar{x}+z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)=P\left(-\infty\le\mu\le1.888+1.644\frac{0.671}{\sqrt{24}}\right)$$

$$P(-\infty\le\mu\le2.112)=0.95$$

```
> horny <- mean(cas) + abs(qnorm(0.05))*(0.45/length(cas))^0.5
> horny
[1] 2.112731
```

Všimnime si, že dolná hranica intervalu je $-\infty$ čo je vzhľadom na charakter meraného javu stav nemožný. Interval spoľahlivosti je teda iba určitou abstrakciou a v niektorých situáciách môže viesť aj k absurdným výsledkom. V tomto prípade môžeme pri interpretácii zobrať do úvahy, že s pravdepodobnosťou 0.95 očakávame strednú hodnotu času stráveného so zákazníkom v intervale od 0 do 2.112 minút.

3.3.2 Konfidenčný interval pre μ ak nepoznáme σ² a máme dostatočne početný výberový súbor

Oproti predchádzajúcemu prípadu sa môžeme častejšie stretnúť so situáciou, keď máme početnejšiu výberovú vzorku a rozptyl σ^2 nepoznáme. V uvedenej situácii vychádzame z platnosti centrálnej limitnej vety. Všimnime si, že nie je potrebný predpoklad o normalite populácie. Na druhej strane, aby sme sa mohli oprieť o centrálnu limitnú vetu, tak sa vyžaduje veľkosť vzorky aspoň $n \geq 30$. Keďže však nepoznáme σ^2 populácie, je potrebné ho nahradiť výberovým rozptylom (resp. výberovou smerodajnou odchýlkou). Voľba výberového rozptylu nemá pre početnejšie vzorky zásadný vplyv na rozdelenie pravdepodobnosti náhodnej premennej:

$$Z = \frac{\left(\overline{X} - \mu\right)}{\frac{s}{\sqrt{n}}} \qquad (3.21)$$

Z náhodnej premennej Z je možné odvodiť konfidenčný interval. V rôznej literatúre sa môžeme stretnúť s odporúčaním použiť túto variantu výpočtu konfidenčného intervalu, ak je $n \geq 40$ (Montgomery – Runger, 2011), avšak vo väčšine sa uvádza $n \geq 30$ (napr. Tkáč, 2001)[13].

Tabuľka 2: Konfidenčné intervaly pre μ ak nepoznáme σ^2 a $n \geq 40$ ($n \geq 30$)

Predpoklady

- μ nie je známe, σ^2 nie je známe,
- μ odhadujeme na základe náhodnej vzorky o veľkosti $n \geq 40$ ($n \geq 30$).

Obojstranný interval

$$P\left(\bar{x} - \left|z_{\frac{\alpha}{2}}\right|\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + \left|z_{\frac{\alpha}{2}}\right|\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Pravostranný interval

$$P\left(-\infty \leq \mu \leq \bar{x} + \left|z_{\alpha}\right|\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Ľavostranný interval

$$P\left(\bar{x} - \left|z_{\alpha}\right|\frac{s}{\sqrt{n}} \leq \mu \leq \infty\right) = 1 - \alpha$$

$\bar{x}$ – výberový aritmetický priemer, $z_{\alpha/2}$ a z_{α} – je kvantil normovaného normálneho rozdelenia, n – rozsah výberového súboru, s – výberová smerodajná odchýlka.

Zdroj:vlastné spracovanie z použitej literatúry

[13] Dôležitá je pritom konvergencia rozdelenia tejto štatistiky. Taktiež je vhodné ak rozdelenie z ktorého hodnoty súboru pochádzajú sa čo najviac podobá na normálne rozdelenie pravdepodobnosti.

Príklad 3.4

Vychádzajme z rovnakého zadania, ako v predchádzajúcom príklade. Teraz však namiesto 24 pozorovaní máme nasledujúcich 48. Údaje sú uvedené v minútach.

```
cas <- c(2.1, 3.4, 2.3, 2.1, 1.9, 1.8, 1.6, 1.5, 1.3, 2.2, 1.3,
  1.4, 1.5, 1.8, 1.9, 2.1, 2.6, 1.4, 1.7, 1.5, 2.5, 2, 2, 1.4,
  1.8, 1.9, 1.9, 1.9, 2.1, 2.4, 3.8, 3.4, 2.1, 2, 1.9, 1.8, 1.6,
  1.7, 1.8, 1.5, 1.6, 1.7, 3, 1.9, 1.9, 1.7, 2, 1.8)
```

Čo ak sa vplyvom zmeny rozptyl času predsa len zmenil? Z nameraných hodnôt potrebujeme vypočítať výberovú smerodajnú odchýlku a dosadíme do vzťahu pre konfidenčný interval.

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = 0.531$$

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = P\left(1.968 - 1.96\frac{0.531}{\sqrt{48}} \le \mu \le 1.968 + 1.96\frac{0.531}{\sqrt{48}}\right) =$$

$$= P\left(1.818 \le \mu \le 2.119\right) = 0.95$$

$$P\left(0 \le \mu \le \bar{x} + z_{\alpha}\frac{s}{\sqrt{n}}\right) = P\left(0 \le \mu \le 1.968 + 1.644\frac{0.531}{\sqrt{48}}\right) = P\left(0 \le \mu \le 2.094\right) = 0.95$$

```
> spodny <- mean(cas) - abs(qnorm(0.025)) *
  sd(cas)/sqrt(length(cas)); spodny
[1] 1.818357
> horny <- mean(cas) + abs(qnorm(0.025)) *
  sd(cas)/sqrt(length(cas)); horny
[1] 2.119143
> horny <- mean(cas) + abs(qnorm(0.05)) *
  sd(cas)/sqrt(length(cas)); horny
[1] 2.094964
```

3.3.3 Konfidenčný interval pre μ ak nepoznáme σ² a máme málo početný výberový súbor

V predchádzajúcom prípade sme spomenuli možný vplyv použitia výberovej smerodajnej odchýlky na rozdelenie pravdepodobnosti náhodnej premennej Z s tým, že pre $n \geq 40$ (resp. 30) je vplyv použitia výberovej smerodajnej odchýlky malý. To však neplatí ak $n < 40$ (resp. 30). V takom prípade nevieme zaručiť, že náhodná premenná Z bude mať približne normované normálne rozdelenie a je potrebné použiť odlišný postup, pri ktorom využívame Studentovo t rozdelenie pravdepodobnosti. V praxi sa s týmto postupom môžeme

stretnúť pomerne často. Upozorňujeme však na dôležitú podmienku. Ak je početnosť výberového súboru menšia ako 40 (30), musíme predpokladať, že náhodná premenná pochádza z normálneho rozdelenia pravdepodobnosti (prípadne tento predpoklad overiť dodatočným testovaním). Ak použijeme tento postup pre rozsah výberového súboru väčší ako 40 (30), predpoklad o normalite nie je nutný. V ekonometrii je zvykom aj pre súbory o rozsahu väčšom ako 40 (30) používať kvantily zo Studentovho *t* rozdelenia.

Tabuľka 3: Konfidenčné intervaly pre μ ak nepoznáme σ^2 a $n < 40$

Predpoklady

- hodnoty populácie majú normálne rozdelenie pravdepodobnosti,
- μ nie je známe, σ^2 nie je známe,
- μ odhadujeme na základe náhodnej vzorky o veľkosti $n < 40$.

Obojstranný interval

$$P\left(\bar{x} - \left|t_{\frac{\alpha}{2},n-1}\right|\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + \left|t_{\frac{\alpha}{2},n-1}\right|\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Pravostranný interval

$$P\left(-\infty \le \mu \le \bar{x} + \left|t_{\alpha,n-1}\right|\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Ľavostranný interval

$$P\left(\bar{x} - \left|t_{\alpha,n-1}\right|\frac{s}{\sqrt{n}} \le \mu \le \infty\right) = 1-\alpha$$

$\bar{x}$ – výberový aritmetický priemer, $t_{\alpha/2,\ (n\ -\ 1)}$ a $t_{\alpha,\ (n\ -\ 1)}$ – kvantil Studentovho *t* rozdelenia s $n - 1$ stupňami voľnosti, n – rozsah výberového súboru, s – výberová smerodajná odchýlka.

Zdroj: vlastné spracovanie z použitej literatúry

Príklad 3.5

Vráťme sa k prvému zadaniu o konfidenčných intervaloch, kde vedenie získalo údaje s rozsahom 24 pozorovaní. Ide o vzorku menšiu ako 40 a dokonca menšiu ako 30. Podobne ako v prvom prípade predpokladáme, že čas trávený so zákazníkmi sa riadi normálnym rozdelením pravdepodobnosti. Inú informáciu už však nemáme a nevieme, aký je skutočný rozptyl základného súboru. Ak vedenie zaujíma konfidenčný interval pre strednú hodnotu času tráveného so zákazníkmi, použijeme pri tom vzťahy z predchádzajúcej tabuľky (Tabuľka 3).

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = 0.493$$

$$P\left(\bar{x} - \left|t_{\frac{\alpha}{2},n-1}\right|\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + \left|t_{\frac{\alpha}{2},n-1}\right|\frac{s}{\sqrt{n}}\right) = P\left(1.888 - 2.068\frac{0.493}{\sqrt{24}} \le \mu \le 1.888 + 2.068\frac{0.493}{\sqrt{24}}\right) =$$

$$= P(1.679 \le \mu \le 2.095) = 0.95$$

$$P\left(0 \le \mu \le \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right) = P\left(0 \le \mu \le 1.888 + 1.713\frac{0.493}{\sqrt{24}}\right) = P\left(0 \le \mu \le 1.715\right) = 0.95$$

Hodnotu kvantilu Studentovho *t* rozdelenia môžeme získať v programe R pomocou funkcie `qt()` a `abs()`. V príklade vyššie sme pre obojstranné 95 % konfidenčné intervaly mohli použiť nasledovné príkazy:

```
> cas <- c(2.1, 3.4, 2.3, 2.1, 1.9, 1.8, 1.6, 1.5, 1.3, 2.2,
  1.3, 1.4, 1.5, 1.8, 1.9, 2.1, 2.6, 1.4, 1.7, 1.5, 2.5, 2, 2,
  1.4)
> spodny <- mean(cas) - abs(qt(0.05/2, df  = 23)) *
  (var(cas)/24)^0.5
> spodny
[1] 1.679397
> horny <- mean(cas) + abs(qt(0.05/2, df  = 23)) *
  (var(cas)/24)^0.5
> horny
[1] 2.095603
```

Inou možnosťou je použiť priamo funkciu `t.test()`:

```
> t.test(cas)

                    One Sample t-test

data:  cas
t = 18.7627, df = 23, p-value = 1.936e-15
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
1.679397 2.095603
sample estimates:
mean of x
   1.8875
```

V tejto súvislosti je vhodné pripomenúť, že v prípade, ak nevieme zabezpečiť dodržanie podmienky o normalite populácie, je možné výpočet konfidenčných intervalov uskutočniť pomocou neparametrických postupov (napr. bootstrap, prípadne iné metódy, kde sa vytvára konfidenčný interval napr. pre medián). Parametrické postupy využívajú vlastnosti známych rozdelení pravdepodobnosti a vo všeobecnosti platí, že ak ich môžeme oprávnene použiť, je vhodné ich uprednostniť pred neparametrickými metódami. Neskôr budeme prezentovať použitie bootstrapingovej metódy, ktorá sa využíva v situáciách, kde nepoznáme analytický tvar rozdelenia pravdepodobnosti nami odhadovaného parametra, prípadne nevieme zaručiť splnenie podmienok parametrických metód.

3.3.4 Konfidenčný interval pre σ^2

Často je cieľom manažérskych rozhodnutí znížiť v procesoch mieru variability (zníženie variability doby vybavenia zákazníkov, vymáhania pohľadávok, splatenia záväzkov, vybavenia žiadosti na poskytnutie úveru, atď.). Z tohto dôvodu nás môže zaujímať, v akom konfidenčnom intervale môžeme očakávať variabilitu procesov. Variabilitu pritom meriame spravidla rozptylom, resp. smerodajnou odchýlkou. Pre výpočet konfidenčného intervalu pre rozptyl využívame známe χ^2 rozdelenie pravdepodobnosti.

Tabuľka 4: Konfidenčné intervaly pre σ^2

<table>
<tr><td colspan="2">Predpoklady
• hodnoty populácie majú normálne rozdelenie pravdepodobnosti,
• μ nie je známe, σ^2 nie je známe,
• σ^2 odhadujeme na základe náhodnej vzorky.</td></tr>
<tr><td rowspan="2">Obojstranný interval
$P\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right) = 1-\alpha$</td><td>Pravostranný interval
$P\left(0 \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha,n-1}}\right) = 1-\alpha$</td></tr>
<tr><td>Ľavostranný interval
$P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \le \sigma^2 \le \infty\right) = 1-\alpha$</td></tr>
<tr><td colspan="2">$\chi^2_{1-\alpha/2,(n-1)}$ a $\chi^2_{1-\alpha,(n-1)}$ – kvantil χ^2 rozdelenia s $n-1$ stupňami voľnosti, n – rozsah výberového súboru, s – výberová smerodajná odchýlka.</td></tr>
</table>

Zdroj: vlastné spracovanie z použitej literatúry

Jednoduchým odmocnením výrazov v uvedenej tabuľke dostaneme konfidenčné intervaly pre smerodajnú odchýlku.

Z údajov v predchádzajúcom príklade sme vypočítali bodový odhad rozptylu ako:

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 0.242$$

Dosadením do vzťahu pre obojstranný a pravostranný konfidenčný interval si vypočítame hranice intervalov. Vedenie tak získa informáciu o miere variability času, ktorý trávi predajca so zákazníkom. Menšia variabilita spravidla znamená väčšiu mieru istoty, a teda aj lepšie plánovanie.

$$P\left(\frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi_{\frac{\alpha}{2},n-1}}\right) = P\left(\frac{(24-1)0.242}{38.076} \le \sigma^2 \le \frac{(24-1)0.242}{11.689}\right) =$$

$$= P\left(0.146 \le \sigma^2 \le 0.477\right) = 0.95$$

$$P\left(0 \le \sigma^2 \le \frac{(n-1)s^2}{\chi_{\alpha,n-1}}\right) = P\left(0 \le \sigma^2 \le \frac{(24-1)0.242}{13.09051}\right) = P\left(0 \le \sigma^2 \le 0.425\right) = 0.95$$

Výpočet kvantilu χ^2 rozdelenia je možný v programe R pomocou funkcie `qchisq()`, napr.:

```
> qchisq(0.025, df = 23)
[1] 11.68855
```

3.3.5 Konfidenčný interval pre populačný podiel

Častým parametrom záujmu je podiel, π. Môže nás zaujímať, aký podiel zákazníkov: je bezdetných, je ženského pohlavia, je z vyššej príjmovej skupiny, už si náš výrobok raz kúpilo a pod. Výberové rozdelenie podielu $\hat{p}$ má približne normálne rozdelenie so strednou hodnotou π a variabilitou $\pi\,(1-\pi)/n$. Táto aproximácia platí, ak $n\pi \ge 5$ a zároveň $\pi(1-p) \ge 5$. Túto informáciu využívame pri tvorbe konfidenčných intervalov. Keďže pri tvorbe konfidenčného intervalu vychádzame z výberového súboru, hodnotu parametra π nahrádzame výberovým podielom $\hat{p}$. Všimnime si, že dolná hranica konfidenčných intervalov by nemala byť menšia ako 0.

Tabuľka 5: Konfidenčné intervaly pre podiel π

Predpoklady

- $n\hat{p} \ge 5$ a zároveň $n(1-\hat{p}) \ge 5$
- π odhadujeme na základe náhodnej vzorky.

Obojstranný interval

$$P\left(\hat{p} - \left|z_{\alpha/2}\right|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \pi \le \hat{p} + \left|z_{\alpha/2}\right|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

Pravostranný interval

$$P\left(0 \le \pi \le \hat{p} + \left|z_{\alpha}\right|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

Ľavostranný interval

$$P\left(\hat{p} - \left|z_{\alpha}\right|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \pi \le \infty\right) = 1-\alpha$$

$\hat{p}$ – výberový podiel, $z_{\alpha/2}$ a z_{α} – kvantil normovaného normálneho rozdelenia, n – rozsah výberového súboru.

Zdroj: vlastné spracovanie z použitej literatúry

Príklad 3.6

Jedným zo všeobecných ukazovateľov lojality zákazníkov je počet nákupov jedného zákazníka u tej istej značky, v tom istom obchode, u toho istého predajcu, a pod. Je to zároveň jeden z kritických obchodných ukazovateľov v automobilovom priemysle. Je známe, že automobilky majú pomerne presnú predstavu o lojalite existujúcich zákazníkov. Jednou z alternatív, ako získať údaje u konkurenčných značiek, je urobiť prieskum. V jednom z týchto prieskumov predajca automobilov zisťoval, či majitelia nových áut (iných ako jeho značky) kupovali nové auto od aktuálnej značky prvýkrát alebo opakovane. Na základe náhodnej vzorky o veľkosti $n = 94$ predajca zistil, že približne 33 % (31) majiteľov nových áut si kúpilo takú istú značku akú vlastnilo predtým. U zákazníkov predajcu je tento parameter približne 40 %. Zdá sa teda, že lojalita zákazníkov u predajcu je vyššia ako u konkurencie. Na druhej strane, hodnota 33 % je náhodná premenná, keďže závisí od náhodného výberu. Ak by sme vypočítali interval spoľahlivosti, mohli by sme s určitou konfidenciou zistiť, v akom intervale môžeme očakávať hodnotu podielu v celkovej populácii.

$$P\left(\hat{p}-\left|z_{\alpha/2}\right|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \pi \le \hat{p}+\left|z_{\alpha/2}\right|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)=$$

$$=P\left(0.33-1.959\sqrt{\frac{0.33(1-0.33)}{94}} \le \pi \le 0.33+1.959\sqrt{\frac{0.33(1-0.33)}{94}}\right)=$$

$$=P(0.23 \le \pi \le 0.42)=0.95$$

Môžeme vidieť, že skutočný podiel je s 95 % konfidenciou v intervale 0.23 až 0.42. Tento výsledok naznačuje, že lojalita zákazníkov predajcu na úrovni 40 % (0.4) zrejme nie je taká výnimočná, keďže lojalita u konkurencie mohla byť aj viac ako 40 %.

V programe R môžeme použiť na výpočet konfidenčného intervalu nasledujúci príkaz `prop.test()`. Výsledky sú veľmi podobné, avšak nie totožné. Dôvodom je používanie mierne odlišnej metódy výpočtu. Rozdiely však nebývajú výrazné (bližšie pozri Newcombe, 1998). Niektoré ďalšie metódy je možné nájsť v programovom balíku `binom`.

```
> prop.test(31, 94, correct = F)

    1-sample proportions test without continuity correction

data:  31 out of 94, null probability 0.5
X-squared = 10.8936, df = 1, p-value = 0.000965
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:
0.2430749 0.4298653
sample estimates:
        p
0.3297872
```

3.4 Konfidenčné intervaly pomocou Bootstrapingu

V tejto časti si zopakujeme základný princíp bootstrapingu a ukážeme si jeho použitie pri skúmaní vlastností štatistík. Spravidla nás zaujímajú vlastnosti rozdelenia pravdepodobnosti skúmaných štatistík. Môže nás napríklad zaujímať, ako vyzerá rozdelenie pravdepodobnosti aritmetického priemeru. Toto rozdelenie je známe a spravidla nie je nutné ho odhadovať inými metódami. Čo však v prípade štatistík ako sú medzikvartilové rozpätie alebo priemerná absolútna odchýlka? Bootstraping je metóda, ktorá nám umožní poznať vlastnosti našich štatistík, napríklad akú majú variabilitu, prípadne priamo, aký majú tvar rozdelenia pravdepodobnosti. Ide o metódu, ktorá na rozdiel od tradičných parametrických metód induktívnej štatistiky vychádza skôr zo simulácií hodnôt ako z teórie. Prax však ukázala, že výsledky, ku ktorým sa pomocou bootstrapingu vieme dopracovať, sú veľmi podobné tým, k akým sa dopracujeme pomocou často náročnejších metód induktívnej štatistiky (niekedy sa tomu hovorí asymptotická štatistika). V prípade malých vzoriek vieme pomocou bootstrapingu dokonca získať presnejšie výsledky, aké získame pomocou induktívnej štatistiky. Keďže v tejto publikácii pracujeme s programom R, bootstraping je prirodzenou voľbou, ktorá sa v programe R implementuje pomerne jednoducho. Pri spracovaní bootstrapingovej metódy si do značnej miery pomôžeme myšlienkami a postupmi prezentovanými v prácach Cheng (2006), El-Shaarawi – Piegorsch (2002) a Efron – Tibshirani (1994).

Určitá štatistika je náhodná premenná a tak nás prirodzene zaujíma, aký má tvar rozdelenia. Bootstraping na to dáva odpoveď nasledujúcim spôsobom. Vytvoríme pomerne veľký počet nezávislých súborov dát z rovnakého empirického rozdelenia pravdepodobnosti. Pre každý z týchto súborov si vypočítame hľadanú štatistiku. Dostávame tak jeden početný súbor štatistík. Z toho vytvorené empirické rozdelenie pravdepodobnosti (prípadne empirická distribučná funkcia) odpovedá na nami hľadanú otázku. Všimnime si, že pri odvodení rozdelenia pravdepodobnosti pre priemer sme sa v predošlých častiach opierali o centrálnu limitnú vetu a zákon veľkých čísel. Pri bootstrapingu to nie je potrebné. Pri niektorých štatistikách, ktoré sú pomerne komplikované tak nemusíme prechádzať náročnými

matematicko-štatistickými metódami, aby sme sa dozvedeli niečo o rozdelení štatistiky[14]. Jednoducho si ju nasimulujeme. V tejto kapitole nás budú zaujímať rozdelenia nasledujúcich štatistík: aritmetický priemer, výberový rozptyl, medián, podiel, šikmosť, špicatosť a medzikvartilové rozpätie.

Zrejme najjednoduchšou metódou počítania konfidenčných intervalov pomocou bootstrapingu je **kvantilová metóda**. Majme empirický súbor *iid* hodnôt X_i, $i = 1, 2, ..., n$. Parameter, ktorý je predmetom nášho záujmu si označme ako θ, pričom je dôležité, aby hodnota tohto parametra závisela od hodnôt v populácii, t. j. aby bola funkciou hodnôt populácie, čo môžeme zapísať ako $\theta = f(x)$, kde x je vektor hodnôt populácie. Keďže máme k dispozícii iba *iid* vzorku X_i, tento parameter θ odhadujeme pomocou $\theta^e = f(X)$, kde index e označuje odhad (z angl. *estimate*) a X je vektor hodnôt empirického súboru.

Pre lepšiu ilustráciu, nech nás zaujíma odhad strednej hodnoty μ. Na odhad strednej hodnoty použijeme aritmetický priemer. Ten, ako z definície aritmetického priemeru vyplýva, sa vypočíta z hodnôt empirického súboru. Našou snahou je pomocou hodnôt empirického súboru odhadnúť distribučnú funkciu odhadu parametra θ, ktorú si môžeme zapísať ako F_θ. Túto distribučnú funkciu odhadujeme kumulatívnou distribučnou funkciou F_θ^e. Prvým krokom pri bootstrapingu je tvorba bootstrap vzoriek. Bootstrap vzorka sa vytvorí tak, že z pozorovaní sa náhodne vyberie n hodnôt so spätným vrátením už raz vybraných štatistických jednotiek. Tieto vzorky si môžeme označiť ako X_j, $j = 1, 2, ..., m$, kde m je počet bootstrap vzoriek. Pre každú z týchto vzoriek si vypočítame odhadovaný parameter θ^e_j. Dostávame tak súbor odhadovaných parametrov o rozsahu m. Kumulatívna distribučná funkcia F_θ^e vytvorená z tohto súboru sa použije na odhad distribučnej funkcie odhadu parametra θ (teda F_θ). Kvantilová metóda na výpočet $(1 - \alpha)$ obojstranného konfidenčného intervalu potom jednoducho vyberie $\alpha/2$ a $(1 - \alpha/2)$ percentil z F_θ^e. Niekedy sa z F_θ^e odhaduje spojité rozdelenie F_θ, z ktorého sa počítajú príslušné kvantily. To by vyžadovalo odhad spojitého rozdelenia pravdepodobnosti. Zrejme jednoduchší spôsob je zabezpečiť dostatočný počet bootstrap vzoriek, povedzme aspoň 1000 (závisí mimo iného od náročnosti počítania odhadovaného parametra θ^e, prípadne od numerickej stability výsledku) a potom neparametrickým spôsobom vybrať príslušný kvantil. Ak nás zaujíma 95 % konfidenčný interval, pri $m = 1000$ po zostavení variačného radu $\theta^e_{(j)}$ jednoducho vyberieme 25-tu a 975-tu hodnotu.

[14] Nie vždy je použitie bootstrapingu tak jednoduché, ako sa to prezentuje v týchto častiach publikácie.

Táto kvantilová metóda sa dá ďalej rozšíriť niekoľkými smermi. Cieľom týchto úprav je zlepšiť štatistické vlastnosti kvantilovej metódy. Druhú metódu, ktorú si predstavíme, označíme ako **BCA metódu** (z angl. ***B**ias **C**orrected **A**ccelerated bootstrap*). Pre výpočet obojstranného konfidenčného intervalu budeme potrebovať nasledujúce vzťahy, ktoré predstavujú príslušné kvantily, ktoré sa z F_θ^e vyberajú:

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{\alpha}\left(\hat{z}_0 + z^{(\alpha)}\right)} \right) \qquad (3.22)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{\alpha}\left(\hat{z}_0 + z^{(1-\alpha)}\right)} \right) \qquad (3.23)$$

kde $\Phi(.)$ je kumulatívna distribučná funkcia normovaného normálneho rozdelenia, $z^{(\alpha)}$ je 100α percentil normovaného normálneho rozdelenia. Napríklad pre $\alpha = 0.05$, ide o 5-ty percentil, t. j. $z^{(5)} = -1.645$ a $\Phi(-1.645) = 0.05$. Pre vzťahy *(3.22)* a *(3.23)* je potrebné vypočítať dva parametre: $\hat{\alpha}$ a $\hat{z}_0$. Pre úplnosť zopakujeme, že v programe R na výpočet hodnôt kumulatívnej distribučnej funkcie normovaného normálneho rozdelenia vieme použiť funkciu `pnorm()`.

```
> pnorm(-1.645)
[1] 0.04998491
```

Nech θ^e je odhad parametra θ vypočítaný na základe pôvodnej vzorky údajov (nie bootstrap vzorky). Potom:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\theta_j^e < \theta^e\right\}}{m} \right) \qquad (3.24)$$

kde vo vzťahu *(3.24)* symbolom # označujeme počet odhadovaných parametrov, ktoré spĺňajú podmienku uvedenú v {}, čiže počet parametrov θ^e_j, ktoré sú menšie ako je pôvodný odhad θ^e. Vydelením číslom m dostaneme podiel parametrov menších ako pôvodný odhad. Vzťah *(3.24)* slúži pre tzv. korekciu skreslenia. Ak je tento podiel 0.5, potom $\Phi^{-1}(0.5) = 0$, ak by bol tento podiel 0.95, dostali by sme sa k hodnote $\Phi^{-1}(0.95) = 1.645$.

```
> qnorm(0.95)
[1] 1.644854
```

Výpočet tzv. akceleračného parametra $\hat{\alpha}$ je o čosi komplikovanejší. Potrebujeme si k tomu zadefinovať ďalšiu metódu vzorkovania (podobne ako bootstrap), ktorá sa nazýva **jackknife**. Majme tak ako na začiatku empirický súbor *iid* hodnôt X_i, $i = 1, 2, ..., n$, pričom

$X = \{X_i,\ i = 1, 2, ..., n\}$. Ďalej nazveme X^{-i} taký súbor hodnôt, kde budú všetky hodnoty súboru X okrem i-tej hodnoty. Následne odhadnime parameter θ zo súboru X^{-i} a označme si príslušný odhad ako $\theta^e_{-i} = f(X^{-i})$. Ak postupne odstránime vždy jednu hodnotu, dostaneme tak spolu n odhadov θ^e_{-i}. Vypočítame si ich priemer:

$$\bar{\theta}^e_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\theta^e_{-i} \qquad (3.25)$$

Potom parameter $\hat{\alpha}$ odhadneme ako:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}\left(\bar{\theta}^e_{(.)} - \theta^e_{-i}\right)^3}{6\left(\sum_{i=1}^{n}\left(\bar{\theta}^e_{(.)} - \theta^e_{-i}\right)^2\right)^{3/2}} \qquad (3.26)$$

Aplikáciu bootstrapingovej metódy si ukážeme na jednoduchom príklade spolu s kódmi v programe R. Použijeme premennú `hp` z databázy `mtcars` z programového balíka `datasets`. Premenná `hp` predstavuje výkon motora s tým, že vyššia hodnota znamená vyšší výkon motora. Spolu máme k dispozícii $n = 32$ výkonov motora. Nejde o veľmi veľkú vzorku, čím sa použitie bootstrapingu pri výpočte konfidenčných intervalov stáva vhodnou alternatívou. Vypočítame si 95 % konfidenčné intervaly pre už vyššie spomínané charakteristiky: priemer, výberový rozptyl, medián, šikmosť, špicatosť a medzi-kvartilové rozpätie. Najprv si však musíme zadefinovať funkcie pre výpočet výberovej šikmosti a špicatosti tak, ako už bolo uvedené vo vzťahoch *(3.14)* a *(3.15)*.

```
> sample_skew <- function(data) {
+ L <- length(data)
+ sample_skew <- (L/((L - 1)*(L - 2))) * sum(((data -
  mean(data))/sd(data))^3)
+ return(sample_skew)
+ }
------------------------------------------------------------------
> sample_kurt <- function(data) {
+ L <- length(data)
+ sample_kurt <- ((L*(L + 1))/((L - 1)*(L - 2)*(L - 3))) *
  sum(((data - mean(data))/sd(data))^4) - (3*(L - 1)^2)/((L -
  2)*(L - 3))
+ return(sample_kurt)
+ }
```

Ďalej si nadefinujeme databázu a objekty, ktoré budeme potrebovať v ďalšom kóde:

```
> attach(mtcars); names(mtcars); priemer <- c(); rozptyl <- c();
  median <- c(); sikmost <- c(); spicatost <- c();
  medzikvartilove_rozpatie <- c(); podiel <- c()
> L <- length(data)
```

```
> B <- 1000
```

V nasledujúcom cykle dochádza k tvorbe bootstrap vzorky a následne k počítaniu odhadovaných štatistík a ich zapisovanie do príslušných objektov:

```
> for (j in 1:B) {
+ x <- sample(hp, size = length(hp), replace = T)
+ priemer <- c(priemer, mean(x))
+ rozptyl <- c(rozptyl, var(x))
+ median <- c(median, median(x))
+ sikmost <- c(sikmost, sample_skew(x))
+ spicatost <- c(spicatost, sample_kurt(x))
+ medzikvartilove_rozpatie <- c(medzikvartilove_rozpatie,
  IQR(x))
+ podiel <- c(podiel, sum(x<100)/L)
+ }
```

Výpočet príslušného konfidenčného intervalu kvantilovou metódou je potom napríklad:

```
> # 95 % Konfidenčný interval pre priemer - spodná a horná
  hranica
> sort(priemer, decreasing = F)[ceiling(B*0.025)]; mean(hp);
  sort(priemer, decreasing = F)[ceiling(B*0.975)]
[1] 124.6562
[1] 146.6875
[1] 169.375
------------------------------------------------------------------
> # 95 % Konfidenčný interval pre rozptyl - spodná a horná
  hranica
> sort(rozptyl, decreasing = F)[ceiling(B*0.025)]; var(hp);
  sort(rozptyl, decreasing = F)[ceiling(B*0.975)]
[1] 2533.66
[1] 4700.867
[1] 6920.862
```

Pre úplnosť všetky výsledky uvádzame v nasledujúcej tabuľke[15]. Pre tie isté dáta si taktiež ukážeme použitie BCA metódy. Až po vytvorenie bootstrap vzoriek je postup rovnaký.

[15] Samozrejme, ide o simulácie, takže ak si to čitateľ zopakuje, výsledky budú odlišné, ale veľkosť „odlišnosti“ bude pomerne malá.

Tabuľka 6: Konfidenčné intervaly pomocou bootstrapingovej kvantilovej metódy

Parameter	Dolná hranica	Bodový odhad	Horná hranica
Priemer	124.65	146.69	169.38
Rozptyl	2533.66	4700.87	6920.86
Medián	109.50	123.00	175.00
Šikmosť	-0.01	0.80	1.45
Špicatosť	-1.31	0.28	2.37
Medzi-kvartilové rozpätie	60.75	83.50	134.00

Zdroj: vlastné spracovanie

```
> boot <- function(data, B = 1000, FUN) {
+ boot_statistics <<- c()
+ for (i in 1:B) {
+ x <- sample(data, size = length(data), replace = T)
+ boot_statistics <<- c(boot_statistics, FUN(x))
+ }
+ }
-----------------------------------------------------------------
# funkcia na tvorbu jackknife vzoriek, na ktorých sa použije
  vhodná funkcia "FUN"
-----------------------------------------------------------------
> jack <- function(data, FUN) {
+ jack_statistics <<- c()
+ for (i in 1:length(data)) {
+ x <- hp[-i]
+ jack_statistics <<- c(jack_statistics, FUN(x))
+ }
+ }
-----------------------------------------------------------------
# funkcia pre podiel
-----------------------------------------------------------------
> prop <- function(data, lower = T, treshold = 0.5) {
+ if (lower == T)
+ a <- sum(data<treshold)/length(data)
+ else
+ a <- sum(data>=treshold)/length(data)
+ print(a)
+ }
-----------------------------------------------------------------
# samotná funkcia pre konfidenčné intervaly BCA.
-----------------------------------------------------------------
> bca_ci <- function(data, B = 1000, FUN, alpha = 0.05) {
+ boot(data, B, FUN)
+ jack(data, FUN)
+ # parameter skreslenia z_hat
+ z_hat =
  qnorm(sum(boot_statistics<FUN(data))/length(boot_statistics))
+ mean_jack <- mean(jack_statistics)
+ # akceleracny parameter a_hat
+ if (sum((jack_statistics - mean_jack))==0) a_hat = 0
+ else a_hat <- sum((jack_statistics - mean_jack)^3) /
  (6*(sum((jack_statistics - mean_jack)^2))^(3/2))
+ # kvantily BCA
```

```
+ lower <- floor(B*pnorm(z_hat + (z_hat + qnorm(alpha/2)) / (1 -
  a_hat*(z_hat + qnorm(alpha/2)))))
+ upper <- ceiling(B*pnorm(z_hat + (z_hat + qnorm(1-alpha/2)) /
  (1 - a_hat*(z_hat + qnorm(1-alpha/2)))))
+ sorted <- sort(boot_statistics, decreasing = F)
+ print(paste("spodna hranica", sorted[lower], "bodovy odhad",
  FUN(data), "horna hranica", sorted[upper]))
+ }
```

Následne je možné vypočítať všetky predošlé konfidenčné intervaly použitím funkcie `bci_ci()`, napríklad:

```
> bca_ci(hp, B = 1000, mean, alpha = 0.05)
```

Tabuľka 7 uvádza konfidenčné intervaly vypočítané pomocou metódy BCA. Všimnime si, že akceleračný parameter v niektorých situáciách nemusí byť definovaný. Takáto situácia vznikla aj v našom ukážkovom príklade pri počítaní konfidenčného intervalu pre medián. Vo vzorkách vytvorených pomocou metódy jackknife bola hodnota mediánu vždy rovnaká, preto menovateľ vo vzťahu pre výpočet akceleračného parametra je rovný 0. V takomto prípade sa akceleračný parameter zvolí ako rovný 0. Táto zmena je zohľadnená aj v kóde vyššie podmienkou `if()`.

Tieto príkazy boli vytvorené pre obojstranné intervaly. Jednoduchými úpravami je možné tieto kódy použiť aj na tvorbu funkcií pre jednostranné intervaly. Pre úplnosť uvádzame, že v prípade metódy BCA je vhodné použiť čo najväčší počet boostrapingových vzoriek (to platí aj vo všeobecnosti). Efron – Tibshirani (1994) odporúčajú viac ako 1000 vzoriek. Existujú aj iné metódy ako kvantilová a BCA metóda, bližšie pozri Efron – Tibshirani (1994) alebo El-Shaarawi – Piegorsch (2002).

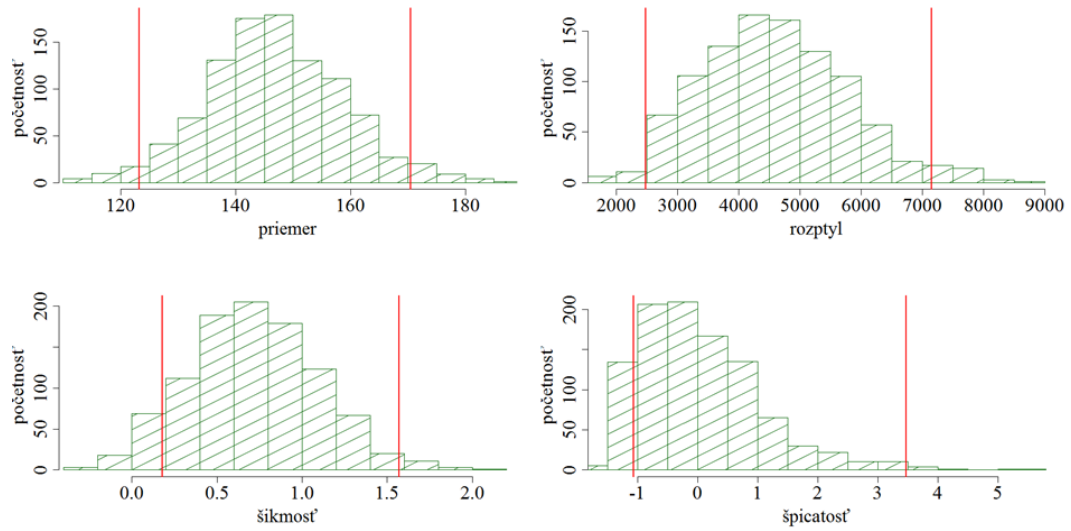
Tabuľka 7: Konfidenčné intervaly pomocou bootstrapingovej BCA metódy

Parameter	Dolná hranica	Bodový odhad	Horná hranica
Priemer	123.22	146.69	170.44
Rozptyl	2480.52	4700.87	7152.44
Medián	105	123	175
Šikmosť	0.18	0.80	1.57
Špicatosť	-1.07	0.28	3.47
Medzi-kvartilové rozpätie	49.75	83.50	121.25

Zdroj: vlastné spracovanie, výstup zo softvéru R

Nasledujúci histogram (pozri Obrázok 3.5) zobrazuje rozdelenia početností priemeru, rozptylu, šikmosti a špicatosti z boostrapových vzoriek. Červenou farbou sú označené hranice konfidenčných intervalov, vypočítané pomocou BCA metódy.

```
> par(mfrow = c(2, 2))
> boot(hp, B = 1000, mean)
> hist(boot_statistics, xlim = c(min(boot_statistics),
  max(boot_statistics)), xlab = "priemer", ylab = "početnosť",
  family = "serif", main = NA, cex.axis = 2, cex.lab = 2,
  density = 10, col = "darkgreen")
> abline(v = c(quantile(boot_statistics, p = 0.025),
  quantile(boot_statistics, p = 0.975)), col = "red", lwd = 2)
 ------------------------------------------------------------
> boot(hp, B = 1000, var)
> hist(boot_statistics, xlim = c(min(boot_statistics),
  max(boot_statistics)), xlab = "rozptyl", ylab = "početnosť",
 family = "serif", main = NA, cex.axis = 2, cex.lab = 2, density
  = 10, col = "darkgreen")
> abline(v = c(quantile(boot_statistics, p = 0.025),
  quantile(boot_statistics, p = 0.975)), col = "red", lwd = 2)
 ------------------------------------------------------------
> boot(hp, B = 1000, sample_skew)
> hist(boot_statistics, xlim = c(min(boot_statistics),
  max(boot_statistics)), xlab = "šikmosť", ylab = "početnosť",
  family = "serif", main = NA, cex.axis = 2, cex.lab = 2,
  density = 10, col = "darkgreen")
> abline(v = c(quantile(boot_statistics, p = 0.025),
  quantile(boot_statistics, p = 0.975)), col = "red", lwd = 2)
 ------------------------------------------------------------
> boot(hp, B = 1000, sample_kurt)
> hist(boot_statistics, xlim = c(min(boot_statistics),
  max(boot_statistics)), xlab = "špicatosť", ylab = "početnosť",
  family = "serif", main = NA, cex.axis = 2, cex.lab = 2,
  density = 10, col = "darkgreen")
> abline(v = c(quantile(boot_statistics, p = 0.025),
  quantile(boot_statistics, p = 0.975)), col = "red", lwd = 2)
```



Obrázok 3.5: Bootstrap štatistiky s hranicami 95 % konfidenčného intervalu

Zdroj: vlastné spracovanie, výstup zo softvéru R

4 Testovanie štatistických hypotéz

Tvrdenie o jednom alebo viacerých parametroch jednej alebo viacerých populácií budeme nazývať **štatistickou hypotézou** (Montgomery – Runger, 2011). Proces, pomocou ktorého zisťujeme, či môžeme zamietnuť štatistickú hypotézu (teda tvrdiť, že neveríme, že hypotéza platí), nazývame **testovaním štatistických hypotéz**. Pomocou testovania štatistických hypotéz môžeme rozlíšiť, či nesúlad medzi tvrdením v hypotéze a výsledkami pozorovaného javu je **náhodný** alebo **systematický**. V jednom z predchádzajúcich prípadov sme pri výpočte konfidenčného intervalu pre podiel zistili, že až 40 % zákazníkov predajcu je lojálnych hoci u konkurencie je to len 33 %. Rozdiel by mohol byť spôsobený náhodou (náhodným výberom), keďže horná hranica intervalu spoľahlivosti pre skutočný podiel lojality u konkurencie bola až 42 %. Už na tomto mieste sme naznačili princíp testovania štatistických hypotéz, ktorý vychádza zo skutočnosti, že náš bodový odhad je náhodnou premennou.

Príklad 4.1
Objem odpadu v meste sa za posledný týždeň zvýšil. Môžeme vychádzajúc z údajov, ktoré máme k dispozícii o objeme odpadu za posledné roky, tento nárast považovať za náhodný? Alebo ide o **systematický** nárast a musíme uvažovať o nápravných opatreniach?

Jednou z nosných myšlienok Deminga (2000) bolo, že neraz sa manažéri správajú k náhodným výkyvom tak, ako keby išlo o výkyvy systematické. Investujú množstvo ekonomických zdrojov na vyriešenie problémov, ktoré v skutočnosti neexistujú. Variabilita je a bude existovať. Raz bude výskyt chýb vo výrobe vyšší a inokedy nižší. Je však potrebné rozlíšiť medzi náhodnou variabilitou a systematickou variabilitou. Rovnako je pre manažéra hrozbou, ak sa k systematickej variabilite správa, ako keby išlo o variabilitu náhodnú (neurobí opatrenia tam, kde ich je treba). Testovanie štatistických hypotéz umožňuje rozlíšiť náhodnú variabilitu od systematickej a tým napomáha manažérom uskutočňovať lepšie rozhodnutia.

4.1 Formulácia štatistických hypotéz

Vo všeobecnosti rozoznávame **základnú** alebo **nulovú hypotézu** a tzv. **alternatívnu hypotézu**. Základnú hypotézu formálne označujeme ako H_0. Cieľom je pokúsiť sa túto hypotézu spochybniť. Alternatívna hypotéza (teda to voči čomu porovnávame základnú hypotézu) predstavuje často presný opak k základnej hypotéze (formálne však nie je vždy jej

negáciou) a formálne sa označuje ako H_1. Tvrdenie „stredná hodnota výšky basketbalistov v univerzitnej lige **je** 192 *cm*“ je nulovou hypotézou. Jedna z možných alternatívnych hypotéz je, že stredná hodnota výšky basketbalistov v univerzitnej lige **nie je** 192 *cm*. Uvedené hypotézy by sme formálne mohli zapísať ako:

$$H_0: \mu = 192\ cm$$
$$H_1: \mu \neq 192\ cm \qquad (4.1)$$

Z populácie basketbalistov hrajúcich v univerzitnej lige náhodne vyberieme 10 hráčov a na základe tejto vzorky chceme overiť nulovú hypotézu. Priemerná výška týchto 10-tich náhodne vybraných basketbalistov nech je 183 *cm*. Z kapitoly o bodovom odhade vieme, že na odhad strednej hodnoty populácie je výberový priemer $\bar{x}$ našim najlepším „nástrojom“ – štatistikou. Tak je potrebné hodnotu 183 *cm* aj interpretovať. Vychádzajúc z výberového súboru odhadujeme, že stredná hodnota populácie je 183 *cm*. Vráťme sa teraz naspäť k hypotéze. Môže sa stať, že rozdiel medzi 192 *cm* – 183 *cm* = 9 *cm* budeme považovať za tak veľký, že hypotézu H_0 zamietneme. Všimnime si, že v takom prípade **netvrdíme**, že skutočná stredná hodnota výšky populácie je menej ako 192 *cm*. Čisto z technického hľadiska alternatívna hypotéza hovorí, že stredná hodnota sa nerovná 192, a teda výška basketbalistov v populácii môže byť vyššia alebo nižšia. Pri takto formulovaných hypotézach hovoríme o **obojstrannej hypotéze**. Poznáme aj jednostranné štatistické hypotézy. Ak by sme uvažovali o jednostrannej hypotéze, tak v našom prípade by jej tvar bol $H_0: \mu \geq 192\ cm$ alebo $H_0: \mu \leq 192\ cm$, kde alternatívnou hypotézou by bolo $H_1: \mu < 192\ cm$, resp. $H_1: \mu > 192\ cm$.

Štatistické hypotézy by mali byť naformulované tak, aby boli kvantifikovateľné, overiteľné a mali štatistický charakter (Good – Hardin, 2009). Ako príklad nevhodnej štatistickej hypotézy uvedieme: „*niektorí naši najbohatší zákazníci sú nevďační*“. Výrazy ako „*bohatí*“ a „*nevďační*“ je potrebné exaktne zadefinovať tak, aby sme tieto štatistické znaky mohli merať. O to dôležitejšie je zadefinovať výraz „*niektorí*“. Ide o väčšinu, tretinu, pätinu, jednu tisícinu, ..., zákazníkov? Prečo práve niektorí? Neraz sa v hypotézach vyskytujú výrazy ako: „*nie všetky*“, „*iba niektoré*“, ktoré tiež nemajú štatistický charakter. Uvedená hypotéza by mohla byť testovateľná ak[16]:

- „*niektorí naši najbohatší*“ by znamenalo 25 % zákazníkov s najvyšším disponibilným príjmom,
- „*nevďační*“ by znamenalo, že našej hotelovej obsluhe platia menšie prepitné ako zvyšných 75 % zákazníkov.

[16] V oboch prípadoch by konštanty bolo vhodné zdôvodniť.

Ak by sme mali takto zadefinované pojmy, mohli by sme z uvedeného výroku naformulovať nasledovnú hypotézu: „*najbohatší zákazníci platia menšie prepitné našej obsluhe ako ostatní zákazníci*“. Na záver tejto podkapitoly si uvedieme jednu zaujímavú myšlienkovú fikciu z knihy Coolican (1999), ktorá predstavuje rozhovor dvoch osôb.

Tabuľka 8: Preklad neformálneho textu do jazyka štatistických hypotéz

Každodenné zmýšľanie	***Formálne zmýšľanie vo výskume***
A: *Na tomto pracovisku nemajú ženy šancu byť povýšené. Z posledných štyroch pohovorov vybrali vždy muža, a to sa zakaždým na pozíciu uchádzali dve ženy a dvaja muži.* **B:** *To fakt? Zistime, koľko mužov by mali vybrať, aby si sa mýlil!* **A:** *Ako to myslíš?* **B:** *Medzi kandidátmi bol rovnaký počet žien aj mužov. A teda na danej pozícii by mal byť rovnaký počet mužov ako aj žien. To znamená dve!* **A:** *Presne na to som myslel. Na tej pozícii by mali byť z tých štyroch pohovorov aspoň dve ženy.* **B:** *Nuž presnejšie iba dve, pokiaľ neberieme do úvahy zvýhodňovanie mužov v minulosti. Otázkou ostáva, či je nula zo štyroch pohovorov dostatočne veľký rozdiel oproti dvom zo štyroch, aby sme mohli tvrdiť, že výberová komisia je zaujatá.*	**Predmet záujmu:** *Do povýšenia sú vyberaní viac muži.* **Tvorba základnej (nulovej) hypotézy:** *Čo by sa muselo stať, aby naša teória nebola pravdivá?* **Vyjadrite nulovú hypotézu štatisticky:** *Podiel vybraných mužov sa rovná podielu vybraných žien.* **Vytvorte štatistický test na zhodnotenie nulovej hypotézy.**

Zdroj: vlastné spracovanie podľa Coolican (1999)

4.2 Postup testovania štatistických hypotéz

K tomu, aby sme poznali postup testovania štatistických hypotéz, potrebujeme poznať jeho základné atribúty (podľa Tkáč, 2001):

- podmienky jeho realizovateľnosti (napríklad niektoré testy možno vykonať iba pre konkrétne typy rozdelení),
- hypotézy H_0, H_1 a hladinu významnosti α,
- testovaciu charakteristiku vypočítanú na základe výberového súboru,

- kritický obor C_α určený na základe hladiny významnosti a kvantilov rozdelenia testovacej charakteristiky, na základe ktorého rozhodneme o (ne)zamietnutí nulovej hypotézy.

Každému typu štatistického testu zodpovedá určitý postup výpočtu testovacej charakteristiky, na základe ktorej rozhodujeme o hypotéze. Rôzne typy štatistických testov závisia jednak od hypotézy, ktorú chceme overiť a zároveň od podmienok, za ktorých hypotézu môžeme overiť. Niektoré štatistické testy sa odporúčajú pri väčších vzorkách a iné pri vzorkách s menším rozsahom. Pri testovaní štatistických hypotéz sa nevyhneme výpočtu testovacej charakteristiky na základe údajov zo vzorky, vypočítaniu kritického oboru a potom pomocou testovacej charakteristiky a kritického oboru môžeme uskutočniť rozhodnutie o základnej hypotéze. Vypočítanú testovaciu charakteristiku porovnávame s tzv. **kritickou hodnotou** a rozhodneme o zamietnutí alebo **neschopnosti zamietnuť** stanovenú nulovú hypotézu. Základnú hypotézu môžeme buď (A) zamietnuť (ak testovacia štatistika patrí do kritického oboru) alebo ju (B) nie sme schopní zamietnuť (ak nespadá). Nulovú hypotézu nemôžeme prijať. Toto tvrdenie si vysvetlíme na nasledujúcej rozsiahlej analógii, ktorá nám zároveň ozrejmí niektoré základné princípy testovania štatistických hypotéz.

Predstavte si, že ste sudcom a predvedú vám obžalovaného. Nulová hypotéza je, že obžalovaný je nevinný (vo väčšine vyspelých krajín platí prezumpcia neviny). Alternatívna hypotéza je, že obžalovaný je vinný. Štatistický test pre sudcu predstavuje metódy spracovania dôkazného materiálu, výpovedí svedkov a iné. V tejto analógii nie je ľahké určiť, čo predstavuje kritickú hodnotu, no pre jej účel to nie je podstatné. Ako sudca môžete urobiť nasledujúce dve rozhodnutia:

- obžalovaného odsúdiť,
- obžalovaného neodsúdiť.

V skutočnosti sa však môžete vo vašom rozhodnutí aj pomýliť. Môžete obžalovaného odsúdiť (zamietneme nulovú hypotézu), pričom v skutočnosti je obžalovaný nevinný. Hovoríme tomu aj „*ľahkovernosť*“ (anglicky sa označuje ako *excessive credulity*), falošný poplach (anglicky sa označuje ako *false alarm* alebo najčastejšie túto skutočnosť označujeme ako *false positive*). V štatistike sa tejto chybe hovorí **chyba prvého druhu** (anglicky *type I. error*). Pravdepodobnosť, že túto chybu urobíme označujeme ako $\boldsymbol{\alpha}$ a nazývame **hladinou významnosti**. Na druhej strane, môžeme obžalovaného neodsúdiť (nevieme zamietnuť nulovú hypotézu), pričom v skutočnosti je obžalovaný vinný. Hovoríme tomu **chyba druhého druhu** (anglicky *type II. error*, tejto skutočnosti sa zvykne hovoriť anglicky aj *false negative*). Pravdepodobnosť, že túto chybu urobíme označujeme ako $\boldsymbol{\beta}$. Prirodzene, našou

snahou by mala byť minimalizácia výskytu jednej aj druhej chyby. Uvedené chyby môžu manažéri urobiť pri každom rozhodnutí.

Ako by sme minimalizovali pravdepodobnosť vzniku oboch chýb? Chybu prvého druhu by sme mohli úplne eliminovať tak, že každého obvineného prepustíme. Ide o extrémne riešenie, ale potom sa nemôže stať, že by sme obžalovaného odsúdili napriek tomu, že je nevinný. Chybu druhého druhu by sme mohli úplne eliminovať tak, že každého obvineného odsúdime. Znova ide o extrémne riešenie, ale potom sa nemôže stať, že by sme v skutočnosti vinného neodsúdili. Zrejme si čitateľ všimol, že chybu prvého ako aj druhého druhu nevieme eliminovať súčasne. Nemôžeme obžalovaných zakaždým odsúdiť a zároveň prepustiť. Svojim spôsobom ide o spojené nádoby. Napriek tomu existujú určité spôsoby ako **minimalizovať** pravdepodobnosť výskytu týchto chýb. Vo všeobecnosti sa chyba I. druhu zvolí a minimalizuje sa chyba II. druhu.

Prvým spôsobom minimalizácie chýb je zbieranie čo najväčšieho počtu dôkazov. Ak by sme opustili našu analógiu, mohli by sme povedať, že zvyšovaním rozsahu výberového súboru (n) môžeme minimalizovať vznik chyby prvého aj druhého druhu. Na tomto mieste môžeme vidieť, nakoľko je dostatočný rozsah n výberového súboru dôležitý. Jednak zvyšuje pravdepodobnosť, že naša vzorka bude reprezentatívna a zároveň platí, že pri väčšom rozsahu výberového súboru môžeme dosiahnuť menšiu pravdepodobnosť výskytu chybných rozhodnutí.

Tým druhým spôsobom je použiť vhodnú testovaciu charakteristiku, čiže použiť vhodnú metódu, pomocou ktorej rozhodneme o hypotéze. Ako sme už uviedli, pri riešení rovnakých problémov môžeme neraz použiť rôzne metódy. V tejto publikácii budeme prezentovať niekoľko známych aj menej známych štatistických testov, s ktorými sa môžeme stretnúť pri riešení základných štatistických hypotéz.

Vráťme sa ešte na chvíľu k problému nemožnosti prijať nulovú hypotézu. Ak sudca povie, že obžalovaný je nevinný, znamená to, že aj v skutočnosti nevinný je? Nie nutne. S istotou by to mohol povedať iba vtedy, ak by poznal všetky možné informácie. V štatistike to platí obdobne. Základnú hypotézu môžeme prijať len vtedy, **ak poznáme hodnoty štatistických znakov u všetkých štatistických jednotiek, teda populácie**. Keďže takáto situácia pri vzorke nevzniká, jediným racionálnym tvrdením v prípade nedokázania viny je, že základnú hypotézu nezamietame, keďže **nemáme dostatok dôkazov na to, aby sme mohli základnú hypotézu zamietnuť**. Ide o významný koncept induktívnej štatisticky a vôbec metódy vedeckej práce.

Tabuľka 9: Tabuľka možných výsledkov pri testovaní hypotéz

	SKUTOČNÝ STAV (ak by sme poznali všetky hodnoty populácie)	
ROZHODNUTIE (na základe vzorky)	H_0 **platí**	H_0 **neplatí**
H_0 **zamietnutá**	Získali sme nepravdivo pozitívny výsledok (**FALSE POSITIVE**; α)	Získali sme pravdivo pozitívny výsledok (**TRUE POSITIVE**; $1 - \beta$)
H_0 **nezamietnutá**	Získali sme pravdivo negatívny výsledok (**TRUE NEGATIVE**)	Získali sme nepravdivo negatívny výsledok (**FALSE NEGATIVE**; β)

Zdroj: vlastné spracovanie podľa Banerjee et al. (2009)

Koncept testovania štatistických hypotéz je kľúčový, a preto si ukážeme ešte jeden príklad. Z predchádzajúceho textu je zrejmé, že účelom štatistického testovania hypotéz je rozhodnutie o zamietnutí H_0, respektíve nezamietnutí H_0 (v prospech H_1).

Majme základný súbor, ktorý pozostáva zo 400 súčiastok umiestnených v prepravnom kontajneri. Nech je štatistickým znakom, ktorý je predmetom skúmania, ľubovoľný parameter, ktorý chceme, aby bol v zhode s našimi požiadavkami. Keďže inšpekcia každej súčiastky je príliš nákladná, manažér kvality sa rozhodol náhodne vybrať $n = 4$ $n < 40$ súčiastky, ktoré podrobil inšpekcii. Zaujíma nás, či sú **všetky** súčiastky v dodávke (v základnom súbore) v zhode s našimi požiadavkami. Všimnime si nasledovné:

- K tomu, aby sme mohli dané tvrdenie vyvrátiť, stačí, ak nájdeme v našom výberovom súbore **jednu jedinú** súčiastku, ktorá nebude spĺňať naše požiadavky.
- Čím **väčšiu** máme **vzorku**, tým ľahšie vyvrátime naše tvrdenie (alebo inak, tým máme väčšiu istotu, že ak existuje nezhoda, tak ju aj odhalíme.)
- K tomu, aby sme mohli dané tvrdenie potvrdiť s istotou, museli by sme inšpekcii podrobiť **všetky súčiastky**. Keďže však operujeme s výberovým súborom, k takému záveru prísť nemôžeme. V najlepšom prípade môžeme iba predpokladať, že ak sa vo výberovom súbore nenašla súčiastka, ktorá by nesplnila naše požiadavky, tak ani v dodávke nenájdeme chybnú súčiastku. Ale upozorňujeme, že to nie je fakt – je to len kvalifikovaný predpoklad.

Na podobnom princípe je založená aj skutočnosť, že nulovú hypotézu môžeme buď zamietnuť alebo nezamietnuť, čo nie je to isté ako ju prijať.

V obidvoch z týchto prípadov vzniká riziko, že v dôsledku výberového súboru a použitých štatistických metód sa dopustíme pri rozhodovaní chyby (prvého alebo druhého druhu). Cieľom matematicko-štatistickej teórie je nachádzať také rozhodovacie kritériá, aby

sa pri dopredu určenej prijateľnej pravdepodobnosti výskytu chyby prvého druhu minimalizovala pravdepodobnosť výskytu chyby druhého druhu (Tkáč, 2001).

Výberové charakteristiky ktoré použijeme pri testovaní, nazveme **testovacími charakteristikami alebo testovacími štatistikami**.[17] Technicky spôsob výpočtu funguje nasledovne. Ak platí nulová hypotéza, potom **sa testovacia charakteristika riadi určitým (nám známym) rozdelením pravdepodobnosti**. Túto informáciu využijeme pri rozhodovaní o prijatí, resp. zamietnutí hypotézy: na základe nulovej hypotézy zostrojíme rozdelenie pravdepodobnosti. Ak je na základe tohto rozdelenia veľmi nepravdepodobné, že by nastal jav, ktorý pozorujeme vo vzorke, môže to mať dva dôvody. Buď je chybná vzorka – čo však v prípade, ak sme vzorku vytvárali korektne (napr. uskutočnili sme náhodný výber, meranie štatistických znakov prebehlo správne) môžeme zamietnuť. Ak sa tak stane, tak nepravdepodobnosť získaného výsledku môže byť len dôsledkom chybného predpokladu, ktorým je tvrdenie nulovej hypotézy. Ak teda pri predpoklade platnosti nulovej hypotézy dostávame vysoko nepravdepodobné výsledky, zrejme je nulová hypotéza nepravdivá. V mnohých štatistických testoch budeme rozlišovať medzi: ľavostrannou, pravostrannou a obojstrannou hypotézou.

Pri **ľavostrannom** teste nás zaujíma, aká je pravdepodobnosť, že ak z rozdelenia testovacej charakteristiky náhodne vyberieme jednu hodnotu, tak táto hodnota bude **menšia**, ako nami vypočítaná hodnota. Ak táto pravdepodobnosť bude menšia ako α, hypotézu H_0 zamietame.

Pri **pravostrannom** teste nás zaujíma, aká je pravdepodobnosť, že ak z rozdelenia testovacej charakteristiky náhodne vyberieme jednu hodnotu, tak táto hodnota bude **väčšia**, ako nami vypočítaná hodnota. Ak táto pravdepodobnosť bude menšia ako α, hypotézu H_0 zamietame.

Pri **obojstrannom** teste nás zaujíma, aká je pravdepodobnosť, že ak z rozdelenia testovacej charakteristiky náhodne vyberieme jednu hodnotu, tak táto hodnota bude: **a)** menšia ako nami vypočítaná hodnota testovacej charakteristiky, s pravdepodobnosťou menšou ako $\alpha/2$, hypotézu H_0 zamietame alebo **b)** väčšia ako nami vypočítaná hodnota testovacej charakteristiky, s pravdepodobnosťou menšou ako $\alpha/2$, hypotézu H_0 tiež zamietame.

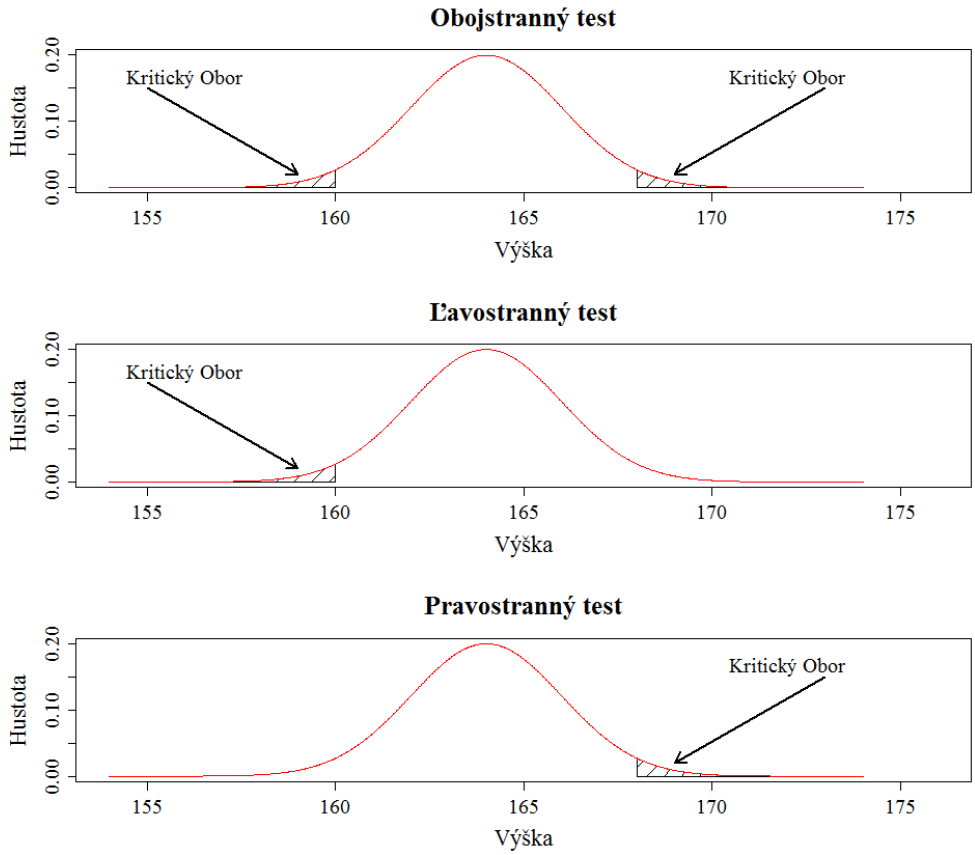
Týmto pravdepodobnostiam hovoríme ***p*-hodnota.** *P*-hodnota je najčastejší spôsob prezentácie výsledku testovania, ktorý softvérové balíky ponúkajú pri rozhodovaní

[17] Tieto dva pojmy budeme v ďalšom texte voľne zamieňať.

o zamietnutí, resp. nezamietnutí štatistickej hypotézy. Rozhodnutie o nezamietnutí, respektíve zamietnutí H_0 potom vykonáme takto:

- ak $\alpha \geq p$ tak H_0 zamietame,
- ak $\alpha < p$, tak H_0 nevieme zamietnuť.

Kritická hodnota, pri ktorej dochádza k zamietnutiu nulovej hypotézy určuje **kritický obor**, C_α. Pre lepšiu ilustráciu si môžeme pozrieť Obrázok 4.1. V prvej časti obrázku je znázornené rozdelenie pravdepodobnosti testovacej charakteristiky a obojstranný kritický obor (pre obojstranné testovanie hypotéz). Za predpokladu, že platí nulová hypotéza, tak testovacia charakteristika sa riadi týmto rozdelením. Ak vypočítaná hodnota testovacej charakteristiky bude spadať do kritického oboru, hypotézu H_0 zamietame.



Obrázok 4.1: Rozdelenie pravdepodobnosti testovacej charakteristicky a kritický obor

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(3, 1))
> x <- seq(154, 174, length = 1000)
> xh <- dnorm(x, mean = 164, sd = 2)
> data <- data.frame(x, xh)
-----------------------------------------------------------------
> plot(data, type = "l", lty = 1, xlab = "Výška", ylab =
  "Hustota", xlim = c(154, 176), family = "serif", cex.axis =
  1.5, cex.lab = 1.7, cex.main = 1.9, main = "Obojstranný test")
```

```
> lb1 = 154; ub1 = 160; lb2 = 168; ub2 = 174;
> i <- (x >= lb1 & x <= ub1)
> polygon(c(lb1, x[i], ub1), c(0, xh[i], 0), density = 10, angle
  = 45, col = "black")
> i <- (x >= lb2 & x <= ub2)
> polygon(c(lb2, x[i], ub2), c(0, xh[i], 0), density = 10, angle
  = 45, col = "black")
> lines(data, type = "l", col = "red")
> arrows(155, 0.15,159, 0.02, length = 0.10, lwd = 2)
> text(156, 0.165, "Kritický Obor", family = "serif", cex = 1.5)
> arrows(173, 0.15, 169, 0.02, length = 0.10, lwd = 2)
> text(172, 0.165, "Kritický Obor", family = "serif", cex = 1.5)
-----------------------------------------------------------------
> plot(data, type = "l", lty = 1, xlab = "Výška", ylab =
  "Hustota", xlim = c(154, 176), family = "serif", cex.axis =
  1.5, cex.lab = 1.7, cex.main = 1.9, main = "Ľavostranný test")
> i <- (x >= lb1 & x <= ub1)
> polygon(c(lb1, x[i], ub1), c(0, xh[i], 0), density = 10, angle
  = 45, col = "black")
> lines(data, type = "l", col = "red")
> arrows(155, 0.15,159, 0.02, length = 0.10, lwd = 2)
> text(156, 0.165, "Kritický Obor", family = "serif", cex = 1.5)
-----------------------------------------------------------------
> plot(data, type = "l", lty = 1, xlab = "Výška", ylab =
  "Hustota", xlim = c(154, 176), family = "serif", cex.axis =
  1.5, cex.lab = 1.7, cex.main = 1.9, main = "Pravostranný
  test")
> i <- (x >= lb2 & x <= ub2)
> polygon(c(lb2, x[i], ub2), c(0, xh[i], 0), density = 10, angle
  = 45, col = "black")
> lines(data, type = "l", col = "red")
> arrows(173, 0.15, 169, 0.02, length = 0.10, lwd = 2)
> text(172, 0.165, "Kritický Obor", family = "serif", cex = 1.5)
```

Číslo α určuje mieru rizika výskytu chyby I. druhu, ktorú sme ochotní podstúpiť. V tejto súvislosti sa často používa číslo γ, nazývané tiež **konfidenčnou pravdepodobnosťou**. V tomto prípade ide o akúsi mieru dôvery, že nenastane chyba prvého druhu. Platí vzťah:

$$\alpha = 1 - \gamma \qquad (4.2)$$

Neschopnosť nájsť zmenu, resp. zamietnuť nulovú hypotézu ak v skutočnosti by sme zmenu mali nájsť, resp. mali zamietnuť nulovú hypotézu, označujeme ako (už spomínanú) chybu druhého druhu β. Opačná pravdepodobnosť $1-\beta$ sa nazýva **sila testu**. Je to pravdepodobnosť, že správne zamietneme nulovú hypotézu a odhalíme efekt, ktorý v skutočnosti existuje. Závislosť sily testu od hodnôt analyzovaného parametra sa označuje **silofunkcia**. Jednou z najčastejších alternatív je modelovanie závislosti sily testu od veľkosti vzorky alebo výberového rozptylu, prípadne s veľkosťou efektu, ktorý chceme odhaliť. Pri voľbe hladiny významnosti nesmieme zabúdať, že so znižovaním hladiny významnosti α rastie výskyt chyby druhého druhu β.

Verejne známym prípadom sú diskusie ohľadom detektora lži pomocou polygrafu. Pomocou tohto prístroja sa dá odhaliť klamstvo s pomerne vysokou pravdepodobnosťou, ak subjekt klame. Kľúčový predpoklad je „*subjekt klame*“. Ak je nulová hypotéza, že subjekt neklame, potom môžeme povedať, že detektor lži má celkom dobré skóre „*true positive*“ (pravdivo zamietne nulovú hypotézu o neklamaní). Na druhej strane pravdepodobnosť, že detektor lži bude signalizovať klamanie aj vtedy, ak v skutočnosti subjekt neklame je pomerne vysoká. Tu je kľúčovým predpokladom „*subjekt neklame*“**.** Detektor lži má tak pre pravdovravných pomerne slabé skóre, keďže sa pomerne často dopúšťa chyby I. druhu „*false positive*“. Detektor vyhlási hypotézu H_0 o neklamaní za nepravdivú, aj keď v skutočnosti nie je nepravdivá. Preto sa neoplatí absolvovať test polygrafom pre subjekty, ktoré neklamú, ale hovoria pravdu.

V nasledujúcich častiach stručne opíšeme niekoľko najčastejšie používaných štatistických hypotéz. Prezentované nie sú zďaleka všetky štatistické testy. Vybrali sme ich vzhľadom na ich použiteľnosť v ekonómii a hospodárskej praxi. Niektoré sú prezentované pomocou ich základných atribútov formou tabuľky, iné sú opísané detailnejšie.

Teraz, keď sme už oboznámení so základnými pojmami testovania štatistických hypotéz a princípov induktívnej štatistiky, môžeme charakterizovať postup testovania štatistických hypotéz do nasledujúcich krokov (podľa Montgomery – Runger, 2011):

1. Oboznámte sa s problémom a identifikujte parameter populácie, resp. populáciu, ktorá je predmetom riešeného problému.
2. Definujte nulovú hypotézu H_0.
3. Špecifikujte vhodnú alternatívnu hypotézu.
4. Vyberte hladinu významnosti α.
5. Vyberte vhodnú testovaciu charakteristiku.
6. Nájdite kritický obor.
7. Vypočítajte hodnotu testovacej charakteristiky.
8. Rozhodnite, či hypotézu H_0 zamietnuť alebo či ju nevieme zamietnuť.

4.3 Základné štatistické testy

V tejto podkapitole budeme prezentovať štatistické testy, s ktorými sa stretávame najčastejšie, prípadne také štatistické testy, ktoré sa využívajú aj v iných štatistických postupoch (napr. regresná analýza):

- Test strednej hodnoty oproti konštante pri známom rozptyle.

- Test strednej hodnoty oproti konštante pri neznámom rozptyle.
- Test dvoch stredných hodnôt: nezávislé súbory.
- Test dvoch stredných hodnôt: závislé súbory.
- Test rozptylu voči konštante.
- Test dvoch rozptylov.
- Test podielu voči konštante.
- Test dvoch podielov: nezávislé súbory.
- Test dvoch podielov: závislé súbory.

4.3.1 Test strednej hodnoty oproti konštante pri známom rozptyle

Majme *iid* vzorku X_i, $i = 1, 2, ..., n$. Tieto hodnoty sú náhodnými realizáciami z normálneho rozdelenia pravdepodobnosti, ktorého rozptyl σ^2 je známy. Testovacou charakteristikou je:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (4.3)$$

kde $\bar{x}$ je aritmetický priemer nameraných hodnôt (náš najlepší odhad strednej hodnoty), μ_0 predstavuje konštantu, resp. strednú hodnotu, voči ktorej aritmetický priemer porovnávame, σ je odmocnina zo známeho rozptylu populácie σ^2 a n je rozsah štatistického súboru. Túto testovaciu charakteristiku budeme označovať ako Z.

Testovacia charakteristika Z sa v prípade platnosti nulovej hypotézy riadi normovaným normálnym rozdelením pravdepodobnosti, formálne $Z \sim N(0, 1)$. Z toho vyplýva, že ak nám nami vypočítaná hodnota testovacej štatistiky zo vzťahu *(4.3)* vyjde veľmi veľké alebo veľmi malé číslo, zrejme nebudeme veriť tomu, že nami namerané hodnoty pochádzajú z takéhoto rozdelenia, kde je stredná hodnota $\mu = \mu_0$. Aby sme vedeli určiť, kedy bude pre nás nameraná hodnota predstavovať dostatočne „veľkú“, resp. „malú“ hodnotu na to, aby sme nulovú hypotézu zamietli, vypočítame si kritický obor. Jeho výpočet závisí od toho, či testujeme obojstrannú alebo jednostranné hypotézy. Vzťahy pre výpočet kritického oboru sú nasledovné:

$H_0: \mu = \mu_0$	Hypotézu H_0 zamietame, ak $\lvert Z\rvert > \lvert z_{(\alpha/2)}\rvert$
$H_1: \mu \neq \mu_0$	

kde $z_{(\alpha/2)}$ je kvantil normovaného normálneho rozdelenia. Jeho hodnotu môžeme získať pomocou kvantilovej funkcie kumulatívnej distribučnej funkcie. V programe R vieme

príslušnú hodnotu odčítať použitím funkcie `qnorm()`. Obdobne postupujeme pri jednostranných hypotézach.

Pravostranný test:

H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$	Hypotézu H_0 zamietame, ak $Z > z_{(1-\alpha)}$

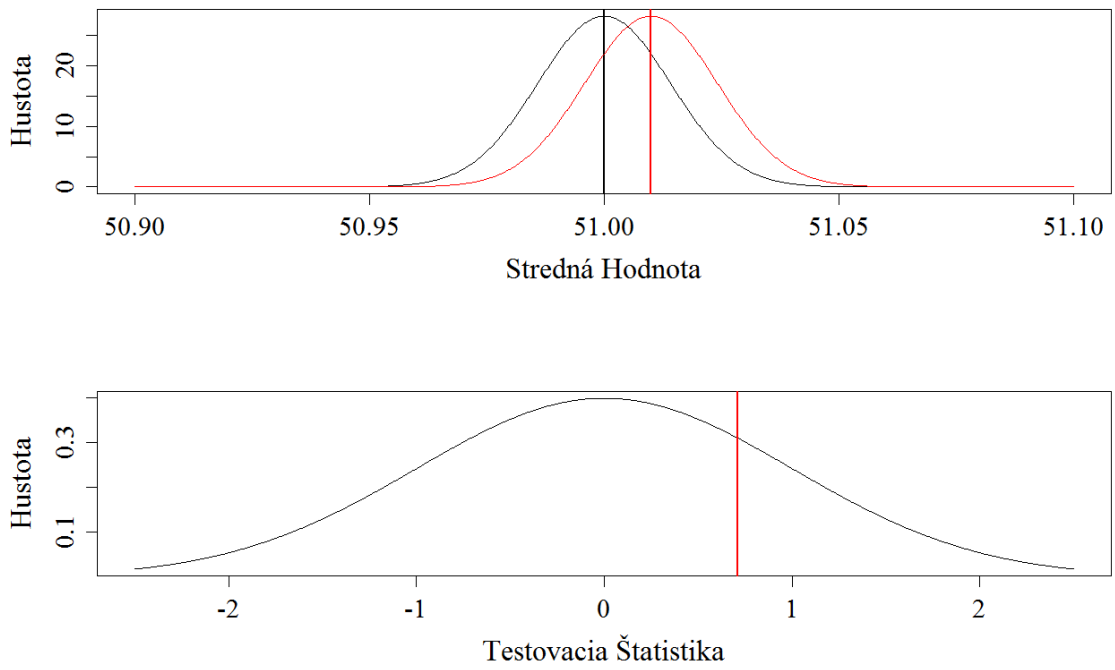
Ľavostranný test:

H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	Hypotézu H_0 zamietame, ak $Z < z_{(\alpha)}$

Na ukážku si realizáciu tohto testu predvedieme na nasledujúcom príklade.

Príklad 4.2

Dodávateľ automobilového priemyslu vyrába súčiastku, ktorej kritický parameter predstavuje priemer hriadeľa, ktorého hodnota podľa odberateľskej zmluvy musí byť 51 *mm*. Určitá tolerancia je samozrejme prípustná. Pri prijímaní dodávky odberateľ z paliet náhodným spôsobom vyberie 25 súčiastok. Pre každú z týchto súčiastok odmeria priemer hriadeľa a vypočíta celkový priemer, $\bar{x} = 51.01$. Vie, že rozptyl, s ktorým výrobca pracuje je na úrovni $\sigma^2 = 0.005$. Taktiež je známe, že hodnoty priemeru hriadeľa sa správajú podľa normálneho rozdelenia pravdepodobnosti. Odberateľa zaujíma, či je na základe získanej vzorky možné tvrdiť, že stredná hodnota hriadeľov je rôzna od 51 *mm*. Ide o obojstrannú hypotézu: H_0: $\mu = 51$ *mm*, H_1: $\mu \neq 51$ *mm*, $Z = 0.71$, $|z_{(\alpha)}| = 1.96$. Keďže $|Z|$ nie je väčšie ako 1.96, hypotézu H_0: $\mu = 51$ *mm* nevieme zamietnuť. Inak povedané, na základe získaných údajov nie je možné tvrdiť, že stredná hodnota hriadeľov dodávky je rôzna od 51 *mm*. Túto situáciu si ukážeme ešte vizuálne.



Obrázok 4.2: Príklad rozdelenia parametra populácie a testovacej štatistiky

Zdroj: vlastné spracovanie, výstup zo softvéru R

Prvá časť obrázku (pozri Obrázok 4.2) predstavuje dve hustoty. Čiernou farbou je označená hustota pravdepodobnosti strednej hodnoty pre prípad, ak je stredná hodnota μ = 51 *mm*, červenou farbou je posunutá stredná hodnota na μ‘ = 51.01 *mm*. Z tohto obrázku je zrejmé, že tieto dve rozdelenia sa pomerne výrazným spôsobom prekrývajú, t. j. že majú spoločnú väčšiu časť plochy, ktorú s osou *x* opisujú. Čím je tejto spoločnej plochy viac, o to máme väčšiu tendenciu tvrdiť, že tieto dve rozdelenia pravdepodobnosti sú približne rovnaké. Spodná časť obrázku zachytáva rozdelenie testovacej charakteristiky. Červenou farbou je zvýraznená nami vypočítaná hodnota testovacej charakteristiky, pričom kritická hodnota je ±1.96. Keďže $|Z| \leq |z_{(\alpha/2)}|$, hypotézu H_0 nevieme zamietnuť.

```
> par(mfrow = c(2, 1))
> x <- seq(50.9, 51.1, length = 1000)
> xh <- dnorm(x, mean = 51, sd = sqrt(0.005)/sqrt(25))
> data <- data.frame(x, xh)
-----------------------------------------------------------------
> plot(data, type = "l", lty = 1, xlab = "Stredná Hodnota", ylab
  = "Hustota", xlim = c(50.9, 51.1), family = "serif", cex.axis
  = 1.5, cex.lab = 1.7, cex.main = 1.9)
> xhh <- dnorm(x, mean = 51.01, sd = sqrt(0.005)/sqrt(25))
> abline(v = c(51, 51.01), col = c("black", "red"), lwd = 1.8)
> data <- data.frame(x, xhh)
> lines(data, type = "l", col = "red")
-----------------------------------------------------------------
> x <- seq(-2.5, 2.5, length = 1000)
> xh <- dnorm(x)
```

```
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1, xlab = "Testovacia
  Štatistika", ylab = "Hustota", xlim = c(-2.5, 2.5), family =
  "serif", cex.axis = 1.5, cex.lab = 1.7, cex.main = 1.9)
> abline(v = 0.71, col = "red", lwd = 1.8)
```

4.3.2 Test strednej hodnoty oproti konštante pri neznámom rozptyle

Majme *iid* vzorku X_i, $i = 1, 2, ..., n$, ktorá predstavuje hodnoty z náhodného výberu. Na rozdiel od predchádzajúceho prípadu uvažujme najprv o situácii, kde $n \leq 30$ a rozptyl hodnôt populácie nie je známy. V tomto prípade potrebujeme, aby X_i boli náhodnými realizáciami z normálneho rozdelenia pravdepodobnosti. Potom testovacia charakteristika má nasledujúci tvar:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \qquad (4.4)$$

Neznámy rozptyl odhadujeme pomocou výberového rozptylu s^2. Testovacia charakteristika t sa (za predpokladu platnosti nulovej hypotézy) riadi Studentovým t rozdelením s $(n - 1)$ stupňami voľnosti. Vzťahy pre výpočet kritického oboru sú nasledovné:

H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	Hypotézu H_0 zamietame, ak $\lvert t \rvert > \lvert t_{\alpha/2,\,(n-1)} \rvert$

pričom $t_{\alpha/2,\,(n-1)}$ predstavuje kvantil Studentovho t rozdelenia, ktorý si vieme vypočítať pomocou kvantilovej funkcie, prípadne v programe R vieme príslušnú hodnotu odčítať použitím funkcie `qt()`. Obdobne postupujeme pri jednostranných hypotézach.

Pravostranný test:

H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$	Hypotézu H_0 zamietame, ak $t > t_{(1-\alpha),\,(n-1)}$

Ľavostranný test:

H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	Hypotézu H_0 zamietame, ak $t < t_{(\alpha),\,(n-1)}$

Tomuto testu hovoríme aj t-test, ktorý patrí zrejme k najrozšírenejšiemu testu, ktorý sa v praxi používa pri overovaní strednej hodnoty a to aj v prípadoch ak $n > 30$ (pripomenieme, že pre dosť veľké n, budú rozdiely medzi Studentovým t rozdelením testovacej štatistiky a normálnym rozdelením pravdepodobnosti prakticky zanedbateľné). Pre väčšie vzorky

predpoklad o normalite nie je kľúčový. V takom prípade sa výsledky z *t*-testu opierajú o centrálnu limitnú vetu.

Príklad 4.3

Praktickú ukážku použitia testu predvedieme priamo na príklade od Verzani (2005, s. 218). V databáze `babies` (programový balík `UsingR`) je premenná `dht`, ktorá predstavuje výšku mužov. Predpokladajme, že hodnoty predstavujú realizácie náhodného výberu. Úlohou je overiť, že výška mužov je 68 palcov, pričom alternatívnou hypotézou je, že výška mužov je väčšia. Úlohu riešime na hladine významnosti $\alpha = 0.05$. Prvým krokom je odstrániť z databázy pozorovania s hodnotou 99, keďže tie predstavujú chýbajúce pozorovania.

```
> library(UsingR)
> attach(babies)
> babies_n <- subset(babies, subset = dht != 99)
> detach(babies); attach(babies_n)
```

Následne vieme test vykonať pomocou funkcie zabudovanej priamo v programe R, kde pri `conf.level` volíme hodnotu $1 - \alpha = \gamma$, čo je už raz spomínaná konfidenčná pravdepodobnosť.

```
> t.test(dht, alternative = c("greater"), mu = 68, conf.level =
  0.95)

                    One Sample t-test

data:  dht
t = 20.7957, df = 743, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 68
95 percent confidence interval:
70.02973      Inf
sample estimates:
mean of x
  70.2043
```

Pre rozhodnutie našej hypotézy nám stačí výsledok v podobe *p*-hodnoty. Ak je táto hodnota menšia ako hladina významnosti, potom hypotézu H_0 zamietame. Tak je tomu aj v tomto prípade, keďže *p*-hodnota je veľmi malé kladné číslo. Z uvedených výsledkov vieme zistiť aj hodnotu testovacej štatistiky, ktorá je označená ako $t = 20.7957$ a počet stupňov voľnosti $df = 743$ (z angl. *degrees of freedom*). Tieto hodnoty si vieme aj manuálne overiť.

```
> (mean(dht)-68)/(sd(dht)/sqrt(length(dht)))
[1] 20.79567
> length(dht)-1
[1] 743
```

Kritickú hodnotu si vypočítame ako:

```
> qt(0.95,743)
[1] 1.646907
```

Výsledky z funkcie `t.test()` nám vrátia aj príslušné konfidenčné intervaly. Vo všeobecnosti platí, že pokiaľ sa testovaná hodnota (hodnota, voči ktorej náš bodový odhad porovnávame v hypotézach) nenachádza v konfidenčnom intervale, hypotézu H_0 zamietame. Z tohto dôvodu sa niekedy aj konfidenčné intervaly vytvorené pomocou bootstrapu zvyknú používať pre potreby štatistického testovania hypotéz.

Príklad 4.4

Zopakujme si predchádzajúci príklad s tým, že budeme overovať hypotézu: H_0: $\mu \leq 70.0$ oproti alternatíve H_1: $\mu > 70.0$.

```
> t.test(dht, alternative = c("greater"), mu = 70, conf.level =
  0.95)

                    One Sample t-test

data:  dht
t = 1.9274, df = 743, p-value = 0.02716
alternative hypothesis: true mean is greater than 70
95 percent confidence interval:
70.02973      Inf
sample estimates:
mean of x
  70.2043
```

Všimnime si *p*-hodnotu, ktorá je 0.02716. Hypotézu H_0 teda na hladine významnosti $\alpha = 0.05$ zamietame, ale ak by sme od začiatku uvažovali s hladinou významnosti $\alpha = 0.01$, tak by sme hypotézu H_0 nemohli zamietnuť. Výpočet *p*-hodnoty si taktiež vieme overiť, stačí nám k tomu poznať rozdelenie pravdepodobnosti testovacej štatistiky a parametre tohto rozdelenia.

```
> 1 - pt(1.9274,743)
[1] 0.02715536
```

Príklad 4.5

Vykonajme predchádzajúci príklad pre ďalšie hypotézy. Prípad A) H_0: $\mu \geq 70.0$ oproti alternatíve H_1: $\mu < 70.0$.

```
> t.test(dht, alternative = c("less"), mu = 70, conf.level =
  0.95)

                    One Sample t-test

data:  dht
t = 1.9274, df = 743, p-value = 0.9728
alternative hypothesis: true mean is less than 70
95 percent confidence interval:
-Inf 70.37887
sample estimates:
mean of x
  70.2043
```

Hypotézu H_0 nevieme zamietnuť. Overme *p*-hodnotu:

```
> pt(1.9274, 743)
[1] 0.9728446
```

Prípad B) H_0: $\mu = 70.0$ oproti alternatíve H_1: $\mu \neq 70.0$.

```
> t.test(dht, alternative = c("two.sided"), mu = 70, conf.level
  = 0.95)

                    One Sample t-test

data:  dht
t = 1.9274, df = 743, p-value = 0.05431
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
69.99621 70.41239
sample estimates:
mean of x
  70.2043
```

Hypotézu H_0 nevieme zamietnuť na hladine významnosti $\alpha = 0.05$, avšak na hladine významnosti $\alpha = 0.1$ by sme ju už zamietli. Kritická hodnota z oboch strán je pritom:

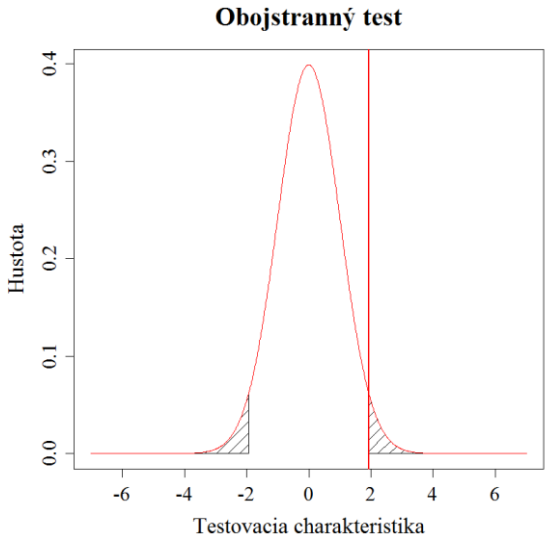
```
> qt(0.025,743)
[1] -1.963162
> qt(0.975,743)
[1] 1.963162
```

Ide zjavne o „tesný“ výsledok. Postup výpočtu *p*-hodnoty „manuálne“ je pre prípad obojstranného testu nasledujúci:

```
> 1 - pt(1.9274,743) + pt(-1.9274,743)
[1] 0.05431073
```

Všimnime si, že výraz `1 - pt(1.9274,743)` predstavuje obsah plochy pod funkciou hustoty Studentovho *t* rozdelenia na pravo od hodnoty 1.9274 a výraz `pt(-1.9274,743)` obsah plochy pod tou istou funkciou hustoty na ľavo od hodnoty –1.9274. Na nasledujúcom obrázku si túto situáciu znázorníme.

```
> x <- seq(-7, 7, length = 1000)
> xh <- dt(x, df = 743)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1, xlab = "Testovacia
  charakteristika", ylab = "Hustota", xlim = c(-7, 7), family =
  "serif", cex.axis = 1.5, cex.lab = 1.7, cex.main = 1.9, main =
  "Obojstranný test")
> lb1 <- -7; ub1 <- -1.9274; lb2 <- 1.9274; ub2 <- 7;
> i <- (x >= lb1 & x <= ub1)
> polygon(c(lb1, x[i], ub1), c(0, xh[i], 0), density = 10, angle
  = 45, col = "black")
> i <- (x >= lb2 & x <= ub2)
> polygon(c(lb2, x[i], ub2), c(0, xh[i], 0), density = 10, angle
  = 45, col = "black")
> lines(data, type = "l", col = "red")
> abline(v = 1.9274, col = "red", lwd = 1.7)
```



Obrázok 4.3: Hustota testovacej charakteristiky *t*-testu

Zdroj: vlastné spracovanie, výstup zo softvéru R

Viackrát sme spomínali, že pre vzorky $n \leq 30$ je potrebné, aby boli namerané hodnoty realizáciami z normálneho rozdelenia pravdepodobnosti. Uskutočníme jednoduchý experiment, ktorého cieľom je sledovať, nakoľko je *t*-test robustný voči porušeniu tohto predpokladu. Pomocou generátora náhodných čísel vyberieme 15 hodnôt z normovaného normálneho rozdelenia pravdepodobnosti $N(\mu = 0, \sigma^2 = 1)$. Vykonáme testovanie hypotéz na hladine významnosti $\alpha = 0.05$, H_0: $\mu = 0$ oproti alternatíve H_1: $\mu \neq 0$. Vieme, že túto hypotézu H_0 by sme nemali zamietnuť. Tento pokus realizujeme 10000 krát. Uvidíme, v koľkých prípadoch danú hypotézu zamietneme. Teoreticky by sa podiel zamietnutých hypotéz H_0 mal pohybovať okolo 5 %, t. j. 500 krát. V tejto variante simulácie dodržiavame podmienky *t*-testu.

```
> pvalues <- c()
> for (i in 1:10000) {
+ x <- rnorm(15)
+ pvalues <- c(pvalues, t.test(x, alternative = c("two.sided"),
  mu = 0, conf.level = 0.95)$p.value)
+ }
> sum(pvalues<0.05)/10000
[1] 0.0498
```

Náš výsledok 0.0498 zodpovedá 4.98 %, čo je dosť blízko k tzv. nominálnej hodnote 0.05, t. j. 5 %. V ďalšom kroku zmeníme rozdelenie na exponenciálne so strednou hodnotou 1, teda $\mu = 1/\lambda = 1$, t. j. $\lambda = 1$. Vykonáme testovanie hypotéz na hladine významnosti $\alpha = 0.05$, H_0: $\mu = 1$ oproti alternatíve H_1: $\mu \neq 1$.

```
> pvalues <- c()
> for (i in 1:10000) {
+ x <- rexp(15, rate = 1)
+ pvalues <- c(pvalues, t.test(x, alternative = c("two.sided"),
  mu = 1, conf.level = 0.95)$p.value)
+ }
> sum(pvalues<=0.05)/10000
[1] 0.0833
```

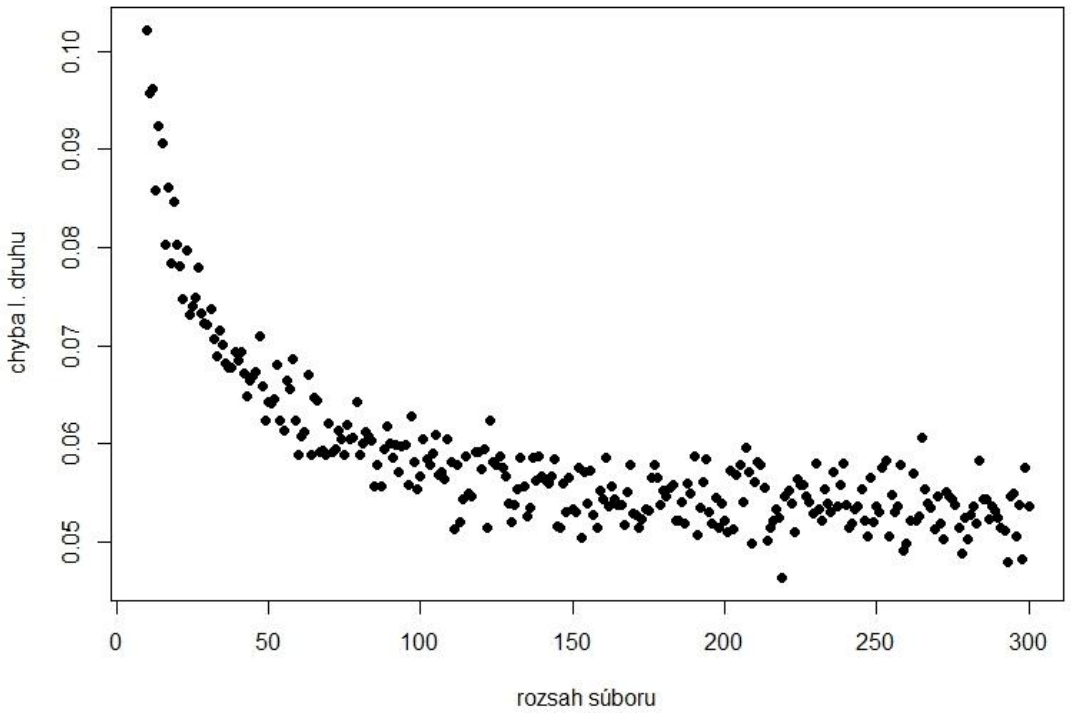
Môžeme vidieť, že 8.33 % už je výrazne viac ako nominálnych 5 %. Presnejšie, chyba prvého druhu je o ≈ 66 % väčšia ako je prípustná nominálna hranica. Vyskúšajme pokus ešte raz pri spojitom rovnomernom rozdelení s min = –2, max = 2. Vykonáme testovanie hypotéz na hladine významnosti $\alpha = 0.05$, H_0: $\mu = 0$ oproti alternatíve H_1: $\mu \neq 0$.

```
> pvalues <- c()
> for (i in 1:10000) {
+ x <- runif(15, min = -2, max = 2)
+ pvalues <- c(pvalues, t.test(x, alternative = c("two.sided"),
  mu = 0, conf.level = 0.95)$p.value)
```

```
+ }
> sum(pvalues<=0.05)/10000
[1] 0.052
```

Výsledok 5.2 % je pomerne dobrý. Táto ukážka nám slúžila k tomu, aby sme videli, že keď podmienky *t*-testu neboli dodržané, chyba I. druhu sa môže zvýšiť a závisí od formy rozdelenia pravdepodobnosti, aký veľký bude tento nárast. Z tohto dôvodu, aj keď nevieme z akého rozdelenia pochádzajú nami získané realizácie, je vhodné údaje pozrieť formou histogramu alebo iného grafického nástroja, prípadne ich otestovať pomocou štatistického testu alebo prístupom k inej informácii zistiť, akého charakteru sú namerané údaje. Napríklad v tomto prípade sa zdá, že pri *t*-teste preferujeme symetrické rozdelenia.

Nakoniec vyskúšame uskutočniť pokus ešte raz s exponenciálnym rozdelením, ale teraz si zvolíme veľkosť vzorky nie $n = 15$, ale $n = 20$. V našej simulácii nám výsledok vyšiel 8.05 %. Aby sme videli závislosť chyby I. druhu (pre exponenciálne rozdelenie) od veľkosti vzorky, tento pokus si zopakujeme pre $n = 10, 11, ..., 300$ a výsledok znázorníme na *x*-*y* obrázku. Upozorňujeme, že táto simulácia pri daných parametroch je časovo náročná. Na nasledujúcom obrázku (pozri Obrázok 4.4) je zrejmý vzťah medzi veľkosťou vzorky a chybou I. druhu. S nárastom veľkosti vzorky sa chyba I. druhu blíži k nominálnej hodnote.



Obrázok 4.4: Závislosť chyby I. druhu od veľkosti vzorky pri *t*-teste

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> typeIerror <- c()
> for (j in 10:300) {
+ pvalues <- c()
+ for (i in 1:10000) {
+ x <- rexp(j, rate = 1)
+ pvalues <- c(pvalues, t.test(x, alternative = c("two.sided"),
  mu = 1, conf.level = 0.95)$p.value)
+ }
+ typeIerror <- c(typeIerror,sum(pvalues<=0.05)/10000)
+ }
-----------------------------------------------------------------
> library(ggplot2)
> plot(10:300, typeIerror, type = "p", pch = 19, xlab = "rozsah
  súboru", ylab = "chyba I. druhu")
```

4.3.3 Test dvoch stredných hodnôt: nezávislé súbory

Majme dve *iid* vzorky X_i, $i = 1, 2, ..., n_x$ a Y_j, $j = 1, 2, ..., n_y$, kde ak je n_x alebo $n_y \leq 30$ je vhodné, aby realizácie náhodného výberu boli z normálneho rozdelenia pravdepodobnosti. Ďalej nech platí nezávislosť medzi hodnotami dvoch vzoriek, čo si môžeme interpretovať tak, že výber hodnôt do jednej vzorky neovplyvňuje výber hodnôt do druhej vzorky. Zaujíma nás, či sú (či sú ich stredné hodnoty rozdielne) rozdiely v stredných hodnotách, prípadne aké veľké sú rozdiely, teda $\mu_x - \mu_y = \mu_0$, $\mu_x - \mu_y \leq \mu_0$, $\mu_x - \mu_y \geq \mu_0$.

V niektorých publikáciách sa môžeme stretnúť s tým, že nasledujúce testy platia iba za predpokladu, že hodnoty sú realizáciami náhodného výberu z normálneho rozdelenia pravdepodobnosti bez ohľadu na veľkosť vzorky. Je známe, že ak tento predpoklad platí, tieto testy dosahujú najlepšie výsledky (v zmysle nízkej chyby I. druhu). Na druhej strane, porušenie predpokladu o normalite spravidla nemá výrazný dopad na výsledky, ak sú vzorky dostatočne veľké (bližšie pozri Panik, 2005). Na nešťastie, čo je dostatočne veľká vzorka sa nedá presne určiť, keďže to môže závisieť od:

- tvaru rozdelenia jednotlivých súborov,
- či sú tieto rozdelenia rovnaké,
- či sú veľkosti vzoriek dvoch súborov rovnaké (preferuje sa rovnaká veľkosť),
- či sú rozptyly dvoch vzoriek odlišné,
- aký veľký rozdiel medzi strednými hodnotami považujeme za zmysluplný.

Naše odporúčanie $n_x, n_y > 30$ je pomerne konzervatívne. Testy dvoch stredných hodnôt sú robustnejšie voči porušeniu predpokladu normality ako testy jednej strednej hodnoty voči konštante. Odporúčania týkajúce sa veľkosti vzorky pri testoch jednej strednej hodnoty (ak je porušený predpoklad normality) sú $n > 30$ (prípadne $n > 40$). Pri teste o dvoch stredných

hodnotách by malo stačiť $n_x + n_y > 30$ (Moore et al., 2009). Ak je rozumné predpokladať, že realizácie náhodného výberu pochádzajú z normálneho rozdelenia pravdepodobnosti, výsledky testov zhody dvoch stredných hodnôt sú pomerne spoľahlivé aj pre malé vzorky.

Existuje niekoľko testovacích charakteristík, ktorých použitie závisí od určitých predpokladov. Uvažujme teda najprv o situácií, kde **rozptyly σ^2_x a σ^2_y sú známe a sú rovné,** teda $\sigma^2_x = \sigma^2_y$. Ďalej veľkosť vzoriek je dostatočne veľká na to, aby sa výberové rozdelenie pravdepodobnosti výberových priemerov dalo aproximovať normálnym rozdelením pravdepodobnosti. Testovacia charakteristika má potom tvar:

$$Z_0 = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \qquad (4.5)$$

Táto testovacia charakteristika sa za predpokladu platnosti nulovej hypotézy riadi normálnym rozdelením pravdepodobnosti, $Z_0 \sim N(0,1)$. Následne rozhodnutie o hypotéze uskutočníme nasledovne:

H_0: $\mu_x - \mu_y = \mu_0$ H_1: $\mu_x - \mu_y \neq \mu_0$	Hypotézu H_0 zamietame, ak $\lvert Z_0\rvert > \lvert z_{(\alpha/2)}\rvert$
H_0: $\mu_x - \mu_y \leq \mu_0$ H_1: $\mu_x - \mu_y > \mu_0$	Hypotézu H_0 zamietame, ak $Z_0 > z_{(1-\alpha)}$
H_0: $\mu_x - \mu_y \geq \mu_0$ H_1: $\mu_x - \mu_y < \mu_0$	Hypotézu H_0 zamietame, ak $Z_0 < z_{(\alpha)}$

kde $z_{(\alpha)}$ je kvantil normovaného normálneho rozdelenia. V prípade, ak **rozptyly σ^2_x a σ^2_y sú známe a nie sú rovné**, teda $\sigma^2_x \neq \sigma^2_y$, tak testovacia charakteristika má tvar:

$$Z_0' = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \qquad (4.6)$$

Znovu platí, že táto testovacia štatistika Z_0' sa riadi normálnym rozdelením pravdepodobnosti $N(0,1)$. Rozhodovanie o hypotéze je obdobné ako vyššie:

H_0: $\mu_x - \mu_y = \mu_0$ H_1: $\mu_x - \mu_y \neq \mu_0$	Hypotézu H_0 zamietame, ak $\lvert Z_0'\rvert > \lvert z_{(\alpha/2)}\rvert$
H_0: $\mu_x - \mu_y \leq \mu_0$ H_1: $\mu_x - \mu_y > \mu_0$	Hypotézu H_0 zamietame, ak $Z_0' > z_{(1-\alpha)}$
H_0: $\mu_x - \mu_y \geq \mu_0$ H_1: $\mu_x - \mu_y < \mu_0$	Hypotézu H_0 zamietame, ak $Z_0' < z_{(\alpha)}$

V oboch predchádzajúcich prípadoch je kľúčovým predpokladom skutočnosť, že poznáme oba rozptyly (parametre populácie σ^2_x a σ^2_y). V praxi je to pomerne zriedkavý

prípad. Častejšie sa stretávame so situáciou, kde rozptyly nie sú známe. V prípade, ak **rozptyly σ^2_x a σ^2_y nie sú známe, ale existuje rozumný dôvod predpokladať, že sú rovné ($\sigma^2_x = \sigma^2_y$)**, tak testovacia charakteristika má tvar:

$$T_0 = \frac{(\bar{x} - \bar{y}) - \mu_0}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \tag{4.7}$$

kde:

$$S_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{k}} \tag{4.8}$$

S_p predstavuje odhad spoločnej smerodajnej odchýlky oboch súborov, kde k sú stupne voľnosti a platí $k = n_x + n_y - 2$. Výberové rozptyly sú označené ako s^2_x a s^2_y. Testovacia charakteristika T_0 sa riadi Studentovým t rozdelením s k stupňami voľnosti. Rozhodovanie o hypotéze je nasledovné:

H_0: $\mu_x - \mu_y = \mu_0$ H_1: $\mu_x - \mu_y \neq \mu_0$	Hypotézu H_0 zamietame, ak $\lvert T_0 \rvert > \lvert t_{(\alpha/2),\,k} \rvert$
H_0: $\mu_x - \mu_y \leq \mu_0$ H_1: $\mu_x - \mu_y > \mu_0$	Hypotézu H_0 zamietame, ak $T_0 > t_{(1-\alpha),\,k}$
H_0: $\mu_x - \mu_y \geq \mu_0$ H_1: $\mu_x - \mu_y < \mu_0$	Hypotézu H_0 zamietame, ak $T_0 < t_{(\alpha),\,k}$

Zrejme najčastejšou situáciu je stav, keď **rozptyly σ^2_x a σ^2_y nie sú známe a nie sú rovné ($\sigma^2_x \neq \sigma^2_y$)**. Testovacia charakteristika má v takom prípade tvar:

$$T_0' = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \tag{4.9}$$

Táto charakteristika T_0' sa riadi Studentovým t rozdelením s υ stupňami voľnosti, ktoré sa vypočítajú podľa vzťahu:

$$\upsilon = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\left(\frac{s_x^2}{n_x}\right)^2 \left(\frac{1}{n_x - 1}\right) + \left(\frac{s_y^2}{n_y}\right)^2 \left(\frac{1}{n_y - 1}\right)} \tag{4.10}$$

pričom výsledok je potrebné zaokrúhliť na najbližšie celé číslo. Môžeme sa stretnúť aj s nasledujúcou aproximáciou: $\upsilon = \min\{(n_x - 1), (n_y - 1)\}$ (Moore et al., 2009). Rozhodovanie o hypotéze je potom:

H_0: $\mu_x - \mu_y = \mu_0$ H_1: $\mu_x - \mu_y \neq \mu_0$	Hypotézu H_0 zamietame, ak $\lvert T_0' \rvert > \lvert t_{(\alpha/2),\upsilon} \rvert$
H_0: $\mu_x - \mu_y \leq \mu_0$ H_1: $\mu_x - \mu_y > \mu_0$	Hypotézu H_0 zamietame, ak $T_0' > t_{(1-\alpha),\upsilon}$
H_0: $\mu_x - \mu_y \geq \mu_0$ H_1: $\mu_x - \mu_y < \mu_0$	Hypotézu H_0 zamietame, ak $T_0' < t_{(\alpha),\upsilon}$

Predošlé testy si vyskúšame aplikovať na modelovom príklade. Použijeme pritom databázu `chickwts` z programového balíka `datasets`. Databáza obsahuje dve premenné: váhu kurčiat („weight“) a druh krmiva („feed“), ktoré sú: bôb obyčajný („horsebean“), ľan siaty („linseed“), sója fazuľová („soybean“), slnečnica obyčajná („sunflower“), mäsová múčka („meat meal“), kasein („casein“). Zaujíma nás, či strednú hodnotu váhy kurčiat, ktorých potrava pozostávala zo sóje fazuľovej, môžeme považovať za rovnakú so strednou hodnotou váhy kurčiat, ktorých potrava pozostávala z ľanu siateho. Overujeme teda hypotézu: H_0: $\mu_x - \mu_y = 0$, H_1: $\mu_x - \mu_y \neq 0$. Predpokladáme, že váha kurčiat sa v oboch vzorkách riadi normálnym rozdelením pravdepodobnosti. V prvej verzii príkladu predpokladáme, že rozptyly váh nie sú známe, avšak sú rovné. Prv ako pristúpime k príkladu si údaje upravíme tak, aby sme mali dva samostatné dátové vektory v programe R, ktoré budú zodpovedať váham príslušnej kategórie kurčiat. Potom použijeme funkciu `t.test()`.

```
> library(datasets); attach(chickwts)
> soja <- chickwts[feed == "soybean", 1]
> lan <- chickwts[feed == "linseed", 1]
------------------------------------------------------------
> t.test(soja, lan, alternative = c("two.sided"), mu = 0,
  var.equal = T, conf.level = 0.95)

                        Two Sample t-test

data:  soja and lan
t = 1.3208, df = 24, p-value = 0.199
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-15.57282  70.92996
sample estimates:
mean of x mean of y
 246.4286  218.7500
```

Z výsledkov môžeme vyčítať priemernú váhu kurčiat podľa krmiva: `soja` = 246.43, `lan` = 218.75. Ďalej vidíme tiež konfidenčný interval pre rozdiel medzi strednými hodnotami. Tu platí, že ak sa v konfidenčnom intervale nachádza 0, potom hypotézu H_0 nebudeme vedieť na danej hladine významnosti zamietnuť. To je aj náš prípad, čo potvrdzuje aj p-hodnota = 0.199 > α. Informáciu o veľkosti vzorky priamo nemáme uvedenú, iba stupne

voľnosti, z ktorých sa dá veľkosť vzorky odhadnúť (alternatívne môžeme použiť príkazy `length(soja)` a `length(lan)`). Pokúsme sa tento výsledok zreprodukovať bez použitia funkcie `t.test()`.

```
> citatel <- mean(soja) - mean(lan)
> nx <- length(soja); ny <- length(lan)
> Sp <- sqrt(((nx - 1)*var(soja) + (ny - 1)*var(lan)) / (nx + ny
  - 2))
> menovatel <- Sp*(sqrt(1/nx + 1/ny))
> test_stat <- citatel/menovatel
> test_stat
[1] 1.320785
```

Následne si vieme overiť aj výsledok pre *p*-hodnotu:

```
> 1 - pt(test_stat, 24) + pt(-test_stat, 24)
[1] 0.1990296
```

Použime teraz rovnaký príklad s tým rozdielom, že nepredpokladáme rovnosť populačných rozptylov. Tento prípad je zrejme najčastejšie používaným. Použijeme funkciu `t.test()`.

```
> t.test(soja, lan, alternative = c("two.sided"), mu = 0,
  var.equal = F, conf.level = 0.95)

                  Welch Two Sample t-test

data:  soja and lan
t = 1.3246, df = 23.63, p-value = 0.198
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-15.48547  70.84262
sample estimates:
mean of x mean of y
 246.4286  218.7500
```

Kvalitatívne nedošlo k odlišným výsledkom. Pre úplnosť uvádzame prepočet vzťahov *(4.9)* a *(4.10)*, ktorý potvrdzuje výsledok získaný pomocou funkcie `t.test()`.

```
> test_stat <- citatel / (sqrt(var(soja)/nx + var(lan)/ny))
> test_stat
[1] 1.324556
> v <- (var(soja)/nx + var(lan)/ny)^2/((var(soja)/nx)^2 * (1/(nx
  - 1)) + (var(lan)/ny)^2 * (1/(ny - 1)))
> v
[1] 23.62952
> round(v)
[1] 24
> 1 - pt(test_stat, v) + pt(-test_stat, v)
```

```
[1] 0.1979871
```

4.3.4 Test dvoch stredných hodnôt: závislé súbory (párový t-test)

Doteraz sme uvažovali o dvoch vzorkách, ktoré sú navzájom nezávislé. Ak by sme merali krvný tlak basketbalistov pred zápasom (vzorka A) a po zápase (vzorka B) a chceli porovnať strednú hodnotu krvného tlaku vzorky A a B, nemohli by sme na to použiť predchádzajúce postupy. V oboch vzorkách totiž vystupujú rovnakí športovci. To znamená, že nejde o nezávislé vzorky. Ak má jeden z basketbalistov tendenciu mať vyšší krvný tlak, je dosť možné, že po zápase bude mať krvný tlak ešte vyšší ako ostatní hráči len v dôsledku tejto svojej vrodenej vlastnosti. V praxi sa môžeme s takouto situáciou stretnúť pomerne často. Ide spravidla o merania „predtým – potom“, ale aj o rôzne merania na tých istých osobách (alebo iných objektoch štatistického skúmania). Uvažujme teda o meraniach, ktoré pre nás predstavujú usporiadané dvojice (X_i, Y_i), $i = 1, 2, ..., n$. Uvedené si môžeme predstaviť na príklade, kde máme jedného pacienta, ktorého váhu X_1 zmeriame na začiatku mesiaca a Y_1na konci mesiaca, dostaneme tak dvojicu meraní jedného pacienta (X_1, Y_1). Súbor takýchto meraní označíme ako (X_i, Y_i), $i = 1, 2, ..., n$. Uvažujme ďalej o náhodnej premennej $R_i = X_i - Y_i$, $i = 1, 2, ..., n$, ktorá predstavuje rozdiel v nameraných hodnotách. Populačný rozdiel si označíme ako μ. Test potom vykonáme na rozdieloch R_i, čím sa vlastne testovanie strednej hodnoty dvoch súborov redukuje na *t*-test o strednej hodnote **rozdielov** μ, ktoré predstavujú iba jeden súbor. Testovanie **dvoch stredných hodnôt závislých súborov** potom vieme vykonať pomocou nasledujúcej testovacej charakteristiky:

$$T_0'' = \frac{\overline{R} - \mu_0}{\frac{s_R}{\sqrt{n}}} \qquad (4.11)$$

kde $\overline{R}$ je priemerný rozdiel R_i, μ_0 je konštanta, voči ktorej tento rozdiel overujeme, s_R je výberová smerodajná odchýlka rozdielov R_i. Testovacia charakteristika T_0'' sa riadi Studentovým *t* rozdelením pravdepodobnosti (za platnosti nulovej hypotézy). Rozhodovanie o hypotéze môžeme vykonať nasledovne:

$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	Hypotézu H_0 zamietame, ak $\lvert T_0'' \rvert > \lvert t_{\alpha/2,\,(n-1)} \rvert$
$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	Hypotézu H_0 zamietame, ak $T_0'' > t_{(1-\alpha),\,(n-1)}$
$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	Hypotézu H_0 zamietame, ak $T_0'' < t_{(\alpha),\,(n-1)}$

Podobne, ako pri teste strednej hodnoty s konštantou, sú $t_{\alpha/2,\,(n-1)}$, $t_{(1-\alpha),\,(n-1)}$ a $t_{(\alpha),\,(n-1)}$ kvantily Studentovho t rozdelenia pravdepodobnosti.

Príklad 4.6

Na ukážku použijeme príklad od Moore et al., (2009), ktorého znenie si mierne upravíme. Overuje sa hypotéza, že spln mesiaca má vplyv na správanie sa ľudí. Vzorku predstavuje 15 pacientov, ktorí majú psychické problémy, pričom majú rovnakú diagnózu (demencia). Pre každého z pacientov sa zaznamenáva priemerný počet agresívnych prejavov správania sa pre dva typy dní. Za prvý typ dňa sa považuje deň pred, deň po a deň splnu mesiaca. Druhý typ sú všetky ostatné dni. Overuje sa teda hypotéza H_0: $\mu = \mu_0$ k alternatíve H_1: $\mu \neq \mu_0$, kde μ predstavuje rozdiel v priemernom počte agresívnych prejavov správania sa pacientov počas dní splnu a obyčajných dní. Vstupné údaje a výsledky testu sú nasledovné:

```
> spln <- c(3.33, 3.67, 2.67, 3.33, 3.33, 3.67, 4.67, 2.67,
  6.00, 4.33, 3.33, 0.67, 1.33, 0.33, 2.00)
> bezny_den <- c(0.27, 0.59, 0.32, 0.19, 1.26, 0.11, 0.30, 0.40,
  1.59, 0.60, 0.65, 0.69, 1.26, 0.23, 0.38)
------------------------------------------------------------
> t.test(spln, bezny_den, alternative = c("two.sided"), paired =
  T, conf.level = 0.95)

                        Paired t-test

data:  spln and bezny_den
t = 6.4518, df = 14, p-value = 1.518e-05
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:

1.623968 3.241365
sample estimates:
mean of the differences
               2.432667
```

Hypotézu H_0 na hladine významnosti $\alpha = 0.05$ zamietame, čo je vo výstupe funkcie možné vidieť jednak z pomerne veľkej hodnoty testovacej štatistiky ($T_0'' = 6.45$), ako aj z nízkej p-hodnoty. Zdá sa, že spln mesiaca môže mať vplyv na správanie sa určitej skupiny ľudí (ľudí s demenciou). Táto vzorka ľudí však nebola vyberaná náhodne. Napriek tomu sa aj v takýchto situáciách často uskutočňujú štatistické testy. Takýto postup nie je a priori chybný, avšak takéto výsledky je samozrejme problematické zovšeobecniť na celú populáciu.

4.3.5 Test rozptylu voči konštante

V predchádzajúcich *t*-testoch sme v niektorých prípadoch uvažovali o rovnosti prípadne nerovnosti populačných rozptylov a na základe toho odporúčali vhodný test. Jednou z možností ako rozhodnúť o (ne)rovnosti rozptylov je použiť štatistický test, kde sa porovnávajú dva rozptyly. Najprv začneme štandardným testom zhody rozptylu s konštantou. Majme *iid* vzorku X_i, $i = 1, 2, ..., n$, ktorej hodnoty sú náhodnými realizáciami z normálneho rozdelenia pravdepodobnosti, formálne $X \sim N(\mu, \sigma^2)$. Cieľom je na základe vzorky overiť, či je možné neznámy rozptyl populácie σ^2 považovať za rovný (rovný menší, rovný väčší) konštante σ_0^2. Testovacou charakteristikou je:

$$\dot{\chi}^2 = \frac{(n-1)s^2}{\sigma_0^2} \qquad (4.12)$$

kde s^2 je výberový rozptyl. Testovacia charakteristika $\dot{\chi}^2$ sa riadi Chí-kvadrát (χ^2) rozdelením pravdepodobnosti s $(n - 1)$ stupňami voľnosti. Rozhodnutie o hypotéze H_0 potom vykonáme nasledovne:

H_0: $\sigma^2 = \sigma_0^2$	Hypotézu H_0 zamietame, ak $\dot{\chi}^2 < \chi^2_{(n-1),\,\alpha/2}$ alebo
H_1: $\sigma^2 \neq \sigma_0^2$	H_0 zamietame, ak $\dot{\chi}^2 > \chi^2_{(n-1),\,(1-\alpha/2)}$
H_0: $\sigma^2 \leq \sigma_0^2$	Hypotézu H_0 zamietame, ak $\dot{\chi}^2 > \chi^2_{(n-1),\,(1-\alpha)}$
H_1: $\sigma^2 > \sigma_0^2$	
H_0: $\sigma^2 \geq \sigma_0^2$	Hypotézu H_0 zamietame, ak $\dot{\chi}^2 < \chi^2_{(n-1),\,\alpha}$
H_1: $\sigma^2 < \sigma_0^2$	

kde $\chi^2_{(n-1),\,\alpha/2}$, $\chi^2_{(n-1),\,1-\alpha/2}$, $\chi^2_{(n-1),\,\alpha}$ a $\chi^2_{(n-1),\,(1-\alpha)}$ predstavujú kvantily Chí-kvadrát (χ^2) rozdelenia pravdepodobnosti s príslušnými stupňami voľnosti.

Príklad 4.7

Na ukážku mierne upravíme príklad od Allen (1990). Výsledky testov programátorskej spôsobilosti sa riadia normálnym rozdelením pravdepodobnosti. Historické údaje naznačujú, že parametrami populácie sú $\mu = 82.6$ a $\sigma^2 = 19.78$. Za posledné roky sa na vzorke 150 respondentov priemer dostal na úroveň 83.2 a výberový rozptyl na 27.3. Cieľom je na hladine významnosti $\alpha = 0.05$ rozhodnúť, či došlo k zvýšeniu rozptylu. Overujeme teda hypotézu H_0: $\sigma^2 \leq 19.78$ H_1: $\sigma^2 > 19.78$. Testovacia charakteristika je:

$$\dot{\chi}^2 = \frac{(150-1)27.3}{19.78} = 205.6471$$

Kritická hodnota $\chi^2_{(n-1),(1-\alpha)}$ predstavuje 178.48, pričom v programe R ju vieme získať pomocou funkcie `qchisq()`.

```
> qchisq(0.95,149)
[1] 178.4854
```

Keďže platí $\dot{\chi}^2 > \chi^2_{(n-1),(1-\alpha)}$ hypotézu H_0 na hladine významnosti 0.05 zamietame v prospech hypotézy H_1. Zdá sa teda, že k zvýšeniu rozptylu naozaj došlo.

V programe R vieme test vykonať na danej vzorke pomocou funkcie `sigma.test()`, ktorý je súčasťou programového balíka `TeachingDemos`.

4.3.6 Test zhody dvoch rozptylov

Častejšie ako s testom rozptylu oproti konštante sa stretávame s potrebou porovnať dva rozptyly. Majme dve *iid* nezávislé vzorky X_i, $i = 1, 2, ..., n_x$ a Y_j, $j = 1, 2, ..., n_y$, pričom obe sú realizáciami náhodného výberu z normálneho rozdelenia pravdepodobnosti, teda $X \sim N(\mu_x, \sigma^2_x)$ a $Y \sim N(\mu_y, \sigma^2_y)$. Ďalej definujme vzorky tak, aby vždy platilo $s^2_x \geq s^2_y$. Testovacia charakteristika má potom tvar:

$$F' = \frac{s_x^2}{s_y^2} \qquad (4.13)$$

Táto testovacia štatistika sa riadi F rozdelením s $\upsilon_1 = n_x - 1$, $\upsilon_2 = n_y - 1$ stupňami voľnosti. Rozhodnutie o hypotézach potom vykonáme nasledovne:

$H_0: \sigma^2_x = \sigma^2_y$ $H_1: \sigma^2_x \neq \sigma^2_y$	Hypotézu H_0 zamietame, ak $F' > F_{\upsilon 1, \upsilon 2, (1-\alpha/2)}$
$H_0: \sigma^2_x \leq \sigma^2_y$ $H_1: \sigma^2_x > \sigma^2_y$	Hypotézu H_0 zamietame, ak $F' > F_{\upsilon 1, \upsilon 2, (1-\alpha)}$

kde $F_{\upsilon 1, \upsilon 2, (1-\alpha/2)}$ a $F_{\upsilon 1, \upsilon 2, (1-\alpha)}$ sú kvantily F rozdelenia pravdepodobnosti. V programe R ich vieme získať pomocou funkcie `qf()`. Všimnime si, že si vystačíme s jedným obojstranným a jedným pravostranným testom. Je to dôsledok toho, že si vzorky definujeme tak, aby platilo $s^2_x \geq s^2_y$.

Príklad 4.8

Použijeme znovu príklad s krmivom pre kurčatá. Cieľom je porovnať rozptyl váh dvoch vzoriek. Zaujíma nás, či môžeme predpokladať, že rozptyl váh kurčiat, ktorých hlavnou

stravou je sója fazuľová, je rovnaký ako rozptyl váh kurčiat, ktorých hlavnou stravou je ľan siaty. Overujeme teda hypotézu H_0: $\sigma^2_x = \sigma^2_y$ oproti alternatíve H_1: $\sigma^2_x \neq \sigma^2_y$. Na výpočet použijeme funkciu `var.test()`.

```
> var.test(soja, lan, alternative = c("two.sided"), conf.level =
  0.95)

             F test to compare two variances

data:  soja and lan
F = 1.0738, num df = 13, denom df = 11, p-value = 0.9172
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
0.3165959 3.4334943
sample estimates:
ratio of variances
          1.073807
```

Hypotézu H_0 nevieme zamietnuť. Výpočet kritickej hodnoty si overíme výpočtom pomocou funkcie `qf()` a overíme si aj *p*-hodnotu, ktorú sme vypočítali cez `var.test()`.

```
> qf(0.975, 13, 11)
[1] 3.391728
> (1 - pf(1.0738, 13, 11))*2
[1] 0.9172031
```

Testy porovnávajúce rozptyl s konštantou a testy na zhodu dvoch rozptylov sú citlivé na porušenie predpokladu o normalite. Na rozdiel od *t*-testu nepomáhajú veľké vzorky ani symetrické rozdelenia. Z tohto dôvodu sa tieto dva testy v praxi veľmi často nevyskytujú. Ak uskutočňujeme *t*-test a rozhodujeme sa, či je rozumné predpokladať rovnaké rozptyly dvoch vzoriek, môže byť vhodnejšie (ak nemáme istotu o normalite hodnôt v oboch vzorkách) porovnať rozdelenia vzoriek pomocou vhodného grafu, prípadne vykonať neparametrické testy na zhodu rozptylov (napr. Levenov test alebo Brown – Forsythov test, pozri Kapitoly 4.6.8 a 4.6.9).

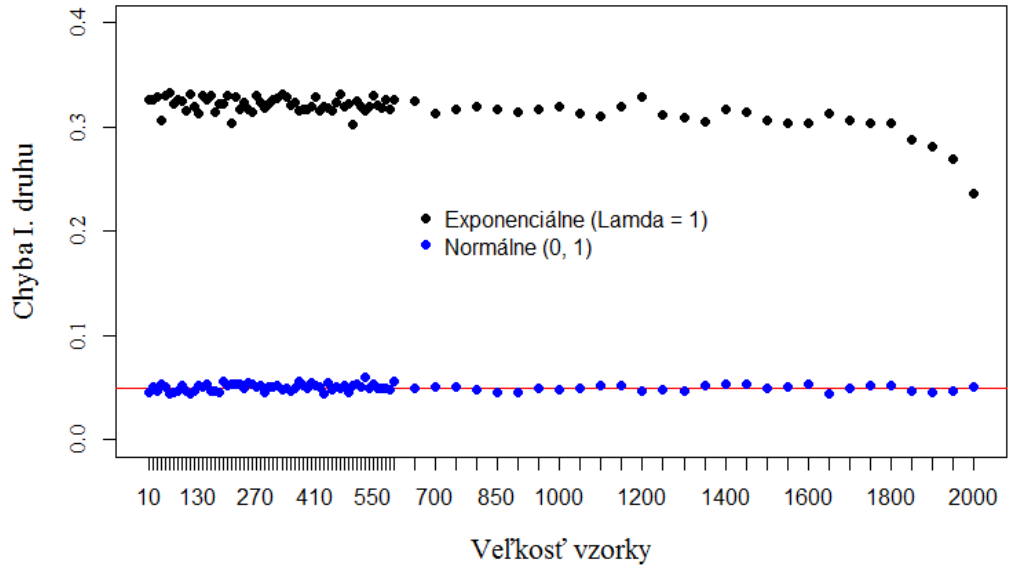
Na ukážku citlivosti testu o rozptyle sme uskutočnili znova jednoduchý experiment. Najprv sme uvažovali o hodnotách z normálneho rozdelenia pravdepodobnosti. Pre rôzne veľkosti vzoriek (vektor „`samples`“) sme generovali po 5000 vzoriek, ktorých hodnoty sme vyberali z normálneho rozdelenia pravdepodobnosti $N(\mu = 0,\ \sigma^2 = 1)$ a vykonali test H_0: $\sigma^2 = 1$ oproti H_1: $\sigma^2 \neq 1$, ktorý sme vyhodnotili na hladine významnosti $\alpha = 0.05$. Následne sme vypočítali podiel chybných zamietnutí hypotézy H_0. Výsledky sme pre jednotlivé veľkosti vzorky naniesli na *x*-*y* graf. Podobne sme postupovali aj v druhom pokuse, kde sme však

uvažovali o exponenciálnom rozdelení s parametrom $\lambda = 1$, t. j. rozptyl tohto rozdelenia je tiež 1. Následne sme vykonali rovnakú štruktúru experimentu ako v predošlom prípade a výsledky naniesli na *x-y* graf.

```
> library(TeachingDemos)
> ratio_norm <- c()
> samples <- c(seq(from = 10, to = 600, by = 10), seq(from =
  650, to = 2000, by = 50))
> for (j in samples) {
+ pval <- c()
+ for (i in 1:5000) {
+ x <- rnorm(j)
+ p <- sigma.test(x, sigma = 1, alternative = c("two.sided"),
  conf.level = 0.95)$p.value
+ pval <- c(pval, p)
+ }
+ ratio_norm <- c(sum(pval<=0.05)/5000, ratio_norm)
+ }
-----------------------------------------------------------------
> ratio_exp <- c()
> for (j in samples) {
+ pval <- c()
+ for (i in 1:5000) {
+ x <- rexp(j, rate = 1)
+ p <- sigma.test(x, sigma = 1, alternative = c("two.sided"),
  conf.level = 0.95)$p.value
+ pval <- c(pval, p)
+ }
+ ratio_exp <- c(sum(pval<=0.05)/5000, ratio_exp)
+ }
-----------------------------------------------------------------
> data <- data.frame(samples, ratio_exp)
> plot(data, type = "p", xaxt = "n", pch = 19, xlab = "Veľkosť
  vzorky", ylab = "Chyba I. druhu", ylim = c(0, 0.4), family =
  "serif", cex.axis = 1.1, cex.lab = 1.3, cex.main = 1.5)
> abline(h = 0.05, col = "red", lwd = 1.8)
> axis(side = 1, at = samples, labels = samples)
> data <- data.frame(samples, ratio_norm)
> points(data, pch = 19, col = "blue")
> legend("center", legend = c("Exponenciálne (Lamda = 1)",
  "Normálne (0, 1)"), pch = 19, col = c("black", "blue"), inset
  = 0.08, bty = "n")
```

Na nasledujúcom obrázku (pozri Obrázok 4.5) zjavne vidno veľký rozdiel v chybe prvého druhu. Kým v prípade normálneho rozdelenia sa chyba pohybuje v okolí nominálnych 5 %, v prípade exponenciálneho rozdelenia výrazne nepomáha ani zvyšovanie veľkosti vzorky, keďže sa chyba prvého druhu pohybuje na úrovni 30 %. Výsledky z tohto malého experimentu samozrejme nie je možné zovšeobecňovať. Mohli sme vybrať aj iné rozdelenia pravdepodobnosti ako exponenciálne. V skutočnosti je nekonečne veľa možností. Zámerne sme vybrali exponenciálne rozdelenie ako pravostranne zošikmené rozdelenie, keďže je

jednoduchým reprezentantom zošikmených nesymetrických rozdelení, ktoré sa značne líšia od normálneho a pritom patria k častým rozdeleniam.



Obrázok 4.5: Citlivosť testu zhody rozptylu s konštantou

Zdroj: vlastné spracovanie, výstup zo softvéru R

4.3.7 Test podielu voči konštante

Neraz je predmetom nášho záujmu počet výskytu určitého javu. Môže nás zaujímať, aký počet potenciálnych zákazníkov, ktorí vošli do obchodu, si niečo v obchode aj kúpili. Aby sme sa vyhli niektorým týždňovým a denným efektom, v priebehu štyroch týždňov vykonáme náhodný výber zákazníkov. Všimnime si, že tento jav (nákup zákazníka) vieme modelovať pomocou binomického rozdelenia pravdepodobnosti. Počet všetkých zákazníkov vo vzorke je n. Existuje určitá pravdepodobnosť, že jeden náhodne vybraný potenciálny zákazník vykoná nákup a tú si označíme ako π. Binomickým rozdelením by sme potom mohli napríklad riešiť úlohu: aká je pravdepodobnosť, že z 50 potenciálnych zákazníkov v 40 prípadoch dôjde ku konverzii na kupujúceho zákazníka? Existuje formálny spôsob ako ukázať, že za určitých podmienok je možné normálnym rozdelením pravdepodobnosti aproximovať binomické rozdelenie pravdepodobnosti. Táto skutočnosť sa využíva pri formovaní testovacej charakteristiky, pomocou ktorou chceme overiť či podiel π zodpovedá určitej konštante p_0. Pokračujúc v predchádzajúcom príklade, zamestnanci predaja tvrdia, že 70 % zákazníkov, ktorí vstúpia do predajne vykonajú aj nákup. Uskutočníme výber a vykonáme test: H_0: $\pi = p_0$ oproti H_1: $\pi \neq p_0$. Predpokladáme, že *iid* vzorka X_i, $i = 1, 2, ..., n$ predstavuje náhodné realizácie, ktoré nadobúdajú hodnotu 0 alebo 1 v závislosti či dôjde

k sledovanému javu, tzv. úspešnému pokusu. Náhodná premenná X zodpovedá počtu úspešných pokusov, pričom sa táto náhodná premenná riadi binomickým rozdelením pravdepodobnosti. V našej vzorke o veľkosti $n = 100$ bolo $x = 65$ kupujúcich zákazníkov. Našim najlepším odhadom podielu π je potom $\hat{p} = x/n$. Následne je testovacia charakteristika:

$$Z^p = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \qquad (4.14)$$

Z^p sa riadi normovaným normálnym rozdelením pravdepodobnosti. Táto aproximácia binomického rozdelenia normálnym platí iba za určitých predpokladov. Vo všeobecnosti sa uvádza, že táto aproximácia je vhodná, ak platí $n\pi \geq 5$ a zároveň $n(1 - \pi) \geq 5$ (napr. Tkáč, 2001; Panik, 2005). Prípadne by malo vždy platiť $n \geq 25$ a v prípade ak si myslíme, že skutočný podiel π je veľmi blízko k 0 alebo 1, tak $n \geq 50$ (Panik, 2005). Rozhodnutie o hypotéze je potom nasledovné:

$H_0: \pi = p_0$ $H_1: \pi \neq p_0$	Hypotézu H_0 zamietame, ak $\lvert Z^p \rvert > \lvert z_{(\alpha/2)} \rvert$
$H_0: \pi \leq p_0$ $H_1: \pi > p_0$	Hypotézu H_0 zamietame, ak $Z^p > z_{(1-\alpha)}$
$H_0: \pi \geq p_0$ $H_1: \pi < p_0$	Hypotézu H_0 zamietame, ak $Z^p < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú kvantily normovaného normálneho rozdelenia pravdepodobnosti (v programe R získané funkciou `qnorm()`). V programe R vykonáme test použitím funkcie `binom.test()`.

```
> binom.test(65, 100, p = 0.7, alternative = c("two.sided"),
  conf.level = 0.95)

                    Exact binomial test

data:  65 and 100
number of successes = 65, number of trials = 100, p-value =
  0.2764
alternative hypothesis: true probability of success is not equal
  to 0.7
95 percent confidence interval:
0.5481506 0.7427062
sample estimates:
probability of success
                  0.65
```

Výpočet sa dá ľahko overiť:

```
> (0.65 - 0.70) / (sqrt(0.7*(1 - 0.7)/100))
[1] -1.091089
> 1 - pnorm(1.091089)+pnorm(-1.091089)
[1] 0.2752337
```

4.3.8 Test dvoch podielov: nezávislé súbory

Podobne ako pri testoch so strednými hodnotami, aj v tomto prípade vieme porovnávať dva podiely navzájom, teda zisťovať, či sú dva podiely rovnaké alebo či je jeden podiel väčší prípadne menší ako druhý. Majme dve nezávislé *iid* vzorky realizácií náhodných premenných X_i, $i = 1, 2, ..., n_x$ a Y_j, $j = 1, 2, ..., n_y$, ktoré podobne ako v predchádzajúcej Kapitole 4.3.7, nadobúdajú hodnoty 0 alebo 1 v závislosti od toho, či došlo k sledovanému javu, tzv. úspešnému pokusu. Náhodné premenné X a Y zodpovedajú počtu úspešných pokusov, pričom sa tieto náhodné premenné riadia binomickým rozdelením pravdepodobnosti. Podiel úspešných pokusov v populáciách si označíme ako π_x a π_y a výberové podiely ako $\hat{p}_x$ a $\hat{p}_y$. Testovaciu charakteristiku, pomocou ktorej môžeme overovať rozdiely v podieloch môžeme zapísať ako:

$$Z'^{p} = \frac{\left(\hat{p}_x - \hat{p}_y\right) - \left(p_x - p_y\right)}{\sqrt{\frac{\hat{p}_x\left(1-\hat{p}_x\right)}{n_x} + \frac{\hat{p}_y\left(1-\hat{p}_y\right)}{n_y}}} \qquad (4.15)$$

Testovacia charakteristika Z'^{π} sa riadi normovaným normálnym rozdelením pravdepodobnosti. Ak Δ_0 použijeme na označenie rozdielu medzi podielmi, rozhodnutie o hypotéze je potom nasledovné:

H_0: $\pi_x - \pi_y = \Delta_0$ H_1: $\pi_x - \pi_y \neq \Delta_0$	Hypotézu H_0 zamietame, ak $\lvert Z'^{\pi} \rvert > \lvert z_{(\alpha/2)} \rvert$
H_0: $\pi_x - \pi_y \leq \Delta_0$ H_1: $\pi_x - \pi_y > \Delta_0$	Hypotézu H_0 zamietame, ak $Z'^{\pi} > z_{(1-\alpha)}$
H_0: $\pi_x - \pi_y \geq \Delta_0$ H_1: $\pi_x - \pi_y < \Delta_0$	Hypotézu H_0 zamietame, ak $Z'^{\pi} < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú kvantily normovaného normálneho rozdelenia pravdepodobnosti (v programe R získané funkciou `qnorm()`). Pri tomto teste znovu platí, že je vhodné mať vzorky čo najpočetnejšie, aby platila aproximácia binomického rozdelenia pravdepodobnosti normálnym. Taktiež by malo pre každú z nezávislých vzoriek platiť už spomínané pravidlo: $n\pi \geq 5$ a zároveň $n(1-\pi) \geq 5$.

V prípade ak overujeme hypotézy, kde $\Delta_0 = 0$, potom sa za vhodnejšiu testovaciu charakteristiku považuje nasledujúca:

$$Z''^{\pi} = \frac{\left(\hat{p}_x - \hat{p}_y\right)}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}} \qquad (4.16)$$

kde Z''^{π} sa riadi normovaným normálnym rozdelením pravdepodobnosti, pričom pre $\hat{p}$ platí:

$$\hat{p} = \frac{\hat{p}_x n_x + \hat{p}_y n_y}{n_x + n_y} \qquad (4.17)$$

Rozhodnutie o hypotéze je potom obdobné ako v predošlom prípade:

H_0: $\pi_x - \pi_y = 0$ H_1: $\pi_x - \pi_y \neq 0$	Hypotézu H_0 zamietame, ak $\lvert Z''^{\pi} \rvert > \lvert z_{(\alpha/2)} \rvert$
H_0: $\pi_x - \pi_y \leq 0$ H_1: $\pi_x - \pi_y > 0$	Hypotézu H_0 zamietame, ak $Z''^{\pi} > z_{(1-\alpha)}$
H_0: $\pi_x - \pi_y \geq 0$ H_1: $\pi_x - \pi_y < 0$	Hypotézu H_0 zamietame, ak $Z''^{\pi} < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú kvantily normovaného normálneho rozdelenia pravdepodobnosti (v programe R získané funkciou `qnorm()`).

Príklad 4.9

Na ilustráciu týchto testov použijeme údaje zo stránky www.focus-research.sk (dostupné online 25.01.2012), ktoré sú verejne dostupné a boli zverejnené v mnohých médiách na Slovensku. Ide o volebné preferencie, konkrétne o dokumenty: Volebné preferencie politických strán – november 2011 a Volebné preferencie politických strán – január 2012. Budeme porovnávať volebné preferencie jednej strany v dvoch prieskumoch. Predpokladáme, že vzorky sú nezávislé. V prvom prieskume strana A získala p_x = 7.5 % v druhom p_y = 6.4 %, pričom v prvom prieskume bola vzorka respondentov, ktorí si vybrali jednu z politických strán n_x = 678 a v druhom n_y = 737. Zaujíma nás nasledovná dvojica hypotéz: H_0: $\pi_x - \pi_y \leq 0$ a H_1: $\pi_x - \pi_y > 0$. Pri platnosti alternatívnej hypotézy by bolo možné pokles preferencií považovať za štatisticky významný. Upozorňujeme, že ak pri výpočtoch použijeme funkciu `prop.test()`, dostaneme odlišné výsledky ako použitím vzorcov, ktoré sme uviedli vyššie (bližšie pozri Newcombe, 1998).

```
> pocet_A_november <- round(678*0.075)
> pocet_A_januar <- round(737*0.064)
```

```
> prop.test(x = c(pocet_A_november, pocet_A_januar), n = c(678,
  737), alternative = c("greater"), conf.level = 0.95, correct =
  F)

2-sample test for equality of proportions without continuity
                          correction

data:  c(pocet_A_november, pocet_A_januar) out of c(678, 737)
X-squared = 0.7181, df = 1, p-value = 0.1984
alternative hypothesis: greater
95 percent confidence interval:
-0.01083912  1.00000000
sample estimates:
    prop 1     prop 2
0.07522124 0.06377205
```

Z výsledkov je zrejmé, že nulovú hypotézu nevieme zamietnuť, inak povedané, na danej hladine významnosti nemáme dostatok dôkazov, aby sme mohli tvrdiť, že došlo k štatisticky významnému zníženiu preferencií pre politickú stranu A. Rozdiely mohli byť spôsobené náhodným výberom.

Jednoduchú funkciu, ktorá by počítala podľa vzťahov vyššie, si vieme ľahko zostrojiť. Vytvoríme pritom funkciu, v ktorej platí $\Delta_0 = 0$.

```
> tse_prop <- function(x = c(), n = c(), conf.level = 0.95) {
+ p_hat <- sum(x)/sum(n)
+ px <- x[1] / n[1]
+ py <- x[2] / n[2]
+ statistics <- (px - py) / sqrt(p_hat*(1 -
  p_hat)*(1/n[1]+1/n[2]))
+ print(paste("proportion x:", round(px, 4), "proportion y:",
  round(py, 4), "statistics:", round(statistics, 4)))
+ print(paste("critical values:", "lower", qnorm((1 -
  conf.level)/2), "upper", qnorm((1 + conf.level)/2)))
+ print(paste("p-value:", (1 - pnorm(abs(statistics)))*2))
+ print(paste("critical values:", "lower", qnorm(1 -
  conf.level)))
+ print(paste("p-value:", pnorm(statistics)))
+ print(paste("critical values:", "upper", qnorm(conf.level)))
+ print(paste("p-value:", 1 - pnorm(statistics)))
+ }
```

Vstupmi funkcie sú vektor *x*, ktorý predstavuje počet úspešných pokusov vo vzorkách, napr. *x* = c(50, 60) a vektor *n*, ktorý predstavuje počet pozorovaní, napr. *n* = c(100, 100). Posledná možnosť predstavuje možnosť zvoliť si určitú konfidenciu[18].

[18] Necháme na čitateľovi, aby si funkciu upravil tak, aby výsledky boli zobrazované v prehľadnejšej podobe. Môže použiť ako objekt výstupu z funkcie maticu, zoznam, prípadne zaokrúhliť počet desatinných miest.

```
> x <- c(50, 60)
> n <- c(100, 100)
> tse_prop(x = x, n = n, conf.level = 0.95)
[1] "proportion x: 0.5 proportion y: 0.6 statistics: -1.4213"
[1] "critical values: lower -1.95996398454005 upper
  1.95996398454005"
[1] "p-value: 0.155218489684684"
[1] "critical values: lower -1.64485362695147"
[1] "p-value: 0.0776092448423421"
[1] "critical values: upper 1.64485362695147"
[1] "p-value: 0.922390755157658"
> tse_prop(x = x, n = n, conf.level = 0.99)
[1] "proportion x: 0.5 proportion y: 0.6 statistics: -1.4213"
[1] "critical values: lower -2.5758293035489 upper
  2.5758293035489"
[1] "p-value: 0.155218489684684"
[1] "critical values: lower -2.32634787404084"
[1] "p-value: 0.0776092448423421"
[1] "critical values: upper 2.32634787404084"
[1] "p-value: 0.922390755157658"
```

4.3.9 Test dvoch podielov: závislé súbory

Okrem predchádzajúcich prípadov existuje možnosť štatisticky overiť podiel dvoch závislých vzoriek. Predpokladajme, že máme **dve závislé vzorky** náhodných premenných X_i, $i = 1, 2, ..., n_x$ a Y_i, $i = 1, 2, ..., n_y$. Pri závislých vzorkách máme dvojice pozorovaní (X_i, Y_i) a zjavne bude platiť $n_x = n_y$. Najprv si zostrojíme nasledujúcu kontingenčnú tabuľku:

Tabuľka 10: Kontingenčná tabuľka – test zhody podielov závislých vzoriek

		Y_i	
		$Y = 0$	$Y = 1$
X_i	$X = 1$	A	B
	$X = 0$	C	D

Zdroj: vlastné spracovanie podľa Kirk (2008)

kde A, B, C, D sú príslušné početnosti v tabuľkách. Napríklad A predstavuje počet štatistických jednotiek, pre ktoré sa v prvej vzorke namerala hodnota 1 (úspešný pokus) a v druhej vzorke hodnota 0 (neúspešný pokus). Ďalej nech $n = n_x + n_y$ a π_x odhadneme ako $\hat{p}_x = (A + B)/n$ a π_y ako $\hat{p}_y = (B + D)/n$. Potom rozdiel $\pi_x - \pi_y$ odhadujeme ako $(A + B)/n - (B + D)/n = A - D$. Testovacou charakteristikou je:

$$Z'''^{p} = \frac{A - D}{\sqrt{A + D}} \qquad (4.18)$$

Ak platí $(A + D) \geq 10$ pre obojstranný test a $(A + D) \geq 30$ pre jednostranný test, testovacia charakteristika Z'''^{π} sa riadi normovaným normálnym rozdelením pravdepodobnosti. Rozhodnutie o hypotéze potom uskutočníme nasledovne:

H_0: $\pi_x - \pi_y = 0$ H_1: $\pi_x - \pi_y \neq 0$	Hypotézu H_0 zamietame, ak $\lvert Z'''^{\pi} \rvert > \lvert z_{(\alpha/2)} \rvert$
H_0: $\pi_x - \pi_y \leq 0$ H_1: $\pi_x - \pi_y > 0$	Hypotézu H_0 zamietame, ak $Z'''^{\pi} > z_{(1-\alpha)}$
H_0: $\pi_x - \pi_y \geq 0$ H_1: $\pi_x - \pi_y < 0$	Hypotézu H_0 zamietame, ak $Z'''^{\pi} < z_{(\alpha)}$

Nasledujúci príklad využívajú údaje prevzaté od Everitt – Hothorn (2009). Nainštalovaním programového balíka `HSAUR` a jeho spustením (pomocou príkazu `library(HSAUR)`) sa vieme dostať priamo ku kontingenčnej tabuľke:

Tabuľka 11: Proces s mladistvými

		Súd pre mladistvých	
		Nebol znovu zatknutý	*Znovu zatknutý*
Súd pre dospelých	*Znovu zatknutý*	*515*	*158*
	Nebol znovu zatknutý	*1134*	*290*

Zdroj: upravené podľa Everitt – Hothorn (2009)

Údaje v tomto príklade predstavujú párové vzorky mladistvých, ktorí boli v roku 1987 v štáte Florida (USA) zatknutí. Pár je vytvorený podľa podobnosti trestného činu, ktorého sa mladiství mali dopustiť. Máme teda situáciu takej závislej vzorky, kde v jednotlivých vzorkách vystupujú rôzne subjekty. Závislosť spočíva v podobnosti trestného činu. Ak máme dvoch mladistvých, ktorí tvoria usporiadanú dvojicu, pričom prvý bol súdený súdom pre dospelých a do roka bol znovu zatknutý, kým druhý bol súdený súdom pre mladistvých a nebol do roka zatknutý, tak sa táto dvojica ocitla v prvej bunke (A). Spolu bolo takýchto dvojíc 515. Cieľom je zistiť, či je podiel znovu zatknutých mladistvých rovnaký, ak boli súdení súdom pre dospelých, resp. súdom pre mladistvých. Manuálne výpočet uskutočníme nasledovne:

```
> test_statistics <- (515 - 290)/sqrt(515 + 290)
> 1 - pnorm(test_statistics) + (pnorm(-test_statistics))
[1] 2.204119e-15
```

Možné je použiť aj funkciu `mcnemar.test()`, ktorá používa Chí-kvadrát test. Výsledky sú podobné.

```
> table <- matrix(c(515, 1134, 158, 290), nrow = 2)
> table
     [,1] [,2]
[1,]  515  158
[2,] 1134  290
-----------------------------------------------------------------
> mcnemar.test(table, correct = F)

                 McNemar's Chi-squared test

data:  table
McNemar's chi-squared = 737.2879, df = 1, p-value < 2.2e-16
```

4.4 Testy dobrej zhody

V predchádzajúcich testoch o strednej hodnote sme v niektorých prípadoch predpokladali, že nami pozorované premenné pochádzajú z normálneho rozdelenia pravdepodobnosti. Ak neexistuje iný, deduktívny dôvod, je možné tento predpoklad formálne overiť pomocou štatistických testov. V tejto časti sa budeme venovať niekoľkým takýmto testom, ktoré overujú tvar rozdelenia pravdepodobnosti nami pozorovaných hodnôt s teoretickým rozdelením alebo s iným empirickým rozdelením pravdepodobnosti. Hypotézy, ktoré budú predmetom šetrenia, môžeme vo všeobecnosti definovať nasledovne:

- H_0: rozdelenie pravdepodobnosti výberového súboru je totožné s teoretickým rozdelením pravdepodobnosti (prípadne s rozdelením pravdepodobnosti iného výberového súboru).
- H_1: rozdelenie pravdepodobnosti výberového súboru je odlišné od teoretického rozdelenia pravdepodobnosti (prípadne od rozdelenia pravdepodobnosti iného výberového súboru).

Začneme pomerne všeobecnými testami a neskôr sa zameriame na najčastejšie používané testy, ktoré boli navrhnuté za účelom testovania, či výberový súbor mohol pochádzať priamo z normálneho rozdelenia pravdepodobnosti. Testom na overovanie normality existuje nepomerne viac ako prezentujeme v tejto kapitole. V nasledujúcom krátkom zozname niektorých **ďalších** testov, ktorými sa bližšie zaoberať nebudeme, uvádzame aj príslušné funkcie a programové balíky, v ktorých sa testy nachádzajú, a pomocou ktorých je možné testy realizovať v programe R:

- Bonett – Seier test: `bonett.test()`, knižnica `moments`.
- SJ test: `sj.test()`, knižnica `lawstat`.
- Shapiro – Francia test: `sf.test()`, knižnica `nortest`.

- Cramer – von Mises test: `cvm.test()`, knižnica `nortest`.
- Lilliefors test: `lillie.test()`, knižnica `nortest`.

4.4.1 Pearsonov Chí-kvadrát test dobrej zhody

Vychádzajme z náhodnej *iid* vzorky X_i, $i = 1, 2, ..., n$. Ďalej predpokladajme, že hodnoty tejto náhodnej vzorky vieme usporiadať do frekvenčnej tabuľky s k číselnými intervalmi tak, že každú jednu hodnotu X_i vieme priradiť práve do jedného z číselných intervalov. Postup vytvorenia týchto intervalov môže byť obdobný, aký sa používa pri tvorbe frekvenčnej tabuľky s číselnými intervalmi.[19] Z nasledujúcich hodnôt sme vytvorili tabuľku.

```
data <- c(2400, 3100, 7200, 3100, 5200, 6200, 4000, 3200, 5300,
  3900, 3600, 5300, 5400, 3300, 4700, 3000, 3300, 4400, 4200,
  5700, 4700, 4900, 3900, 4300, 3900, 2000, 4800, 4100, 4100,
  3800, 6400, 2500, 4100, 4400, 3800, 2600, 5200, 4900, 4500)
```

Tabuľka 12: Intervalové triedy, pozorovaná a očakávaná početnosť

Interval	Triedny znak	Pozorovaná početnosť	Teoretická pravdepodobnosť	Očakávaná početnosť
<2000;3040)	2520	5	0.10	4
<3040;4080)	3560	13	0.25	10
<4080;5120)	4600	13	0.30	12
<5120;6160)	5640	6	0.25	10
<6160;7200>	6680	3	0.10	4
Spolu (n)		40	1	40

Zdroj: Lyócsa et al. (2013)

Stĺpec „Interval“ predstavuje číselné intervaly, do ktorých vieme priradiť pozorované hodnoty X_i. Stĺpec „Triedny znak“ je priemer dolnej a hornej hranice príslušného intervalu. Stĺpec „Pozorovaná početnosť“ predstavuje skutočný počet hodnôt, ktoré sa nachádzajú v príslušnom číselnom intervale.

Princíp Pearsonovho Chí-kvadrát testu spočíva v odhade očakávanej početnosti v príslušných číselných intervaloch. Očakávaná početnosť dáva odpoveď na otázku: Koľko hodnôt by sa malo nachádzať v príslušnom intervale, ak by platilo, že hodnoty pochádzajú z teoretického rozdelenia pravdepodobnosti (rozdelenia, ktoré testujeme)? Čím je väčší rozdiel medzi pozorovanou početnosťou a očakávanou početnosťou, o to menej budeme mať tendenciu veriť tomu, že nami pozorované hodnoty pochádzajú z teoretického rozdelenia pravdepodobnosti. Stĺpec „Teoretická pravdepodobnosť“ predstavuje pravdepodobnosť, že ak náhodne vyberieme hodnotu z teoretického rozdelenia pravdepodobnosti, tak bude patriť do

[19] Pozri Lyócsa et al. (2013).

príslušného intervalu. V stĺpci „Očakávaná početnosť“ dostaneme počty tak, že teoretickú pravdepodobnosť vynásobíme celkovým rozsahom štatistického súboru. Pozorovanú početnosť si označíme ako o_j, teoretickú pravdepodobnosť ako p_j, celkový rozsah súboru ako n, očakávanú početnosť si označíme ako e_j, pričom platí $e_j = np_j$. Overujeme hypotézu[20]:

H_0: rozdelenie pravdepodobnosti výberového súboru je totožné s teoretickým rozdelením pravdepodobnosti.

H_1: rozdelenie pravdepodobnosti výberového súboru je odlišné od teoretického rozdelenia pravdepodobnosti.

S týmto zápisom hypotéz si v tejto publikácii vystačíme. Pre úplnosť uvedieme formálnejší tvar: H_0: $o_j = e_j$, $j = 1, 2, ..., k$, oproti H_1: $\exists j, o_j \neq e_j$.

Testovacia charakteristika má tvar:

$$\chi_P^2 = \sum_{j=1}^{k} \frac{\left(o_j - np_j\right)^2}{np_j} \qquad (4.19)$$

kde $j = 1, 2, ..., k$ predstavuje príslušný číselný interval. Testovacia charakteristika χ_P^2 sa pre dostatočne veľké n dá aproximovať Chí-kvadrát (χ^2) rozdelením pravdepodobnosti. Rozhodnutie o hypotéze vykonáme nasledovne:

H_0: $o_j = e_j$ H_1: $\exists j, o_j \neq e_j$	Hypotézu H_0 zamietame, ak $\chi_P^2 > \chi^2_{(1-\alpha),(k-1-r)}$

kde r je počet parametrov, ktoré je potrebné z pozorovaných údajov odhadnúť, aby sme vypočítali p_j. Ak by sme predpokladali, že teoretické rozdelenie pravdepodobnosti sa riadi Poissonovým rozdelením $r = 1$ (parameter λ), pre normálne rozdelenie pravdepodobnosti $r = 2$ (parameter μ a σ^2). Predtým ako uskutočníme samotný test, je vo všeobecnosti potrebné zabezpečiť, aby platilo $e_j = np_j > 5$. Za určitých podmienok existujú aj menej prísne kritériá (pozri napr. Tkáč, 2001). V prípade, ak táto podmienka nie je splnená, môžeme dva alebo viac susediacich intervalov spojiť tak, aby súčet ich početností spĺňal túto podmienku. V prípade, ak je teoretické rozdelenie pravdepodobnosti diskrétne, „intervaly“ môžu predstavovať jedno číslo. V prípade spojitých rozdelení je štandardným postupom tvorba číselných intervalov.

[20] Pri tomto teste nie je nutné, aby sme porovnávali rozdelenie pravdepodobnosti výberového súboru s teoretickým rozdelením. Môžeme porovnávať aj s ľubovoľnými „očakávanými“ početnosťami. Dôvod prečo sme takto naformulovali hypotézy spočíva v účele tejto kapitoly.

Príklad 4.10

Aplikáciu Pearsonovho Chí-kvadrát testu si ukážeme na príklade od Panik (2005). Majme nasledujúce údaje:

```
umrtia <- c(rep(0, 56), rep(1, 68), rep(2, 33), rep(3, 11),
  rep(4, 2), rep(5, 2), rep(6, 1))
```

Údaje predstavujú počet úmrtí z dopravných nehôd na 173 cestách za obdobie 100 dní. Zaujíma nás, či je možné pozorovaný počet úmrtí modelovať Poissonovým rozdelením pravdepodobnosti. V nasledujúcom postupe sme najprv vytvorili z údajov tabuľku, potom vypočítali parameter Poissonovho rozdelenia λ. Následne sme vypočítali očakávanú pravdepodobnosť použitím funkcie `dpois()`. Zvyšná časť predstavuje výpočet testovacej charakteristiky, pričom sme zlúčili 4 intervaly do jedného, aby sme zachovali podmienku $e_j = np_j > 5$. Keďže testovacia štatistika nevyšla väčšia ako kritická hodnota, hypotézu H_0 nevieme zamietnuť, a preto modelovanie počtu úmrtí Poissonovým rozdelením budeme považovať za možné. V prípade, ak by sme Pearsonovým Chí-kvadrát testom overovali prítomnosť normálneho rozdelenia, môžeme použiť funkcie `pearson.test()` z programového balíka nortest.

```
> umrtia <- c(rep(0, 56), rep(1, 68), rep(2, 33), rep(3, 11),
  rep(4, 2), rep(5, 2), rep(6, 1))
> oj <- table(umrtia); oj
umrtia
 0  1  2  3  4  5  6
56 68 33 11  2  2  1
> lambda <- mean(umrtia)
> hodnoty <- as.numeric(names(oj))
> ej <- dpois(hodnoty, lambda = lambda)*sum(oj)
> ej <- c(ej[1:3], ej[4] + ej[5] + ej[6] + ej[6])
> oj <- c(oj[1:3], oj[4] + oj[5] + oj[6] + oj[6])
> statistics <- sum((ej - oj)^2/ej)
> statistics
[1] 0.5406547
> qchisq(0.95, df = 2)
[1] 5.991465
```

4.4.2 Kolmogorov – Smirnovov test dobrej zhody jedného výberového súboru

Hlavnou výhodou Pearsonovho Chí-kvadrát testu dobrej zhody je jeho univerzálnosť. Empirické rozdelenie môžeme porovnávať s teoretickým diskrétnym, spojitým rozdelením, alebo dokonca s kategorickými početnosťami. Rozdelenia, s ktorými výberový súbor porovnávame, pritom môžu a nemusia byť definované úplne. To znamená, že nie je nutné,

aby boli špecifikované parametre teoretického rozdelenia. Tie sa môžu odhadnúť pomocou pozorovaných údajov. Na druhej strane, výsledky z Pearsonovho Chí-kvadrát testu vedia byť ovplyvnené počtom a šírkou číselných intervalov, ktoré sa viac menej volia subjektívne. Pri spojitých premenných vzniká ďalej problém so spájaním číselných intervalov. Cenou za univerzálnosť sú slabšie štatistické vlastnosti testu (chyba I. a II. druhu) voči špecializovaným, menej univerzálnym testom. Ak je náhodná premenná, z ktorej vyberáme výberový súbor X_i spojitá[21], potom vhodnejším testom je tzv. Kolmogorov – Smirnovov test dobrej zhody. Na rozdiel od Pearsonovho Chí-kvadrát testu dobrej zhody však potrebujeme, aby teoretické rozdelenie pravdepodobnosti bolo špecifikované vrátane svojich parametrov. Ak sa overuje, či výberový súbor môže pochádzať z normálneho rozdelenia pravdepodobnosti, je potrebné špecifikovať jeho strednú hodnotu a rozptyl. Ďalej je potrebné, aby teoretické rozdelenie malo definovanú distribučnú funkciu $F(X)$ a taktiež, aby z výberového súboru bolo možné zostrojiť empirickú distribučnú funkciu $\hat{F}_n(X)$. Overuje sa hypotéza H_0: $F = F_t$ oproti alternatíve H_1: $F \neq F_t$, kde F_t je distribučná funkcia teoretického rozdelenia pravdepodobnosti. Princíp výpočtu Kolmogorov – Smirnovovho testu spočíva v porovnávaní distribučných funkcií. Čím je rozdiel medzi nimi väčší, o to máme väčšiu tendenciu sa prikloniť k záveru, že ide o dve rôzne rozdelenia. Testovacia štatistika má nasledujúci tvar:

$$D_n = \sup_x \left| \hat{F}_n(X) - F_t(X) \right| \qquad (4.20)$$

Zo zápisu je zrejmé, že sa vyberá najmenšie horné ohraničenie (supremum) rozdielov (vzdialeností) distribučných funkcií. Vzhľadom na to, že empirická distribučná funkcia sa spravidla definuje ako spojitá sprava, supremum sa dosiahne nasledujúcim výrazom:

$$D_n = \max_x \left\{ \left| \hat{F}_n(X) - F_t(X) \right|, \left| \hat{F}_n(X - \varepsilon) - F_t(X) \right| \right\} \qquad (4.21)$$

kde $\varepsilon > 0$ a výraz $\hat{F}_n(X - \varepsilon)$ predstavuje funkčnú hodnotu pre najbližšiu menšiu hodnotou k X, nachádzajúcu sa v empirickom štatistickom súbore. Rozhodnutie o hypotéze uskutočníme nasledovne:

[21] Číselné hodnoty výberového súboru by mali byť aspoň poradového charakteru. Napríklad pri dotazníkovom šetrení respondenti odpovedajú na škále od 1 – 7. Namerané hodnoty majú charakter poradí. To však neznamená, že v skutočnosti konštrukt, ktorý sa otázkou meral, nie je možné považovať za spojitú náhodnú premennú.

H_0: $F = F_t$	Hypotézu H_0 zamietame, ak $D_n > d_{n,(1-\alpha)}$
H_1: $F \neq F_t$	

Za predpokladu platnosti nulovej hypotézy sa testovacia štatistika *(4.20)* a *(4.21)* riadi známym rozdelením. Keďže však ide o pomerne komplikované rozdelenie, kritické hodnoty sa pre rôzne veľkosti vzorky a hladinu významnosti α uvádzajú spravidla v tabuľkách. Pre vzorky $n > 40$, je možné použiť nasledujúcu aproximáciu pre $\alpha = 0.20, 0.15, 0.10, 0.05, 0.01$: $d_{n,(1-\alpha)} = 1.07/c$, $d_{n,(1-\alpha)} = 1.14/c$, $d_{n,(1-\alpha)} = 1.22/c$, $d_{n,(1-\alpha)} = 1.36/c$, $d_{n,(1-\alpha)} = 1.63/c$, kde pre parameter c platí:

$$c = \sqrt{n + \sqrt{n/10}} \qquad (4.22)$$

V programe R slúži na výpočet testu funkcia `ks.test()`. Upozorňujeme, že Kolmogorov – Smirnovov test nepredpokladá rovnosť hodnôt. To znamená, že žiadne dve hodnoty v rámci toho istého výberového súboru by nemali byť rovnaké. Ak je táto podmienka porušená, kritické hodnoty (ktoré sú uvádzané v štatistických tabuľkách) nie sú presné. Porušenie tohto predpokladu sa však nepovažuje za kritické. V programe R nás výstup z funkcie `ks.test()` upozorní na možný problém s rovnakými hodnotami. Alternatívou je použiť funkciu `ks.boot()` z programového balíka `Matching`, kde sa počíta upravená verzia Kolmogorov – Smirnovovho testu, kde sa pre potreby počítania kritických hodnôt (a *p*-hodnoty) využíva bootstraping.

Príklad 4.11

Najprv si ilustrujeme postup výpočtu bez použitia funkcie `ks.test()`. Uvažujme o údajoch, ktoré predstavujú denné výnosy akcií (v príklade sme sa inšpirovali prácou od van den Honert, 1997). Našim cieľom je zistiť, či tieto denné výnosy môžeme považovať za realizácie z normálneho rozdelenia pravdepodobnosti so strednou hodnotou 0 a smerodajnou odchýlkou 2.2.

```
> return <- c(rep(-5, 5), rep(-3, 28), rep(-1, 77), rep(1, 196),
  rep(3, 62), rep(5, 32))
```

Najprv si vytvoríme tabuľku, ktorú sme si označili ako „`table`“. Následne sme vytvorili vektor kumulatívnej relatívnej početnosti „`Fex`“ ($\hat{F}_n(X)$). Vektor „`Fex_e`“ ($\hat{F}_n(X-\varepsilon)$) predstavuje vektor „`Fex`“ posunutý o jedno pozorovanie. Vektor „`Fx`“ je vektorom teoretickej kumulatívnej početnosti.

```
> table <- table(return)/400
> Fex <- as.numeric(cumsum(table)); Fex
[1] 0.0125 0.0825 0.2750 0.7650 0.9200 1.0000
> Fex_e <- c(0, Fex[1:(length(Fex)-1)]); Fex_e
[1] 0.0000 0.0125 0.0825 0.2750 0.7650 0.9200
> Fx <- c(pnorm(c(-5, -3, -1, 1, 3, 5), mean = 0, sd = 2.2)); Fx
[1] 0.01152131 0.08634102 0.32471814 0.67528186 0.91365898
  0.98847869
> Dn <- max(abs(Fex - Fx), abs(Fex_e - Fx)); Dn
[1] 0.4002819
```

Hodnota testovacej charakteristiky je 0.4002819, pričom kritickú hodnotu pre hladinu významnosti $\alpha = 0.05$ a našu vzorku o veľkosti $n = 400$ si vieme vypočítať ako:

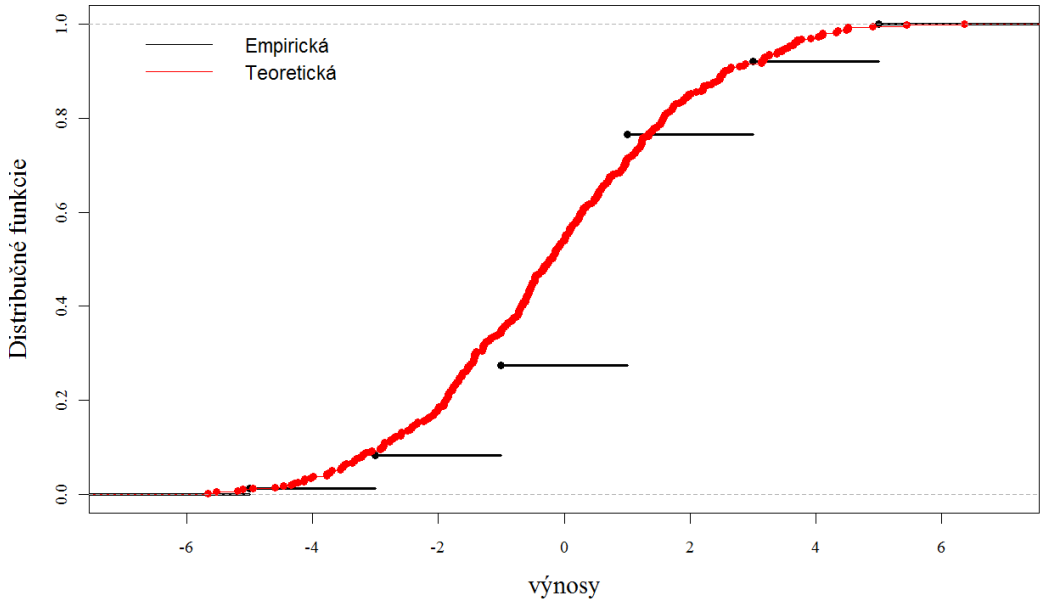
```
> n <- length(return)
> c <- sqrt(n+sqrt(n/10))
> critical <- 1.36/c; critical
[1] 0.0674687
```

Hypotézu H_0 teda zamietame. Na ilustráciu uvádzame nasledujúci obrázok (pozri Obrázok 4.6), ktorý znázorňuje obe distribučné funkcie, empirickú aj teoretickú pre $N(0, 2.2^2)$.

```
> plot.ecdf(ecdf(return), col.01line = "black", lwd = 2.5, pch =
  19, xlab = "výnosy", family = "serif", cex.lab = 1.5, main
  ="", ylab = "Distribučné funkcie")
> lines(ecdf(rnorm(400, mean = 0, sd = 2.2)), col = "red", lwd =
  0.5)
> legend("topleft", legend = c("Empirická", "Teoretická"), col =
  c("black", "red"), inset = 0.025, bty = "n", cex = 1.2, lty =
  1)
```

Použitím funkcie `ks.test()` dostávame podobné výsledky, avšak namiesto kritickej hodnoty nám funkcia vráti priamo *p*-hodnotu. Zároveň sa objaví varovanie týkajúce sa rovnakých hodnôt.

```
> ks.test(return, "pnorm", mean = 0, sd = 2.2)

            One-sample Kolmogorov-Smirnov test

data:  return
D = 0.4003, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(return, "pnorm", mean = 0, sd = 2.2) :
cannot compute correct p-values with ties
```



Obrázok 4.6: Empirická a teoretická distribučná funkcia – porovnanie

Zdroj: upravené podľa van den Honert (1997)

Pre úplnosť uvedieme ešte použitie funcie `ks.boot()`. Prvým argumentom sú vstupné údaje, druhým sú náhodné realizácie z normálneho rozdelenia pravdepodobnosti s požadovanými parametrami. Argument `nboots` predstavuje počet bootstrap vzoriek. Výsledky sú obdobné ako v predošlých prípadoch.

```
> library(Matching)
> ks.boot(return, rnorm(400, mean = 0, sd = 2.2), nboots = 1000)
$ks.boot.pvalue
[1] 0

$ks

                Two-sample Kolmogorov-Smirnov test

data:  Tr and Co
D = 0.385, p-value < 2.2e-16
alternative hypothesis: two.sided

$nboots
[1] 1000

attr(,"class")
[1] "ks.boot"
```

4.4.3 Kolmogorov – Smirnovov test dobrej zhody dvoch výberových súborov

Majme dva nezávislé výberové súbory Z_i, $i = 1, 2, ..., n_z$ a Y_j, $j = 1, 2, ..., n_y$, ktorých hodnoty sú náhodnými realizáciami z určitého spojitého rozdelenia pravdepodobnosti s distribučnými funkciami G a H. Kolmogorov – Smirnovov test dobrej zhody dvoch výberových súborov overuje hypotézu (zaujímajú nás teraz iba obojstranné hypotézy) H_0: $G = H$ oproti alternatíve H_1: $G \neq H$. Všimnime si, že pri *t*-teste dvoch stredných hodnôt porovnávame miery polohy rozdelení, kým pri Kolmogorov – Smirnovovom teste nás zaujíma nie len poloha, ale aj tvar rozdelenia (rozptyl, šikmosť, špicatosť, ...), Hill – Lewicki (2006). Preto sa v niektorých aplikáciách považuje tento test za vhodnú alternatívu k obojstrannému *t*-testu. Ďalej definujme empirické distribučné funkcie $\hat{G}_{nz}(x)$ a $\hat{H}_{ny}(x)$, pričom n_z, $n_y \geq 50$ a x predstavujú konkrétne realizácie Z_i a Y_j. Podobne ako v predošlom prípade, aj tu potrebujeme, aby boli pozorované hodnoty merané aspoň na poradovej škále. Testovacia charakteristika má nasledujúci tvar:

$$D_{nz,ny} = \max_{x \in R} \left\{ \left| \hat{G}_{nz}(x) - \hat{H}_{ny}(x) \right| \right\} \qquad (4.23)$$

Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $G = H$ H_1: $G \neq H$	Hypotézu H_0 zamietame, ak $D_{nz,\,ny} > d_{nz,ny,\,(1-\alpha)}$

kde $d_{nz,ny,\,(1-\alpha)}$ sú kritické hodnoty spravidla uvádzané v štatistických tabuľkách. V programe R na výpočet použijeme funkciu `ks.test()`, ktorá nám vráti priamo *p*-hodnotu.

Príklad 4.12

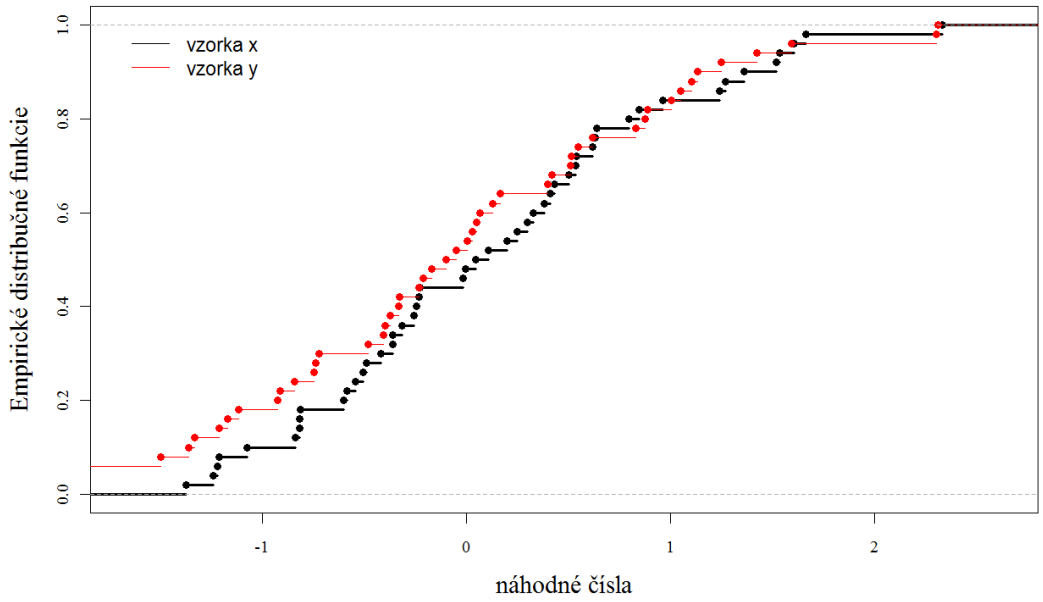
Uvažujme o dvoch súboroch, ktoré vyberáme z normovaného normálneho rozdelenia, oba o rozsahu $n = 50$. Na ich vygenerovanie použijeme v programe R funkciu `rnorm()`.

```
> x <- rnorm(50)
> y <- rnorm(50)
> ks.test(x, y, alternative = c("two.sided"))

              Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.16, p-value = 0.5487
alternative hypothesis: two-sided
```

Hypotézu H_0 nevieme zamietnuť. Keďže sme oba súbory „umelo“ vytvorili z rovnakého rozdelenia, vieme, že ide o správne rozhodnutie.

V niektorých publikáciách sa stretneme s používaním tejto podoby testu aj pre vzorky, kde n_z, n_y < 50. V týchto prípadoch je však možné použiť presnejšie štatistiky, pozri napr. Tkáč (2001). Vo všeobecnosti považujeme Kolmogorov – Smirnovov test pre dva výberové súbory za vhodný pre použitie viac početných vzoriek. Na záver uvádzame empirické kumulatívne distribučné funkcie pre obe vzorky (pozri Obrázok 4.7).

```
> plot.ecdf(ecdf(x), col.01line = "black", lwd = 2.5, pch = 19,
  xlab = "náhodné čísla", family = "serif", cex.lab = 1.5, main
  ="", ylab = "Empirické distribučné funkcie")
> lines(ecdf(y), col = "red", lwd = 0.5)
> legend("topleft", legend = c("vzorka x", "vzorka y"), col =
  c("black", "red"), inset = 0.025, bty = "n", cex = 1.2, lty =
  1)
```



Obrázok 4.7: Empirické distribučné funkcie – zhoda dvoch rozdelení

Zdroj: vlastné spracovanie, výstup zo softvéru R

4.4.4 Anderson – Darlingov test

Anderson – Darlingov test používame na overovanie hypotézy, že hodnoty z *iid* náhodnej vzorky X_i, i = 1, 2, ..., n pochádzajú z teoretického rozdelenia pravdepodobnosti $F_t(X)$, pričom budeme uvažovať, že dané teoretické rozdelenie je normálne rozdelenie pravdepodobnosti, teda $F_N(X)$. Ďalej nech platí $n \geq 8$. Postup výpočtu Anderson –

Darlingovho testu môžeme zhrnúť do štyroch bodov (pozri Cullen – Frey, 1999; Kvam – Vidakovic, 2007):

1) Vytvorme variačný rad $X_{(i)}$.
2) Vypočítajme štandardizované hodnoty $p_{(i)}$ z $X_{(i)}$.
3) Vypočítajme testovaciu štatistiku Anderson – Darlingovho testu A^2 a jej modifikáciu A^*.
4) Porovnajme modifikovanú štatistiku s kritickou hodnotou a rozhodnime o hypotéze.

Štandardizácia sa uskutoční pomocou:

$$Y_{(i)} = \frac{X_{(i)} - \overline{X}}{s} \qquad (4.24)$$

Následne sa vypočíta kumulatívna pravdepodobnosť pre každé $Y_{(i)}$:

$$p_{(i)} = \Phi\left(Y_{(i)}\right) \qquad (4.25)$$

kde $\Phi(.)$ je kumulatívna distribučná funkcia normovaného normálneho rozdelenia pravdepodobnosti. Výpočet testovacej štatistiky je potom:

$$A^2 = -\sum_{i=1}^{n} \frac{(2i-1)\left(\ln\left(p_{(i)}\right) + \ln\left(1 - p_{(n+1-i)}\right)\right)}{n} - n \qquad (4.26)$$

Posledným krokom je výpočet modifikovanej štatistiky:

$$A^* = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \qquad (4.27)$$

Rozhodnutie o hypotéze je potom nasledovné:

H_0: $F(X) = F_N(X)$ H_1: $F(X) \neq F_N(X)$	Hypotézu H_0 zamietame, ak $A^* > Ac_\alpha$

kde Ac_α je kritická hodnota, ktorá pre α = 0.10, 0.05, 0.025, 0.01, 0.005 zodpovedá hodnotám 0.631, 0.752, 0.873, 1.035, 1.159. V programe R môžeme test uskutočniť pomocou funkcie `at.test()`, ktorá je súčasťou programového balíka `nortest`. Anderson – Darlingov test je možné použiť aj pre porovnanie s inými rozdeleniami ako s normálnym, avšak vyžaduje si to prepočítať kritické hodnoty.

Príklad 4.13

Aplikáciu Anderson – Darglingovho testu si ukážeme na dvoch vzorkách. Použijeme pritom databázu `babies` a premenné `age` (vek prvorodičiek) a `wt` (váhu novorodencov). Váhu novorodencov si prevedieme na *kg*.

```
> library(UsingR); attach(babies)
> names(babies)
 [1] "id"        "pluralty"  "outcome"   "date"      "gestation"
  "sex"
 [7] "wt"        "parity"    "race"      "age"       "ed"
  "ht"
[13] "wt1"       "drace"     "dage"      "ded"       "dht"
  "dwt"
[19] "marital"   "inc"       "smoke"     "time"      "number"
> age_new <- subset(babies$age, subset = age != 99)
> wt_new <- wt*28.3495231/1000
```

Zaujíma nás, či je možné považovať vek prvorodičiek za realizácie z normálneho rozdelenia pravdepodobnosti. Histogram a kumulatívna distribučná funkcia naznačujú, že tomu tak nie je. Rozdelenie sa javí ako pravostranne zošikmené. Pre vek si vyskúšame zopakovať postup výpočtu testovacej štatistiky A^* priamo pomocou definovaných vzťahov. V prvom kroku zoradíme vek do variačného radu, potom vek štandardizujeme a vypočítame kumulatívnu pravdepodobnosť.

```
> a <- sort(age_new, decreasing = F)
> a <- (a - mean(a))/sd(a)
> p <- pnorm(a)
> length(p)
[1] 1234
```

Následne vypočítame testovaciu štatistiku.

```
> b <- c()
> for (i in 1:length(p)) {
+ c <- (2*i-1)*(log(p[i]) + log(1 - p[length(p)+1-i]))/length(p)
+ b <- c(b, c)
+ }
> -sum(b)-length(p)
[1] 13.39824
```

Porovnaním s kritickou hodnotou pre $\alpha = 0.05$ je zrejmé, že hypotézu H_0 zamietame, teda na základe našich dát, nie je možné predpokladať, že vek prvorodičiek je realizáciou z normálneho rozdelenia pravdepodobnosti. Výsledok si môžeme overiť použitím funkcie `ad.test()`.

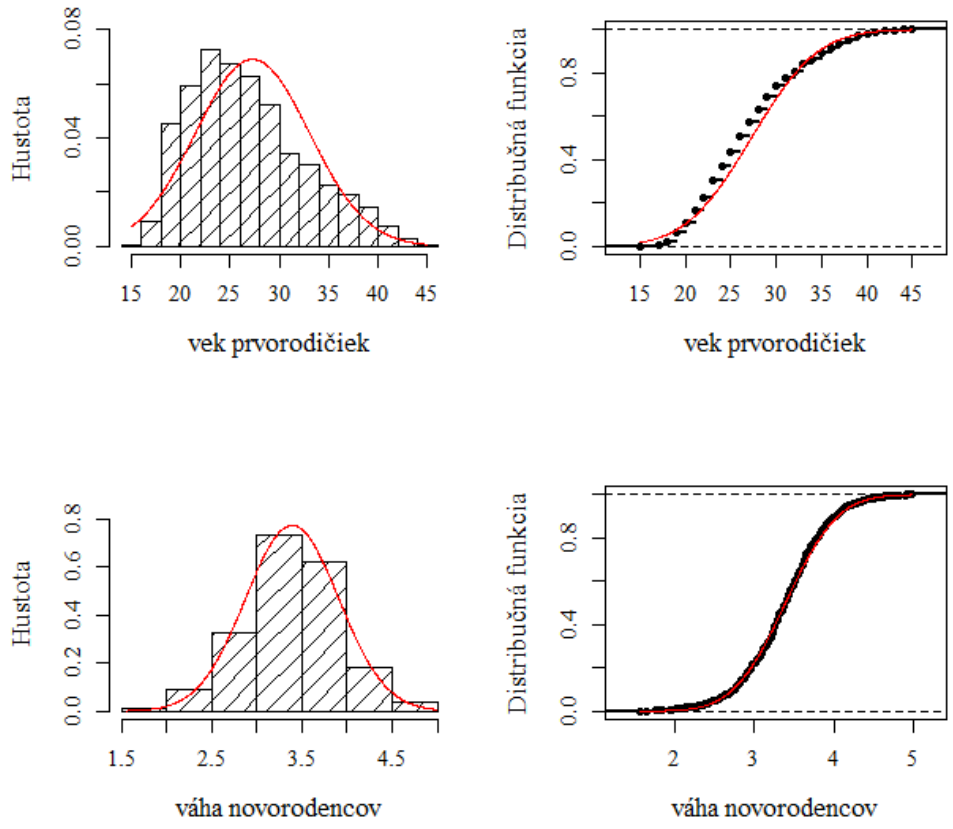
```
> library(nortest)
> ad.test(age_new)

        Anderson-Darling normality test

data:  age_new
```

```
A = 13.3982, p-value < 2.2e-16
```

Testovacia štatistika je prakticky totožná, pričom *p*-hodnota je veľmi nízka a menšia ako hladina významnosti $\alpha = 0.05$. Na základe histogramu a empirickej distribučnej funkcie sa javí váha novorodencov ako viac symetrické rozdelenie. Test však znova (aj keď nie tak výrazne) zamieta nulovú hypotézu.

```
> ad.test(wt_new)

                Anderson-Darling normality test

data:  wt_new
A = 1.4677, p-value = 0.0008669
```

Na záver uvádzame kódy pre vizualizáciu, ktorá neraz pomáha odhadnúť správny pravdepodobnostný model.



Obrázok 4.8: Histogram a distribučná funkcia veku prvorodičiek a váhy novorodencov

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(2, 2))
> hist(age_new, density = 10, col = "black", main = NA, ylim =
  c(0, 0.08), cex.lab = 1.5, cex.axis = 1.3, freq = FALSE, ylab
  = "Hustota", family = "serif", xlab = "vek prvorodičiek")
> x <- seq(min(age_new), max(age_new), length = 1000)
> xh <- dnorm(x, mean = mean(age_new), sd = sd(age_new))
> data <- data.frame(x,xh)
> lines(data, type = "l", col = "red")
> plot(ecdf(age_new), col.01line = "black", lwd = 2.5, pch = 19,
  xlab = "vek prvorodičiek", family = "serif", cex.lab = 1.5,
  cex.axis = 1.3, main ="", ylab = "Distribučná funkcia")
> xh <- pnorm(x, mean = mean(age_new), sd = sd(age_new))
> data <- data.frame(x,xh)
> lines(data, col = "red", lwd = 0.5)
> hist(wt_new, density = 10, col = "black", main = NA, ylim =
  c(0, 0.90), cex.lab = 1.5, cex.axis = 1.3, freq = FALSE, ylab
  = "Hustota", family = "serif", xlab = "váha novorodencov")
> x <- seq(min(wt_new), max(wt_new), length = 1000)
> xh <- dnorm(x, mean = mean(wt_new), sd = sd(wt_new))
> data <- data.frame(x,xh)
> lines(data, type = "l", col = "red")
> plot(ecdf(wt_new), col.01line = "black", lwd = 2.5, pch = 19,
  xlab = "váha novorodencov", family = "serif", cex.lab = 1.5,
  cex.axis = 1.3, main ="", ylab = "Distribučná funkcia")
> xh <- pnorm(x, mean = mean(wt_new), sd = sd(wt_new))
> data <- data.frame(x,xh)
> lines(data, col = "red", lwd = 0.5)
```

4.4.5 Shapiro – Wilkov test

Shapiro – Wilkov test bol vyvinutý za špecifickým účelom testovania, či *iid* náhodná vzorka X_i, i = 1, 2, ..., n pochádza z normálneho rozdelenia pravdepodobnosti. Na jednej strane sa hodí na testovanie iba normality, na strane druhej má lepšie štatistické vlastnosti ako Kolmogorov – Smirnovov test. Výpočet Shapiro – Wilkovho testu je komplikovanejší v porovnaní s predchádzajúcimi testami. Ukážeme postup od Royston (1992), Güner – Johnson (2007) a Kvam – Vidakovic (2007). Pri výpočte testovacej charakteristiky Shapiro – Wilkovho testu potrebujeme získať hodnoty koeficientov a_i. Tie sa uvádzajú pre rôzne veľkosti vzoriek v tabuľkách. Royston (1992) navrhol aproximáciu, pomocou ktorej je možné tieto hodnoty odhadnúť.

Vychádzajme z variačného radu $X_{(i)}$. Následne si vypočítajme:

$$m_i = \Phi^{-1}\left((i-3/8)/(n+1/4)\right) \qquad (4.28)$$

kde Φ^{-1} je inverzná funkcia distribučnej funkcie normovaného normálneho rozdelenia (kvantilová funkcia). Ďalej:

$$\varepsilon = \frac{\sum_{i=1}^{n} m_i^2 - 2m_n^2 - 2m_{n-1}^2}{1 - 2a_n^2 - 2a_{n-1}^2} \qquad (4.29)$$

Ak $u = n^{-1/2}$, potom koeficienty a_i vypočítame ako:

$$a_n = -2.7060556u^5 + 4.434685u^4 - 2.071190u^3 - 0.147981u^2 + 0.221157u + c_n \qquad (4.30)$$

$$a_{n-1} = -3.582633u^5 + 5.682633u^4 - 1.752461u^3 - 0.293762u^2 + 0.042981u + c_{n-1} \qquad (4.31)$$

$$a_i = \varepsilon^{-1/2} m_i \qquad (4.32)$$

s tým, že vzťah *(4.32)* platí pre $i = 3, 4, ..., n - 2$. Posledným pomocným prepočtom sú koeficienty c_n, c_{n-1}:

$$c_i = \frac{m_i}{\sqrt{\sum_{i=1}^{n} m_i^2}} \qquad (4.33)$$

Pôvodnou testovacou charakteristikou bol výraz:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_{(i)} - \overline{X}\right)^2} \qquad (4.34)$$

Keďže sa však testovacia charakteristika W pri platnosti nulovej hypotézy neriadi známym rozdelením, kritické hodnoty boli uvádzané v tabuľkách. Po nasledujúcej transformácii testovacej charakteristiky:

$$Z_W = \frac{\ln(1 - W) - \mu_Z}{\sigma_Z} \qquad (4.35)$$

kde:

$$\mu_Z = 0.0038915(\ln(n))^3 - 0.083751(\ln(n))^2 - 0.31082\ln(n) - 1.5861 \qquad (4.36)$$

$$\ln(\sigma_Z) = 0.0030302(\ln(n))^2 - 0.082676\ln(n) - 0.4803 \qquad (4.37)$$

a Z_W sa riadi (pre $12 \leq n \leq 2000$) normovaným normálnym rozdelením pravdepodobnosti. Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $F(X) = F_n(X)$ H_1: $F(X) \neq F_n(X)$	Hypotézu H_0 zamietame, ak $\lvert Z_W \rvert > \lvert z_{(1-\alpha)} \rvert$

kde $z_{(1-\alpha)}$ je kvantil normovaného normálneho rozdelenia (v programe R ho získame pomocou funkcie `qnorm()`). V programe R slúži na výpočet Shapiro – Wilkovho testu funkcia `shapiro.test()` z programového balíka `nortest`.

Príklad 4.14

Uvažujme o predchádzajúcom príklade váhy novorodencov.

```
> shapiro.test(wt_new)

                 Shapiro-Wilk normality test

data:  wt_new
W = 0.9956, p-value = 0.001195
```

Hypotézu o normalite podobne ako v prípade Anderson – Darlingovho testu na hladine významnosti $\alpha = 0.05$ zamietame. Keďže celý postup výpočtu testovacej charakteristiky Shapiro – Wilkovho testu je pomerne komplikovaný, overíme si ho aj „manuálne“, aby sme sa presvedčili o správnosti postupu. Výsledky testovacej štatistiky vychádzajú rovnaké a rozdiely v *p*-hodnote sú spôsobené tým, že sme použili obojstranný test. V prípade jednostranného testu, tj. $Z_W > z_{(1-\alpha)}$ (pre tento konkrétny príklad) budú výsledky rovnaké.

```
> n <- length(wt_new)
> x <- sort(wt_new, decreasing = F)
> m <- qnorm(((1:n)-3/8)/(n+1/4))
> c <- m / sqrt(sum(m^2))
> u <- 1/sqrt(n)
> aN <- -2.706056*u^5 + 4.434685*u^4 - 2.07119*u^3 -
  0.147981*u^2 + 0.221157*u + c[n]
> aN1 <- -3.582633*u^5 + 5.682633*u^4 - 1.752461*u^3 -
  0.293762*u^2 + 0.042981*u + c[n-1]
> e <- (sum(m^2) - 2*m[n]^2 - 2*m[n-1]^2)/(1 - 2*aN^2 - 2*aN1^2)
> ai <- m[3:(n-2)]/sqrt(e)
> ai <- c(-aN, -aN1, ai, aN1, aN)
> pok <- ai[1:618]*(x[n-(1:618)+1]-x[1:618])
> pokk <- sum(pok)^2
> men <- sum((x-mean(x))^2)
> W <- pokk/men
> W <- (sum(ai*x))^2/sum((x-mean(x))^2); W
[1] 0.9955888
> mu <- 0.0038915*(log(n))^3 - 0.083751*(log(n))^2 -
  0.31082*log(n) - 1.5861
> sig <- exp(0.0030302*(log(n))^2 - 0.082676*log(n) - 0.4803)
> Zw <- (log(1-W) - mu)/sig
> 1-pnorm(Zw)+pnorm(-Zw)
[1] 0.002383422
```

4.4.6 Jarque – Berov test

Zrejme najpoužívanejším testom normality v ekonometrii je Jarque – Berov test. Majme *iid* náhodnú vzorku X_i, $i = 1, 2, ..., n$. Overujeme, či je možné predpokladať, že hodnoty pochádzajú z normálneho rozdelenia pravdepodobnosti. Test je založený na

momentoch tretieho a štvrtého rádu (šikmosť a špicatosť). Definujme si odhad šikmosti *S* a špicatosti *K* nasledovne (ide o odlišný odhad ako výberová šikmosť a špicatosť):

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{3/2}} \qquad (4.38)$$

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^{2}} \qquad (4.39)$$

Testovacia charakteristika má potom tvar:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^3}{4}\right) \qquad (4.40)$$

Za predpokladu platnosti nulovej hypotézy má štatistika *JB* Chí-kvadrát rozdelenie pravdepodobnosti s 2 stupňami voľnosti. Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $F(X) = F_n(X)$ H_1: $F(X) \neq F_n(X)$	Hypotézu H_0 zamietame, ak $JB > \chi^2_{2,\,(1-\alpha)}$

kde $\chi^2_{2,\,(1-\alpha)}$ je kvantil Chí-kvadrát rozdelenia s 2 stupňami voľnosti. Jarque – Berov test je pomerne citlivý na odľahlé hodnoty (čo je spôsobené použitím momentov tretieho a štvrtého rádu, ktoré sú citlivejšie ako napríklad priemer alebo rozptyl). Sila *JB* testu (schopnosť zamietnuť nulovú hypotézu, ak je v skutočnosti alternatívna hypotéza pravdivá) sa ukazuje ako vhodná najmä pre väčšie vzorky, kým v malých vzorkách je sila *JB* testu menšia (pozri Dufour et al., 1998). Z týchto dvoch dôvodov sa občas používa modifikovaný Jarque – Berov test (označovaný ako *JBU*, podľa Urzúa, 1996; prípadne *RBU* skr. z Robust Jaruqe – Bera). Pri výpočte *JBU* budeme najprv potrebovať vypočítať nasledujúce tri vzťahy (Thadewald – Büning, 2007; Urzúa, 1996):

$$vs = \frac{6(n-2)(n+1)}{(n+1)(n+3)} \qquad (4.41)$$

$$ek = \frac{3(n-1)}{(n+1)} \qquad (4.42)$$

$$vk = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \qquad (4.43)$$

Testovacia štatistika má potom tvar:

$$JBU = \left(\frac{S^2}{vs} + \frac{(K-ek)^2}{vk} \right) \qquad (4.44)$$

Rozhodnutie o hypotéze vykonáme rovnako ako v predošlom prípade:

H_0: $F(X) = F_N(X)$	Hypotézu H_0 zamietame, ak $JBU > \chi^2_{2,\,(1-\alpha)}$
H_1: $F(X) \neq F_N(X)$	

V programe R vieme na výpočet *JB* a *JBU* testu použiť funkciu `rjb.test()` z programového balíka `lawstat`. Charakteristiky *JB* a *JBU* je možné modelovať Chí-kvadrát rozdelením len asymptoticky, pričom rýchlosť konvergencie rozdelení (v závislosti od veľkosti vzorky) sa považuje za pomerne nízku. Z tohto dôvodu sa odporúča použiť simuláciu Monte Carlo (Hui et al., 2008).

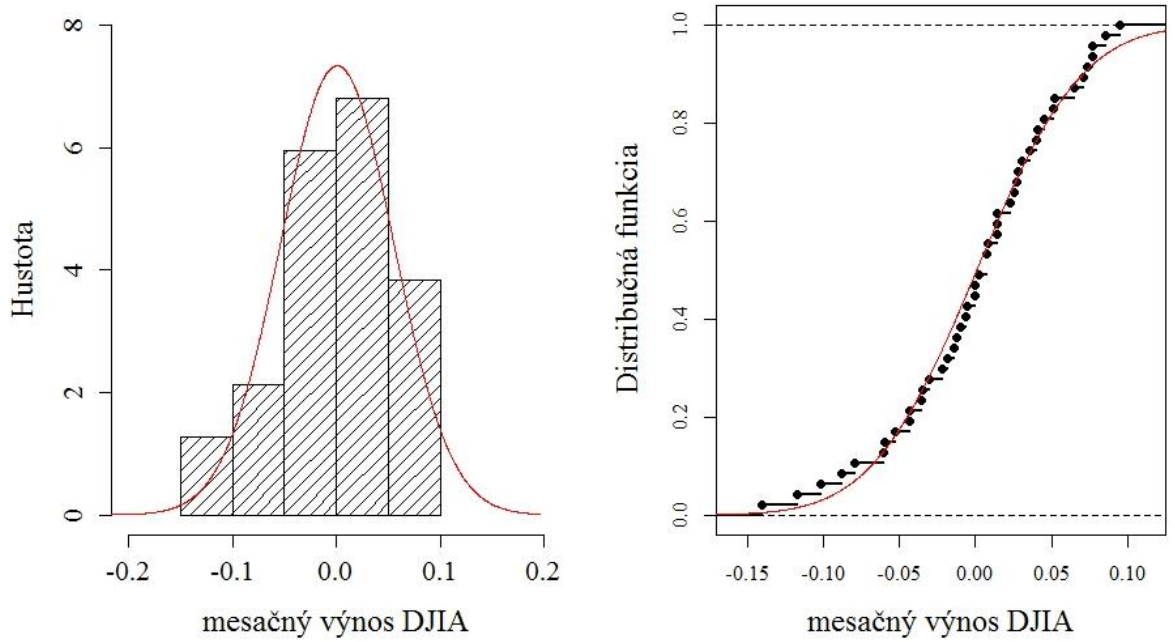
Príklad 4.15

Na ukážku sme vybrali mesačné spojité výnosy[22] z akciového indexu Dow Jones Industrial Average a to za obdobie od 02.01.2008 do 01.12.2011, teda $n = 47$. Ak si označíme cenu akcie ako P_t, potom spojitý výnos vypočítame ako $ln(P_t/P_{t-1})$. Nasledujúce obrázky (pozri Obrázok 4.9) znázorňujú histogram a distribučnú funkciu – empirickú ako aj teoretickú zodpovedajúcu normálnemu rozdeleniu s parametrami odhadnutými priamo z údajov.

```
> djia <- c(-0.0304, -0.0003, 0.0454, -0.0142, -0.1019, 0.0025,
  0.0145, -0.06, -0.1406, -0.0532, -0.006, -0.0884, -0.1172,
  0.0773, 0.0735, 0.0407, -0.0063, 0.0858, 0.0354, 0.0227, 0,
  0.0651, 0.008, -0.0346, 0.0256, 0.0515, 0.014, -0.0792, -
  0.0358, 0.0708, -0.0431, 0.0772, 0.0306, -0.0101, 0.0519,
  0.0272, 0.0281, 0.0076, 0.0398, -0.0188, -0.0124, -0.0218, -
  0.0436, -0.0603, 0.0954, 0.0076, 0.0143)
-------------------------------------------------------------
> par(mfrow = c(1, 2))
> hist(djia, density = 10, col = "black", main = NA, cex.lab =
  1.5, ylim = c(0,8), xlim = c(-0.2, 0.2), cex.axis = 1.3, freq
  = FALSE, ylab = "Hustota", family = "serif", xlab = "ročný
  výnos DJIA")
> x <- seq(min(djia) - 0.2, max(djia) + 0.1, length = 2000)
> xh <- dnorm(x, mean = mean(djia), sd = sd(djia))
> data <- data.frame(x,xh)
```

[22] V skutočnosti pri tomto príklade predpokladáme, že tento časový rad je *iid*, čo je v praxi pri časových radoch veľmi zriedkavé.

```
> lines(data, type = "l", col = "red")
> plot(ecdf(djia), col.01line = "black", lwd = 2.5, pch = 19,
  xlab = "ročný výnos DJIA", family = "serif", cex.lab = 1.5,
  main ="", ylab = "Distribučná funkcia")
> xh <- pnorm(x, mean = mean(djia), sd = sd(djia))
> data <- data.frame(x,xh)
> lines(data, col = "red", lwd = 0.5)
```



Obrázok 4.9: Výnosy z DJIA – histogram a EDF

Zdroj: vlastné spracovanie, výstup zo softvéru R

Počet pozorovaní nepovažujeme za veľký, preto za najvhodnejší spôsob výpočtu budeme považovať ten, v ktorom sa kritické hodnoty budú odhadovať pomocou simulácií. Vo funkcii `rjb.test()` tomu zodpovedá možnosť `crit.values = c("empirical")` a počet simulovaných vzoriek sa zvolí v možnosti `"N = "` (napr. `N = 1000`). V oboch prípadoch sme hypotézu H_0 nevedeli zamietnuť.

```
> library(lawstat)
> rjb.test(djia, option = c("JB"), crit.values = c("empirical"),
  N = 1000)

                    Jarque Bera Test

data:  djia
X-squared = 2.0335, df = 2, p-value = 0.1997
------------------------------------------------------------------
> rjb.test(djia, option = c("RJB"), crit.values =
  c("empirical"), N = 1000)
```

```
                    Robust Jarque Bera Test

data:  djia
X-squared = 2.2937, df = 2, p-value = 0.1801
```

4.5 Testy extrémnych hodnôt

Intuitívne môžeme vnímať extrémne hodnoty v určitom štatistickom súbore ako tie, ktoré sa svojou hodnotou javia byť výrazne nižšie, resp. výrazne vyššie ako ostatné hodnoty v štatistickom súbore. Dôvod výskytu extrémnych hodnôt môže byť pomerne pestrý, čo do značnej miery komplikuje rozhodnutie, čo urobiť v prípade, ak sú v štatistickom súbore prítomné. Účelom tejto kapitoly je ponúknuť niektoré základné štatistické testy na odhalenie extrémnych hodnôt. Problematika je však podstatne širšia. Extrémnou hodnotou môže byť také meranie, ktoré nebolo technicky ani možné. Napríklad pri použití dotazníkového šetrenia mali respondenti označiť v odpovedi celé číslo od 1 po 7. Ak sa v súbore vyskytuje číslo mimo tohto rozsahu, testy nie sú nutné. Ďalším príkladom extrémnej hodnoty je situácia, kde hodnota síce teoreticky mohla nastať, avšak je pre daný súbor veľmi málo pravdepodobná. Ak je predmetom nášho záujmu výška žien a v náhodnej vzorke s rozsahom $n = 30$ pozorovaní máme 2 alebo 3 hodnoty väčšie ako 195 *cm*, je na mieste zisťovať, či sa do vzorky nedostala výška mužov. Môže ísť teda o hodnoty z inej populácie. Zdanlivo extrémna hodnota nemusí vôbec byť extrémnou . Ak merania pochádzajú zo silne zošikmeného rozdelenia pravdepodobnosti, veľké extrémy sa s určitou pravidelnosťou vyskytujú úplne prirodzene. Výskyt viacerých extrémnych hodnôt tak môže byť signálom, že pravdepodobnostný model, z ktorého hodnoty pochádzajú, má túto vlastnosť. Neraz je neželaným dôsledkom extrémnych hodnôt ich výrazný vplyv na výsledky štatistickej analýzy. Pri aritmetickom priemere je pomerne jednoduché ukázať, aký vplyv na neho vie mať extrémna hodnota. V akademických publikáciách sa stretávame s takým prístupom riešenia extrémnych hodnôt, v ktorom sa výsledky prepočítavajú s celým súborom a so súborom bez extrémnych hodnôt. V prípade kvalitatívne zhodných výsledkov je celkový záver z analýzy dôveryhodnejším.

V tejto časti publikácie sa budeme venovať iba niektorým základným testom na zisťovanie výskytu odľahlých hodnôt. Predtým však pripomenieme, že azda najjednoduchším nástrojom identifikácie extrémnych hodnôt je vhodná vizualizácia dát. Častou pomôckou je box – plot. Pri box – plote sa hodnoty X_i (pre ktoré platí, že sú $X_i < X_{(\lceil n0.25 \rceil)} - 1.5R_Q$ alebo $X_i > X_{(\lceil n0.75 \rceil)} + 1.5R_Q$) osobitne označujú a tieto hodnoty je možné považovať za extrémy. Pri konštrukcii box – plotov sú tieto hodnoty zobrazované ako body mimo úsečiek.

4.5.1 Grubbsov test

Majme náhodný *iid* súbor hodnôt X_i, i = 1, 2, ..., n. Predpokladáme, že hodnoty X_i sú realizáciami z normálneho rozdelenia pravdepodobnosti. Grubbsovým testom overujeme nulovú hypotézu, že v súbore nie sú extrémne hodnoty voči alternatíve, že v súbore je aspoň jedna extrémna hodnota. Testovacia charakteristika Grubbsovho testu má tvar (spracované podľa Blischke et al., 2011; Yu et al., 2009; Thompson – Lowthian, 2011):

$$G = \frac{\max_{i=1,2,\ldots,n} \left| X_i - \overline{X} \right|}{s} \tag{4.45}$$

Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: X^* nie je extrém H_1: X^* je extrém	Hypotézu H_0 zamietame, ak $G > g_{n,\alpha}$

kde kritickú hodnotu vypočítame zo Studentovho *t* rozdelenia nasledovne:

$$g_{n,\alpha} = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}} \tag{4.46}$$

pričom $t_{(\alpha/(2n),n-2)}$ je kvantil Studentovho *t* rozdelenia. Tento test sa realizuje naraz iba pre práve jednu hodnotou. Nevýhody Grubbsovho testu by sme zhrnuli do dvoch bodov. Po prvé, predpoklad normality je pomerne reštriktívny. Pred samotnou realizáciou Grubbsovho testu je vhodné vykonať test na normalitu údajov, čo môže zvýšiť celkovú chybu rozhodovacieho procesu. Zároveň je možné, že ak je v súbore prítomná extrémna hodnota, tak práve v dôsledku nej test normalitu zamietne. Po druhé, ak je v štatistickom súbore viac ako jedna extrémna hodnota, Grubbsov test môže mať tendenciu tieto hodnoty nenájsť. Dôvod je ten, že sa overuje vždy iba jedna hodnota. To znamená, že ostatné hodnoty (prípadné ďalšie extrémne hodnoty) ovplyvňujú výpočet priemeru aj výberovej smerodajnej odchýlky. To spôsobí, že sa maximálny absolútny rozdiel (čitateľ) nebude zdať byť príliš veľký vzhľadom na variabilitu údajov vo vzorke (menovateľ), ktorú ale spoluvytvárajú aj extrémne hodnoty.

V programe R môžeme Grubbsov test realizovať pomocou funkcie `grubbs.test()` z knižnice `outliers`.

Príklad 4.16

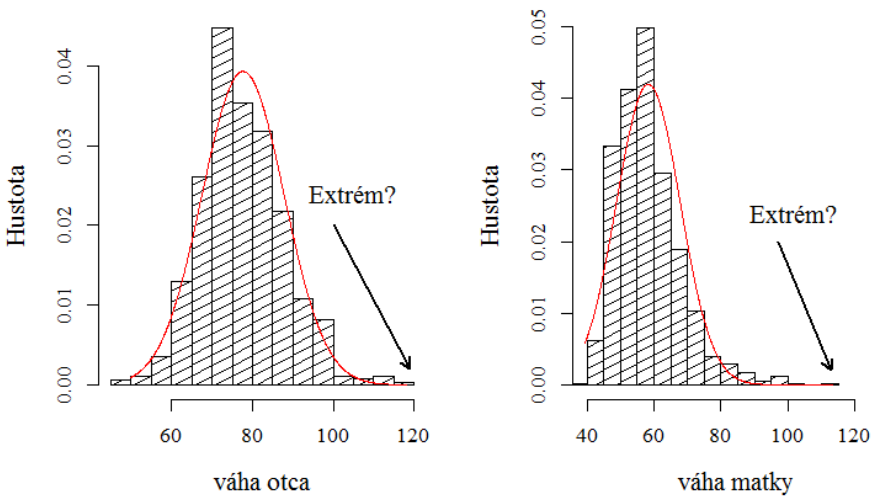
Použijeme databázu `babies` z programového balíka `UsingR`. Budeme predpokladať, že váha otcov aj matiek pochádza z normálneho rozdelenia pravdepodobnosti. Bude nás zaujímať hypotéza, že v súbore váha otcov (matiek) nie je extrémna hodnota. Najprv

z údajov odstránime chýbajúce záznamy (označené číslom 999) a potom si váhu transformujeme na *kg*.

```
> library(UsingR)
> attach(babies)
> dwt <- sort(subset(babies$dwt, subset = dwt != 999),
  decreasing = T)
> dwt <- 0.45359237*dwt
> summary(dwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  49.90   70.31   77.11   77.66   83.91  117.90
> wt1 <- sort(subset(babies$wt1, subset = wt1 != 999),
  decreasing = T)
> wt1 <- 0.45359237*wt1
> summary(wt1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.46   52.05   56.70   58.34   63.05  113.40
```

Všimnime si, že rozdiel medzi minimálnou váhou mužov (`dwt`) a prvým kvartilom je 20 *kg*, kým medzi maximálnou váhou a horným kvartilom je približne 34 *kg*. Maximálna váha sa javí ako kandidát na extrémnu hodnotu. V prípade žien je tento rozdiel ešte väčší. Rozdiel medzi minimálnou váhou a prvým kvartilom je niečo vyše 12 *kg*, kým medzi maximálnou váhou a horným kvartilom skoro 50 *kg*. Ide zrejme o pravostranné zošikmené súbory, čo je možné vidieť aj z ďalšieho obrázku (pozri Obrázok 4.10).

```
> par(mfrow = c(1, 2))
> hist(dwt, density = 10, col = "black", main = NA, cex.lab =
  1.5, cex.axis = 1.3, freq = FALSE, ylab = "Hustota", family =
  "serif", xlab = "váha otca")
> x <- seq(min(dwt), max(dwt), length = 1000)
> xh <- dnorm(x, mean = mean(dwt), sd = sd(dwt))
> data <- data.frame(x,xh)
> lines(data, type = "l", col = "red")
> arrows(100, 0.02,119, 0.002, length = 0.10, lwd = 2)
> text(105, 0.024, "Extrém?", family = "serif", cex = 1.3)
> hist(wt1, density = 10, col = "black", main = NA, xlim =
  c(min(wt1), 130), cex.lab = 1.3, cex.axis = 1.1, freq = FALSE,
  ylab = "Hustota", family = "serif", xlab = "váha matky")
> x <- seq(min(wt1), max(wt1), length = 1000)
> xh <- dnorm(x, mean = mean(wt1), sd = sd(wt1))
> data <- data.frame(x,xh)
> lines(data, type = "l", col = "red")
> arrows(97, 0.02,113, 0.002, length = 0.10, lwd = 2)
> text(102, 0.024, "Extrém?", family = "serif", cex = 1.3)
```



Obrázok 4.10: Histogram váhy otcov a matiek s možnými extrémnymi hodnotami

Zdroj: vlastné spracovanie, výstup zo softvéru R

Následne sme vykonali Grubbsov test, ktorý v oboch prípadoch odhalil prítomnosť extrémnej hodnoty. Týmto sa však úloha nekončí. Alternatívna hypotéza v Grubbsovom teste hovorí o prítomnosti jedného až viacerých extrémnych hodnôt. Preto je potrebné súbor preveriť na prítomnosť ďalších extrémov.

```
> library(outliers)
> grubbs.test(dwt)

        Grubbs test for one outlier

data:  dwt
G = 3.9666, U = 0.9786, p-value = 0.02468
alternative hypothesis: highest value 117.9340162 is an outlier
------------------------------------------------------------
> grubbs.test(wt1)

        Grubbs test for one outlier

data:  wt1
G = 5.7875, U = 0.9720, p-value = 3.379e-06
alternative hypothesis: highest value 113.3980925 is an outlier
```

V prípade váhy mužov sme viac extrémnych hodnôt neidentifikovali. V prípade váhy žien sme museli odstrániť až 8 pozorovaní, kým nám Grubbsov test nevedel zamietnuť nulovú hypotézu na $\alpha = 0.05$. Pre porovnanie by mohol čitateľ zostrojiť histogram bez „extrémov“ a rozdelenie navzájom porovnať.

4.5.2 Dixonov test

Vychádzajme z náhodnej *iid* vzorky hodnôt X_i, i = 1, 2, ..., n, ktoré sú realizáciami z normálneho rozdelenia pravdepodobnosti. Usporiadame ich do variačného radu $X_{(i)}$ vzostupne, to znamená $X_{(1)} = \min_i \{ X_i \}$ a $X_{(n)} = \max_i \{ X_i \}$. Overujeme hypotézu, že hodnota $X_{(n)}$ je extrémnou hodnotou. Testovacia štatistika má nasledujúci tvar:

$$Dx = \begin{cases} \dfrac{X_n - X_{n-1}}{X_n - X_1}, & 3 \le n \le 7 \\ \dfrac{X_n - X_{n-1}}{X_n - X_2}, & 8 \le n \le 10 \\ \dfrac{X_n - X_{n-2}}{X_n - X_2}, & 11 \le n \le 13 \\ \dfrac{X_n - X_{n-2}}{X_n - X_3}, & 14 \le n \le 30 \end{cases} \qquad (4.47)$$

Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $X_{(n)}$ nie je extrém	Hypotézu H_0 zamietame, ak $Dx > Dx_{n,\alpha}$
H_1: $X_{(n)}$ je extrém	

kde $Dx_{n,\alpha}$ je kritická hodnota, ktorá sa uvádza v štatistických tabuľkách. V programe R na výpočet Dixonovho testu slúži funkcia `dixon.test()` z programového balíka `outliers`. Všimnime si, že testovacia charakteristika je definovaná iba pre $n \le 30$. Pre $n > 30$ je možné použiť postup (vrátane kritických hodnôt) od Verma – Quiroz-Ruiz (2006, 2008) a Verma et al. (2008). V prípade, ak by sme overovali hypotézu o tom, či je $X_{(1)}$ extrémna hodnota, stačilo by vytvoriť variačný rad X_i tak, aby platilo $X_{(1)} = \max_i \{ X_i \}$ a $X_{(n)} = \min_i \{ X_i \}$ a postupovať obdobne ako vyššie. Pri funkcii `dixon.test()` stačí použiť možnosť `opposite = T`. Podobne ako pri Grubbsovom teste, aj tu je hlavnou nevýhodou predpoklad normality a možná prítomnosť viac ako jednej extrémnej hodnoty v štatistickom súbore.

Príklad 4.17

V nasledujúcom príklade použijeme databázu `Cars93` z programového balíka `MASS`. Z databázy 93 áut vyberieme také, ktoré nemajú americký pôvod, zároveň majú viac ako 3 valce (+ Wankelov rotačný motor) a majú výkon motora viac ako 80, ale menej ako 150 koní. Zaujíma nás, či existuje extrémna hodnota v spotrebe v meste (premenná `MPG.city`) a váhe (`Weight`). Použitím oboch testov sme identifikovali aspoň jednu

extrémnu hodnotu. Necháme na čitateľovi aby rozhodol, či je v súbore prítomných aj viac extrémnych hodnôt.

```
> library(MASS)
> attach(Cars93)
> b <- subset(Cars93, subset = Origin != "USA" & Cylinders != 3
  & Horsepower > 80 & Horsepower < 150)
> MPG.city <- sort(b$MPG.city, decreasing = F)
> dixon.test(MPG.city)

                Dixon test for outliers

data:  MPG.city
Q = 0.5455, p-value = 0.002054
alternative hypothesis: highest value 42 is an outlier
-----------------------------------------------------------------
> Weight <- sort(b$Weight, decreasing = F)
> dixon.test(Weight)

                Dixon test for outliers

data:  Weight
Q = 0.5224, p-value = 0.01047
alternative hypothesis: highest value 3960 is an outlier
```

4.5.3 Hampelov test

Hampelov test patrí do kategórie neparametrických testov. Pri jeho realizácii nie je potrebný predpoklad o normalite a taktiež neprekáža prítomnosť viacerých extrémnych hodnôt. Už sme spomenuli, že prítomnosť viacerých extrémnych hodnôt môže spôsobiť problémy pri ich identifikácii v prípade Grubbsovho aj Dixonovho testu. Dôvodom je, že aritmetický priemer aj rozptyl (a teda aj smerodajná odchýlka) môžu byť do značnej miery ovplyvnené hoci aj jednou odľahlou hodnotou (prípadne viacerými). Napríklad v prípade týchto dvoch štatistík stačí jedna odľahlá hodnota k tomu, aby sme vedeli posunúť priemer a rozptyl nad (pod) ľubovoľnú hodnotu. V prípade viacerých odľahlých hodnôt tak nemusíme nájsť ani jednu. Tomuto efektu sa hovorí „*masking*“ (voľne preložené ako prekrývanie). Všimnime si, že v prípade mediánu to neplatí. Hampelov test využíva tento princíp. Nejde o štatistickú hypotézu v klasickom slova zmysle. Vypočíta sa charakteristika pre každú jednu hodnotu v štatistickom súbore a ak táto prekročí určitú hodnotu, danú(é) hodnotu(y) označíme za extrémnu(e) (bližšie pozri Reichenbächer – Einax, 2011). Majme výberový súbor X_i, i = 1, 2, ..., n, ktorý v ideálnom prípade je realizáciou zo spojitého rozdelenia pravdepodobnosti. Táto podmienka nie je nutná. Hodnoty môžu byť aj z diskrétneho rozdelenia, ale je vhodné,

aby bola čo najmenšia početnosť rovnakých hodnôt v súbore. Vytvoríme si najprv premennú r_i (postup podľa práce Wilcox, 2012):

$$r_i = \left| X_i - \tilde{X} \right| \tag{4.48}$$

Následne si vypočítame charakteristiku *MAD*:

$$MAD = \tilde{r} \tag{4.49}$$

Potom MADN má tvar:

$$MADN = MAD / 0.6745 \tag{4.50}$$

Testovacia charakteristika je:

$$H_i = \frac{\left| X_i - \tilde{X} \right|}{MADN} \tag{4.51}$$

Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $X_{(i)}$ nie je extrém H_1: $X_{(i)}$ je extrém	Hypotézu H_0 zamietame, ak $H_i > \sqrt{\chi^2_{0.975,1}}$

Z predchádzajúceho príkladu (Príklad 4.17) použijeme dátové vektory a uskutočníme Hampelov test (nazývaný tiež angl. *Hampel identifier*). Procedúru nadefinujeme ako funkciu.

```
> hampel_identifier <- function(data) {
+ ri <- abs(data - median(data))
+ mad <- median(ri)
+ madn <- mad/0.6745
+ hi <- ri/madn
+ critical <- sqrt(qchisq(0.975,1))
+ data[hi>critical]
+ }
```

Výpočtom dôjdeme k obdobným záverom ako pri použití Dixonovho testu. Funkcia nám vráti extrémne hodnoty z daného dátového vektora.

```
> hampel_identifier(Weight)
[1] 3785 3960
> hampel_identifier(MPG.city)
[1] 42
```

4.6 Vybrané neparametrické a parametrické testy

4.6.1 Test náhodnosti

Definujme si najprv pojem blok. Pod blokom budeme rozumieť postupnosť rovnakých prvkov, ktoré nasledujú alebo predchádzajú iným prvkom alebo žiadnym prvkom. Na ukážku

si zoberme nasledujúcu postupnosť písmen, ktoré oddeľujeme do blokov pomocou symbolu „|“.

aaa|bb|aaaaaaa|bbbbbbb|aaaaaa|bbbbbbb|aaaaa *(4.52)*

Takúto postupnosť vieme vyskladať aj z konkrétnych meraní. Ak si stanovíme medián ako strednú hodnotu a označíme všetky hodnoty nad mediánom ako „*a*“ a hodnoty menšie (prípadne rovné) ako medián označíme ako „*b*“ a zaujíma nás postupnosť, v akej sa hodnoty získavali, tak môžeme zostrojiť postupnosť podobnú ako je vo výraze *(4.52)*. Vo všeobecnosti je však nasledujúci postup vhodný v situáciách, kde pozorovaná premenná má dichotomický charakter (jav nastal / jav nenastal).

V mnohých štatistických testoch predpokladáme, že namerané hodnoty sú od seba nezávislé. Môžeme to vnímať aj tak, že to, akú hodnotu nameriame v jednom pozorovaní neovplyvní to, akú hodnotu nameriame neskôr. Preto nás neraz zaujíma, či je možné postupnosť nami nameraných hodnôt vnímať ako náhodnú alebo nie.

Ak máme dva symboly, tak je v postupnosti *(4.52)* celkový počet blokov maximálne rovný *n*. Takúto situáciu dosiahneme vtedy, ak jedno „*a*“ strieda ďalšie „*b*“ a potom nasleduje znovu „*a*“, atď. Považovali by sme takúto postupnosť za náhodnú? Takáto postupnosť by bola možná, ale aby vznikla úplne náhodou, to budeme zrejme považovať za málo pravdepodobné. Prvky „*a*“ a „*b*“ sa opakujú s veľkou pravidelnosťou. Príliš veľký počet blokov naznačuje nenáhodnosť. Neraz je to signálom, že je prítomný cyklus. V postupnosti *(4.52)* je 7 blokov. Je zrejmé, že symboly „*a*“ a „*b*“ sa vzájomne zhlukujú. Zrejme aj príliš malý počet blokov naznačuje nenáhodnosť, ktorá sa často interpretuje ako prítomnosť trendu. Ak sa každý symbol vyskytne v postupnosti aspoň raz, potom najmenší počet blokov v postupnosti je dva. Ak táto podmienka neplatí, tak najmenší počet blokov je jeden. Tieto princípy sa využívajú aj pri testovaní náhodnosti.

Označme si n_a ako počet symbolov *a* a n_b ako počet symbolov *b*, takže $n_a + n_b = n$. Označme ďalej *R* ako náhodnú premennú, ktorá označuje počet blokov v postupnosti. Ďalej uvažujme, že buď $n_a \geq 15$ alebo $n_b \geq 15$. Potom testovacia charakteristika má tvar (Panik, 2005):

$$Z_R = \frac{R - E(R)}{\sqrt{V(R)}} \qquad (4.53)$$

kde $E(R)$ je stredná hodnota a $V(R)$ je rozptyl, pre ktoré platí:

$$E(R) = \frac{2n_a n_b}{n} + 1 \qquad (4.54)$$

$$V(R) = \frac{2n_a n_b (2n_a n_b - n)}{n^2 (n-1)} \qquad (4.55)$$

Testovacia charakteristika sa riadi normovaným normálnym rozdelením pravdepodobnosti $Z_R \sim N(0, 1)$ (za predpokladu platnosti nulovej hypotézy o náhodnosti postupnosti). Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: Postupnosť je náhodná H_1: Postupnosť nie je náhodná	Hypotézu H_0 zamietame, ak $\|Z_R\| > \|z_{(\alpha/2)}\|$
H_0: Postupnosť je náhodná H_1: V postupnosti je príliš veľa blokov	Hypotézu H_0 zamietame, ak $Z_R > z_{(1-\alpha)}$
H_0: Postupnosť je náhodná H_1: V postupnosti je príliš málo blokov	Hypotézu H_0 zamietame, ak $Z_R < z_{(\alpha)}$

pričom $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú kvantily normovaného normálneho rozdelenia pravdepodobnosti.

Príklad 4.18

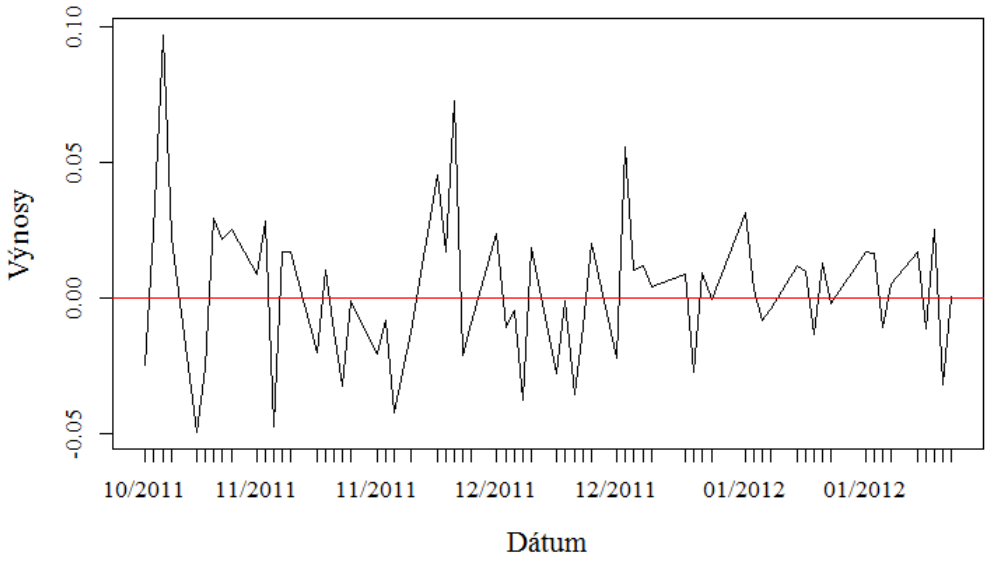
Uvažujme o denných kapitálových výnosoch akcií spoločnosti Occidental Petroleum Corporation (tiker: OXY) od 24.09.2011 do 27.01.2012. Jednou z tradičných ekonomických teórií je teória o efektívnosti finančných trhov. Zjednodušene, ak táto teória platí (v slabej forme), výnosy z akcií by nemali byť predpovedateľné, mali by byť náhodné. V prípade denných údajov to znamená, že postupnosť „kladných" a „záporných" výnosov by mala byť čisto náhodná. Údaje si najprv usporiadame tak, aby sme vytvorili tzv. vektor faktorov. Použijeme na to funkciu `sign()` a `factor()`. Prvá funkcia `sign()` zoberie vektor výnosov a vytvorí nový vektor, ktorý bude mať rovnaký rozsah s tým, že ak dôjde ku kladnému kapitálovému výnosu priradí hodnotu 1 a v prípade záporného kapitálového výnosu priradí hodnotu 0. Funkcia `factor()` potom definuje tento vektor ako vektor, ktorého prvky sú rôzne úrovne jedného faktora. Následne môžeme použiť funkciu `runs.test()` z programového balíka `tseries`.

```
> oxy <- c(-0.02486, 0.02419, 0.09702, 0.02206, -0.04943, -
  0.02573, 0.02918, 0.02167, 0.02554, 0.00895, 0.02815, -
  0.04722, 0.01697, 0.01699, -0.02003, 0.01017, -0.0122, -
  0.03253, -0.00117, -0.02035, -0.00826, -0.04221, -0.01225,
  0.04543, 0.01718, 0.07281, -0.02093, -0.00934, 0.02388, -
  0.01074, -0.00445, -0.03739, 0.01867, -0.02775, -0.00109, -
  0.03556, -0.01018, 0.02, -0.02184, 0.05554, 0.0102, 0.01181,
  0.00435, 0.00877, -0.02724, 0.00937, -0.00021, 0.03127, 0.003,
  -0.00794, -0.00395, 0.01211, 0.0098, -0.01328, 0.01284, -
  0.00204, 0.01721, 0.01641, -0.0106, 0.00511, 0.01684, -
  0.01117, 0.02517, -0.03161, 0.0006)
```

```
> oxy_f <- factor(sign(oxy))
> oxy_f
 [1] -1 1  1  1  -1 -1 1  1  1  1  1  -1 1  1  -1 1  -1 -1 -1 -1
  -1 -1 -1 1  1
[26] 1  -1 -1 1  -1 -1 -1 1  -1 -1 -1 -1 1  -1 1  1  1  1  1  -1
  1  -1 1  1  -1
[51] -1 1  1  -1 1  -1 1  1  -1 1  1  -1 1  -1 1
Levels: -1 1
-----------------------------------------------------------------
> library(tseries)
> runs.test(oxy_f, alternative = c("two.sided"))

                        Runs Test

data:  oxy_f
Standard Normal = 0.1426, p-value = 0.8866
alternative hypothesis: two.sided
```

Hypotézu H_0 nevieme zamietnuť, takže postupnosť výnosov sa nám javí ako náhodná. Uvedenú postupnosť si môžeme zobraziť na nasledujúcom obrázku (pozri Obrázok 4.11).

```
> date <- c("10.25.2011", "10.26.2011", "10.27.2011",
  "10.28.2011", "10.31.2011", "11.1.2011", "11.2.2011",
  "11.3.2011", "11.4.2011", "11.7.2011", "11.8.2011",
  "11.9.2011", "11.10.2011", "11.11.2011", "11.14.2011",
  "11.15.2011", "11.16.2011", "11.17.2011", "11.18.2011",
  "11.21.2011", "11.22.2011", "11.23.2011", "11.25.2011",
  "11.28.2011", "11.29.2011", "11.30.2011", "12.1.2011",
  "12.2.2011", "12.5.2011", "12.6.2011", "12.7.2011",
  "12.8.2011", "12.9.2011", "12.12.2011", "12.13.2011",
  "12.14.2011", "12.15.2011", "12.16.2011", "12.19.2011",
  "12.20.2011", "12.21.2011", "12.22.2011", "12.23.2011",
  "12.27.2011", "12.28.2011", "12.29.2011", "12.30.2011",
  "1.3.2012", "1.4.2012", "1.5.2012", "1.6.2012", "1.9.2012",
  "1.10.2012", "1.11.2012", "1.12.2012", "01.13.2012",
  "01.17.2012", "01.18.2012", "01.19.2012", "01.20.2012",
  "01.23.2012", "01.24.2012", "01.25.2012", "01.26.2012",
  "01.27.2012")
> x.Date <- as.Date(date,format="%m.%d.%Y")
> plot(oxy~x.Date, type="l", lwd=1.5, xaxt = "n", lty = 1,
  family = "serif", xlab = "Dátum", ylab = "Výnosy", cex.axis =
  1.1, cex = 1.3, cex.lab = 1.3)
> abline(h = 0, col = "red", lwd = 1.5)
> axis(1, x.Date, format(x.Date, "%m/%Y"), cex.axis = 1.1,
  family = "serif", tick = T)
```



Obrázok 4.11: Denné výnosy OXY od 24.09.2011 do 27.01.2012

Zdroj: vlastné spracovanie, výstup zo softvéru R

4.6.2 Bartelsov test nezávislosti

Podobne ako v predchádzajúcom prípade nás zaujíma, či postupnosť môžeme považovať za nenáhodnú alebo nevieme vyvrátiť, že je náhodná. Presnejšie, či sú za sebou nasledujúce hodnoty náhodné (uvedené vyplýva z konštrukcie Bartelovho testu). Majme náhodnú vzorku X_i, $i = 1, 2, ..., n$, takú, kde jednotlivé hodnoty X_i sú získavané s určitým časovým odstupom – ide o diskrétny časový rad. Ďalej predpokladajme, že $n > 10$ (pre $4 \leq n \leq 10$, pozri Bartels, 1982). Princíp testu spočíva v overovaní, či za sebou nasledujúce hodnoty sú závislé. Ak by existovala závislosť, hovorili by sme o autokorelácii prvého rádu, a teda by nešlo o náhodnú postupnosť. Závislosť od predchádzajúceho pozorovania môže byť pozitívna. V takom prípade hovoríme o pozitívnej autokorelácii. Zjednodušene si takýto stav môžeme predstaviť tak, že ak v predchádzajúcom pozorovaní nastala vysoká hodnota X_i, potom aj v ďalšom pozorovaní je väčšia pravdepodobnosť, že nastane vysoká hodnota X_{i+1} a vice versa. Pri negatívnej autokorelácii hodnoty X_i majú tendenciu alternovať, t. j. veľká hodnota X_i je nasledovaná malou hodnotou X_{i+1}.

Pri testovaní najprv vytvoríme z výberového súboru X_i súbor poradí hodnôt Xp_i. Napríklad, ak sme v čase namerali nasledujúce hodnoty X_i: 4, 6, 7, 2, 3, 5, potom Xp_i je: 3, 5, 6, 1, 2, 4. Testovacia charakteristika má potom tvar (postup podľa Gibbons – Chakraborti, 2003, s. 97):

$$RBT = \frac{\sum_{i=1}^{n-1}\left(Xp_i - Xp_{i+1}\right)^2}{\sum_{i=1}^{n}\left(Xp_i - \frac{(n+1)}{2}\right)^2} \qquad (4.56)$$

kde testovacia štatistika *RBT* sa v prípade platnosti nulovej hypotézy asymptoticky riadi normálnym rozdelením pravdepodobnosti so strednou hodnotou:

$$E(RBT) = 2 \qquad (4.57)$$

a rozptylom:

$$D(RBT) = \frac{4(n-2)(5n^2 - 2n - 9)}{5n(n+1)(n-1)^2} \approx \frac{20}{(5n+7)} \qquad (4.58)$$

Testovaciu štatistiku si ďalej môžeme štandardizovať:

$$RBTs = \frac{RBT - E(RBT)}{\sqrt{D(RBT)}} \qquad (4.59)$$

Nízke hodnoty *RBTs* znamenajú prítomnosť trendu, kým veľké hodnoty prítomnosť alternujúcich hodnôt. Rozhodnutie o hypotéze potom vieme vykonať nasledovne:

H_0: Autokorelácia nie je prítomná H_1: Autokorelácia je prítomná	Hypotézu H_0 zamietame, ak $\|RBTs\| > \|z_{(\alpha/2)}\|$
H_0: Autokorelácia nie je prítomná H_1: Negatívna autokorelácia je prítomná	Hypotézu H_0 zamietame, ak $RBTs > z_{(1-\alpha)}$
H_0: Autokorelácia nie je prítomná H_1: Kladná autokorelácia je prítomná	Hypotézu H_0 zamietame, ak $RBTs < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ predstavujú kvantily normovaného normálneho rozdelenia pravdepodobnosti. Na realizáciu tohto testu existuje v programe R funkcia `bartels.test()` z programového balíka `lawstat`. Nakoľko je tento test presný si vieme odskúšať použitím simulácií. Zároveň si tak ilustrujeme použitie tohto testu. Majme rovnaké výnosy z akcií ako v predchádzajúcom príklade. Vykonaním testu dospejeme k obdobnému výsledku ako predtým, t. j. údaje sa javia byť ako náhodné (samozrejme nulovú hypotézu neprijímame, len ju nevieme zamietnuť). Všimnime si, že štatistická (ne)závislosť pri tomto teste sa týka iba nezávislosti za sebou nasledujúcich hodnôt. Môže sa stať, že ak je v časovom rade prítomný cyklus, tak existuje štatistická závislosť nie za sebou idúcich hodnôt, ale štatistická závislosť medzi hodnotami v čase t a $t + k$, kde[23] $k > 2$.

[23] Uvedené tvrdenie by bolo vhodné vysvetliť presnejšie. Ak by platilo, že existuje štatistická závislosť medzi pozorovaniami v čase t a $t + 2$, potom sa táto závislosť prejaví aj medzi hodnotami t a $t + 1$, ale ako výrazne už je otázka veľkosti tejto závislosti.

```
> library(lawstat)
> bartels.test(oxy, alternative = c("negative.correlated"))

            Bartels Test - Negative Correlated

data: oxy
Standardized Bartels Statistic = 0.5065, RVN Ratio = 2.126,
p-value = 0.3062
```

Veľkosť vzorky je $n = 65$. Budeme teda simulovať také časové rady, ktoré budú mať dĺžku $n = 65$ a pritom budeme o nich vedieť, že v nich nie je autokorelácia (prvého rádu). Použijeme na to funkciu `rnorm()`, ktorá generuje náhodné čísla z normálneho rozdelenia pravdepodobnosti.

```
> p.val <- c()
> for (i in 1:10000) {
+ p.val[i] <- bartels.test(rnorm(65), alternative =
  c("negative.correlated"))$p.value
+ }
> sum(p.val<0.05)/10000
[1] 0.0491
```

Vidíme, že podiel chybných zamietnutí je pri 10000 pokusoch pomerne blízky k nominálnym 5 %. Túto analýzu (riešenie problematiky chyby I. druhu) posunieme o malý krok dopredu. Položíme si nasledujúcu otázku. Ako sa správa test v prípade, ak v časovom rade je prítomná autokorelácia? Test by ju mal byť schopný nájsť a to tak, že hypotézu o štatistickej nezávislosti zamietne. Ak to nespraví, dochádza k chybe druhého druhu. Ak nulovú hypotézu o štatistickej nezávislosti správne zamietne, hovoríme o sile testu. K tejto simulácii si budeme potrebovať vygenerovať také časové rady, o ktorých hodnotách vieme, že nie sú nezávislé. Použijeme nasledujúci vzťah:

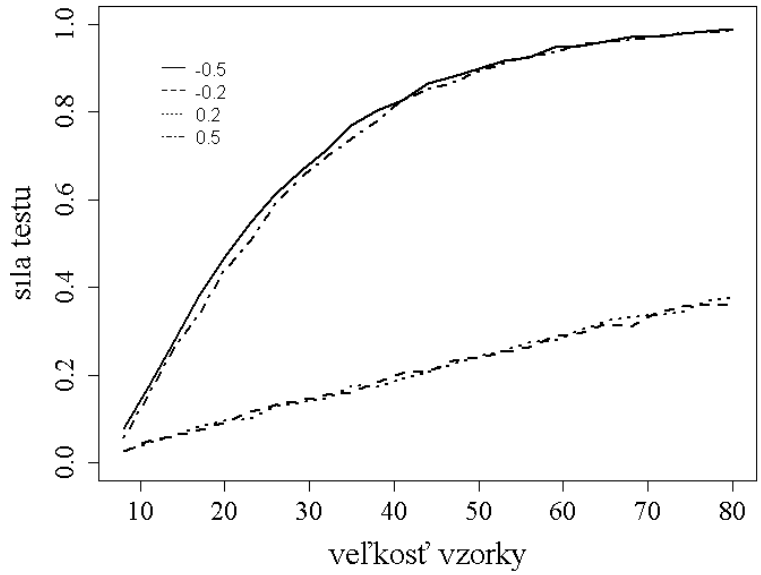
$$X_i = \rho X_{i-1} + e_i \qquad (4.60)$$

kde e_i sú náhodne generované čísla z normálneho rozdelenia pravdepodobnosti. Všimnime si, že hodnoty X_i závisia od predchádzajúcej hodnoty X_{i-1}, a to v závislosti od hodnoty koeficientu ρ. Predpokladajme (bližšie k tejto problematike sa môžeme dočítať v publikáciách venujúcich sa ekonometrii a analýze časových radov), že $|\rho| < 1$. Potom, ak je napr. $\rho = 0.5$, tak predošlá hodnota sa prenásobí koeficientom 0.5 a pripočíta sa náhodné číslo. Zjavne tak táto nová hodnota do určitej miery závisí od predchádzajúcej hodnoty a nie je nezávislá od predchádzajúcich hodnôt. Na druhej strane, ak $\rho = 0$, tak dostávame iba postupnosť náhodných čísel. V našom malom experimente budeme predpokladať hodnoty parametra $\rho = -0.5, -0.2, 0.2, 0.5$. Pre každú z týchto hodnôt parametra ρ budeme simulovať

proces o dĺžke n pozorovaní, pričom $n = 30, 31, ..., 100$. Pre každú kombináciu parametra ρ a n vytvoríme spolu 10000 časových radov pomocou vzťahu *(4.60)* a generátora náhodných čísel, kde náhodné zložky e_i sa generujú z normovaného normálneho rozdelenia. Pre každý z týchto časových radov uskutočníme Bartelsov test a rozhodneme o hypotéze. Potom pre každú kombináciu parametra (ρ, n) zistíme podiel úspešných zamietnutí nulovej hypotézy o neexistencii autokorelácie prvého rádu. Na obrázku znázorníme silu testu v závislosti od veľkosti vzorky. Na os y-ovú si znázorníme silu testu a na os x-ovú veľkosť vzorky n. Tieto silofunkcie zostrojíme pre každý z parametrov ρ. Nasledujúci kód pre lepšiu prehľadnosť okomentujeme.

```
> library(lawstat)
# rho je vektor autokorelačných koeficientov
> rho <- c(-0.5, -0.2, 0.2, 0.5)
# size je vektor, ktorý zodpovedá rôznym veľkostiam vzorky
> size <- seq(from = 8, to = 80, by = 3)
# power je matica, v ktorej sa bude ukladať sila testu
# riadky zodpovedajú sile testu pre veľkosť vzorky n
# stĺpce zodpovedajú sile testu pre autoregresný parameter rho
> power <- matrix(nrow = length(size), ncol = length(rho))
> colnames(p.vals) <- rho
# pomocná premenná, ktorá bude značiť stĺpec, do ktorého sa budú
# zapisovať výsledky v matici power
> s <- 1
# prvá úroveň cyklu vykonáva analýzu pre rôzne parametre rho
> for (i in rho) {
# pomocná premenná, ktorá bude značiť riadok, do ktorého sa budú
# zapisovať výsledky v matici power
+   r <- 1
# druhú úroveň cyklu vykonáva analýzu pre rôzne n
+   for (j in size) {
# pomocný vektor, do ktorého sa ukladajú p-hodnoty Bartelsovho
# testu
+     p.val <- c()
# tretia úroveň cyklu vykonáva pre každú kombináciu rho a n
# Bartelsov test a ukladá si p-hodnotu do vektora p.val
+     for (k in 1:5000) {
+       x <- arima.sim(n = j, list(ar = c(i)))
+       p.val[k] <- bartels.test(x, alternative =
  c("two.sided"))$p.value
+     }
# podiel úspešných zamietnutí sa zapíše na príslušné miesto
# v matici power
+     power[r, s] <- sum(p.val < 0.05)/5000
# mení sa parameter n, preto je potrebné zmeniť aj riadok, do
# ktorého sa budú zapisovať výsledky
+     r <- r + 1
+   }
# mení sa parameter rho, preto je potrebné zmeniť aj stĺpec, do
# ktorého sa budú zapisovať výsledky
+   s <- s + 1
```

```
+ }
```



Obrázok 4.12: Silofunkcie Bartelsovho testu

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> plot(power[, 1] ~ size, ylim = c(0, 1), type="l", col =
  "black", lwd = 2, lty = 1, family = "serif", xlab = "veľkosť
  vzorky", ylab="sila testu", cex.axis = 1.5, cex = 1.5, cex.lab
  = 1.8)
> lines(power[, 2] ~ size, type = "l", lwd = 2, lty = 2)
> lines(power[, 3] ~ size, type = "l", lwd = 2, lty = 3)
> lines(power[, 4] ~ size, type = "l", lwd = 2, lty = 4)
> legend("topleft", legend = rho, lty = 1:length(rho), inset =
  0.08, bty = "n")
```

Z predchádzajúceho obrázku (pozri Obrázok 4.12) môžeme pozorovať dva javy. S veľkosťou vzorky rastie sila testu, čo je dobrá vlastnosť. Zároveň platí, že je našim cieľom, aby táto sila testu rástla čo najrýchlejšie. S tým súvisí druhé pozorovanie, že rast aj sila testu je väčšia pre väčšiu hodnotu parametra $|\rho|$. Uvedené je v celku logické. Predchádzajúce hodnoty sa v nových hodnotách (relatívne k chybovým členom) prejavujú o to silnejšie, o čo väčšia je absolútna hodnota parametra ρ.

K tejto téme je potrebné dodať ešte aspoň dva body. Po prvé, test rieši iba otázku autokorelácie prvého rádu. Ak by bol model, ktorý by generoval hodnoty podľa nasledujúceho vzťahu:

$$X_i = \rho X_{i-2} + e_i \qquad (4.61)$$

potom takýto model by vytváral postupnosti hodnôt X_i, ktoré by boli autokorelované druhého rádu. Nie je možné jednoznačne určiť, ako by sa v takejto situácii Bartelsov test „*správal*“ (v podobe v akej sme ho uviedli). Sila testu by bola zrejme výrazne nižšia, keďže

„*efekt*“ predchádzajúcej hodnoty slabne (uvedené sa dá dokázať jednoduchými rekurzívnymi úpravami výrazu *(4.61)*). Druhým bodom je skutočnosť, že existujú výrazne lepšie testy (čo do chyby I. aj II. druhu) ako je tento test. Tieto testy (napr. Ljung – Box test a Box – Pierce test) sa používajú najmä v ekonometrii a v analýze časových radov, ale nie sú predmetom tejto publikácie.

4.6.3 Jednovzorkový Wilcoxonov znamienkový test

Wilcoxonov znamienkový test slúži na overovanie hypotézy o mediáne. Nevýhodou je predpoklad o symetrickosti rozdelenia (nemusí ísť o normálne rozdelenie, dôležité je aby nebolo zošikmené) a taktiež o spojitosti rozdelenia, z ktorého je realizovaný náhodný *iid* výber X_i, $i = 1, 2, ..., n$ (Panik, 2005; Salkind, 2007). Na druhej strane tento test zohľadňuje nie len to, či sú jednotlivé hodnoty väčšie alebo menšie ako medián ale aj poradie hodnôt, čím zohľadňuje aj nakoľko sú hodnoty „ďaleko“ od mediánu.

Najprv sa vytvorí náhodná premenná $R_i = |X_i - m_0|$, kde m_0 je medián (konštanta), ktorý predstavuje charakteristiku polohy, voči ktorej chceme porovnať výberový súbor. Ak pre ľubovoľné pozorovanie i platí $R_i = 0$, potom sa dané pozorovanie zo vzorky vylúči a zníži sa celkový počet pozorovaní n na $s < n$. Hodnoty R_i sa usporiadajú do variačného radu a priradí sa im poradie, kde najmenšia hodnota má najmenšie poradové číslo (ak nie je zhoda v poradiach, tak je to 1)[24]. V prípade zhody v hodnotách R_i, sa priradí daným hodnotám priemerné poradie. Vytvorí sa tak súbor poradí P_i. Potom nech P_i^+ sú všetky poradia, ktorých zodpovedajúci rozdiel je $X_i - m_0 > 0$ a P_i^- všetky poradia, ktorých zodpovedajúci rozdiel je $X_i - m_0 < 0$. Ďalej nech platí:

$$P^+ = \sum_{i=1}^{n} P_i^+ \qquad (4.62)$$

$$P^- = \sum_{i=1}^{n} P_i^- \qquad (4.63)$$

Ak je celkový počet hodnôt, ktoré po úpravách ostali v súbore väčší ako $s \geq 15$, potom testovacia charakteristika je (rovnakú testovaciu charakteristiku použijeme aj pri Wilcoxonovom znamienkovom teste pre dva závislé súbory):

[24] Pod zhodou v poradiach budeme rozumieť nasledujúce situácie: 1, 2, 2, 2, 5, 6, 6, 8, 9 prípadne 1, 4, 4, 4, 5, 7, 7, 8, 9. V oboch prípadoch sa zvykne voliť priemerné poradie nasledovne: 1, 3, 3, 3, 5, 6.5, 6.5, 8, 9.

$$Z_{WZ} = \frac{P^+ - \frac{s(s+1)}{4}}{\sqrt{\frac{s(s+1)(2s+1)}{24}}} \qquad (4.64)$$

Ak nebolo potrebné vykonať úpravu v empirickom súbore, potom namiesto *s* nechávame vo vzťahoch *n*. Ak si populačný medián označíme ako *m*, potom rozhodnutie o hypotézach je nasledovné:

H_0: $m = m_0$ H_1: $m \neq m_0$	Hypotézu H_0 zamietame, ak $\lvert Z_{WZ}\rvert > \lvert z_{(\alpha/2)}\rvert$
H_0: $m \leq m_0$ H_1: $m > m_0$	Hypotézu H_0 zamietame, ak $Z_{WZ} > z_{(1-\alpha)}$
H_0: $m \geq m_0$ H_1: $m < m_0$	Hypotézu H_0 zamietame, ak $Z_{WZ} < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú príslušné kvantily normovaného normálneho rozdelenia pravdepodobnosti $N(0, 1)$.

Príklad 4.19

V ďalšom príklade použijeme údaje z databázy `Mroz` z programového balíka `car`. Zaujíma nás premenná `income`, ktorá reprezentuje ročný príjem vybraných domácností (predpokladajme ich náhodný výber) v USA bez príjmu manželky (v 80-tych rokoch). Zaujíma nás, či môžeme tvrdiť, že medián príjmu domácností v danom období bol 19000,- USD (t. j. 19.00 v danej databáze) oproti obojstrannej alternatíve (pri $\alpha = 0.05$).

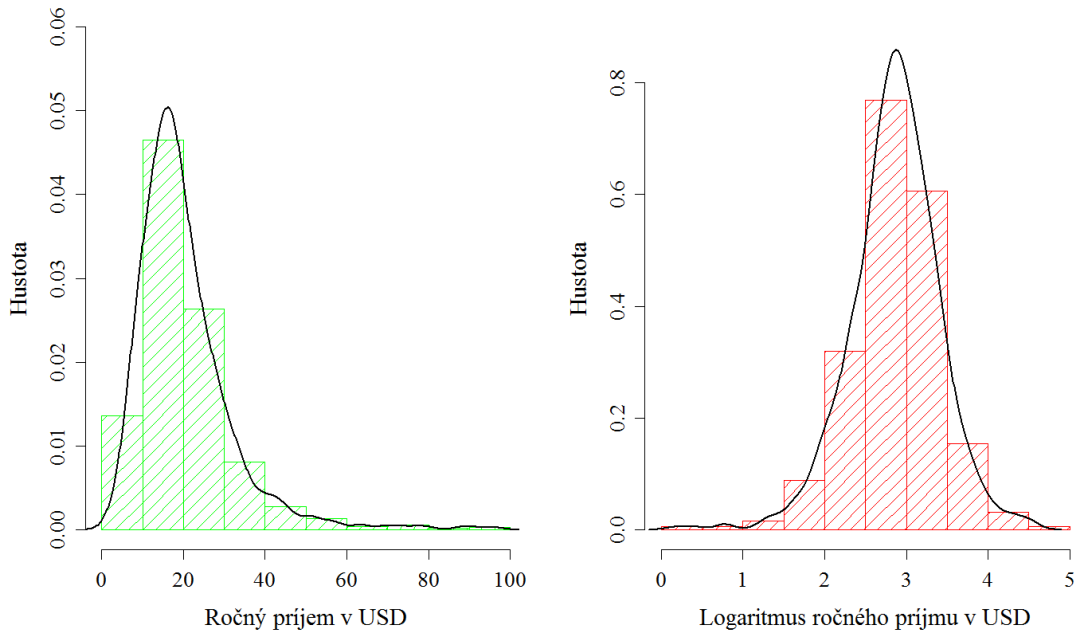
Vypočítaním základných opisných štatistík sa javí rozdelenie príjmu ako pravostranne zošikmené. Všimnime si jednak kladné šikmosti a zároveň pomerne veľký rozdiel medzi tretím kvartilom a mediánom v porovnaní s mediánom a prvým kvartilom. Zrejme by nebolo vhodné považovať toto rozdelenie za symetrické. Pri výpočte šikmosti sme použili funkciu `skewness()` z programového balíka `moments`.

```
> library(car); library(moments)
> income <- subset(Mroz$inc, subset = Mroz$inc > 0)
> summary(income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.20   13.06   17.74   20.16   24.47   96.00
> skewness(income)
[1] 2.222405
```

Vizualizácia údajov potvrdila naše podozrenie o silnom zošikmení. Uvedený tvar rozdelenia sa v ekonómii vyskytuje pomerne často. Jednou z možností je previesť transformáciu vstupných údajov. Z tohto dôvodu sme všetky hodnoty zlogaritmovali.

Týmto spôsobom sa väčšie príjmy domácnosti viac „priblížia“ k ostatným príjmom a nejavia sa ako veľmi extrémne. Rozdelenie je na pravej strane (pozri Obrázok 4.13).

```
> par(mfrow = c(1, 2))
> hist(income, density = 10, col = "green", main = NA, cex.lab =
  1.5, ylim = c(0, 0.06), cex.axis = 1.3, freq = FALSE, ylab =
  "Hustota", family = "serif", xlab = "Ročný príjem v USD")
> lines(density(income), col = "black", lwd = 2)
> income <- log(income)
> hist(income, density = 10, col = "red", main = NA, cex.lab =
  1.5, ylim = c(0, 0.9), cex.axis = 1.3, freq = FALSE, ylab =
  "Hustota", family = "serif", xlab = "Logaritmus ročného príjmu
  v USD")
> lines(density(income), col = "black", lwd = 2)
```



Obrázok 4.13: Histogram ročných príjmov pred a po logaritmickej transformácii hodnôt

Zdroj: vlastné spracovanie, výstup zo softvéru R

Po zlogaritmovaní pôvodných hodnôt vyzerá rozdelenie viac symetrické. Opisné štatistiky tento názor podporujú (existuje aj formálny test na výpočet významnosti šikmosti, ktorý však nie je predmetom tejto publikácie). Rozhodneme sa tak použiť Wilcoxonov znamienkový test, kde medián porovnáme s hodnotou ln(19000) = 9.852194 (na našej škále ln(19) = 2.944439). Upozorňujeme, že po transformácií údajov je interpretácia spravidla problematická. Vzhľadom na to, že sa v tomto teste pracuje s poradiami to však nepovažujeme za zásadný interpretačný problém (po logaritmickej transformácii sa poradia

hodnôt nemenia). Keďže *p*-hodnota je menšia ako hladina významnosti $\alpha = 0.05$, hypotézu H_0 zamietame.

```
> skewness(income)
[1] -0.4187914
> wilcox.test(income, mu = 2.944439, alternative =
  ("two.sided"))

      Wilcoxon signed rank test with continuity correction

data:  income
V = 118996, p-value = 0.0001524
alternative hypothesis: true location is not equal to 2.944439
```

4.6.4 Mann – Whitney – Wilcoxonov (Mann – Whitney U) test pre dve nezávislé vzorky

Tento test slúži ako určitá alternatíva k *t*-testu dvoch nezávislých súborov, prípadne je možné ho vnímať tiež ako alternatívu ku Kolmogorov – Smirnovovmu testu. Majme dva nezávislé *iid* súbory X_i, $i = 1, 2, ..., n_x$ a Y_j, $j = 1, 2, ..., n_y$. Overujeme nulovú hypotézu, že rozdelenia pravdepodobnosti, z ktorého sa tieto súbory vyberali, sú rovnaké. Namerané hodnoty X_i a Y_j spojíme do jedného súboru, ktorý si označíme ako R a vytvoríme z neho variačný rad $R_{(l)}$, $l = 1, 2, ..., n_x + n_y$. Potom každej hodnote priradíme poradie. Označme R_x sumu poradí výberového súboru X_i a R_y sumu poradí výberového súboru Y_j. Intuitívne, ak sú obe vzorky rovnako početné a pochádzajú z rovnakého rozdelenia, potom rozdiely v týchto sumách by nemali byť výrazné (len náhodné). Tento princíp využíva Mann – Whitney – Wilcoxonov test pre dve nezávislé vzorky. Predpokladajme, že veľkosť oboch vzoriek je $n_x, n_y \geq 10$, potom štatistika U má tvar:

$$U_x = R_x - \frac{n_x(n_x+1)}{2} \tag{4.65}$$

$$U_y = R_y - \frac{n_y(n_y+1)}{2} \tag{4.66}$$

kde $n_x(n_x + 1) / 2$ je hypotetický súčet poradí vo vzorke X_i, ak by boli všetky hodnoty X_i menšie ako hodnoty Y_j. Podobne, $n_y(n_y + 1)/2$ je hypotetický súčet poradí vo vzorke Y_j, ak by boli všetky hodnoty Y_j menšie ako hodnoty X_i. Ak uvažujeme o početnejšej vzorke, potom je jedno ktorú štatistiku U použijeme. Testovaciu štatistiku definujeme nasledovne:

$$Z_U = \frac{U - E(U)}{\sqrt{D(U)}} \tag{4.67}$$

pričom pri platnosti nulovej hypotézy sa riadi normálnym rozdelením pravdepodobnosti $N(0, 1)$, kde stredná hodnota U:

$$E(U) = \frac{n_x n_y}{2} \tag{4.68}$$

a rozptyl U:

$$D(U) = \frac{n_x n_y (n_x + n_y + 1)}{12} \tag{4.69}$$

Definujme si rozdelenia, z ktorých sú výberové súbory vyberané ako $f(x)$ a $f(y)$. Rozhodnutie o hypotéze je potom nasledovné:

H_0: $f(x) = f(y)$ H_1: $f(x) \neq f(y)$	Hypotézu H_0 zamietame, ak $\lvert Z_U \rvert > \lvert z_{(1-\alpha/2)} \rvert$
H_0: $f(x) \leq f(y)$ H_1: $f(x) >_{SCH} f(y)$	Hypotézu H_0 zamietame, ak $\lvert Z_U \rvert > z_{(1-\alpha)}$

Všimnime si, že alternatívna hypotéza pri obojstrannom teste nie je stanovená štandardným spôsobom. To je častý problém neparametrických testov akým je aj tento. Problém spočíva v stanovení nulovej hypotézy, ktorá tvrdí, že ide o rovnaké rozdelenia, avšak takéto rozdelenia môžu mať viac ako len jeden parameter. Týmto spôsobom sa tak overuje zhoda v niekoľkých parametroch. Ak však hypotéza neplatí, môže to byť prejavom odlišnosti v jednom z parametrov rozdelenia. Z povahy testu potom vyplýva, že ak je rozdiel medzi R_x a R_y (pri rovnakom počte pozorovaní v oboch súboroch) kladný a dostatočne veľký, tak hodnoty X_i sú väčšie ako hodnoty Y_j, a teda rozdelenie $f(x)$ by malo byť „posunuté“ vpravo od rozdelenia $f(y)$. Ak platí predpoklad, že tvar rozdelení $f(x)$ a $f(y)$ je rovnaký, potom je možná interpretácia o „posune“ rozdelení. V prípade, ak je tento predpoklad porušený, test je stále možné použiť avšak mení sa interpretácia. V prípade platnosti alternatívnej hypotézy, $f(x) >_{SCH} f(y)$, je možné výsledok interpretovať tak, že $f(x)$ stochasticky dominuje $f(y)$. Formálnejšie $P(X > a) > P(Y > a)$, kde a je ľubovoľná realizácia z náhodnej premennej X_i a Y_j. V programe R môžeme na výpočet tohto testu použiť funkciu `wilcox.test()`.

Príklad 4.20

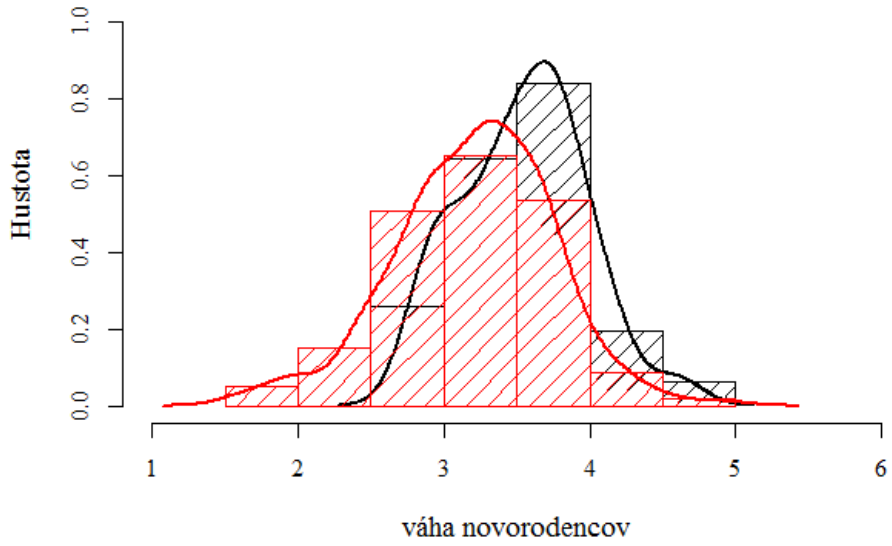
V nasledujúcom príklade použijeme databázu `babies` z programového balíka `UsingR`. Zaujíma nás, či váha novorodencov (premenná `wt`) pochádza z rovnakého rozdelenia, ak porovnáme dve skupiny novorodencov. V prvej skupine sú obaja rodičia Afroameričania (súbor označujeme ako `wt_b`) a v druhej sú obaja rodičia mexického pôvodu (súbor

označujeme ako `wt_x`). Alternatívnou hypotézou je, že novorodenci v prvej skupine sú ťažší ako v druhej skupine. Najprv si váhu prevedieme na jednotky *kg*.

```
> library(UsingR); attach(babies)
> wt_b <- subset(wt, subset = race == 7 & drace ==
  7)*28.3495231/1000
> wt_x <- subset(wt, subset = race == 6 & drace ==
  6)*28.3495231/1000
------------------------------------------------------------
> wilcox.test(wt_b, wt_x, alternative = c("greater"), correct =
  F)

                    Wilcoxon rank sum test

data:  wt_b and wt_x
W = 2265.5, p-value = 0.9997
alternative hypothesis: true location shift is greater than 0
```

Nulovú hypotézu sme neboli schopní zamietnuť. Nemáme teda dostatok dôkazov na to, aby sme mohli povedať, že novorodenci majú odlišnú váhu, voči alternatíve, že v prvej skupine sú deti ťažšie. Aby sme mali možnosť posúdiť vzťah týchto dvoch rozdelení, tak si ich tvar zobrazíme na ďalšom obrázku (pozri Obrázok 4.14). Z obrázku je vidno, že sa histogramy do značnej miery prekrývajú. Zrejme tak ide o realizácie z rovnakého rozdelenia pravdepodobnosti.

```
> hist(wt_x, density = 10, col = "black", main = NA, cex.lab =
  1.2, cex.axis = 1.0, freq = FALSE, ylab = "Hustota", family =
  "serif", xlab = "váha novorodencov", xlim = c(1, 6), ylim =
  c(0, 1.1))
> lines(density(wt_x), col = "black", lwd = 2)
> hist(wt_b, add = TRUE, density = 10, col = "red", main = NA,
  freq = FALSE, xaxt = "n", yaxt = "n", xlab = "", ylab = "")
> lines(density(wt_b), col = "red", lwd = 2)
```



Obrázok 4.14: Histogram váh novorodencov – dve skupiny

Zdroj: vlastné spracovanie, výstup zo softvéru R

4.6.5 Kruskal – Wallisov test pre nezávislé vzorky

Rozšírením Mann – Whitney – Wilcoxonovho testu pre dve nezávislé vzorky na k vzoriek sa dostávame ku Kruskal – Wallisovmu testu. Predpokladajme, že máme k náhodných *iid* vzoriek, ktorých hodnoty si označíme ako $X_{i,j}$, kde $i = 1, 2, ..., n_j$ a $j = 1, 2, ..., k$. Celkový počet pozorovaní je tak $n = n_1 + n_2 + ... n_k$. Ďalej predpokladajme, že $k \geq 3$ a zároveň pre všetky j platí $n_j > 5$ a nakoniec, že náhodné premenné $X_{i,j}$ sú realizáciami zo spojitých rozdelení $f(x_j)$. Predpokladá sa, že rozdelenia $f(x_j)$ sú si okrem charakteristiky polohy inak rovnaké. Formulácia hypotéz znovu nie je jednoznačná. Overujeme hypotézu, že tieto rozdelenia sú si rovné. Presnejšie sa porovnáva rovnosť poradí hodnôt (v niektorých publikáciách sa uvádza nulová hypotéza ako rovnosť k populačných mediánov oproti alternatíve, že aspoň jeden populačný medián je odlišný). Alternatívnu hypotézu si môžeme naformulovať ako prítomnosť aspoň jedného rozdelenia, ktoré je odlišné od ostatných (sleduje sa tým posun polohy rozdelenia).

Postup výpočtu je podobný ako v prípade Mann – Whitney – Wilcoxonovho testu. Hodnotám $X_{i,j}$, sa priradí poradie, kde najmenšie poradie sa priradí najmenšej hodnote (najmenším hodnotám). V prípade zhody poradí sa priradí priemerné poradie. Vytvorí sa tak náhodná premenná poradí $R_{i,j}$. Potom označíme R_j ako sumu poradí j-tého súboru. V prípade, ak v súbore neexistovala žiadna zhoda v poradiach, testovacia charakteristika sa vypočíta ako:

$$KW = \frac{12}{n(n+1)}\left(\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3(n+1) \qquad (4.70)$$

V prípade, ak sa v súbore vyskytujú zhody v poradiach, testovacia charakteristika *KW* sa môže upraviť nasledovne:

$$KW_A = \frac{KW\left(n^2 - n\right)}{\left(n^2 - n - \sum_{p=1}^{P} T_p\right)} \qquad (4.71)$$

kde $p = 1, 2, ..., P$ predstavuje jednotlivé skupiny zhodných poradí, ďalej $T_p = t_p^3 - t_p$, kde t_p je počet zhodných poradí v skupine p zhodných poradí. Ak sa v súbore vyskytuje poradie 25.5 dvakrát (dosiahlo sa spriemerovaním poradia 25 a 26) a poradie 32 trikrát, potom $p = 1, 2$, t.j. $P = 2$ a $T_1 = 6$, $T_2 = 24$.

Testovacia charakteristika zo vzťahu *(4.70)* a *(4.71)* sa za predpokladu platnosti nulovej hypotézy riadi Chí-kvadrát rozdelením s $(k - 1)$ stupňami voľnosti. Rozhodnutie o hypotéze je potom nasledovné:

H_0: $f(x_1) = f(x_2) = ... = f(x_k)$	Hypotézu H_0 zamietame, ak
H_1: aspoň v jednom prípade $f(x_j) \neq f(x_w)$, kde $j \neq w$	$KW > \chi^2_{\alpha, (k-1)}$

V programe R získame kritickú hodnotu použitím funkcie `qchisq()`. Výpočet Kruskal – Wallisovho testu v programe R vieme vykonať použitím funkcie `kruskal.test()`. Použitie testu si predvedieme na údajoch z databázy `States` z programového balíka `car`.

Príklad 4.21

Databáza obsahuje výsledky z testovania študentov v jednotlivých štátoch USA. Zaujímajú nás rozdiely vo výsledkoch testov z matematiky (premenná `SATM`). Použijeme štyri z deviatich regiónov, do ktorých patria jednotlivé štáty USA. Ide o regióny: `MTN` („Mountain), `NE` („New Englang“), `SA` („South Atlantic“), `WNC` („West North Central“). Overujeme hypotézu, že rozdelenia výsledkov v týchto regiónoch sú navzájom rovnaké, oproti hypotéze, že v aspoň jednom regióne je rozdelenie odlišné (v strednej hodnote). Najprv prepočet uskutočníme priamo v programe R bez použitia funkcie `kruskal.test()`.

```
> library(car); attach(States)
> table(region)
region
```

```
ENC ESC  MA MTN  NE PAC  SA WNC WSC
  5   4   3   8   6   5   9   7   4
> region
 [1] ESC PAC MTN WSC PAC MTN NE  SA  SA  SA  SA  PAC MTN ENC ENC
  WNC WNC ESC WSC
[20] NE  SA  NE  ENC WNC ESC WNC MTN WNC MTN NE  MA  MTN MA  SA
  WNC ENC WSC PAC
[39] MA  NE  SA  WNC ESC WSC MTN NE  SA  PAC SA  ENC MTN
Levels: ENC ESC MA MTN NE PAC SA WNC WSC
> math_scores <- subset(States, region %in% c("MTN", "NE", "SA",
  "WNC"))
> detach(States)
> attach(math_scores)
> n <- dim(math_scores)[1]
> R <- rank(SATM)
> sort(R)
 [1]  1.0  2.0  3.0  4.0  5.0  6.0  7.5  7.5  9.5  9.5 11.0 12.0
  13.0 14.0 15.0
[16] 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 26.0 27.0
  28.0 29.0 30.0
> T1 <- 4; T2 <- 4; T <- T1 + T2 #prepočet potrebný pre úpravu
  zhodných poradí
> Rj <- split(R, region)
> MTN <- sum(Rj$MTN); NE <- sum(Rj$NE); SA <- sum(Rj$SA); WNC <-
  sum(Rj$WNC)
> KW <- (12/(n*(n + 1))) * (MTN^2/sum(region == "MTN") +
  NE^2/sum(region == "NE") + SA^2/sum(region == "SA") +
  WNC^2/sum(region == "WNC")) - 3*(n + 1); KW
[1] 23.81103
> KWa <- (KW*(n^2 - n))/(n^2 - n - T); KWa
[1] 24.03202
> qchisq(0.95, df = 3)
[1] 7.814728
> 1 - pchisq(KWa, df = 3)
[1] 2.459843e-05
```

Hodnota testovacej charakteristiky KW_A je výrazne vyššia ako kritická hodnota. Pre úplnosť sme vypočítali aj *p*-hodnotu, ktorá je výrazne nižšia ako hladiny významnosti 5 %. Výsledky z funkcie `kruskal.test()` sú podobné (rozdielnosť spočíva v tom, že `kruskal.test()` nerobí korekciu v prípade zhody poradí). Taktiež na hladine významnosti α = 5 % hypotézu o zhode rozdelení zamietame. Zrejme existuje aspoň jeden región, kde sú výsledky výrazne odlišné od ostatných. Ďalším krokom by mohlo byť porovnanie mier polohy pre jednotlivé regióny, prípadne vykonať neparametrické testy (ako Mann – Whitney – Wilcoxonov test), pomocou ktorých by sa zistilo, v ktorých regiónoch sa dosiahli vyššie (nižšie) bodové výsledky.

```
> kruskal.test(math_scores$SATM, math_scores$region)

                Kruskal-Wallis rank sum test
```

```
data:  math_scores$SATM and math_scores$region
Kruskal-Wallis chi-squared = 23.8216, df = 3,
p-value = 2.722e-05
```

4.6.6 Wilcoxonov znamienkový test pre dve závislé vzorky

Wilcoxonovým znamienkovým testom sa overujú rozdiely v závislých vzorkách. Ide o alternatívu k párovému *t*-testu, ktorý je vhodné použiť najmä pri malých vzorkách. Majme usporiadané dvojice pozorovaní (X_i, Y_i), $i = 1, 2, ..., n$, ktoré predstavujú náhodnú *iid* vzorku. Overuje sa hypotéza o rovnosti rozdelení, z ktorých výberové súbory X_i a Y_i pochádzajú. Ďalej definujme náhodnú premennú $R_i = |X_i - Y_i|$, $i = 1, 2, ..., n$, ktorá predstavuje rozdiel v nameraných hodnotách (podobne ako pri párovom *t*-teste). Ak pre nejaké pozorovanie i platí $R_i = 0$, potom uvedené pozorovanie sa vylúči a veľkosť súboru sa adekvátne zníži z n na $m < n$. Z R_i sa vytvorí variačný rad, v ktorom sú hodnoty usporiadané od najmenšej po najväčšiu. Následne sa každej hodnote priradí poradie a vytvorí sa tak súbor s poradiami P_i. V prípade, že niektoré hodnoty sú zhodné, priradí sa im priemerné poradie. Hodnoty v súbore P_i sa ďalej rozdelia na tie, kde bol rozdiel $R_i = X_i - Y_i > 0$ a na tie, kde bol rozdiel $R_i = X_i - Y_i < 0$. Prvú skupinu poradí označíme ako P_i^+ a druhú ako P_i^-. Nakoniec si vypočítame nasledujúce dva súčty:

$$P^+ = \sum_{i=1}^{n} P_i^+ \tag{4.72}$$

$$P^- = \sum_{i=1}^{n} P_i^- \tag{4.73}$$

Potom pre väčšie vzorky (spravidla $m \geq 25$) má testovacia charakteristika nasledujúci tvar:

$$Z_M = \frac{P^+ - \frac{m(m+1)}{4}}{\sqrt{\frac{m(m+1)(2m+1)}{24}}} \tag{4.74}$$

kde m je konečný počet pozorovaní (v prípade, ak sa vyskytli nulové rozdiely R_i, tak $m < n$). Testovacia charakteristika Z_M sa riadi normovaným normálnym rozdelením pravdepodobnosti. Definujme si rozdelenia, z ktorých sú výberové súbory vyberané ako $f(x)$ a $f(y)$. Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $f(x) = f(y)$ H_1: $f(x) \neq f(y)$	Hypotézu H_0 zamietame, ak $\lvert Z_M \rvert > \lvert z_{(\alpha/2)} \rvert$
H_0: $f(x) \leq f(y)$ H_1: $f(x) >_{SCH} f(y)$	Hypotézu H_0 zamietame, ak $Z_M > z_{(1-\alpha)}$
H_0: $f(x) \geq f(y)$ H_1: $f(x) <_{SCH} f(y)$	Hypotézu H_0 zamietame, ak $Z_M < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú príslušné kvantily normovaného normálneho rozdelenia pravdepodobnosti.

Príklad 4.22

Máme skupinu žien, ktoré sa zúčastnili programu zdravej výživy. Zaujíma nás, ako sa zmenila hmotnosť žien, t.j. rozdiel hmotnosti pred a po absolvovaní programu. Overujeme nulovú hypotézu o rovnosti (rovný väčší) rozdelení oproti alternatíve, že hmotnosť žien je menšia po programe ako pred programom. Výpočet vykonáme priamo v programe R a potom pomocou funkcie `wilcox.test()`.

```
> predtym <- c(90.1, 70.5, 75.9, 74.4, 80.8, 74.9, 91.1, 87.5,
  68.1, 83.7, 69.6, 82.6, 76.9, 72.9, 81.8, 85.6, 79.4, 84.8,
  88.5, 81.5, 89.3, 86.0, 68.1, 85.4, 73.1, 77.0, 83.0, 88.2,
  76.3, 82.9, 84.1, 68.4, 70.6, 70.1, 76.2, 83.7, 86.3, 91.5,
  88.1, 78.1, 72.3, 73.6, 81.3, 84.1, 74.7, 72.6, 80.3, 68.3,
  70.7, 89.2, 73.9, 88.3, 72.8, 77.6, 82.5, 87.4)
> potom <- c(81.4, 82.1, 86.8, 68.4, 85.8, 81.7, 86.3, 79.2,
  76.4, 84.7, 83.9, 80.5, 81.5, 87.0, 74.0, 78.1, 78.3, 76.0,
  62.4, 77.6, 77.4, 77.9, 71.8, 66.8, 67.5, 72.2, 90.1, 79.9,
  70.5, 80.9, 73.1, 67.2, 81.5, 77.8, 70.8, 85.6, 78.0, 73.4,
  67.6, 73.9, 82.8, 67.2, 67.6, 86.1, 74.0, 78.9, 74.9, 82.6,
  74.7, 74.0, 87.3, 69.6, 79.7, 67.7, 87.9, 87.2)
> Ri <- abs(predtym - potom)
> Pi <- rank(Ri)
> Si <- sign(predtym - potom)
> data <- data.frame(predtym, potom, Ri, Pi, Si)
> P_plus <- subset(data$Pi, subset = Si == 1)
> P_plus_s <- sum(P_plus)
> P_plus_s
[1] 976
> m <- length(Si)
> ZM <- (P_plus_s - (m*(m + 1))/4)/sqrt((m*(m + 1)*(2*m +
  1))/24); ZM
[1] 1.451961
> p_value <- 1 - pnorm(ZM); p_value
[1] 0.07325622
```

Hypotézu H_0 nevieme zamietnuť na hladine významnosti $\alpha = 0.05$, avšak boli by sme ju schopní zamietnuť na hladine $\alpha = 0.1$. Prakticky rovnaký výsledok nám vyšiel použitím funkcie `wilcox.test()`.

```
> wilcox.test(predtym, potom, paired = T, exact = F, alternative
  = c("greater"))

     Wilcoxon signed rank test with continuity correction

data:  predtym and potom
V = 976, p-value = 0.07382
alternative hypothesis: true location shift is greater than 0
```

4.6.7 Friedmanov test pre závislé vzorky

V predchádzajúcom teste sa porovnávali dve závislé vzorky. Friedmanov test sa používa na overenie zhody rozdelení viac ako dvoch vzoriek (spracované podľa Corder – Foreman, 2009; Ramachandran – Tsokos, 2009; Sprent – Smeeton; 2000). Predpokladajme, že máme *k* náhodných *id* vzoriek (jedno *i* sme vynechali, keďže nejde o nezávislé vzorky, t. j. *Independent*, ale ide o závislé vzorky), ktorých hodnoty si označíme ako $X_{i,j}$, kde $i = 1, 2, ..., n_j$ a $j = 1, 2, ..., k$. Ďalej predpokladajme, že náhodné premenné $X_{i,j}$ sú realizáciami zo spojitých rozdelení $f(x_j)$ a rozdelenia $f(x_j)$ sú okrem charakteristiky polohy inak rovnaké. V niektorých publikáciách sa môžeme stretnúť s formuláciou nulovej a alternatívnej hypotézy vo forme zhody mediánov voči alternatíve, že aspoň jedna dvojica mediánov je navzájom odlišná. Budeme sa držať pôvodnej formulácie hypotéz, keďže charakter výpočtu testu skôr hovorí o rozdelení poradí. Na rozdiel od Kruskal – Wallisovho testu sa poradia nevytvárajú tak, že sa všetky údaje spoja do jednej vzorky, ale poradia sa vytvárajú pre každý prvok X_i. Ilustrujeme si to na príklade. Premenná, ktorá nás zaujíma, je počet úmyselného ublíženia na zdraví v jednotlivých krajoch na Slovensku v jednotlivých rokoch 2001 až 2010. Údaje sú v nasledujúcej tabuľke . Zaujíma nás, či došlo k zmenám rozdelenia úmyselného ublíženia na zdraví medzi jednotlivými rokmi. Túto hypotézu môžeme vnímať aj tak, či došlo k zmenám (populačných) mediánov medzi jednotlivými rokmi.

Tabuľka 13: Počet úmyselného ublíženia na zdraví

Kraj	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Bratislavský kraj	334	422	472	402	458	359	308	273	314	243
Trnavský kraj	362	372	423	351	364	321	298	279	203	187
Trenčiansky kraj	306	357	320	306	323	264	256	235	199	172
Nitriansky kraj	421	450	469	455	429	366	367	294	342	315
Žilinský kraj	593	755	628	633	618	549	467	417	422	417
Banskobystrický kraj	607	710	554	557	546	462	408	377	374	397
Prešovský kraj	504	546	489	482	523	387	330	335	345	282
Košický kraj	536	691	665	591	584	471	519	432	391	395

Zdroj: vlastné spracovanie, údaje z www.statistics.sk

V tejto tabuľke index i = 1, 2, ..., 8, n = 8 predstavuje kraje a index j = 1, 2, ..., 10 predstavuje roky, teda k = 10. Keďže sa sledoval počet úmyselného ublíženia na zdraví v rovnakých krajoch, ide o závislé vzorky. Predpokladáme, že tieto čísla sú náhodné realizácie. Vytvoríme si najprv poradia pre prvky X_i – výsledky sú v nasledujúcej tabuľke (Tabuľka 14).

Tabuľka 14: Pomocné poradia k Friedmanovmu testu

Kraj	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Bratislavský kraj	5	8	10	7	9	6	3	2	4	1
Trnavský kraj	7	9	10	6	8	5	4	3	2	1
Trenčiansky kraj	6.5	10	8	6.5	9	5	4	3	2	1
Nitriansky kraj	6	8	10	9	7	4	5	1	3	2
Žilinský kraj	6	10	8	9	7	5	4	1.5	3	1.5
Banskobystrický kraj	9	10	7	8	6	5	4	2	1	3
Prešovský kraj	8	10	7	6	9	5	2	3	4	1
Košický kraj	6	10	9	8	7	4	5	3	1	2
S_j (súčet skupiny j)	53.5	75.0	69.0	59.5	62.0	39.0	31.0	18.5	20.0	12.5

Zdroj: vlastné spracovanie

Pre každý rok tieto poradia sčítame a súčet si označíme ako S_j. Friedmanov test využíva nasledujúci princíp. Ak by boli v jednotlivých rokoch počty úmyselného ublíženia na zdraví rovnaké, potom by sme mali sledovať približne konštantné súčty poradí S_j. Už z predchádzajúcej tabuľky (Tabuľka 14) je zrejmý prudký pokles. Je otázne, či je spôsobený tak razantnou zmenou v správaní sa obyvateľstva alebo ide o metodologickú zmenu v evidencii. Formálny štatistický test prevedieme nasledovne. Ak by v poradiach neboli žiadne zhody, potom je testovacia štatistika:

$$FR = \frac{12\sum_{j=1}^{k} S_j^2}{nk(k+1)} - 3n(k+1) \qquad (4.75)$$

V prípade, ak v poradiach existujú zhody, potom sa testovacia štatistika vypočíta ako:

$$FR_A = \frac{n(k-1)\left(\sum_{j=1}^{k}\frac{S_j^2}{n} - \left((1/4)nk(k+1)^2\right)\right)}{\sum_{i=1}^{n}\sum_{j=1}^{k} r_{i,j}^2 - \left((1/4)nk(k+1)^2\right)} \qquad (4.76)$$

kde r_{ij} sú jednotlivé poradia. Testovacia charakteristika FR a FR_A sa riadi Chí-kvadrát rozdelením pravdepodobnosti s (k – 1) stupňami voľnosti. Rozhodnutie o hypotéze je potom nasledovné:

H_0: $f(x_1) = f(x_2) = ... = f(x_k)$	Hypotézu H_0 zamietame, ak
H_1: aspoň v jednom prípade $f(x_j) \neq f(x_w)$, kde $j \neq w$	$FR > \chi^2_{\alpha,\,(k-1)}$

Zamietnutie nulovej hypotézy sa chápe ako zmena v polohe rozdelenia. Dokončíme začatý príklad, na ktorom si ukážeme použitie tohto testu. Manuálnym prepočtom sme dospeli k hodnote testovacej charakteristiky FR = 63.45. Keďže však boli prítomné zhody v poradiach, vypočítali sme FR_A = 63.55 a následne priamo p-hodnotu, ktorá bola výrazne nižšia ako nominálnych 5 %. Hypotézu H_0 teda zamietame. Inak povedané, máme dostatok dôkazov, aby sme mohli tvrdiť, že počty úmyselného ublíženia na zdraví sa medzi aspoň dvoma rokmi zmenili – došlo k posunu strednej hodnoty.

```
> data <- matrix(c(334, 362, 306, 421, 593, 607, 504, 536, 422,
  372, 357, 450, 755, 710, 546, 691, 472, 423, 320, 469, 628,
  554, 489, 665, 402, 351, 306, 455, 633, 557, 482, 591, 458,
  364, 323, 429, 618, 546, 523, 584, 359, 321, 264, 366, 549,
  462, 387, 471, 308, 298, 256, 367, 467, 408, 330, 519, 273,
  279, 235, 294, 417, 377, 335, 432, 314, 203, 199, 342, 422,
  374, 345, 391, 243, 187, 172, 315, 417, 397, 282, 395), ncol =
  10)
> for (i in 1:8) {
+ data[i,] <- rank(data[i,])
+ }
> Sj <- apply(data, 2, sum)
> rij <- sum(data^2)
> n <- dim(data)[1]
> k <- dim(data)[2]
> FR <- (12*sum(Sj^2))/(n*k*(k + 1)) - 3*n*(k + 1); FR
[1] 63.45
> FRA <- (n*(k - 1) * (sum(Sj^2)/n - ((1/4)*n*k*(k +
  1)^2)))/(rij-((1/4)*n*k*(k + 1)^2)); FRA
[1] 63.54628
> 1 - pchisq(FRA, df = 9)
[1] 2.764325e-10
```

Pri výpočte môžeme použiť aj priamo funkciu `friedman.test()`. Vyžaduje si to však správne definovanie databázy. Všetky hodnoty sme zoradili do jednej premennej a k nej sme vytvorili ďalšie dve premenné. Jedna zodpovedala roku (premenná `rok`), v ktorom bolo pozorovanie uskutočnené a druhá kraju (premenná `kraj`), v ktorom bolo pozorovanie zaznamenané. Takto vytvorené premenné sme spojili do jednej databázy, ktorú sme si nazvali `untc` (ide o panelové dáta). Výsledky sú prakticky identické s predošlým manuálnym prepočtom. Všimnime si, že funkcia `friedman.test()` používa priamo korekciu poradí tak, ako sme si ju definovali aj my.

```
> untc <- data.frame( pocty <- c(334, 362, 306, 421, 593, 607,
  504, 536, 422, 372, 357, 450, 755, 710, 546, 691, 472, 423,
  320, 469, 628, 554, 489, 665, 402, 351, 306, 455, 633, 557,
  482, 591, 458, 364, 323, 429, 618, 546, 523, 584, 359, 321,
  264, 366, 549, 462, 387, 471, 308, 298, 256, 367, 467, 408,
```

```
  330, 519, 273, 279, 235, 294, 417, 377, 335, 432, 314, 203,
  199, 342, 422, 374, 345, 391, 243, 187, 172, 315, 417, 397,
  282, 395), rok <- factor(sort(rep(2001:2010, 8))), kraj <-
  factor(rep(1:8, 10)))
------------------------------------------------------------------
> friedman.test(pocty ~ rok | kraj, untc)

                  Friedman rank sum test

data:  pocty and rok and kraj
Friedman chi-squared = 63.5463, df = 9, p-value = 2.764e-10
```

4.6.8 Levenov test zhody rozptylov

Ide o alternatívu k *F*-testu zhody dvoch rozptylov. Levenov test slúži na overovanie hypotézy, že variabilita v *k* skupinách je rovnaká oproti alternatíve, že medzi aspoň jednou dvojicou skupín existuje rozdiel vo variabilite. Spomínali sme pomerne nevhodné štatistické vlastnosti *F*-testu, ktorý je citlivý na porušenie predpokladu o normalite. Levenov test je alternatíva v prípade, ak je rozdelenie symetrické, nie však nutne normálne.

Majme náhodný *iid* súbor $X_{i,j}$, $i = 1, 2, ..., n$ a $j = 1, 2, ..., k$, kde i sú jednotlivé pozorovania a j predstavujú skupiny. Ak porovnávame variabilitu v dvoch skupinách, potom $k = 2$, početnosť hodnôt v jednotlivých skupinách si označíme ako n_j. Ďalej si vytvoríme novú náhodnú premennú $Z_{i,j}^a = \left| X_{i,j} - \overline{X}_j \right|$. Testovacia charakteristika má potom tvar:

$$L = \frac{(n-k)\sum_{j=1}^{k} n_j \left(\overline{Z}_j^a - \overline{Z}^a\right)^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(Z_{i,j}^a - \overline{Z}_j^a\right)^2} \qquad (4.77)$$

kde $\overline{Z}_j^a$ je priemer v skupine j, $\overline{Z}^a$ je celkový priemer. Testovacia charakteristika L sa riadi F rozdelením pravdepodobnosti so stupňami voľnosti $df_1 = (k - 1)$, $df_2 = (n - k)$. Pri formulovaní hypotéz sme použili označenie smerodajných odchýlok. V skutočnosti sa testuje zhoda vo variabilite, ktorá sa však nemeria cez rozptyl ani smerodajnú odchýlku (pozri vzorec vyššie). Rozhodovanie o hypotéze je potom nasledovné:

H_0: $\sigma_1 = \sigma_2 = ... = \sigma_k$ H_1: aspoň v jednom prípade $\sigma_j \neq \sigma_p,\ j \neq p$	Hypotézu H_0 zamietame, ak $L > F_{(1-\alpha),(k-1),(n-k)}$

Príklad 4.23

Použitie Levenovho testu si ukážeme na databáze `babies`, na premennej váha novorodencov. Vytvoríme dve vzorky váh novorodencov tak, ako sme to urobili v kapitole venujúcej sa Mann – Whitney – Wilcoxonovmu testu (Príklad 4.20). Zaujíma nás, či môžeme predpokladať, že variabilita váh je v oboch skupinách rovnaká oproti alternatíve, že je rôzna. Na základe výberových rozptylov sa zdá, že vo vzorke novorodencov, kde sú rodičia mexického pôvodu, je variabilita váh novorodencov menšia. Ide o štatisticky významný rozdiel? Najprv uskutočníme manuálny prepočet priamo v programe R, ktorý je pomerne komplikovaný, preto budeme vytvárať rôzne predbežné výsledky. Následne prepočet uskutočníme pomocou funkcie `leveneTest()`, ktorá je súčasťou programového balíka `car`.

```
> library(UsingR); attach(babies)
> wt_b <- subset(wt, subset = race == 7 & drace ==
  7)*28.3495231/1000
> wt_x <- subset(wt, subset = race == 6 & drace ==
  6)*28.3495231/1000
-----------------------------------------------------------------
> var(wt_b); var(wt_x)
[1] 0.2958725
[1] 0.1803544
> wt_b_z <- abs(wt_b - mean(wt_b))
> wt_x_z <- abs(wt_x - mean(wt_x))
> wt_bx_z <- c(wt_b_z, wt_x_z)
> n <- length(wt_bx_z)
> k <- 2
> n_b <- length(wt_b)
> n_x <- length(wt_x)
> a <- n_b * (mean(wt_b_z) - mean(wt_bx_z))^2
> b <- n_x * (mean(wt_x_z) - mean(wt_bx_z))^2
> m <- (n - k) * (a + b)
> c <- sum((wt_b_z - mean(wt_b_z))^2)
> d <- sum((wt_x_z - mean(wt_x_z))^2)
> ci <- (k - 1) * (c + d)
> L <- m / ci; L
[1] 1.818195
> length(c(wt_b, wt_x))
[1] 267
> qf(0.95, df1 = 1, df2 = 265)
[1] 3.876789
> 1 - pf(L, df1 = 1, df2 = 265)
[1] 0.1786795
```

Použitím funkcie `leveneTest()` dosiahneme podobný výsledok, na základe ktorého nevieme zamietnuť nulovú hypotézu. Naše výsledky tak naznačujú, že váha novorodencov u oboch typov rodičov má podobnú variabilitu. Funkcia `leveneTest()`

požaduje, aby údaje zo všetkých skupín boli uvedené v jednom dátovom vektore. Zároveň je potrebné nadefinovať nový vektor (nazvali sme ho „g“), ktorý bude predstavovať indikátorovú premennú, ktorá bude signalizovať, či sa daná váha týka prvej alebo druhej vzorky novorodencov.

Výsledky naznačujú, že na základe našich údajov neexistuje štatisticky významný rozdiel vo variabilite váh novorodencov.

```
> library(car)
> wt_bx <- c(wt_b, wt_x)
> g <- as.factor(c(rep(1, n_b), rep(0, n_x)))
-------------------------------------------------------------
> leveneTest(wt_bx, g, center = mean)
   Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   1  1.8182 0.1787
      265
```

4.6.9 Brown – Forsythov test zhody dvoch rozptylov

Brown – Forsythov test sa niekedy nazýva aj modifikovaný Levenov test. Rozdiel oproti Levenovmu testu spočíva len vo zvolení odlišnej miery polohy pri tvorbe náhodnej premennej *Z*. Budeme rozlišovať medzi dvoma variantmi:

$$Z_{i,j}^{m} = \left| X_{i,j} - \tilde{X}_{j} \right| \qquad (4.78)$$

$$Z_{i,j}^{m} = \left| X_{i,j} - \overline{X}_{j}^{TRIM=t} \right| \qquad (4.79)$$

kde vo vzťahu *(4.78)* sa za charakteristiku polohy volí medián danej skupiny a vo vzťahu *(4.79)* tzv. *trimmed mean*, čo je aritmetický priemer vypočítaný bez *t* % najmenších a najväčších hodnôt. Prvý variant je vhodný najmä v prípade zošikmených súborov a druhý, ak je podozrenie, že sa populácia riadi rozdelením s tzv. tučnými koncami (angl. *heavy tails*). Ide o rozdelenia, kde je výskyt extrémnych hodnôt pravdepodobnejší. Ktorú verziu testu použiť je subjektívna voľba, v každom prípade sa tieto testy považujú za vždy lepšie, ak si nie sme istí predpokladom normality pri *F*-teste zhody dvoch rozptylov.

Testovacia charakteristika a postup je ďalej obdobný ako pri Levenovom teste. Na výpočet v programe R vieme použiť funkciu `leveneTest()`, kde si voľbu metódy podľa *(4.78)* zvolíme možnosťou `center = median` a voľbu metódy *(4.79)* možnosťou `center = mean` a následne `trim = t`. Všimnime si, že Levenov a Brown – Forsythov test umožňujú overovať zhodu rozptylov vo viac ako len v dvoch skupinách. Tieto testy sa často používajú v súvislosti s metódou ANOVA, ktorej jedným z dôležitých predpokladov je

práve predpoklad rovnosti rozptylov a slúži na overovanie zhody v stredných hodnotách viac ako len dvoch skupín.

Uskutočníme si niekoľko simulácií, pri ktorých porovnáme výskyt chyby I. a II. druhu medzi nasledujúcimi testami: *F*-test zhody dvoch rozptylov, Levenov test, Brown – Forsythov test (s mediánom, ktorý budeme v obrázkoch označovať ako B – F test).

Simulácia A

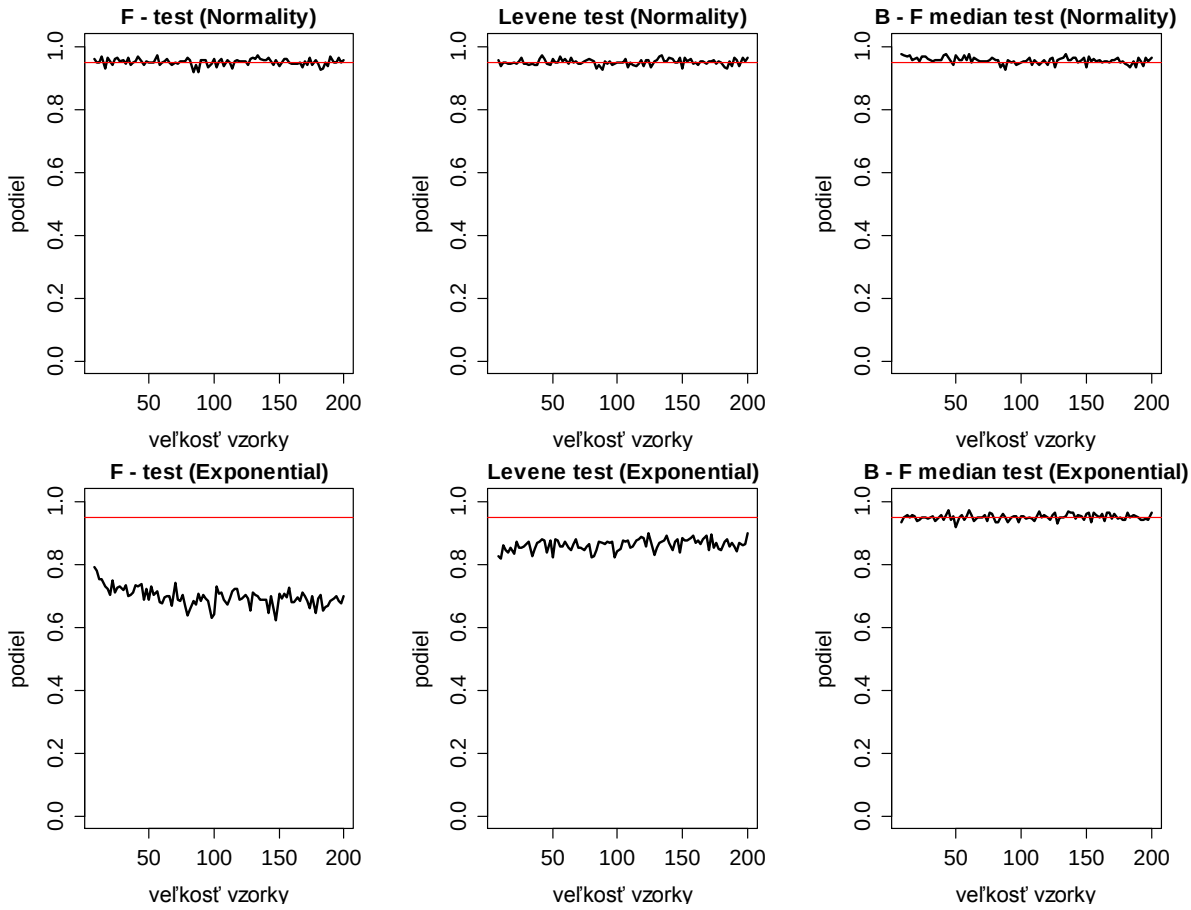
V prvej simulácii nás bude zaujímať výskyt chyby I. druhu, ak budú hodnoty v oboch súboroch generované z normovaného normálneho rozdelenia pravdepodobnosti (t.j. s rovnakým rozptylom). Budeme uvažovať o dvoch vzorkách o rovnakej veľkosti n_x, n_y. Pre každú uvažovanú veľkosť vzoriek n_x, n_y = 8, 10, 12, ..., 200, vygenerujeme súbor hodnôt z uvedeného rozdelenia a otestujeme rozptyly použitím troch testov (hladinu významnosti si zvolíme na úrovni 0.05). Pre každú veľkosť vzorky zopakujeme tento postup (generovanie hodnôt, testovanie) spolu 500 krát. Po tejto iterácii zistíme podiel prípadov, kde sme správne nezamietli nulovú hypotézu o zhode rozptylov. Potom postup zopakujeme pre ďalšiu vzorku. Na konci si výsledky zobrazíme na *x*-*y* grafe tak, že na osi *x*-ovej budeme mať veľkosť vzorky a na osi *y*-ovej podiel úspešného nezamietnutia nulovej hypotézy.

Simulácia B

V druhej simulácii nás bude zaujímať výskyt chyby I. druhu, ak budú hodnoty v oboch súboroch generované z exponenciálneho rozdelenia pravdepodobnosti s parametrom $\lambda = 1$, podobne ako sme už raz uskutočnili pri *F*-teste zhody dvoch rozptylov (tento pokus tak opakujeme). Podmienky sú obdobné ako predtým, budeme uvažovať o dvoch vzorkách o veľkosti n_x, n_y = 8, 10, 12, ..., 200. Vygenerujeme súbor hodnôt z exponenciálneho rozdelenia a otestujeme rozptyly použitím troch testov ($\alpha = 0.05$). Pre každú veľkosť vzorky zopakujeme tento postup (generovanie hodnôt, testovanie) spolu 500 krát. Po tejto iterácii zistíme podiel prípadov, kde sme správne nezamietli nulovú hypotézu o zhode rozptylov. Potom postup zopakujeme pre ďalšiu vzorku. Na konci si výsledky zobrazíme na *x*-*y* grafe podobne ako v predošlom príklade.

```
> ratios <- function(data, value, size) {
+  sum(data>value)/size
+ }
> size <- seq(from = 8, to = 200, by = 2)
> results <- matrix(nrow = length(size), ncol = 3)
> r <- 1
> for (n in size) {
```

```
+   temps <- matrix(nrow = 500, ncol = 3)
+   for (i in 1:500) {
+     g <- as.factor(c(rep(0, n), rep(1, n)))
+     x <- rnorm(n)
+     y <- rnorm(n)
+     xy <- c(x, y)
+     temps[i,1] <-  var.test(x, y, alternative =
  c("two.sided"), conf.level = 0.95)$p.value
+     temps[i,2] <-  leveneTest(xy, g, center = mean)[[3]][1]
+     temps[i,3] <-  leveneTest(xy, g, center = median)[[3]][1]
+   }
+   results[r, ] <- apply(temps, 2, ratios, value = 0.05, size =
  500)
+   r <- r + 1
+ }
-----------------------------------------------------------------
> results_e <- matrix(nrow = length(size), ncol = 3)
> r <- 1
> for (n in size) {
+   temps <- matrix(nrow = 500, ncol = 3)
+   for (i in 1:500) {
+     g <- c(rep(0, n), rep(1, n))
+     x <- rexp(n, rate = 1)
+     y <- rexp(n, rate = 1)
+     xy <- c(x, y)
+     temps[i,1] <-  var.test(x, y, alternative =
  c("two.sided"), conf.level = 0.95)$p.value
+     temps[i,2] <-  leveneTest(xy, g, center = mean)[[3]][1]
+     temps[i,3] <-  leveneTest(xy, g, center = median)[[3]][1]
+   }
+   results_e[r, ] <- apply(temps, 2, ratios, value = 0.05, size
  = 500)
+   r <- r + 1
+ }
-----------------------------------------------------------------
> tests <- c("F - test (Normality)", "Levene test (Normality)",
  "B - F median test (Normality)", "F - test (Exponential)",
  "Levene test (Exponential)", "B - F median test
  (Exponential)")
> rtest <- cbind(results, results_e)
> par(mfrow = c(2, 3), mar = c(4, 5, 2, 2))
> for (i in 1:6) {
+   plot(rtest[,i] ~ size, type = "l", col = "black", lwd = 2,
  ylab = "podiel", xlab = "veľkosť vzorky", ylim = c(0, 1),
  cex.lab = 1, main = tests[i])
+   abline(h = 0.95, col = "red")
+ }
-----------------------------------------------------------------
> par(mfrow = c(2, 3), mar = c(4, 5, 2, 2))
> for (i in 1:6) {
+   plot(rtest[,i] ~ size, type = "l", col = "black", lwd = 2,
  ylab = "podiel", xlab = "veľkosť vzorky", ylim = c(0, 1),
  cex.lab = 1, main = tests[i], cex.main = 1)
+   abline(h = 0.95, col = "red")
+ }
```



Obrázok 4.15: Podiel správne nezamietnutých nulových hypotéz (zhoda variability v dvoch vzorkách) v závislosti od veľkosti vzoriek a rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie, výstup zo softvéru R

Testy zhody dvoch rozptylov by pri hladine významnosti $\alpha = 0.05$ mali správne nezamietnuť nulovú hypotézu o rovnosti rozptylov v 95 % prípadov. Výsledky zo simulácií A a B naznačujú, že táto vlastnosť testu je dodržaná v prípade *F*-testu v situácii, keď sú hodnoty, z ktorých pozorovania pochádzajú, realizáciami normálneho rozdelenia pravdepodobnosti. V prípade, ak hodnoty pochádzali z exponenciálneho rozdelenia, správne nezamietnutie nulovej hypotézy bolo výrazne nižšie a dokonca sa s rastúcou veľkosťou vzorky vlastnosť testu nezlepšila. Levenov test dopadol len o niečo lepšie. V prípade normality boli výsledky blízke teoretickým očakávaniam, ale v prípade exponenciálneho rozdelenia bol podiel správne nezamietnutých nulových hypotéz nižší (aj keď nie tak nízky ako v prípade *F*-testu). S rastúcou veľkosťou vzorky sa výrazne tento podiel nezmenil. Z tohto porovnania vyšiel najlepšie B – F test (s mediánom), kde v oboch prípadoch sa „veľký“ podiel správne nezamietnutých nulových hypotéz pohyboval okolo hodnoty 0.95.

Simulácia C

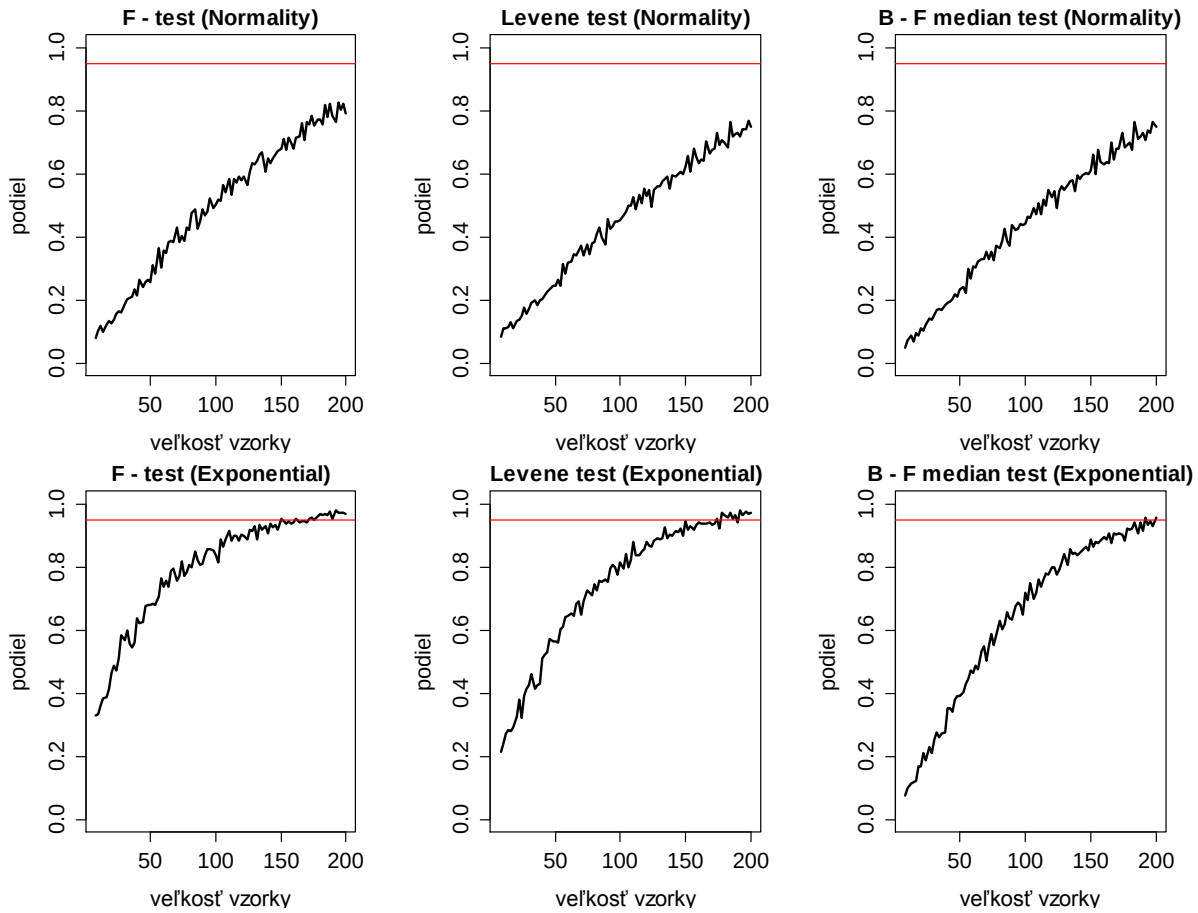
V tretej simulácii nás bude zaujímať výskyt chyby II. druhu v situácii, kde budú hodnoty v oboch súboroch generované z normálneho rozdelenia pravdepodobnosti so strednou hodnotou $\mu = 0$, avšak v jednom prípade bude rozptyl $\sigma^2 = 1$ a v druhom $\sigma^2 = 1.5$. Veľkosť vzorky budeme meniť rovnako ako v simuláciách A a B. Počet iterácií tiež ostáva nezmenený. Zaujímať nás však bude podiel prípadov, kde sme správne zamietli nulovú hypotézu. Výsledky znovu nanesieme do obrázku.

Simulácia D

V poslednej simulácii nás bude zaujímať taktiež výskyt chyby II. druhu, ak budú hodnoty v oboch súboroch generované z exponenciálneho rozdelenia pravdepodobnosti, pričom v prvom súbore bude hodnota parametra $\lambda = 1$ a v druhom $\lambda = 2/3$.

```
> size <- seq(from = 8, to = 200, by = 2)
> results <- matrix(nrow = length(size), ncol = 3)
> r <- 1
> for (n in size) {
+   temps <- matrix(nrow = 500, ncol = 3)
+   for (i in 1:500) {
+     g <- as.factor(c(rep(0, n), rep(1, n)))
+     x <- rnorm(n, mean = 0, sd = 1)
+     y <- rnorm(n, mean = 0, sd = sqrt(1.5))
+     xy <- c(x, y)
+     temps[i,1] <-  var.test(x, y, alternative =
  c("two.sided"), conf.level = 0.95)$p.value
+     temps[i,2] <-  leveneTest(xy, g, center = mean)[[3]][1]
+     temps[i,3] <-  leveneTest(xy, g, center = median)[[3]][1]
+   }
+   results[r, ] <- 1 - apply(temps, 2, ratios, value = 0.05,
  size = 500)
+   r <- r + 1
+ }
-----------------------------------------------------------------
> results_e <- matrix(nrow = length(size), ncol = 3)
> r <- 1
> for (n in size) {
+   temps <- matrix(nrow = 500, ncol = 3)
+   for (i in 1:500) {
+     g <- as.factor(c(rep(0, n), rep(1, n)))
+     x <- rexp(n, rate = 1)
+     y <- rexp(n, rate = 2/3)
+     xy <- c(x, y)
+     temps[i,1] <-  var.test(x, y, alternative =
  c("two.sided"), conf.level = 0.95)$p.value
+     temps[i,2] <-  leveneTest(xy, g, center = mean)[[3]][1]
+     temps[i,3] <-  leveneTest(xy, g, center = median)[[3]][1]
+   }
```

```
+   results_e[r, ] <- 1 - apply(temps, 2, ratios, value = 0.05,
  size = 500)
+   r <- r + 1
+ }
-------------------------------------------------------------------
> tests <- c("F - test (Normality)", "Levene test (Normality)",
  "B - F median test (Normality)", "F - test (Exponential)",
  "Levene test (Exponential)", "B - F median test
  (Exponential)")
> rtest <- cbind(results, results_e)
-------------------------------------------------------------------
> par(mfrow = c(2, 3), mar = c(4, 5, 2, 2))
> for (i in 1:6) {
+   plot(rtest[,i] ~ size, type = "l", col = "black", lwd = 2,
  ylab = "podiel", xlab = "veľkosť vzorky", ylim = c(0, 1),
  cex.lab = 1, main = tests[i], cex.main = 1)
+   abline(h = 0.95, col = "red")
+ }
```



Obrázok 4.16: Podiel správne zamietnutých nulových hypotéz v závislosti od testu zhody dvoch rozptylov, veľkosti vzoriek a rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie, výstup zo softvéru R

Simulácie C a D naznačujú, že sila testu je podobná pri všetkých alternatívach testu, mierne vyššia v prípade *F*-testu. Tu je však potrebné upozorniť, že dôvod prečo v *F*-teste zamietame nulovú hypotézu nemusí byť nutne v tom, že rozptyly sú v skutočnosti od seba odlišné. Dôvod môže byť aj ten, že nie sú splnené predpoklady *F*-testu.

Táto simulácia nám naznačila, že existujú reálne pochybnosti o používaní *F* – testu pre zhodu dvoch rozptylov. V prípade, ak sú podmienky *F*-testu splnené, Levenov test a Brown – Forsythov test dosahujú porovnateľné výsledky. Ak podmienky splnené nie sú, *F*-test dosahuje výrazne vyššiu chybu I. druhu ako alternatívne testy, pričom medzi alternatívnymi testami mal v našom experimente nižšiu chybu I. druhu Brown – Forsythov test. Chyba II. druhu bola v experimentoch pre všetky testy porovnateľná.

4.7 Stručný úvod do metód ANOVA

Metóda ANOVA (z angl. *Analysis of Variance*) je prirodzeným rozšírením testovania zhody dvoch stredných hodnôt pomocou parametrického *t*-testu. Pri týchto testoch sme uvažovali o jednej (závislej) premennej, ktorej pozorovania sme získali pre dve rôzne skupiny. Potom bolo našim cieľom zistiť, či existujú štatisticky významné rozdiely v stredných hodnotách týchto dvoch skupín. Napríklad nás môže zaujímať spokojnosť so životom zvlášť pre vysoko príjmové skupiny obyvateľstva a zvlášť pre ostatné skupiny obyvateľstva. Sú ľudia s vyššími príjmami v priemere viac spokojní so svojím životom? Povedzme, že pomocou dotazníkového šetrenia získame údaje a na tento problém odpovieme použitím *t*-testu zhody dvoch stredných hodnôt.

Zrejme však vieme vytvoriť aj viac ako len dve príjmové skupiny. Ako sa zachovať, ak sa nám zdá byť vhodné rozlíšiť medzi: vysoko príjmovou, stredne príjmovou a nízko príjmovou skupinou? Jednou z možných stratégií sa javí použiť *t*-test na zhodu dvoch stredných hodnôt pre všetky dvojice, t. j. medzi vysoko – stredne, vysoko – nízko, stredne – nízko príjmovými skupinami. Následne identifikujeme, medzi ktorými dvojicami existujú štatisticky významné rozdiely. Dochádza tu však k tzv. inflácii chyby I. druhu. Spolu vykonáme tri testy a v každom z týchto troch testov prijímame možnosť chyby I. druhu na úrovni α. Dá sa ukázať, že za určitých okolností (napr. nezávislosť testov) je pri troch testoch pravdepodobnosť výskytu chyby I. druhu spolu $1 - (1 - \alpha)^u$, kde u je počet testov, v našom prípade 3, t. j. pri $\alpha = 0.05$ je táto chyba 0.142625. Jedným z riešení tohto problému sa javí vhodne znížiť hladinu významnosti α. Ďalšou alternatívou môže byť použitie takých metód,

ktoré vykonajú iba jeden test. V našom prípade, ak chceme vedieť, či priemerná spokojnosť sa mení v závislosti od rôznej úrovne príjmu, môžeme použiť metódu ANOVA.

Príjmovú skupinu budeme nazývať **faktor**. Tento faktor nadobúda tri rôzne stavy, ktoré budeme nazývať **úrovne**. Uvažujme ďalej o ďalšom faktore, povedzme pohlavie. Môže nás zaujímať, aký je vzťah medzi spokojnosťou so životom a faktormi príjmová skupina a pohlavie. Okrem toho nás môže zaujímať aj interakcia faktora príjem s faktorom pohlavie. Všetky tieto základné problémy vieme riešiť použitím metódy ANOVA.

Metóda ANOVA zahŕňa pomerne širokú skupinu metód, z ktorých budeme prezentovať iba tie základné. Začneme členením na:

- jednofaktorovú ANOVA,
- dvojfaktorovú ANOVA,
- viacfaktorovú ANOVA.

Pri jednofaktorovej ANOVA vychádzame z existencie iba jedného faktora. Tento faktor má určitý počet úrovní. Ak by mal len dve úrovne, dostali by sme situáciu ako pri porovnávaní zhody dvoch stredných hodnôt. Dvojfaktorová ANOVA uvažuje s dvoma faktormi s rôznym počtom úrovní (minimálne dve), pričom medzi týmito faktormi môžeme skúmať aj ich interakcie. V prípade, ak uvažujeme o viac ako dvoch faktoroch, môžeme si ich označiť ako viacfaktorovú ANOVA.

Ďalej poznáme modely ANOVA:

- s fixnými faktormi,
- s náhodnými faktormi,
- zmiešané modely.

Ak považujeme faktory za fixné, zaujíma nás priamo vzťah konkrétnych úrovní faktora k závislej premennej. Ak nás zaujíma, či muži alebo ženy sú lepšími predajcami investičných produktov v banke, za závislú premennú si zvolíme počet (prípadne objem) obchodov a za faktor (fixný) pohlavie s dvoma úrovňami: žena, muž.

Ak považujeme faktory za náhodné, zaujíma nás vzťah faktora (úrovne nie sú tak podstatné) k závislej premennej. V príklade vyššie by nás mohla zaujímať úspešnosť predaja v závislosti od okresu, v ktorom predajca pôsobí. Keďže týchto okresov (k roku 2012 bolo na Slovensku 79 okresov) je pomerne veľa, náhodne vyberieme 5 okresov, ktoré budú predstavovať úroveň faktora „miesto“. Všimnime si, že oproti predchádzajúcemu príkladu sme jednotlivé úrovne faktora vyberali náhodne. Pri zmiešaných modeloch máme dva a viac faktorov, z ktorých niektoré sú náhodné a niektoré fixné. Pri náhodných faktoroch chceme

odpovedať na otázku o vzťahu náhodného faktora na závislú premennú, a to aj napriek tomu, že nemáme k dispozícii všetky možné úrovne tohto faktora. V tejto publikácii sa budeme venovať iba fixným faktorom, s používaním náhodných faktorov sme sa stretli zriedkavejšie.

Nakoniec budeme rozlišovať medzi vyváženou a nevyváženou ANOVA. V prvom prípade je počet pozorovaní v každej úrovni faktora totožný, kým v prípade nevyváženej sa počty pozorovaní v jednotlivých úrovniach môžu líšiť. V tejto publikácii sa však budeme venovať iba tzv. vyváženej jedno- a dvojfaktorovej metóde ANOVA s fixnými faktormi.

4.7.1 Jednofaktorová ANOVA s fixnými úrovňami

Uvažujme o pozorovaniach $X_{i,j}$, kde $i = 1, 2, ..., n_j$ predstavujú jednotlivé nezávislé pozorovania, pričom $n = n_1 + n_2 + ... + n_k$ a $j = 1, 2, ..., k$ sú jednotlivé úrovne faktora. Pričom pri vyváženej ANOVA platí $n_1 = n_2 = ... = n_k$. Tieto pozorovania predstavujú náhodné a nezávislé vzorky (vzorky podľa indexu j). Uvažujme o hodinovej mzde respondentov (použijeme databázu `SLID` z programového balíka `car`). Zaujíma nás, či existujú štatisticky významné rozdiely v hodinovej mzde v závislosti od veku. Premennú vek môžeme považovať za skôr intervalovú premennú ako za premennú s niekoľkými úrovňami. Preto si v tomto príklade rozdelíme vek na tri úrovne: do 30 rokov (vrátane), od 30 rokov do 60 rokov, od 60 (vrátane) viac rokov. Hodinová mzda (v kanadských dolároch) pre jednotlivých respondentov je premenná $X_{i,j}$, kde i predstavuje index zodpovedajúci respondentom a j index zodpovedajúci príslušnej vekovej skupine (úrovni faktora). Napríklad $X_{1,1}$ bude zodpovedať hodinovej mzde prvého respondenta v prvej vekovej skupine (do 30 rokov), $X_{1,2}$ bude zodpovedať hodinovej mzde prvého respondenta v druhej vekovej skupine (od 30 do 60 rokov) a $X_{1,3}$ bude zodpovedať hodinovej mzde tretieho respondenta v tretej vekovej skupine (viac ako 60 rokov).

Jednofaktorový model ANOVA s fixným faktorom si môžeme zapísať nasledovne (bližšie pozri napr. Kirk, 2008; Cohen – Cohen, 2008):

$$x_{i,j} = \mu + \alpha_j + u_{i,j} \qquad (4.80)$$

Model zapísaný vo vzťahu *(4.80)* je populačný model. Strednú hodnotu hodinovej mzdy v populácii si označíme ako μ a odhadneme ju cez nasledujúci vzťah:

$$\mu = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{i,j} = \overline{\overline{X}} \qquad (4.81)$$

Efekt úrovne j označujeme ako α_j, pričom platí:

$$\sum_{j=1}^{k} \alpha_j = 0 \qquad (4.82)$$

Ďalej platí, že $\mu + \alpha_j$ považujeme za strednú hodnotu hodinovej mzdy v skupine (úrovni) j. To znamená, že k celkovej strednej hodnote pripočítame efekt skupiny (úrovne) j. Ak platí, že vo veku 30 až 60 rokov je hodinová mzda vyššia ako v ostatných skupinách, potom bude efekt a_j zrejme > 0. Hodnotu $\mu + \alpha_j$ si môžeme odhadnúť nasledovne:

$$\mu + \alpha_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{i,j} = \overline{X}_j \qquad (4.83)$$

Vo vzťahu *(4.80)* predstavuje $u_{i,j}$ chybový člen, tzv. náhodný efekt. Jeho význam si ilustrujeme nasledovne. Náhodne sme vybrali jedného respondenta, ktorého hodinová mzda bola 4.29 a mal 42 rokov. Ak by sme vychádzali z odhadu modelu *(4.80)* (použijeme odhady zo vzťahov *(4.81) – (4.83)*), potom by malo platiť: $x_{i,j} = \mu + \alpha_j + u_{i,j}$. Vychádzajme teraz z odhadovaného modelu, kde celkový priemer je 15.55308 a priemerná hodinová mzda v druhej skupine (vek od 30 do 60 rokov) je 17.74210, t. j. 4.29 = 15.55308 + (17.74210 – 15.55308) + $e_{i,j}$. Aby platila rovnosť ľavej a pravej strany, je potrebné pripočítať $e_{i,j}$ = 13.4521, kde $e_{i,j}$ je odhad chybového člena $u_{i,j}$. Chybový člen by tak mal predstavovať efekty vlastné danému respondentovi. Môže ísť o úroveň vzdelania, o oblasť pracovného zaradenia, zdravotný stav, rodinný stav a mnoho ďalších faktorov, ktoré sa v jednofaktorovom modeli nezohľadňujú.

Model *(4.80)* tak hovorí, že hodinovú mzdu respondenta vieme predpovedať na základe strednej hodnoty hodinovej mzdy, plus strednej hodnoty skupiny (úrovne). Model *(4.80)* predpokladá, že pozorovania medzi skupinami sú nezávislé, že rozptyl hodnôt medzi skupinami je konštantný, a že chybový člen je realizáciou z normálneho rozdelenia so strednou hodnotou 0 a rozptylom σ^2. Vzhľadom na to, že jednotlivé hodnoty sa majú dať vysvetliť pomocou stredných hodnôt (celkovej a aj tej v rámci skupiny) a chybového člena z normálneho rozdelenia, potom aj tieto závislé premenné majú pochádzať z normálneho rozdelenia pravdepodobnosti. Presnejšie, v každej skupine j majú byť pozorovania realizáciami z normálneho rozdelenia.

Predpoklad náhodného výberu je tu pomerne dôležitý, keďže zabezpečuje, aby sa špecifické vlastnosti štatistických jednotiek (napr. respondentov), ktoré by mohli jednostranne ovplyvniť analýzu, rovnomerne rozložili medzi všetky úrovne daného faktora. Týmto spôsobom nedochádza k skresleniu štatistík. V prípade, ak sa táto podmienka nedá zabezpečiť (v spoločenských vedách veľmi častý problém), potom aj interpretácia výsledkov musí byť

výrazne opatrnejšia. Nedá sa totiž vylúčiť, že nameraný výsledok je dôsledkom určitých nepoznaných efektov a nie v dôsledku pôsobenia faktora, ktorý skúmame.

Predpoklad normality je náročné dodržať a v skutočnosti aj overiť. Platí, že ak sú rozdelenia symetrické a počty pozorovaní v jednotlivých skupinách rovnaké, potom porušenie tohto predpokladu nemá zásadný vplyv na chybu I. druhu (podobne ako v prípade *t*-testu). Predpoklad normality sa však týka $u_{i,j}$, nie $e_{i,j}$. Keďže však $e_{i,j}$ predstavuje odhad, a teda priamo súvisí s $u_{i,j}$, tak sa overuje normalita $X_{i,j}$, ktorých normalita vyplýva z $e_{i,j}$. Možno dôležitejšie ako samotný test normality je poznať charakter nameraných veličín. Na reálnych údajoch je predpoklad normality zrejme nereálny. Hodnoty z normálneho rozdelenia pochádzajú z intervalu $\pm\infty$, čo na reálnych údajoch je skôr nemožné. Taktiež je potrebné, aby išlo o spojitú premennú. Preto pred samotným testovaním normality je vhodné skúmať, či je normálne rozdelenie vhodná aproximácia k skutočnému rozdeleniu.

Ďalší predpoklad sa týka rozptylov medzi jednotlivými skupinami, ktoré majú byť rovnaké, t. j. $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$. Tento predpoklad považujeme za najdôležitejší a neraz aj za najväčšiu prekážku pri používaní metódy ANOVA. Pokiaľ nie je rozumné predpokladať jeho dodržanie, je vhodnejšie voliť iné metódy (neparametrické, napr. Kruskal – Wallisov test), ktorých výsledky závisia menej od zmien medzi rozptylmi vzoriek. Vo všeobecnosti taktiež platí, že je vhodné mať rovnako veľké počty pozorovaní v jednotlivých skupinách.

Pomocou odhadu modelu *(4.80)* budeme overovať nasledujúcu hypotézu:

H_0: $\alpha_i = 0$, pre všetky i	Hypotézu H_0 zamietame, ak
H_1: Existuje aspoň jedno i také, aby $\alpha_i \neq 0$	$F_{ANOVA} > F_{(k-1),(n-k),(1-\alpha)}$

Štatistiku F_{ANOVA} získame nasledujúcim postupom. Rozdiel medzi nameranou hodnotou a celkovým priemerom vieme rozložiť na rozdiel v rámci skupiny a medzi-skupinový rozdiel nasledovne:

$$X_{i,j} - \overline{\overline{X}} = X_{i,j} - \overline{X}_j + \overline{X}_j - \overline{\overline{X}} \qquad (4.84)$$

Celkový súčet štvorcov odchýlok si zapíšeme nasledovne (TSS, z angl. *Total Sum of Squares*):

$$TSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{i,j} - \overline{\overline{X}} \right)^2 \qquad (4.85)$$

Súčet štvorcov odchýlok v skupinách je (WSS, z angl. *Within Sum of Squares*):

$$WSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{i,j} - \overline{X}_j \right)^2 \qquad (4.86)$$

Súčet štvorcov odchýlok medzi skupinami je (BSS, z angl. *Between Sum of Squares*):

$$BSS = \sum_{j=}^{k} \sum_{i=1}^{n_j} \left(\overline{X}_j - \overline{\overline{X}} \right)^2 \qquad (4.87)$$

pričom platí:

$$TSS = WSS + BSS \qquad (4.88)$$

Vzťah *(4.87)* je možné zapísať aj nasledovne:

$$BSS = \sum_{j=1}^{k} n_j \left(\overline{X}_j - \overline{\overline{X}} \right)^2 \qquad (4.89)$$

Vzťah *(4.89)* si môžeme vysvetliť aj tak, že každú jednu hodnotu danej skupiny nahradíme priemernou hodnotou danej skupiny. Potom odpočítame a umocníme tieto „upravené hodnoty“ od celkového priemeru a sčítame ich. Tento postup vykonáme pre každú jednu skupinu a súčty sčítame.

Variabilitu vo vnútri skupín (WSS) môžeme chápať ako náhodnú variabilitu, vlastnú danej skupine (úrovni faktora). Variabilitu medzi skupinami navzájom (BSS) môžeme chápať ako efekt skupín. Budeme pokračovať v našom predchádzajúcom príklade. Ak je variabilita medzi skupinami veľká, potom zjavne existujú veľké rozdiely medzi priemernou hodinovou mzdou rôznych vekových skupín. Na druhej strane, ak je veľká variabilita priemernej hodinovej mzdy v rámci jednej vekovej skupiny, potom existujú veľké rozdiely medzi priemernou hodinovou mzdou už v rámci tejto skupiny. Samozrejme je potrebné definovať, čo to znamená veľká variabilita. K tomu slúži podiel variability medzi skupinami k variabilite vo vnútri skupín. Čím je variabilita medzi skupinami relatívne väčšia k variabilite vo vnútri skupín, o to väčšie sú rozdiely medzi skupinami, a o to je väčšia tendencia tvrdiť, že daný faktor má vplyv (nie nutne v kauzálnom slova zmysle) na závislú premennú. Po zohľadnení počtu skupín a počtu pozorovaní má testovacia štatistika nasledujúci tvar:

$$F = \frac{\dfrac{BSS}{(k-1)}}{\dfrac{WSS}{(n-k)}} \qquad (4.90)$$

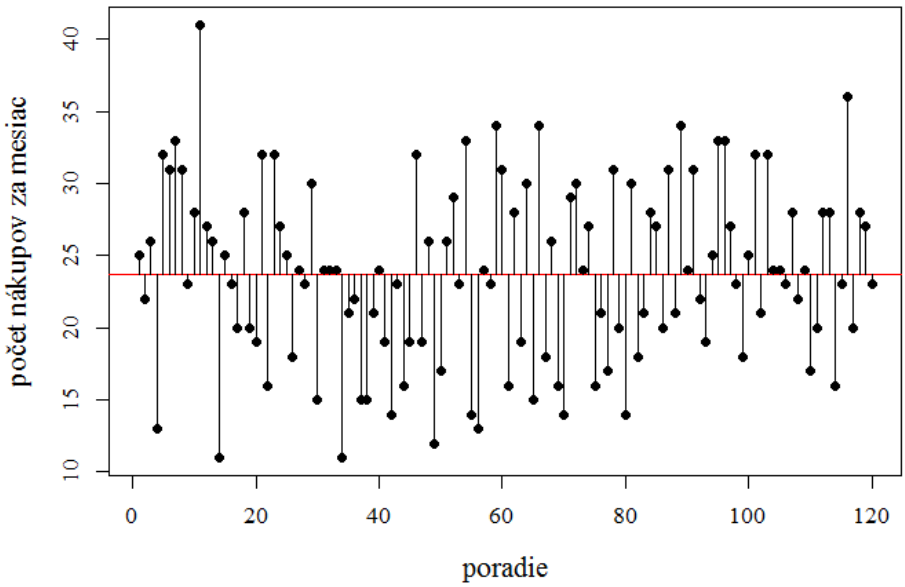
Vychádzajme z nasledujúceho zadania. V databázovom objekte sa nachádzajú tri premenné. Prvá premenná `nakup`, zodpovedá počtu nákupov v potravinách, ktoré náhodne

vybraný respondent za mesiac uskutočnil. Druhá premenná (`pohlavie`) zaznamenáva pohlavie respondenta, kde 1 zodpovedá ženám a 0 zodpovedá mužom. Posledná premenná (`vek`) zaznamenáva vek, kde 1 zodpovedá kategórii do 30 rokov (vrátane) a 0 zodpovedá kategórii veku nad 30 rokov. V jednofaktorovej ANOVA nás bude zaujímať, či existuje rozdiel medzi počtom nákupov žien a mužov. Túto úlohu by sme mohli riešiť aj pomocou Studentovho *t*-testu zhody dvoch nezávislých súborov, prípadne pomocou neparametrických testov. Necháme na čitateľovi, aby výsledky porovnal.

```
> nakup <- c(25, 22, 26, 13, 32, 31, 33, 31, 23, 28, 41, 27, 26,
  11, 25, 23, 20, 28, 20, 19, 32, 16, 32, 27, 25, 18, 24, 23,
  30, 15, 24, 24, 24, 11, 21, 22, 15, 15, 21, 24, 19, 14, 23,
  16, 19, 32, 19, 26, 12, 17, 26, 29, 23, 33, 14, 13, 24, 23,
  34, 31, 16, 28, 19, 30, 15, 34, 18, 26, 16, 14, 29, 30, 24,
  27, 16, 21, 17, 31, 20, 14, 30, 18, 21, 28, 27, 20, 31, 21,
  34, 24, 31, 22, 19, 25, 33, 33, 27, 23, 18, 25, 32, 21, 32,
  24, 24, 23, 28, 22, 24, 17, 20, 28, 28, 16, 23, 36, 20, 28,
  27, 23)
> pohlavie <- c(1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0,
  1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0,
  0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0,
  1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0,
  0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1,
  1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1)
> vek <- c(0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1,
  1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0,
  0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0,
  1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0,
  0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0,
  1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0)
> obchod <- data.frame(nakup, pohlavie, vek)
```

Naším prvým krokom je vizualizácia údajov. Použijeme pritom jednoduchý *x*-*y* graf, kde na *x*-ovej osi zaznačíme poradie, v ktorom sa údaje získali a na *y*-ovej osi počet nákupov. Na nasledujúcom obrázku me si vyznačili aj celkový priemerný počet nákupov (červená horizontálna čiara). Zároveň sme vyznačili čiary, ktoré spájajú jednotlivé body (počty nákupov) s priemerným počtom nákupov. Týmto spôsobom si vieme urobiť určitú vizuálnu predstavu o variabilite nameraných hodnôt.

```
> plot(1:length(nakup), nakup, xlab = "poradie", ylim =
  c(min(nakup), max(nakup)), ylab = "počet nákupov za mesiac",
  pch = 19, cex.lab = 1.3, cex.axis = 1.1, family = "serif")
> abline(h = mean(nakup), col = "red")
> for (i in 1:length(nakup)) lines(c(i, i), c(mean(nakup),
  nakup[i]))
```



Obrázok 4.17: *x-y* graf počtu nákupov v závislosti od poradia zberu údajov

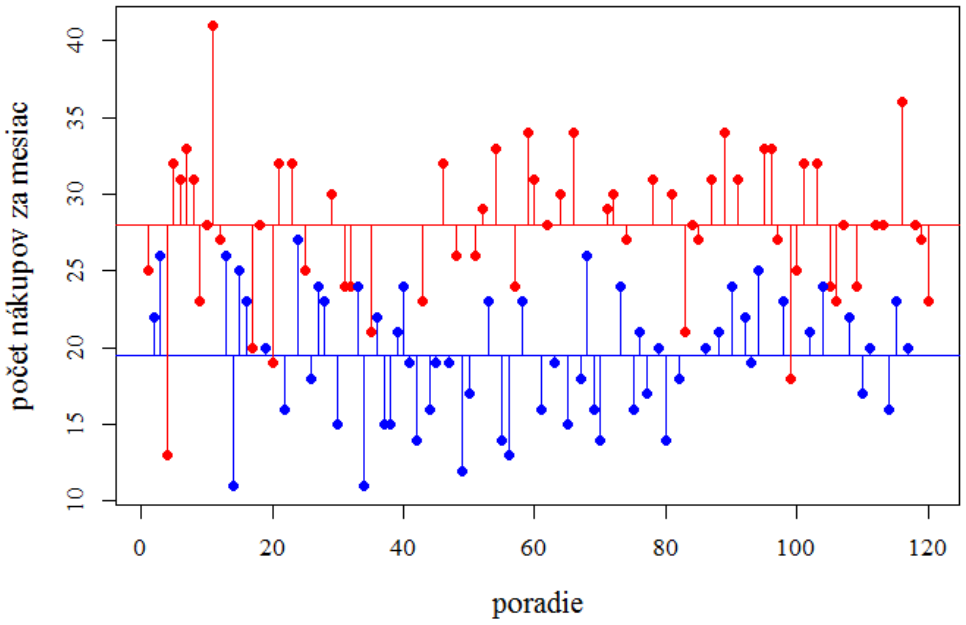
Zdroj: vlastné spracovanie, výstup zo softvéru R

Upravme si ďalej obrázok nasledovne. Označme červenou farbou tie počty nákupov, ktoré zodpovedajú ženám a modrou označíme nákupy zodpovedajúce mužom. Ďalej si vyznačíme vzdialenosti bodov (počtu nákupov) od nie celkového priemeru, ale od skupinového priemeru. Napríklad počet nákupov vybranej ženy budeme porovnávať s priemerným počtom nákupov žien. Túto horizontálnu čiaru si zakreslíme do obrázku a označíme červenou farbou. Priemerný počet nákupov pre mužov si označíme modrou farbou.

```
> x_women <- c(); x_men <- c()
> for (i in 1:length(pohlavie)) if(pohlavie[i] == 1) x_women =
  c(x_women, i)
> for (i in 1:length(pohlavie)) if(pohlavie[i] == 0) x_men =
  c(x_men, i)
> plot(1:length(nakup), type = "n", ylim = c(min(nakup),
  max(nakup)), ylab = "počet nákupov za mesiac", xlab =
  "poradie", cex.lab = 1.3, cex.axis = 11, family = "serif")
> abline(h = mean(nakup[pohlavie == 1]), col = "red")
> abline(h = mean(nakup[pohlavie == 0]), col = "blue")
> points(x_women, nakup[pohlavie == 1], pch = 19, col = "red")
> points(x_men, nakup[pohlavie == 0], pch = 19, col = "blue")
> for (i in x_women) lines(c(i, i), c(mean(nakup[pohlavie ==
  1]), nakup[i]), col = "red")
> for (i in x_men) lines(c(i, i), c(mean(nakup[pohlavie == 0]),
  nakup[i]), col = "blue")
```

Z nasledujúceho obrázku (pozri Obrázok 4.18) sú viditeľné dve dôležité skutočnosti. Priemerné hodnoty počtu nákupov medzi mužmi a ženami sú odlišné. Ďalej je zrejmé, že

variabilita počtom nákupov medzi ženami (červené bodky k červenej horizontále) je menšia, ako keby sa počty nákupov porovnávali k celkovému priemeru. Podobne aj v prípade mužov je variabilita počtu nákupov medzi mužmi nižšia (modré bodky k modrej horizontále), ako keby sa počty nákupov porovnávali k celkovému priemeru. Ak je rozdiel medzi priemermi dostatočne veľký, budeme mať tendenciu vyvodiť z toho záver, že existuje štatisticky významný rozdiel aj v stredných hodnotách počtu nákupov medzi mužmi a ženami. Z tohto obrázku (Obrázok 4.18) je taktiež možné sledovať, že modré a červené body sa výrazne neprekrývajú. Ako keby pochádzali z iných populácií. Ak by sa výraznejšie prekrývali (a to aj napriek odlišným priemerom mužov a žien), naznačovalo by to, že rozdielny priemer počtu nákupov vznikol len v dôsledku náhodného výberu a nie v dôsledku systematických rozdielov medzi nákupným správaním sa mužov a žien. Porovnanie variability medzi skupinami (červená a modrá, Obrázok 4.18) k celkovej variabilite (predošlý Obrázok 4.17 – čierna) predstavuje kľúčový princíp pri výpočtoch metódy ANOVA.



Obrázok 4.18: *x-y* graf počtu nákupov v závislosti od poradia zberu údajov zvlášť pre mužov a zvlášť pre ženy

Zdroj: vlastné spracovanie, výstup zo softvéru R

V našom príklade najprv vykonáme prepočet manuálne pomocou vzťahov, ktoré sme si definovali vyššie. Z výsledkov (ku ktorým dospejeme) pomerne jednoznačne vyplýva, že hypotézu H_0 zamietame. Medzi skupinami existuje štatisticky významný rozdiel. Zdá sa teda, že nie je jedno, či ide o muža alebo o ženu. Ich nákupné správanie sa (z hľadiska počtu nákupov za mesiac) je odlišné.

```
> BSS <- sum(60*(tapply(nakup, pohlavie, mean) -
  mean(nakup))^2); BSS
[1] 2159
> WSS <- sum((nakup[pohlavie == 1] - mean(nakup[pohlavie ==
  1]))^2) + sum((nakup[pohlavie == 0] - mean(nakup[pohlavie ==
  0]))^2); WSS
[1] 2415.8
> TSS <- sum((nakup - mean(nakup))^2); TSS
[1] 4574.8
> #skúška správnosti
> BSS + WSS
[1] 4574.8
> #F(ANOVA)
> F <- (BSS/(2 - 1)) / (WSS/(length(nakup) - 2)); F
[1] 105.46
> #kritická hodnota
> qf(0.95, 1, 118)
[1] 3.9215
```

Štandardný spôsob publikovania výsledkov pri použití metódy ANOVA je tzv. ANOVA tabuľka. V programe R výpočet ANOVA a zároveň príslušnú ANOVA tabuľku získame pomocou funkcie `aov()` a všeobecnej funkcie `summary()`.

```
> summary(aov(nakup ~ pohlavie))
             Df Sum Sq Mean Sq F value Pr(>F)
pohlavie      1   2159  2159.0   105.5 <2e-16 ***
Residuals   118   2416    20.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

V tejto tabuľke máme uvedené v stĺpcoch nasledovné informácie: `Df` – predstavuje počet stupňov voľnosti, `Sum Sq` – predstavuje variabilitu medzi skupinami a variabilitu v skupinách, `Mean Sq` – priemernú variabilitu, `F value` – hodnotu testovacej štatistiky a `Pr(>F)` – pravdepodobnosť, že náhodná premenná z *F* rozdelenia s príslušnými stupňami voľnosti, bude väčšia ako hodnota testovacej štatistiky (teda tzv. *p*-hodnota). Na záver úlohy ešte vykonáme dva testy, ktorými by sme radi zistili, či nami namerané pozorovania (ne)odporujú predpokladom metódy ANOVA. Použijeme Brown – Forsythov test na zhodu variability dvoch súborov (`leveneTest()` z knižnice `car`) a Shapiro – Wilkov test na overenie normality zvlášť pre počet nákupov u žien a zvlášť pre počet nákupov u mužov (`shapiro.test()` z knižnice `nortest`). Empirické údaje nenaznačujú porušenie predpokladov.

```
> library(car)
> leveneTest(nakup, group = as.factor(pohlavie), center =
  median)
  Levene's Test for Homogeneity of Variance (center = median)
```

```
       Df F value Pr(>F)
group   1   0.259 0.6118
      118
-----------------------------------------------------------------
> library(nortest)
> shapiro.test(nakup[pohlavie == 1])

                 Shapiro-Wilk normality test

data:  nakup[pohlavie == 1]
W = 0.9754, p-value = 0.2667
-----------------------------------------------------------------
> shapiro.test(nakup[pohlavie == 0])

                 Shapiro-Wilk normality test

data:  nakup[pohlavie == 0]
W = 0.9672, p-value = 0.1059
```

4.7.2 Dvojfaktorová ANOVA s fixnými úrovňami

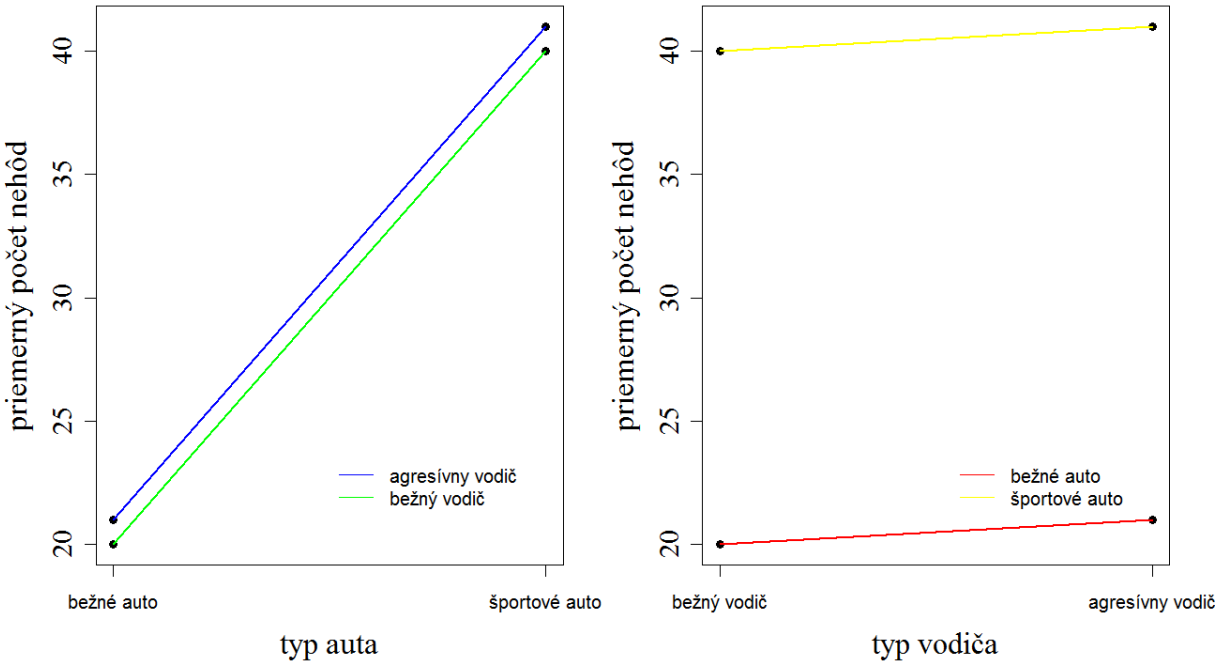
Majme *iid* vzorku $X_{i,j,k}$, kde index j predstavuje úroveň faktora α a platí $j = 1, 2, ..., J$, index k predstavuje úroveň faktora β a platí $k = 1, 2, ..., K$. Pri vyváženom modeli ANOVA ďalej platí, že ak index i zodpovedá príslušnej štatistickej jednotke v j-tej úrovni faktorov α a β, potom $i = 1, 2, ..., n_{j,k}$, pričom $n_{j,k} = c$ (konštanta, ide o vyvážený model ANOVA) pre všetky j, k, a teda celkový počet pozorovaní je $n = JKc$. Na rozdiel od predchádzajúceho prípadu máme dva fixné faktory, prvý s J úrovňami a druhý s K úrovňami. Oproti predchádzajúcemu príkladu ďalej okrem efektu jednotlivých faktorov na závislú premennú $x_{i,j,k}$ môžeme skúmať aj efekt interakcie medzi faktorom α a β.

Analýzu s interakciami neskôr prevedieme na príklade z predošlej kapitoly. Teraz si ilustrujeme niekoľko príkladov, ktoré môžu nastať pri dvoch fixných faktoroch s dvoma úrovňami. Majme počet smrteľných dopravných nehôd na rôznych cestách kraja ako závislú premennú. Skúmame efekt dvoch faktorov. Prvý faktor α predstavuje typ osobného auta, pričom môže ísť buď o auto športové (prvá úroveň faktora α) alebo o tzv. bežné auto (druhá úroveň faktora α). Druhý faktor β predstavuje typ vodiča, pričom môže ísť buď o agresívneho vodiča (prvá úroveň faktora β) alebo o bežného vodiča (druhá úroveň faktora β). Spolu tak vznikajú štyri rôzne situácie: športové auto a agresívny vodič, športové auto a bežný vodič, bežné auto a agresívny vodič, bežné auto a bežný vodič.

Možné efekty si znázorníme na tzv. interakčných grafoch. Najprv ilustrujeme situáciu, kde zjavne existuje efekt typu auta, avšak neexistuje efekt vodiča. Pre tieto účely sme si

vygenerovali priemerné hodnoty nehodovosti, ktoré zodpovedajú rôznym kombináciám áut a vodičov:

V ľavej časti nasledujúceho obrázku (pozri Obrázok 4.19) je zobrazený priemerný počet nehôd, ktorý je výrazne vyšší v prípade športových áut ako v prípade bežných áut. Tento nárast je pritom rovnako výrazný bez ohľadu na to, či ide o vodičov, ktorí boli agresívni alebo tzv. bežní. Tejto situácii hovoríme efekt jedného faktora: efekt typu osobného auta. Efekt vodiča tu nie je výrazný (ak vôbec je). Graf v pravej časti obrázku tento „efekt" jedného faktora ďalej zvýrazňuje.



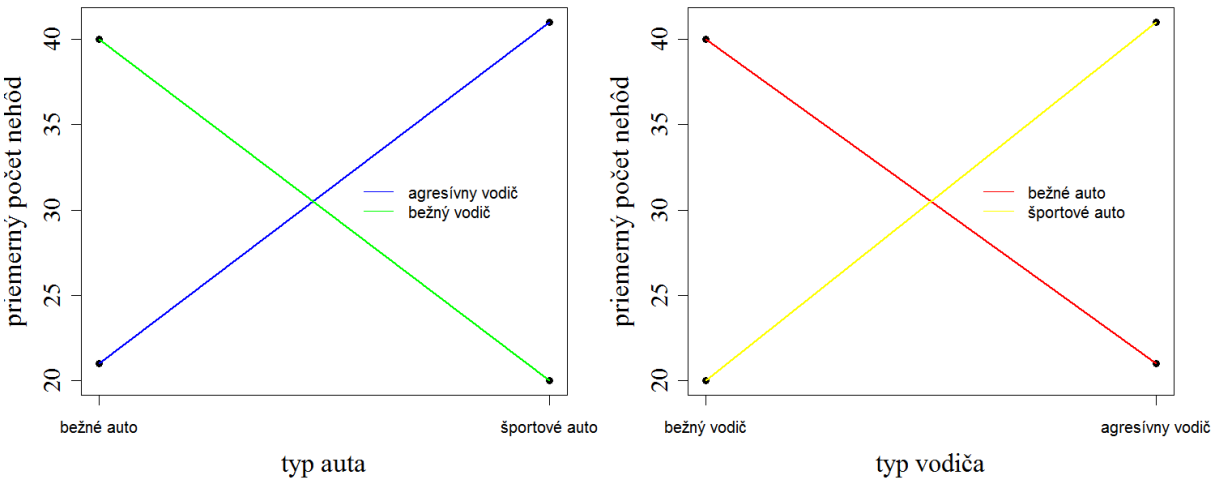
Obrázok 4.19: Interakčný graf – efekt typu auta

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(1, 2))
> plot(c(0, 0, 1, 1), c(20, 21, 40, 41), type = "p", pch = 19,
  xaxt = "n", family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "priemerný počet nehôd", xlab = "typ auta")
> axis(side = 1, at = c(0, 1), labels = c("bežné auto",
  "športové auto"))
> lines(c(0, 1), c(21, 41), type = "l", col = "blue", lwd = 2)
> lines(c(0, 1), c(20, 40), type = "l", col = "green", lwd = 2)
> legend("bottomright", legend = c("agresívny vodič", "bežný
  vodič"), col = c("blue", "green"), lty = 1, inset = 0.08, bty
  = "n")
> plot(c(0, 0, 1, 1), c(20, 40, 21, 41), type = "p", pch = 19,
  xaxt = "n", family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "priemerný počet nehôd", xlab = "typ vodiča")
> axis(side = 1, at = c(0, 1), labels = c("bežný vodič",
  "agresívny vodič"))
> lines(c(0, 1), c(20, 21), type = "l", col = "red", lwd = 2)
> lines(c(0, 1), c(40, 41), type = "l", col = "yellow", lwd = 2)
```

```
> legend("bottomright", legend = c("bežné auto", "športové
  auto"), col = c("red", "yellow"), lty = 1, inset = 0.08, bty =
  "n")
```

Ako by vyzeral tento obrázok v prípade, ak by existovala interakcia faktora „typ auta“ a „typ vodiča“? Nasledujúci obrázok takúto možnú situáciu znázorňuje (pozri Obrázok 4.20). Nie je jedno, aký vodič sa nachádza v akom type auta. Agresívny vodič v športovom aute má väčší počet dopravných nehôd, ako keď ten istý typ vodiča má bežné auto. Na druhej strane (v tomto príklade je to zrejme paradox), bežný vodič v bežnom aute má podobný počet nehôd. Oproti týmto výsledkom sú v silnom kontraste situácie, kde bežný vodič má športové auto, keďže priemerný počet nehôd je výrazne nižší, ako keď má bežné auto.

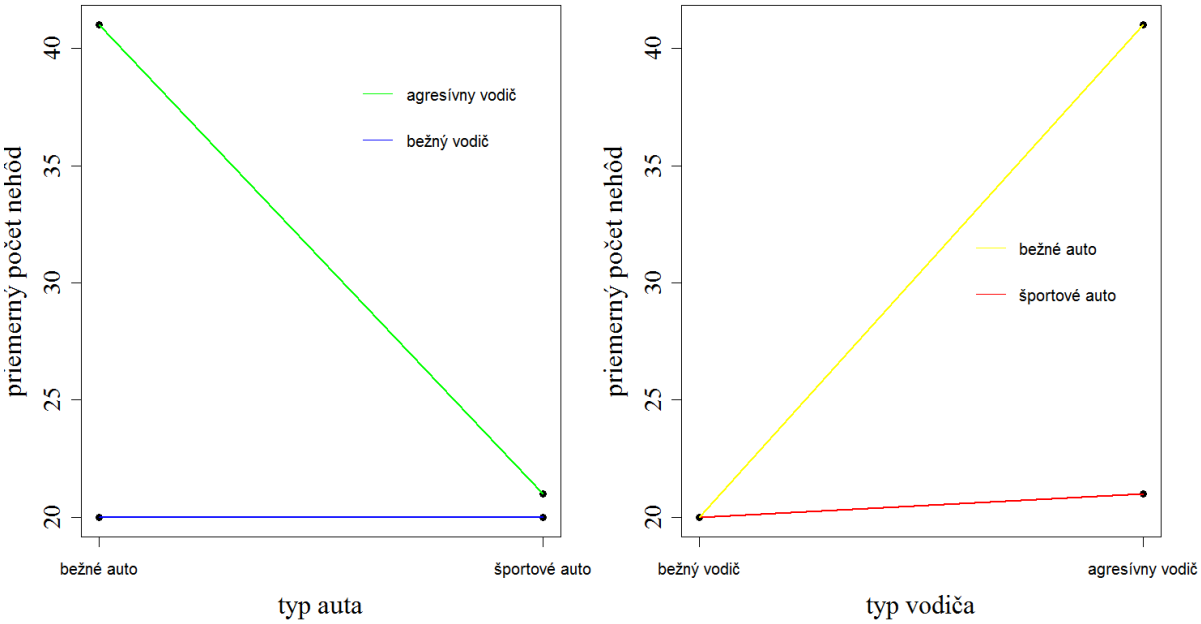


Obrázok 4.20: Interakčný graf – interakcia dvoch faktorov

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(1, 2))
> plot(c(0, 0, 1, 1), c(21, 40, 41, 20), type = "p", pch = 19,
  xaxt = "n", family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "priemerný počet nehôd", xlab = "typ auta")
> axis(side = 1, at = c(0, 1), labels = c("bežné auto",
  "športové auto"))
> lines(c(0, 1), c(21, 41), type = "l", col = "blue", lwd = 2)
> lines(c(0, 1), c(40, 20), type = "l", col = "green", lwd = 2)
> legend("right", legend = c("agresívny vodič", "bežný vodič"),
  col = c("blue", "green"), lty = 1, inset = 0.08, bty = "n")
-------------------------------------------------------------
> plot(c(0, 0, 1, 1), c(20, 40, 41, 21), type = "p", pch = 19,
  xaxt = "n", family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "priemerný počet nehôd", xlab = "typ vodiča")
> axis(side = 1, at = c(0, 1), labels = c("bežný vodič",
  "agresívny vodič"))
> lines(c(0, 1), c(40, 21), type = "l", col = "red", lwd = 2)
> lines(c(0, 1), c(20, 41), type = "l", col = "yellow", lwd = 2)
> legend("right", legend = c("bežné auto", "športové auto"), col
  = c("red", "yellow"), lty = 1, inset = 0.08, bty = "n")
```

Uvažujme ďalej o nasledujúcej hypotetickej situácii, ak je vodič agresívny a zároveň má športové auto, potom je počet nehôd nižší. Ak je vodič agresívny a zároveň má bežné auto, potom je počet nehôd vyšší. Ak ide o bežného vodiča, potom je počet nehôd nízky bez ohľadu na typ auta. Túto situáciu znázorňuje nasledujúci obrázok (pozri Obrázok 4.21).



Obrázok 4.21: Interakčný graf: interakcia úrovní faktorov

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(1, 2))
> plot(c(0, 0, 1, 1), c(41, 20, 21, 20), type = "p", pch = 19,
  xaxt = "n", family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "priemerný počet nehôd", xlab = "typ auta")
> axis(side = 1, at = c(0, 1), labels = c("bežné auto",
  "športové auto"))
> lines(c(0, 1), c(20, 20), type = "l", col = "blue", lwd = 2)
> lines(c(0, 1), c(41, 21), type = "l", col = "green", lwd = 2)
> legend("topright", legend = c("agresívny vodič", "bežný
  vodič"), col = c("green", "blue"), lty = 1, inset = 0.08, bty
  = "n")
-----------------------------------------------------------------
> plot(c(0, 0, 1, 1), c(20, 20, 21, 41), type = "p", pch = 19,
  xaxt = "n", family = "serif", cex.lab = 1.7, cex.axis = 1.5,
  ylab = "priemerný počet nehôd", xlab = "typ vodiča")
> axis(side = 1, at = c(0, 1), labels = c("bežný vodič",
  "agresívny vodič"))
> lines(c(0, 1), c(20, 21), type = "l", col = "red", lwd = 2)
> lines(c(0, 1), c(20, 41), type = "l", col = "yellow", lwd = 2)
> legend("right", legend = c("bežné auto", "športové auto"), col
  = c("yellow", "red"), lty = 1, inset = 0.08, bty = "n")
```

Dvojfaktorový model, ktorý je predmetom nášho záujmu v tejto časti, si môžeme formálne zapísať podobne ako pri jednofaktorovom modeli v predošlej časti (Montgomery – Runger, 2011):

$$x_{i,j,k} = \mu + \alpha_j + \beta_k + \alpha_j\beta_k + u_{i,j,k} \qquad (4.91)$$

Zároveň predpokladáme:

$$\sum_{j=1}^{J} \alpha_j = 0 \qquad (4.92)$$

$$\sum_{k=1}^{K} \beta_k = 0 \qquad (4.93)$$

$$\sum_{j=1}^{J} \alpha_j\beta_k = 0 \qquad (4.94)$$

$$\sum_{k=1}^{K} \alpha_j\beta_k = 0 \qquad (4.95)$$

Urobíme podobný rozklad variability ako pri jednofaktorovom modeli, avšak najprv si zadefinujeme niektoré premenné: $X_{\bullet j\bullet}$ je celkový súčet pozorovaní j-tej úrovne faktora α, $X_{\bullet\bullet k}$ je celkový súčet pozorovaní k-tej úrovne faktora β, $X_{\bullet jk}$ je celkový súčet pozorovaní j-tej a zároveň k-tej úrovne oboch faktorov (α a β) a $X_{\bullet\bullet\bullet}$ je celkový súčet všetkých pozorovaní. Zodpovedajúce priemerné hodnoty si označíme ako: $\overline{X}_{\bullet j\bullet}$, $\overline{X}_{\bullet\bullet k}$, $\overline{X}_{\bullet jk}$ a $\overline{X}_{\bullet\bullet\bullet}$. Overujeme nasledujúce tri hypotézy:

H_0: $\alpha_j = 0$, pre všetky j	Hypotézu H_0 zamietame, ak $F_{ANOVA\text{-}\alpha} > F_{(J-1),\,JK(c-1),\,(1-\alpha)}$
H_1: Existuje aspoň jedno j také, aby $\alpha_j \neq 0$	
H_0: $\beta_k = 0$, pre všetky k	Hypotézu H_0 zamietame, ak $F_{ANOVA\text{-}\beta} > F_{(K-1),\,JK(c-1),\,(1-\alpha)}$
H_1: Existuje aspoň jedno k také, aby $\beta_k \neq 0$	
H_0: $\alpha_j\beta_k = 0$, pre všetky j,k	Hypotézu H_0 zamietame, ak $F_{ANOVA\text{-}\alpha\beta} > F_{(J-1)(K-1),\,JK(c-1),\,(1-\alpha)}$
H_1: Existuje aspoň jedno j,k také, aby $\alpha_j\beta_k \neq 0$	

K jednotlivým hypotézam potrebujeme vypočítať testovacie charakteristiky, pri ktorých budeme vychádzať z celkovej sumy štvorcov *TSS*:

$$\begin{aligned}\sum_{j}\sum_{k}\sum_{i}\left(X_{i,j,k} - \overline{X}_{\bullet\bullet\bullet}\right)^2 &= Kc\sum_{j=1}^{J}\left(\overline{X}_{\bullet j\bullet} - \overline{X}_{\bullet\bullet\bullet}\right)^2 + Jc\sum_{k=1}^{K}\left(\overline{X}_{\bullet\bullet k} - \overline{X}_{\bullet\bullet\bullet}\right)^2 + \\ &+ c\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\overline{X}_{\bullet jk} - \overline{X}_{\bullet j\bullet} - \overline{X}_{\bullet\bullet k} + \overline{X}_{\bullet\bullet\bullet}\right)^2 + \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{c}\left(X_{i,j,k} - \overline{X}_{\bullet jk}\right)^2\end{aligned} \qquad (4.96)$$

Pripomenieme, že c zodpovedá počtu pozorovaní v jednotlivých skupinách, čo je nejaké celé kladné číslo rovnaké pre každú skupinu, keďže uvažujeme o vyváženom modeli ANOVA. Tento zápis vieme tiež napísať ako:

$$TSS = SS_\alpha + SS_\beta + SS_{(\alpha\beta)} + SSE \qquad (4.97)$$

kde SS_α je suma štvorcov faktora α, SS_β je suma štvorcov faktora β, $SS_{(\alpha\beta)}$ je suma štvorcov z interakcií a *SSE* je suma štvorcov rezíduí. Následne si definujeme priemerné sumy štvorcov ako:

$$MSS_\alpha = \frac{SS_\alpha}{(J-1)} \tag{4.98}$$

$$MSS_\beta = \frac{SS_\beta}{(K-1)} \tag{4.99}$$

$$MSS_{\alpha\beta} = \frac{SS_{(\alpha\beta)}}{(J-1)(K-1)} \tag{4.100}$$

$$MSSE = \frac{SSE}{JK(c-1)} \tag{4.101}$$

$$TSS = SS_\alpha + SS_\beta + SS_{(\alpha\beta)} + SSE \tag{4.102}$$

Testovacie charakteristiky potom majú nasledujúci tvar:

$$\mathrm{F}_{ANOVA-\alpha} = \frac{MSS_\alpha}{MSSE} \tag{4.103}$$

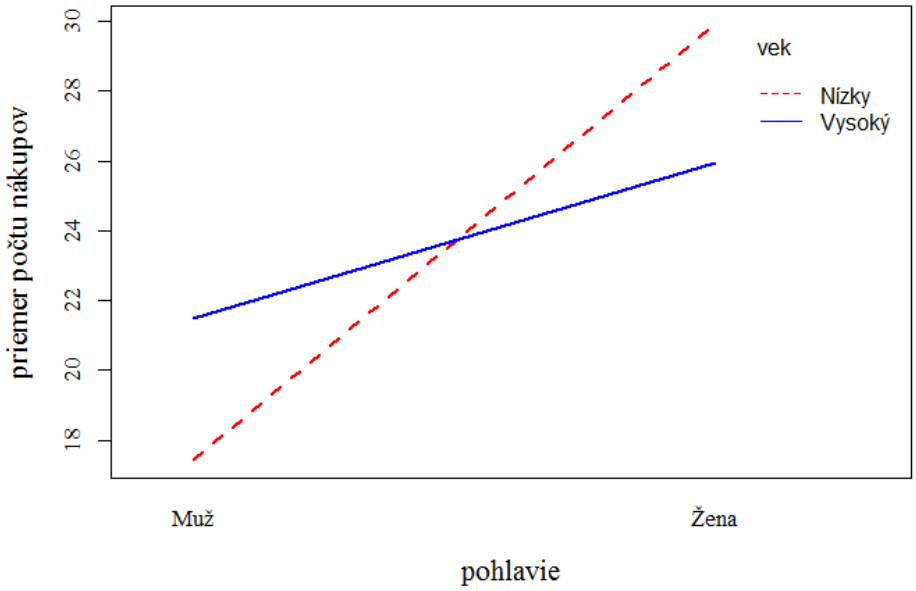
$$\mathrm{F}_{ANOVA-\beta} = \frac{MSS_\beta}{MSSE} \tag{4.104}$$

$$\mathrm{F}_{ANOVA-\alpha\beta} = \frac{MSS_{\alpha\beta}}{MSSE} \tag{4.105}$$

Kvantily $F_{(J-1),\,JK(c-1),\,(1-\alpha)}$, $F_{(K-1),\,JK(c-1),\,(1-\alpha)}$ a $F_{(J-1)\,(K-1),\,JK(c-1),\,(1-\alpha)}$ je možné odčítať z Fisherovho F rozdelenia pravdepodobnosti (prvé dva indexy predstavujú počet stupňov voľnosti a posledný konfidenciu, pri ktorej nás zaujíma príslušná hodnota kvantilu). V programe R k tomu slúži funkcia `qf()`. Aplikáciu dvojfaktorovej metódy ANOVA si ukážeme na príklade z predošlej kapitoly. Okrem pohlavia nás bude zaujímať aj faktor vek a samozrejme aj interakcia medzi faktorom vek a faktorom pohlavia. Budeme overovať všetky tri hypotézy. Najprv sa však zvykne overovať hypotéza o interakcii. Ak je táto interakcia významná, potom nie vždy má praktický význam venovať sa ďalej aj jednotlivým faktorom. Významná interakcia totiž znamená, že interpretovanie faktorov oddelene ignoruje interakciu, ktorá medzi faktormi existuje.

Začneme zostrojením obrázku interakcií, pri ktorého tvorbe použijeme funkciu `interaction.plot()`. Premenné sme si najprv upravili takým spôsobom, že namiesto indikátorovej premennej 0/1 sme v premenných „`pohlavie`“ a „`vek`“ použili textové reťazce „muž“/“žena a „nízky“/“vysoký“. Z nasledujúceho obrázku (pozri Obrázok 4.22) sa

zdá byť významným faktorom interakcia medzi pohlavím a počtom nákupov, ako aj samotné pohlavie. Jedným z možných vysvetlení môže byť, že ženy chodia častejšie nakupovať, takže bez ohľadu na vek, priemerný počet nákupov bude vyšší ako u mužov. Význam faktora pohlavie je možné vidieť tak, že si „vizuálne“ spriemerujeme konce čiar nad pohlavím „žena“ a nad pohlavím „muž“. Rozdiely medzi týmito „virtuálnymi“ priemermi sú zjavné. Z tohto obrázku sa dá taktiež vyčítať informácia, že medzi priemerným počtom nákupov u respondentov s nízkym a vysokým vekom nebude veľký rozdiel (podľa osi *y*-ovej spriemerujeme konce modrej čiary a porovnáme s priemerom koncov červenej čiary).



Obrázok 4.22: Interakčný graf: Interakcia medzi pohlavím a vekom pri nákupe

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> for (i in 1:length(nakup)) {
+ if (pohlavie[i] == 1) pohlavie[i] = "Žena"
+ if (pohlavie[i] == 0) pohlavie[i] = "Muž"
+ if (vek[i] == 1) vek[i] = "Nízky"
+ if (vek[i] == 0) vek[i] = "Vysoký"
+ }
> interaction.plot(pohlavie, vek, nakup, ylab = "priemer počtu
  nákupov", family = "serif", cex.lab = 1.3, cex.axis = 1.1, lwd
  = 2.5, col = c("red", "blue"))
```

Najprv sme vykonali prepočet manuálne. Uvádzame priebežné výsledky, pomocou ktorých je možné celý prepočet ľahšie skontrolovať.

```
> J <- 2; K <- 2; c <- 30; table(pohlavie, vek);
        vek
pohlavie Nízky Vysoký
    Muž     30     30
    Žena    30     30
```

```
> interaction <- paste(vek, pohlavie, sep = " ", collapse =
  NULL)
> SSalpha <- K*c*sum((tapply(nakup, pohlavie, mean) -
  mean(nakup))^2); SSalpha
[1] 2159.008
> SSbeta <- J*c*sum((tapply(nakup, vek, mean) - mean(nakup))^2);
  SSbeta
[1] 0.075
> SSinterakcia <- c*sum((tapply(nakup, interaction, mean) -
  rep(tapply(nakup, pohlavie, mean), 2) - c(mean(nakup[vek ==
  "Nízky"]), mean(nakup[vek == "Vysoký"]), mean(nakup[vek ==
  "Nízky"]), mean(nakup[vek == "Vysoký"])) + rep(mean(nakup),
  4))^2); SSinterakcia
[1] 484.1583
> SSE <- sum((nakup[pohlavie == "Muž" & vek == "Nízky"] -
  rep(tapply(nakup, interaction, mean)[1], 30))^2,
  (nakup[pohlavie == "Muž" & vek == "Vysoký"] -
  rep(tapply(nakup, interaction, mean)[3], 30))^2,
  (nakup[pohlavie == "Žena" & vek == "Nízky"] -
  rep(tapply(nakup, interaction, mean)[2], 30))^2,
  (nakup[pohlavie == "Žena" & vek == "Vysoký"] -
  rep(tapply(nakup, interaction, mean)[4], 30))^2); SSE
[1] 1931.7
> MSSalpha <- SSalpha / (J - 1); MSSalpha
[1] 2159.008
> MSSbeta <- SSbeta / (K - 1); MSSbeta
[1] 0.075
> MSSinterakcia = SSinterakcia / ((J - 1)*(K - 1));
  MSSinterakcia
[1] 484.1583
> MSSE <- SSE / (J*K*(c - 1)); MSSE
[1] 16.65259
> FANOVA_alpha <- MSSalpha / MSSE; FANOVA_alpha;
[1] 129.65
> FANOVA_beta <- MSSbeta / MSSE; FANOVA_beta
[1] 0.004503805
> FANOVA_interakcia = MSSinterakcia / MSSE; FANOVA_interakcia
[1] 29.07406
> # Kritické hodnoty sú v tomto prípade rovnaké
> qf(0.95, (J - 1), J*K*(c - 1)); qf(0.95, (K - 1), J*K*(c -
  1)); qf(0.95, (J - 1)*(K - 1), J*K*(c - 1))
[1] 3.922879
[1] 3.922879
[1] 3.922879
```

Porovnaním testovacích charakteristík s kritickou hodnotou zisťujeme, že nevieme zamietnuť hypotézu o nevýznamnosti faktora vek (čo obrázok interakcií už naznačoval skôr). Na druhej strane zamietame hypotézu o nevýznamnosti faktora pohlavie a interakcie. Tieto výsledky nás privádzajú k nasledujúcemu záveru. Priemerný mesačný nákup žien je väčší ako u mužov. Zároveň starší muži nakupujú viac ako mladší muži a staršie ženy nakupujú menej

ako ženy mladšie. Interakcia môže spočívať v tom, že s vekom narastá potreba mužov nakupovať a u žien táto potreba naopak s vekom klesá (aj keď je stále celkovo vyššia).

Prepočet môžeme vykonať aj použitím funkcie `aov()` a `summary()`. Výsledky, vrátane jednotlivých súm štvorcov, sú obdobné ako pri manuálnom prepočítavaní.

```
> summary(aov(nakup ~ pohlavie + vek + pohlavie : vek))
             Df  Sum Sq Mean Sq  F value    Pr(>F)
pohlavie      1 2159.01 2159.01 129.6500 < 2.2e-16 ***
vek           1    0.08    0.08   0.0045    0.9466
pohlavie:vek  1  484.01  484.01  29.0651 3.734e-07 ***
Residuals   116 1931.70   16.65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Na záver ešte overíme predpoklady. Najprv použijeme Brown – Forsythov test zhody rozptylov (máme spolu štyri skupiny) a potom Shapiro – Wilkov test normality. Vo všetkých prípadoch empirické údaje nenaznačujú, že predpoklad homoskedasticity rozptylov a normality pozorovaní by mali byť porušené.

```
> library(car)
> leveneTest(nakup, group = interaction, center = median)
  Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  1.2375 0.2994
      116
-----------------------------------------------------------------
> library(nortest)
> shapiro.test(nakup[interaction=="Nízky Žena"])

               Shapiro-Wilk normality test

data:  nakup[interaction == "Nízky Žena"]
W = 0.964, p-value = 0.3895
-----------------------------------------------------------------
> shapiro.test(nakup[interaction=="Vysoký Muž"])

               Shapiro-Wilk normality test

data:  nakup[interaction == "Vysoký Muž"]
W = 0.9615, p-value = 0.3377
-----------------------------------------------------------------
> shapiro.test(nakup[interaction=="Vysoký Žena"])

               Shapiro-Wilk normality test

data:  nakup[interaction == "Vysoký Žena"]
W = 0.9497, p-value = 0.1664
-----------------------------------------------------------------
> shapiro.test(nakup[interaction=="Nízky Muž"])

  Shapiro-Wilk normality test
```

```
data:  nakup[interaction == "Nízky Muž"]
W = 0.9641, p-value = 0.3923
```

Našu krátku diskusiu o metóde ANOVA týmto končíme. V danej problematike je toho samozrejme podstatne viac. V praxi sa napríklad často môžeme stretnúť s nevyváženým modelom ANOVA. Tomu sme sa nevenovali. V programe R môžeme použiť rovnaké funkcie na výpočet ako doteraz. Z praktického hľadiska sa ako dôležitejšia téma môže javiť tzv. post-hoc analýza modelov ANOVA. Výsledok z analýzy pomocou metódy ANOVA je v prípade zamietnutia nulovej hypotézy tvrdenie, že existuje štatisticky významný vzťah medzi závislou premennou a špecifickou hladinou (skupinou) určitého faktora. Otázkou ostáva, ktorá je to hladina daného faktora, ktorá tento rozdiel spôsobuje. Okrem vizualizácie údajov si môžeme pomôcť formálnejším prístupom, ktorý sa realizuje práve v tzv. post-hoc ANOVA analýze. Za účelom ďalšieho štúdia v danej oblasti odporúčame literatúru Tkáč (2001) a Montgomery – Runger (2011).

5 Meranie závislostí

Pod mierami závislostí si môžeme predstaviť číselné charakteristiky, ktoré slúžia na kvantifikáciu sily vzťahu medzi dvoma, prípadne viacerými premennými. Pri meraní závislostí nás v empirických aplikáciách spravidla zaujímajú odpovede na dve otázky:

- Aká silná je závislosť medzi premennými?
- Je závislosť medzi premennými štatisticky významná?

Na prvú otázku odpovedáme pomocou koeficientov (mier) závislostí, na druhú pomocou testovania štatistických hypotéz. V skutočnosti býva odpoveď na prvú otázku náročnejšia, keďže výsledkom z výpočtu určitého koeficientu závislosti je konkrétna hodnota, avšak táto hodnota môže znamenať rôznu silu vzťahu v závislosti od situácie, ktorej sa analýza týka.

Koeficienty závislosti sú spravidla konštruované tak, aby ich hodnoty patrili do intervalu od –1 do 1, kde 0 znamená žiadnu závislosť medzi premennými, 1 dokonalú (deterministickú) priamu závislosť a –1 dokonalú nepriamu závislosť. Pod priamou závislosťou medzi dvoma premennými si môžeme predstaviť situáciu, keď sú väčšie hodnoty jednej premennej sprevádzané väčšími hodnotami druhej premennej a menšie menšími. A naopak, pod nepriamou závislosťou situáciu, keď sú väčšie hodnoty jednej premennej sprevádzané menšími hodnotami druhej premennej.

Na tomto mieste pripomenieme, že v tejto časti **pod pojmom závislosť**, **resp. nezávislosť**, myslíme iba takú formu závislosti, ktorú vieme pomocou konkrétneho koeficientu odhadnúť. Niekedy sa tejto forme chápania závislostí hovorí aj **korelovanosť** (často sa korelovanosť spája s lineárnou závislosťou medzi premennými). Predstavme si nasledujúci deterministický vzťah medzi dvoma premennými $y = x^2$. Zjavne existuje vzťah medzi y a x. Napriek tomu pre prípady, kde $x \sim U(-a, a)$, alebo $x \sim N(0, \sigma^2)$ (kde a patrí do R) nájdeme pomocou Pearsonovho korelačného koeficientu veľmi malú závislosť (teoreticky žiadnu závislosť). Dôvod je ten, že Pearsonov korelačný koeficient (predstavíme si ho za chvíľu) meria lineárnu závislosť medzi premennými, kým vzťah medzi y a x nie je lineárny. Ak sa nenájde použitím Pearsonovho korelačného koeficientu závislosť, môžeme sa stretnúť s pomenovaním **nekorelovanosť**. Použitím vybraného koeficientu nemusíme nájsť závislosť (v ponímaní použitého koeficientu), avšak v skutočnosti premenné môžu byť **závislé**.

V empirických aplikáciách sa ďalej môžeme stretnúť so situáciou, keď vyhodnotíme koeficient závislosti ako silný, avšak nevyjde štatisticky významný. Opačný prípad nastáva podľa našej skúsenosti častejšie, teda keď vyjde nízka hodnota koeficientu závislosti, avšak

v následnom testovaní významnosti bude štatisticky významná. Interpretácia takýchto výsledkov závisí vždy od kontextu analýzy, na účel ktorej sa koeficienty závislosti a ich významnosti počítajú.

Ďalej považujeme za dôležité pripomenúť, že ak sa medzi premennými vyskytne závislosť, nemusí to nutne znamenať **kauzalitu.** Spravidla je výskyt závislosti nutným predpokladom ku kauzalite, nie však postačujúcim. Vzťah medzi závislosťou a kauzalitou si ilustrujeme na nasledujúcom triviálnom príklade. Za minútu sa na planéte Zem narodí približne 60 detí. Predstavte si, že približne každú minútu lúsknete prstami. Ak by sme chceli zmerať závislosť medzi lúskaním a narodením detí, zrejme by sme našli silnú závislosť. Každú minútu medzi lúskaním prstov sa totiž narodí skoro 60 detí. Znamená to, že sa deti rodia preto, lebo lúskame prstami?

Existuje pomerne mnoho koeficientov závislostí, z ktorých si na opis vyberieme tie najčastejšie sa vyskytujúce. Pôjde o: Pearsonov koeficient korelácie a Spearmanov poradový koeficient. Tieto koeficienty dopĺňame o Kendallove koeficienty a o koeficienty, ktoré sa používajú, ak počítame závislosť z kategorických (nominálnym) premenných, prípadne priamo z kontingenčných tabuliek (týmto koeficientom sa taktiež hovorí kontingenčné koeficienty).

5.1.1 Pearsonov korelačný koeficient

Majme usporiadané dvojice pozorovaní (X_i, Y_i), i = 1, 2, ..., n. Pearsonovým korelačným koeficientom charakterizujeme vzájomnú lineárnu závislosť medzi premennou X a Y. Na jeho výpočet použijeme nasledujúci formálny vzťah:

$$r_{x,y} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} \qquad (5.1)$$

$r_{x,y}$ nadobúda hodnoty v intervale $r_{x,y} \in \langle -1, 1 \rangle$, pričom platí, že ak $r_{x,y} = 0$ tvrdíme, že medzi premennými neexistuje lineárna závislosť, $r_{x,y} = 1$ existuje dokonalá priama lineárna závislosť a pre $r_{x,y} = -1$ dokonalá nepriama lineárna závislosť. Čím je hodnota $|r_{x,y}|$ bližšie k 1, o to je väčšia (silnejšia) lineárna závislosť medzi skúmanými premennými.

Pearsonov korelačný koeficient meria iba lineárnu závislosť. Nemusí byť vhodný pri premenných, ktoré majú sezónne alebo cyklické zložky (napr. $y = \sin(x)$). V prípade, ak je dôvodné podozrenie, že by medzi premennými mal existovať nelineárny vzťah, Pearsonov

korelačný koeficient tento vzťah „podhodnotí“ a je vhodné použiť alternatívny spôsob merania závislostí (iný koeficient, prípadne použiť regresnú analýzu).

Pearsonov korelačný koeficient je ďalej nevhodné používať, ak pre usporiadané dvojice pozorovaní neplatí homoskedasticita. Pre účely tejto kapitoly si homoskedasticitu môžeme predstaviť nasledovne. Predpokladajme meranie závislosti medzi váhou *X* a výškou *Y* náhodne vybraných osôb. Nech sa váha pohybuje v intervale od 40 *kg* do 120 *kg*. Vyberieme si iba osoby, ktoré majú váhu v intervale od 40 *kg* do 50 *kg* (vrátane) a vypočítame si variabilitu ich výšky, označíme ako $\sigma^2_{(40\text{-}50>}$. Ďalej si vyberieme osoby s váhou od 50 *kg* do 60 *kg* (vrátane) a tiež si vypočítame variabilitu ich výšky $\sigma^2_{(50\text{-}60>}$. Takto môžeme postupovať až po 110 *kg* do 120 *kg*. V prípade homoskedasticity očakávame rovnosť týchto rozptylov, $\sigma^2_{(40\text{-}50>} = \sigma^2_{(50\text{-}60>} = ... = \sigma^2_{(110\text{-}120>}$. Samozrejme, mohli sme vybrať aj iné intervaly váh a taktiež by sme mohli vybrať rôzne intervaly výšky a vypočítať zodpovedajúce variability váh, ktoré by mali byť rovnaké. Ak neplatí homoskedasticita, hovoríme o heteroskedasticite. V takom prípade sa **môže** stať, že na určitom úseku bude závislosť výrazne odlišná ako na inom. Inak povedané, Pearsonov korelačný koeficient na celom intervale bude zavádzajúcim ukazovateľom závislosti.

Pearsonov korelačný koeficient je ďalej nevhodné použiť v situácii, kde existujú extrémne hodnoty (samozrejme príčiny takýchto hodnôt je vždy nutné zhodnotiť). Dá sa pomerne ľahko ukázať, že ide o koeficient, ktorého výsledok do značnej miery vieme ovplyvniť jednou extrémnou hodnotou.

Ak *X* a *Y* sú náhodné premenné pochádzajúce zo spojitých rozdelení, potom vzťah *(5.1)* predstavuje odhad populačného korelačného koeficientu $\rho_{x,y}$. V prípade menších vzoriek sa vzťah *(5.1)* považuje za skreslený odhad korelačného koeficientu $\rho_{x,y}$ a namiesto toho sa môže používať aj nasledujúci vzťah:

$$r^a_{x,y} = \sqrt{1 - \frac{\left(1 - r^2_{x,y}\right)(n-1)}{(n-2)}} \qquad (5.2)$$

Následne môžeme testovať hypotézy o $\rho_{x,y}$. Najčastejšie sa overuje nulová hypotéza $\rho_{x,y} = 0$ oproti alternatívnej hypotéze $\rho_{x,y} \neq 0$, keďže v prípade zamietnutia nulovej hypotézy existuje štatistický dôkaz o prítomnosti lineárnej závislosti medzi premennými. Testovacia charakteristika (v prípade ak H_0: $\rho_{x,y} = 0$) má nasledujúci tvar:

$$t_r = \frac{r_{x,y}\sqrt{(n-2)}}{\sqrt{1 - r^2_{x,y}}} \qquad (5.3)$$

Táto štatistika sa riadi Studentovým *t* rozdelením pravdepodobnosti s $(n - 2)$ stupňami voľnosti. Rozhodnutie o hypotéze potom vykonáme nasledovne:

H_0: $\rho_{x,y} = 0$	Hypotézu H_0 zamietame, ak $\lvert t_s \rvert > \lvert t_{(1-\alpha/2),(n-2)} \rvert$
H_1: $\rho_{x,y} \neq 0$	

Uvažujme o nasledujúcom príklade, kde meriame závislosť medzi rýchlosťou akou operátor nastaví stroj a množstvom chýb počas zmeny. Ak nám vyjde korelačný koeficient povedzme $r_{x,y} = 0.05$ a je štatisticky významný, neznamená to, že existuje zmysluplný vzťah medzi tým, ako rýchlo operatér nastavil stroj a koľko chýb sa vyskytlo (pre zjednodušenie tu predpokladáme kauzalitu). Významnosť je teda len určitým štatistickým indikátorom. Avšak vždy závisí od podstaty riešeného problému, ako sa k interpretácii výsledkov postavíme, či hodnotu korelačného koeficientu budeme považovať za zmysluplnú alebo nie. Hodnota koeficientu na úrovni 0.05 v zásade nemusí znamenať žiaden vzťah.

Pri použití štatistického testu pripomenieme niektoré problémy, s ktorými sa môžeme stretnúť. Testovanie významnosti korelačného koeficientu predpokladá, že náhodné premenné *X* a *Y* pochádzajú z dvojrozmerného normálneho rozdelenia pravdepodobnosti. To znamená, že *X* aj *Y* musia pochádzať z normálneho rozdelenia pravdepodobnosti (marginálne rozdelenia pravdepodobnosti by mali byť normálne). Okrem toho, pre každú realizáciu náhodnej premennej *X*, rozdelenie náhodnej premennej *Y* musí tiež pochádzať z normálneho rozdelenia. Taktiež naopak, pre každú realizáciu náhodnej premennej *Y*, rozdelenie náhodnej premennej *X* musí tiež pochádzať z normálneho rozdelenia (t.j. podmienené rozdelenia majú byť normálne). K overovaniu tohto predpokladu spravidla slúžia štatistické testy viacrozmernej normality (ktoré si opíšeme v ďalších častiach), prípadne vizualizácia dát. Ak je *n* dostatočne veľké (aby platila viacrozmerná centrálna limitná veta), porušenie tohto predpokladu nemá závažné dôsledky. Ak aj je *n* dostatočne veľké, je vhodné, aby boli rozdelenia pravdepodobnosti, z ktorých *X* a *Y* pochádzajú, čo najviac podobné. V opačnom prípade môže dôjsť k podhodnocovaniu lineárnej závislosti medzi premennými.

Interpretácia sily lineárnej väzby medzi premennou *X* a *Y* nie je jednoznačná. V literatúre sa môžeme stretnúť s rôznymi intervalmi, ktoré naznačujú ako interpretovať korelačný koeficient. Napríklad ak je hodnota $|r_{x,y}| > 0.7$ môže sa to považovať za silný vzťah. Uvedené však neplatí vždy. Pri riadenom experimente v laboratórnych podmienkach by $r_{x,y} = 0.71$ mohlo znamenať neúspešný experiment, kým v spoločenských vedách je táto hodnota Pearsonovho korelačného koeficientu považovaná často za veľmi silnú, najmä ak sa porovnávajú nepriamo merané faktory, napr. korelácia medzi „stresom“ a „kvalitou života“.

Zaujímavým faktom je skutočnosť, že ak rozdelenia, z ktorých X a Y pochádzajú nie sú normálne, je možné že $r_{x,y}$ bude nadobúdať ešte menší rozsah hodnôt ako $\langle -1, 1 \rangle$ (pozri Shih – Huang, 1992). To samozrejme znemožňuje interpretáciu skoro akékoľvek odporúčania ohľadom sily korelačných koeficientov.

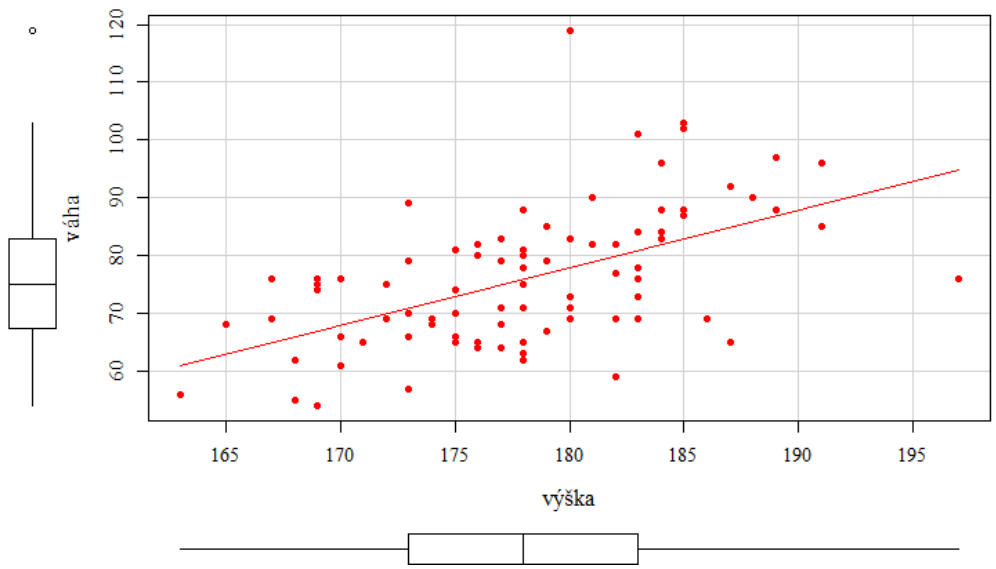
Všimnime si, že s predpokladom normality Pearsonov korelačný koeficient taktiež predpokladá, že premenné X a Y sú spojité náhodné premenné. V praxi sa však často používa koeficient aj pri vyhodnocovaní dotazníkov, kde sa používajú hodnoty namerané na škále poradového charakteru. V tomto prípade je otázne formálne splnenie podmienky o normalite. Napriek týmto skutočnostiam, ak je škála dostatočne početná, použitím Pearsonovho korelačného koeficientu sa nemusíme dopustiť veľkého skreslenia (pozri tzv. *floor* a *ceiling* efekty). V prípade, ak rozdelenia, z ktorých náhodné premenné X a Y pochádzajú sa javia ako veľmi zošikmené, jednou z možností je ich transformovať. Jednou z najčastejších transformácií (ak je to možné) je logaritmovať všetky hodnoty. Ak meriame spokojnosť zákazníka na škále od 1 po 10, tak ide o určitú „virtuálnu" jednotku, ktorej logaritmovaním nemusíme zmeniť podstatu analýzy.

Použitie korelačného koeficientu si ukážeme na príklade, kde nás bude zaujímať, aký silný je lineárny vzťah medzi váhou a výškou respondentov, a či je štatisticky významný (rôzny od 0). Budeme predpokladať, že namerané hodnoty sú realizáciami náhodného výberu. Použijeme databázu `Davis` z programového balíka `car`.

Príklad 5.1

Keďže v databáze sa vyskytujú údaje pre mužov ako aj pre ženy, vytvoríme si najprv databázu s výškou a váhou mužov. Následne si vzájomný vzťah vizualizujeme pomocou *x-y* grafu, pričom použijeme novú funkciu `scatterplot()` (pozri Obrázok 5.1).

```
> library(car); attach(Davis)
> men <- Davis[sex == "M",]; detach(Davis)
> names(men) <- c("pohlavie", "vaha", "vyska", "vaha_s",
  "vyska_s")
> attach(men)
> scatterplot(vaha ~ vyska, smooth = FALSE, xlab = "výška", ylab
  = "váha", col = "red", pch = 19, cex.lab = 1.3, cex.axis =
  1.1, family = "serif")
```



Obrázok 5.1: *x-y* graf závislosti výšky a váhy mužov

Zdroj: vlastné spracovanie, výstup zo softvéru R

Z vizualizácie sa zdá byť zrejmé, že v nameraných údajoch sa vyskytujú určité extrémne hodnoty, čo naznačuje aj box – plot popri *y*-ovej osi. Tieto hodnoty môžu skresliť hodnotu korelačného koeficientu. Taktiež je na mieste otázka, či sú pozorovania z normálneho rozdelenia. Po vypočítaní základných opisných charakteristík sme uskutočnili Shapiro – Wilkov test na normalitu jednotlivých vzoriek (určite lepšou alternatívou je uskutočnisť test viacrozmerného normálneho rozdelenia, ktorý je však prezentovaný neskôr). Pri váhe sme zamietli nulovú hypotézu o normalite údajov. Keďže veľkosť vzorky (`dim(men)[1]`) považujeme za dostatočnú, rozhodli sme sa vypočítať korelačný koeficient a vyhodnotiť jeho významnosť.

```
> summary(vaha); summary(vyska)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  54.00   67.75   75.00   75.90   83.00  119.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    163     173     178     178     183     197
> shapiro.test(vaha); shapiro.test(vyska)

                Shapiro-Wilk normality test

data:  vaha
W = 0.9613, p-value = 0.01001


                Shapiro-Wilk normality test

data:  vyska
W = 0.992, p-value = 0.872
------------------------------------------------------------------
```

```
> cor.test(vaha, vyska)

            Pearson's product-moment correlation

data:  vaha and vyska
t = 5.9388, df = 86, p-value = 5.922e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.3718488 0.6727460
sample estimates:
      cor
0.5392906
```

Namerali sme silne štatisticky významný korelačný koeficient[25] (*p*-hodnota < 0.000), ktorého hodnota je približne 0.539. Pri zohľadnení rôznorodosti ľudí, považujeme túto závislosť za zmysluplnú. Zároveň si všimnime, že je kladná, čo podporuje intuíciu, že s väčšou výškou by sme vo väčšine prípadov mali pozorovať ťažších mužov a naopak. Aby sme boli schopní posúdiť vplyv extrémnych hodnôt na tieto výsledky, musíme najprv tieto extrémne hodnoty identifikovať. Použili sme Hampelov test, konkrétne funkciu, ktorú sme si vytvorili v kapitole 4.5.3. Následne sme zopakovali výpočet opisných štatistík a znovu vykonali iba Shapiro – Wilkovov test.

```
> index <- (1:length(vaha))[vaha %in% hampel_identifier(vaha)]
> index <- c(index, (1:length(vyska))[vyska %in%
  hampel_identifier(vyska)])
> men_no <- men[-index,]; rm(men); detach(men); attach(men_no)
> summary(vaha); summary(vyska)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  54.00   67.00   74.00   74.75   82.00  101.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  163.0   173.0   178.0   177.6   182.0   191.0
-------------------------------------------------------------
> shapiro.test(vaha); shapiro.test(vyska)

                Shapiro-Wilk normality test

data:  vaha
W = 0.9818, p-value = 0.2808


                Shapiro-Wilk normality test

data:  vyska
W = 0.9897, p-value = 0.7476
```

[25] Pri nasledovnom značení *p*-hodnota < 0.000 sa na pravej strane pripúšťajú číslice aj za štvrtým desatinným miestom, ale z hľadiska rozhodnutia o hypotéze už nie je presný výsledok podstatný.

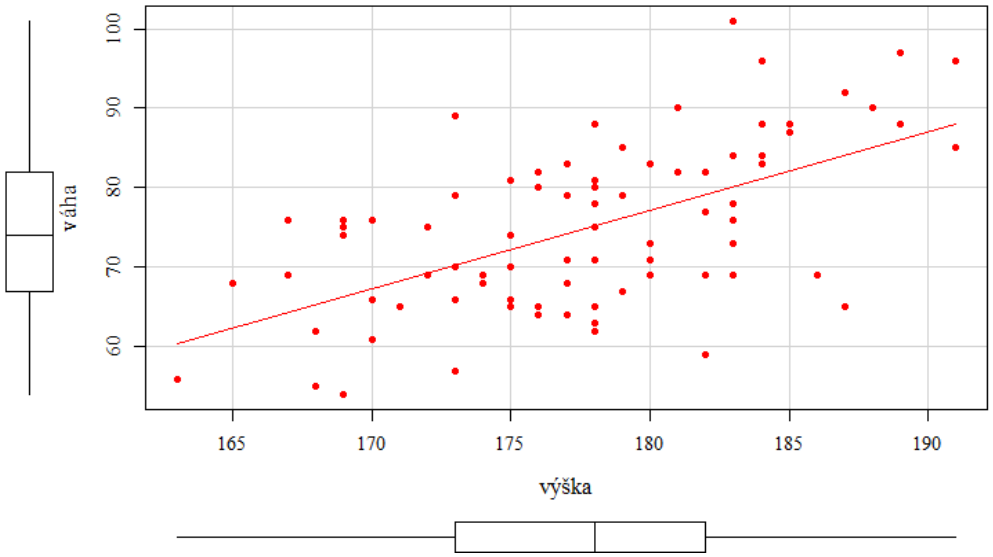
V kóde vyššie sme použili funkciu `hampel_identifier()`, ktorej výstupom sú extrémne hodnoty. K tomu, aby sme ich vylúčili z databázy potrebujeme poznať, v ktorých riadkoch sa tieto extrémne hodnoty nachádzajú. K tomu slúžil vektor `index`.

Štatistické testy naznačujú, že nevieme zamietnuť hypotézu o normalite už ani pri váhe. Upozorňujeme, že k tomu, aby váha a výška tvorili dvojrozmerné normálne rozdelenie nestačí, aby boli oba súbory z normálneho rozdelenia (je to nutná, nie postačujúca podmienka). Keďže však formálne testy na viacrozmerné normálne rozdelenie sú predmetom tejto publikácie až v neskorších častiach, budeme v analýze napriek tomu pokračovať a vystačíme si s jednorozmerným Shapiro – Wilkovým testom. Necháme na čitateľovi aby overil prítomnosť viacrozmernej normality.

Vzťah medzi váhou a výškou si ešte raz zobrazíme, tentoraz však bez extrémnych hodnôt a zopakujeme prepočet a testovanie koeficientu korelácie (bližšie pozri Obrázok 5.2).

```
> scatterplot(vaha ~ vyska, smooth = FALSE, xlab = "výška", ylab
  = "váha", col = "red", pch = 19, cex.lab = 1.3, cex.axis =
  1.1, family = "serif")
> cor.test(vaha, vyska)

            Pearson's product-moment correlation

data:  vaha and vyska
t = 6.5383, df = 82, p-value = 4.963e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.4242363 0.7105998
sample estimates:
     cor
0.585388
```

Koeficient vyšiel znovu významný (čo nebolo v tejto situácii prekvapujúce) a zároveň došlo k miernemu nárastu sily lineárneho vzťahu medzi váhou a výškou mužov.

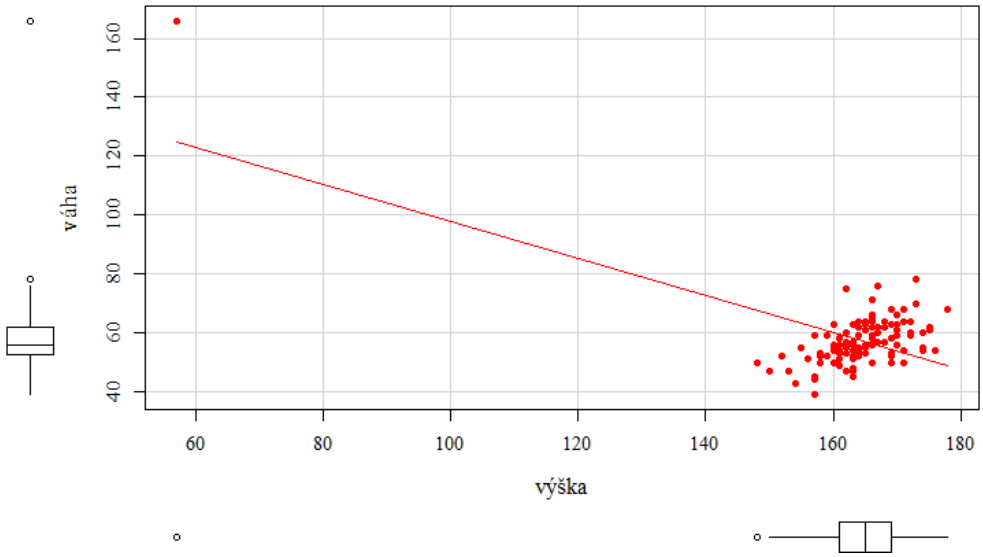


Obrázok 5.2: *x-y* graf závislosti výšky a váhy mužov (bez extrémnych hodnôt)

Zdroj: vlastné spracovanie, výstup zo softvéru R

Analýzu lineárneho vzťahu medzi výškou a váhou si zopakujeme aj na vzorke žien, kde je vplyv extrémnej hodnoty výraznejší. Podobne ako v predošlom prípade si najprv vypočítame základné opisné charakteristiky celého súboru a vzťah zobrazíme v nasledujúcom obrázku (pozri Obrázok 5.3).

```
> attach(Davis)
> women <- Davis[sex == "F",]; detach(Davis)
> names(women) <- c("pohlavie", "vaha", "vyska", "vaha_s",
  "vyska_s")
> attach(women)
> summary(vaha); summary(vyska)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   52.75   56.00   57.87   62.00  166.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   57.0   161.0   165.0   163.7   169.0   178.0
------------------------------------------------------------
> shapiro.test(vaha); shapiro.test(vyska)
                Shapiro-Wilk normality test

data:  vaha
W = 0.5564, p-value < 2.2e-16


                Shapiro-Wilk normality test

data:  vyska
W = 0.4821, p-value < 2.2e-16
------------------------------------------------------------
> scatterplot(vaha ~ vyska, smooth = F, xlab = "výška", ylab =
  "váha", col = "red", pch = 19, cex.lab = 1.3, cex.axis = 1.1,
  family = "serif")
```



Obrázok 5.3: *x-y* graf závislosti výšky a váhy žien

Zdroj: vlastné spracovanie, výstup zo softvéru R

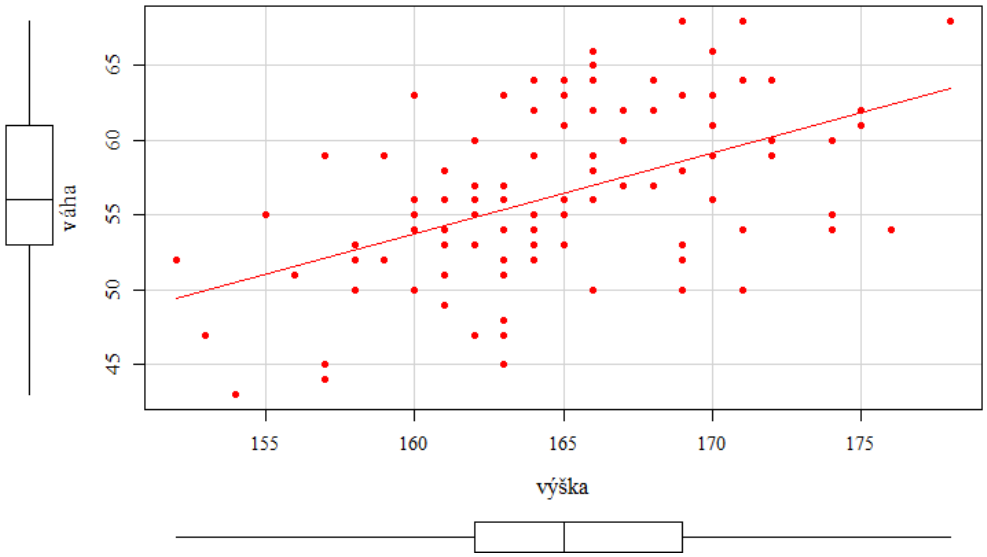
Oba box – ploty signalizujú prítomnosť niekoľkých extrémnych hodnôt. Testy na normalitu silne zamietajú nulovú hypotézu a taktiež sú zjavne veľké rozdiely medzi dolným kvartilom výšky a minimálnou hodnotou, ako aj medzi horným kvartilom váhy a maximálnou hodnotou.

```
> cor.test(vaha, vyska)

            Pearson's product-moment correlation

data:  vaha and vyska
t = -7.6516, df = 110, p-value = 8.125e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.6985368 -0.4534035
sample estimates:
       cor
-0.5893743
```

Po výpočte korelačného koeficientu a príslušného testu sú výsledky prekvapujúce (resp. bez vizualizácie vzťahu premenných by boli prekvapujúce). Medzi výškou a váhou žien bol identifikovaný záporný vzťah, t. j. nepriama lineárna závislosť, ktorá vyšla štatisticky významná. Z grafickej vizualizácie vzťahu premenných je nám však jasné, že prítomnosť extrémnych hodnôt skresľuje dosiahnuté výsledky. Z tohto dôvodu sme sa rozhodli vylúčiť extrémne hodnoty podobným spôsobom ako pri vzorke mužov.

```
> index <- (1:length(vaha))[vaha %in% hampel_identifier(vaha)]
> index <- c(index, (1:length(vyska))[vyska %in%
  hampel_identifier(vyska)])
> women_no <- women[-index,]; rm(women); detach(women);
> attach(women_no)
> summary(vaha); summary(vyska)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    53.0    56.0    56.4    61.0    68.0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  152.0   162.0   165.0   164.9   169.0   178.0
-----------------------------------------------------------------
> shapiro.test(vaha); shapiro.test(vyska)

                Shapiro-Wilk normality test

data:  vaha
W = 0.9855, p-value = 0.3247

                Shapiro-Wilk normality test

data:  vyska
W = 0.9916, p-value = 0.7744
-----------------------------------------------------------------
> scatterplot(vaha ~ vyska, smooth = FALSE, xlab = "výška", ylab
  = "váha", col = "red", pch = 19, cex.lab = 1.3, cex.axis =
  1.1, family = "serif")
-----------------------------------------------------------------
> cor.test(vaha, vyska)

            Pearson's product-moment correlation

data:  vaha and vyska
t = 5.8812, df = 101, p-value = 5.307e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3452963 0.6363919
sample estimates:
      cor
0.5050705
```

Efekt vylúčenia extrémnych hodnôt je výrazný. Po úprave dát je korelačný koeficient kladný už aj pri vzorke žien. Všimnime si výraznú zmenu v nasledujúcom obrázku (Obrázok 5.4). Box – ploty neidentifikujú žiadne extrémne hodnoty a Shapiro – Wilkove testy na normalitu nezamietajú nulovú hypotézu o normalite výšky ani váhy. Tento príklad slúžil na ilustráciu toho, ako výrazne vedia byť výsledky ovplyvnené niekoľkými extrémnymi hodnotami. Vo vzorke žien pritom v jednom pozorovaní zjavne došlo ku chybnému zápisu, keď do premennej výška sa napísala váha a vice versa. Použitie testu na extrémne hodnoty preto v zásade nebolo nutné. Našim cieľom však bolo prezentovať určitý všeobecný postup.



Obrázok 5.4: *x-y* graf závislosti výšky a váhy žien bez extrémnych hodnôt

Zdroj: vlastné spracovanie, výstup zo softvéru R

5.1.2 Spearmanov poradový koeficient

Podobne ako v prípade Pearsonovho korelačného koeficientu vychádzajme z usporiadaných dvojíc pozorovaní (X_i, Y_i), $i = 1, 2, ..., n$. Spearmanov poradový koeficient korelácie ρ_S nadobúda hodnoty z intervalu $\langle -1, 1 \rangle$, pričom hodnoty bližšie k ±1 signalizujú väčšiu monotónnu závislosť medzi premennou X a Y. V prípade, ak $\rho_S > 0$ hovoríme o monotónne priamej závislosti, pri $\rho_S < 0$ o monotónnej nepriamej závislosti, pričom pri $\rho_S = 0$ hovoríme, že medzi premennými X a Y neexistuje monotónny vzťah. Uvažujme ďalej, že X_i aj Y_i sú hodnoty pochádzajúce zo spojitých rozdelení pravdepodobnosti. Prvým krokom výpočtu Spearmanovho poradového koeficientu korelácie je vytvoriť poradia nezávisle pre vzorku X_i aj Y_i. V prípade, ak neexistuje zhoda medzi hodnotami, najmenšej hodnote priradíme 1 a najväčšej n. Túto vzorku poradí, ktorá zodpovedá X_i si označíme ako $P_{x,i}$ a pre Y_i ako $P_{y,i}$. Ďalej si nadefinujme rozdiel medzi poradiami $D_i = P_{x,i} - P_{y,i}$. Spearmanov poradový koeficient odhadneme ako:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)} \qquad (5.4)$$

Ak existuje zhoda v nameraných hodnotách alebo poradiach, volí sa priemerné poradie, pričom namiesto *(5.4)* sa na odhad ρ_S použijú poradia $P_{x,i}$ a $P_{y,i}$ a následne aplikuje vzťah *(5.1)*. Ak sú usporiadané dvojice (X_i, Y_i) náhodnými *iid* realizáciami (druh

dvojrozmerného rozdelenia nie je podstatný), potom môžeme overovať hypotézu o štatistickej významnosti $\rho_S = 0$ oproti alternatíve $\rho_S \neq 0$ pomocou nasledujúcej testovacej charakteristiky:

$$Z_S = r_S \sqrt{n-1} \qquad (5.5)$$

Rozhodnutie o hypotéze vykonáme nasledovne:

H_0: $\rho_S = 0$ H_1: $\rho_S \neq 0$	Hypotézu H_0 zamietame, ak $\lvert Z_S \rvert > \lvert z_{(1-\alpha/2)} \rvert$

kde $z_{(1-\alpha/2)}$ je kvantil normovaného normálneho rozdelenia. Pri menších vzorkách sa odporúča použiť radšej nasledujúcu štatistiku:

$$t_S = r_S \sqrt{\frac{n-2}{1-r_S^2}} \qquad (5.6)$$

Rozhodnutie o hypotéze je obdobné ako pri Pearsonovom korelačnom koeficiente:

H_0: $\rho_S = 0$ H_1: $\rho_S \neq 0$	Hypotézu H_0 zamietame, ak $\lvert t_r \rvert > \lvert t_{(1-\alpha/2),\,(n-2)} \rvert$

kde $t_{(1-\alpha/2),\,(n-2)}$ je kvantil Studentovho t rozdelenia s $(n-2)$ stupňami voľnosti. Výhodou Spearmanovho poradového koeficientu korelácie je menšia náročnosť na predpoklady ako aj skutočnosť, že je schopný nájsť monotónnu závislosť medzi premennými, kým lineárna závislosť je len jeden špecifický prípad monotónnej závislosti. Na druhej strane použitím Spearmanovho poradového koeficientu nevieme bližšie charakterizovať druh závislosti (exponenciálna, logaritmická, lineárna,....).

V programe R môžeme na výpočet vrátane testovania použiť funkciu `cor.test()`, kde si zvolíme možnosť `method = c("spearman")`. Porovnáme si výsledky s predchádzajúcim príkladom, kde použijeme váhu a výšku mužov. Výsledky sú porovnateľné, čo naznačuje, že lineárny vzťah medzi výškou a váhou sa javí ako vhodný spôsob opísania závislosti týchto dvoch premenných. Keďže sa v oboch premenných vyskytujú zhody v poradiach, výstup z funkcie `cor.test()` nás upozorňuje, že *p*-hodnota je len približná. Po prepočítaní vzťahu použitím poradí a Pearsonovho korelačného koeficientu sa výsledky nemenia (tento prepočet neuvádzame).

```
> cor.test(vaha, vyska, method = c("spearman"))

                Spearman's rank correlation rho

data:  vaha and vyska
S = 92016.13, p-value = 1.087e-07
alternative hypothesis: true rho is not equal to 0
```

```
sample estimates:
      rho
0.4947056

Warning message:
In cor.test.default(vaha, vyska, method = c("spearman")) :
  Cannot compute exact p-values with ties
```

5.1.3 Kendallov τ koeficient

Ak premenné X_i a Y_i majú intervalový alebo poradový charakter, okrem Spearmanovho koeficientu je možné použiť aj Kendall τ koeficient. Majme usporiadanú dvojicu pozorovaní (X_i, Y_i), i = 1, 2, ..., n. Pre ľubovoľné dve dvojice (X_i, Y_i) a (X_j, Y_j), $i \neq j$ definujme zhodu *Zh* medzi usporiadanými dvojicami, ak platí:

$$Zh = \left(\left(X_i < X_j\right) \wedge \left(Y_i < Y_j\right)\right) \vee \left(\left(X_i > X_j\right) \wedge \left(Y_i > Y_j\right)\right) \qquad (5.7)$$

Táto situácia nastáva vtedy, ak pri dvojici *i* sú obe hodnoty menšie ako pri dvojici *j*, prípadne ak sú obe hodnoty dvojice *i* väčšie ako hodnoty *j*. Ak by sme porovnali všetky dvojice a pre každú z nich by platila zhoda, išlo by o silnú priamu závislosť. Podobne si definujme nezhodu *Nh* medzi usporiadanými dvojicami ako:

$$Nh = \left(\left(X_i < X_j\right) \wedge \left(Y_i > Y_j\right)\right) \vee \left(\left(X_i > X_j\right) \wedge \left(Y_i < Y_j\right)\right) \qquad (5.8)$$

Ak by sme porovnali všetky dvojice navzájom a pre každú z nich by platila nezhoda, zrejme by išlo o silnú nepriamu závislosť. Kendall τ koeficient využíva práve princíp týchto zhôd, resp. nezhôd. V prípade, ak sú zhodné poradia, tak nedochádza ani k zhode ani k nezhode. Všimnime si, že ak platí zhoda, potom $(X_i - X_j)(Y_i - Y_j) > 0$ a ak platí nezhoda, potom $(X_i - X_j)(Y_i - Y_j) < 0$. Definujme si indikátorovú premennú $a_{i,j}$ (podľa Gibbons – Chakraborti, 2003):

$$a_{i,j} = \operatorname{sgn}\left(X_i - X_j\right)\operatorname{sgn}\left(Y_i - Y_j\right) \qquad (5.9)$$

pričom:

$$\operatorname{sgn}(I) = \begin{cases} -1, u < 0 \\ 0\ , u = 0 \\ 1\ , u > 0 \end{cases} \qquad (5.10)$$

Zjavne, ak $a_{i,j} = 1$ potom je pár usporiadaných dvojíc zhodný, pre $a_{i,j} = -1$ je pár usporiadaných dvojíc nezhodný a pre $a_{i,j} = 0$ pár usporiadaných dvojíc nie je zhodný ani nezhodný (aspoň v jednom prípade je zhoda v poradí). Hodnotu Kendall τ koeficientu potom vypočítame ako:

$$\tau = \frac{2\sum\sum_{1\le i<j\le n} a_{i,j}}{\left(n(n-1)\right)} \qquad (5.11)$$

Vo vzťahu *(5.11)* sa uskutoční súčet pre všetky $a_{i,j}$ ($a_{i,j}$ a $a_{j,i}$ je totožné, preto dochádza iba k jednému započítavaniu, čo zabezpečuje podmienka $1 \le i < j \le n$) a vydelí sa celkovým počtom porovnávaných dvojíc, ktorý je $n(n-1)/2$. Všimnime si, že ak sú všetky porovnania zhodné, potom $\Sigma\Sigma a_{i,j} = n(n-1)/2$ a $\tau = 1$. V prípade, ak sú všetky porovnania nezhodné, tak $\Sigma\Sigma a_{i,j} = -n(n-1)/2$ a $\tau = -1$. Ak vyjde $\tau = 0$, potom počet zhodných a nezhodných porovnaní je totožný. Interpretácia Kendall τ koeficientu je obdobná ako v predošlých prípadoch. Koeficient je definovaný v uzavretom intervale od –1 až po 1, kde hodnoty bližšie k ±1 predstavujú silnejšiu závislosť.

V prípade, ak sa v súbore vyskytujú situácie, kde $u = 0$, potom je vhodné na výpočet Kendall τ koeficientu použiť nasledujúci vzťah (Siegel, 1956):

$$\tau_a = \frac{\sum\sum_{1\le i<j\le n} a_{i,j}}{\sqrt{\frac{1}{2}n(n-1)-T_x}\sqrt{\frac{1}{2}n(n-1)-T_y}} \qquad (5.12)$$

kde:

$$T_x = \frac{1}{2}\sum_{k=1}^{gx} t_k\left(t_k - 1\right) \qquad (5.13)$$

$$T_y = \frac{1}{2}\sum_{p=1}^{gy} t_p\left(t_p - 1\right) \qquad (5.14)$$

pričom t je počet zhodných poradí pre premennú X (prípadne Y vo vzťahu *(5.14)*) a gx je celkový počet skupín so zhodnými poradiami pre premennú X a pre premennú Y je celkový počet zhodných skupín označený ako gy. Ak máme nasledujúce poradia pre premennú X: 1, 2.5, 2.5, 2.5, 5, 6, 7.5, 7.5, 9, 10, potom $gx = 2$ a $t_1 = 3$, $t_2 = 2$ (pozri Kapitolu 4.6.5).

Za predpokladu, že usporiadané dvojice (X_i, Y_i) sú realizáciami náhodného výberu, je ďalším prirodzeným postupom overovanie štatistickej významnosti Kendall τ koeficientu. Dá sa ukázať, že ak je n dostatočne veľké ($n \ge 11$), za predpokladu platnosti nulovej hypotézy H_0: $\tau = 0$ sa nasledujúca testovacia charakteristika riadi normovaným normálnym rozdelením pravdepodobnosti:

$$Z_\tau = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \qquad (5.15)$$

Rozhodnutie o hypotéze je potom nasledovné:

H_0: $\tau = 0$ H_1: $\tau \neq 0$	Hypotézu H_0 zamietame, ak $\lvert Z_\tau \rvert > \lvert z_{(1-\alpha/2)} \rvert$

kde $z_{(1-\alpha/2)}$ je kvantil normovaného normálneho rozdelenia. V programe R vieme na výpočet použiť znova funkciu `cor.test()`, kde si zvolením možnosti `method = c("kendall")` vypočítame hodnotou Kendall τ koeficientu vrátane testu významnosti.

```
> cor.test(vaha, vyska, method = c("kendall"))

                Kendall's rank correlation tau

data:  vaha and vyska
z = 5.1774, p-value = 2.25e-07
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3617643
```

5.1.4 Kontingenčné tabuľky

Kým Pearsonov korelačný koeficient sa hodí pre výpočet závislosti dvoch spojitých premenných, Spearmanov a Kendallov τ koeficient je vhodné použiť tak pri spojitých premenných, ako aj pri premenných, ktoré sú merané na poradovej škále. Ani jeden z týchto koeficientov nie je možné použiť, ak nás zaujíma závislosť medzi dvoma premennými, ktoré sú merané na nominálnej škále[26]. Pri tejto problematike budeme postupovať odlišne ako v predošlých prípadoch. V tejto kapitole sa budeme venovať štatistickej závislosti dvoch nominálnych premenných a v ďalších meraniu sily tejto závislosti. Štatistickú závislosť dvoch nominálnych premenných môžeme vyhodnotiť pomocou Chí-kvadrát testu podobne, ako tomu bolo pri testovaní dobrej zhody, konkrétne pri Pearsonovom Chí-kvadrát teste dobrej zhody. Pri tomto teste sme náhodnú vzorku pozorovaní rozdelili do intervalov a následne sme porovnávali pozorovanú a teoretickú početnosť v týchto intervaloch.

Pri meraní štatistickej závislosti dvoch nominálnych premenných je postup podobný. Pozorovania si nerozdelíme na základe jedného kritéria, ale na základe dvoch kritérií. Pre ilustráciu uvažujme o dvoch nominálnych premenných. Prvou nech je rodinný stav mužov, ktorý môže (v našej prípadovej štúdii) nadobúdať tri stavy: slobodný, ženatý, rozvedený. Druhou je druh pracovného zaradenia s nasledujúcimi stavmi: zamestnaný, nezamestnaný,

[26] Za určitých predpokladov je možné na tieto problémy použiť aj iné varianty Kendallovho koeficientu. Tým sa však v tejto publikácii nebudeme bližšie venovať.

podnikateľ (živnostník). Spolu tak máme 3 x 3 = 9 kombinácií. Zaujíma nás, či existuje štatisticky významná závislosť medzi rodinným stavom a pracovným zaradením. Týchto 9 kombinácií si môžeme utriediť do tzv. dvojrozmernej kontingenčnej tabuľky.

```
> ps <- matrix(c(14, 20, 60, 15, 10, 5, 40, 21, 11), nrow = 3)
> rownames(ps) <- c("zenaty", "slobodny", "rozvedeny")
> colnames(ps) <- c("podnikatel", "nezamestnany", "zamestnany")
> ps
          podnikatel nezamestnany zamestnany
zenaty            14           15         40
slobodny          20           10         21
rozvedeny         60            5         11
```

Ak si označíme i = 1, 2, ..., r ako riadky a j = 1, 2, ..., s ako stĺpce, potom tabuľka predstavuje pozorované početnosti $o_{i,j}$. Princíp výpočtu nezávislosti vyžaduje definovane tzv. očakávaných početností $e_{i,j}$. Keďže sa bude overovať hypotéza H_0: medzi pracovným zaradením a rodinným stavom neexistuje štatisticky významná závislosť, oproti alternatíve H_1: medzi pracovným zaradením a rodinným stavom existuje štatisticky významná závislosť, tak nás bude zaujímať, aká je očakávaná početnosť v tabuľke, ak by medzi premennými neexistovala žiadna závislosť. Pre tieto účely si nadefinujeme marginálne početnosti.

```
> ps <- addmargins(ps)
> ps
          podnikatel nezamestnany zamestnany Sum
zenaty            14           15         40  69
slobodny          20           10         21  51
rozvedeny         60            5         11  76
Sum               94           30         72 196
```

Riadkové súčty budeme označovať $n_{i\bullet}$ a stĺpcové súčty ako $n_{\bullet j}$. V našom prípade tak máme riadkové súčty: $n_{1\bullet}$ = 69, $n_{2\bullet}$ = 51, $n_{3\bullet}$ = 76 a stĺpcové súčty $n_{\bullet 1}$ = 94, $n_{\bullet 2}$ = 30, $n_{\bullet 3}$ = 72. Zároveň platí, že celkový počet pozorovaní je n = 196.

Vrátime sa k našej otázke. Ako by vyzerala táto tabuľka, ak by rodinný stav a pracovné zaradenie boli nezávislé? Ak uvažujeme o štatistickej závislosti, potom môžeme vychádzať z definície, že jav A a jav B sú nezávislé, ak platí $P(A \cap B) = P(A)P(B)$. Predstavme si ďalej, že na základe nami nameraných údajov (teraz budeme predpokladať, že ide o náhodný výber) chceme zistiť, aká je pravdepodobnosť, že ak náhodne vyberieme muža, tak bude ženatý (jav A) a zároveň bude zamestnaný (jav B)? Zaujíma nás teda $P(A \cap B)$. Z údajov v tabuľke vieme, že našim najlepším odhadom pravdepodobnosti nastania javu A bude $P(A)$ = 69/196, keďže ženatých mužov je vo vzorke 69 z celkového počtu 196. Podobne vieme zistiť, že najlepším odhadom pravdepodobnosti nastania javu B je

$P(B) = 72/196$. Teraz už vieme, že ak náhodne vyberieme z našej populácie muža, tak pravdepodobnosť, že bude ženatý a zároveň zamestnaný je $P(A \cap B) = P(A)P(B) = (69/196)(72/196)$. Vychádzajúc z definície štatistickej nezávislosti táto pravdepodobnosť platí iba v prípade, ak medzi rodinným stavom a pracovným zaradením neexistuje štatistická závislosť.

Našim cieľom je získať očakávané početnosti v prípade nezávislosti dvoch javov, pričom výraz $P(A \cap B)$ predstavuje pravdepodobnosť a nie početnosť. Jednoduchým prenásobením celkovou početnosťou dostávame výraz $nP(A \cap B) = 196(69/196)(72/196) = (69*72)/196$. Týmto spôsobom môžeme postupovať pre každú jednu z 9 kombinácií v kontingenčnej tabuľke. Formálnejšie môžeme očakávané početnosti (v prípade nezávislosti dvoch premenných) definovať nasledovne:

$$e_{i,j} = \frac{n_{i\bullet} n_{\bullet j}}{n} \tag{5.16}$$

Tieto očakávané početnosti si vieme v programe R vypočítať použitím funkcie `chisq.test()`, ktorú budeme používať aj pri testovaní štatistickej významnosti.

```
> chisq.test(ps, correct=F)$expected
          podnikatel nezamestnany zamestnany Sum
zenaty      33.09184    10.561224   25.34694  69
slobodny    24.45918     7.806122   18.73469  51
rozvedeny   36.44898    11.632653   27.91837  76
Sum         94.00000    30.000000   72.00000 196
```

Následne má testovacia charakteristika nasledujúci tvar (podobne ako pri Pearsonovom Chí-kvadrát teste dobrej zhody):

$$\chi^2_{CS} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}} \tag{5.17}$$

Za predpokladu platnosti nulovej hypotézy sa táto štatistika riadi Chí-kvadrát rozdelením s $(r - 1)(s - 1)$ stupňami voľnosti. Rozhodnutie o hypotéze je nasledovné:

H_0: premenné sú nezávislé H_1: premenné sú závislé	Hypotézu H_0 zamietame, ak $\chi^2_{CS} > \chi^2_{(1-\alpha),(r-1)(s-1)}$

V prípade, ak premenné nadobúdajú iba dva rôzne stavy, máme v kontingenčnej tabuľke $s = 2$ a $r = 2$, odporúčaná testovacia charakteristika má tvar:

$$\chi^2_{CS}\grave{} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(|o_{i,j} - e_{i,j}| - 0.5)^2}{e_{i,j}} \tag{5.18}$$

V oboch prípadoch (*(5.17)* a *(5.18)*) platí, že $e_{i,j} \geq 5$. V prípade väčších tabuliek je dodržanie tejto podmienky často problematické. Dôsledkom je nepresná aproximácia testovacej charakteristiky Chí-kvadrát rozdelením pravdepodobnosti. Napriek tomu, podľa niektorých zdrojov (Finkelstein – Levin, 2001) pre tabuľky väčšie ako 2 x 2 sa odporúča mať v aspoň 80 % prípadoch očakávanú početnosť väčšiu ako 5 alebo nemať očakávanú početnosť menej ako 1 vo viac ako 10 % prípadov. V prípade, ak sa tieto podmienky nevedia splniť, v niektorých prípadoch pomôže spájanie tried (napr. spojenie rozvedených a slobodných mužov do kategórie „nie ženatý“ muž). Tejto operácii hovoríme *binning*.

Pokračujúc v našej prípadovej štúdii, použijeme funkciu `chisq.test(ps)`. Výsledok silne zamieta nulovú hypotézu o nezávislosti. Máme tak dostatok dôkazov, aby sme mohli tvrdiť, že medzi pracovným zaradením a rodinným stavom mužov existuje štatisticky významný vzťah. V ďalších kapitolách sa budeme venovať tomu, akým spôsobom merať nakoľko je tento vzťah „silný“.

```
> chisq.test(ps)

                  Pearson's Chi-squared test

data:  ps
X-squared = 52.3061, df = 9, p-value = 3.956e-08
```

5.1.5 Phi a Cramerov koeficient

V prípade 2 x 2 tabuľky môžeme použiť *phi* (*Φ*) koeficient miery asociácie, ktorý využíva Chí-kvadrát štatistiku χ^2_{CS} z predošlej analýzy. Vychádzajúc zo vzťahu (*(5.17)* a *(5.18)*) môžeme vidieť, že čím sú rozdiely medzi očakávanou a pozorovanou početnosťou väčšie, o to väčšia bude hodnota χ^2_{CS}, a teda o to väčší vzťah by mal existovať medzi dvoma premennými. Absolútna hodnota χ^2_{CS} však môže závisieť aj od počtu pozorovaní. *Phi* koeficient túto skutočnosť využíva a „normuje“ χ^2_{CS} štatistiku:

$$|\phi| = \sqrt{\frac{\chi^2_{CS}}{n}} \qquad (5.19)$$

|*Φ*| koeficient vypočítaný podľa *(5.19)* môže nadobúdať hodnoty v uzavretom intervale od 0 po 1, kde hodnoty bližšie k 1 znamenajú väčšiu mieru závislosti. Vrátime sa k diskusii o „sile“ vzťahu medzi premennými. Howell (2010) uvádza niekoľko zaujímavých príkladov interpretácie sily vzťahu pomocou *Φ* koeficientu. Prvý príklad je od Smith – Glass (1977), rozvinutý v Rosenthal – Rubin (1982). Smith – Glass (1977) interpretujú hodnotu |*Φ*| = 0.32 najprv ako pomerne nízku. Rosenthal – Rubin (1982) však uvádzajú, že ak si

zoberieme 200 pacientov, ktorí buď absolvovali psychoterapeutickú liečbu (100 pacientov) alebo nie (taktiež 100 pacientov) a zároveň buď v ich liečbe došlo alebo nedošlo k zlepšeniu, tak koeficient $|\Phi| = 0.32$ zodpovedá takej 2 x 2 tabuľke, kde 66 pacientov absolvujúcich psychoterapeutickú liečbu dosiahlo zlepšenie svojho zdravotného stavu, kým iba 34 pacientov dosiahlo zlepšenie v skupine pacientov, ktorí túto liečbu neabsolvovali. Tento rozdiel sa z praktického hľadiska javí ako pomerne významný (skoro dvojnásobný rozdiel v počte pacientov so zlepšeným stavom). Preto hodnota koeficientu na úrovni 0.32 v tomto prípade znamená veľmi silnú závislosť.

```
> hrr <- matrix(c(34, 66, 66, 34), nrow = 2)
> rownames(hrr) <- c("psychoterapeuticka", "kontrolna")
> colnames(hrr) <- c("stav sa nezlepsil", "stav sa zlepsil")
> hrr <- addmargins(hrr); hrr
                   stav sa nezlepsil stav sa zlepsil Sum
psychoterapeuticka                34              66 100
kontrolna                         66              34 100
Sum                              100             100 200
> phi <- sqrt(chisq.test(hrr)$statistic[[1]]/200); phi
[1] 0.32
```

Koeficient Φ sa uvádza aj v nasledujúcom tvare (napr. Tkáč, 2001):

$$\phi = \frac{n_{1,1}n_{2,2} - n_{1,2}n_{2,1}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}} \qquad (5.20)$$

V tomto tvare koeficient Φ nadobúda hodnoty v uzavretom intervale od –1 do 1. Z čitateľa vo vzťahu *(5.20)* je zrejmé, že záporné hodnoty bude Φ nadobúdať vtedy, ak súčin hodnôt mimo hlavnej diagonály bude väčší ako súčin hodnôt na hlavnej diagonále. Aby bolo možné uskutočniť zmysluplnú interpretáciu, je vhodné už pri tvorbe kontingenčnej tabuľky na túto skutočnosť brať ohľad. V programe môžeme použiť na výpočet koeficientu Φ podľa vzťahu *(5.20)* funkciu `phi()` z programového balíka `psych`. Zároveň upozorňujeme, že určité rozdiely vo výsledkoch medzi funkciou `phi()` a `chisq.test()` môžu vznikať v dôsledku korekcie (pozri predchádzajúca kapitola), ktorá sa v prípade funkcie `chisq.test()` uskutočňuje pre tabuľky malého rozsahu.

Nasledujúci príklad ilustruje situáciu so štatisticky významnou závislosťou medzi dvoma premennými, kde však sila tejto väzby je pomerne nízka. Údaje sme prevzali z internetovej stránky[27] Úradu práce sociálnych vecí a rodiny za december 2011. Zaujímal nás počet nezamestnaných v Bratislavskom a Prešovskom kraji a ich pohlavie, teda závislosť medzi týmito dvoma premennými. V prípade existencie závislosti môžeme výsledok

[27] http://www.upsvar.sk/statistiky/nezamestnanost-mesacne-statistiky/2011.html?page_id=31010

interpretovať tak, že z hľadiska nezamestnanosti nie je jedno, akého pohlavia a v akom kraji sa obyvateľ nachádza.

```
> nez <- matrix(c(10033, 9384, 37768, 44112), ncol = 2)
> colnames(nez) <- c("Bratislava", "Presov")
> rownames(nez) <- c("Zeny", "Muzi")
> nez <- addmargins(nez); nez
     Bratislava Presov    Sum
Zeny      10033  37768  47801
Muzi       9384  44112  53496
Sum       19417  81880 101297
-----------------------------------------------------------------
> chisq.test(nez)

                Pearson's Chi-squared test

data:  nez
X-squared = 193.6552, df = 4, p-value < 2.2e-16
-----------------------------------------------------------------
> phi <- sqrt(chisq.test(nez)$statistic[[1]]/101297); phi
[1] 0.04372364
```

Štatistický vzťah medzi premennými sa nám ukázal byť významný, avšak hodnota Φ koeficientu je iba 0.04. Aby sme si urobili určitú predstavu o sile vzťahu, uvažujme o nasledujúcej tabuľke.

```
> teo <- matrix(c(52, 48, 48, 52), ncol = 2)
> colnames(teo) <- c("Bratislava", "Presov")
> rownames(teo) <- c("Zeny", "Muzi")
> teo <- addmargins(teo); teo
     Bratislava Presov Sum
Zeny         52     48 100
Muzi         48     52 100
Sum         100    100 200
-----------------------------------------------------------------
> chisq.test(nez)

                Pearson's Chi-squared test

data:  nez
X-squared = 193.6552, df = 4, p-value < 2.2e-16
-----------------------------------------------------------------
> phi <- sqrt(chisq.test(teo)$statistic[[1]]/200); phi
[1] 0.04
```

Počet nezamestnaných obyvateľov sme normovali na 200 tak, aby bol počet žien 100 a počet nezamestnaných v Bratislavskom kraji tiež 100. Zároveň si všimnime, že Φ koeficient je veľmi blízky hodnote, ktorá nám vyšla zo skutočne nameraných údajov ($|\Phi| = 0.04$). Nakoľko silný je vzťah medzi premennými? Ak je nezamestnaný obyvateľ žena, potom v Bratislavskom kraji je pravdepodobnosť, že nebude zamestnaná 0.52 (52/100), kým

v Prešovskom kraji 0.48 (48/100). Z praktického hľadiska sa to nejaví ako veľký rozdiel. Preto aj napriek štatistickej významnosti sa táto závislosť javí ako veľmi slabá.

Rozšírením Φ koeficientu na tabuľky väčšieho rozmeru ako 2 x 2 sa dostaneme ku Cramerovmu-V koeficientu. Znovu vychádzame z χ^2_{CS} štatistiky:

$$V = \sqrt{\frac{\chi^2_{CS}}{n(k-1)}} \qquad (5.21)$$

kde $k = min\{s, r\}$. Pre $k = 2$ dostávame rovnaký koeficient ako v prípade Φ koeficientu.

5.1.6 Testovanie rovnosti dvoch korelačných koeficientov: nezávislé skupiny

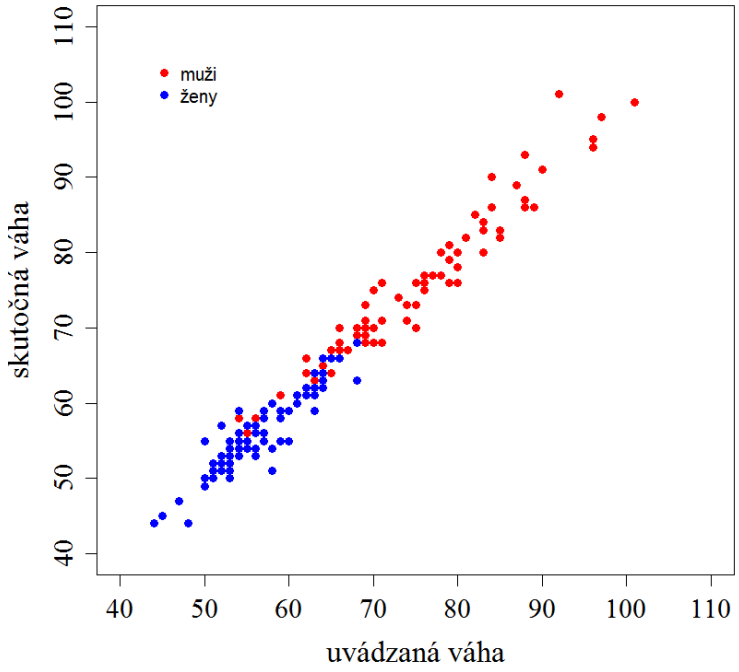
Podobne ako sme overovali zhodu stredných hodnôt, rozptylov, rozdelení alebo rozdelenia poradí, môžeme porovnávať aj rozdiely dvoch korelačných koeficientov. Zaujímať nás pritom bude najpoužívanejší Pearsonov korelačný koeficient. Vychádzajme najprv zo situácie, kde nás zaujímajú korelačné koeficienty dvoch nezávislých skupín. Pomôžeme si príkladom o výške a váhe mužov a žien. Prvou skupinou sú muži a druhou sú ženy. Zaujíma nás korelácia medzi skutočnou váhou a váhou, ktorú respondenti o sebe uviedli. Vyššia hodnota korelačného koeficientu naznačuje väčšiu zhodu medzi uvádzanou a skutočnou váhou. V prípade, ak je korelácia nízka, môže to naznačovať buď to, že respondenti nemajú prehľad o svojej váhe alebo, že ak ide o ich váhu, majú tendenciu nehovoriť pravdu. V tomto príklade použijeme databázu `Davis` z programového balíka `car`. Keďže sa v nej nachádzajú jednak extrémne hodnoty a zároveň chýbajúce údaje (označené ako NA z angl. *Not Available*), databázu si najprv upravíme. Použijeme pritom časť kódu, ktorý sme už prezentovali vyššie.

```
> attach(Davis)
> davis_new <- na.omit(Davis); detach(Davis)
> men <- davis_new[davis_new$sex == "M",]
> names(men) <- c("pohlavie", "vaha", "vyska", "vaha_s",
  "vyska_s")
> women <- davis_new[davis_new$sex == "F",]
> names(women) = c("pohlavie", "vaha", "vyska", "vaha_s",
  "vyska_s")
> index <- (1:length(men$vaha))[men$vaha %in%
  hampel_identifier(men$vaha)]
> index <- c(index, (1:length(men$vyska))[men$vyska %in%
  hampel_identifier(men$vyska)])
> men = men[-index,]
> index <- (1:length(women$vaha))[women$vaha %in%
  hampel_identifier(women$vaha)]
```

```
> index <- c(index, (1:length(women$vyska))[women$vyska %in%
  hampel_identifier(women$vyska)])
> women <- women[-index,]
```

Vypočítame si koreláciu medzi skutočnou a uvádzanou váhou zvlášť pre vzorku mužov a zvlášť pre vzorku žien. Túto závislosť si znázorníme na *x-y* grafe (pozri Obrázok 5.5). Obe korelácie sú štatisticky významné a hodnoty 0.97 (muži) a 0.92 (ženy) naznačujú veľkú zhodu. Tento obrázok (Obrázok 5.5) taktiež naznačuje veľmi obdobný priebeh závislosti. Predsa však u mužov sa javí byť zhoda väčšia. Otázkou je, či ide o štatisticky významný rozdiel.

```
> cor.test(men$vaha, men$vaha_s)

            Pearson's product-moment correlation

data:  men$vaha and men$vaha_s
t = 36.9779, df = 76, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9583599 0.9829479
sample estimates:
      cor
0.9733167
------------------------------------------------------------------
> cor.test(women$vaha, women$vaha_s)

            Pearson's product-moment correlation

data:  women$vaha and women$vaha_s
t = 21.8008, df = 82, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8841664 0.9498349
sample estimates:
      cor
0.9235017
```



Obrázok 5.5: *x-y* graf závislosti skutočnej a uvádzanej váhy mužov a žien

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> plot(men$vaha_s ~ men$vaha, xlab = "uvádzaná váha", xlim =
  c(40, 110), ylim = c(40, 110), ylab = "skutočná váha", col =
  "red", pch = 19, cex.lab = 1.7, cex.axis = 1.5, family =
  "serif")
> lines(women$vaha_s ~ women$vaha, type = "p", pch = 19, col =
  "blue")
> legend("topleft", legend = c("muži", "ženy"), pch = 19, col =
  c("red", "blue"), inset = 0.08, bty = "n")
```

Pre potreby tvorby testovacej charakteristiky si potrebujeme najprv korelačné koeficienty transformovať. Použijeme nasledujúcu transformáciu (tzv. Fisherovu transformáciu korelačného koeficientu, spracované podľa autora Clark-Carter, 2009):

$$r'_{x,y} = \frac{1}{2}\ln\left(\frac{1+r_{x,y}}{1-r_{x,y}}\right) \qquad (5.22)$$

Spätná transformácia je:

$$r_{x,y} = \frac{e^{2r'_{x,y}} - 1}{e^{2r'_{x,y}} + 1} \qquad (5.23)$$

Ak si označíme korelačný koeficient v prvej skupine ako $r`_{x,y}$ a v druhej ako $r``_{x,y}$, pričom n_1 a n_2 predstavujú početnosti prvej a druhej skupiny, potom testovacia charakteristika má tvar:

$$z_{\rho\rho} = \frac{r_{x,y}^{`} - r_{x,y}^{``}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \qquad (5.24)$$

kde testovacia charakteristika $z_{\rho\rho}$ sa riadi normovaným normálnym rozdelením. Rozhodnutie o hypotéze vykonáme nasledovne:

$H_0: \rho_{x,y}^{`} = \rho_{x,y}^{``}$ $H_1: \rho_{x,y}^{`} \neq \rho_{x,y}^{``}$	Hypotézu H_0 zamietame, ak $\lvert Z_{\rho\rho}\rvert > \lvert z_{(\alpha/2)}\rvert$
$H_0: \rho_{x,y}^{`} \leq \rho_{x,y}^{``}$ $H_1: \rho_{x,y}^{`} > \rho_{x,y}^{``}$	Hypotézu H_0 zamietame, ak $Z_{\rho\rho} > z_{(1-\alpha)}$
$H_0: \rho_{x,y}^{`} \geq \rho_{x,y}^{``}$ $H_1: \rho_{x,y}^{`} < \rho_{x,y}^{``}$	Hypotézu H_0 zamietame, ak $Z_{\rho\rho} < z_{(\alpha)}$

kde $z_{(\alpha/2)}$, $z_{(1-\alpha)}$ a $z_{(\alpha)}$ sú kvantilmi normovaného normálneho rozdelenia pravdepodobnosti. Výpočet môžeme previesť v programe R manuálne alebo si vytvoríme jednoduchú funkciu, ktorá nám podľa zvolenej alternatívnej hypotézy vráti príslušnú *p*-hodnotu. V tejto funkcii je potrebné zadať vstupné premenné tak, aby v nich nevznikali chýbajúce údaje. Výsledok potvrdzuje, že na danej vzorke je korelácia u mužov štatisticky významne vyššia ako korelácia u žien. Jednou z interpretácii tak môže byť, že muži majú menšiu tendenciu zavádzať pri udávaní svojej hmotnosti ako ženy.

```
> rho_indcomp <- function(var11, var12, var21, var22,
  alternative = c("two.sided", "greater", "less")) {
+ a <- cor(var11, var12)
+ b <- cor(var21, var22)
+ c <- 0.5*log((1 + a)/(1 - a))
+ d <- 0.5*log((1 + b)/(1 - b))
+ zrho <- (c - d)/(sqrt(1/(length(var11)-3) +1/(length(var21)-
  3)))
+ if (alternative == "two.sided") {
+ cat("two.sided p-value is")
+ print((1-pnorm(abs(zrho)))*2)
+ }
+ if (alternative == "greater") {
+ cat("p-value for r1 > r2 alternative is")
+ print(1-pnorm(zrho))
+ }
+ if (alternative == "less") {
+ cat("p-value for r1 < r2 alternative is")
+ print(pnorm(zrho))
+ }
+ }
------------------------------------------------------------------
> rho_indcomp(men$vaha, men$vaha_s, women$vaha, women$vaha_s,
  alternative = "greater")
p-value for r1 > r2 alternative is[1] 0.0003812558
```

Transformáciu vo vzťahu *(5.22)* je možné použiť aj pri testovaní hypotéz o významnosti korelačného koeficientu oproti ľubovoľnej konštante (nie len H_0: $\rho_{x,y} = 0$). Testovacia charakteristika bude mať potom tvar:

$$z_\rho = \frac{r_{x,y} - \rho_{x,y}}{\sqrt{\frac{1}{n-3}}} \qquad (5.25)$$

O hypotézach rozhodujeme obdobne ako v predošlom prípade, keďže aj testovacia charakteristika z_ρ sa riadi normovaným normálnym rozdelením.

5.1.7 Testovanie rovnosti dvoch korelačných koeficientov: závislé skupiny

Uvažujme najprv o nasledujúcej situácii. Majme náhodný výber, ktorého realizácie si označíme ako X_i, Y_i a Z_i, kde i = 1, 2, ..., n. Zaujíma nás korelácia (v zmysle Pearsonovho korelačného koeficientu) medzi $\rho_{x,y}$ a $\rho_{x,z}$, ktoré chceme navzájom porovnať. Tento prípad je špecifický tým, že v oboch korelačných koeficientoch vystupuje náhodná premenná X. Ak by sme mali jednu skupinu zákazníkov a X by predstavovalo ich spokojnosť, Y počet nákupov a Z priemerný objem nákupov, potom by nás mohlo zaujímať, či je väčšia korelácia medzi spokojnosťou a počtom nákupov $\rho_{x,y}$, alebo spokojnosťou a priemerným objemom nákupov $\rho_{x,z}$. Vznikla by tak situácia, kde tieto dva korelačné koeficienty sú špecifickým spôsobom závislé.

Na základe vzorky pozorovaní (X_i, Y_i, Z_i) ktoré tvoria usporiadanú trojicu odhadneme korelačné koeficienty pomocou Pearsonovho korelačného koeficientu $r_{x,y}$, $r_{x,z}$. Pre potreby testovacej charakteristiky si najprv vypočítame determinant korelačnej matice (spracované podľa autora Clark-Carter, 2009):

$$|C| = \left(1 - \left(r_{x,y}\right)^2 - \left(r_{x,z}\right)^2 - \left(r_{z,y}\right)^2\right) + \left(2 r_{x,y} r_{x,z} r_{z,y}\right) \qquad (5.26)$$

a priemernú koreláciu z porovnávaných korelácií:

$$\bar{r} = \frac{\left(r_{x,y} + r_{x,z}\right)}{2} \qquad (5.27)$$

Testovacia charakteristika má potom tvar:

$$t_{d-\rho} = \left(r_{x,y} - r_{x,z}\right) \sqrt{\frac{(n-1)(1+r_{y,z})}{\left(2|C|\left(\frac{n-1}{n-3}\right)\right) + \left(\bar{r}^2\left(1 - r_{y,z}\right)^3\right)}} \qquad (5.28)$$

Charakteristika $t_{d\text{-}\rho}$ sa riadi Studentovým t rozdelením pravdepodobnosti s $(n - 3)$ stupňami voľnosti. Rozhodnutie o hypotéze tak vykonáme nasledovne:

H_0: $\rho_{x,y} = \rho_{x,z}$	Hypotézu H_0 zamietame, ak $\lvert t_{d\text{-}\rho}\rvert > \lvert t_{(n-3),(\alpha/2)}\rvert$
H_1: $\rho_{x,y} \neq \rho_{x,z}$	
H_0: $\rho_{x,y} \leq \rho_{x,z}$	Hypotézu H_0 zamietame, ak $t_{d\text{-}\rho} > t_{(n-3),(1-\alpha)}$
H_1: $\rho_{x,y} > \rho_{x,z}$	
H_0: $\rho_{x,y} \geq \rho_{x,z}$	Hypotézu H_0 zamietame, ak $t_{d\text{-}\rho} < t_{(n-3),(\alpha)}$
H_1: $\rho_{x,y} < \rho_{x,z}$	

kde $t_{(n-3),(\alpha/2)}$, $t_{(n-3),(1-\alpha)}$ a $t_{(n-3),(\alpha)}$ sú kvantily Studentovho t rozdelenia pravdepodobnosti.

Pri druhej situácii už uvažujme o náhodnom výbere, ktorého realizácie si označíme ako X_i, Y_i, Z_i a W_i, kde $i = 1, 2, ..., n$. Zaujíma nás korelácia (v zmysle Pearsonovho korelačného koeficientu) medzi $\rho_{x,y}$ a $\rho_{z,w}$, ktoré chceme navzájom porovnať. V príkladoch o výške a váhe by táto situácia nastala vtedy, ak by sme chceli porovnať koreláciu medzi výškou a váhou mužov s koreláciou medzi uvádzanou výškou a uvádzanou váhou tých istých mužov. Tak isto ide o závislé skupiny.

Postup výpočtu testovacej charakteristiky je pomerne prácny, avšak podobne ako v predošlom prípade, pomerne ľahko naprogramovateľný. Definujme si najprv priemernú koreláciu porovnávaných korelácií:

$$\bar{r} = \frac{\left(r_{x,y} + r_{z,w}\right)}{2} \tag{5.29}$$

Ďalej si definujme nasledujúcu premennú (výraz v zátvorke je reálne číslo):

$$\mathrm{cov}_{x,y;z,w} = \frac{1}{2}\left(\begin{array}{l}\left(r_{x,z} - \bar{r}r_{y,z}\right)\left(r_{y,w} - \bar{r}r_{y,z}\right) + \left(r_{x,w} - \bar{r}r_{x,z}\right)\left(r_{y,z} - \bar{r}r_{x,z}\right) + \\ \left(r_{x,z} - \bar{r}r_{x,w}\right)\left(r_{y,w} - \bar{r}r_{x,w}\right) + \left(r_{x,w} - \bar{r}r_{y,w}\right)\left(r_{y,z} - \bar{r}r_{y,w}\right)\end{array}\right) \tag{5.30}$$

Ďalej:

$$\bar{s}_{x,y;z,w} = \frac{\mathrm{cov}_{x,y;z,w}}{\left(1 - \bar{r}^2\right)^2} \tag{5.31}$$

Testovacia charakteristika má potom tvar:

$$z_{x,y;z,w} = \left(r'_{x,y} - r'_{z,w}\right)\left(\frac{\sqrt{n-3}}{\sqrt{2 - \left(2\bar{s}_{x,y;z,w}\right)}}\right) \tag{5.32}$$

kde $r'_{x,y}$ a $r'_{z,w}$ sú korelačné koeficienty po Fisherovej transformácii korelačných koeficientov. Testovacia charakteristika $z_{x,y;z,w}$ sa riadi normovaným normálnym rozdelením pravdepodobnosti. Rozhodnutie o hypotéze je nasledovné:

H_0: $\rho_{x,y} = \rho_{z,w}$ H_1: $\rho_{x,y} \neq \rho_{z,w}$	Hypotézu H_0 zamietame, ak $\lvert z_{x,y;z,w}\rvert > \lvert z_{(\alpha/2)}\rvert$
H_0: $\rho_{x,y} \leq \rho_{z,w}$ H_1: $\rho_{x,y} > \rho_{z,w}$	Hypotézu H_0 zamietame, ak $z_{x,y;z,w} > z_{(1-\alpha)}$
H_0: $\rho_{x,y} \geq \rho_{z,w}$ H_1: $\rho_{x,y} < \rho_{z,w}$	Hypotézu H_0 zamietame, ak $z_{x,y;z,w} < z_{(\alpha)}$

6 Viacrozmerné metódy

Podobne, ako môžeme v prípade pravdepodobnostných rozdelení hovoriť o jednorozmerných rozdeleniach a viacrozmerných rozdeleniach, môžeme taktiež uvažovať aj o viacrozmerných metódach induktívnej štatistiky. Kým väčšina metód, ktoré sme popisovali v predchádzajúcich častiach bola založená na práci s náhodnými premennými, viacrozmerné metódy spravidla využívajú náhodné vektory a ich lineárne kombinácie.

Na tomto mieste by sme chceli pripomenúť, že nie každá analýza s viac ako jedným objektom je viacrozmernou analýzou. Ak máme napríklad náhodné premenné X_1, X_2, ..., X_k pre $k \in \mathbb{N}$ o ktorých predpokladáme, že majú to isté rozdelenie pravdepodobnosti a sú navzájom nezávislé, môže sa zdať, že ide o viacrozmernú analýzu – máme predsa *k* objektov. Napriek tomu môže ísť o jednorozmernú analýzu, pokiaľ tieto náhodné premenné využijeme napríklad pre odhad parametrov ich spoločného jednorozmerného pravdepodobnostného rozdelenia.

Metódy popísané v predchádzajúcich odsekoch umožňujú skúmanie veľmi veľkej množiny rôznych problémov – ukázali sme si aplikácie testov na stredné hodnoty, rozptyly, ako aj testovanie závislostí medzi premennými. Môže vyvstávať otázka, aký zmysel má rozšírenie týchto metód o viacrozmernú analýzu a aký je teda jej očakávaný prínos.

Viacrozmerné metódy slúžia podobne ako v prípade jednorozmerných pre deskriptívnu a induktívnu štatistiku. V prípade deskriptívnych metód spravidla uvažujeme a popisujeme vzťahy medzi zložkami tvoriacimi náhodný vektor. V prípade induktívnych metód často realizujeme podobné testy ako v jednorozmernej štatistike, ale tentoraz zohľadňujeme aj vzájomné vzťahy testovaných premenných. V neposlednom rade slúžia metódy viacrozmernej štatistiky aj na zjednodušenie (simplifikáciu) popisovaného súboru – v tomto prípade sa snažíme väčšie množstvo premenných nahradiť menším počtom objektov (komponentov, faktorov), ktoré zachovávajú informáciu obsiahnutú v pôvodných dátach.

V praxi dochádza často k situácii, keď pracujeme so súbormi dát, charakterizujúcimi niekoľko atribútov, resp. vlastností meraných u tých istých subjektov. Príkladom môže byť napríklad databáza žiadateľov o pôžičku, u ktorých by sme mohli evidovať pohlavie, vek, mesačný príjem, vysokoškolské vzdelanie a výšku úspor. Pri takýchto dátach je zrejmé, že medzi premennými môžu existovať vzájomné závislosti – výška mesačného príjmu môže napríklad súvisieť s vekom a dosiahnutým vzdelaním. Ak testujeme na tomto súbore dát niekoľko hypotéz (a zaujíma nás viac než jedna), vzniká tým potenciálne priestor na využitie viacrozmerných štatistických metód.

6.1 Odhad parametrov viacrozmerného normálneho rozdelenia

Podobne ako v prípade jednorozmerného normálneho rozdelenia pravdepodobnosti je v štatistike veľmi dôležitým viacrozmerné normálne rozdelenie. Ak má náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ pre $n \in \mathbb{N}$, ktorého zložkami sú náhodné premenné združené normálne rozdelenie pravdepodobnosti, potom je možné združenú hustotu viacrozmerného normálneho rozdelenia zapísať nasledovne:

$$f_{\mathbf{X}}(x_1, x_2, ..., x_n) = \frac{1}{\sqrt{(2\pi)^n \mid \boldsymbol{\Sigma}_{\mathbf{X}} \mid}} \exp\left[-\frac{1}{2} \left(\mathbf{X} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \left(\mathbf{X} - E(\mathbf{X})\right) \right] \qquad (6.1)$$

kde $\boldsymbol{\Sigma}_{\mathbf{X}}$ predstavuje variančno-kovariančnú maticu, $|\boldsymbol{\Sigma}_{\mathbf{X}}|$ predstavuje determinant variančno-kovariančnej matice a $(\mathbf{X} - E(\mathbf{X}))^T$ transponovanú maticu k $(\mathbf{X} - E(\mathbf{X}))$.

Pripomeňme, že parametrami viacrozmerného normálneho rozdelenia sú vektor stredných hodnôt $E(\mathbf{X})$, ktorého prvkami sú stredné hodnoty náhodných premenných $X_1, X_2, \ldots, X_n$ a variančno-kovariančná matica $\boldsymbol{\Sigma}_{\mathbf{X}}$, ktorej prvkami sú kovariancie medzi týmito premennými. Vzhľadom na vlastnosti kovariancie sa na hlavnej diagonále tejto matice nachádzajú rozptyly premenných $X_1, X_2, \ldots, X_n$.

Otázkou je, ako na základe vzorky odhadnúť parametre viacrozmerného normálneho rozdelenia. V prvom rade si musíme uvedomiť, že vzorku v našom prípade predstavuje $m \in \mathbb{N}$ náhodných vektorov $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m$, kde každý z nich pozostáva z n náhodných premenných (každý z náhodných vektorov $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m$ má n prvkov). Odhad parametrov združeného normálneho rozdelenia je možné uskutočniť napríklad na základe odhadu pomocou metódy maximálnej vierohodnosti.

Ak sú náhodné vektory $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m$ nezávislé, potom môžeme funkciu vierohodnosti zostrojiť veľmi jednoducho – bude predstavovať súčin hustôt pravdepodobnosti viacrozmerného normálneho rozdelenia. Upozornime ešte, že nezávislosť požadujeme medzi náhodnými vektormi, nie medzi ich zložkami – tie korelované byť môžu.

Ak dosadíme i-té pozorovanie (pre $i \in \{1, 2, \ldots m\}$) do funkcie hustoty, dostaneme:

$$f(\mathbf{X}_i) = \frac{1}{\sqrt{(2\pi)^n \mid \boldsymbol{\Sigma} \mid}} \exp\left[-\frac{1}{2} \left(\mathbf{X}_i - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{X}_i - E(\mathbf{X})\right) \right] \qquad (6.2)$$

V poslednom výraze predstavuje $E(\mathbf{X})$ vektor stredných hodnôt a $\boldsymbol{\Sigma}$ je variančno-kovariančná matica. Keďže predpokladáme, že náhodné vektory $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m$ majú to isté rozdelenie pravdepodobnosti, vektor $E(\mathbf{X})$ a matica $\boldsymbol{\Sigma}$ je pre ne spoločná. Predpokladáme teda, že platí:

$$\mathbf{X}_i \sim N_n(E(\mathbf{X}), \boldsymbol{\Sigma}) \tag{6.3}$$

Otázkou je, ako na základe vzorky odhadnúť parametre tohto spoločného rozdelenia pravdepodobnosti. Najčastejšie sa v tomto prípade využíva metóda maximálnej vierohodnosti (angl. *maximum likelihood estimation*, MLE). Princíp metódy je pomerne jednoduchý – definuje sa tzv. funkcia vierohodnosti, ktorá závisí na údajoch zo vzorky, ako aj na neznámych parametroch hľadaného rozdelenia. Keďže jedinou neznámou v takto definovanej funkcii vierohodnosti sú parametre, v ďalšom kroku sa využívajú optimalizačné metódy, ktoré hľadajú také ich hodnoty, pri ktorých funkcia vierohodnosti nadobúda maximum.

Logiku tohto postupu je možné lepšie pochopiť, ak dodáme, že za funkciu vierohodnosti by sme pri diskrétnych premenných mohli použiť funkciu združenej pravdepodobnosti. Optimalizačná úloha spomínaná vyššie by potom zodpovedala maximalizovaniu pravdepodobnosti – hľadali by sme také hodnoty parametrov, pri ktorých by bola pravdepodobnosť výskytu údajov obsiahnutých vo vzorke najväčšia. V istom zmysle by takto získané hodnoty mali „najväčšiu pravdepodobnosť“ toho, že údaje v skúmanej vzorke majú im zodpovedajúce rozdelenie. Navyše, keďže väčšinou predpokladáme, že údaje, ktoré máme vo vzorke, boli do nej vybrané ako nezávislé, predstavuje združenú pravdepodobnosť súčin individuálnych pravdepodobností pre jednotlivé prvky vzorky.

Ak skúmame viacrozmerné normálne rozdelenie, bolo by možné namietať, že tu nemáme možnosť pracovať priamo s funkciou pravdepodobnosti, ktorá pre spojité rozdelenia nie je definovaná. Máme však k dispozícii analogickú funkciu, ktorou je hustota pravdepodobnosti. V jej prípade síce už nemôžeme hovoriť o pravdepodobnostiach, ale základná myšlienka postupu zostáva zachovaná.

Funkciu vierohodnosti pre viacrozmerné normálne rozdelenie môžeme definovať nasledovne (Rencher, 2002, s. 110):

$$L(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m) = \prod_{i=1}^{m} f(\mathbf{X}_i) \tag{6.4}$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^n \mid \boldsymbol{\Sigma} \mid}} \exp\left[-\frac{1}{2} \left(\mathbf{X}_i - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{X}_i - E(\mathbf{X})\right) \right] \tag{6.5}$$

$$= \frac{1}{\sqrt{(2\pi)^{nm} \mid \boldsymbol{\Sigma} \mid^m}} \exp\left(-\frac{1}{2} \sum_{i=1}^{m} \left[\left(\mathbf{X}_i - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{X}_i - E(\mathbf{X})\right) \right] \right) \tag{6.6}$$

Definujme vektor $\bar{\mathbf{X}}$ ako vektor priemerov náhodných premenných:

$$\overline{\mathbf{X}} = \sum_{i=1}^{m} \mathbf{X}_i \tag{6.7}$$

Exponent vo vzťahu *(6.6)* môžeme potom upraviť aj nasledovne:

$$-\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - E(\mathbf{X})\right)\right] \tag{6.8}$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - E(\mathbf{X}) + \overline{\mathbf{X}} - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - E(\mathbf{X}) + \overline{\mathbf{X}} - \overline{\mathbf{X}}\right)\right] \tag{6.9}$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - E(\mathbf{X})\right)\right] \tag{6.10}$$

Predposledná úprava predstavovala len pripočítanie nulového vektora $\bar{\mathbf{X}}$ - $\bar{\mathbf{X}}$. Výrazy v zátvorkách môžeme ďalej roznásobiť, čím dostaneme:

$$= -\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - E(\mathbf{X})\right) + \left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - E(\mathbf{X})\right)\right] \tag{6.11}$$

$$\begin{aligned} = -\frac{1}{2}\sum_{i=1}^{m}\Big[&\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right) + \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right) \\ &+ \left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right) + \left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)\Big] \end{aligned} \tag{6.12}$$

$$\begin{aligned} = &-\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right] - \frac{1}{2}\left[\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T\right]\boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right) \\ &-\frac{1}{2}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right] - \frac{1}{2}\sum_{i=1}^{m}\left[\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)\right] \end{aligned} \tag{6.13}$$

V poslednom výraze dostávame v druhom a treťom sčítanci výraz:

$$\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right) \tag{6.14}$$

pre ktorý ale platí:

$$\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right) = \sum_{i=1}^{m}\mathbf{X}_i - m\overline{\mathbf{X}} = \sum_{i=1}^{m}\mathbf{X}_i - m\frac{1}{m}\sum_{i=1}^{m}\mathbf{X}_i = \sum_{i=1}^{m}\mathbf{X}_i - \sum_{i=1}^{m}\mathbf{X}_i = \mathbf{0} \tag{6.15}$$

Platí teda analógia k jednorozmernej štatistike, v ktorej súčet odchýlok od priemeru bol nulový – v našom prípade súčet odchýlok od vektora priemerov je nulový vektor. Po zohľadnení tejto skutočnosti pre exponent vo vzťahu *(6.13)* dostávame:

$$= -\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right] - \frac{1}{2}\left[\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T\right]\boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)$$
$$-\frac{1}{2}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right] - \frac{1}{2}\sum_{i=1}^{m}\left[\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)\right] \quad (6.16)$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right] - \frac{m}{2}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right) \quad (6.17)$$

Vráťme sa ale teraz k funkcii vierohodnosti *(6.6)*. Tú po dosadení vieme napísať ako

$$\frac{1}{\sqrt{(2\pi)^{nm} \mid \boldsymbol{\Sigma} \mid^m}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\left[\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right] - \frac{m}{2}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right)\right) \quad (6.18)$$

Daný výraz by sme chceli maximalizovať. Uvedomme si niekoľko skutočností. V prvom rade ak by nás zaujímal odhad strednej hodnoty, stačí nám skúmať len exponent, keďže inde sa nám stredná hodnota nevyskytuje. Ďalej, samotný exponent nemôže byť nikdy kladný, keďže zo vzorca *(6.5)* je vidieť, že exponent tvorí kvadratickú formu, založenú na variančno-kovariančnej matici, ktorá je pozitívne definitná. Ak maximalizujeme funkciu vierohodnosti, hľadáme vhodné hodnoty parametrov $E(\mathbf{X})$ a $\boldsymbol{\Sigma}$. Výrazy v exponente obsahujúce $\mathbf{X}_i$ a $\overline{\mathbf{X}}$ sú fixne dané – vypočítavame ich zo vzorky. Kľúčový je preto posledný výraz v exponente. Zrejme ak by platilo:

$$\overline{\mathbf{X}} = E(\mathbf{X}) \quad (6.19)$$

bol by posledný výraz v exponente nulový. Odhadom strednej hodnoty viacrozmerného normálneho rozdelenia metódou maximálnej vierohodnosti je preto vektor priemerov.

Podobne by bolo možné ukázať, že odhadom variančno-kovariančnej matice je matica, ktorú môžeme zo vzorky vypočítať nasledovne:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \quad (6.20)$$

Podobne, ako tomu bolo pri odhade rozptylu pri jednorozmernej štatistike, je tento odhad skreslený. Neskresleným odhadom je:

$$\frac{m}{m-1}\hat{\boldsymbol{\Sigma}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \quad (6.21)$$

6.2 Testovanie viacrozmernej normality

V predchádzajúcej časti sme si ukázali, ako je pomocou metódy maximálnej vierohodnosti možné odhadnúť parametre viacrozmerného normálneho rozdelenia. Odhad týchto parametrov však predpokladá, že náhodné vektory, ktoré skúmame, naozaj majú viacrozmerné normálne rozdelenie. Ak však vychádzame z empirických hodnôt, u ktorých nepoznáme ich pravdepodobnostné rozdelenie, môže aj vyvstať otázka o opodstatnenosti tohto predpokladu. Ak skúmané náhodné vektory nemajú združené normálne rozdelenie, potom aj odhady parametrov môžu byť skreslené a nekonzistentné.

Testovanie predpokladu normality v prípade viacrozmerných rozdelení je viac problematické, ako tomu je v jednorozmernom prípade. Pripomeňme, že v prípade, ak by sme chceli testovať zhodu empirického rozdelenia s teoretickým v prípade náhodných premenných, máme na to viacero testov, ktoré môžeme využiť. Jedným zo základných testov je test dobrej zhody, v ktorom porovnávame teoretické rozdelenie s empirickým pomocou testovacej štatistiky, ktorá má χ^2 rozdelenie pravdepodobnosti.

Ako uvádza Rencher (2002), v prípade spojitej náhodnej premennej by sme mohli postupovať tak, že by sme rozdelili interval hodnôt, ktoré náhodná premenná nadobúda na dizjunktné intervaly a porovnať relatívne početnosti s očakávanými pravdepodobnosťami dané očakávaným rozdelením. Pre jednoduchosť si predstavme, že by sme rozdelili rozpätie hodnôt, ktoré náhodná premenná nadobúda na desať častí.

Ak by sme sa pokúsili postupovať analogicky pri náhodných vektoroch, rýchlo zistíme, že situácia sa trochu komplikuje. Ak by sme mali náhodný vektor pozostávajúci z dvoch náhodných premenných, v takom prípade by delenie každej z nich na desať častí viedlo k 100 rôznym dizjunktným množinám ich kombinácií. Predstaviť si to môžeme tak, ako keby sme štvorec o rozmere 10 x 10 vyskladali z dlaždíc o rozmere 1 x 1 – dokopy by sme spotrebovali 100 dlaždíc. V prípade náhodného vektora pozostávajúceho z troch náhodných premenných je analógiou vyplnenie kocky o stranách veľkosti 10 malými kockami o veľkosti 1 x 1 x 1. Potrebujeme tak 10 x 10 x 10 = 1000 rôznych množín (malých kociek).

Bežné testy na dobrú zhodu vyžadujú, aby sme na každej podmnožine porovnali relatívne početnosti a očakávané pravdepodobnosti. Pre veľký počet rôznych kombinácií (v našom prípade ich je 10^k, kde $k \in \mathbb{N}$ predstavuje počet náhodných premenných v náhodnom vektore) je dosť pravdepodobné, že pri empirickej analýze, v mnohých z nich nebudeme mať žiadne pozorovania. Napríklad pri spomínaných troch náhodných premenných je tisíc rôznych

podmnožín – ak by sme v každej z nich chceli mať aspoň jedno pozorovanie, potrebujeme na to vzorku s minimálne 1000 pozorovaniami.

Testy na overenie viacrozmernej normality preto väčšinou vychádzajú z iných prístupov. Jedným z najrozšírenejších testov je tzv. Mardiov test, ktorý využíva viacrozmernú alternatívu šikmosti a špicatosti.

Mardia (1970) sa považuje za jedného z prvých autorov, ktorý sa zaoberal problematikou viacrozmernej šikmosti a špicatosti. Definoval ich nasledovne:

$$\hat{\beta}_{1,k} = \frac{1}{n^2} \sum_{i,j=1}^{n} \left[\left(\mathbf{X}_i - \overline{\mathbf{X}} \right)^T \hat{\mathbf{\Sigma}}^{-1} \left(\mathbf{X}_j - \overline{\mathbf{X}} \right) \right]^3 \qquad (6.22)$$

$$\hat{\beta}_{2,k} = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\mathbf{X}_i - \overline{\mathbf{X}} \right)^T \hat{\mathbf{\Sigma}}^{-1} \left(\mathbf{X}_i - \overline{\mathbf{X}} \right) \right]^2$$

pre $n \in \mathbb{N}$ označujúce veľkosť vzorky, $\mathbf{X}_i$ jednotlivé vektory pozorované vo vzorke a odhad variančno-kovariančnej matice $\hat{\mathbf{\Sigma}}$, ktorý sme si popísali v časti o odhade parametrov viacrozmerného normálneho rozdelenia metódou maximálnej vierohodnosti.

Mardia (1970) ukázal, že ak $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ majú viacrozmerné normálne rozdelenie, potom má výraz:

$$\frac{n\hat{\beta}_{1,k}}{6} = \frac{1}{6n} \sum_{i,j=1}^{n} \left[\left(\mathbf{X}_i - \overline{\mathbf{X}} \right)^T \hat{\mathbf{\Sigma}}^{-1} \left(\mathbf{X}_j - \overline{\mathbf{X}} \right) \right]^3 \qquad (6.23)$$

pravdepodobnostné rozdelenie χ^2 s $k(k + 1)(k + 2)/6$ stupňami voľnosti. Tento výraz už nemôžeme nazývať výberovou šikmosťou, ale vďaka tomu, že poznáme jeho pravdepodobnostné rozdelenie, môžeme s jeho pomocou realizovať štatistický induktívny test. V prípade, ak nadobúda väčšiu hodnotu, ako je kritická hodnota z rozdelenia χ^2 s daným počtom stupňov voľnosti, zamietame nulovú hypotézu o združenom normálnom rozdelení $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$.

Podobne v prípade výrazu:

$$\left(\hat{\beta}_{2,k} - k(k+2) \right) \sqrt{\frac{n}{8k(k+2)}} \qquad (6.24)$$

je možné dokázať, že má normované normálne rozdelenie, s priemerom rovným nule a rozptylom rovným jednej. Aj tento výraz môžeme využiť na testovanie štatistickej hypotézy o normalite $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$.

Príklad 6.1

V programe R vygenerujme 100 náhodných hodnôt dvojrozmerného náhodného vektora $\mathbf{X}$ so strednou hodnotou $E(\mathbf{X}) = (140, 144)$, rozptylom prvej náhodnej premennej rovným

441, rozptylom druhej premennej 400 a ich kovarianciou rovnou 336. Na tejto vzorke realizujme Mardiov test na normalitu.

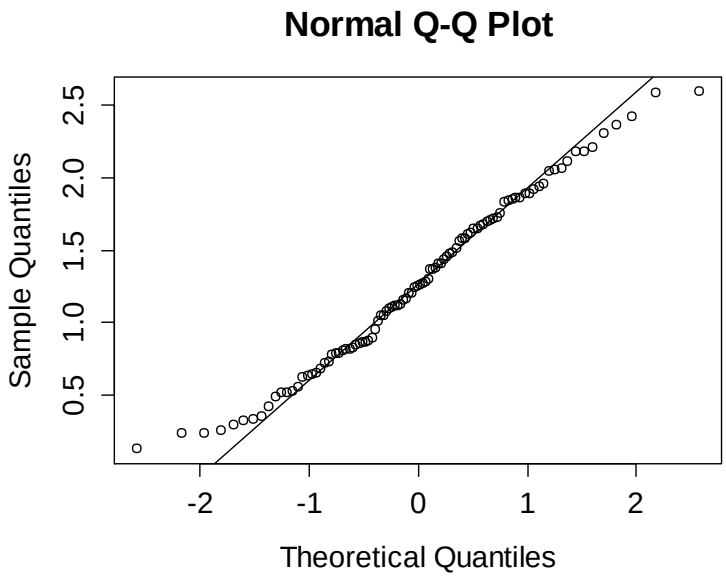
V programe R použijeme na generovanie náhodných hodnôt knižnicu `mvtnorm` a pre testovanie normality knižnicu `psych`.

```
> library(psych)
> library(mvtnorm)
> set.seed(12345)
> data <- rmvnorm(100, mean = c(140, 144), sigma = matrix(c(441,
  336, 336, 400), byrow = TRUE, nrow = 2))
> data[1:10,]
          [,1]     [,2]
 [1,] 153.1158 153.3454
 [2,] 142.8131 129.9444
 [3,] 141.8047 150.5084
 [4,] 119.2936 116.2935
 [5,] 152.7415 152.0592
 [6,] 100.7366 117.8144
 [7,] 149.0515 144.2330
 [8,] 149.0379 169.1344
 [9,] 130.5257 133.4314
[10,] 125.5705 141.2864
```

Nakoniec realizujeme Mardiov test normality.

```
>  mardia(data)
Call: mardia(x = data)

Mardia tests of multivariate skew and kurtosis
Use describe(x) the to get univariate tests
n.obs = 100   num.vars =  2
b1p =  0.12   skew =  2.05  with probability =  0.73
 small sample skew =  2.16  with probability =  0.71
b2p =  6.51   kurtosis =  -1.87  with probability =  0.062
```

Z výsledkov vidíme, že ani v prípade testovania pomocou šikmosti (angl. *skewness*), ani pomocou testovania špicatosti (angl. *kurtosis*) nulovú hypotézu o normalite nezamietame. Tento výsledok však bolo možné očakávať, keďže hodnoty boli naozaj vygenerované z viacrozmerného normálneho rozdelenia. Funkcia `mardia()` zobrazuje aj kvantilový graf, z ktorého je možné vizuálne posúdiť normalitu dát.



Obrázok 6.1: Porovnanie teoretických a empirických kvantilov pomocou funkcie `mardia()`

Zdroj: vlastné spracovanie, výstup zo softvéru R

V knižniciach programu `R` je možné nájsť aj iné testy pre viacrozmernú normalitu, napr. rozšírenie jednorozmerného Shapiro – Wilkovho testu na viacrozmerné normálne rozdelenie. Tento test je možné využiť po inštalácii balíka `mvnormtest`.

Príklad 6.2

V programe R vygenerujme 100 náhodných hodnôt trojrozmerného náhodného vektora **X** so strednou hodnotou $E(\mathbf{X})$ = (140, 144, 150), rozptylom prvej náhodnej premennej rovným 441, druhej 400 a tretej 420. Nech kovariancia medzi prvou a druhou premennou je rovná 336 a ostatné sú nulové. Na týchto dátach uskutočnime Shapiro – Wilkov test pre viacrozmerné normálne rozdelenie.

Najprv vygenerujeme požadované dáta:

```
> set.seed(12345)
> sigma = matrix(c(441, 336, 0, 336, 400, 0, 0, 0, 420), byrow =
  TRUE, nrow = 3)
> sigma
     [,1] [,2] [,3]
[1,]  441  336    0
[2,]  336  400    0
[3,]    0    0  420
> data <- rmvnorm(100, mean = c(140, 144, 150), sigma = sigma)
```

Následne môžeme vykonať požadovaný test. Ešte upozorníme, že funkcia `mshapiro.test` požaduje dáta v opačnom poradí, t.j. jednotlivé premenné sa nachádzajú v riadkoch (ak by sme mali $m \in \mathbb{N}$ premenných a za každú z nich $n \in \mathbb{N}$ pozorovaní,

argumentom funkcie `mshapiro.test` by mala byť matica o rozmeroch $m \times n$). Aby sme našu maticu s údajmi dostali do požadovaného formátu, pôvodnú maticu transponujeme s pomocou funkcie `t()`.

```
> mshapiro.test(t(data))

                  Shapiro-Wilk normality test

data:  Z
W = 0.9852, p-value = 0.3272
```

Z výsledkov vidíme, že nulovú hypotézu o združenom normálnom rozdelení dát nezamietneme.

Predchádzajúci príklad demonštroval testovanie normality v situácii, keď sme si vopred vygenerovali hodnoty z viacrozmerného normálneho rozdelenia – vopred sme teda poznali očakávaný výsledok. V tejto kapitole však budeme používať aj klasický príklad založený na reálnych dátach, s ktorými sa môžeme stretnúť vo väčšine učebníc o viacrozmernej štatistike. Ide o údaje, ktoré po prvý krát použili Flury – Riedwyl (1988) a sú voľne dostupné na internete. Výhodou týchto dát je, že je na nich veľmi jasne možné demonštrovať viacero metód viacrozmernej štatistiky, pričom výsledky sú veľmi výstižné a intuitívne. Pre svoju názornosť je táto databáza (s tými istými údajmi) implementovaná aj v niekoľkých knižniciach programu `R` – napríklad `alr3`, `FRB`, `tclust`, `gclus`. My budeme používať knižnicu `alr3`.

Dáta samotné obsahujú údaje o rozmeroch bankoviek – tisícfrankových švajčiarskych bankovkách. Polovicu z nich tvorili pravé bankovky, druhú polovicu tvorili falzifikáty. V databáze sa celkovo nachádzajú údaje o 200 bankovkách. Pri každej z nich sú uvedené nasledovné miery:

- dĺžka bankovky – premenná `Length`,
- šírka bankovky meraná na ľavom okraji – premenná `Left`,
- šírka bankovky meraná na pravom okraji – premenná `Right`,
- šírka dolného okraja – premenná `Bottom`,
- šírka horného okraja – premenná `Top`,
- dĺžka meraná po uhlopriečke bankovky – premenná `Diagonal`,
- údaj o tom, či ide o pravú (0) alebo falošnú bankovku (1) – premenná `Y`.

Údaje o bankovkách si môžeme priblížiť aj v programe R:

```
> library(alr3)
> data(banknote)
> dim(banknote)
[1] 200   7
> names(banknote)
[1] "Length"   "Left"     "Right"    "Bottom"   "Top"
  "Diagonal" "Y"
> banknote[1:10,]
   Length  Left Right Bottom  Top Diagonal Y
1   214.8 131.0 131.1    9.0  9.7    141.0 0
2   214.6 129.7 129.7    8.1  9.5    141.7 0
3   214.8 129.7 129.7    8.7  9.6    142.2 0
4   214.8 129.7 129.6    7.5 10.4    142.0 0
5   215.0 129.6 129.7   10.4  7.7    141.8 0
6   215.7 130.8 130.5    9.0 10.1    141.4 0
7   215.5 129.5 129.7    7.9  9.6    141.6 0
8   214.5 129.6 129.2    7.2 10.7    141.7 0
9   214.9 129.4 129.7    8.2 11.0    141.9 0
10  215.2 130.4 130.3    9.2 10.0    140.7 0
> summary(banknote)
     Length           Left           Right           Bottom
 Min.   :213.8   Min.   :129.0   Min.   :129.0   Min.   : 7.200
 1st Qu.:214.6   1st Qu.:129.9   1st Qu.:129.7   1st Qu.: 8.200
 Median :214.9   Median :130.2   Median :130.0   Median : 9.100
 Mean   :214.9   Mean   :130.1   Mean   :130.0   Mean   : 9.418
 3rd Qu.:215.1   3rd Qu.:130.4   3rd Qu.:130.2   3rd Qu.:10.600
 Max.   :216.3   Max.   :131.0   Max.   :131.1   Max.   :12.700
      Top           Diagonal           Y
 Min.   : 7.70   Min.   :137.8   Min.   :0.0
 1st Qu.:10.10   1st Qu.:139.5   1st Qu.:0.0
 Median :10.60   Median :140.4   Median :0.5
 Mean   :10.65   Mean   :140.5   Mean   :0.5
 3rd Qu.:11.20   3rd Qu.:141.5   3rd Qu.:1.0
 Max.   :12.30   Max.   :142.4   Max.   :1.0
```

Príklad 6.3

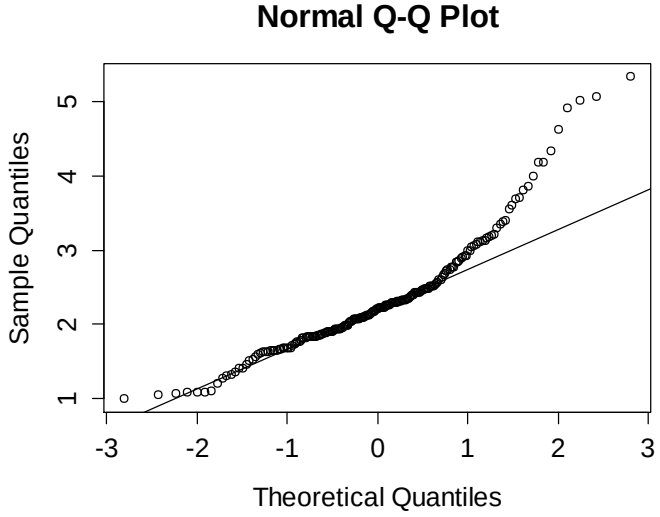
Otestujme dáta o rozmeroch švajčiarskych bankoviek na združenú normalitu. Opäť využijeme funkcie `mardia()` a `mshapiro.test()`.

```
> mardia(banknote[,1:6])
Call: mardia(x = banknote[, 1:6])

Mardia tests of multivariate skew and kurtosis
Use describe(x) the to get univariate tests
n.obs = 200   num.vars =  6
b1p =  6.98   skew =  232.75  with probability =  0
 small sample skew =  237.26  with probability =  0
b2p =  55.27   kurtosis =  5.25  with probability =  1.5e-07
-----------------------------------------------------------------
> mshapiro.test(t(as.matrix(banknote[,1:6])))

                Shapiro-Wilk normality test
```

```
data:  Z
W = 0.9595, p-value = 1.758e-05
```

Výsledky oboch testov naznačujú, že nulovú hypotézu o združenej normalite dát zamietame. Aj obrázok porovnávajúci kvantily teoretického rozdelenia a empirických hodnôt naznačujú odchýlku od normality.

V prípade týchto dát by sme nemali používať žiaden test, ktorý obsahuje tento predpoklad. Pripomeňme však, že aj v tomto prípade je vzhľadom na veľkosť vzorky možné uskutočňovať vďaka viacrozmernej verzii centrálnej limitnej vety testy týkajúce sa vektora stredných hodnôt.



Obrázok 6.2: Porovnanie teoretických a empirických kvantilov v prípade švajčiarskych bankoviek

Zdroj: vlastné spracovanie, výstup zo softvéru R

6.3 Testovanie hypotéz o vektore stredných hodnôt

6.3.1 Testovanie hypotéz o vektore stredných hodnôt pri známej Σ

Uvažujme o situácii, v ktorej by sme poznali variančno-kovariančnú maticu $\mathbf{\Sigma}$ náhodného vektora $\mathbf{X}$, a chceli by sme otestovať hypotézu, či jeho stredná hodnota je rovná zvolenému vektoru $\mathbf{X}^0$. Išlo by teda o testovanie hypotézy:

$$H_0: E(\mathbf{X}) = \mathbf{X}^0 \qquad (6.25)$$

oproti alternatívnej hypotéze:

$$H_1: E(\mathbf{X}) \neq \mathbf{X}^0 \qquad (6.26)$$

Pre ukážku predpokladajme, že zo vzorky pozostávajúcej zo sto pozorovaní sme odhadli vektor priemerov:

$$\bar{\mathbf{X}} = (41, 28)^T \qquad (6.27)$$

a taktiež predpokladajme, že je nám známa variančno-kovariančná matica $\mathbf{\Sigma}$ náhodného vektora X v tvare:

$$\mathbf{\Sigma} = \begin{pmatrix} 326 & 216 \\ 216 & 225 \end{pmatrix} \qquad (6.28)$$

Ide teda o prípad dvoch premenných. Otestujme, či je pravdepodobné, že vektor:

$$\mathbf{X}^0 = (37, 30)^T \qquad (6.29)$$

môže byť vektorom stredných hodnôt pre **X**.

Testovacia charakteristika je v prípade viacrozmerného testu rovná (Rencher, 2002, s. 114):

$$u = n\,(\bar{\mathbf{X}} - \mathbf{X}^0)^T\,\mathbf{\Sigma}^{-1}\,(\bar{\mathbf{X}} - \mathbf{X}^0) \qquad (6.30)$$

V našom prípade to je:

$$100\left[(41,28)-(37,30)\right]\begin{pmatrix} 326 & 216 \\ 216 & 225 \end{pmatrix}^{-1}\left[\begin{pmatrix} 41 \\ 28 \end{pmatrix}-\begin{pmatrix} 37 \\ 30 \end{pmatrix}\right] = 31.82442 \qquad (6.31)$$

Táto testovacia charakteristika má rozdelenie pravdepodobnosti χ^2 s dvoma stupňami voľnosti. Kritická hodnota je $\chi^2_{0,05}(2)$ je rovná 5.991465. Keďže naša testovacia štatistika nadobúda väčšiu ako kritickú hodnotu, nulovú hypotézu v tomto prípade zamietame.

Zostáva otázne, či tento postup bol jediným možným. Na základe poznatkov z predchádzajúcich kapitol by sme mohli nadobudnúť dojem, že by postačovalo použiť dvakrát jednorozmerný test pre strednú hodnotu – za predpokladu, že poznáme variančno kovariančnú maticu $\mathbf{\Sigma}$ by sme dokonca namiesto *t*-testu mohli používať priamo *z*-testy, vychádzajúce z normálneho rozdelenia (keďže z diagonály variančno-kovariančnej matice vieme získať informáciu o rozptyloch jednotlivých náhodných premenných).

Príklad 6.4

Uvažujme o prípade, ak by sme merali dve premenné X_1 a X_2, ktoré by spolu vytvárali náhodný vektor $\mathbf{X} = (X_1, X_2)$. Na základe vzorky, ktorú máme k dispozícii, vypočítame odhad vektora stredných hodnôt $E(\mathbf{X})$ ako vektor priemerov $\bar{\mathbf{X}} = (X_1, X_2)$. Predpokladajme ďalej, že poznáme skutočnú variančno-kovariančnú maticu vektora **X**:

$$\mathbf{\Sigma} = \begin{pmatrix} 441 & 336 \\ 336 & 400 \end{pmatrix}$$

Overme hypotézu, že vektor stredných hodnôt $E(X)$ je rovný vektoru $\mathbf{X}_{H0}$ = (140, 144), ak $\bar{\mathbf{X}}$ = (134, 148) pri počte pozorovaní n = 35.

Zo zadania vyplýva, že náhodná premenná X_1 má štandardnú odchýlku rovnú $441^{1/2}$ = 21 a náhodná premenná X_2 má štandardnú odchýlku rovnú $400^{1/2}$ = 20. Vzájomná kovariancia medzi náhodnými premennými X_1 a X_2 je rovná 336, takže ich vzájomný korelačný koeficient je 336 / (20 * 21) = 0.80. Vstupné údaje môžeme vložiť aj do programu R:

```
> xAvg <- c(134, 148)
> sigma <- matrix(c(441, 336, 336, 400), nrow = 2, byrow = TRUE)
> sigma
     [,1] [,2]
[1,]  441  336
[2,]  336  400
> n <- 35
> xH0 <- c(140, 144)
```

V predchádzajúcom kóde sme vektor priemerov označili ako „`xAvg`“, variančno-kovariančnú maticu „`sigma`“ a vektor čísel, ktoré podľa nulovej hypotézy predstavujú stredné hodnoty ako „`xH0`“.

Úlohou príkladu je testovanie stredných hodnôt. Ak by sme ho chceli realizovať prostredníctvom jednorozmernej induktívnej štatistiky, mohli by sme využiť tzv. z-test, založený na normálnom rozdelení. Tento test by bolo možné využiť preto, lebo predpokladáme známu variančno-kovariančnú maticu $\boldsymbol{\Sigma}$ a nie je potrebné ju odhadovať. Testovaciu štatistiku pre tento test dostaneme zo vzťahu:

$$z_1 = \frac{\bar{x}_1 - x_{1,H0}}{\sigma_1 / \sqrt{n}} \qquad z_2 = \frac{\bar{x}_2 - x_{2,H0}}{\sigma_2 / \sqrt{n}}$$

V programe R po dosadení dostávame:

```
> z1 <- (xAvg[1]-xH0[1])*sqrt(n)/sqrt(sigma[1,1])
> z2 <- (xAvg[2]-xH0[2])*sqrt(n)/sqrt(sigma[2,2])
> z1
[1] -1.690309
> z2
[1] 1.183216
```

Hodnoty získaných testovacích štatistík porovnáme s kvantilom normovaného normálneho rozdelenia.

```
> z <- qnorm(0.975,0,1)
```

```
> z
[1] 1.959964
```

Z výsledku vidíme, že v prípade prvej premennej nulovú hypotézu $E(X_1) = 140$ nevieme zamietnuť, pretože -1.690309 > -1.959964. Podobne v prípade druhej premennej nevieme zamietnuť hypotézu $E(X_2) = 144$, pretože 1.183216 < 1.959964. Skutočné *p*-hodnoty sú (pripomeňme, že počítame obojstranný test):

```
> 2*pnorm(z1, lower.tail=TRUE)
[1] 0.09096895
> 2*pnorm(z2, lower.tail=FALSE)
[1] 0.2367236
```

Namiesto testovania dvoch hypotéz môžeme testovať aj viacrozmerný prípad, ak by sme chceli vedieť, či má náhodný vektor **X** vektor stredných hodnôt $E(\mathbf{X}) = (140, 144)$. V tomto prípade je testovacia štatistika v tvare:

$$Z = n(\bar{\mathbf{X}} - \mathbf{X}_{H0})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \mathbf{X}_{H0})$$

V programe R dostávame:

```
> Z <- n * t(xAvg - xH0) %*% solve(sigma) %*% (xAvg - xH0)
> Z
         [,1]
[1,] 20.71429
```

Hodnotu testovacej štatistiky porovnávame s kvantilom χ^2 rozdelenia s dvoma stupňami voľnosti.

```
> chi <- qchisq(0.95, df = length(xAvg))
> chi
[1] 5.991465
```

Vidíme, že hodnota testovacej štatistiky je väčšia, ako príslušný kvantil, nulovú hypotézu preto v tomto prípade zamietame. *P*-hodnota je pri tomto teste:

```
> pchisq(Z2,length(xAvg), lower.tail=FALSE)
             [,1]
[1,] 3.176508e-05
```

Predchádzajúci príklad poukazuje na zaujímavú skutočnosť – pomerne často je možné pri hľadaní odpovede na určitú otázku použiť viacero štatistických testov. U nás tomu bolo tak v tom zmysle, že sme mohli využiť tak metódy jednorozmernej, ako aj viacrozmernej induktívnej štatistiky. Problémom je, že nás tieto metódy doviedli k dvom odlišným

výsledkom. Kým jednorozmerné testy nezamietali nulovú hypotézu o hodnotách vektora stredných hodnôt náhodného vektora **X**, v prípade viacrozmerného testu sme dostali štatisticky vysoko významný výsledok a nulovú hypotézu sme zamietli.

V takejto situácii je namieste otázka, ktorý z výsledkov je „správny“, resp. ktorým výsledkom sa máme riadiť. Vo všeobecnosti sa spravidla odporúča využitie viacrozmerného testu, z nasledovných dôvodov.

Musíme si uvedomiť, že v našom príklade sme využívali opakovane jednorozmerný *z*-test, v ktorom sme pracovali s obvyklou hladinou významnosti 5 %. Pri testovaní náhodnej premennej X_1 sme použili $\alpha = 0.05$. Podobne aj v prípade druhého testu sme položili $\alpha = 0.05$. Aká je však celková hladina významnosti za *všetky* vykonané testy?

Hladina významnosti $\alpha = 0.05$ znamená, že chybu prvého druhu, t.j. neoprávnené zamietnutie nulovej hypotézy by malo nastať v jednom prípade z dvadsiatich (to je práve spomínaných 5 %). Ak by sme vykonali napríklad desať takýchto testov, celková chyba prvého druhu by vyjadrovala pravdepodobnosť, s akou aspoň v jednom prípade z desiatich nesprávne zamietneme nulovú hypotézu.

Ak by sme na chvíľu predpokladali, že jednotlivé premenné sú nezávislé, potom túto pravdepodobnosť dokážeme vypočítať. Ak α predstavuje chybu prvého druhu, potom $(1 - \alpha)$ predstavuje pravdepodobnosť, že k tejto chybe nedôjde. Pravdepodobnosť toho, že chybu prvého druhu neurobíme pri ani jednom z desiatich nezávislých testov je (pri kapitole o metóde ANOVA je obdobný vzťah, kde nás však zaujímala pravdepodobnosť, že dôjde k chybe I. druhu):

$$(1 - \alpha)^{10} = 0.95^{10} = 0.5987 \qquad (6.32)$$

Z uvedeného je zrejmé, že čím väčší počet jednorozmerných testov realizujeme, tým väčšia bude celková pravdepodobnosť chyby prvého druhu (budeme mať tendenciu zamietať nulovú hypotézu častejšie aj v prípadoch, keď v skutočnosti platí).

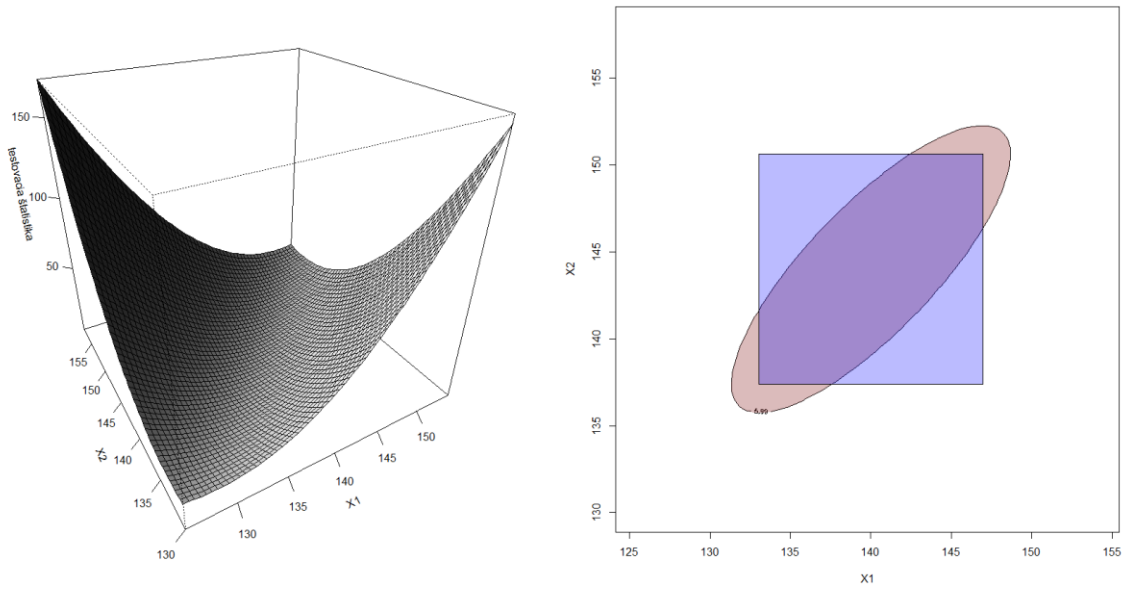
Jedným z jednoduchých pravidiel, ktoré sa snažia tento problém riešiť, je takzvaná Bonferroniho korekcia. Tá spočíva v tom, že sa v jednotlivých jednorozmerných testoch vydelí požadovaná hladina významnosti počtom vykonávaných testov. V našom prípade, keď $\alpha = 0.05$ a uskutočňujeme 10 testov, mali by sme požadovať $\alpha = 0.005$ v každom individuálnom teste. Celkovo by tak bola hladina významnosti za všetky testy približne rovná 0.05.

Problém s Bonferroniho korekciu je ten, že kým v prípade chyby prvého druhu vedie k pomerne konzervatívnym výsledkom, deje sa tak na úkor sily testu. Základný princíp tejto korekcie spočíva v tom, že znižuje hladinu významnosti pre individuálny test – zamietnuť

nulovú hypotézu je preto ťažšie, keďže musí byť splnená prísnejšia podmienka. Na druhej strane to ale znamená, že sa zvyšuje aj pravdepodobnosť, že nedokážeme zamietnuť nulovú hypotézu vtedy, keď by sme to mali urobiť. Inak povedané, zvyšuje sa pravdepodobnosť chyby druhého druhu, čo je to isté, ako keď povieme, že test má menšiu silu.

Alternatívu k tomuto postupu predstavuje využívanie viacrozmerného testu pre strednú hodnotu náhodného vektora. Tento prístup má niektoré výhody – nevzniká tu problém s viacnásobným porovnávaním, keďže sa testuje len jedna štatistická hypotéza. Z tohto dôvodu nie je potrebná ani žiadna korekcia – nevzniká tu teda ani problém s chybou prvého druhu.

Pre ilustráciu vzťahu jednorozmerných a viacrozmerných testov si môžeme ukázať ešte dva obrázky, vychádzajúce z predchádzajúceho príkladu.



Obrázok 6.3: Testovacia štatistika (vľavo) a hodnoty X pre nezamietnutie normality (vpravo)

Zdroj: vlastné spracovanie, výstup z programu R

```
> par(mfrow=c(1,2))
> par(mar=c(2, 2, 2, 2) + 0.1)
> persp(x, y, z, phi = 30, theta = -30, shade = 0.75, xlab =
  "X1", ylab = "X2", zlab = "testovacia štatistika", ticktype =
  "detailed")
> par(mar = c(4, 4, 1, 1) + 0.1)
> contour(x, y, z, levels = c(5.99), xlab = "X1", ylab = "X2")
> clines <- contourLines(x, y, z, levels = c(5.99))
> polygon(clines[[1]]$x, clines[[1]]$y, col = "#77000044",
  border = NA)
> x1 <- xH0[1] - 1.96 * sqrt(sigma[1,1]) / sqrt(n)
> x2 <- xH0[1] + 1.96 * sqrt(sigma[1,1]) / sqrt(n)
> y1 <- xH0[2] - 1.96 * sqrt(sigma[2,2]) / sqrt(n)
> y2 <- xH0[2] + 1.96 * sqrt(sigma[2,2]) / sqrt(n)
> rect(x1, y1, x2, y2, col="#0000ff44")
```

Na ľavej strane predchádzajúceho obrázku vidíme priebeh testovacej štatistiky. Zaujímavý je však pravý obrázok – v ňom sú vyznačené dve oblasti. Modrý obdĺžnik predstavuje hodnoty, ktoré by mohol nadobúdať vektor priemerov, aby pri zvolenej nulovej hypotéze nedochádzalo k jej zamietaniu. Ak by vektor priemerov $\bar{\mathbf{X}}$ nadobúdal hodnoty z modrého obdĺžnika, nulovú hypotézu by sme nezamietali ani v jednom z jednorozmerných testov. Dostávame ho teda ako karteziánsky súčin jednorozmerných 95 %-ných intervalov okolo stredných hodnôt $E(X_1)$ a $E(X_2)$.

Červená elipsa v obrázku zasa predstavuje také hodnoty vektora $\bar{\mathbf{X}}$, pri ktorých by sme nezamietali hypotézu o normalite v prípade viacrozmerného testu. Je vidieť, že aj keď modrý obdĺžnik a červená elipsa nemajú prázdny prienik, množiny nie sú totožné. Pre hodnoty $\bar{\mathbf{X}}$ nachádzajúce sa vo vnútri prieniku platí, že by sme normalitu nezamietali ani v jednorozmerných, ani vo viacrozmerných testoch. Pre hodnoty, ktoré ležia v modrom obdĺžniku, ale nie v elipse, by sme dochádzali k rozdielnym záverom – jednorozmerné testy by normalitu zamietnuť nedokázali, ale viacrozmerný test áno (túto situáciu nazvime „prvým prípadom“). Existuje aj opačný prípad – hodnoty obsiahnuté v elipse, ale nie v obdĺžniku by viedli k zamietnutiu pomocou jednorozmerných testov, ale nie pomocou viacrozmerného (túto situáciu nazvime „druhým prípadom“).

Dôvodom pre tieto odlišnosti je skutočnosť, že viacrozmerný test berie do úvahy aj koreláciu medzi zložkami náhodného vektora. Červená elipsa v obrázku by bola tým užšia (jej vedľajšia os by bola tým kratšia), čím vyššia korelácia by bola medzi zložkami. Odlišnosť záverov v „prvom“ prípade vzniká preto, že jednorozmerné testy ignorujú vzájomný vzťah premenných. Naproti tomu viacrozmerný test konštatuje, že pri danej korelácii medzi premennými by hodnoty v modrom obdĺžniku nemali s danou pravdepodobnosťou nastať – ich výskyt preto znamená odklon od normality. Podobne v „druhom“ prípade hodnôt, ktoré patria červenej elipse, ale nie obdĺžniku, jednorozmerné testy zamietajú normalitu. Avšak viacrozmerný test pripúšťa, že vzhľadom na koreláciu premenných by podobné hodnoty mohli nastať aj pri dátach, ktoré majú združené normálne rozdelenie.

Jedinou nevýhodou tohto viacrozmerného testu je, že v prípade, ak zamietame nulovú hypotézu o zhode vektora stredných hodnôt so zvoleným konštantným vektorom, nedokážeme z výsledkov priamo určiť, ktoré jeho zložky sú štatisticky významne odlišné. Ak dospejeme k takémuto výsledku, je nutná ďalšia analýza na ich identifikáciu.

V tejto podkapitole sme sa zaoberali jednovzorkovým testom pre vektor stredných hodnôt. Jej význam je skôr pedagogický, ako praktický. Vychádzalo sa totiž z predpokladu, že poznáme skutočnú variančno-kovariančnú maticu náhodného vektora $\mathbf{X}$. Pri riešení

skutočných problémov však musíme odhadovať aj variančno-kovariančnú maticu, čo potom vedie k nutnosti využívať iné testovacie charakteristiky, s iným pravdepodobnostným rozdelením. Napriek tomu však majú predchádzajúce odseky zmysel – predstavujú najjednoduchšiu formu testovania štatistických hypotéz o polohe (danej vektorom stredných hodnôt), a problémy, ku ktorým pri ich využití dochádza, je možno s ich pomocou vysvetliť najjednoduchšou formou. Sem patrí hlavne problém s voľbou medzi jednorozmernými a viacrozmernými štatistickými metódami, otázkou, kedy ktoré použiť, ako aj s otázkou interpretácie výsledkov testov, ktoré si navzájom odporujú.

6.3.2 Testovanie hypotéz o vektore stredných hodnôt pri neznámom Σ

Najbežnejší prípad, s ktorým sa môžeme v praxi stretnúť je ten, v ktorom nepoznáme variančno-kovariančnú maticu **Σ**. V takomto prípade nezostáva nič iné, ako ju odhadnúť z výberového súboru, ktorý máme k dispozícii. Využijeme pritom neskreslený odhad variančno-kovariančnej matice v tvare:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T \qquad (6.33)$$

kde n je počet pozorovaní vo vzorke, $\mathbf{X}_i$ je i-té pozorovanie (vektor) a $\bar{\mathbf{X}}$ je vektor priemerov.

Ako sme uviedli vyššie, v prípade známeho **Σ** sme pracovali s testovacou štatistikou v tvare:

$$n\,(\bar{\mathbf{X}} - \mathbf{X}^0)^T\,\boldsymbol{\Sigma}^{-1}\,(\bar{\mathbf{X}} - \mathbf{X}^0) \qquad (6.34)$$

Ak maticu **Σ** nepoznáme, odhadujeme ju pomocou matice **S** a testovacia štatistika nadobúda tvar:

$$T^2 = n\,(\bar{\mathbf{X}} - \mathbf{X}^0)^T\,\mathbf{S}^{-1}\,(\bar{\mathbf{X}} - \mathbf{X}^0) \qquad (6.35)$$

Rozdelenie tejto testovacej štatistiky je za predpokladu, že $\mathbf{X}_i$ majú rovnaké združené rozdelenie pravdepodobnosti, dané Hotellingovým rozdelením pravdepodobnosti s parametrami $p \in \mathbb{N}$ (počet zložiek vektorov $\mathbf{X}_i$) a $n - 1$ (Rencher, 2002, s. 118).

Pre veľké n je možné kritické hodnoty namiesto Hotellingovho rozdelenia aproximovať χ^2 rozdelením s p stupňami voľnosti. V prípade, ak by sme nepracovali s odhadovanou variančno-kovariančnou maticou **S**, ale by sme poznali skutočnú variančno-kovariančnú maticu **Σ**, testovacia štatistika by mala rozdelenie totožné s χ^2 rozdelením s p stupňami voľnosti. Pri malých vzorkách však táto asymptotická vlastnosť vedie k nepresnostiam, spôsobených odhadom matice **S**. Z tohto dôvodu je v týchto prípadoch

vhodnejšie používať aproximáciu podľa F rozdelenia s p a $n - p$ stupňami voľnosti, ktorú dostaneme nasledovnou úpravou testovacej štatistiky:

$$n\,(\bar{\mathbf{X}} - \mathbf{X}^0)^T\,\mathbf{S}^{-1}\,(\bar{\mathbf{X}} - \mathbf{X}^0)\,(n-p)\,/\,(p(n-1)) \qquad (6.36)$$

$$T^2\,(n-p)\,/\,(p(n-1)) \qquad (6.37)$$

Príklad 6.5

Uvažujme o prípade, ak by sme merali dve premenné X_1 a X_2, ktoré by spolu vytvárali náhodný vektor $\mathbf{X} = (X_1, X_2)$. Na základe vzorky overme hypotézu, či $E(\mathbf{X}) = (140, 150)$, ak máme k dispozícii nasledovné hodnoty.

X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
113	107	170	159	154	167	114	126
136	150	133	140	137	133	113	125
154	154	152	168	161	163	145	162
149	161	144	164	118	136	128	133
141	141	119	143	147	175	112	133

Prvým krokom je vloženie príslušných dátových vektorov do programu R.

```
> X1 <- c(113, 136, 154, 149, 141, 170, 133, 152, 144, 119, 154,
  137, 161, 118, 147, 114, 113, 145, 128, 112)
> X2 <- c(107, 150, 154, 161, 141, 159, 140, 168, 164, 143, 167,
  133, 163, 136, 175, 126, 125, 162, 133, 133)
> X <- cbind(X1,X2)
```

Ďalším krokom je výpočet vektora priemerov a odhad variančno – kovariančnej matice.

```
> cmeans <- colMeans(X)
> cmeans
 X1  X2
137 147
> S <- cov(X)
> S
         X1       X2
X1 312.1053 261.5789
X2 261.5789 322.5263
```

Vytvoríme ďalej aj vektor hodnôt zodpovedajúcich nulovej hypotéze.

```
> Xh0 <- c(140, 150)
```

Nakoniec vytvoríme testovaciu štatistiku a odhadneme jej p-hodnotu pomocou F rozdelenia.

```
> n <- dim(X)[1]
> n
[1] 20
> p <- dim(X)[2]
> p
[1] 2
> T2 <- n * (cmeans - Xh0) %*% solve(S) %*% (cmeans - Xh0)
> T2
          [,1]
[1,] 0.6223984
> T2f <- T2 * (n - p) / ((p * (n - 1)))
> T2f
          [,1]
[1,] 0.2948203
> pf(T2f, p, n-p)
          [,1]
[1,] 0.2518071
```

Z výsledku vidíme, že nulovú hypotézu nevieme zamietnuť, keďže *p*- hodnota je rovná približne 0.2518.

Príklad 6.6

Na vzorke pravých a falošných bankoviek z knižnice `alr3` (Flury – Riedwyl, 1988) realizujme test o rovnosti stredných hodnôt rozmerov tisícfrankových bankoviek z prvej série, pričom predpokladáme šírku 215 a výšku 132 milimetrov.

Prvým krokom je vloženie príslušných dátových vektorov do programu R. Pripomeňme, že dátový objekt `banknote` obsahuje stĺpce, v ktorých je uložená šírka bankovky, výška bankovky meraná na ľavom okraji, výška bankovky meraná na pravom okraji, šírka dolného okraja, šírka horného okraja, dĺžka meraná po uhlopriečke bankovky a binárny údaj o tom, či ide o pravú (0) alebo falošnú bankovku (1). Keďže náš zaujíma šírka a výška, budeme pracovať len s prvými tromi stĺpcami. Je dobrú si všimnúť, že informáciu o výške bankovky máme meranú tak na pravom, ako aj ľavom okraji – do analýzy zahrnieme obidve tieto hodnoty. Vzorku si tak isto rozdelíme na pravé a falošné bankovky.

```
> library(alr3)
> prave   <- banknote[banknote[,7]==0, 1:3]
> falosne <- banknote[banknote[,7]==1, 1:3]
> bankovky <- prave
> head(bankovky)
  Length  Left Right
1  214.8 131.0 131.1
2  214.6 129.7 129.7
3  214.8 129.7 129.7
```

```
4  214.8 129.7 129.6
5  215.0 129.6 129.7
6  215.7 130.8 130.5
```

Do vektora priemery si uložíme priemerné hodnoty meraní, a do vektora `skutocneRozmery` uložíme teoretické hodnoty, zodpovedajúce nulovej hypotéze.

```
> priemery <- colMeans(bankovky)
> priemery
 Length    Left   Right
214.969 129.943 129.720
> skutocneRozmery <- c(215,132,132)
```

Podobne ako v predchádzajúcom prípade, ďalej postupujeme výpočtom testovacej štatistiky a stanovením jej významnosti.

```
> n <- dim(bankovky)[1]
> n
[1] 100
> p <- length(priemery)
> p
[1] 3
> S <- cov(bankovky)
> S
           Length       Left      Right
Length 0.15024141 0.05801313 0.05729293
Left   0.05801313 0.13257677 0.08589899
Right  0.05729293 0.08589899 0.12626263
>
> T2 <- n * (t(priemery - skutocneRozmery) %*% solve(S) %*%
   (priemery - skutocneRozmery))
> T2
         [,1]
[1,] 5538.687
> pchisq(T2, df = p, lower.tail=FALSE)
     [,1]
[1,]    0
```

Vypočítaná hodnota testovacej štatistiky ďaleko presahuje kritickú hodnotu, čomu zodpovedá aj nízka *p*-hodnota získaná pomocou aproximácie rozdelením χ^2.

Namiesto prácneho výpočtu Hotellingovho T^2 testu môžeme použiť aj knižnice programu R – tento test implementuje napríklad balík ICSNP.

```
> HotellingsT2(bankovky, mu = skutocneRozmery, test="chi")

             Hotelling's one sample T2-test

data:  bankovky
```

```
T.2 = 5538.687, df = 3, p-value < 2.2e-16
alternative hypothesis: true location is not equal to
  c(215,132,132)
```

Všimnime si, že hodnota testovacej štatistiky je totožná s hodnotou, ktorú sme získali priamym výpočtom.

Test môžeme vykonať aj prostredníctvom aproximácie F rozdelením. V tomto prípade musíme upraviť testovaciu štatistiku.

```
> T2f <- T2 * (n - p) / (p * (n-1))
> T2f
         [,1]
[1,] 1808.931
> pf(T2f, df1=p, df2=n-p, lower.tail = FALSE)
             [,1]
[1,] 5.685955e-85
-----------------------------------------------------------------
> HotellingsT2(bankovky, mu = skutocneRozmery, test="f")

               Hotelling's one sample T2-test

data:  bankovky
T.2 = 1808.931, df1 = 3, df2 = 97, p-value < 2.2e-16
alternative hypothesis: true location is not equal to
  c(215,132,132)
```

Opäť dochádzame k tomu istému záveru – nulovú hypotézu zamietame.

Výsledok pre falošné bankovky je veľmi podobný.

```
> HotellingsT2(falosne, mu = skutocneRozmery, test = "f")

            Hotelling's one sample T2-test

data:  falosne
T.2 = 1784.083, df1 = 3, df2 = 97, p-value < 2.2e-16
alternative hypothesis: true location is not equal to
  c(215,132,132)
```

Zaujímavejší prípad nastáva, ak máme namiesto jednovzorkového testu potrebu vykonávať dvojvzorkový test. Ide o situáciu, keď namiesto porovnávania vektora stredných hodnôt s dopredu známym vektorom konštánt porovnávame medzi sebou dva vektory stredných hodnôt. Tento typ problému vyvstáva napríklad v prípade, ak chceme porovnať viacero premenných na častiach vzorky, ktorú máme k dispozícii (napr. ak vzorku rozdelíme na mužov a ženy, a podobne). Postup, ktorý môžeme v tomto prípade použiť, je dvojvzorkový Hotellingov T^2 test.

V tomto prípade ide formálne o testovanie hypotézy:

$$H_0\text{: } E(\mathbf{X}_1) = E(\mathbf{X}_2) \tag{6.38}$$

$$H_1\text{: } E(\mathbf{X}_1) \neq E(\mathbf{X}_2) \tag{6.39}$$

Na základe údajov používaných v predchádzajúcom príklade by takejto hypotéze mohla zodpovedať otázka, či pravé a falošné bankovky majú v priemere rovnaké rozmery – čo by umožňovalo odhaliť falzifikáty pomocou ich presného merania.

Uvažujme o prípade, kedy máme dve vzorky – pozorovania v rámci prvej označme $\mathbf{X}_{1,i}$ a druhej $\mathbf{X}_{2,j}$, pričom $i = 1, 2, \ldots, n_1$ a $j = 1, 2, \ldots, n_2$, kde n_1 a n_2 sú veľkosti týchto vzoriek ($n_1, n_2 \in \mathbb{N}$). O hodnotách náhodných vektorov $\mathbf{X}_{1,i}$ budeme predpokladať, že majú združené normálne rozdelenie pravdepodobnosti s vektorom stredných hodnôt $\mathrm{E}(\mathbf{X}_1)$ a náhodné vektory $\mathbf{X}_{2,j}$ združené normálne rozdelenie pravdepodobnosti s vektorom stredných hodnôt $\mathrm{E}(\mathbf{X}_2)$. Test ďalej predpokladá, že vektory z oboch vzoriek pochádzajú z rozdelenia s tou istou, aj keď neznámou variančno-kovariančnou maticou $\boldsymbol{\Sigma}$. Predpokladá sa teda, že pozorovania vo vzorkách sú homoskedastické – a to nie len v rámci tej istej vzorky, ale aj medzi vzorkami.

Definujme ďalej vektory priemerov $\bar{\mathbf{X}}_1$ a $\bar{\mathbf{X}}_2$ podobne, ako sme to urobili v predchádzajúcej časti. Vychádzajúc z predpokladu homoskedasticity môžeme pre odhad variančno-kovariančnej matice zlúčiť pozorovania z obidvoch vzoriek, čím je možné vylepšiť štatistické vlastnosti odhadu $\boldsymbol{\Sigma}$. Tento odhad môžeme zapísať v tvare (Rencher, 2002):

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1} \left(\mathbf{X}_{1,i} - \overline{\mathbf{X}}_1\right)\left(\mathbf{X}_{1,i} - \overline{\mathbf{X}}_1\right)^T + \sum_{j=1}^{n_2} \left(\mathbf{X}_{2,j} - \overline{\mathbf{X}}_2\right)\left(\mathbf{X}_{2,j} - \overline{\mathbf{X}}_2\right)^T \right) \tag{6.40}$$

$$= \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2} \tag{6.41}$$

V predchádzajúcom vzťahu predstavujú $\mathbf{S}_1$ a $\mathbf{S}_2$ výberové variančno-kovariančné matice odhadnuté na skúmaných vzorkách a platí $E(\mathbf{S}) = \boldsymbol{\Sigma}$. Testovacia štatistika sa pre overenie požadovanej hypotézy vypočíta podľa vzťahu:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)^T \mathbf{S}^{-1}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) \tag{6.42}$$

Testovacia štatistika T^2 má Hotellingovo rozdelenie pravdepodobnosti s parametrami k a $n_1 + n_2 - 2$.

6.4 Testovanie hypotéz o variančno-kovariančných maticiach

V predchádzajúcej časti sme sa zaoberali testovaním vektora stredných hodnôt. Podobne ako v jednorozmernom prípade najčastejšie testujeme strednú hodnotu a rozptyl,

v prípade viacrozmernej štatistiky môžeme okrem vektora stredných hodnôt testovať variančno-kovariančnú maticu. Pripomeňme, že variančno-kovariančná matica $\mathbf{\Sigma}$ náhodného vektora $\mathbf{X}$ obsahuje na hlavnej diagonále rozptyly zložiek $\mathbf{X}$ a mimodiagonálne prvky predstavujú kovariancie medzi zložkami vektora $\mathbf{X}$.

Skôr ako pristúpime k testovaniu matice variančno-kovariančných matíc, je potrebné vyjadriť sa ešte ku koncepcii počtu stupňov voľnosti. Podobný koncept existuje aj pri výpočte variančno-kovariančných matíc. Ako sme už spomínali v predchádzajúcej podkapitole, pri výpočte neskresleného odhadu $\mathbf{\Sigma}$, ktorý označujeme $\mathbf{S}$ sa vo vzorci vyskytuje výraz $n - 1$, kde $n \in \mathbb{N}$ je počet pozorovaní. Pripomeňme, že pre náhodné vektory $\mathbf{X}_i$ používame na výpočet matice $\mathbf{S}$ vzťah:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \qquad (6.43)$$

Pri práci s variančno-kovariančnými maticami niekedy postupujeme tak, že pre ich výpočet zoskupujeme súbory pozorovaní z rôznych vzoriek do jednej súhrnnej. Takúto operáciu môžeme uskutočniť napríklad vtedy, ak sa máme dôvod domnievať, že variančno-kovariančná matica (poprípade celé rozdelenia pravdepodobnosti) sú vo všetkých vzorkách rovnaké – majme napríklad tri vzorky po 100 študentov vysokých škôl. Ak sa domnievame, že premenné skúmané na vzorkách majú rovnakú populačnú variančno-kovariančnú maticu pre všetky vzorky, potom môžeme pre odhad populačnej variančno-kovariančnej matice použiť všetky údaje. Namiesto 100 údajov (pre každú vzorku samostatne) je potom odhad založený na 300 pozorovaniach.

Ak je predpoklad o totožnej populačnej variančno-kovariančnej matici oprávnený, zoskupením vytvárame súbor, ktorý má vyšší počet pozorovaní (angl. *pooled dataset*), a s jeho použitím by sme mali získať lepší odhad spoločnej variančno-kovariančnej matice. Ak by sme postupovali za každú z $k \in \mathbb{N}$ vzoriek samostatne, menovateľ výrazu pre výpočet $\mathbf{S}$ by bol pre každú vzorku rovný $n - 1$. Dostali by sme k matíc. Ak by sme chceli vypočítať jednu maticu založenú na združení všetkých pozorovaní do jednej vzorky, vzorec pre výpočet spoločnej variančno-kovariančnej matice $\mathbf{S}^p$ by sme zmenili tak, že v menovateli by sa nachádzal výraz $n - k$:

$$\mathbf{S}^p = \frac{1}{n-k}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \qquad (6.44)$$

kde k je počet spájaných vzoriek a n je počet pozorovaní po združení všetkých vzoriek dohromady. Ak by počet pozorovaní v jednotlivých vzorkách bol označený ako n_i pre

$i = 1, 2, ..., k$, potom $n = n_1 + n_2 + ... + n_k$. Variančno-kovariančná matica $\mathbf{S}^p$ je potom odhadom spoločnej populačnej matice $\mathbf{\Sigma}$.

Príklad 6.7

Tri skupiny po desať respondentov odpovedajú na dve otázky na škále od 1 po 10. Odpovede sa nachádzajú v nasledujúcej tabuľke. Vypočítajme odhad variančno-kovariančných matíc pre každú skupinu zvlášť, a následne odhad spoločnej variančno-kovariančnej matice, ak by sme združili pozorovania zo všetkých skupín.

Skupina A		*Skupina B*		*Skupina C*	
$\mathbf{X}_{11}$	$\mathbf{X}_{12}$	$\mathbf{X}_{21}$	$\mathbf{X}_{22}$	$\mathbf{X}_{31}$	$\mathbf{X}_{32}$
6	6	6	8	6	7
9	6	8	9	8	7
6	5	5	6	6	8
6	3	7	9	5	8
8	7	3	6	5	10
4	7	6	8	6	7
5	9	7	6	5	5
4	4	7	8	7	5
6	7	7	7	7	9
9	8	7	4	7	9

Najprv vložíme vektory zodpovedajúce odpovediam respondentov do programu R.

```
> X11 <- c(6,9,6,6,8,4,5,4,6,9)
> X12 <- c(6,6,5,3,7,7,9,4,7,8)
> X21 <- c(6,8,5,7,3,6,7,7,7,7)
> X22 <- c(8,9,6,9,6,8,6,8,7,4)
> X31 <- c(6,8,6,5,5,6,5,7,7,7)
> X32 <- c(7,7,8,8,10,7,5,5,9,9)
```

Vektory môžeme spájať do matíc po stĺpcoch pomocou funkcie `cbind()`. Variančno-kovariančnú maticu môžeme odhadnúť pomocou funkcie `cov()`. Je dobré si všimnúť, že táto funkcia používa vzorec, ktorý sme definovali vyššie – v menovateli zlomku používa číslo $n - 1$.

```
> cov(cbind(X11,X12))
          X11       X12
X11 3.3444444 0.7111111
X12 0.7111111 3.2888889
```

```
> cov(cbind(X21,X22))
          X21       X22
X21 2.0111111 0.7444444
X22 0.7444444 2.5444444

> cov(cbind(X31,X32))
           X31        X32
X31  1.0666667 -0.1111111
X32 -0.1111111  2.7222222
```

Ak by sme chceli odhadnúť variančno-kovariančnú maticu na základe združených údajov, je potrebné spojiť pozorovania zo všetkých troch skupín. Vychádzame pritom z párovania, pri ktorom zlučujeme vektory $\mathbf{X}_{11}$, $\mathbf{X}_{21}$ a $\mathbf{X}_{31}$ do vektora $\mathbf{X}_1$. Podobne zlúčime aj vektory $\mathbf{X}_{12}$, $\mathbf{X}_{22}$ a $\mathbf{X}_{32}$ do vektora $\mathbf{X}_2$. Je treba pripomenúť, že v novovytvorených vektoroch si údaje za jednotlivé pozorovania (riadky) musia zodpovedať – vektory za spájané tri skupiny musíme zlučovať v tom istom poradí.

```
> X1 <- c(X11,X21,X31)
> X1
 [1] 6 9 6 6 8 4 5 4 6 9 6 8 5 7 3 6 7 7 7 7 6 8 6 5 5 6 5 7 7 7
> X2 <- c(X12,X22,X32)
> X2
 [1]  6  6  5  3  7  7  9  4  7  8  8  9  6  9  6  8  6  8  7  4
  7  7  8  8 10  7  5  5  9  9
```

Spoločnú variančno-kovariančnú maticu by sme teoreticky mohli počítať obvyklým spôsobom.

```
> cov(cbind(X1,X2))
          X1        X2
X1 1.9954023 0.3977011
X2 0.3977011 2.9609195
```

Problémom je, že funkcia `cov()` v tomto prípade počíta podľa nesprávneho vzorca – v menovateli je stále počet pozorovaní znížený o 1, aj keď ide o zlúčené údaje. Počet zlúčených pozorovaní je:

```
> length(X1)
[1] 30
```

Neskreslený odhad variančno-kovariančnej matice pri zlúčených údajov môžeme získať tak, že výsledok funkcie `cov()` prenásobíme 30 – 1, a vydelíme 30 – 3 (keďže sme spájali

3 vzorky). Prenásobením sa zbavíme nesprávneho čísla z menovateľa a následne výsledok vydelíme správnym menovateľom.

```
> cov(cbind(X1,X2))*(length(X1)-1)/(length(X1)-3)
          X1        X2
X1 2.1432099 0.4271605
X2 0.4271605 3.1802469
```

Väčšinu testov, ktoré budú uvedené v nasledujúcich podkapitolách, je možné realizovať tak na individuálnych variančno-kovariančných maticiach, ako aj na maticiach založených na združení vzoriek. Postup pritom zostáva vždy nezmenený, mení sa len spomínaný počet stupňov voľnosti. Z dôvodu zrozumiteľnosti nebudeme na tento aspekt ďalej upozorňovať – zavedieme však konvenciu, že počet stupňov voľnosti označíme ako $\nu \in \mathbb{N}$. V prípade, ak pracujeme s individuálnou vzorkou, kladieme $\nu = n - 1$, v prípade združených údajov použijeme $\nu = n - k$.

6.4.1 Testovanie hypotézy o zhode variančno-kovariančnej matice s $\boldsymbol{\Sigma}_0$

Kým v prípade jednorozmernej štatistiky sa ako prvé spravidla realizovali testy zvoleného populačného parametra voči konštante, analógiou, ktorou sa budeme zaoberať v tejto časti, je testovanie hypotézy, či je variančno-kovariančná matica rovná nejakej konkrétnej matici, ktorú označíme $\boldsymbol{\Sigma}_0$. Prvkami tejto matice sú reálne čísla, a nie náhodné premenné – je to teda matica konštánt. V ďalšom texte uvažujeme o p zložkovom náhodnom vektore $\mathbf{X}$, kde $p \in \mathbb{N}$.

Na prvý pohľad sa môže zdať ťažké nájsť situáciu, v ktorej by sme mali apriórnu informáciu o celej variančno-kovariančnej matici $\boldsymbol{\Sigma}_0$. V skutočnosti takéto prípady môžu nastať pomerne často. Stačí si predstaviť prípad, ak by sme položili:

$$\boldsymbol{\Sigma}_0 = \mathbf{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{6.45}$$

V uvedenom prípade by v skutočnosti išlo o test, či je variančno-kovariančná matica rovná jednotkovej matici. Znamenalo by to, že zložky náhodného vektora $\mathbf{X}$ sú nekorelované náhodné premenné, s rozptylom rovným jednej. Ak by sme uvoľnili predpoklad o jednotkovom rozptyle, inou možnosťou je testovať maticu:

$$\boldsymbol{\Sigma}_0 = \sigma^2 \mathbf{I}_p = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} \qquad (6.46)$$

Takáto definícia $\boldsymbol{\Sigma}_0$ by znamenala, že zložky náhodného vektora **X** sú nekorelované, a majú rovnaký rozptyl, rovný σ^2. Podobne by sme mohli definovať prípad, kedy by sme každej náhodnej premennej z náhodného vektora **X** umožnili mať vlastný rozptyl, odlišný od ostatných.

Testovacia štatistika je definovaná nasledovne (Timm, 2002):

$$W = v\,[\,\ln(|\boldsymbol{\Sigma}_0|) - \ln(|\mathbf{S}|) + \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}_0^{-1}) - p\,] \qquad (6.47)$$

V predchádzajúcom vzťahu označuje $|\boldsymbol{\Sigma}_0|$ determinant hypotetickej variančno-kovariančnej matice, ktorá je stanovená v nulovej hypotéze. Podobne $|\mathbf{S}|$ je determinant variančno-kovariančnej matice odhadnutej na základe dostupnej vzorky. Výraz $\boldsymbol{\Sigma}_0^{-1}$ predstavuje inverznú maticu k $\boldsymbol{\Sigma}_0$ a funkcia $\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}_0^{-1})$ predstavuje stopu matice $\mathbf{S}\boldsymbol{\Sigma}_0^{-1}$ (stopa štvorcovej matice je definovaná ako súčet prvkov na jej hlavnej diagonále). Podrobnejší popis použitých maticových operácií presahuje rozsah tejto publikácie – samotný výpočet je však ľahké realizovať v prostredí R, pretože všetky potrebné funkcie sú v ňom už implementované. V prípade, ak je v veľké a platí H_0, má testovacia štatistika W približne rozdelenie χ^2 s $p(p + 1)/2$ stupňami voľnosti.

H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$	Hypotézu H_0 zamietame, ak $W > \chi^2_{(1-\alpha),\,(p(p+1)/2)}$
H_1: $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$	

Pre menšie hodnoty v je možné využiť korekciu (Timm, 2002):

$$C = \frac{1}{6v-1}\frac{2p^2+3p-1}{p+1} \qquad (6.48)$$

S pomocou korekčného člena C je možné definovať modifikovanú testovaciu štatistiku nasledovne:

$$W' = W(1-C) \qquad (6.49)$$

Testovacia štatistika W' má to isté pravdepodobnostné rozdelenie, ako W, takže kritérium pre zamietnutie nulovej hypotézy je podobné:

H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$	Hypotézu H_0 zamietame, ak $W' > \chi^2_{(1-\alpha),\,(p(p+1)/2)}$
H_1: $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$	

Namiesto testu založenom na rozdelení χ^2 je možné použiť aj štatistiku s Fisherovým rozdelením. Pre jej zostrojenie definujme nasledujúce premenné:

$$C_0 = \frac{(p-1)(p+2)}{6v} \tag{6.50}$$

$$f_0 = \frac{\frac{p(p+1)}{2}+2}{\left|C_0 - C^2\right|} \tag{6.51}$$

$$a = \frac{\frac{p(p+1)}{2}}{1 - C - \frac{p(p+1)}{2 f_0}} \tag{6.52}$$

Testovacia štatistika má F rozdelenie s $f = (p\ (p\ +\ 1)/2)$ a f_0 stupňami voľnosti, a definujeme ju ako podiel:

$$F = \frac{W}{a} \tag{6.53}$$

Testovanie hypotézy sa v tomto prípade realizuje nasledovným spôsobom:

H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$	Hypotézu H_0 zamietame, ak $F > F_{(1-\alpha),(f,f0)}$
H_1: $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$	

Príklad 6.8

Na vzorke pravých bankoviek z knižnice `alr3` (Flury – Riedwyl, 1988) realizujme test pre premenné šírka bankovky, výška bankovky meraná na ľavom okraji, výška bankovky meraná na pravom okraji, šírka dolného okraja, šírka horného okraja o rovnosti ich variančno-kovariančnej matice matici:

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.15 & 0 & 0 & 0 & 0 \\ 0 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \end{pmatrix}$$

Požadované údaje o bankovkách získame tak ako doposiaľ z knižnice `alr3`.

```
> library(alr3)
> data(banknote)
> data <- banknote[banknote$Y==0,1:5]
> head(data)
  Length  Left Right Bottom  Top
1  214.8 131.0 131.1    9.0  9.7
2  214.6 129.7 129.7    8.1  9.5
3  214.8 129.7 129.7    8.7  9.6
4  214.8 129.7 129.6    7.5 10.4
```

```
5  215.0 129.6 129.7   10.4  7.7
6  215.7 130.8 130.5    9.0 10.1
```

Do premennej data sme vložili len údaje z požadovaných stĺpcov (1:5) z riadkov, ktoré zodpovedali pravým bankovkám (banknote$Y==0).

Ďalej pokračujeme výpočtom výberovej variančno-kovariančnej matice.

```
> S <- cov(data)
> S
           Length       Left      Right      Bottom         Top
Length 0.15024141 0.05801313 0.05729293  0.05712626  0.01445253
Left   0.05801313 0.13257677 0.08589899  0.05665152  0.04906667
Right  0.05729293 0.08589899 0.12626263  0.05818182  0.03064646
Bottom 0.05712626 0.05665152 0.05818182  0.41320707 -0.26347475
Top    0.01445253 0.04906667 0.03064646 -0.26347475  0.42118788
```

Pre výpočty ďalej potrebujeme maticu $\mathbf{\Sigma}_0$, ktorá je podľa zadania diagonálna.

```
> Sigma0 <- array(0,dim=c(5,5))
> diag(Sigma0) <- c(0.15, 0.15, 0.15, 0.4, 0.4)
> Sigma0
     [,1] [,2] [,3] [,4] [,5]
[1,] 0.15 0.00 0.00  0.0  0.0
[2,] 0.00 0.15 0.00  0.0  0.0
[3,] 0.00 0.00 0.15  0.0  0.0
[4,] 0.00 0.00 0.00  0.4  0.0
[5,] 0.00 0.00 0.00  0.0  0.4
```

Testovanie si ukážeme s pomocou všetkých troch spôsobov uvedených vyššie. Vzhľadom na diskusiu o stupňoch voľnosti z predchádzajúcej kapitoly definujeme:

```
> v <- dim(data)[1]-1
> v
[1] 99
> p <- dim(data)[2] # počet premenných
> p
[1] 5
```

Ak by sme použili základný vzorec pre veľké vzorky, vypočítali by sme testovaciu štatistiku *W* nasledovne:

```
> W <- v*(log(det(Sigma0)) - log(det(S)) + matrix.trace(S %*%
  solve(Sigma0)) - p)
> W
[1] 170.3029
```

Pre posúdenie významnosti potrebujeme túto hodnotu porovnať s rozdelením χ^2, ktorého počet stupňov voľnosti je:

```
> f <- p*(p+1)/2
> f
[1] 15
```

Kritická hodnota by v tomto prípade bola:

```
> qchisq(0.95, df = f, lower.tail = FALSE)
[1] 7.260944
```

Nulovú hypotézu preto na základe asymptotického testu zamietame. V prípade, ak by sme chceli postupovať na základe vzorca určeného pre menšie vzorky, vypočítali by sme korekčný faktor:

```
> C <- (1/(6*v-1)) * (2*p*p+3*p-1)/(p+1)
> C
[1] 0.01798763
```

S jeho pomocou by sme vypočítali upravenú testovaciu štatistiku:

```
> (1-C)*W
[1] 167.2395
```

Kritická hodnota zostáva aj v tomto prípade rovnaká, nulovú hypotézu opäť zamietame. Poslednou alternatívou je využitie testu založeného na *F* rozdelení. Dosadením do vzorcov dostávame nasledujúce hodnoty:

```
> C0 <- (p-1)*(p+2) / (6*v)
> C0
[1] 0.04713805
> f0 <- ((p*(p+1))/2+2) / abs(C0-C*C)
> f0
[1] 363.1354
> a  <- (p*(p+1)/2) / (1-C-(p*(p+1))/(2*f0))
> a
[1] 15.94548
> F  <- W/a
> F
[1] 10.68032
```

Kritická hodnota je v tomto prípade:

```
> qf(0.95, f, f0, lower.tail=FALSE)
```

```
[1] 0.4802989
```

Podobne ako v predchádzajúcich dvoch prípadoch, testovacia štatistika je väčšia ako kritická hodnota a nulovú hypotézu zamietame.

6.4.2 Test sféricity

V predchádzajúcej podkapitole sme uviedli test, ktorý je možné využiť pre testovanie zhody variančno-kovariančnej matice s vopred stanovenou maticou konštánt. Uvedený postup je možné využiť pre ľubovoľnú konštantnú variančno-kovariančnú maticu, uviedli sme si aj niektoré príklady. Jeden z týchto prípadov je však natoľko dôležitý, že sú mu venované samostatné testy. Ide o testovanie hypotézy sféricity, kde testujeme zhodu variančno-kovariančnej matice s maticou, ktorá má nulové mimodiagonálne prvky a to isté číslo na hlavnej diagonále.

Uvedená štruktúra variančno-kovariančnej matice zodpovedá situácii, v ktorej majú všetky náhodné premenné tvoriace náhodný vektor rovnaké rozptyly a sú vzájomne nekorelované. Ak má náhodný vektor navyše aj viacrozmerné združené normálne rozdelenie, nekorelovanosť zložiek náhodného vektora implikuje aj ich nezávislosť.

Testom sa pôvodne zaoberal Mauchly (1940), niekedy sa nazýva aj Mauchlyho test sféricity. Formálne, nulová hypotéza testuje rovnosť:

$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p \tag{6.54}$$

Test je založený na podiele vierohodnosti (angl. likelihood ratio), ktorý vedie k testovacej štatistike v tvare (Rencher, 2002):

$$W = -v \ln\left(\frac{p^p |\mathbf{S}|}{(\mathrm{tr}(\mathbf{S}))^p} \right) \tag{6.55}$$

Testovacia štatistika W má asymptoticky rozdelenie χ^2 s $(p - 1)(p + 2)/2$ stupňami voľnosti.

H_0: $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ H_1: $\mathbf{\Sigma} \neq \sigma^2 \mathbf{I}_p$	Hypotézu H_0 zamietame, ak $W > \chi^2_{(1-\alpha),\,(p-1)(p+2)/2}$

Box (1949) ukázal, že konvergencia k rozdeleniu χ^2 je rýchlejšia, ak sa namiesto testovacej štatistiky W použije štatistika:

$$W' = -\left(v - \frac{2p^2 + p + 2}{6p} \right) \ln\left(\frac{p^p |\mathbf{S}|}{(\mathrm{tr}(\mathbf{S}))^p} \right) \tag{6.56}$$

ktorá má taktiež asymptoticky rozdelenie χ^2 s $(p - 1)(p + 2)/2$ stupňami voľnosti.

H_0: $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p$ H_1: $\boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}_p$	Hypotézu H_0 zamietame, ak $W' > \chi^2_{(1-\alpha),\,(p-1)(p+2)/2}$

Aproximácia rozdelením χ^2 funguje pomerne dobre v prípade, ak $n > 20$ a $p < 6$.

Príklad 6.9

Na vzorke pravých bankoviek z knižnice `alr3` (Flury – Riedwyl, 1988) realizujme test sféricity pre premenné šírka bankovky, výška bankovky meraná na ľavom okraji a výška bankovky meraná na pravom okraji pre falošné bankovky.

Požadované údaje o bankovkách získame tak ako doposiaľ z knižnice `alr3`.

```
library(alr3)
data <- banknote[banknote$Y==1, c(1,2,3)]
```

Ďalej si definujeme počet pozorovaní a počet skúmaných premenných:

```
> n <- dim(data)[1]
> n
[1] 100
> p <- dim(data)[2]
> p
[1] 3
```

Okrem výpočtu výberovej variančno-kovariančnej matice pre ďalšie výpočty potrebujeme poznať aj jej determinant a stopu:

```
> covariance <- cov(data)
> covariance
          Length       Left      Right
Length 0.12401111 0.03151515 0.02400101
Left   0.03151515 0.06505051 0.04676768
Right  0.02400101 0.04676768 0.08894040
> detCovar <- det(covariance)
> detCovar
[1] 0.0003911833
> trCovar <- matrix.trace(covariance)
> trCovar
[1] 0.278002
```

S pomocou týchto hodnôt môžeme pristúpiť k výpočtu samotnej testovacej štatistiky:

```
> u <- (p^p)*detCovar / (trCovar^p)
> u
[1] 0.4915869
> W <- -(n-1)*log(u)
> W
[1] 70.30155
```

Počet stupňov voľnosti máme daný ako $(p - 1)(p + 2)/2$, čo je to isté, ako $p(p + 1)/2 - 1$.

```
> 0.5*p*(p+1)-1
[1] 5
```

Testovaciu štatistiku porovnávame s rozdelením χ^2 s piatimi stupňami voľnosti. Pre p-hodnotu máme približne:

```
> chiPval <- pchisq(W, df = 0.5*p*(p+1)-1, lower.tail=FALSE)
> chiPval
[1] 8.869043e-14
```

Ak by sme chceli postupovať podľa presnejšieho vzorca, dosadíme:

```
> W2 <- -(n-1 - (2*p*p+p+2)/(6*p))*log(u)
> W2
[1] 69.39418
> chiPval <- pchisq(W2, df = 0.5*p*(p+1)-1, lower.tail=FALSE)
> chiPval
[1] 1.369891e-13
```

Vidíme, že v oboch prípadoch dostávame rovnaké závery – nulovú hypotézu o sféricite zamietame.

6.4.3 Test o rovnosti viacerých variančno-kovariančných matíc

V predchádzajúcom texte sme skúmali a testovali zhodu variančno-kovariančnej matice a určitej vopred danej variančno-kovariančnej matice konštánt. Overovali sme aj zhodu so špeciálnym prípadom variančno-kovariančnej matice pri teste sféricity. V oboch prípadoch sme uskutočňovali jednovzorkové testovanie – mali sme k dispozícii len jednu vzorku, z ktorej sme vypočítavali výberovú variančno-kovariančnú maticu. V nasledujúcej časti budeme pracovať s viacerými vzorkami. Podobne, ako v jednorozmernej štatistike môžeme porovnávať rozptyly dvoch súborov, vo viacrozmernom prípade môžeme porovnávať rovnosť variančno-kovariančných matíc.

Ak uvažujeme o $k \in \mathbb{N}$ vzorkách, potom základnou nulovou hypotézou, ktorá nás zaujíma, je hypotéza o tom, že platí $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$.

Je dobré si uvedomiť, že ak by testovaná nulová hypotéza bola pravdivá, a variančno-kovariančné matice vo všetkých vzorkách boli naozaj rovnaké, mohli by sme spoločnú variančno-kovariančnú maticu, ktorú označíme $\boldsymbol{\Sigma}$, odhadnúť presnejšie spojením všetkých

vzoriek do jednej spoločnej (resp. združenej), na základe ktorej by sme mohli vypočítať výberovú charakteristiku. Pripomeňme diskusiu zo začiatku tejto kapitoly, že spoločnú variančno-kovariančnú maticu môžeme odhadnúť ako:

$$\mathbf{S} = \frac{1}{\sum_{i=1}^{k} n_i - k} \sum_{i=1}^{k} (n_i - 1)\mathbf{S}_i = \frac{1}{v} \sum_{i=1}^{k} (n_i - 1)\mathbf{S}_i \qquad (6.57)$$

Ak n_i sú veľkosti vzoriek pre i = 1, 2, ..., k, potom v definujeme ako súčet počtu pozorovaní vo všetkých vzorkách, od ktorých odpočítame počet vzoriek k.

Test je založený na štatistike M, ktorú vypočítame nasledovne:

$$M = \sqrt{\frac{|\mathbf{S}_1|^{(n_1-1)} |\mathbf{S}_2|^{(n_2-1)} \cdots |\mathbf{S}_k|^{(n_k-1)}}{|\mathbf{S}|^{v}}} \qquad (6.58)$$

V predchádzajúcom vzorci predstavuje výraz $|\mathbf{S}_i|$ pre i = 1, 2, ..., k determinanty výberových variančno-kovariančných matíc za jednotlivé vzorky, a $|\mathbf{S}|$ determinant variančno-kovariančnej matice vypočítanej zo združených vzoriek.

Aproximácie pre štatistiku M pomocou rozdelení χ^2 a F navrhol Box (1949), tomuto testu sa preto hovorí aj Boxov M-test.

Najprv sa zameriame na aproximáciu rozdelením χ^2. V tomto prípade sa využíva koeficient:

$$c = \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{v} \right) \left(\frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \right) \qquad (6.59)$$

kde p je počet premenných vo vzorkách. Testovacia štatistika založená na M má tvar:

$$u = -2\,(1 - c)\,\ln M \qquad (6.60)$$

Ekvivalentne to môžeme napísať aj takto:

$$u = -(1 - c) \left(\sum_{i=1}^{k} (n_i \ln|\mathbf{S}_i| - v \ln|\mathbf{S}| \right) \qquad (6.61)$$

Ak označíme:

$$W = v \ln |\mathbf{S}| - \sum_{i=1}^{k} (n_i - 1) \ln |\mathbf{S}_i| \qquad (6.62)$$

potom má testovacia štatistika tvar:

$$u = (1 - c) W \qquad (6.63)$$

Testovacia štatistika u má približne rozdelenie χ^2 s $f = (k - 1)\, p\, (p + 1)$ stupňami voľnosti. Kritérium pre testovanie hypotézy je preto nasledovné (Rencher, 2002, s. 257):

H_0: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$ H_1: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2 \neq \cdots \neq \boldsymbol{\Sigma}_k$	Hypotézu H_0 zamietame, ak $u > \chi^2_{(1-\alpha),\,(k-1)\,p\,(p+1)}$

Aproximácia pomocou rozdelenia χ^2 sa používa v prípade, ak n_i je aspoň 20 pre $i = 1, 2, \ldots, k$ a súčasne počet premenných p vo vzorkách, ako aj počet vzoriek k je menší ako 6. V prípade, ak tieto podmienky nie sú splnené, je obvyklé použiť aproximáciu pomocou F rozdelenia.

Pre výpočet testovacej štatistiky s približne F rozdelením je potrebné definovať nasledovné koeficienty (Timm, 2002):

$$C_2 = \frac{(p-1)(p+2)}{6(k-1)} \left(\sum_{i=1}^{k} \frac{1}{(n_i - 1)^2} - \frac{1}{v^2} \right) \qquad (6.64)$$

$$a_1 = \frac{1}{2}(k-1)p(p+1) \qquad (6.65)$$

$$a_2 = \frac{a_1 + 2}{|C_0 - C^2|} \qquad (6.66)$$

$$b_1 = \frac{1 - C - \dfrac{a_1}{a_2}}{a_1}$$

$$b_2 = \frac{1 - C + \dfrac{2}{a_2}}{a_2} \qquad (6.67)$$

V prípade, ak $C_2 - C^2 > 0$ je testovacia štatistika daná vzorcom:

$$F = -2\, b_1 \ln M \qquad (6.68)$$

V opačnom prípade použijeme testovaciu štatistiku:

$$F = \frac{2 a_2 b_2 \ln M}{a_1 (1 + 2 b_2 \ln M)} \qquad (6.69)$$

V oboch prípadoch má testovacia štatistika približne F rozdelenie s a_1 a a_2 stupňami voľnosti. Kritériu pre testovanie hypotézy je potom:

H_0: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$ H_1: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2 \neq \cdots \neq \boldsymbol{\Sigma}_k$	Hypotézu H_0 zamietame, ak $F > F_{1-\alpha,a_1,a_2}$

Príklad 6.10

Otestujme, či pravé a falošné bankovky, z databázy využitej v predchádzajúcom príklade majú rovnakú variančno-kovariančnú maticu pre premenné šírka bankovky, výška bankovky meraná na ľavom okraji a výška bankovky meraná na pravom okraji. Najprv uskutočníme výber požadovaných premenných.

```
library(alr3)
data1 <- banknote[banknote$Y==0,c(1,2,3)]
data2 <- banknote[banknote$Y==1,c(1,2,3)]
```

Pravé bankovky sú v premennej `data1`, falošné v premennej `data2`.

```
> k
[1] 2
> n1 <- dim(data1)[1]
> n1
[1] 100
> p1 <- dim(data1)[2]
> p1
[1] 3
> n2 <- dim(data2)[1]
> n2
[1] 100
> p2 <- dim(data2)[2]
> p2
[1] 3
```

Predchádzajúcimi príkazmi sme si definovali základné konštanty. Keďže porovnávame dve vzorky, $k = 2$. Ďalej vidíme, že $n_1 = n_2 = 100$, teda vzorky sú rovnako veľké. Počet premenných v prvej aj druhej vzorke je $p_1 = p_2 = 3$. Všimnime si, že ak by bol počet premenných rôzny, úloha by nedávala zmysel. V ďalšom kroku vypočítame variančno-kovariančné matice pre jednotlivé vzorky, ako aj pre združenú vzorku:

```
> covar1 <- cov(data1)
> covar1
           Length       Left      Right
Length 0.15024141 0.05801313 0.05729293
Left   0.05801313 0.13257677 0.08589899
Right  0.05729293 0.08589899 0.12626263
```

```
> covar2 <- cov(data2)
> covar2
           Length       Left      Right
Length 0.12401111 0.03151515 0.02400101
Left   0.03151515 0.06505051 0.04676768
Right  0.02400101 0.04676768 0.08894040
> covarPool <- ((n1 - 1) * covar1 + (n2 - 1) * covar2)/(n1 +n2 -
  2)
> covarPool
           Length       Left      Right
Length 0.13712626 0.04476414 0.04064697
Left   0.04476414 0.09881364 0.06633333
Right  0.04064697 0.06633333 0.10760152
```

Testovacia štatistika je založená na determinantoch týchto matíc, ktoré vypočítame v nasledujúcom kroku:

```
> detCovar1 <- det(covar1)
> detCovar1
[1] 0.00111728
> detCovar2 <- det(covar2)
> detCovar2
[1] 0.0003911833
> detCovarPool <- det(covarPool)
> detCovarPool
[1] 0.0007171418
```

Ukážme si teraz, ako by sa postupovalo pri výpočte pomocou aproximácie rozdelením χ^2. V tomto prípade by sme vypočítali:

```
> lnM <- 0.5*((n1-1)*log(detCovar1)+(n2-1)*log(detCovar2))-
  0.5*(n1+n2-2)*log(detCovarPool)
> lnM
[1] -8.054571
> coefChi <- (1/(n1-1)+1/(n2-1)-1/(n1+n2-2))*(2*p1*p1+3*p1-
  1)/(6*(p1+1)*(k-1))
> coefChi
[1] 0.01641414
```

Celá testovacia štatistika by potom mala tvar:

```
> u <- -2*(1-coefChi)*lnM
> u
[1] 15.84472
```

Po stanovení počtov stupňov voľnosti:

```
> dfChi <- 0.5*(k-1)*p1*(p1+1)
> dfChi
```

```
[1] 6
```

Môžeme vypočítať *p*-hodnotu:

```
> chiPval <- pchisq(u, df = dfChi, lower.tail=FALSE)
> chiPval
[1] 0.01461211
```

Nulovú hypotézu by sme preto zamietali. Pre výpočet pomocou *F* aproximácie by sme museli vypočítať nasledovné pomocné koeficienty:

```
> coefF <- (1/((n1-1)^2) + 1/((n2-1)^2) - 1/((n1+n2-2)^2))*(p1-
  1)*(p1+2)/(6*(k-1))
> coefF
[1] 0.0002975887
> a1 <- dfChi
> a1
[1] 6
> a2 <- (a1+2)/abs(coefF-coefChi*coefChi)
> a2
[1] 284044.1
> b1 <- (1-coefChi-a1/a2)/a1
> b1
[1] 0.1639275
> b2 <- (1- coefChi+2/a2)/a2
> b2
[1] 3.462818e-06
```

Nakoniec vypočítame testovaciu štatistiku a *p*-hodnotu:

```
> if (coefF>coefChi*coefChi) statF <- -2*b1*lnM else statF <- -
  2*a2*b2*lnM/(a1+2*a1*b2*lnM)
> coefF
[1] 0.0002975887
> fPval <- pf(statF,a1,a2, lower.tail=F)
> fPval
[1] 0.01461593
```

Aj v prípade aproximácie *F* rozdelením nulovú hypotézu na $\alpha = 0.05$ zamietame. Máme teda dôvod domnievať sa, že variančno-kovariančné matice pre pravé a falošné bankovky v sledovaných charakteristikách nie sú rovnaké. Líšiť sa môžu rozptyly hodnôt meraných veličín (čo by mohlo znamenať aj menšiu precíznosť pri výrobe bankoviek), ale aj rozdiely medzi vzťahmi meraných veličín.

V súvislosti s Boxovým *M* testom je potrebné spomenúť jedno varovanie. Tento test je veľmi citlivý, a aj pri malých rozdieloch má tendenciu zamietať nulovú hypotézu. Podobne

má problémy aj v prípade, ak pozorovania nepochádzajú z normálneho rozdelenia pravdepodobnosti (viacrozmerného). Z tohto dôvodu je jeho použitie obmedzené – až do tej miery, že test nie je implementovaný v štandardných knižniciach programu R.

6.4.4 Testy o nezávislosti

V tejto časti si priblížime testy, ktoré sa často označujú ako testy nezávislosti. V skutočnosti je tento názov presný, len ak sa použije dodatočný predpoklad o združenom viacrozmernom rozdelení skúmaných premenných. Testy sú totiž založené na skúmaní prvkov variančno-kovariančnej matice, konkrétne vzájomných kovariancií. Platí, že v prípade, ak majú premenné združené viacrozmerné rozdelenie, nekorelovanosť (nulová kovariancia) implikuje nezávislosť skúmaných premenných.

Na začiatok uvažujme o situácii, v ktorej si skúmané premenné rozdelíme na dve skupiny. Ak by sme vo vzorke mali celkovo $p + q$ premenných, kde $p, q \in \mathbb{N}$, potom označme prvých p premenných ako skupinu **x** a zvyšných q premenných ako skupinu **y**.

Celkovú variančno-kovariančnú maticu $\mathbf{\Sigma}$ rádu $(p + q) \times (p + q)$ je možné napísať v blokovej forme nasledovne:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}} \\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}} \end{pmatrix} \tag{6.70}$$

V predchádzajúcom zápise majú matice $\mathbf{\Sigma_{xx}}$, $\mathbf{\Sigma_{xy}}$, $\mathbf{\Sigma_{yx}}$ a $\mathbf{\Sigma_{yy}}$ rozmery v tomto poradí $p \times p$, $p \times q$, $q \times p$ a $q \times q$. Maticu $\mathbf{\Sigma}$ teda „rozdelíme“ podľa skupín premenných tak po riadkoch, ako aj po stĺpcoch.

Hypotéza o nezávislosti skupiny premenných **x** a **y** hovorí, že matica $\mathbf{\Sigma}$ má tvar:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{yy}} \end{pmatrix} \tag{6.71}$$

To je ekvivalentné tvrdeniu, že $\mathbf{\Sigma_{xy}} = (\mathbf{\Sigma_{yx}})^T = \mathbf{0}$, kde $\mathbf{0}$ je matica rozmeru $p \times q$, ktorá má všetky prvky nulové.

Hypotézu o nezávislosti skupín premenných **x** a **y** overujeme pomocou testovacej štatistiky:

$$\Lambda = \frac{|\mathbf{S}|}{|\mathbf{S_{yy}}||\mathbf{S_{xx}}|} \tag{6.72}$$

Táto testovacia štatistika má Wilksovo rozdelenie lambda s parametrami q, p a $n - 1 - p$. Podľa Renchera (2002, s. 163) je možné toto rozdelenie aproximovať F rozdelením.

Z princípu konštrukcie hypotézy je zrejmé, že by mohli vzniknúť situácie, kde by sme uvažovali s viac než len dvomi skupinami premenných. Všeobecný prípad dostávame, ak by sme sledované premenné rozdelili do $k \in \mathbb{N}$ skupín premenných, ktoré označíme $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k$. Variančno-kovariančnú maticu $\mathbf{\Sigma}$ potom môžeme rozdeliť na nasledovné bloky:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1k} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{\Sigma}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{k1} & \mathbf{\Sigma}_{k2} & \cdots & \mathbf{\Sigma}_{kk} \end{pmatrix} \tag{6.73}$$

Nulová hypotéza predpokladá nekorelovanosť medzi skupinami premenných, čo by znamenalo:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{kk} \end{pmatrix} \tag{6.74}$$

V predchádzajúcom zápise si je dobré všimnúť, že matice núl v rôznych blokoch nie sú nutne totožnými nulovými maticami – všetky síce obsahujú nulové prvky, ale rozmery matíc môžu byť rôzne, v závislosti od počtu premenných v jednotlivých skupinách premenných $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k$.

Základná testovacia štatistika je daná vzťahom podobným tomu, ktorý sme používali pri delení na dve skupiny.

$$u = \frac{|\mathbf{S}|}{|\mathbf{S}_{11}||\mathbf{S}_{22}|\cdots|\mathbf{S}_{kk}|} \tag{6.75}$$

Opäť je teda počítaná na základe determinantov výberových variančno-kovariančných matíc. Táto testovacia štatistika nemá jednoduché pravdepodobnostné rozdelenie, takže sa používa aproximácia založená na rozdelení χ^2.

Pre výpočet modifikovanej štatistiky je potrebné definovať nasledovné pomocné premenné:

$$a_2 = p^2 - \sum_{i=1}^{k} p_i^{\,2} \tag{6.76}$$

$$a_3 = p^3 - \sum_{i=1}^{k} p_i^{\,3} \tag{6.77}$$

$$f = a_2 / 2 \tag{6.78}$$

$$c = 1 - \frac{1}{12\,fv}(2a_3 + 3a_2) \qquad (6.79)$$

Modifikovaná testovacia štatistika má potom f stupňov voľnosti a tvar:

$$u' = -\,vc \ln u \qquad (6.80)$$

O hypotéze rozhodujeme nasledovne:

$H_0: \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix}$	Hypotézu H_0 zamietame, ak $u' > \chi^2_{1-\alpha,f}$
$H_1: \quad \boldsymbol{\Sigma} \neq \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix}$	

Príklad 6.11

Pokračujme v našom príklade s bankovkami. Rozdeľme hodnoty merané na bankovkách na nasledovné skupiny:

- dĺžka bankovky,
- šírka bankovky meraná na pravom a ľavom okraji,
- šírka dolného a horného okraja,
- dĺžka meraná po uhlopriečke bankovky.

Zistime, či sú uvedené skupiny premenných nekorelované v skupine falošných bankoviek.

Po výbere požadovaných premenných z databázy je potrebné vypočítať variančno-kovariančné matice tak za všetky premenné spolu (**S**), ako aj za čiastkové matice $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$ a $\mathbf{S}_4$.

```
> data <- banknote[banknote$Y==1, c(1:6)]
> n <- dim(data)[1]
> n
[1] 100
> covMatica <- cov(data)
> covMatica
          Length     Left     Right  Bottom       Top Diagonal
Length    0.1240  0.03152  0.024001 -0.1006  0.019435  0.01157
Left      0.0315  0.06505  0.046768 -0.0240 -0.011919 -0.00505
Right     0.0240  0.04677  0.088940 -0.0186  0.000132  0.03419
Bottom   -0.1006 -0.02404 -0.018576  1.2813 -0.490192  0.23848
Top       0.0194 -0.01192  0.000132 -0.4902  0.404456 -0.02207
```

```
Diagonal  0.0116 -0.00505  0.034192  0.2385 -0.022071  0.31121
> cov11 <- covMatica[1:1,1:1]
> cov11
[1] 0.1240111
> cov22 <- covMatica[2:3,2:3]
> cov22
            Left      Right
Left  0.06505051 0.04676768
Right 0.04676768 0.08894040
> cov33 <- covMatica[4:5,4:5]
> cov33
           Bottom        Top
Bottom  1.2813131 -0.4901919
Top    -0.4901919  0.4044556
> cov44 <- covMatica[6:6,6:6]
> cov44
[1] 0.3112121
```

Ako vyplýva zo zadania, dokopy máme šesť premenných, v skupinách po jednej, dvoch, dvoch a jednej premennej.

```
> p1 <- 1
> p2 <- 2
> p3 <- 2
> p4 <- 1
> p <- p1 + p2 + p3 + p4
> p
[1] 6
```

Základ pre testovaciu štatistiku predstavuje podiel determinantu celej variančno-kovariančnej matice a súčinu determinantov jej blokov:

```
> u <- det(covMatica)/(cov11*det(cov22)*det(cov33)*cov44)
> u
[1] 0.5333891
```

Všimnime si, že v príkaze v programe R sme v menovateli počítali len dva determinanty – za prvú a štvrtú skupinu premenných sme dosadili len vypočítané kovariancie (reálne čísla). Je tomu tak preto, lebo ide o jednoprvkové skupiny premenných a v tomto prípade by sme mali počítať determinant z jedného čísla (matice rádu 1×1). Determinant je v tomto prípade rovný vypočítanej kovariancii. Keďže pravdepodobnostné rozdelenie vypočítanej štatistiky nepoznáme, vypočítame modifikovanú štatistiku. Najprv si pripravíme pomocné premenné:

```
> a2 <- p*p - (p1*p1 + p2*p2 + p3*p3 + p4*p4)
> a2
```

```
[1] 26
> a3 <- p*p*p - (p1*p1*p1 + p2*p2*p2 + p3*p3*p3 + p4*p4*p4)
> a3
[1] 198
> f <- 0.5 * a2
> f
[1] 13
> v <- n-1
> v
[1] 99
> a2 <- p*p - (p1*p1 + p2*p2 + p3*p3 + p4*p4)
> a2
[1] 26
> a3 <- p*p*p - (p1*p1*p1 + p2*p2*p2 + p3*p3*p3 + p4*p4*p4)
> a3
[1] 198
> f <- 0.5 * a2
> f
[1] 13
> v <- n-1
> v
[1] 99
> c <- 1 - (2*a3 + 3*a2)/(12*f*v)
> c
[1] 0.9693085
```

A napokon vypočítame samotnú štatistiku:

```
> uStar <- -v*c*log(u)
> uStar
[1] 60.31221
```

Jej významnosť získame z rozdelenia χ^2:

```
> pchisq(uStar,df = f, lower.tail=F)
[1] 4.619416e-08
```

Ako vidíme, nulovú hypotézu zamietame – medzi skupinami premenných existuje vzťah.

Ako poslednú alternatívu môžeme uviesť špeciálny prípad, pri ktorom by sme testovali nekorelovanosť, resp. nezávislosť všetkých premenných (po dvojiciach). V takomto prípade by sme v rámci nulovej hypotézy vychádzali z nasledovného tvaru variančno-kovariančnej matice pre $p \in \mathbb{N}$ premenných:

$$H0: \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp}^2 \end{pmatrix} \tag{6.81}$$

Takáto nulová hypotéza môže pôsobiť dojmom, že ide o špeciálny prípad testovania zhody variančno-kovariančnej matice s maticou konštánt – teda o prípad, ktorý sme skúmali na začiatku tejto podkapitoly. V skutočnosti ide o inú situáciu. Keďže na začiatku kapitoly sme skúmali zhodu s pevne danou variančno-kovariančnou maticou, bolo nutné poznať aj hodnoty rozptylov na hlavnej diagonále. Test nezávislosti, ktorý prezentujeme v tejto kapitole, však takúto informáciu nepožaduje. Hypotézu, ktorú skúmame, by sme mohli preformulovať do tvaru „variančno-kovariančná matica je diagonálna“. K hodnotám na diagonále sa nevyjadrujeme, a z hľadiska testu ani pre nás nie sú zaujímavé – jediné, čo overujeme, sú nulové vzájomné kovariancie premenných.

Keďže spomínaná forma variančno-kovariančnej matice je špeciálnym prípadom situácie popísanej vyššie (každá premenná je skupinou, resp. každá skupina premenných je jednoprvková), postup pri testovaní sa nemení. Výpočet sa však môže značne zjednodušiť, ak použijeme nasledujúce vzťahy:

$$u = \frac{|\mathbf{S}|}{s_{11}^2 s_{22}^2 \cdots s_{pp}^2} \tag{6.82}$$

$$u' = -\left(v - \frac{2p+5}{6}\right) \ln u \tag{6.83}$$

Testovaciu štatistiku u môžeme vypočítať aj pomocou korelačnej matice, ktorú označme **R**. Jej prvkami nie sú kovariance, ale korelácie medzi hodnotami skúmaných premenných. V prípade **R** je štatisika u daná ako determinant korelačnej matice v tvare:

$$u = |\mathbf{R}| \tag{6.84}$$

Testovacia štatistika u' má pravdepodobnostné rozdelenie χ^2 s $p(p-1)/2$ stupňami voľnosti. Hypotézu potom testujeme nasledovne:

$H_0: \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp}^2 \end{pmatrix}$	Hypotézu H_0 zamietame, ak $u' > \chi^2_{1-\alpha, p(p-1)/2}$

$H_1: \quad \boldsymbol{\Sigma} \neq \begin{pmatrix} \sigma_{11}{}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{22}{}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp}{}^2 \end{pmatrix}$	

Príklad 6.12

Využime premenné o falošných bankovkách z predchádzajúceho príkladu a realizujem test hypotézy o nezávislosti všetkých šiestich premenných.

V príklade používame $n = 100$ pozorovaní a $p = 6$ premenných.

```
> data <- banknote[banknote$Y==1, c(1:6)]
> n <- dim(data)[1]
> n
[1] 100
> p <- dim(data)[2]
> p
[1] 6
```

Na základe údajov vypočítame tentoraz korelačnú namiesto variančno-kovariančnej matice:

```
> R <- cor(data)
> R
          Length    Left     Right  Bottom       Top Diagonal
Length    1.0000  0.3509  0.228533 -0.2524  0.086781   0.0589
Left      0.3509  1.0000  0.614852 -0.0833 -0.073483  -0.0355
Right     0.2285  0.6149  1.000000 -0.0550  0.000698   0.2055
Bottom   -0.2524 -0.0833 -0.055026  1.0000 -0.680931   0.3777
Top       0.0868 -0.0735  0.000698 -0.6809  1.000000  -0.0622
Diagonal  0.0589 -0.0355  0.205516  0.3777 -0.062209   1.0000
> u <- det(R)
> u
[1] 0.178
```

Pre modifikovanú testovaciu štatistiku a počet stupňov voľnosti dostávame:

```
> uStar <- -(n-1-(2*p+5)/6)*log(u)
> uStar
[1] 166
> f <- p*(p-1)/2
> f
[1] 15
```

To nás vedie k výpočtu *p*-hodnoty pomocou rozdelenia χ^2:

```
> pchisq(uStar,df = f, lower.tail=F)
[1] 1.54e-27
```

Podobne ako v predchádzajúcich prípadoch, nulovú hypotézu zamietame.

7 Príklady

Táto kapitola pozostáva zo zadaní k príkladom, ako aj s riešením pre väčšinu z týchto príkladov. Príklady boli formulované aplikačne. Pri formuláciách sme sa snažili úlohy zostaviť tak, aby úloha mala charakter neštruktúrovaného zadania, pri ktorom je zrejme najnáročnejšou úlohou pre čitateľa pochopiť, aký typ analýzy sa od neho vyžaduje.

7.1 Zadania príkladov

Príklad 7.1

V tomto príklade budeme počítať s údajmi dostupnými online [22.06.2013] na stránke <http://lib.stat.cmu.edu/DASL/Datafiles/Shoppers.html>. Ide o údaje týkajúce sa výdavkov 50-tich spotrebiteľov v obchode s potravinami. Predpokladáme, že spotrebitelia boli oslovení na základe náhodného výberu. Odhadnite strednú hodnotu výdavkov spotrebiteľov tohto obchodu spolu s príslušnými intervalmi spoľahlivosti ($\alpha = 0.05$).

Tabuľka 15: Výdavky 50-tich spotrebiteľov v USD/1 nákup

P.č.	výdavky	P.č.	výdavky	P.č.	výdavky	P.č.	výdavky	P.č.	výdavky
1	2.32	11	13.67	21	18.30	31	27.07	41	37.52
2	6.61	12	13.72	22	18.71	32	28.76	42	39.28
3	6.90	13	14.35	23	19.54	33	29.15	43	40.80
4	8.04	14	14.52	24	19.55	34	30.54	44	43.97
5	9.45	15	14.55	25	20.58	35	31.99	45	45.58
6	10.26	16	15.01	26	20.89	36	32.82	46	52.36
7	11.34	17	15.33	27	20.91	37	33.26	47	61.57
8	11.63	18	16.55	28	21.13	38	33.8	48	63.85
9	12.66	19	17.15	29	23.85	39	34.76	49	64.30
10	12.95	20	18.22	30	26.04	40	36.22	50	69.49

Zdroj: Moore – McCabe (1989) (dostupné online [22.06.2013] *na <http://lib.stat.cmu.edu/DASL/Datafiles/Shoppers.html>)*

Príklad 7.2

K dispozícii máme ročné údaje o HDP a spotrebe na obyvateľa (CONS), v trhových cenách v Slovenskej republike za obdobie 1995 – 2010. Údaje sú uvedené v nasledujúcej tabuľke[28] (Tabuľka 16). Vypočítajte intervaly spoľahlivosti pre stredné hodnoty týchto dvoch premenných a pre stredné hodnoty prvých diferencií HDP a spotreby ($\alpha = 0.05$).

[28] Keďže však ide o časové rady, bolo by vhodné pracovať skôr s diferenciami, lebo pôvodné hodnoty časového radu určite nie sú nezávislé. Pracujeme teda za predpokladu, že namerané hodnoty sú *iid*.

Tabuľka 16: HDP a spotreba za obdobie 1995 – 2010 v SR

Rok	HDP	CONS
1995	7.0	5.1
1996	7.7	5.9
1997	8.3	6.3
1998	8.8	6.8
1999	9.0	6.9
2000	9.5	7.3
2001	10.4	8.1
2002	11.1	8.7
2003	11.5	8.9
2004	12.3	9.4
2005	13.5	10.3
2006	15.0	11.4
2007	16.9	12.4
2008	18.2	13.6
2009	17.0	13.8
2010	18.0	14.0

Zdroj: Eurostat

Pozn.: údaje sú v parite kúpnej sily na jedného obyvateľa v trhových cenách

---

Príklad 7.3

Koncom januára 2012 sme z internetovej stránky www.reality.sk stiahli ceny a rozlohy náhodne vybraných bytov z mesta Košice. Zaujíma nás, aká je stredná hodnota cien bytov v Košiciach, ktorých rozloha je viac ako 50 m^2. Vypočítajte tiež interval spoľahlivosti pre túto strednú hodnotu ($\alpha = 0.05$). Údaje sú uvedené v nasledujúcej tabuľke (Tabuľka 17).

Tabuľka 17: Cena a rozloha vybraných bytov v Košiciach

Byt	Cena v EUR	m^2
1	113500	96
2	64000	68
3	36500	43
4	63000	56
5	69000	51
6	48400	37
7	66900	67
8	86000	92
9	61900	68
10	63000	64
11	81500	74
12	72500	71
13	109800	99.6
14	89900	80
15	52700	42

Zdroj: reality.sk

Príklad 7.4

V tomto príklade využijeme databázu `datasets`, konkrétne dáta `cars`. Tieto dáta obsahujú 50 pozorovaní a dve premenné – rýchlosť sledovaného auta (v míľach za hodinu) a jeho brzdnú dráhu (v stopách). Vypočítajte priemernú brzdnú dráhu áut, ktoré dosiahli rýchlosť 15 míľ za hodinu a vyššiu. Vytvorte *x-y* graf, na ktorom by bolo možné vidieť prípadnú závislosť rýchlosti auta a jeho brzdnej dráhy.

Príklad 7.5

Otvorte si databázu `Rabbit` z knižnice `MASS`. Odhadnite strednú hodnotu zmeny krvného tlaku zajacov podľa toho, či im bolo aplikované placebo (skupina `Control`) alebo skutočný liek (skupina `MDL`). Vypočítajte intervaly spoľahlivosti pre tieto stredné hodnoty a pokúste sa zhodnotiť (samozrejme len z pohľadu štatistiky, nie medicíny), či sa preukázala v danom výskume opodstatnenosť tohto lieku.

Príklad 7.6

V tomto príklade sa pozrieme na údaje o zadlženosti krajín EÚ. Z voľne dostupných zdrojov je možné stiahnuť údaje za ukazovateľ D/HDP, teda pomer štátneho dlhu a hrubého domáceho produktu krajiny (v angl. *debt-to-GDP ratio*). My budeme pracovať s údajmi z databázy Eurostat, ktoré uvádzame v tabuľke. K dispozícii máme údaje za obdobie od 4. kvartálu 2000 do 3. kvartálu 2010. Pracovať budeme s krajinami, o ktorých sa predpokladá výskyt problémov s výškou štátneho dlhu. Tieto krajiny sú známe pod mierne degradujúcim akronymom „PIIGGS“ – **P**ortugal (Portugalsko), **I**reland (Írsko), **I**taly (Taliansko), **G**reece (Grécko), **G**reat Britain (Veľká Británia), **S**pain (Španielsko).

Podľa tzv. Paktu stability a rastu (v angl. *Stability and Growth Pact*) je pre členské krajiny EÚ stanovená hranica verejného dlhu v pomere k HDP na 60 %. Vypočítajte s 95 % konfidenciou, či uvedené krajiny spĺňali v priemere za sledované obdobie toto kritérium.

```
obs <- c("2000Q4", "2001Q1", "2001Q2", "2001Q3", "2001Q4",
  "2002Q1", "2002Q2", "2002Q3", "2002Q4", "2003Q1", "2003Q2",
  "2003Q3", "2003Q4", "2004Q1", "2004Q2", "2004Q3", "2004Q4",
  "2005Q1", "2005Q2", "2005Q3", "2005Q4", "2006Q1", "2006Q2",
  "2006Q3", "2006Q4", "2007Q1", "2007Q2", "2007Q3", "2007Q4",
  "2008Q1", "2008Q2", "2008Q3", "2008Q4", "2009Q1", "2009Q2",
  "2009Q3", "2009Q4", "2010Q1", "2010Q2", "2010Q3")
time <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
  17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31,
  32, 33, 34, 35, 36, 37, 38, 39, 40)
Portugal <- c(48.5, 48.2, 49.1, 50.3, 51.2, 50.3, 51.1, 52.7,
  53.8, 55.3, 54.4, 56, 55.9, 55.6, 57.1, 57.1, 57.6, 58.1,
```

```
  59.3, 61.8, 62.8, 62.5, 64.4, 63.5, 63.9, 62, 62.1, 62.3,
  62.7, 61.6, 62.7, 63.1, 65.3, 67.7, 73.1, 72.7, 76.1, 77,
  80.8, 84.2)
Ireland <- c(37.1, 35.7, 34.2, 34.6, 34.5, 34.4, 34.4, 34.2,
  30.7, 31.7, 32.7, 32.4, 31, 31.5, 32.3, 31.1, 29.5, 29.7,
  29.9, 29.2, 27.3, 27.8, 27.5, 26.1, 24.8, 24.9, 24.6, 28.8,
  25, 27.5, 32.4, 38.9, 44.3, 51.9, 60.1, 62.4, 65.5, 78.7,
  79.7, 90.5)
Italy <- c(109.2, 110.7, 111, 109, 108.8, 111.3, 110.8, 110,
  105.7, 108, 108.1, 108.8, 104.4, 107.2, 108.8, 108.2, 103.9,
  108, 110.4, 108.6, 105.9, 108.2, 109.9, 109.1, 106.6, 107.3,
  107.2, 105.9, 103.6, 106.1, 105.7, 105.1, 106.3, 112.2, 114.2,
  117.1, 116, 118.1, 119, 119.6)
Greece <- c(103.4, 103.3, 102.2, 100.9, 103.7, 101.7, 101.2,
  100.6, 101.7, 99.4, 99.2, 97.2, 97.4, 98.9, 102.2, 101.5,
  98.9, 98.6, 102, 102.5, 109, 107.6, 107.8, 106.9, 106.8,
  108.2, 106.8, 105.3, 105.8, 106.5, 107.2, 108.8, 110.9, 118.6,
  123.2, 125.7, 127.9, 133.2, 135.6, 140.1)
Great.Britain <- c(41, 38.9, 38.7, 37.7, 37.7, 37, 37.3, 37,
  37.5, 36.7, 37.4, 38.2, 39, 38.9, 39.3, 39.4, 40.9, 40.1,
  41.4, 41.8, 42.5, 42, 43.3, 43.2, 43.4, 42.9, 43.5, 43.6,
  44.5, 43.6, 44.6, 45.8, 52.1, 55.8, 59.3, 63.6, 68.1, 71.2,
  73.2, 75.1)
Spain <- c(59.3, 57.9, 57.4, 56.4, 55.5, 54.6, 54.2, 53.1, 52.5,
  51.3, 51.4, 49.6, 48.7, 48.7, 48, 46.7, 46.2, 45.8, 44.7,
  43.6, 43, 42.2, 40.9, 40.6, 39.6, 39.5, 38.9, 37.5, 36.1,
  35.3, 35.8, 36.7, 39.8, 43.1, 47.1, 49.6, 53.2, 54.9, 56.8,
  57.7)
```

Príklad 7.7

V tomto príklade využijeme databázu `survey` (z knižnice `MASS`), ktorá obsahuje odpovede 237 študentov z Univerzity v Adelaide na rôzne otázky. Predpokladajme, že sme týchto študentov vybrali na základe náhodného výberu. Odhadnite podiel mužov na danej univerzite spolu s intervalom spoľahlivosti pre tento podiel.

Príklad 7.8

Univerzita v Montane realizovala ekonomický prieskum na náhodnej vzorke 209 respondentov, ktorých sa pýtali otázky ohľadne ekonomickej situácie – a či sa ich finančná situácia zlepšila, zhoršila alebo je rovnaká ako pred rokom. Údaje sú dostupné online [22.06.2013] na stránke <http://lib.stat.cmu.edu/DASL/Datafiles/montanadat.html> (za rok 1992). Vypočítajte podiel rezidentov v Montane, ktorí si myslia, že ich finančná situácia sa zhoršila oproti predchádzajúcemu roku a vypočítajte interval spoľahlivosti ($\alpha = 0.05$).

Príklad 7.9

Databáza `samhda` (knižnica `UsingR`) obsahuje rôzne údaje o 600 žiakoch základných a stredných škôl (6. ročník, 8. ročník a 10. ročník), týkajúce sa ich zdravotných návykov. Prieskum bol realizovaný v roku 1996 organizáciou WHO. Okrem iného, táto databáza obsahuje údaje o skúsenostiach respondentov s marihuanou – či túto drogu aspoň raz vyskúšali (kód 1) alebo nie (kód 2). Ak predpokladáme, že vzorka bola vyberaná náhodne, viete odhadnúť v akom intervale sa pohybuje percento amerických študentov, ktorí majú skúsenosti s marihuanou?

Príklad 7.10

V tomto príklade opäť využijeme databázu `survey` (z knižnice `MASS`), ktorá obsahuje odpovede 237 študentov z Univerzity v Adelaide na rôzne otázky. Odhadnite strednú hodnotu veku študentov na univerzite a následne otestujte, či priemerný vek študentov univerzity je 18 rokov a viac. Vychádzame z predpokladu, že študenti boli do výskumnej vzorky vyberaní náhodne.

Príklad 7.11

V databáze `exec.pay` (knižnica `UsingR`) máme k dispozícii mzdy riaditeľov 199 podnikov z USA v 10 000 USD za rok 2000. Vypočítajte, či stredná hodnota ročného príjmu riaditeľa v USA je viac ako 300 000 USD. Pre odhad strednej hodnoty mzdy vypočítajte aj obojstranné intervaly spoľahlivosti. Predpokladáme, že vzorka bola vyberaná náhodne.

Príklad 7.12

K dispozícii máme 1000 pozorovaní o finančnej situácii respondentov prieskumu realizovaného v roku 2001 americkým rezervným systémom FED (dáta `cfb`, knižnica `UsingR`)[29]. Vypočítajte, či je rozumné predpokladať, že ľudia zarábajúci ročne 40 000 USD a viac (premenná `INCOME`) majú strednú hodnotu úspor väčšiu ako 7 000 USD (premenná `SAVING`).

Príklad 7.13

V tomto príklade využijeme dáta zo 4 európskych akciových indexov, ktoré sú dostupné v databáze `EuStockMarkets` (knižnica `datasets`). Ide o denné uzatváracie

[29] Uvedené dáta prešli určitou transformáciou (bootstrap), čo však pre nás nie je podstatné. Podstatné je, že túto vzorku môžeme považovať za náhodnú.

ceny indexov DAX (Nemecko), SMI (Švajčiarsko), CAC (Francúzsko) a FTSE (Anglicko) za obdobie od roku 1991 do roku 1998. Najprv z údajov vypočítajte tzv. spojité výnosy podľa vzťahu $\ln(P_{t+1}/P_t)$, kde P_t je uzatváracia cena indexu v čase t. Následne zistite, či priemerný denný výnos týchto indexov je menší ako 0.1 %.

Príklad 7.14

Vráťme sa k údajom o zadlženosti krajín EÚ. Pracovať budeme opäť s krajinami, o ktorých sa predpokladá výskyt problémov s výškou štátneho dlhu (uvažujeme o rokoch prvej dekády druhého milénia). Tieto krajiny sú známe pod mierne degradujúcim akronymom „PIIGGS“ – **P**ortugal (Portugalsko), **I**reland (Írsko), **I**taly (Taliansko), **G**reece (Grécko), **G**reat Britain (Veľká Británia), **S**pain (Španielsko). Z voľne dostupných zdrojov je možné stiahnuť údaje za ukazovateľ D/HDP, teda pomer štátneho dlhu (D) a hrubého domáceho produktu (HDP) krajiny (v angl. *debt-to-GDP ratio*).

Ako už bolo uvedené vyššie, podľa tzv. Paktu stability a rastu je pre členské krajiny EÚ stanovená hranica verejného dlhu v pomere k HDP na 60 %. Ak predpokladáme, že tieto pozorovania sú pre každú z krajín *iid*, na hladine významnosti 5 % zistite, či pre jednotlivé krajiny môžeme predpokladať, že ich zadlženosť je nižšia ako 60 % HDP.

Príklad 7.15

Opäť využijeme databázu `survey` (z knižnice `MASS`), ktorá obsahuje odpovede 237 študentov z Univerzity v Adelaide na rôzne otázky. V tomto príklade uvažujeme, že ide o náhodný výber študentov. Populáciu tak predstavujú všetci študenti Univerzity v Adelaide. Zaujíma nás, či na základe náhodnej vzorky je možné predpokladať, že stredná hodnota výšky mužov je rovnaká ako stredná hodnota výšky žien.

Príklad 7.16

K dispozícii máme databázu obsahujúcu údaje o baseballových hráčoch z roku 1986 a niektoré údaje z roku 1987 (knižnica `vcd`, databáza `Baseball`). Zistite, či strednú hodnotu mzdy hráčov z východnej divízie môžeme považovať za väčšiu, ako strednú hodnotu mzdy hráčov zo západnej divízie. Mzdy sú uvedené za rok 1987 v tis. USD (premenná `sal87`) a príslušnosť hráča k divízii je za rok 1986 (premenná `div86`). Používajú sa mzdy z roku 1987 aby sme prípadne mohli sledovať, či odrážajú výkony hráčov v predošlom súťažnom ročníku. Pri oboch vzorkách predpokladáme, že tieto hodnoty sú *iid*.

Príklad 7.17

Spoločnosť prevádzkujúca predajne s oblečením má zastúpenie v každom z krajských miest Slovenska. Pre potreby kvartálneho hodnotenia predajní uskutočnila náhodný výber kupujúcich zákazníkov, pričom sa okrem iného zaujímala o spokojnosť zákazníkov s predajcami v prevádzke. Mieru tejto spokojnosti zákazníci vyjadrili na škále od 1 (nespokojný) do 10 (spokojný). V každej prevádzke sa vybralo 33 zákazníkov. Údaje sú v nasledujúcich vektoroch, `satisfc` (spokojnosť), `gend` (pohlavie), `city` (mesto). Overte, či je rozumné predpokladať rovnosť spokojnosti v nasledujúcich dvojiciach miest: Banská Bystrica – Bratislava, Banská Bystrica – Košice, Prešov – Košice, Žilina – Trenčín, Žilina – Trnava.

```
satisfc <- c(7, 1, 6, 5, 7, 5, 10, 6, 10, 1, 5, 5, 7, 5, 5, 7,
  5, 10, 9, 9, 8, 3, 2, 4, 7, 10, 1, 7, 5, 6, 4, 7, 6, 4, 10, 6,
  3, 3, 5, 7, 6, 5, 7, 10, 6, 6, 1, 5, 7, 6, 7, 9, 3, 9, 5, 10,
  5, 9, 5, 1, 6, 6, 7, 6, 5, 9, 7, 3, 6, 7, 7, 7, 7, 6, 3, 3, 1,
  10, 7, 10, 1, 10, 5, 7, 6, 5, 7, 7, 7, 6, 7, 5, 4, 6, 3, 7, 6,
  8, 7, 6, 3, 6, 7, 9, 6, 7, 1, 7, 7, 6, 6, 7, 5, 7, 8, 2, 3, 4,
  7, 7, 8, 2, 5, 5, 7, 5, 6, 5, 5, 5, 7, 7, 7, 4, 6, 9, 8, 4, 5,
  5, 7, 6, 6, 4, 10, 3, 3, 3, 4, 2, 4, 7, 2, 4, 7, 5, 5, 9, 4,
  5, 2, 4, 5, 6, 6, 5, 5, 7, 7, 7, 5, 6, 6, 9, 5, 7, 2, 8, 6, 5,
  5, 7, 1, 1, 7, 3, 10, 6, 6, 6, 7, 6, 8, 3, 5, 7, 6, 7, 9, 6,
  2, 8, 8, 8, 6, 5, 7, 1, 6, 10, 8, 3, 5, 8, 6, 7, 1, 5, 5, 6,
  5, 7, 7, 5, 2, 7, 4, 7, 7, 5, 6, 8, 6, 8, 5, 7, 2, 10, 10, 9,
  9, 4, 8, 9, 9, 9, 7, 3, 9, 9, 6, 10, 10, 1, 1, 10, 10, 9, 10,
  9, 10, 2, 10, 9)
city <- c("ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba",
  "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba",
  "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba", "ba",
  "ba", "ba", "ba", "ba", "po", "po", "po", "po", "po", "po",
  "po", "po", "po", "po", "po", "po", "po", "po", "po", "po",
  "po", "po", "po", "po", "po", "po", "po", "po", "po", "po",
  "po", "po", "po", "po", "po", "po", "po", "tr", "tr", "tr",
  "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr",
  "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr",
  "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr", "tr",
  "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn",
  "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn",
  "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn", "tn",
  "tn", "tn", "tn", "zi", "zi", "zi", "zi", "zi", "zi", "zi",
  "zi", "zi", "zi", "zi", "zi", "zi", "zi", "zi", "zi", "zi",
  "zi", "zi", "zi", "zi", "zi", "zi", "zi", "zi", "zi", "zi",
  "zi", "zi", "zi", "zi", "zi", "zi", "ni", "ni", "ni", "ni",
  "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni",
  "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni",
  "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ni", "ke",
  "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke",
  "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke",
  "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke", "ke",
  "ke", "ke", "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb",
```

```
  "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb",
  "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb", "bb",
  "bb", "bb", "bb", "bb", "bb")
gend <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1,
  1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
  1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1,
  1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1,
  0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0,
  1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1,
  0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1,
  1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1)
```

Príklad 7.18

Na internetovej stránke máme k dispozícii údaje o zamestnanosti žien (v percentách voči celkovej zamestnanosti) vo vybraných mestách USA za rok 1968 a 1972 (dostupné online [22.06.2013] <http://lib.stat.cmu.edu/DASL/Datafiles/LaborForce.html>). Namerané hodnoty považujeme za hodnoty získané z náhodného výberu. Variabilita zamestnanosti žien je charakteristika, ktorá nám hovorí o tom, nakoľko je medzi vybranými štátmi rovnorodá. Vyššia variabilita naznačuje, že medzi jednotlivými štátmi existujú značné rozdiely v zamestnanosti žien. V ďalšej etape by nás mohlo zaujímať čo je dôsledkom tejto rôznorodosti. Prvou úlohou je zistiť, či je možné považovať variabilitu (populačnú) v zamestnanosti medzi rokmi 1968 a 1972 za rovnakú. Druhou úlohou je zistiť, či došlo k zmene zamestnanosti žien v USA v roku 1972 oproti roku 1968. Údaje sú uvedené v nasledujúcej tabuľke (Tabuľka 18).

Tabuľka 18: Zamestnanosť žien vo vybraných mestách USA v rokoch 1968 a 1972

Mesto	1968	1972	Mesto	1968	1972
N.Y.	0.42	0.45	Wash., D.C.	0.42	0.52
L.A.	0.5	0.5	Cinn.	0.51	0.53
Chicago	0.52	0.52	Baltimore	0.49	0.57
Philadelphia	0.45	0.45	Newark	0.54	0.53
Detroit	0.43	0.46	Minn/St. Paul	0.5	0.59
San Francisco	0.55	0.55	Buffalo	0.58	0.64
Boston	0.45	0.6	Houston	0.49	0.5
Pitt.	0.34	0.49	Patterson	0.56	0.57
St. Louis	0.45	0.35	Dallas	0.63	0.64
Connecticut	0.54	0.55			

Zdroj: United States Department of Labor Statistics

Príklad 7.19

Malá pekáreň sa pokúša získať priazeň zákazníkov s novým druhom pečiva. Ak pekáreň predá aspoň 1/3 z toho čo predá zo starého pečiva, tak sa výroba nového oplatí. Ak predá aspoň 1/2, predaj bude považovaný za úspech. Za účelom vyhodnotenia potenciálnej úspešnosti si majiteľ pripravil jednoduchý experiment. Spomedzi zákazníkov, ktorí si kupovali starý druh pečiva, náhodne vybral 50 a na ochutnávku im ponúkol v náhodnom poradí aj nové pečivo. Takýchto experimentov realizuje spolu 3, pre každý z troch nových druhov pečiva. Výsledky sú nasledovné:

- V porovnaní so starým druhom pečiva by nové pečivo A uprednostnilo 25 zákazníkov.
- V porovnaní so starým druhom pečiva by nové pečivo B uprednostnilo 22 zákazníkov.
- V porovnaní so starým druhom pečiva by nové pečivo C uprednostnilo 30 zákazníkov.

Zistite, pri ktorých druhoch pečiva si s 95 % konfidenciou môžeme byť istí, že pekáreň predá požadované množstvo pečiva.

Príklad 7.20

Výrobca detských hračiek sa rozhoduje o farbe jedného z jeho produktov, pričom do úvahy pripadá červená a modrá farba. Z jednoduchého experimentu (v ktorom si 30 detí malo vybrať z dvoch totožných hračiek avšak rôznej farby) vyplynulo, že 21 detí si vybralo hračku modrej farby. Zistite, či tieto údaje poukazujú na preferencie detí ohľadne farby danej hračky. Hladinu významnosti stanovte na 1 %.

Príklad 7.21

Developerská spoločnosť zvažuje výstavbu nového bytového komplexu a menšieho obchodného centra na sídlisku. Obáva sa, že ak začne s realizáciou investičného zámeru, občianske združenia spustia petíciu proti tejto výstavbe. Na základe vlastných skúseností vie, že získať potrebný počet podpisov nie je pre občianske združenie jednoduché. Taktiež vie, že ak viac ako 75 % ľudí sa v prieskume vyjadrí, že nesúhlasí s touto výstavbou, podpisová akcia má reálnu šancu uspieť. Z týchto dôvodov sa spoločnosť rozhodla uskutočniť prieskum, kde na náhodne vybranej vzorke 200 respondentov zistila, že 132 by v danom čase nesúhlasila

s výstavbou. Zistite, či spoločnosť má dôvod sa obávať, že prípadná petičná akcia by mohla skončiť úspešne. Uvažujte o 1 % hladine významnosti.

Príklad 7.22

V tomto príklade sa vrátime k prieskumu, ktorý bol realizovaný na Univerzite v Montane. Ide o ekonomický prieskum na náhodnej vzorke 209 respondentov, ktorých sa pýtali, ako vnímajú ekonomickú situáciu a či sa ich finančná situácia zlepšila, zhoršila alebo je rovnaká ako pred rokom. Zistite, či muži a ženy vnímajú rozdielne, či sa ich finančná situácia oproti predchádzajúcemu roku zhoršila (premenná `FIN = 1`). Potom zistite, či rozdielne vnímajú zlepšenie ekonomickej situácie v krajine (premenná `STAT = 1`). Potrebné údaje (pripomíname, že ide o údaje za rok 1992) sú voľne dostupné online [22.06.2013] na stránke <http://lib.stat.cmu.edu/DASL/Datafiles/montanadat.html>.

Príklad 7.23

K dispozícii máme dáta zo 4 európskych akciových indexov, ktoré sú dostupné v databáze `EuStockMarkets` (knižnica `datasets`). Ide o denné uzatváracie ceny indexov DAX (Nemecko), SMI (Švajčiarsko), CAC (Francúzsko) a FTSE (Anglicko) za obdobie od roku 1991 do roku 1998. Z týchto údajov sa použitím nasledujúcich transformácii dostaneme k tzv. nadmerným spojitým výnosom.

```
> library(datasets)
> attach(data.frame(EuStockMarkets))
-----------------------------------------------------------------
> rCAC <- ar.ols(diff(log(CAC)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rDAX <- ar.ols(diff(log(DAX)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rFTSE <- ar.ols(diff(log(FTSE)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rSMI <- ar.ols(diff(log(SMI)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
```

Spôsob úpravy týchto údajov nie je v tejto situácii tak dôležitý. Pripomenieme, že operácia `diff(log())` vytvorila z uzatváracích cien spojité výnosy. Potom došlo k ďalšej transformácii, po ktorej je rozumné predpokladať, že takto získané výnosy spĺňajú podmienku *iid*. Tieto výnosy budeme označíme ako nadmerné výnosy. Nadmerný výnos si môžeme interpretovať nasledovne: ide o výnos, ktorý dosiahneme po odpočítaní očakávaného výnosu. Ak napríklad očakávame výnos 0.01 a skutočný výnos bude –0.01, rozdiel bude –0.02 a bude predstavovať nadmerný výnos.

Vašou úlohou je zistiť, či môžeme niektoré z týchto nadmerných výnosov považovať za realizácie z toho istého rozdelenia pravdepodobnosti. Rovnaké rozdelenia nám dajú informáciu o podobnosti štruktúry výnosov na rôznych trhoch. Všimnime si, že sa neporovnávajú len stredné hodnoty alebo len rozptyly (prípadne iné parametre), ale priamo celé rozdelenie, ktoré je determinované skupinou rôznych parametrov (napr. aj šikmosťou, špicatosťou, ...).

Príklad 7.24

Pokračujme v analýze akciových indexov z databázy `EuStockMarkets` (knižnica `datasets`). Mnoho ekonomických modelov predpokladá, že výnosy pochádzajú z normálneho rozdelenia pravdepodobnosti (napr. Markowitzova teória portfólia, Black-Scholesov model oceňovania opcií). V tomto príklade sa pokúste zistiť, či nadmerné výnosy z jednotlivých akciových trhov je možné považovať za realizácie z normálneho rozdelenia (teoreticky môžu vo všetkých prípadoch nadmerné výnosy pochádzať z rovnakého typu rozdelenia, avšak s rôznymi parametrami).

Príklad 7.25

V tomto príklade budeme pracovať s údajmi z databázy RegDat Štatistického úradu Slovenskej republiky, konkrétne s reálnou mzdou a nezamestnanosťou po okresoch v rámci SR za rok 2010. Jednotlivé okresy sú priradené ku krajom prostredníctvom premennej `kraj` (Bratislavský [BA] = 1, Trnavský [TT] = 2, Trenčiansky [TN] = 3, Nitriansky [NR] = 4, Žilinský [ZA] = 5, Banskobystrický [BB] = 6, Prešovský [PO] = 7, Košický [KE] = 8). Zrejme môžeme očakávať, že reálna mzda v Bratislavskom kraji je vyššia ako priemerná reálna mzda v SR. Taktiež nezamestnanosť v Banskobystrickom a Prešovskom kraji bude zrejme vyššia ako je celoslovenský priemer. Pokúste sa zistiť, ktoré úrovne nezamestnanosti a reálnej mzdy môžeme v okresoch považovať za odľahlé hodnoty – zaujímajú nás extrémne veľké hodnoty. Uvažujeme o hladine významnosti 5 %.

```
okres <- c("OkresBratislavaI", "OkresBratislavaII",
  "OkresBratislavaIII", "OkresBratislavaIV", "OkresBratislavaV",
  "OkresMalacky", "OkresPezinok", "OkresSenec",
  "OkresDunajskáStreda", "OkresGalanta", "OkresHlohovec",
  "OkresPiešťany", "OkresSenica", "OkresSkalica", "OkresTrnava",
  "OkresBánovcenadBebravou", "OkresIlava", "OkresMyjava",
  "OkresNovéMestonadVáhom", "OkresPartizánske",
  "OkresPovažskáBystrica", "OkresPrievidza", "OkresPúchov",
  "OkresTrenčín", "OkresKomárno", "OkresLevice", "OkresNitra",
  "OkresNovéZámky", "OkresŠaľa", "OkresTopoľčany",
```

```
  "OkresZlatéMoravce", "OkresTvrdošín", "OkresŽilina",
  "OkresBytča", "OkresČadca", "OkresDolnýKubín",
  "OkresKysuckéNovéMesto", "OkresLiptovskýMikuláš",
  "OkresMartin", "OkresNámestovo", "OkresRužomberok",
 "OkresTurčianskeTeplice", "OkresVeľkýKrtíš", "OkresZvolen",
  "OkresŽarnovica", "OkresŽiarnadHronom", "OkresBanskáBystrica",
  "OkresBanskáŠtiavnica", "OkresBrezno", "OkresDetva",
  "OkresKrupina", "OkresLučenec", "OkresPoltár", "OkresRevúca",
  "OkresRimavskáSobota", "OkresStaráĽubovňa", "OkresStropkov",
  "OkresSvidník", "OkresVranovnadTopľou", "OkresBardejov",
  "OkresHumenné", "OkresKežmarok", "OkresLevoča",
  "OkresMedzilaborce", "OkresPoprad", "OkresPrešov",
  "OkresSabinov", "OkresSnina", "OkresSpišskáNováVes",
  "OkresTrebišov", "OkresGelnica", "OkresKošiceI",
  "OkresKošiceII", "OkresKošiceIII", "OkresKošiceIV",
  "OkresKošice-okolie", "OkresMichalovce", "OkresRožňava",
  "OkresSobrance")
kraj <- c(1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3,
  3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5,
  5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7,
  7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8)
real_mzda <- c(1136.262365, 1182.604046, 1045.361376,
  945.5485251, 956.2427591, 803.8499242, 611.3537118,
  691.560467, 597.0947331, 647.0011585, 726.3167276,
  654.1306479, 616.7008288, 726.3167276, 800.2851796,
  569.4679619, 657.6953926, 566.7944034, 691.560467,
  571.2503342, 616.7008288, 644.3276, 713.8401212, 689.7780946,
  519.5615364, 712.0577489, 684.4309776, 573.0327065,
  663.0425096, 571.2503342, 603.3330363, 560.5561002,
  732.5550307, 619.3743873, 544.5147491, 630.0686213,
  702.254701, 602.4418501, 756.6170573, 531.1469566,
  726.3167276, 556.1001693, 555.2089832, 675.5191159,
  708.4930042, 652.3482756, 696.0163978, 553.4266108,
  581.0533821, 653.2394617, 509.7584885, 539.1676321,
  485.696462, 564.1208448, 517.7791641, 553.4266108,
  499.9554407, 537.3852598, 513.3232332, 438.463595,
  562.3384725, 544.5147491, 598.8771054, 509.7584885,
  657.6953926, 652.3482756, 552.5354247, 482.1317173,
  581.0533821, 548.97068, 517.7791641, 805.6322966, 877.8183763,
  622.939132, 786.0262009, 606.0065948, 606.8977809, 612.244898,
  626.5038767)
nezamestnanost <- c(3.18, 4.6, 3.8, 3.58, 3.98, 7.46, 5.96,
  5.71, 11.01, 6.29, 7.87, 6.98, 10.92, 8.76, 6.15, 9.34, 6.72,
  8.76, 7.63, 12.27, 12.49, 12.05, 7.51, 7.25, 16.34, 14.55,
  7.52, 12.48, 10.11, 11, 10.31, 13.02, 8.41, 14.87, 11.04,
  13.87, 12.23, 11.18, 9.15, 13.58, 11.17, 11.47, 23.71, 9.24,
  20.23, 14.39, 8.95, 17.1, 18, 16.16, 19.95, 23.2, 22.06,
  28.83, 33.64, 13.63, 17.14, 18.8, 19.68, 19.43, 15.66, 26.18,
  18.41, 19.4, 10.65, 16.6, 25.71, 19.38, 16.28, 24.42, 19.14,
  8.07, 9.27, 8.59, 7.82, 21.27, 17.21, 26.82, 20.34)
```

Príklad 7.26

Vráťme sa opäť k databáze `EuStockMarkets` (knižnica `datasets`), v ktorej máme k dispozícii dáta zo 4 európskych akciových indexov. Ide o denné uzatváracie ceny indexov DAX (Nemecko), SMI (Švajčiarsko), CAC (Francúzsko) a FTSE (Anglicko) za obdobie od roku 1991 do roku 1998. Tak ako už bolo načrtnuté v texte, po sebe idúce výnosy finančných aktív na trhu by mali byť v zmysle teórie efektívnych trhov nezávislé (čo je nutná podmienka k tomu, aby boli náhodné). Rozhodnite, či je možné považovať za sebou nasledujúce výnosy skúmaných akciových indexov za nezávislé.

Príklad 7.27

Je známe, že na Slovensku je reálna mzda v Bratislavskom kraji vyššia ako priemerná reálna mzda v SR. Nezamestnanosť je vyššia ako priemer v Banskobystrickom a Prešovskom kraji (pracujeme s predchádzajúcou databázou, ktorá je dostupná z databázy RegDat Štatistického úradu Slovenskej republiky, konkrétne ide o reálnu mzdu a nezamestnanosť po okresoch v rámci SR za rok 2010).

Na základe Wilcoxonovho znamienkového testu zistite, či môžeme považovať hodnotu 550 EUR za medián reálnej mzdy dosahovanej v okresoch SR. Taktiež overte, či je v okresoch SR mediánová nezamestnanosť nižšia ako 10 %.

Príklad 7.28

V tomto príklade overte hypotézu, že akciové indexy dosahujú mediánovú nadmernú výnosnosť vyššiu ako 0. Pracovať budeme s už známou databázou `EuStockMarkets` (knižnica `datasets`), v ktorej máme k dispozícii dáta zo 4 európskych akciových indexov. Ide o denné uzatváracie ceny indexov DAX (Nemecko), SMI (Švajčiarsko), CAC (Francúzsko) a FTSE (Anglicko) za obdobie od roku 1991 do roku 1998. Na výpočet nadmerného výnosu použijeme obdobný postup ako v jednom z predošlých príkladov:

```
> library(datasets)
> attach(data.frame(EuStockMarkets))
-----------------------------------------------------------------
> rCAC <- ar.ols(diff(log(CAC)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rDAX <- ar.ols(diff(log(DAX)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rFTSE <- ar.ols(diff(log(FTSE)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rSMI <- ar.ols(diff(log(SMI)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
```

Zopakujeme, že nadmerný výnos si interpretujeme ako výnos, ktorý dosiahneme po odpočítaní očakávaného výnosu. Ak napríklad očakávame výnos 0.01 a skutočný výnos bude 0.03, rozdiel bude 0.02 a bude predstavovať nadmerný výnos.

Príklad 7.29

K dispozícii máme priemerné ročné mzdy mužov a žien z 51 štátov USA za rok 2009 (dostupné online [22.06.2013) <http://www.census.gov/prod/2010pubs/acsbr09-3.pdf>). Spolu ide o 102 prierezových údajov, ktoré však zrejme nemajú normálne rozdelenie (overte). Zistite, či mzdy mužov (označené ako 1) je možné považovať vo všeobecnosti za vyššie ako mzdy žien.

```
stat <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
  17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31,
  32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46,
  47, 48, 49, 50, 51, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
  14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28,
  29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43,
  44, 45, 46, 47, 48, 49, 50, 51)
mzda <- c(61993, 48389, 41916, 49174, 43425, 39122, 55116,
  40359, 40621, 49439, 47983, 56902, 39509, 42667, 36465, 48492,
  48038, 45911, 50236, 39648, 45234, 39516, 39562, 36977, 41660,
  49336, 42156, 44572, 57738, 51019, 42494, 37528, 46747, 44563,
  39174, 51305, 44812, 40748, 41331, 59387, 50837, 42634, 40693,
  43631, 40440, 48066, 39830, 40231, 45800, 44174, 47828, 54698,
  40019, 34651, 40584, 35691, 32109, 44937, 32576, 32578, 39248,
  38058, 45062, 31222, 33665, 28640, 38025, 37645, 35977, 39354,
  31010, 35276, 30562, 30578, 28515, 31993, 37841, 32314, 34121,
  44166, 39017, 32341, 28506, 35301, 33616, 29413, 38521, 33611,
  30481, 30658, 43900, 37527, 31431, 29742, 31762, 29122, 34542,
  28461, 27855, 31186, 29350, 31308)
pohlavie <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0)
```

Príklad 7.30

Máme k dispozícii údaje o priemerných mzdách učiteľov v tis. USD a výdavkoch na študentov v tis. USD za jednotlivé štáty v USA (údaje sú prevzaté z Moore – McCabe, 1989, dostupné online [*22.06.2013*] na <http://lib.stat.cmu.edu/DASL/Datafiles/Shoppers.html>). Overte hypotézu, že (A) rozdelenie miezd ako aj (B) rozdelenie výdavkov na študentov sú vo všetkých regiónoch v USA rovnaké (máme 9 regiónov označených premennou `Region`: `NE` – Northeast, `MA` – Mid-Atlantic, `SA` – South Atlantic, `ESC` – East South Central, `ENC` – East

North Central, WSC – West South Central, WNC – West North Central, MN – Mountain States, PA – Pacific States).

```
stat <- c("Maine", "Vt.", "R.I.", "N.Y.", "Pa.", "Ind.",
  "Mich.", "Minn.", "Mo.", "S.Dak", "Kan.", "Md.", "Va.",
  "N.C.", "Ga.", "Ky.", "Ala.", "Ark.", "Okla.", "Mont.",
  "Wyo.", "N.Mex.", "Utah", "Wash.", "Calif.", "Hawaii", "N.H.",
  "Mass.", "Conn.", "N.J.", "Ohio", "Ill.", "Wis.", "Iowa",
"N.Dak", "Neb.", "Del.", "D.C.", "W.Va.", "S.C.", "Fla.",
  "Tenn.", "Miss.", "La.", "Tex.", "Idaho", "Colo.", "Ariz.",
  "Nev.", "Oreg.", "Alaska")
region <- c("NE", "NE", "NE", "MA", "MA", "ENC", "ENC", "WNC",
  "WNC", "WNC", "WNC", "SA", "SA", "SA", "SA", "ESC", "ESC",
  "WSC", "WSC", "MN", "MN", "MN", "MN", "PA", "PA", "PA", "NE",
  "NE", "NE", "MA", "ENC", "ENC", "ENC", "WNC", "WNC", "WNC",
  "SA", "SA", "SA", "SA", "SA", "ESC", "ESC", "WSC", "WSC",
  "MN", "MN", "MN", "MN", "PA", "PA")
prijem <- c(19.6, 20.3, 29.5, 30.7, 25.9, 24.3, 30.2, 27.4, 22,
  18.1, 27.6, 27.2, 23.4, 22.8, 22.1, 20.9, 22.9, 19.5, 21.4,
  22.5, 27.2, 22.6, 22.3, 26, 29.1, 25.8, 20.3, 26.8, 26.6,
  27.2, 24.5, 27.2, 26.5, 21.7, 20.8, 20.9, 24.6, 34, 20.6,
  21.6, 22.3, 21.8, 18.4, 20.5, 25.2, 21, 25.9, 26.6, 25.6,
  25.8, 41.5)
vydavky <- c(3.35, 3.55, 4.67, 5.71, 4.17, 3.16, 3.78, 3.98,
  3.16, 2.97, 3.91, 4.35, 3.59, 3.37, 2.98, 2.85, 2.73, 2.64,
  2.75, 3.95, 5.44, 3.4, 2.3, 3.71, 3.61, 3.77, 3.11, 4.64,
  4.89, 5.54, 3.55, 3.62, 4.25, 3.57, 3.06, 3.29, 4.52, 5.02,
  2.82, 2.92, 3.73, 2.53, 2.31, 3.12, 3.43, 2.51, 4.04, 2.83,
  2.93, 4.12, 8.35)
```

Príklad 7.31

V jednom z predchádzajúcich príkladov sme riešili, či existuje rozdiel v mzde baseballových hráčov v závislosti od príslušnosti ich tímu k divízii. Za týmto účelom sme využili jednostranný *t*-test. Nulovú hypotézu sme zamietli v prospech alternatívnej, že mzdy hráčov z východnej divízie sú v priemere vyššie ako mzdy hráčov zo západnej divízie. V tomto príklade overte rovnakú hypotézu, avšak tentoraz s použitím neparametrických testov (Kruskal – Wallisov test a Mann – Whitney – Wilcoxonov test). Pracujeme s databázou o baseballových hráčoch z roku 1986 a niektoré údaje z roku 1987 (knižnica `vcd`, databáza `Baseball`). Mzdy sú uvedené za rok 1987 v tis. USD (premenná `sal87`) a príslušnosť hráča k divízii je za rok 1986 (premenná `div86`).

Príklad 7.32

Vráťme sa k údajom z príkladu Príklad 7.30, ktorý obsahuje údaje o priemerných mzdách učiteľov v tis. USD a výdavkoch na študentov v tis. USD za jednotlivé štáty v USA.

Mali sme možnosť z grafickej vizualizácie týchto dvoch premenných vidieť, že ich variabilita je v jednotlivých regiónoch USA rozdielna. Prostredníctvom neparametrických testov (Levenov test a Brown – Forsythov test) overte, či sú tieto rozdiely štatisticky významné.

Príklad 7.33

Pokračujme s databázou o baseballových hráčoch (knižnica `vcd`, databáza `Baseball`). V jednom z predchádzajúcich príkladov sme ukázali (s využitím Mann – Whitney – Wilcoxonovho a Kuskal – Wallisovho testu), že je možné pozorovať štatisticky významné rozdiely v mzdách hráčov v závislosti od divízie. V tomto príklade sa pokúste o overenie tej istej hypotézy, ale využite metódu ANOVA.

Príklad 7.34

V predchádzajúcich príkladoch sme riešili, či existujú rozdiely (A) v mzdách učiteľov a (B) výdavkoch na študentov v rámci jednotlivých regiónov USA, rovnako ako rozdiely vo variabilite v týchto dvoch premenných. Tentoraz sa pokúste zistiť, či existuje lineárna závislosť medzi výškou mzdy učiteľov a výdavkami na študentov po jednotlivých štátoch USA. Pracujte so súborom údajov, ktorý sme poskytli v jednom z predchádzajúcich príkladov (Príklad 7.30).

Príklad 7.35

V tabuľke nižšie máme k dispozícii údaje o akciových indexoch a HDP krajín V4 (Česká republika, Maďarsko, Poľsko a Slovensko) za obdobie od Q1:1996 do Q4:2009. Zaujíma nás, či dochádza k spoločnému vývoju medzi akciovými trhmi a reálnou ekonomikou. Zjednodušene je možné myšlienku vzájomného vzťahu formulovať nasledovne. Investori na trhu majú určité očakávania o budúcom vývoji hodnoty akciových spoločností. Ak tieto očakávania hovoria o rastúcej hodnote, investori tieto očakávania premietnu do nákupu akcií a spôsobia tým nárast cien. Zároveň to ale znamená, že ak dochádza k nárastu vnímanej hodnoty podnikov, zrejme by malo dôjsť aj k vyššiemu ekonomickému rastu. Akciové trhy sa tak často považujú za určitý indikátor budúceho vývoja reálnej ekonomiky. Našim cieľom je v prehľadnej forme zobraziť (teda iba vizualizovať) vzťahy medzi vývojom akciových trhov a vývojom reálnej ekonomiky. Hodnoty v tabuľke sú zaznamenané v podobe tzv. spojitých výnosov.

Tabuľka 19: Vývoj HDP a akciových výnosov

hdp_cz	px	hdp_hu	bux	hdp_pl	wig	hdp_sk	sax
0.01245	0.14037	0.00246	0.3754	0.02179	0.20242	0.01721	0.1724
0.00782	0.03372	0.00511	0.0827	0.01178	0.05139	0.01634	-0.06689
0.00121	-0.04405	0.01271	0.14625	-0.03301	-0.05097	0.01535	-0.13375
-0.00300	0.03389	0.01157	0.26961	0.0617	0.13367	0.00295	0.15075
-0.01173	-0.13235	0.01310	0.22725	0.01373	-0.10052	0.00948	-0.21656
-0.00887	0.09159	0.01260	0.12411	0.01184	0.14954	0.01777	0.04096
0.00320	-0.07878	0.01298	0.03903	0.01598	-0.15169	-0.00021	0.04943
-0.00216	0.01939	0.01440	0.07895	0.01789	0.15739	0.01802	-0.27281
0.00064	-0.07823	0.01254	-0.10342	0.00886	-0.07539	-0.00485	-0.25170
-0.00278	-0.26078	0.01031	-0.53510	0.01201	-0.31968	-0.00823	-0.06949
0.00220	0.09131	0.00702	0.32201	-0.01616	0.05686	0.08315	-0.06935
0.00369	-0.01560	0.00860	-0.13890	0.01959	0.12275	-0.03981	-0.12470
0.00071	0.22041	0.00994	0.16675	0.01888	0.14680	-0.01507	0.01376
0.01156	0.03792	0.01481	0.03950	0.02045	-0.16827	-0.00890	-0.01799
0.00558	-0.02580	0.01424	0.26784	0.01208	0.26407	-0.00271	-0.06966
0.00825	0.27422	0.00826	0.12569	0.00774	0.24077	0.00922	0.01084
0.01732	-0.20499	0.01559	-0.18419	0.01251	-0.13020	0.01358	-0.05609
0.01257	-0.04521	0.00798	-0.00577	0.00066	-0.16822	0.00520	0.19456
0.00388	-0.04715	0.01454	-0.05220	0.01010	0.07296	0.00746	0.02669
0.00059	-0.11739	0.00916	-0.16549	0.00145	-0.27764	0.01059	-0.11532
0.00707	-0.01850	0.00951	0.00649	-0.00274	-0.07912	0.00771	0.20359
0.01000	-0.22993	0.01102	-0.08115	0.00659	-0.21761	0.00113	0.17024
-0.00072	0.17304	0.00700	0.14414	-0.00212	0.16688	0.02282	0.01468
0.00554	0.08288	0.01719	0.12896	0.00422	0.10271	0.00270	-0.07378
0.00500	-0.03828	0.00598	-0.11206	0.00494	-0.09611	0.01121	0.02152
-0.00028	0.06996	0.01227	-0.02099	0.00640	-0.15194	0.02657	-0.01176
0.01116	0.04031	0.00870	0.09353	0.00785	0.11790	0.00317	0.21157
0.01226	0.06736	0.01035	-0.04928	0.00676	-0.07243	0.01498	0.19437
0.01235	0.08235	0.01058	0.04716	0.01560	0.13538	0.01243	-0.07726
0.00153	0.1178	0.01113	0.13761	0.01350	0.16542	-0.00257	0.07266
0.01432	0.09062	0.01208	0.04919	0.01180	0.06346	0.02116	0.04845
0.00670	0.22305	0.01202	0.15862	0.01672	0.11301	0.00852	0.02321
0.01398	-0.03747	0.01121	0.04836	0.01319	-0.02037	0.00738	0.07348
0.01551	0.09823	0.01003	0.09183	0.00213	0.04546	0.02741	0.17690
0.01091	0.16457	0.00845	0.15335	0.01135	0.08148	0.01394	0.33560
0.01809	0.12414	0.00614	0.14879	0.00893	0.01908	0.01227	0.31750
0.01570	0.03507	0.01270	0.09117	0.00500	0.02421	0.02103	-0.02844
0.01620	0.18341	0.00709	0.20257	0.01538	0.20723	0.01639	0.05277
0.01763	0.01319	0.01266	-0.09904	0.01275	0.05267	0.00929	-0.10646
0.01603	0.03397	0.01087	0.10410	0.01544	0.07608	0.02787	0.00927
0.01999	-0.09168	0.01125	-0.07460	0.01693	0.00863	0.02370	-0.10067
0.01603	0.04025	0.00602	0.01613	0.01936	0.01004	0.01540	0.07478
0.01313	0.09320	0.00663	0.13400	0.01428	0.11833	0.03074	0.02216
0.02456	0.07474	-0.00368	-0.06011	0.01741	0.06901	0.02019	0.00624
0.00323	0.08231	-0.00021	0.21114	0.0156	0.0657	0.02155	-0.02022
0.01448	-0.02329	0.00333	-0.01802	0.01373	-0.03399	0.02558	0.04983
0.01013	-0.00066	0.00541	-0.07973	0.02026	-0.05011	0.05717	0.03394
0.00286	-0.15666	0.01009	-0.18932	0.01267	-0.14784	-0.01881	0.03278
0.00693	-0.04508	-0.00272	-0.06287	0.00935	-0.1402	0.01521	-0.05064
0.00164	-0.20817	-0.00926	-0.07741	0.00668	-0.08321	0.01210	0.03499
-0.00684	-0.33915	-0.02166	-0.43267	-0.00252	-0.28681	0.00347	-0.23284
-0.04216	-0.13516	-0.02943	-0.10044	0.00454	-0.16873	-0.07603	-0.07011
-0.00273	0.18072	-0.01399	0.32476	0.00568	0.20851	0.00792	-0.00647
0.00576	0.25328	-0.00573	0.27781	0.00618	0.16314	0.01146	-0.09837
0.00728	-0.03500	0.00247	0.04829	0.01080	0.08577	0.01663	-0.12173

Zdroj: vlastné spracovanie údajov z Eurostatu

Príklad 7.36

Vráťme sa k databáze o baseballových hráčoch (knižnica `vcd`, databáza `Baseball`). Už v predchádzajúcom príklade sme ukázali (s využitím Mann – Whitney – Wilcoxonovho a Kuskal – Wallisovho testu), že je možné pozorovať štatisticky významné rozdiely v mzdách hráčov v závislosti od divízie. Táto databáza však obsahuje množstvo iných údajov. Vypočítajte korelačné koeficienty medzi nasledujúcimi premennými: `atbat86`, `hits86`,

`homer86`, `runs86`, `rbi86`, `walks86`, `years`, `atbat`, `hits`, `homeruns`, `runs`, `rbi`, `walks`, `outs86`, `assist86`, `error86`, `sal87` (v zásade nás v tomto bode nemusí zaujímať, čo konkrétne premenné znamenajú, ide o technické cvičenie, ale ich presný opis je samozrejme možné nájsť po zadaní príkazu `?Baseball`). Pokúste sa korelačnú maticu medzi všetkými premennými vizualizovať nejakým prehľadným spôsobom.

Príklad 7.37

Uvažujme o nasledujúcom zjednodušenom príklade. Z normovaného normálneho rozdelenia si vygenerujte premennú x s počtom pozorovaní 1000. Následne vygenerujte premennú y dvoma jednoduchými spôsobmi:

1. $y = x^2$
2. $y = e^x$

Táto premenná je teda stanovená deterministicky v závislosti od hodnôt náhodnej premennej x. Pokúste sa zistiť, aká je závislosť medzi premennou x a y prostredníctvom vhodne zvoleného korelačného koeficientu (v príklade predpokladáme, že nepoznáme akým spôsobom boli premenné vytvorené a nazeráme na nich ako na dve náhodné premenné).

Príklad 7.38

Svetová banka zverejňuje každý rok prieskum o stave podnikateľského prostredia v 183 krajinách sveta pod názvom *Doing Business* (pri riešení môže čitateľ použiť ľubovoľný ročník prieskumu). Z tejto databázy sme vybrali nasledujúce premenné:

- `Rank` – ide o poradie krajiny v rámci tzv. *Ease of Doing Business* indexu (Index podnikateľského prostredia).
- `Start` – poradie krajiny podľa jednoduchosti s akou sa založí podnik.
- `Proced` – počet procedúr, ktoré je nutné vykonať pri založení podniku.
- `Time` – ako dlho trvá založenie podniku, merané v dňoch.

Zistite, či existuje závislosť medzi umiestnením krajiny v rámci indexu *Ease of Doing Business* a ostatnými zvolenými premennými. Tu je potrebné si uvedomiť, že celkové poradie (v premennej `Rank`) je ovplyvňované aj hodnotami v ostatných premenných. Niekedy však nevieme povedať akým spôsobom, resp. nakoľko. Korelačná analýza v tomto prípade slúži ako určitá analýza citlivosti.

Príklad 7.39

Máme k dispozícii údaje z prieskumu podnikateľských subjektov, ktorý bol realizovaný Svetovou bankou za rok 2009 v 30 krajinách východnej Európy a strednej Ázie (bližšie informácie na www.enterprisesurveys.org). Celkovo máme k dispozícii pozorovania za 11668 podnikateľských subjektov. Z množstva údajov z tohto prieskumu sme pre naše potreby vybrali nasledujúce premenné:

- `ownership_W` – v tejto premennej je odpoveď na otázku, či je jedným z vlastníkov podniku žena. Premenná nadobúda tri úrovne: áno, nie, neviem.
- `management_W` – táto premenná obsahuje odpoveď na otázku, či je vo vedení podniku žena.
- `country` – v ktorej krajine má podnikateľský subjekt sídlo.
- `industry` – zaradenie podniku do odvetvia podľa Svetovej banky.
- `sales` – celkové tržby za posledný fiškálny rok vyjadrené v domácej mene.
- `employees` – počet zamestnancov (na plný úväzok) ku koncu posledného fiškálneho roka.
- `labor_costs` – kompletné mzdové náklady za posledný fiškálny rok.
- `sales_3y` – celkové tržby dosiahnuté za fiškálny rok tri roky dozadu.

V tomto príklade overte, či existuje závislosť medzi pohlavím vlastníkov podniku a pohlavím vedenia podniku. Konkrétnejšie, či existuje vzťah medzi (ne)prítomnosťou žien vo vlastníctve a vedení podnikov. Túto hypotézu overte za celú vzorku a následne aj za subjekty zo Slovenska.

Príklad 7.40

Pokračujme v predchádzajúcom príklade a údajoch, avšak už len s odpoveďami od respondentov zo Slovenska. Overovať budeme takú istú hypotézu o nezávislosti medzi pohlavím vedenia podniku a vlastníkov. Tentoraz však odpovede „neviem“ z databázy odstráňte a vypočítajte *Phi* koeficient a Cramerov-*V* koeficient (tieto dva koeficienty asociácie by sa mali v tomto príklade rovnať, keďže pracovať budeme s kontingenčnou tabuľkou o rozmeroch 2 x 2).

Príklad 7.41

V databáze `survey` (z knižnice `MASS`) máme k dispozícii odpovede 237 študentov z Univerzity v Adelaide na rôzne otázky. Zistite, či existuje závislosť medzi pohlavím

študentov (premenné `Sex`) a tým, či píšu ľavou alebo pravou rukou (premenné `W.Hnd`). Kontingenčnú tabuľku aj vizualizujte.

Príklad 7.42

Máme k dispozícii rôzne ukazovatele za európske podniky obchodované na akciových trhoch (údaje sú prevzaté zo stránky pages.stern.nyu.edu/~adamodar, dostupné online [22.06.2013]). Databáza je oproti pôvodnému zdroju očistená o chýbajúce dáta, pričom obsahuje nasledujúce premenné:

- `Ticker` – ide identifikačnú skratku podniku, pod ktorou sa obchoduje na burze.
- `Sub.Group` – táto premenná rozdeľuje obsiahnuté podniky na tri úrovne: UK, EU a Eastern Europe & Russia.
- `Beta` – je štandardný ukazovateľ z modelu CAPM (z angl. *Capital Asset Pricing Model*), ktorý charakterizuje volatilitu akcií daného podniku. V tomto príklade ho budeme využívať ako aproximáciu rizikovosti akcií.
- `PB` – je pomerový ukazovateľ trhovej hodnoty podniku. Dáva do pomeru trhovú cenu akcie a účtovnú hodnotu akcie.
- `Payout` – je výplatný pomer, ktorý sa počíta ako pomer dividendy na akciu a zisku pripadajúceho na jednu akciu. Hovorí teda o tom, aký pomer sa vypláca akcionárom v podobe dividend z vytvoreného zisku.
- `Growth` – je historická miera rastu tržieb za posledné tri roky.
- `ROE` – je rentabilita vlastného kapitálu.

Vypočítajte korelácie medzi týmito fundamentálnymi ukazovateľmi podniku v rámci skupín EU a UK (UK je skratka pre Anglicko, ktoré samozrejme patrí do Európskej únie, avšak v tejto databáze sa údaje oddelili). Korelačné koeficienty (použite Pearsonov koeficient), ktoré budú významné na hladine 5 % v oboch skupinách, otestujte na ich rovnosť.

Príklad 7.43

Uskutočnili sme jednoduchý experiment. Na hodine štatistiky sme každému študentovi odmerali jeho výšku a dĺžku ruky (v cm). Zároveň sme si zaznačili pohlavie študenta. Údaje je možné nájsť pod zadaním tohto príkladu. Chceme zistiť, či je rozumné predpokladať, že medzi výškou ľudí a dĺžkou ruky existuje štatistický vzťah.

a) Vytvorte vhodný obrázok na vizualizáciu vzťahu medzi výškou a dĺžkou ruky.

b) Skúste do obrázku zakresliť lineárnu regresnú priamku (použite funkciu `abline(lm(y ~ x))`.

c) Skúste na základe obrázku odhadnúť, či je vzťah vhodné opísať lineárnou alebo prípadne inou (rýdzo) monotónnou funkciou.

d) Vypočítajte silu vzťahu medzi výškou študenta a dĺžkou ruky. Použite ľubovoľný koeficient na meranie závislosti. Otestujte jeho významnosť (predpokladajme, že ide o realizácie z náhodného výberu).

e) Pomocou vhodného obrázku zobrazte silu vzťahu zvlášť pre mužov a zvlášť pre ženy. Do obrázkov naneste aj regresné priamky.

f) Vypočítajte silu vzťahu pomocou Pearsonovho korelačného koeficientu. V ktorej skupine je vzťah medzi výškou študenta a dĺžkou ruky väčší? Otestujte významnosť koeficientov.

g) Overte, či je vo všeobecnosti možné silu vzťahu v týchto dvoch skupinách považovať za štatisticky rovnakú. Pre tento účel vytvorte funkciu.

h) Pokúste sa úlohu g) vyriešiť aj pomocou štandardného bootstrapingu.

```
vyska <- c(170, 168, 177, 183, 191, 171, 163, 160, 172, 165,
  168, 187, 170, 190, 192, 171, 178, 198, 178, 179, 166, 174,
  181, 194, 184, 175)
dlzka <- c(74, 72, 79, 75, 83, 73, 70, 66, 70, 74, 73, 80, 73,
  81, 84, 72, 73, 87, 72, 78, 70, 75, 79, 83, 75, 74)
pohlavie <- c("z", "z", "m", "m", "m", "z", "z", "z", "m", "z",
  "z", "m", "z", "m", "m", "z", "m", "m", "z", "m", "z", "z",
  "m", "m", "m", "m")
```

Príklad 7.44

Ekonóm v podniku dostal nariadenie predať firemné autá. Ekonóm na základe skúseností vie, že medzi cenou daného modelu auta a počtom najazdených kilometrov existuje nepriama úmera. Z databázy ojazdených automobilov si ekonóm náhodne vybral 30 áut rovnakej značky, modelu a motorizácie. Zaznamenal si cenu, za ktorú chcú predávajúci auto predať (samotná predajná cena je spravidla nižšia), počet najazdených kilometrov a rok výroby auta.

a) Pomocou vizualizácie overte tvrdenie, že medzi počtom najazdených kilometrov a ponúkanou cenou vozidla existuje nejaká forma vzťahu, ktorú by bolo možné opísať pomocou (rýdzo) monotónnej funkcie.

b) Vypočítajte silu vzťahu medzi ponúkanou cenou a počtom najazdených kilometrov. Overte štatistickú významnosť tohto vzťahu.

c) Analogicky ako v a) pomocou vizualizácie overte tvrdenie, že medzi rokom výroby a počtom najazdených kilometrov existuje nejaká forma (rýdzo) monotónneho vzťahu.

d) Vypočítajte silu vzťahu medzi ponúkanou cenou a počtom najazdených kilometrov. Overte štatistickú významnosť tohto vzťahu.

e) Aká by bola korelácia medzi cenou a počtom najazdených kilometrov, ak by bola premenná rok výroby konštantná? Použite pritom parciálny korelačný koeficient, ktorý sa dá vypočítať nasledovne:

$$\hat{\rho}_{AB,C} = \frac{\hat{\rho}_{AB} - \hat{\rho}_{AC}\hat{\rho}_{BC}}{\sqrt{\left(1-\hat{\rho}_{AC}^2\right)\left(1-\hat{\rho}_{BC}^2\right)}}$$

kde $\hat{\rho}$ predstavujú odhady korelačných koeficientov medzi jednotlivými premennými *A*, *B* a *C*.

```
km_v_10tis <- c(8.2, 12.1, 5.9, 2.1, 10.7, 3.6, 11.5, 4.8, 5.2,
  8.4, 8.6, 7.7, 6, 9.5, 10.4, 3.9, 8.4, 5, 8, 11, 6.6, 16.1,
  4.4, 4.2, 7.1, 1.5, 1.9, 1.3, 5.8, 9.9)
rok_vyroby <- c(2004, 2003, 2004, 2005, 2004, 2005, 2003, 2006,
  2005, 2005, 2005, 2004, 2005, 2005, 2004, 2006, 2005, 2006,
  2004, 2003, 2005, 2003, 2006, 2006, 2006, 2006, 2006, 2006,
  2005, 2004)
cena_v_tis_eur <- c(6.81, 5.95, 6.61, 7.61, 5.61, 6.94, 5.61,
  6.94, 6.28, 7.47, 6.48, 6.94, 7.14, 6.48, 5.95, 6.28, 6.28,
  6.61, 6.28, 6.28, 7.27, 7.94, 7.14, 7.61, 5.61, 6.94, 8.14,
  8.27, 7.94, 6.94)
```

Príklad 7.45

Zozbierali sme údaje o výške disponibilného nominálneho ročného dôchodku zamestnanca, ktorý je slobodný a zarába 50 % z priemernej mzdy v Slovenskej republike. Tieto údaje sme doplnili o ročné údaje inflácie merané pomocou harmonizovaného indexu spotrebiteľských cien. Údaje sú dostupné v databáze Eurostat (skratky premenných v databáze sú [earn_nt_net] a [prc_hicp_aind]). V nižšie uvedených troch vektoroch sú zaznamenané tieto premenné, pričom pri disponibilnom dôchodku sme uviedli percentuálny nárast (pokles), teda zmenu oproti predošlému roku.

a) Zaujíma nás, či je možné vychádzať z toho, že ak došlo k nárastu cenovej hladiny, tak zároveň došlo aj k nárastu disponibilného dôchodku. Zobrazte a kvantifikujte tento vzťah.

b) Posúďte aj významnosť tohto vzťahu a použite rôzne koeficienty.

c) Z údajov je vidieť, že v roku 2000 došlo k prudkému nárastu disponibilného dôchodku. Kvantifikujte tento vzťah len pre údaje od roku 2001 až po rok 2012. Výsledky porovnajte.

d) Je možné, že vzťah medzi cenovou hladinou a výškou mzdy existuje, avšak najprv dochádza k nárastu cenovej hladiny a mzdy reagujú s omeškaním 1 alebo až 2 rokov. Overte toto tvrdenie. Najprv tieto vzťahy zobrazte, potom naneste do obrázku lineárnu regresnú priamku (analogicky použite príkaz v tvare `abline(lm(y ~ x))`, kvantifikujte vzťah a overte štatistickú významnosť týchto vzťahov.

```
rok <- c(1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005,
  2006, 2007, 2008, 2009, 2010, 2011, 2012)
disp_dochodok <- c(3.9, -1.17, -10.7, 67.34, 0.05, 6.11, 9.14,
  19, 12.34, 11.21, 18.35, 17.77, 9.97, 2.52, 0.71, 1.81)
hicp <- c(6, 6.7, 10.4, 12.2, 7.2, 3.5, 8.4, 7.5, 2.8, 4.3, 1.9,
  3.9, 0.9, 0.7, 4.1, 3.7)
```

Príklad 7.46

Chystáte sa kúpiť si byt. Uvažujete nad dvojizbovými bytmi na vybranom sídlisku v meste Košice. Z internetovej stránky ste získali údaje o ponúkanej cene (podobne ako pri ojazdených autách, cena pri ktorej dôjde k realizácii obchodu je spravidla nižšia) a obývateľnej ploche bytov.

a) Je podľa Vás pravda, že čím je väčšia plocha bytu, o to je menšia cena za jeden meter štvorcový? Váš názor podoprite dôkazmi vo forme vhodných výpočtov a vhodnej vizualizácie.

b) Zrejme by malo platiť, že ak ide o byty v podobnej lokalite a s rovnakým počtom izieb, potom vo všeobecnosti by byt s väčšou obývateľnou plochou mal byť aj celkovo drahší. Podobne ako v a) toto tvrdenie overte.

```
plocha <- c(48, 50, 60, 52, 55, 50, 58, 49, 52, 50, 72, 55, 50,
  56, 56, 55, 50, 56, 50, 90, 54, 54, 54)
cena <- c(37842, 41493, 54770, 41493, 42157, 41493, 41161,
  44812, 48132, 40829, 81658, 44812, 41161, 39501, 46140, 42821,
  39335, 49625, 42821, 96263, 46472, 38174, 48132)
```

Príklad 7.47

V rámci rýchleho prieskumu vzdelávací inštitút kontaktoval 43 náhodne vybraných obchodných zástupcov v Košickom kraji, ktorí majú licenciu poskytovať finančné

poradenstvo. Jedným zo sledovaných údajov bol počet sprostredkovaných povinných zmluvných poistení (PZP) za jeden kvartál.

a) Zaujíma nás, či je možné očakávať, že stredná hodnota počtu uzavretých PZP je v danom kraji za jeden kvartál 50.
b) Zostrojte histogram z počtu uzavretých PZP a do histogramu naneste hustotu rozdelenia strednej hodnoty za predpokladu platnosti nulovej hypotézy (pri odhade variability strednej hodnoty použite výberový rozptyl), ako aj hustotu rozdelenia strednej hodnoty za predpokladu, že stredná hodnota je rovná výberovému aritmetickému priemeru.

```
PZP <- c(51, 52, 53, 53, 54, 52, 51, 48, 49, 48, 48, 48, 50, 50,
  51, 52, 53, 53, 51, 48, 48, 48, 48, 48, 50, 50, 47, 48, 46,
  49, 50, 51, 52, 53, 54, 49, 48, 47, 47, 48, 48, 48, 49)
```

Príklad 7.48

Marketingová spoločnosť uskutočnila v roku 2011 podľa požiadavky zadávateľa prieskum trhu spokojnosti zákazníkov. Medzi inými mali zákazníci ohodnotiť, nakoľko sú spokojní s po predajnou podporou. Svoju spokojnosť hodnotili na Likertovej škále od 1 do 7. Kde 1 znamená: „*bol som veľmi spokojný*“ a 7 znamená: „*bol som veľmi nespokojný*“. Spomedzi všetkých 2500 zákazníkov spoločnosti, agentúra náhodne oslovila 50-tich.

a) Vytvorte (vertikálny) box – plot a do obrázku vložte čiary, ktoré na osi *y*-ovej budú zodpovedať hodnotám 5.2 a 4.5.
b) V roku 2009 bola odhadovaná stredná hodnota spokojnosti zákazníkov s po predajnými službami 4.5 a v roku 2010 5.2. Vašou úlohou je overiť, či je rozumné predpokladať, že v roku 2011 došlo k poklesu strednej hodnoty spokojnosti s po predajnými službami všetkých zákazníkov, a to tak oproti roku 2010, ako aj oproti roku 2009.
c) Zostrojte obojstranné 95 % konfidenčné intervaly pre strednú hodnotu a porovnajte ich s bootstrapingovanými konfidenčnými intervalmi (štandardný jednoduchý bootstrap).
d) Zostrojte obojstranné 95 % konfidenčné intervaly pre rozptyl základného súboru (predpokladáme, že spokojnosť je realizáciou z normálneho rozdelenia pravdepodobnosti) a analogicky ako v predošlom prípade ich porovnajte s bootstrapingovanými konfidenčnými intervalmi.

e) Je podľa Vás rozumné očakávať, že viac ako polovica všetkých zákazníkov by hodnotila spokojnosť vyššie ako 4?

f) Zostrojte histogram spokojnosti a naneste do neho vertikálne čiary označujúce hranice konfidenčného intervalu pre strednú hodnotu. Tento postup vykonajte pre rôzne miery konfidencie, povedzme: 80 %, 90 %, 95 %, 99 % a 99.9 %. Čo môžeme pozorovať?

```
spokojnost <- c(5, 4, 6, 7, 7, 5, 4, 2, 3, 1, 5, 3, 4, 2, 2, 3,
  3, 5, 2, 4, 2, 2, 3, 1, 2, 2, 4, 2, 2, 3, 1, 2, 3, 3, 5, 2, 4,
  2, 2, 3, 1, 2, 3, 3, 7, 7, 5, 4, 3, 1)
```

Príklad 7.49

K dispozícii máme údaje o tržbách obchodného reťazca za vybraný mesiac v 23 približne rovnako veľkých predajniach, v štyroch krajoch na Slovensku.

a) V ktorom regióne sme namerali najväčšiu variabilitu tržieb? Diskutujte o tom, čo môže väčšia variabilita tržieb v danom kraji napovedať.

b) Vypočítajte priemer a medián tržieb. Výsledky porovnajte a zaznačte ich do histogramu tržieb. Je podľa Vás možné rozdelenie početnosti tržieb charakterizovať ako pravostranné?

c) Predpokladajme, že výška tržieb pochádza z normálneho rozdelenia pravdepodobnosti. Ak by sme náhodne vybrali predajňu, aká je pravdepodobnosť, že výška tržieb bude menšia ako 3.4mil. EUR? Aká je pravdepodobnosť, že výška tržieb bude v intervale od 2.0mil. EUR do 4.0mil. EUR? Záleží na tom, či ide o otvorené alebo uzavreté intervaly? Vašu odpoveď zdôvodnite.

d) Ak by sme považovali namerané tržby za náhodné realizácie určitého dáta generujúceho procesu, akú najmenšiu strednú hodnotu tržieb na jednu pobočku by sme mohli očakávať?

```
trzby <- c(1.35, 3.74, 4.26, 1.2, 1.22, 6.57, 1.56, 6.85, 1.03,
  6.76, 2.6, 3.42, 4.59, 3.29, 6.25, 3.96, 6.7, 2.95, 3.71,
  3.22, 1.28, 2.56, 5.64)
kraj <- c("KOŠICE", "BANSKÁ BYSTRICA", "PREŠOV", "ŽILINA",
  "KOŠICE", "BANSKÁ BYSTRICA", "PREŠOV", "ŽILINA", "KOŠICE",
  "BANSKÁ BYSTRICA", "PREŠOV", "ŽILINA", "KOŠICE", "BANSKÁ
  BYSTRICA", "PREŠOV", "ŽILINA", "KOŠICE", "BANSKÁ BYSTRICA",
  "PREŠOV", "ŽILINA", "KOŠICE", "BANSKÁ BYSTRICA", "PREŠOV")
```

Príklad 7.50

Reklamná agentúra na podnet zadávateľa zisťovala, ako vníma cieľová skupina práve prebiehajúce reklamné aktivity zadávateľa. Prieskum sa uskutočnil na náhodných vzorkách cieľovej skupiny v dvoch lokalitách: Bratislava (BA) a mimo Bratislavy (MBA). Mimo iných boli v osobných rozhovoroch zisťované odpovede na nasledujúce otázky (uvádzame aj možné odpovede, z ktorých si mohli respondenti/obyvateľstvo vyberať).

1) Aký je Váš vek? (v rokoch)
2) Aké je Vaše pohlavie? (Muž/Žena)
3) Aký je Váš názor na logo spoločnosti? (Páči sa mi/Nepáči sa mi)
4) Použili ste niekedy produkt spoločnosti? (Áno/Nie)
5) Poskytuje pre Vás prebiehajúca reklamná kampaň dostatočné množstvo relevantných informácií? (číslo na škále od 1 do 10, kde väčšie číslo znamená viac relevantných informácií a menšie číslo menej relevantných informácií).

Úlohy:

a) Môže byť podľa Vás pravda, že v Bratislave sa menšiemu podielu obyvateľstva páči logo, ako v lokalite mimo Bratislavy?
b) Predpokladajme, že vek žien sa riadi normálnym rozdelením pravdepodobnosti. Sú podľa Vás vo vzorke žien extrémne hodnoty veku?
c) Manažér zo spoločnosti, ktorá zadávala reklamu tvrdí, že si myslí, že existuje silná závislosť medzi pohlavím a názorom na logo. Skúste overiť jeho tvrdenie.
d) Môže byť podľa Vás pravda, že stredná hodnota veku obyvateľov, ktorým sa logo nepáči je väčšia, ako stredná hodnota veku obyvateľov, ktorým sa logo páči?
e) Manažér zo spoločnosti, ktorá zadávala reklamu tvrdí, že si myslí, že existuje silná závislosť medzi vekom respondentov a ich názorom na množstvo relevantných informácií reklamnej kampane. Skúste overiť jeho tvrdenie a popíšte danú závislosť.
f) Zobrazte vzťah medzi vekom a názorom na množstvo relevantných informácií v reklame v závislosti od toho, či sa cieľovej skupine logo spoločnosti páči, alebo nie.
g) V akom intervale môžeme očakávať, že sa nachádza stredná hodnota veku cieľovej skupiny zákazníkov?
h) Predpokladajme, že názor zákazníkov na množstvo relevantných informácií sa riadi normálnym rozdelením pravdepodobnosti, a to tak v Bratislave, ako aj mimo nej. Je podľa Vás rozumné predpokladať, že variabilita názorov na množstvo relevantných informácií z reklamnej kampane je v oboch oblastiach rovnaká? Je podľa Vás rozumné predpokladať, že stredná hodnota názoru na množstvo relevantných

informácií je v cieľovej skupine mimo Bratislavy odlišná od názoru cieľovej skupiny v Bratislave?

i) Zadávateľa reklamy zaujíma, či môže predpokladať, že podiel ľudí vo vzorke z Bratislavy, ktorým sa logo páči, je väčší ako 0.6.

```
region <- c("BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA",
  "BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA",
  "BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA", "BA",
  "BA", "BA", "BA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA",
  "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA",
  "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA",
  "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA", "MBA",
  "MBA")
vek <- c(19, 23, 24, 22, 26, 30, 74, 65, 49, 22, 30, 53, 54, 23,
  40, 44, 50, 69, 43, 52, 19, 46, 40, 33, 32, 35, 46, 45, 44,
  58, 27, 43, 23, 30, 53, 54, 58, 43, 26, 24, 22, 26, 54, 58,
  74, 65, 64, 50, 34, 23, 52, 19, 24, 22, 43, 52, 19, 92, 74,
  65, 43, 21, 25, 19, 19)
pohlavie <- c("M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
  "M", "M", "M", "M", "M", "M", "Z", "Z", "Z", "Z", "Z", "Z",
  "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "M", "M", "M",
  "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
  "M", "M", "M", "M", "Z", "Z", "Z", "Z", "Z", "Z", "Z", "Z",
  "Z", "Z", "Z", "Z", "Z", "Z", "Z")
logo <- c("paci sa mi", "nepaci sa mi", "paci sa mi", "paci sa
  mi", "paci sa mi", "paci sa mi", "nepaci sa mi", "nepaci sa
  mi", "nepaci sa mi", "nepaci sa mi", "nepaci sa mi", "nepaci
  sa mi", "nepaci sa mi", "nepaci sa mi", "nepaci sa mi",
  "nepaci sa mi", "paci sa mi", "paci sa mi", "paci sa mi",
  "paci sa mi", "paci sa mi", "paci sa mi", "paci sa mi",
  "nepaci sa mi", "nepaci sa mi", "paci sa mi", "paci sa mi",
  "paci sa mi", "paci sa mi", "nepaci sa mi", "nepaci sa mi",
  "paci sa mi", "nepaci sa mi", "paci sa mi", "paci sa mi",
  "nepaci sa mi", "nepaci sa mi", "nepaci sa mi", "paci sa mi",
  "nepaci sa mi", "nepaci sa mi", "nepaci sa mi", "nepaci sa
  mi", "paci sa mi", "nepaci sa mi", "nepaci sa mi", "nepaci sa
  mi", "paci sa mi", "nepaci sa mi", "paci sa mi", "paci sa mi",
  "nepaci sa mi", "paci sa mi", "paci sa mi", "paci sa mi",
  "paci sa mi", "paci sa mi", "paci sa mi", "paci sa mi", "paci
  sa mi", "nepaci sa mi", "nepaci sa mi", "paci sa mi", "paci sa
  mi", "paci sa mi")
pouz_prod <- c("áno", "nie", "áno", "nie", "nie", "nie", "nie",
  "nie", "áno", "nie", "áno", "nie", "nie", "nie", "nie", "nie",
  "nie", "nie", "áno", "áno", "áno", "áno", "nie", "áno", "áno",
  "nie", "nie", "áno", "áno", "nie", "áno", "áno", "nie", "nie",
  "áno", "nie", "nie", "áno", "áno", "nie", "nie", "nie", "nie",
  "áno", "nie", "nie", "áno", "áno", "nie", "áno", "nie", "nie",
  "áno", "áno", "áno", "áno", "áno", "áno", "áno", "áno", "áno",
  "nie", "áno", "áno", "áno")
informacna_hodnota <- c(9, 8, 8, 8, 7, 6, 1, 3, 5, 7, 7, 3, 4,
  8, 4, 6, 4, 6, 4, 3, 8, 6, 4, 3, 10, 6, 5, 6, 6, 5, 8, 5, 9,
  8, 5, 5, 6, 4, 8, 8, 7, 6, 4, 5, 3, 4, 5, 5, 7, 8, 5, 7, 10,
  9, 5, 6, 8, 9, 2, 3, 4, 10, 9, 8, 6)
```

Príklad 7.51

Podnikateľ sa rozhodol uskutočniť malý prieskum u konkurenčných predajní. Počas jedného týždňa náhodne vybral vzorku 24 zákazníkov. U zákazníkov si zaznačil veľkosť ich nákupu, teda koľko v obchode minuli peňazí (nákup v EUR), ich priemerný mesačný zárobok (príjem), ich pohlavie (1 – muž, 0 – žena), a či nakupujú cez víkendy (1 – áno, 0 – nie).

a) Zošikmenie veľkosti nákupov nám môže napovedať, či sa extrémne veľké alebo extrémne malé nákupy vyskytujú častejšie alebo menej často vzhľadom k výskytu ostatných hodnôt v empirickom súbore. Pomocou mediánu a aritmetického priemeru rozhodnite o zošikmení hodnôt veľkosti nákupov.

b) V akom intervale môžeme očakávať, že sa vo všeobecnosti bude nachádzať stredná hodnota objemu nákupu všetkých zákazníkov? Predpokladáme, že hodnoty pochádzajú z normálneho rozdelenia pravdepodobnosti.

c) Akú najväčšiu strednú hodnotu objemu nákupov všetkých zákazníkov obchodu môžeme očakávať? Predpokladáme, že hodnoty pochádzajú z normálneho rozdelenia pravdepodobnosti.

d) Je podľa Vás rozumné vychádzať z toho, že vo všeobecnosti je stredná hodnota objemu nákupov u zákazníkov 50,- EUR? Predpokladáme, že hodnoty pochádzajú z normálneho rozdelenia pravdepodobnosti.

e) Skúste pomocou jednočíselnej charakteristiky overiť nasledujúce tvrdenie: zákazníci s vyšším mesačným príjmom majú tendenciu realizovať väčšie nákupy.

f) Skúste pomocou jednočíselnej charakteristiky overiť nasledujúce tvrdenie: muži a ženy nenakupujú v rovnaké dni.

g) Je podľa Vás na základe vzorky možné tvrdiť, že vo všeobecnosti (v populácii) muži v priemere uskutočňujú nákupy rovnako veľké (v EUR) ako ženy? Predpokladáme, že hodnota nákupov (v EUR) u mužov aj žien sa riadi normálnym rozdelením pravdepodobnosti.

h) Je podľa Vás na základe vzorky možné tvrdiť, že vo všeobecnosti (v populácii) je priemerný mesačný príjem zákazníkov na úrovni 600,- EUR? Predpokladáme, že výška mesačného príjmu sa riadi normálnym rozdelením pravdepodobnosti.

i) Overte, či je možné mesačný príjem mužov považovať za realizácie z rovnakého rozdelenia pravdepodobnosti ako mesačný príjem žien.

```
nakup <- c(50, 10, 70, 60, 50, 45, 75, 35, 42, 28, 16, 18, 8,
  60, 50, 40, 35, 75, 50, 30, 37, 23, 25, 25)
```

```
prijem <- c(730, 600, 670, 700, 625, 700, 640, 320, 420, 310,
  420, 410, 300, 370, 500, 480, 400, 700, 625, 350, 540, 270,
  370, 480)
pohlavie <- c(1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1,
  0, 1, 0, 1, 0, 0, 0)
den <- c(0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0,
  0, 0, 1, 0, 0, 1)
```

Príklad 7.52

Podnikateľ vlastní menšie potraviny s dvoma pokladňami. Vo vybrané dni počas dvoch mesiacov (v náhodné hodiny) náhodne vybral 135 zákazníkov a odmeral čas zdržania sa zákazníkov pri pokladni (v sekundách).

a) Overte, či čas zdržania sa pri pokladni je možné modelovať pomocou normálneho rozdelenia pravdepodobnosti.
b) Za predpokladu, že sa tieto hodnoty riadia normálnym rozdelením pravdepodobnosti, aká je pravdepodobnosť, že náhodne vybraný zákazník by sa pri pokladni zdržal menej ako 30 sekúnd?
c) Pomocou mediánu a aritmetického priemeru rozhodnite o zošikmení hodnôt. Zostrojte histogram, hodnoty naneste do grafu. Porovnajte Váš záver s koeficientom šikmosti.
d) V akom intervale (konfidencia 0.95) sa podľa Vás bude nachádzať stredná hodnota času zdržania sa zákazníkov pri pokladni?

```
cas <- c(5, 15, 25, 35, 45, 55, 65)
pocetnost <- c(10, 15, 30, 35, 25, 15, 5)
```

Príklad 7.53

Z databázy zákazníkov si importér nechal zobraziť údaje o veku zákazníkov a druhu kúpeného auta. Údaje sú v nasledujúcej kontingenčnej tabuľke. Vašou úlohou je zistiť, aký silný je vzťah medzi vekom zákazníka a druhom auta.

```
druh <- c("sportove","rodinne","terenne")
nizky_vek <- c(20, 30, 50)
stredny_vek <- c(60, 260, 45)
vysoky_vek <- c(85, 220, 110)
```

Príklad 7.54

Máme k dispozícii údaje vo forme korelačnej tabuľky. Ide o údaje z prieskumu medzi náhodne vybranými zákazníkmi stravovacieho zariadenia. Respondenti odpovedali na otázky

týkajúce sa spokojnosti so stravou a obsluhou. Svoju odpoveď zaškrtli na škále od 1 po 7. V stĺpcoch sú odpovede respondentov ohľadom ich spokojnosti s obsluhou (1 – veľmi spokojný, 7 – veľmi nespokojný). V riadkoch sú odpovede respondentov ohľadom ich spokojnosti so stravou (1 – veľmi spokojný, 7 – veľmi nespokojný).

a) Zobrazte možnú závislosť medzi spokojnosťou so stravou a obsluhou.
b) Kvantifikujte závislosť medzi premennými.
c) Zistite, či je možné nameraný vzťah považovať za štatisticky významný.

```
prieskum <- matrix(c(3, 2, 0, 0, 0, 1, 0, 0, 2, 1, 2, 0, 1, 0,
  1, 2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 2, 1, 1, 0, 0, 0, 1, 2, 2, 2,
  0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 2, 0, 4), ncol = 7)
colnames(prieskum) <- c("obsluha - 1", "obsluha - 2", "obsluha -
  3", "obsluha - 4", "obsluha - 5", "obsluha - 6", "obsluha -
  7")
rownames(prieskum) <- c("strava - 1", "strava - 2", "strava -
  3", "strava - 4", "strava - 5", "strava - 6", "strava - 7")
```

Príklad 7.55

Máme dva súbory údajov. Veľkosť nákupov v dvoch obchodoch (*A*, *B*). Vytvorte jednoduchú funkciu, ktorej vstupom bude vektor údajov a výstupom bude tabuľka (napr. vo forme matice) so základnými opisnými charakteristikami: priemer, medián, kvartily, minimum, maximum, výberová smerodajná odchýlka, výberový rozptyl, medzi-kvartilové rozpätie, variačný koeficient, šikmosť, špicatosť a vybraný test normality.

```
A <- c(15, 18, 16, 14, 11, 8, 9, 19, 20, 18, 16, 24, 17, 16, 16,
  4, 5, 21, 24, 10, 8, 9, 24, 7, 16, 20, 12, 16, 15, 14, 11, 19,
  7, 22, 8, 9, 18, 15, 16, 18, 17, 13, 15, 12, 14)
B <- c(11, 7, 9, 10, 11, 8, 12, 14, 16, 18, 19, 15, 14, 17, 17,
  12, 13, 14, 16, 15, 14, 17, 16, 15, 13, 14, 15, 14, 19, 16, 8,
  10, 12, 14, 16, 18, 16, 17, 15, 16, 15, 14, 10, 19, 16)
```

Príklad 7.56

Výrobca mlieka sa rozhoduje urobiť novu reklamnú kampaň. V rámci nej by sa chcel zamerať na tých zákazníkov, ktorí si jeho výrobky doteraz veľmi nekupovali. Výrobca sa preto rozhodol vo vlastnej réžií uskutočniť prieskum zákazníkov. Ten sa uskutočnil na vzorke 24 respondentov, ktorí nakúpili ich tovar vo vybraných obchodných domoch. Spomedzi rôznych údajov o respondentoch sa zisťoval aj ich vek.

a) Vypočítajte základne charakteristiky polohy premennej vek: priemer a medián.

b) Vypočítajte kvartily a spolu s charakteristikami polohy skúste z hľadiska veku opísať zákazníkov výrobcu mlieka.

c) Uvažujme o tom, že vzorka 24 respondentov bola získaná na základe náhodného výberu. Je podľa vás rozumné tvrdiť, že vo všeobecnosti sú viac ako tri štvrtiny (všetkých) zákazníkov mladších ako 50 rokov?

```
vek <- c(45, 49, 21, 38, 32, 62, 64, 58, 50, 37, 33, 27, 24, 25,
  37, 42, 49, 52, 73, 82, 37, 65, 28, 50)
```

Príklad 7.57

Zákazníci potravín sa sťažovali na správanie sa predajného personálu. Majiteľ sa rozhodol, že vedúci prevádzky absolvuje manažérske školenie a predajný personál osobitné školenie o predaji a komunikácii. Aby majiteľ zistil, či došlo k požadovanej zmene, uskutočnil dva prieskumy. Prvý prieskum sa uskutočnil pred školením. Náhodne vybraní zákazníci vyjadrovali svoju spokojnosť s personálom na škále od 1 (nespokojný) po 10 (veľmi spokojný). Tieto výsledky porovnal s prieskumom, ktorý majiteľ uskutočnil dva mesiace po školení.

a) Sú vo všeobecnosti zákazníci viac jednoznační vo svojich odpovediach po školení, ako boli pred školením?

b) Overte, či môžeme tvrdiť, že vo všeobecnosti došlo k zlepšeniu hodnotenia správania sa personálu.

c) Porovnajte dolné kvartily spokojnosti a skúste výsledok interpretovať.

d) Pochádzajú hodnoty z rovnakého rozdelenia pravdepodobnosti? Zostrojte najprv dva prekrývajúce sa histogramy (porozmýšľajte, či je možné pomocou riešenia tejto úlohy riešiť aj úlohu a)

```
pred <- c(3, 7, 8, 4, 6, 2, 9, 7, 8, 9, 6, 5, 2, 4, 9, 8, 5, 7,
  8, 9, 6)
potom <- c(5, 7, 8, 8, 4, 6, 5, 7, 7, 9, 8, 5, 6, 6, 8, 8, 9, 8,
  7, 6, 8)
```

Príklad 7.58

Výrobný závod vedie evidenciu počtu nesprávne vyrobených výrobkov. Vo vektore `chyby` sú uvedené počty chybných výrobkov, ktoré zodpovedajú jednotlivým pracovným dňom.

a) Je podľa vás väčšina hodnôt menších ako aritmetický priemer? Čo to hovorí o šikmosti súboru? Ktorá hodnota je najpočetnejšia?
b) Zostrojte box – plot a naneste do neho čiaru označujúcu hodnotu 10. Pred zlepšovaním výrobného procesu bol priemerný počet chýb na úrovni 10. Predpokladajme, že počty chybných výrobkov sú realizácie náhodného výberu. Je podľa Vás možné tvrdiť, že došlo k zníženiu chybovosti?

```
chyby <- c(3, 4, 1, 2, 3, 4, 5, 2, 6, 8, 5, 3, 4, 5, 2, 7, 4, 8,
  5, 6, 5, 4, 3, 5, 4, 7, 5, 10, 13, 12, 10, 8, 7, 6, 8, 7, 10,
  12, 13, 8, 6, 5, 7, 8, 9, 5, 8, 9, 8, 7, 9, 12, 14, 5, 6, 18,
  17, 16, 15, 14, 13, 10, 8, 9, 7, 5, 6, 12, 18, 13, 15, 1, 14,
  15, 12, 8, 9, 12, 7, 6, 5, 12, 18, 2, 13, 14, 15, 10, 8, 6, 8,
  7, 5)
```

Príklad 7.59

Prepravca vlastní dva druhy nákladných vozidiel. Máme namerané priemerné spotreby pre obe nákladné vozidlá, pričom ide o hodnoty pre rôzne trasy.

a) Zostrojte prekrývajúce sa histogramy pre obe spotreby.
b) Je podľa Vás možné tvrdiť, že spotreby sú podobné?
c) Zostrojte box – ploty pre obe spotreby.
d) Je pravda, že vozidlo 2 dosahuje v priemere menšiu spotrebu?
e) Vozidlo 1 a vozidlo 2 má vlastného vodiča. Jednotlivé hodnoty však predstavujú rôzne trasy. Skúste overiť, či je možné tvrdiť, že existuje závislosť medzi spotrebou a trasou vozidla.

```
vozidlo_1 <- c(15, 14, 13, 16, 15, 17, 16, 18, 17, 14, 13, 15,
  14, 12, 16, 17, 14, 13, 19)
vozidlo_2 <- c(16, 14, 14, 15, 14, 17, 17, 16, 15, 13, 13, 14,
  14, 12, 14, 15, 14, 13, 16)
```

Príklad 7.60

Menší strojárenský podnik uskutočnil každoročný prieskum spokojnosti zamestnancov. Spomedzi všetkých 420 zamestnancov, bolo náhodne vybraných 43. Jednou z otázok bolo zistiť mieru spokojnosti zamestnancov so stravovaním sa (čas na jedlo, vzdialenosť stravovacieho zariadenia od pracoviska, kvalita jedla, spokojnosť s obsluhou, spokojnosť s prostredím,....). Zamestnanci hodnotili svoju celkovú spokojnosť so stravovaním na škále od 1 (veľmi nespokojný) po 10 (veľmi spokojný).

a) Vypočítajte priemernú spokojnosť zamestnancov vo vzorke a variabilitu tejto spokojnosti (rozptylom).
b) Odhadnite strednú hodnotu spokojnosti zamestnancov v podniku a variabilitu tejto spokojnosti?
c) Vypočítajte 95 % konfidenčný interval strednej hodnoty spokojnosti so stravovaním sa všetkých zamestnancov. Vypočítajte 95 % konfidenčný interval rozptylu spokojnosti so stravovaním sa všetkých zamestnancov podniku. (predpokladáme normálne rozdelenie pravdepodobnosti nameraných hodnôt).
d) Vypočítajte 99 % konfidenčný interval podielu zamestnancov v podniku, ktorých spokojnosť bola nižšia ako 6.
e) Je podľa Vás reálne očakávať, že priemerná spokojnosť so stravovaním sa všetkých zamestnancov podniku je nie menej ako 6.0? Počítajte pre $\alpha = 0.05$.
f) Akú najvyššiu strednú hodnotu spokojnosti je podľa vás reálne očakávať? Počítajte pri $\alpha = 0.05$.
g) Aký najmenší rozptyl spokojnosti so stravovaním sa všetkých zamestnancov podniku je možné očakávať? (predpokladáme normálne rozdelenie pravdepodobnosti nameraných hodnôt) - počítajte pri 95 % konfidencii.

```
spokojnost <- c(7, 6, 8, 5, 5, 8, 5, 6, 8, 8, 10, 10, 8, 2, 8,
  5, 6, 5, 7, 5, 5, 5, 8, 6, 7, 8, 5, 2, 2, 5, 6, 7, 5, 6, 8, 9,
  10, 8, 7, 8, 8, 8, 8)
```

Príklad 7.61

V závode ľudia pracujú na dve zmeny: rannú a poobedňajšiu. Keďže je meranie nákladné, za posledný pol rok sa náhodne vybralo 18 dní, v rámci ktorých sa meral priemerný čas výroby jedného výrobku (v minútach). Predpokladáme, že namerané hodnoty pochádzajú z normálneho rozdelenia pravdepodobnosti.

a) Odhadnite strednú hodnotu času výroby pre obe zmeny a porovnajte.
b) Pre obe zmeny zostrojte konfidenčné intervaly pre stredné hodnoty ($\alpha = 0.05$). Čo na základe intervalov viete povedať o priemernom čase výroby?
c) Akú najmenšiu strednú hodnotu času výroby v oboch zmenách môžeme očakávať? ($\alpha = 0.01$).
d) Vytvorte box – ploty pre obe zmeny a porovnajte ich.

```
ranna <- c(5, 5, 4, 3, 3, 5, 2, 2, 3, 5, 6, 2, 5, 3, 2, 4, 5, 5)
```

```
poobednajsia <- c(4, 5, 6, 6, 5, 5, 2, 5, 3, 5, 4, 4, 5, 3, 5,
  4, 6, 5)
```

Príklad 7.62

Manažér skladu preberá tovar na sklad. Prepravná spoločnosť mala doniesť 150000 kusov jedného druhu súčiastok. Manažér je zodpovedný za kontrolu, či došiel správny počet súčiastok v požadovanej kvalite. Súčiastky sú balené po 1000 kusov v 150 krabiciach. Kritickým parametrom súčiastky je priemer hriadeľa, ktorý má mať priemer 3 cm. Kontrola každej krabice a každej súčiastky je časovo veľmi náročná. Manažér preto náhodne vyberie 30 krabíc, v ktorých spočíta počet kusov výrobkov. Zároveň náhodne vyberie 26 súčiastok, pri ktorých odmeria priemer hriadeľa. Výsledky sú uvedené v nasledujúcich vektoroch.

a) Zistite, či manažér môže predpokladať, že v celej dodávke je stredná hodnota počtu výrobkov v jednej krabici 1000 ks.
b) Zistite, či manažér môže predpokladať, že zo všetkých hriadeľov v dodávke je stredná hodnota hriadeľa 3 cm.
c) Overte podozrenie manažéra, že v celej dodávke je podiel hriadeľov s priemerom väčším ako 3 cm rôzny od 50 %.
d) Je podľa vás pravda, že vo vzorke hriadeľov je priemer rovný 3.01 cm?
e) Priemerný priemer hriadeľa je iba jeden z parametrov, ktorý manažéra zaujíma. Druhý s parametrov je rozptyl priemeru hriadeľa v dodávke. Overte či je rozumné predpokladať, že rozptyl priemerov hriadeľa v dodávke je rovný 0.001.

```
pocet_kusov <- c(998, 1000, 1010, 1006, 1002, 1015, 1008, 1002,
  998, 1001, 995, 1010, 1003, 1000, 1004, 997, 1014, 999, 1015,
  1002, 980, 1008, 1002, 994, 1020, 1010, 1003, 992, 1000, 1002)
priemer_hriadela <- c(2.98, 2.97, 3.02, 3.05, 2.98, 3.02, 3.01,
  3.04, 3.02, 2.99, 3.14, 3.03, 3.01, 3.00, 3.00, 3.04, 3.05,
  3.02, 3.01, 3.02, 3.03, 3.02, 3.01, 3.02, 3.00, 3.01)
```

Príklad 7.63

Obchod vykonal prieskum na náhodnej vzorke 25 svojich zákazníkoch. K dispozícii máme nasledujúce premenné: `nakup`: veľkosť nákupu v EUR; `prijem`: výška príjmu zákazníka v EUR; `vek`: vek zákazníka; `pohlavie`: ak ide o muža, premenná nadobúda hodnotu 1 (v opačnom prípade 0); `auto`: ak má auto, premenná nadobúda hodnotu 1 (v opačnom prípade 0); `nehnutelnost`: ak vlastní nehnuteľnosť, premenná nadobúda hodnotu 1 (v opačnom prípade 0); `rodinný stav`: premenná nadobúda hodnotu 1, ak je

zákazník slobodný/á (v opačnom prípade 0). Vedúceho prevádzky zaujímajú niektoré vlastnosti zákazníkov, ktoré formuloval do nasledujúcich úloh:

a) Overte, či všetky veľkosti nákupov je možné považovať za bežné, alebo sú vo vzorke niektoré tzv. extrémne nákupy. Predpokladáme, že veľkosť nákupov pochádza z normálneho rozdelenia pravdepodobnosti.
b) Vypracujte úlohu a) bez predpokladu normality.
c) Aká je sila štatistického vzťahu medzi pohlavím zákazníka a tým, či vlastní auto, resp. medzi tým, či zákazník má auto a tým, či vlastní nehnuteľnosť. Taktiež zistite, či je možné tieto vzťahy považovať za štatisticky významné.
d) Zobrazte vzťah medzi veľkosťou nákupu a veľkosťou príjmu zákazníka. Zistite silu lineárneho vzťahu medzi premennými. Overte významnosť tohto vzťahu. Postupujte rovnako aj v prípade, ak by nás zaujímala sila monotónnej závislosti (inej ako lineárnej) medzi týmito dvoma premennými. Do obrázku naneste lineárnu priamku popisujúcu daný vzťah.
e) Existuje štatisticky významný rozdiel vo veľkosti strednej hodnoty príjmu mužov a žien v populácii všetkých zákazníkov (predpokladáme, že výška príjmu sa riadi normálnym rozdelením pravdepodobnosti)?
f) Je podiel všetkých zákazníkov, ktorí vlastnia auto väčší, ako podiel zákazníkov, ktorí vlastnia nehnuteľnosť?
g) Je variabilita príjmov mužov väčšia, ako variabilita príjmov žien (u všetkých zákazníkov, pričom predpokladáme, že výška príjmu sa riadi normálnym rozdelením pravdepodobnosti)?

```
nakup <- c(15, 14, 13, 12, 13, 15, 28, 14, 10, 8, 9, 1, 11, 13,
  15, 18, 19, 15, 11, 27, 12, 13, 14, 12, 14)
prijem <- c(450, 550, 525, 600, 800, 400, 900, 550, 600, 450,
  400, 450, 500, 550, 600, 700, 750, 650, 550, 1050, 450, 450,
  500, 450, 500)
vek <- c(25, 35, 35, 45, 50, 45, 50, 45, 60, 35, 25, 20, 20, 25,
  30, 35, 40, 35, 30, 55, 40, 35, 40, 40, 40)
pohlavie <- c(1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1,
  1, 0, 1, 0, 0, 1, 0, 1)
auto <- c(0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1,
  1, 1, 1, 1, 0, 0, 1)
nehnutelnost <- c(0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
  1, 1, 0, 0, 1, 1, 0, 0, 1, 1)
rodinny_stav <- c(0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1,
  1, 0, 1, 1, 0, 1, 1, 0, 1, 1)
```

Príklad 7.64

Študenti písali písomku z predmetu Finančné trhy. V dvoch vektoroch máme zaznačené body z dvoch skupín. Skupina C a skupina E. Považujme tieto body za realizácie náhodného výberu.

a) V ktorej skupine (C alebo E) je spodná hranica štvrtiny najúspešnejších študentov vyššia? Na výpočet použite vhodné kvantily.
b) Zostrojte box – plot a skúste pomocou neho rozhodnúť, ktorá zo skupín získala v priemere väčší počet bodov. Váš názor potom otestujte pomocou vhodného neparametrického testu.
c) V ktorej skupine bola väčšia rôznorodosť počtu získaných bodov? Skúste interpretovať, o čom to vypovedá?
d) V ktorej skupine (C alebo E) je spodná hranica polovice najúspešnejších študentov vyššia?

```
body_C <- c(10, 10, 10, 10, 10, 10, 10, 9, 9, 9, 9, 9, 9, 9, 8,
  8, 8, 8, 8, 8, 8, 8, 8, 7, 7, 6, 6, 5, 5, 5, 4 ,4)
body_E <- c(10, 10, 8, 8, 7, 7, 6, 6, 5, 5, 5, 5, 4, 4, 4, 4, 3,
  3, 3)
```

Príklad 7.65

Zákazníci výrobného podniku sa sťažovali, že plechy ktoré spoločnosť dodáva boli ťažšie ako hmotnosť dohodnutá v zmluve, t.j. ťažšie ako 15kg. Vedúci výroby sa preto rozhodol z výroby náhodne vybrať 36 plechov na preskúšanie.

a) Zistite, v akom intervale je možné očakávať, že sa stredná hodnota hmotnosti všetkých vyrobených plechov bude nachádzať ($\alpha = 0.10$).
b) Zistite, v akom intervale je možné očakávať, že sa rozptyl hmotnosti všetkých vyrobených plechov bude nachádzať ($\alpha = 0.01$).
c) Overte pomocou štatistického testu, či je stredná hodnota váhy všetkých plechov rovná 15kg. Zobrazte funkciu hustoty testovacej štatistiky za predpokladu nulovej hypotézy a do obrázku znázornite (vertikálnou čiarou) kritickú hodnotu z *t*-testu.

```
plechy_vaha <- c(15.1, 15.2, 15.2, 15.2, 15.4, 15.2, 15.2, 15.4,
  15.3, 15.3, 15.3, 15.1, 15.0, 15.5, 15.2, 15.2, 15.2, 15.1,
  14.8, 14.8, 15.1, 15.1, 15.1, 15.3, 14.8, 14.9, 15.0, 15.3,
  15.2, 15.2, 15.1, 15.2, 15.2, 15.2, 15, 15.3)
```

Príklad 7.66

Z dodávky uhlia sa náhodne vybralo 12 vzoriek uhlia. Pre každú vzorku sa zaznamenala vlhkosť (meraná v %). Predpokladáme, že namerané hodnoty pochádzajú z normálneho rozdelenia pravdepodobnosti.

a) V akom intervale sa s pravdepodobnosťou 0.99, bude nachádzať stredná hodnota vlhkosti uhlia v dodávke?

b) Zostrojte dva obrázky. Na prvom bude na os *y*-ovej horná a dolná hranica konfidenčného intervalu strednej hodnoty, vypočítaná pomocou Studentovho *t* rozdelenia. Na os *x*-ovej bude konfidenčná pravdepodobnosť a to od 0.50 do 0.99 (po stotinách). Do tohto obrázku taktiež naneste konfidenčné intervaly vypočítané pomocou jednoduchého bootstrappingu.

c) Aká je maximálna hodnota 60 % najmenších vlhkostí. Aká je minimálna hodnota 30 % najväčších vlhkostí. Aká je maximálna hodnota 50 % najväčších vlhkostí.

```
uhlie <- c(25, 27, 32, 43, 29, 30, 27, 26, 25, 55, 29, 32)
```

Príklad 7.67

Z prieskumu sme získali údaje o 29 respondentoch, ktorí predstavujú náhodnú vzorku z celkovej populácie zákazníkov. Medzi iným sme zisťovali pohlavie zákazníkov a to, ktorú farbu výrobku preferujú (bielu, čiernu). Výsledky sú v nasledujúcej kontingenčnej tabuľke zaznamenanej ako matica. Vypočítajte silu vzťahu medzi premennou farba a pohlavie. Skúste túto štatistickú závislosť vizualizovať.

```
contt <- matrix(c(9, 5, 6, 9), ncol = 2)
rownames(contt) <- c("biela", "cierna")
colnames(contt) <- c("muz", "zena")
```

Príklad 7.68

Slovenská banka náhodne vybrala 10 zamestnancov, na ktorých testovala, ako rýchlo spracujú požiadavku zákazníka v novom informačnom systéme. K dispozícii máme namerané dve premenné. Premenná "rýchlosť" zaznamenáva čas [min] spracovania a premenná "hodiny" počet hodín, ktoré zamestnanec s danou aplikáciou strávil. Je podľa vás rozumné tvrdiť (vo všeobecnosti), že zamestnanci, ktorí trávili s danou aplikáciou viac hodín, tak zároveň spracovali požiadavku zákazníka rýchlejšie?

```
rychlost <- c(2, 4, 6, 5, 7, 9, 11, 13, 12, 15)
```

```
hodiny <- c(11, 8, 7, 6, 6, 6, 5, 4, 4, 2)
```

Príklad 7.69

Spoločnosť AXY s.r.o. sa venuje plneniu plastových fliaš. Jej zákazníkov predstavujú výrobcovia nealkoholických nápojov. Jeden zo zákazníkov podal manažérovi sťažnosť, ktorá sa týkala nedostatočného objemu nápoja v plastovej fľaši. Dohodnutý objem predstavuje 0.330 [l.]. Manažér výroby sa rozhodol sťažnosť overiť. Z výroby náhodne vybral 10 plastových fliaš a odmeral objem nápoja. Vašou úlohou je zistiť, či je pravda, že objem nápoja je vo všeobecnosti odlišný od dohodnutého objemu 0.330 [l.]. Predpokladáme, že objem nápoja sa riadi normálnym rozdelením pravdepodobnosti.

```
objem <- c(0.328, 0.329, 0.331, 0.330, 0.325, 0.327, 0.326,
  0.328, 0.326, 0.322)
```

Príklad 7.70

Vo vektore "nákup" predstavuje číslo 1 situáciu, keď zákazník kúpil produkt, 0 ak nekúpil. V stĺpci "pohlavie" predstavuje číslo 1 situáciu, ak zákazník je muž, 0 ak žena. V stĺpci "prvý" predstavuje číslo 1 situáciu, ak ide o prvý nákup zákazníka v predajni, 0 ak opakovaný nákup. Je vzťah medzi premennými "nákup" a "pohlavie" silnejší ako vzťah medzi "nákup" a "prvý"? Odpoveď podložte výsledkami

```
nakup <- c(1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1)
pohlavie <- c(0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1)
prvy <- c(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1)
```

Príklad 7.71

Vo vektore "predtým" je počet kusov vybraného výrobku, ktorý sa predal pred spustením reklamnej kampane. Vo vektore "potom" je počet kusov vybraného výrobku, ktorý sa predal po skončení reklamnej kampane. Považujme údaje za realizácie náhodného výberu a taktiež za realizácie z normálneho rozdelenia pravdepodobnosti. Je podľa vás možné tvrdiť, že vo všeobecnosti došlo k zvýšeniu predaja výrobkov po skončení reklamnej kampane

```
predtym <- c(101, 102, 52, 42, 35, 45, 65, 15, 8, 7)
potom <- c(120, 105, 45, 50, 40, 42, 80, 21, 2, 12)
```

7.2 Riešenia k príkladom

Príklad 7.1

Keďže ide o viac početný súbor ($n > 30$), na výpočet intervalov spoľahlivosti môžeme použiť kvantily normálneho rozdelenia (predpokladáme platnosť centrálnej limitnej vety).

```
> vydavky <- c(2.32, 6.61, 6.9, 8.04, 9.45, 10.26, 11.34, 11.63,
  12.66, 12.95, 13.67, 13.72, 14.35, 14.52, 14.55, 15.01, 15.33,
  16.55, 17.15, 18.22, 18.3, 18.71, 19.54, 19.55, 20.58, 20.89,
  20.91, 21.13, 23.85, 26.04, 27.07, 28.76, 29.15, 30.54, 31.99,
  32.82, 33.26, 33.8, 34.76, 36.22, 37.52, 39.28, 40.8, 43.97,
  45.58, 52.36, 61.57, 63.85, 64.3, 69.49)
> mean (vydavky)
[1] 25.8364
> mean (vydavky) - abs(qnorm(0.05/2)) * (var(vydavky)/50)^0.5
[1] 21.35900
> mean (vydavky) + abs(qnorm(0.05/2)) * (var(vydavky)/50)^0.5
[1] 30.31380
```

Priemerné výdavky spotrebiteľov sú teda 25.8364 USD, čo je náš najlepší bodový odhad strednej hodnoty. Pri $\alpha = 0.05$ je dolná hranica intervalu spoľahlivosti 21.35900 a horná na úrovni 30.31380.

Príklad 7.2

Údaje potrebné na výpočet je možné do softvéru R vložiť cez príkaz `read.csv` alebo ich môžeme zadať aj manuálne vo forme vektora. Keďže ide o nízky počet pozorovaní, zvolíme na ukážku druhý spôsob. Taktiež intervaly spoľahlivosti môžeme vypočítať manuálne alebo cez funkciu `t.test()`. Tu opäť zvolíme na ukážku zložitejší postup, teda vypočítame intervaly spoľahlivosti manuálne.

```
> HDP <- c(7.0, 7.7, 8.3, 8.8, 9.0, 9.5, 10.4, 11.1, 11.5, 12.3,
  13.5, 15.0, 16.9, 18.2, 17.0, 18.0)
> mean(HDP)
[1] 12.1375
> mean(HDP) - abs(qt(0.05/2, df  = 15)) * (var(HDP)/16)^0.5
[1] 10.09319
> mean(HDP) + abs(qt(0.05/2, df  = 15)) * (var(HDP)/16)^0.5
[1] 14.18181
-------------------------------------------------------------
> Spotreba <- c(5.1, 5.9, 6.3, 6.8, 6.9, 7.3, 8.1, 8.7, 8.9,
  9.4, 10.3, 11.4, 12.4, 13.6, 13.8, 14.0)
> mean(Spotreba)
[1] 9.30625
> mean(Spotreba) - abs(qt(0.05/2, df  = 15)) *
  (var(Spotreba)/16)^0.5
[1] 7.725349
```

```
> mean(Spotreba) + abs(qt(0.05/2, df  = 15)) *
  (var(Spotreba)/16)^0.5
[1] 10.88715
```

Výberový priemer HDP na jedného obyvateľa v SR bolo v danom období 12.1375. Spodný interval spoľahlivosti pre strednú hodnotu je 10.09319 a horný interval spoľahlivosti je 14.18181. Výberový priemer pre spotrebu na jedného obyvateľa v SR je 9.30625, s dolným intervalom spoľahlivosti strednej hodnoty 7.725349 a horným 10.88715.

Pre korektnosť uvádzame aj postup, v ktorom transformujeme premenné pomocou funkcie `diff()` na diferencie (pracujeme teda so zmenami HDP a zmenami v spotrebe). Po takejto transformácii môžeme predpokladať, že premenné sú *iid*.

```
> HDP <- c(7.0, 7.7, 8.3, 8.8, 9.0, 9.5, 10.4, 11.1, 11.5, 12.3,
  13.5, 15.0, 16.9, 18.2, 17.0, 18.0)
> dHDP <- diff(HDP)
> mean(dHDP)
[1] 0.7333333
> mean(dHDP) - abs(qt(0.05/2, df  = 15)) * (var(dHDP)/16)^0.5
[1] 0.3605110
> mean(dHDP) + abs(qt(0.05/2, df  = 15)) * (var(dHDP)/16)^0.5
[1] 1.106156
-----------------------------------------------------------------
> Spotreba <- c(5.1, 5.9, 6.3, 6.8, 6.9, 7.3, 8.1, 8.7, 8.9,
  9.4, 10.3, 11.4, 12.4, 13.6, 13.8, 14.0)
> dSpotreba <- diff(Spotreba)
> mean(dSpotreba)
[1] 0.5933333
> mean(dSpotreba) - abs(qt(0.05/2, df  = 15)) *
  (var(dSpotreba)/16)^0.5
[1] 0.4039002
> mean(dSpotreba) + abs(qt(0.05/2, df  = 15)) *
  (var(dSpotreba)/16)^0.5
[1] 0.7827664
```

Príklad 7.3

To čo musíme spraviť najprv, je oddeliť tie ceny skúmaných bytov, ktorých rozloha je viac ako 50 m^2. Za týmto účelom využijeme funkciu `subset()` (funkciu je vhodné použiť bez ohľadu na spôsob importu dát). Potom postupujeme tak ako v predchádzajúcom príklade, avšak keďže z daných dát vyberáme len určitú podmnožinu, bude sa meniť počet pozorovaní a aj počet stupňov voľnosti. Na porovnanie vypočítame intervaly spoľahlivosti pre strednú hodnotu aj s využitím funkcie `t.test()`.

```
> Cena <- c(113500, 64000, 36500, 63000, 69000, 48400, 66900,
  86000, 61900, 63000, 81500, 72500, 109800, 89900, 52700)
```

```
> Rozloha <- c(96, 68, 43, 56, 51, 37, 67, 92, 68, 64, 74, 71,
  99.6, 80, 42)
-----------------------------------------------------------------
> Cena_x <- subset(Cena, subset = Rozloha > 50)
> length(Cena)
[1] 15
> length(Cena_x)
[1] 12
> mean(Cena_x)
[1] 78416.67
> mean(Cena_x) - abs(qt(0.05/2, df  = length(Cena_x) - 1)) *
  (var(Cena_x)/length(Cena_x))^0.5
[1] 66879.47
> mean(Cena_x) + abs(qt(0.05/2, df  = length(Cena_x) - 1)) *
  (var(Cena_x)/length(Cena_x))^0.5
[1] 89953.86
-----------------------------------------------------------------
> t.test(Cena_x, alternative = "two.sided", conf.level = 0.95)

                    One Sample t-test

data:  Cena_x
t = 14.9598, df = 11, p-value = 1.172e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
66879.47 89953.86
sample estimates:
mean of x
78416.67
```

Našim najlepším odhadom strednej hodnoty ceny bytov v Košiciach s rozlohou viac ako 50 m^2 je 78416.67 EUR s dolným intervalom spoľahlivosti 66879.47 a s horným 89953.86. Výpočet intervalov spoľahlivosti je rovnaký bez ohľadu na to, či sme ich počítali manuálne alebo s využitím funkcie `t.test()`. Pre zjednodušenie preto v ďalších príkladoch budeme využívať už len túto funkciu, ale čitateľ si samozrejme môže prepočítať intervaly spoľahlivosti vždy aj manuálne (pre lepšie nadobudnutie rutiny).

Príklad 7.4

Postupovať budeme podobne ako v predchádzajúcom príklade. Autá rozdelíme podľa rýchlosti na tie, ktoré dosiahli rýchlosť viac ako 15 míľ za hodinu (vrátane). Pre zjednodušenie už intervaly spoľahlivosti nebudeme počítať manuálne, ale len s využitím funkcie `t.test()`.

```
> library(datasets)
> rychlost_15 = subset(cars$dist, subset = cars$speed >= 15)
-----------------------------------------------------------------
```

```
> t.test(rychlost_15, alternative = "two.sided", conf.level =
  0.95)

                      One Sample t-test

data:  rychlost_15
t = 12.489, df = 26, p-value = 1.721e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
47.46379 66.16584
sample estimates:
mean of x
56.81481
```

Z výsledkov môžeme vidieť, že v roku 1920 mali autá do značnej miery dlhú brzdnú dráhu. Pri rýchlosti 15 míľ za hodinu (približne 24 *km*/*h*) a viac, mali autá priemernú brzdnú dráhu od 47.46 až do 66.17 stôp (teda približne od 14 *m* až do 20 *m*). Nízke rýchlosti a vysoké brzdné dráhy je zrejme možné vysvetliť tým, že tieto merania sa uskutočnili v roku 1920. Ostatné úlohy necháme na čitateľa.

Príklad 7.5

V tomto príklade musíme opäť vypočítať oddelene priemery za dve skupiny v rámci jednej údajovej databázy. Samotný výpočet týchto priemerov je možné realizovať pomocou funkcie `aggregate()`. Na výpočet intervalov spoľahlivosti však opäť použijeme rozdelenie dát do dvoch skupín pomocou funkcie `subset()`. Ak by sme použili funkciu `t.test(Rabbit$BPchange ~ Rabbit$Treatment)` vykonali by sme *t*-test zhody dvoch stredných hodnôt.

```
> library(MASS)
> aggregate(Rabbit$BPchange, list(gp = Rabbit$Treatment), mean)
       gp         x
1 Control 13.558333
2     MDL  8.878333
-----------------------------------------------------------------
> MDL <- subset(Rabbit$BPchange, subset = Rabbit$Treatment ==
  "MDL")
> t.test(MDL, alternative = "two.sided", conf.level = 0.95)

                      One Sample t-test

data:  MDL
t = 4.6478, df = 29, p-value = 6.744e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
4.971516 12.785151
sample estimates:
mean of x
```

```
8.878333
-----------------------------------------------------------------
> Control <- subset(Rabbit$BPchange, subset = Rabbit$Treatment
  == "Control")
> t.test(Control, alternative = "two.sided", conf.level = 0.95)

                        One Sample t-test

data:  Control
t = 6.1199, df = 29, p-value = 1.147e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
9.027264 18.089403
sample estimates:
mean of x
13.55833
```

Môžeme vidieť, že priemerná zmena krvného tlaku bola nižšia pri skupine MDL (aj keď intervaly spoľahlivosti sú dosť široké), teda liek môžeme považovať za účinný. Aj bez podrobnejších znalostí z medicíny je zrejme čitateľovi jasné, že využitie štatistiky nemá obmedzenia podľa oblasti výskumu.

Príklad 7.6

Pri riešení otázky, či krajiny známe pod akronymom PIIGGS dodržiavajú hranicu verejného dlhu v pomere k HDP pod úrovňou 60 % (podľa tzv. Paktu stability a rastu), využijeme funkciu `t.test()` a konfidenčné intervaly, ktoré sú uvedené vo výstupe tejto funkcie.

```
> data <- read.csv(file = "...cesta k súboru...\\debt_gdp.csv",
  sep = ";", dec = ".", header = T)
-----------------------------------------------------------------
> results <- matrix(ncol = 5, nrow = 6)
> rownames(results) <- names(data)[3:8]
> colnames(results) <- c("priemer", "dolná hr.", "horná hr.",
  "štatis.", "p-hod.")
> for (i in 1:6) {
+ a <- t.test(data[,i+2], alternative = "two.sided", conf.level
  = 0.95)
+ results[i, 1] <- round(a$estimate, 2)
+ results[i, 2] <- round(a$conf.int[1], 2)
+ results[i, 3] <- round(a$conf.int[2], 2)
+ }
-----------------------------------------------------------------
> results
             priemer dolná hr. horná hr.
Portugal       60.85     58.03     63.66
Ireland        37.99     32.79     43.19
Italy         109.35    108.07    110.63
Greece        107.96    104.48    111.44
```

```
Great.Britain   45.43     42.00     48.86
Spain           47.35     45.04     49.66
```

Z výsledkov vyplýva, že výšku verejného dlhu v pomere k HDP pod úrovňou 60 % nedodržiavajú Grécko, Portugalsko a Taliansko. V priemere za sledované obdobie dosiahli obe krajiny zadlženosť viac ako 100 %. Taliansko 109.35 % s dolným intervalom spoľahlivosti 108.07 % a s horným 110.63 %. Grécko malo priemernú výšku verejného dlhu k HDP na úrovni 107.96 % s intervalmi spoľahlivosti 104.48 % – 111.44 %. Zrejme teda práve tieto krajiny môžeme vyhodnotiť ako problematické vzhľadom na ich veľmi vysokú zadlženosť.

Príklad 7.7

V danej databáze je jeden respondent, pri ktorom nie je uvedené jeho pohlavie (z akých príčin je nepodstatné). Toto pozorovanie teda do úvahy nemôžeme brať a odstránime ho pomocou funkcie `na.omit()`. Intervaly spoľahlivosti môžeme vypočítať manuálne, prípadne vhodnou alternatívou, ktorá výpočet do značnej miery zjednodušuje, je funkcia `prop.test()`.

```
> library(MASS)
> gender <- na.omit(survey$Sex)
> n <- length(gender)
> M <- sum(gender == "Male")
> podiel = M/n; podiel
[1] 0.5
> I <- abs(qnorm(0.05/2)) * sqrt(podiel * (1 - podiel)/n)
> CI <- podiel + c(-I, I); CI
[1] 0.4362086 0.5637914
-----------------------------------------------------------------
> prop.test(M, n, alternative = "two.sided", conf.level = 0.95)

   1-sample proportions test without continuity correction

data:  M out of n, null probability 0.5
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.4367215 0.5632785
sample estimates:
p
0.5
```

Na základe tejto vzorky môžeme povedať, že na danej univerzite je podiel mužov a žien vyvážený (0.5), ale ak vezmeme do úvahy aj intervaly spoľahlivosti, tak sa podiel mužov môže pohybovať od 0.4367 do 0.5633. Vo funkcii `prop.test()` je možné nastaviť

ešte jeden parameter – `correct`. Tento parameter je vhodné použiť pri menších vzorkách a slúži na úpravu Pearsonovej χ^2 štatistiky. Táto korekcia však má svoje nevýhody (vysoká chyba II. typu), a preto sa neodporúča jej použitie. Výsledky dosiahnuté funkciou `prop.test()` sú mierne odlišné od tých, ktoré sú vypočítané manuálne[30].

Príklad 7.8

Keďže opäť niektoré údaje nie sú dostupné, využijeme funkciu `na.omit()`. V premennej `FIN` sú odpovede respondentov kódované nasledovne: finančná situácia sa zhoršila (1), ostala nezmenená (2) a je lepšia ako pred rokom (3). Nás budú zaujímať len hodnoty `FIN` = 1.

```
> data <- read.csv(file = "...cesta
  k súboru...\\montana_survey.csv", sep = ";", dec = ".", header
  = T)
-----------------------------------------------------------------
> FIN_1 <- sum(na.omit(data$FIN == "1")); FIN_1
[1] 61
> n <- length(na.omit(data$FIN)); n
[1] 208
> podiel <- FIN_1/n; podiel
[1] 0.2932692
> I <- abs(qnorm(0.05/2)) * sqrt(podiel * (1-podiel)/n)
> CI <- podiel + c(-I,I); CI
[1] 0.2313997 0.3551387
-----------------------------------------------------------------
> prop.test(FIN_1, n, alternative = "two.sided", conf.level =
  0.95, correct = FALSE)

    1-sample proportions test without continuity correction

data:  FIN_1 out of n, null probability 0.5
X-squared = 35.5577, df = 1, p-value = 2.476e-09
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2355975 0.3584385
sample estimates:
p
0.2932692
```

Z uvedených výsledkov môžeme dospieť k záveru, že s 95 % konfidenciou sa finančná situácia obyvateľov Montany oproti predchádzajúcemu roku zhoršila u približne 23 % až 36 % obyvateľov.

---

[30] Bližšie pozri `http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf` (s. 62).

Príklad 7.9

Z uvedenej databázy nás bude zaujímať len premenná `marijuana`. Táto premenná nadobúda hodnoty 1 (ak respondent má skúsenosti s drogou) a 2 (ak respondent nemá skúsenosti s drogou).

```
> library(UsingR)
> MAR <- sum(samhda$marijuana == "1"); MAR
[1] 134
> n <- length(samhda$marijuana); n
[1] 600
> podiel <- MAR/n; podiel
[1] 0.2233333
> I <- abs(qnorm(0.05/2)) * sqrt(podiel * (1-podiel)/n)
> CI <- podiel + c(-I,I); CI
[1] 0.1900086 0.2566581
------------------------------------------------------------
> prop.test(MAR, n, alternative = "two.sided", conf.level =
  0.95, correct = FALSE)

   1-sample proportions test without continuity correction

data:  MAR out of n, null probability 0.5
X-squared = 183.7067, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.1918283 0.2583586
sample estimates:
p
0.2233333
```

Podľa dostupných údajov sme dospeli k záveru, že v roku 1996 malo s 95 % konfidenciou skúsenosti s marihuanou približne 19 % až 26 % amerických študentov.

Príklad 7.10

Strednú hodnotu veku študentov aj s intervalmi spoľahlivosti už odhadnúť vieme. Na zodpovedanie otázky, či je stredná hodnota veku študentov univerzity 18 rokov a viac, využijeme funkciu `t.test()`. Nulová hypotéza v tomto prípade je H_0: $\mu \leq 18.0$ oproti alternatívnej H_1: $\mu > 18.0$. Ak nulovú hypotézu budeme vedieť zamietnuť na stanovenej hladine významnosti (0.05), tak môžeme predpokladať, že priemerný vek študentov univerzity je viac ako 18 rokov. Pre lepšiu kontrolu premenných, ktoré sú uvedené vo výstupe funkcie `t.test()`, uvádzame aj manuálne výpočty.

```
> library(MASS)
> Vek <- survey$Age
> mean(Vek)
[1] 20.37451
```

```
> tstat <- (mean(Vek) - 18)/(sd(Vek)/sqrt(length(Vek))); tstat
[1] 5.646169
> df <- length(Vek) - 1; df
[1] 236
> p_value <- 1 - pt(tstat,df); p_value
[1] 2.346504e-08
> I <- abs(qt(0.95, df = length(Vek)-1)) *
  (sd(Vek)/sqrt(length(Vek)))
> CI <- mean(Vek)+c(-I, I); CI
[1] 19.68004 21.06899
-----------------------------------------------------------------
> t.test(Vek, alternative = c("greater"), mu = 18, conf.level =
  0.95)

                    One Sample t-test

data:  Vek
t = 5.6462, df = 236, p-value = 2.347e-08
alternative hypothesis: true mean is greater than 18
95 percent confidence interval:
19.68004      Inf
sample estimates:
mean of x
20.37451
```

Nulovú hypotézu môžeme zamietnuť na hladine významnosti 0.01 (za určitých okolností môžeme na p-hodnotu nazerať, ako na najmenšiu možnú hladinu významnosti, na ktorej by sme mohli zamietnuť nulovú hypotézu, obvykle nás však zaujímajú iba hladiny 0.1, 0.05 a 0.01. Ak vieme zamietnuť nulovú hypotézu na hladine významnosti 2.347e-08, určite vieme aj na 0.1, 0.05, a 0.01. Z tejto trojice si pri interpretácii vyberieme najmenšiu možnú hodnotu, čo je v našom prípade 0.01). Z toho vyplýva, že stredná hodnota veku študentov univerzity je (aj pri $\alpha = 0.05$) 18 rokov alebo viac. Mohli sme použiť aj obojstranný *t*-test, kde nulová hypotéza je H_0: $\mu = 18.0$ a alternatívna H_1: $\mu \neq 18.0$. Avšak *p*-hodnotu z jednostranného testu by sme potom museli vydeliť na polovicu.

```
> t.test(Vek, alternative = c("two.sided"), mu = 18, conf.level
  = 0.95)

                    One Sample t-test

data:  Vek
t = 5.6462, df = 236, p-value = 4.693e-08
alternative hypothesis: true mean is not equal to 18
95 percent confidence interval:
19.54600 21.20303
sample estimates:
mean of x
20.37451
```

Príklad 7.11

Na zodpovedanie otázky, či riaditeľ podniku v USA zarobí v priemere viac ako 300 000 USD ročne, použijeme jednostranný *t*-test s nulovou hypotézou H_0: $\mu \leq 30.0$ a alternatívnou H_1: $\mu > 30.0$ (hodnoty v databáze sú uvedené v 10 000 USD).

```
> library(UsingR)
> mean(exec.pay)
[1] 59.88945
> t.test(exec.pay, alternative = c("greater"), mu = 30,
  conf.level = 0.95)

                      One Sample t-test

data:  exec.pay
t = 2.0365, df = 198, p-value = 0.02152
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
35.63457      Inf
sample estimates:
mean of x
59.88945
```

Nulovú hypotézu môžeme zamietnuť na hladine 5 % (keďže *p*-hodnota = 0.02152), teda môžeme prijať alternatívnu hypotézu, že stredná hodnota ročného príjmu riaditeľa podniku v USA je viac ako 300 000 USD.

Na výpočet obojstranných intervalov spoľahlivosti pre priemernú mzdu (598 894.5) môžeme tiež využiť *t*-test, ale ako alternatívnu hypotézu vo funkcii zvolíme `two.sided` – teda obojstranný test.

```
> library(UsingR)
> mean(exec.pay)
[1] 59.88945
> t.test(exec.pay, alternative = c("two.sided"), mu = 30,
  conf.level = 0.95)

                      One Sample t-test

data:  exec.pay
t = 2.0365, df = 198, p-value = 0.04303
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
30.94629 88.83260
sample estimates:
mean of x
59.88945
```

Stredná hodnota ročného príjmu riaditeľa podniku v USA je od 309 462.9 USD až 888 326 USD (pri $\alpha = 0.05$).

Príklad 7.12

Na zodpovedanie otázky, či ľudia zarábajúci 40 000 USD ročne a viac majú strednú hodnotu úspor vyššiu ako 7 000 USD, použijeme najprv funkciu `subset()`, aby sme vyselektovali len potrebné pozorovania a následne vykonáme test strednej hodnoty oproti konštante = 7 000 USD. Keďže nás však zaujíma, či je možné strednú hodnotu považovať za väčšiu ako 7000, je výhodné úlohu formulovať tak, aby uvedené tvrdenie bolo v alternatívnej hypotéze. Tú na rozdiel od nulovej hypotézy vieme totiž prijať. Nulová hypotéza teda je H_0: $\mu \leq 7\ 000$ a alternatívna H_1: $\mu > 7\ 000$. Vo funkcii `t.test()` vhodne upravíme možnosť `alternative`.

```
> library(UsingR)
> Saving_40000 <- subset(cfb$SAVING, subset = cfb$INCOME >=
  40000)
> length(Saving_40000)
[1] 486
> t.test(Saving_40000, alternative = c("greater"), mu = 7000,
  conf.level = 0.95)

                      One Sample t-test

data:  Saving_40000
t = 1.815, df = 485, p-value = 0.03507
alternative hypothesis: true mean is greater than 7000
95 percent confidence interval:
7210.17     Inf
sample estimates:
mean of x
9284.3
```

Nulovú hypotézu zamietame na hladine 5 % pri *p*-hodnote 0.03507. Ak by sme chceli odhadnúť, v akom intervale sa pohybuje stredná hodnota úspor tejto časti respondentov (ide nám pre zmenu o obojstranné intervaly spoľahlivosti), postupovali by sme podobne ako v predchádzajúcom príklade.

```
> library(UsingR)
> Saving_40000 <- subset(cfb$SAVING, subset = cfb$INCOME >=
  40000)
> length(Saving_40000)
[1] 486
> t.test(Saving_40000, alternative = c("two.sided"), mu = 7000,
  conf.level = 0.95)

                      One Sample t-test
```

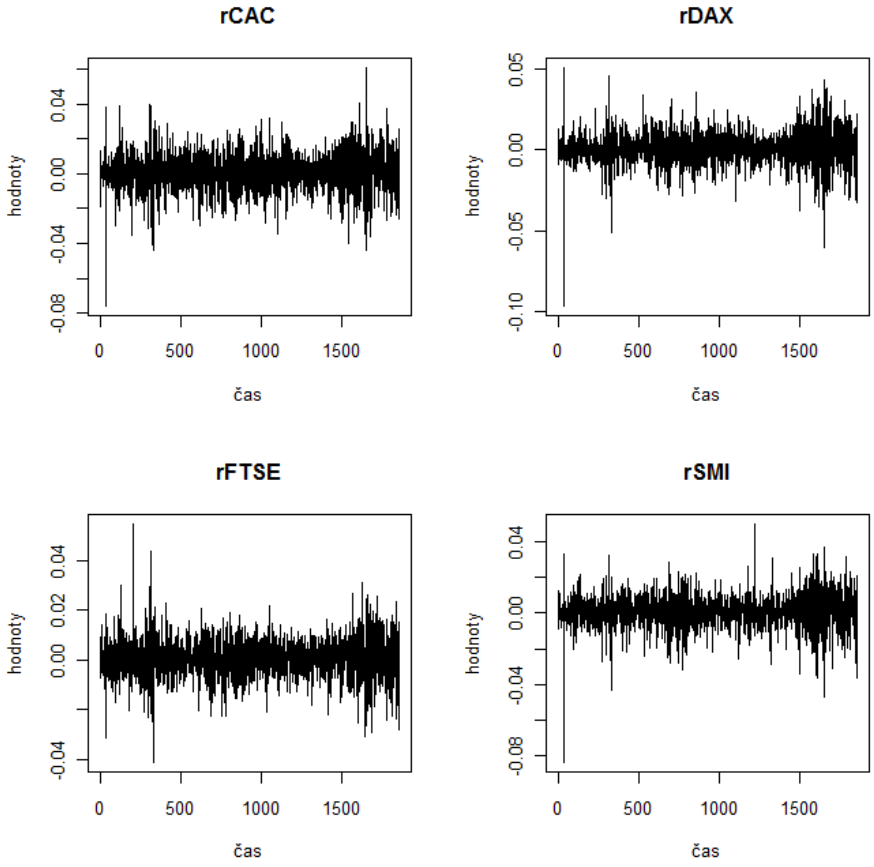
```
data:  Saving_40000
t = 1.815, df = 485, p-value = 0.07014
alternative hypothesis: true mean is not equal to 7000
95 percent confidence interval:
6811.372 11757.229
sample estimates:
mean of x
9284.3
```

Stredná hodnota výšky úspor obyvateľov, ktorí zarobia ročne viac ako 40 000 USD sa s 95 % konfidenciou pohybuje v rozmedzí od 6 811.372 USD do 11 757.229 USD. Vzhľadom na predchádzajúce zadanie by však bolo logickejšie vypočítať jednostranný, ľavostranný interval spoľahlivosti, ktorého výsledok je však súčasťou výstupu z predchádzajúcej funkcie.

Príklad 7.13

Uzatváracie ceny indexov najprv transformujeme na spojité výnosy. Pre lepšiu predstavu ako takéto výnosy vyzerajú si ich zobrazíme aj graficky.

```
> library(datasets)
> attach(data.frame(EuStockMarkets))
> rCAC <- diff(log(CAC))
> rDAX <- diff(log(DAX))
> rFTSE <- diff(log(FTSE))
> rSMI <- diff(log(SMI))
-------------------------------------------------------------
> par(mfrow = c(2,2))
> plot(rCAC, type = "l", main = "rCAC", ylab = "hodnoty", xlab =
  "čas")
> plot(rDAX, type = "l", main = "rDAX", ylab = "hodnoty", xlab =
  "čas")
> plot(rFTSE, type = "l", main = "rFTSE", ylab = "hodnoty", xlab
  = "čas")
> plot(rSMI, type = "l", main = "rSMI", ylab = "hodnoty", xlab =
  "čas")
```



Obrázok 7.1: Spojité výnosy akciových indexov

Zdroj: výstup zo softvéru R

Z uvedeného obrázku (Obrázok 7.1) môžeme vidieť, že priemer bude blízky nule. Zo skúseností však vieme, že tieto časové rady nie je možné považovať za nezávislé náhodné realizácie. Inak povedané, výška výnosu v čase t bude štatisticky závisieť od výšky výnosov v čase $t - k$, kde k je počet oneskorení. Pri použití štandardných testov však potrebujeme, aby pozorovania boli štatisticky nezávislé. Preto by v praxi nebolo vhodné v tomto príklade pokračovať štandardným postupom. Pre ilustráciu však budeme v tomto príklade ďalej pokračovať a predpokladať (aj keď nesprávne), že ide o *iid* pozorovania. Uskutočníme nasledujúce testovanie, kde nás zaujíma, či priemerný denný výnos týchto indexov je menší ako 0.1 %. Nulová hypotéza teda je H_0: $\mu \geq 0.001$ a alternatívna H_1: $\mu < 0.001$. Znovu sme do alternatívnej hypotézy umiestnili tvrdenie, ktorého prípadná pravdivosť nás zaujíma.

```
> data <- data.frame(rCAC, rDAX, rFTSE, rSMI)
> results <- matrix(ncol = 5, nrow = 4)
> rownames(results) <- names(data)
> colnames(results) <- c("priemer", "dolná hr.", "horná hr.",
  "štatis.", "p-hod.")
> for (i in 1:4) {
```

```
+   a <- t.test(data[,i], alternative = "less", mu = 0.001,
  conf.level = 0.95)
+   results[i, 1] <- round(a$estimate, 4)
+   results[i, 2] <- round(a$conf.int[1], 4)
+   results[i, 3] <- round(a$conf.int[2], 4)
+   results[i, 4] <- round(a$statistic, 2)
+   results[i, 5] <- round(a$p.value, 2)
+ }
------------------------------------------------------------------
> results
      priemer dolná hr. horná hr. štatis. p-hod.
rCAC    4e-04      -Inf    0.0009   -2.20   0.01
rDAX    7e-04      -Inf    0.0010   -1.46   0.07
rFTSE   4e-04      -Inf    0.0007   -3.08   0.00
rSMI    8e-04      -Inf    0.0012   -0.85   0.20
```

Uvedenú hypotézu môžeme zamietnuť na hladine 5 % v prípade indexov CAC a FTSE. Tieto dva indexy teda vykazujú priemerný denný výnos menší ako 0.1 %. Nemecký DAX a švajčiarsky SMI zrejme dosahujú priemernú dennú výnosnosť rovnú alebo väčšiu ako 0.1 %, keďže pri týchto indexoch sme nulovú hypotézu na stanovenej hladine významnosti nevedeli zamietnuť.

Príklad 7.14

Za účelom zistenia, či krajiny známe pod akronymom PIIGGS dosahujú zadlženosť štatisticky významne nižšiu ako 60 % HDP použijeme funkciu `t.test()`. Upozorňujeme, že podobne ako v niektorých predošlých príkladoch predpokladáme nezávislosť a náhodnosť pozorovaní, ktoré pochádzajú z rovnakého rozdelenia pravdepodobnosti (tento predpoklad nie je v tomto prípade správny, ale urobíme ho z pedagogických dôvodov). Testujeme nulovú hypotézu H_0: $\mu \geq 60$ voči alternatívnej hypotéze H_1: $\mu < 60$.

```
> data <- read.csv(file = "...cesta k súboru...\\debt_gdp.csv",
  sep = ";", dec = ".", header = T)
------------------------------------------------------------------
> results <- matrix(ncol = 5, nrow = 6)
> rownames(results) <- names(data)[3:8]
> colnames(results) <- c("priemer", "dolná hr.", "horná hr.",
  "štatis.", "p-hod.")
> for (i in 1:6) {
+  a <- t.test(data[,i+2], alternative = "less", mu = 60,
  conf.level = 0.95)
+  results[i, 1] <- round(a$estimate, 2)
+  results[i, 2] <- round(a$conf.int[1], 2)
+  results[i, 3] <- round(a$conf.int[2], 2)
+  results[i, 4] <- round(a$statistic, 2)
+  results[i, 5] <- round(a$p.value, 2)
+ }
------------------------------------------------------------------
```

```
> results
              priemer dolná hr. horná hr. štatis. p-hod.
Portugal        60.85      -Inf     63.19    0.61   0.73
Ireland         37.99      -Inf     42.32   -8.56   0.00
Italy          109.35      -Inf    110.42   77.71   1.00
Greece         107.96      -Inf    110.86   27.89   1.00
Great.Britain   45.43      -Inf     48.29   -8.59   0.00
Spain           47.35      -Inf     49.27  -11.07   0.00
```

V prípade krajín Írsko, Veľká Británia a Španielsko môžeme zamietnuť nulovú hypotézu na stanovenej hladine významnosti. Pre tieto krajiny môžeme prijať alternatívnu hypotézu, že pomer verejného dlhu k HDP je menší ako 60 %, a teda tieto krajiny dodržiavajú podmienku stanovenú v Pakte stability a rastu. Pre krajiny Portugalsko, Grécko a Taliansko nulovú hypotézu zamietnuť nevieme, máme teda určitý dôvod predpokladať, že tieto krajiny stanovenú úroveň verejného dlhu k HDP nedodržiavajú.

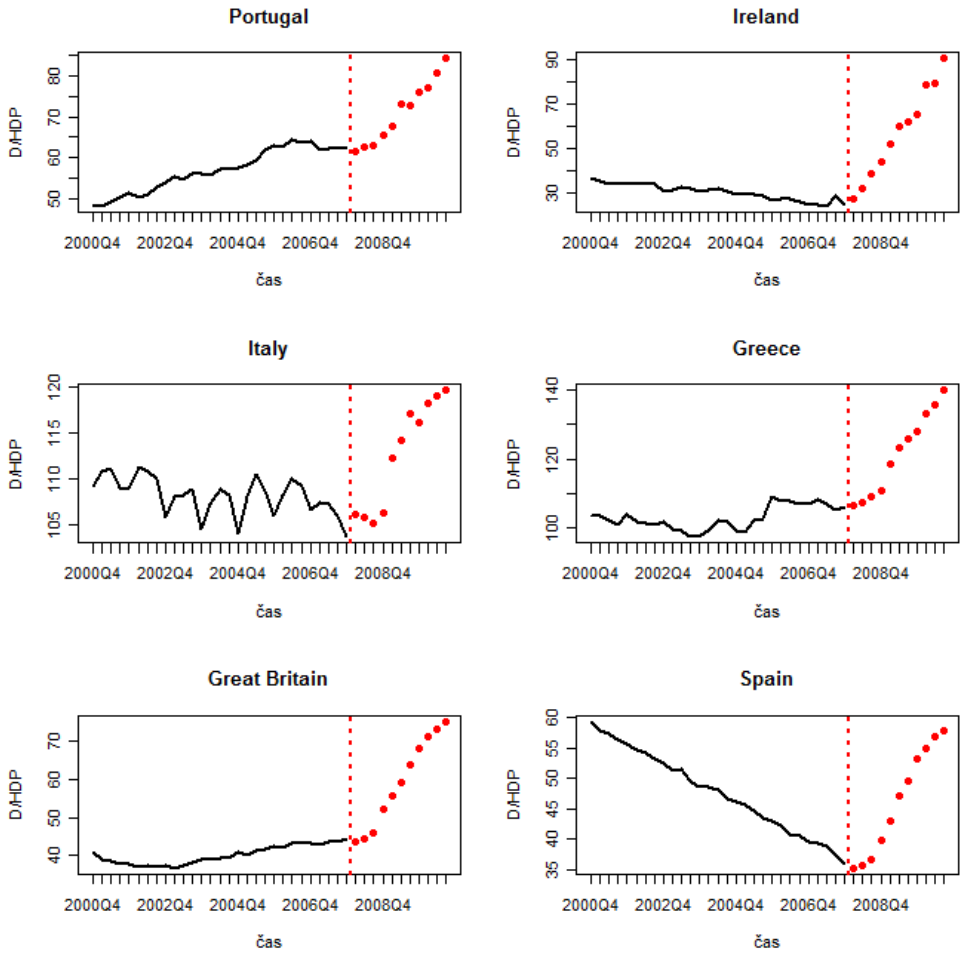
Ak by sme využili obojstranné intervaly (výsledky neuvádzame) spoľahlivosti pre priemerné hodnoty, potom by sme v takomto prípade považovali za problematické krajiny (z pohľadu ich zadlženosti) len Grécko a Taliansko. Pri Portugalsku by sme dostali priemernú hodnotu dlhu k HDP na úrovni 60.84 % (čiže mierne prevyšujúcu dané kritérium), avšak so spodným intervalom spoľahlivosti na úrovni 58.03 %.

Na základe induktívnej štatistiky sme dospeli k záveru, že problematické krajiny (nespĺňajúce kritérium verejného dlhu k HPD na úrovni 60 %) sú Portugalsko, Taliansko a Grécko. Bez použitia induktívnej štatistiky by sme vyhodnotili za problematické len Taliansko a Grécko. Pre lepšie pochopenie skutočnosti, prečo aj ostatné krajiny sú zaradené medzi problematické a označené mierne degradujúcim akronymom, je vhodné pozrieť sa na grafickú podobu skúmaných dát.

```
> data = read.csv(file = "...cesta k súboru...\\debt_gdp.csv",
  sep = ";", dec = ".", header = T)
> par(mfrow = c(3,2))
-----------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Portugal[1:29], type = "l",
  ylab = "D/HDP", main = "Portugal", xlab = "čas", col =
  "black", lwd=2, xaxt = "n", xlim = c(1, max(data$time)), ylim
  = c(min(data$Portugal), max(data$Portugal)))
> points(x = data$time[30:40], y = data$Portugal[30:40], col =
  "red", pch = 19)
> axis(side = 1, at = data$time, labels = data$obs)
> abline(v = 29.5, lty=3, lwd=2, col = "red")
-----------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Ireland[1:29], type = "l",
  ylab = "D/HDP", main = "Ireland", xlab = "čas", col = "black",
  lwd=2, xaxt = "n", xlim = c(1, max(data$time)), ylim =
  c(min(data$Ireland), max(data$Ireland)))
```

```
> points(x = data$time[30:40], y = data$Ireland[30:40], col =
  "red", pch = 19)
> axis(side = 1, at = data$time, labels = data$obs)
> abline(v = 29.5, lty=3, lwd=2, col = "red")
-------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Italy[1:29], type = "l",
  ylab = "D/HDP", main = "Italy", xlab = "čas", col = "black",
  lwd=2, xaxt = "n", xlim = c(1, max(data$time)), ylim =
  c(min(data$Italy), max(data$Italy)))
> points(x = data$time[30:40], y = data$Italy[30:40], col =
  "red", pch = 19)
> axis(side = 1, at = data$time, labels = data$obs)
> abline(v = 29.5, lty=3, lwd=2, col = "red")
-------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Greece[1:29], type = "l",
  ylab = "D/HDP", main = "Greece", xlab = "čas", col = "black",
  lwd=2, xaxt = "n", xlim = c(1, max(data$time)), ylim =
  c(min(data$Greece), max(data$Greece)))
> points(x = data$time[30:40], y = data$Greece[30:40], col =
  "red", pch = 19)
> axis(side = 1, at = data$time, labels = data$obs)
> abline(v = 29.5, lty=3, lwd=2, col = "red")
-------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Great.Britain[1:29], type =
  "l", ylab = "D/HDP", main = "Great Britain", xlab = "čas", col
  = "black", lwd=2, xaxt = "n", xlim = c(1, max(data$time)),
  ylim = c(min(data$Great.Britain), max(data$Great.Britain)))
> points(x = data$time[30:40], y = data$Great.Britain[30:40],
  col = "red", pch = 19)
> axis(side = 1, at = data$time, labels = data$obs)
> abline(v = 29.5, lty=3, lwd=2, col = "red")
-------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Spain[1:29], type = "l",
  ylab = "D/HDP", main = "Spain", xlab = "čas", col = "black",
  lwd=2, xaxt = "n", xlim = c(1, max(data$time)), ylim =
  c(min(data$Spain), max(data$Spain)))
> points(x = data$time[30:40], y = data$Spain[30:40], col =
  "red", pch = 19)
> axis(side = 1, at = data$time, labels = data$obs)
> abline(v = 29.5, lty=3, lwd=2, col = "red")
```

Z grafickej vizualizácie dát môžeme vidieť, že problémom nie je ani tak samotná výška ukazovateľov D/HDP a nedodržanie podmienok Paktu stability a rastu (pomer verejného dlhu k HDP na úrovni 60 %), ale rastúci trend zadlženosti od roku 2008. Tento rastúci trend je najvýraznejší v prípade Írska, ktoré pred rokom 2008 vykazovalo priemernú úroveň zadlženosti k HDP na veľmi nízkej úrovni (približne 30 %) a dokonca s klesajúcim trendom. Po roku 2008 vplyvom finančnej krízy museli krajiny prijať opatrenia, ktoré zvyšovali verejný dlh, avšak zároveň dochádzalo k poklesu HDP. Ukazovateľ D/HDP teda súbežne narastal kvôli zvyšovaniu čitateľa zlomku (dlh) a súčasne kvôli znižovaniu menovateľa (HDP).



Obrázok 7.2: Vývoj verejného dlhu k HDP krajín PIIGGS pred krízou a počas krízy

Zdroj: výstup zo softvéru R

Príklad 7.15

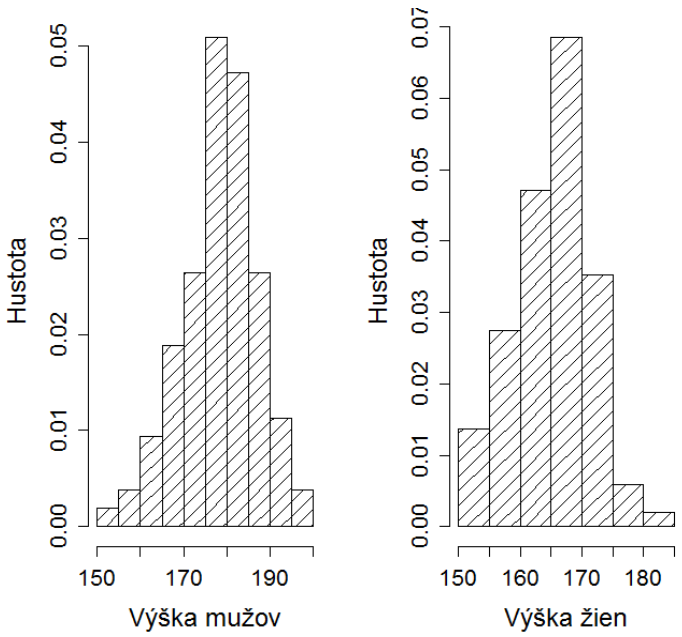
Najprv si z celej databázy oddelíme výšku mužov a žien s využitím funkcie `subset()`. Keďže niektoré pozorovania chýbajú, použijeme funkciu `na.omit()`. Po odstránení chýbajúcich údajov máme k dispozícii 106 pozorovaní pre mužov a 102 pozorovaní pre ženy. Vypočítané priemery pre tieto dve skupiny sú odlišné (178.8 *cm* u mužov a 165.7 *cm* u žien) a predstavujú bodové odhady stredných hodnôt v populácii. Napriek tomu, že tieto bodové odhady sú odlišné, je dôležité si uvedomiť, že ide o štatistiky, ktorých hodnota závisí od toho, ktorí študenti sa dostali do našej vzorky. Keďže túto vzorku sme vybrali náhodne, aj tieto odhady predstavujú náhodné premenné. Preto nezodpovedanou otázkou ostáva, či tento rozdiel v bodových odhadoch stredných hodnôt výšok je natoľko odlišný, aby sme mohli s dostatočnou istotou tvrdiť, že neveríme, že aj skutočné stredné

hodnoty výšky mužov a žien na univerzite sú rovnaké, teda, že je rozdiel v stredných hodnotách štatisticky významne odlišný od nuly.

```
> library(MASS)
> attach(survey)
> vyska_M <- na.omit(subset(Height, subset = Sex == "Male"));
  length(vyska_M)
[1] 106
> vyska_Z <- na.omit(subset(Height, subset = Sex == "Female"));
  length(vyska_Z)
[1] 102
> mean(vyska_M)
[1] 178.8260
> mean(vyska_Z)
[1] 165.6867
```

Skôr ako použijeme funkciu `t.test()` pre dve stredné hodnoty musíme zistiť, či môžeme populačné rozptyly považovať za rovnaké alebo nie. Pre lepšiu prehľadnosť vytvoríme najprv histogram.

```
> par(mfrow = c(1, 2))
> hist(vyska_M, density = 10, col = "black", main = NA, cex.lab
  = 1.5, cex.axis = 1.3, freq = FALSE, ylab = "Hustota", xlab =
  "Výška mužov")
> hist(vyska_Z, density = 10, col = "black", main = NA, cex.lab
  = 1.5, cex.axis = 1.3, freq = FALSE, ylab = "Hustota", xlab =
  "Výška žien")
```



Obrázok 7.3: Histogram výšky mužov a žien

Zdroj: výstup zo softvéru R

Na prvý pohľad vidíme, že rozptyl vo výške mužov je väčší ako vo výške žien. Aby však naše rozhodnutie bolo korektné, realizujeme test na zhodu dvoch populačných rozptylov.

```
> var(vyska_M)
[1] 70.22862
> var(vyska_Z)
[1] 37.84436
> var.test(vyska_M, vyska_Z, ratio = 1, alternative =
  "two.sided", conf.level = 0.95)

                F test to compare two variances

data:  vyska_M and vyska_Z
F = 1.8557, num df = 105, denom df = 101, p-value = 0.001951
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
1.257430 2.734627
sample estimates:
ratio of variances
1.855722
```

Nulovú hypotézu o rovnosti populačných rozptylov (pomer rozptylov je rovný 1) môžeme zamietnuť a prijímame alternatívnu hypotézu (pomer rozptylov nie je rovný 1). Vo funkcii `t.test()` teda zvolíme možnosť `var.equal = F`.

```
> t.test(vyska_M, vyska_Z, alternative = "two.sided", mu = 0,
  var.equal = F, conf.level = 0.95)
```

```
                    Welch Two Sample t-test

data:  vyska_M and vyska_Z
t = 12.9243, df = 192.703, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
11.13420 15.14454
sample estimates:
mean of x mean of y
178.8260  165.6867
```

Na hladine významnosti 1 % môžeme zamietnuť nulovú hypotézu H_0: $\mu_x - \mu_y = 0$ a prijať alternatívnu H_1: $\mu_x - \mu_y \neq 0$. Z uvedeného vyplýva, že nemôžeme považovať strednú hodnotu výšky mužov a žien na danej univerzite za rovnakú (keďže vzorka pozostávala zo študentov jednej univerzity, všeobecné tvrdenie o populačnom priemere sa týka len tejto jednej univerzity). Na tomto mieste sa ešte vrátime k použitiu *F*-testu. Tento test na zhodu dvoch rozptylov predpokladá, že hodnoty v oboch vzorkách pochádzajú z normálneho rozdelenia pravdepodobnosti. Normalitu sme v predošlom prípade predpokladali. Na druhej strane, ak nemáme istotu, môžeme vykonať *t*-test pre oba prípady: a) ak uvažujeme o zhode rozptylov a b) ak uvažujeme o nezhode rozptylov. Kvalitatívne sme nenamerali žiadny rozdiel vo výsledkoch.

```
> t.test(vyska_M, vyska_Z, alternative = "two.sided", mu = 0,
  var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  vyska_M and vyska_Z
t = 12.8497, df = 206, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
11.12338 15.15536
sample estimates:
mean of x mean of y
178.8260  165.6867
```

Príklad 7.16

Postupovať budeme rovnako ako v predchádzajúcom príklade. Vzorku rozdelíme podľa príslušnosti hráčov k divízii na `East` (východná divízia) a `West` (západná divízia). Následne overíme, či je možné považovať populačné rozptyly za rovnaké pomocou funkcie `var.test()`. Ak áno, nastavíme argument funkcie `t.test()` na `var.equal = T`. Ak nie, tak nastavíme argument `var.equal = F`. Keďže riešime otázku, či v jednej divízii sú

stredné hodnoty miezd vyššie ako v druhej, tak ďalší argument vo funkcii `t.test()` nastavíme na `alternative = "greater"`. Testujeme teda nulovú hypotézu H_0: $\mu_x - \mu_y \leq 0$ oproti alternatívnej H_1: $\mu_x - \mu_y > 0$.

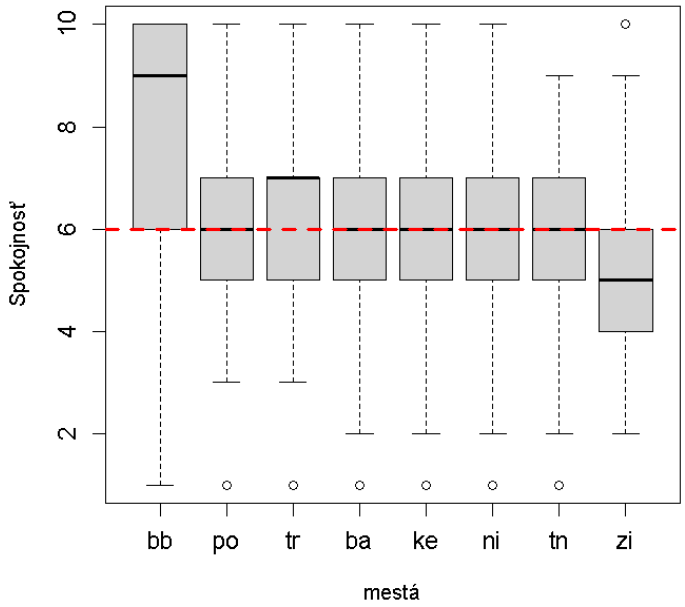
```
> library(vcd)
> attach(Baseball)
-----------------------------------------------------------------
> East <- na.omit(subset(sal87, subset = div86 == "E"));
  length(East)
[1] 129
> West <- na.omit(subset(sal87, subset = div86 == "W"));
  length(West)
[1] 134
> var(East)
[1] 278812.3
> var(West)
[1] 117707.1
-----------------------------------------------------------------
> var.test(East, West, ratio = 1, alternative = "two.sided",
  conf.level = 0.95)

                F test to compare two variances

data:  East and West
F = 2.3687, num df = 128, denom df = 133, p-value = 1.197e-06
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
1.679014 3.345529
sample estimates:
ratio of variances
2.368696
-----------------------------------------------------------------
> t.test(East, West, alternative = "greater", mu = 0, var.equal
  = F, conf.level = 0.95)

                    Welch Two Sample t-test

data:  East and West
t = 3.145, df = 218.458, p-value = 0.000946
alternative hypothesis: true difference in means is greater than
  0
95 percent confidence interval:
82.3211     Inf
sample estimates:
mean of x mean of y
624.2714  450.8769
```

Nulovú hypotézu môžeme zamietnuť na 1 % hladine významnosti, a teda mzdy hráčov z východnej divízie môžeme považovať za v priemere vyššie ako mzdy hráčov zo západnej divízie (resp. boli vyššie v roku 1987).

Príklad 7.17

Na odhad zhody variability v tomto príklade nepoužijeme najprv žiadny test, ale skúsime sa spoľahnúť na vizualizáciu údajov. Porovnávame box – ploty pre každé z uvedených miest. Následne si náš názor overíme pomocou Brown – Forsythovho testu na zhodu variability v dvoch súboroch.

```
> data <- data.frame(satisfc, as.factor(city))
> names(data) <- c("satisfaction", "city")
> bymeans <- with(data, reorder(data$city, -data$satisfaction,
  mean))
> boxplot(data$satisfaction ~ bymeans, cex.axis = 1.2, col =
  "lightgray", xlab = "mestá", ylab = "Spokojnosť")
> abline(h = mean(data$satisfaction), lwd = 3, lty = 2, col =
  "red")
```



Obrázok 7.4: Box – ploty spokojnosti podľa miest

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> library(car)
> sat_1 <- subset(data, subset = city == c("ba", "ke"))
> sat_2 <- subset(data, subset = city == c("bb", "ke"))
> sat_3 <- subset(data, subset = city == c("po", "ke"))
> sat_4 <- subset(data, subset = city == c("zi", "tr"))
> sat_5 <- subset(data, subset = city == c("zi", "tn"))
-------------------------------------------------------------
> leveneTest(sat_1$satisfaction, sat_1$city, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.2507 0.6202
      31
> leveneTest(sat_2$satisfaction, sat_2$city, center = median)
Levene's Test for Homogeneity of Variance (center = median)
```

```
      Df F value Pr(>F)
group  1  0.5195 0.4766
      30
> leveneTest(sat_3$satisfaction, sat_3$city, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0649 0.8006
      30
> leveneTest(sat_4$satisfaction, sat_4$city, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0305 0.8624
      31
> leveneTest(sat_5$satisfaction, sat_5$city, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.4545  0.505
      32
```

Z box – plotov sa javí rozptyl odlišný iba v prípade krajského mesta Banská Bystrica. Vykonali sme sériu štatistických testov, kde sme porovnali zhodu vo variabilite jednotlivých dvojíc súborov, ktoré budeme v ďalšej analýze porovnávať. Ani v jednom prípade sme nevedeli zamietnuť hypotézu o zhode variability. Preto v testoch porovnávajúcich stredné hodnoty budeme všade uvažovať o alternatíve s rovnakými rozptylmi.

```
> t.test(sat_1$satisfaction ~ sat_1$city, alternative =
  "two.sided", mu = 0, var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  sat_1$satisfaction by sat_1$city
t = -0.0091, df = 31, p-value = 0.9928
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-1.658724  1.644018
sample estimates:
mean in group ba mean in group ke
        6.117647         6.125000
------------------------------------------------------------
> t.test(sat_2$satisfaction ~ sat_2$city, alternative =
  "two.sided", mu = 0, var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  sat_2$satisfaction by sat_2$city
t = 1.6535, df = 30, p-value = 0.1087
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-0.3526424  3.3526424
sample estimates:
```

```
mean in group bb mean in group ke
           7.625            6.125
-----------------------------------------------------------------
> t.test(sat_3$satisfaction ~ sat_3$city, alternative =
  "two.sided", mu = 0, var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  sat_3$satisfaction by sat_3$city
t = 0.8372, df = 30, p-value = 0.4091
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-0.8995966  2.1495966
sample estimates:
mean in group ke mean in group po
           6.125            5.500
-----------------------------------------------------------------
> t.test(sat_4$satisfaction ~ sat_4$city, alternative =
  "two.sided", mu = 0, var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  sat_4$satisfaction by sat_4$city
t = 1.7945, df = 31, p-value = 0.08249
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-0.1822115  2.8513292
sample estimates:
mean in group tr mean in group zi
        6.687500         5.352941
-----------------------------------------------------------------
> t.test(sat_5$satisfaction ~ sat_5$city, alternative =
  "two.sided", mu = 0, var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  sat_5$satisfaction by sat_5$city
t = 0.5198, df = 32, p-value = 0.6068
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-1.030184  1.736066
sample estimates:
mean in group tn mean in group zi
        5.705882         5.352941
```

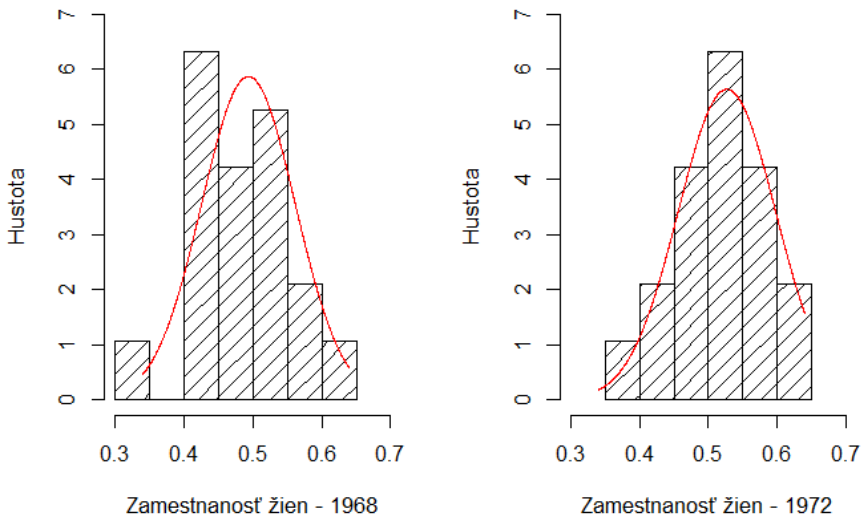
Prekvapujúco sme danej hladine významnosti nevedeli ani v jednom prípade zamietnuť nulovú hypotézu o zhode stredných hodnôt. Zdá sa tak, že vyššie hodnoty namerané v Banskej Bystrici mohli byť spôsobené aj našim výberom respondentov do vzorky. Do istej miery sa zdá, že dôvod, prečo sme nevedeli zamietnuť nulovú hypotézu pri testoch stredných hodnôt (kde vystupovali respondenti z Banskej Bystrice) je, že naše vzorky boli

málo početné. Viac početné vzorky totiž (za inak nezmenených podmienok) zvyšujú silu testov – teda pravdepodobnosť, že správne zamietneme nulovú hypotézu.

Príklad 7.18

Na overovanie zhody rozptylov použijeme testy zhody dvoch rozptylov. Najprv overíme predpoklad normality, ktorý je potrebný pre použitie *F*-testu. Zatiaľ však nepoužijeme formálne testy, ale vystačíme si s histogramom. Pri riešení otázky, či došlo k zmene zamestnanosti žien v USA v rokoch 1968 a 1972, využijeme párový *t*-test. Je totiž úplne zrejmé, že ide „párové dáta“, keďže merania sa týkajú tých istých miest len v rôznych rokoch. Najprv však potrebujeme zistiť, či je možné považovať populačné rozptyly v týchto dvoch vzorkách za rovnaké alebo nie.

```
> X1968 <- c(0.42, 0.50, 0.52, 0.45, 0.43, 0.55, 0.45, 0.34,
  0.45, 0.54, 0.42, 0.51, 0.49, 0.54, 0.50, 0.58, 0.49, 0.56,
  0.63)
> X1972 <- c(0.45, 0.50, 0.52, 0.45, 0.46, 0.55, 0.60, 0.49,
  0.35, 0.55, 0.52, 0.53, 0.57, 0.53, 0.59, 0.64, 0.50, 0.57,
  0.64)
> year <- as.factor(c(rep(1968, length(X1968)), rep(1972,
  length(X1972))))
> un_usa <- data.frame(c(X1968, X1972), year)
> names(un_usa) <- c("zamestnanost","rok"); attach(un_usa)
> x <- seq(min(zamestnanost), max(zamestnanost), length = 1000)
> xh <- dnorm(x, mean = mean(X1968), sd = sd(X1968))
> xhh <- dnorm(x, mean = mean(X1972), sd = sd(X1972))
> n_1968 <- data.frame(x, xh)
> n_1972 <- data.frame(x, xhh)
-----------------------------------------------------------------
> par(mfrow = c(1, 2))
> hist(X1968, density = 10, col = "black", main = NA, xlim =
  c(0.3, 0.7), ylim = c(0, 7), cex.lab = 1.1, cex.axis = 1.0,
  freq = FALSE, ylab = "Hustota", xlab = "Zamestnanosť žien -
  1968")
> lines(n_1968, type = "l", col = "red")
> hist(X1972, density = 10, col = "black", main = NA, xlim =
  c(0.3, 0.7), ylim = c(0, 7), cex.lab = 1.1, cex.axis = 1.0,
  freq = FALSE, ylab = "Hustota", xlab = "Zamestnanosť žien -
  1972")
> lines(n_1972, type = "l", col = "red")
```



Obrázok 7.5: Histogram zamestnanosti

Zdroj: vlastné spracovanie, výstup zo softvéru R

Zdá sa, že kým v prípade zamestnanosti žien z roku 1972 by sme ešte mohli uvažovať o normálnom rozdelení, pri údajoch z roku 1968 sa predpoklad normality zdá byť menej presvedčivý. Z tohto dôvodu uskutočníme Levenov test (za charakteristiku miery polohy sa vybral aritmetický priemer) a dve varianty Brown – Forsythovho testu (za charakteristiku miery polohy sa vybral medián a upravený priemer[31]). Ani v jednom z testov sme neboli schopní zamietnuť nulovú hypotézu o rovnosti rozptylov. Zdá sa, že odlišnosti medzi štátmi sa v priebehu štyroch rokov výrazne nezmenili.

```
> leveneTest(zamestnanost, rok, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1   6e-04 0.9813
      36
-----------------------------------------------------------------
> leveneTest(zamestnanost, rok, center = mean, trim = 0.1)
Levene's Test for Homogeneity of Variance (center = mean: 0.1)
      Df F value Pr(>F)
group  1   5e-04 0.9827
      36
-----------------------------------------------------------------
> leveneTest(zamestnanost, rok, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0013 0.9709
      36
```

[31] Parameter „trim“ označuje, koľko percent najmenších a najväčších hodnôt sa pred výpočtom aritmetického priemeru zo vzorky odstráni.

Pri riešení otázky, či došlo k zmene zamestnanosti žien v USA v rokoch 1968 a 1972, využijeme párový *t*-test. Je totiž zrejmé, že ide „párové dáta“, keďže merania sa týkajú tých istých miest len v rôznych rokoch.

```
> t.test(X1968, X1972, alternative = "two.sided", mu = 0,
  var.equal = T, conf.level = 0.95, paired = TRUE)

                        Paired t-test

data:  X1968 and X1972
t = -2.4577, df = 18, p-value = 0.02435
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-0.062478527 -0.004889895
sample estimates:
mean of the differences
-0.03368421
```

Z výsledkov môžeme vidieť, že nulovú hypotézu môžeme zamietnuť na 5 % hladine významnosti a vieme prijať alternatívnu hypotézu, teda že rozdiel v priemeroch nie je rovný nule. Inými slovami, na základe týchto výsledkov zrejme došlo k zmene v zamestnanosti žien v USA. Ak by sme tieto výsledky porovnali s *t*-testom, kde by sa chybne predpokladala nezávislosť vzoriek, došli by sme ku kvalitatívne odlišnému záveru.

```
> t.test(X1968, X1972, var.equal = T, alternative = "two.sided",
  mu = 0, conf.level = 0.95, paired = F)

                      Two Sample t-test

data:  X1968 and X1972
t = -1.4959, df = 36, p-value = 0.1434
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-0.07935368  0.01198526
sample estimates:
mean of x mean of y
0.4931579 0.5268421
```

Príklad 7.20

V tomto príklade nás zaujíma, či na základe údajov zo vzorky vieme rozhodnúť, ktorú farbu deti preferujú. Konkrétne nás zaujíma, či môžeme považovať výsledky experimentu (v ktorom 21 detí z 30 zvolilo modrú farbu hračky) za dostatočne presvedčujúce a či môžeme prijať tvrdenie, že deti preferujú na hračkách modrú farbu. Jednou z možností ako toto overiť, je testovať hypotézu H_0: $p \leq 0.5$ oproti alternatívnej H_1: $p > 0.5$.

```
> binom.test(21, 30, p = 0.5, alternative = "greater",
  conf.level = 0.99)

                    Exact binomial test

data:  21 and 30
number of successes = 21, number of trials = 30,
p-value = 0.02139
alternative hypothesis: true probability of success is greater
  than 0.5
99 percent confidence interval:
0.4725554 1.0000000
sample estimates:
probability of success
0.7
```

Na hladine významnosti 1 % nevieme zamietnuť nulovú hypotézu, a teda nemôžeme prijať žiadne tvrdenie o jednoznačnosti preferencií detí o farbách použitých na hračkách. Ak by sme však testovali na hladine významnosti 5 %, tak by sme mohli prijať alternatívnu hypotézu o tom, že podiel je väčší ako 0.5. Výrobcovi hračiek by sme potom mohli odporučiť vyrábať hračky v modrej farbe.

Príklad 7.21

V tomto príklade máme za úlohu zistiť, či developerská spoločnosť má dôvod sa obávať, že by prípadná petičná akcia proti ich projektu mohla skončiť úspešne (75 % ľudí musí vyjadriť nesúhlas). Zaujíma nás jednostranná hypotéza H_0: $p \geq 0.75$ oproti alternatívnej H_1: $p < 0.75$. Ak by sme boli schopní prijať alternatívnu hypotézu, znamenalo by to pomerne silný signál (nikdy nie istotu) k tomu, že petičná akcia nebude mať úspech. Pri výpočte v programe R postupujeme podobne ako v predošlých príkladoch.

```
> binom.test(132, 200, p = 0.75, alternative = "less",
  conf.level = 0.99)

                    Exact binomial test

data:  132 and 200
number of successes = 132, number of trials = 200,
p-value = 0.00275
alternative hypothesis: true probability of success is less than
  0.75
99 percent confidence interval:
0.0000000 0.7365346
sample estimates:
probability of success
0.66
```

Pre investičnú spoločnosť sú výsledky pomerne povzbudivé. Nulovú hypotézu sme na hladine významnosti $\alpha = 0.01$ zamietli. Tento záver môžeme odčítať jednak z konfidenčného intervalu, ktorý je v celom rozsahu menší ako 0.75 ako aj z *p*-hodnoty, ktorá je na úrovni 0.00275, teda menej ako nominálnych 0.01.

Príklad 7.22

V tomto príklade máme za úlohu zistiť, či sú rozdiely medzi mužmi a ženami vo vnímaní ich finančnej situácie, konkrétne či sa zhoršila v roku 1992 oproti predchádzajúcemu roku. Testujeme teda nulovú hypotézu o rovnosti podielov vnímania finančnej situácie medzi mužmi a ženami H_0: $p_M - p_Z = 0$ oproti alternatívnej hypotéze H_1: $p_M - p_Z \neq 0$.

```
> data <- read.csv(file = "...cesta
  k súboru...\\montana_survey.csv", sep = ";", dec = ".", header
  = T)
-----------------------------------------------------------------
> M <- na.omit(subset(data$FIN, subset = data$SEX == "0"));
  length(M)
[1] 106
> Z <- na.omit(subset(data$FIN, subset = data$SEX == "1"));
  length(Z)
[1] 102
-----------------------------------------------------------------
> FIN_1_M <- sum(na.omit(M == "1")); FIN_1_M
[1] 31
> FIN_1_Z <- sum(na.omit(Z == "1")); FIN_1_Z
[1] 30
-----------------------------------------------------------------
> prop.test(x = c(FIN_1_M, FIN_1_Z), n = c(length(M),
  length(Z)), alternative = "two.sided", conf.level = 0.95,
  correct = FALSE)

  2-sample test for equality of proportions without continuity
                        correction

data:  c(FIN_1_M, FIN_1_Z) out of c(length(M), length(Z))
X-squared = 7e-04, df = 1, p-value = 0.979
alternative hypothesis: two.sided
95 percent confidence interval:
-0.1254305  0.1221008
sample estimates:
prop 1    prop 2
0.2924528 0.2941176
```

Zo 106 mužov vo vzorke 31 vníma zhoršenie ich finančnej situácie a zo 102 žien vníma zhoršenie 30 žien. Pri *p*-hodnote 0.979 nevieme zamietnuť nulovú hypotézu, a teda môžeme považovať podiel mužov a žien, ktorí vnímajú zhoršenie ich finančnej situácie za rovnaký (istí si nemôžeme byť, avšak výsledky naznačujú skôr túto situáciu).

Obdobne postupujeme aj pri ďalšej úlohe, v ktorej máme zistiť, či celkové zlepšenie ekonomickej situácie v krajine vnímajú muži a ženy rozdielne.

```
> M <- na.omit(subset(STAT, subset = SEX == "0")); length(M)
[1] 107
> Z <- na.omit(subset(STAT, subset = SEX == "1")); length(Z)
[1] 102
-----------------------------------------------------------------
> STAT_M <- sum(na.omit(M == "1")); STAT_M
[1] 41
> STAT_Z <- sum(na.omit(Z == "1")); STAT_Z
[1] 22
-----------------------------------------------------------------
> prop.test(x = c(STAT_M, STAT_Z), n = c(length(M), length(Z)),
  alternative = "two.sided", conf.level = 0.95, correct = FALSE)

  2-sample test for equality of proportions without continuity
                          correction

data:  c(STAT_M, STAT_Z) out of c(length(M), length(Z))
X-squared = 6.957, df = 1, p-value = 0.00835
alternative hypothesis: two.sided
95 percent confidence interval:
0.04560439 0.28937820
sample estimates:
prop 1    prop 2
0.3831776 0.2156863
```

Z týchto výsledkov vyplýva, že zo 107 mužov vníma zlepšenie celkovej situácie v krajine 41 mužov a zo 102 žien vníma zlepšenie 22 žien. Na základe testu o zhode dvoch podielov môžeme zamietnuť nulovú hypotézu o ich rovnosti s *p*-hodnotou 0.00835. Môžeme sa tak prikloniť k záveru, že muži a ženy vnímajú rozdielne zlepšenie ekonomickej situácie v krajine.

Príklad 7.23

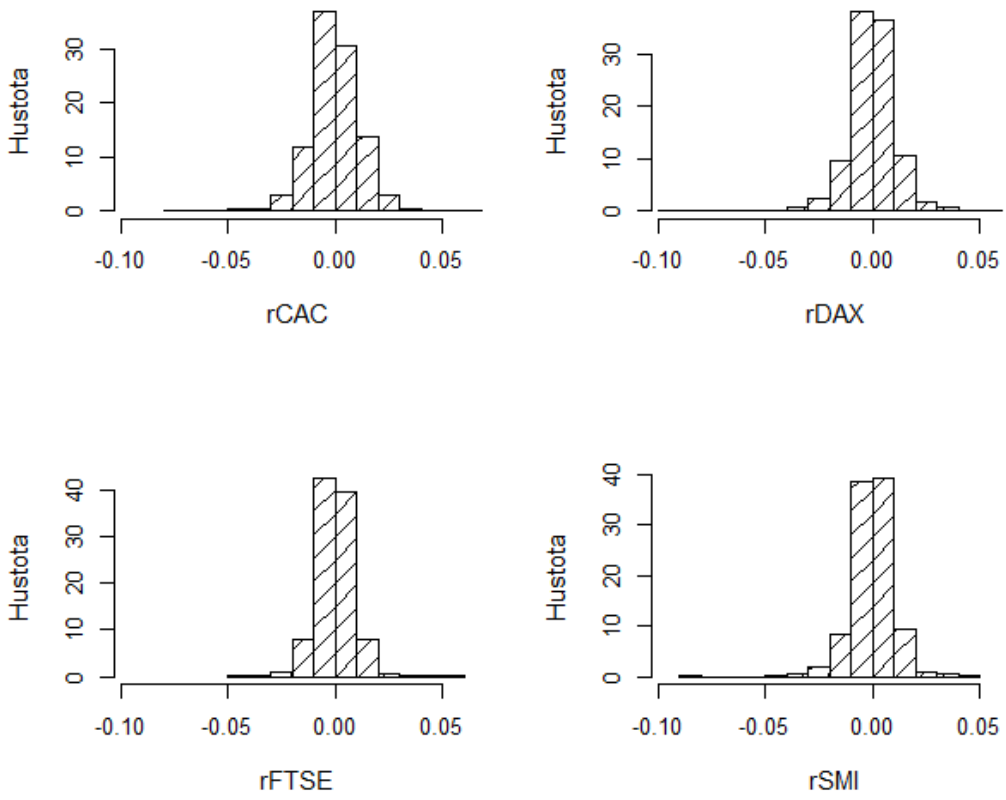
Po transformácii uzatváracích cien si pre lepšiu predstavu o rozdelení nadmerných výnosov zobrazíme histogramy (pozri Obrázok 7.6). Môžeme očakávať, že pôjde o rozdelenia s tzv. tučnými koncami (z angl. *fat tails*), čiže s extrémnou šikmosťou alebo špicatosťou a častejším výskytom extrémnych hodnôt. Kvôli lepšej porovnateľnosti histogramov sme rozsah osi *x*-ovej a osi *y*-ovej prispôsobili tak, aby bolo možné tieto histogramy priamo porovnať.

```
> library(datasets)
> attach(data.frame(EuStockMarkets))
-----------------------------------------------------------------
> rCAC <- ar.ols(diff(log(CAC)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
```

```
> rDAX <- ar.ols(diff(log(DAX)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rFTSE <- ar.ols(diff(log(FTSE)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
> rSMI <- ar.ols(diff(log(SMI)), aic = FALSE, order.max = 1,
  demean = TRUE)$resid
-----------------------------------------------------------------
> low <- min(c(rCAC, rDAX, rFTSE, rSMI), na.rm = TRUE); high <-
  max(c(rCAC, rDAX, rFTSE, rSMI), na.rm = TRUE);
> par(mfrow = c(2, 2))
> hist(rCAC, density = 10, col = "black", main = NA, cex.lab =
  1.1, cex.axis = 1.0, freq = FALSE, ylab = "Hustota", xlab =
  "rCAC", xlim = c(low, high))
> hist(rDAX, density = 10, col = "black", main = NA, cex.lab =
  1.1, cex.axis = 1.0, freq = FALSE, ylab = "Hustota", xlab =
  "rDAX", xlim = c(low, high))
> hist(rFTSE, density = 10, col = "black", main = NA, cex.lab =
  1.1, cex.axis = 1.0, freq = FALSE, ylab = "Hustota", xlab =
  "rFTSE", xlim = c(low, high))
> hist(rSMI, density = 10, col = "black", main = NA, cex.lab =
  1.1, cex.axis = 1.0, freq = FALSE, ylab = "Hustota", xlab =
  "rSMI", xlim = c(low, high))
```

Vizuálne sa zdajú byť rozdelenia dosť podobné, zrejme niektoré dvojice nadmerných výnosov bude možné považovať za realizácie z rovnakého rozdelenia.

Za účelom formálneho overenia, či môžeme niektoré z nadmerných výnosov považovať za realizácie z toho istého rozdelenia pravdepodobnosti, využijeme Kolmogorov – Smirnovov test (funkcia `ks.test()`). Tento test však nepredpokladá rovnosť hodnôt, čiže žiadne dve hodnoty vo výberovom súbore by nemali byť rovnaké. Ak je táto podmienka porušená, kritické hodnoty (ktorú sú uvádzané v tabuľkách) nie sú presné. Funkcia `ks.test()` nás upozorní na možný problém s rovnakými hodnotami chybovým hlásením. Alternatívou je použiť funkciu `ks.boot()` z knižnice `Matching`, kde sa počíta upravená verzia Kolmogorov – Smirnovovho testu, v ktorej sa pre potreby počítania kritických hodnôt (a *p*-hodnoty) využíva bootstraping. Upozorňujeme, že pri použití funkcie `ks.boot()` môže v závislosti od počtu bootstrap vzoriek (parameter `nboots`) analýza trvať pomerne dlho. Na druhej strane väčší počet týchto vzoriek zaručuje väčšiu stabilitu výsledkov. Rozhodli sme sa použiť 10000 bootstrap vzoriek.



Obrázok 7.6: Histogramy nadmerných výnosov akciových indexov

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> library(Matching)
> ks.test(rCAC, rDAX)

                Two-sample Kolmogorov-Smirnov test

data:  rCAC and rDAX
D = 0.0452, p-value = 0.04485
alternative hypothesis: two-sided

Warning message:
In ks.test(rCAC, rDAX) : cannot compute correct p-values with
  ties
------------------------------------------------------------------
> ks.boot(rCAC, rDAX, nboots = 10000)
$ks.boot.pvalue
[1] 0.0431

$ks

                Two-sample Kolmogorov-Smirnov test

data:  Tr and Co
D = 0.0452, p-value = 0.04485
alternative hypothesis: two.sided

$nboots
```

```
[1] 10000
attr(,"class")
[1] "ks.boot"
```

V prípade indexov CAC a DAX môžeme zamietnuť nulovú hypotézu o rovnosti distribučných funkcií na hladine významnosti 5 %. Postup zopakujeme pre všetky dvojice akciových indexov[32], ale uvádzať budeme už len výsledky z funkcie `ks.boot()`.

```
> data <- data.frame(rCAC, rDAX, rSMI, rFTSE)
> results <- matrix(ncol = 3, nrow = 6)
> colnames(results) <- c("D statistics", "KS p-val.", "Bootstrap
  p-val.")
> rownames(results) <- c("CAC - DAX", "CAC - SMI", "CAC - FTSE",
  "DAX - SMI","DAX - FTSE", "SMI - FTSE")
> for (i in 1:4) {
+ for (j in min((i+1), 4):4) {
+ a <-ks.boot(data[i], data[j], nboots = 10000)
+ results[i+j-2, 1] <- round(a$ks$statistic, 4);
+ results[i+j-2, 2] <- round(a$ks$p.value, 4);
+ results[i+j-2, 3] <- round(a$ks.boot.pvalue, 4);
+ }
+ }
-----------------------------------------------------------------
> results
          D statistics KS p-val. Bootstrap p-val.
CAC - DAX        0.0452    0.0448           0.0415
CAC - SMI        0.0705    0.0002           0.0003
CAC - FTSE       0.0323    0.2873           0.2859
DAX - SMI        0.0463    0.0373           0.0362
DAX - FTSE       0.0285    0.4363           0.4360
SMI - FTSE       0.0000    1.0000           1.0000
```

Na hladine významnosti 5 % nevieme zamietnuť nulovú hypotézu v prípade dvojíc indexov CAC – FTSE, DAX – FTSE a FTSE – SMI.

Príklad 7.24

Na testovanie normality využijeme v tomto príklade tri testy, a to Anderson – Darlingov (funkcia `ad.test()` z knižnice `nortest`), Shapiro – Wilkov (funkcia `shapiro.test()` z knižnice `stats`) a Jarque – Berov test (funkcia `rjb.test()` z knižnice `lawstat`). Pri Jarque – Berovom teste budeme vychádzať z empirických (simulovaných) kritických hodnôt a použijeme obe jeho verzie, t. j. klasickú aj modifikovanú. Zostrojili sme pritom jednoduchú funkciu, ktorej argumentom je databáza údajov (stĺpce

[32] Je zrejmé, že ak zopakujeme test aj pre dvojicu indexov CAC a DAX, tak výsledok môže byť mierne odlišný.

predstavujú premenné) a počet bootstrap vzoriek pre Jarque-Bera testy. Výstupom je matica *p*-hodnôt z jednotlivých štatistických testov normality.

```
> library(nortest)
> library(stats)
> library(lawstat)
-----------------------------------------------------------------
> excess_returns <- data.frame(rCAC, rDAX, rFTSE, rSMI)
> norm_check <- function(dataframe, B) {
+ normality_tests <- function(data, B) {
+ temp <- c()
+ temp[1] <- ad.test(data)$p.value
+ temp[2] <- shapiro.test(data)$p.value
+ temp[3] <- rjb.test(na.omit(data), option = c("JB"),
  crit.values = c("empirical"), N = B)$p.value
+ temp[4] <- rjb.test(na.omit(data), option = c("RJB"),
  crit.values = c("empirical"), N = B)$p.value
+ return(temp)
+ }
+ results <- apply(dataframe, 2, normality_tests, B)
+ rownames(results) <- c("Anderson-Darling", "Shapiro-Wilk",
  "Jarque-Bera", "Robust Jarque-Bera")
+ colnames(results) <- colnames(dataframe)
+ return(results)
+ }
-----------------------------------------------------------------
> norm_check(excess_returns, B = 1000)
                         rCAC         rDAX         rFTSE         rSMI
Anderson-Darling   1.817153e-12 1.530857e-31 2.355213e-09 6.160279e-29
Shapiro-Wilk       1.518707e-14 8.411550e-24 1.302503e-14 4.297013e-23
Jarque-Bera        0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
Robust Jarque-Bera 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
```

Z výsledkov vyplýva, že v každom jednom prípade sme mohli zamietnuť nulovú hypotézu o normálnom rozdelení. Nadmerné výnosy akciových indexov teda zrejme nemôžeme považovať za realizácie z normálneho rozdelenia pravdepodobnosti.

Príklad 7.25

Na identifikáciu odľahlých hodnôt v reálnej mzde využijeme najprv Grubssov test dostupný cez funkciu `grubbs.test()` v knižnici `outliers`[33]. Nevýhodou tohto testu je, že v nulovej hypotéze testuje prítomnosť vždy len jednej odľahlej hodnoty. Preto po identifikovaní prvého outliera je nutné túto hodnotu odstrániť a test zopakovať[34]. Jeho ďalšou

[33] Dixonov test pre nás v tomto príklade nie je vhodný, keďže máme až 79 pozorovaní a nemáme vhodnú softvérovú implementáciu pre početnejšie vzorky.

[34] Vybrané hodnoty odstraňujeme na základe ich pozície v dátovom vektore. Ďalšou možnosťou by bolo odstrániť konkrétne hodnoty (napr. hodnotu 1182.604046 identifikovanú ako prvý outlier) cez príkaz `real_mzda[!real_mzda == 1182.604046]`. Pri tomto postupe si ale musíme skontrolovať, na koľko desatinných miest sú údaje zaokrúhlené – napr. pri Hampelovom teste (uvedenom nižšie) dostávame

nevýhodou je, že po odstránení jednej extrémnej hodnoty sa štatistické vlastnosti (napr. priemer, rozptyl) súboru zmenia a tým sa ovplyvňujú testovacie kritériá na vyhodnotenie odľahlých hodnôt.

```
> data <- read.csv(file = "...cesta k súboru...", sep = ";", dec
  = ".", header = T)
> attach(data)
-----------------------------------------------------------------
> library(outliers)
> grubbs.test(real_mzda, opposite = FALSE)

                Grubbs test for one outlier

data:  real_mzda
G = 3.8696, U = 0.8056, p-value = 0.001879
alternative hypothesis: highest value 1182.604046 is an outlier
-----------------------------------------------------------------
> real_mzda2 <- real_mzda[-2]
> grubbs.test(real_mzda2, opposite = FALSE)

                Grubbs test for one outlier

data:  real_mzda2
G = 3.9678, U = 0.7929, p-value = 0.001106
alternative hypothesis: highest value 1136.262365 is an outlier
-----------------------------------------------------------------
> real_mzda3 <- real_mzda2[-1]
> grubbs.test(real_mzda3, opposite = FALSE)

                Grubbs test for one outlier

data:  real_mzda3
G = 3.6728, U = 0.8202, p-value = 0.004673
alternative hypothesis: highest value 1045.361376 is an outlier
-----------------------------------------------------------------
> real_mzda4 <- real_mzda3[-1]
> grubbs.test(real_mzda4, opposite = FALSE)

                Grubbs test for one outlier

data:  real_mzda4
G = 3.2089, U = 0.8609, p-value = 0.03439
alternative hypothesis: highest value 956.2427591 is an outlier
-----------------------------------------------------------------
> real_mzda5 <- real_mzda4[-2]
```

---

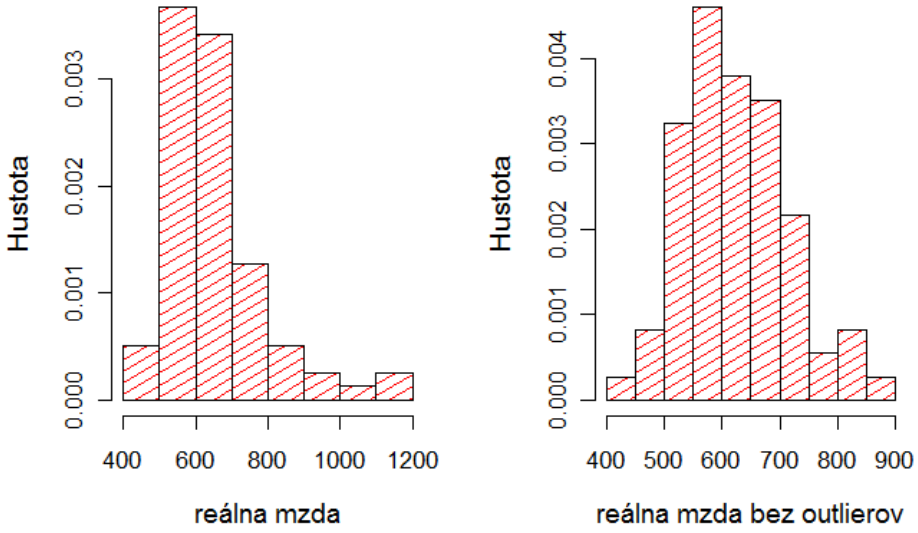
výsledky zaokrúhlené len na 4 desatinné miesta. V tom prípade dostávame odľahlú hodnotu 1182.6040 a po použití príkazu `real_mzda[!real_mzda == 1182.6040]` by tam daná hodnota ostala, keďže softvér R by túto hodnotu nerozpoznal. Riešením by bolo zaokrúhliť najprv dátový vektor na 4 desatinné miesta pomocou príkazu `round(real_mzda, 4)`.

```
> grubbs.test(real_mzda5, opposite = FALSE)

                Grubbs test for one outlier

data:  real_mzda5
G = 3.3691, U = 0.8445, p-value = 0.01743
alternative hypothesis: highest value 945.5485251 is an outlier
------------------------------------------------------------------
> real_mzda6 <- real_mzda5[-1]
> grubbs.test(real_mzda6, opposite = FALSE)

                Grubbs test for one outlier

data:  real_mzda6
G = 2.9229, U = 0.8814, p-value = 0.09825
alternative hypothesis: highest value 877.8183763 is an outlier
```

Môžeme vidieť, že celý tento postup je pomerne zbytočne rozsiahli. Vhodnejšie by bolo zrejme vytvoriť funkciu, ktorá tieto iterácie „nájsť extrémnu hodnotu“ -> „odstrániť extrémnu hodnotou“ vykoná za nás. Zostrojenie tejto funkcie necháme na čitateľa.

Na hladine významnosti 5 % sme postupne identifikovali 5 extrémnych hodnôt. Konkrétne ide o okresy Bratislava I až Bratislava V. Pre lepšiu prehľadnosť uvádzame aj histogram rozdelenia reálnej mzdy v SR pred a po odstránení extrémnych hodnôt.

```
> par(mfrow = c(1, 2))
> hist(real_mzda, density = 10, col = "red", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "reálna mzda", family = "serif")
> hist(real_mzda6, density = 10, col = "red", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "reálna mzda bez outlierov", family =
  "serif")
```



Obrázok 7.7: Histogram reálnej mzdy v okresoch SR

Zdroj: vlastné spracovanie, výstup zo softvéru R

Okrem vyššie spomínanej nevýhody tohto testu (identifikácia len jednej extrémnej hodnoty) má Grubbsov test ešte jednu nevýhodu, a to predpoklad normálneho rozdelenia údajov. Hypotézu o normálnom rozdelení však možno zamietneme práve z dôvodu výskytu extrémnych hodnôt. Z tohto dôvodu je celý postup testovania pomerne otázny a vhodné je tento test používať pri údajoch, kde z deduktívnych dôvodov predpokladáme normalitu (napríklad na základe povahy procesu, ktorý údaje generuje). Ak nemáme žiadne deduktívne dôvody predpokladať, či hodnoty sú alebo nie sú z normálneho rozdelenia, neostáva nám nič iné ako tento predpoklad overiť pomocou testovania normality. Na testovanie normality využijeme v tomto príklade Anderson – Darglingov test, Shapiro – Wilkov test a robustný Jarque – Berov test.

```
> library(nortest)
> library(stats)
> library(lawstat)
------------------------------------------------------------------
> ad.test(real_mzda)

               Anderson-Darling normality test

data:  real_mzda
A = 3.2456, p-value = 3.329e-08
------------------------------------------------------------------
> shapiro.test(real_mzda)

                 Shapiro-Wilk normality test

data:  real_mzda
W = 0.8414, p-value = 9.695e-08
```

```
------------------------------------------------------------------
> rjb.test(real_mzda, option = c("RJB"), crit.values =
  c("empirical"), N = 1000)

                    Robust Jarque Bera Test

data:  real_mzda
X-squared = 225.7613, df = 2, p-value < 2.2e-16
```

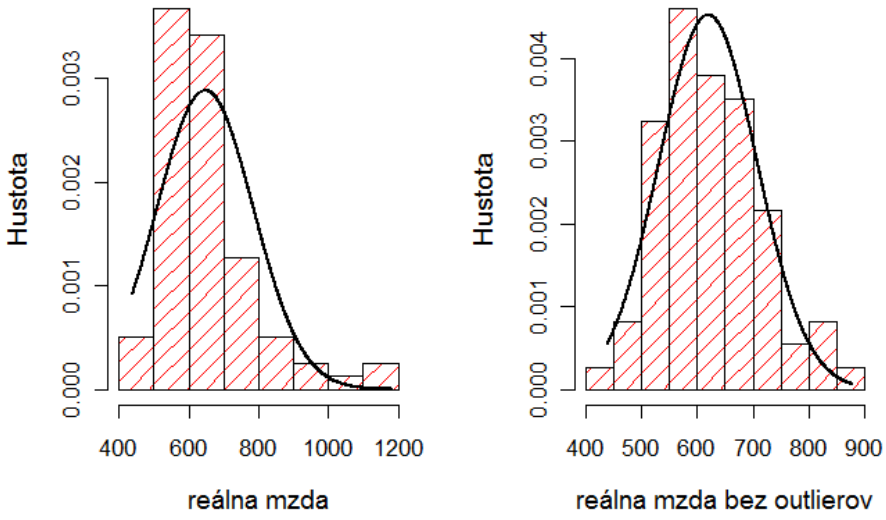
Na pôvodných dátach o reálnych mzdách v okresoch SR sme pri všetkých testoch zamietli nulovú hypotézu o normalite. Použitie Grubbsovho testu je tak spochybniteľné. Keď použijeme tieto testy normality na dátach očistených o extrémne hodnoty, nulovú hypotézu nevieme zamietnuť na hladine významnosti 5 % ani pri jednom teste. Je teda zrejmé, že reálne mzdy v okresoch SR nemôžeme považovať za realizácie z normálneho rozdelenia práve kvôli výskytu extrémnych hodnôt.

```
> ad.test(real_mzda6)

               Anderson-Darling normality test

data:  real_mzda6
A = 0.6344, p-value = 0.09453
------------------------------------------------------------------
> shapiro.test(real_mzda6)

                  Shapiro-Wilk normality test

data:  real_mzda6
W = 0.9727, p-value = 0.1098
------------------------------------------------------------------
> rjb.test(real_mzda6, option = c("RJB"), crit.values =
  c("empirical"), N = 1000)

                    Robust Jarque Bera Test

data:  real_mzda6
X-squared = 3.6029, df = 2, p-value = 0.1235
```

Ak sa pozrieme ešte raz na histogram hodnôt reálnych miezd v okresoch SR, ale pridáme do grafu aj funkcie hustoty rozdelenia pravdepodobnosti, tak je úplne zrejmé, že pôvodné dáta sú pravostranne zošikmené práve kvôli extrémnym hodnotám Bratislavského kraja.

```
> par(mfrow = c(1, 2))
> hist(real_mzda, density = 10, col = "red", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "reálna mzda")
> x <- seq(min(real_mzda), max(real_mzda), length = 1000)
> xh <- dnorm(x, mean = mean(real_mzda), sd = sd(real_mzda))
```

```
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
> hist(real_mzda6, density = 10, col = "red", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "reálna mzda bez outlierov")
> x <- seq(min(real_mzda6), max(real_mzda6), length = 1000)
> xh <- dnorm(x, mean = mean(real_mzda6), sd = sd(real_mzda6))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
```



Obrázok 7.8: Histogram reálnej mzdy v okresoch SR

Zdroj: vlastné spracovanie, výstup zo softvéru R

Keďže si pri reálnych mzdách nemôžeme byť istí predpokladom normality, môžeme použiť neparametrický Hampelov test.

```
> hampel_identifier <- function(data) {
+  ri <- abs(data - median(data))
+  mad <- median(ri)
+  madn <- mad/0.6745
+  hi <- ri/madn
+  critical <- sqrt(qchisq(0.975,1))
+  data[hi>critical]
+ }
> hampel_identifier(real_mzda)
[1] 1136.2624 1182.6040 1045.3614 945.5485 956.2428 877.8184
```

Pri Hampelovom teste môžeme vidieť, že za extrémne hodnoty sú považované všetky okresy Bratislava I až Bratislava V, rovnako ako pri použití Grubbsovho testu, ale jeden okres je v tomto prípade navyše. Ide o okres Košice II, ktorý pri Grubbsovom teste nebol významný na hladine 5 % (pri nulovej hypotéze, že ide o outlier, bola *p*-hodnota 0.09825) a pri tomto okrese sme ukončili testovanie ďalších extrémnych hodnôt v prípade Grubbsovho testu.

Môžeme teda zhodnotiť, že v 6 okresoch v rámci SR sú reálne mzdy vyhodnotené ako extrémne vysoké. Jednoduchým spôsobom by sme mohli testovať aj extrémne nízke hodnoty reálnej mzdy v okresoch SR. Stačí, ak vo funkcii `grubbs.test()` nastavíme parameter `opposite = TRUE`. Pre názornosť ukážeme jeden takýto príklad.

```
> grubbs.test(real_mzda, opposite = TRUE)

                Grubbs test for one outlier

data:  real_mzda
G = 1.5097, U = 0.9704, p-value = 1
alternative hypothesis: lowest value 438.463595 is an outlier
```

Okres Bardejov bol vyhodnotený ako okres, v ktorom je reálna mzda najnižšia (približne 438 EUR). Nejde však o extrém, keďže nulovú hypotézu nevieme zamietnuť (pozri *p*-hodnotu).

Podľa zadania príkladu musíme ešte celý postup zopakovať pre premennú nezamestnanosť.

```
> grubbs.test(nezamestnanost, opposite = FALSE)

                Grubbs test for one outlier

data:  nezamestnanost
G = 3.0649, U = 0.8780, p-value = 0.06348
alternative hypothesis: highest value 33.64 is an outlier
------------------------------------------------------------------
> nezamestnanost2 = nezamestnanost[-55]
> grubbs.test(nezamestnanost2, opposite = FALSE)

                Grubbs test for one outlier

data:  nezamestnanost2
G = 2.5146, U = 0.9168, p-value = 0.4073
alternative hypothesis: highest value 28.83 is an outlier
```

V prípade nezamestnanosti Grubbsov test vyhodnotil ako extrémnu hodnotu nezamestnanosť v okrese Rimavská Sobota (33.64 %) na hladine významnosti 10 % (s *p*-hodnotou 0.06348). Pri ďalšej hodnote podozrivej na extrém sme už nulovú hypotézu zamietnuť nevedeli ani na hladine 10 % (ide o okres Revúca s nezamestnanosťou 28.83 %). Keďže však zo zadanie príkladu vyplýva, že máme nájsť extrémne hodnoty na hladine významnosti 5 %, v prípade premennej nezamestnanosť musíme možno prekvapujúco zhodnotiť, že ani jednu hodnotu nemôžeme považovať za extrémnu. Pozrime sa ešte, či údaje o nezamestnanosti môžeme považovať za realizácie z normálneho rozdelenia.

```
> library(nortest)
> library(stats)
> library(lawstat)
-----------------------------------------------------------------
> ad.test(nezamestnanost)

               Anderson-Darling normality test

data:  nezamestnanost
A = 1.0353, p-value = 0.009518
-----------------------------------------------------------------
> shapiro.test(nezamestnanost)

                 Shapiro-Wilk normality test

data:  nezamestnanost
W = 0.9552, p-value = 0.007362
-----------------------------------------------------------------
> rjb.test(nezamestnanost, option = c("RJB"), crit.values =
  c("empirical"), N = 1000)

                   Robust Jarque Bera Test

data:  nezamestnanost
X-squared = 5.8784, df = 2, p-value = 0.06451
```

Keďže si opäť nemôžeme byť istí, či údaje o nezamestnanosti majú normálne rozdelenie, použijeme neparametrický Hampelov test.

```
> hampel_identifier <- function(data) {
+  ri <- abs(data - median(data))
+  mad <- median(ri)
+  madn <- mad/0.6745
+  hi <- ri/madn
+  critical <- sqrt(qchisq(0.975,1))
+  data[hi>critical]
+ }
> hampel_identifier(nezamestnanost)
[1] 28.83 33.64
```

Tento test identifikoval za extrémne hodnoty nezamestnanosti v okresoch Rimavská Sobota a Revúca. Tieto výsledky môžeme z ekonomického hľadiska považovať za reálnejšie. Z technického hľadiska sme si pri Grubbsovom teste nemohli byť istí predpokladom o normalite, takže výsledky z neparametrického Hampelovho testu budeme považovať za relevantnejšie.

Príklad 7.26

Pri rozhodovaní o nezávislosti premenných sme uviedli dva testy – test náhodnosti dostupný cez funkciu `runs.test()` v knižnici `tseries` a Bartelov test nezávislosti dostupný cez funkciu `bartels.test()` v knižnici `lawstat`[35]. Prvý z nich vieme použiť len pri dichotomických premenných. Pri spojitých výnosoch sa však môže stať, že zmena z jedného dňa na druhý je nulová, čím nám vznikajú tri úrovne faktora (rast, pokles, žiadna zmena). Ak sa pozrieme na údaje za index CAC, tak môžeme vidieť, že 87 pozorovaní je práve tohto typu – teda spojitý výnos je nulový.

```
> library(datasets)
> attach(data.frame(EuStockMarkets))
-----------------------------------------------------------------
> rCAC <- diff(log(CAC))
> rDAX <- diff(log(DAX))
> rFTSE <- diff(log(FTSE))
> rSMI <- diff(log(SMI))
-----------------------------------------------------------------
> rCAC_f <- factor(sign(rCAC), exclude = 0)
> sum(is.na(rCAC_f))
[1] 87
```

V našom príklade teda test náhodnosti `runs.test()` použiť nevieme, keďže nepracujeme s dichotomickou premennou. Môžeme však použiť Bartelov test, pri ktorom budeme testovať závislosť spojitých výnosov, presnejšie povedané, či je prítomná autokorelácia v danom časovom rade (H_0: autokorelácia nie je prítomná, H_1: autokorelácia je prítomná). Príslušný parameter vo funkcii `bartels.test()` teda nastavíme na `alternative = "two.sided"`. V zmysle zadania príkladu nám je totiž jedno, či ide o pozitívnu (`alternative = "positive.correlated"`) alebo negatívnu autokoreláciu (`alternative = "negative.correlated"`).

```
> library(lawstat)
> bartels.test(rCAC, alternative = "two.sided")

                Bartels Test - Two sided

data:  rCAC
Standardized Bartels Statistic = -1.6963, RVN Ratio = 1.921,
p-value = 0.08982
-----------------------------------------------------------------
> bartels.test(rDAX, alternative = "two.sided")
```

[35] Považujeme za dôležité pripomenúť, že v ekonometrii časových radov máme k dispozícii lepšie testy, na základe ktorých môžeme rozhodovať o náhodnosti premenných, resp. o predpovedateľnosti výnosov finančných aktív.

```
                  Bartels Test - Two sided

data:  rDAX
Standardized Bartels Statistic = 1.2116, RVN Ratio = 2.056,
p-value = 0.2257
------------------------------------------------------------
> bartels.test(rFTSE, alternative = "two.sided")

                  Bartels Test - Two sided

data:  rFTSE
Standardized Bartels Statistic = -2.7397, RVN Ratio = 1.873,
p-value = 0.006149
------------------------------------------------------------
> bartels.test(rSMI, alternative = "two.sided")

                  Bartels Test - Two sided

data:  rSMI
Standardized Bartels Statistic = -2.4449, RVN Ratio = 1.887,
p-value = 0.01449
```

Na hladine významnosti 10 % môžeme nulovú hypotézu zamietnuť v prípade spojitých výnosov indexu CAC, na hladine 5 % nulovú hypotézu zamietame pri indexe SMI a na hladine významnosti 1 % môžeme nulovú hypotézu zamietnuť pri indexe FTSE. Každopádne pri nemeckom indexe DAX nulovú hypotézu zamietnuť nevieme ani na hladine 10 %. Hodnoty spojitých výnosov tohto indexu sa tak javia ako náhodné. Ak by sme opomenuli možnú štatistickú závislosť vyššieho rádu, cykly alebo iné formy nelineárnej závislosti, potom v zmysle teórie efektívnych trhov je nemecký akciový trh možné považovať za informačne efektívny (v tzv. slabej forme efektívnosti). Skutočnosť, že zmeny v cenách indexu DAX považujeme za náhodné, je v súlade s modelom náhodnej prechádzky (v angl. *random walk*).

Pri modeli náhodnej prechádzky $P_t = P_{t-1} + \varepsilon_t$ môžeme **zmenu** ceny akcie definovať ako $\Delta P_t = P_t - P_{t-1} = P_{t-1} + \varepsilon_t - P_{t-1} = \varepsilon_t$, kde ε_t je náhodná chyba, a preto zmena ceny akcie závisí len od náhodného člena a je nezávislá od ceny akcie v čase *t-1*.

Je však úplne zrejmé, že vývoj uzatváracích cien na trhu náhodný nebude a určitú perzistentnosť v ich vývoji môžeme očakávať.

```
> bartels.test(CAC, alternative = "two.sided")

                  Bartels Test - Two sided

data:  CAC
Standardized Bartels Statistic = -42.9266, RVN Ratio = 0.009,
p-value < 2.2e-16
------------------------------------------------------------
```

```
> bartels.test(DAX, alternative = "two.sided")

                  Bartels Test - Two sided

data:  DAX
Standardized Bartels Statistic = -43.0712, RVN Ratio = 0.003,
p-value < 2.2e-16
-----------------------------------------------------------------
> bartels.test(FTSE, alternative = "two.sided")

                  Bartels Test - Two sided

data:  FTSE
Standardized Bartels Statistic = -43.0859, RVN Ratio = 0.002,
p-value < 2.2e-16
-----------------------------------------------------------------
> bartels.test(SMI, alternative = "two.sided")

                  Bartels Test - Two sided

data:  SMI
Standardized Bartels Statistic = -43.1018, RVN Ratio = 0.001,
p-value < 2.2e-16
```

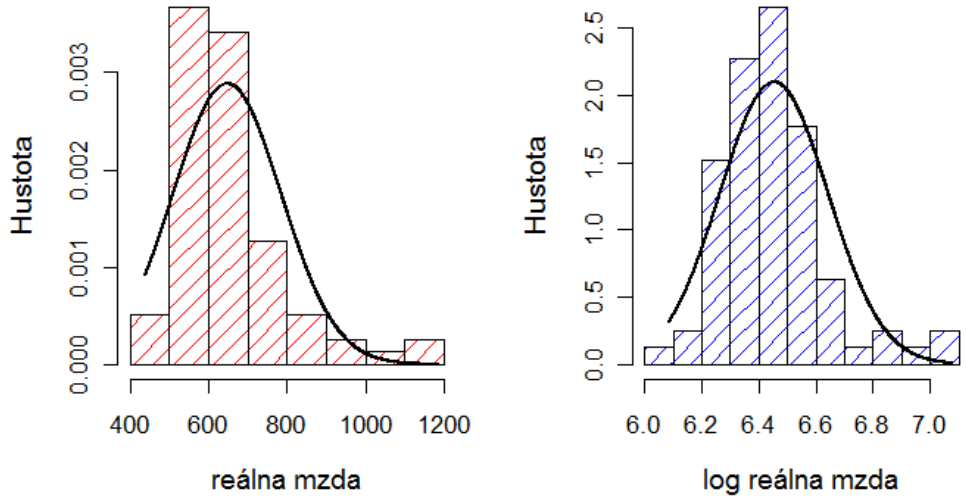
Pri uzatváracích cenách skúmaných akciových indexov môžeme nulovú hypotézu o neprítomnosti autokorelácie zamietnuť, a to vo všetkých prípadoch na hladine významnosti 1 %. Uzatváracie ceny vykazujú zjavné trendy, preto o náhodnosti v tomto prípade nemôžeme hovoriť.

Príklad 7.27

Pri Wilcoxonovom znamienkovom teste sme už uviedli, že rozdelenie hodnôt by malo byť symetrické. V texte sme si už raz pomohli logaritmickou transformáciou premennej, čím sme dosiahli symetrickejšie rozdelenie. V tomto príklade nám logaritmická transformácia reálnej mzdy veľmi nepomôže, čo je viditeľné z nasledujúcich histogramov.

```
> data <- read.csv(file = "...cesta k súboru...\\sk_data.csv",
  sep = ";", dec = ".", header = T)
> attach(data)
-----------------------------------------------------------------
> par(mfrow = c(1, 2))
> hist(real_mzda, density = 10, col = "red", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "reálna mzda")
> x <- seq(min(real_mzda), max(real_mzda), length = 1000)
> xh <- dnorm(x, mean = mean(real_mzda), sd = sd(real_mzda))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
```

```
> hist(log(real_mzda), density = 10, col = "blue", border =
  "black", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE, ylab = "Hustota", xlab = "log reálna mzda")
> x <- seq(min(log(real_mzda)), max(log(real_mzda)), length =
  1000)
> xh <- dnorm(x, mean = mean(log(real_mzda)), sd =
  sd(log(real_mzda)))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
```



Obrázok 7.9: Histogram reálnej mzdy v okresoch SR a jej logaritmická transformácia

Zdroj: vlastné spracovanie, výstup zo softvéru R

Aj po logaritmickej transformácii reálnej mzdy je rozdelenie tejto premennej mierne pravostranne zošikmené. Túto skutočnosť si môžeme ľahko overiť s použitím funkcie `skewness()` z knižnice `moments`.

```
> library(moments)
> skewness(real_mzda)
[1] 1.796726
> skewness(log(real_mzda))
[1] 1.10642
```

Zrejme nám neostáva iná možnosť ako odstrániť odľahlé hodnoty. Grubbsov test nemôžeme použiť kvôli porušenému predpokladu normality – čo je zapríčinené samotným výskytom extrémnych hodnôt. Použijeme teda opäť neparametrický Hampelov test a extrémne hodnoty zo súboru odstránime.

```
> hampel_identifier <- function(data) {
+   ri <- abs(data - median(data))
+   mad <- median(ri)
+   madn <- mad/0.6745
+   hi <- ri/madn
```

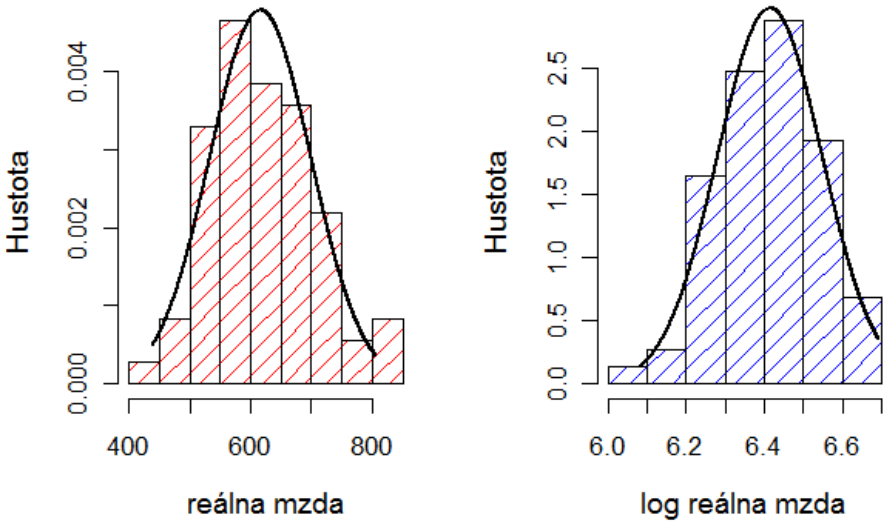
```
+  critical <- sqrt(qchisq(0.975,1))
+  data[hi>critical]
+ }
> hampel_identifier(real_mzda)
[1] 1136.2624 1182.6040 1045.3614 945.5485 956.2428 877.8184
-------------------------------------------------------------
> x <- round(real_mzda, 4)
> real_mzda_out <- x[!(x %in% c(1136.2624, 1182.6040, 1045.3614,
  945.5485, 956.2428, 877.8184))]
```

Keď už máme údaje očistené o extrémne hodnoty, ich rozdelenie by už nemalo byť zošikmené. Po vypočítaní šikmosti sa javí ako vhodné použiť logaritmické hodnoty reálnych miezd.

```
> library(moments)
> skewness(real_mzda_out)
[1] 0.3784237
> skewness(log(real_mzda_out))
[1] 0.09269228
```

Pre lepšiu prehľadnosť sa pozrieme ešte aj na histogram reálnej mzdy očistenej o extrémne hodnoty a jej hodnoty po logaritmickej transformácii.

```
> par(mfrow = c(1, 2))
> hist(real_mzda_out, density = 10, col = "red", border =
  "black", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE, ylab = "Hustota", xlab = "reálna mzda")
> x <- seq(min(real_mzda_out), max(real_mzda_out), length =
  1000)
> xh <- dnorm(x, mean = mean(real_mzda_out), sd =
  sd(real_mzda_out))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-------------------------------------------------------------
> hist(log(real_mzda_out), density = 10, col = "blue", border =
  "black", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE, ylab = "Hustota", xlab = "log reálna mzda")
> x <- seq(min(log(real_mzda_out)), max(log(real_mzda_out)),
  length = 1000)
> xh <- dnorm(x, mean = mean(log(real_mzda_out)), sd =
  sd(log(real_mzda_out)))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
```



Obrázok 7.10: Histogram reálnej mzdy v okresoch SR a jej logaritmická transformácia (bez outlierov)

Zdroj: vlastné spracovanie, výstup zo softvéru R

V tomto príklade máme overiť, či môžeme hodnotu 550 EUR považovať za mediánovú reálnu mzdu v okresoch SR. Keďže chceme pracovať s logaritmickými hodnotami, tak ide o hodnotu 6.309918.

```
> log(550)
[1] 6.309918
```

Hypotézu o rovnosti mediánovej reálnej mzdy v okresoch SR overíme pomocou funkcie `wilcox.test()`. Nulová hypotéza hovorí o tom, že populačný medián je rovný 550 EUR (resp. 6.309918) a alternatívna, že medián nie je rovný tejto hodnote.

```
> wilcox.test(real_mzda_out, mu = log(550), alternative =
  "two.sided")

      Wilcoxon signed rank test with continuity correction

data:  real_mzda_out
V = 2701, p-value = 1.155e-13
alternative hypothesis: true location is not equal to 6.309918
```

Z uvedených výsledkov vyplýva, že mediánová reálna mzda v okresoch SR nie je rovná 550 EUR. Zrejme pôjde o vyššiu hodnotu (čo môžeme ľahko zistiť z deskriptívnej štatistiky). Formálne môžeme overovať alternatívnu hypotézu, kde tvrdíme, že mediánová mzda je po transformácii väčšia ako 6.309918. Vo funkcii `wilcox.test()` môžeme zmeniť alternatívnu hypotézu na `alternative = "greater"`.

```
> wilcox.test(real_mzda_out, mu = log(550), alternative =
  "greater")

      Wilcoxon signed rank test with continuity correction

data:  real_mzda_out
V = 2701, p-value = 5.773e-14
alternative hypothesis: true location is greater than 6.309918
```

Môžeme teda vidieť, že mediánová výška reálnej mzdy v okresoch SR nie je 550 EUR, ale ide o vyššiu hodnotu.

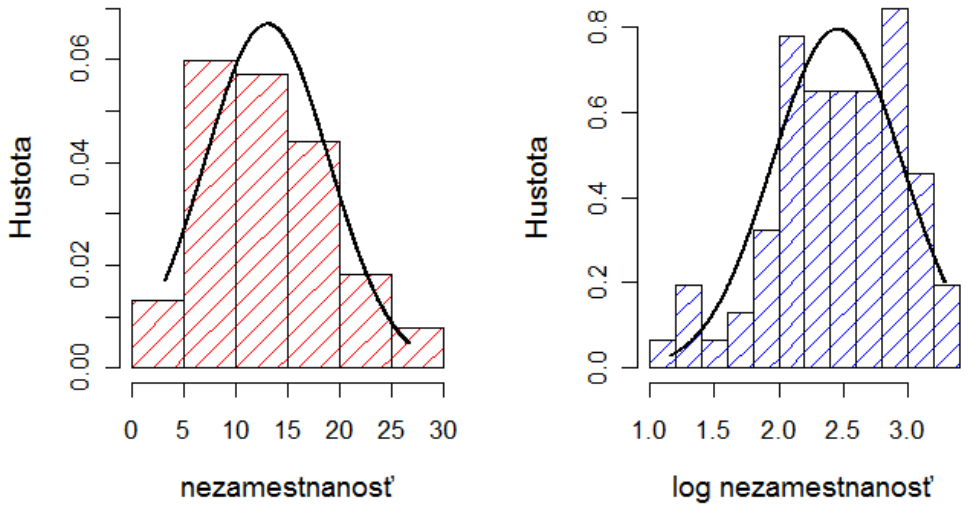
V ďalšej časti tohto príkladu musíme ešte overiť, či sa v okresoch SR vyskytuje mediánová nezamestnanosť nižšia ako 10 %. V prvom rade opäť pristúpime k odstráneniu odľahlých hodnôt pomocou neparametrického Hampelovho testu.

```
> hampel_identifier <- function(data) {
+  ri <- abs(data - median(data))
+  mad <- median(ri)
+  madn <- mad/0.6745
+  hi <- ri/madn
+  critical <- sqrt(qchisq(0.975,1))
+  data[hi>critical]
+ }
> hampel_identifier(nezamestnanost)
[1] 28.83 33.64
-----------------------------------------------------------------
x <- round(nezamestnanost, 2)
nezamestnanost_out <- x[!(x %in% c(28.83, 33.64))]
```

Identifikované sú dve odľahlé hodnoty (okres Revúca s nezamestnanosťou na úrovni 28.83 % a okres Rimavská Sobota s nezamestnanosťou 33.64 %). Tieto hodnoty zo súboru odstránime a údaje o nezamestnanosti (ako aj ich logaritmy) zobrazíme vo forme histogramu kvôli lepšej predstave o ich rozdelení.

```
> par(mfrow = c(1, 2))
> hist(nezamestnanost_out, density = 10, ylim = c(0, 0.07), col
  = "red", border = "black", main = NA, cex.lab = 1.2, cex.axis
  = 1.1, freq = FALSE, ylab = "Hustota", xlab =
  "nezamestnanosť")
> x <- seq(min(nezamestnanost_out), max(nezamestnanost_out),
  length = 1000)
> xh <- dnorm(x, mean = mean(nezamestnanost_out), sd =
  sd(nezamestnanost_out))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
> hist(log(nezamestnanost_out), density = 10, col = "blue",
  border = "black", main = NA, cex.lab = 1.2, cex.axis = 1.1,
  freq = FALSE, ylab = "Hustota", xlab = "log nezamestnanosť")
```

```
> x <- seq(min(log(nezamestnanost_out)),
  max(log(nezamestnanost_out)), length = 1000)
> xh <- dnorm(x, mean = mean(log(nezamestnanost_out)), sd =
  sd(log(nezamestnanost_out)))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
```



Obrázok 7.11: Histogram nezamestnanosti v okresoch SR a jej logaritmická transformácia (bez outlierov)

Zdroj: vlastné spracovanie, výstup zo softvéru R

Údaje bez extrémnych hodnôt sa javia ako symetrické, aj keď v prvom prípade ide o mierne pravostranné zošikmenie a v druhom prípade (zlogaritmované dáta) o mierne ľavostranné zošikmenie. Exaktnejšie to môžeme vidieť po použití funkcie `skewness()`.

```
> library(moments)
> skewness(nezamestnanost_out)
[1] 0.4506697
> skewness(log(nezamestnanost_out))
[1] -0.4701261
```

Wilcoxonov test preto aplikujeme na pôvodné dáta, ako aj na ich logaritmické hodnoty. Pre istotu pripomíname, že pri práci s transformovanými údajmi treba meniť aj argument vo funkcii `wilcox.test()`. V našom prípade, kde sledovaná hodnota je 10 %, ide o hodnotu 2.302585.

```
> log(10)
[1] 2.302585
```

Môžeme už pristúpiť k overeniu hypotézy, či mediánová hodnota nezamestnanosti v okresoch SR je hodnota nižšia ako 10 %.

```
> wilcox.test(nezamestnanost_out, mu = 10, alternative = "less")

      Wilcoxon signed rank test with continuity correction

data:  nezamestnanost_out
V = 2210.5, p-value = 0.9998
alternative hypothesis: true location is less than 10
------------------------------------------------------------------
> wilcox.test(log(nezamestnanost_out), mu = log(10), alternative
  = "less")

      Wilcoxon signed rank test with continuity correction

data:  log(nezamestnanost_out)
V = 2034, p-value = 0.9966
alternative hypothesis: true location is less than 2.302585
```

Nulovú hypotézu zamietnuť nevieme, a teda mediánová nezamestnanosť v okresoch SR je zrejme vyššia ako 10 %. Aby sme toto tvrdenie vedeli dokázať, využijeme opačnú alternatívnu hypotézu.

```
> wilcox.test(nezamestnanost_out, mu = 10, alternative =
  "greater")

      Wilcoxon signed rank test with continuity correction

data:  nezamestnanost_out
V = 2210.5, p-value = 0.0001607
alternative hypothesis: true location is greater than 10
------------------------------------------------------------------
> wilcox.test(log(nezamestnanost_out), mu = log(10), alternative
  = "greater")

      Wilcoxon signed rank test with continuity correction

data:  log(nezamestnanost_out)
V = 2034, p-value = 0.003454
alternative hypothesis: true location is greater than 2.302585
```

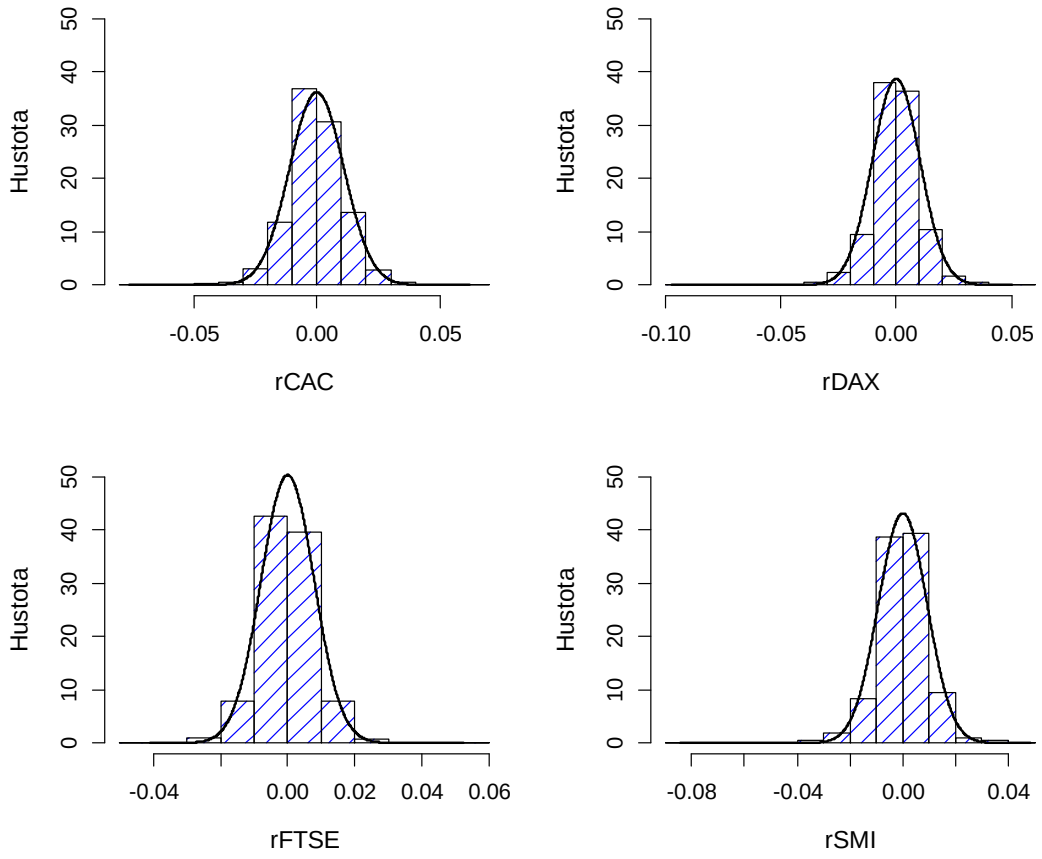
Nulové hypotézy môžeme v oboch prípadoch (na pôvodných dátach ako aj logaritmoch nezamestnanosti) zamietnuť na hladine významnosti 1 %. Z týchto výsledkov už môžeme prijať tvrdenie, že mediánová hodnota nezamestnanosti v okresoch SR je vyššia ako 10 %.

Príklad 7.28

V prvom kroku pristúpime k transformácii uzatváracích cien na spojité výnosy tak, ako v predchádzajúcich príkladoch, v ktorých sme pracovali s touto databázou (tento postup už na tomto mieste neuvádzame).

V predchádzajúcom príklade sme zistili, že vo všeobecnosti nadmerné výnosy skúmaných akciových indexov nemôžeme považovať za realizácie z normálneho rozdelenia. Pre použitie Wilcoxonovho testu však údaje nemusia byť z normálneho rozdelenia, ale stačí, ak bude rozdelenie symetrické. Aby sme mohli posúdiť, či údaje nie sú zošikmené, zostrojíme si najprv histogram (prejdeme na trochu efektívnejší spôsob generovania obrázkov).

```
> par(mfrow = c(2, 2))
> data_markets <- data.frame(rCAC, rDAX, rFTSE, rSMI)
> xlabs <- c("rCAC", "rDAX", "rFTSE", "rSMI"); j = 1
> for (srs in data_markets) {
+ srs <- na.omit(srs)
+ hist(srs, density = 10, col = "blue", border = "black", ylim =
  c(0, 50), main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE, ylab = "Hustota", xlab = names(data_markets)[j])
+ x <- seq(min(srs), max(srs), length = 1000)
+ xh <- dnorm(x, mean = mean(srs), sd = sd(srs))
+ dx <- data.frame(x, xh)
+ lines(dx, type = "l", col = "black", lwd = 2)
+ j = j + 1
+ }
```



Obrázok 7.12: Histogram nadmerných výnosov akciových indexov

Zdroj: vlastné spracovanie, výstup zo softvéru R

Na základe vizualizácie dát prostredníctvom histogramu sa rozdelenie použitých dát javí ako symetrické. Pozrieme sa ešte na deskriptívnu štatistiku, kde pomocou vybraných percentilov vieme určiť mieru zošikmenia. Všimnime si, že vzdialenosti 10-teho percentilu od mediánu a 90-teho percentilu od mediánu sú prakticky totožné, podobne aj vzdialenosti 25-teho percentilu a 75-teho percentilu od mediánu. Rozdelenia môžeme považovať za pomerne symetrické.

```
> skew_markets <- matrix(nrow = 4, ncol = 7);
  colnames(skew_markets) <- c("mean", "sd", "10th %", "25th %",
  "50th %", "75th %", "90th %")
> rownames(skew_markets) <- names(data_markets); j = 1
>  for (srs in data_markets) {
+   srs <- round(na.omit(srs), 3)
+  descriptives <- c(round(mean(srs), 3), round(sd(srs), 3),
  quantile(srs, probs = c(0.10, 0.25, 0.5, 0.75, 0.9)))
+   skew_markets[j,] <- descriptives
+   j = j + 1
+  }
> skew_markets
       mean    sd 10th % 25th % 50th % 75th % 90th %
rCAC  0.000 0.011 -0.012 -0.006  0.000  0.007  0.014
rDAX  0.001 0.010 -0.011 -0.005  0.000  0.006  0.013
rFTSE 0.000 0.008 -0.009 -0.004  0.000  0.005  0.010
rSMI  0.001 0.009 -0.010 -0.004  0.001  0.006  0.011
```

Môžeme pristúpiť k testovaniu prostredníctvom Wilcoxonovho testu, pomocou ktorého sa pokúsime zistiť, či mediánovou hodnotou nadmerných výnosov je hodnota vyššia ako 0 – inými slovami, či indexy vykazujú častejšie kladné výnosy ako záporné.

```
> wilcox.test(rCAC, mu = 0, alternative = "greater")

      Wilcoxon signed rank test with continuity correction

data:  rCAC
V = 865768, p-value = 0.461
alternative hypothesis: true location is greater than 0
------------------------------------------------------------
> wilcox.test(rDAX, mu = 0, alternative = "greater")

      Wilcoxon signed rank test with continuity correction

data:  rDAX
V = 875025, p-value = 0.3092
alternative hypothesis: true location is greater than 0
------------------------------------------------------------
> wilcox.test(rFTSE, mu = 0, alternative = "greater")

      Wilcoxon signed rank test with continuity correction

data:  rFTSE
```

```
V = 863333, p-value = 0.503
alternative hypothesis: true location is greater than 0
------------------------------------------------------------
> wilcox.test(rSMI, mu = 0, alternative = "greater")

      Wilcoxon signed rank test with continuity correction

data:  rSMI
V = 884943, p-value = 0.177
alternative hypothesis: true location is greater than 0
```

Ani pri jednom z týchto indexov sme nulovú hodnotu nevedeli zamietnuť. Pripomenieme, že nadmerný výnos je výnos oproti očakávanému výnosu (z tzv. autoregresného modelu prvého rádu). V tomto teste nás teda zaujíma, či oproti očakávanému výnosu dosahujeme častejšie vyšší výnos. Odpoveďou je, že nami získané údaje túto myšlienku nepotvrdzujú.

Pred samotným testovaním sme mohli ešte odstrániť extrémne hodnoty (keďže spojité výnosy finančných aktív sa vyznačujú tzv. tučnými koncami), ktoré sa pri takýchto časových radoch samozrejme vyskytujú. Takto by sme mohli posúdiť odolnosť nami dosiahnutých výsledkov na extrémne hodnoty. Zopakujeme ešte celý postup pre dané časové rady po odstránení extrémnych hodnôt. Keďže však údaje nemajú normálne rozdelenie, opäť použijeme neparametrický Hampelov test.

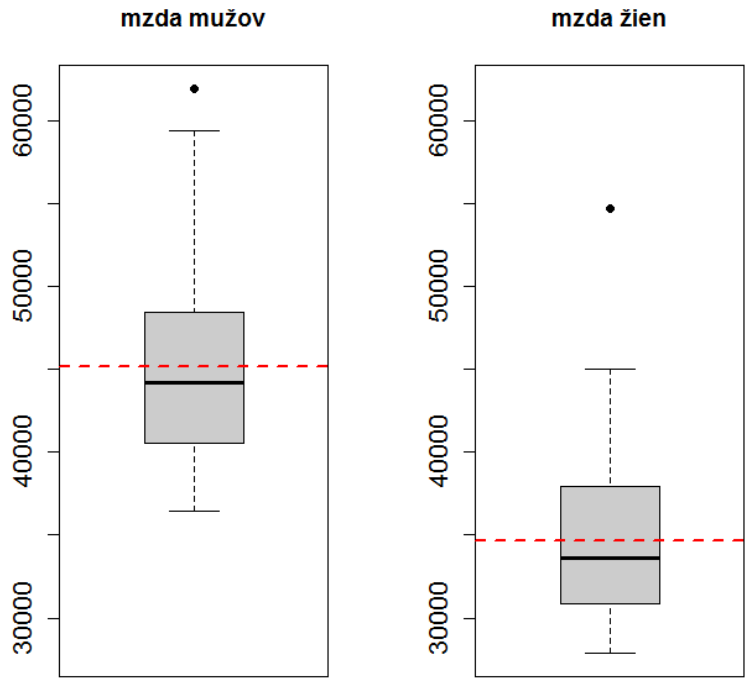
```
> data_markets <- round(data_markets, 5)
> wilk_markets <- matrix(nrow = 4, ncol = 2);
  colnames(wilk_markets) <- c("test statistics", "p-value")
> rownames(wilk_markets) <- names(data_markets); j <- 1
> for (srs in data_markets) {
+ srs_out <- srs[!(x %in% c(hampel_identifier(srs)))]
+ temp <- wilcox.test(srs_out, mu = 0, alternative = "greater")
+ wilk_markets[j,] <- c(temp[[1]], temp[[3]])
+ j <- j + 1
+ }
------------------------------------------------------------------
> wilk_markets
      test statistics   p-value
rCAC          863997 0.4594957
rDAX          873173 0.3088607
rFTSE         863316 0.5032772
rSMI          884016 0.1767846
```

Podobne ako v predošlých prípadoch, ani teraz sme nevedeli zamietnuť nulovú hypotézu. Tento záver nie je veľmi prekvapivý, keďže tieto testy vychádzajú z poradí a tie sú na extrémne hodnoty málo náchylné. Ide o jednu z výhod neparametrických testov založených na poradiach hodnôt.

Príklad 7.29

V tomto príklade máme zistiť, či mzdy mužov je možné považovať vo všeobecnosti za vyššie ako mzdy žien. K dispozícii máme priemerné ročné mzdy mužov a žien z 51 štátov USA za rok 2009. Na začiatok si zobrazíme údaje tak, aby sme mohli vidieť rozdiely v polohách a variabilite príjmov mužov a žien. Preto si najprv zostrojíme box – plot.

```
> mzda_M <- data$Mzda[data$Pohlavie == 1]
> mzda_Z <- data$Mzda[data$Pohlavie == 0]
> par(mfrow = c(1, 2))
> minimum <- min(data$Mzda)
> maximum <- max(data$Mzda)
> boxplot(mzda_M, main = "mzda mužov", col = gray(0.8), pch =
  19, cex.axis = 1.3, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(mzda_M), lwd = 2, lty = 2, col = "red")
> boxplot(mzda_Z, main = "mzda žien", col = gray(0.8), pch = 19,
  cex.axis = 1.3, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(mzda_Z), lwd = 2, lty = 2, col = "red")
```



Obrázok 7.13: Box – plot priemerných miezd mužov a žien

Zdroj: vlastné spracovanie, výstup zo softvéru R

Z uvedených box – plotov sa javí, že priemerné mzdy mužov sú skutočne vyššie ako priemerné mzdy žien. Všeobecným pravidlom je, že pokiaľ sa „krabice“ neprekrývajú, tak medzi premennými existujú významné rozdiely v charakteristike polohy. Samozrejme je to

len určité pravidlo a nezohľadňuje úplne variabilitu. Otázkou teda ostáva, či tieto vizuálne rozdiely sú aj štatisticky významné a kvôli takému zisteniu už je nutné exaktnejšie testovanie.

V prvom kroku overíme, či rozdelenie údajov môžeme považovať za normálne. Tento krok je pre nás podstatný najmä z toho dôvodu, že pokiaľ bude predpoklad normality porušený, budeme musieť zvoliť alternatívu k *t*-testu zhody dvoch stredných hodnôt v podobe neparametrického testu.

Na testovanie normality využijeme tri testy (s ktorými sme už pracovali), a to Anderson – Darlingov (funkcia `ad.test()` z knižnice `nortest`), Shapiro – Wilkov (funkcia `shapiro.test()` z knižnice `stats`) a Jarque – Berov test (funkcia `rjb.test()` z knižnice `lawstat`). Pri Jarque – Berovom teste budeme vychádzať z empirických (simulovaných) kritických hodnôt a použijeme iba jeho modifikovanú verziu (`option = "RJB"`).

```
> library(nortest); library(stats); library(lawstat)
> data <- read.csv(file = "...cesta k suboru...\\genders.csv",
  sep = ";", dec = ".", header = T)
-----------------------------------------------------------------
> mzda_M <- subset(data$Mzda, subset = data$Pohlavie == 1)
> mzda_Z <- subset(data$Mzda, subset = data$Pohlavie == 0)
> mzda <- data.frame(mzda_Z, mzda_M)
-----------------------------------------------------------------
> norm_results <- matrix(nrow = 3, ncol = 4)
> colnames(norm_results) <- c("Ženy stat.", "Ženy p-val.", "Muži
  stat.", "Muži p-val.")
> rownames(norm_results) <- c("Anderson - Darling", "Shapiro -
  Wilk", "Jarque - Bera")
> j = 1
> for (srs in mzda) {
+  a <- ad.test(srs); b <- shapiro.test(srs); c <- rjb.test(srs)
+   norm_results[, j] <- round(c(a$statistic, b$statistic,
  c$statistic), 5)
+   norm_results[, j+1] <- round(c(a$p.value, b$p.value,
  c$p.value), 4)
+   j = j + 2
+ }
-----------------------------------------------------------------
> norm_results
                  Ženy stat. Ženy p-val. Muži stat. Muži p-val.
Anderson - Darling    1.33340      0.0017    1.08593      0.0069
Shapiro - Wilk        0.89345      0.0003    0.92702      0.0038
Jarque - Bera        27.78325      0.0000    7.63246      0.0220
```

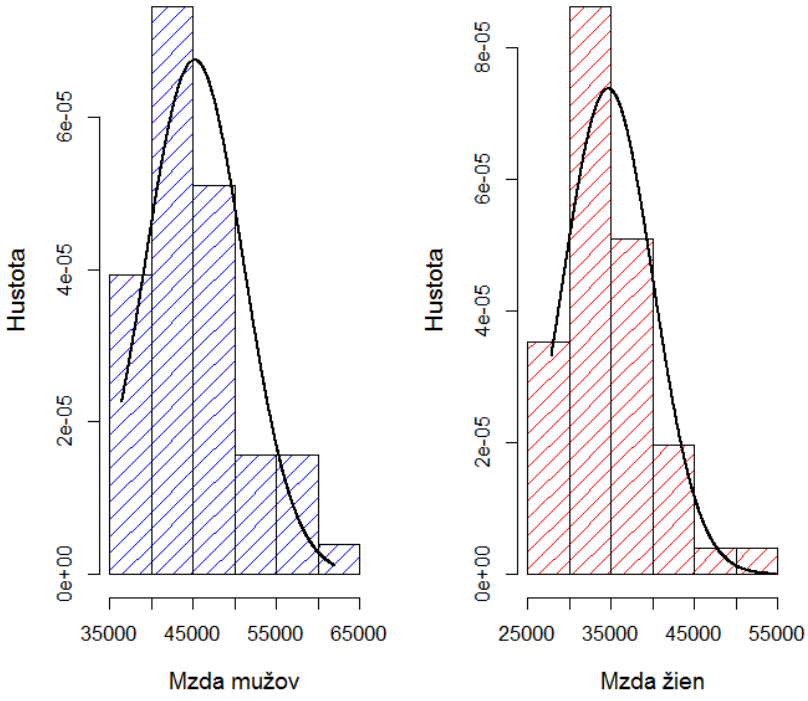
Normalitu miezd mužov aj žien môžeme zamietnuť na základe výsledkov zo všetkých použitých testov. V ďalšom kroku teda nebudeme používať parametrické testy, ale ako

vhodná alternatíva sa javí Mann – Whitney – Wilcoxonov test (v softvéri R dostupný cez funkciu `wilcox.test()`).

```
> wilcox.test(mzda_M, mzda_Z, paired = F, alternative =
  "greater")

        Wilcoxon rank sum test with continuity correction

data:  mzda_M and mzda_Z
W = 2394, p-value = 1.286e-13
alternative hypothesis: true location shift is greater than 0
```

Z výsledkov vyplýva, že mzdy mužov sú vo všeobecnosti vyššie ako mzdy žien, keďže nulovú hypotézu o rovnosti rozdelení (v miere polohy) môžeme zamietnuť v prospech alternatívnej hypotézy.

Pri tomto teste je vhodné overiť, či tvar rozdelení je rovnaký, aby sme sa mohli držať interpretácie o posunutí jedného rozdelenia voči druhému. Z tohto dôvodu si zobrazíme dáta vo forme histogramov, pričom môžeme vidieť, že tvar rozdelení sa javí ako veľmi podobný.



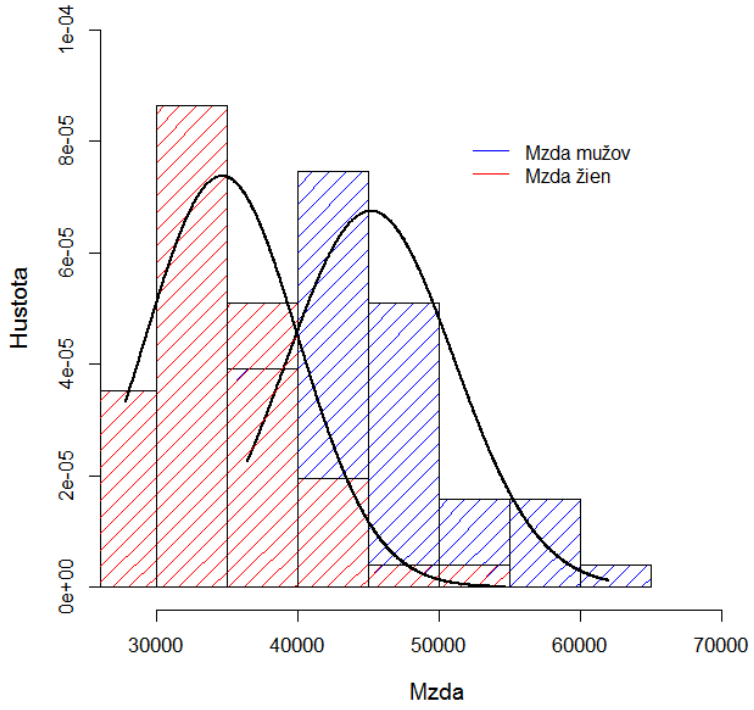
Obrázok 7.14: Histogram priemerných miezd mužov a žien

Zdroj: vlastné spracovanie, výstup zo softvéru R

```
> par(mfrow = c(1, 2))
> hist(mzda_M, density = 10, col = "blue", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "Mzda mužov")
> x <- seq(min(mzda_M), max(mzda_M), length = 1000)
> xh <- dnorm(x, mean = mean(mzda_M), sd = sd(mzda_M))
```

```
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
> hist(mzda_Z, density = 10, col = "red", border = "black", main
  = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "Mzda žien")
> x <- seq(min(mzda_Z), max(mzda_Z), length = 1000)
> xh <- dnorm(x, mean = mean(mzda_Z), sd = sd(mzda_Z))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
```

Keďže z grafickej vizualizácie sme dospeli k záveru, že rozdelenia sú si veľmi podobné, môžeme ostať pri interpretácii použitého testu o posunutí rozdelenia miezd mužov „doprava“ (po reálnej osi miezd x) od rozdelenia miezd žien. Posun rozdelení bolo možné vidieť už z uvedených box – plotov. Zobraziť tento rozdiel môžeme aj iným spôsobom, a to použitím vzájomne sa prekrývajúcich histogramov.



Obrázok 7.15: Prekrývajúce sa histogramy priemerných miezd mužov a žien

Zdroj: vlastné spracovanie, výstup zo softvéru R

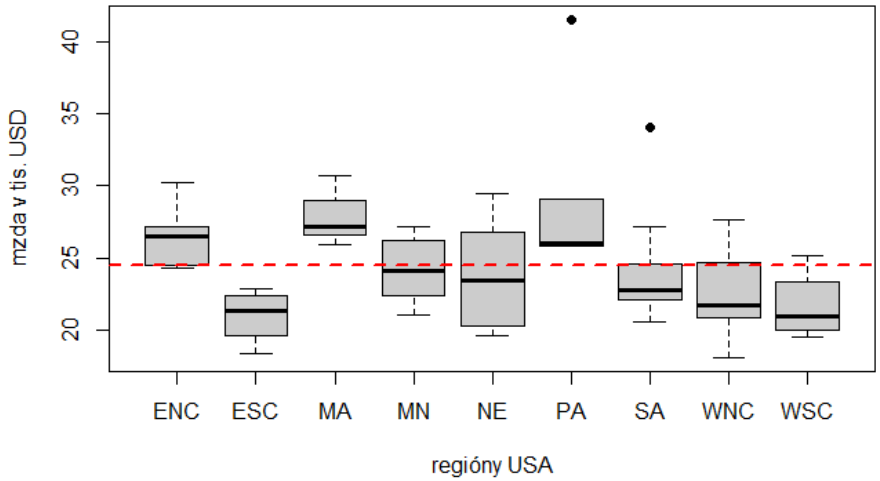
```
> hist(mzda_M, density = 10, col = "blue", border = "black",
  main = NA, xlim = c(min(Mzda), max(Mzda)+10000), ylim = c(0,
  0.0001), cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "Mzda")
> x <- seq(min(mzda_M), max(mzda_M), length = 1000)
> xh <- dnorm(x, mean = mean(mzda_M), sd = sd(mzda_M))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
```

```
> hist(mzda_Z, add = TRUE, density = 10, col = "red", border =
  "black", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE)
> x <- seq(min(mzda_Z), max(mzda_Z), length = 1000)
> xh <- dnorm(x, mean = mean(mzda_Z), sd = sd(mzda_Z))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
> legend("topright", legend = c("Mzda mužov", "Mzda žien"), lty
  = 1.5, col = c("blue", "red"), inset = 0.2, bty = "n")
```

Príklad 7.30

Za účelom overenia hypotézy o rovnakých rozdeleniach miezd učiteľov (premenná `Pay`) medzi regiónmi v USA využijeme Kruskal – Wallisov test. Alternatívou by mohol byť neparametrický Mann – Whitney – Wilcoxonov test medzi všetkými dvojicami regiónov (čo by bolo ale časovo náročnejšie). Pozrime sa najprv na vizualizáciu dát vo forme box – plotov.

```
> data <- read.csv(file = "...cesta
  k súboru...\\educ_spending.csv", sep = ";", dec = ".", header
  = T)
> attach(data)
------------------------------------------------------------------
> table(Region)
Region
ENC ESC  MA  MN  NE  PA  SA WNC WSC
  5   4   3   8   6   5   9   7   4
------------------------------------------------------------------
> boxplot(Pay ~ Region, ylab = "mzda v tis. USD", xlab =
  "regióny USA", col = gray(0.8), pch = 19, cex.axis = 1,
  cex.lab = 1)
> abline(h = mean(Pay), lwd = 2, lty = 2, col = "red")
```



Obrázok 7.16: Mzdy učiteľov po regiónoch v USA

Zdroj: vlastné spracovanie, výstup zo softvéru R

Z grafickej vizualizácie na prvý pohľad vidíme, že aj keď v mnohých prípadoch sú rozdiely medzi regiónmi minimálne, v niektorých sú pomerne veľké. Napríklad v regiónoch Pacific a Mid-Atlantic sú mzdy výrazne vyššie ako v regióne East South Central alebo West South Central. Červenou prerušovanou čiarou je zvýraznený priemer vypočítaný pre všetky štáty (ktoré sú prvkami regiónov). Ku konkrétnym priemerným hodnotám miezd učiteľov po jednotlivých regiónoch sa tiež vieme jednoducho dostať a môžeme pozorovať značné rozdiely. Pre lepšiu predstavu o dátach zobrazíme tiež smerodajnú odchýlku, minimum a maximum.

```
> xbar <- tapply(Pay, Region, mean)
> s <- tapply(Pay, Region, sd)
> min <- tapply(Pay, Region, min)
> max <- tapply(Pay, Region, max)
> cbind(priemer = xbar, št.odchýlka = s, minimum = min, maximum
  = max)

       priemer   št.odchýlka   minimum   maximum
ENC   26.54000      2.398541      24.3      30.2
ESC   21.00000      1.916594      18.4      22.9
MA    27.93333      2.482606      25.9      30.7
MN    24.21250      2.357624      21.0      27.2
NE    23.85000      4.276798      19.6      29.5
PA    29.64000      6.776651      25.8      41.5
SA    24.28889      4.115047      20.6      34.0
WNC   22.64286      3.549111      18.1      27.6
WSC   21.65000      2.490649      19.5      25.2
```

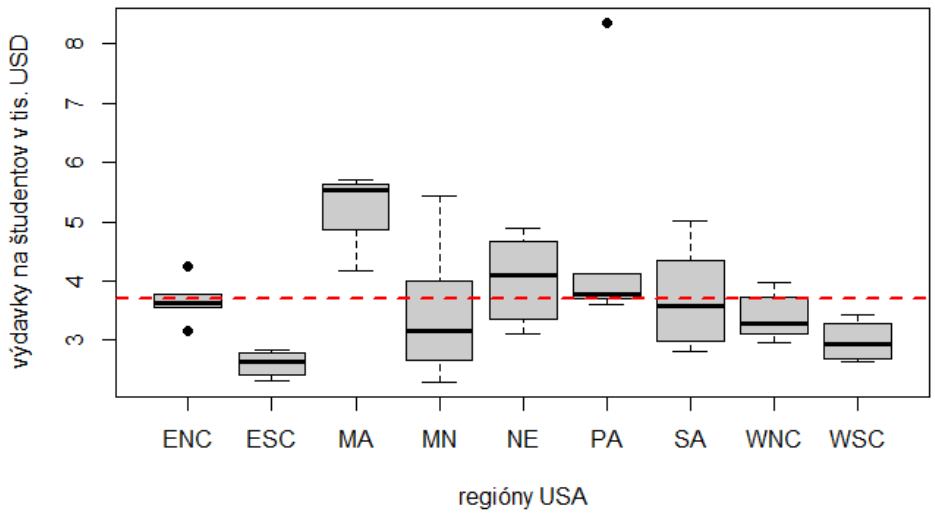
Z vypočítaných priemerov a uvedených box – plotov vidíme, že evidentne existujú rozdiely medzi mzdami učiteľov v jednotlivých regiónoch USA. Kruskal – Wallisov test nám poslúži na formalizovanie tohto tvrdenia.

```
> kruskal.test(Pay, Region)

                Kruskal-Wallis rank sum test

data:  Pay and Region
Kruskal-Wallis chi-squared = 16.5766, df = 8, p-value = 0.03483
```

Nulovú hypotézu o rovnosti rozdelení miezd učiteľov môžeme zamietnuť na hladine významnosti 5 % a prijať alternatívnu hypotézu, že v aspoň jednom prípade je rozdelenie odlišné od ostatných rozdelení (v miere polohy).

V ďalšom kroku sa pozrieme, či existujú rozdiely aj vo výdavkoch na študentov, resp. overíme hypotézu, že rozdelenia výdavkov na študentov vo všetkých regiónoch v USA sú rovnaké. Celú analýzu začneme opäť vizualizáciou dát a základnou deskriptívnou štatistikou premennej `Spend`.

```
> boxplot(Spend ~ Region, ylab = "výdavky na študentov v tis.
  USD", xlab = "regióny USA", col = gray(0.8), pch = 19,
  cex.axis = 1, cex.lab = 1)
> abline(h = mean(Spend), lwd = 2, lty = 2, col = "red")
```



Obrázok 7.17: Výdavky na študentov po regiónoch v USA

Zdroj: vlastné spracovanie, výstup zo softvéru R

Podobne ako v predchádzajúcom prípade nám vizualizácia údajov základnej opisnej štatistiky naznačuje, že aj pri tejto premennej existujú značné rozdiely medzi regiónmi v USA. Zaujímavá je tiež rozdielna variabilita vo vnútri regiónu, teda medzi jednotlivými štátmi.

```
> xbar <- tapply(Spend, Region, mean)
> s <- tapply(Spend, Region, sd)
> min <- tapply(Spend, Region, min)
> max <- tapply(Spend, Region, max)
> cbind(priemer = xbar, št.odchýlka = s, minimum = min, maximum
  = max)

      priemer   št.odchýlka   minimum   maximum
ENC     3.672     0.3954365      3.16      4.25
ESC     2.605     0.2368544      2.31      2.85
MA      5.140     0.8443341      4.17      5.71
MN      3.425     1.0293271      2.30      5.44
NE      4.035     0.7823490      3.11      4.89
PA      4.712     2.0427481      3.61      8.35
SA      3.700     0.7794549      2.82      5.02
WNC     3.420     0.4067759      2.97      3.98
WSC     2.985     0.3607862      2.64      3.43
```

Zrejme rozdelenia výdavkov na študentov sú dokonca viac rozdielne ako mzdy učiteľov. Za účelom exaktného overenia hypotézy o rovnosti rozdelení výdavkov medzi všetkými regiónmi v USA použijeme opäť Kruskal – Wallisov test.

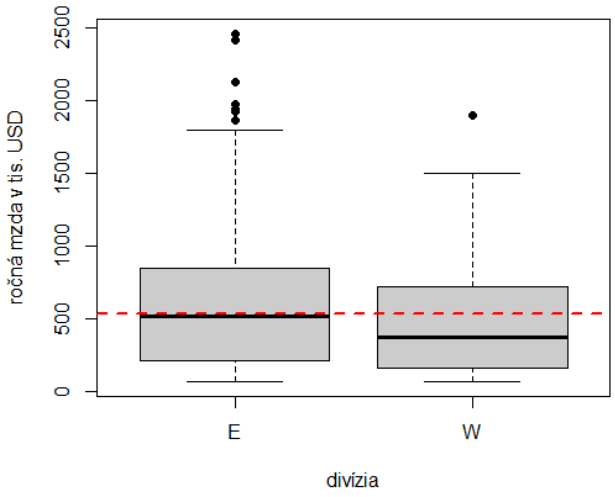
```
> kruskal.test(Spend, Region)

                Kruskal-Wallis rank sum test

data:  Spend and Region
Kruskal-Wallis chi-squared = 21.4083, df = 8, p-value = 0.006138
```

Nulovú hypotézu môžeme v prípade tejto premennej zamietnuť na hladine významnosti 1 %, a teda prikloníme sa k alternatívnej hypotéze, že aspoň jeden región vykazuje odlišné rozdelenie výdavkov na študentov (v strednej hodnote).

Príklad 7.31

Ideme overovať hypotézu, že mzdy hráčov východnej divízie sú väčšie ako mzdy hráčov zo západnej divízie. Za týmto účelom budeme používať neparametrické testy. Najprv sa však pozrieme na dáta vo vizuálnej podobe.

```
> library(vcd)
> attach(Baseball)
-----------------------------------------------------------------
> boxplot(sal87 ~ div86, ylab = "ročná mzda v tis. USD", xlab =
  "divízia", col = gray(0.8), pch = 19, cex.axis = 1, cex.lab =
  1)
> abline(h = mean(na.omit(sal87)), lwd = 2, lty = 2, col =
  "red")
```



Obrázok 7.18: Mzdy hráčov vo východnej (E) a západnej divízii (W)

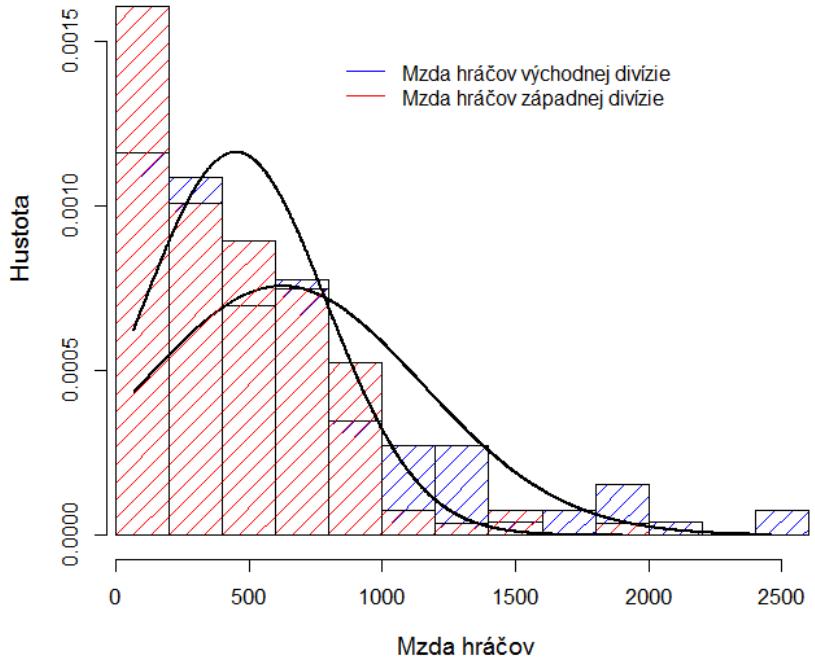
Zdroj: vlastné spracovanie, výstup zo softvéru R

Mzdy hráčov vo východnej divízii sa zdajú byť o čosi vyššie (vzhľadom na medián). Taktiež môžeme pozorovať výskyt viacerých odľahlých hodnôt. Týmito extrémami sa zatiaľ zaoberať nebudeme a hypotézu overíme bez ich odstránenia. Najprv rozdelíme premennú `sal87` (mzdy hráčov) podľa príslušnosti k divízii a využijeme Mann – Whitney – Wilcoxonov test (funkcia `wilcox.test()` pracuje len s numerickými premennými).

```
> East <- na.omit(subset(sal87, subset = div86 == "E"))
> West <- na.omit(subset(sal87, subset = div86 == "W"))
------------------------------------------------------------
> wilcox.test(East, West, paired = F, alternative = "greater")

        Wilcoxon rank sum test with continuity correction

data:  East and West
W = 10109.5, p-value = 0.008717
alternative hypothesis: true location shift is greater than 0
```

Nulovú hypotézu o rovnosti rozdelení môžeme zamietnuť na hladine významnosti 1 % a prijať alternatívnu hypotézu, že mzdy hráčov z východnej divízie sú vo všeobecnosti väčšie ako miery polohy miezd hráčov zo západnej divízie. Skutočnosť, že rozdelenie miezd hráčov z východnej divízie je posunuté doprava, môžeme pozorovať po zostrojení prekrývajúceho sa histogramu skúmaných dát.

```
> hist(East, density = 10, col = "blue", border = "black", main
  = NA, xlim = c(min(na.omit(sal87)), max(na.omit(sal87))+100),
  ylim = c(0, 0.0018), cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE, ylab = "Hustota", xlab = "Mzda hráčov")
> x <- seq(min(East), max(East), length = 1000)
> xh <- dnorm(x, mean = mean(East), sd = sd(East))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
------------------------------------------------------------
> hist(West, add = TRUE, density = 10, col = "red", border =
  "black", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE)
> x <- seq(min(West), max(West), length = 1000)
> xh <- dnorm(x, mean = mean(West), sd = sd(West))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
> legend("topright", legend = c("Mzda hráčov východnej divízie",
  "Mzda hráčov západnej divízie"), lty = 1.8, col = c("blue",
  "red"), inset = 0.2, bty = "n")
```



Obrázok 7.19: Histogram mzdy hráčov vo východnej a západnej divízii

Zdroj: vlastné spracovanie, výstup zo softvéru R

Stredná hodnota mzdy hráčov z východnej divízie je mierne vyššia a celé rozdelenie je posunuté doprava (zrejme aj vplyvom extrémnych hodnôt).

S cieľom potvrdiť dosiahnuté výsledky by sme mohli analýzu zopakovať bez extrémnych hodnôt. Na tomto mieste je potrebné zdôrazniť, že vplyv extrémnych hodnôt sa preukáže ani nie tak veľkosťou extrémov v jednotlivých divíziách, ale skôr skutočnosťou, že dvakrát viac extrémnych hodnôt sa vyskytuje v súbore východnej divízie. Tieto neparametrické testy totiž konvertujú údaje na poradia. Preto veľkosť extrémnej hodnoty nemá tak výrazný vplyv na výsledky ako počet extrémnych hodnôt. Za týmto účelom aplikujeme (nám už známy) neparametrický Hampelov test a k nemu vytvorenú funkciu v softvéri R `hampel_identifier()`.

```
> hampel_identifier <- function(data) {
+   ri <- abs(data - median(data))
+   mad <- median(ri)
+   madn <- mad/0.6745
+   hi <- ri/madn
+   critical <- sqrt(qchisq(0.975,1))
+   data[hi>critical]
+ }
------------------------------------------------------------------
> x <- round(East, 4)
> East_out <- x[!(x %in% c(hampel_identifier(x)))];
  length(hampel_identifier(x))
[1] 10
```

```
> x <- round(West, 4)
> West_out = x[!(x %in% c(hampel_identifier(x)))];
  length(hampel_identifier(x))
[1] 5
```

V mzdách hráčov východnej divízie bolo identifikovaných 10 odľahlých hodnôt a v mzdách hráčov západnej divízie 5 odľahlých hodnôt (pre naše účely nás v tomto bode nezaujíma, o ktoré hodnoty presne ide, preto sme si vypísali len počet a nie konkrétne hodnoty). Po identifikovaní a odstránení extrémnych hodnôt opäť využijeme Mann – Whitney – Wilcoxonov test.

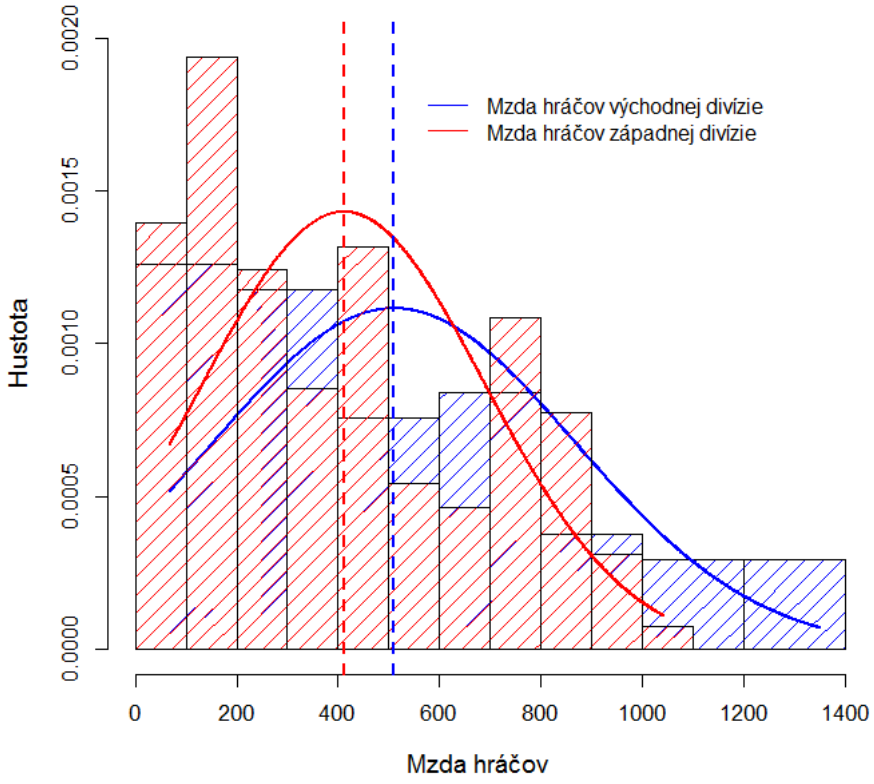
```
> wilcox.test(East_out, West_out, paired = F, alternative =
  "greater")

        Wilcoxon rank sum test with continuity correction

data:  East_out and West_out
W = 8765.5, p-value = 0.02677
alternative hypothesis: true location shift is greater than 0
```

Nulovú hypotézu môžeme stále zamietnuť. Došlo k zvýšeniu *p*-hodnoty (a zníženiu testovacej štatistiky). Uvedené je spôsobené tým, že: 1) extrémne hodnoty sú iba maximálne, 2) viac extrémnych hodnôt sa nachádza vo východnej divízii, kde sa javia byť mzdy vyššie. Výsledky tak aj po zohľadnení odľahlých hodnôt ostávajú významné pri $\alpha = 5$ %.

Na záver sa ešte pozrieme ako vyzerajú histogramy miezd po odstránení extrémnych hodnôt a či stále je možné pozorovať posunutie rozdelenia miezd hráčov východnej divízie aj na základe grafickej vizualizácie. Priemerné mzdy pre jednotlivé divízie zobrazíme vertikálnymi prerušovanými čiarami.

```
> hist(East_out, density = 10, col = "blue", border = "black",
  main = NA, xlim = c(0, max(East_out)), ylim = c(0, 0.002),
  cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab = "Hustota",
  xlab = "Mzda hráčov")
> x <- seq(min(East_out), max(East_out), length = 1000)
> xh <- dnorm(x, mean = mean(East_out), sd = sd(East_out))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "blue", lwd = 2)
------------------------------------------------------------
> hist(West_out, add = TRUE, density = 10, col = "red", border =
  "black", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE)
> x <- seq(min(West_out), max(West_out), length = 1000)
> xh <- dnorm(x, mean = mean(West_out), sd = sd(West_out))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "red", lwd = 2)
```

```
> legend("topright", legend = c("Mzda hráčov východnej divízie",
  "Mzda hráčov západnej divízie"), lty = 1.8, col = c("blue",
  "red"), inset = 0.1, bty = "n")
-----------------------------------------------------------------
> abline(v = mean(East_out), lwd = 2, lty = 2, col = "blue")
> abline(v = mean(West_out), lwd = 2, lty = 2, col = "red")
```



Obrázok 7.20: Histogram mzdy hráčov vo východnej a západnej divízii (bez extrémov)

Zdroj: vlastné spracovanie, výstup zo softvéru R

Príklad 7.32

V tomto príklade máme overiť, či sú rozdiely vo variabilite miezd učiteľov (premenná `Pay`) a vo výdavkoch na študentov (premenná `Spend`) štatisticky významné. Takáto analýza nám umožňuje rozhodnúť o tom, v ktorých častiach USA je väčšia rovnosť medzi mzdami učiteľov (a druhou sledovanou premennou – výdavkami na študentov). V danej databáze máme údaje za jednotlivé štáty USA, ktoré sú pridelené do jednotlivých regiónov (premenná `Region`). Pozrime sa najprv na hodnoty smerodajných odchýlok v daných regiónoch.

```
> data <- read.csv(file = ""...cesta
  k súboru...\\educ_spending.csv", sep = ";", dec = ".", header
  = T)
> attach(data)
-----------------------------------------------------------------
```

```
> s_1 <- tapply(Pay, Region, sd)
> s_2 <- tapply(Spend, Region, sd)
> cbind("št.odchýlka v mzdách" = s_1, "št.odchýlka vo výdavkoch"
  = s_2)
          št.odchýlka v mzdách   št.odchýlka vo výdavkoch
ENC              2.398541                0.3954365
ESC              1.916594                0.2368544
MA               2.482606                0.8443341
MN               2.357624                1.0293271
NE               4.276798                0.7823490
PA               6.776651                2.0427481
SA               4.115047                0.7794549
WNC              3.549111                0.4067759
WSC              2.490649                0.3607862
```

V niektorých prípadoch sú rozdiely vo variabilite skúmaných premenných medzi jednotlivými regiónmi troj- až štvornásobné, je však otázne, či budú aj štatisticky významné. Na tomto mieste považujeme za vhodné upozorniť, že pracujeme s málo početnými súbormi (z tohto pohľadu preto ide skôr len o ukážkový príklad).

Levenov test je v softvéri R dostupný prostredníctvom funkcie `leveneTest()`. Jeho modifikáciu (Brown – Forsythov test) je možné realizovať taktiež cez túto funkciu, ale je nutné nastaviť argument funkcie `center = median`.

```
> library(car)
> leveneTest(Pay, Region, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  8  1.3946  0.227
      42
-----------------------------------------------------------------
> leveneTest(Pay, Region, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  8  0.4881 0.8577
      42
```

Pri prvej skúmanej premennej (mzdy učiteľov) nemôžeme nulovú hypotézu zamietnuť ani pri jednom teste. Z dostupných údajov sa tak prikláňame k záveru, že variabilitu v mzdách učiteľov medzi jednotlivými regiónmi v USA je možné považovať za rovnakú. Ak sa pozrieme na grafickú vizualizáciu dát (Obrázok 7.16 a Obrázok 7.17), tak väčšie rozdiely vo variabilite môžeme pozorovať pri druhej skúmanej premennej (výdavky na študentov).

```
> library(car)
> leveneTest(Spend, Region, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value   Pr(>F)
group  8  2.9868 0.009634 **
      42
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-----------------------------------------------------------------
> leveneTest(Spend, Region, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  8  0.7766 0.6254
      42
> mu <- tapply(Pay, Region, mean)
> p10 <- tapply(Pay, Region, quantile, probs = 0.10)
> p25 <- tapply(Pay, Region, quantile, probs = 0.25)
> p50 <- tapply(Pay, Region, quantile, probs = 0.50)
> p75 <- tapply(Pay, Region, quantile, probs = 0.75)
> p90 <- tapply(Pay, Region, quantile, probs = 0.90)
> cbind(mu, p10, p25, p50, p75, p90)
          mu   p10    p25   p50    p75   p90
ENC 26.54000 24.38 24.500 26.50 27.200 29.00
ESC 21.00000 19.15 20.275 21.35 22.075 22.57
MA  27.93333 26.16 26.550 27.20 28.950 30.00
MN  24.21250 21.91 22.450 24.10 26.075 26.78
NE  23.85000 19.95 20.300 23.45 26.750 28.15
PA  29.64000 25.80 25.800 26.00 29.100 36.54
SA  24.28889 21.40 22.100 22.80 24.600 28.56
WNC 22.64286 19.72 20.850 21.70 24.700 27.48
WSC 21.65000 19.80 20.250 20.95 22.350 24.06
```

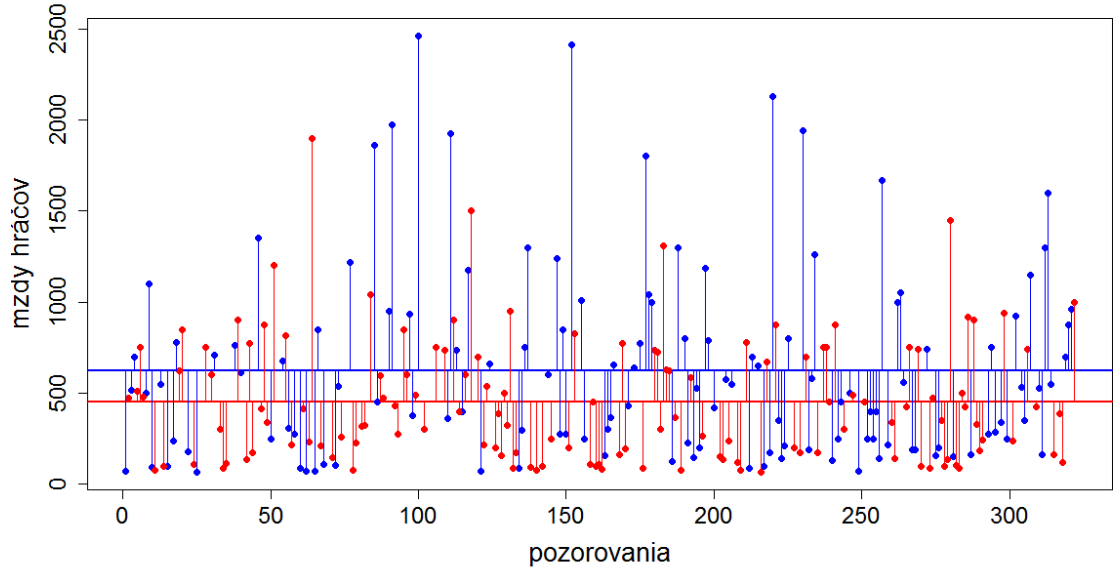
V prípade výdavkov na študentov dostávame konfliktné výsledky. Preto sme si chceli porovnať mieru šikmosti nameraných hodnôt a zostrojili sme tabuľku s príslušnými percentilmi. Z nej je možné vyčítať, že nepomerne väčšie rozdiely bývajú medzi rozdielom mediánu a 90-tym percentilom a rozdielom mediánu a 10-tym percentilom. To naznačuje pravostranné zošikmenie, v ktorom zrejme lepšou charakteristikou polohy je medián ako priemer. Ak by sme sa museli prikloniť k jednému z týchto testov, vybrali by sme si test založený na mediánoch.

Príklad 7.33

V tomto príklade máme overiť, či existujú rozdiely v mzdách hráčov podľa ich príslušnosti k divízii, pričom sa pokúsime aplikovať metódu ANOVA. Konkrétne pôjde o jednofaktorovú ANOVU s fixným faktorom, keďže máme len jeden faktor (divízia). Tento faktor (premenná `div86`) nadobúda len dve úrovne (východná divízia – E a západná divízia – W), a teda v zásade je postup obdobný ako v prípade *t*-testov alebo neparametrických testov na porovnávanie dvoch stredných hodnôt (resp. celých rozdelení). Účelom tohto príkladu je skôr precvičenie postupu. Metóda ANOVA je v softvéri R dostupná prostredníctvom funkcie `aov()`. Pred samotným výpočtom sa tradične pozrieme na vizuálnu prezentáciu dát.

```
> East <- c(); West <- c()
```

```
> for (i in 1:length(div86)) if(div86[i] == "E") East = c(East,
  i)
> for (i in 1:length(div86)) if(div86[i] == "W") West = c(West,
  i)
> plot(1:length(sal87), type = "n", ylim =
  c(min(na.omit(sal87)), max(na.omit(sal87))), ylab = "mzdy
  hráčov", xlab = "pozorovania", cex.lab = 1.7, cex.axis = 1.5)
> abline(h = mean(na.omit(sal87[div86 == "E"])), lwd = 2, col =
  "blue")
> abline(h = mean(na.omit(sal87[div86 == "W"])), lwd = 2, col =
  "red")
> points(East, sal87[div86 == "E"], pch = 19, col = "blue")
> points(West, sal87[div86 == "W"], pch = 19, col = "red")
> for (i in East) lines(c(i, i), c(mean(na.omit(sal87[div86 ==
  "E"])), sal87[i]), col = "blue")
> for (i in West) lines(c(i, i), c(mean(na.omit(sal87[div86 ==
  "W"])), sal87[i]), col = "red")
```



Obrázok 7.21: Graf mzdy hráčov v závislosti od divízie

Zdroj: vlastné spracovanie, výstup zo softvéru R

Do *x-y* grafu sme naniesli aj priemerné mzdy v rámci dvoch úrovní faktora, ktorého efekt sledujeme. Jednotlivé body predstavujú naše pozorovania/hráčov a každé pozorovanie je spojené s priemerom podľa príslušnosti hráča k danej divízii (modrou farbou je označená východná divízia a červenou farbou západná). Z takto zostrojeného grafu môžeme na prvý pohľad vidieť, že priemerné mzdy v jednotlivých divíziách sú odlišné. Taktiež môžeme pozorovať, že: 1) spravidla modré bodky sa nachádzajú nad červenými, 2) že vo východnej divízii je variabilita miezd väčšia. Pozrime sa, či metóda ANOVA vyhodnotí rozdiely v mzdách hráčov podľa divízie ako štatisticky významné.

```
> summary(aov(sal87~div86))
```

```
             Df   Sum Sq Mean Sq F value  Pr(>F)
div86         1  1976102 1976102   10.04 0.00171 **
Residuals   261 51343011  196717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
59 observations deleted due to missingness
```

Nulovú hypotézu o rovnosti stredných hodnôt môžeme zamietnuť na hladine významnosti 1 %. Rozdiely v mzdách hráčov z východnej a západnej divízie sa teda javia ako štatisticky významné.

Jednofaktorová ANOVA v prípade faktora s dvoma úrovňami poskytuje rovnaké výsledky ako dosiahneme pri použití *t*-testu na zhodu stredných hodnôt dvoch nezávislých súborov (uvedené je možné aj formálne ukázať). Rozdelíme si preto mzdy hráčov do dvoch skupín (podľa divízie) a toto tvrdenie overíme s využitím funkcie `t.test()`. Argument funkcie týkajúci sa rovnosti rozptylov nastavíme na `var.equal = T`, keďže ide o predpoklad metódy ANOVA (tento predpoklad o rovnosti rozptylov medzi súbormi overíme neskôr).

```
> East <- na.omit(subset(sal87, subset = div86 == "E"))
> West <- na.omit(subset(sal87, subset = div86 == "W"))
-------------------------------------------------------------
> t.test(East, West, alternative = "two.sided", mu = 0,
  var.equal = T, conf.level = 0.95)

                     Two Sample t-test

data:  East and West
t = 3.1695, df = 261, p-value = 0.001709
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
65.66921 281.11977
sample estimates:
mean of x mean of y
624.2714  450.8769
```

Môžeme vidieť, že výsledky sú naozaj rovnaké. Ak si porovnáme výsledky aj s neparametrickými testami, tak rozhodnutie o zamietnutí nulovej hypotézy bolo vo všetkých aplikovaných testoch totožné. Nevýhodou metódy ANOVA sú však do značnej miery obmedzujúce predpoklady, najmä teda predpoklad o normalite a rovnosti rozptylov (výhodou je, že ak máme dôvod domnievať sa, že predpoklady sú dodržané, metóda ANOVA poskytuje menšiu chybu II. druhu ako neparametrické testy). Na záver otestujeme ešte tieto predpoklady o vhodnosti použitia metódy ANOVA.

Na testovanie normality využijeme tri testy (s ktorými sme už pracovali v predchádzajúcich príkladoch), a to Anderson – Darlingov (funkcia `ad.test()` z knižnice `nortest`), Shapiro – Wilkov (funkcia `shapiro.test()` z knižnice `stats`) a Jarque – Berov test (funkcia `rjb.test()` z knižnice `lawstat`). Pri Jarque – Berovom teste budeme vychádzať z empirických (simulovaných) kritických hodnôt a použijeme len jeho modifikovanú verziu (`option = "RJB"`).

```
> library(nortest)
> ad.test(East)
                 Anderson-Darling normality test
data:  East
A = 4.8798, p-value = 3.851e-12

> ad.test(West)
                 Anderson-Darling normality test
data:  West
A = 3.6204, p-value = 4.378e-09
------------------------------------------------------------------
> library(stats)
> shapiro.test(East)
                   Shapiro-Wilk normality test
data:  East
W = 0.8608, p-value = 1.155e-09

> shapiro.test(West)
                   Shapiro-Wilk normality test
data:  West
W = 0.888, p-value = 1.278e-08
------------------------------------------------------------------
> library(lawstat)
> rjb.test(East, option = "RJB", crit.values = "empirical", N =
  1000)
                    Robust Jarque Bera Test
data:  East
X-squared = 83.4598, df = 2, p-value = 0.0007728

> rjb.test(West, option = "RJB", crit.values = "empirical", N =
  1000)
                    Robust Jarque Bera Test
data:  West
X-squared = 42.9384, df = 2, p-value < 2.2e-16
```

Nulovú hypotézu o normálnom rozdelení môžeme zamietnuť pri všetkých použitých testoch. Už porušením tohto predpokladu sa stávajú výsledky metódy ANOVA spochybniteľné. Len pre úplnosť ešte otestujeme aj predpoklad o zhode variability prostredníctvom Levenovho a Brown – Forsythovho testu.

```
> library(car)
> leveneTest(sal87, div86, center = mean)
```

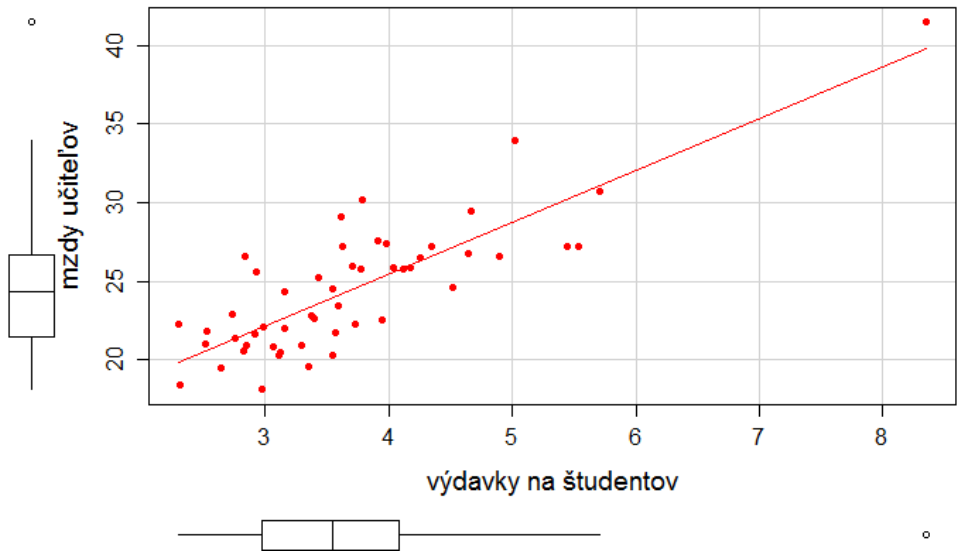
```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)
group   1  14.299 0.0001934 ***
      261
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-------------------------------------------------------------
> leveneTest(sal87, div86, center = median)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   1  10.503 0.001347 **
      261
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Aj predpoklad o konštantných rozptyloch je porušený, keďže na základe výsledkov z Levenovho testu a aj jeho modifikácie (Brown – Forsythovho testu) môžeme zamietnuť nulovú hypotézu o zhode smerodajných odchýlok medzi mzdami hráčov z východnej divízie a mzdami hráčov zo západnej divízie.

Príklad 7.34

Za účelom zistenia existencie lineárnej závislosti medzi mzdami učiteľov (premenná `prijem`) a výdavkami na študentov (premenná `vydavky`) v jednotlivých štátoch USA využijeme Pearsonov korelačný koeficient. Najprv je však vždy vhodné dáta vizualizovať (ako sme už viackrát spomenuli, minimálne je to vhodné kvôli výskytu prípadných odľahlých hodnôt). V tomto prípade, keďže ideme skúmať závislosť dvoch podielových premenných, sa ako vhodná forma vizualizácie javí *x-y* graf. Pre jeho zostrojenie využijeme funkciu `scatterplot()` z knižnice `car`.

```
> data <- read.csv(file = "...cesta
  k súboru...\\educ_spending.csv", sep = ";", dec = ".", header
  = T)
> attach(data)
-------------------------------------------------------------
> library(car)
> scatterplot(Pay ~ Spend, smooth = F, xlab = "výdavky na
  študentov", ylab = "mzdy učiteľov", col = "red", pch = 19,
  cex.lab = 1.4, cex.axis = 1.2)
```



Obrázok 7.22: Graf závislosti mzdy učiteľov a výdavkov na študentov

Zdroj: vlastné spracovanie, výstup zo softvéru R

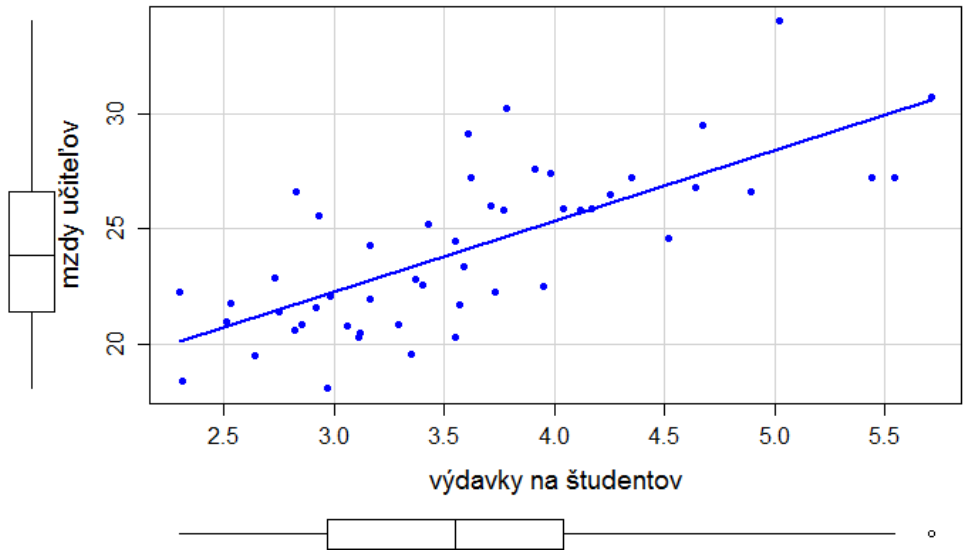
Z uvedeného obrázku (Obrázok 7.22) je zjavne viditeľná lineárna závislosť medzi mzdami učiteľov a výdavkami na študentov. Taktiež môžeme pozorovať výskyt extrémnej hodnoty (bez exaktného testovania je minimálne táto jedna hodnota evidentne odľahlá). Ide o posledné pozorovanie v danej databáze, konkrétne rozlohou najväčší štát z USA, Aljašku. Korelačný koeficient najprv vypočítame za všetky údaje.

```
> cor.test(Pay, Spend)

            Pearson's product-moment correlation

data:  Pay and Spend
t = 10.2824, df = 49, p-value = 7.927e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7136105 0.8977100
sample estimates:
      cor
0.8266297
```

Korelačný koeficient je vo výške 0.8266, čo naznačuje vysokú lineárnu závislosť medzi skúmanými premennými. Tento korelačný koeficient je štatisticky významný, keďže nulovú hypotézu $\rho = 0$ môžeme zamietnuť pri p-hodnote výrazne nižšej ako 0.01 (na testovanie hypotézy sa využíva obojstranný t-test). Uvedené je možné vidieť aj z toho, že konfidenčný interval nezahŕňa 0. Pozrime sa, ako sa zmení vypočítaný korelačný koeficient po odstránení odľahlej hodnoty (identifikovanej len na základe vizualizácie dát).

```
> Pay_out <- Pay[-51]
> Spend_out <- Spend[-51]
```

```
> scatterplot(Pay_out ~ Spend_out, lwd = 2, smooth = F, xlab =
  "výdavky na študentov", ylab = "mzdy učiteľov", col = "blue",
  pch = 19, cex.lab = 1.4, cex.axis = 1.2)
```



Obrázok 7.23: Graf závislosti mzdy učiteľov a výdavkov na študentov (bez extrému)

Zdroj: vlastné spracovanie, výstup zo softvéru R

Po odstránení jednej extrémnej hodnoty sa znížil korelačný koeficient na úroveň 0.7303. Stále však ide o silnú lineárnu závislosť a je zrejmé, že pokiaľ sú v jednotlivých štátoch vyššie výdavky na študentov, rovnako sú vyššie aj mzdy učiteľov (a vice versa). Bolo by možné pokračovať aj odstránením ďalších hodnôt identifikovaných ako extrémne, napr. s použitím Hampelovho testu. Tento krok už ale necháme na čitateľa.

```
> cor.test(Pay_out, Spend_out)

            Pearson's product-moment correlation

data:  Pay_out and Spend_out
t = 7.4074, df = 48, p-value = 1.75e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.5673186 0.8382742
sample estimates:
      cor
0.7303374
```

V závere sa pokúsime overiť, nakoľko je možné namerané hodnoty považovať za realizácie z normálneho rozdelenia. Pripomíname, že k tomu, aby vektor náhodných premenných X a Y predstavoval realizácie z dvojrozmerného normálneho rozdelenia pravdepodobnosti nestačí, aby boli oba súbory z normálneho rozdelenia (je to nutná, nie postačujúca podmienka).

Príklad 7.35

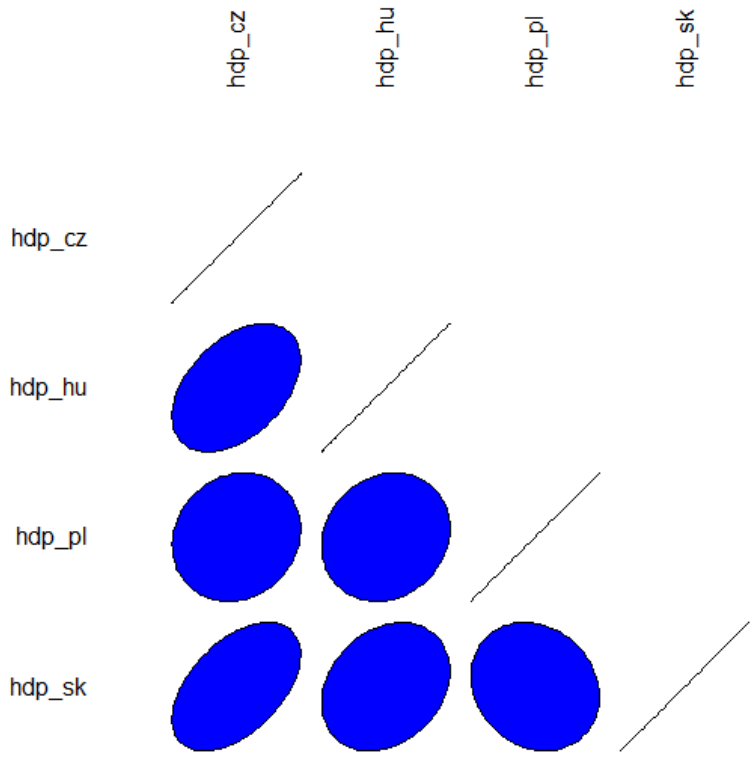
Pri výpočte korelácií medzi viacerými premennými je vhodné využiť funkciu `cor()`, ktorá vráti korelačnú maticu (avšak bez štatistickej významnosti korelačných koeficientov). Pre naše účely sme z danej databázy (zo súboru `granger_gdp.csv`) brali do úvahy len údaje za jednotlivé HDP (údaje za akciové indexy využijeme neskôr).

```
> data <- read.csv(file = "...cesta
  k súboru...\\granger_gdp.csv", sep = ";", dec = ".", header =
  T)
> attach(data)
> gdp <- data.frame(hdp_cz, hdp_hu, hdp_pl, hdp_sk)
------------------------------------------------------------
> correl <- round(cor(gdp), 2); correl
       hdp_cz hdp_hu hdp_pl hdp_sk
hdp_cz   1.00   0.44   0.12   0.50
hdp_hu   0.44   1.00   0.14   0.25
hdp_pl   0.12   0.14   1.00  -0.16
hdp_sk   0.50   0.25  -0.16   1.00
```

Najvyššia korelácia je zaznamenaná medzi vývojom slovenského a českého HDP (0.50) a medzi vývojom českého a maďarského HDP (0.44). Ostatné korelácie sú výrazne nižšie a v prípade slovenského a poľského HDP je dosiahnutá dokonca záporná korelácia. To v žiadnom prípade neznamená, že by medzi týmito ekonomikami neexistoval vzťah v ich vývoji. Do úvahy sme však brali iba vývoj v tom istom období, pričom v skutočnosti presnejšia analýza by musela brať do úvahy spoločný vývoj v širšom spektre ako len v jednom kvartály.

Ak by sme pracovali s väčším počtom premenných, korelačná matica sa môže stať menej prehľadnou. V softvéri R však máme viacero možností ako korelačnú maticu vizualizovať do prehľadnejšej podoby. Jedným zo spôsobov je využiť funkciu `plotcorr()` z knižnice `ellipse`. Táto funkcia zobrazuje prvky korelačnej matice (korelačné koeficienty) vo forme elíps, pričom čím je korelácia medzi dvoma premennými vyššia, tým je elipsa užšia. Taktiež smer elipsy poukazuje na smer lineárnej závislosti. Pri korelácii rovnej 1 dostávame vo vizualizácii priamku (ktorej výskyt môžeme vidieť na hlavnej diagonále).

```
> library(ellipse)
> plotcorr(correl, diag = TRUE, col = "blue", type = "lower")
```



Obrázok 7.24: Vizualizácia korelačnej matice pomocou `plotcorr()`

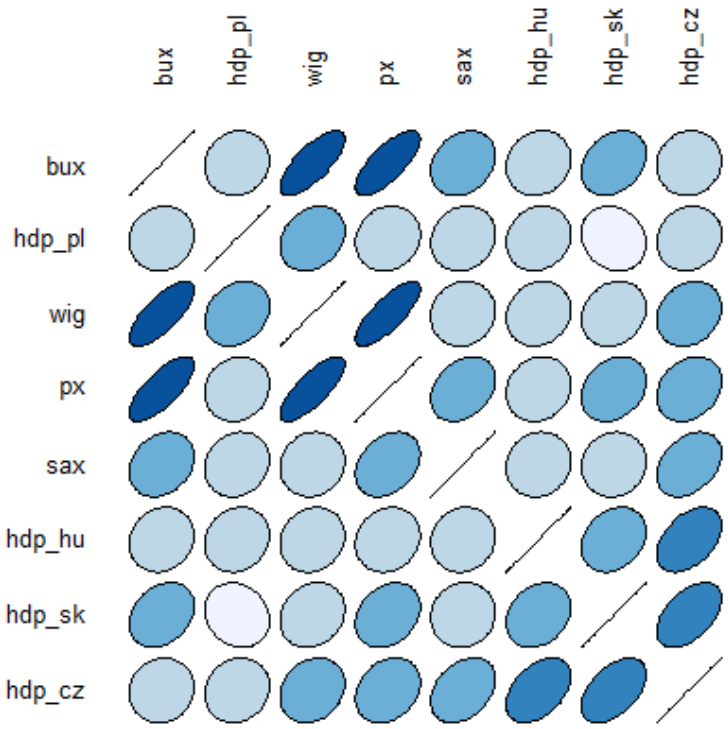
Zdroj: vlastné spracovanie, výstup zo softvéru R

Menšou úpravou sa vieme dopracovať k takej vizualizácii korelačnej matice, ktorá bude rôznej výške korelačných koeficientov priraďovať aj rôzne farby[36]. Tentoraz vypočítame korelácie medzi všetkými premennými v databáze, čiže aj pre akciové indexy z krajín V4. Urobíme tak len z dôvodu, že chceme demonštrovať výhody tejto formy vizualizácie korelačnej matice na väčšom počte premenných.

```
> correl <- round(cor(data), 2); correl
       hdp_cz   px hdp_hu  bux hdp_pl  wig hdp_sk  sax
hdp_cz   1.00 0.25   0.44 0.08   0.12 0.20   0.50 0.35
px       0.25 1.00   0.14 0.74   0.12 0.78   0.26 0.31
hdp_hu   0.44 0.14   1.00 0.13   0.14 0.14   0.25 0.10
bux      0.08 0.74   0.13 1.00   0.09 0.74   0.29 0.24
hdp_pl   0.12 0.12   0.14 0.09   1.00 0.24  -0.16 0.14
wig      0.20 0.78   0.14 0.74   0.24 1.00   0.18 0.10
hdp_sk   0.50 0.26   0.25 0.29  -0.16 0.18   1.00 0.08
sax      0.35 0.31   0.10 0.24   0.14 0.10   0.08 1.00
-----------------------------------------------------------------
> ord <- order(correl[1,])
> xc <- correl[ord, ord]
```

[36] Kód na túto úpravu je možné získať priamo z pomocných stránok ku knižnici `ellipse` (v softvéri R po použití príkazu `?plotcorr`): `http://127.0.0.1:27099/library/ellipse/html/plotcorr.html`.

```
> colors <- c("#A50F15", "#DE2D26", "#FB6A4A", "#FCAE91",
  "#FEE5D9", "white", "#EFF3FF", "#BDD7E7", "#6BAED6",
  "#3182BD", "#08519C")
> plotcorr(xc, col = colors[5*xc + 8], cex.lab = 0.8)
```



Obrázok 7.25: Vizualizácia korelačnej matice – rôzne farby

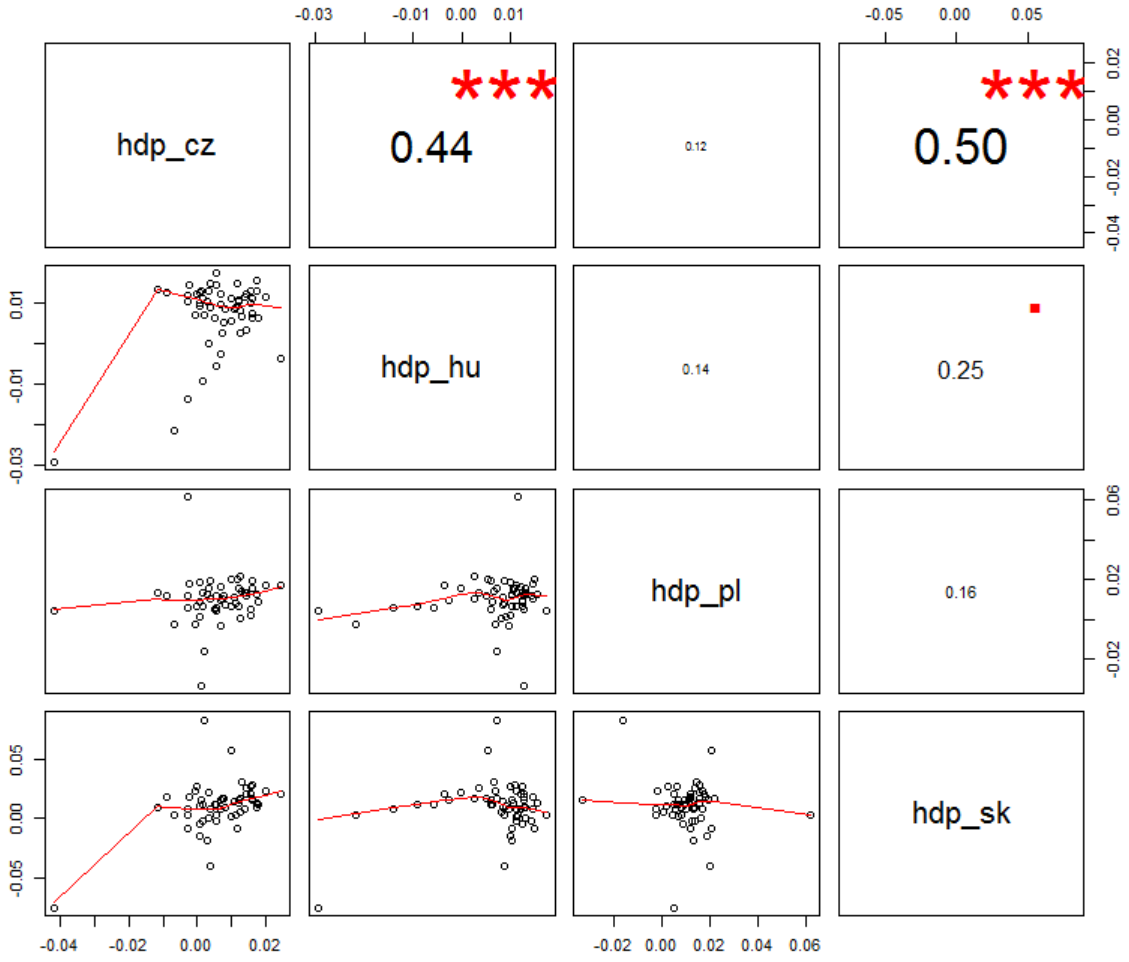
Zdroj: vlastné spracovanie, výstup zo softvéru R

Posledná forma vizualizácie korelačnej matice, ktorú si ukážeme, využíva funkciu `panel.cor()`[37]. Pri tejto vizualizácii ide o najkomplexnejší pohľad na skúmané lineárne závislosti medzi rôznym počtom premenných. Prostredníctvom tejto funkcie dostávame pod hlavnou diagonálou matice *x*-*y* grafy, nad hlavnou diagonálou získame vypočítané korelačné koeficienty (pričom veľkosť písma zobrazuje rôzne úrovne korelácie) a taktiež vo výstupe môžeme vidieť štatistickú významnosť vypočítaných korelačných koeficientov. Významnosť koeficientov je počítaná klasicky cez *t*-test, teda prostredníctvom funkcie `cor.test()`. Ak by nás zaujímali významnosti, tie sú zobrazené štandardne vo forme hviezdičiek (viď samotný kód funkcie).

```
> panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
+ {
+     usr <- par("usr"); on.exit(par(usr))
+     par(usr = c(0, 1, 0, 1))
```

[37] Táto funkcia je dostupná na:
`http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=137`

```
+      r <- abs(cor(x, y))
+      txt <- format(c(r, 0.123456789), digits=digits)[1]
+      txt <- paste(prefix, txt, sep="")
+      if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
+
+      test <- cor.test(x,y)
+      Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
+                       cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
+                       symbols = c("***", "**", "*", ".", " "))
+
+      text(0.5, 0.5, txt, cex = cex * r)
+      text(.8, .8, Signif, cex=cex, col=2)
+ }
> pairs(gdp, lower.panel=panel.smooth, upper.panel=panel.cor)
```



Obrázok 7.26: Vizualizácia závislosti medzi premennými

Zdroj: zdrojový kód prevzatý z http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=137

Príklad 7.36

V tomto príklade budeme opäť pracovať s databázou o baseballových hráčoch (knižnica `vcd`, databáza `Baseball`). Urobíme si prehľad dát (funkcia `head`) a vytvoríme si novú databázu.

```
> library(vcd)
> attach(Baseball)
> names(Baseball)
 [1] "name1"    "name2"    "atbat86"  "hits86"   "homer86"
  "runs86"
 [7] "rbi86"    "walks86"  "years"    "atbat"    "hits"
  "homeruns"
[13] "runs"     "rbi"      "walks"    "league86" "div86"
  "team86"
[19] "posit86"  "outs86"   "assist86" "error86"  "sal87"
  "league87"
[25] "team87"
> head(Baseball[,c(3:15, 20:23)])
  .....
> data <- data.frame(Baseball[,c(3:15, 20:23)])
```

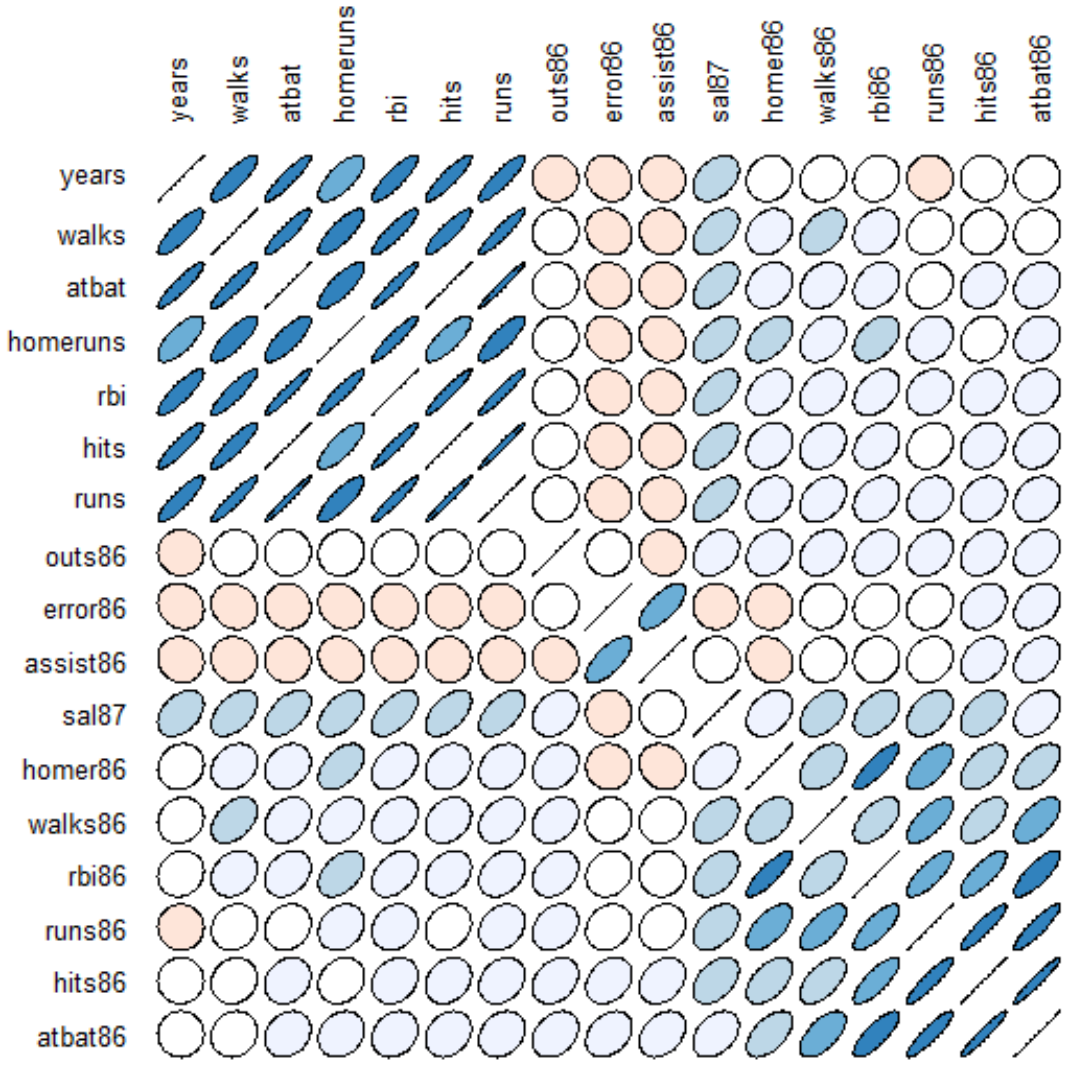
Zo zadania vyplýva, že máme vypočítať korelačné koeficienty medzi väčším počtom premenných a korelačnú maticu vizualizovať nejakým vhodným spôsobom. Ako sme už spomínali vyššie, takáto vizualizácia môže pri väčšom počte premenných výrazne pomôcť pri identifikácii vzájomných vzťahov medzi nimi. Na výpočet Pearsonovho korelačného koeficientu opäť využijeme funkciu `cor()`, v ktorej nastavíme parameter `use = "complete.obs"`. Niektoré premenné obsahujú chýbajúce pozorovania a toto je jeden zo spôsobov, ktorým zabezpečíme, že pre každý pár premenných budú do výpočtu vstupovať len kompletné údaje za hráčov.

```
> correl <- round(cor(data, use = "complete.obs"), 2); correl

         atbat86 hits86 homer86 runs86 rbi86 walks86 years
atbat86     1.00   0.96    0.56   0.90  0.80    0.62  0.01
hits86      0.96   1.00    0.53   0.91  0.79    0.59  0.02
homer86     0.56   0.53    1.00   0.63  0.85    0.44  0.11
runs86      0.90   0.91    0.63   1.00  0.78    0.70 -0.01
rbi86       0.80   0.79    0.85   0.78  1.00    0.57  0.13
walks86     0.62   0.59    0.44   0.70  0.57    1.00  0.13
years       0.01   0.02    0.11  -0.01  0.13    0.13  1.00
atbat       0.21   0.21    0.22   0.17  0.28    0.27  0.92
hits        0.23   0.24    0.22   0.19  0.29    0.27  0.90
homeruns    0.21   0.19    0.49   0.23  0.44    0.35  0.72
runs        0.24   0.24    0.26   0.24  0.31    0.33  0.88
rbi         0.22   0.22    0.35   0.20  0.39    0.31  0.86
walks       0.13   0.12    0.23   0.16  0.23    0.43  0.84
outs86      0.31   0.30    0.25   0.27  0.31    0.28 -0.02
assist86    0.34   0.30   -0.16   0.18  0.06    0.10 -0.09
```

```
error86     0.33   0.28   -0.01   0.19  0.15    0.08 -0.16
sal87       0.39   0.44    0.34   0.42  0.45    0.44  0.40
          atbat  hits homeruns  runs   rbi walks outs86 assist86
atbat86    0.21  0.23     0.21  0.24  0.22  0.13   0.31     0.34
hits86     0.21  0.24     0.19  0.24  0.22  0.12   0.30     0.30
homer86    0.22  0.22     0.49  0.26  0.35  0.23   0.25    -0.16
runs86     0.17  0.19     0.23  0.24  0.20  0.16   0.27     0.18
rbi86      0.28  0.29     0.44  0.31  0.39  0.23   0.31     0.06
walks86    0.27  0.27     0.35  0.33  0.31  0.43   0.28     0.10
years      0.92  0.90     0.72  0.88  0.86  0.84  -0.02    -0.09
atbat      1.00  1.00     0.80  0.98  0.95  0.91   0.05    -0.01
hits       1.00  1.00     0.79  0.98  0.95  0.89   0.07    -0.01
homeruns   0.80  0.79     1.00  0.83  0.93  0.81   0.09    -0.19
runs       0.98  0.98     0.83  1.00  0.95  0.93   0.06    -0.04
rbi        0.95  0.95     0.93  0.95  1.00  0.89   0.10    -0.10
walks      0.91  0.89     0.81  0.93  0.89  1.00   0.06    -0.07
outs86     0.05  0.07     0.09  0.06  0.10  0.06   1.00    -0.04
assist86  -0.01 -0.01    -0.19 -0.04 -0.10 -0.07  -0.04     1.00
error86   -0.07 -0.07    -0.17 -0.09 -0.12 -0.13   0.08     0.70
sal87      0.53  0.55     0.52  0.56  0.57  0.49   0.30     0.03
          error86 sal87
atbat86      0.33  0.39
hits86       0.28  0.44
homer86     -0.01  0.34
runs86       0.19  0.42
rbi86        0.15  0.45
walks86      0.08  0.44
years       -0.16  0.40
atbat       -0.07  0.53
hits        -0.07  0.55
homeruns    -0.17  0.52
runs        -0.09  0.56
rbi         -0.12  0.57
walks       -0.13  0.49
outs86       0.08  0.30
assist86     0.70  0.03
error86      1.00 -0.01
sal87       -0.01  1.00
```

Pri väčšom počte pozorovaní je korelačná matica do značnej miery neprehľadná. Z tohto dôvodu využijeme jeden zo spôsobov jej vizualizácie. Konkrétne sa ako vhodná voľba javí byť vizualizácia korelačnej matice vo forme elíps.

```
> library(ellipse)
> ord <- order(correl[1,])
> xc <- correl[ord, ord]
> colors <- c("#A50F15", "#DE2D26", "#FB6A4A", "#FCAE91",
  "#FEE5D9", "white", "#EFF3FF", "#BDD7E7", "#6BAED6",
  "#3182BD", "#08519C")
> plotcorr(xc, col = colors[5*xc + 6], cex.lab = 0.8)
```



Obrázok 7.27: Vizualizácia korelačnej matice

Zdroj: vlastné spracovanie, výstup zo softvéru R

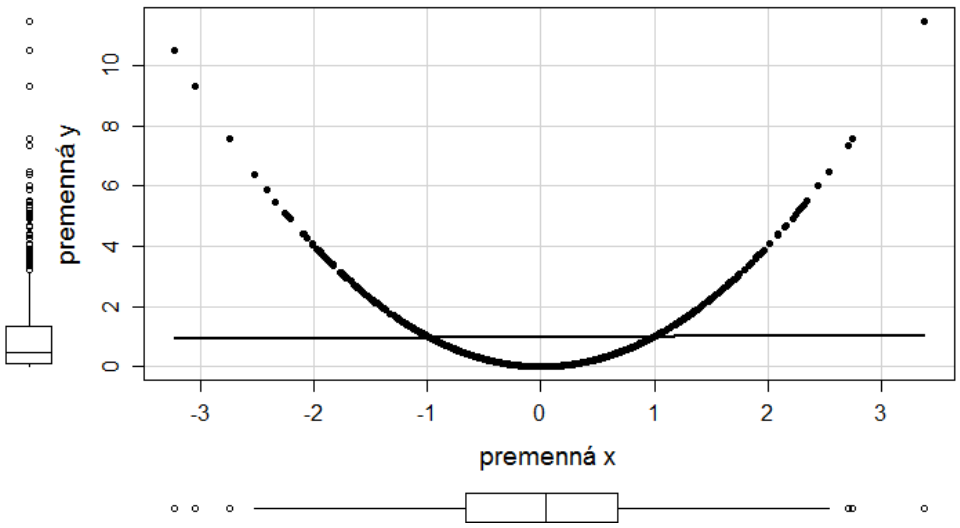
Z takejto formy korelačnej matice môžeme na prvý pohľad vidieť, medzi ktorými premennými je korelácia vyššia (a prípadne stojí za bližšie skúmanie), resp. medzi ktorými premennými je korelácia nižšia. Do istej miery je zaujímavá korelácia premenných so mzdou hráčov (`sal87`). Výkonnostné charakteristiky hráčov korelujú s ich mzdou len do istej miery, keďže korelačné koeficienty dosahujú úroveň približne 0.5.

Príklad 7.37

V tomto príklade máme zistiť vzťah medzi dvoma premennými, ktoré si vygenerujeme. Premennú x vygenerujeme ako realizáciu z normovaného normálneho rozdelenia (teda so strednou hodnotou 0 a rozptylom 1). K tejto premennej vygenerujeme deterministicky premennú y. Cieľom tohto príkladu je ukázať rozdiel medzi Pearsonovým

a Spearmanovým korelačným koeficientom pri odhaľovaní závislosti medzi dvoma premennými. V prvom prípade si premennú y vygenerujeme ako kvadratickú funkciu $y = x^2$.

```
> P_x <- rnorm(1000, mean = 0, sd = 1)
> P_y <- (P_x)^2
------------------------------------------------------------------
> library(car)
> scatterplot(P_y ~ P_x, smooth = F, col = "black", xlab =
  "premenná x", ylab = "premenná y", lwd = 2, lty = 1, pch = 19,
  cex.lab = 1.4, cex.axis = 1.2)
```



Obrázok 7.28: Graf závislosti premennej generovanej ako $y = x^2$

Zdroj: vlastné spracovanie, výstup zo softvéru R

Vzťah medzi takto vygenerovanými premennými zobrazuje Obrázok 7.28. Je úplne zrejmé, že nejde o lineárnu závislosť, avšak nejde ani o monotónnu závislosť. Pearsonov a Spearmanov korelačný koeficient by nám pri odhalení závislosti medzi takýmito dvoma premennými nemuseli pomôcť. Problém spočíva v tom, že výsledky z týchto dvoch testov budú v tomto prípade výrazne skreslené, avšak v empirickej analýze si to nemáme ako overiť. Pri používaní týchto dvoch koeficientov totiž predpokladáme, že ak existuje vzťah medzi premennými, potom by mal byť lineárny (monotónny). V našom prípade medzi premennými vzťah existuje, ale nie takého druhu, aký vieme pomocou týchto koeficientov zmerať. V našom príklade sme použitím Pearsonovho korelačného koeficientu dosiahli významnú negatívnu závislosť. Môže sa teda stať, že si budeme (na vzorke $n = 1000$) pomerne výrazne istí, že sa nám podarilo nájsť určitý prejav lineárnej nepriamej závislosti medzi premennými.

```
> cor.test(P_x, P_y, method = "pearson", alternative =
  c("two.sided"))

            Pearson's product-moment correlation
```

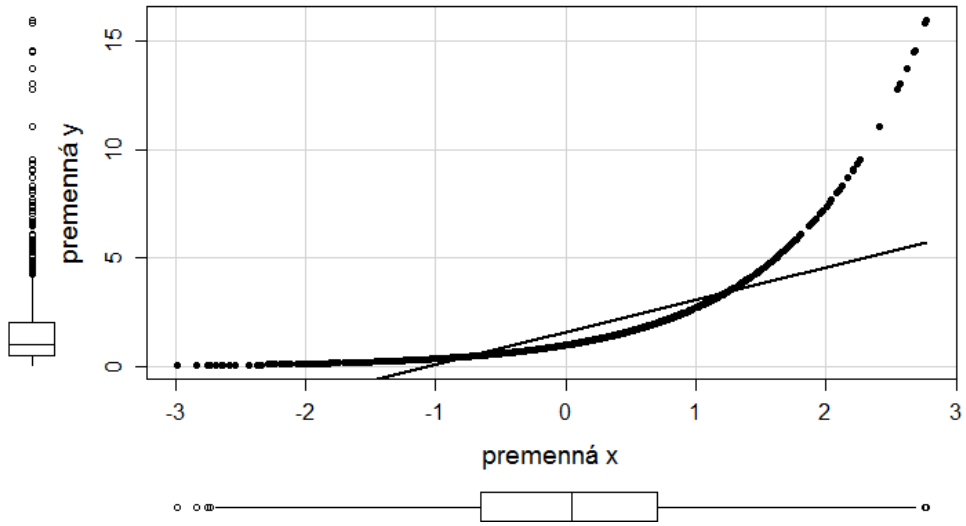
```
data:  P_x and P_y
t = -3.1768, df = 998, p-value = 0.001535
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.16104863 -0.03829895
sample estimates:
       cor
-0.1000545
-----------------------------------------------------------------
> cor.test(P_x, P_y, method = "spearman", alternative =
  c("two.sided"))

              Spearman's rank correlation rho

data:  P_x and P_y
S = 168919114, p-value = 0.6694
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.0135157
```

Výsledky z oboch korelačných koeficientov sú podobné a poukazujú len na veľmi nízku (lineárnu resp. monotónnu) závislosť. Keďže poznáme skutočný vzťah medzi týmito premennými, tak môžeme povedať, že výsledky **sú zavádzajúce**. Je zrejmé, že takýto „prekvapujúci“ výsledok mohol byť dôsledkom náhodne vygenerovaných hodnôt. Preto sme celý pokus zopakovali spolu 10000 krát a vypočítali si podiel prípadov, v ktorých *p*-hodnota bola väčšia ako 0.05 (t. j. správne nezamietnutie nulovej hypotézy, keďže medzi premennými nie je lineárny ani monotónny vzťah). Celkový podiel správne nezamietnutých nulových hypotéz nebol tak veľký ako by sme čakali, iba 62.54 % prípadov pri Pearsonovom teste a 85.94 % prípadov pri Spearmanovom teste. Jednoduchým spôsobom ako sa vyhnúť tomuto záveru (z empirického hľadiska nesprávnemu) je vizualizácia dát.

```
> pvalsI <- c(); pvalsII <- c();
> for (i in 1:10000) {
+ P_x <- rnorm(1000, mean = 0, sd = 1)
+ P_y <- (P_x)^2
+ a <- cor.test(P_x, P_y, method = "pearson", alternative =
  c("two.sided"))
+ b <- cor.test(P_x, P_y, method = "spearman", alternative =
  c("two.sided"))
+ pvalsI[i] <- a$p.val
+ pvalsII[i] <- b$p.val
+ }
> sum(pvalsI >= 0.05)/10000; sum(pvalsII >= 0.05)/10000
[1] 0.6254
[1] 0.8594
```

Ak si premennú *y* vytvoríme druhým spôsobom, teda ako $y = e^x$, situácia by už mohla byť odlišná. Ako je zrejmé aj z nasledujúceho obrázku (Obrázok 7.29), tak v tomto prípade ide o monotónnu závislosť.

```
> P_x <- rnorm(1000,0,1)
> P_y <- exp(P_x)
-----------------------------------------------------------------
> library(car)
> scatterplot(P_y ~ P_x, smooth = F, col = "black", xlab =
  "premenná x", ylab = "premenná y", lwd = 2, lty = 1, pch = 19,
  cex.lab = 1.4, cex.axis = 1.2)
```



Obrázok 7.29: Graf monotónnej závislosti premennej *x* a *y*

Zdroj: vlastné spracovanie, výstup zo softvéru R

Pearsonov korelačný koeficient stále nebude vhodnou voľbou, keďže prostredníctvom neho vieme kvantifikovať len lineárnu závislosť, ktorá úplne zrejme nebude dobre popisovať daný typ vzťahu medzi vygenerovanými premennými (pre lepšiu predstavu sú v grafoch zobrazené aj lineárne trendy). Pozrime sa, aká je odhadovaná závislosť medzi premennými s použitím Pearsonovho a Spearmanovho korelačného koeficientu.

```
> cor.test(P_x, P_y, method = "pearson")

            Pearson's product-moment correlation

data:  P_x and P_y
t = 33.9883, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7023643 0.7599501
sample estimates:
      cor
0.7324646
```

```
-----------------------------------------------------------------
> cor.test(P_x, P_y, method = "spearman")

              Spearman's rank correlation rho

data:  P_x and P_y
S = 0, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

Pearsonov korelačný koeficient je vo výške 0.73 a je aj štatisticky významný. Naproti tomu Spearmanov korelačný koeficient určil presne úroveň závislosti a vypočítaný korelačný koeficient je rovný 1. Dôvod prečo nájdeme pomerne silnú lineárnu závislosť pomocou Pearsonovho korelačného koeficientu medzi premennými spočíva v tom, že na určitom intervale exponenciálna funkcia je celkom vhodne aproximovateľná lineárnou funkciou.

Je zrejmé, že Spearmanov korelačný koeficient zachytáva monotónnu závislosť, pričom lineárna je špeciálnym prípadom tejto závislosti. Porovnaním týchto dvoch koeficientov a doplnením pomocou vizualizácie je možné si urobiť presnejšiu predstavu o vzťahu medzi premennými.

V tomto príklade sme chceli poukázať na skutočnosť, že ak medzi premennými existuje iná ako lineárna závislosť (avšak monotónna), tak Pearsonov korelačný koeficient nezachytí túto závislosť správne. Spearmanov korelačný koeficient sa tak javí ako vhodný doplnok ku klasickej korelačnej analýze (realizovanej s využitím Pearsonovho korelačného koeficientu). Jednak z dôvodu, že je možné zachytiť širšie spektrum závislosti medzi premennými a nie len lineárne, ale aj z dôvodu, že pri lineárnej závislosti poskytuje porovnateľné výsledky ako Pearsonov korelačný koeficient. So zrýchlením výpočtovej techniky sa ešte vhodnejšou alternatívou k Pearsonovmu korelačnému koeficientu javí Kendall τ koeficient.

Príklad 7.38

Za účelom overenia vzájomnej závislosti medzi indexom *Ease of Doing Business* a ostatnými charakteristikami podnikateľského prostredia v krajinách využijeme (vzhľadom na charakter dát) Spearmanov a Kendallov korelačný koeficient.

```
> all <- read.csv(file = "...cesta k súboru...\\EoDB.csv", sep =
  ";", dec = ".", header = T)
> attach(all)
> data <- data.frame(Rank, Start, Proced, Time)
-----------------------------------------------------------------
```

```
> spearman <- cor(data, use = "everything", method =
  c("spearman")); spearman
           Rank     Start    Proced      Time
Rank   1.0000000 0.7351018 0.5515306 0.5644200
Start  0.7351018 1.0000000 0.8244014 0.7798218
Proced 0.5515306 0.8244014 1.0000000 0.7399849
Time   0.5644200 0.7798218 0.7399849 1.0000000
```

Pri Spearmanovom korelačnom koeficiente sa preukázal najsilnejší vzťah medzi indexom *Ease of Doing Business* a premennou `Start` (čo je tiež poradová premenná, ktorá hovorí o jednoduchosti založenia podniku v danej krajine). Zdá sa, že ak je v krajine pomerne ľahké založiť nový podnik, budú aj ostatné atribúty podnikateľského prostredia priaznivé a celkové hodnotenie bude tak pozitívne. Spomedzi skúmaných premenných je najvyššia korelácia medzi počtom procedúr potrebných na založenie podniku a počtom dní na to potrebných, čo nie až tak prekvapujúce. Evidentne čím viac procedúr je nutné vykonať na založenie podniku, tým viac dní to aj potrvá (a vice versa).

Pri využití Kendallovho korelačného koeficientu sú výsledky obdobné.

```
> kendall <- cor(data, use = "everything", method =
  c("kendall")); kendall
           Rank     Start    Proced      Time
Rank   1.0000000 0.5442476 0.4019737 0.3984295
Start  0.5442476 1.0000000 0.6513795 0.5891683
Proced 0.4019737 0.6513795 1.0000000 0.5666390
Time   0.3984295 0.5891683 0.5666390 1.0000000
```

Získané korelačné matice si pre zopakovanie môžeme aj vizualizovať.

```
> par(mfrow = c(1,2))
> library(ellipse)
> ord <- order(spearman[1,])
> xc <- correl[ord, ord]
> colors <- c("#A50F15", "#DE2D26", "#FB6A4A", "#FCAE91",
  "#FEE5D9", "white", "#EFF3FF", "#BDD7E7", "#6BAED6",
  "#3182BD", "#08519C")
> plotcorr(xc, col = colors[5*xc + 6], cex.lab = 0.8)
------------------------------------------------------------
> library(ellipse)
> ord <- order(kendall[1,])
> xc <- correl[ord, ord]
> colors <- c("#A50F15", "#DE2D26", "#FB6A4A", "#FCAE91",
  "#FEE5D9", "white", "#EFF3FF", "#BDD7E7", "#6BAED6",
  "#3182BD", "#08519C")
> plotcorr(xc, col = colors[5*xc + 6], cex.lab = 0.8)
```

Obrázok 7.30: Vizualizácia korelačných matíc (Spearman – vľavo, Kendall – vpravo)

Zdroj: vlastné spracovanie, výstup zo softvéru R

Aj keď nás v tomto príklade významnosť korelačných koeficientov nezaujíma, využijeme tento príklad na prezentovanie funkcie `rcor.test()` z knižnice `ltm`.

```
> library(ltm)
> rcor.test(data, method = "spearman")

       Rank   Start  Proced Time
Rank    *****  0.735  0.552  0.564
Start  <0.001  *****  0.824  0.780
Proced <0.001 <0.001  *****  0.740
Time   <0.001 <0.001 <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

Výstupom z funkcie je matica, ktorá nad hlavnou diagonálou obsahuje korelačné koeficienty a pod hlavnou diagonálou sú *p*-hodnoty k *t*-testu. V našom prípade môžeme vidieť, že všetky vypočítané Spearmanove korelačné koeficienty sú štatisticky významné na hladine 1 %. Pre Kendallov korelačný koeficient stačí nastaviť argument funkcie `method = "kendall"`. Aj tieto korelačné koeficienty sú štatisticky významne rôzne od nuly.

```
> rcor.test(data, method = "kendall")
       Rank   Start  Proced Time
Rank    *****  0.544  0.402  0.398
Start  <0.001  *****  0.651  0.589
Proced <0.001 <0.001  *****  0.567
Time   <0.001 <0.001 <0.001  *****
upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

Príklad 7.39

Máme overiť hypotézu, že existuje závislosť medzi tým, že žena je jedným z vlastníkov podniku a žena je taktiež vo vedení podniku. Pracujeme so súborom `enerprise_survey.csv`, v ktorom máme dostatočný počet pozorovaní za krajiny východnej Európy a strednej Ázie. Najprv túto hypotézu overíme na celej vzorke a v závere príkladu na vzorke podnikov zo Slovenska.

V prvom kroku si zostrojíme kontingenčnú tabuľku.

```
> data <- read.csv(file = "...cesta
  k súboru...\\enterprise_survey.csv", sep = ";", dec = ".",
  header = T)
> attach(data)
-----------------------------------------------------------------
> kont <- table(ownership_W, management_W); kont
           management_W
ownership_W  ano neviem  nie
     ano    1772     13 2823
     neviem   35     12  210
     nie     409     12 6382
```

Na prvý pohľad nám početnosti v kontingenčnej tabuľke napovedajú, že závislosť medzi pohlavím vlastníka a manažmentu podniku zrejme existuje. Pridajme do tabuľky ešte marginálne početnosti prostredníctvom funkcie `addmargins()`.

```
> marg <- addmargins(kont); marg
           management_W
ownership_W   ano neviem   nie   Sum
     ano     1772     13  2823  4608
     neviem    35     12   210   257
     nie      409     12  6382  6803
     Sum     2216     37  9415 11668
```

Ak by neexistovala závislosť medzi pohlavím vlastníkov a vedením, tak pri náhodnom výbere podniku môžeme pravdepodobnosť, že vlastníkom a zároveň aj manažérom je žena, vypočítať ako 4608/11668 × 2214/11668 = 0.0749:

```
> marg[13]/marg[16] * marg[4]/marg[16]
[1] 0.0749
```

Marginálne početnosti nám slúžia na výpočet tzv. očakávaných početností. Ak by sme v našom príklade chceli vypočítať očakávanú početnosť pre prípad, že žena nevystupuje ani ako vlastník a ani ako manažér, môžeme postupovať nasledujúcim spôsobom.

```
> marg[12]*marg[15]/marg[16]
[1] 5489.394
```

Číslo 5489.394 nám hovorí aký počet podnikov, v ktorých žena nie je vo vlastníckej ani v manažérskej štruktúre, môžeme očakávať, ak medzi pohlavím vlastníkov a manažérov neexistuje závislosť. Túto očakávanú početnosť si môžeme vypočítať pre každý jeden prvok kontingenčnej tabuľky. Každopádne však výpočet očakávaných početností je jednoduchšie vypočítať v softvéri R prostredníctvom funkcie `chisq.test()`.

```
> chisq.test(kont, correct = F, simulate.p.value = TRUE, B =
  10000)$expected
                    management_W
ownership_W         ano     neviem       nie
     ano      875.15667 14.612273 3718.2311
     neviem    48.80974  0.814964  207.3753
     nie     1292.03360 21.572763 5489.3936
```

V texte sme uviedli, že očakávané početnosti by mali byť väčšie alebo rovné 5, avšak podľa odporúčaní (Finkelstein – Levin, 2001) je pri väčších kontingenčných tabuľkách postačujúce, ak máme v aspoň 80 % prípadoch očakávanú početnosť väčšiu ako 5, prípadne nemáme očakávanú početnosť menej ako 1 vo viac ako 10 % prípadoch. Preto nízku očakávanú početnosť pre odpovede „neviem“ nepovažujeme za problematickú. Samotnú hypotézu o nezávislosti dvoch nominálnych premenných preto overíme prostredníctvom Pearsonovho Chí-kvadrát testu. Vo funkcii `chisq.test()` nastavíme parameter `simulate.p-value = TRUE`, čím získame výpočet *p*-hodnôt prostredníctvom Monte Carlo simulácie. Počet iterácií (parameter `B`) sme zvolili na 10000.

```
> chisq.test(kont, correct = F, simulate.p.value = TRUE, B =
  10000)

        Pearson's Chi-squared test with simulated p-value
                  (based on 10000 replicates)

data:  kont
X-squared = 2045.134, df = NA, p-value = 9.999e-05
```

Nulovú hypotézu môžeme zamietnuť na hladine významnosti 1 %. Zrejme teda existuje vzťah medzi tým, či je žena vo vedení a či žena je jedným z vlastníkov podniku. Spôsob, akým sa tieto simulácie uskutočňujú si len rámcovo naznačíme. Z empirickej kontingenčnej tabuľky získame pravdepodobnosti pre každý prvok tejto tabuľky za predpokladu nezávislosti premenných tak, ako sme to popísali vyššie. Následne si vygenerujeme novú tabuľku, v ktorej platí nezávislosť medzi premennými. Použijeme k tomu predchádzajúce pravdepodobnosti a multinomické rozdelenie pravdepodobnosti (čo je

zovšeobecnením binomického rozdelenia pravdepodobnosti na výskyt početností väčšieho počtu nezávislých javov, pozri `rmultinom()`).

```
> ex_prop <- chisq.test(kont)$expected/11668
> prob <- c(ex_prop)
> simulated <- addmargins(matrix(c(rmultinom(1, 11668, prob)),
  ncol = 3))
```

Vypočítame testovaciu štatistiku a celý postup s generovaním tabuľky opakujeme povedzme 10000 krát. Získame tak rozdelenie pravdepodobnosti testovacej štatistiky za predpokladu platnosti nulovej hypotézy. Následne zistíme $1 - \alpha$ kvantil a táto hodnota bude pre nás predstavovať príslušnú kritickú hodnotu.

Analyzovanú hypotézu v závere príkladu overíme aj na vzorke podnikateľských subjektov len zo SR. Po použití príkazu `table(country)` môžeme vidieť, že vo vzorke zo SR je 275 respondentov.

```
> table(country)
country
                Albania                 Armenia
                    175                     374
             Azerbaijan                 Belarus
                    380                     273
Bosnia and Herzegovina                Bulgaria
                    361                     288
                Croatia          Czech Republic
                    159                     250
                Estonia           Fyr Macedonia
                    273                     366
                Georgia                 Hungary
                    373                     291
             Kazakhstan                  Kosovo
                    544                     270
       Kyrgyz Republic                  Latvia
                    235                     271
              Lithuania                 Moldova
                    276                     363
               Mongolia              Montenegro
                    362                     116
                 Poland                 Romania
                    455                     541
                 Russia                  Serbia
                   1004                     388
       Slovak Republic                Slovenia
                    275                     276
             Tajikistan                  Turkey
                    360                    1152
                Ukraine              Uzbekistan
                    851                     366
```

Najprv si oddelíme skúmané nominálne premenné podľa príslušnosti respondenta ku krajine.

```
> SR_ownership_W <- subset(ownership_W, subset = country ==
  "Slovak Republic")
> SR_management_W <- subset(management_W, subset = country ==
  "Slovak Republic")
```

Následne môžeme pristúpiť k vytvoreniu kontingenčnej tabuľky a marginálnych početností.

```
> kont_SR <- table(SR_ownership_W, SR_management_W); kont_SR
              SR_management_W
SR_ownership_W ano neviem nie
        ano     40      0  44
        neviem   2      0   7
        nie     27      1 154
-------------------------------------------------------------
> marg_SR <- addmargins(kont_SR); marg_SR
              SR_management_W
SR_ownership_W ano neviem nie Sum
        ano     40      0  44  84
        neviem   2      0   7   9
        nie     27      1 154 182
        Sum     69      1 205 275
```

Podľa kontingenčnej tabuľky to vyzerá tak, že aj v podnikoch zo SR zrejme existuje závislosť medzi pohlavím vedenia a vlastníkov. Keď sa však pozrieme na výpočet očakávaných početností, tak môžeme vidieť, že v menej ako 80 % prípadov máme početnosti väčšie ako 5 a taktiež početnosti menšie ako 1 sa vyskytujú vo viac ako 10 % prípadov. Preto testom významnosti nemôžeme úplne dôverovať.

```
> chisq.test(kont_SR, correct = F, simulate.p.value = TRUE, B =
  10000)$expected
                    SR_management_W
SR_ownership_W       ano     neviem        nie
        ano    21.076364 0.30545455  62.618182
        neviem  2.258182 0.03272727   6.709091
        nie    45.665455 0.66181818 135.672727
-------------------------------------------------------------
> chisq.test(kont_SR, correct = F, simulate.p.value = TRUE, B =
  10000)

       Pearson's Chi-squared test with simulated p-value
                 (based on 10000 replicates)

data:  kont_SR
X-squared = 33.1847, df = NA, p-value = 0.0038
```

Z vyššie uvedených dôvodov zopakujeme celý výpočet, ale odpovede „neviem“ nahradíme odpoveďou „nie“ (spájame triedy). Získame tak kontingenčnú tabuľku s rozmermi 2 x 2, v ktorej sa problém s nízkymi očakávanými početnosťami už vyskytovať nebude.

```
> bin <- data.frame(SR_ownership_W, SR_management_W)
> NEVIEM <- bin == "neviem"
> bin_ok <- replace(bin, NEVIEM, "nie")
-----------------------------------------------------------------
> kont_SR_bin <- table(bin_ok, exclude = "neviem"); kont_SR_bin
              SR_management_W
SR_ownership_W ano nie
           ano  40  44
           nie  29 162
-----------------------------------------------------------------
> marg_SR_bin <- addmargins(kont_SR_bin); marg_SR_bin
              SR_management_W
SR_ownership_W ano nie Sum
           ano  40  44  84
           nie  29 162 191
           Sum  69 206 275
-----------------------------------------------------------------
> chisq.test(kont_SR_bin, correct = F, simulate.p.value = TRUE,
  B = 10000)$expected
              SR_management_W
SR_ownership_W      ano       nie
           ano 21.07636  62.92364
           nie 47.92364 143.07636
```

Problém s nízkym počtom očakávaných početností sme odstránili spojením dvoch tried a môžeme pristúpiť k samotnému overeniu nulovej hypotézy o nezávislosti medzi pohlavím vedenia a vlastníkov podnikov v SR.

```
> chisq.test(kont_SR_bin, correct = F, simulate.p.value = TRUE,
  B = 10000)

        Pearson's Chi-squared test with simulated p-value
                  (based on 10000 replicates)

data:  kont_SR_bin
X-squared = 32.6572, df = NA, p-value = 9.999e-05
```

Nulovú hypotézu aj na vzorke podnikateľských subjektov zo SR zamietame na hladine významnosti 1 %. Evidentne teda existuje závislosť medzi prítomnosťou žien vo vlastníctve a vedení podnikov.

Príklad 7.40

V tomto príklade ideme opäť overovať hypotézu o nezávislosti medzi pohlavím manažmentu a vlastníkov podniku, tentoraz však len na vzorke respondentov zo Slovenska a bez spájania tried (odpovede „neviem“ z databázy vylúčime).

```
> SR <- read.table(file = "...cesta
  k súboru...\\enterprise_survey.csv", sep = ";", dec = ".",
  header = T)
> SR <- subset(SR, subset = country == "Slovak Republic", select
  = c(ownership_W, management_W))
> attach(SR)
-----------------------------------------------------------------
> kont_SR <- table(SR, exclude = "neviem"); kont_SR
           management_W
ownership_W ano nie
        ano  40  44
        nie  27 154
-----------------------------------------------------------------
> kont_SR_margin <- addmargins(table(SR, exclude = "neviem"));
  kont_SR_margin
           management_W
ownership_W ano nie Sum
        ano  40  44  84
        nie  27 154 181
        Sum  67 198 265
```

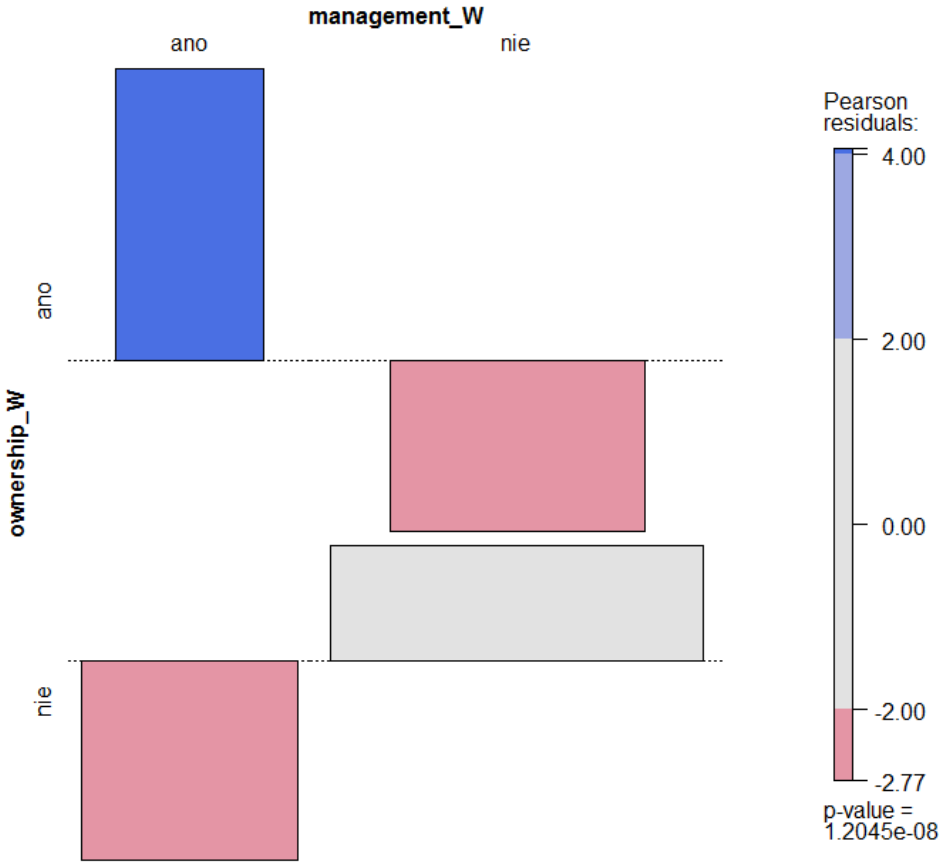
Máme teda zostavenú kontingenčnú tabuľku o rozmeroch 2 x 2. Oproti predchádzajúcemu príkladu môžeme vidieť, že máme o 10 pozorovaní menej. V texte sme uviedli, že na výpočet *Phi* koeficientu je možné využiť funkciu `phi()` z knižnice `psych`. K dispozícii však na výpočet koeficientov asociácie medzi nominálnymi premennými máme v softvéri R aj iné možnosti. Vhodné je napríklad použiť funkciu `assocstats()` z knižnice `vcd`, prostredníctvom ktorej vieme veľmi jednoducho vypočítať Pearsonov Chí-kvadrát test, *Phi* koeficient, kontingenčný koeficient alebo Cramerov-*V* koeficient.

```
> library(psych)
> phi(kont_SR)
[1] 0.35
-----------------------------------------------------------------
> library(vcd)
> assocstats(kont_SR)
                    X^2 df   P(> X^2)
Likelihood Ratio 30.917  1 2.6934e-08
Pearson          32.480  1 1.2045e-08

Phi-Coefficient   : 0.35
Contingency Coeff.: 0.33
Cramer's V        : 0.35
```

Silu vzájomného vzťahu medzi skúmanými nominálnymi premennými môžeme vidieť podľa vypočítaných koeficientov a významnosť (výsledok testovania nulovej hypotézy o nezávislosti) prostredníctvom Pearsonovho Chí-kvadrát testu (k dispozícii je aj tzv. *Likelihood Ratio* test, ktorým sme sa bližšie nezaoberali). Keďže sme pracovali s kontingenčnou tabuľkou s rozmermi 2 x 2, tak *Phi* koeficient a Cramerov-*V* koeficient sú totožné.

Zaujímavým doplnkom z knižnice `vcd` sú rôzne formy vizualizácie vzťahu medzi nominálnymi (resp. kategorickými) premennými. Na ukážku uvedieme dva spôsoby vizualizácie. Prvým je tzv. asociačný graf, ktorý zobrazuje odchýlky od modelu nezávislosti medzi premennými.

```
> library(vcd)
> assoc(kont_SR, shade = TRUE)
```



Obrázok 7.31: Asociačný graf medzi nominálnymi premennými

Zdroj: vlastné spracovanie, výstup zo softvéru R

Každý prvok z kontingenčnej tabuľky je v asociačnom grafe zobrazený vo forme obdĺžnika, ktorého obsah zobrazuje rozdiely v pozorovaných a očakávaných početnostiach[38]. Samotné umiestnenie obdĺžnikov je realizované vzhľadom na nezávislosť medzi premennými. Ak je obdĺžnik umiestnený nad deliacou čiarou (v riadku), znamená to, že pozorované početnosti v kontingenčnej tabuľke sú vyššie ako očakávané (v prípade nezávislosti premenných). V opačnom prípade je obdĺžnik umiestnený pod čiarou. V našom prípade sú pozorované početnosti vyššie ako očakávané (pri nezávislosti premenných) v prípade odpovedí áno/áno (žena je vo vedení a aj je jedným z vlastníkov) a nie/nie (žena nie je vo vedení a nie je ani jedným z vlastníkov). V asociačnom grafe máme zobrazené tiež rezíduá (odchýlky) od modelu nezávislosti, ktorých farba sa mení podľa smeru odchýlky a ich veľkosti. Spočítaním týchto odchýlok dostávame Pearsonov Chí-kvadrát test a uvedená *p*-hodnota korešponduje s výsledkami tohto testu, ktorú sme dostali prostredníctvom funkcie `assocstats()`, resp. s výsledkami, ktoré by sme dostali s použitím funkcie `chisq.test()` avšak bez simulovaných *p*-hodnôt.

```
> chisq.test(kont_SR, correct = F)

                  Pearson's Chi-squared test

data:  kont_SR
X-squared = 32.4796, df = 1, p-value = 1.205e-08
```

Druhý spôsob vizualizácie nominálnych premenných, ktorý si ukážeme, je tzv. mozaikový graf (oba uvedené spôsoby vizualizácie je možné použiť aj na väčšie kontingenčné tabuľky).

```
> library(vcd)
> mosaic(kont_SR, shade = TRUE, legend = TRUE)
```

[38] Bližšie napr. na `http://cran.r-project.org/web/packages/vcd/vcd.pdf` alebo na `http://datavis.ca/online/mosaics/about.html`.

Obrázok 7.32: Mozaikový graf medzi nominálnymi premennými

Zdroj: vlastné spracovanie, výstup zo softvéru R

Princíp mozaikového grafu je veľmi podobný princípu zostavenia asociačného grafu. Opäť máme k dispozícii aj Pearsonove rezíduá, pričom v našom prípade môžeme vidieť, že tmavomodrou farbou sú rezíduá väčšie ako 4, čo indikuje oveľa väčšiu početnosť žien vo vedení a súčasne žien majiteliek, ako by to bolo v prípade nezávislosti medzi týmito dvoma premennými. Tmavšou červenou sú rezíduá menšie ako 2, čo znamená, že ide skôr o zriedkavú kombináciu skúmaných premenných.

Príklad 7.41

Ideme overovať hypotézu o nezávislosti medzi pohlavím študentov a tým, ktorou rukou píšu. Pracujeme s údajmi z databázy `survey` (knižnica `MASS`). Najprv si zostrojíme kontingenčnú tabuľku medzi skúmanými premennými.

```
> library(MASS)
> library(vcd)
> attach(survey)
> data <- data.frame(Sex, W.Hnd)
-------------------------------------------------------------
> kont <- table(data); kont
        W.Hnd
Sex      Left Right
  Female    7   110
  Male     10   108
```

Z kontingenčnej tabuľky môžeme vidieť, že študentov ľavákov vo vzorke máme len 17. Ak chceme vidieť, aké by mal byť početnosti v prípade nezávislosti, je nutné vypočítať očakávané početnosti.

```
> chisq.test(kont, correct = F)$expected
        W.Hnd
Sex          Left    Right
  Female 8.46383 108.5362
  Male   8.53617 109.4638
```

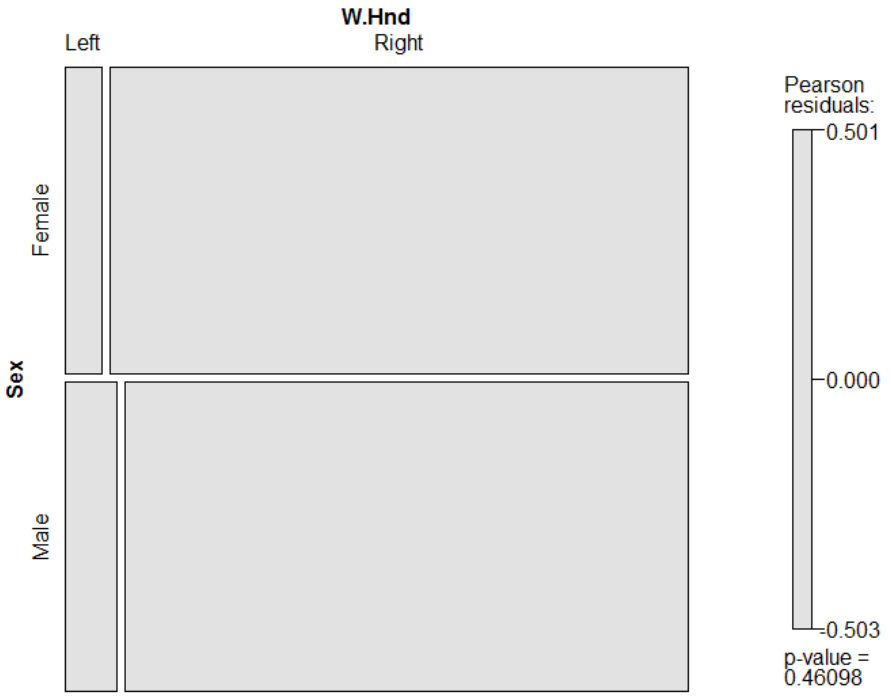
O závislosti medzi premennými by sme mohli uvažovať v prípade, ak očakávané početnosti sú výrazne odlišné od pozorovaných početností v kontingenčnej tabuľke. V tomto príklade tomu však nie je tak, takže zrejme neexistuje vzťah medzi pohlavím a tým, ktorou rukou študenti píšu. Uvedenú skutočnosť napriek tomu overíme s využitím príslušných testov.

```
> chisq.test(kont, correct = F)

                 Pearson's Chi-squared test

data:  kont
X-squared = 0.5435, df = 1, p-value = 0.461
-----------------------------------------------------------------
> assocstats(kont)
                     X^2 df P(> X^2)
Likelihood Ratio 0.54629  1  0.45984
Pearson          0.54351  1  0.46098

Phi-Coefficient   : 0.048
Contingency Coeff.: 0.048
Cramer's V        : 0.048
```

V rámci opakovania sme uviedli aj výpočet Pearsonovho Chí-kvadrát testu pomocou funkcie `chisq.test()`, ale je zrejmé, že tieto výsledky budú totožné s výstupom z funkcie `assocstats()`.

Náš predpoklad o nezávislosti premenných sa potvrdil, čo môžeme vidieť na asociačných koeficientoch (ktoré dosahujú hodnoty veľmi blízke nule), ale taktiež aj na nevýznamnosti Pearsonovho Chí-kvadrát testu. Aby sme mali možnosť vidieť, ako vyzerá grafická vizualizácia nezávislých premenných, na záver tohto príkladu sa môžeme pozrieť na mozaikový graf.

```
> mosaic(kont, shade = TRUE, legend = TRUE, )
```



Obrázok 7.33: Mozaikový graf – nezávislé premenné

Zdroj: vlastné spracovanie, výstup zo softvéru R

Z uvedené grafu je znovu zrejmé, že medzi skúmanými premennými závislosť nie je. Vizualizácia je veľmi často vhodným doplnkom formálnejšej analýzy dát. Napríklad v tomto prípade sa na prvý pohľad vieme rozhodnúť, či má zmysel ďalšie skúmanie vzťahu medzi zvolenými premennými.

Príklad 7.42

V tomto príklade máme vypočítať Pearsonov korelačný koeficient medzi rôznymi fundamentálnymi ukazovateľmi podnikov z databázy Eurocompfirm v rámci skupín podnikov z EU a UK (rozdelenie podľa premennej `Sub.Group`). Korelačné koeficienty, ktoré budú významné v oboch skupinách na hladine 5 %, máme ešte otestovať na ich rovnosť. Najprv si pripravíme požadované dáta, teda rozdelíme si celú databázu ukazovateľov podľa príslušnosti ku skupine z EU alebo UK.

```
> data <- read.csv(file = "...cesta
  k súboru...\\Eurocompfirm.csv", sep = ";", dec = ".", header =
  T)
> attach(data)
-----------------------------------------------------------------
> EU <- subset(data, Sub.Group=="EU")
> data_EU = data.frame(EU$Beta, EU$PB, EU$Payout, EU$Growth,
  EU$ROE)
-----------------------------------------------------------------
> UK <- subset(data, Sub.Group=="UK")
```

```
> data_UK = data.frame(UK$Beta, UK$PB, UK$Payout, UK$Growth,
  UK$ROE)
```

Na výpočet korelačných koeficientov využijeme funkciu `rcor.test()` z knižnice `ltm`. Táto funkcia je vhodná z toho dôvodu, že jej výstupom je štvorcová matica, kde nad hlavnou diagonálou máme možnosť vidieť vypočítané korelačné koeficienty a pod hlavnou diagonálou sú umiestnené prislúchajúce významnosti.

```
> library(ltm)
> rcor.test(data_EU, method = "pearson")

           EU.Beta EU.PB  EU.Payout EU.Growth EU.ROE
EU.Beta     *****  -0.055  0.014    -0.038    -0.016
EU.PB       0.016   *****  0.006     0.012     0.230
EU.Payout   0.539   0.801  *****    -0.015    -0.040
EU.Growth   0.090   0.593  0.502     *****     0.019
EU.ROE      0.468  <0.001  0.081     0.397     *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
------------------------------------------------------------
> rcor.test(data_UK, method = "pearson")

           UK.Beta UK.PB  UK.Payout UK.Growth UK.ROE
UK.Beta     *****  -0.039 -0.066     0.046     0.011
UK.PB       0.338   ***** -0.027     0.038     0.171
UK.Payout   0.105   0.513  *****    -0.108    -0.047
UK.Growth   0.258   0.352  0.008     *****    -0.057
UK.ROE      0.793  <0.001  0.249     0.163     *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

V skupine podnikov z EU sú významné na hladine 5 % vzťahy medzi ukazovateľom P/B a ukazovateľmi Beta a ROE. Z podvzorky podnikov z UK sú na hladine 5 % významné vzťahy medzi výplatným pomerom (premenná `Payout`) a mierou rastu tržieb (premenná `Growth`). Ďalší významný vzťah je medzi ukazovateľmi P/B a ROE. Z toho vyplýva, že jediné korelačné koeficienty v rámci oboch skupín, ktoré má zmysel testovať na vzájomnú rovnosť, sú koeficienty medzi P/B a ROE. V skupine podnikov z EU dosiahol tento korelačný koeficient výšku 0.230 a v skupine podnikov z UK dosiahol výšku 0.171. V oboch prípadoch teda ide skôr o nižšiu závislosť a čo do samotnej výšky korelačných koeficientov, tak tie sú si dosť podobné. Avšak uvidíme, či je možné považovať tieto koeficienty za štatisticky porovnateľné. Na testovanie rovnosti korelačných koeficientov medzi týmito dvoma nezávislými skupinami využijeme vlastnú funkciu `rho_indcomp()`.

```
> rho_indcomp <- function(var11, var12, var21, var22,
  alternative = c("two.sided", "greater", "less")) {
+ a <- cor(var11, var12)
+ b <- cor(var21, var22)
+ c <- 0.5*log((1 + a)/(1 - a))
+ d <- 0.5*log((1 + b)/(1 - b))
+ zrho <- (c - d)/(sqrt(1/(length(var11)-3) +1/(length(var21)-
  3)))
+ if (alternative == "two.sided") {
+ cat("two.sided p-value is")
+ print((1-pnorm(abs(zrho)))*2)
+ }
+ if (alternative == "greater") {
+ cat("p-value for r1 > r2 alternative is")
+ print(1-pnorm(zrho))
+ }
+ if (alternative == "less") {
+ cat("p-value for r1 < r2 alternative is")
+ print(pnorm(zrho))
+ }
+ }
-----------------------------------------------------------------
> rho_indcomp(EU$ROE, EU$PB, UK$ROE, UK$PB, alternative =
  "two.sided")
two.sided p-value is[1] 0.1872588
```

Na základe uvedenej *p*-hodnoty nemôžeme zamietnuť nulovú hypotézu o rovnosti dvoch korelačných koeficientov zo skúmaných skupín podnikov medzi ukazovateľmi P/B a ROE. Aj napriek tomu, že korelačné koeficienty sa javili ako podobné (čo do ich výšky), mohli by sme prijať tvrdenie, že vzťah medzi P/B a ROE je silnejší v rámci skupiny podnikov z EU. Ak však toto tvrdenie otestujeme prostredníctvom tej istej funkcie, v ktorej však zmeníme nulovú hypotézu (na $\rho^1_{x,y} \leq \rho^2_{x,y}$) a alternatívnu hypotézu (na $\rho^1_{x,y} > \rho^2_{x,y}$), dosiahneme *p*-hodnotu na úrovni 0.0936. Alternatívnu hypotézu, že korelačný koeficient medzi premennými P/B a ROE je vyšší v skupine podnikov z EU, v porovnaní s korelačným koeficientom zo skupiny podnikov z UK, vieme prijať len na hladine významnosti 10 %.

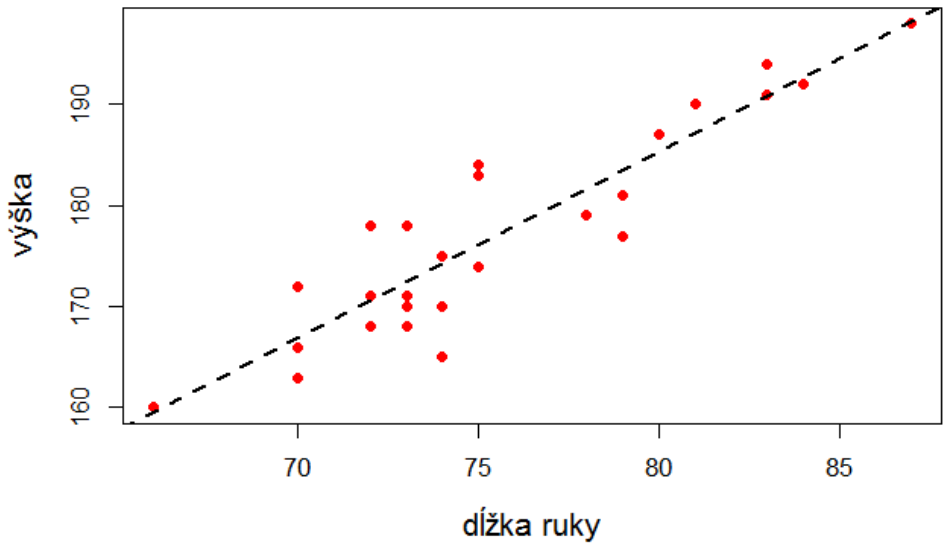
```
> rho_indcomp(EU$ROE, EU$PB, UK$ROE, UK$PB, alternative =
  "greater")
p-value for r1 > r2 alternative is[1] 0.09362938
```

Príklad 7.43

V tomto príklade máme k dispozícii 26 študentov, pričom za každého študenta máme pozorovania o výške, dĺžke ruky a pohlaví. V prvej úlohe máme zostrojiť vhodný obrázok, ktorý by vizuálne zobrazoval vzťah medzi výškou a dĺžkou ruky. Za týmto účelom sa štandardne využíva *x-y* graf. V programe R môžeme takýto graf vytvoriť viacerými spôsobmi,

v tomto príklade si ukážeme dva zrejme najjednoduchšie. Prvý spôsobom je štandardný, bez použitia akejkoľvek knižnice:

```
> plot(vyska ~ dlzka, xlab = "dĺžka ruky", ylab = "výška", col =
  "red", pch = 19, cex.axis = 1.1, cex.lab = 1.3)
> abline(lm(vyska ~ dlzka),lty = 2, lwd = 2)
```



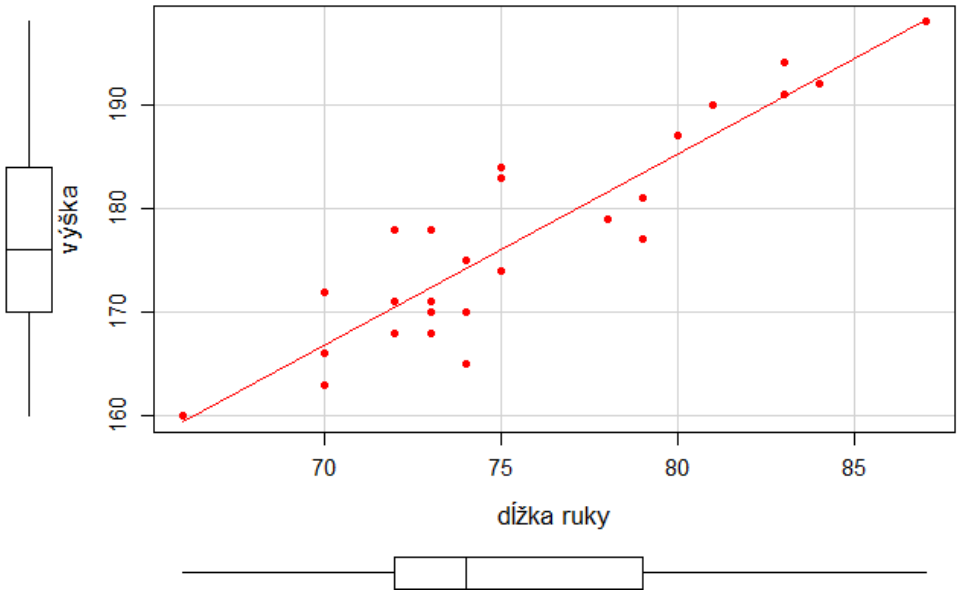
Obrázok 7.34: Graf závislosti medzi výškou študentov a dĺžkou ich ruky – prvý spôsob

Zdroj: výstup zo softvéru R

V súvislosti s druhou úlohou, v ktorej máme do grafu zobraziť aj lineárnu regresnú priamku, sme ju do tohto obrázku naniesli prostredníctvom funkcie `abline()` (hrúbku meníme cez parameter `lwd` a typ čiary cez parameter `lty`).

Pri druhom spôsobe vytvorenia *x-y* grafu využijeme knižnicu `car` a funkciu `scatterplot()`:

```
> library(car)
> scatterplot(vyska ~ dlzka, smooth = F, xlab = "dĺžka ruky",
  ylab = "výška", col = "red", pch = 19, cex.lab = 1.3, cex.axis
  = 1.1)
```



Obrázok 7.35: Graf závislosti medzi výškou študentov a dĺžkou ich ruky – druhý spôsob

Zdroj: výstup zo softvéru R

Pri tomto type vytvorenia *x-y* grafu je tzv. regresná priamka načrtnutá do grafu priamo, bez nutnosti použitia akýchkoľvek ďalších príkazov, rovnako ako box – ploty vedľa osí grafu. Z uvedených obrázkov môžeme vidieť, že vzťah medzi skúmanými premennými popisuje lineárna priamka dobre, a teda zrejme nie je nutné uvažovať o inom ako lineárnom vzťahu medzi výškou a dĺžkou ruky (zrejme nemusíme pripomínať, že ide značne zjednodušené rozhodnutie o type funkčnej závislosti, ale pre naše potreby zatiaľ postačujúce).

V ďalšej úlohe máme vypočítať silu závislosti medzi skúmanými premennými. Za týmto účelom využijeme Pearsonov korelačný koeficient, ktorý meria silu lineárnej závislosti, ale mohli by sme použiť aj koeficienty, ktoré zachytávajú aj iný typ závislosti ako lineárnu (Spearman alebo Kendall). V programe `R` môžeme na výpočet všetkých troch korelačných koeficientov použiť funkciu `cor()`, pričom pre získanie konkrétneho typu koeficientu stačí meniť argument funkcie `method = c("pearson", "kendall", "spearman")`:

```
> data=data.frame(vyska, dlzka)
> cor(data, use = "everything", method = "pearson")
          vyska     dlzka
vyska 1.0000000 0.9094061
dlzka 0.9094061 1.0000000
-------------------------------------------------------------
> cor(data, use = "everything", method = "spearman")
          vyska     dlzka
vyska 1.0000000 0.8511588
dlzka 0.8511588 1.0000000
-------------------------------------------------------------
> cor(data, use = "everything", method = "kendall")
```

```
          vyska     dlzka
vyska 1.0000000 0.7126929
dlzka 0.7126929 1.0000000
```

Všetky tri koeficienty dosahujú relatívne vysokých hodnôt. Vzťah medzi dĺžkou ruky a výškou je pomerne silný. Pri testovaní významnosti korelačných koeficientov využijeme funkciu `cor.test()`, ktorá má taktiež argument `method = c("pearson", "kendall", "spearman")`, ktorý meníme podľa toho, ktorý korelačný koeficient je predmetom nášho záujmu. Ukážeme si to na príklade Pearsonovho korelačného koeficientu, pričom testovať budeme nulovú hypotézu $\rho = 0$, oproti alternatívnej $\rho \neq 0$ (testujeme či ide o štatisticky významný vzťah a keďže ide o obojstranný test, nastavíme parameter funkcie `alternative = "two.sided"`).

```
> cor.test(vyska, dlzka, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  vyska and dlzka
t = 10.7119, df = 24, p-value = 1.262e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8059629 0.9589560
sample estimates:
      cor
0.9094061
```

Vypočítaný Pearsonov korelačný koeficient vo výške 0.909 je významný na hladine 1 %, keďže *p*-hodnota pri obojstrannom *t*-teste je nižšia ako stanovená hladina významnosti (číslo blízke nule). Vhodnou alternatívou na výpočet významnosti korelačného koeficientu je funkcia `rcor.test()` z knižnice `ltm`, ktorej výsledkom je matica, v ktorej nad hlavnou diagonálou je korelačný koeficient a pod hlavnou diagonálou jeho významnosť:

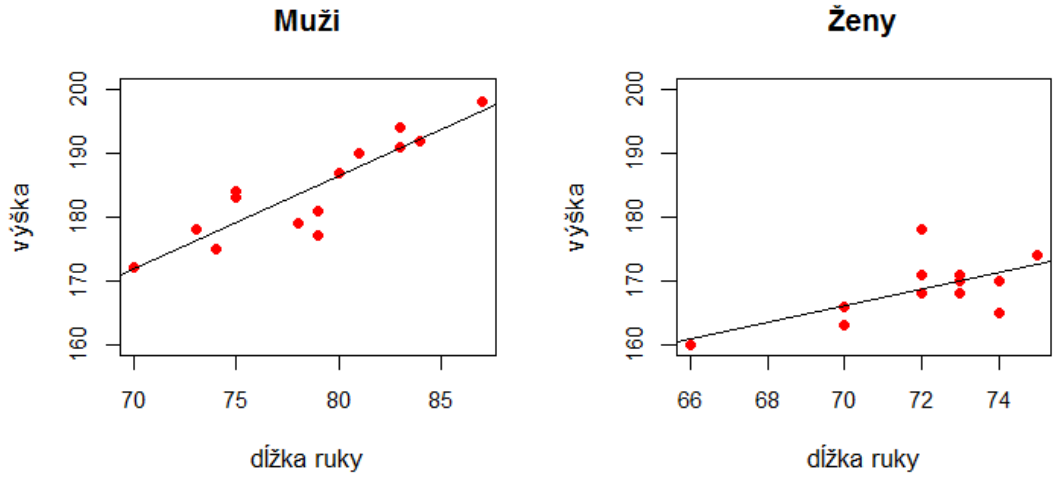
```
> library(ltm)
> rcor.test(data, method = "pearson")

      vyska  dlzka
vyska  *****  0.909
dlzka <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

V ďalších otázkach tohto príkladu máme pracovať zvlášť s údajmi pre mužov a zvlášť pre ženy. Za týmto účelom si vytvoríme pomocné premenné.

```
> vyska_M <- subset(vyska, subset = pohlavie == "m")
> vyska_Z <- subset(vyska, subset = pohlavie == "z")
> dlzka_M <- subset(dlzka, subset = pohlavie == "m")
> dlzka_Z <- subset(dlzka, subset = pohlavie == "z")
```

Graf závislosti pre skúmané premenné zostrojíme podľa prvého (vyššie uvedeného) spôsobu s tým, že nastavíme rovnakú škálu na osi *y*.

```
> par(mfrow = c(1, 2))
> plot(vyska_M ~ dlzka_M, main = "Muži", ylim = c(160, 200),
  xlab = "dĺžka ruky", ylab = "výška", col = "red", pch = 19,
  cex.axis = 0.9, cex.lab = 1.1)
> abline(lm(vyska_M ~ dlzka_M),lty=1, lwd=1)
> plot(vyska_Z ~ dlzka_Z, main = "Ženy", ylim = c(160, 200),
  xlab = "dĺžka ruky", ylab = "výška", col = "red", pch = 19,
  cex.axis = 0.9, cex.lab = 1.1)
> abline(lm(vyska_Z ~ dlzka_Z),lty = 1, lwd = 1)
```



Obrázok 7.36: Graf závislosti podľa pohlavia

Zdroj: výstup zo softvéru R

Keď už máme vytvorené pomocné premenné, ktoré rozdeľujú celú vzorku podľa pohlavia, môžeme tiež pristúpiť k ďalšej úlohe, v ktorej máme vypočítať korelačné koeficienty zvlášť podľa pohlavia. Na výpočet Pearsonovho korelačného koeficientu a jeho významnosti využijeme už len funkciu `cor.test()`.

```
> cor.test(vyska_M, dlzka_M, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  vyska_M and dlzka_M
t = 6.7697, df = 12, p-value = 1.987e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6815624 0.9649986
```

```
sample estimates:
      cor
0.8902184
-----------------------------------------------------------------
> cor.test(vyska_Z, dlzka_Z, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  vyska_Z and dlzka_Z
t = 2.6697, df = 10, p-value = 0.0235
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.1130305 0.8896319
sample estimates:
      cor
0.6450864
```

Môžeme vidieť, že korelačný koeficient vypočítaný medzi výškou a dĺžkou ruky na vzorke mužov je väčší (0.890), ako korelačný koeficient vypočítaný na údajoch vzorky žien (0.645). Prvý korelačný koeficient je významný na hladine 1 %, pričom druhý je významný už len na 5 % hladine významnosti.

Ďalšiu otázku ktorú potrebujeme zodpovedať je, či silu vzťahu medzi skúmanými premennými na týchto dvoch vzorkách môžeme považovať za štatisticky rovnakú alebo nie. Za účelom porovnania dvoch korelačných koeficientov máme vytvorenú funkciu `rho_indcomp()`:

```
> rho_indcomp <- function(var11, var12, var21, var22,
  alternative = c("two.sided", "greater", "less")) {
+ a <- cor(var11, var12)
+ b <- cor(var21, var22)
+ c <- 0.5*log((1 + a)/(1 - a))
+ d <- 0.5*log((1 + b)/(1 - b))
+ zrho <- (c - d)/(sqrt(1/(length(var11)-3) +1/(length(var21)-
  3)))
+ if (alternative == "two.sided") {
+ cat("two.sided p-value is")
+ print((1-pnorm(abs(zrho)))*2)
+ }
+ if (alternative == "greater") {
+ cat("p-value for r1 > r2 alternative is")
+ print(1-pnorm(zrho))
+ }
+ if (alternative == "less") {
+ cat("p-value for r1 < r2 alternative is")
+ print(pnorm(zrho))
+ }
+ }
```

Prostredníctvom tejto funkcie vieme overiť uvedenú hypotézu o rovnosti korelačných koeficientov, teda ideme overiť nulovú hypotézu: $\rho^{M}_{x,y} = \rho^{Ž}_{x,y}$, voči alternatívnej: $\rho^{M}_{x,y} \neq \rho^{Ž}_{x,y}$.

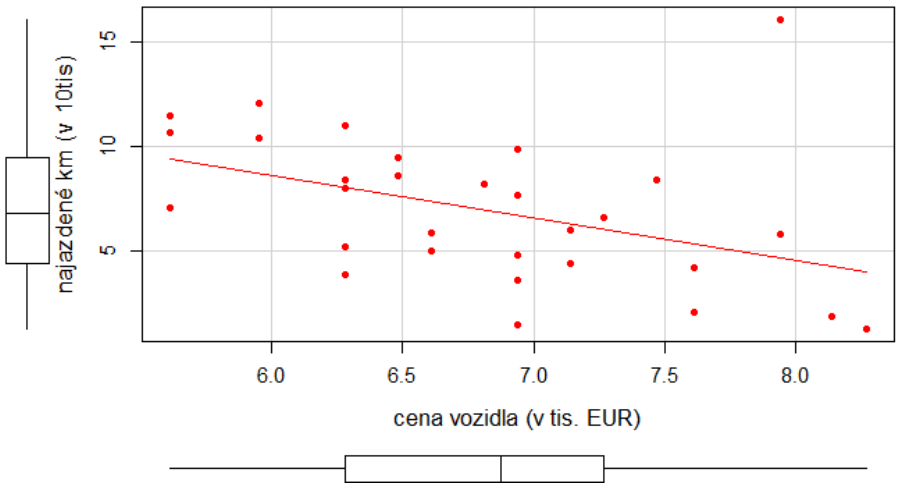
```
> rho_indcomp(vyska_M, dlzka_M, vyska_Z, dlzka_Z, alternative =
  "two.sided")
two.sided p-value is[1] 0.1443394
```

Na základe uvedeného testu rovnosti dvoch korelačných koeficientov môžeme vidieť, že nulovú hypotézu nevieme zamietnuť (*p*-hodnota je 0.144) ani na hladine významnosti 10 %, a teda nemáme dostatok dôkazov k tomu, aby sme tieto korelačné koeficienty mohli považovať za štatisticky odlišné. Tento výsledok je zaujímavý najmä preto, lebo rozdiel medzi korelačnými koeficientmi je dosť výrazný, no napriek tomu nie je štatisticky významný. Zrejme by výsledok testu bol odlišný, ak by sme k dispozícii mali väčší počet pozorovaní v jednotlivých vzorkách mužov a žien.

Porovnávanie korelačných koeficientov pomocou bootstrappingu necháme na čitateľovi. Napovieme, že vzorkovanie musí prebiehať na usporiadaných *n*-ticiach.

Príklad 7.44

V tomto príklade pracujeme s údajmi o automobilových vozidlách. V prvej úlohe máme formou vizualizácie údajov o cene vozidla a počtu najazdených kilometrov zistiť, či medzi týmito premennými existuje nejaká forma vzťahu. Na zobrazenie vzájomného vzťahu použijeme *x*-*y* graf z knižnice `car`, podobne ako v predošlom prípade.



Obrázok 7.37: Vzťah medzi počtom najazdených kilometrov a cenou vozidla

Zdroj: výstup zo softvéru R

```
> library(car)
> scatterplot(km_v_10tis ~ cena_v_tis_eur, smooth = F, xlab =
  "cena vozidla (v tis. EUR)", ylab = "najazdené km (v 10tis)",
  col = "red", pch = 19, cex.lab = 1.3, cex.axis = 1.1)
```

Na uvedenom obrázku môžeme vidieť, že ekonóm v podniku predpokladal správne, keď považoval vzťah medzi cenou vozidla a počtom najazdených kilometrov za nepriamo úmerný. Zrejme by bol tento nepriamy vzťah ešte výraznejší, ak by sme nebrali do úvahy jednu zjavnú extrémnu úlohu.

Exaktnejšie máme tento vzťah kvantifikovať v druhej zadanej úlohe, na čo nám zatiaľ bude postačovať korelačný koeficient. Zároveň máme zistiť, či je vzťah medzi skúmanými premennými aj štatisticky významný. Za týmto účelom si z rôznych dostupných možností v programe `R` zvolíme funkciu `cor.test()`, ktorá nám okrem hodnoty korelačného koeficientu vráti aj výsledky *t*-testu významnosti tohto koeficientu.

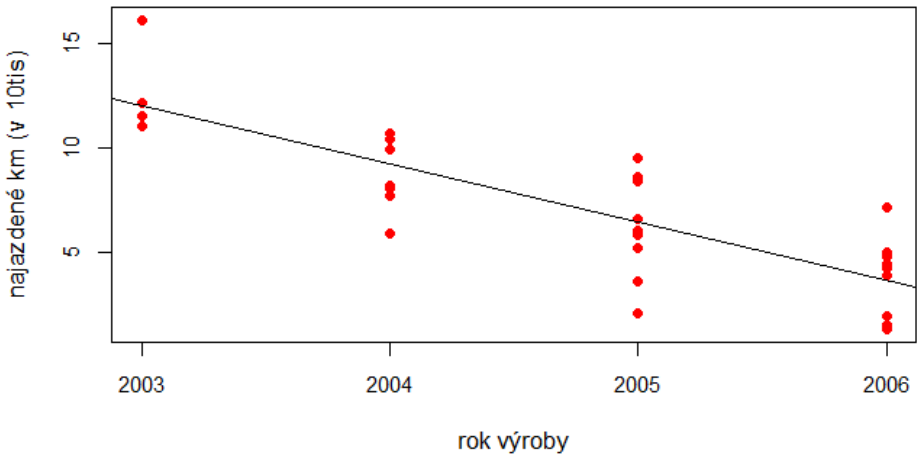
```
> cor.test(km_v_10tis, cena_v_tis_eur, method = "pearson",
  alternative = "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  km_v_10tis and cena_v_tis_eur
t = -2.5004, df = 28, p-value = 0.01853
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.68246214 -0.07915185
sample estimates:
      cor
-0.427238
```

Môžeme vidieť, že existuje negatívna korelácia medzi cenou vozidla a počtom najazdených kilometrov, keďže korelačný koeficient je záporný ($\rho = -0.427$). S rastom počtu najazdených kilometrov klesala aj cena, ktorú si predávajúci za autá pýtali. Tento korelačný koeficient je významný na hladine 5 % (*p*-hodnota je 0.018), resp. povedané inými slovami, nulovú hypotézu o nulovom korelačnom koeficiente medzi skúmanými premennými môžeme zamietnuť v prospech alternatívnej hypotézy, teda korelačný koeficient môžeme považovať za rôzny od nuly.

V úlohách c) a d) máme postupovať analogicky, avšak predmetom záujmu je skúmanie vzťahu medzi počtom najazdených kilometrov a rokom výroby vozidla. Vzťah medzi týmito premennými si najprv zobrazíme graficky.

```
> plot(km_v_10tis ~ rok_vyroby, xlab = "rok výroby", ylab =
  "najazdené km (v 10tis)", xaxt = "n", col = "red", pch = 19,
  cex.axis = 0.9, cex.lab = 1.1)
```

```
> grid <- c(2003,2004,2005,2006)
> axis(side=1, las = "1", at = grid, cex.axis = 1, tick = TRUE,
  labels = c(2003,2004,2005,2006), cex.axis = 0.9)
> abline(lm(km_v_10tis ~ rok_vyroby),lty = 1, lwd = 1)
```



Obrázok 7.38: Vzťah medzi počtom najazdených kilometrov a rokom výroby vozidla

Zdroj: výstup zo softvéru R

Pri vizualizácii vidno, že rok výroby má pomerne nízku rôznorodosť hodnôt. V takomto prípade je možné použiť aj iné metódy skúmania závislostí. Jednou by bolo vykonanie neparametrického Kurskal – Wallisovho testu, prípadne použitie kontingenčných koeficientov, alebo tzv. determinačného pomeru, ktorý však nie je predmetom tejto publikácie (pozri Tkáč, 2001). Rozhodli sme sa však ostať pri korelačných koeficientoch, keďže rok výroby je zjavne intervalová premenná.

Po vizualizácii vzťahu medzi týmito dvoma premennými môžeme pristúpiť k výpočtu korelačného koeficientu a overeniu jeho štatistickej významnosti prostredníctvom obojstranného testu.

```
> cor.test(km_v_10tis, rok_vyroby, method = "pearson",
  alternative = "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  km_v_10tis and rok_vyroby
t = -7.5638, df = 28, p-value = 3.072e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.9107937 -0.6514253
sample estimates:
       cor
-0.8193925
```

Pearsonov korelačný koeficient nám hovorí o značne silnej negatívnej korelácii medzi počtom najazdených kilometrov a rokom výroby vozidla (čo je očakávaný výsledok). Tento vzťah môžeme považovať aj za štatisticky významný.

Na záver tohto príkladu ešte máme vypočítať, aká by bola korelácia medzi cenou a počtom najazdených kilometrov, ak by bola premenná rok výroby konštantná. Na výpočet takto zadanej úlohy použijeme parciálny korelačný koeficient. Pre jednoduchosť si najprv vytvoríme korelačnú maticu zo všetkých troch premenných (označíme ju `correl`) a následne podľa v zadaní uvedeného vzťahu na výpočet parciálneho korelačného koeficientu pristúpime k jeho výpočtu jednoduchým adresovaním na prvky danej korelačnej matice.

```
> data <- data.frame(km_v_10tis, cena_v_tis_eur, rok_vyroby)
> correl <- cor(data, use = "everything", method = "pearson");
  correl
               km_v_10tis cena_v_tis_eur rok_vyroby
km_v_10tis      1.0000000     -0.4272380 -0.8193925
cena_v_tis_eur -0.4272380      1.0000000  0.3453184
rok_vyroby     -0.8193925      0.3453184  1.0000000
-------------------------------------------------------------------
> rho <- (correl[2] - correl[3]*correl[6])/sqrt((1-correl[3]^2)*
  (1-correl[6]^2)); rho
[1] -0.2682053
```

Po eliminácii vplyvu tretej premennej (rok výroby) môžeme vidieť, že vzťah medzi počtom najazdených kilometrov a cenou vozidla sa znížil, korelácia však stále ostáva záporná. Evidentne vysoká záporná korelácia medzi počtom najazdených kilometrov a rokom výroby do značnej miery ovplyvňovala aj koreláciu medzi počtom najazdených kilometrov a cenou vozidla.

Príklad 7.46

Podľa zadania príkladu máme v prvej úlohe zistiť, či pri bytoch s väčšou rozlohou je cena za jeden meter štvorcový nižšia. Najprv si musíme vytvoriť premennú, ktorá bude zodpovedať cene za jeden meter štvorcový, označme ju ako `cena_1m2`. Následne môžeme pristúpiť k výpočtu korelačného koeficientu, ktorý by nám mal poskytnúť odpoveď na danú otázku.

```
> cena_1m2 <- cena/plocha
> cor.test(plocha, cena_1m2, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  plocha and cena_1m2
```

```
t = 3.8441, df = 21, p-value = 0.0009425
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.3135382 0.8339554
sample estimates:
      cor
0.6426797
```

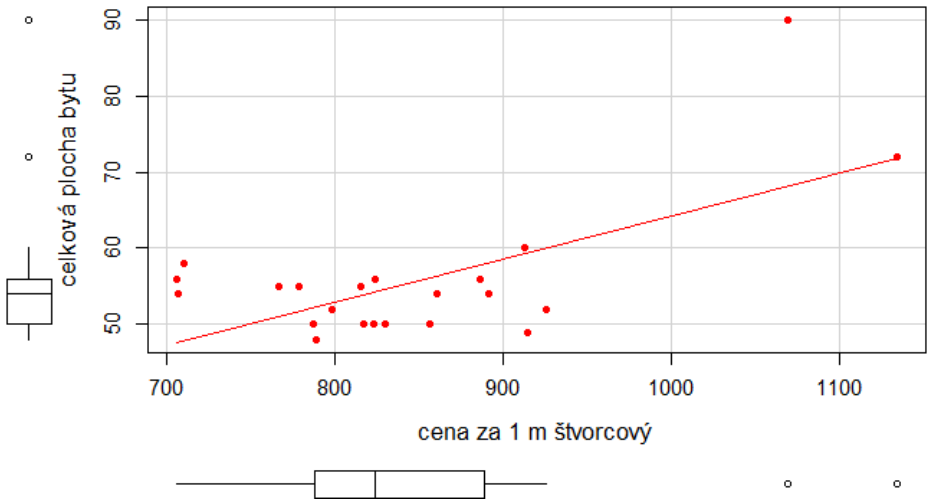
Z uvedeného môžeme vidieť, že existuje pozitívna lineárna závislosť medzi plochou bytu a cenou za jeden meter štvorcový. V tomto prípade teda nie je pravdou, že čím je väčšia plocha bytu, o to je menšia cena za jeden meter štvorcový, práve naopak, cena by mala byť vyššia.

Obdobne v druhej úlohe, v ktorej máme zistiť vzťah medzi plochou byt a celkovou cenou, taktiež platí pozitívna korelácia. Je zrejmé, že čím má byt vyššiu rozlohou, tým je drahší.

```
> cor.test(plocha, cena, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  plocha and cena
t = 12.6293, df = 21, p-value = 2.821e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8617340 0.9745939
sample estimates:
      cor
0.9400296
```

V tomto príklade sme mali za úlohu dané vzťahy zobraziť aj graficky, čo odporúčame vykonať pri každej analýze, kde to len je možné a to najmä z toho dôvodu, že pri vizuálnom zobrazení údajov môžeme veľmi ľahko identifikovať extrémne hodnoty, ktoré následne môžu ovplyvniť celý výsledok nasledujúcej analýzy. Na nasledujúcom obrázku je zobrazený vzťah medzi rozlohou bytu a cenou za jeden meter štvorcový.

```
> library(car)
> scatterplot(plocha ~ cena_1m2, smooth = F, xlab = "cena za 1 m
  štvorcový", ylab = "celková plocha bytu", col = "red", pch =
  19, cex.lab = 1.3, cex.axis = 1.1)
```

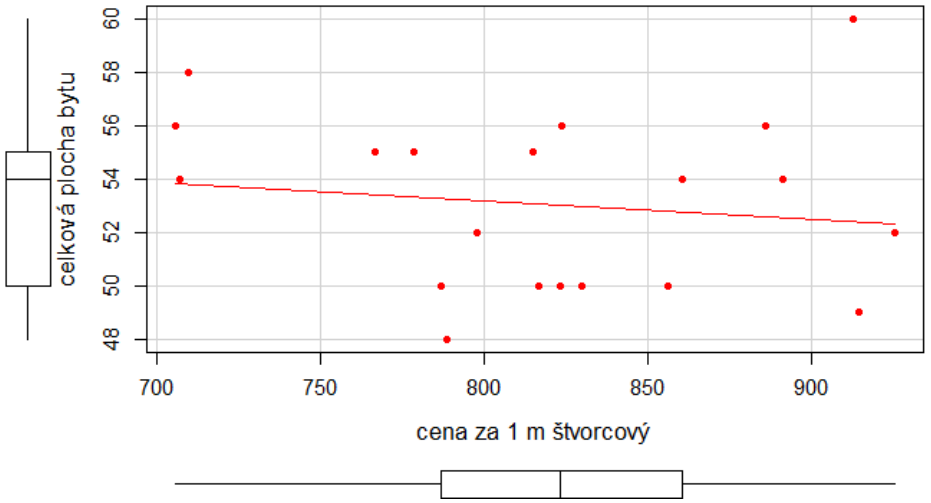


Obrázok 7.39: Vzťah medzi plochou bytu a cenou za 1 m^2

Zdroj: výstup zo softvéru R

Z uvedeného grafu môžeme pozorovať, že v našich údajoch sú prítomné minimálne dve extrémne hodnoty (čo je viditeľné najmä na zobrazených box – plotoch). Bolo by možno zaujímavé vidieť, aký majú extrémne pozorovania vplyv na naše rozhodnutie. Z tohto dôvodu by bolo vhodné pred analýzou obe pozorovania odstrániť.

Ak sa pozrieme na vizualizáciu vzťahu medzi plochou bytu a cenou za jeden meter štvorcový, tak môžeme vidieť, že tento vzťah sa zásadne zmenil po odstránení extrémnych hodnôt.



Obrázok 7.40: Vzťah medzi plochou bytu a cenou za 1 m^2 po odstránení extrémov

Zdroj: výstup zo softvéru R

```
> plocha_2 <- plocha[!(plocha %in%
  c(hampel_identifier(plocha)))]
> cena_1m2_2 <- cena_1m2[!(cena_1m2 %in%
  c(hampel_identifier(cena_1m2)))]
-------------------------------------------------------------
```

```
> scatterplot(plocha_2 ~ cena_1m2_2, smooth = F, xlab = "cena za
  1 m štvorcový", ylab = "celková plocha bytu", col = "red", pch
  = 19, cex.lab = 1.3, cex.axis = 1.1)
```

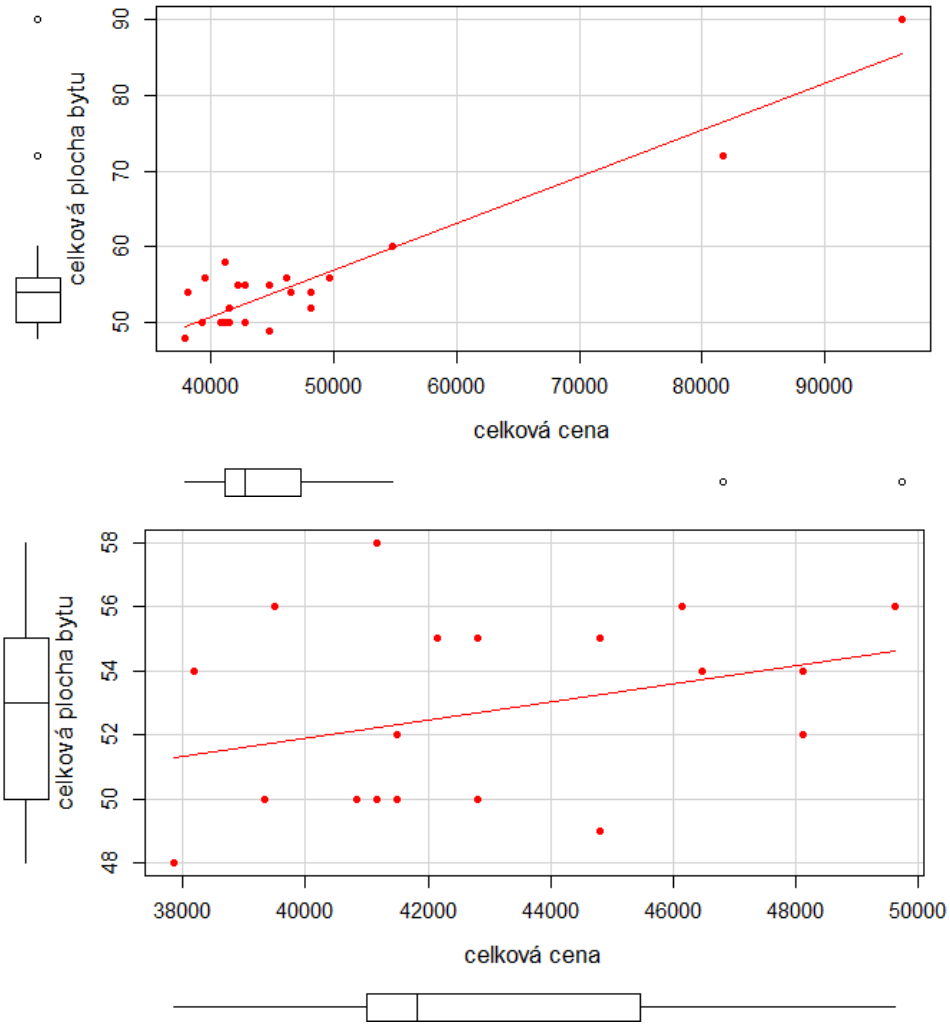
Rozptýlenosť údajov je po odstránení odľahlých hodnôt lepšie viditeľná, pričom regresná priamka už má záporný sklon. Pri výpočte Pearsonovho korelačného koeficientu by sme mali dostať negatívnu koreláciu, avšak zrejme kvôli rozptýlenosti údajov by tento korelačný koeficient nemal byť významný.

```
> cor.test(plocha_2, cena_1m2_2, method = "pearson", alternative
  = "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  plocha_2 and cena_1m2_2
t = -0.6238, df = 19, p-value = 0.5402
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.5403146  0.3089064
sample estimates:
       cor
-0.1416725
```

Z výsledkov naozaj vidíme, že korelácia je po odstránení extrémnych hodnôt záporná, avšak korelačný koeficient nie je štatisticky rozdielny od nuly. Rozhodnutie o tom, že čím je väčšia plocha bytu, o to je menšia cena za jeden meter štvorcový, teda nie je také jednoznačné ako sa na začiatku riešenia príkladu mohlo zdať.

Takmer rovnaká situácia nastáva pri riešení druhej úlohy, v ktorej sme skúmali vzťah medzi rozlohou bytu a jeho celkovou cenou. Intuitívne by medzi týmito dvoma premennými mala existovať pozitívna korelácia. Ak sa však pozrieme na grafické znázornenie ich vzťahu vo forme *x-y* grafu, tak opäť je výskyt odľahlých hodnôt úplne zjavný.

```
> cena_2 <- cena[!(cena %in% c(54770, 81658, 96263))]
> plocha_2 <- plocha[!(plocha %in% c(60, 72, 90))]
------------------------------------------------------------
> scatterplot(plocha ~ cena, smooth = F, xlab = "celková cena",
  ylab = "celková plocha bytu", col = "red", pch = 19, cex.lab =
  1.3, cex.axis = 1.1)
> scatterplot(plocha_2 ~ cena_2, smooth = F, xlab = "celková
  cena", ylab = "celková plocha bytu", col = "red", pch = 19,
  cex.lab = 1.3, cex.axis = 1.1)
```



Obrázok 7.41: Vzťah medzi plochou bytu a celkovou cenou pred odstránením extrémov (hore) a po odstránení extrémov (dole)

Zdroj: výstup zo softvéru R

Pri výpočte korelačného koeficientu medzi rozlohou bytu a jeho celkovou cenou sa nám výsledky predošlej analýzy opäť výrazne zmenili. Korelačný koeficient po odstránení extrémov značne poklesol (z 0.940 na 0.329) a nie je už ani štatisticky významný.

```
> cor.test(plocha_2, cena_2, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  plocha_2 and cena_2
t = 1.4759, df = 18, p-value = 0.1572
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1333435  0.6732040
sample estimates:
      cor
0.328565
```

Príklad 7.47

V prvej úlohe tohto príkladu máme zistiť, či je možné očakávať, že stredná hodnota počtu uzavretých PZP je v danom kraji za jeden kvartál 50. Ak vypočítame priemerný počet PZP prostredníctvom funkcie `mean()`, tak získame číslo 49.7907. Keďže však chceme výsledok zovšeobecniť, vhodnejšie je využiť *t*-test.

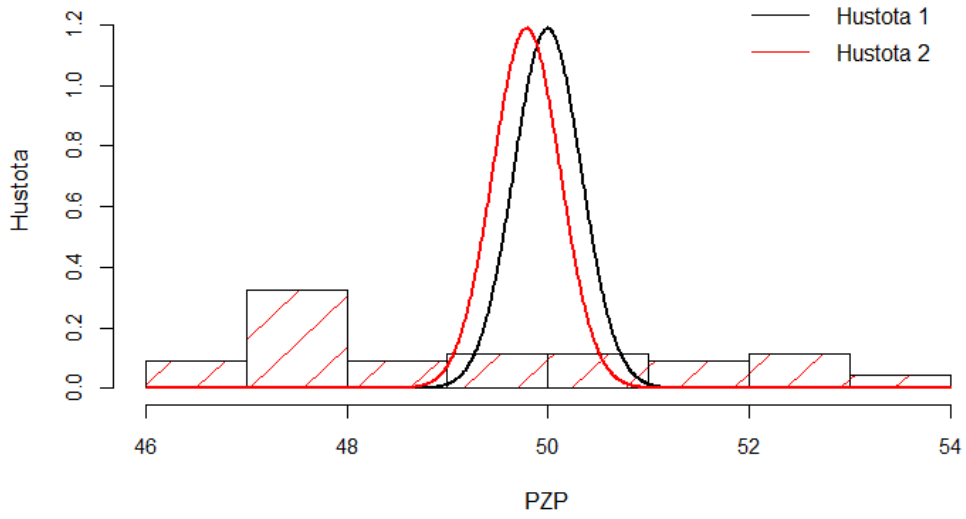
```
> t.test(PZP, alternative = "two.sided", mu = 50, conf.level =
  0.95)

                    One Sample t-test

data:  PZP
t = -0.6303, df = 42, p-value = 0.5319
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
49.12061 50.46079
sample estimates:
mean of x
  49.7907
```

Na základe výsledkov *t*-testu môžeme tvrdiť, že je možné očakávať strednú hodnotu počtu uzatvorených PZP v danom kraji na úrovni 50. Inak povedané, nevieme zamietnuť nulovú hypotézu H_0: $\mu = 50$, oproti alternatívnej H_1: $\mu \neq 50$.

V rámci druhej úlohy máme podľa zadanie príkladu zostrojiť histogram z počtu uzavretých PZP a do tohto histogramu zobraziť dve hustoty rozdelenia strednej hodnoty, pri jednej použijeme strednú hodnotu 50 (označíme ju ako „Hustota 1“) a pri druhej budeme vychádzať z predpokladu, že je stredná hodnota rovná výberovému aritmetickému priemeru (označíme ju ako „Hustota 2“). Pri odhade variability použijeme pri oboch hustotách výberový rozptyl. Z obrázku je potom vidno, že prienik týchto dvoch hustôt je pomerne veľký, čo naznačuje, že môže ísť jedno a to isté rozdelenie, resp. že neexistujú rozdiely v týchto rozdeleniach, pričom jediným rozdielom bola stredná hodnota hustoty. Tento výsledok bol potvrdený aj formálnym štatistickým testom.

```
> hist(PZP, density = 5, col = "red", border = "black", ylim =
  c(0, 1.3), main = NA, cex.lab = 1.1, cex.axis = 0.9, freq =
  FALSE, ylab = "Hustota", xlab = "PZP")
> x <- seq(min(PZP), max(PZP), length = 1000)
> xh <- dnorm(x, mean = 50, sd = sqrt(var(PZP)/(length(PZP)-1)))
> dx <- data.frame(x, xh)
> xh_2 <- dnorm(x, mean = mean(PZP), sd =
  sqrt(var(PZP)/(length(PZP)-1)))
> dx_2 <- data.frame(x, xh_2)
> lines(dx, type = "l", col = "black", lwd = 2)
```

```
> lines(dx_2, type = "l", col = "red", lwd = 2)
> legend("topright", legend = c("Hustota 1", "Hustota 2"), lty =
  1.5, col = c("black", "red"), inset = 0, bty = "n")
```



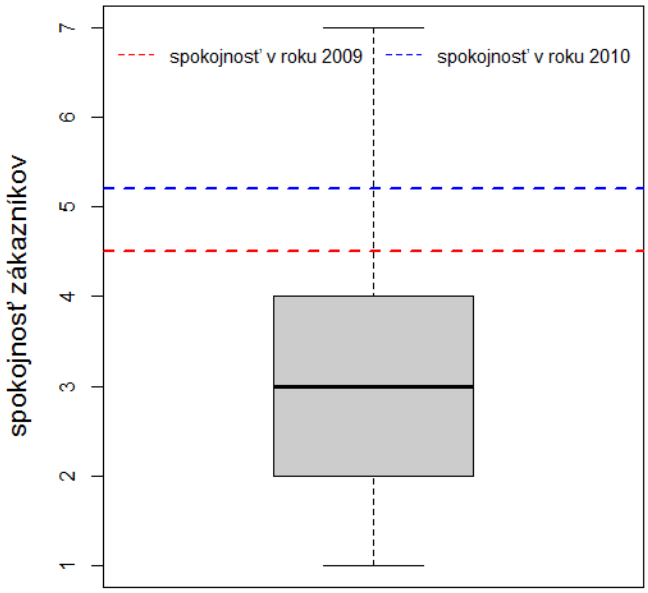
Obrázok 7.42: Histogram rozdelenia PZP

Zdroj: výstup zo softvéru R

Príklad 7.48

V tomto príklade sa budeme zaoberať údajmi z prieskumu spokojnosti zákazníkov. Najprv máme za úlohu vytvoriť graf box – plot a do tohto grafu naniesť stredné hodnoty spokojnosti zákazníkov za rok 2009 (hodnota 4.5) a za rok 2010 (hodnota 5.2).

```
> boxplot(spokojnost, ylab = "spokojnosť zákazníkov", col =
  gray(0.8), pch = 19, cex.axis = 1.1, cex.lab = 1.3)
> abline(h = 4.5, lwd = 2, lty = 2, col = "red")
> abline(h = 5.2, lwd = 2, lty = 2, col = "blue")
> legend("top", legend = c("spokojnosť v roku 2009", "spokojnosť
  v roku 2010"), cex = 0.9, ncol = 2, lty = 2, col = c("red",
  "blue"), inset = 0.05, bty = "n")
```



Obrázok 7.43: Box – plot spokojnosti zákazníkov

Zdroj: výstup zo softvéru R

Z vizualizácie dát pomocou Box – plotu je zrejmé, že došlo k výraznému poklesu spokojnosti zákazníkov. Následne máme zistiť, či v roku 2011 (v čase realizácie prieskumu spokojnosti) došlo k poklesu strednej hodnoty spokojnosti s po predajnými službami všetkých zákazníkov, a to tak oproti roku 2009, ako aj oproti roku 2010. Za týmto účelom použijeme jednostranný *t*-test s nulovou hypotézou pre rok 2009 H_0: $\mu \geq 4.5$ a alternatívnou H_1: $\mu < 4.5$ a pre rok 2010 H_0: $\mu \geq 5.2$ a alternatívnou H_1: $\mu < 5.2$.

```
> t.test(spokojnost, alternative = "less", mu = 4.5, conf.level
  = 0.95)

                    One Sample t-test

data:  spokojnost
t = -5.2704, df = 49, p-value = 1.521e-06
alternative hypothesis: true mean is less than 4.5
95 percent confidence interval:
-Inf 3.654453
sample estimates:
mean of x
     3.26
------------------------------------------------------------------
> t.test(spokojnost, alternative = "less", mu = 5.2, conf.level
  = 0.95)

                    One Sample t-test

data:  spokojnost
t = -8.2456, df = 49, p-value = 4.033e-11
alternative hypothesis: true mean is less than 5.2
95 percent confidence interval:
```

```
-Inf 3.654453
sample estimates:
mean of x
     3.26
```

Prvú nulovú hypotézu o tom, že stredná hodnota spokojnosti zákazníkov v roku 2011 je nižšia ako spokojnosť zákazníkov v roku 2009, môžeme zamietnuť na hladine významnosti 1 %. Keďže spokojnosť zákazníkov v roku 2010 bola vyššia ako v roku 2009, tak aj druhú testovanú hypotézu môžeme zamietnuť. Odhad strednej hodnoty spokojnosti zákazníkov v roku 2011 je 3.26.

V ďalšej úlohe máme podľa zadania zostrojiť obojstranné 95 % intervaly spoľahlivosti. Môžeme použiť manuálny výpočet intervalov spoľahlivosti alebo využiť funkciu `t.test()`, ktorá intervaly spoľahlivosti obsahuje vo svojom výstupe.

```
> mean(spokojnost)
[1] 3.26
> mean(spokojnost) - abs(qt(0.05/2, df  = length(spokojnost) -
  1)) * (var(spokojnost)/length(spokojnost))^0.5
[1] 2.787194
> mean(spokojnost) + abs(qt(0.05/2, df  = length(spokojnost) -
  1)) * (var(spokojnost)/length(spokojnost))^0.5
[1] 3.732806
-----------------------------------------------------------------
> t.test(spokojnost, alternative = "two.sided", mu = 0,
  conf.level = 0.95)

                    One Sample t-test

data:  spokojnost
t = 13.856, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
2.787194 3.732806
sample estimates:
mean of x
     3.26
```

Zostrojenie bootstrapingových konfidenčných intervalov necháme na čitateľa. Rovnako konfidenčné intervaly pre rozptyl.

V úlohe e) máme zodpovedať otázku, či je rozumné očakávať, že viac ako polovica všetkých zákazníkov by hodnotila spokojnosť vyššie ako 4. Použijeme test podielu voči konštante, ale najprv si vytvoríme pomocnú premennú (`spokojnost_4`), ktorá bude obsahovať len tých zákazníkov, ktorých spokojnosť bola vyššia ako 4. Samotný test podielu realizujeme prostredníctvom funkcie `binom.test()`.

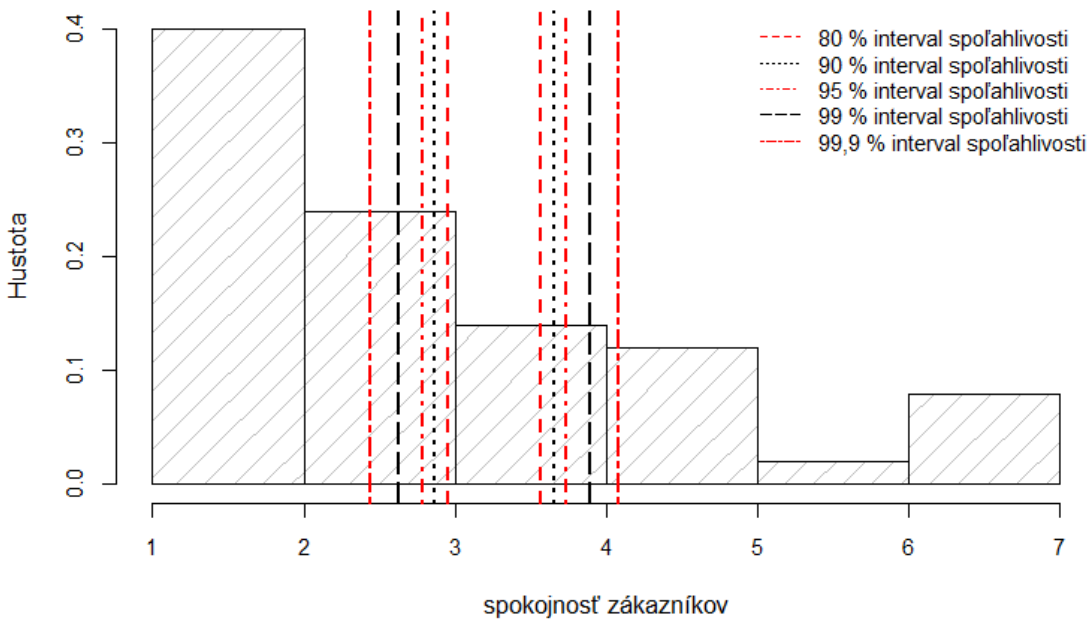
```
> spokojnost_4 <- subset(spokojnost, subset = spokojnost > 4);
> length(spokojnost_4)
[1] 11
-----------------------------------------------------------------
> binom.test(length(spokojnost_4), length(spokojnost), p = 0.5,
  alternative = "greater", conf.level = 0.95)

                    Exact binomial test

data:  length(spokojnost_4) and length(spokojnost)
number of successes = 11, number of trials = 50, p-value = 1
alternative hypothesis: true probability of success is greater
  than 0.5
95 percent confidence interval:
0.1285574 1.0000000
sample estimates:
probability of success
                  0.22
```

Na základe uvedených výsledkov môžeme tvrdiť, že nie je racionálne očakávať, že viac ako polovica všetkých zákazníkov by svoju spokojnosť hodnotila na danej škále vyšším číslom ako 4.

V poslednej úlohe máme zostrojiť histogram spokojnosti zákazníkov s tým, že do tohto grafu máme naniesť obojstranné intervaly spoľahlivosti pre strednú hodnotu vytvorené za predpokladu rôznej konfidencie. Intervaly spoľahlivosti získame z výstupu funkcie `t.test()`. Tieto intervaly sú uložené v objekte `conf.int`, pričom na prvom mieste je spodná hranica intervalu spoľahlivosti a na druhom mieste je horná hranica intervalu spoľahlivosti. Cieľom je sledovať, ako sa mení šírka týchto intervalov s rastúcou konfidenciou.

```
> a <- t.test(spokojnost, alternative = "two.sided", mu = 0,
  conf.level = 0.800)
> b <- t.test(spokojnost, alternative = "two.sided", mu = 0,
  conf.level = 0.900)
> c <- t.test(spokojnost, alternative = "two.sided", mu = 0,
  conf.level = 0.950)
> d <- t.test(spokojnost, alternative = "two.sided", mu = 0,
  conf.level = 0.990)
> e <- t.test(spokojnost, alternative = "two.sided", mu = 0,
  conf.level = 0.999)
-----------------------------------------------------------------
> hist(spokojnost, density = 7, col = "grey", border = "black",
  main = NA, cex.lab = 1.1, cex.axis = 0.9, freq = FALSE, ylab =
  "Hustota", xlab = "spokojnosť zákazníkov")
> abline(v = a$conf.int[1], lwd = 2, lty = 2, col = "red")
> abline(v = a$conf.int[2], lwd = 2, lty = 2, col = "red")
> abline(v = b$conf.int[1], lwd = 2, lty = 3, col = "black")
> abline(v = b$conf.int[2], lwd = 2, lty = 3, col = "black")
> abline(v = c$conf.int[1], lwd = 2, lty = 4, col = "red")
```

```
> abline(v = c$conf.int[2], lwd = 2, lty = 4, col = "red")
> abline(v = d$conf.int[1], lwd = 2, lty = 5, col = "black")
> abline(v = d$conf.int[2], lwd = 2, lty = 5, col = "black")
> abline(v = e$conf.int[1], lwd = 2, lty = 6, col = "red")
> abline(v = e$conf.int[2], lwd = 2, lty = 6, col = "red")
> legend("topright", legend = c("80 % interval spoľahlivosti",
  "90 % interval spoľahlivosti", "95 % interval spoľahlivosti",
  "99 % interval spoľahlivosti", "99,9 % interval
  spoľahlivosti"), cex = 0.9, ncol = 1, lty = c(2,3,4,5,6), col
  = c("red", "black", "red", "black", "red", "black", "red",
  "black", "red", "black"), inset = 0, bty = "n")
```



Obrázok 7.44: Histogram spokojnosti zákazníkov s rôznymi intervalmi spoľahlivosti pre strednú hodnotu

Zdroj: výstup zo softvéru R

Na uvedenom obrázku môžeme vidieť, že čím vyšší je interval spoľahlivosti, tým je aj širší. Inak povedané, vymieňame našu konfidenciu za presnosť.

Príklad 7.50

V tomto príklade budeme opäť pracovať s údajmi z marketingového prieskumu. V prvej úlohe máme zistiť, či sa logo páči menšiemu podielu obyvateľov v Bratislave, v porovnaní s obyvateľmi mimo Bratislavy. Pomocou testu dvoch podielov teda ideme overiť nulovú hypotézu H_0: $p_{BA} - p_{MBA} \geq 0$, oproti alternatívnej H_1: $p_{BA} - p_{MBA} < 0$.

```
> BA <- subset(logo, subset = region == "BA"); length(BA)
[1] 31
> MBA <- subset(logo, subset = region == "MBA"); length(MBA)
[1] 34
> sum_BA <- sum(BA == "paci sa mi"); sum_BA
[1] 16
> sum_MBA <- sum(MBA == "paci sa mi"); sum_MBA
[1] 19
------------------------------------------------------------
> prop.test(x = c(sum_BA, sum_MBA), n = c(length(BA),
  length(MBA)), alternative = "less", conf.level = 0.95, correct
  = FALSE)

  2-sample test for equality of proportions without continuity
                          correction

data:  c(sum_BA, sum_MBA) out of c(length(BA), length(MBA))
X-squared = 0.1189, df = 1, p-value = 0.3651
alternative hypothesis: less
95 percent confidence interval:
-1.0000000  0.1608112
sample estimates:
   prop 1    prop 2
0.5161290 0.5588235
```

Z 31 respondentov z Bratislavy sa 16 vyjadrilo tak, že sa im logo páči a z 34 respondentov mimo Bratislavy sa logo páčilo 19. Pri teste dvoch podielov sme nevedeli zamietnuť nulovú hypotézu, a teda nemôžeme tvrdiť, že by sa logo malo páčiť menšiemu podielu obyvateľov Bratislavy ako obyvateľom mimo Bratislavy (čo je viditeľné aj zo samotných podielov).

V druhej úlohe máme podľa zadania zistiť, či sa vo veku žien nevyskytujú nejaké odľahlé hodnoty. Keďže podľa zadania môžeme vychádzať z predpokladu normálneho rozdelenia veku žien, môžeme použiť parametrický Grubbsov test. Ten je dostupný prostredníctvom funkcie `grubbs.test()` v knižnici `outliers`.

```
> zeny <- subset(vek, subset = pohlavie == "Z"); length(zeny)
[1] 30
------------------------------------------------------------
> library(outliers)
> grubbs.test(zeny)
```

```
                Grubbs test for one outlier

data:  zeny
G = 2.7536, U = 0.7295, p-value = 0.04831
alternative hypothesis: highest value 92 is an outlier
------------------------------------------------------------
> zeny_2 <- zeny[-23]
> grubbs.test(zeny_2)

                Grubbs test for one outlier

data:  zeny_2
G = 2.1605, U = 0.8273, p-value = 0.3622
alternative hypothesis: highest value 74 is an outlier
```

V celej vzorke respondentov máme k dispozícii 30 žien. Z nich jedna hodnota bola testom vyhodnotená ako extrémna, a to hodnota 92. Nevýhodou tohto testu je samozrejme skutočnosť, že v nulovej hypotéze sa testuje prítomnosť vždy len jednej odľahlej hodnoty. Preto po identifikovaní prvého extrému je nutné túto hodnotu odstrániť a test zopakovať. Ďalšia najvyššia hodnota bola 74, avšak pri danej *p*-hodnote sme túto hodnotu ako extrémnu už nevyhodnotili.

Výsledok overiť aj s použitím neparametrického Hampelovho testu, ktorý, ako sme už viackrát spomínali, je dostupný prostredníctvom funkcie `hampel_identifier()`. Predsa len, pri predošlom teste sme vychádzali z predpokladu o normalite, ktorý nemusí byť správnym.

```
> hampel_identifier <- function(data) {
+  ri <- abs(data - median(data))
+  mad <- median(ri)
+  madn <- mad/0.6745
+  hi <- ri/madn
+  critical <- sqrt(qchisq(0.975,1))
+  data[hi>critical]
+ }
> hampel_identifier(zeny)
[1] 92
```

Aj s použitím neparametrického Hampelovho testu bola len jedna hodnota vyhodnotená ako extrémna, a to hodnota 92.

V ďalšej úlohe zadania máme overiť tvrdenie manažéra, ktorý si myslí, že existuje závislosť medzi pohlavím respondenta a jeho názorom na logo. Najprv overíme hypotézu o existencii závislosti dvoch nominálnych premenných prostredníctvom Pearsonovho Chí-kvadrát testu.

```
> kont <- table(pohlavie, logo)
```

```
> marg <- addmargins(kont); marg
        logo
pohlavie nepaci sa mi paci sa mi Sum
     M             23         12  35
     Z              7         23  30
     Sum           30         35  65
-----------------------------------------------------------------
> chisq.test(kont, correct = F, simulate.p.value = TRUE, B =
  10000)

       Pearson's Chi-squared test with simulated p-value
                 (based on 10000 replicates)

data:  kont
X-squared = 11.6749, df = NA, p-value = 0.0008999
```

Na základe výsledkov testu môžeme považovať tvrdenie manažéra za opodstatnené, keďže nulovú hypotézu o nezávislosti skúmaných dvoch premenných môžeme zamietnuť. Zrejme teda naozaj existuje vzťah medzi pohlavím respondenta a jeho názorom na logo. Otázne ostáva, aký silný tento vzťah je. Keďže pracujeme s nominálnymi premennými, na výpočet závislosti môžeme použiť *Phi* koeficient, Cramer-*V* alebo kontingenčný koeficient. Všetky tieto tri alternatívy sú dostupné prostredníctvom funkcie `assocstats()` z knižnice `vcd`[39].

```
> library(vcd)
> assocstats(kont)
                    X^2 df   P(> X^2)
Likelihood Ratio 12.124  1 0.00049780
Pearson          11.675  1 0.00063347

Phi-Coefficient   : 0.424
Contingency Coeff.: 0.39
Cramer's V        : 0.424
```

Na základe uvedených koeficientov asociácie môžeme vidieť silu vzťahu medzi pohlavím respondenta a jeho názorom na logo. Keďže ide o stredne silnú závislosť, avšak štatisticky významnú, môžeme potvrdiť názor manažéra, že medzi týmito dvoma skúmanými premennými existuje vzťah.

V ďalšej úlohe potrebujeme overiť, či stredná hodnota veku obyvateľov, ktorým sa logo nepáči je väčšia, ako stredná hodnota veku obyvateľov, ktorým sa logo páči. Takto formulovanú hypotézu môžeme overiť prostredníctvom *t*-testu. Najskôr však potrebujeme

[39] Vo funkcii `chisq.test()` sme nastavili parameter `simulate.p-value = TRUE`, čím sme získali výpočet *p*-hodnôt prostredníctvom Monte Carlo simulácie. Počet iterácií (parameter `B`) sme zvolili na 10000. Z tohto dôvodu sa výsledky z tejto funkcie úplne nezhodujú s výsledkami funkcie `assocstats()`.

overiť, či je možné považovať populačné rozptyly za rovnaké pomocou funkcie `var.test()`. Samozrejme, predpokladáme pritom že vek obyvateľov sa riadi normálnym rozdelením pravdepodobnosti. Necháme na čitateľa aby túto skutočnosť overil a prípadne vykonal analýzu bez použitia tohto predpokladu.

```
> vek_logo_1 <- subset(vek, subset = logo == "paci sa mi")
> vek_logo_2 <- subset(vek, subset = logo == "nepaci sa mi")
------------------------------------------------------------
> var.test(vek_logo_1, vek_logo_2, ratio = 1, alternative =
  "two.sided", conf.level = 0.95)

              F test to compare two variances

data:  vek_logo_1 and vek_logo_2
F = 1.0797, num df = 34, denom df = 29, p-value = 0.839
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
0.5235629 2.1817721
sample estimates:
ratio of variances
          1.079723
```

Nulovú hypotézu o rovnosti rozptylov zamietnuť nevieme, takže ich môžeme považovať za rovnaké. Argument funkcie `t.test()` z tohto dôvodu nastavíme na `var.equal = T`. Keďže riešime otázku, či stredná hodnota veku obyvateľov, ktorým sa logo nepáči (premenná `vek_logo_2`) je väčšia, ako stredná hodnota veku obyvateľov, ktorým sa logo páči (premenná `vek_logo_1`), tak ďalší argument vo funkcii `t.test()` nastavíme na `alternative = "less"`. Testujeme teda nulovú hypotézu H_0: $\mu_{\text{vek\_logo\_1}} - \mu_{\text{vek\_logo\_2}} \geq 0$, oproti alternatívnej H_1: $\mu_{\text{vek\_logo\_1}} - \mu_{\text{vek\_logo\_2}} < 0$.

```
> t.test(vek_logo_1, vek_logo_2, alternative = "less", mu = 0,
  var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  vek_logo_1 and vek_logo_2
t = -0.3914, df = 63, p-value = 0.3484
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf 5.64336
sample estimates:
mean of x mean of y
39.97143  41.70000
```

Nulovú hypotézu na základe výsledkov *t*-testu zamietnuť nevieme, pričom výberové priemery v skupinách respondentov sú veľmi podobné. Zrejme sa môžeme prikloniť skôr

k názoru, že nie je rozdiel v stredných hodnotách veku obyvateľov v týchto dvoch skupinách respondentov (takto stanovenú hypotézu by sme exaktne mohli overiť tak, že nastavíme parameter funkcie `t.test()` na `alternative = "two.sided"`).

V úlohe e) potrebujeme overiť ďalšie tvrdenie manažéra, ktorý si myslí, že existuje silná závislosť medzi vekom respondentov a ich názorom na množstvo relevantných informácií reklamnej kampane. Tvrdenie tohto typu môžeme overiť prostredníctvom korelačných koeficientov. Interpretácia korelačných koeficientov však môže byť nejednoznačná, pokiaľ nezadefinujeme hranice sily závislosti. Na tomto mieste budeme za týmto účelom postupovať podľa Hendla (2006), ktorý definuje pásma závislosti nasledovným spôsobom:

- Slabá závislosť – absolútna hodnota korelačného koeficientu je z intervalu <0.1;0.3).
- Stredná závislosť – absolútna hodnota korelačného koeficientu je z intervalu <0.3;0.7).
- Silná závislosť – absolútna hodnota korelačného koeficientu je z intervalu <0.7;1.0>.

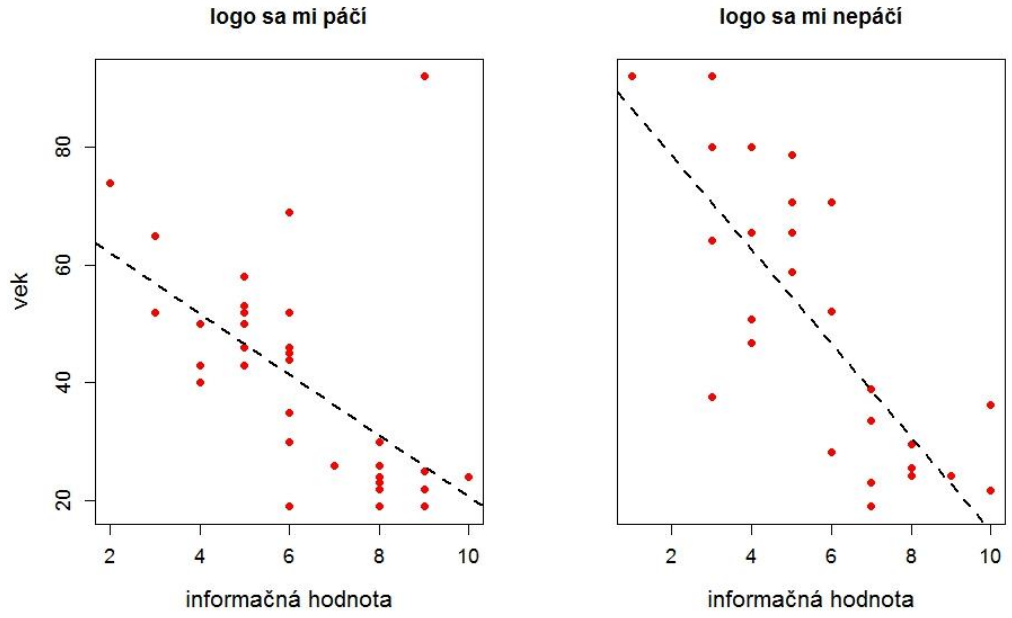
Na výpočet korelačných koeficientov použijeme v tomto príklade funkciu `rcor.test()` z knižnice `ltm`.

```
> library(ltm)
> data <- data.frame(vek, informacna_hodnota)
> rcor.test(data, method = "pearson")

                   vek    informacna_hodnota
vek                 ***** -0.669
informacna_hodnota <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
------------------------------------------------------------
```

Medzi skúmanými premennými existuje záporná korelácia, čo znamená, že čím starší respondent, tým považoval informácie za menej relevantné (a vice versa). Manažér tvrdil, že medzi týmito dvoma premennými existuje silná závislosť, aj keď nepovedal aká.

V ďalšej úlohe máme zostrojiť graf, ktorý by zobrazoval vzťah medzi vekom a názorom na množstvo relevantných informácií v reklame v závislosti od toho, či sa respondentovi páčilo logo alebo nie. Pre jednoduchosť si najprv vytvoríme niekoľko pomocných premenných a až tak prejdeme k samotnej vizualizácii daného vzťahu prostredníctvom *x-y* grafu.

```
> vek_logo_1 <- subset(vek, subset = logo == "paci sa mi")
> vek_logo_2 <- subset(vek, subset = logo == "nepaci sa mi")
> informacna_hodnota_1 <- subset(informacna_hodnota, subset =
  logo == "paci sa mi")
> informacna_hodnota_2 <- subset(informacna_hodnota, subset =
  logo == "nepaci sa mi")
-----------------------------------------------------------------
> par(mfrow = c(1, 2))
> plot(vek_logo_1 ~ informacna_hodnota_1, main = "logo sa mi
  páčí", xlab = "informačná hodnota", ylab = "vek", col = "red",
  pch = 19, cex.axis = 1.1, cex.lab = 1.3)
> abline(lm(vek_logo_1 ~ informacna_hodnota_1),lty = 2, lwd = 2)
> plot(vek_logo_2 ~ informacna_hodnota_2, main = "logo sa mi
  nepáčí", yaxt = "n", xlab = "informačná hodnota", ylab = "",
  col = "red", pch = 19, cex.axis = 1.1, cex.lab = 1.3)
> abline(lm(vek_logo_2 ~ informacna_hodnota_2),lty = 2, lwd = 2)
```



Obrázok 7.45: Graf závislosti veku a vnímaním informačnej hodnoty v reklame

Zdroj: výstup zo softvéru R

Môžeme vidieť, že regresná priamka popisujúca závislosť medzi vekom a názorom na množstvo relevantných informácií má väčší sklon v prípade, ak sa respondentom logo nepáči. Evidentne sa dané logo nepáčilo skôr starším respondentom, ktorí informačnú hodnotu reklamnej kampane tiež nevnímali ako postačujúcu. V oboch prípadoch však stále ide o zápornú koreláciu, čo naznačuje, že informačná hodnota v reklame je viac dostatočné pre mladšiu cieľovú skupinu ako pre staršiu.

V úlohe g) máme zistiť, v akom intervale by sa mala nachádzať stredná hodnota veku zákazníkov. Intervaly pre strednú hodnotu sú súčasťou výstupu funkcie `t.test()`, preto pre jednoduchosť využijeme práve túto funkciu.

```
> t.test(vek, alternative = "two.sided", mu = 0, conf.level =
  0.95)

                     One Sample t-test

data:  vek
t = 18.6433, df = 64, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
36.40059 45.13787
sample estimates:
mean of x
 40.76923
```

Priemerná hodnota veku zákazníkov je 40.77, pričom spodná hranica 95 % konfidenčného intervalu má hodnotu 36.40 a horná 45.14. V tomto intervale by sa teda mala pohybovať stredná hodnota veku zákazníkov s 95 % pravdepodobnosťou.

V predposlednej úlohe máme zodpovedať otázku, či je možné považovať variabilitu názorov na množstvo relevantných informácií z reklamnej kampane v Bratislave a mimo Bratislavy za rovnakú. Keďže v zadaní sa predpokladá normálne rozdelenie údajov, na testovanie zhody dvoch rozptylov môžeme využiť štandardný *F*-test.

```
> info_BA <- subset(informacna_hodnota, subset = region == "BA")
> info_MBA <- subset(informacna_hodnota, subset = region ==
  "MBA")
-------------------------------------------------------------
> var.test(info_BA, info_MBA, alternative = c("two.sided"),
  conf.level = 0.95)

             F test to compare two variances

data:  info_BA and info_MBA
F = 0.9782, num df = 30, denom df = 33, p-value = 0.9556
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
0.482838 2.005900
sample estimates:
ratio of variances
         0.9781658
```

Keďže *p*-hodnota je 0.956, nevieme zamietnuť nulovú hypotézu o rovnosti rozptylov. Môžeme teda súhlasiť s tvrdením zo zadania príkladu, že variabilitu názorov na relevantnosť poskytnutých informácií z reklamy v Bratislave a mimo Bratislavy môžeme považovať za rovnakú.

Ak by sme nepracovali s predpokladom normálneho rozdelenia, môžeme za účelom testovania rovnosti rozptylov využiť neparametrické testy: Levenov a Brown – Forsythov test, ktoré sú obe dostupné prostredníctvom funkcie `leveneTest()` v knižnici `car`.

```
> library(car)
> region_F <- as.factor(region)
-------------------------------------------------------------
> leveneTest(informacna_hodnota, region_F, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  0.1447  0.705
      63
-------------------------------------------------------------
> leveneTest(informacna_hodnota, region_F, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.1507 0.6992
      63
```

Výsledok z neparametrických testov je však rovnaký, v oboch testoch nevieme zamietnuť nulovú hypotézu o rovnosti rozptylov.

Ďalej potrebujeme ešte zistiť, či je rozumné predpokladať, že stredná hodnota názoru na množstvo relevantných informácií je v cieľovej skupine mimo Bratislavy odlišná od názoru cieľovej skupiny v Bratislave. Za týmto účelom môžeme využiť *t*-test, pričom na základe vyššie uvedeného testu rozptylov budeme považovať rozptyly v týchto dvoch skupinách respondentov za rovnaké.

```
> t.test(info_BA, info_MBA, alternative = "two.sided", mu = 0,
  var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  info_BA and info_MBA
t = -0.9905, df = 63, p-value = 0.3257
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-1.5774421  0.5319013
sample estimates:
mean of x mean of y
 5.741935  6.264706
```

Na základe výsledkov *t*-testu môžeme vidieť, že nulovú hypotézu o rovnosti stredných hodnôt sme zamietnuť nevedeli. Stredná hodnota názorov na informačnú hodnotu reklamy teda nie je odlišná v Bratislave a mimo nej.

V poslednej úlohe tohto príkladu nás zaujíma, či podiel ľudí v Bratislave, ktorým sa logo páči, je väčší ako 0.6. Na vyriešenie tejto otázky použijeme test na podiel.

```
> logo_BA <- subset(logo, subset = region != "MBA");
> length(logo_BA)
[1] 31
> logo_BA_p <- subset(logo, subset = region != "MBA" & logo
  !="nepaci sa mi");
> length(logo_BA_p)
[1] 16
-----------------------------------------------------------------
> binom.test(length(logo_BA_p), length(logo_BA), p = 0.6,
  alternative = "greater", conf.level = 0.95)

                    Exact binomial test

data:  length(logo_BA_p) and length(logo_BA)
number of successes = 16, number of trials = 31, p-value =
  0.8716
alternative hypothesis: true probability of success is greater
  than 0.6
95 percent confidence interval:
0.3565721 1.0000000
sample estimates:
probability of success
              0.516129
```

Zo všetkých respondentov žijúcich v Bratislave (31) sa logo páčilo 16. Alternatívnu hypotézu o tom, že podiel ľudí z Bratislavy, ktorým sa logo páči, je väčší ako 0.6, prijať nevieme.

Príklad 7.51

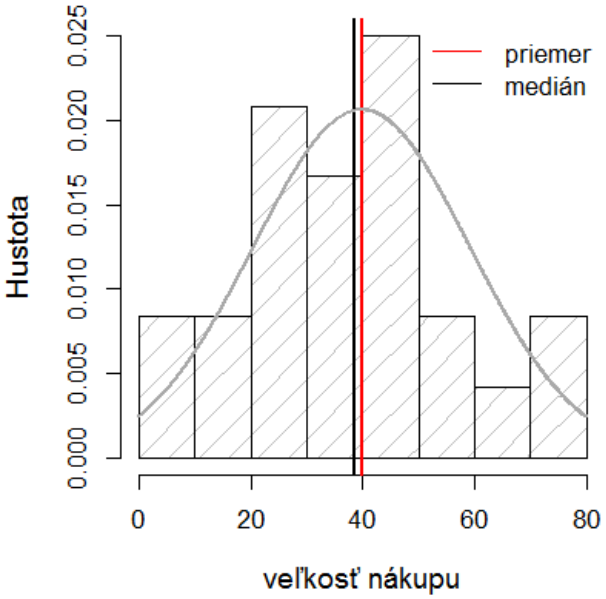
V prvej úlohe zadania tohto príkladu máme rozhodnúť o zošikmení hodnôt veľkosti nákupov na základe priemeru a mediánu. Empirický súbor by mal mať kladnú šikmosť v prípade, že priemer je väčší ako medián, a naopak, ak je priemer menší ako medián súbor by mal mať zápornú šikmosť. Priemer vypočítame prostredníctvom funkcie `mean()`, medián cez funkciu `median()` a šikmosť funkciou `skewness()` z knižnice `moments`.

```
> library(moments)
> mean(nakup)
[1] 39.875
> median(nakup)
[1] 38.5
> skewness(nakup)
[1] 0.2283344
```

Ak by sme teda mali rozhodnúť o zošikmení údajov len na základe priemeru a mediánu, tak môžeme vidieť, že keďže je priemer väčší ako medián, empirický súbor je pravostranne zošikmený. Extrémne kladných hodnôt však nie je príliš veľa, keďže nejde

o výrazné zošikmenie. Túto skutočnosť môžeme pozorovať aj z uvedeného histogramu rozdelenia veľkosti nákupov.

```
> hist(nakup, density = 7, col = "grey", border = "black", main
  = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "veľkosť nákupu")
> abline(v = mean(nakup), lwd = 2, lty = 1, col = "red")
> abline(v = median(nakup), lwd = 2, lty = 1, col = "black")
> legend("topright", legend = c("priemer", "medián"), cex = 1.1,
  ncol = 1, lty = 1, col = c("red", "black"), inset = 0, bty =
  "n")
> x <- seq(0, 80, length = 1000)
> xh <- dnorm(x, mean = mean(nakup), sd = sd(nakup))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "darkgrey", lwd = 2)
```



Obrázok 7.46: Histogram rozdelenia veľkosti nákupov

Zdroj: výstup zo softvéru R

V úlohe b) potrebujeme zistiť, v akom intervale môžeme očakávať, že sa bude nachádzať stredná hodnota nákupu všetkých zákazníkov. Za účelom zistenie intervalov spoľahlivosti pre strednú hodnotu využijeme funkciu `t.test()`.

```
> t.test(nakup, alternative = "two.sided", mu = 0, conf.level =
  0.95)

                    One Sample t-test

data:  nakup
t = 10.0996, df = 23, p-value = 6.339e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
31.70758 48.04242
sample estimates:
mean of x
   39.875
```

Strednú hodnotu nákupu všetkých zákazníkov môžeme očakávať v intervale od 31.71 do 48.04 EUR, čo predstavuje 95 % interval spoľahlivosti. V súvislosti s ďalšou úlohou, v ktorej máme zistiť akú najväčšiu strednú hodnotu objemu nákupov očakávame, by sme mohli vytvoriť pravostranný konfidenčný interval. Túto časť úlohy necháme na čitateľovi.

V úlohe d) máme zistiť, či môžeme predpokladať, že stredná hodnota objemu nákupov u zákazníkov je 50,- EUR. Opäť teda môžeme využiť *t*-test s tým, že keďže už poznáme intervaly spoľahlivosti pre strednú hodnotu, zrejme túto hypotézu budeme môcť zamietnuť. Formálne testujeme H_0: $\mu = 50$ oproti alternatívnej H_1: $\mu \neq 50$.

```
> t.test(nakup, alternative = "two.sided", mu = 50, conf.level =
  0.95)

                    One Sample t-test

data:  nakup
t = -2.5645, df = 23, p-value = 0.01733
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
31.70758 48.04242
sample estimates:
mean of x
   39.875
```

Hypotézu o strednej hodnote rovnej 50,- EUR môžeme zamietnuť na hladine významnosti 5 %. Čo ako sme už uviedli bolo zrejmé, keďže táto hodnota sa nenachádza v 95 % intervale spoľahlivosti pre strednú hodnotu.

Následne potrebujeme zistiť, či zákazníci s vyšším mesačným príjmom majú tendenciu realizovať nákupy vo vyššej hodnote. Inými slovami nás zaujíma vzťah medzi týmito dvoma premennými, ktorý kvantifikujeme s využitím Pearsonovho korelačného koeficientu.

```
> cor.test(nakup, prijem, method = "pearson", alternative =
  "two.sided", exact = TRUE)

            Pearson's product-moment correlation

data:  nakup and prijem
t = 3.9086, df = 22, p-value = 0.0007535
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.3192141 0.8293873
```

```
sample estimates:
      cor
0.6401755
```

Z výsledkov funkcie `cor.test()` vyplýva, že existuje vzťah medzi príjmom zákazníkov a celkovou hodnotou ich nákupov. Ide o priamo úmerný vzťah, keďže korelačný koeficient (jednočíselná charakteristika) je 0.64, pričom tento koeficient je aj štatisticky významný na hladine 1 %.

V ďalšej úlohe máme overiť tvrdenie, že muži a ženy nenakupujú v rovnaké dni. Keďže premenné "`pohlavie`" a "`den`" majú nominálny charakter, toto tvrdenie overíme pomocou kontingenčných korelačných koeficientov.

```
> library(vcd)
> kont <- table(pohlavie, den)
> assocstats(kont)
                     X^2 df P(> X^2)
Likelihood Ratio 4.6409  1 0.031218
Pearson          4.4444  1 0.035015

Phi-Coefficient   : 0.43
Contingency Coeff.: 0.395
Cramer's V        : 0.43
```

Zistili sme, že existuje štatisticky významný pozitívny vzťah medzi pohlavím zákazníkov a dňom kedy realizujú svoje nákupy. Z toho vyplýva, že muži a ženy naozaj nakupujú v rozdielne dni. Aby sme boli presnejší tak treba pripomenúť, že premenná `den` je kódovaná číslom 1, ak sa nákupy realizujú cez víkend a 0, ak nie. Jedno z pohlaví teda nakupuje skôr cez víkend a druhé mimo víkendu. Aby sme zistili, či cez víkend nakupujú skôr ženy alebo skôr muži, stačí sa pozrieť na kontingečnú tabuľku (evidentne muži nakupujú skôr cez víkend a ženy mimo víkendu:

```
> kont
        den
pohlavie  0  1
       0 10  2
       1  5  7
```

V úlohe g) by sme mali zistiť, či je možné tvrdiť na základe našej vzorky, že vo všeobecnosti muži v priemere realizujú rovnako veľké nákupy (v EUR) ako ženy. Tvrdenie tohto typu môžeme overiť pomocou dvojvzorkového *t*-testu, pri ktorom najprv musíme zistiť, či populačné rozptyly môžeme považovať za rovnaké, alebo nie.

```
> nakup_M <- subset(nakup, subset = pohlavie == "1")
> nakup_Z <- subset(nakup, subset = pohlavie == "0")
```

```
------------------------------------------------------------------
> var.test(nakup_M, nakup_Z, alternative = c("two.sided"),
  conf.level = 0.95)

                F test to compare two variances

data:  nakup_M and nakup_Z
F = 1.2872, num df = 11, denom df = 11, p-value = 0.6828
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
0.3705453 4.4712164
sample estimates:
ratio of variances
          1.287163
```

Nulovú hypotézu o rovnosti populačných rozptylov nevieme zamietnuť, takže ich budeme považovať za rovnaké. Len pripomíname, že pracujeme s predpokladom normálneho rozdelenie údajov, avšak ak by tento predpoklad dodržaný nebol, mohli by sme využiť neparametrické testy. Môžeme pristúpiť k testovaniu rovnosti stredných hodnôt nákupov mužov a žien:

```
> t.test(nakup_M, nakup_Z, alternative = "two.sided", mu = 0,
  var.equal = T, conf.level = 0.95)

                     Two Sample t-test

data:  nakup_M and nakup_Z
t = 0.9149, df = 22, p-value = 0.3702
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-9.184361 23.684361
sample estimates:
mean of x mean of y
    43.50     36.25
```

Na základe uvedených výsledkov dvojvzorkového *t*-testu môžeme tvrdiť, že muži v priemere realizujú rovnako veľké nákupy (v EUR) ako ženy, keďže nulovú hypotézu o rovnosti stredných hodnôt sme zamietnuť nevedeli – prikláňame sa teda k platnosti nulovej hypotézy. Presnejšie by samozrejme bolo, že na základe našej vzorky nemôžeme potvrdiť, že existuje rozdiel vo veľkosti nákupov medzi mužmi a ženami.

V ďalšej úlohe budeme opäť využívať *t*-test na overenie tvrdenia, že priemerný mesačný príjem zákazníkov je vo všeobecnosti (teda v populácii) na úrovni 600,- EUR.

```
> t.test(prijem, alternative = "two.sided", mu = 600, conf.level
  = 0.95)
```

```
                    One Sample t-test

data:  prijem
t = -3.3745, df = 23, p-value = 0.002615
alternative hypothesis: true mean is not equal to 600
95 percent confidence interval:
433.9925 560.1742
sample estimates:
mean of x
 497.0833
```

Keďže horná hranica 95 % intervalu spoľahlivosti pre strednú hodnotu príjmu zákazníkov je 560.17 EUR, je úplne zrejmé, že hypotézu o priemernom príjme na úrovni 600 ,- EUR môžeme zamietnuť. Z *p*-hodnoty vidíme, že nulovú hypotézu by sme zamietli aj na hladine významnosti 1 %. Priemerná hodnota príjmu zákazníkov v našej vzorke je na úrovni približne 497,- EUR.

V poslednej úlohe tohto príkladu máme overiť, či je možné mesačný príjem mužov považovať za realizácie z rovnakého rozdelenia pravdepodobnosti ako mesačný príjem žien. Za účelom overenia tohto tvrdenia môžeme využiť Kolmogorov – Smirnovov test (funkcia `ks.test()` z knižnice `Matching`).

```
> prijem_M <- subset(prijem, subset = pohlavie == "1")
> prijem_Z <- subset(prijem, subset = pohlavie == "0")
-------------------------------------------------------------------
> library(Matching)
> ks.test(prijem_M, prijem_Z)

              Two-sample Kolmogorov-Smirnov test

data:  prijem_M and prijem_Z
D = 0.3333, p-value = 0.5176
alternative hypothesis: two-sided

Warning message:
In ks.test(prijem_M, prijem_Z) : cannot compute exact p-values
  with ties
```

Ako však môžeme vidieť, pri výsledkoch z testu dostávame aj výstražné hlásenie. Tento test totiž nepredpokladá rovnosť hodnôt, čiže žiadne dve hodnoty vo výberovom súbore by nemali byť rovnaké. Ak je táto podmienka porušená, kritické hodnoty nie sú presné. Funkcia `ks.test()` nás upozorní na možný problém s rovnakými hodnotami práve týmto výstražným hlásením. Alternatívou je použiť funkciu `ks.boot()` z tej istej knižnice, v ktorej sa počíta upravená verzia Kolmogorov – Smirnovovho testu, kde sa pre potreby počítania kritických hodnôt (a teda aj *p*-hodnoty) využíva bootstraping.

```
> ks.boot(prijem_M, prijem_Z, nboots = 1000)
$ks.boot.pvalue
[1] 0.447

$ks

              Two-sample Kolmogorov-Smirnov test

data:  Tr and Co
D = 0.3333, p-value = 0.5176
alternative hypothesis: two.sided

$nboots
[1] 1000

attr(,"class")
[1] "ks.boot"
```

Výsledky sa nám ani po použití upravenej verzie testu nezmenili, a keďže nevieme zamietnuť nulovú hypotézu o rovnosti distribučných funkcií, môžeme mesačný príjem mužov a žien považovať za realizácie z rovnakého rozdelenia pravdepodobnosti.

Príklad 7.52

V tomto príklade máme zadaný čas zdržania sa zákazníkov pri pokladni v obchode, avšak cez početnosť. Najskôr si teda musíme pripraviť databázu (premenná `data`), s ktorou budeme ďalej pracovať.

```
cas <- c(5, 15, 25, 35, 45, 55, 65)
pocetnost <- c(10, 15, 30, 35, 25, 15, 5)
data <- c(rep(5, 10), rep(15, 15), rep(25, 30), rep(35, 35),
  rep(45, 25), rep(55, 15), rep(65, 5))
```

V prvej úlohe potrebujeme zistiť, či je možné považovať čas zdržania sa pri pokladni za realizáciu z normálneho rozdelenia. Na testovanie normality využijeme tri testy, a to Anderson – Darlingov (funkcia `ad.test()` z knižnice `nortest`), Shapiro – Wilkov (funkcia `shapiro.test()` z knižnice `stats`) a Jarque – Berov test (funkcia `rjb.test()` z knižnice `lawstat`). Pri Jarque – Berovom teste budeme vychádzať z empirických (simulovaných) kritických hodnôt a použijeme obe jeho verzie, t.j. klasickú (`option = "JB"`) aj modifikovanú (`option = "RJB"`).

```
> library(nortest)
> library(stats)
> library(lawstat)
-----------------------------------------------------------------
> ad.test(data)
```

```
                Anderson-Darling normality test

data:  data
A = 2.4671, p-value = 2.883e-06
------------------------------------------------------------------
> shapiro.test(data)

                  Shapiro-Wilk normality test

data:  data
W = 0.9517, p-value = 0.0001121
------------------------------------------------------------------
> rjb.test(data, option = "JB", crit.values = "empirical", N =
  1000)

                      Jarque Bera Test

data:  data
X-squared = 1.731, df = 2, p-value = 0.3458
------------------------------------------------------------------
> rjb.test(data, option = "RJB", crit.values = "empirical", N =
  1000)

                   Robust Jarque Bera Test

data:  data
X-squared = 0.3415, df = 2, p-value = 0.8129
```

Nulovú hypotézu o normálnom rozdelení údajov nevieme zamietnuť pri oboch verziách Jarque – Berovho testu. Ak by sme nepozerali výsledky iných testov, mohli by sme predpokladať, že údaje sú realizáciami z normálneho rozdelenia pravdepodobnosti. Ako naznačili výsledky z ďalších dvoch testov, výsledky nie sú veľmi jednoznačné. Aj keď tieto dva testy sú veľmi citlivé, bezpečnejšie by bolo nepovažovať údaje za realizácie z normálneho rozdelenia.

V ďalšej úlohe máme na základe priemeru a mediánu rozhodnúť o zošikmení údajov a následne zostrojiť histogram.

```
> mean(data)
[1] 33.51852
> median(data)
[1] 35
```

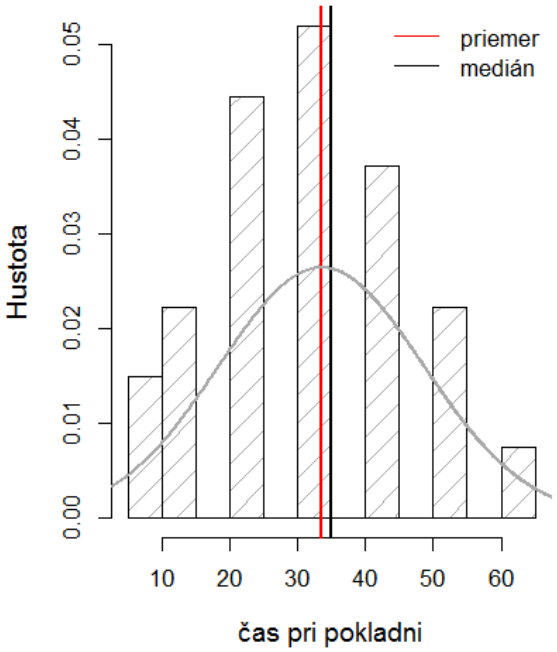
Keďže medián je väčší ako priemer, údaje by mali byť ľavostranne zošikmené, čiže šikmosť by mala byť záporná. Vzhľadom na variabilitu sa rozdiel nezdá byť výrazný. Otázku zošikmenia si vieme overiť aj s použitím funkcie `skewness()`.

```
> library(moments)
> skewness(data)
```

```
[1] -0.006695171
```

Údaje sú teda aj podľa koeficientu šikmosti javia byť ľavostranne zošikmené. Pre lepšiu predstavu si môžeme údaje aj vizualizovať formou histogramu, na ktorom by sme mohli vidieť, že rozdelenie údajov je v zásade symetrické a nemalo by byť veľmi odlišné od normálneho rozdelenia.

```
> hist(data, density = 7, col = "darkgrey", border = "black",
  main = NA, cex.lab = 1.2, cex.axis = 1.1, freq = FALSE, ylab =
  "Hustota", xlab = "čas pri pokladni")
> abline(v = mean(data), lwd = 2, lty = 1, col = "red")
> abline(v = median(data), lwd = 2, lty = 1, col = "black")
> legend("topright", legend = c("priemer", "medián"), cex = 1.1,
  ncol = 1, lty = 1, col = c("red", "black"), inset = 0, bty =
  "n")
> x <- seq(0, 80, length = 1000)
> xh <- dnorm(x, mean = mean(data), sd = sd(data))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "darkgrey", lwd = 2)
```



Obrázok 7.47: Histogram rozdelenia času stráveného pri pokladni

Zdroj: výstup zo softvéru R

Jemné ľavostranné zošikmenie je na danom obrázku pozorovateľné, ale naozaj môžeme vidieť, že rozdelenie údajov je veľmi podobné normálnemu rozdeleniu. Zrejme preto sme dostali aj také nejednoznačné výsledky pri testoch na normalitu.

V poslednej úlohe máme zistiť, v akom intervale by sa mala nachádzať stredná hodnota času zdržania sa zákazníkov pri pokladni. Za týmto účelom využijeme funkciu `t.test()`.

```
> t.test(data, alternative = "two.sided", mu = 0, conf.level =
  0.95)

                    One Sample t-test

data:  data
t = 25.7282, df = 134, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
30.94182 36.09522
sample estimates:
mean of x
 33.51852
```

Stredná hodnota stráveného času pri pokladni všetkých zákazníkov (populácie) by sa mala pohybovať od 30.9 sekúnd do 36.1 sekúnd (čo je hľadaný 95 % interval spoľahlivosti pre strednú hodnotu).

Príklad 7.53

V tomto príklade máme k dispozícii údaje o veku zákazníkov a druhu kúpeného auta, pričom údaje sú pripravené vo forme kontingenčnej tabuľky. Máme zistiť, aký silný je vzťah medzi vekom zákazníkov a druhom auto, ktorý si kúpili.

```
> ctable <- rbind(nizky_vek, stredny_vek, vysoky_vek)
> colnames(ctable) = druh; ctable
            sportove rodinne terenne
nizky_vek         20      30      50
stredny_vek       60     260      45
vysoky_vek        85     220     110
-----------------------------------------------------------------
> library(vcd)
> assocstats(ctable)
                     X^2 df   P(> X^2)
Likelihood Ratio 77.869  4 4.4409e-16
Pearson          79.418  4 2.2204e-16

Phi-Coefficient   : 0.3
Contingency Coeff.: 0.288
Cramer's V        : 0.212
```

Z uvedených koeficientov asociácie je zrejmé, že vzťah medzi vekom a typom vozidla nie je až taký silný. Ak by sme hodnoty považovali za realizácie z náhodného výberu, tak napriek nie veľmi silnému vzťahu by sme mohli tvrdiť, že ide o štatisticky významný vzťah.

Z tabuľky sa zdá, že starší zákazníci preferujú viac rodinné autá. Napríklad v kategórii nízky vek ich je 30 a v kategórii vysoký vek 220, čo je viac ako sedem násobok. Pri športových autách je tento nárast iba štvornásobný. Nárast sa pritom dal očakávať, keďže starší zákazníci zrejme disponujú väčšou kúpnou silou.

Príklad 7.55

V tomto príklade máme zostrojiť funkciu, ktorej výstup bude obsahovať základné opisné charakteristiky (priemer, medián, kvartily, minimum, maximum, výberová smerodajná odchýlka, výberový rozptyl, medzi-kvartilové rozpätie, variačný koeficient, šikmosť, špicatosť) a nejaký zvolený test normality. Možností ako túto funkciu vytvoriť je samozrejme viac, ponúkame jednu z nich (funkciu sme nazvali `summary.stat()`):

```
> library(moments); library(nortest); library(tseries)
> summary.stat <- function(x) {
+
+ vysledky_1 <- matrix(ncol = 1, nrow = 5)
+   vysledky_1[1,1] <- mean(x);
+   vysledky_1[2,1] <- median(x);
+   vysledky_1[3,1] <- quantile(x, prob = 0.25);
+   vysledky_1[4,1] <- quantile(x, prob = 0.75);
+   vysledky_1[5,1] <- IQR(x, na.rm = TRUE, type = 7);
+   rownames(vysledky_1) <- c("priemer", "median",  "dolny
  kvartil", "horny kvartil", "medzi kvart. rozp.")
+   colnames(vysledky_1) <- "hodnota"
+
+ vysledky_2 <- matrix(ncol = 1, nrow = 5)
+   vysledky_2[1,1] <- sd(x);
+   vysledky_2[2,1] <- var(x);
+   vysledky_2[3,1] <- min(x);
+   vysledky_2[4,1] <- max(x);
+   vysledky_2[5,1] <- abs(sd(x)/mean(x))*100;
+   rownames(vysledky_2) <- c("smerod. odch.", "rozptyl",
  "minimum", "maximum", "var. koeficient")
+   colnames(vysledky_2) <- "hodnota"
+
+ vysledky_3 <- matrix(ncol = 1, nrow = 5)
+   vysledky_3[1,1] <- skewness(x);
+   vysledky_3[2,1] <- kurtosis(x);
+   vysledky_3[3,1] <- ad.test(x)$p.value;
+   vysledky_3[4,1] <- shapiro.test(x)$p.value;
+   vysledky_3[5,1] <- jarque.bera.test(x)$p.value[[1]];
+   rownames(vysledky_3) <- c("sikmost", "spicatost","Anderson -
  Darling", "Shapiro - Wilk", "Jarque - Bera")
+   colnames(vysledky_3) <- "hodnota"
+
+   results <- list(); results[["vysledky_1"]] <- vysledky_1;
  results[["vysledky_2"]] <- vysledky_2; results[["vysledky_3"]]
  <- vysledky_3;
+   return(results)
```

```
+ }
```

Ako môžeme vidieť, do funkcie sme zarhnuli všetky požadované opisné charakteristiky a pridali sme tri testy na normalitu (Anderson – Darlingov, Shapiro – Wilkov a Jarque – Berov test), resp. len k nim prislúchajúce *p*-hodnoty. Funkcia si vyžaduje použitie troch knižníc, a to `moments`, `nortest` a `tseries`. Výstup z tejto funkcie sme pre lepšiu prehľadnosť usporiadali do troch objektov (tzv. zoznamov) – `vysledky_1`, `vysledky_2` a `vysledky_3`. Celkový výstup z funkcie vyzerá nasledovne:

```
> summary.stat(A)
$vysledky_1
                    hodnota
priemer            14.57778
median             15.00000
dolny kvartil      11.00000
horny kvartil      18.00000
medzi kvart. rozp.  7.00000

$vysledky_2
                  hodnota
smerod. odch.    5.087815
rozptyl         25.885859
minimum          4.000000
maximum         24.000000
var. koeficient 34.901168

$vysledky_3
                      hodnota
sikmost            -0.1132443
spicatost           2.3737618
Anderson - Darling  0.3041829
Shapiro - Wilk      0.3631088
Jarque - Bera       0.6598423
```

Túto sumarizujúcu funkciu je možné využiť ako vhodnú alternatívu k ostatným funkciám na výpočet deskriptívnej štatistiky v programe `R`, keďže výsledky dostávame v prehľadnejšej forme. Ak by sme chceli vidieť ešte výsledky aj pre druhý vektor hodnôt zo zadania tohto príkladu, stačí nám príkaz `summary.stat(B)`:

```
> summary.stat(B)
$vysledky_1
                    hodnota
priemer            14.15556
median             15.00000
dolny kvartil      12.00000
horny kvartil      16.00000
medzi kvart. rozp.  4.00000

$vysledky_2
```

```
                   hodnota
smerod. odch.     3.037310
rozptyl           9.225253
minimum           7.000000
maximum          19.000000
var. koeficient 21.456665

$vysledky_3
                        hodnota
sikmost              -0.55827519
spicatost             2.67861283
Anderson - Darling    0.03239150
Shapiro - Wilk        0.04874628
Jarque - Bera         0.28206971
```

Príklad 7.56

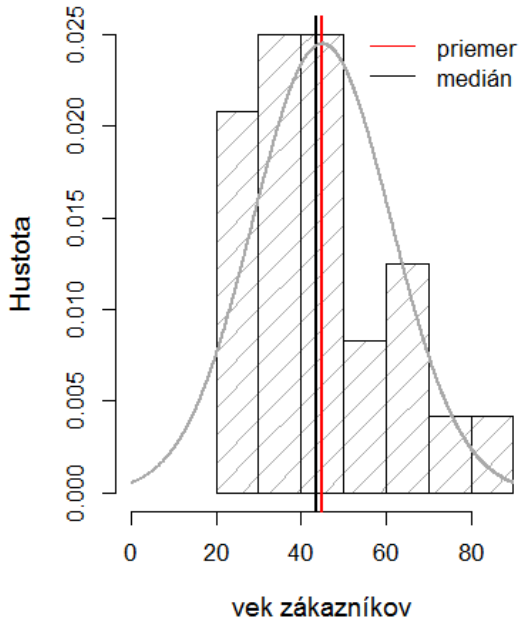
V tomto príklade by sme mali v rámci prvých dvoch úloh vypočítať základné charakteristiky polohy (priemer a medián), kvartily a následne opísať zákazníkov z hľadiska ich veku. Keďže už máme vytvorenú funkciu, ktorá sumarizuje základnú deskriptívnu štatistiku, pri riešení tohto príkladu si pomôžeme práve touto funkciou (ide o funkciu `summary.stat()`, ktorej použitie vyžaduje knižnice, `moments`, `nortest` a `tseries`).

```
> summary.stat(vek)
$vysledky_1
                    hodnota
priemer               45.00
median                43.50
dolny kvartil         32.75
horny kvartil         53.50
medzi kvart. rozp.    20.75

$vysledky_2
                   hodnota
smerod. odch.     16.28016
rozptyl          265.04348
minimum           21.00000
maximum           82.00000
var. koeficient   36.17812

$vysledky_3
                     hodnota
sikmost            0.4902925
spicatost          2.4762617
Anderson - Darling 0.5376181
Shapiro - Wilk     0.4506721
Jarque - Bera      0.5390622
```

Priemerný vek zákazníkov je 45 rokov, 50 % zákazníkov je vo veku 32.75 až 53.5 rokov, jedna štvrtina je mladšia ako 32.75 rokov a jedna štvrtina je staršia ako 53.5 rokov.

Najmladší zákazník vo vzorke mal 21 rokov a najstarší 82 rokov. Priemer a medián sú blízko seba, čo naznačuje značnú symetriu rozdelenia veku zákazníkov, aj keď rozdelenie je mierne pravostranne zošikmené, čo môžeme vidieť aj na histograme. Testy na normalitu sme podľa zadanie príkladu počítať nemuseli, ale môžeme vidieť, že ani v jednom z uvedených testov sme nulovú hypotézu o normalite zamietnuť nevedeli.

```
> hist(vek, density = 7, col = "darkgrey", border = "black",
  xlim = c(0, 90), main = NA, cex.lab = 1.2, cex.axis = 1.1,
  freq = FALSE, ylab = "Hustota", xlab = "vek zákazníkov")
> abline(v = mean(vek), lwd = 2, lty = 1, col = "red")
> abline(v = median(vek), lwd = 2, lty = 1, col = "black")
> legend("topright", legend = c("priemer", "medián"), cex = 1.1,
  ncol = 1, lty = 1, col = c("red", "black"), inset = 0, bty =
  "n")
> x <- seq(0, 90, length = 1000)
> xh <- dnorm(x, mean = mean(vek), sd = sd(vek))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "darkgrey", lwd = 2)
```



Obrázok 7.48: Histogram rozdelenia veku zákazníkov

Zdroj: výstup zo softvéru R

V poslednej úlohe ešte máme overiť, či sú viac ako tri štvrtiny (všetkých) zákazníkov mladších ako 50 rokov. Za týmto účelom využijeme test na podiel, pričom celkový počet zákazníkov mladších ako 50 rokov je v našej vzorke 15 (premenná `vek_50`), z celkového počtu zákazníkov 24.

```
> vek_50 <- sum(vek < "50")
```

```
> binom.test(vek_50, length(vek), p = 0.75, alternative =
  "greater", conf.level = 0.95)

                    Exact binomial test

data:  vek_50 and length(vek)
number of successes = 15, number of trials = 24, p-value =
  0.9453
alternative hypothesis: true probability of success is greater
  than 0.75
95 percent confidence interval:
 0.4371071 1.0000000
sample estimates:
probability of success
                 0.625
```

Na základe realizovaného testu nevieme prijať alternatívnu hypotézu, že viac ako tri štvrtiny všetkých zákazníkov by mali byť mladší ako 50 rokov (nezamietame nulovú hypotézu). Bolo by možné použiť aj inak formulovanú alternatívnu hypotézu, t.j. že viac ako 75 % zákazníkov by malo byť mladších ako 50 rokov. Túto hypotézu by sme prijali. Z aplikačného hľadiska nie je rozdiel v použití jednej alebo druhej formy alternatívnej hypotézy. Formálne by však presnejšia bola (vzhľadom na zadanie) práve druhá formulácia. Dôvody sú nasledujúce. Zaujíma nás, či môžeme považovať viac ako tri štvrtiny zákazníkov za mladších ako 50 rokov. Toto tvrdenie môžeme prijať iba ak ho postavíme do alternatívnej hypotézy. Druhý dôvod je, že ak v alternatívnej hypotéze je ostrá nerovnosť a v zadaní sa pýtame na ostrú nerovnosť, potom by alternatívna hypotéza mala tejto skutočnosti zodpovedať. My sme sa však rozhodli v tomto učebnom texte medzi týmito dvoma alternatívami nerobiť rozdiely.

Príklad 7.57

K dispozícii máme údaje o spokojnosti zákazníkov s predajným personálom, ktorý absolvoval školenie za účelom zlepšenia služieb. V prvej úlohe máme zistiť, či sú zákazníci vo svojich odpovediach viac jednoznační pred tým, ako personál absolvoval školenie, alebo po školení. Potrebujeme teda vykonať test na zhodu rozptylov medzi dvoma realizovanými prieskumami. Najprv však realizujeme test na normalitu, aby sme vedeli, či použijeme parametrické alebo neparametrické testy.

```
> library(nortest)
> library(stats)
> library(lawstat)
> ad.test(pred)
------------------------------------------------------------
                Anderson-Darling normality test
```

```
data:  pred
A = 0.5675, p-value = 0.1235
------------------------------------------------------------------
> shapiro.test(pred)

                  Shapiro-Wilk normality test

data:  pred
W = 0.9133, p-value = 0.06372
------------------------------------------------------------------
> rjb.test(pred, option = "RJB", crit.values = "empirical", N =
  1000)

                    Robust Jarque Bera Test

data:  pred
X-squared = 1.197, df = 2, p-value = 0.2803
```

Na vzorke údajov o spokojnosti zákazníkov pred školením zamestnancov nulovú hypotézu o normálnom rozdelení nevieme zamietnuť na hladine významnosti 5 % pri všetkých troch realizovaných testoch. Jedine v prípade Shapiro – Wilkovho testu by sme nulovú hypotézu mohli zamietnuť na hladine 10 %.

```
> ad.test(potom)

                Anderson-Darling normality test

data:  potom
A = 0.7286, p-value = 0.04856
------------------------------------------------------------------
> shapiro.test(potom)

                  Shapiro-Wilk normality test

data:  potom
W = 0.9228, p-value = 0.09871
------------------------------------------------------------------
> rjb.test(potom, option = "RJB", crit.values = "empirical", N =
  1000)

                    Robust Jarque Bera Test

data:  potom
X-squared = 0.8376, df = 2, p-value = 0.4633
```

Pri testovaní normality vzorky údajov po absolvovanom školení môžeme nulovú hypotézu o normalite zamietnuť na hladine 5 % pri Anderson – Darlingovom teste, na hladine 10 % pri Shapiro – Wilkovom teste a jedine pri Jarque – Berovom teste nulovú hypotézu zamietnuť nevieme. Výsledky testov na normalitu údajov sú teda opäť nejednoznačné. Pre

istotu preto vyskúšame okrem parametrického testu aj neparametrické testy na zhodu rozptylov.

Pri neparametrických testoch (Levenov a Brown – Forsythov test) si najprv vytvoríme pomocnú indikátorovú premennú, ktorá nadobúda hodnoty 1 ak ide o údaje z prvého prieskumu (pred školením) a hodnotu 0 ak ide o údaje z druhého prieskumu (po školení).

```
> library(car)
> data <- c(pred, potom)
> group <- as.factor(c(rep(1, length(pred)), rep(0,
  length(potom))))
-------------------------------------------------------------
> leveneTest(data, group, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  1  5.4655 0.02448 *
      40
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-------------------------------------------------------------
> leveneTest(data, group, center = median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1   3.886 0.05564 .
      40
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pri Levenovom teste môžeme nulovú hypotézu o zhode rozptylov zamietnuť na hladine významnosti 5 % (pri *p*-hodnote 0.024) a pri Brown – Forsythovom teste na hladine 10 % (pri *p*-hodnote 0.056). Zrejme teda rozptyly v týchto dvoch vzorkách rovnaké nie sú. Pri *F*-teste vieme testovať aj jednostrannú hypotézu, na základe ktorej by sme mali vedieť povedať aj to, v ktorej vzorke bol rozptyl väčší.

```
> var.test(pred, potom, ratio = 1, alternative = "greater",
  conf.level = 0.95)

              F test to compare two variances

data:  pred and potom
F = 2.6196, num df = 20, denom df = 20, p-value = 0.0184
alternative hypothesis: true ratio of variances is greater than
  1
95 percent confidence interval:
1.233251      Inf
sample estimates:
ratio of variances
          2.619617
```

Parametrický *F*-test na zhodu rozptylov naznačuje, že rozptyly v týchto dvoch vzorkách nie sú rovnaké. Nulovú hypotézu H_0: $\sigma^2_{pred} / \sigma^2_{potom} \leq 1$ môžeme zamietnuť v prospech alternatívnej H_1: $\sigma^2_{pred} / \sigma^2_{potom} > 1$, z čoho vyplýva, že rozptyl v prvej vzorke je väčší ako rozptyl v druhej vzorke. Môžeme tak prijať tvrdenie, že zákazníci boli vo svojich odpovediach viac jednoznační po tom, ako personál absolvoval školenie.

V ďalšej úlohe máme overiť, či vo všeobecnosti došlo k zlepšeniu hodnotenia správania sa personálu. Potrebujeme teda overiť nulovú hypotézu H_0: $\mu_{pred} - \mu_{potom} \geq 0$, voči alternatívnej H_1: $\mu_{pred} - \mu_{potom} < 0$. Zo zadania príkladu však nie je úplne zrejmé, či respondenti sú tí istí alebo nie. Pre istotu preto použijeme aj párový párový *t*-test (pre prípad, že by sa jednalo o tých istých respondentov). Na základe predchádzajúcej analýzy budeme považovať rozptyly za rozdielne.

```
> t.test(pred, potom, alternative = "less", var.equal = F,
  conf.level = 0.95, paired = FALSE)

                  Welch Two Sample t-test

data:  pred and potom
t = -1.0569, df = 33.327, p-value = 0.1491
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.3719437
sample estimates:
mean of x mean of y
 6.285714  6.904762
-----------------------------------------------------------------
> t.test(pred, potom, alternative = "less", var.equal = F,
  conf.level = 0.95, paired = TRUE)

                      Paired t-test

data:  pred and potom
t = -1.2384, df = 20, p-value = 0.115
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.2431159
sample estimates:
mean of the differences
             -0.6190476
```

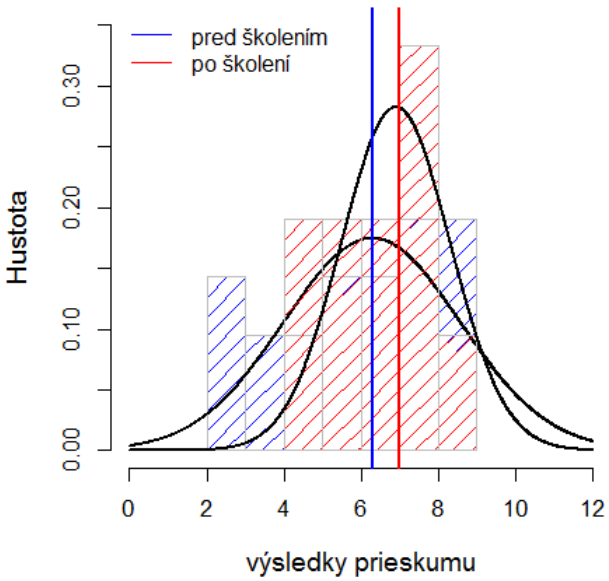
Bez ohľadu na to, či sa oba prieskumy realizovali na tých istých respondentoch, alternatívnu hypotézu prijať nevieme, a preto nemôžeme tvrdiť, že došlo k zlepšeniu hodnotenia personálu po absolvovaní školenia. V predposlednej úlohe máme porovnať dolné kvartily hodnotenia oboch prieskumov a interpretovať výsledky.

```
> quantile(pred, prob = 0.25)
```

```
25%
  5
> quantile(potom, prob = 0.25)
25%
  6
```

V prvom prieskume je jedna štvrtina hodnotení menších ako známka 5 a v druhom prieskume je jedna štvrtina hodnotení menších ako známka 6. Na základe len týchto výsledkov to vyzerá tak, že došlo k zlepšeniu hodnotenia zamestnancov štvrtiny najmenej spokojných zákazníkov. Na záver tohto príkladu ešte máme zodpovedať na otázku, či hodnoty oboch prieskumov pochádzajú z rovnakého rozdelenia pravdepodobnosti. Podľa zadania máme tiež zostrojiť dva prekrývajúce sa histogramy a zistiť, či by sa úloha b) dala riešiť rovnakým spôsobom.

```
> hist(pred, density = 10, col = "blue", border = "grey", main =
  NA, xlim = c(0,12), ylim = c(0, 0.35), cex.lab = 1.2, cex.axis
  = 1.1, freq = FALSE, ylab = "Hustota", xlab = "výsledky
  prieskumu")
> x <- seq(0, 12, length = 1000)
> xh <- dnorm(x, mean = mean(pred), sd = sd(pred))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
> hist(potom, add = TRUE, density = 10, col = "red", border =
  "grey", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE)
> x <- seq(0, 12, length = 1000)
> xh <- dnorm(x, mean = mean(potom), sd = sd(potom))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
> abline(v = mean(pred), lwd = 2, lty = 1, col = "blue")
> abline(v = median(potom), lwd = 2, lty = 1, col = "red")
> legend("topleft", legend = c("pred školením", "po školení"),
  lty = 1.5, col = c("blue", "red"), bty = "n")
```



Obrázok 7.49: Histogram rozdelenia hodnotenia personálu pred školením a po ňom

Zdroj: výstup zo softvéru R

Z histogramov môžeme vidieť, že rozdelenia nie sú úplne rovnaké, aj keď histogramy sa do značnej miery prekrývajú. Po absolvovanom školení je priemerné hodnotenie zamestnancov o niečo lepšie, čo môžeme vidieť aj z mierneho posunutia rozdelenia údajov po školení. Toto posunutie však zrejme nie je dostatočne veľké na to, aby sme vedeli povedať, že došlo k štatisticky významnému zlepšeniu v hodnotení zamestnancov. V úlohe b) sme riešili v zásade rovnakú otázku s využitím *t*-testu.

V kontexte tejto úlohy môžeme tiež použiť Mann – Whitney – Wilcoxonov test (v softvéri `R` dostupný cez funkciu `wilcox.test()`), ak vychádzame z predpokladu, že rozdelenia sú rovnaké. Potom môžeme test použiť na zistenie toho, či jedno rozdelenie je „dostatočne posunuté“ od iného rozdelenia.

```
> wilcox.test(pred, potom, paired = F, alternative = "less")

        Wilcoxon rank sum test with continuity correction

data:  pred and potom
W = 196, p-value = 0.2698
alternative hypothesis: true location shift is less than 0

Warning message:
In wilcox.test.default(pred, potom, paired = F, alternative =
  "less") : cannot compute exact p-value with ties
```

V tomto prípade nemôžeme zamietnuť nulovú hypotézu o rovnosti rozdelení v prospech alternatívy, že miery polohy rozdelenia hodnotenia personálu po absolvovaní školenia sú väčšie, ako miery polohy rozdelenia pred školením.

Alternatívou pri testovaní zhody rozdelení môže byť tiež Kruskal – Wallis test, ktorý tiež vychádza z porovnávania mier polohy.

```
> kruskal.test(data, group)

                Kruskal-Wallis rank sum test

data:  data and group
Kruskal-Wallis chi-squared = 0.3923, df = 1, p-value = 0.5311
```

Výsledky tohto testu však taktiež poukazujú na to, že rozdelenia údajov z oboch prieskumov sú zrejme rovnaké. Poslednou alternatívou na testovanie zhody rozdelení, ktorú si ukážeme v tomto príklade, je Kolmogorov – Smirnovov test.

```
> library(Matching)
> ks.test(pred, potom, alternative = "two.sided")

            Two-sample Kolmogorov-Smirnov test

data:  pred and potom
D = 0.1905, p-value = 0.8407
alternative hypothesis: two-sided

Warning message:
In ks.test(pred, potom, alternative = "two.sided") :
  cannot compute exact p-values with ties
-----------------------------------------------------------------
> ks.boot(pred, potom, alternative = "two.sided", nboots = 1000)
$ks.boot.pvalue
[1] 0.551

$ks

            Two-sample Kolmogorov-Smirnov test

data:  Tr and Co
D = 0.1905, p-value = 0.8407
alternative hypothesis: two.sided

$nboots
[1] 1000

attr(,"class")
[1] "ks.boot"
```

Prostredníctvom rôznych testov sme ukázali, že rozdelenia pravdepodobnosti hodnotenia zamestnancov pred školením a po školení môžeme považovať za rovnaké. Okrem

iného sme tiež zistili, že nedošlo k zlepšeniu v hodnotení personálu zákazníkmi po absolvovaní školenia.

Príklad 7.58

V tomto príklade máme k dispozícii údaje o chybovosti vyrobených výrobkov. V prvej úlohe máme zistiť, či je väčšina hodnôt menších ako aritmetický priemer. Potrebujeme teda vypočítať priemer a medián, na základe ktorých sa vieme vyjadriť aj o zošikmení údajov.

```
> mean(chyby)
[1] 8.225806
> median(chyby)
[1] 8
------------------------------------------------------------------
> library(moments)
> skewness(chyby)
[1] 0.5273668
```

Môžeme vidieť, že väčšina údajov je menšia ako aritmetický priemer, keďže medián je väčší ako priemer. Zároveň vieme povedať aj to, že výberový súbor je pravostranne zošikmený, čo sme pre istotu overili aj výpočtom šikmosti. Ďalej by sme v rámci tejto úlohy mali vypočítať aj modus, teda najpočetnejšiu hodnotu. V programe `R` neexistuje na výpočet modusu žiadna základná funkcia (ako v prípade priemeru alebo mediánu), preto si pomôžeme funkciou[40] `modus()`:

```
> modus <- function(x) {
+   ux <- unique(x)
+   ux[which.max(tabulate(match(x, ux)))]
+ }
> modus(chyby)
[1] 5
```

Alternatívou na výpočet modusu môže byť funkcia `mlv()` z knižnice `modeest`. Pri oboch funkciách dostávame najpočetnejšiu hodnotou v empirickom súbore hodnotu 5.

```
> library(modeest)
> mlv(chyby, method = "mfv")
Mode (most likely value): 5
Bickel's modal skewness: 0.516129
Call: mlv.default(x = chyby, method = "mfv")
```

[40] Funkcia je prevzatá z diskusného fóra na <http://stackoverflow.com/questions/2547402/standard-library-function-in-r-for-finding-the-mode>, dostupná online [22.06.2013].

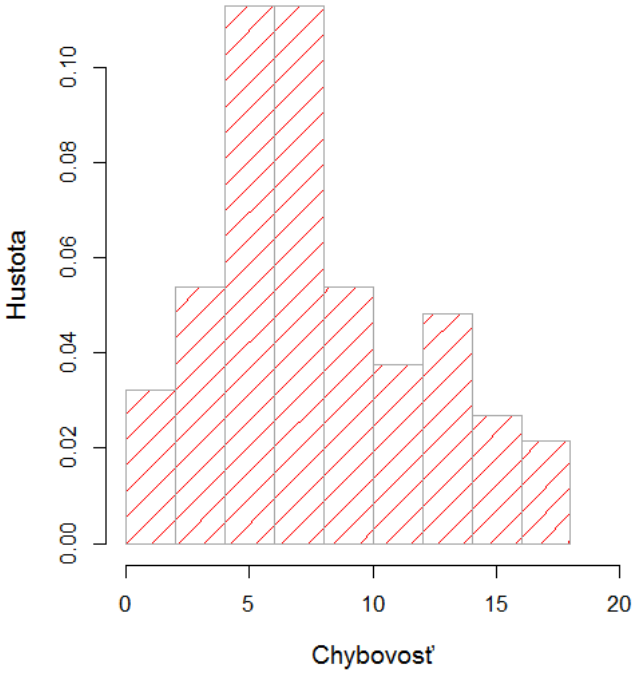
Samozrejme môže sa stať, že rozdelenie údajov nie je unimodálne, čo znamená, že obsahuje viac modusov (tzv. multimodálne). Na overenie tejto skutočnosti môžeme využiť viacero testov, napr. Hartiganov dip test, ktorý je dostupný v knižnici `diptest`.

```
> library(diptest)
> dip.test(chyby, simulate.p.value = TRUE, B = 1000)

   Hartigans' dip test for unimodality with simulated p-value
                (based on 1000 replicates)

data:  chyby
D = 0.0645, p-value = 0.005
alternative hypothesis: non-unimodal, i.e., at least bimodal
```

Na základe výsledkov tohto testu môžeme vidieť, že nulovú hypotézu o unimodálnom rozdelení môžeme zamietnuť na hladine významnosti 1 %, čiže ide o minimálne bimodálne rozdelenie. Možno jednoduchšie riešenie ako zistiť, či súbor je unimodálny alebo nie, je zostrojiť histogram.

```
> hist(chyby, density = 7, col = "red", xlim = c(0, 20), border
  = "darkgrey", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE, ylab = "Hustota", xlab = "Chybovosť")
```

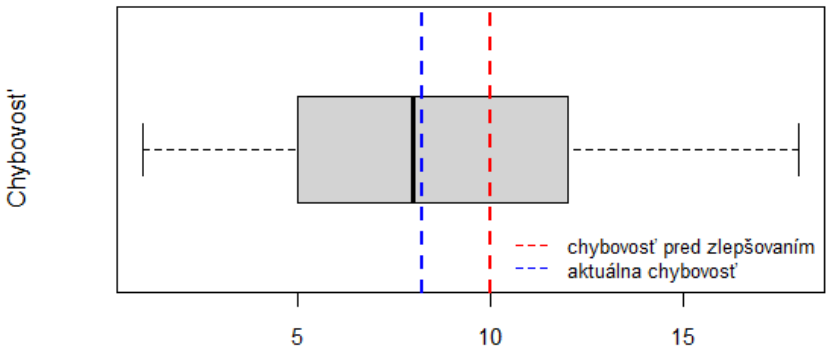
Z uvedeného histogramu je viditeľné, že empirický súbor môže obsahovať aj viac ako len jeden z modusov. Histogram je samozrejme určitým zjednodušením, keďže sa netvorí z pôvodných údajov.



Obrázok 7.50: Histogram rozdelenia chybovosti výrobkov

Zdroj: výstup zo softvéru R

V druhej úlohe máme zostrojiť box – plot a naniesť do neho hodnotu 10, ktorá predstavuje priemerný počet chýb pred zlepšovaním výrobného procesu.



Obrázok 7.51: Box – plot chybovosti výrobného procesu

Zdroj: výstup zo softvéru R

```
> boxplot(chyby, ylab = "Chybovosť", border = "black",
  horizontal = TRUE, col = "lightgrey", pch = 19, cex.axis =
  0.9, cex.lab = 1)
> abline(v = 10, lwd = 2, lty = 2, col = "red")
> abline(v = mean(chyby), lwd = 2, lty = 2, col = "blue")
> legend("bottomright", legend = c("chybovosť pred zlepšovaním",
  "aktuálna chybovosť"), cex = 0.8, ncol = 1, lty = 2, col =
  c("red", "blue"), bty = "n")
```

Z uvedeného box – plotu môžeme vidieť, že aktuálna priemerná chybovosť je nižšia (hodnota 8.23), ako bola priemerná chybovosť pred zlepšovaním výrobného procesu (hodnota 10). Aby sme toto tvrdenie mohli overiť, môžeme použiť jednostranný *t*-test s nulovou hypotézou H_0: $\mu \geq 10$ a alternatívnou H_1: $\mu < 10$.

```
> t.test(chyby, alternative = "less", mu = 10, conf.level =
  0.95)

                    One Sample t-test

data:  chyby
t = -4.0546, df = 92, p-value = 5.251e-05
alternative hypothesis: true mean is less than 10
95 percent confidence interval:
    -Inf 8.95288
sample estimates:
mean of x
 8.225806
```

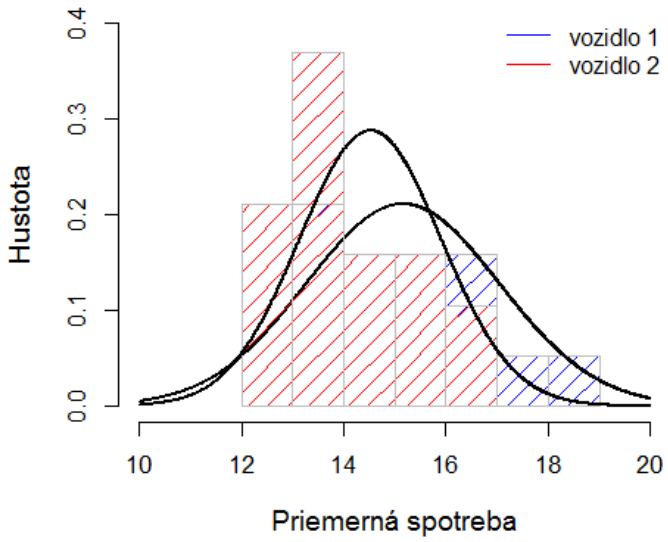
Nulovú hypotézu zamietame na hladine významnosti 1 % v prospech alternatívnej hypotézy, ktorá hovorí o tom, že aktuálna priemerná chybovosť je nižšia ako 10. Môžeme teda prijať tvrdenie, že došlo k zníženiu priemernej chybovosti výrobkov po zlepšení výrobného procesu.

Príklad 7.59

V tomto príklade sa budeme zaoberať priemernými spotrebami dvoch nákladných vozidiel. Pri prvej úlohe máme zostrojiť prekrývajúce sa histogramy.

```
> hist(vozidlo_1, density = 10, col = "blue", border = "grey",
  main = NA, xlim = c(10,20), ylim = c(0, 0.4), cex.lab = 1.2,
  cex.axis = 1.1, freq = FALSE, ylab = "Hustota", xlab =
  "Priemerná spotreba")
> x <- seq(10, 20, length = 1000)
> xh <- dnorm(x, mean = mean(vozidlo_1), sd = sd(vozidlo_1))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
-----------------------------------------------------------------
> hist(vozidlo_2, add = TRUE, density = 10, col = "red", border
  = "grey", main = NA, cex.lab = 1.2, cex.axis = 1.1, freq =
  FALSE)
> x <- seq(10, 20, length = 1000)
> xh <- dnorm(x, mean = mean(vozidlo_2), sd = sd(vozidlo_2))
> dx <- data.frame(x, xh)
> lines(dx, type = "l", col = "black", lwd = 2)
> legend("topright", legend = c("vozidlo 1", "vozidlo 2"), lty =
  1.5, col = c("blue", "red"), bty = "n")
```

Z prekrývajúcich sa histogramov môžeme na prvý pohľad vidieť, že rozdelenia priemernej spotreby oboch nákladných vozidiel sú do značnej miery podobné. V druhej úlohe by sme mali odpovedať na otázku, či sú spotreby podobné. Aby bolo tvrdenie tohto typu overené exaktnejšie (nie len na základe vizualizácie údajov), mali by sme použiť test na zhodu dvoch rozdelení. Keďže rozdelenia priemerných spotrieb sa môžu líšiť v rôznych parametroch, ako vhodná možnosť na vyriešenie takto položenej otázky sa javí Kolmogorov – Smirnovov test. Využijeme priamo alternatívu, ktorá *p*-hodnotu počíta na základe bootsrapingu, keďže pri použití funkcie `ks.test()` by sme dostali výstražné hlásenie kvôli rovnakým hodnotám vo vzorke. Týmto zároveň predpokladáme, že nami namerané hodnoty sú realizáciami z náhodného výberu.



Obrázok 7.52: Histogram rozdelenia priemerných spotrieb nákladných vozidiel

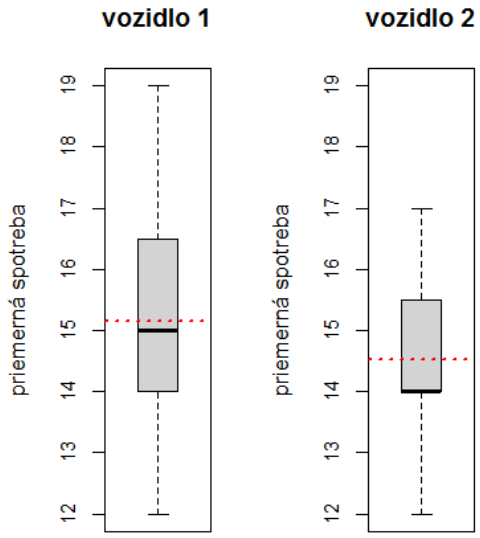
Zdroj: výstup zo softvéru R

```
> library(Matching)
> ks.boot(vozidlo_1, vozidlo_2, nboots = 1000)
$ks.boot.pvalue
[1] 0.798

$ks

                Two-sample Kolmogorov-Smirnov test

data:  Tr and Co
D = 0.1579, p-value = 0.9718
alternative hypothesis: two.sided
$nboots
[1] 1000
attr(,"class")
[1] "ks.boot"
```

Nulovú hypotézu o rovnosti dvoch rozdelení zamietnuť nevieme, čo znamená, že aj na základe testu zhody dvoch rozdelení môžeme tvrdiť, že priemerné spotreby sú podobné. Aj keby sme použili neparametrický Mann – Whitney – Wilcoxonov test, dospeli by sme k rovnakému záveru. A to napriek tomu, že histogram rozdelenia priemernej spotreby prvého vozidla je mierne posunutý (jeho stredná hodnota je vyššia, ako stredná hodnota priemernej spotreby druhého vozidla).

V rámci vizualizácie údajov máme (podľa ďalšej úlohy) zostrojiť ešte box – ploty pre obe priemerné spotreby[41].

```
> par(mfrow = c(1, 2))
> boxplot(vozidlo_1, border = "black", ylab = "priemerná
  spotreba", ylim = c(12, 19), main = "vozidlo 1", col =
  "lightgrey", pch = 19, cex.axis = 0.9, cex.lab = 1.1)
> abline(h = mean(vozidlo_1), lwd = 2, lty = 3, col = "red")
> boxplot(vozidlo_2, border = "black", ylab = "priemerná
  spotreba", ylim = c(12, 19), main = "vozidlo 2", col =
  "lightgrey", pch = 19, cex.axis = 0.9, cex.lab = 1.1)
> abline(h = mean(vozidlo_2), lwd = 2, lty = 3, col = "red")
```



Obrázok 7.53: Box – ploty priemerných spotrieb nákladných vozidiel

Zdroj: výstup zo softvéru R

Pri box – plotoch môžeme vidieť, že vozidlo dva dosahuje o niečo nižšiu priemernú spotrebu (ktoré sú vyznačené červenou prerušovanou čiarou). Táto skutočnosť bola tiež viditeľná už v zobrazených histogramoch. V úlohe d) máme zistiť, či druhé vozidlo naozaj

[41] V tejto publikácii pri vizualizácii údajov používame len niekoľko základných farieb. V programe R ich je samozrejme na výber oveľa viac. V prípade záujmu môže čitateľ nájsť ich názvy napr. v dokumente <http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>, dostupné online [22.06.2013].

dosahuje nižšiu spotrebu (v priemere). Najprv musíme ale overiť, či údaje môžeme považovať za realizácie z normálneho rozdelenia, aby sme podľa toho prispôsobili výber patričných testov. Na testovanie normality využijeme už len Shapiro – Wilkov a modifikovaný Jarque – Berov test.

```
> library(stats)
> library(lawstat)
-----------------------------------------------------------------
> shapiro.test(vozidlo_1)

                  Shapiro-Wilk normality test

data:  vozidlo_1
W = 0.9634, p-value = 0.6418
-----------------------------------------------------------------
> rjb.test(vozidlo_1, option = c("RJB"), crit.values =
  c("empirical"), N = 1000)

                   Robust Jarque Bera Test

data:  vozidlo_1
X-squared = 0.5277, df = 2, p-value = 0.6586
-----------------------------------------------------------------
> shapiro.test(vozidlo_2)

                  Shapiro-Wilk normality test

data:  vozidlo_2
W = 0.9328, p-value = 0.1948
-----------------------------------------------------------------
> rjb.test(vozidlo_2, option = c("RJB"), crit.values =
  c("empirical"), N = 1000)

                   Robust Jarque Bera Test

data:  vozidlo_2
X-squared = 0.3153, df = 2, p-value = 0.8154
```

Pri oboch testoch nevieme zamietnuť nulovú hypotézu o normálnom rozdelení priemerných spotrieb daných vozidiel. Údaje preto môžeme považovať za realizácie z normálneho rozdelenia a na zistenie toho, či druhé vozidlo dosahuje nižšiu spotrebu (v priemere) ako prvé, môžeme použiť štandardný *t*-test. Pred samotným uskutočnením *t*-testu musíme ešte zistiť, či je možné populačné rozptyly považovať za rovnaké.

```
> var.test(vozidlo_1, vozidlo_2, ratio = 1, alternative =
  "two.sided", conf.level = 0.95)

              F test to compare two variances

data:  vozidlo_1 and vozidlo_2
```

```
F = 1.8576, num df = 18, denom df = 18, p-value = 0.1986
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
0.7156655 4.8215092
sample estimates:
ratio of variances
          1.857576
```

Keďže nulovú hypotézu o rovnosti rozptylov zamietnuť nevieme, môžeme ich považovať za rovnaké. Vo funkcii `t.test()` preto nastavíme parameter funkcie na `var.equal = T`, táto možnosť sa prejaví lepším odhadom variability, keďže pri odhade variability dôjde k spojeniu oboch vzoriek.

```
> t.test(vozidlo_1, vozidlo_2, alternative = "greater",
  var.equal = T, conf.level = 0.95)

                    Two Sample t-test

data:  vozidlo_1 and vozidlo_2
t = 1.1723, df = 36, p-value = 0.1244
alternative hypothesis: true difference in means is greater than
  0
95 percent confidence interval:
-0.2779783        Inf
sample estimates:
mean of x mean of y
 15.15789  14.52632
```

Keďže máme zistiť, či druhé vozidlo dosahuje v priemere nižšiu spotrebu, testujeme jednostrannú nulovú hypotézu H_0: $\mu_{vozidlo\_1} - \mu_{vozidlo\_2} \leq 0$, oproti alternatívnej H_1: $\mu_{vozidlo\_1} - \mu_{vozidlo\_2} > 0$. Nulovú hypotézu sme však zamietnuť nevedeli, preto uvedené tvrdenie potvrdiť nemôžeme. V priemere nie sú spotreby týchto dvoch vozidiel štatisticky rozdielne.

V poslednej úlohe máme overiť tvrdenie, že existuje závislosť medzi spotrebou a trasou vozidla. Keďže jednotlivé hodnoty (priemerné spotreby) sú za rôzne trasy, toto tvrdenie môžeme overiť prostredníctvom korelačného koeficientu. Vysoká korelácia by hovorila o tom, že na rovnakých trasách mávajú vozidlá vyššie priemerné spotreby (a vice versa).

```
> library(ltm)
> data <- data.frame(vozidlo_1, vozidlo_2)
> rcor.test(data, method = "pearson")

          vozidlo_1 vozidlo_2
vozidlo_1  *****     0.790
vozidlo_2 <0.001     *****
```

```
upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

Pearsonov korelačný koeficient naznačuje silnú závislosť medzi priemernými spotrebami, ktoré sú usporiadané podľa trás, a teda môžeme potvrdiť, že existuje štatisticky významná (pozitívna) korelácia v priemerných spotrebách medzi jednotlivými trasami.

Príklad 7.61

V tomto príklade máme k dispozícii údaje z dvoch pracovných zmien o priemernom čase výroby jedného výrobku (v minútach). V prvej úlohe máme odhadnúť stredné hodnoty času výroby výrobku pre obe zmeny a porovnať ich.

```
> mean(ranna)
[1] 3.833333
> mean(poobednajsia)
[1] 4.555556
```

Stredné hodnoty sme odhadli prostredníctvom aritmetického priemeru (ako neskresleného estimátora strednej hodnoty). V prvej zmene je priemerná hodnota času potrebného na jeden výrobok nižšia, ako čas potrebný na výrobu výrobku v druhej zmene. Ďalej máme podľa zadania zostrojiť 95 % intervaly spoľahlivosti pre stredné hodnoty. Za týmto účelom využijeme funkciu `t.test()`.

```
> t.test(ranna, alternative = "two.sided", mu = 0, conf.level =
  0.95)

                    One Sample t-test

data:  ranna
t = 12.1419, df = 17, p-value = 8.404e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
3.167242 4.499425
sample estimates:
mean of x
 3.833333
------------------------------------------------------------------
> t.test(poobednajsia, alternative = "two.sided", mu = 0,
  conf.level = 0.95)

                    One Sample t-test

data:  poobednajsia
t = 17.6244, df = 17, p-value = 2.338e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
4.010210 5.100901
```

```
sample estimates:
mean of x
 4.555556
```

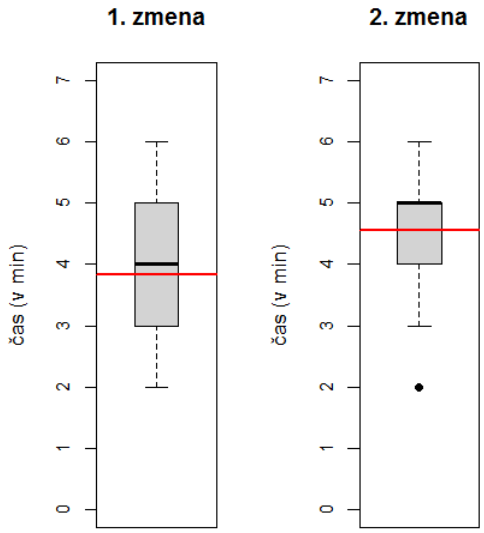
V prvej zmene je dolný 95 % interval spoľahlivosti 3.17 a horný je 4.50, čo znamená, že v priemere môžeme očakávať najnižšiu strednú hodnotu času potrebného na výrobu jedného výroku 3.17 minút a najvyššiu 4.50. V druhej zmene trvá výroba jedného výrobku o niečo dlhšie, dolný interval spoľahlivosti pre strednú hodnotu času je 4.01 a horný 5.10.

V ďalšej otázke by sme mali zodpovedať na otázku, akú najmenšiu strednú hodnotu času výroby v oboch zmenách môžeme očakávať pri $\alpha = 0.01$. Potrebujeme teda zostrojiť dolný 99 % interval spoľahlivosti. Za účelom zopakovania si výpočtu intervalov spoľahlivosti, v tejto úlohe si dolný interval vypočítame manuálne.

```
> mean(ranna) - abs(qt(0.01, df  = length(ranna) - 1)) *
  (var(ranna)/length(ranna))^0.5
[1] 3.022925
> mean(poobednajsia) - abs(qt(0.01, df  = length(poobednajsia) -
  1)) * (var(poobednajsia)/length(poobednajsia))^0.5
[1] 3.892054
```

Pri 99 % konfidencii očakávame najmenšiu strednú hodnotu času výroby v prvej zmene na hodnote 2.91 minút a pri druhej zmene 3.81 minút. Aby boli rozdiely v čase potrebnom na výrobu jedného výroku medzi dvoma zmenami lepšie pozorovateľné, v poslednej úlohe máme zostrojiť box – ploty pre obe zmeny.

```
> par(mfrow = c(1, 2))
> boxplot(ranna, border = "black", ylab = "čas (v min)", ylim =
  c(0, 7), main = "1. zmena", col = "lightgrey", pch = 19,
  cex.axis = 0.9, cex.lab = 1.1)
> abline(h = mean(ranna), lwd = 2, lty = 1, col = "red")
> boxplot(poobednajsia, border = "black", ylab = "čas (v min)",
  ylim = c(0, 7), main = "2. zmena", col = "lightgrey", pch =
  19, cex.axis = 0.9, cex.lab = 1.1)
> abline(h = mean(poobednajsia), lwd = 2, lty = 1, col = "red")
```



Obrázok 7.54: Box – ploty výrobného času jedného výrobku v dvoch zmenách

Zdroj: výstup zo softvéru R

Z box – plotov môžeme vidieť, že mediánová hodnota potrebného času na výrobu jedného výrobku pri prvej zmene je rovná 4 minútam, zatiaľ čo pri druhej zmene je to 5 minút. Pri druhej zmene je spodný kvartil rovný mediánu pri prvej zmene, z čoho vyplýva, že zatiaľ čo 50 % času potrebného na výrobu jedného výrobku pri prvej zmene je kratší ako 4 minúty, tak pri druhej zmene je to len 25 % času. Pri druhej zmene bol čas výroby jedného výrobku 2 minúty vyhodnotený ako odľahlá hodnota.

Príklad 7.62

V tomto príklade sa budeme zaoberať kontrolou dodaných súčiastok, pričom súčiastky sú balené po 1000 kusov v 150 krabiciach. Manažér náhodne vybral 20 krabíc, pri ktorých potrebuje skontrolovať počet súčiastok a 26 súčiastok, pri ktorých kontroluje ich priemer. V prvej úlohe máme zistiť, či stredná hodnota počtu výrobkov v jednej krabici je 1000 ks. Prostredníctvom *t*-testu teda ideme overiť H_0: $\mu = 1000$, oproti alternatívnej H_1: $\mu \neq 1000$.

```
> t.test(pocet_kusov, alternative = "two.sided", mu = 1000,
  conf.level = 0.95)

                    One Sample t-test

data:  pocet_kusov
t = 2.064, df = 29, p-value = 0.04807
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
1000.027 1005.973
sample estimates:
mean of x
     1003
```

Nulovú hypotézu môžeme zamietnuť na hladine významnosti 5 %, preto môžeme predpokladať, že v celej dodávke nie je priemerný počet súčiastok v jednej krabici 1000 kusov.

Analogicky potrebujeme v úlohe b) zistiť, či stredná hodnota všetkých súčiastok v dodávke je 3 cm.

```
> t.test(priemer_hriadela, alternative = "two.sided", mu = 3,
  conf.level = 0.95)

                    One Sample t-test

data:  priemer_hriadela
t = 3.011, df = 25, p-value = 0.005881
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
3.005955 3.031737
sample estimates:
mean of x
 3.018846
```

Z výsledkov vyplýva, že nulovú hypotézu o strednej hodnote rovnej 3 cm zamietame na hladine významnosti 1 %.

V úlohe c) máme overiť podozrenie manažéra, že v celej dodávke je podiel hriadeľov s priemerom väčším ako 3 cm rôzny od 50 %. Inými slovami máme overiť, či polovica hriadeľov má priemer väčší ako 3 cm. Prostredníctvom testu na podiel overujeme H_0: $p_{\text{viac ako 3cm}} = 0.5$, oproti alternatívnej H_1: $p_{\text{viac ako 3cm}} \neq 0.5$.

```
> binom.test(length(priemer_hriadela[(priemer_hriadela>3)]),
  length(priemer_hriadela), p = 0.5, alternative = "two.sided",
  conf.level = 0.95)

                    Exact binomial test

data:  length(priemer_hriadela[(priemer_hriadela > 3)]) and
  length(priemer_hriadela)
number of successes = 19, number of trials = 26, p-value =
  0.02896
alternative hypothesis: true probability of success is not equal
  to 0.5
95 percent confidence interval:
0.5221252 0.8842678
sample estimates:
probability of success
             0.7307692
```

Nulovú hypotézu zamietame, a teda môžeme prijať podozrenie manažéra, že podiel hriadeľov s priemerom väčším ako 3 cm nie je 50 %.

V ďalšej úlohe máme zistiť, či je pravda, že vo vzorke hriadeľov je priemer rovný 3.01 cm. Keďže táto úloha sa netýka celej dodávky (populácie), ale len vzorky, pre jej vyriešenie nám stačí vypočítať priemer.

```
> mean(priemer_hriadela)
[1] 3.018846
```

Keďže priemer hriadeľa nebol jediný parameter, ktorý manažér kontroloval, v poslednej úlohe máme zistiť, či rozptyl priemerov hriadeľa v dodávke je rovný 0.001. Za týmto účelom použijeme funkciu `sigma.test()` z knižnice `TeachingDemos`, prostredníctvom ktorej vieme vypočítať Chí-kvadrát test rozptylu voči konštante.

```
> library(TeachingDemos)
> sigma.test(priemer_hriadela, sigmasq = 0.001, alternative =
  "two.sided", conf.level = 0.95)

         One sample Chi-squared test for variance

data:  priemer_hriadela
X-squared = 25.4654, df = 25, p-value = 0.8731
alternative hypothesis: true variance is not equal to 0.001
95 percent confidence interval:
0.0006265091 0.0019410006
sample estimates:
var of priemer_hriadela
            0.001018615
```

Nulovú hypotézu zamietnuť nevieme, čiže rozptyl priemerov súčiastok v celej dodávke môžeme považovať za 0.001.

Príklad 7.63

K dispozícii máme v tomto príklade údaje o zákazníkoch obchodu z realizovaného prieskumu. V prvej úlohe vedúceho prevádzky zaujímalo, či všetky namerané objemy nákupov je možné považovať za bežné, ale sa vo vzorke vyskytujú aj nejaké extrémne hodnoty. Pri tejto prvej úlohe máme pracovať s predpokladom normality údajov, takže využijeme Grubbsov test na identifikáciu odľahlých hodnôt.

```
> library(outliers)
> grubbs.test(nakup)

                Grubbs test for one outlier

data:  nakup
G = 2.6421, U = 0.6970, p-value = 0.05447
alternative hypothesis: highest value 28 is an outlier
```

Nákup v hodnote 28 EUR bol vyhodnotený na základe tohto testu ako extrémne veľký. Grubbsov test (ako už vieme) však identifikuje vždy len jednu extrémnu hodnotu, takže test by bolo vhodné zopakovať dovtedy, pokiaľ nulovú hypotézu o existencii odľahlej hodnoty nebudeme vedieť zamietnuť.

```
> nakup_2 <- nakup[-7]
> grubbs.test(nakup_2)

                Grubbs test for one outlier

data:  nakup_2
G = 3.0083, U = 0.5894, p-value = 0.00891
alternative hypothesis: highest value 27 is an outlier
-----------------------------------------------------------------
> nakup_3 <- nakup_2[-19]
> grubbs.test(nakup_3)

                Grubbs test for one outlier

data:  nakup_3
G = 3.2476, U = 0.4988, p-value = 0.001809
alternative hypothesis: lowest value 1 is an outlier
-----------------------------------------------------------------
> nakup_4 <- nakup_3[-11]
> grubbs.test(nakup_4)

                Grubbs test for one outlier

data:  nakup_4
G = 2.2432, U = 0.7490, p-value = 0.1929
alternative hypothesis: highest value 19 is an outlier
```

Spolu sme identifikovali 3 extrémne hodnoty, nákupy v hodnote 28, 27 a 1 EUR. V rámci druhej úlohy by sme mali spraviť to isté, avšak bez predpokladu normálneho rozdelenia údajov. Použijeme teda neparametrický Hampelov test.

```
> hampel_identifier <- function(data) {
+  ri <- abs(data - median(data))
+  mad <- median(ri)
+  madn <- mad/0.6745
+  hi <- ri/madn
+  critical <- sqrt(qchisq(0.975,1))
+  data[hi>critical]
+ }
> hampel_identifier(nakup)
[1] 28  1 27
```

Aj bez predpokladu normality sme identifikovali tri extrémne hodnoty, rovnaké ako pri použití Grubbsovho testu.

Ďalej vedúceho prevádzky zaujímalo, aká je sila štatistického vzťahu medzi pohlavím zákazníka a tým, či vlastní auto, resp. medzi tým, či zákazník má auto a tým, či vlastní nehnuteľnosť. Čiže potrebujeme zostrojiť dve kontingečné tabuľky: v prvej budeme skúmať premenné `pohlavie` a `auto`, v druhej premenné `auto` a `nehnutelnost`.

```
> kont_1 <- table(pohlavie, auto)
> kont_2 <- table(auto, nehnutelnost)
> marg_1 <- addmargins(kont_1); marg_1
        auto
pohlavie  0  1 Sum
     0    3  9  12
     1    8  5  13
     Sum 11 14  25
> marg_2 <- addmargins(kont_2); marg_2
     nehnutelnost
auto   0  1 Sum
  0    9  2  11
  1    7  7  14
  Sum 16  9  25
```

Z vytvorených kontingenčných tabuliek potom vypočítame koeficienty asociácie a aj štatistickú významnosť prostredníctvom funkcie `assocstats()` z knižnice `vcd`.

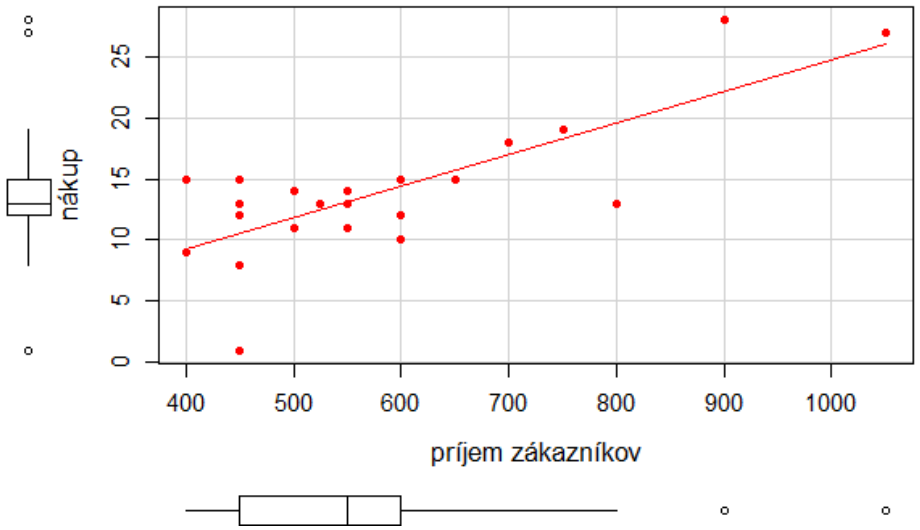
```
> library(vcd)
> assocstats(kont_1)
                    X^2 df P(> X^2)
Likelihood Ratio 3.4772  1 0.062220
Pearson          3.3810  1 0.065952

Phi-Coefficient   : 0.368
Contingency Coeff.: 0.345
Cramer's V        : 0.368
-----------------------------------------------------------------
> assocstats(kont_2)
                    X^2 df P(> X^2)
Likelihood Ratio 2.8317  1 0.092419
Pearson          2.7068  1 0.099924

Phi-Coefficient   : 0.329
Contingency Coeff.: 0.313
Cramer's V        : 0.329
```

Miera závislosti je mierne vyššia v prvej kontingenčnej tabuľke. V oboch prípadoch ide o stredne silné vzťahy, avšak významné len na hladine 10 % (čo je dôsledkom najmä menšieho počtu pozorovaní).

V úlohe d) máme tiež zistiť mieru závislosti medzi premennými (veľkosť nákupu a príjem zákazníka), avšak keďže v tomto prípade už nejde o premenné na nominálnej škále, môžeme použiť štandardné korelačné koeficienty. Najskôr máme zistiť silu lineárneho

vzťahu, čiže použijeme Pearsonov korelačný koeficient. Vzťah medzi skúmanými premennými vizualizujeme prostredníctvom *x-y* grafu.

```
> library(car)
> scatterplot(nakup ~ prijem, smooth = F, xlab = "príjem
  zákazníkov", ylab = "nákup", col = "red", pch = 19, cex.lab =
  1.3, cex.axis = 1.1)
```



Obrázok 7.55: Graf závislosti medzi veľkosťou nákupu a príjmom zákazníkov

Zdroj: výstup zo softvéru R

Keďže sme zvolili vytvorenie *x-y* grafu prostredníctvom funkcie `scatterplot()`, môžeme tiež vidieť na zobrazených box – plotoch identifikované extrémne hodnoty. Pri zisťovaní vzájomnej lineárnej závislosti medzi týmito dvoma premennými sme sa rozhodli tieto hodnoty zo vzorky odstrániť. Upozorňujeme, že to nemusí byť správny postup. Extrémna hodnota pri jednej premennej neimplikuje extrému hodnotu pri usporiadanej *n*-tici, ktorej súčasťou sú aj realizácie premennej, na ktorých sme identifikovali extrémne hodnoty. Ak už ale extrémne hodnoty odstraňujeme, potom ich odstraňujeme ako usporiadanú *n*-ticu (t.j. v našom prípade dvojicu).

```
> data <- data.frame(nakup, prijem)
> data_out <- data[-c(7, 12, 20),];
> data_out
   nakup prijem
1     15    450
2     14    550
3     13    525
4     12    600
5     13    800
6     15    400
8     14    550
9     10    600
```

```
10     8    450
11     9    400
13    11    500
14    13    550
15    15    600
16    18    700
17    19    750
18    15    650
19    11    550
21    12    450
22    13    450
23    14    500
24    12    450
25    14    500
```

Keď už máme odstránené extrémne hodnoty (ktoré sa nachádzali na pozíciách 7, 12 a 20), môžeme pristúpiť k výpočtu Pearsonovho korelačného koeficientu. Keďže potrebujeme overiť aj štatistickú významnosť korelačného koeficientu, na jeho výpočet zvolíme funkciu `rcor.test()` z knižnice `ltm`.

```
> library(ltm)
> rcor.test(data, method = "pearson")

       nakup  prijem
nakup   *****  0.768
prijem <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
------------------------------------------------------------
> rcor.test(data_out, method = "pearson")

       nakup  prijem
nakup   *****  0.488
prijem  0.021  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

Na ukážku sme vypočítali aj korelačný koeficient medzi príjmom a veľkosťou nákupu zákazníkov pred odstránením extrémnych hodnôt. Korelačný koeficient je v tomto prípade 0.768 (významný na hladine 1 %). Po odstránení extrémnych hodnôt môžeme pozorovať výrazné zníženie korelačného koeficientu, až na úroveň 0.488. Významnosť nám taktiež klesla, keďže tento korelačný koeficient je významný už len na hladine 5 %. V tejto úlohe máme tiež overiť silu inej ako lineárnej (monotónnej) závislosti. Pre tento účelom využijeme Spearmanov korelačný koeficient.

```
> rcor.test(data_out, method = "spearman")
```

```
       nakup  prijem
nakup   *****  0.347
prijem  0.113  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values

Warning message:
In cor.test.default(mat[, index[i, 1]], mat[, index[i, 2]], ...)
  : Cannot compute exact p-values with ties
```

Spearmanov korelačný koeficient je nižší ako Pearsonov a zároveň nie je ani štatisticky významný. V úlohe e) vedúceho prevádzky zaujíma, či existuje štatisticky významný rozdiel vo veľkosti strednej hodnoty príjmu mužov a žien v populácii všetkých zákazníkov. Túto úlohu máme riešiť za predpokladu normálneho rozdelenia, čo znamená, že môžeme použiť štandardné parametrické testy: *F*-test na zistenie toho, či môžeme populačné rozptyly považovať za rovnaké alebo nie a potom samotný *t*-test na overenie rozdielov v stredných hodnotách.

```
> prijem_M <- subset(prijem, subset = pohlavie == 1)
> prijem_Z <- subset(prijem, subset = pohlavie == 0)
------------------------------------------------------------
> var.test(prijem_M, prijem_Z, ratio = 1, alternative =
  "two.sided", conf.level = 0.95)

                F test to compare two variances

data:  prijem_M and prijem_Z
F = 1.5678, num df = 12, denom df = 11, p-value = 0.4646
alternative hypothesis: true ratio of variances is not equal to
  1
95 percent confidence interval:
0.4571371 5.2074295
sample estimates:
ratio of variances
          1.567803
```

Na základe výsledkov *F*-testu môžeme považovať populačné rozptyly za rovnaké (nulovú hypotézu sme nezamietli), čomu prispôsobíme argument `var.equal` vo funkcii `t.test()`.

```
> t.test(prijem_M, prijem_Z, alternative = "two.sided", mu = 0,
  var.equal = T, conf.level = 0.95)

                     Two Sample t-test

data:  prijem_M and prijem_Z
t = -0.6184, df = 23, p-value = 0.5424
```

```
alternative hypothesis: true difference in means is not equal to
  0
95 percent confidence interval:
-174.09507   93.96686
sample estimates:
mean of x mean of y
 555.7692  595.8333
```

Nulovú hypotézu o rovnosti stredných hodnôt príjmov mužov a žien zamietnuť nevieme, takže skôr sa prikloníme k názoru, že neexistuje štatisticky významný rozdiel v stredných hodnotách.

V predposlednej úlohe máme zistiť, či podiel všetkých zákazníkov, ktorí vlastnia auto je väčší, ako podiel zákazníkov, ktorí vlastnia nehnuteľnosť. Prostredníctvom testu dvoch podiel overujeme hypotézu H_0: $p_{\text{auto}} - p_{\text{nehnuteľnosť}} \leq 0$, oproti alternatívnej H_1: $p_{\text{auto}} - p_{\text{nehnuteľnosť}} > 0$.

```
> sum_auto <- sum(auto == 1)
> sum_nehnutelnost <- sum(nehnutelnost == 1)
------------------------------------------------------------
> prop.test(x = c(sum_auto, sum_nehnutelnost), n =
  c(length(auto), length(nehnutelnost)), alternative =
  "greater", conf.level = 0.95, correct = FALSE)

        2-sample test for equality of proportions without
                    continuity correction

data:  c(sum_auto, sum_nehnutelnost) out of c(length(auto),
  length(nehnutelnost))
X-squared = 2.0129, df = 1, p-value = 0.07798
alternative hypothesis: greater
95 percent confidence interval:
-0.02715661  1.00000000
sample estimates:
prop 1 prop 2
  0.56   0.36
```

Podiel zákazníkov, ktorí vlastnia auto (0.56) je o niečo vyšší, ako podiel zákazníkov, ktorí vlastnia nehnuteľnosť (0.36). Stanovenú nulovú hypotézu ale môžeme zamietnuť len na hladine významnosti 10 %.

V poslednej úlohe vedúceho prevádzky zaujíma, či vo všeobecnosti (u všetkých zákazníkov) je variabilita príjmov mužov väčšia, ako variabilita príjmov žien. Túto úlohu máme riešiť za predpokladu normálneho rozdelenia, čiže opäť môžeme použiť štandardný *F*-test na zhodu dvoch populačných rozptylov.

```
> var.test(prijem_M, prijem_Z, ratio = 1, alternative =
  "greater", conf.level = 0.95)
```

```
             F test to compare two variances

data:  prijem_M and prijem_Z
F = 1.5678, num df = 12, denom df = 11, p-value = 0.2323
alternative hypothesis: true ratio of variances is greater than
  1
95 percent confidence interval:
0.5624267       Inf
sample estimates:
ratio of variances
          1.567803
```

Už v úlohe e) sme testovali populačné rozptyly príjmov mužov a žien, pričom sme ich vyhodnotili ako rovnaké (používali sme obojstrannú hypotézu). V tomto prípade testujeme jednostrannú nulovú hypotézu H_0: $\sigma^2_{\text{príjem\_M}} / \sigma^2_{\text{príjem\_Z}} \leq 1$, oproti alternatívnej H_1: $\sigma^2_{\text{príjem\_M}} / \sigma^2_{\text{príjem\_Z}} > 1$. Potvrdilo sa nám, že rozptyly príjmov sú zrejme naozaj rovnaké, pretože ani jednostrannú hypotézu sme zamietnuť nevedeli.

Zoznam literatúry

[1] A Free Software Project. Dostupné na: <http://cran.r-project.org/doc/html/interface98-paper/paper_2.html>.

[2] ALLEN, A. O. 1990. *Probability, Statistics, and Queueing Theory with Computer Science Applications. Second Edition*. San Diego : Academic Press, 1990. 768 s. ISBN 978-012051-051-1.

[3] BACON, C. R. 2008. *Practical Portfolio Performance Measurement and Attribution*. *Second Edition*. Chichester : John Wiley & Sons, 2008. 400 s. ISBN: 978-0-470-05928-9.

[4] BANERJEE, A. – CHITNIS, U. B. – JADHAV, S. L. – BHAWALKAR, J. S. – CHAUDHURY, S. 2009. Hypothesis testing, type I and type II errors. In: *Industrial Psychiatry Journal*, 2009, vol. 18, n. 2, p. 127 – 131. ISSN 0972-6748.

[5] BARTELS, R. 1982. The Rank Version of von Neumann's Ratio Test for Randomness. *In: Journal of the American Statistical Association*, 1982, vol. 77, no. 377, p. 40 – 46. ISSN 0162-1459.

[6] BLISCHKE, W. R. – REZAUL KARIM, M. – PRABHAKAR MURTHY, D. N. 2011. *Warranty Data Collection and Analysis*. London : Springer, 2011. 613 s. ISBN 978-085729-646-7.

[7] BOX, G. E. P. 1949. A General Distribution Theory for a Class of Likelihood Criteria. In: *Biometrika*, 1949, vol. 36, n. 3/4, p. 317–346. ISSN 0006-3444.

[8] CLARK-CARTER, D. 2009. *Quantitative Psychological Research. The Complete Student Companion*. *3rd Edition*. New York : Psychology press, 2009. 712 s. ISBN 978-184169-691-1.

[9] COHEN, Y. – COHEN, J. Y. 2008. *Statistics and Data with R: An Applied Approach Through Examples*. Chichester : John Wiley & Sons, 2008. 599 s. ISBN: 978-0-470-75805.

[10] COMREY, A. L. – LEE, H. B. 2009. *Elementary Statistics: A Problem Solving Approach 4th Edition*. Morrisville : lulu.com, 2009, 188 s. ISBN 978-1-4116-6616-0.

[11] COOLICAN, H. 1999. *Research Methods and Statistics (Aspects of Psychology)*. Hodder Education, 1999, 208 s. ISBN 0340748990.

[12] CORDER, G. W. – FOREMAN, D. I. 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Hoboken : John Wiley & Sons, 2009. 247 s. ISBN 978-0-470-45461-9.

[13] CORDER, W. G. – FOREMAN, D. I. 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Hoboken : John Wiley & Sons, 2009, 264 s. ISBN 978-0-470-45461-9.

[14] CRAWLEY, M. J. 2007. *The R Book*. Chichester : John Wiley & Sons, 2007. 950 s. ISBN 978-0-470-51024-7.

[15] CULLEN, A. C. – FREY, H. CH. 1999. *Probabilistic techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York : Plenum Press, 1999. 352 s. ISBN 0-306-45956-6.

[16] DALGAARD, P. 2008. *Introductory Statistics with R: Second Edition*. New York : Springer Science+Business Media, 2008. 363 s. ISBN 978-0-387-79054-1.

[17] DEMING, W. E. 2000. *Out of the Crisis*. Cambridge : The MIT Press, 2000. s. 507. ISBN 978-026254-115-2.

[18] DUFOUR, J. M. – FARHAT, A. – GARDIOL, L. – KHALAF, L. 1998. Simulation-based Finite Sample Normality Tests in Linear Regressions. In: *The Econometrics Journal*, 1998, vol. 1, no. 1, p. 154 – 173. ISSN 1368-423X.

[19] EFRON, B. – TIBSHIRANI, R. J. 1994. *An Introduction to Bootstrap*. New York : Chapman and Hall/CRC, 1994. 456 s. ISBN 978-041204-231-7.

[20] EL-SHAARAWI, A. H. – PIEGORSCH. 2002. *Encyclopedia of Environmetrics. Vol. 2*. New York : John Wiley & Sons, 2002. 2800 s. ISBN 978-047189-997-6.

[21] EVERITT, B. S. – HOTHORN, T. 2005. *HSAUR: A Handbook of Statistical Analyses Using R*. Dostupné na: < http://cran.r-project.org/web/packages/HSAUR/>.

[22] FAY, M. P. – PROSCHAN, M. A. 2010. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. In: *Statistics Surveys*, 2010, vol. 4, p. 1 – 39. ISSN 1935-7516.

[23] FINKELSTEIN, M. O. – LEVEN, B. 2001. *Statistics for Lawyers. Second Edition*. New York : Springer, 2001. 552 s. ISBN 10-038795-007-9.

[24] FISHER, R. A. 1920. A mathematical examination of the methods of determining, by the mean error and by the mean square error. In: *Monthly Notices of the Royal Astronomical Society*, 1920, vol. 80, p. 758 – 770. Dostupné na: <http://articles.adsabs.harvard.edu/full/1920MNRAS..80..758F>.

[25] FLURY, B. – RIEDWYL, H. 1988. *Multivariate Statistics: A practical approach*. London : Chapman & Hall, 1988, 312 s. ISBN 9401070415.

[26] FOX, J. – WEISBERG, S. 2011. *An {R} Companion to Applied Regression. Second Edition*. Thousand Oaks : Sage, 2011. 472 s. ISBN 1-412-97514-X.

[27] FOX, J. 2005. The R Commander: A Basic Statistics Graphical User Interface to R. In: *Journal of Statistical Software*, 2005, vol. 14, no. 9, p. 1 – 45. ISSN 1548-766.

[28] GIBBONS, J. D. – CHAKRABORTI, S. 2003. *Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded*. Basel : CRC Press, 2003. 680 s. ISBN 978-082474-052-8.

[29] GOOD, P. I. – HARDIN, J. 2009. *Common Errors in Statistics (and How to Avoid Them). Third Edition*. New York : John Wiley & Sons, 2009. 288 s. ISBN 978-047045-798-6.

[30] GÜNER, B. – JOHNSON, J. T. 2007. *Comparison of the Shapiro-Wilk and Kurtosis Tests for the Detection of Pulsed Sinusoidal Radio Frequency Interference*. Dostupné na: <http://esl.eng.ohio-state.edu/~rstheory/iip/shapiro.pdf>.

[31] HENDL, J. 2006. *Přehled statistických metod zpracování dat*. Praha : Portál, 2006, 583 s. ISBN 80-7367-123-9.

[32] HERZBERG, F. 1964. The Motivation-Hygiene Concept and Problems of Manpower. In: *Personnel Administration*, 1964, vol. 27, no. 1, p. 3 – 7. ISSN 0031-5729

[33] HILL, T. – LEWICKI, P. 2006. *Statistics: methods and applications. A comprehensive reference for science, industry, and data mining*. Tulsa : StatSoft, 2006. 800 s. ISBN 1-884233-59-7.

[34] HOWELL, D. C. 2010. *Statistical Methods for Psychology. 7th Edition*. Belmont : Cengage Wadsworth, 2010. 768 s. ISBN 10-0-495-59786-4.

[35] HUI, W. – GEL, Y. R. – GASTWIRTH, J. L. 2008. lawstat: An R Package for Law, Public Policy and Biostatistics. In: *Journal of Statistical Software*, 2008, vol. 28, no. 3, p. 1 – 26. ISSN 1548-766.

[36] CHENG, C. H. 2006. *Resampling Methods*. In: S.G. Henderson – B. L. Nelson. 2006. Simulation. Amsterdam : North-Holland, 2006. s. 415 – 454. ISBN 978-0-444-51428-8.

[37] KIRK, R. E. 2008. *Statistics an Introduction. 5th Edition*. Belmont : Thomson Wadsworth, 2008, 649 s. ISBN 0-534-56478-X.

[38] KVAM, P. H. – VIDAKOVIC, B. 2007. *Nonparametric Statistics with Applications to Science and Engineering*. Hoboken : John Wiley & Sons, 2007. 420 s. ISBN 978-0-470-08147-1.

[39] LEHMANN, E. L. 1999. „Student“ and small-sample theory. In: *Statistical Science*, 1999, vol. 14, no. 4, p. 418 – 426. ISSN 0883-4237.

[40] LYÓCSA, Š. – BAUMÖHL, E. – VÝROST, T. 2013. *Kvantitatívne metódy v ekonómii I.* Košice : Elfa, 2013, 201 s. ISBN 978-80-8086-209-1.

[41] MARDIA, K. V. 1970. Measures of Multivariate Skewness and Kurtosis with Applications. In: *Biometrika*, 1970, vol. 57, n. 3, 519 – 530, ISSN 0006-3444.

[42] MAUCHLY, J. W. 1940. Significance Test for Sphericity of a Normal n-Variate Distribution. In: *The Annals of Mathematical Statistics*, 1940, vol. 11, n. 2, s. 204 – 209, ISSN 0003-4851.

[43] MONTGOMERY, D. C. – RUNGER, G. C. 2011. *Applied Statistics and Probability for Engineers. 5th Edition*. New York : John Wiley & Sons, 2011. 768 s. ISBN 978-0-470-05304-1.

[44] MOORE, D. S. – McCABE, G. P. 2009. *Introduction to the Practice of Statistics. 6th Edition*. New York : W. H. Freeman and Company, 2009. 709 s. ISBN 978-1-429-21622-7.

[45] NEWCOMBE, R. G. 1998. Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. In: *Statistics in Medicine*, 1998, vol. 17, no. 8, p. 857 – 872. ISSN 0277-6715.

[46] PANIK, M. J. 2005. *Advanced Statistics from Elementary Point of View*. London : Elsevier Academic Press, 2005. 824 s. ISBN 978-0-120-88494-0.

[47] Program R. Dostupný na: <http://cran.at.r-project.org/bin/windows/base/>.

[48] RAMACHANDRAN, K. M. – TSOKOS, CH. P. 2009. *Mathematical Statistics with Applications*. London : Elsevier Academic Press, 2009. 824 s. ISBN 978-0-12-374848-5.

[49] REICHENBÄCHER, M. – EINAX, J. W. 2011. Challenges in Analytical Quality Assurance. London : Springer, 2011. 375 s. ISBN 978-364216-594-8.

[50] RENCHER, A. C. 2002. *Methods of multivariate analysis*. Hoboken : John Wiley & Sons, 2002, 715 s. ISBN 0-471-41889-7.

[51] ROSENTHAL, R. – RUBIN, D. B. 1982. A simple, general purpose display of magnitude of experimental effect. In: *Journal of Educational Psychology*, 1982, vol. 74, no. 2, p. 166 – 169. ISSN 0022-0663.

[52] ROYSTON, P. 1992. Approximating the Shapiro-Wilk W-test for non-normality. In: *Statistics and Computing*, 1992, vol. 3, no. 3, p. 117 – 119. ISSN 0960-3174.

[53] SALKIND, N. J. 2007. *Encyclopedia of Measurement and Statistics. Vol. 2*. London : SAGE Publications, 2007. 1136 s. ISBN 10-1-412-91611-9.

[54] SHIH, W. J. – HUANG, W. M. 1992. Evaluating Correlation with Proper Bounds. In: *Biometrics*, 1992, vol. 48, no. 4, p. 1207 – 1213. ISSN 0006-341X.

[55] SIEGEL, J. 1956. *Nonparametric Statistics for The Behavioral Sciencies*. London : McGraw-Hill Book Company, 1956. 312 s. ISBN 10-0-070-57348-4.

[56] SMITH, M. L. – GLASS, G. V. 1977. Meta-Analysis of Psychotheraphy Outcome Studies. In: *American Psychologist*,1977, vol. 32, no. 9, p. 752 – 760. ISSN 0003-066X.

[57] SPRENT, P. – SMEETON, N. C. 2000. *Applied Nonparametric Statistical Methods. 3rd Edition*. London : Chapman & Hall/CRC, 2000. 480 s. ISBN 1-58488-145-3.

[58] STUDENT (GOSSET, W. S.). 1908. The Probable Error of Mean. In: *Biometrika*, 1908, vol. 6, no. 1, p. 1 – 25. ISSN 0006-3444.

[59] THADEWALD, T. – BÜNING, H. 2007. Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison. In: *Journal of Applied Statistics*, 2007, vol. 34, no. 1, p. 87 – 105. ISSN 0266-4763.

[60] THOMPSON, M. – LOWTHIAN, P. J. 2011. *Notes on Statistics and Data Quality for Analytical Chemists*. London : Imperial College Press, 2011. 260 s. ISBN 978-184816-617-2.

[61] TIMM, N. H. 2002. *Applied multivariate analysis*. New York : Springer Verlag, 2002, 719 s. ISBN 0387953477.

[62] TKÁČ, M. 2001. *Štatistické riadenie kvality*. Bratislava : Ekonóm, 2001. 312 s. ISBN 80-225-0145-X.

[63] URZÚA, C. M. 1996. On the correct use of omnibus tests for normality. In: *Economics Letters*, 1996, vol. 53, no. 3, p. 247 – 251. ISSN 0165-1765.

[64] VAN DEN HONERT, R. 1997. *Intermediate Statistical Methods for Business and Economics*. Cape Town : University of Cape Town Press, 1997. 390 s. ISBN 978-191971-309-0.

[65] VERMA, S. P. – QUIROZ-RUIZ, A. – DÍAZ-GONZÁLEZ, L. 2008. Critical values for 33 discordancy test variants for outliers in normal samples up to sizes 1000, and applications in quality control and Earth Sciences. In: *Revista Mexicana de Ciencias Geológicas*, 2008, vol. 25, no. 1, p. 82 – 96. ISSN 1026-8774.

[66] VERMA, S. P. – QUIROZ-RUIZ, A. 2006. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and application in sciene and engineering. In: *Revista Mexicana de Ciencieas Geológicas*, 2006, vol. 23, n. 2, p. 133 – 161. ISSN 1026-8774.

[67] VERMA, S. P. – QUIROZ-RUIZ, A. 2008. Critical values for 33 discordancy test variantas for outliers in normal samples of very large sizes from 1,000 to 30,000 and evaluation in different regression models for the interpolation and extrapolation of

critical values. In: *Revista Mexicana de Ciencias Geológicas*, 2008, vol. 25, no. 3, p. 369 – 381. ISSN 1026-8774.

[68] VERZANI, J. 2004. *Using R for Introductory Statistics*. London : Chapman and Hall/CRC, 2004, 432 s. ISBN 978-158488-450-7.

[69] WILCOX, R. R. 2012. *Introduction to Robust Estimation and Hypothesis Testing. 3rd Edition*. Waltham : Academic Press, 2012. 608 s. ISBN 978-012386-983-8.

[70] YU, K. – SHARP, I. – JAY GUO, Y. 2009. *Ground-Based Wireless Positioning*. Chichester : John Wiley & Sons, 2011. 450 s. ISBN 978-047074-704-9.

Zoznam obrázkov

Zoznam tabuliek

Zoznam programových knižníc

[1] strucchange 1.4-7
[2] MASS 7.3-19
[3] UsingR 0.1-18
[4] ggplot2 0.9.1
[5] lattice 0.20
[6] triangle 0.5
[7] mvtnorm 0.9-9992
[8] car 2.0-12
[9] lmtest 0.9-30
[10] psych 1.2.4
[11] zoo
[12] datasets 2.14.0
[13] moments_0.13
[14] TeachingDemos_2.8
[15] Matching_4.8-0
[16] nortest_1.0-2
[17] lawstat_2.3
[18] outliers_0.14
[19] tseries_0.10-29
[20] vcd_1.2-13
[21] ellipse_0.3-7
[22] HSAUR_1.3-2
[23] alr_2.0.0
[24] stats_2.15.2
[25] ltm_0.9-9
[26] modeest_2.1
[27] diptest_0.75-4

Názov: Kvantitatívne metódy v ekonómii II.
Autori: Štefan Lyócsa, Eduard Baumöhl, Tomáš Výrost
Vydavateľstvo: elfa, s.r.o., Park Komenského 7, 040 01 Košice
Vydanie: prvé
Tlač: elfa, s.r.o., Park Komenského 7, 040 01 Košice

ISBN 978-80-8086-210-7

ISBN 978-80-8086-210-7