Štefan Lyócsa – Eduard Baumöhl – Tomáš Výrost

Kvantitatívne metódy v ekonómii I.

ELFA 2013

Štefan Lyócsa
Eduard Baumöhl
Tomáš Výrost

Kvantitatívne metódy v ekonómii I.

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

doc. Ing. Vladimír Gazda, PhD.

Katedra financií, Ekonomická fakulta, Technická univerzita v Košiciach

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: 9.7 AH

ISBN 978-80-8086-209-1

O B S A H

Úvod

Cieľom týchto skrípt je poskytnúť študijný materiál študentom predmetu Kvantitatívne metódy v ekonómii I (KMvE I, ktorý sa v súčasnosti ponúka ako predmet pre študentov programu Honoris na inžinierskom stupni štúdia, Podnikovohospodárskej fakulty so sídlom v Košiciach, Ekonomickej univerzity v Bratislave), ako aj našim bakalárom a diplomantom, ktorí sa rozhodli vo svojej práci aplikovať niektoré základné kvantitatívne metódy používané v ekonómii. Pri tvorbe obsahovej náplne predmetu KMvE I, ako aj tejto publikácie, bolo cieľom prezentovať študentom jednoduché metódy spracovania a vizualizácie dát, a taktiež ich zároveň pripraviť na prácu v programe R.

Pri tvorbe materiálov sme vychádzali z potreby poskytnúť študentom aplikačný pohľad na problematiku spracovania údajov, pričom sme sa snažili formálnu matematickú štatistiku obmedziť na nevyhnutné minimum. Táto snaha na mnohých miestach nevyhnutne vedie k určitým matematickým nepresnostiam, ktoré snáď budú vyvážené intuitívnejším pohľadom do problematiky kvantitatívneho spracovania údajov. Ak čitateľ hľadá formálnejší prístup k spracovaniu údajov, musí siahnuť po iných publikáciách.

Túto publikáciu je potrebné chápať ako dokument, ktorý je určený potrebám pomerne úzkej skupine študentov. Výklad je na hodinách dopĺňaný a komentovaný. Preto by tieto skriptá nemali vzbudzovať dojem ucelenej učebnice. Ide len o skriptá – pomocný študijný materiál – a aj keď súhrne predstavujú originálny text, v mnohých prípadoch sme sa nechali inšpirovať inými publikáciami (uvedené v literatúre) a myšlienkami, ktoré nám ešte ako študentom boli vštepované našimi mentormi.

Skriptá sa skladajú zo štyroch hlavných častí a príkladov. V prvej vymedzíme niektoré základné pojmy tak, ako ich budeme používať v publikácii. Druhá je štandardným úvodom do opisu dát.

V tretej časti sa začíname venovať programu R. Zaujímajú nás spôsoby importovania dát, zadávania príkazov, manipulácie s rôznymi objektmi, logické operátory, používanie základných metód opisu dát a významný priestor venujeme vizualizácií dát. Jednotlivé kódy z programu R sa snažíme písať čo najjednoduchšie, priam naivne, keďže náročnejšie úlohy nás čakajú až v ďalších publikáciách, ktoré na tieto skriptá nadväzujú a zručnosti chceme budovať radšej pomaly. Z hľadiska používania programu R je cieľom tejto publikácie, aby bol študent po absolvovaní tejto publikácie v programe R v prvom rade zorientovaný.

V štvrtej časti poskytujeme stručný úvod do teórie pravdepodobnosti. Dôraz kladieme najmä na rôzne druhy rozdelení, s ktorými sa pri kvantitatívnom spracovaní údajov najčastejšie stretávame.

V poslednej časti sú zadania a riešenia k vybraným príkladom. Tieto príklady sme sa snažili koncipovať pomerne všeobecne. Nebolo cieľom zostaviť čo najväčší počet príkladov s jednoznačným zadaním a krátkym riešením. Cieľom bolo vytvoriť príklady z údajov, ktoré majú svoj príbeh tak, aby sa študent mohol v budúcnosti v podobných situáciách ľahšie zorientovať. Aby pochopil, ako zobraziť údaje, a ktoré opisné charakteristiky mu pomôžu pochopiť problém, ktorý práve rieši. Tento charakter príkladov sme sa snažili dodržať aj v celom texte, keďže aplikácia kvantitatívnych metód je hlavným motívom týchto skrípt.

1 Základné pojmy

Všeobecne sa na štatistiku môžeme pozerať ako na vedný odbor, ktorý poskytuje nástroje na výber, zber, organizovanie, triedenie, prezentáciu, ale najmä analýzu údajov. Pre potreby ekonómov slúži štatistika na získavanie informácií s cieľom lepšie pochopiť ekonomickým situáciám, s ktorými sa stretávajú. Môžeme rozlišovať medzi dátami (údajmi) a informáciami. Pod **dátami** budeme rozumieť každú správu, bez ohľadu na to, či má pre prijímateľa správy význam alebo nie. Na rozdiel od údajov majú **informácie** pre pozorovateľa určitý význam, alebo možno presnejšie, hodnotu.

Príklad 1.1

Ak 10 zákazníkov vyjadrilo mieru svojej spokojnosti s novým produktom nasledovnými slovnými výrazmi:

spokojný, veľmi spokojný, nespokojný, nespokojný, spokojný, veľmi spokojný, veľmi spokojný, spokojný, veľmi nespokojný, spokojný,

potom tie predstavujú pre nás dáta, z ktorých vhodnou úpravou (vizualizácia prostredníctvom vhodného obrázku: histogram, koláčový graf, stĺpcový graf, bodový graf, *x-y* graf) alebo transformáciou (priradením číselnej hodnoty k slovám: veľmi nespokojný – 1, nespokojný – 2, spokojný – 3, veľmi spokojný – 4) a následnými výpočtami (priemer, medián, variabilita) získame informácie, ktoré rozširujú naše poznanie o danej problematike, a tak môžu prispieť k prijímaniu lepších rozhodnutí.

Predmetom štatistiky budeme nazývať **štatistické jednotky**, ktoré sú nositeľmi javov, ktorých skúmanie je spravidla zámerom štatistickej analýzy. Pri sledovaní vývoja HDP na Slovensku v určitom období budeme pod štatistickou jednotkou chápať Slovensko. Všimnime si, že náš problém sme v predošlom príklade vymedzili z troch hľadísk: **vecného**, **priestorového** a **časového hľadiska**. V našom prípade vecnému vymedzeniu zodpovedá HDP (teda to, čo meriame), priestorovému Slovensko (kde) a časovému sledované obdobie (kedy).

Štatistickým jednotkám môžeme priradiť rôzne **štatistické znaky**, ktoré reprezentujú ich vlastnosť. Zoberme si napríklad podnik v určitom čase a odvetví pôsobiaci na Slovensku. Jeho štatistickým znakom môže byť: vybraný finančný ukazovateľ, počet zamestnancov, trhová pozícia, konkurenti, produkty a produktové portfólio, úroveň ponúkaných služieb a podobne. Ak nás zaujíma správanie sa spotrebiteľov (štatistická jednotka), pod ich štatistickými znakmi môžeme rozumieť: veľkosť nákupu, čas nákupu, obsah nákupu,

pohlavie, vek, miesto nákupu, rodinný stav, príjmová skupina a iné. V tejto publikácii sa nebudeme detailnejšie venovať postupom zbierania a úpravy dát. Pre tieto účely odporúčame publikácie zaoberajúce sa induktívnou štatistikou.

Dáta získané zo štatistického pozorovania tak, ako boli namerané, budeme nazývať **prvotné**. Hodnoty jednotlivých štatistických znakov budeme súhrne nazývať **štatistickými súbormi** daných štatistických znakov. Počet hodnôt v štatistickom súbore predstavuje **rozsah štatistického súboru**. Pri meraní HDP za obdobie 15 po sebe nasledujúcich rokov, je rozsah štatistického súboru absolútnej výšky HDP $n = 15$. Malým písmenom n budeme označovať rozsah štatistického súboru (niekedy sa zvykne používať aj N a pri časových radoch aj T). Ak by sme merali percentuálne zmeny za sebou nasledujúcich pozorovaní za dané obdobie, potom je rozsah súboru $n = 14$. Hodnoty štatistického znaku budeme všeobecne označovať ako X a individuálne hodnoty štatistického znaku ako X_i, kde index i reprezentuje konkrétne hodnoty, pričom $i = 1, 2, ..., n$ (keďže v štatistickom súbore je n hodnôt). Pri veľkom množstve údajov je neraz vhodná úprava zotriedenie prvotných údajov do tzv. **variačného radu**. Vtedy namerané hodnoty X_i zotriedime podľa nami definovaného pravidla, napríklad od najmenšej hodnoty po najväčšiu hodnotu. Takto usporiadané hodnoty označujeme ako $X_{(i)}$. Index i je v zátvorke, čo znamená, že ide o usporiadanie hodnôt štatistického súboru. Pokiaľ nebude povedané inak, ide o vzostupné usporiadanie. V našom príklade vzostupného usporiadania tak platí $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$.

Zjednodušene môžeme tvrdiť, že ak sa hodnoty štatistického znaku menia, potom ide o **premennú**. Aj keď to nie je úplne presné, pojem štatistický znak budeme v ďalšom texte často voľne zamieňať s pojmom **premenná**. Špecifickým prípadom premennej je tzv. indikátorová premenná, ktorá nadobúda iba dva stavy: 0 ak k danému javu nedochádza a 1 ak áno. Tradičným príkladom je modelovanie nákupu. Návštevník webovej stránky môže uskutočniť nákup tovaru (čo si označíme číslom 1) alebo nemusí (čo si pre zmenu označíme ako 0). Ďalším príkladom je pohlavie zákazníka (0 – muž a 1 – žena). Našou snahou môže byť potom zistenie, ktoré iné premenné súvisia s kúpou tovaru. Ak zistíme, o ktoré premenné ide, potom naše ďalšie kroky môžu viesť k snahe tieto premenné ovplyvňovať – riadiť. Premenná pritom môže byť kvantitatívneho alebo kvalitatívneho charakteru. O **kvantitatívnych** premenných budeme hovoriť vtedy, ak štatistický znak nadobúda číselné hodnoty. V takom prípade sa zvykne štatistický znak ďalej členiť na spojitú, resp. diskrétnu premennú. O **spojitej** premennej hovoríme, ak hodnoty štatistického znaku môžu nadobúdať ľubovoľnú hodnotu medzi dvoma reálnymi číslami, potom na danom intervale je premenná spojitá. O **diskrétnej** premennej budeme hovoriť, ak medzi dvoma reálnymi číslami

premenná nemôže nadobúdať ľubovoľnú hodnotu. V takom prípade je na danom intervale premenná diskrétna. Ide o pomerne voľné definície. Presnejšie o diskrétnej premennej budeme hovoriť, ak dokážeme hodnoty, ktoré nadobúda, očíslovať prirodzenými číslami (napr. 1, 2, 3, …). Formálne by sme mohli povedať, že premenná je diskrétna, ak je jej obor hodnôt konečný, alebo má rovnakú mohutnosť ako množina prirodzených čísel ℕ (je ich "rovnako veľa"). Môžeme si všimnúť, že prirodzenými číslami nie je možné očíslovať všetky čísla z ľubovoľného neprázdneho intervalu – reálnych čísel je viac ako čísel prirodzených. Takéto rozdelenie premenných na spojité a diskrétne je preto jednoznačné. Môžu existovať aj také situácie, kde na určitom intervale je premenná spojitá a na inom diskrétna. Týmto prípadom sa nebudeme venovať.

Ak premenná má **kvalitatívny** charakter, potom hodnoty štatistického znaku nadobúdajú spravidla slovné (grafické) premenné, ktoré nie je možné rozumne transformovať na číselné charakteristiky (v rámci škál merania tomu môžu zodpovedať hodnoty nominálnej škály merania, pozri Kapitolu 1.1) a nemá s nimi význam používať ani elementárne matematické operátory „+“, „–“. Napríklad výšku respondenta vieme uviesť ako reálne číslo. Podobne by sme si však mohli dohodnúť aj kódovanie pre farbu očí: číslo 1 by zodpovedalo modrej, 2 zelenej, 3 hnedej a podobne. Ak premenná má kvalitatívny charakter, potom na hodnotách štatistického znaku nemá zmysel vykonávať aritmetické operácie, aj keď sú kódované ako čísla. V prípade kvalitatívneho znaku – farby očí – nemá zmysel vykonávať operáciu (1 + 2) / 2 = 1.5. Touto operáciou sa snažíme vypočítať „priemernú farbu očí“, a aj keď aritmeticky je možné uvedený výpočet realizovať (výsledok je 1.5), tento výsledok je neinterpretovateľný a nedáva zmysel.

Veľmi dôležitým a v skutočnosti netriviálnym konceptom je rozdeľovanie premenných na **náhodné** alebo **deterministické**. Zatiaľ si vystačíme s menej presnou definíciou náhodnej premennej. Pokiaľ nie sme schopní dopredu povedať akú hodnotu bude premenná nadobúdať, budeme hovoriť o náhodných premenných. V opačnom prípade, teda ak nie je žiadna neistota ohľadom nadobúdanej hodnoty (zväčša ak je zmena premennej „pod kontrolou“) hovoríme o deterministických premenných. Hádzanie kocky je situácia, kde pred hodom nevieme s istotou určiť, aká hodnota padne. Preto číslo – výsledok hodu kockou, o ktorom uvažujeme skôr, ako kocku hodíme, budeme nazývať náhodnou premennou. Presnejšiu definíciu, ktorá bude rozlišovať medzi výsledkom náhodného pokusu a náhodnou premennou, si stanovíme pri definovaní pravdepodobnosti. Socioekonomické javy majú veľmi často náhodný charakter. Spokojnosť zamestnancov, zákazníkov, spoľahlivosť produktov a ponúkaných služieb, úspech reklamnej kampane, to všetko má spravidla náhodný

charakter. Do určitej miery môžeme ovplyvňovať „charakteristiky“ náhodnosti, ale presnú mieru úspechu reklamnej kampane nepoznáme.

Čitateľ sa už možno stretol s pojmami ako **vzorka** a **populácia**. Pod populáciou budeme rozumieť množinu všetkých štatistických jednotiek, ktorých vlastnosti nás zaujímajú. Neraz však pre nás jednoducho nie je možné získať hodnoty od všetkých štatistických jednotiek. Ak však aj napriek tomu chceme skúmať vlastnosti všetkých štatistických jednotiek, za určitých podmienok môžeme v našej analýze vychádzať z údajov o časti tejto populácie. Tejto časti štatistických jednotiek budeme hovoriť **vzorka** pochádzajúca z populácie všetkých štatistických jednotiek. Spôsobom, ako túto vzorku vybrať, sa v tejto publikácii nebudeme bližšie venovať a je skôr predmetom publikácií venujúcich sa induktívnej štatistike.

Neraz je našou snahou na základe vzorky povedať niečo o všetkých štatistických jednotkách. Populáciu nie je však vždy ľahké definovať. Ako by sme definovali populáciu pri pozorovaní hodnôt HDP na Slovensku za sledované obdobie?

Pri analýze je samozrejme vhodné, aby pri všetkých štatistických analýzach boli čo najpresnejšie definované tak populácia, ako aj vzorka.

Príklad 1.2

Univerzita zakúpila elektronickú knižnicu pre potreby svojich zamestnancov a študentov. Po menšej propagačnej akcii sa vedenie univerzity zaujíma, či študenti vedia, že škola disponuje takou databázou a či im sú známe situácie, pri ktorých im môžu zdroje z tejto databázy pomôcť pri ich štúdiu. Populáciou však nie sú iba študenti, ale najmä akademickí zamestnanci školy, ktorí prístup k elektronickým knižniciam potrebujú pre svoj vedecký výskum. Keďže organizačne je veľmi náročné opýtať sa každého študenta a zamestnanca, zamestnanec zodpovedný za implementáciu elektronickej knižnice sa rozhodne určitým spôsobom vybrať časť študentov a zamestnancov a odpovede získané od nich zovšeobecniť. Táto skupina študentov a zamestnancov, ktorí budú oslovení, predstavujú vzorku.

1.1 Škály merania

Kľúčovým vstupom do procesu štatistického spracovania sú dáta. Dáta môžu byť pritom vo forme číselných vstupov, slov alebo vizuálnych vstupov: grafy, ikony, obrázky, a podobne. Spravidla sa zaujímame o určitú premennú. Povedzme výšku osoby. Túto premennú môžeme na určitom intervale považovať za spojitú v zmysle, že teoreticky môžu

existovať ľudia s ľubovoľnou výškou od určitej minimálnej až po určitú maximálnu hodnotu. Na meranie výšky si použijeme meter, na ktorom je najmenšia jednotka dĺžky milimeter. Pomocou tohto nástroja tak môžeme dostať iba určitý konečný a spočítateľný počet rôznych výšok. Ak by sme použili presnejší laserový merač výšky, mohli by sme dostať väčší počet rôznych výšok (napr. minimálna dĺžka je v mikrometroch). Pri meraní rovnakej premennej používame rôzne škály merania. Podľa toho, akú škálu merania sme použili (resp. mali k dispozícii), môžeme zvoliť vhodnú metódu na spracovanie údajov v štatistickej analýze. Dáta tak môžeme členiť na rôzne **škály merania** v závislosti od toho, ktoré základné matematické operácie (+, –, ×, /) s nimi má význam vykonávať. Klasické členenie škál merania pochádza zo 40-tych rokov od Stevensa (1946), ktorý rozlišuje medzi tzv.: nominálnou, poradovou (ordinálnou), intervalovou a podielovou škálou merania.

Nominálna škála – štatistické znaky štatistických jednotiek sa vytvárajú vo forme a) číslovania, ktorého účelom je identifikácia príslušnej štatistickej jednotky, alebo b) pomenovania špecifických skupín štatistických jednotiek.

Príklad 1.3

Označenie automobilového motora výrobným číslom môže slúžiť na neskoršiu identifikáciu kradnutých motorových vozidiel. Štatistickou jednotkou je v takom prípade motor a jeho vybraným štatistickým znakom je jeho označenie (výrobné číslo). Môžeme spočítať počet týchto označení, prípadne počet rovnakých označení (ktoré by sa ale nemali vyskytovať), nemá však význam robiť medzi týmito označeniami rozdiel, sčítanie, násobenie alebo delenie.

Ďalšími príkladmi nominálnej škály, kde namiesto číslovania použijeme slovné premenné, sú: obľúbená farba respondenta, typ strednej školy respondenta, pohlavie respondenta a podobne. Z praktického hľadiska rozdiel medzi číslovaním a slovným označením premennej v tomto prípade nevidíme. Druh školy môžem označiť číslom: 1, 2, 3,... alebo slovným spojením stredná priemyselná škola elektrotechnická, stredná priemyselná škola strojnícka, obchodná akadémia, gymnázium, ... V oboch prípadoch však nemá význam robiť iné matematické operácie, ako je sčítanie celkovej početnosti jednotlivých kategórií. Napr. koľko študentov navštevovalo strednú priemyselnú školu strojnícku.

Poradová (ordinálna) škála – štatistické znaky štatistických jednotiek sa udávajú vo forme poradí. Na údajoch z týchto škál merania je možné uskutočniť matematické operácie, ktoré zachovajú poradie a význam poradovej škály to nezmení (napríklad všetky poradia prenásobíme číslom 2). Vieme tak určiť, ktoré hodnoty na škále sú väčšie (viac preferované),

ktoré menšie (menej preferované). Nevieme však povedať o koľko, vieme len povedať o koľko poradí. V prieskumoch trhu alebo v psychologických testoch sa poradová škála vyskytuje pomerne často.

Príklad 1.4

Poradie, v ktorom atléti dobehnú do cieľa je premenná, ktorej škála merania je poradová. Na základe poradia nevieme povedať, nakoľko bol prvý v cieli rýchlejší ako druhý, tretí, ... Teda okrem rovnosti vieme rozhodnúť o tom, čo je viac (lepšie) a čo menej (horšie). Sčítanie a odčítanie nemusí mať vždy význam. Z toho vyplýva, že ani počítanie niektorých častých štatistík nemusí byť vhodné (priemer, rozptyl,...).

Častým prípadom v dotazníkoch je používanie tzv. Likertovej škály. Respondent má možnosť označiť odpovede v otázke a týmto odpovediam sú priradené čísla. Napríklad na škále od 1 – 5 označte nakoľko sa vám páči biela farba, kde 1 – vôbec sa mi nepáči, 2 – nepáči sa mi, 3 – aj sa mi páči aj sa mi nepáči, 4 – páči sa mi, 5 – veľmi sa mi páči. Nie je jednoznačné, či s takto získanými údajmi narábať ako s poradovou alebo intervalovou (podielovou) premennou. V tomto prípade sme sa snažili zvoliť také slovné charakteristiky, aby respondent videl v odpovediach určitú symetriu. Dôvod je ten, že respondentom vnímané rozdiely medzi odpoveďami by sme chceli chápať ako ekvidistantné. Ak je škála postavená tak, že platí (alebo je rozumné predpokladať) ekvidistantnosť, potom sa pri spracovaní takto získaných dát často používajú metódy navrhnuté pre dáta získané z intervalových alebo podielových škál. Vo väčšine prípadov je bezpečnejšie považovať odpovede respondentov za poradovú premennú. V tejto problematike doteraz neexistuje jednoznačný názor.

Intervalová škála – štatistické znaky štatistických jednotiek sa vytvárajú vo forme čísel, pomocou ktorých vieme nie len rozhodnúť o preferencii medzi štatistickými jednotkami, ale zároveň je rozdiel medzi dvoma rôznymi hodnotami pochádzajúcich z intervalovej škály zmysluplný. Pri intervalovej škále má rozdiel medzi dvoma hodnotami rovnaký význam, bez ohľadu na veľkosť hodnôt na škále. Rozdiel medzi 10°C a 8°C sú 2°C, rovnako ako rozdiel medzi 100°C a 98°C. Ak by sme teplotu merali na poradovej škále, kde jednotke by zodpovedala najmenšia teplota a 100 najväčšia, potom rozdiel medzi 10 a 8 by nemusel znamenať rovnaký posun ako medzi 100 a 98.[1]

[1] K rozdielom hodnôt meraných na intervalovej škále môžeme pristupovať ako k hodnotám podielovej škály.

Príklad 1.5

Tradičným príkladom je meranie teploty v stupňoch Celzia. Rozdiel medzi 24.0°C a 8.0°C je 16.0°C. Má význam tvrdiť, že **rozdiel** medzi 8.0°C a 16.0°C (t.j. rozdiel 8°C) je dva krát väčší ako medzi 24.0°C a 8.0°C (t.j. rozdiel 16°C). Nemá však význam tvrdiť, že 24.0°C je tri krát väčšia **teplota** ako 8.0°C, keďže 0.0°C na škále tepla počítajúcej so stupňami Celzia, neznamená absolútnu nulu – teda nulovú (žiadnu) teplotu. Ale teplotu, pri ktorej dochádza k zmene skupenstva vody (aj to iba pri určitom atmosférickom tlaku). Iným príkladom by bolo, keby sme počítali na Fahrenheitovej škále merania tepla. Intervalová škála teda nemá definovanú absolútnu nulu. Preto nemá význam hovoriť o násobkoch. Ďalším príkladom je meranie času pomocou kalendárov. Ľudia si zvolili 0-tý rok podľa určitej udalosti. To neznamená, že v roku 0 nebol čas.

Podielová škála – štatistické znaky štatistických jednotiek sa vytvárajú vo forme čísel, pomocou ktorých vieme determinovať: rovnosť, preferencie, rovnosť intervalov (porovnávanie rozdielov pri intervalovej škále) ako aj rovnosť podielov. Absolútna nula je implicitne vždy zahrnutá. V štatistickej analýze sa preferuje používanie podielových škál z praktických dôvodov:

- Ako sme už vyššie naznačili, niektoré matematické operácie nie je možné použiť, ak je nami získaný údaj sledovaný na nominálnej škále. Štatistické metódy, ktoré využívajú údaje na podielovej škále sú najrozšírenejšie. Existujú tak pestrejšie možnosti v štatistickej analýze.
- Všetky štatistické znaky s podielovou škálou je v prípade potreby možné redukovať na intervalovú, poradovú alebo nominálnu škálu (aj keď pritom dochádza k strate informácie).

Členenie škál podľa Stevensa (1946) na nominálnu, poradovú, intervalovú a podielovú škálu je dôsledkom pedagogickej snahy priblížiť štatistické metódy širšej odbornej (aj vedeckej) verejnosti, najmä výskumníkom z oblasti spoločenských vied. Na základe vybraných škál merania štatistických znakov a cieľa je často možné odporučiť vhodnú štatistickú metódu (Stevens, 1951).

Príklad 1.6

Predstavme si, že manažéra predaja zaujíma, či predajcovia – ženy, majú systematicky lepšie výsledky predajnosti v priebehu roka ako muži. Štatistickou jednotkou sú predajcovia. Štatistickým znakom je pohlavie predajcov a obrat ich predaja. Tento predaj

pritom meriame v EUR. Na akom type škály meriame predaj v EUR? Ide o podielovú škálu merania, a keďže našim cieľom je zistiť, či rozdiel medzi dvoma predajmi je systematický, za určitých predpokladov môžeme použiť tzv. *t*-test zhody dvoch stredných hodnôt. Ak by nás zaujímalo, či častejšie je v daný deň úspešnejšia žena, potom je štatistickým znakom úspech / neúspech za daný deň a išlo by o binárnu (indikátorovú alebo alternatívne nominálnu) premennú. V tom prípade by nás zaujímalo, či je väčší podiel úspešnejších dní u žien ako u mužov, na čo by sme mohli za určitých podmienok použiť test o zhode dvoch podielov.

Členenie škál podľa Stevensa (1946, 1951) prevzalo mnoho autorov, ktorí následne použitie jednotlivých štatistických metód odporúčali pre jednotlivé druhy škál. Nesporne vie byť uvedené členenie prínosom. Na druhej strane, členenie škál podľa Stevensa (1946, 1951) zahŕňa zopár úskalí, vďaka ktorým, ak by sme slepo nasledovali odporúčané metódy, by sme mohli použiť nesprávne metódy a následne prijať nepresné alebo dokonca chybné závery. Kritiku tejto klasifikácie škál merania si zosumarizujeme do dvoch bodov:

- Neraz nie je možné jednoznačne určiť škálu merania, ktorá môže závisieť aj od cieľa štatistickej analýzy.

Príklad 1.7

Nasledujúci príklad je dobre známou ukážkou nejednoznačnej interpretácie používania škál. Profesor na univerzite zodpovedný za prideľovanie futbalových čísiel bol obvinený z toho, že nováčikom dáva nezvyčajne nízke čísla.

Profesor protestoval, že tieto čísla majú význam na nominálnej škále merania. Štatistik zasa predpokladal, že čísla nevyjadrujú nominálnu škálu, čo mu umožňovalo vypočítať priemer a vykonať všetky potrebné operácie na to, aby mohol overiť tieto obvinenia. Keď profesor protestuje, že ide o futbalové čísla v zmysle nominálnej premennej, štatistik mu odpovedá: „*čísla nevedia odkiaľ pochádzajú*“. V závislosti od voľby škály by sme mohli súhlasiť s profesorom ako aj so štatistikom.

- Uvedené členenie škál neraz vedie k snahe výskumníkov používať poradové škály merania a týmto sa vyhnúť tzv. parametrickým testom, ktoré sú považované za vhodnejšie, avšak náročnejšie na podmienky ich používania.

Vzhľadom na túto kritiku nebudeme pri jednotlivých štatistických metódach dávať špecifické odporúčania v súvislosti so škálami merania.

2 Opis dát

Pri štatistickom spracovaní, ekonóm, pracuje so súborom dát, ktoré získal ako výstup z merania. Podobne ako výrobca nábytku svoj výstup (nábytok – napr. skriňu) môže charakterizovať prostredníctvom jeho účelu, rozmerov, použitých materiálov, ekonóm charakterizuje súbor dát, resp. ich opisuje pomocou určitých ukazovateľov, ktorých snahou je v pomerne jednoduchej podobe (jedno číslo, jeden graf) poskytnúť čo najviac relevantných informácií o **štatistickom súbore**. Keďže metódy obsiahnuté v tejto časti majú plniť práve túto úlohu, nazývame ich opisnými metódami a zahrňujeme ich pod **opisnú štatistiku** (označovanú tiež ako **deskriptívna štatistika**). Účelom opisnej štatistiky je charakterizovať empiricky namerané dáta. Pre zjednodušenie a znázornenie daného konceptu si predstavíme jednoduchý príklad.

Príklad 2.1

V období blížiacich sa parlamentných volieb sa stretávame so stále častejšími správami o preferenciách jednotlivých politických strán. Tieto preferencie sa získavajú dopytovaním vzorky respondentov. V našich podmienkach ide zvyčajne o telefonický prieskum na vzorke približne 1000 respondentov. Ak spracujeme výsledky sčítaním a následným percentuálnym vyjadrením preferencie respondentov voči politickým stranám, potom tieto výsledky je možné považovať za jednoznačné len v prípade práve tých 1000 respondentov, ktorých sme dopytovali. Aj tie platia len k danému času, ku ktorému sa prieskum uskutočnil, keďže nevieme vylúčiť, že si to respondenti rozmyslia. Z tohto dôvodu sa pri prezentovaní výsledkov z týchto prieskumov zvykne písať: „*Ako by dopadli voľby, ak by sa konali v minulom týždni*“. Samozrejme, taktiež nevieme vylúčiť, že nám nehovoria pravdu. Či by dopadli v danom momente voľby presne podľa výsledkov nameraných u 1000 respondentov s istotou povedať nevieme. Na základe týchto údajov, pomocou opisnej štatistiky nevieme získať znalosti o celkových politických preferenciách v populácii (všetci oprávnení voliči). Ak teda chceme vedieť rozloženie síl politických strán u všetkých relevantných voličov – populácii (a to je cieľom týchto prieskumov), potom sa na výsledky z opisnej štatistiky nemôžeme spoliehať. V takom prípade používame metódy **induktívnej štatistiky**, kde by sme na základe vzorky 1000 respondentov mohli **s určitou presnosťou** povedať, aké je rozloženie politických strán u všetkých voličov na Slovensku v čase získavania údajov, a to vychádzajúc len z nami dostupných, obmedzených dát. Aby sme uverili tomu, že výsledky z prieskumu nie je možné považovať **s istotou** za smerodajné,

stačí, ak si porovnáme výsledky rôznych agentúr, ktoré sa takýmto prieskumom venujú – výsledky sú takmer vždy odlišné.

V ďalšej časti textu sa budeme venovať metódam opisnej štatistiky, ktoré v závislosti od opisovaného problému rozdeľujeme na miery: polohy, variability a tvaru. V tejto časti sa nebudeme venovať tzv. mieram asociácií alebo závislostí, ktoré sa pri opisnej štatistike taktiež používajú.

2.1 Miery polohy

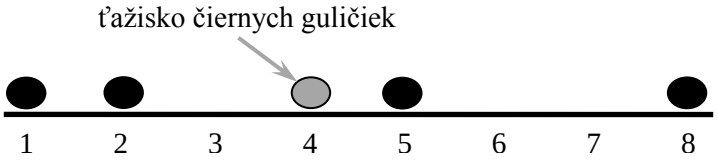
K základným mieram polohy môžeme zaradiť: priemer, medián a modus. Kvantily budeme považovať za osobitnú kategóriu. Vo všeobecnosti je účelom mier polohy charakterizovať, okolo akej hodnoty sa sústreďujú hodnoty štatistického súboru. Poskytujú nám informáciu, ktorú by sme mohli voľne charakterizovať ako informáciu o strede alebo ťažisku, kde sa údaje nachádzajú. Použitie konkrétnej miery polohy pritom závisí prevažne od situácie a cieľa štatistickej analýzy. Hneď v úvode Kapitoly 5 uvažujeme o sérii príkladov, ktoré na prvý pohľad vyžadujú rovnaký postup riešenia. V skutočnosti je v prvom príklade (Príklad 5.1) vhodné použiť tzv. aritmetický priemer, v druhom príklade (Príklad 5.2) harmonický, v treťom príklade (Príklad 5.3) geometrický a v poslednom štvrtom príklade (Príklad 5.4) chronologický priemer. Riešenia sú uvedené v spomínanej Kapitole 5.

2.1.1 Aritmetický priemer

Ide o najpoužívanejšiu mieru polohy, ktorej aj v ďalšom texte budeme hovoriť jednoducho priemer. Obrazne si aritmetický priemer môžeme predstaviť ako ťažisko (pozri Obrázok 2.1). Jednoduchý tvar na výpočet aritmetického priemeru je:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (2.1)$$

Kde veľké grécke písmeno sigma Σ je všeobecne používané na vyjadrenie sčítania jednotlivých hodnôt premenných, X je množina hodnôt, z ktorých chceme vypočítať aritmetický priemer, n je počet hodnôt v tejto množine a index i vyjadruje jednotlivé hodnoty z tejto množiny, resp. pozorovania.



Obrázok 2.1: Aritmetický priemer

Zdroj: vlastné spracovanie

2.1.2 Geometrický priemer

Na rozdiel od aritmetického priemeru, kde sa *n* hodnôt sčíta a následne vydelí ich počtom *n*, pri geometrickom priemere hodnoty vzájomne násobíme a potom *n*-tou odmocninou získame hodnotu geometrického priemeru. Formálne je výpočet jednoduchého geometrického priemeru nasledovný:

$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} X_i} \qquad (2.2)$$

Pričom gréckym symbolom veľké pí (Π) označujeme násobenie hodnôt pozorovanej premennej. Uvedený vzťah platí, ak $X_i > 0$.

Všimnime si, že platí:

$$\ln\left(\overline{X}_G\right) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(X_i\right) \qquad (2.3)$$

Voľba aritmetického alebo geometrického priemeru sa z praktického hľadiska nemusí javiť ako jednoduché rozhodnutie. V ekonómii sa s geometrickým priemerom môžeme stretnúť najmä pri počítaní priemerného rastu a zmeny.

Príklad 2.2

Zoberme si pokles produktivity práce zamestnancov predaja v priebehu dvoch dní. Na začiatku bola produktivita 10000,- EUR/deň, o dva pracovné dni 9025,- EUR/deň. Aký bol priemerný pokles za deň? Nech $g > 0$ je pokles za deň. Potom vieme, že musí platiť $10000g^2 = 9025$, takže $g = 0.95$. To zodpovedá 5 % poklesu predaja za deň oproti predchádzajúcemu dňu. V skutočnosti mohlo ísť aj o 20 % nárast v prvý deň a približne 24.79 % pokles v druhý deň (10000 x 1.2 x 0.7521) ≈ 9025. Nás však zaujíma priemerný rast (pokles), teda akou hodnotou musíme vynásobiť každý deň, aby sme sa za dva dni dostali z čísla 10000 k číslu 9025. Ak by sme počítali aritmetický priemer týchto zmien: (1.2 + 0.7521)/2 = 0.97605, tak by náš výsledok neprešiel jednoduchou skúškou správnosti: (10000 x 0.97605 x 0.97605) = 9526.73603 ≠ 9025.

Príklad 2.3

Uvažujme, že máme k dispozícii portfólio akcií s hodnotou Y,- EUR. Ak v jeden deň dôjde k poklesu hodnoty portfólia povedzme o – 50 % a nasledovný deň dôjde ku korekcii o + 50 %, tak celkovo sme prišli o 25 % z pôvodnej hodnoty portfólia Y, 1.5(0.5Y) = 0.75Y, Y – 0.75Y = 0.25Y. Ak by sme počítali priemerný rast pomocou aritmetického priemeru, tak by sme si mohli myslieť, že portfólio nič nezarobilo a nič neprerobilo: (1.5 + 0.5)/2 = 1. Takéto rozhodnutie by však bolo chybné. V skutočnosti bol priemerný rast (resp. pokles) ≈ 0.866, t.j. $0.866^2 = 0.75$.

2.1.3 Harmonický priemer

Harmonický priemer sa používa na výpočet priemernej intenzity akými sú rýchlosť, vzdialenosť za určitý čas a podobne. Na výpočet sa používa vzťah:

$$\overline{X}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \tag{2.4}$$

V spoločenských vedách sa s harmonickým priemerom môžeme stretnúť len zriedkakedy. Patrí do základnej trojice tzv. Pytagorových priemerov. Aritmetický priemer je vždy väčší alebo rovný geometrickému priemeru a geometrický priemer je vždy väčší alebo rovný harmonickému priemeru, teda platí:

$$\overline{X} \geq \overline{X}_G \geq \overline{X}_H \tag{2.5}$$

2.1.4 Chronologický priemer

Chronologický priemer je špecifickým prípadom aritmetického priemeru, s ktorým sa môžeme stretnúť najmä v oblasti evidencie zásob, odbytu, nákupu. Používa sa na výpočet priemerných hodnôt v určitom časovom rozmedzí, kde máme k dispozícii **usporiadané údaje** k určitému okamihu. Vzťah na výpočet chronologického priemeru je nasledovný:

$$\overline{X}_{CH} = \frac{\frac{X_1 + X_2}{2} + \frac{X_2 + X_3}{2} + \ldots + \frac{X_{n-1} + X_n}{2}}{(n-1)} \tag{2.6}$$

2.1.5 Medián

Používanie aritmetického priemeru má výhodu v tom, že používa na svoj výpočet všetky namerané údaje[2], je jednoduchý na výpočet a okrem toho má iné, vhodné štatisticko-

[2] Intuitívne si to môžeme vysvetliť tak, že berie do úvahy informáciu zo všetkých pozorovaní.

matematické vlastnosti (len heslovite spomenieme, že má známe rozdelenie pravdepodobnosti a je spravidla vhodným odhadom strednej hodnoty). Jeho hlavnou nevýhodou je, že ho nie je vždy možné dobre interpretovať a jedna extrémna hodnota môže výrazne ovplyvniť konečnú hodnotu priemeru.

Príklad 2.4

Na ilustráciu možných nedostatkov aritmetického priemeru uvažujme o nasledujúcom súbore údajov, ktoré zodpovedajú výške ľudí (v cm):

192, 162, 189, 171, 159, 166, 203, 172, 194, 196.

Jednoduchým výpočtom zistíme priemernú výšku v skupine ľudí, ktorá má hodnotu 180.4 cm. V skutočnosti nikto nie je v skupine vysoký presne 180.4 cm a dokonca z údajov vyplýva, že nikto v skupine nemá ani len podobnú výšku. Aritmetický priemer je takto v určitých situáciách málo reprezentatívnou mierou polohy. V skutočnosti sme do tejto vzorky 10 meraní vybrali dve skupiny ľudí – športovcov. Prvú skupinu s priemernou výškou 166 cm tvoria gymnasti a druhú skupinu s výškou 194.8 cm basketbalisti. Obe skupiny tak voči tej druhej predstavujú extrémne hodnoty.

V niektorých situáciách môže byť vhodnou alternatívou k aritmetickému priemeru medián. Za medián môžeme považovať tú hodnotu empirického štatistického súboru, od ktorej je $\lfloor n/2 \rfloor$ hodnôt väčších alebo rovných, a teda aj $\lfloor n/2 \rfloor$ menších alebo rovných, kde n je **rozsah štatistického súboru.** Zátvorkou $\lfloor ... \rfloor$ označujeme zaokrúhlenie na celé číslo nadol (tzv. dolná celá časť), kým zátvorkou $\lceil ... \rceil$ by sme označili zaokrúhľovanie nahor (tzv. horná celá časť). Výhodou mediánu je, že je ľahko interpretovateľný a „*robustný*“[3] voči extrémnym hodnotám. Pre potreby výpočtu mediánu sa hodnoty štatistického súboru zoradia vzostupne do variačného radu, kde ak n je párne číslo, medián $\tilde{X}$ vypočítame ako:

$$\tilde{X} = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} \tag{2.7}$$

Vychádzajúc z uvedenej definície mediánu, v skutočnosti by sme za mediánovú hodnotu mohli považovať ľubovoľné číslo v intervale:

$$X_{\left(\frac{n}{2}\right)} \leq \tilde{X} \leq X_{\left(\frac{n}{2}+1\right)} \tag{2.8}$$

keďže také číslo bude spĺňať našu definíciu mediánu. Ak n je nepárne číslo, tak medián zodpovedá hodnote:

[3] Robustný sa v uvedenom prípade rozumie to, že extrémnymi hodnotami nie je výrazne ovplyvnený.

$$\tilde{X} = X_{\left(\frac{n+1}{2}\right)} \tag{2.9}$$

Príklad 2.5

Ak by sme pokračovali v predchádzajúcom príklade, potom po vzostupnom zoradení do tzv. variačného radu: 159, 162, 166, 171, 172, 189, 192, 194, 196, 203 by predstavovala hodnota $\tilde{X} = \frac{172+189}{2} = 180.5\,cm$ medián, čo prirodzene znova nerieši problém so skutočnosťou, že sa okolo mediánovej hodnoty nenachádzajú športovci z nášho rozsahu údajov. Použitie mediánu[4] ako „všelieku“ v prípade extrémnych hodnôt evidentne tiež nie je úplne vhodné a vhodnosť jeho použitia bude závisieť od konkrétnej situácie.

Príklad 2.6

Ostro sledovaný vývoj priemernej mzdy v hospodárstve sa neraz dostáva do kontrastu s tvrdením o rozširovaní tzv. sociálnych nožníc. Rast priemernej mzdy v hospodárstve ako argument zvyšovania životnej úrovne obyvateľstva nemusí uspieť. Ak sa 90 % obyvateľom s menším príjmom výška mzdy nezmení, ale horným 10 % áno, potom dochádza k nárastu rozdielu medzi najbohatšími a zvyškom obyvateľstva. Zároveň dochádza k zvyšovaniu priemernej mzdy. Je teda otázne, nakoľko vývoj priemernej mzdy bude hovoriť o životnej úrovni obyvateľstva. Ak by sme namiesto aritmetického priemeru použili medián, jeho hodnota by sa v našom príklade nezmenila.

2.1.6 Modus

Najpočetnejšej hodnote v empirickom štatistickom súbore hovoríme modus. Podobne ako ostatné miery polohy, je jeho účelom pomôcť určiť, pri ktorej hodnote sa dáta koncentrujú[5]. Z praktického hľadiska je najzaujímavejšou situáciou ak má štatistický súbor viac ako jeden modus[6]. Vtedy hovoríme, že ide o **multimodálny** štatistický súbor (resp. ak má dve modálne hodnoty, ide o **bimodálny**). Výhodou modusu oproti mediánu a aritmetickému priemeru je skutočnosť, že sa môže použiť aj na popísanie takých

[4] Samozrejme, toto nie je tvrdenie, ktorého sa držíme v každej situácii. Je to len určité pravidlo. Možno trochu presnejšie tvrdenie by bolo, že medián je vhodným ukazovateľom miery polohy pri existencii extrémnych hodnôt, pri ktorých nie je záujem tieto hodnoty odstrániť.

[5] Toto je do určitej miery zjednodušená predstava. V prípade spojitej premennej naša definícia neplatí.

[6] Máme na mysli najmä taký štatistický súbor, v ktorom sa nevyskytuje veľa rôznorodých hodnôt. Ak máme v štatistickom súbore viac ako jeden modus, je to minimálne zvláštna situácia a snažíme sa prísť na to, čo je toho dôvodom. Či je to vlastnosť toho čo sme merali alebo tam existuje iný, zaujímavejší (podstatnejší?) faktor, ktorý tento efekt spôsobuje. Pozri Príklad 2.7.

štatistických znakov, ktorých hodnoty sú z nominálnej škály. V praxi môžeme modálne hodnoty použiť v situáciách:

- keď pri opisnej charakteristike štatistického súboru overujeme, či máme údaje z viac ako jednej populácie spojené do jedného súboru. Inak povedané, môže nás to upozorniť, že v jednom súbore porovnávame neporovnateľné veci.

Príklad 2.7

Zoberme si odevný závod, ktorý chce ušiť oblek pre športovcov. Potrebujeme zistiť rozmery, ktoré by sme chceli šiť. Vyberieme si vzorku mužov, avšak na nešťastie sa nám do nej dostali v približne rovnakom počte iba basketbalisti a gymnasti. Ak by sa ako miery zobrali priemerné hodnoty, tak by sa stalo, že by oblek bol pre basketbalistov malý a pre gymnastov príliš veľký. Zrejme by sa veľa produkcie na trhu nerealizovalo. Ak by sme chceli vypočítať modus podľa nami stanovenej definície, potom každá z hodnôt by bola modálnou, keďže ani jedna sa neopakuje viac ako práve jeden krát. V praxi sa v takejto situácii častejšie miesto výpočtu jedného modusu zobrazí histogram (pozri Kapitolu 3.4.4). Ten by v našom prípade naznačil, že máme viac modálnych hodnôt. Existovalo by tak podozrenie, že v skutočnosti máme dva rôzne „typy“ ľudí. A obleky by sme šili zvlášť pre gymnastov a zvlášť pre basketbalistov alebo iba pre jednu z týchto skupín.

- keď hľadáme typických reprezentantov, napr. zákazníkov. Vyberú sa 3 – 4 kľúčové ukazovatele, z ktorých sa vypočíta modálna hodnota a na základe nej sa charakterizuje typický reprezentant skupiny (napr. v obchodnom dome to môže byť: v akom čase, aký veľký nákup, z akej skupiny tovarov nakupuje typický zákazník obchodného domu).

Príklad 2.8

Počas prieskumu trhu sme zisťovali mimo iného základné demografické údaje o existujúcich zákazníkoch spoločnosti: pohlavie, vek, stav. Na základe modálnych hodnôt potom môžeme zostaviť profil typického zákazníka ako napr. slobodného muža v strednom veku.

Nevýhodou modusu je skutočnosť, že podobne ako medián na svoj výpočet nevyužíva všetky údaje. Rovnako v prípade existencie viacerých modusov je jeho vypovedacia schopnosť ako miery polohy otázna.

2.2 Kvantily

Majme namerané hodnoty $X_1, X_2, ..., X_n$, ktoré si zoradíme od najmenšej po najväčšiu hodnotu, t.j. zostavíme variačný rad $X_{(1)}, X_{(2)}, ..., X_{(n)}$. Potom r-tý kvantil predstavuje takú hodnotu štatistického znaku X, pre ktorú platí:

$$X_r = X_{(\lceil np \rceil)} \tag{2.10}$$

pričom $0 < p < 1$, $p = r/\alpha$, kde α udáva na koľko rovnako početných častí sa pomocou kvantilov rozdeľuje empirický štatistický súbor[7] a r je poradie kvantilu pre ktoré platí, $r = 1, 2, ..., \alpha - 1$. Špecifickým prípadom kvantilov, ktoré sa zrejme vyskytujú najčastejšie sú percentily a kvartily (taktiež tu zaraďujeme medián). Percentily rozdeľujú štatistický súbor na 100 rovnako početných častí a kvartily na 4 rovnako početné časti. Keďže medián, kvartily, decily a percentily rozdeľujú empirické štatistické súbory na rovnako početné časti, je možné ich považovať za „deliace“ hodnoty. Kvantily sú užitočné aj pri odhade minimálnych štandardov (napr. z celkového počtu uchádzačov do zamestnania vyberáme všetkých, ktorých výsledky z testov presiahli 95-ty percentil počtu bodov na vstupných testoch), pri porovnávaní kľúčových parametrov výkonnosti procesov a podobne. Vyššie spomínaný vzťah je vhodné používať pri väčších rozsahoch štatistických súborov.

Pri počítaní kvantilov rôzne softvérové produkty môžu postupovať pomocou rôznych metód výpočtu. Nami uvedený vzťah je považovaný za tzv. *neparametrický* výpočet kvantilov. Pre potreby tohto textu považujeme uvedený vzťah za dostačujúci. Je však na mieste pripomenúť, aby si užívatelia pred počítaním kvantilov skontrolovali spôsob výpočtu a posúdili jeho vhodnosť vzhľadom na cieľ realizovanej analýzy.

V súvislosti s kvantilmi sa môžeme stretnúť s tzv. päť číselným zhrnutím štatistického súboru na základe kvantilov, ktoré pozostáva z:

$$X_{(1)}, X_{(\lceil n0.25 \rceil)}, \tilde{X}, X_{(\lceil n0.75 \rceil)}, X_{(n)} \tag{2.11}$$

Teda minimálna hodnota, hodnota tzv. dolného kvartilu, od hodnoty ktorého je 75 % hodnôt väčších, medián, hodnota tzv. horného kvartilu, od ktorého je 25 % hodnôt väčších a maximálna hodnota. Spolu týchto päť hodnôt poskytuje pomerne dosť informácií o polohe, variabilite údajov a tvare rozdelenia početností, ktoré si vysvetlíme v ďalších kapitolách. Uvedené charakteristiky sa používajú aj pri tvorbe tzv. box – plotov (pozri Kapitolu 3.4.5).

[7] Predpokladáme empirický štatistický súbor v zmysle, že ide o štatistický súbor s konkrétne nameranými hodnotami.

Príklad 2.9

Spokojnosť zamestnancov so zabezpečovaním stravovania zo strany zamestnávateľa bola meraná na celočíselnej škále od 1 (veľmi nespokojný) po 7 (veľmi spokojný). Namerané výsledky za oddelenie Nákupu a Ľudských zdrojov boli nasledujúce:

Nákup: 1, 4, 6, 6, 6, 6, 7, 7, 5, 6, 6, 3, 5, 7, 7, 7, 2, 1, 3, 6, 6, 7, 7, 6, 7, 5, 6, 6.

Ľudské zdroje: 2, 3, 7, 5, 7, 4, 5, 7, 3, 4, 4, 6, 6, 1, 1, 5, 1, 1, 4, 6, 6, 5, 5, 5, 4, 3, 5, 7.

Manažment porovnáva spokojnosť na základe hodnoty dolného kvartilu, keďže ich cieľom je, aby v prvom rade bolo čo najmenej nespokojných zamestnancov. Výpočty ukázali, že v prípade oddelenia Nákup, dolnému kvartilu zodpovedá hodnota 5, a teda 25 % zamestnancov vyjadrilo menšiu alebo rovnakú spokojnosť akej zodpovedá bodové hodnotenie 5. Na oddelení Ľudských zdrojov je bodová hodnota dolného kvartilu 3. Z toho manažment usúdil, že akútnejší problém so spokojnosťou so zabezpečením stravovania je nutné riešiť u zamestnancov z oddelenia Ľudských zdrojov.

2.3 Niektoré miery variability

Koncept variability môžeme považovať v štatistike za kľúčový. Ak by neexistovala variabilita, zrejme by neexistovala ani štatistika. Svet bez variability by bol jednotvárny. Všetci ľudia by mali rovnakú výšku, rovnakú váhu, rovnakú farbu očí, a podobne.

Uvažujme o dvoch predajných miestach jednej banky. V predajnom mieste A sa každý predajca špecializuje na inú skupinu produktov. Ak príde zákazník, podľa toho o aký produkt má záujem, musí sa postaviť do radu k príslušnému špecialistovi. Na predajnom mieste B predajcovia nie sú špecialisti. Ak príde zákazník, môže sa postaviť do ľubovoľného radu. V prvom aj druhom prípade môže byť priemerný čas vybavenia zákazníka veľmi podobný. Rozdiel spočíva v tom, že ak sa na predajnom mieste A vytvorí u jedného špecialistu dlhý rad, zákazníci budú čakať príliš dlho, kým na predajnom mieste B sa dlhé rady skoro nikdy nevytvárajú. Ak na predajnom mieste A nie sú rady, zákazník bude vybavený rýchlejšie ako na predajnom mieste B. Výsledkom je, že priemerný čas vybavenia zákazníka je síce podobný, avšak na predajnom mieste A je väčšia variabilita času vybavenia ako na predajnom mieste B. **Hovorí sa, že zákazník nevníma priemernú hodnotu, ale variabilitu.** Problém s predajným miestom A spočíva v tom, že je podstatne náročnejšie naplánovať, koľko času bude zákazník tráviť na predajnom mieste A. Môže to byť veľmi málo aj veľmi veľa. Na predajnom mieste B bude vedieť, že čas odbavenia bude trvať približne stále rovnako dlho.

Príklad 2.10

Uvažujme o dvoch bankách A a B. V banke A existujú dva poradovníky. Jeden pre jedného a druhý pre druhého zamestnanca banky. Doba čakania u vybraných zákazníkov od zapísania sa do poradovníka po prijatie k zamestnancovi banky je: 5, 11, 2, 13, 19 minút. V banke B existuje jeden poradovník, z ktorého sa zákazníci priradzujú zamestnancom banky. Doba čakania zákazníkov od zapísania sa do poradovníka po prijatie k zamestnancovi banky je: 7, 10, 12, 11, 10 minút. Priemerná hodnota čakania je v oboch bankách rovnaká, t.j. 10 minút. Rozdiel je však vo variabilite. Kým v prípade banky A je zákazník pod vplyvom väčšej neistoty prameniacej zo skutočnosti, že na vybavenie môže čakať iba 2 minúty, ale aj 19 minút, v prípade banky B sa môže s väčšou istotou spoľahnúť na to, že sa k vybaveniu dostane približne za 10 minút.

Cieľom predchádzajúceho príkladu bolo poukázať na význam variability a obmedzenia mier polohy. Prezentovať samotnú priemernú hodnotu je neraz skresľujúce. Pri prezentovaní výsledkov z prieskumov trhu (ako aj výskumu) je preto zvykom uvádzať vedľa mier polohy aj príslušnú mieru variability. Pod variabilitou môžeme rozumieť stupeň rôznorodosti, rozptýlenia, kolísania alebo odchýlenia hodnôt od centrálnej (tzv. strednej) hodnoty. Ak bolo cieľom mier polohy merať okolo akej hodnoty sú údaje v štatistickom súbore koncentrované, tak miery variability merajú nakoľko „husto“ sa tieto hodnoty okolo miery polohy koncentrujú. V anglickej literatúre sa pojem variabilita prezentuje v závislosti od kontextu ako: *variability* alebo *dispersion* (v slov. disperzia). Na výpočet variability existuje niekoľko prístupov, ktorých použitie si ukážeme v nasledujúcich častiach. Miery variability majú tri spoločné vlastnosti:

- V prípade, ak je ich hodnota rovná 0, potom sú hodnoty v štatistickom súbore rovnaké – neexistuje variabilita.
- Hodnoty miery variability sú nezáporné.
- Rastúca miera variability znamená väčšie rozptýlenie údajov v štatistickom súbore.

Ďalej si predstavíme najčastejšie používané miery variability: rozpätie štatistického súboru, medzi-kvartilové rozpätie, smerodajnú odchýlku, rozptyl, absolútnu odchýlku a variačný koeficient.

2.3.1 Variačné rozpätie – rozpätie štatistického súboru

K najjednoduchším mieram variability patrí variačné rozpätie. Nesporne je jeho výhodou nenáročnosť výpočtu a ľahká interpretácia. Vypočítame ho ako rozdiel najväčšej a najmenšej hodnoty štatistického súboru:

$$R = X_{(n)} - X_{(1)} \tag{2.12}$$

Variačné rozpätie nám tak hovorí, aký najväčší rozdiel hodnôt existuje v štatistickom súbore. Väčší rozdiel by mal naznačovať väčšiu variabilitu. Variačné rozpätie má však niektoré zjavné nevýhody. V prvom rade neberie do úvahy variabilitu medzi maximálnou a minimálnou hodnotou. Okrem toho môže byť ľahko skresľujúci. Ak v štatistickom súbore existuje extrémna hodnota, výrazne to môže ovplyvniť náš pohľad na variabilitu súboru. Rozpätie nám nedá odpoveď na otázku, či sú hodnoty štatistického súboru zoskupené v blízkosti strednej hodnoty. S variačným rozpätím sa môžeme stretnúť pri meraní variability cien, napr. cien akcií na trhu. Každý obchodný deň sa sledujú tzv. otváracia, minimálna, maximálna a uzatváracia cena akcie. Rozdiel medzi maximálnou a minimálnou cenou akcie je jeden z elementárnych spôsobov merania variability cien. Existuje celá skupina ukazovateľov, ktoré toto rozpätie (aj iné) používajú.

Príklad 2.11

V obchodnom podniku sledovali dĺžku doby splatnosti svojich záväzkov po oficiálnej dobe splatnosti. Z určitých ekonomických dôvodov nebolo cieľom spoločnosti skrátiť dobu splatnosti, ale znížiť variabilitu. Údaje v počtoch dní sú nasledujúce:

3, 7, 6, 2, 5, 6, 1, 9, 7, 3, 11, 9, 13, 14, 23, 2, 6, 9, 28, 14, 2, 5, 3, 8, 5, 3, 9, 10, 11, 8, 12, 4.

Variačné rozpätie zo štatistického súboru je R = 27 dní. Môžeme vidieť, že ak by sme odstránili 28 denné omeškanie, variačné rozpätie by sa znížilo na R = 22 dní a v prípade, ak by sme odstránili aj jedno 23 denné omeškanie, dosiahli by sme hodnotu variačného rozpätia iba R = 13. Účelom bolo nie odstrániť zjavne extrémne hodnoty (to by mohla byť chyba), ale poukázať, nakoľko je variačné rozpätie citlivé na extrémne hodnoty.

2.3.2 Medzi-kvartilové rozpätie

V predchádzajúcich kapitolách sme si definovali kvantily a následne percentily a kvartily. Rozdiel medzi horným a dolným kvartilom predstavuje medzi-kvartilové rozpätie R_Q. V medzi-kvartilovom rozpätí sa tak nachádza 50 % hodnôt štatistického súboru. Aj keď uvedený spôsob výpočtu miery variability nie je náchylný na extrémne hodnoty, podobne ako variačné rozpätie je jeho hlavnou nevýhodou, že neberie do úvahy všetky hodnoty

štatistického súboru, a teda nezvažuje rozptýlenosť hodnôt vo vnútri kvartilového rozpätia ani mimo kvartilového rozpätia:

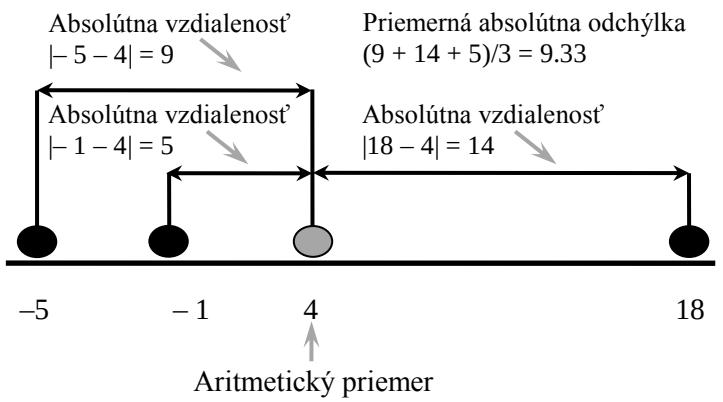
$$R_Q = X_{(\lceil 0.75n \rceil)} - X_{(\lceil 0.25n \rceil)} \tag{2.13}$$

2.3.3 Priemerná absolútna odchýlka

Interpretačne za najvhodnejšiu mieru variability môžeme považovať priemernú absolútnu odchýlku, ktorej hodnotu vypočítame jednoduchým priemerovaním absolútnych odchýlok jednotlivých hodnôt od aritmetického priemeru. Inak povedané, dostávame odpoveď na otázku, nakoľko sú v priemere hodnoty štatistického súboru vzdialené od ich aritmetického priemeru. V praxi sa priemerná absolútna odchýlka využíva zriedkavo v dôsledku nevhodných štatisticko-matematických vlastností absolútnej odchýlky v porovnaní s rozptylom a smerodajnou odchýlkou. Pracovať s absolútnou hodnotou je komplikované. Uvedený argument má však svoju váhu iba v určitých obmedzených prípadoch (bližšie pozri Gorard, 2005), avšak priemerná absolútna odchýlka má z interpretačného hľadiska jasné výhody oproti všetkým ostatným ukazovateľom miery variability. Z tohto dôvodu ju môžeme považovať za neprávom podceňovanú (aspoň čo sa týka deskriptívnej štatistiky). Priemernú absolútnu odchýlku vypočítame z nasledujúceho vzťahu:

$$S_{\bar{d}} = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}\right| \tag{2.14}$$

Výhodou priemernej absolútnej odchýlky (na rozdiel od variačného rozpätia a medzikvartilového rozpätia) je, že berie do úvahy všetky hodnoty štatistického súboru. Čím je hodnota absolútnej odchýlky menšia, tým je aritmetický priemer vo všeobecnosti *zrejme vhodnejšou mierou polohy* štatistického súboru, keďže hodnoty sa sústreďujú bližšie k aritmetickému priemeru. Extrémne odľahlé hodnoty štatistického súboru majú menší vplyv na hodnotu absolútnej odchýlky ako na rozptyl, resp. na smerodajnú odchýlku. Z týchto dôvodov považujeme použitie priemernej absolútnej odchýlky pre potreby opisnej štatistiky za vhodnú alternatívu voči rozptylu a smerodajnej odchýlke, aj keď znova zopakujeme, v praxi je používaná len sporadicky. Princíp výpočtu priemernej absolútnej odchýlky si môžeme znázorniť graficky na nasledujúcom obrázku.



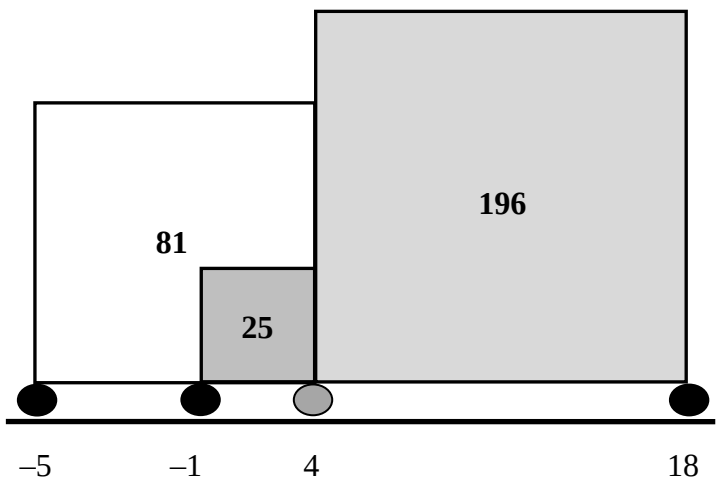
Obrázok 2.2: Ukážka výpočtu priemernej absolútnej odchýlky

Zdroj: vlastné spracovanie

2.3.4 Rozptyl a smerodajná (štandardná) odchýlka

V induktívnej štatistike je dominantnou mierou variability rozptyl a z neho odvodená smerodajná odchýlka. Na rozdiel od absolútnej odchýlky sa rozptyl vypočíta ako priemer štvorcov odchýlok:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad (2.15)$$



Obrázok 2.3: Ukážka výpočtu rozptylu

Zdroj: vlastné spracovanie

Vzhľadom na skutočnosť, že štvorcami rozdielov sa jednotky, v ktorých máme údaje namerané tiež umocňujú, z interpretačného hľadiska je prijateľnejšie namiesto rozptylu za opisnú charakteristiku používať smerodajnú odchýlku, ktorá je druhou odmocninou rozptylu. Ak by sme počítali rozptyl času obsluhy zákazníka, výsledná hodnota by bola v jednotke času t^2, čo je zbytočná interpretačná výzva. Vzťah pre smerodajnú odchýlku je tak:

$$\sigma = \sqrt{\sigma^2} \tag{2.16}$$

Napriek odmocneniu je priamočiara interpretácia smerodajnej odchýlky stále komplikovaná. Ako pri všetkých mierach variability platí, že čím je väčšia hodnota smerodajnej odchýlky (rozptylu) tým je aritmetický priemer menej vypovedajúci, resp. hodnoty štatistického súboru sú vzhľadom na aritmetický priemer viac rozptýlené. Zjavnou nevýhodou smerodajnej odchýlky (rozptylu) oproti absolútnej odchýlke je skutočnosť, že hodnoty vzdialenejšie od aritmetického priemeru majú väčší vplyv na hodnotu smerodajnej odchýlky (rozptylu) ako hodnoty bližšie k aritmetickému priemeru (čo je dané umocnením), ale na rozdiel od priemernej absolútnej odchýlky, rozptyl má vhodné štatisticko-matematické vlastnosti (Fisher,1920).

2.3.5 Variačný koeficient

Variačný koeficient môžeme použiť na porovnávanie variability dvoch a viacerých štatistických súborov. Jeho najbežnejšou formou je využitie smerodajnej odchýlky, teda v tvare:

$$V = \left|\frac{\sigma}{\overline{X}}\right| \times 100\ \% \tag{2.17}$$

Pričom samozrejme musí platiť, že $\overline{X} \neq 0$. Z výrazu vyplýva, že variačný koeficient vyjadruje podiel variability na aritmetickom priemere. Týmto spôsobom, aj napriek skutočnosti, že jednotky, v ktorých dochádza k meraniu sú odlišné, môžeme porovnať vzájomnú variabilitu a rozhodnúť pri opisnej štatistike, ktorý zo štatistických súborov preukazuje vyššiu variabilitu. Na ilustráciu použijeme názorný príklad podľa Triola (2004).

Príklad 2.12

Majme štatistický súbor s nameranou výškou a váhou mužov s nasledujúcimi hodnotami pre aritmetický priemer a smerodajnú odchýlku:

$\overline{X}_{\text{VÝŠKA}}$ = 173.58 cm a $\sigma_{\text{VÝŠKA}}$ = 7.67 cm

$\overline{X}_{\text{VÁHA}}$ = 78.25 kg a $\sigma_{\text{VÁHA}}$ = 11.94 kg

Prirodzene, nemôžeme porovnať smerodajnú odchýlku v jednotke dĺžky s jednotkou váhy. Vypočítaním variačného koeficientu dostaneme:

$$V_{výška} = \left|\frac{\sigma}{\overline{X}}\right| \times 100\ \% = \left|\frac{7.67\ \text{cm}}{173.58\ \text{cm}}\right| \times 100\ \% = 4.42\ \%$$

$$V_{váha} = \left|\frac{\sigma}{\overline{X}}\right| \times 100\ \% = \left|\frac{11.94\ \text{kg}}{78.25\ \text{kg}}\right| \times 100\ \% = 15.26\ \%$$

Môžeme teda vidieť, že výška mužov preukazuje menšiu variabilitu ako váha mužov. Uvedený výsledok nie je prekvapujúci. Neraz môžeme vidieť dvoch mužov s približne rovnakou výškou, ale diametrálne odlišnou hmotnosťou, kým podstatne zriedkavejšie vidíme dvoch mužov, kde jeden z nich je dvakrát vyšší ako druhý.

2.4 Miery tvaru

Poznáme dve základné miery tvaru štatistického súboru: šikmosť a špicatosť. Miery šikmosti je možné interpretovať ako miery asymetrie. Na tomto mieste (t.j. bez definovania rozdelenia početností) nie je jednoduché intuitívne vysvetliť význam mier asymetrie (resp. špicatosti). Pomôžme si týmto všeobecným pravidlom: v prípade, ak je väčšina hodnôt štatistického súboru menšia ako aritmetický priemer (na reálnej osi sa nachádzajú „na ľavo“ od aritmetického priemeru) hovoríme, že ide o *pravostranné zošikmenie* a naopak. Ak je väčšina hodnôt štatistického súboru väčšia ako aritmetický priemer (na reálnej osi sa nachádzajú „na pravo“ od aritmetického priemeru) hovoríme, že ide o *ľavostranné zošikmenie*.

V prípade opisnej charakteristiky štatistického súboru je možné si túto vlastnosť overiť pomocou nasledujúceho vzťahu, tzv. **koeficientu šikmosti**:

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^3}} \qquad (2.18)$$

Pre hodnoty $S > 0$ ide o pravostranne zošikmenie a pre $S < 0$ ide o ľavostranné zošikmenie.

Príklad 2.13

Pravidlo, ktoré sme uviedli vyššie neplatí vždy, resp. jeho platnosť závisí od toho, akým spôsobom meriame asymetriu. Uvažujme o nasledujúcom príklade. Majmä údaje:–3, –3, –3, 3, 3, 3, –4. Medián je –3, priemer –0.57. Podľa všeobecného pravidla vyššie, by malo ísť o pravostranné zošikmenie, keďže väčšina hodnôt je menších ako aritmetický priemer. Vypočítaním koeficientu šikmosti dostaneme hodnotu 0.33 a teda naše pravidlo je potvrdené ak koeficientom šikmosti. Použije teraz nasledujúce údaje: –3, –3, –3, 3, 3, 3, –8, priemer je –1.14, medián ostal –3. Stále platí, že väčšina hodnôt je menších ako aritmetický priemer a teda podľa nášho pravidla by malo ísť o pravostranné zošikmenie.

Vypočítaním koeficientu šikmosti dostávame hodnotu –0.39, čo je však znakom ľavostranného zošikmenia.

Okrem spomínaného vzťahu sa môžeme stretnúť aj s nasledujúcimi vzťahmi asymetrie:

$$S_P = \frac{\overline{X} - \hat{X}}{\sigma} \tag{2.19}$$

čo je Pearsonov koeficient šikmosti. Znamienko udáva podobne ako pri *S* smer šikmosti. Rozdiel spočíva v tom, že absolútna hodnota S nemá vypovedaciu schopnosť, kým v prípade absolútnej hodnoty $|S_P|$ môžeme tvrdiť, že čím je väčšia, tým je štatistický súbor šikmejší (nezávisle od strany zošikmenia). V prípade, ak nie je definovaný modus v empirickom súbore (napr. v situácii, kde každá hodnota sa vyskytuje práve jeden krát), tak na základe empirických skúseností pri približne symetrických štatistických súboroch je možné ho odhadnúť ako:

$$\hat{X} = \overline{X} - 3\left(\overline{X} - \tilde{X}\right) \tag{2.20}$$

Ďalšou alternatívou je S_B, tzv. *Bowleyov koeficient šikmosti*, ktorý vychádza z jednoduchého princípu, že v prípade dokonalej symetrie je absolútna vzdialenosť medzi horným kvartilom $X_{(\lceil 0.75n \rceil)}$ a dolným kvartilom $X_{(\lceil 0.25n \rceil)}$ rovnaká. Bowleyov koeficient šikmosti však využíva iba stredných 50 % hodnôt.

$$S_B = \frac{\left(X_{(\lceil n0,75 \rceil)} - \tilde{X}\right) - \left(\tilde{X} - X_{(\lceil n0,25 \rceil)}\right)}{\left(X_{(\lceil n0,75 \rceil)} - \tilde{X}\right) + \left(\tilde{X} - X_{(\lceil n0,25 \rceil)}\right)} \tag{2.21}$$

Určitým vylepšením je nasledujúci vzťah, tzv. *Kellyho koeficient šikmosti*:

$$S_K = \frac{\left(X_{(\lceil n0,90 \rceil)} - \tilde{X}\right) - \left(\tilde{X} - X_{(\lceil n0,10 \rceil)}\right)}{\left(X_{(\lceil n0,90 \rceil)} - \tilde{X}\right) + \left(\tilde{X} - X_{(\lceil n0,10 \rceil)}\right)} \tag{2.22}$$

Nie je však jasné, prečo sa pre výpočet použil práve 90-ty percentil a nie vyšší, resp. nižší percentil. Pre potreby opisnej štatistiky odporúčame zvážiť použitie konkrétnej miery šikmosti vychádzajúc z jednoduchosti interpretácií.

Ako sme už naznačili vyššie, ako indikátor šikmosti môže poslúžiť skúmanie vzťahu medzi aritmetickým priemerom, mediánom. K tomu teraz pridáme aj modus. Za predpokladu, že hodnoty štatistického súboru majú jednu modálnu hodnotu, hovoríme o ľavostrannom zošikmení, ak platí nasledujúca nerovnosť $\overline{X} < \tilde{X} < \hat{X}$ a o pravostrannom zošikmení, ak platí $\overline{X} > \tilde{X} > \hat{X}$.

Miery špicatosti si môžeme interpretovať ako mieru koncentrácie hodnôt v okolí strednej hodnoty. Čím vyššia je miera koncentrácie, tým je väčšia tendencia hodnôt sústrediť sa okolo aritmetického priemeru. Okrem iného, koeficient špicatosti slúži neraz na odhad „odolnosti“ rozdelenia početností na extrémne hodnoty. Inak povedané, pri vyššej špicatosti môžeme spravidla očakávať vyšší výskyt extrémnych hodnôt[8]. Štatistické súbory sa v závislosti od tvaru a miery špicatosti rozdeľujú na: mezokurtické (normálne, $K = 0$), leptokurtické (špicaté, $K > 0$) a platykurtické (ploché, $K < 0$), kde K je koeficient špicatosti:

$$K = \frac{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} - 3 \qquad (2.23)$$

Kde pre $K = 0$ je štatistický súbor mezokurtický, pre $K > 0$ leptokurtický (t.j. špicatejší ako mezokurtický[9]) a pre $K < 0$ je platykurtický.

[8] Špicatosť sa zvykne zamieňať aj s pojmom „fat tails“ alebo „heavy tails“, čiže tučné alebo ťažké konce. Existujú rozdelenia hodnôt, ktoré predpovedajú väčšiu pravdepodobnosť výskytu extrémnych udalostí (napr. finančná kríza subjektu) – týmto rozdeleniam hodnôt sa zvykne hovoriť rozdelenia s tučnými (resp. ťažkými) koncami.

[9] Resp. špicatejší ako frekvenčná krivka normálneho rozdelenia pravdepodobnosti – pozri bližšie Kapitolu 4.6.2.

3 Úvod do programu R

R je štatistický softvér, ktorý je od roku 1995 (http://cran.r-project.org/doc/html/interface98-paper/paper_2.html) voľne dostupným na stiahnutie ako aj inštaláciu. Program R primárne slúži na spracovanie väčšieho množstva dát pomocou štatistík a vizualizácie. Väčšina metód, ktoré si v tejto publikácii predstavíme, sú realizovateľné v rôznych štatistických programoch, ktorých užívateľské rozhranie je možno intuitívnejšie a najmä ľahšie pochopiteľné. Na druhej strane, program R je veľmi flexibilný. Užívateľovi umožňuje podstatne širšie využitie ako iné štatistické programy. Program R v podstate nemusí slúžiť len na štatistické spracovanie. Je možná jeho spolupráca s inými komerčnými produktmi, nie nutne štatistického charakteru. Môže slúžiť na tvorbu a spracovanie obrázkov. Umožňuje vytvárať programy (spustiteľné v programe R). Možnosť písať príkazy a týmto spôsobom komunikovať s počítačom nám umožňuje aspoň čiastočne nahliadnuť do základov programovania. Vo výskume sa pritom často stretávame s potrebou vytvárať vlastné postupy a algoritmy, ktoré nie sú v iných (spravidla komerčných) programoch dostupné. Pomerne veľkou výhodou sa tak javí široká záujmová skupina používateľov (spravidla vedeckých pracovníkov) , ktorí rozširujú možnosti programu R o rôzne analytické balíky (tzv. *packages*), ktoré obsahujú nové štatistické testy. Neraz sa dokonca stane, že ak štatistik príde s určitou novou metódou, okrem jej publikovania vo vedeckom časopise napíše aj program, pomocou ktorého je možné túto metódu aplikovať. Často je jazyk, v ktorom sa tieto programy píšu, práve R. V ekonómii patria k iným populárnym programovým balíkom: STATA, S-PLUS, EViews, GAUSS alebo OX Metrics. Určitou alternatívou je program Gretl, ktorý je taktiež voľne dostupným a pritom užívateľsky prijateľnejším programom. Okrem toho sa na účely ekonometrickej analýzy často používajú aj moduly z programu MATLAB.

V tomto texte nepôjdeme v programovaní do väčších detailov. Niektoré príkazy sa môžu zdať zbytočne dlhé a komplikované. V mnohých prípadoch je možné dosiahnuť rovnaké výsledky použitím kratších postupov. Našim cieľom však bolo urobiť príkazy čo najviac intuitívne pre úplných začiatočníkov, ktorí s programovaním nemajú žiadne skúsenosti.

Spravidla platí, že najlepšie sa učí práca s programovacím jazykom na reálnych problémoch. Existuje pomerne ľahko dostupná literatúra na internete, dokonca už aj v slovenskom jazyku (Želinský et al., 2010; Pančíková, 2011; Kalina et al., 2010). V tejto kapitole si ukážeme niektoré základné funkcie a postupy, ktoré budeme v ďalšom texte používať. Budeme pritom vychádzať najmä z publikácií Verzani (2004), Crawley (2007),

Cohen – Cohen (2008), Murrell (2006), Dalgaard (2008), Wickham (2009), Everitt – Hothorn (2005) a samozrejme z vlastných skúseností. Štruktúru tohto úvodu (ako aj niektoré vybrané príklady) sme prevzali priamo od Verzani (2004), ktorý na predmete KMvE I a II na Podnikovohospodárskej fakulte so sídlom v Košiciach predstavuje odporúčanú literatúru.

Všetky kódy, ktoré v publikácii uvádzame boli na konci skontrolované vo verzii 2.14.0 z 31.10.2011, 32bitová verzia pre OS Windows. Na záver uvádzame aj celý zoznam použitých programových balíkov, verzie, ktorá bola pri spätnom overovaní kódov použitá.

3.1 Inštalácia programu R

Aktuálnu verziu programu pre operačný systém MS Windows je možné stiahnuť z nasledujúceho odkazu [dostupné online, 16.02.2012]:

`http://cran.at.r-project.org/bin/windows/base/`

Po inštalácii (*run installation* → *full installation* → *install*) sa odporúča aktualizovať vyššie spomínané programové balíky (v angl. *packages*). Tieto programové balíky rozširujú funkcionalitu programu R. S inštaláciou programu R sa inštalujú aj niektoré základné programové balíky. Tie sú spravidla pravidelne aktualizované o nové funkcie, prípadne o opravu starších funkcií. Aktualizáciu je možné uskutočniť nasledovne (je potrebný prístup na internet):

Menu Packages → Update packages → Select mirror (napr. Austria) → „označiť všetky programové balíky“ → prebehne update

Tento postup je vhodné čas od času opakovať, keďže program R nás neupozorní o existencii novších verzií programových balíkov. Zoznam programových balíkov, ktoré je možné si priamo nainštalovať, získame pomocou:

Menu Packages → Install packages

Ako príklad uvedieme programový balík `strucchange`, ktorý sa hodí pri výpočte štrukturálnych zlomov v regresných koeficientoch.

Menu Packages → Install packages → „je potrebné nájsť a vybrať si balík `strucchange`“ → Select mirror (napr. Austria) → prebehne inštalácia

Následne sa **programový balík** aktivuje príkazom:

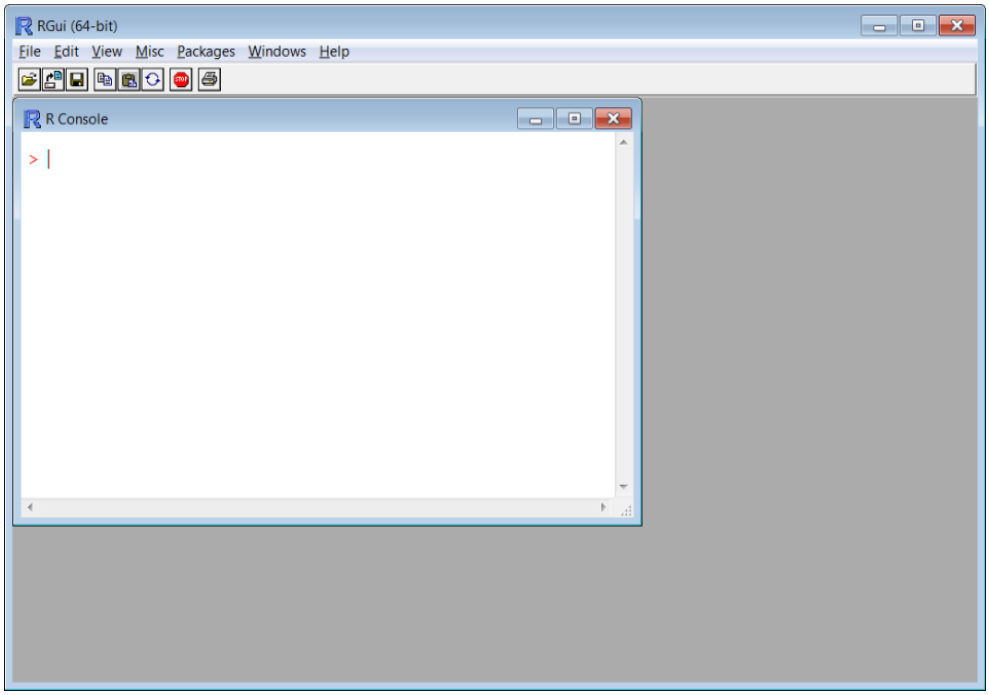
```
library(strucchange)
```

Tieto programové balíky sa častejšie nazývajú **knižnicami**. My budeme toto pomenovanie v texte voľne zamieňať.

3.2 Základné operácie v programe R

Po spustení programu sa zobrazí základné užívateľské rozhranie, ktoré je znázornené na nasledujúcom obrázku.

Príkazy sa píšu do konzoly („R Console“). Priamo do konzoly môžeme zadávať základné operátory ako +, -, /, * a funkcie ako `sqrt()`, `sin()`, `cos()`, `exp()`, `log()`, `pi`[10]. Je vhodné si na začiatku vyskúšať najmä postupnosť výpočtov pri použití zátvoriek.



Obrázok 3.1: Užívateľské rozhranie v programe R

Zdroj: vlastné spracovanie v programe R

```
> 1 + (2/3)*2
[1] 2.333333
> 1 + 2/(3*2)
[1] 1.333333
> 1 + 2/3*2
[1] 2.333333
```

V základnom vybavení programu R sa za oddeľovač desatinných miest považuje bodka, nie čiarka. Z tohto dôvodu používame túto normu aj v tejto publikácii. Keďže sa pri práci v programe R často stretávame s niektorými symbolmi ako napr. $, odporúčame tiež používať anglickú klávesnicu.

Ak nám nie je známe, ako presne použiť určitú funkciu, je možné pozrieť manuál, ku ktorému sa dostaneme nasledovne:

[10] Všetky funkcie v programe R uvádzame spolu so zátvorkami. V tomto prípade jedine `pi` nie je funkcia.

```
?log
```

Príklad 3.1

Napíšte log(10) do konzoly a potvrďte. Následne získajte hodnotu zodpovedajúcu log(10) s použitím log(x, base).

V programe R môžeme číslam (nielen im) priradiť určité označenie (meno), ktoré ich budú definovať. Vytvárame tak určité objekty, ktoré sa potom privolávajú. Používame k tomu operátory =, <-, ->. Napríklad:

```
> x = 2*7+1; x
[1] 15
> x = 2*x+1; x
[1] 31
> x <- 2*x+1; x
[1] 63
> 2*x+1 -> x; x
[1] 127
```

Program R „číta“ príkazy zľava doprava po riadkoch. Spravidla jeden riadok pre neho predstavuje jeden príkaz. Ak chceme v jednom riadku napísať viac ako jeden príkaz, oddeľujeme ho pomocou bodkočiarky.

Príklad 3.2

Zistite, aký je výsledok pre objekt „x“ bez toho, aby ste tieto príkazy zadávali do programu R.

```
x <- 3/2*7/2/2
x <- x^2*2^2
```

Na tvorbu dátových vektorov budeme používať funkciu `c()`, ktorá predstavuje skratku pre „create“ (vytvor). Ak napríklad chceme definovať vektor pozorovaní ako vektor `x`, `y`, `z` alebo `c`, môžeme to uskutočniť ako:

```
> x <- c(1, 2, 3)
> y <- c(7, 8, 9)
> z <- c(4, 5, 6)
> c <- c(x, z, y)
> x; y; z; c
[1] 1 2 3
[1] 7 8 9
[1] 4 5 6
[1] 1 2 3 4 5 6 7 8 9
```

3.3 Práca s údajmi v programe R

Následne môžeme na týchto vektoroch vykonávať rôzne operácie. K tým základným by sme zaradili nasledovné: `sum()`, `length()`, `mean()`, `sort()`, `min()`, `max()`, `range()`, `diff()`, `cumsum()`. Podľa predchádzajúceho poradia ide o: súčet hodnôt, počet hodnôt (dĺžka vektora), aritmetický priemer hodnôt, usporiadanie hodnôt, minimálna hodnota, maximálna hodnota, minimálna a maximálna hodnota, rozdiely za sebou idúcich hodnôt (diferencia), absolútny kumulatívny súčet hodnôt.

```
> x <- c(1, 3, 5, 7, 5, 3, 1)
> sum(x)
[1] 25
> length(x)
[1] 7
> mean(x)
[1] 3.571429
> sort(x)
[1] 1 1 3 3 5 5 7
> min(x)
[1] 1
> max(x)
[1] 7
> range(x)
[1] 1 7
> diff(x)
[1]  2  2  2 -2 -2 -2
> cumsum(x)
[1]  1  4  9 16 21 24 25
```

Príklad 3.3

Vytvorte vektor s názvom „`FL`“, kde budú hodnoty vektora „`c`“ v klesajúcom poradí. Nesmiete však použiť funkciu `c()`, iba dátový vektor `c` a funkcie definované vyššie.

```
> FL <- sort(c, decreasing = T); FL
[1] 9 8 7 6 5 4 3 2 1
```

Sčítavanie vektorov je ekvivalentné sčítavaniu jednotlivých prvkov vektorov, ktoré sú na rovnakých pozíciách. Napríklad:

```
> x <- c(1, 2, 3); y <- c(2, 4, 6); x + y
[1] 3 6 9
```

Príklad 3.4

Uskutočnite nasledujúce operácie `x + z`; `y - z`; `z - x`; `x - z`; kde `z <- c(5, 4, 3, 2, 1)`. Aké výsledky program R vracia?

Príklad 3.5

Vytvorte jednoriadkový príkaz, ktorý vypočíta rozptyl hodnôt {1, 3, 6, 8, 9}. Použite k tomu iba funkcie, ktoré sa doteraz spomínali. Vzorec pre výpočet rozptylu je uvedený v Kapitole 2.3.4.

Ak potrebujeme definovať väčší počet hodnôt, ktoré tvoria určitú postupnosť, môžeme k tomu použiť niektoré predefinované funkcie v R. Ide najmä o základné funkcie `seq()` a `rep()`, ako aj o využívanie operátora „`:`“. Ako príklad si uvedieme tvorbu nasledujúcej postupnosti:

```
> a <- 1; h <- 4; n <- 5;
> a+h*(0:(n-1))
[1]  1  5  9 13 17
> seq(1, 9, by = 2)
[1] 1 3 5 7 9
> seq(1, 10, by = 2)
[1] 1 3 5 7 9
> seq(1, 11, by = 2)
[1]  1  3  5  7  9 11
> rep(1, 3); rep(1:3, 3)
[1] 1 1 1
[1] 1 2 3 1 2 3 1 2 3
```

Príklad 3.6

Vypíšte postupnosť čísel vytvorenú nasledujúcim príkazom bez toho, aby ste to zadali do programu R, `rep(seq(1, 4, by = 2), 2)` a `a + h*(0:n-1)`.

Príklad 3.7

Pomocou príkazov v R vytvorte nasledujúce postupnosti čísel (Verzani, 2004):

1, 3,…, 999

1, 1, 1, 2, 2, 2, 3, 3, 3, ….., 100, 100, 100

1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5

1/1, 1/2, 1/3, …., 1/100; 1/1:10

1, 8, 27, 64, 125, …1000000

0, 25, 50, 75, ….., 1000

Často je potrebné vybrať určitú špecifickú hodnotu z dátového vektora, alebo zapísať určitú hodnotu na určité špecifické miesto v dátovom vektore. Zoberme si nasledujúci vektor `x <- c(1:10)`. Ak chceme zmeniť hodnotu na prvej pozícii vo vektore, potrebujeme programu oznámiť, ktorú hodnotu chceme zmeniť a aká má byť nová hodnota. Indexovanie sa uskutočňuje pomocou hranatých zátvoriek `[]`. Napríklad:

```
> x[1]
[1] 1
> x[1] <- 4; x[10] <- 99
> x[length(x)]
[1] 99
```

Ak je potrebné vybrať určitý širší rozsah údajov z dátového vektora, môžeme to uskutočniť pomocou operátora „`:`“ alebo cez vektor `c()`. Napríklad náhľad prvých štyroch hodnôt vektora `x` vieme urobiť buď cez príkaz `x[1:4]` alebo ako `x[c(1, 2, 3, 4)]`, kde prvá možnosť je zjavne jednoduchšia a úspornejšia. Iným spôsobom, ako zobraziť náhľad (prípadne označiť) všetky hodnoty okrem určitej špecifickej, je použiť operátor „`-`“.

```
> x[-1]; x[-10]
[1]   2 3 4 5 6 7 8 9 99
[1] 4 2 3 4 5 6 7 8 9
```

Za základné logické operátory budeme považovať nasledujúce: `<`, `<=`, `>`, `>=`, `==`, `!=`, `&`, `|`, ktoré znamenajú ostro menší, menší alebo rovný, ostro väčší, väčší alebo rovný, rovný, nerovný, a zároveň, alebo.

```
> x <- c(1:10)
> x > 5
 [1] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
> x <= 5
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
> x >= 5
 [1] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> x == 5
 [1] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
> x != 5
 [1]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
> x <- c(1, 3, 5, 6, 8, 2, 4, 6, 9)
> x > 1 & x < 5
[1] FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE
> x > 5 | x < 2; x
[1]  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
```

```
[1] 1 3 5 6 8 2 4 6 9
> sum(x > 3 & x < 7)
[1] 4
```

Príklad 3.8

Vyskúšajte nasledujúci príkaz a vysvetlite, akú postupnosť (a prečo) hodnôt vracia, `x[seq(from = 1, to = length(x), by = 2)]`.

Ak potrebujeme nahradiť väčší počet hodnôt, ktoré nasledujú za sebou, môžeme ich cez indexovanie označiť a nahradiť v jednej operácii, prípadne ich môžeme do vektora aj pridať. Postup je obdobný. Napríklad:

```
> x
[1] 1 3 5 6 8 2 4 6 9
> x[10:12] <- c(10, 11, 12)
> x
 [1]  1  3  5  6  8  2  4  6  9 10 11 12
```

V tejto súvislosti sa môžeme stretnúť ešte s príkazom `%in%`, ktorý môže slúžiť na hľadanie konkrétnych hodnôt v dátovom vektore, prípadne ak potrebujeme zistiť prienik dvoch vektorov.

```
> x %in% c(2,4)
 [1] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
  FALSE FALSE
```

Mnoho programových balíkov má v sebe zabudované určité súbory dát. Tradičná štruktúra prierezových dát považuje riadky za pozorovania zodpovedajúce štatistickým jednotkám a stĺpce za premenné. Aby sme sa dostali do dátových súborov programových balíkov, musíme si ich najprv aktivovať. Spravidla je to potrebné urobiť stále, keď spustíme program R a chceme s danou databázou pracovať. Použijeme databázu `geyser`, ktorú môžeme nájsť v programovom balíku `MASS`.

```
> library(MASS)
```

Pomocou príkazu `data()` nahráme databázu `geyser`, teda požadovaný príkaz je v tvare: `data(geyser)`. Názvy premenných získame pomocou príkazu `names()`, teda `names(geyser)`. V konzole si databázu môžeme zobraziť vypísaním jej názvu (príkaz `geyser`). Neraz sú však tieto databázy pomerne obsiahle, preto sa zobrazenie celej databázy nezvykne používať. Ak chceme pracovať s určitou premennou musíme ju označiť. Vo

všeobecnosti k tomu používame symbol `$`. Napríklad pri práci s premennou „waiting“ jej vypísanie dosiahneme nasledovným spôsobom – `geyser$waiting`. Prvá časť príkazu označuje databázu a symbolom `$` vyberáme premennú. Iným spôsobom je oznámiť programu, že budeme databázu `geyser` používať priamo. Vtedy môžeme použiť príkaz `attach()`, v tomto konkrétnom prípade napíšeme príkaz `attach(geyser)`. Teraz zobrazenie premenných môžeme uskutočniť priamo napísaním danej premennej. Ak už databázu nepotrebujeme, odhlásime ju príkazom `detach()`.

```
> attach(geyser)
> waiting[1]
[1] 80
> duration[1]
[1] 4.016667
> detach(geyser)
```

Na záver tejto časti si ukážeme ešte niekoľko užitočných príkazov. Ak si chceme uložiť nami vytvorený dátový súbor `x`, aby sme s ním mohli pracovať v inom programe, môžeme k tomu použiť nasledujúci príkaz.

```
> dump("x", "we.txt")
```

Alebo ak chceme uložiť súbor vo formáte .csv na špecifické miesto:

```
> write.csv(x, file = "D:...cesta k súboru...\\udaje.csv", sep =
  ",")
```

Ten sa uloží do adresára, ktorého miesto nájdeme pomocou príkazu:

```
> getwd()
```

Pokiaľ v programe nebude explicitne zadané iné miesto, budú sa súbory a objekty (pozri príkaz `save()`) z programu R ukladať na miesto, ktorého lokalitu zistíme použitím príkazu `getwd()`. Ak chceme zmeniť toto miesto (ide o pracovný adresár, z angl. working directory), tak môžeme použiť príkaz `setwd()`.

Doteraz sme si údaje vytvárali priamo v programe R. Pomerne často sa stretávame s potrebou dostať nami získané dáta do programu R tak, aby sme s nimi mohli pracovať. Ukážeme si jednoduchý postup, ktorý predpokladá, že údaje sú uložené v súbore (s názvom „udaje“) s príponou .csv a majú už vyššie spomínanú štruktúru. Riadky predstavujú pozorovania a stĺpce premenné, pričom v prvom riadku sú názvy premenných a stĺpce sú od seba oddelené bodkočiarkou (to je možné skontrolovať v obyčajnom textovom editore). Príkaz má potom nasledujúci tvar:

```
> read.table("...cesta k súboru...\\udaje.csv", header = T, sep
  = ";", dec = ".")
```

Príklad 3.9

Použite databázu `treering`, nájdite jej opis. Zistite, koľko pozorovaní je väčších ako 1.5 a zároveň menších ako 2.5. Použite databázu `nym.2002` z programového balíka `UsingR`. Koľko rokov mal najstarší účastník maratónu? Aký bol vekový rozdiel medzi najstarším a najmladším účastníkom maratónu?

Niektoré základné opisné charakteristiky súboru sme si už prezentovali. Funkcia `median()` vráti medián, funkcia `mean()` aritmetický priemer, funkcia `IQR()` medzi-kvartilové rozpätie, funkcia `quantile()` hodnotu zodpovedajúcu danému kvantilu (bližšie pozri `?quantile`) a funkcia `summary()` okrem priemeru vráti minimálnu, maximálnu hodnotu súboru, medián, dolný a horný kvartil. Na rozdiel od mnohých iných komerčných štatistických programov, program R ponúka niekoľko spôsobov výpočtu kvantilov.

Aby sme zdôraznili význam kvantilov, ukážeme si na nasledujúcom príklade ich vzťah k podielom. Majme nasledujúci dátový vektor, `a <- c(1, 2, 2, 3, 3, 3, 4, 4, 5, 6, 7, 8, 8, 9, 10)`. Aký podiel z celkového počtu údajov je menších ako 6? Odpoveď môžeme formulovať cez príkaz `sum(a < 6)/length(a)`, ktorý nám vráti hodnotu 0.6, čo zodpovedá 60 % údajom menším ako číslo 6. Pri kvantiloch nás však zaujíma: od ktorej hodnoty usporiadaného štatistického súboru je 60 % údajov menších ako táto hodnota? Ak použijeme základnú funkciu v programe R, odpoveď je `quantile(a, 0.6)` a výsledok, ktorý nám program R vráti, je 5.4. Všimnime si, že túto hodnotu v našom empirickom súbore nemáme. Aj keď výsledok je správny (a správny by bol pre akúkoľvek hodnotu v intervale od 5 do 6), rôzne metódy výpočtu kvantilov používajú rôzne aproximácie, na základe ktorých vyberú jednu z hodnôt v intervale od 5 do 6. Pri väčších empirických súboroch sú rozdiely minimálne.

3.4 Vizualizácia v programe R a triedenie početností

V minulosti (alebo v situáciách, v ktorých nie je k dispozícii štatistický softvér) sa pri väčšom **rozsahu štatistického súboru** vykonávalo tzv. triedenie štatistického súboru. Uvedenému triedeniu sa budeme venovať iba stručne. Na tomto mieste si ho opíšeme, keďže sa používa na tvorbu histogramov – jedného z najčastejšie používaných vizuálnych pomôcok

v štatistike. V kapitole venujúcej sa základným pojmom sme vymedzili pojem **prvotnej tabuľky** a **variačný rad**. V nasledujúcej časti využijeme tieto definície k tomu, aby sme zostavili tzv. **frekvenčnú tabuľku**, pomocou ktorej môžeme získať súhrnný prehľad rozsahovo početnejšieho štatistického súboru. Pri spracovaní pomocou štatistického softvéru existuje možnosť takýto prehľad získať prostredníctvom určitých vizuálnych pomôcok: napríklad pomocou už spomínaného histogramu alebo box – plotu. Upozorňujeme, že po triedení štatistického súboru a následnom počítaní základných charakteristík polohy, variability, kvantilov a mier tvaru vychádzajúc z frekvenčnej tabuľky, je potrebné tieto charakteristiky počítať mierne odlišným spôsobom. V dôsledku triedenia štatistického súboru dochádza k určitej strate informácií a výsledky z takto počítaných charakteristík sa odlišujú od tých, ktoré sme uviedli vyššie. V Kapitole 3.4.2 si vzťahy pre výpočet základných charakteristík z frekvenčnej tabuľky názorne ukážeme. Štatistický softvér pri ľubovoľnom rozsahu štatistického súboru počíta uvedené charakteristiky pri opisnej štatistike spôsobom, aký sme prezentovali v predchádzajúcich častiach[11].

3.4.1 Frekvenčná tabuľka

Majme hodnoty štatistického súboru utriedené do variačného radu $X'_{(1)} \leq X'_{(2)} \leq \ldots \leq X'_{(n)}$. Niektoré hodnoty utriedeného štatistického súboru sa môžu rovnať, napr. $2 \leq 3 \leq 3 \leq 3 \leq 4 \leq 5 \leq 5 \leq 6 \leq 6$. Urobme ďalšiu úpravu, kde variačný rad zotriedime tak, že žiadne dve hodnoty sa vo variačnom rade nemôžu vyskytovať viac ako jeden krát, teda $X_{(1)} < X_{(2)} < \ldots < X_{(s)}$, pričom zjavne $s \leq n$. V našom ilustratívnom prípade tak máme $2 < 3 < 4 < 5 < 6$. Niektoré hodnoty sa však vyskytujú početnejšie a túto početnosť nazývame **absolútnou početnosťou** hodnoty $X_{(j)}$ pre $i, j = 1, 2, \ldots, s$ a označujeme ju ako $n_{(j)}$. Ak by sme chceli pozorovať, ako narastá celkový súčet absolútnych početností v usporiadanom variačnom rade, mohli by sme to pozorovať pomocou absolútnej kumulatívnej početnosti $N_{(i)}$, pre ktorú platí vzťah:

$$N_{(i)} = \sum_{j=1}^{i} n_{(j)} \qquad (3.1)$$

Kde $i \leq s$ a predstavuje kumulovaný súčet početností po i-tu hodnotu radu a zrejme platí, že $N_{(s)} = \sum_{j=1}^{s} n_{(j)} = n$. Takto usporiadaný štatistický súbor sa zvykne zobrazovať vo forme tabuľky (Tabuľka 1).

[11] Mierne odlišnosti vznikajú v prípadoch, kde štatistický softvér namiesto opisných charakteristík polohy, variability a mier tvaru využíva tzv. výberové charakteristiky, bližšie pozri publikáciu venujúcu sa induktívnej štatistike.

Tabuľka 1: Absolútne početnosti

poradie hodnôt	hodnota	absolútna početnosť	absolútna kumulatívna početnosť
1	2	1	1
2	3	3	4
3	4	1	5
4	5	2	7
5	6	2	9

Zdroj: vlastné spracovanie

Absolútna aj kumulatívna početnosť (obe rôznymi spôsobmi) slúžia na sledovanie, ako sú rozdelené (distribuované) početnosti hodnôt štatistického súboru. Rovnaký účel majú odvodené miery, tzv. relatívna početnosť $f(j)$ a kumulatívna relatívna početnosť $F(i)$, ktoré vyjadrujú podiel početnosti (resp. absolútnej kumulatívnej početnosti) na celkovom rozsahu štatistického súboru, teda:

$$f_{(j)} = \frac{n_{(j)}}{n} \tag{3.2}$$

$$F_{(i)} = \sum_{j=1}^{i} f_{(j)} \tag{3.3}$$

Tabuľka 2: Frekvenčná tabuľka – ukážka 1

poradie hodnôt	hodnota	absolútna početnosť	relatívna početnosť	relatívna kumulatívna početnosť
1	2	1	0.11	0.11
2	3	3	0.33	0.44
3	4	1	0.11	0.55
4	5	2	0.22	0.77
5	6	2	0.22	1

Zdroj: vlastné spracovanie

Takto usporiadanej tabuľke (Tabuľka 2) hovoríme **frekvenčná tabuľka**. Niekedy sa k frekvenčnej tabuľke pridáva aj stĺpec triednych znakov (pozri ďalej). Všimnime si, že z pôvodného štatistického súboru v našom ilustratívnom prípade o rozsahu $n = 9$ sme triedením dosiahli rozsah $s = 5$. Z tohto dôvodu sa tento proces spracovania dát nazýva aj tzv. **zhusťovanie**. Napriek tomu pri extrémne väčších rozsahoch údajov s veľkým variačným rozpätím, s ktorými sa môžeme stretnúť, by ani takého zhusťovanie nestačilo. Ak by sme mali $n = 4000$, týmto utriedením by sme sotva mohli očakávať, že sa nám podarí roztriediť štatistické údaje do napríklad $s \leq 20$, a teda aj takto utriedený súbor by bol pomerne neprehľadný. Vzhľadom na to, že štatistické programy nasledujúcu formu triedenia

uskutočňujú v prípade potreby automaticky, v ďalšej časti si ukážeme spôsob rozdeľovania štatistických súborov len rámcovo.

Počet tried *k*, do ktorých si chceme zotriediť štatistický súbor s ľubovoľným rozsahom *n* je subjektívnou voľbou. Existujú však určité odporúčania, ako napr. Sturgessovo pravidlo:

$$k \approx 1 + 3.322 \log n \qquad (3.4)$$

Iným je pravidlo uvádzané v Tkáč (2001), pri ktorom si môžeme vybrať počet tried z intervalu:

$$0.55n^{0.4} \leq k \leq 1.25n^{0.4} \qquad (3.5)$$

K ďalšiemu pravidlu patrí:

$$k \approx \sqrt{n} \qquad (3.6)$$

V programe R je pri tvorbe histogramov základným pravidlom Sturgessovo pravidlo, okrem toho je možnosť použiť aj Scottovo pravidlo:

$$h \approx \frac{3.5s}{n^{1/3}} \qquad (3.7)$$

kde *s* je výberová smerodajná odchýlka, s ktorou sa stretneme v publikáciách venujúcich sa induktívnej štatistike. Posledným menovaným pravidlom (program R ho ponúka) je Freedman-Diaconisovo pravidlo:

$$h \approx \frac{2R_Q}{n^{1/3}} \qquad (3.8)$$

Pri posledných dvoch pravidlách si všimnime, že odhadujú šírku intervalu a nie počet. Počet sa následne dopočíta jednoduchou úpravou vzťahu (3.10). Druhou zaujímavosťou je, že tieto dve pravidlá pracujú nie len s počtom hodnôt, ale aj s ich variabilitou.

Každé z týchto pravidiel je potrebné brať len ako určité odporúčanie. V konečnom dôsledku záleží na zhotoviteľovi analýzy. Niektoré štatistické metódy (ako napr. Chí-kvadrát test dobrej zhody) si vyžadujú, aby boli hodnoty štatistického súboru roztriedené do tried. Niektoré softvérové balíky v týchto prípadoch umožňujú užívateľovi vybrať si počet tried (manuálne). Táto subjektívna úloha nie je úplne triviálna, keďže príliš malý počet tried *k* by mohol agregovať veľký podiel hodnôt do jednej triedy a tým by mohlo dôjsť k výraznej strate informácií. Na druhej strane, voľba príliš veľkého počtu tried *k* by mohla spôsobiť, že niektoré triedy nebudú obsahovať žiadne hodnoty, čím sa takto upravený prehľad údajov stane neprehľadným, prípadne, že by nebolo dosť dobre možné vizuálne odhadnúť druh rozdelenia pravdepodobnosti (bližšie pozri Kapitolu 4.2).

Každá z týchto tried má svoju dolnú a hornú hranicu $K_j = <t_{j-1}, t_j)$, K_j je príslušný interval (trieda), kde $j = 1, 2, ..., k$ a t_{j-1}, t_j je dolná, resp. horná hranica *j*-teho intervalu. Aby sme mohli počítať určité charakteristiky z takto vytvorených intervalov, je potrebné

zadefinovať určitého reprezentanta intervalu. Ak máme napríklad iba 2 intervaly $K_1 = \langle 3, 4)$, $K_2 = \langle 4, 5)$ s nejakou absolútnou triednou (intervalovou) početnosťou, ako vypočítame aritmetický priemer? Vytvoríme tzv. triedny znak $Z_{(j)}$, pre ktorý platí:

$$Z_{(j)} = \frac{1}{2}\left(t_{j-1} + t_j\right) \tag{3.9}$$

Ide teda o aritmetický priemer. V našom prípade $Z_{(1)} = ½ (3 + 4) = 3.5$.

Už vieme, aký počet intervalov chceme vytvoriť, vieme ako vypočítame ich triedny znak. Ešte stále však chýba spôsob výpočtu hraníc intervalov. K tomu potrebujeme vedieť šírku triednych intervalov, ktorá by mala byť konštantná (v prípade prvého a posledného intervalu ale nemusí byť, napr. keď sa uvádza vek „18 a menej“ alebo „60 a viac“). Pri tvorbe intervalov (napríklad pri tvorbe dotazníkov) je vhodné, ak sú splnené dve podmienky:

- každé meranie musí byť jednoznačne zaradené práve do jednej triedy,
- šírka triednych intervalov musí byť konštantná[12].

Príklad 3.10

Častým omylom je tvorba takých intervalov v dotazníkoch, kde jednu hodnotu nie je možné priradiť do intervalu. V dotazníku môže byť položená otázka, v ktorej si respondent má vybrať, do akej vekovej kategórie patrí. Nasledujúce členenie nie je presné: do 18 rokov, od 19 do 25 rokov, od 26 do 40 rokov, 41 a viac rokov (pokiaľ nie je explicitne povedané, že ide o dovŕšený vek). Ak by mal respondent 18.5 rokov nebolo by jasné, do ktorej kategórie by sa zaradil. Druhou nevýhodou je, že intervaly (okrem prvého a posledného) nie sú rovnako široké.

Šírku triedneho intervalu vypočítame pomocou jednoduchého vzťahu:

$$h \approx \frac{R}{k} = \frac{X_{(n)} - X_{(1)}}{k} = \frac{\max\limits_{i=1,2,\ldots,n} X_i - \min\limits_{i=1,2,\ldots,n} X_i}{k} \tag{3.10}$$

Príklad 3.11

Majme nasledujúci rozsah údajov:

```
> 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,
```

[12] Uvedené v určitých špecifických situáciách platiť nemusí, napr. neraz je prvý (najmenší) a posledný (najväčší) interval podstatne väčší, napr. vek $K_s = 65$ a viac.

```
4100, 3800, 6400, 2500, 4100, 4400, 3800, 2600, 5200, 4900,
4500)
> length(data)
[1] 39
```

Pomocou zhusťovania vytvoríme frekvenčnú tabuľku. Počet tried si vypočítame vzťahom *(3.5)*, teda:

$$0.55n^{0.4} \leq k \leq 1.25n^{0.4}$$

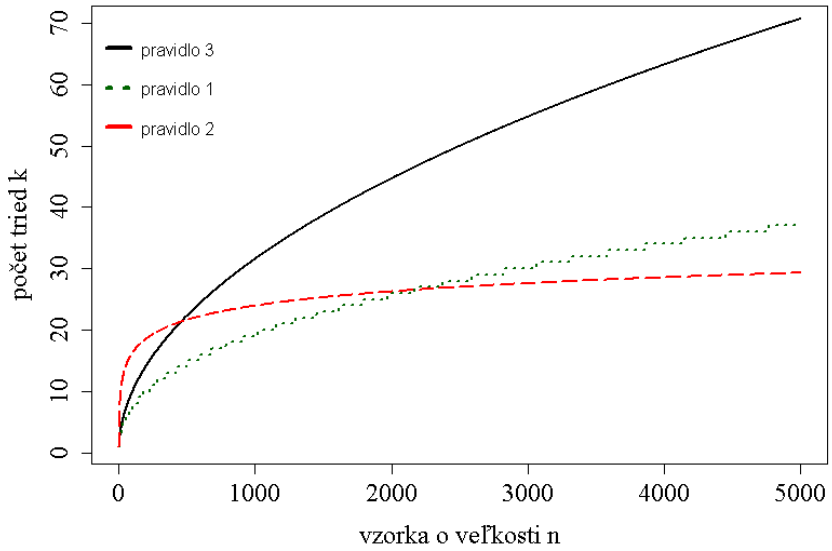
$$2.38 \leq k \leq 5.41$$

Keďže $k \in \mathbb{N}$ je celé prirodzené číslo, tak máme na výber $k \in \{3, 4, 5\}$. Vyberieme si k = 5. Pre porovnanie počítajúc Sturgessovým pravidlom je $k \approx 6.28$ a vzťahom $\sqrt{n}$, $k \approx 6.24$. Na nasledujúcom obrázku pre zaujímavosť môžeme vidieť, ako sa mení počet odporúčaných tried v závislosti od rozsahu štatistického súboru, kde:

pravidlo 1: $\lfloor k \rfloor = 1.25n^{0.4}$.

pravidlo 2: k sa vypočíta ako 1 + 3.322 log(n) zaokrúhlené na celé číslo nadol.

pravidlo 3: k sa vypočíta ako $\sqrt{n}$ zaokrúhlené na celé číslo nadol.



Obrázok 3.2: Počet tried podľa pravidiel

Zdroj: vlastné spracovanie v programe R

Pre úplnosť uvádzame aj príslušný kód v programe R, pomocou ktorého sme obrázok vytvorili. V ďalšej časti tejto publikácie budeme uvádzať kódy k obrázkom a to jednak z dôvodu, aby ich bolo možné reprodukovať a zároveň, aby čitateľ mohol pomocou týchto kódov riešiť svoje vlastné úlohy. Vizualizácii sa budeme venovať podrobnejšie v samostatnej kapitole.

```
> n <- 5000
> pravidla <- list(floor(1.25*(1:n)^0.4), 1 + 3.322*log((1:n)),
  sqrt(1:n))
> plot(x = 1:n, y = pravidla[[3]], type = "l", family = "serif",
  xlab = "vzorka o veľkosti n", ylab = "počet tried k", ylim =
  c(1, 70), cex.axis = 1.5, cex.lab = 1.5, col = "black", lwd =
  2)
> colour = c("darkgreen", "red")
> l_type = c(3, 5)
> labels <- c("pravidlo 3", "pravidlo 1", "pravidlo 2")
> for (i in 1:2) {
+  lines(x = 1:n, y = pravidla[[i]], type = "l", lty =
  l_type[i], col = colour[i], lwd = 2)
+ }
> legend("topleft", labels, lty = c(1, l_type), inset = 0.005,
  bty = "n", cex = 1, col = c("black", colour), lwd = 4)
```

Šírku triedneho intervalu si potom vypočítame dosadením do vzťahu:

$$h \approx \frac{R}{k} = \frac{X_{(n)} - X_{(1)}}{k} = \frac{\max\limits_{i=1,2,\ldots,n} X_i - \min\limits_{i=1,2,\ldots,n} X_i}{k} = \frac{7200 - 2000}{5} = 1040$$

Následne si vypočítame triedne znaky, priradíme hodnoty a vytvoríme frekvenčnú tabuľku:

Tabuľka 3: Frekvenčná tabuľka – ukážka 2

Interval	Triedny znak	absolútna početnosť	relatívna početnosť	absolútna kumulatívna početnosť	relatívna kumulatívna početnosť
<2000; 3040)	2520	5	0.13	5	0.13
<3040; 4080)	3560	12	0.31	17	0.44
<4080; 5120)	4600	13	0.33	30	0.77
<5120; 6160)	5640	6	0.15	36	0.92
<6160; 7200>	6680	3	0.08	39	1

Zdroj: vlastné spracovanie

Všimnime si, že posledný interval je z oboch strán uzavretý. Je to nutné, keďže všetky merania musia byť zaradené práve do jednej triedy. Aj tá najväčšia. Z takto vytvorenej frekvenčnej tabuľky si môžeme vytvoriť tzv. histogram, čo je vlastne špecifický prípad stĺpcového grafu[13], kde šírka stĺpca je ekvivalentná šírke triedneho intervalu a výška stĺpca je ekvivalentná absolútnej (alebo relatívnej) početnosti.

Pri rozhodovaní sa, akým spôsobom zobraziť a prezentovať údaje, si spravidla vystačíme s ich delením na tri kategórie: kategorické, diskrétne a spojité dáta. Pri kategorických dátach ako napríklad: farba auta, názov značky, typ automobilu, nemá význam

[13] Na rozdiel od stĺpcového grafu sú jednotlivé stĺpce zobrazované spolu „bez medzier“ medzi stĺpcami.

počítať priemer, variabilitu alebo iné charakteristiky. Má však význam sledovať počet výskytu tej ktorej kategórie. Pri diskrétnych už vieme povedať, ktoré hodnoty sú väčšie a ktoré menšie (ide napríklad o poradia, početnosť výskytu určitého javu a pod.). Počítať priemer alebo charakteristiky variability vo väčšine prípadov nemá význam, aj keď v empirickom výskume sa s tým stretávame pomerne často. V niektorých prípadoch je to odôvodniteľné a aj o týchto diskrétnych dátach sa dá uvažovať ako o spojitých (napr. ak máme veľa poradí jednej premennej). Pri spojitých má spravidla význam počítať tak priemer, ako aj iné základné štatistické charakteristiky. Nevyčerpali sme síce všetky typy dát a ani náš opis nebol úplne presný, avšak vo väčšine aplikácií si s týmto členením vystačíme.

Typ dát nie je jediným rozhodovacím kritériom. Cieľ prezentácie údajov je nemenej dôležitým. Niekedy sa vyžaduje, aby vizuálna prezentácia údajov bola tak povediac samonosná. To znamená, aby bez bližšieho vysvetlenia bolo prijímateľovi zrejmé, na aké údaje sa pozerá a čo má z danej prezentácie v údajoch vidieť, prípade aký jav sa mu chce ukázať.

3.4.2 Opisné charakteristiky pre frekvenčné tabuľky

Ak Z_j je triedny znak, n_j je absolútna početnosť triedneho intervalu, k je počet tried (intervalov), kde j = 1, 2, ..., k a $n = \sum_{j=1}^{k} n_j$ je rozsah štatistického súboru, potom môžeme vypočítať základné miery polohy a variability pomocou nasledujúcich vzťahov.

- Vážený aritmetický priemer:

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{k} Z_j n_j \tag{3.11}$$

kde n je súčtom n_j cez j = 1, 2, ..., k.

- Vážený geometrický priemer:

$$\overline{X}_G = \sqrt[n]{\prod_{j}^{k} Z_j^{\,n_j}} \tag{3.12}$$

- Vážený harmonický priemer:

$$\overline{X}_H = \frac{n}{\sum_{j=1}^{k} \frac{n_j}{Z_j}} \tag{3.13}$$

Pri váženom chronologickom priemere je situácia komplikovanejšia, keďže rozmer času sa zvyčajne udáva v počtoch dní medzi dvoma za sebou nasledujúcimi obdobiami, a preto je váhou počet dní d_j pre j = 2, 3, ..., k.

$$\overline{X}_{CH} = \frac{\frac{Z_1 + Z_2}{2} d_2 + \frac{Z_2 + Z_3}{2} d_3 + \ldots + \frac{Z_{k-1} + Z_k}{2} d_k}{\sum_{j=2}^{k} d_j} \tag{3.14}$$

Uvažujme o mediánovom intervale ako o tom, v ktorom sa nachádza $X_{(0.5n)}$-tá hodnota variačného radu. Túto hodnotu vieme identifikovať vo frekvenčnej tabuľke na základe absolútnej (relatívnej) kumulatívnej početnosti. Potom nech E je dolná hranica mediánového intervalu, h šírka intervalu, $N_{(i-1)}$ nech je absolútna kumulatívna početnosť triedneho intervalu nachádzajúceho sa pred mediánovým intervalom a $\tilde{n}_{(j)}$ nech je absolútna početnosť mediánového intervalu. Medián vypočítame ako:

$$\tilde{X} = E + h \frac{\frac{n}{2} - N_{(i-1)}}{\tilde{n}_{(j)}} \tag{3.15}$$

Pre ostatné kvantily môžeme použiť obdobný vzťah vychádzajúci z relatívnych početností. Nech a_r je dolnou hranicou r-tého kvantilového intervalu. Ak je predmetom nášho záujmu 5-ty percentil, potom r predstavuje dolnú hranicu takého intervalu, v ktorom sa 5-ty percentil s určitosťou nachádza. To môžeme ľahko zistiť pomocou relatívnej kumulatívnej početnosti. Nech h je už spomínaná šírka intervalu a podobne ako v kapitole venujúcej sa kvantilom $p = r / \alpha$, kde α udáva na koľko častí sa pomocou kvantilov rozdeľuje štatistický súbor a r je poradie kvantilu, pre ktoré platí $r = 1, 2, \ldots, \alpha - 1$. Ďalej nech $F_{(i_r-1)}$ je kumulatívna relatívna početnosť intervalu bezprostredne predchádzajúceho r kvantilového intervalu a $f_{(j_r)}$ je relatívna početnosť r kvantilového intervalu. Potom pre hodnoty kvantilov platí:

$$X_r = a_r + h \frac{p - F_{(i_r-1)}}{f_{(j_r)}} \tag{3.16}$$

Pre frekvenčné tabuľky si modálnu hodnotu vypočítame na základe nasledujúceho vzťahu. Tento postup zopakujeme pre všetky intervaly, ktoré majú najväčšiu početnosť (v prípade, ak sa v štatistickom súbore nachádza viac modálnych hodnôt):

$$\hat{X} = M + h \frac{d_p}{d_p + d_n} \tag{3.17}$$

Kde M je začiatok modálneho intervalu, $d_p = \hat{n}_{(j)} - \hat{n}_{(j-1)}$, kde $\hat{n}_{(j)}$ je absolútna početnosť modálneho intervalu a $d_n = \hat{n}_{(j)} - \hat{n}_{(j+1)}$.

Podobne ako pri mierach polohy, musíme rovnakým princípom upraviť aj miery variability a tvaru. Tieto zmeny si ukážeme na najčastejších mierach variability: priemernej

absolútnej odchýlke a rozptylu a najčastejšie používaných mierach tvaru: šikmosť a špicatosť (počítané cez tzv. momentové charakteristiky). V jednotlivých vzťahoch zodpovedá výrazu n_j početnosť (váha) triedy j.

- Vážená priemerná absolútna odchýlka:

$$s_{\bar{d}} = \frac{1}{n}\sum_{j=1}^{k}\left|Z_j - \bar{X}\right| n_j \tag{3.18}$$

- Vážený rozptyl:

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{k}\left(Z_j - \bar{X}\right)^2 n_j \tag{3.19}$$

- Vážený koeficient šikmosti:

$$S = \frac{\frac{1}{n}\sum_{j=1}^{k}\left(Z_j - \bar{X}\right)^3 n_j}{\sqrt{\left(\frac{1}{n}\sum_{j=1}^{k}\left(Z_j - \bar{X}\right)^2 n_j\right)^3}} \tag{3.20}$$

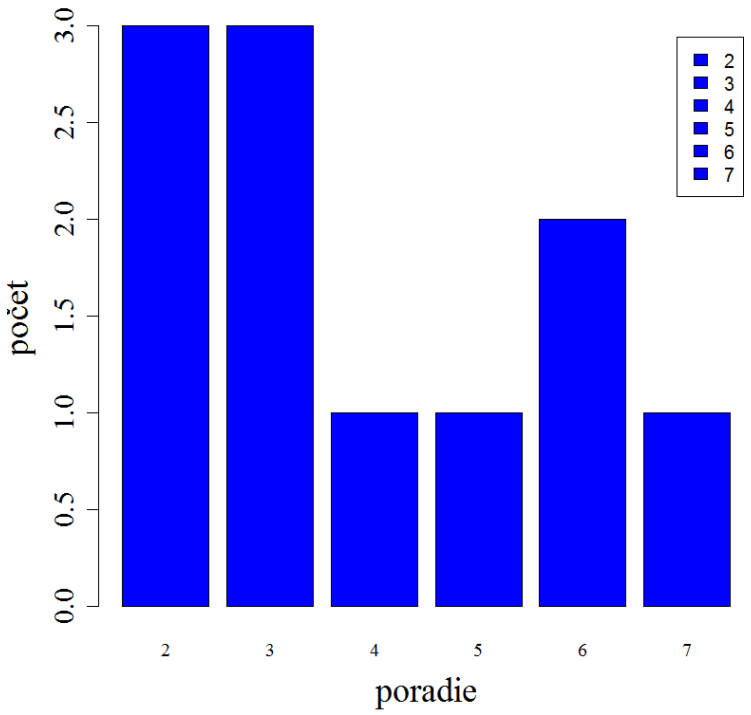
- Vážený koeficient špicatosti:

$$K = \frac{n\sum_{j=1}^{k}\left(Z_j - \bar{X}\right)^4 n_j}{\left(\sum_{j=1}^{k}\left(Z_j - \bar{X}\right)^2 n_j\right)^2} - 3 \tag{3.21}$$

3.4.3 Stĺpcový graf

K zostrojeniu stĺpcového grafu stačí v programe R použiť funkciu `barplot()`. Takto použitá funkcia zobrazí všetky jedinečné hodnoty ako stĺpce, kde na osi x-ovej sú poradia hodnôt tak, ako boli zadané v dátovom vektore a os y-ová predstavuje samotnú veľkosť hodnôt. Podľa tejto veľkosti sa zobrazí aj výška stĺpcov. Pri väčšom množstve hodnôt je stĺpcový graf neprehľadný. Stĺpcový graf sa preto používa v odlišnej podobe, a to kombináciou dvoch funkcií: `table()` a `barplot()`. Funkcia `table()` vytvorí tabuľku a z nej sa následne vytvorí stĺpcový graf pomocou `barplot()`.

```
> res <- c(2, 2, 2, 3, 3, 3, 4, 5, 6, 6, 7)
> table(res)
res
2 3 4 5 6 7
3 3 1 1 2 1
> barplot(table(res), legend = TRUE, xlab = "poradie", ylab =
  "počet", col = "blue", family = "serif", cex.axis = 1.5,
  cex.lab = 1.8)
```



Obrázok 3.3: Stĺpcový graf – ukážka 1

Zdroj: vlastné spracovanie v programe R

Parametre funkcie (občas ich budeme nazývať argumenty) ako `legend`, `xlab`, `ylab` a `col` sa spravidla dajú voliť pre každý typ grafu. Ak je potrebné presnejšie špecifikovať umiestnenie legendy, existuje funkcia `legend()`. Niektoré možnosti si budeme postupne ukazovať spolu s ďalšími obrázkami (v texte sa o týchto možnostiach budeme explicitne vyjadrovať len zriedkakedy). Napríklad ak chceme jednotlivým kategóriám v tabuľke pridať názvy, môžeme to urobiť pomocou funkcie `names()`.

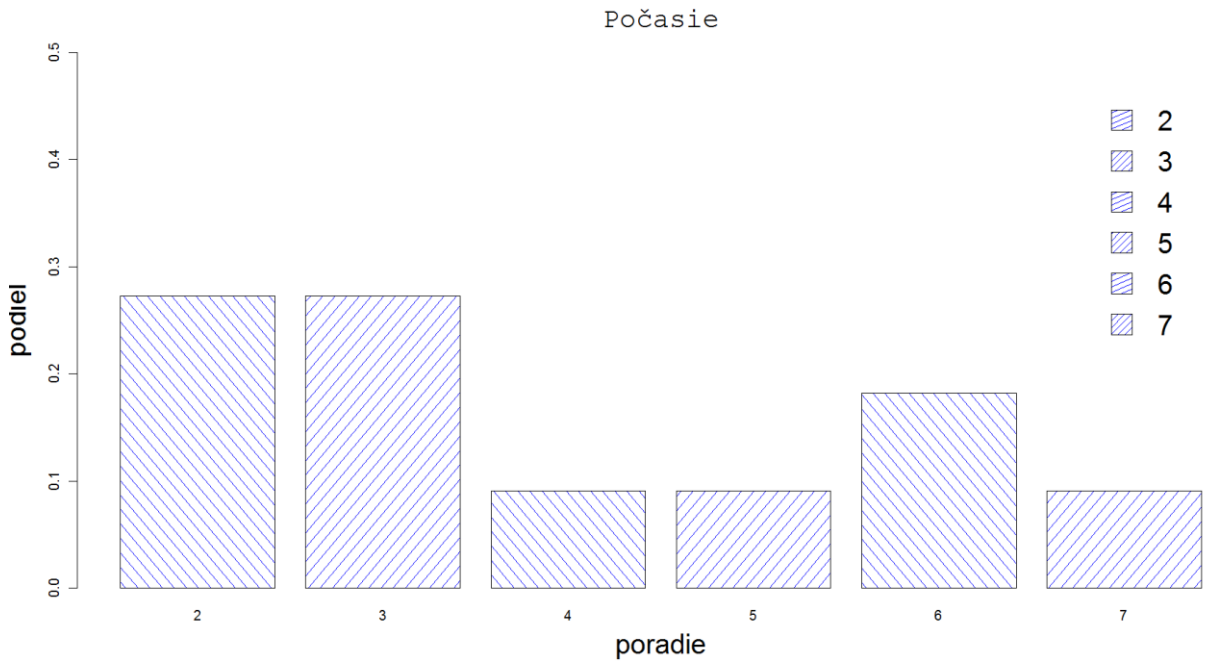
Príklad 3.12

Čo robí nasledovný príkaz? Porozmýšľajte nad tým bez toho, aby ste príkaz v softvéri R spustili.

```
ts <- table(res)/length(res)
```

Potom príkaz spustite, skúste zmeniť niektoré parametre a opíšte, čo jednotlivé možnosti v stĺpcovom grafe menia. Tento stĺpcový graf je znázornený na nasledujúcom obrázku (Obrázok 3.4).

```
> ts <- table(res)/length(res)
> barplot(ts, xlab = "poradie", legend = TRUE, ylab = "podiel",
  cex.lab = 2, col = "blue", angle = c(135,45), density = 10,
  main = list("Počasie", cex = 2, font = 10), ylim = c(0,0.5),
  args.legend = list(x = "topright", inset = 0.05, bty = "n",
  cex = 2, fill = "blue", density = 20, angle = c(20,40)))
```



Obrázok 3.4: Stĺpcový graf – ukážka 2

Zdroj: vlastné spracovanie v programe R

Príklad 3.13

Čo je nevyhovujúce na nasledujúcich stĺpcových grafoch?

```
> barplot(table(res)/length(res), legend = TRUE, xlab =
  "poradie", ylab = "podiel", col = "blue", main = "Wheather",
  ylim = c(0.1,0.5))
> barplot(table(res)/length(res), legend = TRUE, xlab =
  "poradie", ylab = "podiel", col = "blue", main = "Wheather",
  ylim = c(0,0.25))
```

V ďalšom texte budeme potrebovať programový balík `UsingR`. Po nainštalovaní ho spustíme pomocou príkazu `library(UsingR)`. Následne si otvoríme databázu `central.park` pomocou príkazu `data()`. Ilustrujeme si zmenu názvu jednotlivých premenných na x-ovej osi. To sa môže hodiť pri vizualizácii niektorých časových radov.

```
library(UsingR); data(central.park)
> barplot(central.park$MAX, names.arg = 1:31, xlab = "day", ylab
  = "max.temp.")
> barplot(central.park$MAX, names.arg = c("ano", 2:20, "hej",
  22:31), xlab = "day", ylab = "max.temp.")
```

Príklad 3.14

Načítajte nasledujúce údaje do vektora `beer`, ktorý predstavuje, aký typ piva študenti preferujú: 1 – domáce plechovky, 2 – domáce fľaše, 3 – malé varené, 4 – importované. Údaje môžeme do vektora `bear` vložiť aj nasledujúcim spôsobom:

```
> beer <- scan()
1: 3 4 1 1 3 4 3 3 1 3 2 1 2 1 2 3 2 3 1 1 1 1 4 3 1
26:
Read 25 items
```

Vytvorte jednoduchý stĺpcový graf početností odpovedí tak, aby stĺpce postupne klesali. Zopakujte to aj pre klesajúce podiely odpovedí.

```
> barplot(sort(table(beer), decreasing = TRUE), col =
  rainbow(4), main = "Beer preferences")
```

3.4.4 Histogram

Pri číselných dátach (diskrétnych aj spojitých) pomocou vhodne zvoleného obrázku môžeme zobraziť niektoré dôležité charakteristiky údajov. Jedným z najčastejšie používaných obrázkov je histogram. Dobre zvolený histogram naznačuje, okolo ktorej číselnej hodnoty (a či vôbec) sa údaje sústreďujú, aká je ich variabilita, prípadne aký je tvar rozdelenia početností údajov v súbore, alebo či sa v súbore nachádzajú extrémne hodnoty. Vizualizácia hodnôt je forma prehľadného usporiadania hodnôt štatistického súboru, podobne ako frekvenčná tabuľka alebo variačný rad. Ľudský mozog je schopný spracovať najľahšie a najviac informácií práve na základe vizuálnych vstupov, teda pomocou ľudského zmyslu – zraku. Na rozdiel od tabuliek a čísel, obrázky sú pre ľudské oko väčším podnetom. Obrázky vedia byť nositeľom informácií, ktoré sa z jednoduchých charakteristík polôh a variability nedajú vyčítať. Histogram patrí k najpoužívanejším nástrojom na vizualizáciu hodnôt štatistického súboru. Zachytáva množstvo užitočných informácií, ktoré si je možné rýchlo pozrieť a urobiť si tak prvý prehľad o štatistickom súbore (upravené podľa Tkáč, 2001):

- Dáva informáciu o rozdelení početností, teda v ktorých intervaloch sa aký počet hodnôt nachádza. Môžeme tak odhadnúť, v akých intervaloch sú najpravdepodobnejšie hodnoty, resp. môžeme pozorovať určitú tendenciu koncentrácie hodnôt štatistického súboru pri určitých hodnotách.

Príklad 3.15

Ak by sme chceli zmerať dĺžku 12 cm ceruzky meradlom, ktoré meria s presnosťou ± 1 mikrometra a vykonali by sme merania, môžeme očakávať, že dosiahneme rôzne dĺžky v jednotkách mikrometra (vplyvom meracieho procesu, vonkajších vplyvov, teploty atď.). Tieto hodnoty sa zrejme budú koncentrovať v okolí 12 cm. V iných situáciách nie je známa apriórna hodnota, v okolí ktorej (ktorých) by sa hodnoty mali koncentrovať. Tu sme vedeli, že ideme pomerať práve 12 cm ceruzku. Spracovaním hodnôt štatistického súboru do histogramu túto informáciu môžeme získať.

- Dáva informáciu o šikmosti a špicatosti štatistického súboru.
- Môžeme pozorovať, či má štatistický súbor jeden alebo viac modálnych hodnôt.
- Na základe šikmosti môžeme odhadnúť aj vzťah medzi ostatnými mierami polohy.
- Môžeme pozorovať výskyt odľahlých hodnôt, tzv. extrémnych hodnôt.
- Slúži na odhad teoretického rozdelenia pravdepodobností (bližšie pozri Kapitolu 4.2).
- Môžeme pozorovať variabilitu hodnôt v štatistickom súbore.

Okrem analytických účelov je histogram vhodným prezentačným nástrojom využívaným na zobrazovanie rozdelenia početností štatistického súboru pri prezentovaní výsledkov z prieskumov, výskumov a experimentov. Vzájomným porovnaním viacerých histogramov môžeme odhadnúť rozdiel v stredných hodnotách, ako aj mierach variability medzi rôznymi štatistickými súbormi.

Histogram sa v programe R vytvára pomocou príkazu `hist()`.

```
> attach(faithful)
> hist(waiting, breaks = "Scott")
```

Samozrejme, takto koncipovaný graf je len vzorovým a užívateľ si ho môže prispôsobiť pomocou argumentov, ktoré daná funkcia ponúka.

Príklad 3.16

Nasledujúci príkaz vytvorí pomerne **nevhodný** histogram (nie len z estetického hľadiska). Skúste zmeniť argumenty vo funkcii `hist()` tak, aby bol histogram podľa vášho názoru vhodnejší. Aby ste vedeli čo môžete zmeniť, skúšajte náhodne meniť argumenty funkcie.

```
> hist(waiting, breaks = "Scott", density = 20, col = "blue",
  lwd = 3, lty = 2, cex.axis = 1.1, cex.lab = 1.5, cex.main = 2,
  col.axis = "red", col.lab = "grey", col.main = 4, fg =
  "yellow", font.main = 4, font.lab = 2, font.axis = 3, las =
  0.2, pch = 4)
```

Príklad 3.17

Nájdite databázu `OBP`, ktorá obsahuje jednu premennú. Aby ste mohli robiť zmysluplné interpretácie, zistite o akú premennú ide. Zostrojte histogram s 12 intervalmi. Interpretujte výsledok: miery polohy a variability, extrémne hodnoty a tvar rozdelenia početností. Pri riešení úlohy použite argument "`prob = T`".

Do histogramu, ako aj do mnohých iných grafov, je možné vkladať ostatné grafy. V prípade histogramu sa ako obzvlášť vhodná javí funkcia `density()`, pomocou ktorej je možné preložiť cez histogram krivku určitého (z údajov odhadnutého) rozdelenia pravdepodobnosti (prípadne početnosti). Na teraz stačí, ak si pod touto krivkou predstavíme krivku, ktorá vyhladí vrcholy stĺpcov histogramu.

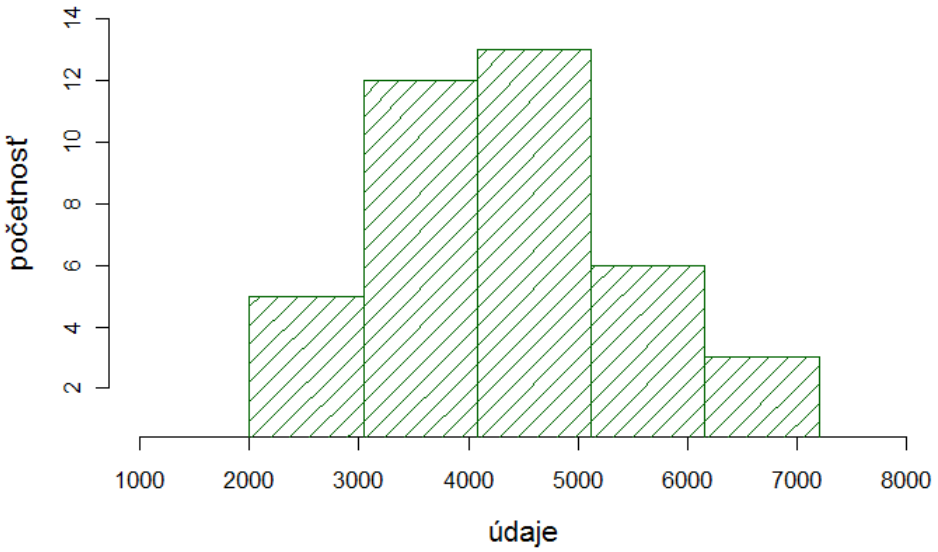
```
> hist(waiting, breaks = "sturges", main = NULL, prob = T,
  ylab=" ")
> lines(density(waiting))
```

Príklad 3.18

Pracujte s databázou `cfb` a zostrojte histogram zo všetkých premenných, kde to je možné. Rozhodnite, ktoré z rozdelení početností sú pravostranne zošikmené.

Na záver uvedieme Obrázok 3.5, ktorý predstavuje jednoduchý histogram z údajov, ktoré sme použili pri tvorbe frekvenčnej tabuľku v Kapitole 3.4.1 (Tabuľka 3).

```
> bins <- seq(from = 2000, to = 7200, by = 1040)
> hist(data, breaks = bins, xlab = "údaje", ylab = "početnosť",
  main = NA, ylim = c(1,15), xlim = c(1000, 8000), cex.axis =
  1.1, cex.lab = 1.3, density = 10, col = "darkgreen")
```

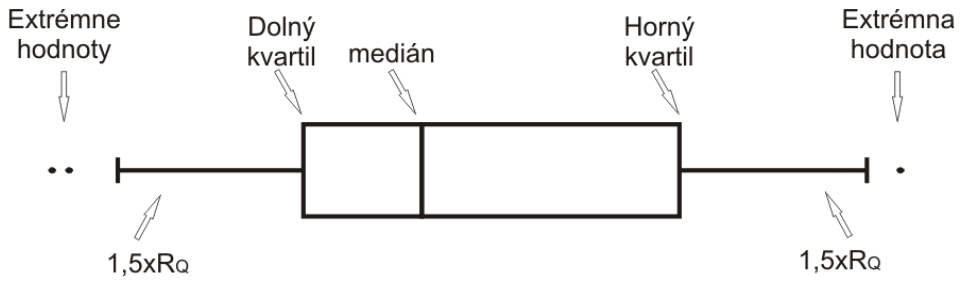


Obrázok 3.5: Histogram – ukážka

Zdroj: vlastné spracovanie v programe R

3.4.5 Box – plot

Box – plot je po histograme možné považovať spolu s *x-y* grafom za jednu z najčastejšie používaných vizuálnych pomôcok. Box – plot využíva kvantily, konkrétne už nami spomínaných päť číselných zhrnutí štatistického súboru: $X_{(1)}, X_{(\lceil n0.25 \rceil)}, \tilde{X}, X_{(\lceil n0.75 \rceil)}, X_{(n)}$. Príslušné kvantily si vieme určiť pomocou funkcie `quantile()`, prípadne ak nás zaujíma medzi-kvartilové rozpätie tak `IQR()`. Pre stručný prehľad základných štatistík súboru si spravidla vystačíme s funkciou `summary()` alebo niekedy aj `fivenum()`, ktoré nám vrátia minimálnu hodnotu, dolný kvartil, medián, horný kvartil a maximálnu hodnotu. Tie sa potom nanesú na graf tak, ako je to znázornené na nasledujúcom obrázku:



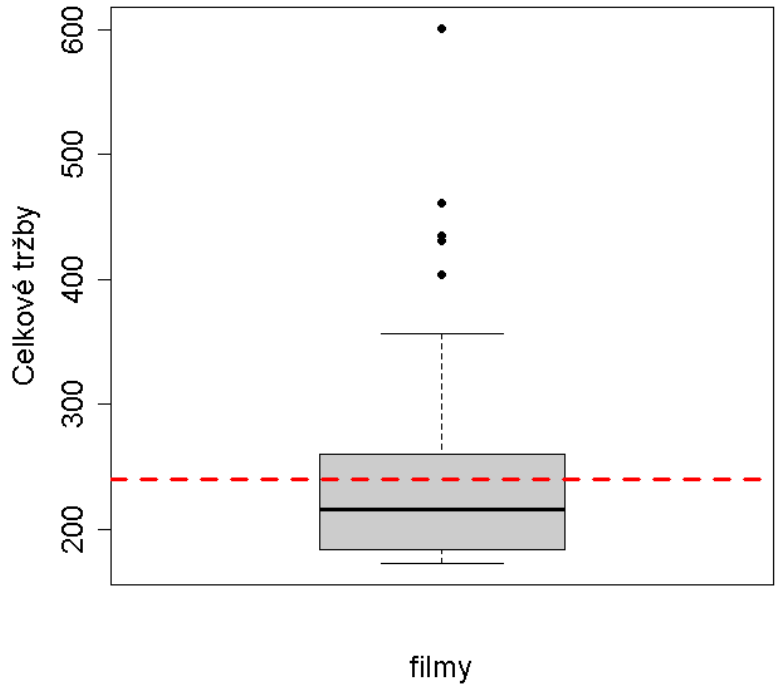
Obrázok 3.6: Všeobecná štruktúra box – plotu

Zdroj: vlastné spracovanie

V tzv. krabici (v Slovenskej literatúre sa môžeme stretnúť s ekvivalentným názvom „Krabicový graf“) sa nachádza 50 % hodnôt. Dĺžka úsečiek vychádzajúcich z krabice má maximálnu dĺžku $1.5R_Q$, prípadne menej. Všetky hodnoty mimo tohto rozsahu, či už maximálne alebo minimálne, sú považované za extrémne hodnoty a označujú sa v grafe ako

bodky (prípadne nejakým iným znakom, napr. krížikom). Box – ploty patria k veľmi užitočným nástrojom používaným napr. v technickej praxi pri stratifikácií variability (Tkáč, 2001). Na nasledujúcom obrázku (Obrázok 3.7), je príklad box – plotu vytvoreného v programe R. Použili sme pritom databázu `alltime.movies` z programového balíka `UsingR`, ktorá zahrňuje celkové tržby z vybraných filmov. Z obrázku je zrejmé, že niektoré filmy sú schopné dosiahnuť extrémne veľké tržby. Z box – plotu sa taktiež zdá, že ide o pravostranné zošikmenie (v ekonómii ide vôbec o veľmi často sa vyskytujúce zošikmenie). Do box – plotu sme taktiež zobrazili horizontálnu čiaru, ktorá predstavuje priemernú hodnotu tržieb. Priemerná hodnota je zjavne väčšia ako medián, čo potvrdzuje našu tézu o pravostrannom zošikmení. Bodky nad úsečkou vychádzajúcou z krabice predstavujú filmové trháky (z pohľadu analýzy – extrémne hodnoty).

```
> library(UsingR); mean(alltime.movies$Gross)
[1] 240.1899
> par(mar = c(5, 5, 2, 2))
> boxplot(alltime.movies$Gross, ylab = "Celkové tržby", xlab =
  "filmy", col = gray(0.8), pch = 19, cex.axis = 1.5, cex.lab =
  1.5)
> abline(h = mean(alltime.movies$Gross), lwd = 3, lty = 2, col =
  "red")
```



Obrázok 3.7: Box – plot z tržieb filmov v mil. USD

Zdroj: vlastné spracovanie v programe R

Príklad 3.19

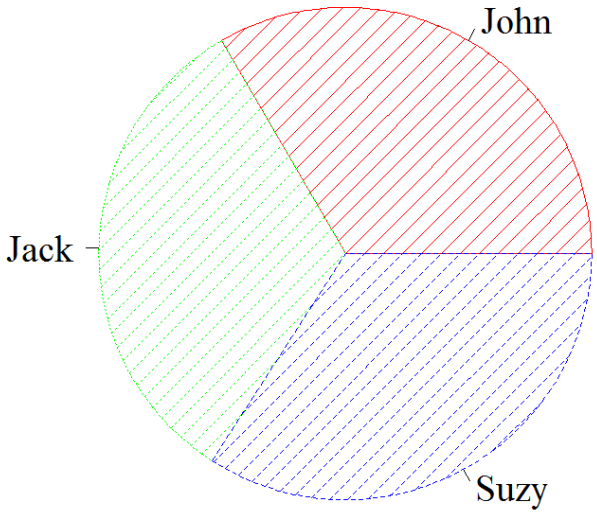
Spustite nasledujúcu sekvenciu príkazov. Akú informáciu nimi získate?

```
> f <- fivenum(alltime.movies$Gross)
> the.names <- rownames(alltime.movies)
> the.names[alltime.movies$Gross>(f[4]+1.5*(f[4]-f[2]))]
[1] "Titanic                                  "
[2] "Star Wars                                "
[3] "E.T.                                     "
[4] "Star Wars: The Phantom Menace            "
[5] "Spider-Man                               "
```

3.4.6 Koláčový a bodový graf

Pomerne obľúbeným typom grafu je koláčový graf, na tvorbu ktorého v programe R slúži funkcia `pie()`. Nasledujúci príkaz vytvorí jednoduchý koláčový graf tak, aby jednotlivé výseky boli označené.

```
> sales <- c(45, 44, 46)
> names(sales) <- c("John", "Jack", "Suzy")
> pie(sales, main = "Predaj", col = rainbow(3), density = 10,
  lty = c(1, 3, 14), edges = 400)
```

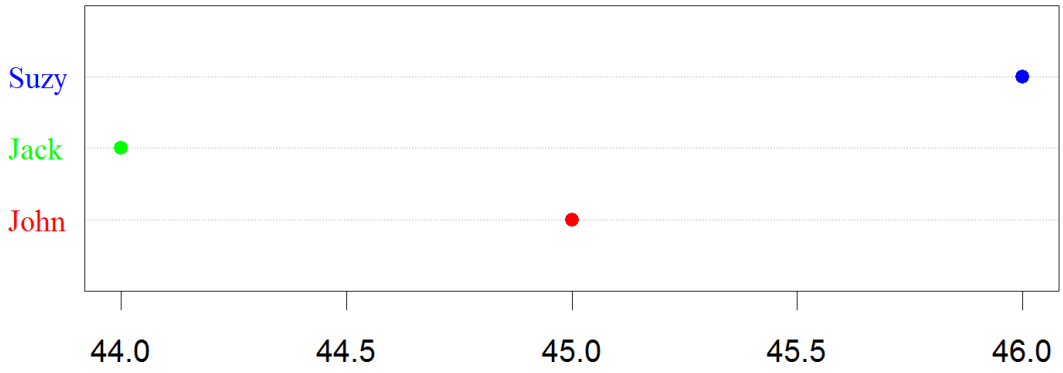


Obrázok 3.8: Koláčový graf predaja troch predajcov

Zdroj: vlastné spracovanie v programe R

Určitou alternatívou ku koláčovému grafu je bodový graf, ktorý získame prostredníctvom funkcie `dotchart()`.

```
> dotchart(sales, col = rainbow(3), lty = c(1, 3, 14), family =
  "serif", cex = 2, pch = 19)
```



Obrázok 3.9: Bodový graf predaja troch predajcov

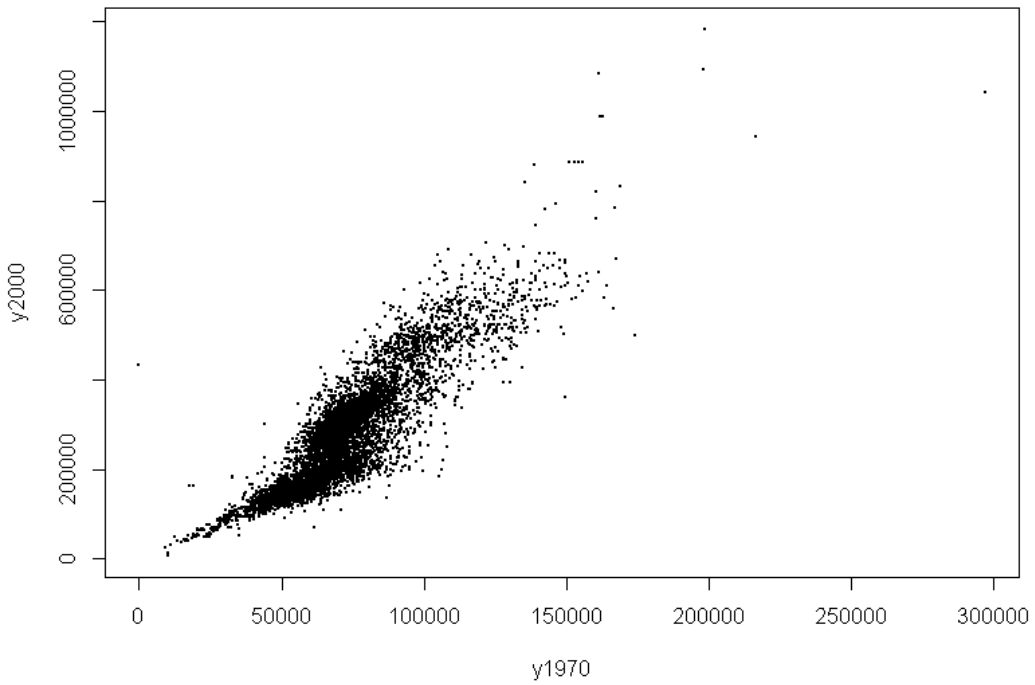
Zdroj: vlastné spracovanie v programe R

3.4.7 x-y graf

V anglickej literatúre sa môžeme stretnúť s ekvivalentným označením „*scatter – plot*“. Ide o užitočný a veľmi rozšírený graf v prípade, ak si chceme vizuálne overiť existenciu možných závislostí medzi dvoma premennými (merané na aspoň poradovej škále). Nutnou podmienkou je, aby namerané hodnoty v dvoch štatistických súboroch predstavovali navzájom usporiadané dvojice. Usporiadanú *n*-ticu si môžeme veľmi jednoducho predstaviť nasledovným spôsobom. Ak máme rôznych zákazníkov, pričom u každého z nich sledujeme jeho výšku a váhu, potom tie tvoria usporiadanú dvojicu, lebo sa týkajú jedného zákazníka, jednej štatistickej jednotky. Všeobecnejšie sa na to môžeme pozerať nasledovne: máme jednu štatistickú jednotku, u ktorej sledujeme dva znaky. Tieto dva znaky potom tvoria usporiadanú dvojicu. Samozrejme, je potrebné tieto merania zaznamenať stále v rovnakom poradí (napr. najprv výšku a potom váhu). Na usporiadaní hodnôt záleží. Do tvorby *x-y* grafu môžeme použiť iba úplné údaje. To znamená iba tých zákazníkov, kde máme zaznačenú aj ich výšku aj váhu. Typ analýzy, ktorý by sme potom pomocou *x-y* grafu mohli uskutočniť, je hľadanie závislostí.

Názorný príklad je v databáze `homedata` v programovom balíku `UsingR`. V databáze sú dve premenné. Ceny nehnuteľností v roku 1970 a ceny tých istých nehnuteľností v roku 2000 (keďže ide o tie isté nehnuteľnosti ide o usporiadanú dvojicu pozorovaní). *x-y* graf zostrojíme pomocou funkcie `plot()`.

```
> attach(homedata)
> names(homedata)
[1] "y1970" "y2000"
> plot(y1970, y2000, pch = 19, cex = 0.2)
```



Obrázok 3.10: *x-y* graf

Zdroj: vlastné spracovanie v programe R

Z obrázku je zrejmé, že v roku 2000 sú ceny nehnuteľností vyššie, a že medzi cenami v roku 1970 a 2000 zrejme existuje určitý lineárny vzťah. To znamená, že ceny nehnuteľností v roku 2000 sú určitým násobkom väčšie ako ceny v roku 1970. Regresnou analýzou by bolo možné túto hypotézu bližšie overiť.

Na ukážku si môžeme pomôcť špecializovaným programovým balíkom `ggplot2`. Necháme na čitateľovi aby si vyskúšal nasledujúci príkaz, ktorý zostrojí podobný obrázok ako predošlý (Obrázok 3.10). Upozorňujeme však, že vo verzii programu R 2.14.0 tento príkaz nefunguje. Je potrebné nainštalovať novšiu alebo staršiu verziu programu.

```
> library(ggplot2)
> qplot(y1970, y2000, data = homedata, size = I(2))
```

Na výpočet sily lineárneho vzťahu medzi dvoma premennými sa používa korelačný koeficient. Hodnoty blízko 1 znamenajú silnú priamu lineárnu závislosť a hodnoty blízko –1 silnú nepriamu lineárnu závislosť. Ak by bola hodnota korelačného koeficientu 0, medzi premennými neexistuje lineárny vzťah. Na výpočet korelačného koeficientu môžeme použiť funkciu `cor()`. V tejto publikácii venujúcej sa opisu dát, sa korelačnej ani regresnej analýze bližšie venovať nebudeme. Len pre úplnosť, v predchádzajúcom príklade by sme za účelom výpočtu korelačného koeficientu (Pearsonovho) mohli použiť príkaz `cor(y1970, y2000)`.

3.4.8 Ďalšie formy vizualizácie dát v programe R

Zostrojme histogram a box – plot z premennej `OBP` (databáza `OBP` je z programového balíka `UsingR`). Následne vytvorme nový vektor, v ktorom bude chýbať 10 % najmenších a 10 % najväčších hodnôt. Z nového vektora zostrojíme nový histogram aj box – plot a porovnáme ich. Bude nás zaujímať, aký bol vizuálny efekt (zmena dvoch obrázkov) po odstránení hodnôt pri histograme prípadne box – plote. Použijeme pritom funkciu `if()` a cyklus `for()`.

```
> library(UsingR)
> data(OBP)
> hist(OBP); boxplot(OBP)
> newOBP <- c()
> for (i in OBP) {
+ if (i > quantile(OBP, 0.1) & i < quantile(OBP, 0.9)) {
+ newOBP <- c(newOBP, i)
+ }
+ }
> hist(newOBP); boxplot(newOBP)
```

Riadok `newOBP <- c()` nám vytvorí nový dátový vektor, do ktorého si postupne budeme ukladať čísla, ktoré budú spĺňať nami požadovanú podmienku, t.j. aby išlo o hodnoty z `OBP` menšie ako 90-ty percentil a zároveň väčšie ako 10-ty percentil. Túto podmienku si v jazyku R môžeme formálne nadefinovať ako `i > quantile(OBP, 0.1) & i < quantile(OBP, 0.9)`, kde `i` bude vybrané pozorovanie z vektora `OBP`. Cyklus `for()` sa používa, ak máme záujem určitú operáciu pravidelne opakovať. Môžeme ho vo výraznej miere používať pri simuláciách. Príkaz funguje spravidla tak, že za `for` musí nasledovať zátvorka `()`, v ktorej sa špecifikuje index (premenná), ktorá sa bude iterovať. V tomto prípade sa zvolil index „`i`“ a časť kódu „`in OBP`“ znamená, že sa zoberú prvky vektora `OBP`. Prvý riadok (ak bude príkaz na viac ako jeden riadok) musí končiť zloženou zátvorkou `{`. Nasledujúci riadok je funkcia `if()`, ktorá špecifikuje podmienku a hovorí o tom, aké príkazy sa majú uskutočniť, ak bude podmienka v zátvorke splnená. Znova sa riadok končí hranatou zátvorkou `{`. Následne príkaz `newOBP <- c(newOBP, i)` vytvára nový dátový vektor, v ktorom sa k predošlému dátovému vektoru `newOBP` pripíše hodnota `i`. Na záver sa príkaz `if()` vnorený do cyklu `for()` končí zátvorkou `}`, ktorou sa ukončí aj cyklus `for()`.

Keďže ide o prvé použitie cyklu `for()`, rozpíšeme si celú iteráciu. V prvom kroku sa za index `i` zoberie prvý prvok vektora `OBP`. Následne sa tento prvok preverí v podmienke vo funkcii `if()`. Ak spĺňa podmienku, tak sa k vektoru `newOBP` pripíše tento prvok. Ak

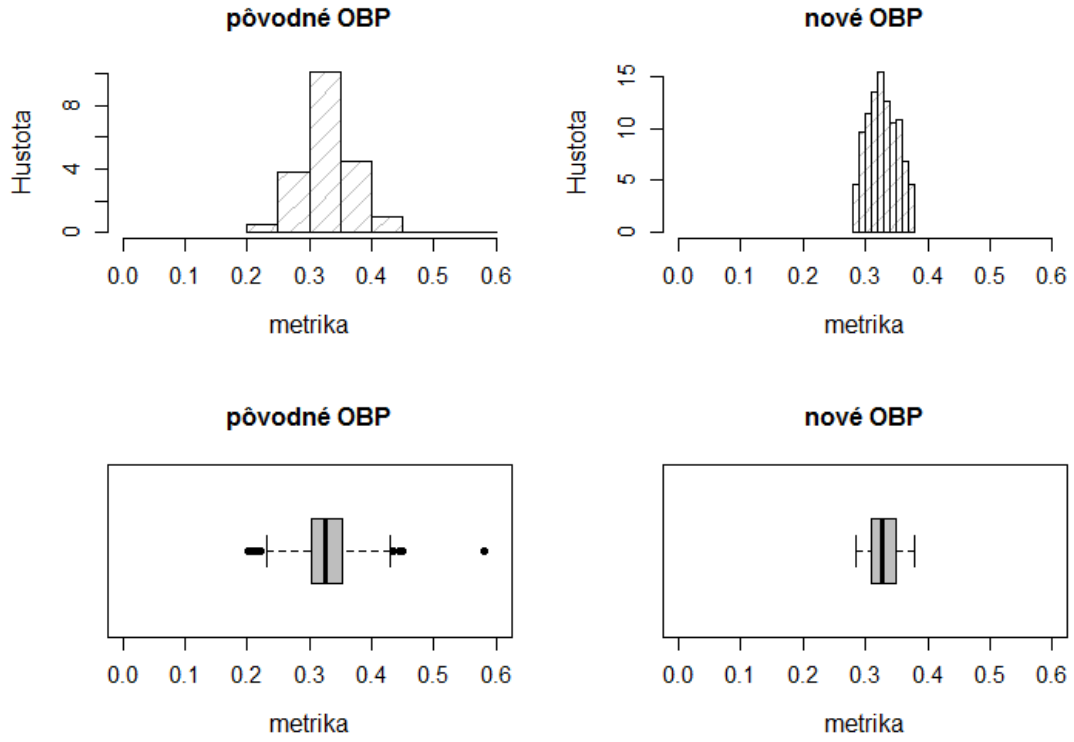
podmienka nie je splnená, potom sa príkaz `newOBP <- c(newOBP, i)` nerealizuje a v ďalšom kroku sa za index `i` dosadí druhá hodnota z vektora `OBP` a postup sa opakuje pokiaľ sa neprejde všetkými prvkami vektora `OBP`. Týmto spôsobom sa každý prvok vektora `OBP` skontroluje voči príslušnej podmienke vo funkcii `if()` a prípadne dopíše do nového vektora, kde budú iba také hodnoty, ktoré našu podmienku spĺňajú.

Pre ilustráciu si ukážeme ekvivalentný zápis.

```
> newOBP <- c()
> for (i in 1:length(OBP)) {
+ if (OBP[i] > quantile(OBP, 0.1) & OBP[i] < quantile(OBP, 0.9))
  {
+ newOBP <- c(newOBP, OBP[i])
+   }
+ }
```

Následne si môžeme porovnať histogram a box – plot (Obrázok 3.11). Pri porovnávaní dvoch histogramov je užitočný nasledujúci postup.

```
> par(mfrow = c(2, 2))
> hist(OBP, main = "pôvodné OBP", density = 10, col = "grey",
  border = "black", cex.lab = 1.2, cex.axis = 1.1, freq = FALSE,
  ylab = "Hustota", xlim = c(0, 0.6), xlab = "metrika")
> hist(newOBP, main = "nové OBP", density = 10, col = "grey",
  border = "black", cex.lab = 1.2, cex.axis = 1.1, freq = FALSE,
  ylab = "Hustota", xlim = c(0, 0.6), xlab = "metrika")
> boxplot(OBP, main = "pôvodné OBP", density = 10, col = "grey",
  border = "black", cex.lab = 1.2, cex.axis = 1.1, pch = 19,
  horizontal = T, ylim = c(0, 0.6), xlab = "metrika")
> abline(h = mean(OBP),lwd = 3,lty = 2,col = "red")
> boxplot(newOBP, main = "nové OBP", density = 10, col = "grey",
  border = "black", cex.lab = 1.2, cex.axis = 1.1, pch = 19,
  horizontal = T, ylim = c(0, 0.6), xlab = "metrika")
> abline(h = mean(newOBP),lwd = 3,lty = 2,col = "red")
```



Obrázok 3.11: Histogram a box – plot rôznych vzoriek

Zdroj: vlastné spracovanie v programe R

Ak máme dve premenné a zaujíma nás ich vzájomný vzťah, na jeho zobrazenie si nevystačíme s jedným histogramom alebo box – plotom. Vhodnou alternatívou v takom prípade sú *x-y* grafy. Ak sú obe premenné kategorické, používame na prezentáciu tzv. kontingenčné tabuľky. Ako príklad si ukážeme situáciu, kde prezentujeme lojalitu voličov k jednej politickej strane. Tabuľku si môžeme nadefinovať rôznymi spôsobmi, napr. spájaním riadkov tabuľky alebo spájaním stĺpcov tabuľky. Vytvoríme si nasledujúcu tabuľku:

```
> lojalita <- rbind(c(80, 45), c(10, 40))
> rownames(lojalita) <- c("predtým:ANO", "predtým:NIE")
> colnames(lojalita) <- c("potom:ANO", "potom:NIE")
> lojalita
            potom:ANO potom:NIE
predtým:ANO        80        45
predtým:NIE        10        40
```

Výsledná tabuľka ukazuje, že počet voličov, ktorý v predošlých voľbách danú politickú stranu volili, a zároveň aj v druhých voľbách volili tú istú stranu, bol 80. Ak v druhých voľbách nevolili tú istú stranu, počet bol iba 45. V druhom prípade ide o nelojálnych voličov, resp. o voličov, ktorý zmenili svoj názor. Na druhej strane, ak v predošlých voľbách danú politickú stranu nevolilo 50 z opýtaných voličov (súčet druhého riadku, tzv. marginálna početnosť), potom v nasledujúcich voľbách zmenilo svoj názor

a zároveň volilo nami analyzovanú politickú stranu 10 voličov. Zvyšných 40 voličov stále nevolilo danú politickú stranu. Tieto výsledky naznačujú určitý odliv voličov, čo vidno aj z marginálnych početností. Iný spôsob zadania tej iste tabuľky cez stĺpce je:

```
> lojalita <- cbind(c(80, 10), c(45, 40))
> rownames(lojalita) <- c("predtým:ANO", "predtým:NIE")
> colnames(lojalita) <- c("potom:ANO", "potom:NIE")
> lojalita
            potom:ANO potom:NIE
predtým:ANO        80        45
predtým:NIE        10        40
```

Tretí spôsob, ktorý si ukážeme, využíva definovanie objektu matice:

```
> lojalita <- matrix(c(80, 10, 45, 40), nrow = 2)
> varnames <- c("ANO", "NIE")
> dimnames(lojalita) <- list(predtým = varnames, potom =
  varnames)
> lojalita
       potom
predtým ANO NIE
    ANO  80  45
    NIE  10  40
```

Funkcia `dimnames()` pomenuje dimenzie matice, riadky a stĺpce. Ak by sa jednotlivé názvy stĺpcov a riadkov od seba líšili, môžeme si jednotlivé riadky a stĺpce definovať ako `rownames()` pre riadky a `colnames()` pre stĺpce.

Použime nasledujúce údaje a vytvorme tabuľku:

```
> library(UsingR)
> attach(grades)
> ?grades
> names(grades)
> znamky <- table(prev, grade)
```

Vytvorením tabuľky si vieme utvoriť názor, či minulá úspešnosť študentov na testoch z matematiky (podľa známok), je dobrým indikátorom budúcej úspešnosti z matematiky. Súčty riadkov a stĺpcov v kontingenčnej tabuľke predstavujú marginálne početnosti. Tie je možné do tabuľky pridať pomocou funkcie `addmargins()`, v našom prípade `addmargins(znamky)`.

```
grade
prev      A        A-    B+    B     B-    C+    C     D     F    Sum
   A     15         3     1     4     0     0     3     2     0     28
   A-     3         1     1     0     0     0     0     0     0      5
   B+     0         2     2     1     2     0     0     1     1      9
   B      0         1     1     4     3     1     3     0     2     15
   B-     0         1     0     2     0     0     1     0     0      4
   C+     1         1     0     0     0     0     1     0     0      3
   C      1         0     0     1     1     3     5     9     7     27
   D      0         0     0     1     0     0     4     3     1      9
   F      1         0     0     1     1     1     3     4    11     22
  Sum    21         9     5    14     7     5    20    19    22    122
```

Všimnime si jednu zaujímavú vlastnosť, ktorú pri kontingenčných tabuľkách môžeme sledovať. Ak sa pozrieme iba na prvý stĺpec, vidíme početnosti známok študentov v predchádzajúcom teste z matematiky. V riadku môžeme vidieť početnosti tých študentov, ktorí dostali v druhom teste z matematiky známku A. Následne sa pozrime iba na druhý stĺpec a postupne na ďalšie. Môžeme tak pozorovať, **ako sa mení rozdelenie početností** v závislosti od toho, akú známku dostali študenti v druhom teste. Podobne môžeme postupovať aj pri riadkoch. Ide o určitý indikátor závislosti medzi dvoma kategorickými premennými. Niektoré koeficienty merajúce formu závislosti medzi kategorickými premennými využívajú práve túto vlastnosť.

Podobne ako pri dátových vektoroch, aj pri tabuľkách si vieme označiť jednotlivé prvky tabuľky. Vytvorme si tabuľku s marginálnymi početnosťami `newznamky <- addmargins(znamky)` a teraz pomocou `[]` môžeme označovať jednotlivé prvky tabuľky. Vyskúšajte: `newznamky[1]`, `newznamky[2]`, `newznamky[3]`, `newznamky[10]`, `newznamky[91]`,`newznamky[100]`.

Príklad 3.20

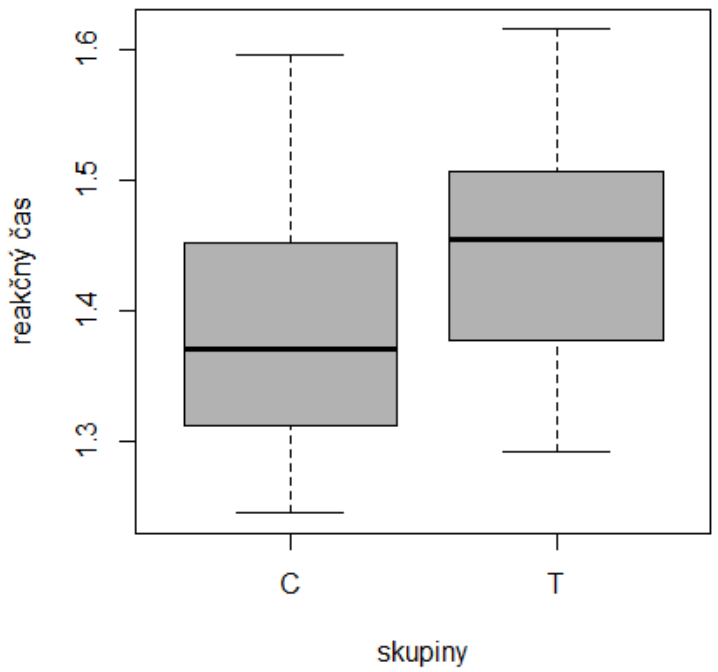
V databáze `UScereal`, `library(MASS)` sú k dispozícii informácie o cereálnych produktoch vo vybraných potravinách v USA. Vytvorte tabuľku, ktorá ukáže vzťah medzi výrobcom a umiestnením produktu v regáloch predajne. Pozrite si najprv databázu a opis premenných. Existuje vzťah medzi výrobcom a umiestnením na regáloch? Skúste vytvoriť tabuľku, v ktorej by boli percentá a nie početnosti. Tabuľka by mala zobrazovať, v koľkých percentách prípadov pre jednotlivé produkty boli výrobky umiestnené v prvom, druhom a treťom rade.

```
> library(MASS)
> attach(UScereal)
> cereal <- table(mfr, shelf)
```

```
> cereal_new <- cereal/rowSums(cereal)
> cereal <- table(mfr, shelf)
> cereal
   shelf
mfr  1  2  3
  G  6  7  9
  K  4  7 10
  N  2  0  1
  P  2  1  6
  Q  0  3  2
  R  4  0  1
> cereal_new <- cereal/rowSums(cereal)
> cereal_new
   shelf
mfr         1         2         3
  G 0.2727273 0.3181818 0.4090909
  K 0.1904762 0.3333333 0.4761905
  N 0.6666667 0.0000000 0.3333333
  P 0.2222222 0.1111111 0.6666667
  Q 0.0000000 0.6000000 0.4000000
  R 0.8000000 0.0000000 0.2000000
```

Porovnávanie dvoch vzoriek sme už uskutočnili, keď sme porovnávali súbor `OBP` s novým súborom `OBP` bez 10 % najmenších a 10 % najväčších hodnôt. V obdobnej situácii by sme sa ocitli, ak by sme mali údaje o spokojnosti zákazníkov s určitým produktom a zaujímalo by nás, či miera tejto spokojnosti je iná pre mužov ako pre ženy, prípadne pre ľudí mladších ako 18 rokov a starších ako 18 rokov a podobne. V programovom balíku `UsingR` je databáza `reaction.time`, v ktorom máme štyri premenné: vek respondenta, jeho pohlavie, či používal mobilný telefón pri jazde autom a jeho reakčný čas. Ak nás zaujíma, či je rozdiel medzi reakčným časom respondentov, ktorí používali mobilný telefón (označíme ako T) a respondentmi, ktorí mobilný telefón nepoužívali (označíme ako C), môžeme použiť nasledovný príkaz:

```
> attach(reaction.time)
> boxplot(reaction.time$time~control, col = grey(0.7), density =
  10, xlab = c("skupiny"), ylab = "reakčný čas")
```



Obrázok 3.12: Box – plot reakčného času vodičov

Zdroj: vlastné spracovanie v programe R

Príklad 3.21

Porovnajte histogram a box – plot reakčného času a porovnajte základné deskriptívne štatistiky medzi dvoma súbormi. Je rozdiel v reakčných časoch dvoch skupín výrazný? Váš záver zdôvodnite.

Jeden zo spôsobov ako rozdeliť súbor podľa jednej premennej je napísať krátky skript:

```
> mobile <- c(); cont <- c()
> for (i in 1:length(reaction.time$control)) {
+ if (reaction.time$control[i] == "T") {
+ mobile <- c(mobile, reaction.time$time[i])
+ }
+ else {
+ cont <- c(cont, reaction.time$time[i])
+ }
+ }
```

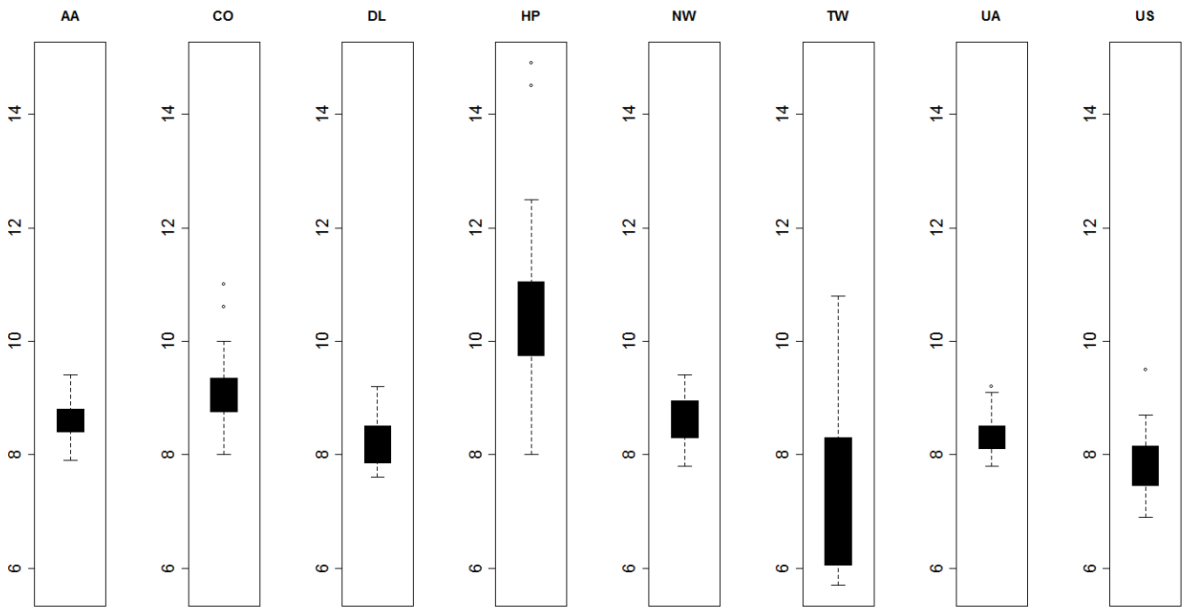
Výrazne jednoduchšou možnosťou však v tomto prípade je použiť logické operátory a indexovanie dátových vektorov:

```
> cont <- reaction.time$time[reaction.time$control == "C"]
> mobile <- reaction.time$time[reaction.time$control == "T"]
```

Už sme si ukazovali ako v programe R zobraziť jeden alebo viac box – plotov. V tejto časti si túto problematiku ďalej rozvinieme a ukážeme si aj zobrazovanie *x-y* grafov pri

viacrozmerných štatistických súboroch. Na začiatku budeme pracovať s databázou `ewr` (`UsingR`). Databáza zobrazuje čas potrebný na prípravu odletu (`taxi-out`) od opustenia brány po samotný vzlet, ako aj čas pristávania (`taxi-in`), čo predstavuje čas od kontaktu s pristávacou dráhou po otvorenie brány lietadla. Premenná `inorout` triedi hodnoty v riadkoch na `taxi-in` a `taxi-out` čas. Čím je tento čas kratší, o to efektívnejšie (z hľadiska letiska) je využitý čas lietadla na letisku. Zobrazíme box – plot `taxi-in` časov pre jednotlivé letecké spoločnosti (Obrázok 3.13). Pri takto zostavenej databáze je jednou z možností nasledujúci (pomerne prácny) postup:

```
> new <- ewr[inorout == "in",]
> new <- new[,3:10]
> par(mfcol = c(1,8))
> minimum <- min(sapply(new, min))
> maximum <- max(sapply(new, max))
> boxplot(new$AA, col = "black", main = "AA", ylim = c(minimum,
  maximum))
> boxplot(new$CO, col = "black", main = "CO", ylim = c(minimum,
  maximum))
> boxplot(new$DL, col = "black", main = "DL", ylim = c(minimum,
  maximum))
> boxplot(new$HP, col = "black", main = "HP", ylim = c(minimum,
  maximum))
> boxplot(new$NW, col = "black", main = "NW", ylim = c(minimum,
  maximum))
> boxplot(new$TW, col = "black", main = "TW", ylim = c(minimum,
  maximum))
> boxplot(new$UA, col = "black", main = "UA", ylim = c(minimum,
  maximum))
> boxplot(new$US, col = "black", main = "US", ylim = c(minimum,
  maximum))
```



Obrázok 3.13: Box – ploty taxi-in času leteckých spoločností
Zdroj: vlastné spracovanie v programe R

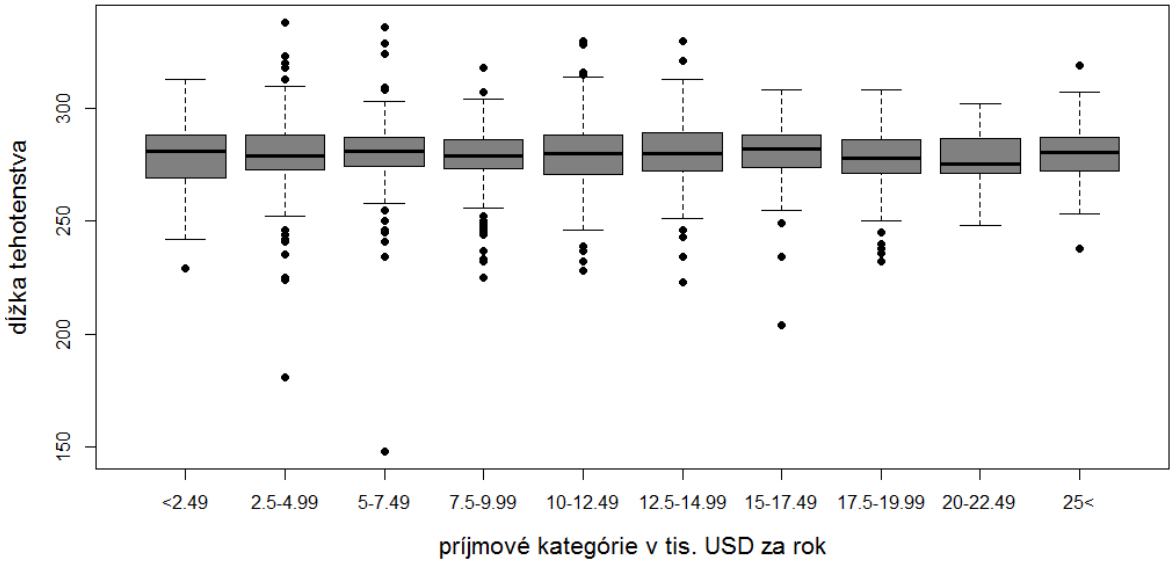
Z takto zostavených box – plotov môžeme vidieť určité rozdiely v časoch, ktoré potrebujú na pristátie rôzne letecké spoločnosti. Podobne môžeme postupovať aj pri taxi-out časoch.

V ďalšom príklade použijeme databázu `babies` (`UsingR`). Najprv si databázu upravíme a zobrazíme box – plot, kde nás zaujíma, či existuje vzťah medzi dobou tehotenstva žien a výšky príjmu domácnosti. Samozrejme, dĺžka tehotenstva žien je daná fyziologicky, avšak určitá odchýlka od 40 týždňov existuje. Za normálne sa spravidla považuje trvanie tehotenstva v intervale od 37 do 42 týždňov. Z toho teda vyplýva, že existuje určitá variabilita v celkovej dĺžke. Je zaujímavé sledovať, či túto variabilitu je možné vysvetliť pomocou životosprávy žien, sociálneho, spoločenského alebo ekonomického prostredia atď.. Prvým krokom k takejto analýze môže byť zobrazenie box-plotov pre rôzne úrovne týchto prostredí. V tomto príklade nás zaujíma výška príjmu domácnosti.

Začneme úpravou databázy, keďže v niektorých údajoch sú čísla ako 999, prípadne 98, ktoré znamenajú, že požadovaný údaje pre daného respondenta nebol nameraný. Tieto hodnoty musíme z údajovej základe vylúčiť, keďže ide len o kódy, nie o skutočné hodnoty premenných.

```
> attach(babies)
> new <- subset(babies, subset = gestation != 999 & inc !=
  98,varwidth = TRUE)
> detach(babies)
> attach(new)
```

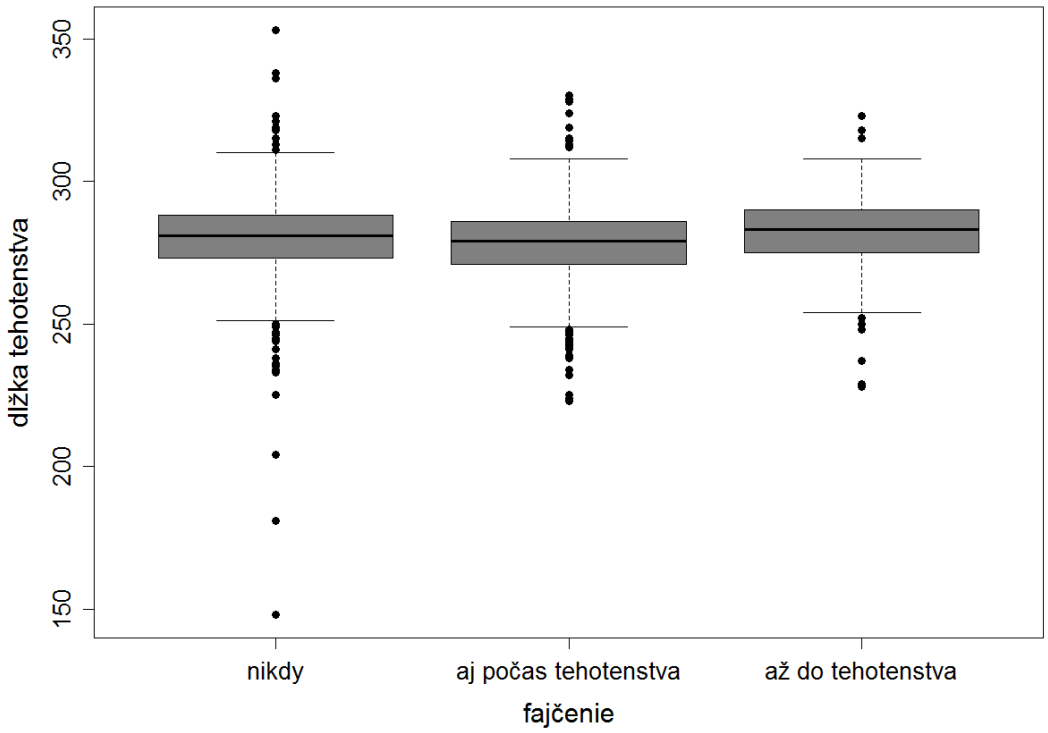
```
> boxplot(gestation~inc, pch = 19, names = c("<2.49", "2.5-
  4.99", "5-7.49", "7.5-9.99", "10-12.49", "12.5-14.99", "15-
  17.49", "17.5-19.99", "20-22.49", "25<"), xlab = "príjmové
  kategórie v tis. USD za rok", ylab = "dĺžka tehotenstva", col
  = grey(0.5), cex.lab = 1.3, cex.axis = 1)
```



Obrázok 3.14: Box – ploty dĺžky tehotenstva (počet dní) v závislosti od príjmu
Zdroj: vlastné spracovanie v programe R

Z predchádzajúceho obrázku nie je vidno výrazný posun v mediánoch. Pre jednotlivé box - ploty sa mení príjmová kategória, ale posun v mediánoch sa javí ako minimálny. Na druhej strane, určité rozdiely badať vo variabilite. Všimnite si, že výskyt extrémnych hodnôt je zrejme častejší pri nižších príjmových kategóriách a to oboma smermi. Skúsme sa pozrieť na ďalší vzťah, ktorý sa javí ako reálnejší. Budeme porovnávať dĺžku tehotenstva žien v závislosti od toho, či ide o fajčiara. Použijeme na to ďalšiu sériu box – plotov.

```
> attach(babies)
> new <- subset(babies, subset = gestation != 999 & smoke != 9 &
  smoke !=3)
> detach(babies)
> boxplot(new$gestation~new$smoke, pch = 19, names = c("nikdy",
  "aj počas tehotenstva", "až do tehotenstva"), xlab =
  "fajčenie", ylab = "dĺžka tehotenstva", family = "serif", col
  = grey(0.5), cex.lab = 1.5, cex.axis = 1.3)
```



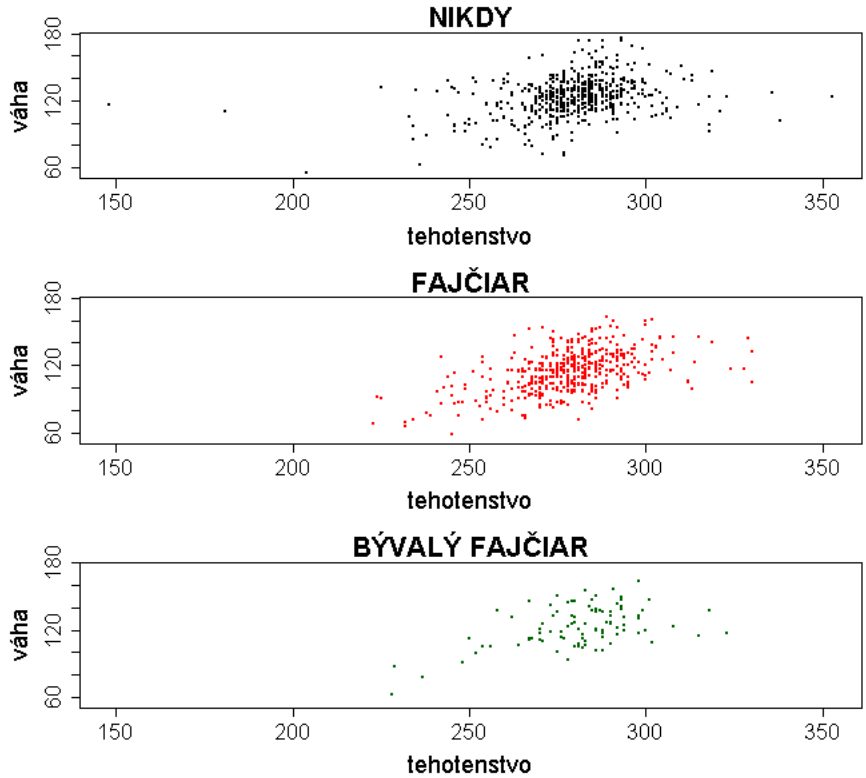
Obrázok 3.15: Box – ploty dĺžky tehotenstva v závislosti od fajčenia

Zdroj: vlastné spracovanie v programe R

Zaujímavé je, že znova nevidno väčšie posuny v mediánoch, ale skôr rozdiely vo variabilite, pričom práve u žien, ktoré nepatrili k fajčiarom je variabilita dĺžky tehotenstva zrejme najvyššia.

Ďalej si príklad rozvinieme o ďalšiu premennú a tou je váha narodeného dieťaťa (meraná v unciach). Zaujímať nás bude vzťah medzi váhou a dĺžkou tehotenstva, teda zostrojíme *x-y* grafy pre rôzne skupiny fajčiarov.

```
> attach(babies)
> new <- subset(babies, subset = gestation != 999 & smoke != 9 &
  smoke != 3)
> detach(babies); attach(new)
> ming <- min(gestation); maxg <- max(gestation);
> minwt <- min(wt); maxwt <- max(wt);
> par(mfrow = c(3, 1))
> plot(gestation[smoke==0], wt[smoke==0], pch = 19, col =
  "black", xlab = "tehotenstvo", ylab = "váha", cex = 0.5,
  cex.lab = 1.7, cex.axis = 1.5, xlim = c(ming, maxg), ylim =
  c(minwt, maxwt), main = "NIKDY", cex.main = 2)
> plot(gestation[smoke==1], wt[smoke == 1], pch = 19, col =
  "red", xlab = "tehotenstvo", ylab = "váha", cex = 0.5, cex.lab
  = 1.7, cex.axis = 1.5, xlim = c(ming, maxg), ylim = c(minwt,
  maxwt), main = "FAJČIAR", cex.main = 2)
> plot(gestation[smoke==2], wt[smoke == 2], pch = 19, col =
  "darkgreen", xlab = "tehotenstvo", ylab = "váha", cex = 0.5,
  cex.lab = 1.7, cex.axis = 1.5, xlim = c(ming, maxg), ylim =
  c(minwt, maxwt), main = "BÝVALÝ FAJČIAR", cex.main = 2)
```



Obrázok 3.16: *x-y* graf dĺžky tehotenstva a váhy narodeného dieťaťa v závislosti od fajčenia

Zdroj: vlastné spracovanie v programe R

V predchádzajúcom obrázku sú všetky osi nastavené na rovnaké veľkosti škál čo uľahčuje porovnanie. Až na extrémne hodnoty však zásadné rozdiely medzi jednotlivými skupinami nebadať. Zrejme by bola potrebná podrobnejšia analýza v podobe regresnej analýzy.

Na záver ukážka z niektorých pokročilejších grafov pomocou programového balíka `lattice`.

```
> library(lattice)
> attach(babies)
> options(lattice.theme = "col.whitebg")
> histogram(~wt|factor(smoke), data = babies, subset = wt != 999
  & smoke != 9)
> densityplot(~wt|factor(smoke), data = babies, subset = wt !=
  999 & smoke != 9)
> bwplot(gestation~factor(inc), data = babies, subset =
  gestation != 999 & inc != 98)
> xyplot(wt~gestation|factor(smoke), data = babies, subset = wt
  != 999 & gestation != 999)
```

3.5 Úvod do práce s databázami v programe R

Pod výrazom databázové objekty v zásade budeme rozumieť tzv. `data frames`. Práca s dátami je v skutočnosti podstatne viac časovo náročná ako samotná analýza (do toho

nepočítame naštudovanie metodológie). Doteraz sme pracovali s databázovými objektmi v programe R, ktoré boli k dispozícií v rôznych programových balíkoch. Pri ukážke, akým spôsobom importovať údaje do programu R, sme sa problematiky databázových objektov už raz dotkli. Na tomto mieste si ukážeme, akým spôsobom tieto objekty vytvárať. Začneme troma objektmi: `x <- 1:10; y <- letters[rep(1:2, 5)]; z <- 1:15`

Vyskúšame tieto tri objekty spojiť do jednej databázy:

```
> data.frame(x,y,z)
Error in data.frame(x, y, z) :
  arguments imply differing number of rows: 10, 15
```

Dostávame chybové hlásenie. Objekty v databáze by mali mať rovnaký rozmer, v tomto prípade rovnakú dĺžku.

```
> data.frame(x,y)
    x y
1   1 a
2   2 b
3   3 a
4   4 b
5   5 a
6   6 b
7   7 a
8   8 b
9   9 a
10 10 b
```

Uvedené pravidlo sa netýka zoznamov (objektov s názvom `list`).

```
> list(x,y,z)
[[1]]
 [1]  1  2  3  4  5  6  7  8  9 10

[[2]]
 [1] "a" "b" "a" "b" "a" "b" "a" "b" "a" "b"

[[3]]
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

Rozdiel medzi `data.frame()` a `list()` spočíva v tom, že kým v `list()` môžeme mať rôzne objekty (vrátane `data.frame`) v `data.frame` by sme mali mať dátové vektory (hodnoty v dátových vektoroch nemusia byť nutne reálne čísla). Na druhej strane, na dátových objektoch môžeme vykonávať niektoré operácie, ktoré na zoznamoch nie. Jednotlivé prvky týchto databázových objektov pomenovať.

```
> a <- list("x je meno" = x,"y je meno" = y)
> a
```

```
$`x je meno`
[1] 1 3 5 6 8 2 4 6 9

$`y je meno`
[1] 2 4 6

> a$x
[1] 1 3 5 6 8 2 4 6 9
> a$"x je meno"
[1] 1 3 5 6 8 2 4 6 9
```

Naopak to môže byť problém, vyskúšajte:

```
> b <- list(x = "x je meno", y = "y je meno")
> b$x
[1] "x je meno"
> b$"x je meno"
NULL
```

Pri `data.frame` je postup podobný ako sme si ukázali skôr, napríklad pri tvorbe tabuliek alebo niektorých grafov (`dotchart, piechart, barplot`).

```
> c <- data.frame(x,y)
> names(c)
[1] "x" "y"
> names(c) <- c("x je meno", "y je meno")
> names(c)
[1] "x je meno" "y je meno"
```

Pre istotu si zopakujeme, že rozmer databázového objektu vieme zistiť pomocou funkcie `dim()` podobne ako pri maticiach, prípadne ak bude stačiť, môžeme použiť funkciu `length()`.

```
dim(a); dim(b); dim(c); length(a); length(b); length(c)
```

Na viacerých miestach sme sa už stretli s odkazom na premenné v databázových objektoch. Pre úplnosť si ich zopakujeme. Ak sa potrebujeme odkázať na premennú určitej databázy, použijeme `$`, napr: `c$"x je meno"`. Iný spôsob je `c[1]`.[14] Takže ak chceme vypočítať napríklad aritmetický priemer prvej premennej, môžeme to uskutočniť ako: `mean(c[1])`. Ukážeme si rôzne príklady na databáze `mtcars`.

```
attach(mtcars)
> mtcars["Honda Civic",] # všimnite si čiarku v hranatej
  zátvorke!
```

[14] Existujú aj úplne odlišné typy objektov, ktoré majú rôzne časti, na ktoré sa môžeme odkázať použitím iných operátorov, napr. @.

```
             mpg cyl disp hp drat    wt  qsec vs am gear carb
Honda Civic 30.4   4 75.7 52 4.93 1.615 18.52  1  1    4    2
> mtcars[,"mpg"]
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4
  17.3 15.2 10.4
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0
  30.4 15.8 19.7
[31] 15.0 21.4
> mtcars[1]
                     mpg
Mazda RX4           21.0
Mazda RX4 Wag       21.0
Datsun 710          22.8
Hornet 4 Drive      21.4
Hornet Sportabout   18.7
Valiant             18.1
Duster 360          14.3
Merc 240D           24.4
Merc 230            22.8
Merc 280            19.2
Merc 280C           17.8
Merc 450SE          16.4
Merc 450SL          17.3
Merc 450SLC         15.2
Cadillac Fleetwood  10.4
Lincoln Continental 10.4
Chrysler Imperial   14.7
Fiat 128            32.4
Honda Civic         30.4
Toyota Corolla      33.9
Toyota Corona       21.5
Dodge Challenger    15.5
AMC Javelin         15.2
Camaro Z28          13.3
Pontiac Firebird    19.2
Fiat X1-9           27.3
Porsche 914-2       26.0
Lotus Europa        30.4
Ford Pantera L      15.8
Ferrari Dino        19.7
Maserati Bora       15.0
Volvo 142E          21.4
> mtcars$mpg
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4
  17.3 15.2 10.4
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0
  30.4 15.8 19.7
[31] 15.0 21.4
> mtcars["Honda Civic", "mpg"]
[1] 30.4
```

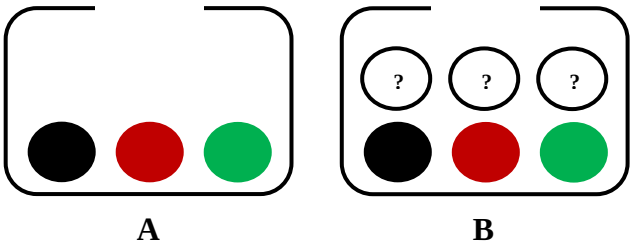
4 Úvod do pravdepodobnosti

4.1 Definovanie pravdepodobnosti

Teória pravdepodobnosti bude podobne ako predošlá a zvyšná časť textu prezentovaná intuitívne, s obmedzeným množstvom formálnych zápisov. Našim cieľom je sústrediť sa na kľúčové myšlienky, ktoré považujeme za dôležité pri empirickom skúmaní ekonomických javov. Pomerne obsiahle sa budeme venovať niektorým známym rozdeleniam, ich vizualizácii a práci s obrázkami v programe R.

Bez toho, aby sme na začiatku uviedli exaktnú definíciu pravdepodobnosti, má zrejme každý z nás nejakú predstavu o tom, čo sa skrýva za týmto pojmom. Aby sme sa dostali k pomerne jednoduchej definícii pravdepodobnosti, predstavme si, že máme jednu čiernu škatuľu s otvorom, do ktorej obsahu z vonku nevidíme. Do tejto prázdnej škatule vložíme 3 guľky rôznej farby. Povedzme zelenú, čiernu a červenú. Aká je pravdepodobnosť, že ak zatrasieme škatuľou a náhodne vyberieme jednu guľku tak jej farba bude zelená? Zrejme každý z nás pozná správnu odpoveď, pravdepodobnosť bude 1/3.

Predstavme si ďalej, že do tejto škatule vložíme ďalšie 3 guľky, pričom nevieme, ktorá má akú farbu. Jediné čo vieme je, že tieto guľky môžu mať znova zelenú, čiernu alebo červenú farbu. Znova sa opýtame rovnakú otázku. Ak zatrasieme touto škatuľou a vyberieme z nej jednu guľku, aká je pravdepodobnosť, že vybraná guľka bude zelená? Zrazu je odpoveď komplikovanejšia, keďže nepoznáme počty farieb v škatuli. Uvedenú situáciu si môžeme znázorniť na nasledujúcom obrázku.



Obrázok 4.1: A) 3 guľky; B) 6 guliek, 3 so známou farbou, 3 s neznámou farbou
Zdroj: vlastné spracovanie

Otázkou teda je, ako zistíme pravdepodobnosť, že náhodne vybraná guľka bude mať zelenú farbu. Jedným riešením je zobrať škatuľu a vybrať z nej jednu guľku a zapísať si

výsledok[15]. Guľku potom vrátime do škatuľky, zatrasieme ňou a pokus opakujeme. Výsledok z každého takéhoto pokusu si pritom zapíšeme do tabuľky. V jednom riadku budú javy, ktoré môžu nastať (farba: *zelená, čierna a červená*) a v druhom počet, koľko krát nastali. Ak by sme tento pokus vykonali veľa krát, povedzme 10000 krát, zakaždým za tých istých podmienok, zrejme by sme pre každú farbu dostali podiel, koľko spomedzi všetkých pokusov sa vybrala práve konkrétna farba. V nasledujúcej tzv. **pravdepodobnostnej tabuľke** sú výsledky z takto simulovaného procesu.

Tabuľka 4: Pravdepodobnostná tabuľka

JAV	Zelená	Čierna	Červená
PODIEL	0.503	0.320	0.177

Zdroj: vlastné spracovanie

Účelom tohto príkladu bolo intuitívne vysvetliť jeden z prístupov k pravdepodobnosti. Podľa výsledkov zapísaných v uvedenej tabuľke môžeme vidieť, že najčastejšie sme vybrali guľku zelenú, a teda zrejme bude najpravdepodobnejším výsledkov ak realizujeme jeden pokus. Potom čiernu a nakoniec červenú guľku. Ide o prístup, ktorý vychádza z empiricky nameraných údajov a je zrejme prirodzenej intuícii najbližšie. Týmto spôsobom vieme odhadnúť, koľko guliek je ktorej farby (3 zelené, 2 čierne a 1 červená). Dôležité je, že jednotlivé pokusy by mali byť realizované za približne rovnakých podmienok. Inak sa môže stať, že výsledok pokusu ovplyvní iná skutočnosť. Ak by sme ako prvú guľku vybrali zelenú, vložili ju naspäť do škatule na vrch, a škatuľou nezatriasli, ľahko by sa mohlo stať, že v ďalšom pokuse znova vyberieme tú istú guľku. Naše výsledky by boli zrejme skreslené. Ide teda o pokus, ktorého výsledkom je **náhodný jav**. Čo je jav, ktorý nevieme s istotou predvídať. Opakom je tzv. **jav deterministický**. Ďalším dôležitým detailom je, že jednotlivé pokusy by mali byť na sebe nezávislé v zmysle, že výsledok jedného pokusu by nemal ovplyvniť výsledok druhého pokusu. V spoločenských vedách je dodržanie týchto podmienok neraz pomerne problematické. Ak robíme analýzu časových radov, neraz sú za sebou idúce pozorovania navzájom závislé (ide o tzv. autokoreláciu). Aj vizuálny pohľad na vývoj HDP krajiny naznačuje, že HDP krajiny sa nemení náhodne, ale má určitú tendenciu kopírovať vývoj z minulosti a mení sa len pomaly.

Vráťme sa ešte raz k pokusu s guľkami. Ak nás zaujíma pravdepodobnosť, že vytiahneme zelenú farbu a túto skutočnosť si symbolicky označíme ako $P(A)$, pričom počet

[15] Samozrejme mohli by sme celú škatuľku vyprázdniť a zistiť čo je v nej, ale to náš myšlienkový experiment nedovoľuje.

pokusov je n a počet výsledkov pokusov, v ktorých nastal jav A, si označíme ako $A(n)$, potom $P(A)$ (ak existuje také reálne číslo) si môžeme zapísať ako:

$$P(A) = \lim_{n \to \infty} \frac{A(n)}{n} \tag{4.1}$$

Samozrejme, nekonečný počet pokusov nikdy nezískame, a teda často nám situácia neumožňuje zistiť ani presnú pravdepodobnosť. Môžeme však rozumne predpokladať, že ak sú pokusy nezávislé a realizované za tých istých podmienok, tak čím viac pokusov realizujeme, tým bude odhad pravdepodobnosti presnejší. S takýmto poznaním si zvyčajne vystačíme.

V spoločenských vedách sa údaje získavajú pre pochopenie spravidla hromadných javov. Teda javov, ktoré sa skladajú z individuálnych javov. Nákup produktu v banke od jedného zákazníka je jav individuálny. Zriedkakedy nás zaujíma prečo došlo k nákupu práve u jedného konkrétneho zákazníka. To čo nás spravidla zaujíma je, prečo zákazníci vo všeobecnosti nakupujú tento produkt. Vlastnosti jedného zákazníka budú iné ako vlastnosti celej skupiny zákazníkov.

Definícia pravdepodobnosti vo vzťahu (4.1) je tzv. **definícia pravdepodobnosti cez početnosti** alebo aj **štatistická definícia pravdepodobnosti**, či frekventistická definícia pravdepodobnosti. Jednoduchá, intuitívne jasná a v praxi často použiteľná. Inou výzvou by bolo určiť pravdepodobnosť, že na zem dopadne meteorit väčší ako je plocha mesta Praha. Ide o jav, ktorý zrejme nie je hromadný. Tu nemôžeme realizovať množstvo pokusov a potom vyhodnotiť početnosť jednotlivých výsledkov. Ide o zriedkavý jav. Tu nám naša definícia pravdepodobnosti nepomôže. Jednou z možností je tzv. subjektívne stanovenie pravdepodobnosti, kde sa napríklad dopytujú experti.

V ďalšej časti stručne zadefinujeme tzv. **axiomatickú** definíciu **pravdepodobnosti**. Axióma je tvrdenie, ktorého pravdivosť sa nedokazuje.

Nech množina F je množina všetkých javov a množina S je množina všetkých elementárnych javov. Množina F zahrňuje všetky podmnožiny množiny S, množinu S samotnú a prázdnu množinu. **Elementárny jav** si definujeme ako jav, ktorý sa nedá rozložiť na menšie javy. Pre zjednodušenie si vzťah množiny F a S vysvetlíme na príklade s hádzaním kocky. Pri hádzaní kocky môže padnúť ľubovoľné číslo z množiny elementárnych javov $S = \{1, 2, 3, 4, 5, 6\}$. Jav (prvok množiny F) je ale udalosť, ktorá je predmetom nášho záujmu. Môže nás napríklad zaujímať, či padne párne alebo nepárne číslo. Či padne prvočíslo

({2, 3, 5} ∈ *F*) alebo číslo menšie ako 3 ({1, 2} ∈ *F*) alebo ľubovoľná kombinácia z týchto a mnoho iných javov[16]. Prvú axiómu si môžeme zapísať ako:

$$\forall A \in F : P(A) \geq 0 \qquad (4.2)$$

Pravdepodobnosť, že nastane nejaký jav *A* ∈ *F* je reálne číslo *p* také, ktoré pre všetky javy *A* ∈ *F* je vždy nezáporné.

$$P(S) = 1 \qquad (4.3)$$

Druhá axióma hovorí, že môžeme s istotou tvrdiť, že nastane nejaký jav z množiny všetkých elementárnych javov. Ak by sme hádzali štandardnou hracou kockou a zaujímalo by nás iba číslo, ktoré padne, potom množina všetkých možných javov *F* obsahuje možné kombinácie prvkov *S*. Táto axióma zároveň hovorí, že každému výsledku pokusu zodpovedá nejaký elementárny jav (prvok z množiny S). Ďalším dôsledkom je, že ak nepoznáme *všetky* možné javy, ktoré môžu nastať, nemôžeme poznať pravdepodobnosť nastania *akéhokoľvek* javu. Majme:

$$A_1, A_2, ..., A_m,\ A_i \cap A_j = \varnothing, i \neq j :\ P(A_1 \cup A_2 \cup ... \cup A_m) = \sum_{i=1}^{m} P(A_i) \qquad (4.4)$$

Tretia axióma hovorí o aditívnej vlastnosti pravdepodobnosti. Pravdepodobnosť, že padne číslo 2 alebo číslo 3, je súčet pravdepodobnosti padnutia čísla 2 a pravdepodobnosti padnutia čísla 3.

Trojici (*S*, *F*, *P*) hovoríme pravdepodobnostný priestor, kde *S* je množina všetkých možných elementárnych javov, *F* je množina javov (ktoré sú predmetom nášho záujmu) a P je pravdepodobnostná miera. V tomto prípade si môžeme ukázať niekoľko základných viet, ktoré je možné odvodiť z týchto axióm.

Pravdepodobnosť, že nastane žiadny jav je nulová:

$$P(\varnothing) = 0 \qquad (4.5)$$

Pravdepodobnosť, že nastane jav *A* je menšia alebo rovná ako 1:

$$P(A) \leq 1 \qquad (4.6)$$

Pravdepodobnosť, že nenastane jav *A* je:

$$P(A^C) = 1 - P(A) \qquad (4.7)$$

Kde horným indexom *C* označujeme doplnok k javu *A*. Ak z nastania javu *A* vyplýva nastanie javu *B* (pričom to nemusí platiť naopak), potom:

$$A \subset B \text{ a } P(A) \leq P(B) \qquad (4.8)$$

Pravdepodobnosť, že nastane jav *A* alebo jav *B* je:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (4.9)$$

[16] Zaujímavým dôsledkom je, že súčet pravdepodobnosti pre všetky javy z množiny *F* **môže** byť väčší ako 1.

Pre ľubovoľné javy A a B platí:

$$P(A \cap B) = P(A) - P(A \cap B^C) \tag{4.10}$$

Pravdepodobnosť, že nastane jav A ak nastane jav B, označujeme ako $P(A|B)$ a hovoríme tomu podmienená pravdepodobnosť. Z axióm ďalej vyplýva:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{4.11}$$

Ide o tzv. **Bayesovu vetu**. Možný príklad, na ktorom je možné túto vetu vysvetľovať je vychádza z diagnostiky pacientov[17]. Majme populáciu žien vo veku 45 rokov. Z tejto populácie má 1 % žien rakovinu prsníka. V 80 % prípadov žien, ktoré majú rakovinu prsníka sa aj v skutočnosti diagnostikuje rakovina prsníka. V 9.6 % prípadov žien, ktorým sa diagnostikuje rakovina prsníka ju v skutočnosti nemá. Aká je pravdepodobnosť, že ak náhodne vybranej žene diagnostikujú rakovinu prsníka, tak ju aj v skutočnosti bude mať? Ide o zaujímavý príklad, kde nastanie javu B (diagnostikovanie rakoviny) spresňuje pravdepodobnosť nastania javu A (rakovina prsníka). Zo zadania je zrejmé, že pravdepodobnosť, že žena bude mať v skutočnosti rakovinu je $P(A) = 0.01$, ďalej pravdepodobnosť, že ak má žena rakovinu, tak test jej chorobu odhalí je $P(B|A) = 0.8$. Ešte je potrebné určiť pravdepodobnosť, že test bude pozitívny (jav B). Ak v 80 % prípadov žien, ktorým sa diagnostikuje rakovina prsníka má v skutočnosti rakovinu prsníka, tak potom 0.8×0.01 hovorí o tom, aká je pravdepodobnosť, že u náhodne vybranej ženy bude test pozitívny, ak táto žena má rakovinu. Ďalej výraz 0.096×0.99 hovorí o tom, aká je pravdepodobnosť, že u náhodne vybranej ženy bude test pozitívny, ak táto žena nemá rakovinu. Spolu je tak pravdepodobnosť, že test bude pozitívny $P(B) = 0.8 \times 0.01 + 0.096 \times 0.99 = 0.10304$. Tento tvar je možné prepísať nasledovne: $P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$. O tento tvar je možné rozšíriť vzťah (4.11). Výsledok je potom približne 0.078, ktorý nám hovorí o tom, že ak sa diagnostikuje žene rakovina prsníka, tak pravdepodobnosť že ju v skutočnosti bude mať je 0.078. Tento výsledok je často vnímaný ako pomerne kontraintuitívny. Jedným z dôvodov prečo ide o tak malé číslo je v tom, že tento test v pomerne veľa prípadoch dáva falošné signály (9.6 % medzi zdravými ženami – predstavte si v koľkých prípadoch budú ženy chybne diagnostikované, ak sa testom podrobí 1000 žien). K tejto problematike sa dostaneme neskôr. Zaujímavou črtou tejto vety je skutočnosť, že nám umožňuje bližšie určiť pravdepodobnosť nastania určitého javu v prípade, že nastane jav iný. V tomto prípade vieme vyhodnotiť, nakoľko nám pomôže diagnostický nástroj nájsť rakovinu. Aj keď nejde o

[17] Prevzali sme ho z http://oscarbonilla.com/2009/05/visualizing-bayes-theorem/; dostupné online [17.02.2012].

dokonalý diagnostický nástroj, v praxi pomáha, keďže z 1 % sme sa dostali na 7.8 %. Pravdepodobnosti $P(A)$ sa hovorí „*prior*“. Ide o dopredu známu informáciu. $P(B|A)$ a $P(B|A^C)$ sú podmienené pravdepodobnosti, ktoré sú tiež dopredu známe. Všimnime si, čo by sa stalo, ak by tieto podmienené pravdepodobnosti boli rovnaké. Znamenalo by to, že použitím testu by sme diagnostikovali rakovinu s rovnakou pravdepodobnosťou u zdravých ako aj u chorých žien. Asi by nešlo o dobrý test, keďže by nebol schopný rozlišovať medzi týmito dvoma stavmi. Uvedené vyplýva aj z jednoduchých formálnych úprav: $P(B|A) = P(B|A^C)$, taktiež platí $P(A) = 1 - P(A^C)$, potom: $P(A|B) = (P(A)P(B|A))/(P(B|A)\,P(A) + P(B|A)\,(1 - P(A^C)) = P(A)$. Z toho teda vyplýva, že nastanie javu B nám nijakým spôsobom nepomáha vysvetliť pravdepodobnosť nastania javu A. Týmto sa dostávame k užitočnej definícii nezávislosti javov.

Môžeme si teraz priamo zadefinovať pojem nezávislosti. Ak máme jav A a B, potom hovoríme, že sú tieto dva javy **štatisticky nezávislé**, ak platí:

$P(A), P(B) > 0 \wedge P(A|B) = P(A)$ alebo $P(A), P(B) > 0 \wedge P(B|A) = P(B)$, resp. ak platí $P(A \cap B) = P(A)P(B)$.

Príklad 4.1

Nasledujúci príklad predstavuje pomerne známy paradox. V televíznej súťaži moderátor súťažiacemu ukáže tri dvere. Za jednými dverami sa skrýva auto, za ostatnými dvoma nič. Moderátor vie, za ktorými dverami sa nachádza auto. Moderátor vyzve súťažiaceho, aby si vybral dvere. Ak si vyberie dvere za ktorými je auto, toto auto vyhráva. Súťažiaci si vyberie dvere č. 1, avšak tie ostávajú ešte stále zatvorené. Z ostávajúcich dvoch dverí, moderátor jedny dvere otvorí. Otvorí samozrejme dvere, za ktorými nič nie je, povedzme dvere č. 3. Následne vyzve súťažiaceho, aby buď zotrval na svojej pôvodnej voľbe (dvere č. 1) alebo zmenil svoju voľbu (na dvere č. 2). Otázkou je, čo má súťažiaci urobiť, resp. inak povedané, akú stratégiu má súťažiaci na začiatku zvoliť: zmenu voľby alebo zotrvanie pri pôvodnej voľbe?

Je zrejmé, že na začiatku je pravdepodobnosť, že súťažiaci vyberie správne dvere 1/3. Aby sme zistili správnu odpoveď na otázku, môžeme si znázorniť všetky možné situácie a voľby súťažiaceho. Predpokladajme, že si vyberie dvere č. 1.

- Za dverami č. 1 je auto. Ak ostane pri svojej voľbe vyhráva auto. Ak zmení svoju voľbu auto nevyhrá.
- Za dverami č. 1 nie je auto. Auto je za dverami č. 2. Moderátor otvorí dvere č. 3. Ak ostane pri svojej voľbe nevyhrá auto. Ak zmení svoju voľbu auto vyhrá.

- Za dverami č. 1 nie je auto. Auto je za dverami č. 3. Moderátor otvorí dvere č. 2. Ak ostane pri svojej voľbe nevyhrá auto. Ak zmení svoju voľbu auto vyhrá.

Ak nebude svoju voľbu meniť, v scenároch vyhrá iba v jednom z troch prípadov, čo zodpovedá pravdepodobnosti 1/3. Ak bude svoju voľbu meniť, v scenároch vyhrá auto v dvoch z troch prípadoch, čo zodpovedá pravdepodobnosti 2/3. Preto sa súťažiacemu oplatí zmeniť svoju voľbu.

4.2 Rozdelenie pravdepodobnosti

V tejto časti sa budeme venovať pojmom ako náhodná premenná a pokúsime sa intuitívne priblížiť význam rozdelenia pravdepodobnosti. Nevyhneme sa určitým formálnejším zápisom ani zjednodušeniam. Pre bežné používanie štatistiky nie je nutné tieto formálne zápisy ovládať (je to nesporne výhodou), na druhej strane, ak čitateľ prostredníctvom nich nájde motiváciu k podrobnejšiemu štúdiu, môže k tomu použiť študijnú literatúru venujúcu sa matematickej štatistike a pravdepodobnosti.

Ak vykonáme nejaký pokus (v spoločenských vedách tu patrí aj dopytovanie sa osôb) a s istotou nevieme povedať, aký bude výsledok tohto pokusu, hovoríme, že to, čo je predmetom pokusu (to čo meriame) je náhodný jav. **Náhodná premenná** je funkcia, ktorá priraďuje výsledku pokusu reálne číslo[18]. Majme pravdepodobnostný priestor (S, F, P), potom za náhodnú premennú budeme považovať funkciu $X: S \rightarrow R$.

Príklad 4.2

Výsledkom pokusu nemusí byť len reálne číslo. Môže to byť farba (aká farba sa páči zákazníkom?), vlastnosť (typ osobnosti pri psychoteste) a podobne. Je však pravdou, že uprednostňujeme číselné vyjadrenie náhodného javu, a preto ak to je možné, výsledky zmysluplne kódujeme. Ak napríklad nejaký jav môže, resp. nemusí nastať, tak nastanie označíme reálnym číslom 1 a nenastanie 0. Je teda vhodné už pri plánovaní experimentov, výskumov a prieskumov myslieť na túto skutočnosť a snažiť sa vymyslieť také premenné, ktoré bude možné merať pomocou reálnych čísel na tzv. podielovej alebo intervalovej škále.

Ak nás pri hádzaní kocky zaujíma, či padne párne alebo nepárne číslo, potom náhodná premenná nám podľa výsledku pokusu vráti hodnotu 1 (párne číslo) alebo 0 (nepárne

[18] Nami použitá definícia náhodnej premennej nie je najvšeobecnejšia. Výsledkom náhodného javu môže byť aj kvalitatívna premenná, kde je zrejme vždy možné urobiť kódovanie do množiny R.

číslo), v závislosti od toho, aký elementárny jav nastane ({1, 3, 5} v prípade nepárneho výsledku a {2, 4, 6} v prípade párneho výsledku).

Budeme rozlišovať medzi dvoma skupinami náhodných premenných. Spojité náhodné premenné a diskrétne náhodné premenné[19]. **Diskrétne náhodné premenné** nadobúdajú iba hodnoty, ktorých počet je konečný alebo majú spočítateľný počet možných hodnôt. **Spojitá náhodná premenná** môže nadobudnúť ľubovoľnú hodnotu medzi ľubovoľnými dvoma rôznymi hodnotami. V niektorej literatúre sa s vymedzením diskrétnej náhodnej premennej môžeme stretnúť aj tak, že pokiaľ náhodná premenná nie je spojitá, potom je diskrétna.

Príklad 4.3

Diskrétna náhodná premenná je napríklad počet autobusov na zastávke v priebehu jednej hodiny alebo počet správnych odpovedí na teste, prípadne počet výhercov v lotérii. Spravidla sa jedná o „počet“. Spojitá náhodná premenná môže byť výška, váha ľudí, objem vody vo fľaši a podobne. Nie vždy je ľahké zistiť, či náhodná premenná je diskrétneho alebo spojitého charakteru[20]. Napríklad peniaze sú spojitá náhodná premenná. To, že nie sú technicky nekonečne deliteľné (keďže najmenší je 1 cent, ak máme na mysli eurá), na tejto skutočnosti nič nemení. Je rozdiel medzi škálou, pomocou ktorej meriame prejavy premennej a samotnou náhodnou premennou.

4.2.1 Rozdelenie početností

Vychádzajme zo situácie, kde máme 10000 pokusov, pri ktorých hádžeme férovou kockou. Po 10000 pokusoch sú výsledky znázornené v nasledujúcej tabuľke. Tejto tabuľke hovoríme tabuľka rozdelenia početností. Rozdelenie početností zaznamenáva, s akou početnosťou nastávajú rôzne javy, keďže každému javu priradí početnosť (rozdelenie) jeho výskytu.

[19] Môže existovať aj prípad, keď náhodná premenná je na nejakom intervale povedzme <0,1> spojitá, na (1,2) nie je definovaná a na <2;3> je znova spojitá. Týmto prípadom sa nebudeme venovať. Pojem spojitej a diskrétnej premennej sme už v predchádzajúcich častiach textu venovali určitý priestor. Na tomto mieste však definujeme náhodné diskrétne a spojité premenné.

[20] V tejto súvislosti sú zaujímavé problémy ohľadom rozdielu medzi spojitosťou / nespojitosťou náhodnej premennej a príslušnej škály merania. Teoreticky ani jednu spojitú premennú nevieme merať tak, aby samotné možné výsledky pomocou meracieho prístroja mohli nadobúdať ľubovoľné hodnoty. Je to prirodzene dané technickým obmedzením. Niektoré štatistické metódy vyžadujú, aby bola náhodná premenná spojitá. Nie však, aby bola škála merania spojitá.

Tabuľka 5: Tabuľka rozdelenia početností

JAV	1	2	3	4	5	6
POČET	1625	1695	1651	1690	1633	1706

Zdroj: vlastné spracovanie

Rozdelenie početností je vhodné použiť iba v situáciách, kde je celková početnosť konečná. S použitím sa teda môžeme stretnúť najmä pri empirických štatistických súboroch. Rozdelenie početností sa vizuálne znázorňuje pomocou histogramu alebo pomocou tzv. empirickej distribučnej funkcie, ktorú si ukážeme v ďalšej časti.

4.2.2 Diskrétne a spojité rozdelenie pravdepodobnosti

Ak prijmeme štatistickú definíciu pravdepodobnosti, potom pravdepodobnostná tabuľka (Tabuľka 4) by pre nás mohla predstavovať odhad rozdelenia pravdepodobnosti, keďže každému možnému javu priraďuje určitú pravdepodobnosť, s ktorou daný jav nastal. Zadefinujeme si termín „rozdelenie pravdepodobnosti“ zvlášť pre diskrétne a pre spojité premenné. V úvode tejto časti sme nespomenuli dôležitý rozdiel medzi spojitou a diskrétnou premennou. Kým v prípade diskrétnej náhodnej premennej má význam pýtať sa, s akou pravdepodobnosťou nastane nejaký elementárny jav (padne číslo 2), tak v prípade spojitej náhodnej premennej musíme jav definovať ako určitý číselný interval (s akou pravdepodobnosťou bude náhodne vybraný človek mať výšku od 160 do 170 cm).

Dôvod vyplýva zo samotnej definície spojitej náhodnej premennej a pravdepodobnosti. Ak môže nastať ľubovoľná hodnota medzi dvoma rôznymi hodnotami, tak zrejme môže nastať nekonečne veľa možností. Keďže pravdepodobnosť, že nastane nejaký jav z množiny možných javov je 1 a týchto javov je ∞, intuitívne z toho vyplýva, že pravdepodobnosť vzniku jedného z týchto javov je nekonečne malá (veľmi blízka nule). Riešením je priradenie pravdepodobnosti intervalu hodnôt namiesto konkrétnym hodnotám ktoré náhodná premenná môže nadobúdať.

Príklad 4.4

Iným myšlienkovým experimentom je nasledujúca situácia. Majme terč, na ktorý hádžeme šípky. V našom myšlienkovom experimente hodíme šípku a určite trafíme terč. Nevieme však, ktorý bod terča trafíme. Týchto bodov je nekonečne veľa, a tak sa zdá byť intuitívne správne uvažovať o pravdepodobnosti zasiahnutia ľubovoľného bodu (vychádzajúc z geometrickej definície pravdepodobnosti) ako o nekonečne malej pravdepodobnosti. Napriek tomu, ak hodíme šípku, tak trafíme nejaký bod na terči. Ak si náhodne vyberieme

bod na terči, hovoríme, že pravdepodobnosť, že ho trafíme je takmer isto nulová. To neznamená, že ten jav nemôže nastať. Môže, avšak jeho pravdepodobnosť je prakticky nulová.

Ak X je diskrétna náhodná premenná, hovoríme, že má diskrétne rozdelenie pravdepodobnosti, ak platí: $\sum_s P(\{s\}) = 1$, kde s sú elementárne javy množiny všetkých možných elementárnych javov S a $P(\{s\}) \geq 0$. Inak povedané, ak sčítame pravdepodobnosť nastania všetkých možných elementárnych javov, dostaneme pravdepodobnosť rovnú 1. Inou definíciou je nasledujúca: *Diskrétne rozdelenie pravdepodobnosti je pravidlo, ktoré každej možnej hodnote diskrétnej náhodnej premennej priradí pravdepodobnosť, že náhodná premenná nadobudne túto hodnotu* (upravené podľa Tkáč, 2001).

Príklad 4.5

Všimnime si prípad štandardnej hracej kocky. Vieme, že pri hádzaní takejto férovej kocky bude pravdepodobnosť, že padne ľubovoľné číslo, vždy rovnaká a to 1/6. To znamená, že pravidlo, ktoré priradí každej možnej hodnote diskrétnej náhodnej premennej (hodnotám 1, 2, 3, 4, 5 a 6) pravdepodobnosť, tak robí „rovnomerne“. Teda ak chceme vedieť akým pravidlom sa riadi padnutie čísla na hracej kocke, odpoveď je: rovnomerným diskrétnym rozdelením pravdepodobnosti. Ak by však kocka nebola férová, bola by rôzne vyvážená, potom nie každý jav by bol rovnako pravdepodobný. Nás by mohlo zaujímať, podľa akého pravidla sa budú jednotlivým javom priraďovať pravdepodobnosti. Môžeme uskutočniť experiment a toto pravidlo odhadnúť po povedzme 10000 pokusoch v hádzaní kocky.

Dôležitý koncept je **kumulatívna distribučná funkcia *F*(*x*)**, ktorá každej číselnej hodnote x diskrétnej náhodnej premennej X s diskrétnym rozdelením pravdepodobnosti priradí pravdepodobnosť, že náhodná premenná X bude menšia (a rovná) ako x:

$$F(x) = P(X \leq x) \tag{4.12}$$

Niektoré dôležité vlastnosti kumulatívnej distribučnej funkcie sú:

$$\lim_{x \to -\infty} F(x) = 0 \text{ a } \lim_{x \to +\infty} F(x) = 1 \tag{4.13}$$

$$\text{ak } a, b \in R \text{ a } a < \text{b}, \text{ potom } F(a) \leq F(b), \tag{4.14}$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) \tag{4.15}$$

Spojitá náhodná premenná X má funkciu **hustoty rozdelenia pravdepodobnosti *f*(*x*)** takú, že platí:

$$f(x) \geq 0, -\infty < x < \infty \tag{4.16}$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{4.17}$$

$$P(X = c) = 0, c \in R \tag{4.18}$$

Potom vieme vypočítať pravdepodobnosť, že náhodná premenná X nadobudne ľubovoľnú hodnotu z číselného intervalu tak, že:

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx, a,b \in R \wedge a \le b \tag{4.19}$$

Pripomenieme, že existujú aj také rozdelenia spojitej náhodnej premennej, ktoré nemajú hustotu. Týmto prípadom sa nebudeme bližšie venovať. Znova využijeme definíciu podľa Tkáč (2001), ktorý spojité rozdelenie pravdepodobnosti definuje jednoducho a veľmi výstižne ako: .... *pravidlo, ktoré množine hodnôt z každého číselného intervalu spojitej náhodnej premennej priradí pravdepodobnosť, že náhodná premenná nadobudne hodnotu z tohto intervalu*.

Príklad 4.6

Zoberme si ako príklad váhu novorodencov, ktorá bude pre nás spojitá náhodná premenná. Keďže nevieme exaktne akým rozdelením pravdepodobnosti sa riadi váha novorodencov, praktický postup je taký, že najprv získame údaje prostredníctvom reprezentatívnej vzorky, pričom odhadneme spojité rozdelenie pravdepodobnosti. Na základe neho potom budeme vedieť rozhodnúť, s akou pravdepodobnosťou bude mať novorodenec váhu menšiu ako 2 kg. Ak by (čisto teoreticky) takáto váha bola málo pravdepodobná a narodil by sa novorodenec s takou váhou, zrejme by to bol dôvod na vyšetrovanie príčiny a osobitej starostlivosti. Existuje aj iný postup, a to pomocou empirickej distribučnej funkcie, ktorú si popíšeme neskôr.

Kumulatívna distribučná funkcia spojitého rozdelenia pravdepodobnosti má rovnaké vlastnosti ako pri diskrétnom rozdelení pravdepodobnosti, avšak jednou dôležitou vlastnosťou je, že derivovaním kumulatívnej distribučnej funkcie vieme získať funkciu hustoty pravdepodobnosti.

Zatiaľ sme si v príkladoch ukazovali iba jedno diskrétne rovnomerné rozdelenie pravdepodobnosti. Teoretických rozdelení pravdepodobnosti je pomerne veľa (zrejme v stovkách, príp. v tisícoch) a spravidla v každom vednom obore sa používa určitá skupina, niektoré však predsa len považujeme za častejšie používané a týmto budeme venovať krátky prehľad v nasledujúcej časti. Najprv si však ešte definujeme dva dôležité koncepty: stredná hodnota a disperzia.

Stredná hodnota diskrétnej náhodnej premennej

Stredná hodnota náhodnej premennej X (nazývaná aj očakávaná hodnota) je:

$$E[X] = \mu = \sum_{s \in S} X(s) P(\{s\}) \tag{4.20}$$

a stredná hodnota funkcie $g(x)$ je:

$$E[g(X)] = \mu_{g(X)} = \sum_{s \in S} g(X(s)) P(\{s\}) \tag{4.21}$$

Ako vyplýva z týchto vzťahov, stredná hodnota má pravdepodobnostný charakter a vyjadruje, akú hodnotu môžeme pri danom pravdepodobnostnom rozdelení očakávať. Druhý vzťah je tiež pomerne často používaným tvarom funkcie, ktorý uvádzame pre úplnosť. V tomto texte sa mu bližšie venovať nebudeme.

Disperzia diskrétnej náhodnej premennej

Disperzia vyjadruje mieru rozptýlenia hodnôt v príslušnom rozdelení pravdepodobnosti. Vzťah na výpočet disperzie diskrétnej náhodnej premennej je nasledovný:

$$D[X] = E\left[(X - \mu)^2\right] = \sum_{s \in S} (s - \mu)^2 P(\{s\}) \tag{4.22}$$

Stredná hodnota spojitej náhodnej premennej

Ak spojitá náhodná premenná má hustotu rozdelenia $f(x)$, potom stredná hodnota náhodnej premennej X je:

$$E[X] = \mu = \int_{-\infty}^{+\infty} x f(x) dx \tag{4.23}$$

Stredná hodnota funkcie $g(x)$ je:

$$E[g(X)] = \mu_{g(X)} = \int_{-\infty}^{+\infty} g(x) f(x) dx \tag{4.24}$$

Disperzia spojitej náhodnej premennej

Pre spojitú náhodnú premennú X platí, že jej disperzia je:

$$D[X] = E\left[(X - \mu)^2\right] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \tag{4.25}$$

Príklad 4.7

Študenti si často dávajú otázku, aký je rozdiel medzi aritmetickým priemerom a strednou hodnotou. Najmä ak uvážime, že matematický vzťah medzi váženým aritmetickým priemerom, kde váhy sú príslušné pravdepodobnosti nastania jednotlivých hodnôt, je matematicky totožný so vzťahom pre výpočet strednej hodnoty diskrétnej náhodnej premennej. Na ilustráciu rozdielu si pomôžeme jednou zjednodušenou analógiou.

Predstavme si, že pri meraní spokojnosti zamestnancov na škále od 1 – 10 v 15-tich pokusoch nameriame nasledujúce hodnoty:
8, 9, 10, 11, 8, 5, 6, 8, 9, 10, 11, 2, 9, 9, 8,
Ak vypočítame aritmetický priemer a budeme ho považovať za strednú hodnotu, potom zjavne predpokladáme, že váha týchto hodnôt je daná ich početnosťou v súbore nameraných hodnôt. Inou alternatívou je predpokladať, že spokojnosť zákazníkov sa riadi určitým vybraným rozdelením pravdepodobnosti. V takom prípade sú pri výpočte strednej hodnoty váhou „pravdepodobnosti“ vypočítané z daného predpokladaného rozdelenia pravdepodobnosti.
S odlišným vysvetlením sa môžeme stretnúť, ak sa rozlišuje či ide o výberový súbor alebo o populáciu, čo je však predmetom publikácií venujúcich sa induktívnej štatistike.

4.2.3 Empirická distribučná funkcia

Aby sme mohli zadefinovať empirickú distribučnú funkciu potrebujeme trochu tematicky predbehnúť. Predpokladajme, že hodnoty náhodnej premennej X sú z reprezentatívnej (náhodnej) vzorky a sú zotriedené tak, že žiadne dve hodnoty vo variačnom rade sa nevyskytujú viac ako jeden krát (podobne ako v Kapitole 3.4.1 a 3.4.4), teda $x_{(1)}$, $x_{(2)}$,..., $x_{(s)}$, pričom $s \leq n$, kde n je rozsah štatistického súboru, pričom n_j je absolútna početnosť j-tej hodnoty vo výberovom súbore, j = 1, 2, ..., s. Potom empirickú distribučnú funkciu $\hat{F}_n(x)$ môžeme formálne definovať nasledovne:

$$\hat{F}_n(x) = \begin{cases} 0, & \text{ak} \quad x < x_{(1)} \\ \dfrac{1}{n} \displaystyle\sum_{x \leq x_{(j)}} n_j, & \text{ak} \quad x_{(1)} \leq x < x_{(s)} \\ 1, & \text{ak} \quad x \geq x_{(s)} \end{cases} \tag{4.26}$$

kde absolútna kumulatívna početnosť je:

$$\sum_{x \leq x_{(j)}} n_j \tag{4.27}$$

Príklad 4.8
Uvažujme o vzorke n = 100 novorodencov, kde náhodnou premennou je váha novorodenca. Ak si zoradíme váhy novorodencov zistíme, že so zaokrúhlením na jednu desatinu máme 12 rôznych váh, t.j. s = 12 (viď. nasledujúcu tabuľku).

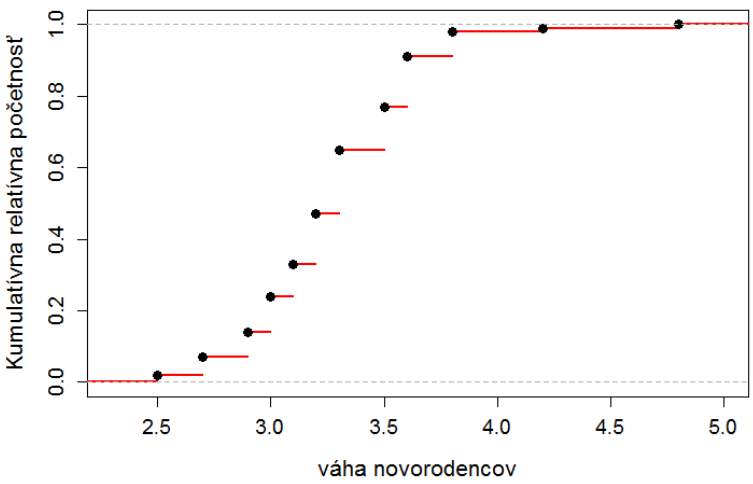
Tabuľka 6: Váha novorodencov

VÁHA	2.5	2.7	2.9	3.0	3.1	3.2	3.3	3.5	3.6	3.8	4.2	4.8
POČET	2	5	7	10	9	14	18	12	14	7	1	1

Zdroj: vlastné spracovanie

Najmenšou váhou je 2.5 kg a najväčšou 4.8 kg. Zo vzťahu (4.26) vyplýva, že hodnota $\hat{F}_n(x)$ pre hmotnosť menšiu ako 2.5 kg je 0 (t.j. nulová pravdepodobnosť) a pre hmotnosť väčšiu ako 4.8 kg je 1. Čo to v skutočnosti znamená je, že pravdepodobnosť, že váha novorodenca je menšia ako 2.5 kg je na základe vzorky 0 a zároveň, že váha novorodenca je menšia ako 4.8 kg s pravdepodobnosťou 1. Pre všetky váhy v tomto intervale platí $\frac{1}{n}\sum_{j=1}^{i} n_j$. Čiže napríklad váha 3.0 kg je 4-tá v poradí, t.j. $i = 4$, $\sum_{j=1}^{i} n_j = \sum_{j=1}^{4} n_j =$ $= 2+5+7+10 = 24$, takže $\hat{F}_{100}(3.0) = 24/100 = 0.24$. Pravdepodobnosť, že náhodne vybrané dieťa bude mať váhu menšiu alebo rovnú 3.0 kg je 0.24. Na nasledujúcom obrázku je znázornený vývoj empirickej distribučnej funkcie.

```
> vaha <- c(2.5, 2.7, 2.9, 3.0, 3.1, 3.2, 3.3, 3.5, 3.6, 3.8,
  4.2, 4.8)
> pocet <- c(2, 5, 7, 10, 9, 14, 18, 12, 14, 7, 1, 1)
> vaha_vsetko <- c()
> for (i in 1:length(vaha)) vaha_vsetko <- c(vaha_vsetko,
  rep(vaha[i], pocet[i]))
> data <- ecdf(vaha_vsetko)
> plot(data, xlab = "váha novorodencov", cex = 1.3,
  col.hor="red", ylab = "Kumulatívna relatívna početnosť", lwd =
  2, cex.axis = 1.2, cex.lab = 1.3, main = NA)
```



Obrázok 4.2: Kumulatívna distribučná funkcia

Zdroj: vlastné spracovanie v programe R

Empirická distribučná funkcia má nasledujúce tri vlastnosti:

- $\lim_{x \to -\infty} \hat{F}_n(x) = 0$ a $\lim_{x \to +\infty} \hat{F}_n(x) = 1$,
- $\hat{F} : R \to \langle 0,1 \rangle$ a je rastúca na R,
- $\hat{F}$ je sprava spojitá, schodovitá funkcia.

Prvá vlastnosť hovorí, že pre x idúce limitne k + ∞ hodnota empirickej distribučnej funkcie má limitu 1 a naopak, pre x idúce limitne k – ∞ hodnota empirickej distribučnej funkcie má limitu 0. Druhá vlastnosť znamená, že empirická distribučná funkcia (ďalej EDF z angl. *Empirical Distribution Function*) nadobúda hodnoty z intervalu $\langle 0,1 \rangle$, čo je zrejmé aj z definície pravdepodobnosti. Tretiu vlastnosť vieme interpretovať aj nasledovne. Zoberme si 3.0 kg ako hodnotu z definičného oboru EDF. Ak by bola EDF spojitá zľava, muselo by platiť, že $\lim_{x \to 3^-} \hat{F}_n(x) = \hat{F}_n(3)$ čo neplatí, ale platí $\lim_{x \to 3^+} \hat{F}_n(x) = \hat{F}_n(3)$, a teda je spojitá sprava. Spojitosť sprava je konvencia.

4.3 Chebyshevova nerovnosť

Chebyshevovu nerovnosť môžeme využiť pri tzv. „zákone veľkých čísel“. Predtým ako si ju formálnejšie vysvetlíme, ilustrujeme si jej pomerne všeobecné použitie, ktoré je bližšie zameraniu tejto publikácie. Chebyshevova nerovnosť tvrdí, že pre všetky štatistické súbory (nezávisle od rozdelenia pravdepodobnosti) je:

- aspoň 3/4 alebo 75 % hodnôt v rozmedzí 2 smerodajných odchýlok od aritmetického priemeru,
- aspoň 8/9 alebo 88.89 % hodnôt v rozmedzí 3 smerodajných odchýlok od aritmetického priemeru,
- aspoň 15/16 alebo 93.75 % hodnôt v rozmedzí 4 smerodajných odchýlok od aritmetického priemeru,
- aspoň 24/25 alebo 96 % hodnôt v rozmedzí 5 smerodajných odchýlok od aritmetického priemeru,
- aspoň 35/36 alebo 97.22 % hodnôt v rozmedzí 6 smerodajných odchýlok od aritmetického priemeru.

Bez toho aby sme použili pojmy ako pravdepodobnosť, si Chebyshevovu nerovnosť môžeme pre praktické použitie definovať nasledovne. Nech X je ľubovoľná náhodná

premenná so strednou hodnotou μ a s rozptylom σ^2. Potom pre ľubovoľné c > 1, aspoň 1 – $1/c^2$ podiel hodnôt sa nachádza v intervale $\mu \pm c\sigma$.

$$\mu - c\sigma < 1 - \frac{1}{c^2} < \mu + c\sigma \tag{4.28}$$

Príklad 4.9

Prieskum trhu ukázal, že zákazníci u konkurencie nakupujú tovar v priemere každých 3.4 dní so smerodajnou odchýlkou 1.1. Na základe Chebyshevovej nerovnosti odhadnime, koľko zákazníkov nakupuje u konkurencie tovar v rozmedzí 2 až 4.8 dní.

$$\mu - c\sigma < 1 - \frac{1}{c^2} < \mu + c\sigma;\ 3.4 - 1.1c < 1 - \frac{1}{c^2} < 3.4 + 1.1c;\ \approx 38.26\,\%$$

A teda aspoň 38.26 % zákazníkov konkurencie si tovar kupuje v rozmedzí 2 až 4.8 dní.

Chebyshevovu nerovnosť si teraz definujeme formálnejšie tak, aby sme ju mohli použiť pri zákone veľkých čísel. Nech je X diskrétna náhodná premenná so strednou hodnotou μ a nech $\varepsilon > 0$. Potom:

$$P(|X - \mu| \geq \varepsilon) \leq \frac{D[X]}{\varepsilon^2} \tag{4.29}$$

Táto veta hovorí, že náhodná premenná X sa bude nachádzať mimo intervalu $\mu \pm \varepsilon$ s pravdepodobnosťou nie väčšou ako D[X] / ε^2. Ľavá strana vzťahu (4.29) sa dá vyjadriť pomocou kumulatívnej distribučnej funkcie $F(x) = P(X \leq x)$. Všimnime si, že nás vlastne zaujíma $F(\mu - \varepsilon)$ a zároveň $(1 - F(\mu + \varepsilon))$. Pre diskrétne rozdelenie pravdepodobnosti ďalej platí, že kumulatívnu distribučnú funkciu dostaneme sčítaním príslušných pravdepodobností. Napríklad, ak nás pri hádzaní férovou kockou zaujíma $F(3)$, potom $P(X = 1) + P(X = 2) + P(X = 3) = P(X \leq 3) = 1/6 + 1/6 + 1/6 = 1/2$. Ak si teda označíme $f(x)$ za distribučnú funkciu, potom pre diskrétne rozdelenie bude platiť:

$$P(|X - \mu| \geq \varepsilon) = \sum_{|x-\mu| \geq \varepsilon} f(x) \tag{4.30}$$

Zároveň vieme, že pre disperziu platí $D[X] = \Sigma\,(x - \mu)^2 f(x)$. Potom sa dá jednoducho ukázať, že platia nasledovné nerovnosti:

$$D[X] = \sum_{x} (x - \mu)^2 f(x) \geq \sum_{|x-\mu| \geq \varepsilon} (x - \mu)^2 f(x) \geq \sum_{|x-\mu| \geq \varepsilon} \varepsilon^2 f(x) \tag{4.31}$$

Ďalej vyberieme konštantu pred sumu a dostaneme Chebyshevovu nerovnosť.

$$\sum_{|x-\mu| \geq \varepsilon} \varepsilon^2 f(x) = \varepsilon^2 \sum_{|x-\mu| \geq \varepsilon} f(x) = \varepsilon^2 P(|X - \mu| \geq \varepsilon) \tag{4.32}$$

$$D[X] \geq P(|X-\mu| \geq \varepsilon)$$
$$\frac{D[X]}{\varepsilon^2} \geq P(|X-\mu| \geq \varepsilon) \tag{4.33}$$

Z tejto definície sa dá ľahko ukázať väzba na našu prvú definíciu. Stačí, ak si položíme $\varepsilon = k\sigma$, potom:

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{k^2\sigma^2} \tag{4.34}$$

Pre spojitú náhodnú premennú je podstata predošlého postupu rovnaká, rozdiel v definícii spočíva v tom, že je potrebné mať definovanú hustotu rozdelenia, strednú hodnotu a rozptyl. Potom pre spojitú náhodnú premennú má Chebyshevova nerovnosť tvar:

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} = \frac{D[X]}{\varepsilon^2} \tag{4.35}$$

Ide o analogický výraz ako pre diskrétnu náhodnú premennú. Označenie disperzie je otázkou konvencie. V našom texte sme pre diskrétne aj spojité rozdelenia používali $D[X]$, ktoré voľne zamieňame s výrazom σ^2.

4.4 Zákon veľkých čísel

Majme náhodné premenné X_i, $i = 1, 2, ..., n$, ktoré sú nezávislé realizácie z rovnakého rozdelenia pravdepodobnosti s definovanou (konštantnou) strednou hodnotou μ. Nech pre $\overline{X}_n$ platí:

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \tag{4.36}$$

Zákon veľkých čísel hovorí o konvergencii tohto aritmetického priemeru k strednej hodnote μ pre zvyšujúce sa n. Existuje niekoľko podôb tohto zákona: slabý zákon veľkých čísel a silný zákon veľkých čísel patria k najčastejšie uvádzaným. Slabý zákon veľkých čísel sa zaoberá limitou pravdepodobnosti týkajúcej sa $\overline{X}_n$ a silný zákon veľkých čísel pravdepodobnosťou limity $\overline{X}_n$. Slabý zákon veľkých čísel si môžeme formálne vyjadriť ako:

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|\overline{X}_n - \mu| \leq \varepsilon) = 1 \tag{4.37}$$

Tento výraz si môžeme interpretovať tak, že $\overline{X}_n$ konverguje v pravdepodobnosti k μ pre $n \to \infty$. Uvedené tvrdenie si môžeme overiť pomocou simulácie takým spôsobom, že ak vykonáme dostatočne veľa pokusov, absolútny rozdiel medzi aritmetickým priemerom a strednou hodnou rozdelenia bude veľmi malý (mal by sa približovať k nule, avšak nie je potrebné, aby toto približovanie bolo monotónne). Presnejšie povedané, pravdepodobnosť, že

tento rozdiel bude menší ako ľubovoľné malé kladné číslo ε, sa bude pre rastúci počet pokusov blížiť k 1. Ekvivalentne sa dá výraz (4.37) zapísať ako:

$$\forall \varepsilon > 0, \lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0 \qquad (4.38)$$

Dôkaz si ukážeme, keďže využíva už nám známu Chebyshevovu nerovnosť. Vychádzame z toho, že keďže X_i sú nezávislé realizácie z rovnakého rozdelenia pravdepodobnosti (tzv. *iid* z angl. *Independent Identicaly Distributed*), potom vieme, že platí:

$$D\left[X_1 + X_2 + ... + X_n\right] = n\sigma^2 \qquad (4.39)$$

Použitím pravidla $D[aX] = a^2D[X]$, kde a je konštanta ďalej dostaneme:

$$D\left[\overline{X}_n\right] = \frac{\sigma^2}{n} \qquad (4.40)$$

Ďalej použitím Chebyshevovej vety dostávame:

$$\forall \varepsilon > 0, P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{n\varepsilon^2} \qquad (4.41)$$

Pre $n \to \infty$ potom z predchádzajúceho výrazu je priamo vidno, že ak je ε konštanta, potom výraz v pravo s rastúcim n sa bude zmenšovať smerom k 0 (sprava), a teda:

$$\forall \varepsilon > 0, \lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0 \qquad (4.42)$$

Pre spojité rozdelenie pravdepodobnosti je postup obdobný.

Silný zákon veľkých čísel tvrdí:

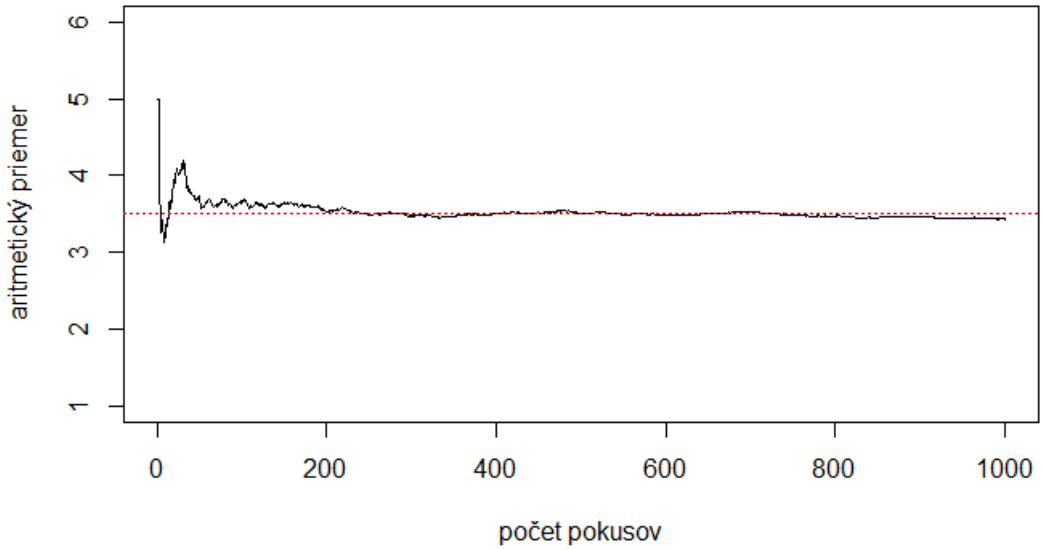
$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1 \qquad (4.43)$$

Inými slovami, $\overline{X}_n$ konverguje takmer isto k μ pre $n \to \infty$. V tomto prípade môžeme zjednodušene vetu vnímať tak, že ak vykonáme nekonečne veľa pokusov, tak pravdepodobnosť, že aritmetický priemer realizácií náhodných pokusov bude totožný so strednou hodnotou rozdelenia z ktorého sa náhodné premenné generujú, bude rovná 1. Intuitívne si rozdiel medzi týmito dvoma formami zákona veľkých čísel môžeme ďalej predstaviť tak, že ak slabý zákon veľkých čísel umožňuje, aby bol rozdiel medzi $\overline{X}_n$ a μ väčší ako $\varepsilon > 0$ pre $n \to \infty$, zákon veľkých čísel tvrdí, že sa tak takmer isto nestane.

Fungovanie tohto zákona si ukážeme na dvoch jednoduchých simuláciách v programe R. V prvom prípade použijeme štandardný pokus hádzania kocky, v druhom použijeme empiricky definované diskrétne rozdelenie pravdepodobnosti.

Vykonajme N = 1000 pokusov hádzania kocky a postupne počítajme aritmetický priemer pre n = 1, 2, ..., 1000, konvergenciu tohto aritmetického priemeru si znázorníme na nasledujúcom obrázku (Obrázok 4.3).

```
> data <- c()
> means <- c()
> for (i in 1:1000) {
+ data <- c(data, sample(1:6, size = 1))
+ means <- c(means, mean(data))
+ }
> plot.ts(means, lwd = 1.5, ylim = c(1, 6), xlab = "počet
  pokusov", ylab = "aritmetický priemer", cex.lab = 1.5,
  cex.axis = 1.3)
> abline(3.5, 0, col = "red", lty = 3)
```



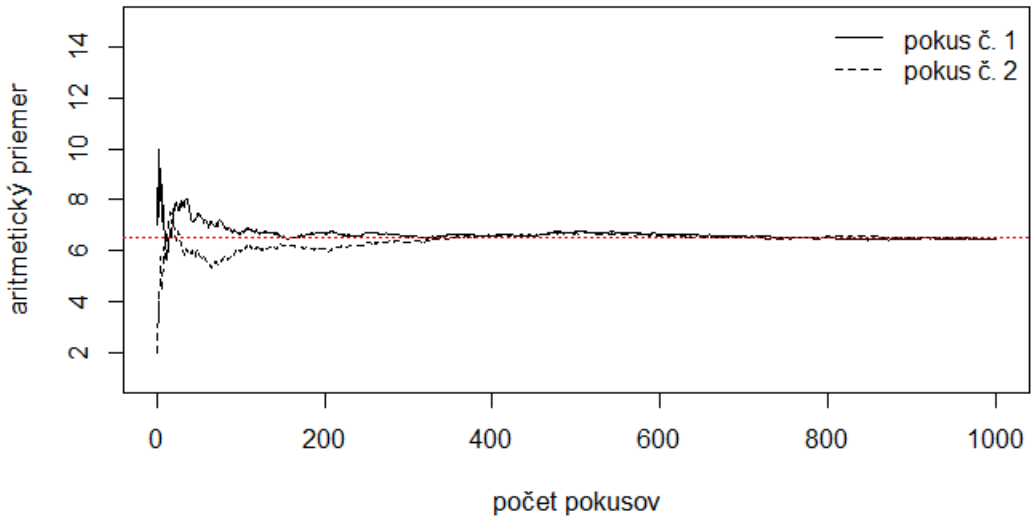
Obrázok 4.3: Približovanie aritmetického priemeru k strednej hodnote – 1. simulácia

Zdroj: vlastné spracovanie v programe R

V druhom prípade vychádzame z toho, že v náhodnom pokuse môže nastať ľubovoľné celé číslo v intervale od 1 do 15 pričom pravdepodobnosť, že náhodná premenná nadobudne hodnotu 2 je 0.3 a pri ostatných hodnotách je táto pravdepodobnosť 0.05. Stredná hodnota je $\mu = 6.5$. Tento pokus sme realizovali dva krát.

```
> pravdepodobnosti <- c(0.05, 0.3, rep(0.05,13))
> premenne <- c(1:15)
> stredna_hodnota <- sum(pravdepodobnosti * premenne)
> data1 <- c(); data2 <- c()
> means1 <- c(); means2 <- c()
> for (i in 1:1000) {
+ data1 <- c(data1, sample(premenne, size = 1, prob =
  pravdepodobnosti))
+ data2 <- c(data2, sample(premenne, size = 1, prob =
  pravdepodobnosti))
+ means1 <- c(means1, mean(data1))
+ means2 <- c(means2, mean(data2))
+ }
> plot(means1, type = "l", lwd = 1.5, ylim = c(1, 15), xlab =
  "počet pokusov", ylab = "aritmetický priemer", cex.lab = 1.1,
  cex.axis = 1)
```

```
> lines(means2, lwd = 1.5, lty = 14)
> abline(stredna_hodnota, 0, col = "red", lty = 3)
> legend("topright", legend = c("pokus č. 1", "pokus č. 2"), bty
  = "n", lty = c(1, 14))
```



Obrázok 4.4: Približovanie aritmetického priemeru k strednej hodnote – 2. simulácia

Zdroj: vlastné spracovanie v programe R

V oboch prípadoch je vidno, ako sa aritmetický priemer blížil s rastom počtu pokusov k strednej hodnote označenej červenou priamkou. Neraz je zaujímavou otázkou skúmať, ako rýchlo sa budú blížiť aritmetické priemery ku skutočnej strednej hodnote.

4.5 Diskrétne rozdelenia pravdepodobnosti

Z nasledujúcich rozdelení pravdepodobností si budeme charakterizovať ich vybrané vlastnosti. Ak to pre dané rozdelenie bude možné, tak nás bude zaujímať:

- Funkcia hustoty rozdelenia pravdepodobnosti[21].
- Parametre funkcie hustoty rozdelenia pravdepodobnosti.
- Kumulatívna distribučná funkcia[22].
- Stredná hodnota.
- Medián.
- Modus.
- Disperzia.

[21] V anglickej literatúre sa zvykne pri diskrétnych rozdeleniach používať pojem *Probability Mass Function* (budeme označovať ako PMF) a pri spojitých *Probability Density Function* (PDF),

[22] Na jej označenie budeme používať skratku CDF (z angl. *Cumulative Distribution Function*).

4.5.1 Bernoulliho rozdelenie pravdepodobnosti

Bernoulliho rozdelenie pravdepodobnosti sa používa na opísanie takých javov, kde po určitej udalosti (napr. pokus) môže nastať jeden z dvoch možných stavov. Zvyčajne ide o typ: úspech / neúspech, kúpil / nekúpil, dobrý výrobok / chybný výrobok a podobne. V takomto prípade môže náhodná premenná X nadobúdať len dve hodnoty: 0 a 1. Ak tieto pokusy sú navzájom nezávislé, hovoríme týmto pokusom Bernoulliho pokusy. Nezávislosť a náhodnosť pokusov (súčasne sa predpokladajú rovnaké podmienky pokusu) je veľmi častým predpokladom rôznych štatistických metód. Aj keď sa tento prípad môže zdať triviálny, Bernoulliho pokus je v skutočnosti situácia, s ktorou sa v rôznych podobách môžeme stretnúť pomerne často.

Tabuľka 7: Tabuľka základných vlastností – Bernoulliho rozdelenie

PRAVDEPODOBNOSTNÁ FUNKCIA		DISTRIBUČNÁ FUNKCIA
$P(X) = p^x(1-p)^{1-x} = \begin{cases} p, & X = 1 \\ (1-p), & X = 0 \end{cases}$		$F(x) = P(X \le x) = \begin{cases} 0, & X < 0 \\ (1-p), & 0 \le X < 1 \\ 1, & X \ge 1 \end{cases}$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = p$		$\hat{\mu} = \begin{cases} 0 & ak\ p < (1-p) \\ 0,1 & ak\ p = (1-p) \\ 1 & ak\ p > (1-p) \end{cases}$
DISPERZIA	$\sigma^2 = p(1-p)$	PARAMETRE p

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 2.3

Podľa worldstat.org (1.12.2011 – údaje za roky 2007-2008) je na Slovensku 48.5 % mužov a 51.5 % žien. Ak by sme sa pýtali, aká je pravdepodobnosť, že ak náhodne vyberieme jednu osobu, tak jej pohlavie bude žena, odpoveď je pomerne jednoduchá aj bez počítania. Pôjde o p = 0.515. Formálne to môžeme overiť použitím pravdepodobnostnej funkcie Bernoulliho rozdelenia pravdepodobnosti. Nastanie javu X = 1 si definujeme ako pohlavie = žena, takže p = 0.515. Po dosadení do vzorca dostaneme:

$$p(X = 1) = 0.515^1(1-0.515)^{1-1} = 0.515$$

4.5.2 Geometrické rozdelenie pravdepodobnosti

Predpokladajme, že realizujeme niekoľko Beurnoulliho pokusov. Ak je pravdepodobnosť, že nastane udalosť v jednom pokuse p, potom geometrické rozdelenie

pravdepodobnosti nám dá odpoveď na otázku, s akou pravdepodobnosťou prvý krát nastane udalosť práve v *x*-tom pokuse (teda po sérii neúspešných pokusov). Uvedený problém sa dá aj „otočiť“ na nenastanie udalosti. Teda s akou pravdepodobnosťou v *x*-tom pokuse udalosť prvý krát nenastane. Budeme uvažovať prvý prípad. Nech X je náhodná premenná, ktorá predstavuje poradie pokusu, pri ktorom nastane takzvaný úspešný pokus. Potom:

Tabuľka 8: Tabuľka základných vlastností – geometrické rozdelenie

PRAVDEPODOBNOSTNÁ FUNKCIA		DISTRIBUČNÁ FUNKCIA
$P(x) = p(1-p)^{x-1},\ x = 1, 2, \ldots$		$F(x) = P(X \le x) = 1-(1-p)^x$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = \frac{1}{p}$	$\tilde{\mu} = \left\lceil \frac{-\log(2)}{\log(1-p)} \right\rceil$	$\hat{\mu} = 1$
DISPERZIA	$\sigma^2 = \frac{(1-p)}{p^2}$	PARAMETRE p

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 2.4

Aká je pravdepodobnosť, že ak budeme náhodne vyberať osoby z populácie Slovenskej republiky, tak v 4-tom pokuse pôjde prvý krát o ženu?

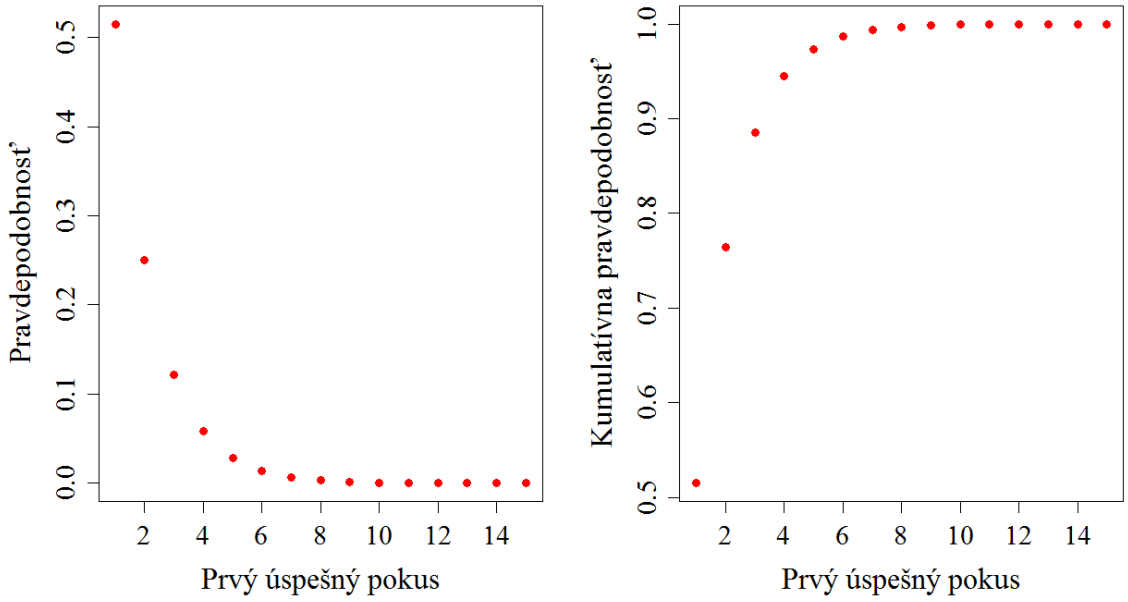
$$p(4) = 0.515(1-0.515)^{4-1} \approx 0.0587$$

V programe R si môžeme funkciu rozdelenia pravdepodobnosti a distribučnú funkciu znázorniť pomocou nasledujúcich kódov. V ďalšom texte budeme používať podobný syntax s tým rozdielom, že si spravidla pomôžeme už existujúcimi funkciami v programe R.

```
> p <- 0.515
> x <- c(1:15)
> xh <- p*(1-p)^(1:15-1)
> data <- data.frame(x, xh)
> par(mfcol = c(1,2))
> plot(data, type= "p", lty = 2, xlab = "Prvý úspešný pokus",
  ylab = "Pravdepodobnosť", pch = 19, col = "red", cex.axis =
  1.5, cex.lab = 1.7)
> xhh <- cumsum(xh)
> data <- data.frame(x, xhh)
> plot(data, type = "p", lty = 2, xlab = "Prvý úspešný pokus",
  ylab = "Kumulatívna pravdepodobnosť", pch = 19, col = "red",
  cex.axis = 1.5, cex.lab = 1.7)
```

Princíp vizualizácie rozdelenia spočíva v jednoduchom výpočte súradníc pre *x*-*y* obrázok. Vektor *x* predstavuje hodnoty na osi *x*-ovej a vektor *xh* príslušné pravdepodobnostné hodnoty vypočítané z uvedenej tabuľky základných vlastností rozdelenia (Tabuľka 8).

Následne tieto body nanesieme do *x-y* grafu pomocou funkcie `plot()`. Rovnaký princíp budeme používať pri tvorbe všetkých obrázkov rozdelení pravdepodobnosti. Obdobne aj pre kumulatívnu distribučnú funkciu, kde príslušné hodnoty kumulatívnej distribučnej funkcie sú vo vektore *xhh*.



Obrázok 4.5: PMF a CDF geometrického rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.5.3 Binomické rozdelenie pravdepodobnosti

Ak realizujeme *n* Bernoulliho pokusov a zaujíma nás pravdepodobnosť, že uspejeme práve *x* krát, potom túto pravdepodobnosť môžeme modelovať pomocou binomického rozdelenia pravdepodobnosti.

Tabuľka 9: Tabuľka základných vlastností – binomické rozdelenie

PRAVDEPODOBNOSTNÁ FUNKCIA		DISTRIBUČNÁ FUNKCIA
$P(x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,1,2,\ldots,n$		$F(x)=P(X\le x)=\sum_{i=0}^{x}\binom{n}{i}p^i(1-p)^{n-i}$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=np$	$\tilde{\mu}=\{\lfloor np\rfloor,\lceil np\rceil\}$	$\hat{\mu}=\lfloor(n+1)p\rfloor$
DISPERZIA	$\sigma^2=np(1-p)$	PARAMETRE p,n

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.10

Uvažujme o teste, kde máme 10 otázok. V každej otázke máme 5 možných odpovedí, ale iba jedna je správna. Ďalej predpokladajme, že sa študent látku nenaučil a o danej problematike vôbec nič nevie. Pravdepodobnosť, že uspeje v jednej otázke je potom p = 0.2. Test úspešne zvládne, ak správne odpovie aspoň na 6 otázok. Aká je pravdepodobnosť, že úspešne zvládne celý test?

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - \sum_{i=0}^{5} \binom{10}{i} 0.2^i (1-0.2)^{10-i} \approx 0.0064$$

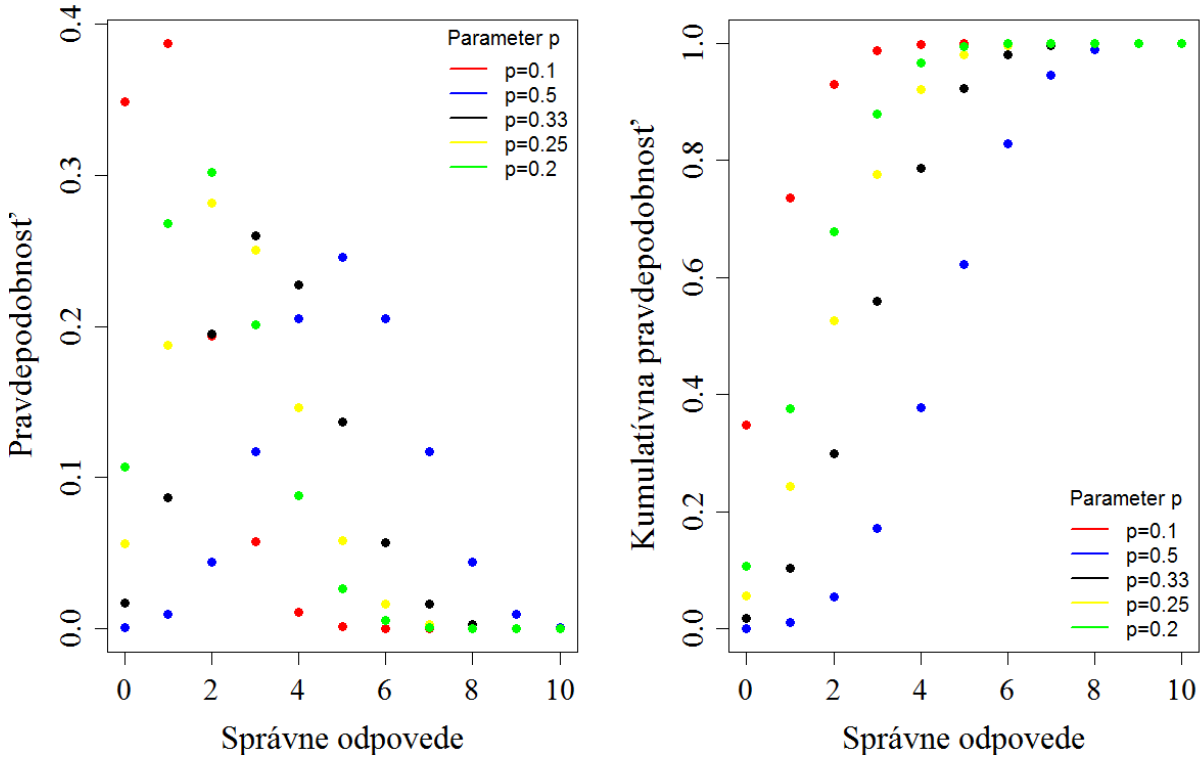
Niektoré často používané rozdelenia pravdepodobnosti je možné v programe R modelovať priamo pomocou už existujúcich funkcií. Binomické rozdelenie k týmto rozdeleniam patrí. Na výpočet pravdepodobnosti vieme použiť funkciu `dbinom()` a na výpočet kumulatívnej distribučnej funkcie `pbinom()`. Pri výpočtoch je tak potrebné poznať iba parametre rozdelenia. V prípade binomického rozdelenia ide o parameter p a n. Predchádzajúci príklad môžeme riešiť nasledovne:

```
> 1 - pbinom(5, 10, prob = 0.2)
[1] 0.006369382
```

Na nasledujúcom obrázku (Obrázok 4.6), sú znázornené rozdelenia pravdepodobnosti, ako aj kumulatívne distribučné funkcie pre rôzne parametre pravdepodobnosti. Princíp vizualizácie je obdobný ako v predošlom príklade.

```
> x <- c(0:10)
> xh <- dbinom(c(0:10), 10, prob = 0.1)
> data <- data.frame(x, xh)
> par(mfcol = c(1, 2))
> plot(data, type = "p", lty = 2, xlab = "Správne odpovede",
  ylab = "Pravdepodobnosť", col = "red", pch = 19, family =
  "serif", cex.axis = 1.5, cex.lab = 1.7)
> col = c("red", "blue", "black", "yellow", "green")
> labels <- c("p = 0.1", "p = 0.5", "p = 0.33", "p = 0.25", "p =
  0.2")
> for (i in 1:4) {points(x, dbinom(c(0:10), 10, prob = 1/(i+1)),
  col = col[i+1], pch = 19)}
> legend("topright", inset = 0.01, title = "Parameter p",
  labels, lwd = 2, col = col, bty = "n")
> xhh <- pbinom(c(0:10), 10, prob = 0.1)
> data <- data.frame(x, xhh)
> plot(data, type = "p", lty = 2, xlab = "Správne odpovede",
  ylab = "Kumulatívna pravdepodobnosť", ylim = c(0,1), pch = 19,
  col = "red", family = "serif", cex.axis = 1.5, cex.lab = 1.7)
> for (i in 1:4) {points(x, pbinom(c(0:10), 10, prob = 1/(i+1)),
  col = col[i+1], pch = 19)}
```

```
> legend("bottomright", inset = 0.01, title = "Parameter p",
  labels, lwd = 2, col = col, bty = "n")
```



Obrázok 4.6: PMF a CDF binomického rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.5.4 Hypergeometrické rozdelenie pravdepodobnosti

Pri binomickom rozdelení pravdepodobnosti sme realizovali niekoľko pokusov a v každom pokuse mohla nastať jedna z dvoch situácií: buď jav nastane alebo nenastane. Pri hypergeometrickom rozdelení pravdepodobnosti sa realizuje jeden pokus, pri ktorom môžeme pozorovať nastanie javu vo viac ako len v jednom prípade. Majme konečnú populáciu N pokusov, u ktorej vieme, že existuje M javov. Z tejto populácie jedným „ťahom“ vyberieme n pokusov. Hypergeometrické rozdelenie opíše pravdepodobnosť počtu úspešných pokusov (nastanie javu) X v jenom „ťahu“.

Tabuľka 10: Tabuľka základných vlastností – hypergeometrické rozdelenie

PRAVDEPODOBNOSTNÁ FUNKCIA		DISTRIBUČNÁ FUNKCIA
$P(x)=\dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$		$F(x)=P(X\le x)=\sum_{i=0}^{x}\dfrac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=n\dfrac{M}{N}$		$\hat{\mu}=\left\lfloor\dfrac{(n+1)(M+1)}{N+2}\right\rfloor$
DISPERZIA	$\sigma^2=\left(\dfrac{N-n}{N-1}\right)n\dfrac{M}{N}\left(1-\dfrac{M}{N}\right)$	PARAMETRE $x,n\in X,\ N=1,2,\ldots$

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.11

Majme výrobnú dávku s 20 súčiastkami, z ktorých predpokladáme, že 3 sú chybné. S odberateľom máme zmluvu, podľa ktorej pri preberaní tovaru odberateľ náhodne skontroluje 5 súčiastok. Ak sú aspoň 2 súčiastky chybné, potom má nárok na reklamáciu celej dávky. Aká je pravdepodobnosť, že dôjde k reklamácii?

$$P(X\ge 2)=1-P(X\le 1)=1-\sum_{i=0}^{1}\frac{\binom{3}{i}\binom{20-3}{5-i}}{\binom{20}{5}}\approx 0.1404$$

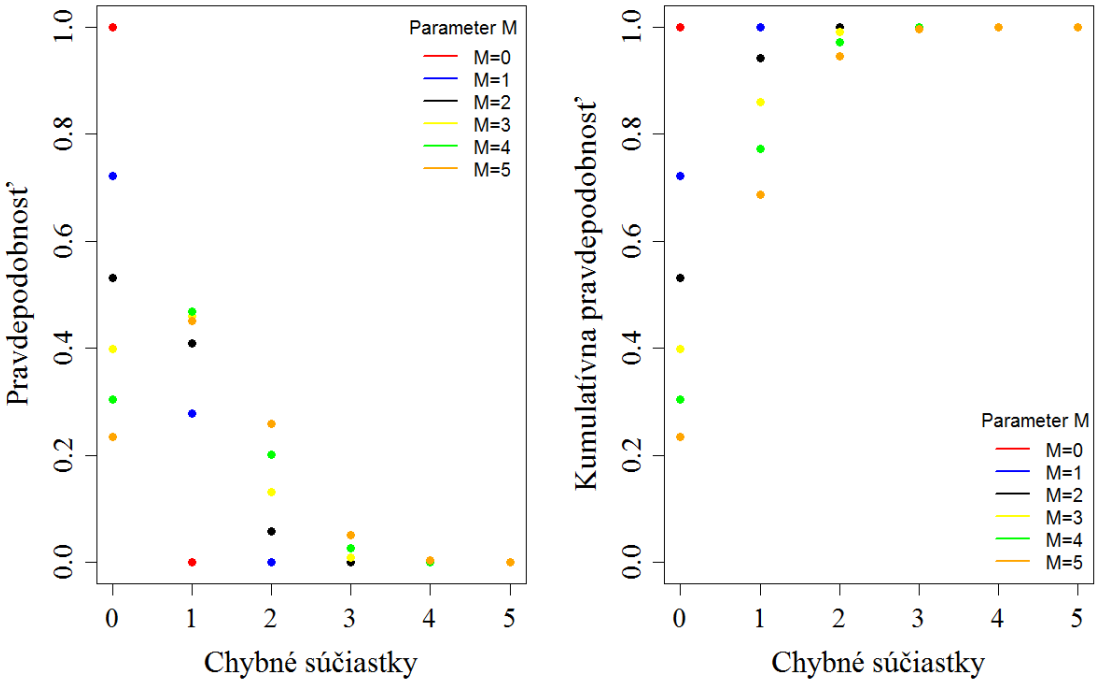
V programe R môžeme použiť funkciu `dhyper()` pre výpočet pravdepodobnosti, prípade `phyper()` pre výpočet hodnôt kumulatívnej distribučnej funkcie. Argumenty funkcie sa však do určitej miery líšia oproti tým, ktoré sme uviedli v predchádzajúcej tabuľke. Obe funkcie uvažujú o parametroch *m*, *n* a *k*, ktorým našej notácii zodpovedajú nasledovné, *m* = *M*, *n* = *N* – *M*, *k* = *n*. Výpočet je potom realizovaný nasledujúcim príkazom:

```
> 1 - phyper(1, 3, 17, 5)
[1] 0.1403509
```

Nasledujúce príkazy korešpondujú vizualizácii rozdelenia pravdepodobnosti a príslušnej kumulatívnej distribučnej funkcii pre *M* = 0, 1, 2, 3, 4, 5.

```
> x <- c(0:5)
> xh <- dhyper(0:5, 0, 17, 5)
> data <- data.frame(x, xh)
> par(mfcol = c(1,2))
```

```
> plot(data, type = "p", lty = 2, xlab = "Chybné súčiastky",
  ylab = "Pravdepodobnosť", pch = 19, col = "red", family =
  "serif", cex.axis = 1.5, cex.lab = 1.7)
> col <- c("red", "blue", "black", "yellow", "green", "orange")
> labels <- c("M=0", "M=1", "M=2", "M=3", "M=4", "M=5")
> for (i in 1:5) points(x, dhyper(0:5, i, 17, 5), col =
  col[i+1], pch = 19)
> legend("topright", inset = 0.01, title = "Parameter M",
  labels, lwd = 2, col = col, bty = "n")
> xhh <- phyper(0:5, 0, 17, 5)
> data <- data.frame(x, xhh)
> plot(data, type = "p", lty = 2, xlab = "Chybné súčiastky",
  ylab = "Kumulatívna pravdepodobnosť", pch = 19, col = "red",
  ylim = c(0,1), family = "serif", cex.axis = 1.5, cex.lab =
  1.7)
> for (i in 1:5) points(x, phyper(0:5, i, 17, 5), col =
  col[i+1], pch = 19)
> legend("bottomright", inset=0.01, title = "Parameter M",
  labels, lwd = 2, col = col, bty = "n")
```



Obrázok 4.7: PMF a CDF hypergeometrického rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.5.5 Rovnomerné rozdelenie pravdepodobnosti

Ide o pomerne jednoduché pravidlo pravdepodobnosti. S rovnomerným rozdelením pravdepodobnosti sme sa stretli v mnohých zmienkach o hádzaní férovej (klasickej) kocky. Každý jav, ktorý môže nastať, môže nastať s rovnakou pravdepodobnosťou. Ak je *a* dolná hranica intervalu, *b* horná hranica intervalu a spolu máme *n* možných javov nachádzajúcich sa

v číselnom intervale $\langle a,b \rangle$, potom základné vlastnosti rovnomerného rozdelenia môžeme vyjadriť tak, ako sú uvedené v nasledujúcej tabuľke.

Tabuľka 11: Tabuľka základných vlastností – rovnomerné diskrétne rozdelenie

PRAVDEPODOBNOSTNÁ FUNKCIA	DISTRIBUČNÁ FUNKCIA
$P(x)=\begin{cases} \frac{1}{n}, & a \le x \le b \\ 0, & ak\ x \notin \langle a,b \rangle \end{cases}$	$F(x)=P(X \le x)=\begin{cases} 0, & k < a \\ \frac{\lfloor x \rfloor - a + 1}{n}, & a \le x \le b \\ 1, & k > b \end{cases}$

STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = \frac{a+b}{2}$	$\tilde{\mu} = \frac{a+b}{2}$	

DISPERZIA	PARAMETRE
$\sigma^2 = \frac{n^2-1}{12}$	a, b, n

Zdroj: upravené podľa zdrojov v použitej literatúre

4.5.6 Poissonovo rozdelenie pravdepodobnosti

S Poissonovým rozdelením pravdepodobnosti sa môžeme v ekonómii stretnúť tiež pomerne často. Modelujeme ním, koľko krát môže nastať jav za určitý časový interval alebo v istej oblasti. Toto rozdelenie je ale pomerne náročné na predpoklady v zmysle, že nie je ľahké rozpoznať, či ho je možné aplikovať v konkrétnej situácii. Obzvlášť dôležité sú nasledujúce dva predpoklady:

- Pravdepodobnosť výskytu javu sa v jednom intervale (oblasti) nemení.
- Pravdepodobnosť výskytu javu v jednom intervale je nezávislá od pravdepodobnosti výskytu udalosti v inom intervale (tieto dva intervaly sa vzájomne neprekrývajú).

Ďalšie dva predpoklady sa nepovažujú za tak dôležité, keďže ich porušenie vedie k nevýrazným (aj keď je to samozrejme relatívny pojem) odchýlkam pri výpočte skutočnej pravdepodobnosti. Ide o:

- Pravdepodobnosť, že nastane jeden jav v určitom intervale (oblasti) je približne proporcionálna k celkovému intervalu (oblasti),
- Pravdepodobnosť výskytu dvoch udalostí v malom intervale je prakticky zanedbateľná.

Poissonovo rozdelenie pravdepodobnosti je možné modelovať pomocou jediného parametra λ, ktorý predstavuje očakávaný počet udalostí za daný časový interval.

Tabuľka 12: Tabuľka základných vlastností – Poissonovo rozdelenie

PRAVDEPODOBNOSTNÁ FUNKCIA		DISTRIBUČNÁ FUNKCIA
$P(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}, x=0,1,2,...,\lambda>0$		$F(x)=P(X\le x)=e^{-\lambda}\sum_{i=0}^{x}\frac{\lambda^{i}}{i!}$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=\lambda$	$\tilde{\mu}=\left\lfloor \lambda+\frac{1}{3}-\frac{0{,}02}{\lambda}\right\rfloor$	$\hat{\mu}=\begin{cases}\lambda-1, & \lambda\in N\\ \lfloor\lambda\rfloor, & \lambda\notin N\end{cases}$
DISPERZIA	$\sigma^2=\lambda$	PARAMETRE λ

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.12

Za posledných 50 nedieľ za sebou prešlo na vybranom úseku cesty v nedeľu 1000 automobilov. Aká je pravdepodobnosť, že v jednu nedeľu prejde práve 50 áut?

$$p(50)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-50}50^{50}}{50!}\approx 0.056$$

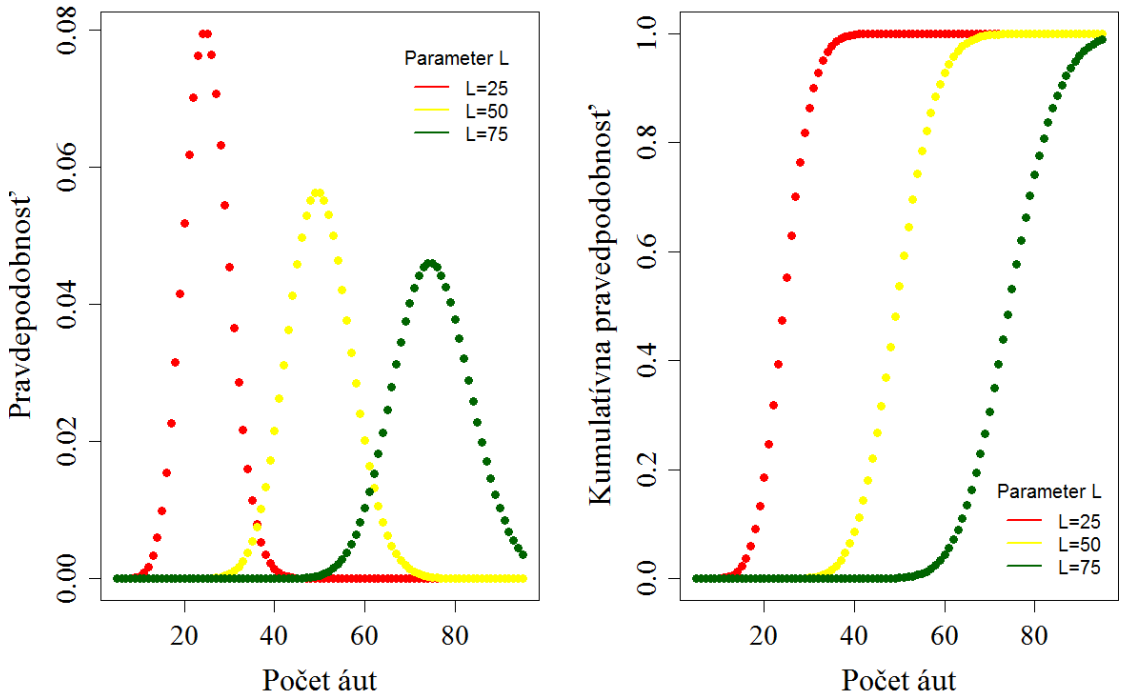
V program R na výpočet pravdepodobnosti môžeme použiť funkciu `dpois()` a na výpočet hodnôt kumulatívnej distribučnej funkcie funkciu `ppois()`. K výsledku z predchádzajúceho príkladu stačí zadať príkaz:

```
> dpois(50, 50)
[1] 0.05632501
```

Pre rôzne parametre λ = 25, 50, 75 je rozdelenie pravdepodobnosti možné zobraziť nasledovne:

```
> x <- c(5:95)
> xh <- dpois(5:95, 25)
> data <- data.frame(x, xh)
> par(mfcol = c(1, 2))
> plot(data, type = "p", lty = 2, xlab = "Počet áut", ylab =
  "Pravdepodobnosť", pch = 19, col = "red", family = "serif",
  cex.axis = 1.5, cex.lab = 1.7)
> col = c("red", "yellow", "darkgreen")
> points(x, dpois(5:95, 50), col = col[2], pch = 19)
> points(x, dpois(5:95, 75), col = col[3], pch = 19)
> labels <- c("L = 25", "L = 50", "L = 75")
> legend("topright", inset = 0.02, title = "Parameter L",
  labels, lwd = 2, col = col, bty = "n")
> xhh <- ppois(5:95, 25)
> data <- data.frame(x, xhh)
> plot(data, type = "p", lty = 2, xlab = "Počet áut", ylab =
  "Kumulatívna pravedpodobnosť", pch = 19, col = "red", family =
  "serif", cex.axis = 1.5, cex.lab = 1.7)
> points(x, ppois(5:95, 50), col =col[2], pch = 19)
> points(x, ppois(5:95, 75), col = col[3], pch = 19)
```

```
> legend("bottomright", inset = 0.02, title = "Parameter L",
  labels, lwd = 2, col = col, bty = "n")
```



Obrázok 4.8: PMF a CDF Poissonovho rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6 Spojité rozdelenia pravdepodobnosti

4.6.1 Rovnomerné spojité rozdelenie pravdepodobnosti

Pokiaľ náhodná premenná môže nadobúdať všetky hodnoty v určitom číselnom intervale a tie sú rovnako pravdepodobné, modelujeme jej rozdelenie pravdepodobnosti pomocou spojitého rovnomerného rozdelenia pravdepodobnosti. V praxi sa spojité rovnomerné rozdelenie môže používať najmä v simuláciách, kde o jednej premennej vieme, že môže nadobudnúť hodnoty z určitého číselného intervalu, ale nevieme rozhodnúť, ktoré hodnoty sú viac a ktoré menej pravdepodobné. Iný pohľad na spojité rozdelenie pravdepodobnosti je, že neponúka analytikovi dokopy žiadnu informáciu okrem intervalu možných hodnôt.

Tabuľka 13: Tabuľka základných vlastností – rovnomerné spojité rozdelenie

HUSTOTA	DISTRIBUČNÁ FUNKCIA
$f(x)=\frac{1}{b-a},\ a\le x\le b, a,b\in R$	$F(x)=\begin{cases} 0, & x\le a \\ \frac{(x-a)}{(b-a)}, & x\in\langle a,b\rangle \\ 1, & x\ge b \end{cases}$

STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=\frac{1}{2}(a+b)$	$\tilde{\mu}=\frac{1}{2}(a+b)$	$\hat{\mu}=\forall x\in\langle a,b\rangle$
DISPERZIA	$\frac{1}{12}(b-a)^2$	PARAMETRE a,b

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.13

Spoločnosť nemá jasnú informáciu o tom, aký bude dopyt po ich výrobku. Predajnú cenu plánujú stanoviť v najhoršom prípade na úroveň celkových nákladov min = 20,- Eur a v najlepšom prípade nepredpokladajú cenu vyššiu ako max = 150,- Eur. Aká je pravdepodobnosť, že cena bude v intervale od 30,- do 120,- Eur?

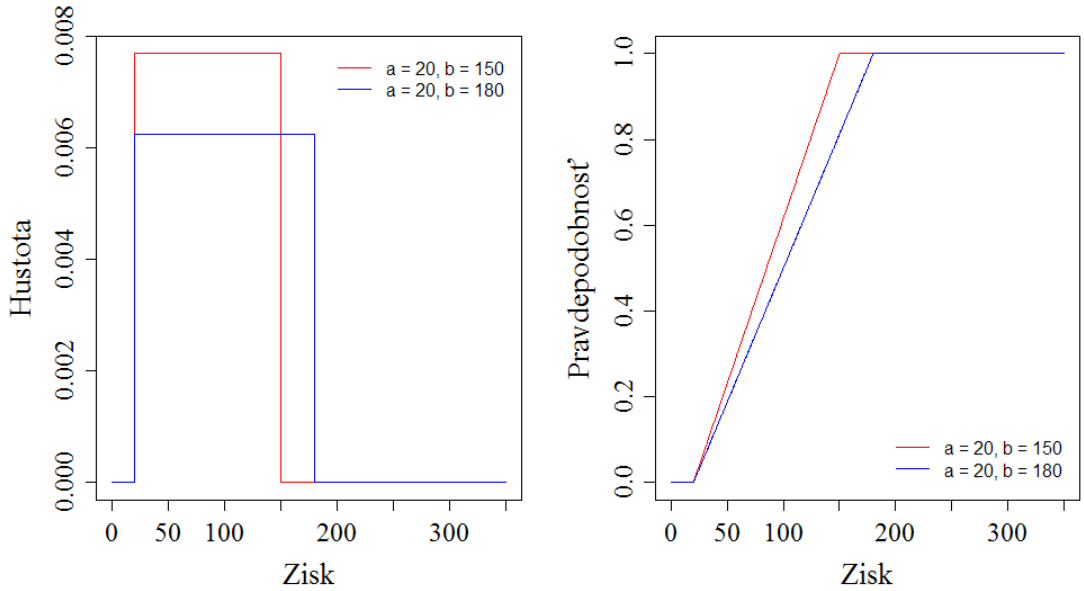
$$F(30\le X\le 120)=F(X\le 120)-F(X\le 30)=\frac{120-20}{150-20}-\frac{30-20}{150-20}=\frac{10}{13}-\frac{1}{13}\approx 0.6923$$

```
> punif(120, min = 20, max = 150) - punif(30, min = 20, max =
  150)
[1] 0.6923077
```

Na nasledujúcom obrázku (Obrázok 4.9) sú uvedené PDF a CDF rovnomerného spojitého rozdelenia pravdepodobnosti pre a = 20, b = 150 a v druhom prípade a = 20, b = 180.

```
> x <- seq(0, 350, length = 1000)
> xh <- dunif(x, min = 20, max = 150)
> data <- data.frame(x, xh)
> par(mfrow = c(1,2))
> plot(data, type = "l", lty = 1, xlab = "Zisk", ylab =
  "Hustota", col = "red", family = "serif", cex.lab = 1.7,
  cex.axis = 1.5)
> lines(x, dunif(x, min = 20, max = 180), lty = 1, col = "blue")
> legend("topright", inset = 0.02, legend = c("a = 20, b = 150",
  "a = 20, b = 180"), lty = 1, col = c("red", "blue"), bty =
  "n")
> xh <- punif(x, min = 20, max = 150)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1, xlab = "Zisk", ylab =
  "Hustota", col = "red", family = "serif", cex.lab = 1.7,
  cex.axis = 1.5)
> lines(x, punif(x, min = 20, max = 180), lty = 1, col = "blue")
```

```
> legend("bottomright", inset = 0.02, legend = c("a = 20, b =
  150", "a = 20, b = 180"), lty = 1, col = c("red", "blue"), bty
  = "n")
```



Obrázok 4.9: PDF a CDF rovnomerného spojitého rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.2 Normálne rozdelenie pravdepodobnosti

Ide o základné rozdelenie s pomerne širokým využitím v praxi (občas sa nazýva aj Gaussovo rozdelenie pravdepodobnosti). Mnoho štatistických metód predpokladá, že hodnoty výberového súboru sa riadia práve normálnym rozdelením pravdepodobnosti. Mnoho javov pozorovateľných v prírode sa riadi normálnym rozdelením. Význam tohto rozdelenia bude zrejme jasnejší pri vysvetlení centrálnej limitnej vety. Nie zriedka sa aj vo výskume stretávame s modelmi, ktoré teoreticky predpokladajú normálne rozdelenie rôznych náhodných premenných – najmä vo finančnej teórii. Bez ohľadu na vhodnosť týchto predpokladov, ide o významné rozdelenie, ktorého objav sa pripisuje Abrahamovi de Moivre. Vysvetlíme si pritom dva druhy rozdelenia: všeobecné normálne rozdelenie pravdepodobnosti a tzv. štandardné (normované) normálne rozdelenie pravdepodobnosti.

Tabuľka 14: Tabuľka základných vlastností – normálne rozdelenie

HUSTOTA		DISTRIBUČNÁ FUNKCIA
$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x,\mu\in R, \sigma>0$		$F(x)=\frac{1}{2}+\frac{1}{2}erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$
STREDNÁ HODNOTA μ	MEDIÁN μ	MODUS μ
DISPERZIA	σ^2	PARAMETRE μ,σ^2

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.14

Fakulta sa rozhodla ponúknuť študentom školské tričká. Aby vedela koľko kusov akých veľkostí má objednať, zaujíma sa o bežné parametre populácie študentov. Zo všetkých študentov sa náhodne vybrala vzorka študentiek a zistilo sa, že priemerná výška študentiek je 164 cm a rozptyl 4 cm^2. Ak vieme normálnym rozdelením pravdepodobnosti opísať výšku študentiek, aká je pravdepodobnosť, že ak náhodne vyberieme jednu študentku, tak jej výška bude menšia ako 160 cm?

$$F(X\le 160)=\frac{1}{2}+\frac{1}{2}erf\left(\frac{164-160}{2\sqrt{2}}\right)=0.0228$$

Pripomenieme, že v predchádzajúcom príklade nemá význam sa pýtať na pravdepodobnosť, že náhodne vyberieme študenta s výškou presne 160 cm, keďže ide o spojitú náhodnú premennú.

Ak uvažujeme o normálnom rozdelení so strednou hodnotou 0 a rozptylom 1, potom hovoríme o štandardnom normálnom rozdelení (označuje sa aj ako normované normálne rozdelenie). Náhodnú premennú *X*, ktorá má normálne rozdelenie, môžeme pomocou jednoduchého vzťahu transformovať na náhodnú premennú so štandardným normálnym rozdelením. Výpočet pravdepodobnosti z hustoty rozdelenia je pomerne komplikovaný. K integrovaniu funkcie hustoty sa používajú numerické metódy, pomocou ktorých vieme nájsť určité približné riešenie. Z tohto dôvodu je výhodné, ak nájdeme čo najpresnejšie riešenie pre štandardné normálne rozdelenie a potom všetky ostatné prípady náhodnej premennej riadiacej sa normálnym rozdelením transformuje na štandardné normálne rozdelenie.

Ak má náhodná premenná *X* normálne rozdelenie $N \sim (\mu, \sigma^2)$, potom náhodná premenná *Z:*

$$Z=\frac{\bar{x}-\mu}{\sigma} \tag{4.44}$$

má štandardné normálne rozdelenie $N \sim (0, 1)$. Vo vzťahoch uvedených v predchádzajúcej tabuľke (Tabuľka 14) stačí nahradiť príslušné parametre hodnotami 0 a 1. Hodnoty distribučnej funkcie štandardného normálneho rozdelenia je možné nájsť v rôznych štatistických tabuľkách.

V programe R sme použili nasledujúci kód pre výpočet predchádzajúceho príkladu:

```
> pnorm(160, mean = 164, sd = 2)
[1] 0.02275013
```

Ak by nás zaujímala hodnota výšky (študentky), pri ktorej je pravdepodobnosť 0.05, že náhodne vybraná študentka bude mať výšku menšiu ako je táto hodnota, mohli by sme použiť kvantilovú funkciu. V programe R týmto funkciám zodpovedajú funkcie rozdelení pravdepodobnosti začínajúce sa na „*q*“, napr.: `qnorm`, `qbinom`, `qpois`,....

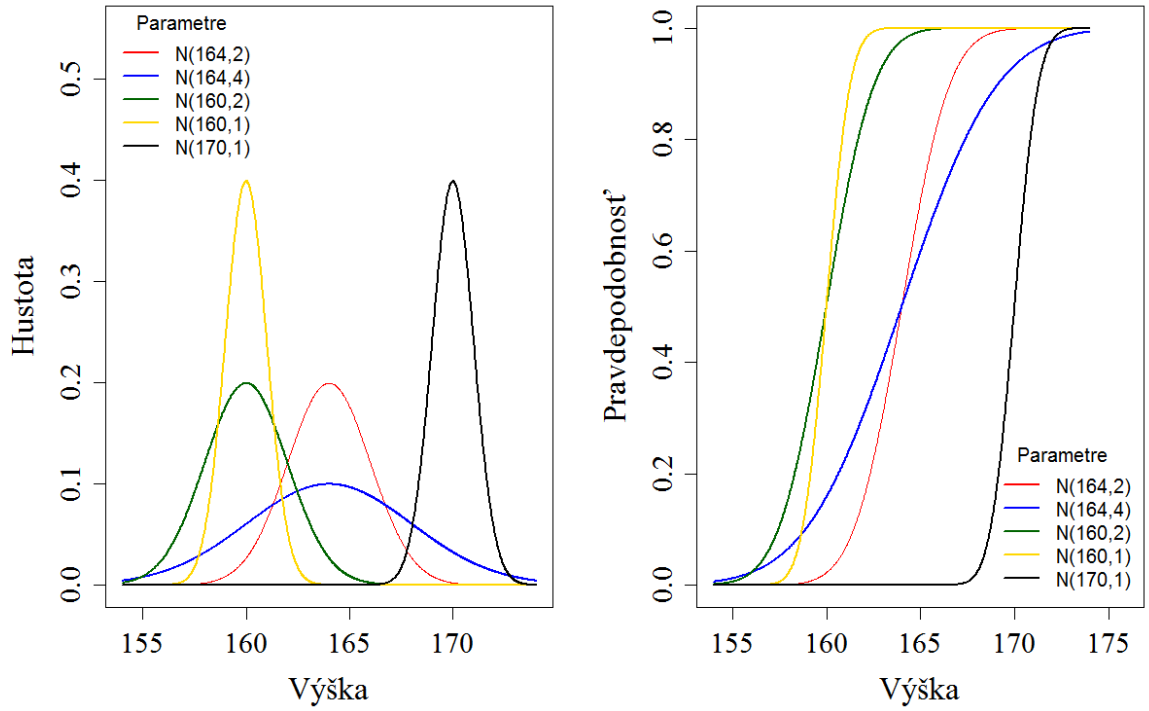
```
> qnorm(0.05, mean = 164, sd = 2)
[1] 160.7103
```

Na nasledujúcom obrázku (Obrázok 4.10), je znázornená funkcia hustoty a kumulatívna distribučná funkcia štandardného normálneho rozdelenia. Pri zobrazovaní spojitého rozdelenia je princíp v programe R obdobný ako v predošlom prípade. Rozdiel spočíva v tom, že jednotlivé body v sústave sú spájané čiarami do polygónu. Ak je tých bodov dostatočne veľa, celkový vizuálny efekt pripomína hladkú krivku.

```
> x <- seq(154, 174, length = 1000)
> xh <- dnorm(x, mean = 164, sd = 2)
> data <- data.frame(x, xh)
> par(mfcol = c(1, 2))
> plot(data, type = "l", lty = 1.1, xlab = "Výška", ylab =
  "Hustota", col = "red", ylim = c(0, 0.55), family = "serif",
  cex.axis = 1.5, cex.lab = 1.7)
> colors <- c("red", "blue", "darkgreen", "gold", "black")
> labels <- c("N(164,2)", "N(164,4)", "N(160,2)", "N(160,1)",
  "N(170,1)")
> lines(x, dnorm(x, mean = 164, sd = 4), lwd = 2, col = "blue")
> lines(x, dnorm(x, mean = 160, sd = 2), lwd = 2, col =
  "darkgreen")
> lines(x,dnorm(x, mean = 160, sd = 1), lwd = 2, col = "gold")
> lines(x,dnorm(x, mean = 170, sd = 1), lwd = 2, col = "black")
> legend("topleft", inset = 0.01, title = "Parametre", labels,
  lwd = 2, col = colors, bty = "n")
> xhh <- pnorm(x, mean = 164, sd = 2)
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1, xlab = "Výška", ylab =
  "Pravdepodobnosť", col = "red", xlim = c(154, 176), family =
  "serif", cex.axis = 1.5, cex.lab = 1.7)
> lines(x, pnorm(x, mean = 164, sd = 4), lwd = 2, col = "blue")
```

```
> lines(x,pnorm(x, mean = 160, sd = 2), lwd = 2, col =
  "darkgreen")
> lines(x,pnorm(x, mean = 160, sd = 1), lwd = 2, col = "gold")
> lines(x,pnorm(x, mean = 170, sd = 1), lwd = 2, col = "black")
> legend("bottomright", inset = 0.01, title = "Parametre",
  labels, lwd = 2, col = colors, bty = "n")
```

Zo zaujímavosti si môžeme vyskúšať spustiť predchádzajúci skript s nasledujúcou zmenou v prvom riadku: `x <- seq(154, 174, length = 10)`.



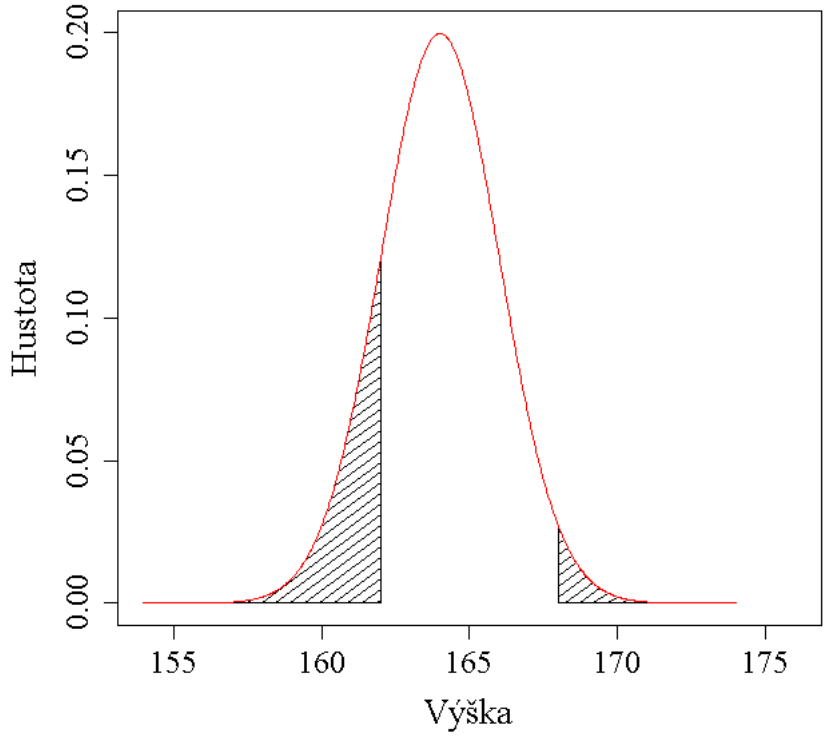
Obrázok 4.10: PDF a CDF normálneho rozdelenia pravdepodobnosti – ukážka 1

Zdroj: vlastné spracovanie v programe R

Pokračujme v predchádzajúcom príklade s tým, že nás bude zaujímať pravdepodobnosť, že v prípade náhodne vybranej študentky bude jej výška v intervale do 162 cm alebo viac ako 168 cm. Riešením je `1 - pnorm(168, mean = 164, sd = 2) + pnorm(162, mean = 164, sd = 2)`. Nasledujúci obrázok (Obrázok 4.11) znázorňuje obsah plochy, ktorý nás v tomto prípade zaujíma:

```
> 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)
> lb1 = 154; ub1 = 162; 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")
```



Obrázok 4.11: PDF normálneho rozdelenia pravdepodobnosti – ukážka 2

Zdroj: vlastné spracovanie v programe R

4.6.3 Centrálna limitná veta

Spomedzi spojitých rozdelení pravdepodobnosti považujeme práve normálne rozdelenie pravdepodobnosti za najdôležitejšie a najpoužívanejšie. Jednak je mnoho iných rozdelení odvodených od normálneho rozdelenia, jednak niektoré významné diskrétne rozdelenia pravdepodobnosti konvergujú k normálnemu rozdeleniu a v neposlednom rade, mnoho javov v prírode sa akoby správalo podľa normálneho rozdelenia pravdepodobnosti. Jeho význam je zrejme najjednoznačnejšie vidno pri centrálnej limitnej vete. V definícii centrálnej limitnej vety sa stretneme s pojmami ako populácia a náhodne vybraný súbor hodnôt (vzorka). Na teraz vystačíme s predstavou, že populácii zodpovedajú všetky možné pozorovania nejakého javu a pod náhodným výberom iba náhodne vybrané pozorovania z populácie.

Centrálna limitná veta hovorí o asymptotickom rozdelení pravdepodobnosti náhodnej premennej Z a T (pozri vzťahy nižšie). Majme náhodné premenné X_i, $i = 1, 2, ..., n$, ktoré sú

nezávislé realizácie z rovnakého rozdelenia pravdepodobnosti so strednou hodnotou μ a s konštantným rozptylom σ^2. Potom pre $n \to \infty$ má náhodná premenná Z:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{4.45}$$

štandardné normálne rozdelenie pravdepodobnosti (čiže so strednou hodnotou $\mu = 0$ a s rozptylom $\sigma^2 = 1$). Upriamme však teraz pozornosť na náhodnú premennú T:

$$T = \sum_{i=1}^{n} x_i \tag{4.46}$$

ktorá má normálne rozdelenie pravdepodobnosti so strednou hodnotou $n\mu$ a rozptylom $n\sigma^2$.

To čo centrálna limitná veta vlastne tvrdí, že aritmetický priemer z výberov o rozsahu n sa riadi normálnym rozdelením so strednou hodnotou μ a rozptylom σ^2/n. Pokúsime sa túto vetu vysvetliť na jednoduchom príklade hádzania kocky. Vykonajme s férovou kockou nasledujúci pokus. Hoďme kocku 10 krát za sebou a sčítajme všetky čísla, ktoré padli tak, ako je to vo vzťahu náhodnej premennej T. Potom pokus zopakujme povedzme 1000 krát. To čo zistíme je, že rozdelenie početnosti pomerne dobre aproximuje normálne rozdelenie pravdepodobnosti, ktoré je charakteristické tým, že hodnoty v „strede“ sú častejšie ako extrémne hodnoty. V prípade kocky, môžeme v 10-tich pokusoch nahádzať maximálne 60 bodov a minimálne 10 bodov. Priemerná bodová hodnota by mala byť 35 bodov. Ak vykonáme spomínaných 1000 pokusov, tak zistíme, že naozaj najviac hodnôt sa bude sústrediť okolo tejto hodnoty a extrémne vysoké a nízke súčty budú zriedkavé. Dôvod pre túto skutočnosť je fakt, že hodnotu 35 dostaneme viacerými spôsobmi ako hodnotu 60 alebo 10. K tomu, aby sme mohli dostať hodnotu 60, potrebujeme 10 krát za sebou hodiť číslo 6, čo je pomerne málo pravdepodobné (p je veľmi malé číslo, presnejšie $p = 0.00000001654$). Naproti tomu, hodnotu 35 môžem dosiahnuť viacerými spôsobmi. Pre zjednodušenie si predstavte, že máme iba 2 hody. Pri súčte je min. 2 a max. 12 a oba vieme dosiahnuť iba 1 spôsobom. Hodnotu 7 môžeme dosiahnuť nasledujúcimi spôsobmi: 1+6, 2+5, 3+4 a ešte v opačnom poradí. Táto vlastnosť je veľmi zaujímavá a našťastie je možné ju využiť pri modelovaní mnohých javov v prírode. Preto sa považuje normálne rozdelenie pravdepodobnosti za tak významné.

To dôležité, čo centrálna limitná veta vysvetľuje, si zhrnieme do nasledujúcich bodov:

- Priemer výberového súboru (výberový priemer) je sám o sebe náhodnou premennou s vlastným rozdelením pravdepodobnosti.

- Priemer z výberového súboru (výberový priemer) má menšiu variabilitu ako jednotlivé pozorovania z výberového súboru.
- Smerodajná odchýlka priemeru z výberového súboru (výberový priemer) je $s/\sqrt{n}$.
- Najväčšou prekážkou pre študentov býva pochopiť, že nie len samotné hodnoty z výberového súboru majú rozdelenie pravdepodobnosti, ale aj charakteristiky, ako napr. priemer, ktoré z tohto výberového súboru počítame.

Spomenieme, že existuje niekoľko verzií centrálnej limitnej vety, z ktorých niektoré za určitých predpokladov pripúšťajú, aby náhodné premenné X_i neboli z rovnakého rozdelenia. Týmto verziám sa v tejto publikácii bližšie venovať nebudeme. Taktiež z centrálnej limitnej vety je možné ukázať, ako normálne rozdelenie pravdepodobnosti za určitých predpokladov aproximuje určité diskrétne rozdelenia: binomické, hypergeometrické a Poissonovo. Ani tejto problematike sa však bližšie venovať nebudeme.

Doteraz sme používali priamo funkcie hustoty a distribučné funkcie rozdelení pravdepodobnosti. V empirických aplikáciách ekonometrických metód, prípadne v manažmente rizika, sa môžeme stretnúť s potrebou simulovania určitých hodnôt s dopredu definovaného rozdelenia pravdepodobnosti. Niektoré takéto aplikácie si budeme v priebehu tohto textu ukazovať. Na tomto mieste využijeme centrálnu limitnú vetu, aby sme si ukázali jednoduché používanie niektorých funkcií v programe R, ktoré umožňujú generovanie náhodných (presnejšie pseudonáhodných) čísel.

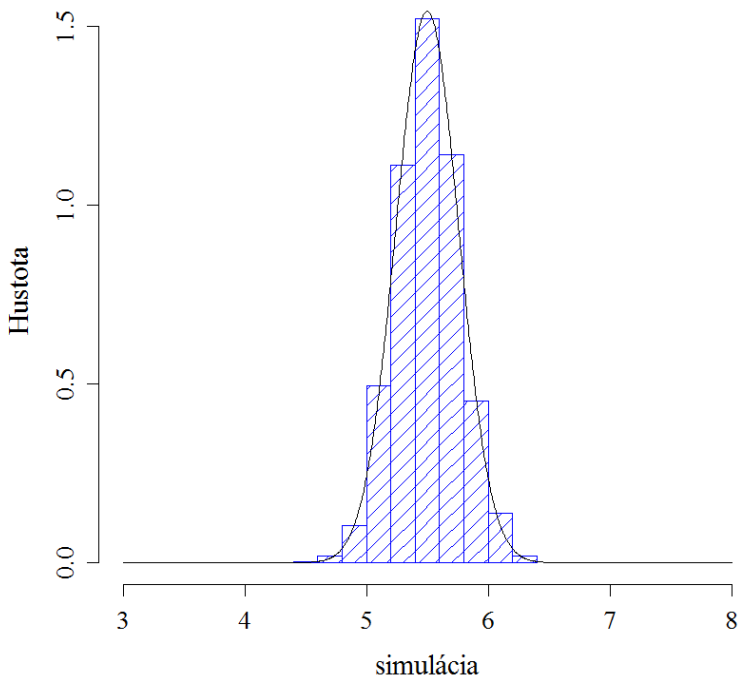
Pokúsme sa tvrdenie centrálnej limitnej vety overiť pomocou simulácií. Centrálna limitná veta predpovedá, že ak zoberieme súbor náhodných premenných z ľubovoľného rozdelenia pravdepodobnosti, potom stredná hodnota z tohto súboru sa bude riadiť normálnym rozdelením so stredou hodnotou μ a rozptylom σ^2/n. V predošlej časti sme si predstavili niekoľko rozdelení. Uvedené tvrdenie sa pokúsime overiť nasledovným spôsobom:

- Vygenerujeme vzorku o veľkosti $n > 30$ z ľubovoľného rozdelenia pravdepodobnosti.
- Vypočítame si aritmetický priemer a hodnotu si uložíme do vektora `means <- c()`.
- Uvedený postup zopakujeme 5000 krát.
- Zobrazíme histogram priemerov a preložíme ním funkciu normálneho rozdelenia.
- Vypočítame si aritmetický priemer a rozptyl z vektora priemerov.

Postup vyskúšame najprv na rovnomernom rozdelení pravdepodobnosti, $X \sim U(3, 8)$, kde sme zvolili $n = 31$. Priemer priemerov nám pri prvej simulácii vyšiel 5.502 a rozptyl priemerov 0.068. Nasledujúci obrázok predstavuje histogram priemerov zostavený tak, aby

obsah stĺpcov bol rovný 1. Zároveň sme na tento histogram naniesli hustotu normálneho rozdelenia so strednou hodnotou (8 + 3) / 2 = 5.5 a rozptylom $((8 – 3)^2 /12) / 31 = 0.067$. Tieto hodnoty sa zvolili nasledovne. Stredná hodnota spojitého rovnomerného rozdelenia pravdepodobnosti je (Tabuľka 13) $(b – a) / 2$, kde b je horná hranica intervalu a a je dolná hranica intervalu, na ktorom je rovnomerné rozdelenie definované. Rozptyl hodnôt pochádzajúcich z takto definovaného rozdelenia sa vypočíta zo vzťahu $(b – a)^2 / 12$. Keďže nás však zaujíma rozptyl strednej hodnoty, vychádzajúc zo vzťahu (4.45) je zrejmé, že uvedený výraz ešte musíme vydeliť počtom pozorovaní, a teda $((b – a)^2 / 12) / n$. Všimnime si, že simulované výsledky pomerne presne odhadujú skutočnú strednú hodnotu a rozptyl strednej hodnoty.

```
> means <- c()
> variances <- c()
> for (i in 1:5000) {
+ simulated <- runif(n = 31, min = 3, max = 8)
+ means <- c(means, mean(simulated))
+ variances <- c(variances, var(simulated))
+ }
> hist(means, density = 10, col = "blue", main = NA, cex.lab =
  1.5, xlim = c(3,8), cex.axis = 1.3, freq = FALSE, ylab =
  "Hustota", family = "serif", xlab = "simulácia")
> x <- seq(3, 8, length = 1000)
> xh <- dnorm(x, mean = 5.5, sd = 0.067^0.5)
> data <- data.frame(x, xh)
> lines(data, type = "l")
```



Obrázok 4.12: Histogram simulovaných priemerov a PDF normálneho rozdelenia

Zdroj: vlastné spracovanie v programe R

V praxi nemáme možnosť simulovať z dopredu zvoleného rozdelenia, keďže neraz nepoznáme ani rozdelenie, a teda ani jeho parametre. Spravidla to vyzerá tak, že máme iba jednu vzorku a následne na základe nej máme odhadnúť rozdelenie hodnôt. Preto vykonajme ešte jednu iteráciu, v ktorej náhodne vyberieme $n = 31$ pozorovaní z rovnakého rozdelenia ako v predošlej simulácii.

Odhad na základe jednej vzorky bol v našom prípade 5.317 pre strednú hodnotu a 0.068 pre rozptyl priemerov (`vzorka_1 <- runif(n = 31, min = 3, max = 8); mean(vzorka_1); var(vzorka_1)/31`). Uvedené hodnoty sú pomerne blízko k hodnotám, ktoré sme získali z 5000 iterácií. Výsledky simulácií tak môžeme porovnať s teoretickými hodnotami, aby sme zistili, nakoľko vieme získať presný údaj o strednej hodnote pri vzorke o veľkosti $n = 31$.

Zdôrazníme, že centrálna limitná veta hovorí o asymptotickej vlastnosti strednej hodnoty, takže existencia odchýlok od prípadnej teoretickej hodnoty je prípustná. Spravidla je podstatné, aby sme používali také štatistické metódy, pomocou ktorých budú naše odhady konvergovať čo najrýchlejšie (napr. v závislosti od veľkosti vzorky) k skutočným teoretickým hodnotám. Tejto problematike je venovaný väčší priestor v publikáciách venujúcich sa induktívnej štatistike.

V predchádzajúcej simulácii, ako aj vo všetkých ostatných simuláciách, nie je možné výsledky zreprodukovať s úplnou presnosťou, čo je spôsobené samotnou podstatou generovania pseudonáhodných čísel. Rozdiely by však nemali byť veľké. Pri ďalšej simulácii si postup zopakujeme pre podstatne menšiu vzorku. Predchádzajúca simulácia uvažovala o vzorke s veľkosťou $n = 31$. V nasledujúcom prípade zvolíme $n = 10$ a porovnáme presnosť výsledkov.

Po 5000 iteráciách a vzorke $n = 10$, bola priemerná hodnota priemerov 5.493, čo je stále pomerne blízko k skutočnej strednej hodnote (t.j. 5.5) a rozptyl priemerov 0.209. Teoreticky je pri danej veľkosti vzorky možné očakávať rozptyl priemerov $((8 - 3)^2 /12) / 10 = 0.208$. Rozdiely sa nejavia ako výrazné. Problém však spočíva v tom, že významnosť veľkosti vzorky závisí aj od rozdelenia, z ktorého sa údaje generujú. Miera dôležitosti veľkosti vzorky sa pred samotným experimentom dá v empirickom výskume len zriedkakedy určiť. Závisí totiž od mnohých parametrov, ktoré nám dopredu nie sú známe. Pri našom experimente sme vedeli, že generujeme vzorku z rovnomerného rozdelenia. V praxi to vedieť nemusíme. Preto platí pravidlo: čím väčšia vzorka, o to lepšie štatistické výsledky vieme získať. Problémy vznikajú v situáciách, keď je získanie väčšieho počtu pozorovaní veľmi nákladné.

4.6.4 Trojuholníkové rozdelenie pravdepodobnosti

V ekonómii, obzvlášť v manažmente, je trojuholníkové rozdelenie pravdepodobnosti jedno z najpoužívanejších. S jeho použitím sa môžeme stretnúť v mnohých aplikáciách projektového manažmentu, manažmente rizika, pri analýze metód rozhodovania sa a mnoho iných. Konkrétnejšie, trojuholníkové rozdelenie pravdepodobnosti sa hodí na modelovanie situácií, kde vieme povedať najpravdepodobnejší, najhorší a najlepší scenár. Zvyčajne ak nemáme iné ako tieto informácie, trojuholníkové rozdelenie pravdepodobnosti je zaujímavou alternatívou. Nech a je minimálna hodnota, b je maximálna hodnota a c je najpravdepodobnejšia hodnota.

Tabuľka 15: Tabuľka základných vlastností – trojuholníkové rozdelenie

HUSTOTA	DISTRIBUČNÁ FUNKCIA
$f(x)=\begin{cases} 0, & x \le a \\ \dfrac{2(x-a)}{(b-a)(c-a)}, & a < x \le c \\ \dfrac{2(b-x)}{(b-a)(b-c)}, & c \le x < b \\ 0, & x \ge b \end{cases} ; a < c < b \in R$	$F(x)=\begin{cases} 0, & x \le a \\ \dfrac{(x-a)^2}{(b-a)(c-a)}, & a < x \le c \\ 1-\dfrac{(b-x)^2}{(b-a)(b-c)}, & c \le x < b \\ 1, & x \ge b \end{cases}$

STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=(a+b+c)/3$	$\tilde{\mu}=\begin{cases} a+\sqrt{\dfrac{(b-a)(c-a)}{2}}, & c \ge \dfrac{(b-a)}{2} \\ b-\sqrt{\dfrac{(b-a)(b-c)}{2}}, & c \le \dfrac{(b-a)}{2} \end{cases}$	$\hat{\mu}=c$

DISPERZIA	PARAMETRE
$\dfrac{a^2+b^2+c^2-ab-ac-bc}{18}$	a,b,c

Zdroj: upravené podľa zdrojov v použitej literatúre

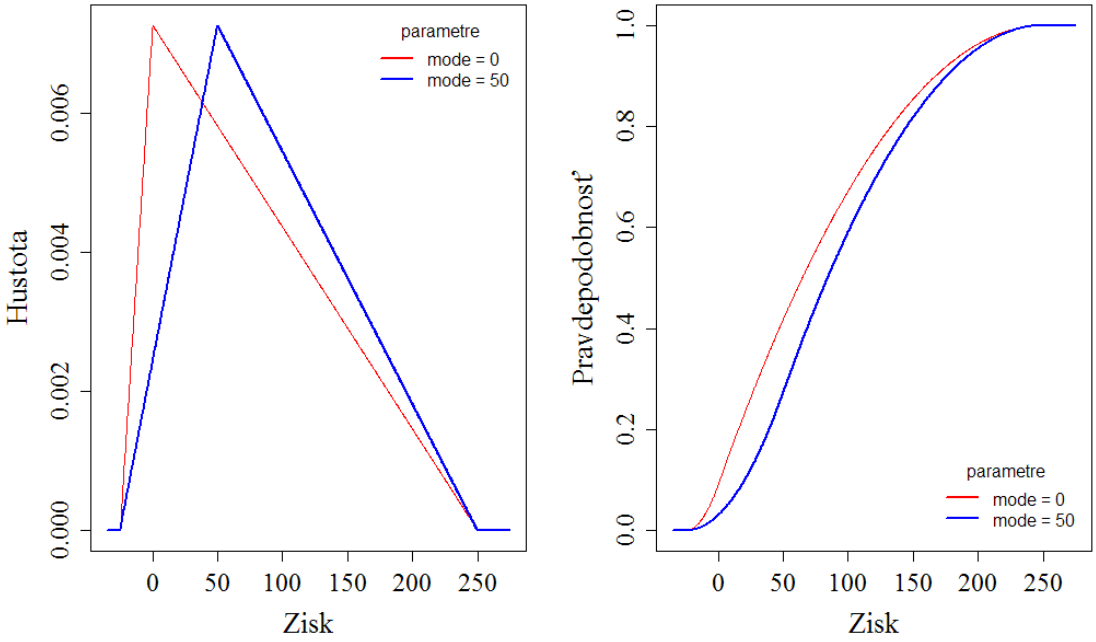
Príklad 4.15

Projekt môže spoločnosti priniesť v najhoršom prípade stratu 25 tis. EUR a v najlepšom prípade zisk 250 tis. EUR. Najpravdepodobnejším scenárom je nulový zisk. Vedúci projektu sa rozhodol ziskový profil modelovať trojuholníkovým rozdelením pravdepodobnosti. Aká je pravdepodobnosť, že zisk bude > 0?

```
> library(triangle)
> 1 - ptriangle(0, a = - 25, b = 250, c = 0)
[1] 0.9090909
```

Na nasledujúcom obrázku je PDF a CDF trojuholníkového rozdelenia pravdepodobnosti pre modálnu hodnotu c = 0 a c = 50.

```
> x <- seq(-35, 275, length  =1000)
> xh <- dtriangle(x, a = -25, b = 250, c = 0)
> par(mfrow = c(1,2))
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.5, xlab = "Zisk", ylab =
  "Hustota",col = "red", cex.lab = 1.7, cex.axis = 1.5, family =
  "serif")
> lines(x, dtriangle(x, a = -25, b = 250, c = 50), lwd = 2,col =
  "blue")
> legend("topright", inset = 0.02, title = "parametre", legend =
  c("mode = 0","mode = 50"), lwd = 2,col = c("red", "blue"), bty
  = "n")
> xh <- ptriangle(x, a = -25, b = 250, c = 0)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.5, xlab = "Zisk", ylab =
  "Hustota", col = "red", cex.lab = 1.7, cex.axis = 1.5, family
  = "serif")
> lines(x, ptriangle(x, a = -25, b = 250, c = 50), lwd = 2, col
  = "blue")
> legend("bottomright", inset = 0.02, title = "parametre",
  legend = c("mode = 0","mode = 50"), lwd = 2, col = c("red",
  "blue"), bty = "n")
```



Obrázok 4.13: PDF a CDF trojuholníkového rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.5 Exponenciálne rozdelenie pravdepodobnosti

Poissonovo rozdelenie pravdepodobnosti modeluje pravdepodobnosť výskytu určitého počtu javov za nejaký časový interval (prípadne v nejakej oblasti). Exponenciálne rozdelenie pravdepodobnosti modeluje časový interval medzi takýmito dvoma Poissonovskými javmi. Obe rozdelenia sa používajú pri tzv. Poissonovom procese. Poissonovský proces má niekoľko

dôležitých predpokladov, ktoré sme popísali pri Poissonovom rozdelení pravdepodobnosti. Udalosti, ktoré sú predmetom nášho záujmu sa môžu vyskytnúť v ľubovoľnom čase sledovaného intervalu a výskyt jednej udalosti je nezávislý na predošlej udalosti. Upozorňujeme, že neraz nevieme zabezpečiť dodržanie týchto predpokladov. Pri interpretácii je potrebné v týchto situáciách brať na túto skutočnosť ohľad. Ak je meraná udalosť časové rozpätie hovorov na zákazníckej linke, zrejme existujú určité trendy, kedy sa zákazníci s touto linkou kontaktujú častejšie (napr. v priebehu obeda) a kedy menej (nad ránom). Pravdepodobnosť kontaktu so zákazníckou linkou tak nie je konštantná. Na druhej strane je aspoň približne konštantná v určitom časovom intervale (napr. počas jednej hodiny).

Podobne ako pri Poissonovom rozdelení pravdepodobnosti, aj v tomto prípade nám stačí na popísanie sledovaného procesu (časový výskyt medzi dvoma javmi) jeden parameter λ, ktorý si vieme vypočítať zo strednej hodnoty náhodnej premennej. Parameter λ by sme mohli interpretovať ako „koľko krát nastane udalosť za jednotku času“.

Tabuľka 16: Tabuľka základných vlastností – exponenciálne rozdelenie

HUSTOTA		DISTRIBUČNÁ FUNKCIA
$f(x)=\lambda e^{-\lambda x},\ x\geq 0,\ \lambda>0$		$F(x)=1-e^{-\lambda x},\ x\geq 0$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=1/\lambda$	$\tilde{\mu}=\frac{\ln(2)}{\lambda}$	$\hat{\mu}=0$
DISPERZIA	$1/\lambda^2$	PARAMETRE λ

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.16

Za posledných 50 nedieľ za sebou prešlo na vybranom úseku cesty v nedeľu 1000 automobilov. Aká je pravdepodobnosť, že medzi dvoma za sebou idúcimi autami bude menej ako 1 hodina času?

Z takto zadaného zadania vyplýva, že na jednu nedeľu pripadá 20 áut a na jednu hodinu teda 20/24 áut za hodinu. To je stredná hodnota času medzi dvoma javmi:

$$\mu=1000/50/24=0.833\wedge\mu=1/\lambda,\ \lambda=1/\mu,\ \lambda=1/0.833=1.2$$

$$F(X<1)=1-e^{-\lambda x}=1-e^{-1.2(1)}\approx 0.6988$$

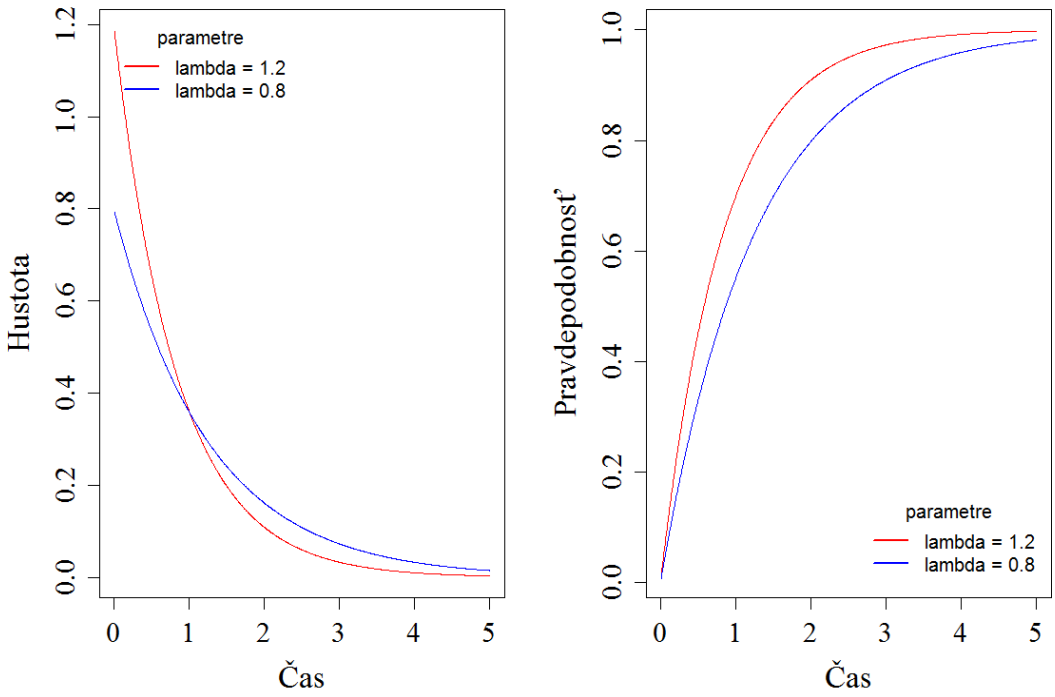
Ak sme subjekt, ktorý rozhoduje o rozšírení cesty, môže nás zaujímať, s akou pravdepodobnosťou bude čas medzi prejazdom dvoch áut viac ako 3 hodiny, prípadne s pravdepodobnosťou 0.95, aký bude časový interval (v hodinách) medzi dvoma za sebou nasledujúcimi autami na ceste.

$$F(X>3)=1-F(X\le 3)=e^{-1.2(3)}\approx 0.0273$$

$$F(X\le x)=0.95;1-e^{-1.2(x)}=0.95;\ -\frac{\ln(1-0.95)}{1.2}\approx 2.4964$$

Na nasledujúcom obrázku (Obrázok 4.14) je PDF a CDF exponenciálneho rozdelenia pravdepodobnosti pre λ = 1.2 a v druhom prípade pre λ = 0.8.

```
> x <- seq(0.01, 5, length = 1000)
> xh <- dexp(x, rate = 1.2)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.2, xlab = "Čas", ylab =
  "Hustota", col = "red", cex.lab = 1.7, cex.axis = 1.5, family
  = "serif")
> lines(x, dexp(x, rate = 0.8), lty = 1.2, col = "blue")
> legend("topleft", inset = 0.02, title = "parametre", legend =
  c("lambda = 1.2", "lambda = 0.8"), lwd = 2, col = c("red",
  "blue"), bty = "n")
> x <- seq(0.01, 5, length = 1000)
> xh <- pexp(x, rate = 1.2)
> data = data.frame(x, xh)
> plot(data, type = "l", lty = 1.2, xlab = "Čas", ylab =
  "Pravdepodobnosť", col = "red", cex.lab = 1.7, cex.axis = 1.5,
  family = "serif")
> lines(x, pexp(x, rate = 0.8),lty = 1.2, col = "blue")
> legend("bottomright", inset = 0.02, title = "parametre",
  legend = c("lambda = 1.2", "lambda = 0.8"), lwd = 2, col =
  c("red", "blue"), bty = "n")
```



Obrázok 4.14: PDF a CDF exponenciálneho rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.6 Lognormálne rozdelenie pravdepodobnosti

Hovoríme, že náhodná premenná X má lognormálne rozdelenie pravdepodobnosti, pokiaľ platí, že $ln\ X$ má normálne rozdelenie pravdepodobnosti. Z toho je zrejme jasné, že aj vzťahy pre lognormálne rozdelenie pravdepodobnosti budú vychádzať z normálneho rozdelenia pravdepodobnosti. V ekonómii ide najčastejšie o modelovanie výnosov, rastov, rôznych finančných a iných podnikových ukazovateľov, ktoré spravidla nemôžu nadobúdať hodnoty menšie ako 0 (vrátane).

Nech W je náhodná premenná, ktorá sa riadi normálnym rozdelením pravdepodobnosti a jej parametre sú μ_W a σ_W^2, potom $X = e^W$, sa riadi lognormálnym rozdelením s nasledujúcimi vlastnosťami:

Tabuľka 17: Tabuľka základných vlastností – lognormálne rozdelenie

HUSTOTA		DISTRIBUČNÁ FUNKCIA
$f(x) = \frac{1}{\sigma_W x\sqrt{2\pi}} e^{-\left(\frac{(\ln x - \mu_W)^2}{2\sigma_W^2}\right)}, x, \sigma_W > 0\, a\, \mu_W \in R$		$F(x) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln x - \mu_W}{\sigma_W \sqrt{2}}\right)$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = e^{\mu_W + \frac{\sigma_W^2}{2}}$	$\tilde{\mu} = e^{\mu_W}$	$\hat{\mu} = e^{\mu_W - \sigma_W^2}$
DISPERZIA	$\left(e^{\sigma_W^2} - 1\right) e^{2\mu_W + \sigma_W^2}$	PARAMETRE μ_w, σ_w^2

Zdroj: upravené podľa zdrojov v použitej literatúre

Príklad 4.17

Nevysvetlené straty vo výrobe sa evidujú určitým podielom z celkového množstva vyrobených výrobkov. Uvedený podiel je pre nás náhodná premenná. Zrejme nemôže nadobudnúť záporné hodnoty a predpokladáme, že má lognormálne rozdelenie pravdepodobnosti. Namerané hodnoty za posledných 20 rokov sú nasledujúce:

0.0003, 0.0002, 0.003, 0.0005, 0.004, 0.002, 0.01, 0.0005, 0.0007, 0.003, 0.0004, 0.0008, 0.001, 0.0002, 0.0006, 0.003, 0.003, 0.0006, 0.002, 0.009.

Za predpokladu, že sa v budúcnosti vzor správania týchto strát nezmení, aká je pravdepodobnosť, že budúci rok sa z výroby stratí viac ako 0.005 podielu výrobkov?

Vypočítame prirodzené logaritmy nameraných hodnôt, ktoré by sa mali riadiť normálnym rozdelením pravdepodobnosti:

–8.12, –8.52, –5.81, –7.61, –5.53, –6.22, –4.61, –7.61, –7.27, –5.81, –7.83, –7.14, –6.91, –8.52, –7.42, –5.81, –5.81, –7.42, –6.22, –4.72.

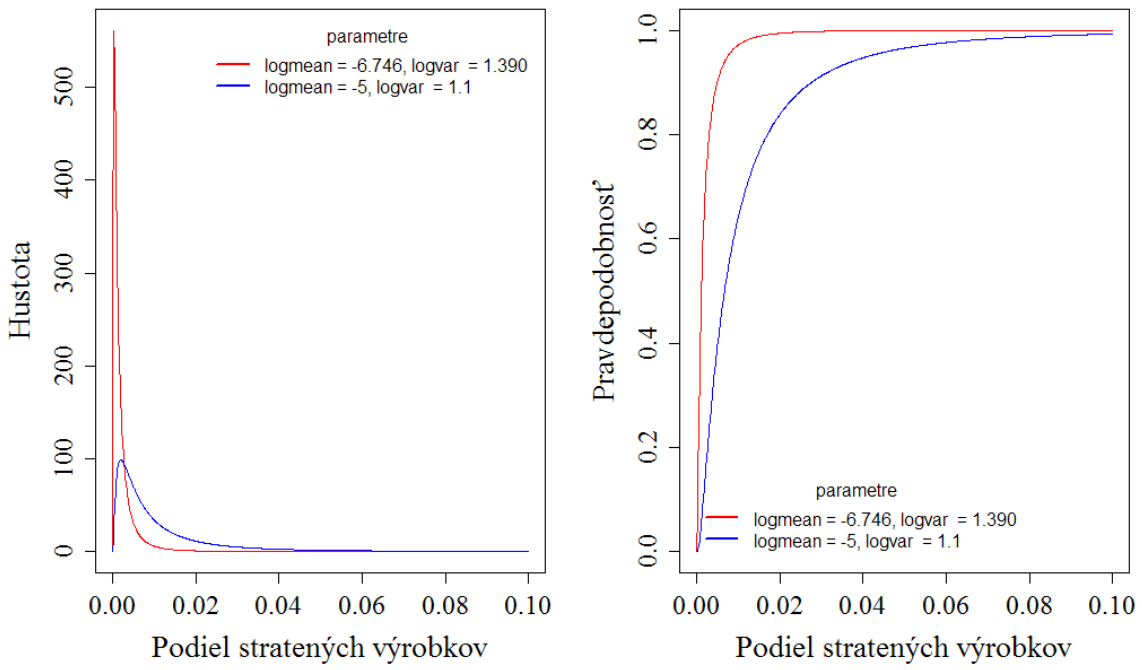
Určíme strednú hodnotu a rozptyl. Následne ich môžeme dosadiť do vzťahov pre výpočet pravdepodobnosti z distribučnej funkcie normálneho rozdelenia.

$$\mu_w = -6.746 \text{ a } \sigma_w^2 = 1.390$$

$$F(X \geq 0.005) = 1 - \left(\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln 0.005 - (-6.746)}{1.179\sqrt{2}}\right)\right) = 1 - (0.5 + 0.5.(0.780)) = 0.109$$

Na nasledujúcom obrázku (Obrázok 4.15) je PDF a CDF lognormálneho rozdelenia pravdepodobnosti pre $\mu_W = -5$ a $\sigma_W^2 = 1.1$, ako aj pre empiricky namerané hodnoty $\mu_W = -1.746$ a $\sigma_W^2 = 1.390$.

```
> scrap <- c(0.0003,  0.0002,  0.003,  0.0005,  0.004,  0.002,
  0.01,  0.0005,  0.0007,  0.003,  0.0004,  0.0008,  0.001,
  0.0002,  0.0006,  0.003,  0.003,  0.0006,  0.002,  0.009)
> x <- seq(0, 0.1, length = 1000)
> xh <- dlnorm(x, meanlog = mean(log(scrap)), sdlog =
  sd(log(scrap))*(length(scrap)-1)/length(scrap))
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.2, xlab = "Podiel stratených
  výrobkov", ylab = "Hustota", col = "red", cex.lab = 1.7,
  cex.axis = 1.5, family = "serif")
> lines(x, dlnorm(x, meanlog = -5, sdlog = 1.1), lty = 1.2, col
  = "blue")
> labels <- c("logmean = -6.746, logvar  = 1.390","logmean = -5,
  logvar  = 1.1")
> legend("topright", inset = 0.02, title = "parametre", legend =
  labels, lwd = 2, col = c("red", "blue"), bty = "n")
> xhh <- plnorm(x, meanlog = mean(log(scrap)), sdlog =
  sd(log(scrap))*(length(scrap)-1)/length(scrap))
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1.2, xlab = "Podiel stratených
  výrobkov", ylab = "Pravdepodobnosť", col = "red", xlim =
  c(0,0.1), cex.lab = 1.7, cex.axis = 1.5, family = "serif")
> lines(x, plnorm(x, meanlog = -5, sdlog = 1.1), lty = 1.2, col
  = "blue")
> legend("bottomleft", inset = 0.02, title = "parametre", legend
  = labels, lwd = 2,col = c("red", "blue"), bty = "n")
```



Obrázok 4.15: PDF a CDF lognormálneho rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.7 Weibullovo rozdelenie pravdepodobnosti

K spojitým rozdeleniam s pomerne pestrými možnosťami využitia patrí Weibullovo rozdelenie pravdepodobnosti, ktoré sa používa často v súvislosti s analýzou životnosti, či už živých organizmov (biológia, farmácia, medicína) alebo produktov a výrobkov (ekonómia, strojárstvo). V týchto situáciách sa predpokladá, že náhodná premenná X je vždy kladná, t.j. $X > 0$ a neraz táto náhodná premenná reprezentuje čas, v ktorom nastane určitá udalosť, napr. dôjde k reklamácii, poruche auta alebo súčiastky a podobne. Ide o rozdelenie s dvoma parametrami, ktorých obmenou je možné dosiahnuť pomerne široké spektrum tvarovo rôznorodých rozdelení. Ako bude možné vidieť z nasledujúcich obrázkov, rozdelenia sa tvarovo môžu podobať na normálne, lognormálne, ale aj na exponenciálne rozdelenie pravdepodobnosti. Táto flexibilita je zrejme dôvodom, prečo ide o jedno z najpopulárnejších rozdelení. Iným rozmerom, ktorému sa však v tejto publikácii bližšie venovať nebudeme, je využitie Weibullovo rozdelenia pravdepodobnosti na modelovanie extrémnych hodnôt.

Weibullovo rozdelenie má dva parametre, ktoré si označíme ako α a β. Prvý parameter α je tzv. parameter miery, ktorý je v rovnakých jednotkách ako náhodná premenná X, ktorú rozdelením modelujeme. Parameter α je dokonca vždy 63.2 percentilom a vyjadruje mieru rozptýlenia hodnôt. Druhý parameter β je bezrozmerný a vyjadruje tvar rozdelenia. Preto sa môžeme stretnúť s jeho pomenovaním ako parameter tvaru. V nasledujúcej tabuľke, máme

uvedené základné vlastnosti Weibullovho rozdelenia pravdepodobnosti, kde sa môžeme stretnúť s tzv. gamma funkciou, ktorú označujeme ako Γ(•). V tejto publikácii sa bližšie tomuto typu funkcií venovať nebudeme, pre úplnosť len uvedieme, že pre $x \in \mathbb{N}$ platí $\Gamma(x) = (x - 1)!$, komplikácie vznikajú pri $x \in C$, kde C je množina komplexných čísel.

Tabuľka 18: Tabuľka základných vlastností – Weibullovo rozdelenie

HUSTOTA		DISTRIBUČNÁ FUNKCIA
$f(x) = \frac{\beta}{\alpha^{\beta}} x^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}}, x > 0$		$F(x) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = \alpha\Gamma\left(1 + \frac{1}{\beta}\right)$	$\tilde{\mu} = \alpha(\ln(2))^{\frac{1}{\beta}}$	$\hat{\mu} = \alpha\left(\frac{\beta - 1}{\beta}\right)^{\frac{1}{\beta}}, \beta > 1$
DISPERZIA	$\alpha^2\Gamma\left(1 + \frac{2}{\beta}\right) - \mu^2$	PARAMETRE α, β

Zdroj: upravené podľa zdrojov v použitej literatúre

Na rozdiel od predchádzajúcich rozdelení býva odhad parametrov rozdelenia (α a β) netriviálny. Použitím vhodného softvérového vybavenia sa dostáva odhad do technickej roviny. Na tomto mieste prvý krát využijeme možnosť prezentovať metódu používanú na odhad parametrov rozdelenia. Ide o tzv. metódu maximálnej vierohodnosti, ktorú budeme v skratke označovať *MLE* (z angl. *Maximum Likelihood Estimation*).

Aj keď pre účely tejto podkapitoly nepovažujeme metódu *MLE* za kľúčovú, v skutočnosti sa využíva pomerne často pri analýze životnosti ako aj v ekonometrii. Využijeme tak prvú možnosť sa s týmto prístupom oboznámiť. Vynechaním technických detailov je možné prejsť priamo na vzťahy (4.51) a (4.52), pomocou ktorých je možné odhadnúť parametre rozdelenia. Princíp postupu je priamočiary a intuitívne pomerne príťažlivý. Majme náhodný výber $X_1, X_2, ..., X_n$, z určitého rozdelenia (v našom prípade Weibullovho) s hustotou $f(x;\theta_1,\theta_2,...,\theta_k)$, kde n je veľkosť vzorky a θ sú parametre rozdelenia f. Funkcia vierohodnosti tejto náhodnej vzorky je spoločná hustota n náhodných premenných. Inak povedané, funkcia vierohodnosti počíta takú hustotu pravdepodobnosti, v ktorej tieto náhodné premenné pochádzajú z rovnakého rozdelenia. Cieľom je potom maximalizovať túto funkciu vzhľadom na hľadané parametre. Zjednodušene by sa dalo povedať, že našim cieľom je maximalizovať pravdepodobnosť, že náhodné premenné pochádzajú z daného rozdelenia s určitými parametrami, pričom tento cieľ sa snažíme dosiahnuť práve prostredníctvom hľadaných parametrov. Postup je teda nasledovný:

$$L = \prod_{i=1}^{n} f\left(x_i; \theta_1, \theta_2, ..., \theta_k\right) \tag{4.47}$$

Keďže budeme funkciu L maximalizovať, za určitých podmienok, ktoré pri tejto hustote platia (a vo väčšine nami používaných prípadov to bude platiť), môžeme funkciu L logaritmovať a až potom derivovať, čím sa výpočet zjednoduší. Hodnotu derivácie položíme rovnú nule a dostaneme:

$$\frac{\partial \ln L}{\partial \theta} = 0 \tag{4.48}$$

V našom prípade je cieľom na základe nameraných údajov odhadnúť parametre Weibullovho rozdelenia pravdepodobnosti. Funkcia f je pritom funkciou hustoty Weibullovho rozdelenia pravdepodobnosti (Tabuľka 18). Funkciu maximálnej vierohodnosti môžeme napísať v tvare:

$$L\left(x_i; \alpha, \beta\right) = \prod_{i=1}^{n} \frac{\beta}{\alpha^{\beta}} x_i^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}} \tag{4.49}$$

$$\frac{\partial \ln L\left(x_i; \alpha, \beta\right)}{\partial \alpha \, \partial \beta} = 0 \tag{4.50}$$

Po výpočte parciálnych derivácií, ich položení rovné nule a zjednodušovaní si najprv vyjadríme parameter β z nasledujúcej rovnice:

$$\frac{\sum_{i=1}^{n} x_i^{\beta} \ln x_i}{\sum_{i=1}^{n} x_i^{\beta}} - \frac{1}{\beta} - \frac{1}{n} \sum_{i=1}^{n} \ln x_i = 0 \tag{4.51}$$

Po vypočítaní parametra β dosadením do nasledujúcej rovnice vypočítame parameter α:

$$\alpha = \left(\frac{\sum_{i=1}^{n} x_i^{\beta}}{n} \right)^{\left(\frac{1}{\beta}\right)} \tag{4.52}$$

Pre úplnosť uvedieme, že existujú aj iné metódy odhadu parametrov tohto rozdelenia. Za porovnateľne presnú sa považuje metóda momentov, za slabšiu metóda najmenších štvorcov. Odpoveďou na otázku, aké kritériá pri posudzovaní metód odhadu použiť, sa budeme čiastočne venovať v nasledujúcich kapitolách.

Príklad 4.18

Spoločnosť predávajúca softvérové riešenia šetrila priebeh reklamácií. Nasledujúce hodnoty vyjadrujú počet dní od nákupu reklamovaného produktu až po jeho samotnú reklamáciu. Spoločnosť z minulých skúseností predpokladá, že čas, ku ktorému od nákupu dôjde u problémových zákazníkov k reklamácii, sa riadi Weibullovým rozdelením pravdepodobnosti. Odhadnite pravdepodobnosť, že čas od nákupu do reklamácie bude menej ako 30 dní.

10, 12, 16, 16, 15, 14, 15, 18, 19, 21, 22, 23, 23, 22, 29, 36, 42, 43, 50, 52, 60, 70, 74, 80, 102, 112.

$$\frac{\sum_{i=1}^{n} x_i^{\beta} \ln x_i}{\sum_{i=1}^{n} x_i^{\beta}} - \frac{1}{\beta} - \frac{1}{n}\sum_{i=1}^{n} \ln x_i = \frac{\sum_{i=1}^{n} x_i^{\beta} \ln x_i}{\sum_{i=1}^{n} x_i^{\beta}} - \frac{1}{\beta} - \frac{1}{26} 88.4 = 0 \Rightarrow \beta = 1.475$$

$$\alpha = \left(\frac{\sum_{i=1}^{n} x_i^{1.475}}{26} \right)^{\left(\frac{1}{1.475}\right)} \Rightarrow \alpha = 42.7$$

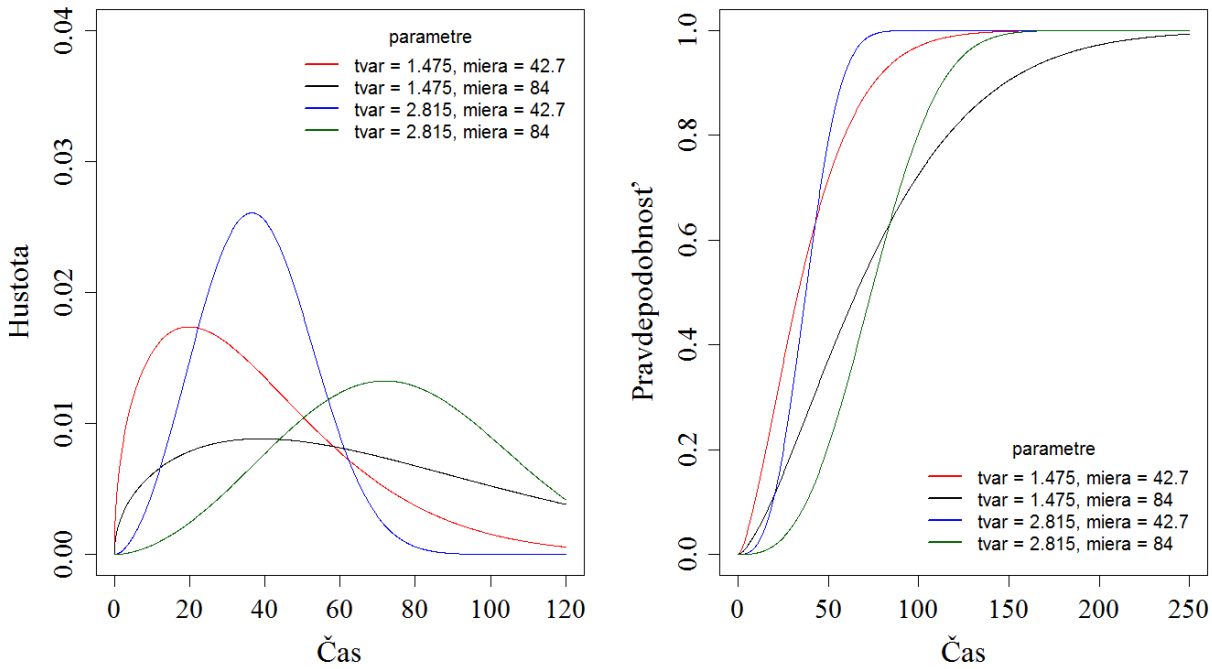
A teraz môžeme prejsť k samotnému výpočtu pravdepodobnosti z distribučnej funkcie Weibullovho rozdelenia pravdepodobnosti:

$$F(X \le x) = 1 - e^{-\left(\frac{30}{42.7}\right)^{1.475}} = 0.4474$$

Na nasledujúcom obrázku (Obrázok 4.16) sú 4 rôzne hustoty a distribučné funkcie Weibullovho rozdelenia pravdepodobnosti zobrazené tak, aby predstavovali zmenu tvaru rozdelenia v závislosti od zmeny vybraných parametrov rozdelenia.

```
> x <- seq(0, 120, length = 1000)
> xh <- dweibull(x, shape = 1.475, scale = 42.7)
> data <- data.frame(x, xh)
> par(mfrow = c(1, 2))
> plot(data, type = "l", lty = 1.2, xlab = "Čas", ylab =
  "Hustota", col = "red", ylim = c(0,0.04), cex.lab = 1.7,
  cex.axis = 1.5, family = "serif")
> lines(x, dweibull(x, shape = 1.475, scale = 84), lty = 1.2,
  col = "black")
> lines(x, dweibull(x, shape = 2.815, scale = 42.7), lty = 1.2,
  col = "blue")
> lines(x, dweibull(x, shape = 2.815, scale = 84), lty = 1.2,
  col = "darkgreen")
```

```
> labels <- c("tvar = 1.475, miera = 42.7", "tvar = 1.475, miera
  = 84", "tvar = 2.815, miera = 42.7", "tvar = 2.815, miera =
  84")
> legend("topright", inset = 0.02, title = "parametre", labels,
  lwd = 2, col = c("red", "black", "blue", "darkgreen"), bty =
  "n")
> x <- seq(0, 250, length = 1000)
> xhh <- pweibull(x, shape = 1.475, scale = 42.7)
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1, xlab = "Čas", ylab =
  "Pravdepodobnosť", col = "red", xlim = c(0,250), cex.lab =
  1.7, cex.axis = 1.5, family = "serif")
> lines(x, pweibull(x, shape = 1.475, scale = 84), lty = 1.2,
  col = "black")
> lines(x, pweibull(x, shape = 2.815, scale = 42.7), lty = 1.2,
  col = "blue")
> lines(x, pweibull(x, shape = 2.815, scale = 84), lty = 1.2,
  col = "darkgreen")
> legend("bottomright", inset = 0.02, title = "parametre",
  labels, lwd = 2, col = c("red", "black", "blue", "darkgreen"),
  bty = "n")
```



Obrázok 4.16: PDF a CDF Weibullovho rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.8 Gamma rozdelenie pravdepodobnosti

Podobne ako Weibullovo rozdelenie pravdepodobnosti, aj gamma rozdelenie môže nadobúdať rôzne tvary, od silne zošikmených rozdelení po symetrické, a to v závislosti od

dvoch parametrov. Parameter miery si znova označíme ako α a parameter tvaru ako[23] β. Pre potreby tejto publikácie si gamma rozdelenie ukazujeme ako všeobecný prípad niektorých odvodených rozdelení, akým je napr. Chí-kvadrát rozdelenie pravdepodobnosti. Využitie gamma rozdelenia je však veľmi podobné Weibullovmu rozdeleniu a ponúka jednu praktickú výhodu. Odhad parametrov gamma rozdelenia (α a β) je podstatne jednoduchší. Na druhej strane výpočet distribučnej funkcie si vyžaduje znalosť neúplnej gamma funkcie γ. Na výpočet pravdepodobnosti odporúčame použiť štatistické softvérové balíky. Rozdelenie uvádzame na ilustračné účely.

Tabuľka 19: Tabuľka základných vlastností – gamma rozdelenie

HUSTOTA		DISTRIBUČNÁ FUNKCIA
$f(x)=\frac{1}{\alpha\Gamma(\beta)}\left(\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\frac{x}{\alpha}}\right), x,\alpha,\beta>0$		$F(x)=\gamma\left(\beta,\frac{x}{\alpha}\right)$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=\alpha\beta$		$\hat{\mu}=\begin{cases}0, \alpha<0\\ \alpha^2-\alpha\end{cases}$
DISPERZIA	$\alpha^2\beta$	PARAMETRE α,β

Zdroj: upravené podľa zdrojov v použitej literatúre

Parameter miery α a tvaru β je možné odhadnúť pomocou tzv. metódy momentov, nasledujúcimi vzťahmi:

$$\alpha=\frac{s^2}{\bar{x}} \tag{4.53}$$

$$\beta=\left(\frac{\bar{x}}{s}\right)^2 \tag{4.54}$$

Príklad 4.19

Nasledujúce hodnoty reprezentujú výšku úverov v tis. EUR, pri ktorých bola vyhlásená neschopnosť splácať a úvery boli klasifikované ako nevymožiteľné. Z minulých skúseností vieme, že výška týchto úverov sa riadi gamma rozdelením pravdepodobnosti. Vedenie banky zaujíma, s akou pravdepodobnosťou bude náhodný nevymožiteľný úver mať hodnotu väčšiu ako 100000,-EUR.

11, 15, 16, 20, 20, 20, 24, 25, 30, 38, 40, 40, 41, 50, 50, 60, 60, 67, 70, 89, 110, 120, 140, 180.

[23] Vo väčšine literatúry sa môžeme stretnúť s opačným označením parametrov. Pri Weibullovom rozdelení sme ale definovali parameter β ako parameter tvaru. Za účelom sprehľadnenia označení sme sa rozhodli toto označenie dodržať aj pri gamma rozdelení.

$$\alpha = \frac{s^2}{\bar{x}} = \frac{43.67^2}{55.67} = 34.26$$

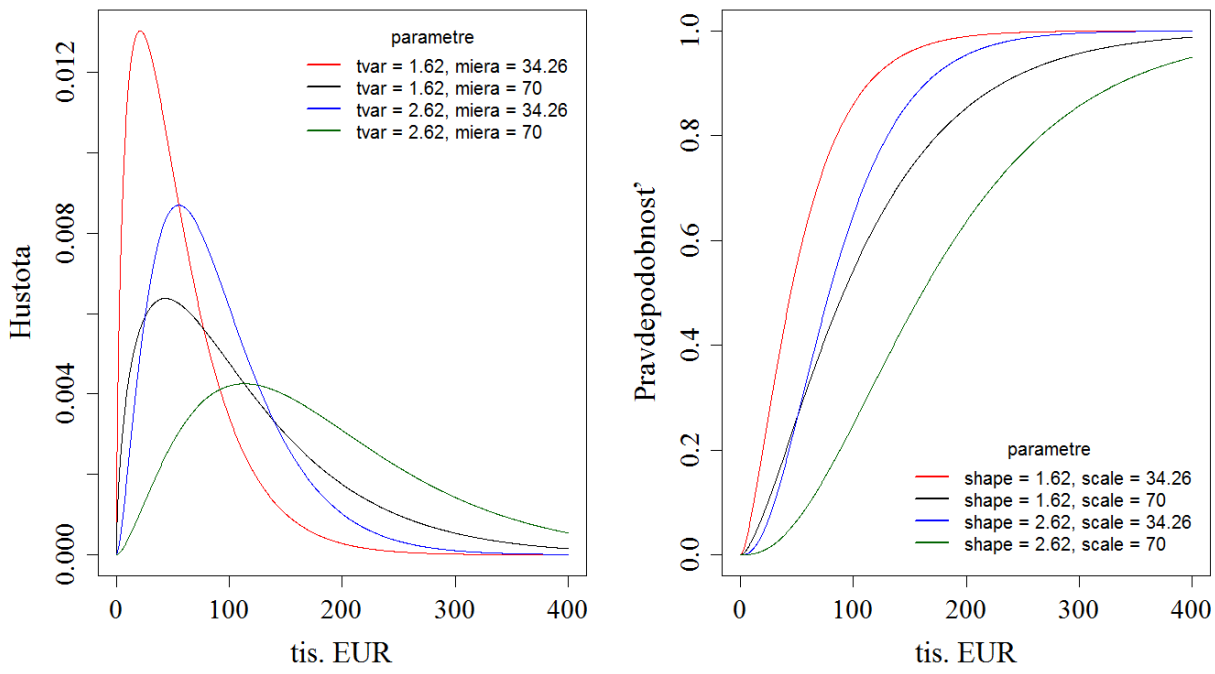
$$\beta = \left(\frac{\bar{x}}{s}\right)^2 = \left(\frac{55.67}{43.67}\right)^2 = 1.62$$

$$F(X \geq 100) = 1 - \gamma\left(1.62, \frac{100}{34.26}\right) = 0.1396$$

```
> 1 - pgamma(100, 1.62, scale = 34.26)
[1] 0.1396292
```

Na nasledujúcom obrázku (Obrázok 4.17) sú zobrazené 4 rôzne hustoty a distribučné funkcie gamma rozdelenia pravdepodobnosti, kde sme menili parameter miery α a tvaru β.

```
> x <- seq(0, 400, length = 1000)
> xh <- dgamma(x, shape = 1.62, rate = 1, scale = 34.26)
> data <- data.frame(x, xh)
> par(mfrow = c(1, 2))
> plot(data, type = "l", lty = 1.2, xlab = "tis. EUR", ylab =
  "Hustota", col = "red", xlim = c(0, 400), cex.lab = 1.7,
  cex.axis = 1.5, family = "serif")
> lines(x, dgamma(x, shape = 1.62, rate = 1, scale = 70), lty =
  1.2, col = "black")
> lines(x, dgamma(x, shape = 2.62, rate = 1, scale = 34.26), lty
  = 1.2, col = "blue")
> lines(x, dgamma(x, shape = 2.62, rate = 1, scale =
  70),lty=1.2, col = "darkgreen")
> labels <-  c("tvar = 1.62, miera = 34.26","tvar = 1.62, miera
  = 70","tvar = 2.62, miera = 34.26","tvar = 2.62, miera = 70")
> legend("topright", inset = 0.02, title = "parametre", labels,
  lwd = 2, col = c("red", "black", "blue", "darkgreen"), bty =
  "n")
> xhh <- pgamma(x, shape = 1.62, rate = 1, scale = 34.26)
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1.2, xlab = "tis. EUR", ylab =
  "Pravdepodobnosť", col = "red", xlim = c(0,400), cex.lab =
  1.7, cex.axis = 1.5, family = "serif")
> lines(x,pgamma(x, shape = 1.62, rate = 1.2, scale = 70), lty =
  1, col = "black")
> lines(x, pgamma(x, shape = 2.62, rate = 1.2, scale = 34.26),
  lty = 1, col = "blue")
> lines(x, pgamma(x, shape = 2.62, rate = 1.2, scale = 70),lty =
  1, col = "darkgreen")
> labels <- c("shape = 1.62, scale = 34.26","shape = 1.62, scale
  = 70","shape = 2.62, scale = 34.26","shape = 2.62, scale =
  70")
> legend("bottomright", inset = 0.02, title = "parametre",
  labels, lwd = 2, col = c("red", "black", "blue", "darkgreen"),
  bty = "n")
```



Obrázok 4.17: PDF a CDF gamma rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.9 Chí-kvadrát rozdelenie pravdepodobnosti

Rozdelenia ako Chí-kvadrát (označované aj prostredníctvom symbolu χ^2), *F*-rozdelenie a *t*-rozdelenie patria do širšej skupiny teoretických rozdelení, s ktorými sa môžeme stretnúť pri vyhodnocovaní mnohých štatistických testov. Chí-kvadrát rozdelenie je na jednej strane rozdelenie odvodené od normálneho rozdelenia pravdepodobnosti (dokonca pri $n > 90$ je možná aproximácia normálnym rozdelením) a zároveň ide o špecifický prípad gamma rozdelenia pravdepodobnosti.

Majme náhodné premenné $X_1, X_2, ..., X_k$, ktoré sú nezávislé a pochádzajú z normálneho rozdelenia so strednou hodnotou 0 a rozptylom 1, teda $N \sim (0,1)$. Potom náhodná premenná $Q = \sum_{i=1}^{k} X_i^2$ má Chí-kvadrát rozdelenie pravdepodobnosti s tzv. *k* stupňami voľnosti. Chí-kvadrát rozdelenie má zopár užitočných vlastností. Napríklad súčet dvoch náhodných premenných s Chí-kvadrát rozdelením pravdepodobnosti má tiež Chí-kvadrát rozdelenie. Ďalšou vlastnosťou je, že náhodná premenná s Chí-kvadrát rozdelením sa dá rozložiť na súčet niekoľkých nezávislých Chí-kvadrát rozdelení. Chí-kvadrát rozdelenie je špecifický typ gamma rozdelenia, kde parameter tvaru $\beta = k / 2$ a parameter miery $\alpha = 2$.

Pre úplnosť uvedieme, že pri štatistických testoch sa využíva nasledujúca vlastnosť:

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (4.55)$$

kde náhodná premenná χ^2 má Chí-kvadrát rozdelenie pravdepodobnosti s $n - 1$ stupňami voľnosti, pričom s^2 je tzv. výberový rozptyl (pozri publikáciu venujúcu sa induktívnej štatistike). Presnejšie, ak náhodná premenná X pochádza z normálneho rozdelenia, výberový rozptyl má Chí-kvadrát rozdelenie a zároveň platí, že podiel dvoch rozptylov bude mať F-rozdelenie (pozri ďalšie rozdelenie). Tieto skutočnosti sa využívajú pri konštrukcii intervalov spoľahlivosti a následne pri konštrukcii štatistických testov v publikáciách venujúcich sa induktívnej štatistike.

Hodnoty distribučnej funkcie pre rôzne hodnoty stupňov voľnosti sa v minulosti spravidla uvádzali v tabuľkách. V súčasnosti je možné tieto hodnoty získať priamo z funkcií rôznych štatistických softvérov.

Tabuľka 20: Tabuľka základných vlastností – Chí-kvadrát rozdelenie

HUSTOTA		DISTRIBUČNÁ FUNKCIA
$f(x) = \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, x > 0$		$F(x) = \dfrac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$
STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = k$	$\tilde{\mu} \approx k - \frac{2}{3} + \frac{4}{27k} - \frac{8}{729k^2}$	$\hat{\mu} = k - 2, k > 2$
DISPERZIA	$2k$	PARAMETRE k

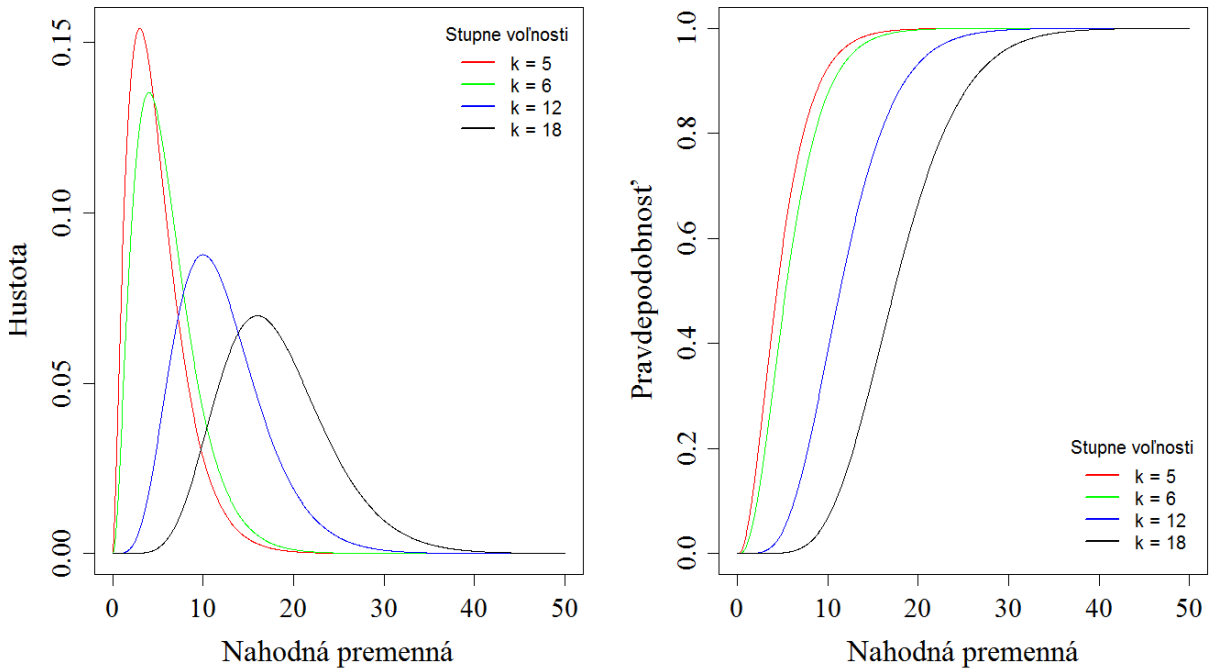
Zdroj: upravené podľa zdrojov v použitej literatúre

Na nasledujúcom obrázku (Obrázok 4.18) sú 4 rôzne hustoty a distribučné funkcie Chí-kvadrát rozdelenia pravdepodobnosti tak, aby sme ukázali ako sa s rôznou hodnotou parametra tvaru mení tvar a rozptýlenie rozdelenia.

Príkazy pre funkcie hustoty:

```
> x <- seq(0, 50, length = 1000)
> xh <- dchisq(x, 5)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.2, xlab = "Nahodná premenná",
  ylab = "Hustota", col = "red", cex.lab = 1.7, cex.axis = 1.5,
  family = "serif")
> col <- c("red", "green", "blue", "black")
> for (i in 1:3) lines(x, dchisq(x, i*6), lty = 1.2, col =
  col[i+1])
> label <- c("k = 5", "k = 6", "k = 12", "k = 18")
> legend("topright", inset = 0.02, title = "Stupne voľnosti",
  label, lwd = 2,col = col, bty = "n")
> xhh <- pchisq(x, 5)
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1.2, xlab = "Nahodná premenná",
  ylab = "Pravdepodobnosť", col = "red", cex.lab = 1.7, cex.axis
  = 1.5, family = "serif")
```

```
> for (i in 1:3) lines(x, pchisq(x, i*6), lty = 1.2, col =
  col[i+1])
> label <- c("k = 5", "k = 6", "k = 12", "k = 18")
> legend("bottomright", inset = 0.02, title = "Stupne voľnosti",
  label, lwd = 2, col = col, bty = "n")
```



Obrázok 4.18: PDF a CDF Chí-kvadrát rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.10 F-rozdelenie pravdepodobnosti

Ak dve náhodné premenné s Chí-kvadrát rozdelením dáme do vzájomného pomeru, výsledné rozdelenie pravdepodobnosti je tzv. *F*-rozdelenie. Pokiaľ Chí-kvadrát rozdelenie sa používa pri testoch o rozptyle, *F*-rozdelenie pri porovnávaní dvoch rozptylov. Ide o ďalšie rozdelenie, ktoré sa používa pri mnohých štatistických hypotézach a môžeme ho považovať spolu s normálnym, Chí-kvadrát a *t*-rozdelením za jedno z najvýznamnejších v štatistickej indukcii. Nech U_1 a U_2 sú náhodné premenné s Chí-kvadrát rozdelením a d_1 a d_2 sú príslušné stupne voľnosti, potom náhodná premenná *F* má *F*-rozdelenie:

$$F = \frac{\frac{U_1}{d_1}}{\frac{U_2}{d_2}} \tag{4.56}$$

V nasledujúcej tabuľke, je výpis základných vzťahov *F*-rozdelenia. Distribučná funkcia vychádza zo špeciálnej neúplnej beta funkcie *I*, ktorej výpočtu sa nebudeme bližšie venovať. V odborných publikáciách sa môžeme stretnúť s mnohými aproximáciami tejto distribučnej funkcie. Pre naše potreby si vystačíme so softvérovým riešením.

Tabuľka 21: Tabuľka základných vlastností – *F*-rozdelenie

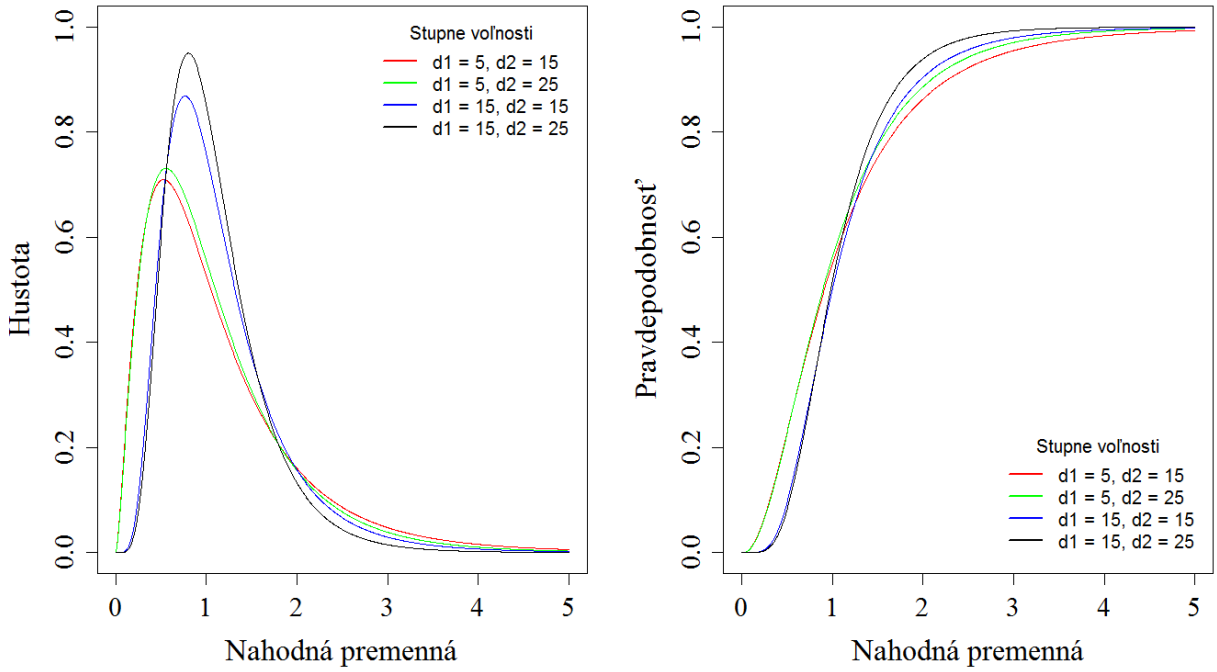
HUSTOTA	DISTRIBUČNÁ FUNKCIA
$f(x)=x^{\frac{d_1}{2}-1}\dfrac{\Gamma\left(\frac{d_1+d_2}{2}\right)\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\left[\left(\frac{d_1}{d_2}\right)x+1\right]^{\frac{(d_1+d_2)}{2}}},\ x>0$	$F(x)=I_{\frac{d_1x}{d_1x+d_2}}(d_1/2,d_2/2)$

STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu=\dfrac{d_2}{d_2-2},\ d_2>2$		$\dfrac{(d_1-2)}{d_1}\dfrac{d_2}{(d_2+2)},\ k>2$
DISPERZIA	$\dfrac{2d_2^2(d_1+d_2-2)}{d_1(d_1-2)^2(d_2-4)},\ d_2>4$	PARAMETRE d_1, d_2

Zdroj: upravené podľa zdrojov v použitej literatúre

Na nasledujúcom obrázku sú 4 rôzne hustoty a distribučné funkcie *F*-rozdelenia pravdepodobnosti zobrazené tak, aby sme ukázali, ako sa s rôznou hodnotou parametrov mení tvar rozdelenia.

```
> x <- seq(0, 5, length = 1000)
> xh <- df(x, 5, 15)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.2, xlab = "Nahodná premenná",
  ylab = "Hustota", col = "red", ylim = c(0, 1), cex.lab = 1.7,
  cex.axis = 1.5, family = "serif")
> col <- c("red", "green", "blue", "black")
> lines(x, df(x, 5, 25), lty = 1.2, col = "green")
> lines(x, df(x, 15, 15), lty = 1.2, col = "blue")
> lines(x, df(x, 15, 25), lty = 1.2, col = "black")
> label <- c("d1 = 5, d2 = 15", "d1 = 5, d2 = 25", "d1 = 15, d2
  = 15", "d1 = 15, d2 = 25")
> legend("topright", inset = 0.02, title = "Stupne voľnosti",
  label, lwd = 2, col = col, bty = "n")
> xhh <- pf(x, 5, 15)
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1.2, xlab = "Nahodná premenná",
  ylab = "Pravdepodobnosť", col = "red", ylim = c(0, 1), cex.lab
  = 1.7, cex.axis = 1.5, family = "serif")
> lines(x, pf(x, 5, 25), lty = 1.2, col = "green")
> lines(x, pf(x, 15, 15), lty = 1.2, col = "blue")
> lines(x, pf(x, 15, 25), lty = 1.2, col = "black")
> legend("bottomright", inset = 0.02, title = "Stupne voľnosti",
  label, lwd = 2, col = col, bty = "n")
```



Obrázok 4.19: PDF a CDF *F*-rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.6.11 Studentovo t-rozdelenie pravdepodobnosti

Majme náhodnú premennú *Z*, ktorá má normálne rozdelenie so strednou hodnotou 0 a s rozptylom 1 a náhodnú premennú χ^2, ktorá má Chí-kvadrát rozdelenie pravdepodobnosti s $(n-1)$ stupňami voľnosti. Potom náhodná premenná *T*, má tzv. Studentovo *t*-rozdelenie:

$$T = \frac{Z}{\chi^2} \tag{4.57}$$

Ak d_1 je počet stupňov voľnosti a F_1 hypergeometrická funkcia, potom funkciu hustotu a distribučnú funkciu môžeme zapísať tak, ako je to uvedené v nasledujúcej tabuľke:

Tabuľka 22: Tabuľka základných vlastností – *t*-rozdelenie

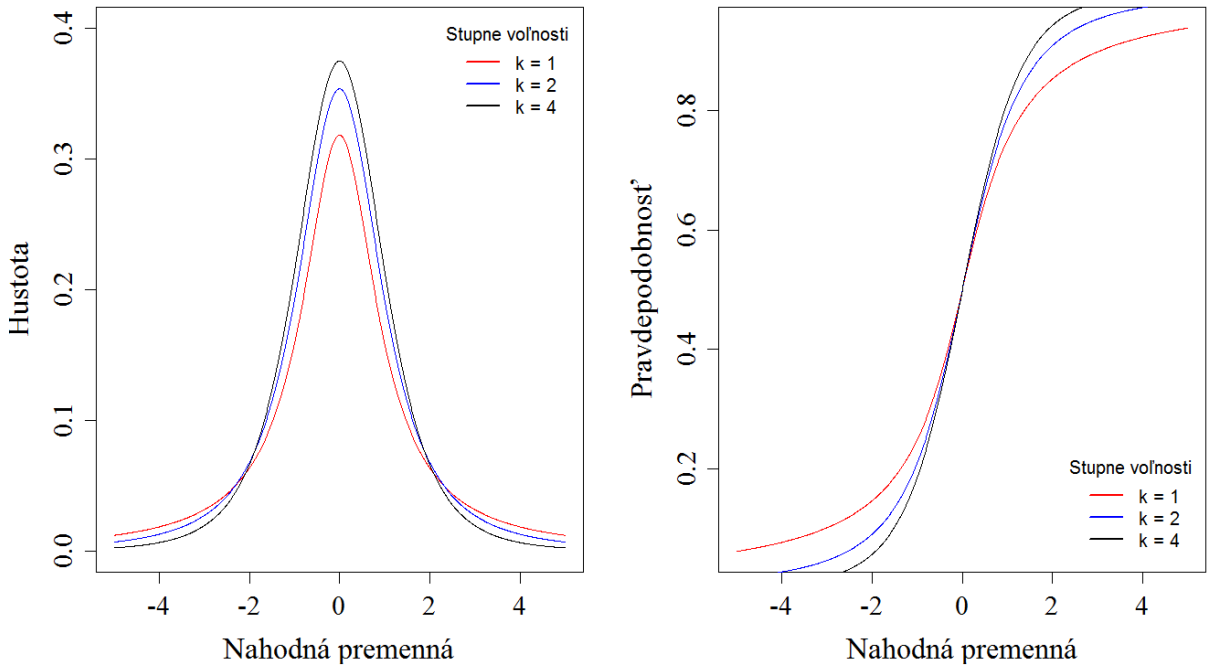
HUSTOTA	DISTRIBUČNÁ FUNKCIA
$f(x) = \frac{\Gamma\left(\frac{d_1+1}{2}\right)}{\sqrt{d_1\pi}\,\Gamma\left(\frac{d_1}{2}\right)}\left(1+\frac{x^2}{d_1}\right)^{-\frac{(d_1+1)}{2}}$ $x \in R,\ d_1 \in N$	$F(x) = \frac{1}{2} + x\Gamma\left(\frac{d_1+1}{2}\right)\frac{2F_1\left(\frac{1}{2},\frac{d_1+1}{2};\frac{3}{2};-\frac{x^2}{d_1}\right)}{\sqrt{\pi d_1}\,\Gamma\left(\frac{d_1}{2}\right)}$

STREDNÁ HODNOTA	MEDIÁN	MODUS
$\mu = 0,\ d_1 \geq 2$	$\tilde{\mu} = 0$	$\hat{\mu} = 0$

DISPERZIA	$\frac{d_1}{d_1-2},\ d_1 > 2$	PARAMETRE d_1

Zdroj: upravené podľa zdrojov v použitej literatúre

```
> x <- seq(-5, 5, length = 1000)
> xh <- dt(x, 1)
> data <- data.frame(x, xh)
> plot(data, type = "l", lty = 1.2, xlab = "Nahodná premenná",
  ylab = "Hustota", col = "red", ylim = c(0, 0.4), cex.lab =
  1.7, cex.axis = 1.5, family = "serif")
> col <- c("red", "blue", "black")
> for (i in 1:2) lines(x, dt(x, i*2), lty = 1.2, col = col[i+1])
> label <- c("k = 1", "k = 2", "k = 4")
> legend("topright", inset = 0.02, title = "Stupne voľnosti",
  label, lwd = 2, col = col, bty = "n")
> xhh <- pt(x, 1)
> data <- data.frame(x, xhh)
> plot(data, type = "l", lty = 1.2, xlab = "Nahodná premenná",
  ylab = "Pravdepodobnosť", col = "red", cex.lab = 1.7, cex.axis
  = 1.5, family = "serif")
> col <- c("red","blue","black")
> for (i in 1:2) lines(x, pt(x,i*2), lty = 1,col = col[i+1])
> legend("bottomright", inset = 0.02, title = "Stupne voľnosti",
  label, lwd = 2, col = col, bty = "n")
```



Obrázok 4.20: PDF a CDF *t*-rozdelenia pravdepodobnosti

Zdroj: vlastné spracovanie v programe R

4.7 Viacrozmerné rozdelenia pravdepodobnosti

Pri analýze ekonomických údajov je niekedy potrebné zaoberať sa aj prípadom, ak skúmame naraz viac než jednu náhodnú premennú. Môže nás zaujímať usporiadaná dvojica veľkosti nákupu a výšky príjmov zákazníkov. Usporiadaná množina náhodných premenných, s akými sme pracovali doposiaľ, vytvára tzv. náhodný vektor. Podobne ako môžeme

definovať strednú hodnotu, disperziu a kovarianciu medzi náhodnými premennými, môžeme definovať aj analogické vlastnosti pre náhodné vektory. Taktiež môžeme definovať aj distribučnú funkciu a viacrozmerné pravdepodobnostné rozdelenia. Vďaka viacrozmernej verzii centrálnej limitnej vety má v teórii významné postavenie hlavne viacrozmerné normálne rozdelenie, ako aj ďalšie rozdelenia od neho odvodené, ktoré si charakterizujeme v tejto časti.

4.7.1 Náhodný vektor, jeho stredná hodnota a variančno-kovariančná matica

Náhodnú premennú sme definovali ako funkciu, ktorá priraďuje výsledku pokusu reálne číslo. Uvažujme o situácii, ak by sme mali viac než jednu takúto náhodnú premennú. Označme si jednotlivé náhodné premenné ako $X_1, X_2, ..., X_n$, $n \in \mathbb{N}$. Náhodným vektorom $\mathbf{X}$ označíme vektor pozostávajúci z týchto náhodných premenných:

$$\mathbf{X} = (X_1, X_2, ..., X_n) \tag{4.58}$$

Strednú hodnotu náhodnej premennej, ktorú sme si už definovali, môžeme využiť pri definícii strednej hodnoty náhodného vektora:

$$\mathrm{E}(\mathbf{X}) = (\mathrm{E}(X_1), \mathrm{E}(X_2), ..., \mathrm{E}(X_n)) \tag{4.59}$$

Strednú hodnotu náhodného vektora teda dostaneme, ak vytvoríme vektor stredných hodnôt náhodných premenných, ktoré ho tvoria.

O niečo zložitejšie je to v prípade disperzie. Zdalo by sa, že by sme mohli postupovať rovnako ako pri strednej hodnote – definovať obyčajný vektor pozostávajúci z individuálnych disperzií. Nebude tomu však tak. Dôvod je jednoduchý – okrem individuálnych rozptylov je totiž veľmi dôležitý aj vzťah (presnejšie závislosť) premenných, ktoré tvoria náhodný vektor.

Ide o situáciu veľmi podobnú diferenciálnemu počtu funkcie viacerých premenných. Ak si spomenieme na poznatky z matematiky, pri funkcii f, ktorá má $n \in \mathbb{N}$ premenných, môžeme spočítať gradient, čiže vektor prvých parciálnych derivácií. Ak by sme však chceli určiť druhé derivácie, nedostali by sme vektor, ale tzv. Hessovu maticu, skladajúcu sa z druhých parciálnych derivácií. Maticu dostávame preto, že druhé derivácie môžeme dostať (krížovým) derivovaním prvých derivácií podľa všetkých premenných. Dostávame tak $n \times n$ derivácií (napr. podľa prvej a znova prvej premennej, podľa prvej a druhej, prvej a tretej, a pod.).

Analógia s naším problémom spočíva v tom, že namiesto „disperzie“ náhodného vektora nebudeme definovať vektor, ale takzvanú variančno-kovariančnú maticu (analogicky k Hessovej matici). Tá bude mať pre náhodný vektor $\mathbf{X}$ rozmer $n \times n$, a prvkami budú kovariancie medzi všetkými dvojprvkovými kombináciami náhodných premenných.

Variančno-kovariančnú maticu náhodného vektora **X** budeme označovať $\mathbf{\Sigma_X}$, poprípade var(**X**).

Aby sme si priblížili prvky $\mathbf{\Sigma_X}$, je najprv potrebné definovať kovarianciu medzi náhodnými premennými. Kovariancia medzi náhodnými premennými *X* a *Y* je daná vzťahom:

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}(X))\,(Y - \mathrm{E}(Y))] \tag{4.60}$$

Znamienko kovariancie nám hovorí o vzťahu medzi týmito náhodnými premennými. Ak je kovariancia kladná, hovoríme, že medzi *X* a *Y* je priama závislosť (s rastom *X* spravidla rastie aj *Y*). Ak je kovariancia záporná, hovoríme, že medzi *X* a *Y* je nepriama závislosť (rastúce hodnoty *X* sú spravidla sprevádzané klesajúcimi hodnotami *Y*). V prípade nulovej kovariancie medzi náhodnými premennými *X* a *Y* hovoríme, že sú **nekorelované**. Na tomto mieste je treba poznamenať, že nekorelovanosť neimplikuje nezávislosť v zmysle, ako sme ju definovali v predchádzajúcich častiach. Kovariancia meria len jednu špecifickú formu závislosti – premenné *X* a *Y* môžu byť závislé, no nekorelované. Platí však iné tvrdenie – nezávislé náhodné premenné *X* a *Y* sú vždy aj nekorelované.

Pomocou kovariancie preto môžeme definovať $\mathbf{\Sigma_X}$. Variančno-kovariančná matica bude predstavovať maticu, ktorej prvkami sú príslušné kovariancie medzi náhodnými premennými.

$$\mathbf{\Sigma_X} = \begin{pmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{cov}(X_n, X_n) \end{pmatrix} \tag{4.61}$$

Všimnime si však, že z definície kovariancie po dosadení tej istej náhodnej premennej dostávame:

$$\begin{aligned} \mathrm{cov}(X,X) &= \mathrm{E}[(X - \mathrm{E}(X))\,(X - \mathrm{E}(X))] \\ &= \mathrm{E}[(X - \mathrm{E}(X))^2] \\ &= \mathrm{D}(X) \end{aligned} \tag{4.62}$$

Kovariancia náhodnej premennej samej so sebou je teda rovná jej rozptylu. Vidíme, že vo variančno-kovariančnej matici sú to prvky na hlavnej diagonále (napr. $\mathrm{cov}(X_1, X_1)$ alebo $\mathrm{cov}(X_n, X_n)$). Môžeme preto napísať:

$$\mathbf{\Sigma_X} = \begin{pmatrix} D(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & D(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & D(X_n) \end{pmatrix} \tag{4.63}$$

Uvedená skutočnosť vysvetľuje aj názov variančno-kovariančnej matice – na hlavnej diagonále obsahuje disperzie (rozptyl sa predkladá v angličtine ako „variance“) a mimodiagonálne prvky sú kovariancie.

Variančno-kovariančná matica nám poskytuje pomerne veľa informácií o náhodných premenných tvoriacich náhodný vektor **X**. Vieme z nej vyčítať disperzie všetkých premenných obsiahnutých v **X**. Navyše, pomocou $\mathbf{\Sigma_X}$ vieme charakterizovať aj vzťahy medzi všetkými náhodnými premennými v **X**, keďže poznáme ich vzájomné kovariancie.

4.7.2 Združené, marginálne a podmienené rozdelenia

V predchádzajúcej časti sme definovali náhodný vektor, strednú hodnotu náhodného vektora a jeho variančno-kovariančnú maticu. Z toho čo sme uviedli je zrejmé, že okrem individuálnych vlastností náhodných premenných sú veľmi dôležité aj ich vzájomné vzťahy, ktoré sme si priblížili na príklade kovariancií.

V tejto časti by sme radi zodpovedali otázku, ako si predstaviť viacrozmerné pravdepodobnostné rozdelenie. V predchádzajúcich kapitolách sme popísali viaceré diskrétne a spojité rozdelenia náhodných premenných. Podobne ako pri strednej hodnote náhodného vektora, kde sme vychádzali zo strednej hodnoty jednotlivých premenných, aj pri definícii viacrozmerného rozdelenia budeme vychádzať práve z jednorozmerných. Je však potrebné vysvetliť, ako z čiastkových pravdepodobnostných rozdelení dostať jedno pravdepodobnostné rozdelenie pre celý náhodný vektor.

Pripomeňme, že ako v prípade diskrétnych, tak aj spojitých rozdelení bolo vždy možné definovať tzv. (kumulatívnu) distribučnú funkciu, pomocou ktorej môžeme popísať pravdepodobnostné rozdelenie. Hodnotu distribučnej funkcie F náhodnej premennej X v bode x sme definovali nasledovne:

$$F(x) = P(X \leq x) \tag{4.64}$$

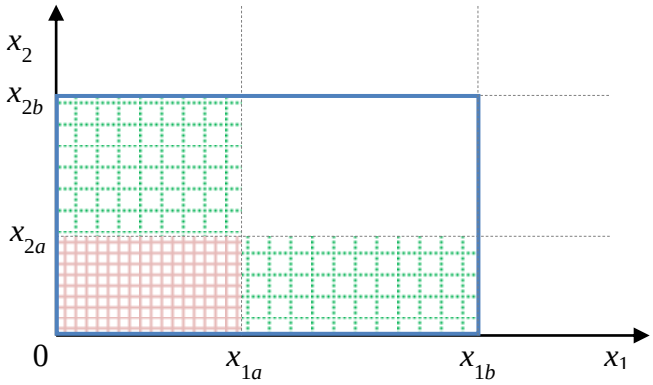
Bez ujmy na všeobecnosti pracujme pre prehľadnosť s najjednoduchším viacrozmerným rozdelením pravdepodobnosti – dvojrozmerným. Uvažujme preto náhodný vektor $\mathbf{X} = (X_1, X_2)$, kde X_1, X_2 sú náhodné premenné. Pre náhodný vektor **X** definujeme **združenú distribučnú funkciu** $F_{\mathbf{X}}(x_1, x_2)$ (angl. *joint distribution function*):

$$F_{\mathbf{X}}(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) \tag{4.65}$$

Tá vyjadruje pravdepodobnosť, že prvá zložka náhodného vektora **X** nadobudne hodnotu menšiu alebo rovnú x_1 a súčasne druhá zložka náhodného vektora **X** nadobudne hodnotu menšiu alebo rovnú x_2.

Z definície združenej distribučnej funkcie platí:

$$P(x_{1a} < X_1 \leq x_{1b}, x_{2a} < X_2 \leq x_{2b}) = F_{\mathbf{X}}(x_{1b}, x_{2b}) - F_{\mathbf{X}}(x_{1b}, x_{2a}) - F_{\mathbf{X}}(x_{1a}, x_{2b}) + F_{\mathbf{X}}(x_{1a}, x_{2a}) \tag{4.66}$$



Obrázok 4.21: Obsah obdĺžnika daného súradnicami (x_{1a}, x_{2a}) a (x_{1b}, x_{2b})

Zdroj: vlastné spracovanie

Intuitívne vysvetlenie predchádzajúcej vlastnosti je zrejmé (Obrázok 4.21). Ak by sme uvažovali o obsahu bieleho obdĺžnika vymedzenom bodmi (x_{1a}, x_{2a}) a (x_{1b}, x_{2b}), môžeme ho vyrátať tak, že najprv spočítame obsah celého modrého obdĺžnika vymedzeného bodmi (0, 0) a (x_{1b}, x_{2b}), od ktorého by sme odpočítali zelené obdĺžniky vymedzené bodmi (0, 0) a (x_{1a}, x_{2b}), ako aj (0, 0) a (x_{1b}, x_{2a}). Problém je, že tieto dva obdĺžniky sa prekrývajú. Červený obdĺžnik vymedzený (0, 0) a (x_{1a}, x_{2a}) by sme odrátali dvakrát (je súčasťou oboch zelených obdĺžnikov), musíme ho preto ešte raz prirátať.

Ak si to zhrnieme, pre výpočet obsahu bieleho obdĺžnika musíme vykonať nasledovné operácie na obdĺžnikoch vymedzených počiatkom súradnicovej sústavy (0, 0):

- vypočítať obsah modrého obdĺžnika, daného koncovým bodom (x_{1b}, x_{2b}),
- odpočítať obsah zeleného obdĺžnika, daného koncovým bodom (x_{1a}, x_{2b}),
- odpočítať obsah zeleného obdĺžnika, daného koncovým bodom (x_{1b}, x_{2a}),
- pripočítať obsah červeného obdĺžnika, daného koncovým bodom (x_{1a}, x_{2a}),

Vlastnosť *(4.66)* je úplne analogická našej predchádzajúcej úvahe s obdĺžnikmi s tým rozdielom, že namiesto obsahov sa za príslušné množiny počítajú hodnoty distribučnej funkcie (a obdĺžniky nepočítame od 0, ale od $-\infty$ v smere oboch osí).

Združená distribučná funkcia má niekoľko vlastností, ktoré sa podobajú na vlastnosti distribučnej funkcie jednej premennej:

$$\lim_{(x_1,x_2)\to(-\infty,-\infty)} F_{\mathbf{X}}(x_1,x_2)=0 \tag{4.67}$$

$$\lim_{(x_1,x_2)\to(\infty,\infty)} F_{\mathbf{X}}(x_1,x_2)=1 \tag{4.68}$$

Združená distribučná funkcia je ohraničená a je neklesajúca vzhľadom na všetky premenné.

Okrem združenej distribučnej funkcie môžeme uvažovať aj o distribučných funkciách jednotlivých náhodných premenných, ktoré vytvárajú náhodný vektor **X**. Ak by sme napríklad

uvažovali len o pravdepodobnostnom rozdelení náhodnej premennej X_1 bez ohľadu na to, aké hodnoty nadobúda náhodná premenná X_2 (resp. presnejšie, po zohľadnení všetkých hodnôt, ktoré môže nadobúdať náhodná premenná X_2), dostávame tzv. **marginálne rozdelenia pravdepodobnosti**.

Marginálne distribučné funkcie pre dve náhodné premenné tvoriace zložky náhodného vektora **X** sú definované nasledovne:

$$F_{X_1}(x_1) = P(X_1 \le x_1) = P(X_1 \le x_1, X_2 \le \infty) = \lim_{x_2 \to \infty} F(x_1, x_2) \tag{4.69}$$

$$F_{X_2}(x_2) = P(X_2 \le x_2) = P(X_1 \le \infty, X_2 \le x_2) = \lim_{x_1 \to \infty} F(x_1, x_2) \tag{4.70}$$

Marginálne distribučné funkcie vyjadrujú pravdepodobnosť, s akou bude príslušná náhodná premenná nadobúdať hodnotu menšiu ako zvolené číslo, ak neberieme do úvahy realizácie druhej náhodnej premennej.

Je zrejmé, že sa pri takomto prístupe dobrovoľne vzdávame určitej informácie. Pokiaľ medzi náhodnými premennými X_1 a X_2 existuje určitý vzťah a poznali by sme hodnotu náhodnej premennej X_2 (napríklad ak by sme vedeli, že $X_2 = 2$), mohlo by nám to pomôcť odhadnúť, aké hodnoty nadobudne náhodná premenná X_1.

Ak by sme takúto informáciu chceli využiť, znamenalo by to, že by sme sa snažili popísať pravdepodobnostné rozdelenie náhodnej premennej X_1, ktoré by záviselo na tom, aké hodnoty nadobudne náhodná premenná X_2. Inak povedané, hodnoty X_1 by boli podmienené hodnotami X_2. Takto získané rozdelenie nazývame **podmieneným rozdelením pravdepodobnosti**. V súvislosti s Bayesovou vetou sme si už v jednej z predchádzajúcich kapitol definovali podmienenú pravdepodobnosť. S jej využitím môžeme definovať **podmienenú distribučnú funkciu** pre náhodné premenné X_1 a X_2. Tieto distribučné funkcie budeme v súlade so spomínanou definíciou podmienenej pravdepodobnosti označovať nasledovne:

$$F_{X_1|X_2}(x_1 \mid x_2) = P(X_1 \le x_1, X_2 = x_2) \tag{4.71}$$

$$F_{X_2|X_1}(x_2 \mid x_1) = P(X_1 = x_1, X_2 \le x_2) \tag{4.72}$$

Vzťah medzi združeným, marginálnym a podmieneným rozdelením pravdepodobnosti si môžeme vysvetliť na prípade diskrétnych, aj spojitých náhodných premenných.

V prípade diskrétnych náhodných premenných X_1 a X_2 máme definované pravdepodobnostné funkcie pre obidve premenné. Marginálnu funkciu pravdepodobnosti pre náhodnú premennú X_1 môžeme definovať podľa podmienenej pravdepodobnosti a marginálnej pravdepodobnosti pre premennú X_2:

$$P(X_1 = x_1) = \sum_{x_2 \in H(X_2)} P(X_1 = x_1, X_2 = x_2) = \sum_{x_2 \in H(X_2)} P(X_1 = x_1 \mid X_2 = x_2) P(X_2 = x_2) \tag{4.73}$$

V predchádzajúcom vzťahu sú súčty realizované pre všetky hodnoty x_2, ktoré nadobúda náhodná premenná X_2 (ide o prvky oboru hodnôt X_2, ktorý označujeme ako $H(X_2)$). Všimnime si, že marginálne rozdelenie pre X_1 môžeme dostať pomocou marginálneho rozdelenia náhodnej premennej X_2 a podmieneného rozdelenia $X_1|X_2$. Čo sa stane v prípade, ak sú náhodné premenné X_1 a X_2 nezávislé?

Z predchádzajúcich kapitol vieme, že z definície pre dve nezávislé náhodné premenné X_1 a X_2 platí:

$$P(X_1 = x_1 \mid X_2 = x_2) = P(X_1 = x_1) \tag{4.74}$$

Po dosadení do vzorca *(4.73)* dostávame

$$\sum_{x_2 \in H(X_2)} P(X_1 = x_1 \mid X_2 = x_2) P(X_2 = x_2) = P(X_1 = x_1) \sum_{x_2 \in H(X_2)} P(X_2 = x_2) \tag{4.75}$$

$$= P(X_1 = x_1) \cdot 1 \tag{4.76}$$

$$= P(X_1 = x_1) \tag{4.77}$$

Vo všeobecnosti by sme mohli pre združenú funkciu pravdepodobnosti napísať pomocou podmienenej a marginálnej funkcie pravdepodobnosti:

$$P(X_1 = x_1, X_2 = x_2) = P(X_2 = x_2 \mid X_1 = x_1)\, P(X_1 = x_1) \tag{4.78}$$

Podľa vyššie uvedenej vlastnosti pre nezávislé náhodné veličiny X_1, X_2 dostávame:

$$P(X_1 = x_1, X_2 = x_2) = P(X_2 = x_2 \mid X_1 = x_1)\, P(X_1 = x_1) \tag{4.79}$$

$$= P(X_2 = x_2)\, P(X_1 = x_1) \tag{4.80}$$

To znamená, že pre nezávislé náhodné veličiny predstavuje združená funkcia pravdepodobnosti súčin jednotlivých marginálnych funkcií pravdepodobnosti.

S pomocou marginálnych a podmienených rozdelení X_1 a X_2 môžeme odvodiť ďalšie veličiny, akými sú napríklad podmienená stredná hodnota, podmienený rozptyl a podobne. Na ukážku si definujme napríklad podmienenú strednú hodnotu, ktorá má v štatistike (ako aj v teórii pravdepodobnosti) zásadný význam. Podmienenou strednou hodnotou nazývame výraz:

$$E(X_1 \mid X_2 = x_2) = \sum_{x_1 \in H(X_1)} x_1 P(X_1 = x_1 \mid X_2 = x_2) \tag{4.81}$$

$$= \sum_{x_1 \in H(X_1)} x_1 \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} \tag{4.82}$$

V prvom prípade sme podmienenú strednú hodnotu zapísali pomocou podmienenej pravdepodobnosti, v druhom pomocou združenej a marginálnej funkcie pravdepodobnosti.

V prípade spojitých náhodných premenných nemá zmysel uvažovať o funkcii pravdepodobnosti, využívame preto spravidla hustoty pravdepodobnosti. V ďalšom texte predpokladajme, že pre skúmané rozdelenie existujú všetky funkcie hustôt pravdepodobnosti, s ktorými budeme pracovať.

Analógiou funkcie marginálnej pravdepodobnosti sú **funkcie marginálnej hustoty pravdepodobnosti**, ktoré označíme $f_{X_1}(x_1)$ a $f_{X_2}(x_2)$. Podobne **podmienené hustoty pravdepodobnosti** označíme $f_{X_1|X_2}(x_1 \mid x_2)$ a $f_{X_2|X_1}(x_2 \mid x_1)$. **Združenú hustotu pravdepodobnosti** definujeme s ich pomocou ako:

$$f_{\mathbf{X}}(x_1, x_2) = f_{X_1|X_2}(x_1 \mid x_2) f_{X_2}(x_2) = f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1) \qquad (4.83)$$

Pre združenú distribučnú funkciu diskrétneho náhodného vektora **X** potom dostávame:

$$F_{\mathbf{X}}(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2) \qquad (4.84)$$

$$= \sum_{x_i \le x_1} \sum_{x_j \le x_2} P(X_1 = x_i, X_2 = x_j) \qquad (4.85)$$

Pre združenú distribučnú funkciu spojitého náhodného vektora **X** máme:

$$F_{\mathbf{X}}(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2) \qquad (4.86)$$

$$= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{\mathbf{X}}(x_i, x_j) dx_j dx_i \qquad (4.87)$$

Príklad 4.20

Popíšme združené, marginálne a podmienené rozdelenia pravdepodobnosti pre náhodné premenné *X* a *Y*, ak obidve môže nadobúdať hodnoty z množiny {1, 2, 3} s nasledovnými pravdepodobnosťami

		Y		
		1	2	3
	1	0.15	0.04	0.03
X	2	0.13	0.19	0.06
	3	0.11	0.12	0.17

Riešenie: V prvom rade si treba uvedomiť, že pravdepodobnosti obsiahnuté v tabuľke nám definujú združené rozdelenie pravdepodobnosti. Je tomu tak preto, lebo každej kombinácii hodnôt, ktoré môže nadobúdať náhodná premenná *X*, ako aj náhodná premenná *Y*, môžeme jednoznačne priradiť pravdepodobnosť. Napríklad je zrejmé, že

$$P(X = 1, Y = 3) = 0.03$$

Zároveň môžeme overiť, že naozaj ide o rozdelenie pravdepodobnosti – súčet všetkých pravdepodobností v tabuľke je

$$0.15 + 0.04 + 0.03 + 0.13 + 0.19 + 0.06 + 0.11 + 0.12 + 0.17 = 1.00$$

Marginálne rozdelenia môžeme zo združeného získať tak, že spočítame riadkové, resp. stĺpcové súčty pravdepodobností:

		Y			
		1	2	3	
	1	0.15	0.04	0.03	0.22
X	2	0.13	0.19	0.06	0.38
	3	0.11	0.12	0.17	0.40
		0.39	0.35	0.26	

Marginálnu pravdepodobnosť $P(X = 1)$ dostaneme takto:

$$P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) + P(X = 1, Y = 3)$$
$$= 0.15 + 0.04 + 0.03$$
$$= 0.22$$

Podobne pre marginálnu pravdepodobnosť $P(Y = 2)$ máme:

$$P(Y = 2) = P(X = 1, Y = 2) + P(X = 2, Y = 2) + P(X = 3, Y = 2)$$
$$= 0.04 + 0.19 + 0.12$$
$$= 0.35$$

Zostáva ešte definovať podmienené pravdepodobnosti.

Charakterizujme teraz podmienené rozdelenie pravdepodobnosti náhodnej premennej X v závislosti na Y, ak $Y = 3$. Zrejme tak ako doposiaľ, náhodná premenná X môže nadobúdať tri hodnoty, 1, 2 a 3.

		Y		
		1	2	3
	1	0.15	0.04	0.03
X	2	0.13	0.19	0.06
	3	0.11	0.12	0.17

Ak platí $Y = 3$, potom do úvahy prichádza len posledný stĺpec tabuľky. Pre ostatné bunky totiž Y nadobúda iné hodnoty ako 3.

Ak by sme mali preto stanoviť pravdepodobnosť javu napríklad $P(X = 1 \mid Y = 3)$, mohli by sme byť v pokušení tvrdiť, že hľadaná pravdepodobnosť je rovná 0.03. V skutočnosti tomu tak ale nebude, čo je pri pohľade na posledný stĺpec tabuľky intuitívne zrejmé (pravdepodobnosť by bola veľmi nízka).

Dôvodom je skutočnosť, že aj podmienené rozdelenie $X \mid Y = 3$ je pravdepodobnostným rozdelením. Súčet pravdepodobností všetkých javov musí byť rovný jednej.

Už z úlohy o marginálnych pravdepodobnostiach vieme, že marginálna pravdepodobnosť $Y = 3$ má hodnotu:

$$P(Y = 3) = P(X = 1, Y = 3) + P(X = 2, Y = 3) + P(X = 3, Y = 3) = 0.26$$

Stĺpec, ktorý sme v predchádzajúcej tabuľke vyznačili zelenou preto neudáva podmienené pravdepodobnosti. Tie dostaneme, ak ich vydelíme marginálnou pravdepodobnosťou $P(Y = 3)$.

Dostaneme tak hľadaný výsledok:

$$P(X = 1 \mid Y = 3) = P(X = 1, Y = 3) / P(Y = 3) = 0.03/0.26 = 0.12$$

$$P(X = 2 \mid Y = 3) = P(X = 1, Y = 3) / P(Y = 3) = 0.06/0.26 = 0.23$$

$$P(X = 3 \mid Y = 3) = P(X = 1, Y = 3) / P(Y = 3) = 0.17/0.26 = 0.65$$

Pre kontrolu ešte môžeme uviesť, že v tomto prípade naozaj dostávame súčet pravdepodobnosti rovný jednej:

$$P(X = 1 \mid Y = 3) + P(X = 2 \mid Y = 3) + P(X = 3 \mid Y = 3) = 0.12 + 0.23 + 0.65 = 1.00$$

4.7.3 Dvojrozmerné normálne rozdelenie

S pomocou rozdelení charakterizovaných v predchádzajúcej časti je možné definovať dvojrozmerné normálne rozdelenie pravdepodobnosti. O náhodných premenných X_1 a X_2 tvoriacich náhodný vektor $\mathbf{X} = (X_1, X_2)$ hovoríme, že majú dvojrozmerné združené rozdelenie pravdepodobnosti, ak všetky lineárne kombinácie zložiek $\mathbf{X}$ v tvare:

$$a_1X_1 + a_2X_2 \tag{4.88}$$

kde $a_1, a_2 \in \mathbb{R}$, majú jednorozmerné normálne rozdelenie pravdepodobnosti.

Vysvetlime si teraz, čo táto definícia znamená. V prvom rade si uvedomme, že kým $\mathbf{X}$ predstavuje náhodný vektor, jeho zložky X_1 a X_2 sú náhodné premenné, ktoré majú jednorozmerné rozdelenia pravdepodobnosti. Ak vytvoríme nejakú ich lineárnu kombináciu, ako sme to popísali vyššie, dostávame novú náhodnú premennú, ktorá má taktiež jednorozmerné rozdelenie pravdepodobnosti.

Pravdepodobnostné rozdelenie lineárnej kombinácie by teoreticky mohlo byť rôzne. Ak však pre všetky $a_1, a_2 \in \mathbb{R}$ dostávame vždy jednorozmerné normálne rozdelenie, hovoríme, že náhodný vektor má združené dvojrozmerné normálne rozdelenie.

Mohlo by sa nám zdať, že súčet (alebo iná lineárna kombinácia) náhodných premenných s normálnym rozdelením pravdepodobnosti má normálne rozdelenie, ako sa požaduje v definícii. Vyvstáva otázka, či táto požiadavka je vždy splnená. V skutočnosti súčet (alebo lineárna kombinácia) normálne rozdelených *nezávislých* náhodných premenných má vždy normálne rozdelenie. Platí preto, že nezávislé náhodné premenné s normálnym rozdelením pravdepodobnosti majú vždy združené normálne rozdelenie. Na tomto mieste je potrebné upozorniť, že dané náhodné premenné musia byť naozaj nezávislé – nie je

postačujúce, aby boli len nekorelované (nestačí, aby ich variančno-kovariančná matica bola diagonálna).

Problémy vznikajú v prípade, ak náhodné premenné X_1 a X_2 nie sú nezávislé. V takomto prípade môžu, ale nemusia mať združené normálne rozdelenie a je potrebné preskúmať platnosť vzťahu uvedeného v definícii.

V predchádzajúcich odsekoch sme sa venovali problému, či náhodné premenné s normálnym rozdelením pravdepodobnosti budú vždy združene normálne rozdelené (odpoveď bola nie). Pravdivé je však nasledovné tvrdenie: ak má náhodný vektor **X** združené normálne rozdelenie, potom jeho zložky majú nutne každá jednorozmerné normálne rozdelenie.

V prípade dvojrozmerného náhodného vektora **X** navyše platí aj tvrdenie, že nekorelované náhodné premenné so združeným normálnym rozdelením pravdepodobnosti sú vždy aj nezávislé (v tomto prípade nekorelovanosť implikuje nezávislosť).

V ďalšej časti by sme chceli pristúpiť k charakteristike dvojrozmerného normálneho rozdelenia.

V jednorozmernom prípade sme charakterizovali normálne rozdelenie náhodnej premennej pomocou dvoch parametrov: strednej hodnoty a rozptylu. Vo viacrozmernom prípade budeme mať taktiež dva parametre. Prvým je vektor stredných hodnôt, ktorý nám charakterizuje polohu v rámci daného rozdelenia. Druhým parametrom bude variančno-kovariančná matica.

Pripomeňme, že pomocou variančno-kovariančnej matice vieme získať o náhodných premenných tvoriacich náhodný vektor pomerne veľa informácií. Jednak vieme zistiť rozptyly jednotlivých náhodných premenných (nachádzajú sa na hlavnej diagonále), ale taktiež vieme zistiť aj vzájomné kovariancie medzi premennými, ktoré charakterizujú ich vzájomný vzťah. Ako sme videli v predchádzajúcich odsekoch, vzájomný vzťah medzi premennými (a jeho absencia, teda nezávislosť) môže byť veľmi dôležitý, nie je preto prekvapivé, že vystupuje aj ako parameter tohto rozdelenia pravdepodobnosti.

Variančno-kovariančná matica pre náhodný vektor $\mathbf{X} = (X_1, X_2)$ môže byť zapísaná nasledovne:

$$\mathbf{\Sigma}_{\mathbf{X}} = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) \end{pmatrix} \tag{4.89}$$

$$= \begin{pmatrix} \sigma_{X_1}^2 & \operatorname{cov}(X_1, X_2) \\ \operatorname{cov}(X_1, X_2) & \sigma_{X_2}^2 \end{pmatrix} \tag{4.90}$$

$$= \begin{pmatrix} \sigma_{X_1}^2 & \rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2} \\ \rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{pmatrix} \tag{4.91}$$

Prvá rovnosť predstavuje definíciu variančno-kovariančnej matice. V druhom kroku sme označili σ_{X_1} a σ_{X_2} rozptyly náhodných premenných (keďže kovariancia náhodnej premennej samej so sebou je rovná jej rozptylu). V poslednom kroku sme využili definíciu korelačného koeficientu, ktorý je daný vzťahom:

$$\rho_{X_1X_2} = \frac{\mathrm{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} \tag{4.92}$$

Teraz už máme všetko potrebné, aby sme definovali združenú hustotu pravdepodobnosti dvojrozmerného normálneho rozdelenia $f_{\mathbf{X}}$:

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho_{X_1X_2}^2}} \exp\left(-\frac{1}{2(1-\rho_{X_1X_2}^2)} H(x_1, x_2)\right) \tag{4.93}$$

$$H(x_1, x_2) = \left(\frac{X_1 - E(X_1)}{\sigma_{X_1}}\right)^2 + \left(\frac{X_2 - E(X_2)}{\sigma_{X_2}}\right)^2 - 2\rho_{X_1X_2}\frac{(X_1 - E(X_1))(X_2 - E(X_2))}{\sigma_{X_1}\sigma_{X_2}} \tag{4.94}$$

Združená hustota sa výrazne podobá na hustotu obyčajného jednorozmerného normálneho rozdelenia.

Zo združenej funkcie hustoty je možné získať aj individuálne marginálne hustoty pravdepodobnosti pre náhodné premenné X_1 a X_2:

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}} \exp\left(-\frac{1}{2}\left(\frac{x_1 - E(X_1)}{\sigma_{X_1}}\right)^2\right) \tag{4.95}$$

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_{X_2}} \exp\left(-\frac{1}{2}\left(\frac{x_2 - E(X_2)}{\sigma_{X_2}}\right)^2\right) \tag{4.96}$$

V prípade marginálnych distribučných funkcií vidíme, že naozaj ide o jednorozmerné hustoty pravdepodobnosti normálneho rozdelenia. Z ich zápisu je zrejmé, že pri marginálnych rozdeleniach strácame informáciu o vzájomnom vzťahu premenných: vzorec *(4.95)* vôbec neobsahuje X_2, a vzorec *(4.96)* vôbec neobsahuje výrazy súvisiace s X_1. Na tomto vidieť aj to, prečo dve normálne rozdelené náhodné premenné nemusia mať vždy združené normálne rozdelenie – okrem marginálnych rozdelení musí určité podmienky spĺňať aj ich vzájomný vzťah, ktorý z marginálnych rozdelení nevidíme.

Podobne je možné odvodiť aj podmienené hustoty pravdepodobnosti:

$$f_{X_2|X_1}(x_2 \mid x_1) = \frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho_{X_1X_2}^2}} \exp\left(-\frac{1}{2} \left(\frac{x_2 - E\left(X_2 - \rho_{X_1X_2} \frac{\sigma_{X_2}}{\sigma_{X_1}} (X_1 - E(X_1)) \right)}{\sigma_{X_2}\sqrt{1-\rho_{X_1X_2}^2}} \right)^2 \right) \quad (4.97)$$

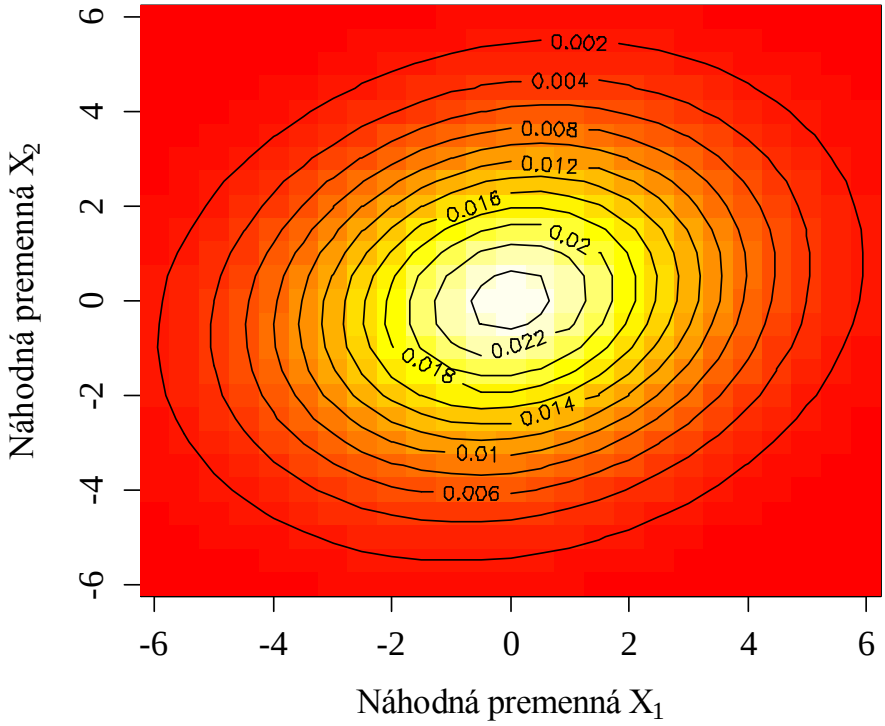
$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}\sqrt{1-\rho_{X_1X_2}^2}} \exp\left(-\frac{1}{2} \left(\frac{x_1 - E\left(X_1 - \rho_{X_1X_2} \frac{\sigma_{X_1}}{\sigma_{X_2}} (X_2 - E(X_2)) \right)}{\sigma_{X_1}\sqrt{1-\rho_{X_1X_2}^2}} \right)^2 \right) \quad (4.98)$$

Pri podmienených hustotách pravdepodobnosti získavame popis pravdepodobnostného rozdelenia pre prípad, že je hodnota jednej premennej podmienená známou hodnotou druhej. Ak vieme, že medzi X_1 a X_2 existuje vzťah, potom ak poznáme hodnotu, ktorú nadobudne X_2, dostávame iné rozdelenie pravdepodobnosti pre hodnoty, ktoré pri danom X_2 môže nadobúdať X_1.

Úlohu a význam jednotlivých pravdepodobnostných rozdelení, ako aj hustôt a distribučných funkcií môžeme zjednodušene popísať takto:

- **Združené rozdelenie pravdepodobnosti** nám hovorí, aká je pravdepodobnosť javu, že náhodný vektor **X** = (X_1, X_2) bude nadobúdať určité hodnoty. Inak povedané, hovorí nám o rozdelení pravdepodobnosti hodnôt, ktoré môžu spoločne nadobúdať náhodné premenné X_1 a X_2.
- **Marginálne rozdelenia pravdepodobnosti** pre náhodné premenné X_1 a X_2 nám hovoria o pravdepodobnostiach, s akými premenné X_1 a X_2 budú nadobúdať rôzne hodnoty, pokiaľ nevyužijeme informáciu o ich vzťahu – teda bez ohľadu na druhú premennú. Dostávame samostatné jednorozmerné rozdelenie pravdepodobnosti pre X_1 a X_2.
- **Podmienené rozdelenia pravdepodobnosti** sú rozdeleniami pravdepodobnosti, ktoré popisujú pravdepodobnosť nadobudnutia určitých hodnôt jednou z náhodných premenných, ak predpokladáme, že druhá náhodná premenná nadobúda konkrétnu hodnotu. Tieto rozdelenia nám dávajú odpoveď na otázku, aké sú pravdepodobnosti pre náhodnú premennú X_1 vzhľadom na jej vzťah k X_2, ak vieme, že X_2 nadobúda známe hodnoty. Rozdelenie sa podobá na združené rozdelenie pravdepodobnosti s tým rozdielom, že náhodná je len jedna premenná (v našom príklade X_1) a hodnota druhej je daná (X_2).

Dvojrozmerné normálne rozdelenie si môžeme priblížiť aj graficky. Pri dvojrozmernom rozdelení bude pre vizualizáciu hustoty pravdepodobnosti nutné využiť trojrozmernú plochu. Jednou z možností, ako tento problém obísť, je zakresliť rôzne úrovne hodnoty združenej funkcie hustoty pomocou „vrstevníc“, teda úrovňových kriviek. V programe R môžeme takýto obrázok nakresliť využitím funkcie `countour`.

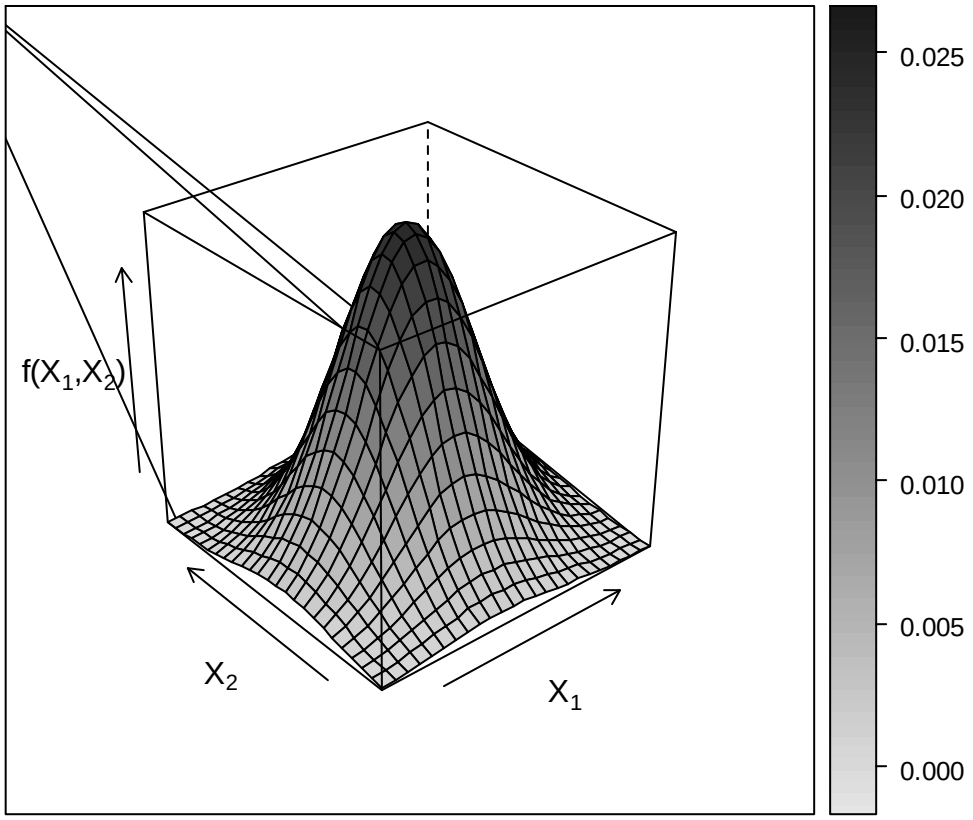


Obrázok 4.22 Úrovňové krivky pre dvojrozmerné združené normálne rozdelenie

Zdroj: vlastné spracovanie v programe R

```
> library(mvtnorm)
> library(lattice)
> rm(list = ls())
-------------------------------------------------------------------
> x <- seq(-6, 6, 0.5)
> y <- x
> density <- function(x, y) { dmvnorm(x = cbind(x, y), sigma =
  matrix(c(7, 1, 1, 6), nrow = 2))}
> z <- outer(x, y, density)
> par(mar = c(4, 5, 1, 0.5))
> image(x, y, z, col = heat.colors(32), xlab =
  expression("Náhodná premenná X"[1]), ylab =
  expression("Náhodná premenná X "[2]), cex.lab = 1, family =
  "serif")
> contour(x, y, z, add = TRUE)
```

V programe R môžeme zobraziť aj trojrozmernú projekciu hustoty pravdepodobnosti združeného normálneho rozdelenia.



Obrázok 4.23 Združená hustota pravdepodobnosti dvojrozmerného normálneho rozdelenia

Zdroj: vlastné spracovanie v programe R

```
> library(mvtnorm)
> library(lattice)
> rm(list=ls())
-----------------------------------------------------------------
> x <- seq(-6, 6, 0.5)
> y <- x
> density <- function(x, y) {dmvnorm(x = cbind(x, y), sigma =
  matrix(c(7, 1, 1, 6), nrow = 2))}
> z <- outer(x, y, density)
> par(mar = c(4, 5, 1, 0.5))
> grid <- expand.grid(x = x,y = y)
> z <- c()
> for (c in 1 : (dim(grid)[1])) grid[["z"]][c] =
  dmvnorm(cbind(grid[c, 1], grid[c, 2]), sigma = matrix(c(7, 1 ,
  1, 6), nrow = 2))
> newcols <- colorRampPalette(c("grey90", "grey10"))
> wireframe(z ~ x * y, grid, colorkey = TRUE, drape = TRUE,
  col.regions = newcols(100), xlab = expression("X"[1]), ylab =
  expression("X"[2]), zlab = expression("f(X"[1]*",X"[2]*")"))
```

Na obrázku je možné vidieť niekoľko skutočností. Ak si všimneme krivky, ktoré v grafe prechádzajú v smere osi X_1, môžeme ľahko vidieť, že ich tvar pripomína

jednorozmerné normálne rozdelenie pravdepodobnosti. Je tomu tak preto, že tento tvar zodpovedá podmienenému rozdeleniu pravdepodobnosti. Prečo je tomu tak? Pre krivky na našom obrázku platí, že sú rovnobežné s osou X_1. Znamená to teda, že pre všetky body na nej je hodnota súradnice X_2 rovnaká. To presne zodpovedá situácii, v ktorej uvažujeme o pravdepodobnostnom rozdelení podmienenej hustoty pravdepodobnosti pre konštantné X_2.

Zároveň ale platí, že daná krivka nie je naozaj podmienenou hustotou. Dôvod je jednoduchý – aby naozaj išlo o pravdepodobnostné rozdelenie, integrál z danej funkcie od mínus nekonečna po nekonečno by mal byť rovný jednej – len tak môže ísť o hustotu pravdepodobnosti. V príklade uvedenom v závere predchádzajúcej podkapitoly sme videli, že hodnoty je nutné upraviť o marginálnu pravdepodobnosť tej náhodnej premennej, ktorá je konštantná. V tomto prípade by sme funkciu s príslušnou krivkou rovnobežnou s osou X_1 museli ešte vydeliť marginálnou hustotou pravdepodobnosti, že náhodná premenná X_2 nadobudne hodnotu zodpovedajúcu konštante. Dostali by sme podmienenú hustotu pravdepodobnosti $f_{X_1|X_2}$.

Podobne sa dá postupovať aj vtedy, ak by sme sledovali krivky rovnobežné s osou X_2, čo znamená, že by sme skúmali body, v ktorých je súradnica v smere osi X_1 konštantná. V tomto prípade by sme skúmali krivku, ktorá súvisí s opačnou podmienenou hustotou pravdepodobnosti ($f_{X_2|X_1}$).

4.7.4 Viacrozmerné normálne rozdelenie

Viacrozmerné normálne rozdelenie je prirodzeným rozšírením dvojrozmerného normálneho rozdelenia, ktoré sme si predstavili v predchádzajúcej časti. Uvažujeme pritom o náhodnom vektore $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ pre $n \in \mathbb{N}$, ktorého zložkami sú náhodné premenné. Ak platí, že všetky lineárne kombinácie náhodných premenných:

$$a_1X_1 + a_2X_2 + \ldots + a_nX_n \tag{4.99}$$

kde $a_1, a_2, \ldots, a_n \in \mathbb{R}$, majú jednorozmerné normálne rozdelenie pravdepodobnosti, potom hovoríme, že náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, pre $n \in \mathbb{N}$ má viacrozmerné normálne rozdelenie.

Vzťahy, ktoré sme si prezentovali pri dvojrozmernom normálnom rozdelení zostávajú v platnosti. Predstavujú špeciálny prípad viacrozmerného normálneho rozdelenia.

Pri porovnaní hustoty pravdepodobnosti jednorozmerného a združeného dvojrozmerného rozdelenia pravdepodobnosti vidíme, že jej zápis sa so zvyšujúcim sa

rozmerom značne komplikuje. Z dôvodu úspornosti je v prípade rozdelení s vyššou dimenziou často využívaný maticový zápis, ktorý je ekvivalentný a nesporne kratší.

S využitím maticového zápisu je možné združenú hustotu viacrozmerného normálneho rozdelenia (ak existuje) zapísať nasledovne:

$$f_{\mathbf{X}}(x_1, x_2, ..., x_n) = \frac{1}{\sqrt{(2\pi)^n \mid \mathbf{\Sigma_X} \mid}} \exp\left[-\frac{1}{2}\left(\mathbf{X} - E(\mathbf{X})\right)^T \mathbf{\Sigma}_{\mathbf{X}}^{-1}\left(\mathbf{X} - E(\mathbf{X})\right)\right] \qquad (4.100)$$

kde $|\mathbf{\Sigma_X}|$ predstavuje determinant variančno-kovariančnej matice a $(\mathbf{X} - E(\mathbf{X}))^T$ transponovanú maticu k $(\mathbf{X} - E(\mathbf{X}))$.

4.7.5 Viacrozmerná centrálna limitná veta

Pri jednorozmerných rozdeleniach pravdepodobnosti sme si predstavili jeden z kľúčových záverov z teórie pravdepodobnosti, nazývaný centrálna limitná veta. Jej podstatou je tvrdenie, že aritmetický priemer nezávislých náhodných premenných s tým istým rozdelením pravdepodobnosti má asymptoticky normálne rozdelenie. Toto tvrdenie je možné rozšíriť aj pre prípad viacrozmerných rozdelení – podobne ako v jednorozmernom prípade, aj tu zohráva kľúčovú úlohu (viacrozmerné) normálne rozdelenie.

V jednorozmernom prípade sme skúmali $n \in \mathbb{N}$ náhodných premenných. Teraz budeme uvažovať o tom, že budeme mať n nezávislých náhodných vektorov $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ s tým istým rozdelením pravdepodobnosti. Predpokladajme ďalej, že každý z týchto vektorov má $k \in \mathbb{N}$ prvkov.

Definujme ďalej vektor stredných hodnôt $E(\mathbf{X})$ ako k-rozmerný vektor, ktorého prvkami sú stredné hodnoty náhodných vektorov. Pripomeňme, že vektory $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ majú podľa predpokladu to isté rozdelenie pravdepodobnosti, takže platí:

$$E(\mathbf{X}_1) = E(\mathbf{X}_2) = ... = E(\mathbf{X}_n) \qquad (4.101)$$

Podobne všetky tieto náhodné vektory majú variančno-kovariančnú maticu $\mathbf{\Sigma_X}$.

Ako posledný definujme vektor priemerov:

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \qquad (4.102)$$

Centrálna limitná veta pre viacrozmerné rozdelenia hovorí, že náhodný vektor:

$$\sqrt{n}\left(\overline{\mathbf{X}} - E(\mathbf{X})\right) \qquad (4.103)$$

má asymptoticky viacrozmerné združené normálne rozdelenie s nulovým vektorom stredných hodnôt a variančno-kovariančnou maticou $\mathbf{\Sigma_X}$.

4.7.6 Wishartovo, Hotellingovo a Wilksovo rozdelenie

V teórii pravdepodobnosti sa po definovaní normálneho rozdelenia pre náhodné premenné postupuje spravidla tým, že sa definujú ďalšie významné pravdepodobnostné rozdelenia odvodené od normálneho. To zahŕňa napr. Chí-kvadrát rozdelenie pre súčet druhých mocnín nezávislých normovaných normálnych náhodných premenných, Studentovo *t*-rozdelenie, ako aj *F*-rozdelenie.

Podobne aj v prípade viacrozmerných rozdelení pravdepodobnosti existujú odvodené rozdelenia, ktoré sa v štatistike využívajú napríklad na testovanie hypotéz v rámci induktívnej štatistiky.

Analógiou jednorozmerného rozdelenia Chí-kvadrát je tzv. Wishartovo rozdelenie. Pripomeňme, že jednorozmerné Chí-kvadrát rozdelenie vznikne ako pravdepodobnostné rozdelenie súčtu:

$$C = \sum_{i=1}^{n} X_i^2 \tag{4.104}$$

kde $n \in \mathbb{N}$ a X_i pre $i = 1, 2, ..., n$ sú nezávislé náhodné premenné s normovaným normálnym rozdelením pravdepodobnosti ($N(0,1)$). X_i majú strednú hodnotu rovnú nule a rozptyl rovný jednej. Definovaná náhodná premenná C má potom Chí-kvadrát rozdelenie pravdepodobnosti s n stupňami voľnosti ($C \sim \chi^2(n)$).

Ak by sme namiesto náhodných premenných X_i pracovali s m-rozmernými náhodnými vektormi $\mathbf{X}_i$ pre $i = 1,2,\ldots, n$ a $n, m \in \mathbb{N}$, budeme predpokladať, že každý z náhodných vektorov bude mať to isté m-rozmerné združené normálne rozdelenie pravdepodobnosti $N_m(\mathbf{0}, \mathbf{\Sigma_X})$ s nulovým vektorom stredných hodnôt a variančno-kovariančnou maticou $\mathbf{\Sigma_X}$.

S pomocou týchto vektorov definujeme maticu obsahujúcu náhodné premenné v tvare:

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \tag{4.105}$$

Ani v tomto prípade sa nevyhneme maticovému zápisu. Analógia s rozdelením Chí-kvadrát je však zrejmá – náhodnú premennú C sme dostali ako súčet druhých mocnín náhodných premenných s normálnym rozdelením pravdepodobnosti. Druhú mocninu X_i^2 je však možné taktiež napísať ako X_iX_i, vďaka čomu by sme mohli vzorec *(4.104)* prepísať nasledovne:

$$C = \sum_{i=1}^{n} X_i X_i \tag{4.106}$$

Matica $\mathbf{W}$ a náhodná premenná C sú preto veľmi podobné – líšia sa len charakterom objektov, ktorých súčin sčítavame (náhodné premenné v prípade C a náhodné vektory

v prípade **W**) a tým, že v prípade **W** je druhá matica v súčine transponovaná. Ak by sme v náhodnej premennej X_i priradili rozmer 1 × 1, potom by dokonca v istom zmysle platilo aj:

$$C = \sum_{i=1}^{n} X_i X_i^T \tag{4.107}$$

a jediný rozdiel by bol naozaj len v charaktere objektov.

Pri definícii **W** sčítavame prvky $\mathbf{X}_i\mathbf{X}_i^T$. Ak by sme na tento súčin aplikovali operátor strednej hodnoty $E\left(\mathbf{X}_i\mathbf{X}_i^T\right)$, dostali by sme variančno-kovariančnú maticu $\mathbf{\Sigma_X}$. Ak by sme považovali náhodné vektory $\mathbf{X}_i$ za výberovú vzorku, potom má matica **W** úzky súvis s výpočtom výberovej variančno-kovariančnej matice. Z tohto dôvodu je užitočné odvodiť si aj jej pravdepodobnostné rozdelenie, ktoré je ďalej možné využívať napríklad pri testovaní štatistických hypotéz. Pravdepodobnostné rozdelenie matice **W** danej vzťahom *(4.105)* nazývame *m*-rozmerným **Wishartovým rozdelením pravdepodobnosti**.

Ak existuje, Wishartovo rozdelenie má hustotu pravdepodobnosti:

$$f(\mathbf{W}) = \frac{|\mathbf{W}|^{\frac{n-m-1}{2}} \exp\left[-\frac{1}{2} tr\left(\mathbf{\Sigma}_\mathbf{X}^{-1}\mathbf{W}\right)\right]}{2^{\frac{nm}{2}} |\mathbf{\Sigma_X}|^{\frac{n}{2}} \Gamma_m\left(\frac{n}{2}\right)} \tag{4.108}$$

Kde *tr* je takzvaná stopa matice (angl. *trace*), $|\mathbf{W}|$ a $|\mathbf{\Sigma_X}|$ sú determinanty príslušných matíc. Výraz Γ_m je viacrozmerná gama funkcia v tvare:

$$\Gamma_m\left(\frac{n}{2}\right) = \pi^{\frac{m(m-1)}{4}} \prod_{j=1}^{m} \frac{n+1-j}{2} \tag{4.109}$$

Wishartovo rozdelenie pravdepodobnosti má dva parametre: počet stupňov voľnosti *n*, ktorý je rovný počtu sčítavaných vektorov $\mathbf{X}_i$ a ich variančno-kovariančnú maticu $\mathbf{\Sigma_X}$.

Pri popise jednorozmerných pravdepodobnostných rozdelení sa spravidla postupuje tak, že po definícii normálneho rozdelenia odvodíme rozdelenie Chí-kvadrát, s pomocou normálneho rozdelenia a rozdelenia Chí-kvadrát sa potom následne definuje Studentovo *t*-rozdelenie. Pripomeňme, že ak má náhodná premenná *X* normované normálne rozdelenie *N*(0, 1), náhodná premenná *V* má pravdepodobnostné rozdelenie Chí-kvadrát s $n \in \mathbb{N}$ stupňami voľnosti $\chi^2(n)$ a *X* a *V* sú vzájomne nezávislé, potom náhodná premenná:

$$\sqrt{n} X V^{-\frac{1}{2}} \tag{4.110}$$

má Studentovo *t*-rozdelenie s *n* stupňami voľnosti (*t*(*n*)).

Tento prípad rozšírime na viacrozmerný, nech $\mathbf{X}_i$ pre $i = 1, 2, ..., n$, $n \in \mathbb{N}$ sú *m*-rozmerné ($m \in \mathbb{N}$) náhodné vektory so združeným normálnym rozdelením

pravdepodobnosti $\mathbf{X}_i \sim N_m(\mathbf{0}, \mathbf{\Sigma_X})$ a $\mathbf{W}$ je matica s Wishartovým rozdelením pravdepodobnosti $\mathbf{W} \sim W(n, \mathbf{\Sigma_X})$. Potom o náhodnej premennej:

$$T^2 = n\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X} \tag{4.111}$$

hovoríme, že má **Hotellingovo T^2** rozdelenie pravdepodobnosti s dvoma parametrami, m a n.

Z praktického hľadiska počítame hodnoty Hotellingovho rozdelenia priamo len málokedy. Existuje totiž vzťah, ktorým môžeme hodnoty distribučnej funkcie Hotellingovho T^2 rozdelenia previesť na hodnoty distribučnej funkcie F-rozdelenia. Ak má náhodná premenná Y Hotellingovo T^2 rozdelenie pravdepodobnosti s parametrami m a n, teda $Y \sim T^2(m, n)$, potom platí:

$$\frac{n-m+1}{mn} Y \sim F(m, n-m+1) \tag{4.112}$$

kde F označuje jednorozmerné Fisherovo F-rozdelenie pravdepodobnosti.

Hotellingovo viacrozmerné pravdepodobnostné rozdelenie T^2 úzko súvisí aj s ďalšou zaujímavou mierou, ktorá sa používa vo viacrozmernej štatistike. Tou je takzvaná **Mahalanobisova vzdialenosť**. Využíva sa v prípade, ak by sme chceli definovať vzdialenosť medzi dvoma náhodnými vektormi $\mathbf{X}$ a $\mathbf{Y}$ z toho istého rozdelenia pravdepodobnosti, ktoré majú variančno-kovariančnú maticu $\mathbf{\Sigma_{XY}}$. Mahalanobisova vzdialenosť $d(\mathbf{X},\mathbf{Y})$ je definovaná nasledovne:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})^T \mathbf{\Sigma_{XY}} (\mathbf{X}-\mathbf{Y})} \tag{4.113}$$

Zo zápisu je zrejmé, že ak sú náhodné premenné tvoriace vektory $\mathbf{X}$ a $\mathbf{Y}$ nekorelované, a všetky majú disperzie rovné jednej, matica $\mathbf{\Sigma_{XY}}$ bude jednotkovou maticou a Mahalanobisova vzdialenosť bude rovná bežnej Euklidovskej vzdialenosti medzi dvoma bodmi. Ak tomu tak nie je, a napríklad zložky vektorov $\mathbf{X}$ a $\mathbf{Y}$ sú korelované, potom vzdialenosť týchto vektorov nebude závisieť len od ich hodnôt, a teda umiestnení v priestore, ale aj od ich vzájomného vzťahu definovanom korelačnou štruktúrou.

Praktické aplikácie Mahalanobisovej vzdialenosti zahŕňajú napríklad klasifikáciu, pri ktorej sa snažíme posúdiť, či nejaký náhodný vektor patrí do pravdepodobnostného rozdelenia spolu s inou skupinou vektorov. S pomocou skupiny vektorov, o ktorých vieme, že tvoria jedno pravdepodobnostné rozdelenie dokážeme pomocou Mahalanobisovej vzdialenosti posúdiť, či je skúmaný vektor od tohto rozdelenia „príliš ďaleko“, a teda či je nepravdepodobné, že patrí do daného rozdelenia. Vo všeobecnosti sa Mahalanobisova vzdialenosť využíva v klasifikačnej a zhlukovej analýze.

Posledné viacrozmerné rozdelenie, ktoré si priblížime, sa nazýva **Wilksovo rozdelenie** lambda (v ďalšom texte ho budeme skrátene nazývať len Wilksovo rozdelenie). Využíva sa jednak v ekonometrii, kde sa používa pri testoch podielov vierohodností (angl. *likelihood ratio tests*), ale aj vo viacrozmernej štatistike pri viacrozmernej analýze rozptylu (MANOVA).

Wilksovo rozdelenie sa svojím pôvodom podobá na jednorozmerné *F*-rozdelenie. Toto jednorozmerné rozdelenie pravdepodobnosti vzniká pri analýze rozdelenia podielu dvoch náhodných premenných, ktoré majú Chí-kvadrát rozdelenie. Využíva sa v prípadoch, ak by sme chceli napríklad porovnávať rozptyly dvoch náhodných premenných, resp. v ekonometrii pri lineárnej regresii, kde pomocou neho porovnávame podiel vysvetlenej a celkovej variability za účelom posúdenia kvality a významnosti kvantifikovaného modelu.

V predchádzajúcich odsekoch sme videli, že viacrozmernou analógiou jednorozmerného Chí-kvadrát rozdelenia je Wishartovo rozdelenie pravdepodobnosti. Namiesto dvoch náhodných premenných s rozdelením Chí-kvadrát preto tentoraz začnime s dvoma premennými, ktoré majú práve Wishartovo rozdelenie. Zachovaná zostane aj ďalšia analógia – podobne ako v jednorozmernom prípade predstavuje Studentovo *t*-rozdelenie špeciálny prípad *F*-rozdelenia, je aj Hotellingovo T^2 rozdelenie špeciálnym prípadom Wilksovho rozdelenia pravdepodobnosti.

Nech majú matice $\mathbf{W}_1$ a $\mathbf{W}_2$ *l*-rozmerné Wishartovo rozdelenie s tou istou variančno-kovariančnou maticou $\mathbf{\Sigma_X}$ a počtom stupňov voľnosti *m* a *n* (teda $\mathbf{W}_1 \sim W(m, \mathbf{\Sigma_X})$ a $\mathbf{W}_2 \sim W(n, \mathbf{\Sigma_X})$). Potom náhodná premenná definovaná ako podiel determinantov:

$$\lambda = \frac{|\mathbf{W}_1|}{|\mathbf{W}_1 + \mathbf{W}_2|} = \frac{1}{|\mathbf{\Sigma_X} + \mathbf{W}_1^{-1}\mathbf{W}_2|} \qquad (4.114)$$

má Wilksovo rozdelenie lambda s parametrami *l*, *m* a *n*. Túto skutočnosť zapisujeme ako:

$$\lambda \sim \Lambda(l, m, n) \qquad (4.115)$$

Pre veľké hodnoty *m* je možné získať aproximáciu tohto rozdelenia pomocou jednorozmerného rozdelenia pravdepodobnosti Chí-kvadrát.

5 Príklady

5.1 Zadania príkladov

Príklad 5.1

Idete s autom pol hodinu rýchlosťou 40 km/h a druhú polhodinu rýchlosťou 80 km/h. Aká bola po jednej hodine vaša priemerná rýchlosť?

Príklad 5.2

Idete s autom prvú polovicu cesty 40 km/h a druhú polovicu cesty rýchlosťou 80 km/h. Aká bola na konci cesty vaša priemerná rýchlosť?

Príklad 5.3

Do banky ste vložili 100,- EUR. V prvom roku sa vaša investícia zhodnotila o 1.05 % v druhom roku o 8.2 % a v treťom roku o 3.1 %. Počas celého obdobia ste vložené peniaze z banky nevyberali. Aký je priemerný ročný rast vašej investície?

Príklad 5.4

K 1.1.2009 je stav nedokončenej výroby 400 ks, k 1.2.2009 je stav nedokončenej výroby 200 ks a k 1.3.2009 je stav nedokončenej výroby 400 ks. Aký je priemerný stav nedokončenej výroby od začiatku roka až po 1.3.2009?

Príklad 5.5

Z databázy EUROBAROMETER (dostupné online [03.10.2011] na `http://ec.europa.eu/public_opinion/cf/index_en.cfm`) je možné získať odpovede respondentov vo vybraných krajinách na nasledujúcu otázku:

Vo všeobecnosti považujete členstvo vašej krajiny v Európskom spoločenstve (v spoločnom trhu): a) za dobrú vec, b) za zlú vec, c) za ani dobrú ani zlú vec, d) neviem. Polročné výsledky pre roky 2005 až prvý polrok 2010 za Slovensko a Českú republiku sú v nasledujúcich vektoroch (ide o časové rady):

```
svk_good <- c(0.54, 0.50, 0.55, 0.61, 0.64, 0.58, 0.57, 0.62, 0.66, 0.68, 0.59)
svk_bad <-  c(0.06, 0.07, 0.06, 0.06, 0.06, 0.06, 0.06, 0.05, 0.03, 0.05, 0.07)
cze_good <- c(0.49, 0.44, 0.52, 0.51, 0.46, 0.45, 0.48, 0.46, 0.42, 0.40, 0.31)
cze_bad <- c(0.11, 0.11, 0.09, 0.10, 0.12, 0.15, 0.11, 0.12, 0.13, 0.13, 0.16)
```

Vašou úlohou je:

- Porovnajte rozdiely medzi Slovenskom a Českou republikou. Kedy boli rozdiely vo vnímaní najväčšie a kedy najmenšie?
- Vypočítajte rozsah hodnôt pre jednotlivé dátové vektory. O čom nám hovoria?
- V koľkých prípadoch viac ako 60 % respondentov na Slovensku považovalo členstvo v Európskom spoločenstve za dobrú vec? V koľkých prípadoch viac ako 50 % respondentov v Českej republike považovalo členstvo v Európskom spoločenstve za dobrú vec?
- Vytvorte dátové vektory, ktoré budú monitorovať zmenu v %-ách. V ktorých krajinách a v ktorých premenných môžeme vidieť najväčšie zmeny a kedy? Počítajte v nominálnych hodnotách, ale aj v percentách.
- Čo vytvára nasledujúci príkaz? (nepoužívajte ho priamo v R, zistite to len jeho vizuálnym skúmaním):

```
p <- c(svk_good[seq(from = 1,to = length(svk_good), by = 2)])
```

Príklad 5.6

Nahrajte databázu Quarterly Earnings per Share spoločnosti Johnson & Johnson z programového balíka „datasets“ (`?datasets`). Zistite si, čo obsahuje táto databáza. Keď budete poznať jej názov, použite operátor „`?`“. Rozdeľte databázu do dvoch dátových vektorov. Prvý by mal obsahovať prvých 42 údajov o zisku na akciu a druhý zvyšné údaje. Porovnajte výšku zisku (napr. cez priemer, medián). Vytvorte štyri dátové vektory. Každý z nich by mal obsahovať (chronologicky) iba zisk z rovnakých kvartálov. Pomocou funkcie `plot.ts()` zobrazte hodnoty kvartálnych ziskov. Vytvorte vektor percentuálnych zmien ziskov. Zobrazte ho pomocou funkcie `plot.ts()`. Vytvorte vektor percentuálnych zmien ziskov kvartálu k rovnakému kvartálu predchádzajúceho roku. Zobrazte ho pomocou funkcie `plot.ts()`. V čom by mal byť rozdiel?

Príklad 5.7

V tomto príklade budeme pracovať s databázou `Anscombe` (knižnica `car`), z ktorej nás budú zaujímať údaje o výdavkoch na vzdelávanie na jednu osobu (`education`) a príjem na jednu osobu (`income`). Údaje sú z roku 1970 za 51 štátov z USA. Zdá sa byť rozumné predpokladať, že výdavky na vzdelávanie a príjmy nebudú vo všetkých štátoch rovnaké. Aby

sme si urobili prehľad o týchto údajoch, zostrojte histogram týchto dvoch premenných a vypočítajte priemer, smerodajnú odchýlku, rozptyl, minimum a maximum.

Príklad 5.8

Databáza `Salaries` (knižnica `car`) obsahuje 397 pozorovaní týkajúcich sa platov vysokoškolských učiteľov z USA. Ide o 9 mesačné platy zo semestra 2008 – 2009. Zostrojte box – plot platov (premenná `salary`) pre mužov a ženy. Pre tieto dve skupiny vypočítajte priemer, smerodajnú odchýlku, minimum a maximum. Zaujíma nás, aké sú rozdiely vo výške platov medzi mužmi a ženami.

Príklad 5.9

Z makroekonomických údajov databázy `unemployment` (knižnica `lmtest`) vypočítajte základnú deskriptívnu štatistiku. Táto databáza obsahuje údaje o nezamestnanosti (`UN`), peňažnom agregáte (`m`), deflátor (`p`), reálny dopyt po tovaroch a službách (`g`), reálny export (`x`). Tieto údaje korešpondujú ročným údajom pre USA v období od 1890 do 1979. Je zrejmé, že všetky makroekonomické ukazovatele sa v čase menia. Nás bude zaujímať nezamestnanosť. Deskriptívna štatistika nám pomôže odpovedať na otázky, aká bola priemerná miera nezamestnanosti v danom období v USA, alebo nakoľko je nezamestnanosť variabilná.

Príklad 5.10

V databáze `UN` z knižnice `car` máme k dispozícii údaje o úmrtnosti detí na 1000 narodených detí a HDP per capita v 207 krajinách sveta za rok 1998. Údaje sú z databázy OSN. Vypočítajte deskriptívnu štatistiku pre tieto dve premenné a vizualizujte údaje tak, aby sme mohli vidieť prípadný vzťah medzi HDP per capita a úmrtnosťou detí. Bez toho, aby sme dopredu vedeli, aký je vzťah medzi týmito premennými, zrejme môžeme očakávať, že v krajinách s vyšším HDP per capita bude úmrtnosť detí nižšia.

Príklad 5.11

V tomto príklade využijeme databázu `survey` (z knižnice `MASS`), ktorá obsahuje odpovede 237 študentov z Univerzity v Adelaide na rôzne otázky. Vypočítajte deskriptívnu štatistiku pre premenné výška (`Height`) a vek (`Age`). Porovnajte výšku mužov a žien v danej vzorke. Zrejme je jasné, že muži sú vyšší, ale ako je to s variabilitou?

Príklad 5.12

Z nasledujúcej postupnosti čísel vypočítajte (bez použitia funkcií v R) aritmetický, geometrický a harmonický priemer: 7, 5, 21, 15, 47, 34, 9, 42, 19, 68, 83.

Príklad 5.13

Z nasledujúcej postupnosti čísel vypočítajte aritmetický, geometrický a harmonický priemer: 2, 15, 28, 7, 43, 31, 19, 84, 13, 76, 37. Použite manuálny (prepočet ako v predchádzajúcom príklade) a pokúste sa výpočet overiť aj s využitím funkcií v programe R.

Príklad 5.14

Ukazovateľ označovaný ako P/E (z angl. *price – earnings ratio*) vyjadruje pomer trhovej ceny akcie k účtovnému zisku (spravidla k ročnému), teda koľko sú investori ochotní zaplatiť za jednu menovú jednotku vykázaného zisku. Z toho teda vyplýva, že udáva, za koľko rokov bude splatená trhová cena akcie prostredníctvom zisku, ktorý generuje (za podmienok ceteris paribus). Tento ukazovateľ môže slúžiť na porovnávanie spoločností, ktoré pochádzajú z pravidla jedného a toho istého odvetvia. Najjednoduchšie rozhodnutie by mohlo vychádzať z tvrdenia, že čím je väčšie P/E, tým je akcia na trhu viac nadhodnotená, a preto je vhodné ju predať, naopak, čím je P/E menšie, tým je akcia podhodnotená trhom, a preto ju treba kúpiť (Baumöhl et al., 2011, s. 220). Pri pomerne zjednodušenom uvažovaní, tak spoločnosti s podpriemerným P/E budeme považovať za podhodnotené a tie, ktoré majú P/E väčšie ako priemer za nadhodnotené. Vypočítajte priemernú výšku ukazovateľa P/E v rámci odvetvia Telekomunikácie. V tomto odvetví sú hodnoty ukazovateľa P/E nasledujúce: 12, 16, 25, 8, 68, 27, 19, 31, 24.

Príklad 5.15

Mesačné výnosy plynúce z vašej investície v priebehu jedného roka boli nasledujúce: 12 %, –16 %, –25 %, –34 %, 8 %, 17 %, 27 %, –19 %, 31 %, 24 %, –40 %, 21 %. Vypočítajte priemerný výnos za daný rok.

Príklad 5.16

V tomto príklade sa bližšie pozrieme na dáta týkajúce sa prebiehajúcej dlhovej krízy v 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 (dáta sú dostupné aj v súbore `debt_gdp.csv` na `www.econometrics.sk`). 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).

Najprv vypočítajte základnú deskriptívnu štatistiku pre tieto krajiny a zistite, ktoré krajiny majú v priemere najvyšší pomer dlhu k HDP. Potom sa pokúste zistiť, či sa tieto priemerné hodnoty zmenili (zvýšili/znížili) od roku 2008. Na záver zobrazte údaje graficky tak, aby bolo možné z grafu jednoducho odlíšiť hodnoty z pred roku 2008 a od roku 2008 (zjednodušene môžeme povedať „pred krízou“ a „počas krízy“).

Príklad 5.17

V databáze `survey` (z knižnice `MASS`), máme k dispozícii odpovede 237 študentov z Univerzity v Adelaide na rôzne otázky. V tomto príklade nás bude zaujímať len výška študentov tejto univerzity. Zobrazte výšku študentov v podobe box – plotov zvlášť pre mužov a zvlášť pre ženy. Do grafov naneste tiež priemernú výšku v rámci danej skupiny študentov. Rozhodnite, či existuje rozdiel vo výške mužov a žien na danej univerzite len na základe vizualizácie dát.

Príklad 5.18

V databáze `EuStockMarkets` (knižnica `datasets`) 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. Z daných uzatváracích cien vypočítajte tzv. spojité výnosy podľa vzťahu $\ln(P_{t+1}/P_t)$, kde P_t je uzatváracia cena indexu v čase t. Vypočítajte deskriptívnu štatistiku pre uzatváracie ceny a taktiež pre spojité výnosy. Aby ste získali lepšiu predstavu o daných časových radoch, zobrazte obe skupiny do vizuálnej podoby. Pre spojité výnosy načrtnite graf box – plot a rozhodnite, ktorý index dosahuje najvyššiu priemernú výnosnosť. Túto priemernú výnosnosť zobrazte aj v box – plotoch.

Príklad 5.19

Z databázy RegDat Štatistického úradu Slovenskej republiky máme k dispozícii údaje za rok 2010 o nominálnych mzdách a nezamestnanosti po okresoch v rámci SR. Mzdy sme upravili o infláciu a získali sme tak reálne mzdy. 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). Vytvorená databáza je dostupná v súbore `sk_data` na `www.econometrics.sk`.

Prostredníctvom grafickej vizualizácie (box – plotov) reálnych miezd a nezamestnanosti zistite, či existujú rozdiely v týchto dvoch makroekonomických premenných medzi jednotlivými krajmi.

Príklad 5.20

Pracujte s databázou `cfb`. Rozhodnite, pri ktorých premenných by ste radšej na výpočet ukazovateľa polohy použili medián a kde skôr aritmetický priemer. Vytvorte tri dátové vektory z premennej `VEHIC`. V prvom bude chýbať 5 % najväčších hodnôt, v druhom 10 % a v treťom 15 %. Pozrite si histogram pre tieto dátové vektory. Porovnajte priemer a medián pri týchto dátových vektoroch. Ako sa mení tvar rozdelenia početností? Okrem histogramu, skúste použiť aj box – plot na porovnanie tvaru rozdelenia.

Príklad 5.21

Pracujte s databázou `normtemp`, ktorá obsahuje rôzne merania vykonané na 130 zdravých, náhodne vybratých ľuďoch. Premenná `temperature` zobrazuje nameranú teplotu tela. Vytvorte dátový vektor, v ktorom bude teplota tela v stupňoch Celzia (v premennej `temperature` sú jednotky Fahrenheit). Následne vypočítajte priemernú teplotu a jej variabilitu. Zmerajte si teplotu a určite, akému kvantilu vami nameraná teplota zodpovedá. Koľko ľudí vo vzorke malo väčšiu teplotu ako vy? Zostrojte histogram a graf box – plot. Súdiac podľa týchto grafov, aký interval teploty by ste považovali za „normálny“?

Príklad 5.22

Premenná `exec.pay` z programového balíka `UsingR` má výrazne pravostranne zošikmené rozdelenie (zostrojte histogram pre získanie tejto informácie). Ide o príjem riaditeľov

vybraných spoločností v USD. Uskutočnite nasledujúcu transformáciu hodnôt: `log(1 + exec.pay, 10)`. Zostrojte histogram a porovnajte s predošlým.

Príklad 5.23

Vytvorte stĺpcový graf, koláčový graf a bodový graf z nasledujúcich údajov o používaní webových prehliadačov [údaje boli dostupné online k dátumu 17.02.2012 na http://www.w3counter.com/globalstats.php?year=2012&month=1]: Internet Explorer 30.9 %, Firefox 24.8 %, Chrome 24.6 %, Safari 6.5 %, Opera 2.5 %. Rovnako pre používanie operačných systémov: Windows 7 38.53 %, Windows XP 30.55 %, Apple OS X 8.89 %, Windows Vista 8.64 %, Apple iOS 4.45 %, Android 1.82 %, Linux 1.6 %, ostatné 5.52 %.

Príklad 5.24

Použite premennú `mpg` z databázy `mtcars` (programový balík `UsingR`). Najprv zobrazte údaje. Každý riadok má názov. Vytvorte bodový graf, kde každé pozorovanie na osi y-ovej bude mať príslušné označenie zodpovedajúce názvu (značky) auta.

Príklad 5.25

Použite databázu `npdb` (programový balík `UsingR`). Databáza obsahuje priestupky lekárov v USA. Premenná ID predstavuje identifikačný údaj lekára. Vytvorte tabuľku, ktorá zobrazí koľko lekárov malo iba jeden priestupok, koľko lekárov malo dva priestupky, atď..

Príklad 5.26

Použite funkciu `attach()` na priradenie databázy `MLBattend` do aktuálnej pracovnej sekcie v programe R. Vytvorte dátový vektor, ktorý bude zobrazovať návštevnosť na zápasoch New York Yankees (v premennej "`franchise`" názov `NYA`). Návštevnosť zoraďte chronologicky a vytvorte bodový graf. Čo ovplyvňovalo návštevnosť zápasov (rozmýšľajte nad možnými ekonomickým dôvodmi)?

Príklad 5.27

Finančná spoločnosť vyhodnocovala svojich predajcov na konci kvartálov, pričom vychádzala z celkových predajných výsledkov za posledný kvartál. Objem predaných služieb vyjadrený v EUR za jednotlivých zamestnancov v okrese A a v okrese B je nasledovný:

Okres A

4500, 4800, 5200, 4100, 4200, 3500, 2500, 4500, 2400, 4700, 3300, 2800, 4700, 5300, 4400, 4700, 4900, 5100, 3800, 4600, 3800, 6400, 2500, 4100, 4400, 3800, 2600, 5200.

Okres B

2400, 3100, 7200, 3100, 5200, 6200, 4000, 3200, 5300, 3900, 3600, 5300, 5400, 3300, 4700, 3700, 3000, 3300, 4400, 4200, 5700, 4700, 4900, 3900, 2000, 4800, 4100, 4100.

Zistite, v ktorom okrese vykazovali zamestnanci priemerne vyšší objem predaných služieb? Porovnajte mieru variability objemu predaných služieb medzi okresmi A a B. Interpretujte šikmosť a špicatosť objemu predaných služieb v oboch okresoch. U ktorých zamestnancov (podľa objemu predaja) sa bude bližšie vyšetrovať príčina nižšieho objemu predaja, ak bodom zvratu je 25-ty percentil z oboch okresov?

Príklad 5.28

Z databázy homedata (`UsingR`) vytvorte histogram (relatívnej početnosti) z premennej `price_ratio = y2000/y1970`. Do tohto histogramu naneste odhadovanú funkciu hustoty. Popíšte tvar rozdelenia. Potom vypočítajte šikmosť a špicatosť.

Príklad 5.29

Použite databázu `state` (programový balík `datasets`). Najprv vytvorte `data.frame()` nasledovným spôsobom: `x77 = data.frame(state.x77)`. Následne vytvorte *x-y* grafy nasledovných kombinácií: `population - frost, population - murder, population - area, income - hs grad, income - murder`. Ktoré z týchto vzťahov vám pripadajú silné? Ak niektoré áno, aké vysvetlenie preto máte?

Príklad 5.30

V programovom balíku `UsingR` je databáza `normtemp`, ktorá obsahuje telesnú teplotu 130 zdravých jednotlivcov. Premenná `gender` predstavuje pohlavie, pričom jej kódovanie je nasledovné: 1 – muži, 2 – ženy.

Majú muži a ženy rovnakú telesnú teplotu? Kto má väčšiu variabilitu telesnej teploty? Okrem numerických štatistík, nájdite vhodnú vizuálnu prezentáciu dát, kde by rozdiely vo variabilite boli čo najviac viditeľné (ak sú).

Sú rozdiely medzi telesnými teplotami žien a mužov významne odlišné? Zdôvodnite svoje rozhodnutie.

Príklad 5.31

Použite databázu `samhda` (`UsingR`), ktorá monitoruje určité návyky tínedžerov. Budú nás zaujímať tri premenné: pohlavie (`gender`), počet dní z posledných 30, počas ktorých študent fajčil (`amt.smoke`) a či niekedy fajčil marihuanu (`marijuana`). Vyberte vhodné vizuálne prostriedky k tomu, aby sme videli, či existujú v týchto premenných určité vzťahy.

Príklad 5.32

Stiahnite si uzatváracie ceny (v angl. *adjusted close*) od 1.1.2000 po posledný piatok aj s dátumami z troch indexov: S&P 500, FTSE 100 and Nikkei 225 (použite údaje zo stránky http://finance.yahoo.com/). Importujte údaje do programu R ako dátové vektory. Všimnite si, že pri niektorých dátumoch existuje údaj v jednom indexe a pri inom indexe údaj nemusí existovať (napríklad štátny sviatok). Uskutočnite tzv. *Listwise deletion* a spojte dátové vektory do databázového objektu (http://en.wikipedia.org/wiki/Listwise_deletion).

Príklad 5.33

Vychádzajme z kvantitatívnej teórie peňazí. Ak je M množstvo peňazí v ekonomike (napr. agregát M2), V je rýchlosť obehu peňazí v ekonomike, P je cenová hladina (pre tieto účely spravidla meraná pomocou HDP deflátora) a Y je reálne HDP, potom základná rovnica má tvar: $MV = PY$. Túto rovnicu je možné prepísať do nasledujúcej podoby $\Delta M\% + \Delta V\% = \Delta P\% + \Delta Y\%$ (ide o približné riešenie, ktoré platí ak tieto percentuálne zmeny nie sú veľmi veľké, rádovo menej ako 10 %). Kvantitatívna teória peňazí hovorí, že ak centrálna banka riadi ponuku peňazí v ekonomike a rýchlosť obehu peňazí v ekonomike je konštantná, potom zvýšenie ponuky peňazí v ekonomike spôsobuje zvýšenie cenovej hladiny (inflácie). Zmena reálneho HDP sa v tomto prípade nepovažuje za podstatnú. Predpokladá sa, že uvedený vzťah platí dlhodobo.

Vašou úlohou je toto tvrdenie formálne (ako aj vizuálne) overiť na vybranej členskej krajine skupiny OECD. Použite agregáty M1 a M3, ktoré je možné získať napr. z http://stats.oecd.org/index.aspx, pričom pre zmenu cenovej hladiny použite spotrebiteľský cenový index (CPI).

Následne skúste overiť uvedenú hypotézu pre vybranú skupinu aspoň 20 krajín za obdobie 1995-2005 a 2000-2010.

Príklad 5.34

Rozlišujme dve úrokové miery: nominálnu úrokovú mieru i a reálnu úrokovú mieru r. Vzťah medzi nominálnou úrokovou mierou a reálnou môžeme zapísať ako: $r = i - \pi$, kde π predstavuje infláciu (Fisherova rovnica). Presnejší zápis je $(1 + r) = (1 + i) / (1 + \pi)$, čo je možné prepísať ako:

$\ln(1 + r) = \ln(1 + i) - \ln(1 + \pi)$

keďže $\ln(1 + r) \approx r$ ak je r dostatočne malé (rádovo menej ako 20 %), tak potom dostávame nami uvedený zjednodušený zápis. Podľa Fisherovej rovnice, nárast v inflácii spôsobuje nárast nominálnej úrokovej miery. Vašou úlohou je overiť správnosť tohto vzťahu pre vybrané krajiny OECD. Za ukazovateľ nominálnej úrokovej miery použite trojmesačné krátkodobé úrokové miery a pre zmenu cenovej hladiny použite spotrebiteľské cenové indexy (CPI). Overte tento vzťah aj pre množinu aspoň 20 krajín OECD pre obdobie 2000-2005.

Príklad 5.35

Nominálny výmenný kurz (*NER*) medzi dvoma krajina (s vlastnou menou) je cena, za ktorú si môžeme jednou menou kúpiť menu druhú. Spravidla domáca mena je v čitateli (teda za jednotku domácej meny) a zahraničná mena v menovateli. Napríklad EUR/CZK 22.1125 znamená, že jedno Euro je ekvivalentné 22.1125 Českým korunám (nepriama kotácia EUR). Ak tento kurz vzrastie, hovoríme, že Euro sa posilnilo a naopak, ak poklesne, tak hovoríme, že Euro sa oslabilo. Posilnenie je spravidla dobré pre exportérov a oslabovanie pomáha importérom tovarov a služieb. Na rozdiel od nominálneho výmenného kurzu, reálny výmenný kurz zohľadňuje cenovú hladinu. Na výpočet môžeme použiť nasledovný vzťah: $RER = (NER * Pd)/ Pf$. Kde P predstavuje cenu vybraného tovaru v domácej krajine (d ako *domestic*) alebo v zahraničnej krajine (f ako *foreign*). Ak je *RER* vysoký, zahraničný tovar je vzhľadom k domácemu tovaru relatívne lacnejší a naopak. Ak je *NER* EUR/CZK 20 a určitý typ auta v Eurozóne stojí 10000 EUR a v Českej republike 150000 CZK, potom je *RER* = (20 * 10000) / 150000 = 1.333 (bez rozmerná jednotka), čo si môžeme interpretovať tak, že za jedno auto v Eurozóne dostaneme v Českej republike 1.333 áut. Vašou úlohou je nájsť 20 – 30 porovnateľných automobilových modelov na Slovensku a v Českej republike a vypočítať (rovnako aj vhodne vizualizovať) *RER*.

Príklad 5.36

Vytvorte dva x-y grafy. Na jednom bude na osi x-ovej veľkosť vzorky a na osi y-ovej priemerná hodnota simulovaných priemerov. Na druhom bude na osi y-ovej rozptyl

priemerov. Simuláciu realizujte pri iteráciách I = 5000, hodnoty vyberajte z ľubovoľného spojitého rozdelenia pravdepodobnosti a pokus realizujte pre vzorky n = 5:50.

Príklad 5.37

Študenti ekonomických odborov sa môžu s vybranými rozdeleniami stretnúť najmä vo finančnom manažmente, manažmente rizika, v projektovom manažmente, vo financiách a v mnohých iných predmetoch. Jedna z tradičných úloh v manažmente rizika je výpočet možného výnosu z investičného projektu. Niektoré kľúčové náhodné premenné (napr. významný variabilný náklad) môžu výrazne ovplyvniť konečný výnos z investície. Tieto kľúčové premenné sa modelujú pomocou vybraného rozdelenia pravdepodobnosti, pomocou čoho sa v konečnom dôsledku získava rozdelenie možných výnosov. Uvažujme o jednoduchom príklade, kde máme náklady C_1 s U ~ (min = 2, max = 6) a C_2 s U ~ (min = 4, max = 5). Všetky tieto premenné sú vzájomne nezávislé a trhová cena výrobku je taktiež neistá. Po prieskume trhu (zákazníci, konkurencia) sa ale zistilo, že trhovú cenu M je možné modelovať ako P ~ (λ = 1.2). Aké rozdelenie pravdepodobnosti má očakávaný výnos z investície? Navrhnite možný postup riešenia.

5.2 Riešenia k vybraným príkladom

Príklad 5.1

Idete s autom pol hodinu rýchlosťou 40 km/h a druhú polhodinu rýchlosťou 80 km/h. Aká bola po jednej hodine vaša priemerná rýchlosť?

Riešenie príkladu si môžeme interpretovať tak, že menovateľ rýchlosti sa nemení (nemení sa báza). Preto stačí na výpočet priemernej rýchlosti použiť aritmetický priemer, teda (40 km/h + 80 km/h) / 2 = 60 km/h.

Príklad 5.2

Idete s autom prvú polovicu cesty 40 km/h a druhú polovicu cesty rýchlosťou 80 km/h. Aká bola na konci cesty vaša priemerná rýchlosť?

Na rozdiel od predchádzajúceho príkladu sa mení menovateľ. Nevieme teda povedať, ako dlho išlo auto 40 km/h, prípadne 80 km/h. Ako budeme vidieť, ide o situáciu použitia harmonického priemeru.

Vieme, že pre rýchlosť platí: $v = s / t$ z čoho vyplýva, že $t = s / v$. Autom prejdeme jednu cestu o vzdialenosti s. Polku cesty prejdeme rýchlosťou v_1, teda $t_1 = (s / 2) / v_1$. Obdobne pre druhú polku cesty platí: $t_2 = (s / 2) / v_2$. Pre celkovú priemernú rýchlosť tak platí:

$$v = \frac{s}{\frac{\frac{s}{2}}{v_1} + \frac{\frac{s}{2}}{v_2}} = \frac{2sv_1v_2}{sv_1 + sv_2} = \frac{2v_1v_2}{(v_1 + v_2)} = 53.33 km/h$$

Príklad 5.3

Vaše investície rástli v prvom roku 1.05 %, v druhom roku 8.2 % a v treťom roku 3.1 %. Aký je priemerný rast vašich investícií?

Keďže sa mení báza, z ktorej sa percentá počítajú, je potrebné to pri výpočte zohľadniť. Vhodný je geometrický priemer. Všimnime si, že ak používame výraz priemer, v zásade tým hovoríme, že ak priemernú hodnotu súboru použijeme n krát, potom dosiahneme súčet (súčin) n čísel daného súboru.

$$\mu_G = \sqrt[3]{1.0105 x 1.082 x 1.031} = 1.0407$$

Inak povedané, ak by vaša investícia rástla o 4.07 % každý rok po dobu troch rokov, konečná hodnota investície by bola rovnaká, ako keby v prvom roku rástla o 1.05 %, v druhom o 8.2 % a v treťom o 3.1 %.

Príklad 5.4

K 1.1.2009 je stav nedokončenej výroby 400 ks, k 1.2.2009 je stav nedokončenej výroby 200 ks a k 1.3.2009 je stav nedokončenej výroby 400 ks. Aký je priemerný stav nedokončenej výroby od začiatku roka k 1.3.2009?

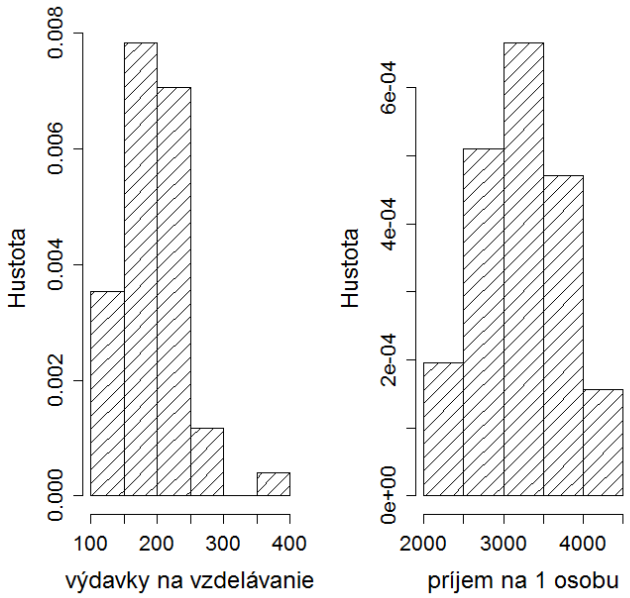
$$\mu_{CH} = \frac{\frac{400 + 200}{2} + \frac{200 + 400}{2}}{2} = 300 \text{ ks}$$

Príklad 5.7

Vyriešme najprv prvú časť zadania, a teda zostrojenie histogramu. V softvéri R je zostrojenie histogramu pomerne jednoduché. Najprv si samozrejme zvolíme danú knižnicu (`car`) a následne pripojíme zvolené dáta (`Anscombe`). Potom histogram vytvoríme pomocou funkcie `hist()`.

```
> library(car)
> attach(Anscombe)
```

```
-----------------------------------------------------------------
> par(mfrow = c(1, 2))
> hist(education, density = 10, col = "black", main = NA,
  cex.lab = 1.5, cex.axis = 1.3, freq = FALSE, ylab = "Hustota",
  xlab = "výdavky na vzdelávanie")
> hist(income, density = 10, col = "black", main = NA, cex.lab =
  1.5, cex.axis = 1.3, freq = FALSE, ylab = "Hustota", xlab =
  "príjem na 1 osobu")
```



Obrázok 5.1: Histogram skúmaných premenných

Zdroj: vlastné spracovanie v programe R

Na osi *x*-ovej sú výdavky na vzdelávanie a na osi *y*-ovej je hustota. Ide vlastne o relatívnu početnosť normovanú tak, aby obsah plochy vo všetkých stĺpcoch spolu bol rovný 1. Či už použijeme početnosť (`freq = TRUE`) alebo hustotu, tvar histogramu to neovplyvní. Všimnime si najprv histogram výdavkov na vzdelávanie. Zdá sa, že v prevažnej väčšine štátov sú výdavky na vzdelávanie menšie ako 250,- USD/osobu, kým v zopár málo štátoch sú výdavky väčšie. Tieto výdavky nevyzerajú byť rozdelené symetricky ani rovnomerne. Zrejme to naznačuje, že zopár málo štátov by mohlo používať inú stratégiu v oblasti vzdelávania. Na zodpovedanie tejto otázky, by bola potrebná dodatočná analýza. Naproti tomu, histogram príjmov na osobu je symetrický. To však zďaleka neznamená, že rovnaký. Väčšina štátov malo výšku príjmu od 2500,- USD do 4000,- USD, čo môžeme považovať za dosť veľké rozpätie. Ak k tomu pridáme, že zopár málo štátov malo príjem na osobu ešte menší ako 2500,- USD a skoro rovnaký počet malo príjem väčší ako 4000,- USD, zdá sa, že medzi niektorými štátmi je rozdiel v príjmoch skoro dvojnásobný.

Základnú deskriptívnu štatistiku môžeme vypočítať viacerými spôsobmi. Prvý spôsob vypočítania spočíva v zadaní požadovanej funkcie (tak ako je uvedené v texte, v kapitole Úvod do programu R). Pre priemer je to funkcia `mean()`, pre smerodajnú odchýlku `sd()`, pre rozptyl `var()`, pre minimum `min()` a pre maximum je to funkcia `max()`. Upozorňujeme, že v programe R funkcia `sd()` v skutočnosti počíta výberovú smerodajnú odchýlku, ktorá na rozdiel od smerodajnej odchýlky má v menovateli (n – 1). Pri vzorkách s väčším rozsahom je rozdiel vo výsledkoch spravidla zanedbateľný. Pre jednoduchosť sme sa preto rozhodli používať v tejto publikácii túto funkciu.

```
> mean(education)
[1] 196.3137
> sd(education)
[1] 46.45449
> var(education)
[1] 2158.020
> min(education)
[1] 112
> max(education)
[1] 372
```

```
> mean(income)
[1] 3225.294
> sd(income)
[1] 560.026
> var(income)
[1] 313629.1
> min(income)
[1] 2081
> max(income)
[1] 4425
```

Ďalší spôsob výpočtu deskriptívnej štatistiky je pomocou funkcie `summary()`, ktorá nám vráti v prehľadnej podobe základnú opisnú štatistiku, ale bez smerodajnej odchýlky a rozptylu.

```
> summary(education)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  112.0   165.0   192.0   196.3   228.5   372.0
> summary(income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2081    2786    3257    3225    3612    4425
```

Obdobné výsledky môžeme dostať, ak použijeme funkciu `summary()` na celú databázu.

```
> summary(Anscombe)
   education         income          young           urban
 Min.   :112.0   Min.   :2081   Min.   :326.2   Min.   : 322.0
 1st Qu.:165.0   1st Qu.:2786   1st Qu.:342.1   1st Qu.: 552.5
 Median :192.0   Median :3257   Median :354.1   Median : 664.0
 Mean   :196.3   Mean   :3225   Mean   :358.9   Mean   : 664.5
 3rd Qu.:228.5   3rd Qu.:3612   3rd Qu.:369.1   3rd Qu.: 790.5
 Max.   :372.0   Max.   :4425   Max.   :439.7   Max.   :1000.0
```

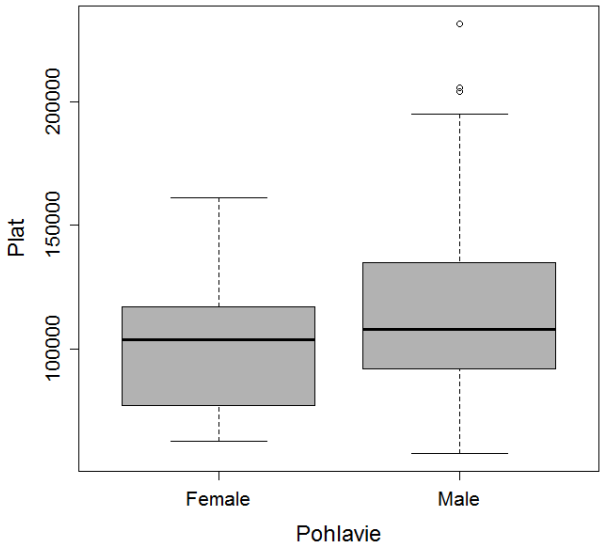
Všimnime si rozdiel medzi mediánom a aritmetickým priemerom. Pri vzdelávaní je priemer mierne väčší ako medián, čo naznačuje pravostranné zošikmenie hodnôt (podľa

obrázku sa to potvrdzuje) a v prípade príjmu je priemer mierne menší ako medián, čo naopak naznačuje ľavostranné zošikmenie hodnôt. Zrejme zaujímavejšie je sledovať variabilitu v oboch súboroch. Ako však interpretovať smerodajnú odchýlku 46.45449 pre výdavky na vzdelávanie? Často je zmysluplnejšie sledovať vývoj variability, napríklad v čase, alebo porovnať variabilitu medzi rôznymi subjektmi (v tomto prípade napríklad s variabilitou výdavkov na vzdelávanie v krajinách Európskej únie). Samotná hodnota variability sa interpretuje pomerne ťažko. Vychádzajme však z toho, že daná premenná má približne symetrické rozdelenie (hodnoty by mali pochádzať z normálneho rozdelenia pravdepodobnosti, ale nasledujúce pravidlá je možné použiť ako hrubý odhad aj pri symetrickom rozdelení). Potom pravidlo „jedna sigma“ hovorí, že v intervale 196.3 ± 46.45449 (priemerná hodnota plus/mínus smerodajná odchýlka) sa nachádza približne 68.3 % všetkých hodnôt. Pravidlo „dve sigmy“ hovorí, že v intervale 196.3 ± 2*46.45449 sa nachádza približne 95.4 % všetkých hodnôt a pravidlo „tri sigmy“ hovorí, že v intervale 196.3 ± 3*46.45449 sa nachádza až 99.7 % všetkých hodnôt. Tieto pravidlá nám pomáhajú pochopiť variabilitu v nameraných údajoch. Napríklad, ak by nám teoreticky vyšla smerodajná odchýlka na úrovni okolo 190, naznačovalo by to, že veľa hodnôt je extrémne blízko nule (ale zároveň kladných, keďže výdavky na vzdelávanie zrejme nemôžu byť záporné).

Príklad 5.8

Rovnako ako zostrojenie histogramu, aj zostrojenie box – plotu je v programe R pomerne jednoduché. Na jeho vytvorenie slúži v programe R funkcia `boxplot()`.

```
> library(car)
> attach(Salaries)
> boxplot(salary ~ sex, col = grey (0.7), density = 10, xlab =
  c("Pohlavie"), ylab = c("Plat"), cex.axis = 1.3, cex.lab =
  1.5)
```



Obrázok 5.2: Box – plot skúmaných premenných

Zdroj: vlastné spracovanie v programe R

Pre výpočet deskriptívnej štatistiky pre rôzne skupiny môžeme opäť postupovať rôznymi spôsobmi. Napríklad s využitím rozhrania z knižnice `Rcmdr` môžeme jednoduchým spôsobom dostať nasledujúce výsledky:

```
> library(Rcmdr)
> numSummary(salary, groups = sex)
           mean       sd    0%   25%    50%      75%   100%   n
Female 101002.4 25952.13 62884 77250 103750 117002.5 161101  39
Male   115090.4 30436.93 57800 92000 108043 134863.8 231545 358
```

V tomto prípade máme namiesto minimálnej a maximálnej hodnoty uvedený 0-tý a 100-tý percentil. Ďalším zo spôsobov je oddelenie požadovaných údajov pomocou funkcie `subset()`. Následne tak vieme vypočítať deskriptívnu štatistiku pre dané skupiny (mužov a ženy) s využitím konkrétnych funkcií na výpočet príslušnej deskriptívnej štatistiky (`max()`, `min()`, `mean()`, `sd()`, `var()`, a pod.). Taktiež môžeme použiť funkciu `summary()`, tá však v tomto prípade zaokrúhľuje výsledky na desiatky.

```
> Ženy <- subset(salary, subset = sex == "Female")
> Muži <- subset(salary, subset = sex == "Male")
```

Posledný spôsobom, ktorý si ukážeme, využíva funkciu `tapply()`. Pomocou tejto funkcie zadefinujeme (A) premennú, ktorá je predmetom nášho skúmania (`salary`); (B) faktor, podľa ktorého rozdeľujeme dáta (`sex`); (C) štatistiku, ktorú chceme vypočítať (počet pozorovaní – `length`, priemer – `mean`, smerodajnú odchýlku – `sd`, minimum – `min` a maximum – `max`).

```
> n <- tapply(salary, sex, length)
> xbar <- tapply(salary, sex, mean)
> s <- tapply(salary, sex, sd)
> min <- tapply(salary, sex, min)
> max <- tapply(salary, sex, max)
> cbind(n = n, priemer = xbar, št.odchýlka = s, minimum = min,
  maximum = max)
         n  priemer št.odchýlka minimum maximum
Female  39 101002.4    25952.13   62884  161101
Male   358 115090.4    30436.93   57800  231545
```

Venujme sa teraz interpretácii. Všimnime si, že rozdiel medzi mediánmi pohlaví je 4293,-USD, kým medzi priemermi cez 14088,- USD. Ak sa pozrieme na box – plot, môžeme vidieť, že vo vzorke mužov je niekoľko extrémne vysoko zarábajúcich učiteľov – mužov (v porovnaní s ostatnými učiteľmi). Zrejme preto je rozdiel medzi priemermi o toľko väčší. Kým medzi mužmi sa vyskytujú extrémne vysoké príjmy, medzi ženami nie. Preto je priemer mužov výrazne väčší. Ak si zostrojíte histogram, zrejme bude rozdelenie príjmov mužov viac pravostranne zošikmené ako rozdelenie príjmov žien. Všimnite si ďalej, že zrejme nie je prekvapujúce, že uvidíme v oboch prípadoch pravostranné zošikmenie. Platy učiteľov majú určitú minimálnu hranicu (floor), pod ktorú sa nedostanú, takže „zľava“ (zdola) sú platy ohraničené, kým „sprava“ (zhora) nie. Otázkou, ktorú ostáva zodpovedať je, čím sú spôsobené tieto rozdiely medzi mužmi a ženami. Je to tým, že existuje diskriminácia pohlaví? Je to tým, že muži častejšie obsadzujú vedúce pozície na školách (čo je spravidla ekvivalentné väčšiemu príjmu)? Je to v dôsledku nevhodne vybranej vzorky (mužov máme vo vzorke rozhodne viac ako žien)?

Príklad 5.9

Z vyššie uvedených spôsobov výpočtu základnej deskriptívnej štatistiky je zrejme najjednoduchším výpočet pomocou funkcie `summary()`. Keďže táto funkcia nevracia výsledky pre smerodajnú odchýlku, jednoduchým príkazom si môžeme túto štatistiku dopočítať.

```
> library(lmtest)
> summary(unemployment)
       UN              m                p                G
Min.   : 1.226  Min.   :  3.92  Min.   :0.1500  Min.   : 16.36
1st Qu.: 3.858  1st Qu.: 15.19  1st Qu.:0.1953  1st Qu.: 26.85
Median : 5.233  Median : 45.70  Median :0.3305  Median : 52.70
Mean   : 6.478  Mean   :147.96  Mean   :0.4642  Mean   :105.17
3rd Qu.: 7.527  3rd Qu.:190.58  3rd Qu.:0.6438  3rd Qu.:180.38
Max.   :22.981  Max.   :914.40  Max.   :1.6280  Max.   :300.40
       x
```

```
Min.   :  5.90
1st Qu.: 10.85
Median : 17.30
Mean   : 32.35
3rd Qu.: 39.03
Max.   :172.82
> sd(unemployment)
       UN           m           p           G           x
   4.3382713   204.9327461  0.3390043   94.0411656  34.6514368
```

Maximálna miera nezamestnanosti bola na úrovni 22.981 %. Takáto vysoká nezamestnanosť nastala počas veľkej hospodárskej krízy v 30-tych rokoch 20-tého storočia. Priemerná nezamestnanosť počas týchto 90 rokov bola 6.478 %. Všimnime si, že ak by sme nebrali do úvahy obdobie od 1930 do 1940, priemerná miera nezamestnanosti za celé obdobie by klesla na iba 5.463608.

```
> mean(unemployment[-c(41:51),1])
[1] 5.463608
```

Príklad 5.10

Ak by sme pri výpočte v programe postupovali tak ako v predchádzajúcom príklade, narazíme na určité problémy, ktoré súvisia s tým, že niektoré údaje pre niektoré krajiny nie sú dostupné.

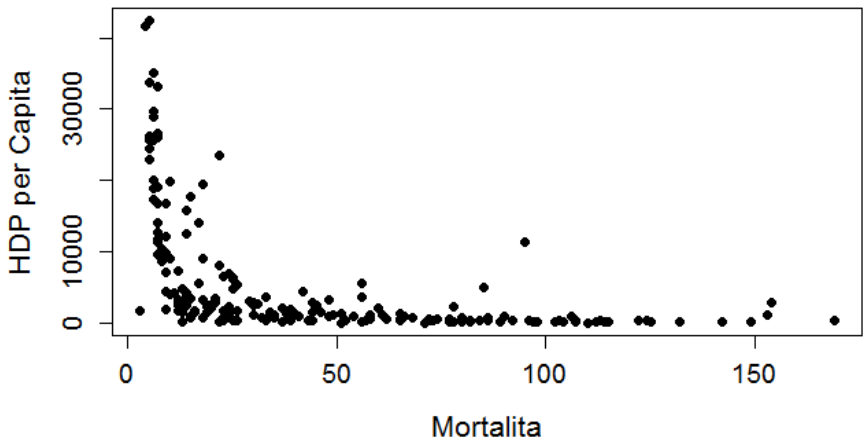
```
> library(car)
> summary(UN)
  infant.mortality      gdp
  Min.   :  2.00   Min.   :   36
  1st Qu.: 12.00   1st Qu.:  442
  Median : 30.00   Median : 1779
  Mean   : 43.48   Mean   : 6262
  3rd Qu.: 66.00   3rd Qu.: 7272
  Max.   :169.00   Max.   :42416
  NA's   :  6.00   NA's   :   10
> sd(UN)
  infant.mortality              gdp
                NA               NA
```

Už vo výstupe z funkcie `summary()` vidíme, že nedostupné údaje sú označené ako NA, z angl. *not available*. Chýbajúce údaje v súbore môžu byť pre niektoré funkcie v programe R problematické, čo môžeme vidieť pri výpočte smerodajnej odchýlky cez funkciu `sd()`. Jednoduchým riešením môže byť využitie funkcie `na.omit()`, pomocou ktorej odstránime pozorovania s chýbajúcimi údajmi.

```
> sd(na.omit(UN))
infant.mortality              gdp
        38.55439       8909.41852
```

Ďalej si môžeme všimnúť, že smerodajná odchýlka je pomerne veľká. Priemerné HDP per capita (gdp) má hodnotu 6262 a smerodajná odchýlka až 8909.41852. Ak by sme uplatnili pravidlo jedna sigma, dosiahli by sme interval, v ktorom by sme mali záporné HDP na osobu. Už tento fakt naznačuje, že použitie tohto pravidla nie je možné, čo následne (bez použitia formálnejších testov, histogramu,...) naznačuje, že údaje o HDP na osobu sa neriadia symetrický rozdelením, a teda určite nie normálnym rozdelením pravdepodobnosti. Ak porovnáme medián a priemer, rozdiel je tiež pomerne veľký. Všetky tieto výsledky naznačujú že v HDP na osobu existuje veľká variabilita, že ide o silne pravostranné zošikmenie, keďže zopár málo krajín má vysoké HDP na osobu a väčšina krajín sveta má nízke HDP na osobu. Podobné úvahy viedli ľudí k zisteniu o nerovnomernom rozdelení bohatstva vo svete. Zo zaujímavosti si môžeme znázorniť vzťah medzi HDP na osobu a úmrtnosťou detí (počet na 1000 narodení) pomocou *x-y* grafu. Zrejme je rozumné očakávať, že v krajinách s nízkou ekonomickou aktivitou bude zdravotná starostlivosť zanedbávaná, čo sa môže premietnuť do vyššej miery mortality detí. Nasledujúci obrázok existenciu nelineárneho vzťahu naznačuje. Od určitej hranice HDP na osobu s jej ďalším poklesom mortalita detí začne prudko stúpať.

```
> UN_1 <- na.omit(UN)
> attach(UN_1)
> plot(UN_1, type = "p", lty = 1, xlab = "Mortalita", ylab =
  "HDP per Capita", family = "serif", cex.axis = 1.5, cex.lab =
  1.7, pch = 19)
```



Obrázok 5.3: *x-y* graf závislosti HDP per capita a mortality detí

Zdroj: vlastné spracovanie v programe R

Príklad 5.11

Predtým ako pristúpime k výpočtu deskriptívnej štatistiky je vhodné zistiť, či nemáme chýbajúce pozorovania v danej vzorke. Za týmto účelom môžeme použiť funkcie `is.na()`, ktorá pre každú hodnotu vráti logický argument `TRUE` alebo `FALSE` podľa toho, či v sa v dátach vyskytuje chýbajúce pozorovanie alebo nie. Keďže výsledkom je vektor 237 logických argumentov, je jednoduchšie pred túto funkciu uviesť ešte funkciu `sum()`, ktorá nám spočíta chýbajúce pozorovania (podľa argumentu `TRUE`).

```
> library(MASS)
> attach(survey)
> sum(is.na(Sex))
[1] 1
> sum(is.na(Height))
[1] 28
> sum(is.na(Age))
[1] 0
```

Môžeme vidieť, že jedno pozorovanie chýba v premennej pohlavie (`Sex`) a 28 pozorovaní chýba v premennej výška (`Height`). Preto v ďalšom výpočte budeme používať funkciu `na.omit()`. Postupovať budeme tak ako v predchádzajúcom príklade, teda využijeme funkciu `summary()` a smerodajnú odchýlku dopočítame pomocou funkcie `sd()`.

```
> vyska_M <- na.omit(subset(Height, subset = Sex == "Male"))
> vyska_Z <- na.omit(subset(Height, subset = Sex == "Female"))
> vek_M <- na.omit(subset(Age, subset = Sex == "Male"))
> vek_Z <- na.omit(subset(Age, subset = Sex == "Female"))
-----------------------------------------------------------------
> summary(vyska_M); sd(vyska_M)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  154.9   172.8   180.0   178.8   185.0   200.0
[1] 8.380252
> summary(vyska_Z); sd(vyska_Z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  150.0   162.6   166.8   165.7   170.0   180.3
[1] 6.151777
> summary(vek_M); sd(vek_M)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.75   17.92   18.88   20.33   20.29   70.42
[1] 6.069863
> summary(vek_Z); sd(vek_Z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.92   17.50   18.42   20.41   19.98   73.00
[1] 6.906053
```

Ako sa dalo čakať, priemerná výška mužov je väčšia ako u žien. Taktiež nie je prekvapením vyššia smerodajná odchýlka. Porovnávanie dvoch smerodajných odchýlok

navzájom je ideálne vtedy, ak obe premenné merajú rovnakú veličinu a majú rovnaký priemer. Preto vhodným doplnkom tejto analýzy je výpočet variačného koeficientu.

```
> mean(vyska_M)/sd(vyska_M)
[1] 21.33898
> mean(vyska_Z)/sd(vyska_Z)
[1] 26.93314
```

Kým smerodajná odchýlka výšky mužov bola vyššia ako u žien, variačný koeficient je vyšší u žien. Kto má teda vyššiu variabilitu závisí od významu analýzy. Odchýlka 1 cm od aritmetického priemeru 178.8 cm je pomerovo (alebo percentuálne) menšia ako odchýlka 1 cm od aritmetického priemeru 165.7 cm. Čiže variabilita je väčšia u mužov, ale relatívne k priemernej výške je väčšia u žien. Ak by sme šili šaty rôznej veľkosti, tak má pre nás zrejme väčší význam sledovať smerodajnú odchýlku.

Príklad 5.12

Za účelom zopakovania vzťahov na výpočet aritmetického, geometrického a harmonického priemeru uvedieme ich manuálny výpočet v programe R. Aritmetický priemer označíme `xbar`, geometrický `geom` a harmonický `harm`. Len pripomenieme, že funkciou `length()` získame počet pozorovaní, funkciou `sum()` získame súčet všetkých prvkov a funkciou `prod()` získame ich súčin.

```
> x <- c(7, 5, 21, 15, 47, 34, 9, 42, 19, 68, 83)
> xbar <- sum(x)/length(x); xbar
[1] 31.81818
> geom <- prod(x)^(1/length(x)); geom
[1] 22.40497
> harm <- 1/mean(1/x); harm
[1] 15.23256
```

Príklad 5.13

Funkcia na výpočet aritmetického priemeru `mean()` patrí medzi základné a nie je nutné inštalovať žiadnu knižnicu. Geometrický a harmonický priemer síce nepatria medzi základné funkcie, ale sú súčasťou viacerých knižníc. My budeme používať knižnicu `psych`, v ktorej na výpočet geometrického priemeru slúži funkcia `geometric.mean()` a na výpočet harmonického priemeru slúži funkcia `harmonic.mean()`.

```
> library(psych)
> x <- c(2, 15, 28, 7, 43, 31, 19, 84, 13, 76, 37)
-----------------------------------------------------------------
```

```
> xbar <- sum(x)/length(x); xbar
[1] 32.27273
> mean(x)
[1] 32.27273
-----------------------------------------------------------------
> geom <- prod(x)^(1/length(x)); geom
[1] 21.5152
> geometric.mean(x)
[1] 21.5152
-----------------------------------------------------------------
> harm <- 1/mean(1/x); harm
[1] 11.19711
> harmonic.mean(x)
[1] 11.19711
```

Môžeme vidieť, že výsledky sú rovnaké bez ohľadu na použitý postup. Taktiež platí, že aritmetický priemer je väčší ako geometrický, a ten je väčší ako harmonický ($\mu \geq \mu_G \geq \mu_H$).

Príklad 5.14

Bežnou chybou pri výpočte priemerných hodnôt z pomerových ukazovateľov, ktoré majú v čitateli cenu, je použitie aritmetického priemeru. Keďže tento ukazovateľ má v čitateli cenu akcie a v menovateli zisk pripadajúci na jednu akciu, tak pri použití aritmetického priemeru dávame väčšiu váhu vyšším hodnotám tohto ukazovateľa. Harmonický priemer sa tak javí ako lepšia voľba, pretože každý údaj bude mať rovnakú váhu.

```
> library(psych)
> PE <- c(12, 16, 25, 8, 68, 27, 19, 31, 24)
> harmonic.mean(PE)
[1] 18.39992
> mean(PE)
[1] 25.55556
```

Pre porovnanie uvádzame aj aritmetický priemer, ktorý ako vieme, bude vyšší ako harmonický.

Príklad 5.15

Pri výpočte priemernej výnosnosti z percentuálnych zmien je vhodné použiť geometrický priemer. Zo spôsobu výpočtu geometrického priemeru je však zrejmé, že zo záporných hodnôt geometrický priemer nevypočítame[24]. Čo však môžeme spraviť je, že uvedené percentuálne zmeny transformujeme na koeficienty zmeny (pripočítaním čísla 1 ku každej zložke vektora percentuálnych zmien).

[24] Taktiež nulové hodnoty predstavujú problém pri výpočte geometrického priemeru. V tomto prípade sa odporúča dané hodnoty nahradiť (napr. číslom 1) alebo vylúčiť z výpočtu.

```
> library(psych)
> percento <- c(0.12, -0.16, -0.25, -0.34, 0.08, 0.17, 0.27, -
  0.19, 0.31, 0.24, -0.4, 0.21)
> vynos <- percento + 1
-----------------------------------------------------------------
> percento
 [1] 0.12 -0.16 -0.25 -0.34  0.08  0.17  0.27 -0.19  0.31  0.24
  -0.40  0.21
> vynos
 [1] 1.12 0.84 0.75 0.66 1.08 1.17 1.27 0.81 1.31 1.24 0.60 1.21
-----------------------------------------------------------------
> geometric.mean(vynos)
[1] 0.972305
> mean(vynos)
[1] 1.005
```

V tomto príklade je použitie správneho priemeru do značnej miery významné, keďže úplne mení konečnú interpretáciu. Pri geometrickom priemere je výsledok záporný (priemerná ročná výnosnosť je približne –2.77 %) a pri aritmetickom je výsledok kladný (priemerná ročná výnosnosť je +0.5 %). Percentuálnu výnosnosť dostaneme, ak naspäť odpočítame od vypočítaných priemerných hodnôt číslo 1 a prenásobíme 100:

```
> (geometric.mean(vynos) - 1)*100
[1] - 2.769496
> (mean(vynos) - 1)*100
[1] 0.5
```

Ukázali sme, akým spôsobom je možné vypočítať geometrický priemer aj zo záporných hodnôt. Je však dôležité uviesť, že priemer z takejto transformácie nie je rovnaký, ako priemer bez transformácie. Uvažujme ten istý príklad avšak so všetkými kladnými hodnotami:

```
> percento <- c(0.12, 0.16, 0.25, 0.34, 0.08, 0.17, 0.27, 0.19,
  0.31, 0.24, 0.4, 0.21)
> vynos <- percento + 1; vynos
 [1] 1.12 1.16 1.25 1.34 1.08 1.17 1.27 1.19 1.31 1.24 1.40 1.21
-----------------------------------------------------------------
> geometric.mean(percento)
[1] 0.2094204
> geometric.mean(vynos) - 1
[1] 0.2251468
```

Môžeme vidieť, že „priemerný výnos“ (v skutočnosti nejde o priemerný výnos) počítaný z percent je rovný 20.94 % a priemerný výnos z koeficientov rastu je rovný 22.51 %.

Príklad 5.16

Výpočet deskriptívnej štatistiky pre zjednodušenie zrealizujeme pomocou funkcie `summary()`. Samozrejme, zaujímať nás budú len vypočítané hodnoty pre ukazovateľ D/HDP. Keďže zo zadania príkladu vyplýva, že porovnať máme priemerné hodnoty, budeme si všímať najmä tie. Pre lepší obraz o zadlženosti týchto krajín sú však zaujímavé aj minimálne a maximálne hodnoty dosahované v rámci sledovaného obdobia. Napríklad pri Grécku môžeme vidieť, že minimálna hodnota ukazovateľa D/HDP od 4. kvartálu roku 2000 bola 97.2 % (údaje D/HDP sú uvádzané priamo v percentách). 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 %. Grécko túto hranicu evidentne v rámci sledovaného obdobia ani raz nedodržalo. Rovnaká situácia je v prípade Talianska, ktoré v rámci sledovaného obdobia malo minimálny pomer dlhu k HDP na úrovni 103.6 %.

```
> data <- read.csv(file = "...cesta k súboru...\\debt_gdp.csv",
  sep <- ";", dec = ".", header = T)
> summary(data)
        Portugal          Ireland          Italy
        Min.   :48.20     Min.   :24.60    Min.   :103.6
        1st Qu.:55.08     1st Qu.:29.10    1st Qu.:106.5
        Median :61.70     Median :32.35    Median :108.7
        Mean   :60.85     Mean   :37.99    Mean   :109.3
        3rd Qu.:63.60     3rd Qu.:36.05    3rd Qu.:110.7
        Max.   :84.20     Max.   :90.50    Max.   :119.6

        Greece            Great.Britain    Spain
        Min.   : 97.2     Min.   :36.70    Min.   :35.30
        1st Qu.:101.4     1st Qu.:38.85    1st Qu.:40.83
        Median :104.5     Median :41.90    Median :47.55
        Mean   :108.0     Mean   :45.43    Mean   :47.35
        3rd Qu.:108.3     3rd Qu.:44.52    3rd Qu.:53.45
        Max.   :140.1     Max.   :75.10    Max.   :59.30
```

Jednoduchým spôsobom vieme tiež vypočítať, ako sa menili priemerné hodnoty ukazovateľa D/HDP pred rokom 2008 a od roku 2008.

```
> Portugal <- c(mean(data$Portugal[1:29]),
  mean(data$Portugal[30:40]))
> Ireland <- c(mean(data$Ireland[1:29]),
  mean(data$Portugal[30:40]))
> Italy <- c(mean(data$Italy[1:29]), mean(data$Italy[30:40]))
> Greece <- c(mean(data$Greece[1:29]), mean(data$Greece[30:40]))
> Great.Britain <- c(mean(data$Great.Britain[1:29]),
  mean(data$Great.Britain[30:40]))
> Spain <- c(mean(data$Spain[1:29]), mean(data$Spain[30:40]))
> rbind(Portugal, Ireland, Italy, Greece, Great.Britain, Spain)
                  [,1]      [,2]
Portugal       56.88276  71.30000
Ireland        30.60690  71.30000
```

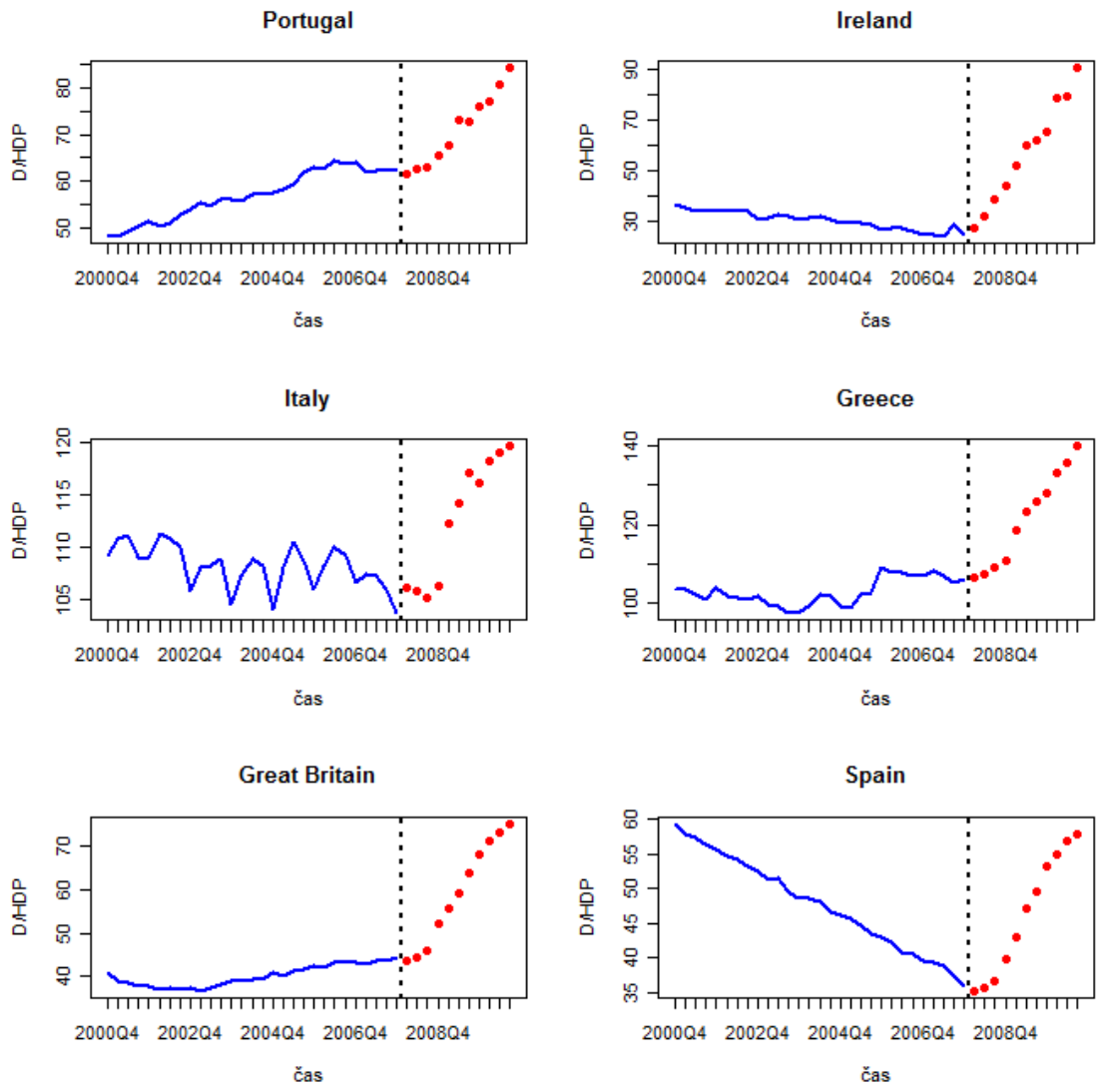
```
Italy         108.08966 112.67273
Greece        102.78276 121.60909
Great.Britain  40.16552  59.30909
Spain          47.72069  46.36364
```

Môžeme vidieť, že vo všetkých krajinách došlo k nárastu v priemerných hodnotách pomerového ukazovateľa D/HDP. Okrem Španielska, kde došlo k miernemu poklesu priemerných hodnôt pomeru dlhu k HDP.

V ďalšom kroku by sme sa mali pokúsiť o vizualizáciu týchto dát tak, aby boli z grafu na prvý pohľad odlíšiteľné hodnoty „pred krízou“ (pred rokom 2008) a „počas krízy“ (od roku 2008). Na výber máme samozrejme viacero možností, ako napríklad skombinovať dva typy grafov (čiarový a bodový), farebne rozlíšiť požadované hodnoty, prípadne použiť vertikálne čiary na oddelenie daných hodnôt. Na ukážku zvolíme kombináciu všetkých týchto možností.

```
> 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 = "blue",
  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$obs)
> abline(v = 29.5, lty=3, lwd=2)
-------------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Ireland[1:29], type = "l",
  ylab = "D/HDP", main = "Ireland", xlab = "čas", col = "blue",
  lwd=2, xaxt = "n", xlim = c(1, max(data$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)
-------------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Italy[1:29], type = "l",
  ylab = "D/HDP", main = "Italy", xlab = "čas", col = "blue",
  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$obs)
> abline(v = 29.5, lty=3, lwd=2)
-------------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Greece[1:29], type = "l",
  ylab = "D/HDP", main = "Greece", xlab = "čas", col = "blue",
lwd=2, xaxt = "n", xlim = c(1, max(data$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)
-----------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Great.Britain[1:29], type =
  "l", ylab = "D/HDP", main = "Great Britain", xlab = "čas", col
  = "blue", 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$obs)
> abline(v = 29.5, lty=3, lwd=2)
-----------------------------------------------------------------
> plot(x = data$time[1:29], y = data$Spain[1:29], type = "l",
  ylab = "D/HDP", main = "Spain", xlab = "čas", col = "blue",
  lwd=2, xaxt = "n", xlim = c(1, max(data$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)
```



Obrázok 5.4: Vývoj verejného dlhu k HDP krajín PIIGGS pred krízou a počas krízy

Zdroj: vlastné spracovanie v programe R

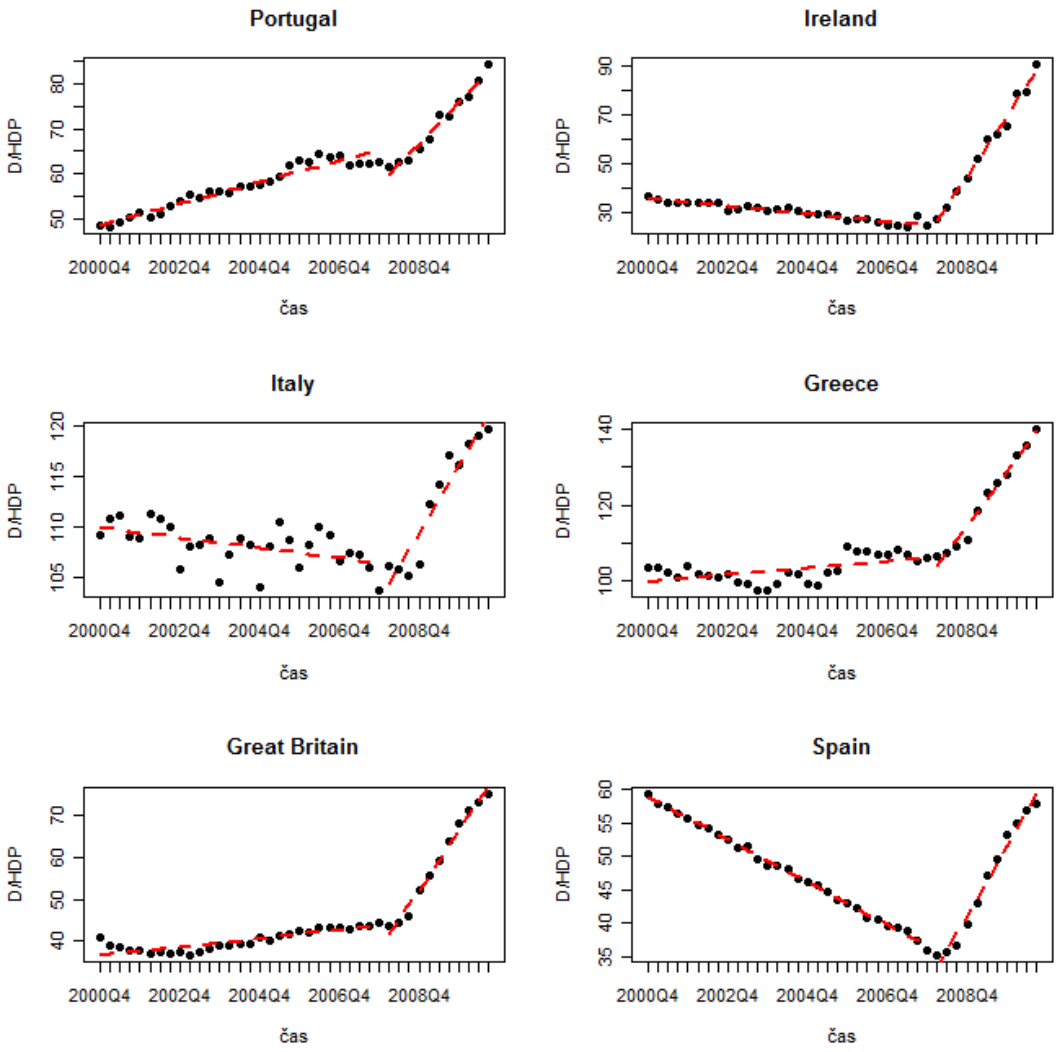
Pri týchto dátach je zaujímavé tiež sledovať zmenu v trende. V tejto publikácii sme sa lineárnej regresii nevenovali. Avšak k vizualizácii lineárnych trendov nie je nutné podrobnejšie poznať teóriu, ktorá sa skrýva za tvorbou (lineárnych) regresných modelov. Ak by sme vypočítali jednoduchú lineárnu regresiu, mohli by sme vidieť, že koeficient pri časovom trende sa do značnej miery zmenil. Inými slovami, rast verejného dlhu v pomere k HDP sa od roku 2008 výrazne zvýšil. Lineárnu regresiu v tvare $D/HDP_t = \beta_0 + \beta_1 t + \varepsilon_{i,t}$ vypočítame pomocou funkcie `lm()` pre každú krajinu vo vzorke a trend nanesieme do bodového grafu (presnejšie tzv. vyrovnané hodnoty, v angl. *fitted values*).

```
> data <- read.csv(file = "...cesta k súboru...\\debt_gdp.csv",
  sep = ";", dec = ".", header = T)
> library(zoo)
```

```
> par(mfrow = c(3,2))
-----------------------------------------------------------------
> plot(x = data$time[1:40], y = data$Portugal[1:40], type = "p",
  main = "Portugal", col = "black", pch = 19, xaxt = "n", ylab =
  "D/HDP", xlab = "čas", xlim = c(1, max(data$time)), ylim =
  c(min(data$Portugal), max(data$Portugal)))
> axis(side = 1, at = data$time, labels = data$obs)
> trend_1 <- lm(data$Portugal[1:29] ~ data$time[1:29])$fitted
> trend_2 <- lm(data$Portugal[30:40] ~ data$time[30:40])$fitted
> line_trend_1 <- zoo(trend_1,data$time[1:29])
> line_trend_2 <- zoo(trend_2,data$time[30:40])
> lines(line_trend_1,type="l",col=2,lty=2,lwd=2)
> lines(line_trend_2,type="l",col=2,lty=2,lwd=2)
-----------------------------------------------------------------
> plot(x = data$time[1:40], y = data$Ireland[1:40], type = "p",
  main = "Ireland", col = "black", pch = 19, xaxt = "n", ylab =
  "D/HDP", xlab = "čas", xlim = c(1, max(data$time)), ylim =
  c(min(data$Ireland), max(data$Ireland)))
> axis(side = 1, at = data$time, labels = data$obs)
> trend_1 <- lm(data$Ireland[1:29] ~ data$time[1:29])$fitted
> trend_2 <- lm(data$Ireland[30:40] ~ data$time[30:40])$fitted
> line_trend_1 <- zoo(trend_1,data$time[1:29])
> line_trend_2 <- zoo(trend_2,data$time[30:40])
> lines(line_trend_1,type="l",col=2,lty=2,lwd=2)
> lines(line_trend_2,type="l",col=2,lty=2,lwd=2)
-----------------------------------------------------------------
> plot(x = data$time[1:40], y = data$Italy[1:40], type = "p",
  main = "Italy", col = "black", pch = 19, xaxt = "n", ylab =
  "D/HDP", xlab = "čas", xlim = c(1, max(data$time)), ylim =
  c(min(data$Italy), max(data$Italy)))
> axis(side = 1, at = data$time, labels = data$obs)
> trend_1 <- lm(data$Italy[1:29] ~ data$time[1:29])$fitted
> trend_2 <- lm(data$Italy[30:40] ~ data$time[30:40])$fitted
> line_trend_1 <- zoo(trend_1,data$time[1:29])
> line_trend_2 <- zoo(trend_2,data$time[30:40])
> lines(line_trend_1,type="l",col=2,lty=2,lwd=2)
> lines(line_trend_2,type="l",col=2,lty=2,lwd=2)
-----------------------------------------------------------------
> plot(x = data$time[1:40], y = data$Greece[1:40], type = "p",
  main = "Greece", col = "black", pch = 19, xaxt = "n", ylab =
  "D/HDP", xlab = "čas", xlim = c(1, max(data$time)), ylim =
  c(min(data$Greece), max(data$Greece)))
> axis(side = 1, at = data$time, labels = data$obs)
> trend_1 <- lm(data$Greece[1:29] ~ data$time[1:29])$fitted
> trend_2 <- lm(data$Greece[30:40] ~ data$time[30:40])$fitted
> line_trend_1 <- zoo(trend_1,data$time[1:29])
> line_trend_2 <- zoo(trend_2,data$time[30:40])
> lines(line_trend_1,type="l",col=2,lty=2,lwd=2)
> lines(line_trend_2,type="l",col=2,lty=2,lwd=2)
-----------------------------------------------------------------
> plot(x = data$time[1:40], y = data$Great.Britain[1:40], type =
  "p", main = "Great Britain", col = "black", pch = 19, xaxt =
  "n", ylab = "D/HDP", xlab = "čas", xlim = c(1,
  max(data$time)), ylim = c(min(data$Great.Britain),
  max(data$Great.Britain)))
```

```
> axis(side = 1, at = data$time, labels = data$obs)
> trend_1 <- lm(data$Great.Britain[1:29] ~
  data$time[1:29])$fitted
> trend_2 <- lm(data$Great.Britain[30:40] ~
  data$time[30:40])$fitted
> line_trend_1 <- zoo(trend_1,data$time[1:29])
> line_trend_2 <- zoo(trend_2,data$time[30:40])
> lines(line_trend_1,type="l",col=2,lty=2,lwd=2)
> lines(line_trend_2,type="l",col=2,lty=2,lwd=2)
-----------------------------------------------------------------
> plot(x = data$time[1:40], y = data$Spain[1:40], type = "p",
  main = "Spain", col = "black", pch = 19, xaxt = "n", ylab =
  "D/HDP", xlab = "čas", xlim = c(1, max(data$time)), ylim =
  c(min(data$Spain), max(data$Spain)))
> axis(side = 1, at = data$time, labels = data$obs)
> trend_1 <- lm(data$Spain[1:29] ~ data$time[1:29])$fitted
> trend_2 <- lm(data$Spain[30:40] ~ data$time[30:40])$fitted
> line_trend_1 <- zoo(trend_1,data$time[1:29])
> line_trend_2 <- zoo(trend_2,data$time[30:40])
> lines(line_trend_1,type="l",col=2,lty=2,lwd=2)
> lines(line_trend_2,type="l",col=2,lty=2,lwd=2)
```

Z takto vytvorených grafov je potom na prvý pohľad zrejmé, že od roku 2008 naozaj dochádzalo k zvýšeniu ukazovateľov D/HDP. Keď si uvedomíme, že počas hospodárskej krízy je bežné, že sa zvyšuje zadlženosť krajiny (zvyšujú sa výdavky na stabilizáciu, podporu rastu ekonomiky) a zároveň HDP danej krajiny klesá, tak je úplne zrejmé, že takto postavený ukazovateľ bude výrazne rásť. Na jeho zvýšení sa súčasne podieľa tak rast čitateľa ako aj pokles menovateľa.



Obrázok 5.5: Vývoj verejného dlhu k HDP v krajinách PIIGGS

Zdroj: vlastné spracovanie v programe R

Pre zaujímavosť sa môžeme pozrieť, ako sa menil koeficient pri lineárnom trende pred rokom 2008 a po ňom. Najmarkantnejšia zmena je pozorovateľná v prípade Írska, kde sa negatívny trend (znižovanie zadlženosti krajiny) stal výrazne kladným od roku 2008.

```
> rm(list=ls())
> data = read.csv(file = "...cesta k súboru...\\debt_gdp.csv",
  sep = ";", dec = ".", header = T)
> Portugal_1 <- lm(data$Portugal[1:29] ~ data$time[1:29])
> Portugal_2 <- lm(data$Portugal[30:40] ~ data$time[30:40])
> Ireland_1 <- lm(data$Ireland[1:29] ~ data$time[1:29])
> Ireland_2 <- lm(data$Ireland[30:40] ~ data$time[30:40])
> Italy_1 <- lm(data$Italy[1:29] ~ data$time[1:29])
> Italy_2 <- lm(data$Italy[30:40] ~ data$time[30:40])
> Greece_1 <- lm(data$Greece[1:29] ~ data$time[1:29])
> Greece_2 <- lm(data$Greece[30:40] ~ data$time[30:40])
```

```
> Great.Britain_1 <- lm(data$Great.Britain[1:29] ~
  data$time[1:29])
> Great.Britain_2 <- lm(data$Great.Britain[30:40] ~
  data$time[30:40])
> Spain_1 <- lm(data$Spain[1:29] ~ data$time[1:29])
> Spain_2 <- lm(data$Spain[30:40] ~ data$time[30:40])
-----------------------------------------------------------------
> cat("Portugal", "\t", Portugal_1$coeff[[2]], "\t",
  Portugal_2$coeff[[2]], "\n", "Ireland", "\t",
  Ireland_1$coeff[[2]], "\t", Ireland_2$coeff[[2]], "\n",
  "Italy", "\t", Italy_1$coeff[[2]], "\t", Italy_2$coeff[[2]],
  "\n", "Greece", "\t", Greece_1$coeff[[2]], "\t",
  Greece_2$coeff[[2]], "\n", "Great.Britain", "\t",
  Great.Britain_1$coeff[[2]], "\t", Great.Britain_2$coeff[[2]],
  "\n", "Spain", "\t", Spain_1$coeff[[2]], "\t",
  Spain_2$coeff[[2]], "\n")

  Portugal         0.5935468       2.306364
  Ireland         -0.3981773       6.15
  Italy           -0.1259606       1.672727
  Greece           0.2241379       3.599091
  Great.Britain    0.2473892       3.526364
  Spain           -0.7912808       2.580909
```

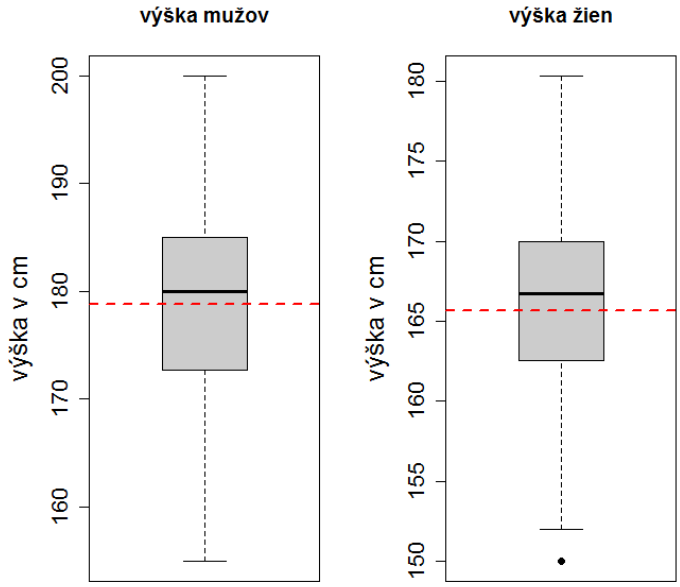
Príklad 5.17

Najprv si z danej 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 (uvedené vieme zistiť pomocou funkcie `length()`). Môžeme vidieť, že priemery sú pre tieto dve skupiny odlišné (178.8 cm u mužov a 165.7 cm u žien).

```
> library(MASS)
> data <- attach(survey)
-----------------------------------------------------------------
> vyska_M <- na.omit(subset(Height, subset = Sex == "Male"));
  mean(vyska_M)
[1] 178.8260
> length(vyska_M)
[1] 106
> vyska_Z <- na.omit(subset(Height, subset = Sex == "Female"));
  mean(vyska_Z)
[1] 165.6867
> length(vyska_Z)
[1] 102
```

Ďalej pristúpime k zostrojeniu box – plotov. Pokračujeme v uvedenom kóde nasledujúcimi príkazmi (musíme mať zadefinované premenné `vyska_M` a `vyska_Z`):

```
> par(mfrow = c(1, 2))
> boxplot(vyska_M, ylab = "výška v cm", main = "výška mužov",
  col = gray(0.8), pch = 19, cex.axis = 1.3, cex.lab = 1.5)
> abline(h = mean(vyska_M), lwd = 2, lty = 2, col = "red")
> boxplot(vyska_Z, ylab = "výška v cm", main = "výška žien", col
  = gray(0.8), pch = 19, cex.axis = 1.3, cex.lab = 1.5)
> abline(h = mean(vyska_Z), lwd = 2, lty = 2, col = "red")
```

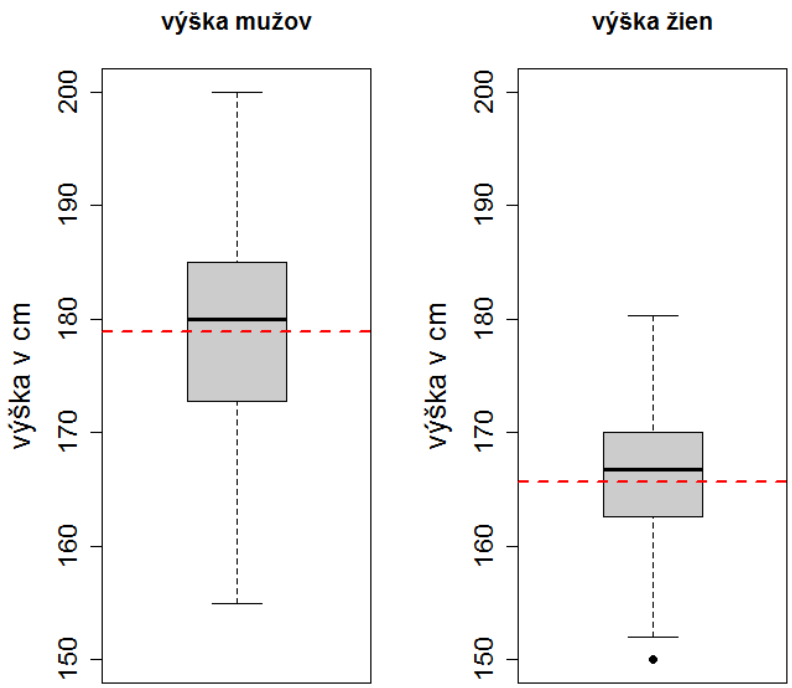


Obrázok 5.6: Box – ploty výšky mužov a žien z databázy `survey` (rozdielne mierky)

Zdroj: vlastné spracovanie v programe R

Z uvedených box – plotov by sme mohli prijať tvrdenie, že v rámci danej univerzity sú muži vyšší ako ženy. Keďže sú však y-ové osi v oboch grafoch rozdielne, toto tvrdenie nemusí byť na prvý pohľad také zrejmé. Z toho dôvodu by bolo vhodné, keby boli grafy prispôsobené na jednotnú mierku. Najvýhodnejšie je zadefinovať minimum a maximum, aby sme si mohli byť istí, že všetky hodnoty sa v grafe box – plot zobrazia.

```
> par(mfrow = c(1, 2))
> minimum <- min(na.omit(data$Height))
> maximum <- max(na.omit(data$Height))
> boxplot(vyska_M, ylab = "výška v cm", main = "výška mužov",
  col = gray(0.8), pch = 19, cex.axis = 1.3, cex.lab = 1.5, ylim
  = c(minimum, maximum))
> abline(h = mean(vyska_M), lwd = 2, lty = 2, col = "red")
> boxplot(vyska_Z, ylab = "výška v cm", main = "výška žien", col
  = gray(0.8), pch = 19, cex.axis = 1.3, cex.lab = 1.5, ylim =
  c(minimum, maximum))
> abline(h = mean(vyska_Z), lwd = 2, lty = 2, col = "red")
```



Obrázok 5.7: Box – ploty výšky mužov a žien z databázy `survey` (rovnaké mierky)

Zdroj: vlastné spracovanie v programe R

Z takto zostavených grafov už je viditeľnejšie, že muži na danej univerzite sú vyšší ako ženy. Dolný kvartil výšky mužov je vyšší, ako horný kvartil výšky žien, čo v predchádzajúcom spôsobe zobrazenie box – plotov nebolo na prvý pohľad zrejmé. Práve porovnanie kvartilov dvoch súborov (v podstate „krabíc“) je neraz silným indikátorom toho, že rozdiely v dvoch súboroch nie sú náhodné, ale pomerne významné.

Príklad 5.18

Deskriptívnu štatistiku vypočítame pre zjednodušenie pomocou funkcie `summary()`. Z uzatváracích cien však toho veľa zistiť nevieme. Nie je totiž také zaujímavé, pri akých cenách sa indexy obchodujú. Podstatnejšie sú zmeny v týchto cenách. Pre tieto účely sa namiesto rozdielov často používajú spojité výnosy.

```
> library(datasets)
> ceny <- data.frame(EuStockMarkets)
> summary(ceny)

      DAX             SMI             CAC             FTSE
 Min.   :1402   Min.   :1587   Min.   :1611   Min.   :2281
 1st Qu.:1744   1st Qu.:2166   1st Qu.:1875   1st Qu.:2843
 Median :2141   Median :2796   Median :1992   Median :3247
 Mean   :2531   Mean   :3376   Mean   :2228   Mean   :3566
 3rd Qu.:2722   3rd Qu.:3812   3rd Qu.:2274   3rd Qu.:3994
 Max.   :6186   Max.   :8412   Max.   :4388   Max.   :6179
```

Zo spojitých výnosov už vieme zistiť viac. Môžeme vidieť, že najväčší denný pokles zaznamenal nemecký DAX (–0.096). Najväčší denný nárast zaznamenal francúzsky CAC (0.06). Zaujímať by nás mohla tiež variabilita výnosov (v podobe smerodajnej odchýlky), ktorú si vieme dopočítať pomocou funkcie `sd()`. Najvyššiu variabilitu vykazujú indexy DAX a CAC. Variabilita výnosov je určitým indikátorom neistoty investorov na trhu.

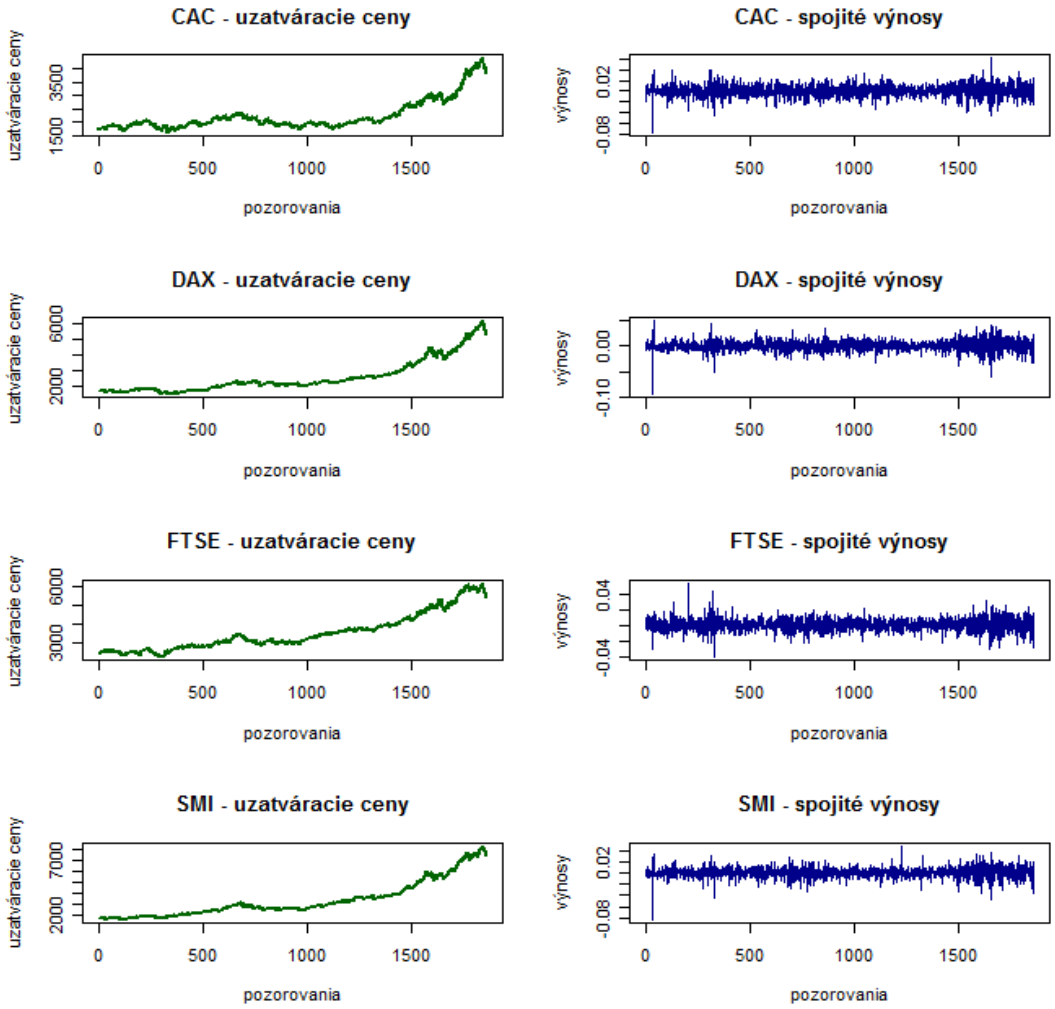
```
> library(datasets)
> attach(data.frame(EuStockMarkets))
> rCAC <- diff(log(CAC))
> rDAX <- diff(log(DAX))
> rFTSE <- diff(log(FTSE))
> rSMI <- diff(log(SMI))
-----------------------------------------------------------------
> vynosy <- data.frame(rCAC, rDAX, rFTSE, rSMI)
> summary(vynosy)

      rCAC                 rDAX                 rFTSE
 Min.   :-0.0757532   Min.   :-0.0962770   Min.   :-4.140e-02
 1st Qu.:-0.0060632   1st Qu.:-0.0046854   1st Qu.:-4.319e-03
 Median : 0.0000000   Median : 0.0004726   Median : 8.021e-05
 Mean   : 0.0004371   Mean   : 0.0006520   Mean   : 4.320e-04
 3rd Qu.: 0.0070965   3rd Qu.: 0.0063553   3rd Qu.: 5.254e-03
 Max.   : 0.0609773   Max.   : 0.0507601   Max.   : 5.440e-02
      rSMI
 Min.   :-0.0838250
 1st Qu.:-0.0038034
 Median : 0.0008858
 Mean   : 0.0008179
 3rd Qu.: 0.0060738
 Max.   : 0.0496798
-----------------------------------------------------------------
> sd(rCAC)
[1] 0.01103088
> sd(rDAX)
[1] 0.01030084
> sd(rFTSE)
[1] 0.007957728
> sd(rSMI)
[1] 0.009250036
```

Pre lepšiu predstavu je vždy vhodné pozrieť sa na vývoj skúmaných premenných v grafe. Vedľa uzatváracích cien môžeme priamo načrtnúť spojité výnosy a na prvý pohľad je potom zrejmé, do akej miery sú tieto dve skupiny časových radov rozdielne.

```
> 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(4, 2))
> plot(CAC, type = "l", ylab = "uzatváracie ceny", main = "CAC -
  uzatváracie ceny", xlab = "pozorovania", col = "darkgreen",
  lwd=2)
> plot(rCAC, type = "l", ylab = "výnosy", main = "CAC - spojité
  výnosy", xlab = "pozorovania", col = "darkblue", lwd = 1)
-----------------------------------------------------------------
> plot(DAX, type = "l", ylab = "uzatváracie ceny", main = "DAX -
  uzatváracie ceny", xlab = "pozorovania", col = "darkgreen",
  lwd=2)
> plot(rDAX, type = "l", ylab = "výnosy", main = "DAX - spojité
  výnosy", xlab = "pozorovania", col = "darkblue", lwd = 1)
-----------------------------------------------------------------
> plot(FTSE, type = "l", ylab = "uzatváracie ceny", main = "FTSE
  - uzatváracie ceny", xlab = "pozorovania", col = "darkgreen",
  lwd = 2)
> plot(rFTSE, type = "l", ylab = "výnosy", main = "FTSE -
  spojité výnosy", xlab = "pozorovania", col = "darkblue", lwd =
  1)
-----------------------------------------------------------------
> plot(SMI, type = "l", ylab = "uzatváracie ceny", main = "SMI -
  uzatváracie ceny", xlab = "pozorovania", col = "darkgreen",
  lwd = 2)
> plot(rSMI, type = "l", ylab = "výnosy", main = "SMI - spojité
  výnosy", xlab = "pozorovania", col = "darkblue", lwd = 1)
```



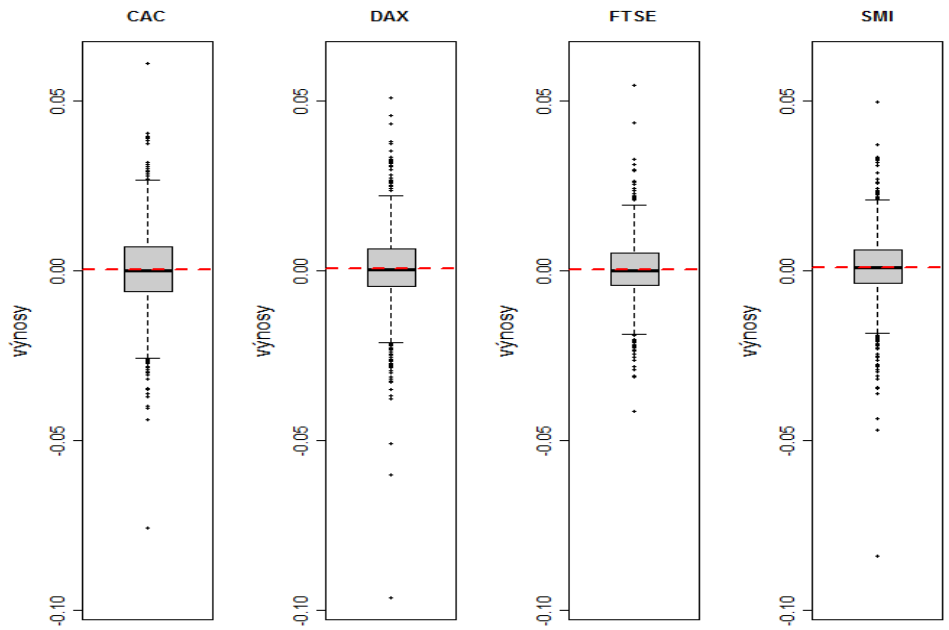
Obrázok 5.8: Uzatváracie ceny a spojité výnosy skúmaných indexov

Zdroj: vlastné spracovanie v programe R

Na záver ešte zostrojíme box – ploty pre spojité výnosy daných indexov. Podľa zadania máme priemernú hodnotu naniesť do týchto grafov a rozhodnúť, ktorý index dosahuje najvyšší priemerný denný výnos. Keďže rozdiely sú minimálne (ako sme mohli vidieť z deskriptívnej štatistiky) a priemery všetkých indexov sú blízke nule, tak získať z grafickej podoby takúto informáciu zrejme nebude možné.

```
> data <- data.frame(rCAC, rDAX, rFTSE, rSMI)
> par(mfrow = c(1, 4))
> minimum <- min(sapply(data, min))
> maximum <- max(sapply(data, max))
-----------------------------------------------------------------
> boxplot(rCAC, ylab = "výnosy", main = "CAC", col = gray(0.8),
  pch = 19, cex.axis = 1.3, cex.lab = 1.5, ylim = c(minimum,
  maximum))
> abline(h = mean(rCAC), lwd = 2, lty = 2, col = "red")
-----------------------------------------------------------------
```

```
> boxplot(rDAX, ylab = "výnosy", main = "DAX", col = gray(0.8),
  pch = 19, cex.axis = 1.3, cex.lab = 1.5, ylim = c(minimum,
  maximum))
> abline(h = mean(rDAX), lwd = 2, lty = 2, col = "red")
-----------------------------------------------------------------
> boxplot(rFTSE, ylab = "výnosy", main = "FTSE", col =
  gray(0.8), pch = 19, cex.axis = 1.3, cex.lab = 1.5, ylim =
  c(minimum, maximum))
> abline(h = mean(rFTSE), lwd = 2, lty = 2, col = "red")
-----------------------------------------------------------------
> boxplot(rSMI, ylab = "výnosy", main = "SMI", col = gray(0.8),
  pch = 19, cex.axis = 1.3, cex.lab = 1.5, ylim = c(minimum,
  maximum))
> abline(h = mean(rSMI), lwd = 2, lty = 2, col = "red")
```



Obrázok 5.9: Box – ploty spojitých výnosov skúmaných indexov

Zdroj: vlastné spracovanie v programe R

Môžeme vidieť, že spojité výnosy každého z indexov vykazujú viacero extrémnych hodnôt. Na prvý pohľad vieme identifikovať, ktorý index dosahuje minimálny denný výnos z danej vzorky (DAX), rovnako taktiež ktorý maximálny (CAC). Vyjadriť sa k otázke, ktorý index dosahuje v priemere najvyšší výnos však z týchto box – plotov nie je veľmi možné.

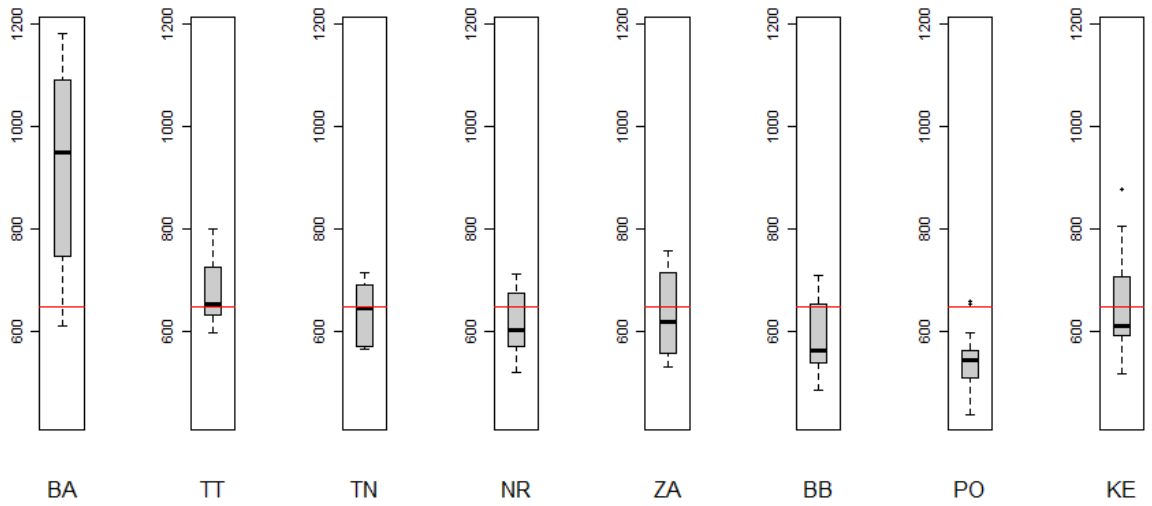
Príklad 5.19

Pri zisťovaní rozdielov medzi jednotlivými krajmi na základe grafickej vizualizácie dát využijeme box – ploty. Najprv sa pozrieme na reálne mzdy a cez funkciu `subset()` vytvoríme z okresov 8 premenných podľa príslušnosti k danému kraju (premenná `kraj`).

```
> data <- read.csv(file = "...cesta k súboru...\\sk_data.csv",
  sep = ";", dec = ".", header = T)
> attach(data)
-----------------------------------------------------------------
> BA <- subset(real_mzda, subset = kraj == 1)
> TT <- subset(real_mzda, subset = kraj == 2)
> TN <- subset(real_mzda, subset = kraj == 3)
> NR <- subset(real_mzda, subset = kraj == 4)
> ZA <- subset(real_mzda, subset = kraj == 5)
> BB <- subset(real_mzda, subset = kraj == 6)
> PO <- subset(real_mzda, subset = kraj == 7)
> KE <- subset(real_mzda, subset = kraj == 8)
```

Následne môžeme pristúpiť k vytvoreniu box – plotov obdobne, ako v predchádzajúcich príkladoch, t.j. zadefinovaním maximálnej/minimálnej hodnoty reálnej mzdy a nastavením rovnakej mierky v grafoch. Takýmto spôsobom by prípadné rozdiely medzi krajmi mali byť viditeľné na prvý pohľad. Ak by sme zostrojili všetky boxploty „zvlášť“, výsledný obrázok by sa pri tak veľkom počte boxplotov stal málo prehľadným.

```
> par(mfrow = c(1, 8))
> minimum <- min(real_mzda)
> maximum <- max(real_mzda)
-----------------------------------------------------------------
> boxplot(BA, xlab = "BA", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(TT, xlab = "TT", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(TN, xlab = "TN", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(NR, xlab = "NR", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(ZA, xlab = "ZA", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(BB, xlab = "BB", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(PO, xlab = "PO", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
> boxplot(KE, xlab = "KE", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
```

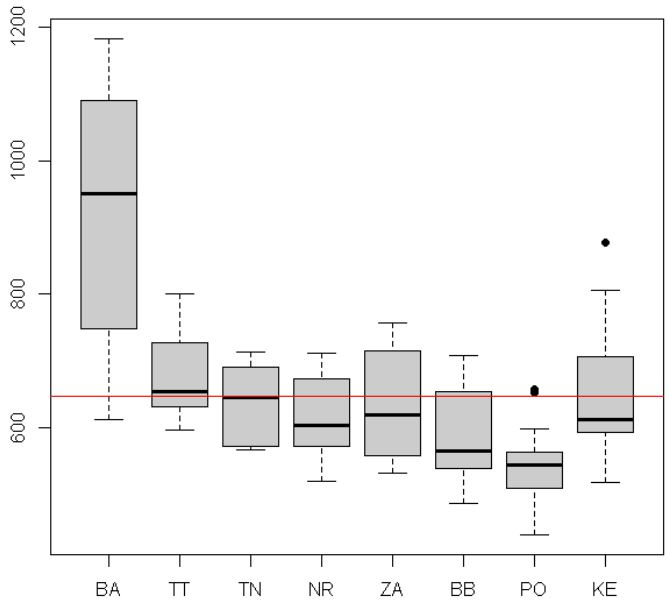


Obrázok 5.10: Box – ploty reálnych miezd v krajoch SR (1 spôsob)

Zdroj: vlastné spracovanie v programe R

Prehľadnejší a zároveň rýchlejšie vytvorený je obrázok, ktorý spája všetky box – ploty do jedného súradnicového systému.

```
> boxplot(real_mzda ~ kraj, col = gray(0.8), pch = 19, cex.axis
  = 1, cex.lab = 1.5, xaxt = "n")
> axis(1, at = c(1:8), labels = c("BA", "TT", "TN", "NR", "ZA",
  "BB", "PO", "KE"))
> abline(h = mean(real_mzda), lwd = 1, lty = 1, col = "red")
```



Obrázok 5.11: Box – ploty reálnych miezd v krajoch SR (2. spôsob)

Zdroj: vlastné spracovanie v programe R

Z uvedených grafov je zrejmé, že reálna mzda v Bratislavskom kraji je vyššia ako v ostatných krajoch SR. Okrem Trnavského kraja, medián reálnej mzdy je vo všetkých

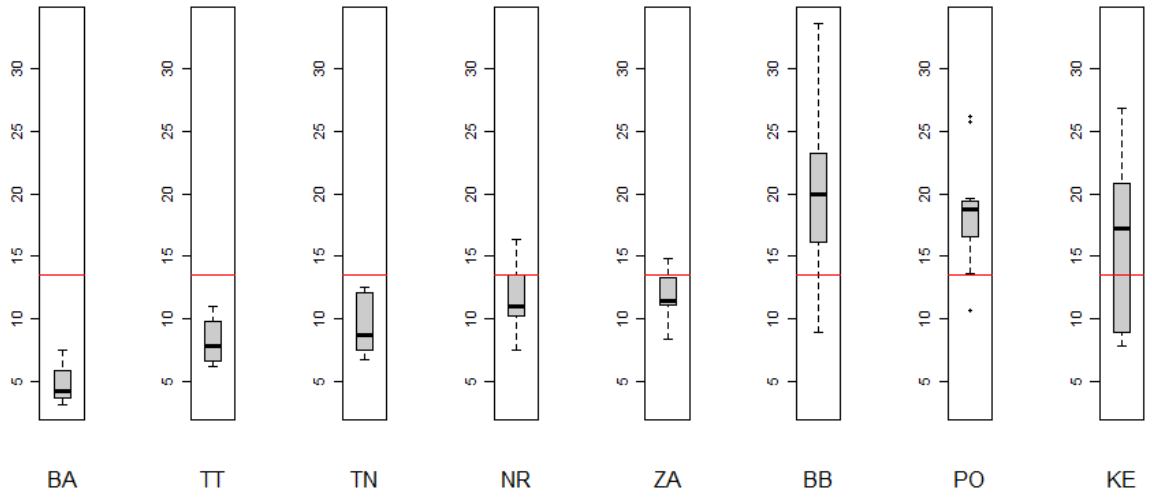
ostatných krajoch nižší ako priemer v SR (v grafoch zvýraznený červenou čiarou). Najnižšie reálne mzdy sa dosahujú v okresoch Prešovského kraja. Môžeme teda konštatovať, že len z grafickej podoby dát o reálnych mzdách v jednotlivých krajoch v SR je možné pozorovať značné rozdiely. Pri vizualizácii druhej premennej (nezamestnanosť v okresoch SR) budeme postupovať rovnakým spôsobom.

```
> data <- read.csv(file = "...cesta k súboru...\\sk_data.csv",
  sep = ";", dec = ".", header = T)
> attach(data)
-----------------------------------------------------------------
> BA <- subset(nezamestnanost, subset = kraj == 1)
> TT <- subset(nezamestnanost, subset = kraj == 2)
> TN <- subset(nezamestnanost, subset = kraj == 3)
> NR <- subset(nezamestnanost, subset = kraj == 4)
> ZA <- subset(nezamestnanost, subset = kraj == 5)
> BB <- subset(nezamestnanost, subset = kraj == 6)
> PO <- subset(nezamestnanost, subset = kraj == 7)
> KE <- subset(nezamestnanost, subset = kraj == 8)
```

Najprv sme si opäť rozdelili údaje za jednotlivé okresy podľa premennej `kraj`, teda podľa príslušnosti okresu ku kraju v SR. Následne môžeme pristúpiť k vizualizácii dát formou box – plotov (prehľadnejšiu formu nechávame na čitateľa).

```
> par(mfrow = c(1, 8))
> minimum <- min(nezamestnanost)
> maximum <- max(nezamestnanost)
-----------------------------------------------------------------
> boxplot(BA, xlab = "BA", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
> boxplot(TT, xlab = "TT", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
> boxplot(TN, xlab = "TN", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
> boxplot(NR, xlab = "NR", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
> boxplot(ZA, xlab = "ZA", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
> boxplot(BB, xlab = "BB", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
```

```
> boxplot(PO, xlab = "PO", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
> boxplot(KE, xlab = "KE", col = gray(0.8), pch = 19, cex.axis =
  1, cex.lab = 1.5, ylim = c(minimum, maximum))
> abline(h = mean(nezamestnanost), lwd = 1, lty = 1, col =
  "red")
```



Obrázok 5.12: Box – ploty nezamestnanosti v krajoch SR

Zdroj: vlastné spracovanie v programe R

Najnižšia nezamestnanosť je zaznamenaná v okresoch Bratislavského kraja a najvyššia v okresoch Banskobystrického kraja. Najvyššiu variabilitu v nezamestnanosti po okresoch vykazuje tiež Banskobystrický kraj. Evidentne aj pri tejto premennej existujú medzi krajmi v SR značné rozdiely. Priemerná nezamestnanosť (vyznačená v grafoch červenou čiarou) je „ťahaná“ smerom hore okresmi z východného Slovenska a z Banskobystrického kraja. Ak by sme chceli získať lepšiu predstavu o skúmanej premennej, môžeme sa tiež pozrieť na deskriptívnu štatistiku, ako príklad uvádzame opisné charakteristiky pre nezamestnanosť.

```
> library(Rcmdr)
> numSummary(nezamestnanost, groups = kraj)
       mean       sd    0%    25%   50%     75%  100%  n
1  4.783750 1.467378  3.18  3.745  4.29  5.7725  7.46  8
2  8.282857 2.041566  6.15  6.635  7.87  9.8400 11.01  7
3  9.335556 2.337018  6.72  7.510  8.76 12.0500 12.49  9
4 11.758571 2.962518  7.52 10.210 11.00 13.5150 16.34  7
5 11.817273 1.962922  8.41 11.105 11.47 13.3000 14.87 11
6 19.650769 6.995498  8.95 16.160 19.95 23.2000 33.64 13
7 18.513077 4.220670 10.65 16.600 18.80 19.4300 26.18 13
8 16.293636 6.892237  7.82  8.930 17.21 20.8050 26.82 11
```

Zoznam použitej literatúry

[1] A Free Software Project. Dostupné na: <http://cran.r-project.org/doc/html/interface98-paper/paper_2.html>.

[2] BAUMÖHL, E. – LYÓCSA, Š. – VÝROST, T. 2011. *Fundamentálna analýza akciových trhov.* Košice : Elfa, 2011. 323 s. ISBN 978-80-8086-191-6

[3] COHEN, Y. – COHEN, J. Y. 2008. *Statistics and Data with R: An Applied Approach Through Examples*. Chichester : John Wiley & Sons, 2008. ISBN: 978-0-470-75805.

[4] CRAWLEY, M. J. 2007. *The R Book*. Chichester : John Wiley & Sons, 2007. ISBN 978-0-470-51024-7.

[5] DALGAARD, P. 2008. *Introductory Statistics with R: Second Edition*. New York : Springer Science+Business Media, 2008. ISBN 978-0-387-79054-1.

[6] DeGROOT, M. – SCHERVISH, M. J. 2011. *Probability and Statistics*. Boston : Pearson Education, 2011. ISBN 978-0-321-50046-5.

[7] DOYLE, P. G. 2006. *Grinstead and Snell's Introduction to Probability*, Version dated 4 July 2006. Dostupné na: < http://www.math.dartmouth.edu/~prob/prob/prob.pdf>.

[8] EVERITT, B. S. – HOTHORN, T. 2005. *HSAUR: A Handbook of Statistical Analyses Using R*. Dostupné na: < http://cran.r-project.org/web/packages/HSAUR/>.

[9] 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>.

[10] FOX, J. – WEISBERG, S. 2011. *An {R} Companion to Applied Regression, Second Edition*. Thousand Oaks : Sage, 2011. ISBN 141297514X.

[11] 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.

[12] FREEDMAN, D. – DIACONIS, P. 1981. On the histogram as a density estimator:L_2 theory. In: *Probability theory and related fields*, 1981, vol. 57, no. 4, p. 453 – 476. ISSN 0178-8051.

[13] GORARD, S. 2005. Revisiting a 90 – year old debate: The advantages of the mean absolute deviation. In: *British Journal of Educational Studies*, 2005, vol. 53, no. 4, p. 417 – 430. ISSN 0007-1005.

[14] KALINA, M. – BACIGÁL, T. – SCHIESSLOVÁ, A. 2010. Základy pravdepodobnosti a matematickej štatistiky. Slovenská technická univerzita v Bratislave, 2010. ISBN 978-80-227-3273-4. Dostupné na: <http://cran.r-project.org/doc/contrib/FundamentalsOfProbAndMStatistics-Slovak.pdf>.

[15] Knižnica datasets. Dostupné na: <http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/datasets-package.html>.

[16] MURELL, P. 2006. *R Graphics*. London : Chapman and Hall/CRC, 2006. ISBN: 1-58488-486-X.

[17] PANČÍKOVÁ, L. 2011. R manual. Dostupné na: <http://fria.fri.uniza.sk/~kmame/drupal/subory/pancikova/Rmanual/index.php>.

[18] Program R. Dostupný na: <http://cran.at.r-project.org/bin/windows/base/>.

[19] REVELLE, W. 2012. *psych: Procedures for Personality and Psychological Research Northwestern University*. Evanston, 2012. Dostupné na: <http://personality-project.org/r/psych.manual.pdf>.

[20] ROUSAS, G. G. 1997. *A Course in Mathematical Statistics, Second Edition*. London : Academic Press, 1997. ISBN 0-12-599315-3.

[21] SCOTT, D. W. 1979. On optimal and data-based histograms. In: *Biometrika*, 1979, vol. 66, no. 3, p. 605 – 610. ISSN 0006-3444.

[22] STEVENS, S. S. 1946. On the theory of scales of measurement. In: *Science*, 1946, vol. 103, no. 2684, p. 677 – 680. ISSN 0036-8075.

[23] STEVENS, S. S. 1951. *Mathematics, measurement, and psychophysics*, p. 1 – 49. In: STEVENS, S. S. *Handbook of Experimental Psychology*, New York : John Wiley, 1951. ISBN 0471823686.

[24] STURGES, H. A. 1926. The Choice of a Class Interval. In: *Journal of the American Statistical Association*, 1926, vol. 21, no. 153, p. 65 – 66. Dostupné na: <http://www.aliquote.org/cours/2012_biomed/biblio/Sturges1926.pdf>.

[25] TKÁČ, M. 2001. *Štatistické riadenie kvality*. Bratislava : Ekonóm, 2001. ISBN 80-225-0145-X.

[26] TRIOLA, M. F. 2004. *Elementary Statistics: Updates for the latest technology, 9th Updated Edition*. MA : Addison-Wesley, 2004, ISBN 0321288394.

[27] VENABLES, W. N. – RIPLEY, B. D. 2002. *Modern Applied Statistics with S, Fourth Edition*. New York : Springer, 2002. ISBN 0-387-95457-0.

[28] VERHOEFF, T. 1993. *The Laws of Large Numbers Compared*. Dostupné na: <http://www.dklevine.com/archive/strong-law.pdf >.

[29] VERZANI, J. 2004. *Using R for Introductory Statistics*. London : Chapman and Hall/CRC, 2004. ISBN 13-978-1584884507.

[30] WICKHAM, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York : Springer Science+Business Media, 2009. ISBN 978-0-387-98141-3.

[31] ZEILEIS, A. – HOTHORN, T. 2002. Diagnostic Checking in Regression Relationships. In: *R News*, vol. 2, no. 3, p. 7 – 10. Dostupné na: <http://cran.r-project.org/doc/Rnews/>.

[32] ZEILEIS, A. – KLEIBER, CH. – KRAMER, W. – HORNIK, CH. 2002, strucchange: An R Package for Testing Structural Change in Linear Regression Models. In: *Journal of Statistical Software*, 2002, vol. 7, no. 2, p. 1 – 38. ISSN 1548-7660.

[33] ŽELINSKÝ, T. – GAZDA, V. – VÝROST, T. 2010. *Ekonometria : študijný materiál pre externé vzdelávanie*. Košice : Technická Univerzita v Košiciach, 2010. ISBN 978-80-553-0389-5.

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] `Rcmdr`

[10] `lmtest 0.9-30`

[11] `psych 1.2.4`

[12] `zoo`

[13] `datasets 2.14.0`

Názov: Kvantitatívne metódy v ekonómii I.
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 978-80-8086-209-1

ISBN 978-80-8086-209-1