Univerzita Pavla Jozefa Šafárika v Košiciach
Lekárska fakulta

Základy (bio)štatistiky pre medikov

Jaroslav Majerník

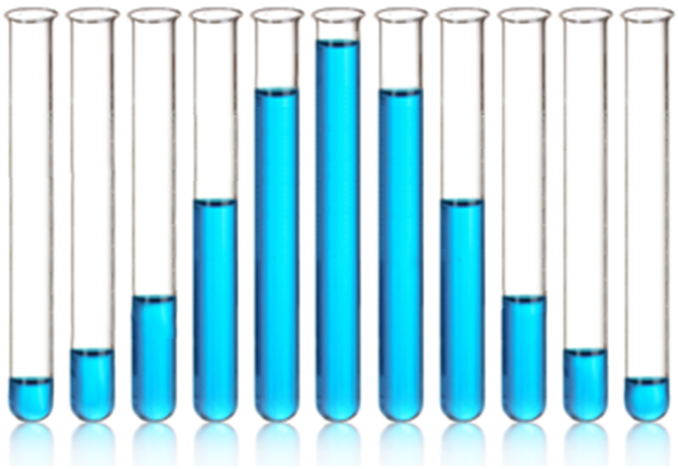
Košice 2021

Univerzita Pavla Jozefa Šafárika v Košiciach
Lekárska fakulta

Základy (bio)štatistiky

pre medikov

Jaroslav Majerník



Košice 2021

Táto vysokoškolská učebnica vznikla s prispením Kultúrnej a edukačnej grantovej agentúry Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky v rámci riešenia projektu KEGA 011UPJŠ-4/2019, „Zvyšovanie kompetencií a úrovne kritického myslenia študentov medicínskych študijných programov s využitím simulačných nástrojov problémovo orientovaného vzdelávania a medicíny založenej na dôkazoch“.

Základy (bio)štatistiky pre medikov

Vysokoškolská učebnica

Pracovisko: Ústav lekárskej informatiky, Lekárska fakulta, UPJŠ v Košiciach

Recenzenti:
doc. Ing. Daniel Schwarz, Ph.D.
Inštitút bioštatistiky a analýz, Lekárska fakulta, Masarykova univerzita

doc. Ing. Teodor Tóth, PhD.
Katedra biomedicínskeho inžinierstva a merania, Strojnícka fakulta, Technická univerzita v Košiciach

Za odbornú a jazykovú stránku vysokoškolskej učebnice zodpovedá autor. Rukopis neprešiel redakčnou ani jazykovou úpravou.

ISBN 978-80-574-0066-0

Predslov

Vysokoškolská učebnica *Základy (bio)štatistiky pre medikov* je určená primárne študentom lekárskych a zdravotníckych študijných programov na lekárskych fakultách, pričom jej hlavným cieľom je priblížiť študentom problematiku využitia štatistiky a jej vybraných metód spracovania a hodnotenia údajov v rôznych klinických výskumných úlohách. Svojim obsahom i rozsahom sa nesnaží nahradiť dostupnú odbornú literatúru venovanú podrobnému popisu, charakteristikám a dokazovaniu jednotlivých štatistických metód a operácií, keďže tieto sú pre neodbornú a laickú verejnosť veľmi často komplikované a nezrozumiteľné. Naopak, poslaním tejto učebnice je uviesť čitateľa nenáročným a názorným spôsobom do základov štatistického uvažovania, a to vysvetlením základných pojmov, metód i interpretácií výsledkov štatistického hodnotenia medicínsky orientovaných štúdií, s ktorými sa poslucháči medicíny môžu stretávať pri štúdiu rôznych odborných i vedeckých prác.

Aj tu je totiž nesmierne dôležité uvedomiť si, že všadeprítomná štatistika prispieva nielen k podpore rozhodovania, napríklad o výbere spoľahlivejšej diagnostickej metódy alebo účinnejšieho liečebného postupu, samozrejme s prihliadnutím na históriu a aktuálny stav konkrétneho pacienta, ale rovnako tak prináša často aj nedorozumenia vyplývajúce z nesprávnej interpretácie štatistických charakteristík. Študenti, ale aj samotní vedci a výskumníci by preto nemali „slepo“ veriť výstupom svojich štatistických aplikácií, zvlášť ak získané výsledky nedávajú racionálny význam, prípadne ak prezentované výsledky nekorešpondujú so závermi prezentovanými autormi prác. Množstvo výpočtových nástrojov pre oblasť štatistiky sa už dnes tak jednoducho ovláda, že potreba štatistických vedomostí sa vytráca. Avšak, tento postoj je nesprávny, keďže mnohí používatelia vykonávajú rôzne analýzy bez toho, aby vedeli čo program robí a tiež bez toho, aby vedeli ako interpretovať výsledky. V tomto kontexte sme potom svedkami neužitočnosti a nepoužiteľnosti veľkého množstva štúdií. Podstata kritického uvažovania v tejto oblasti preto okrem iného vychádza aj z pochopenia základných princípov štatistiky, ktoré napomáhajú rozpoznať, ktoré vedecké výsledky sú relevantné a ktoré sú vo svojej podstate nezmyselné.

Text tejto učebnice sa neopiera o spracovanie a hodnotenie údajov pomocou konkrétneho softvérového vybavenia, keďže štatistických programov a aplikácií umožňujúcich rýchle vykonanie štatistických výpočtov existuje veľké množstvo a spôsob práce s nimi býva navyše súčasťou ich sprievodnej dokumentácie, a často aj rozsiahlych používateľských manuálov. Rovnako tak si čitateľ môže vyberať z početného zastúpenia voľne dostupných i komerčných riešení, ktoré budú vyhovovať jeho požiadavkám a nárokom. Vo vybraných kapitolách tejto učebnice si však čitateľ nájde vložené informácie o niektorých funkciách MS Excel, pomocou ktorých si môže overiť výpočet vybraných štatistických charakteristík.

Snahou autora je teda priblížiť čitateľovi podstatu tej-ktorej štatistickej charakteristiky či štatistickej metódy, a to názorným spôsobom s vysvetlením jej významu a použiteľnosti v aplikačnej oblasti. Aj preto je samotný matematický aparát obmedzený na minimum, ktoré je nevyhnutné pre pochopenie hlavnej podstaty. Učebnica taktiež nepredstavuje sumár rozsiahleho aparátu štatistických testov a metód, ale poskytuje nevyhnutné informácie potrebné pre pochopenie základov štatistického uvažovania.

Nemenej dôležitým príspevkom učebnice je upriamenie pozornosti študentov na význam objektívneho hodnotenia najnovších medicínskych poznatkov, a tým zároveň aj posilnenie budovania kritického myslenia nevyhnutného pre úspešné a efektívne aplikovanie najlepších dostupných dôkazov v procesoch zameraných na zdravotnú starostlivosť o pacienta. Rozšírením obzoru vedeckého štatistického uvažovania tak študent získa základy využiteľné nielen vo svojom odbore, ale aj v rôznych situáciách bežného života.

Jaroslav Majerník

Obsah

Úvod

Princípy štatistického uvažovania nachádzame takmer vo všetkých vedných odboroch, ale stretávame sa s nimi aj v bežnom živote. Kým v prípade zariadení, tovarov, súčiastok a rôznych iných objektov vieme ich charakteristiky veľmi presne popísať, v prípade medicíny a zložitosti charakteristík živých organizmov čelíme okrem individuálnych rozdielov aj skutočnosti, že mnohé charakteristiky nie je možné detailne popísať a je pri nich potrebné brať do úvahy istý stupeň neurčitosti. Našťastie, matematický aparát prepracovaných štatistických metód ponúka možnosti, ako tieto problémy pri splnení konkrétnych predpokladov efektívne vyriešiť. Aj tu však samotné znalosti o existencii jednotlivých štatistických metód nepostačujú a je potrebné vedieť, kedy a ktorý nástroj štatistiky je vhodné použiť. Schopnosť štatistického uvažovania teda zahŕňa aj schopnosť vybrať správny štatistický model a vedieť získané výsledky aj správne interpretovať.

Štatistiku môžeme vo všeobecnosti chápať ako odbor matematiky, ktorý sa zaoberá zberom, organizáciou, klasifikáciou, prezentáciou, analýzou a interpretáciou údajov podporujúcich rozhodovanie na základe ich kvantitatívnych charakteristík. Inými slovami povedané, štatistika predstavuje metodológiu používanú k interpretácii a vyvodzovaniu záverov zo zozbieraných údajov. Aj preto je veľmi dôležité o spracovávaných informáciách, a teda aj o údajoch tvoriacich tieto informácie vedieť:

1. Ako sú získané?
2. Ako sú analyzované?
3. Ako sú interpretované?

Spôsob získavania údajov má vplyv jednak na možnosti ich ďalšieho spracovania, ako aj na relevantonsť odvodených záverov vzhľadom na reálny stav skúmaného problému. Rovnako je pochopiteľné, že nepresné údaje, napríklad údaje získané chybou merania či údaje zaťažené subjektívnymi faktormi na strane pozorovateľa budú viesť aj k väčšiemu riziku chybných záverov. Použitie nevhodnej štatistickej metódy v analýze hoc aj tých najpresnejších a reprezentatívnych údajov môže taktiež končiť znehodnotením celého experimentálneho

výskumu. V neposlednom rade je potrebné si uvedomiť, že je dôležité získať aj skúsenosti správnej interpretácie jednotlivých kvantitatívnych i kvalitatívnych štatistických charakteristík.

Pri práci s údajmi, ktoré získavame v rámci rôznych, viac-či menej vedecky orientovaných úloh, sa pomocou štatistiky a jej metód snažíme dopracovať k pochopeniu skúmaného problému. Samotný proces analýzy údajov pritom prechádza viacerými krokmi. Na začiatku je potrebné formulovať výskumný problém, jeho zámer a ciele, ktoré chceme výskumom dosiahnuť. Nasleduje definovanie základného súboru objektov alebo subjektov, ktoré sú predmetom záujmu a rovnako aj výberového súboru, na ktorom sa výskum bude realizovať (spravidla nie je možné získať údaje od všetkých objektov alebo subjektov, viď nasledujúce kapitoly 1 a 2). Na skupine skúmaných objektov alebo subjektov následne získavame, pozorujeme alebo odmeriavame a zaznamenávame hodnoty požadovaných údajov. Na získanom súbore hodnôt vykonáme popisnú analýzu údajov, ktorá nám poskytne prvý obraz o skúmanom probléme. V mnohých prípadoch prieskumov či dotazníkových úloh sa stretávame s tým, že organizáciou a sumarizáciou zozbieraných informácií pomocou charakteristík popisnej štatistiky sa tu štatistické hodnotenie končí. Vo vedeckých štúdiách a výskumne orientovaných úlohách však výsledky popisnej analýzy údajov ďalej použijeme na výber vhodných štatistických metód a uplatníme ich na riešenie stanoveného problému. Až na základe výsledkov následného štatistického testovania pomocou zvolených metód realizujeme odvodenie a prezentáciu záverov, ku ktorým štatistická analýza údajov dospela.

Z uvedeného vyplýva, že hlavým cieľom štatistiky býva najmä podpora pri robení záverov o základnom súbore, a to na základe informácií obsiahnutých v nám dostupných výberových súboroch, vrátane posúdenia rozsahu či miery neistoty obsiahnutej v týchto záveroch.

Kapitola 1

Získavanie údajov

Údaje, ktoré spracovávame pomocou štatistických metód môžu byť získavané z rôznych zdrojov. Vo všeobecnosti je možné údaje zbierať a následne analyzovať v rámci nových štúdií a experimentov, alebo je možné použiť existujúce zdroje údajov, ktoré boli získané v predošlých obdobiach inými výskumnými tímami. Nové štúdie sú realizované spravidla vtedy, ak nie je možné stanovené ciele naplniť na základe existujúcich údajov. Na druhej strane, nové údaje nie je potrebné získavať, ak už údaje boli získané v predošlých podobných štúdiách či experimentoch, ale niektoré analýzy zamerané na naplnenie cieľov novej štúdie neboli realizované. V mnohých štúdiách sa totiž neanalyzujú všetky získané údaje, prípadne sa nerobia všetky možné analýzy. Údaje, ktoré boli získané prvýkrát v rámci štúdie alebo experimentu, a ktoré neboli predtým analyzované, označujeme aj ako primárne údaje. Údaje, ktoré boli získané v minulosti niekym iným sú zasa označované aj ako sekundárne údaje. K takýmto údajom patria napríklad zdravotné záznamy pacientov, interné záznamy rôznych inštitúcií, internetové zdroje, údaje publikované v časopisoch a pod.

1.1 Metódy zberu údajov

Historicky najznámejším spôsobom získavania informácií pre odvodenie štatistického popisu sledovaných charakteristík bolo úplné sčítavanie všetkých objektov, resp. subjektov, ktoré boli predmetom záujmu. Napríklad, pravidelné sčítanie obyvateľstva žijúceho v konkrétnej oblasti alebo krajine. Aj keď si laická verejnosť práve tento prístup často asociuje s pojmom štatistika, je potrebné poznamenať, že samotné zozbieranie informácií o všetkých pozorovaných objektoch alebo subjektoch ešte nezaručuje správne vyhodnotenie a interpretácie výsledkov. Preto je pri štatistickom uvažovaní potrebné poznať nielen samotné informácie, ale aj to, aké postupy boli použité na ich zber i za akých podmie-

nok boli zozbierané. To nám umožní okrem iného vykonávať opakovanie rôznych výskumne orientovaných štúdií, ale aj objektívne porovnávanie ich výsledkov s výsledkami získanými inými výskumnými tímami pracujúcimi na identických alebo príbuzných úlohách.

Hlavným východiskom pri realizácii rôznych typov výskumných úloh je formulácia skúmaného problému (prípadne skupiny problémov) a stanovenie cieľov, ktoré má plánovaný výskum priniesť. Kým v rámci jednoduchších výskumných úloh sa môže jednať napríklad o dotazníkový zber základných faktov či všeobecný opis aktuálneho stavu skúmaného problému, u komplexných vedecky orientovaných výskumov sú formulované a overované hypotézy, a na základe ich výsledkov sú odvodené praktické odporúčania, ktoré často ovplyvňujú ďalší vývoj a smerovanie v danej oblasti. Avšak, aby boli odvodené závery relevantné a spoľahlivé, jedným zo základných predpokladov je, aby boli spracovávané aj relevantné a správne údaje, a teda je potrebné vedieť ako boli údaje získané, t. j. je potrebné poznať použitú metódu zberu údajov.

Realizácia výskumu si okrem odborných poznatkov riešiteľov a vynaloženia zodpovedajúceho úsilia vyžaduje vždy aj finančné a časové náklady, ktoré by mali zodpovedať získaným informáciám a dosiahnutému výsledku. Zníženie nákladov je možné zabezpečiť napríklad využitím už existujúcich údajov (údaje z predošlých štúdií, existujúce zdravotné záznamy pacientov a pod.), ktoré sú síce rýchlo k dispozícii, ale nemusia byť úplné, aktuálne či zodpovedajúce účelu výskumu, keďže mohli byť získavané s iným zámerom. Rovnako tak môžu pochádzať od rôznych zdrojov, čo jednak zvyšuje riziko vzniku chýb v odvodených záveroch a ich výslednej interpretácii, ako aj znižuje možnosti následného opakovania či porovnávania výskumných štúdií. Na druhej strane, v závislosti od riešenej problematiky, je často nevyhnutné získať aktuálne údaje, napríklad na skupine pacientov liečených novými metódami prípadne novým liekom, kedy musíme počítať tak s finančnými, ako aj s časovými nárokmi. S týmito faktormi, okrem mnohých iných, je potrebné vždy uvažovať pri plánovaní realizácie výskumných úloh. Avšak, táto problematika už nie je predmetom záujmu tejto učebnice.

Vo všeobecnosti je možné definovať nasledovné základné skupiny zberu údajov, v rámci ktorých je možné identifikovať viacero ďalších podskupín (spôsobov realizácie získavania požadovaných informácií).

- **Sčítanie** (populácie) – získava údaje od každého subjektu (člena základného súboru – populácie). Poskytuje najpresnejšie informácie, ale je aj časovo i finančne náročné.

- **Výberový prieskum** – získava údaje z podskupiny (časti) základného súboru (populácie) na odvodenie charakteristík (celej) populácie. Potrebných je menej personálnych, časových i finančných zdrojov.
- **Experiment** – riadená štúdia (napríklad klinická), kde sa vedci snažia pochopiť závislosti medzi jednotlivými skúmanými premennými, t. j. súvislosti medzi príčinou a dôsledkom. Analyzuje sa účinok (vplyv) nezávislej premennej na závislú (napríklad účinok liečby). Systémová chyba (*bias*) býva v porovnaní so štúdiami, u ktorých nie je zber údajov realizovaný podľa presne stanovených podmienok menšia.
- **Pozorovanie** – skúma súvislosti medzi príčinou a dôsledkom, avšak výskumníci neovplyvňujú zaradenie subjektov do jednotlivých skupín a ani to, ktorá skupina podstúpi akú liečbu. Systémová chyba (*bias*) má väčší vplyv na výsledok, t. j. závery môžu byť menej presné.

Metódy zberu údajov patriace pod skupiny sčítania a výberových prieskumov sú používané na pochopenie skúmaných parametrov základného súboru (populácie). Hlavné charakteristiky týchto dvoch kategórií sú uvedené v tabuľke 1.1.

Tabuľka 1.1: Vlastnosti sčítania a výberového prieskumu.

Sčítanie	**Výberový prieskum**
- celá populácia	- podskupina populácie
- presné hodnoty parametrov	- odhaduje parametre populácie
- časovo a finančne náročné	- rýchlejší, lacnejší a jednoduchší

Metódy zberu údajov patriace pod skupiny experimentov a pozorovaní sú používané na pochopenie závislostí medzi skúmanými premennými. Hlavné charakteristiky týchto dvoch kategórií sú uvedené v tabuľke 1.2.

Tabuľka 1.2: Vlastnosti experimentu a pozorovania.

Experiment	**Pozorovanie**
- cielene vytvorené skupiny subjektov	- prirodzene existujúce skupiny subjektov
- liečba kontrolovaná výskumníkmi	- žiaden vplyv výskumníkov na určenie liečby
- menšie systémové chyby	- viac náchylné na systémové chyby

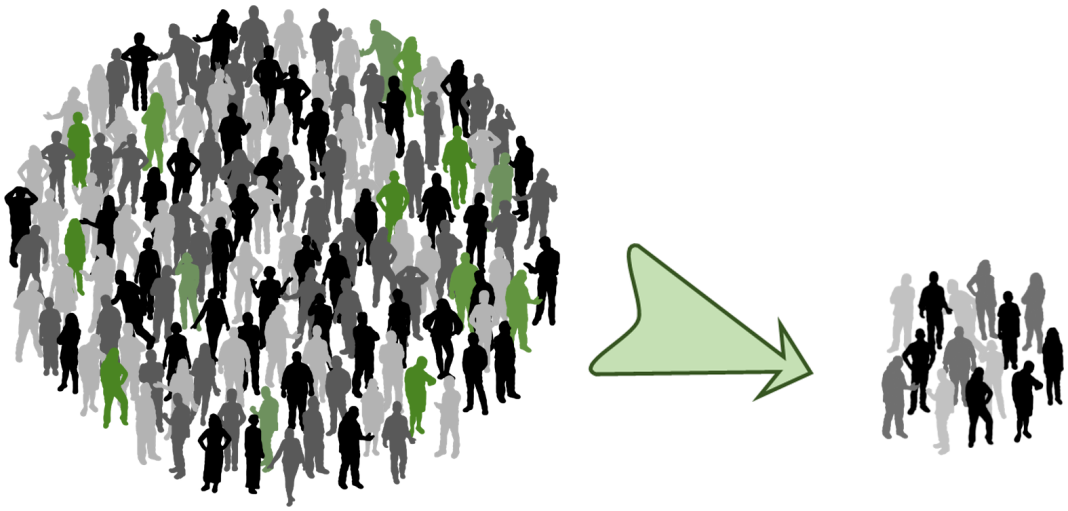
1.2 Základný a výberový súbor

Medzi najzákladnejšie pojmy štatistiky patria základný súbor a výberový súbor. V štatistickej literatúre sa tieto pojmy často označujú aj ekvivalentnými pojmami odvodenými z anglosaskej terminológie, ktorými sú populácia (*population*) a vzorka (*sample*).

Základný súbor je možné charakterizovať ako súbor všetkých objektov alebo subjektov (štatistických jednotiek), ktorých štatistické znaky sú predmetom záujmu (pozorovania, výskumu) výskumníkov v ich štúdii. Predstavuje teda najväčšiu možnú množinu hodnôt náhodnej veličiny – študovanej premennej, na ktorej môžeme realizovať štúdiu, napríklad všetci pacienti s určitou diagnózou v danom čase. Prirodzene, zloženie základného súboru sa v čase mení (novo diagnostikovaní pacienti, vyliečení pacienti, zomrelí a pod.). Ak je základný súbor zložený z konečného počtu hodnôt, potom hovoríme o konečnom základnom súbore. Na druhej strane, ak je základný súbor určený tak, že má nekonečnú postupnosť hodnôt, potom ho označujeme ako nekonečný základný súbor. Veľkosť základného súboru označujeme ako N a je to celkový počet objektov alebo subjektov v základnom súbore.

Málokedy sú však získavané požadované údaje pre všetky objekty alebo subjekty obsiahnuté v základnom súbore. Je totiž takmer nemožné získať údaje od každej štatistickej jednotky v krátkom časovom intervale, disponovať množstvom potrebného vybavenia a zázemia, veľkým počtom personálu s potrebnými skúsenosťami či zodpovedajúcim finančným zabezpečením. Navyše, riziko nezískania všetkých údajov by bolo príliš vysoké, čo by v konečnom dôsledku zvyšovalo systémovú chybu (*bias*). Častejšie, najmä z vyššie uvedených časových, organizačných, personálnych či finančných dôvodov, je však pozorovaný len vybraný reprezentatívny súbor objektov alebo subjektov základného súboru, a takýto súbor potom označujeme ako výberový súbor, t. j. vybraný počet štatistických jednotiek reprezentujúcich základný súbor. Napríklad náhodne vybraná skupina pacientov s danou diagnózou, pacienti z určitého regiónu, pacienti konkrétnej vekovej skupiny a pod. Veľkosť výberového súboru označujeme ako n, pričom platí, že $n < N$ a reprezentuje počet objektov alebo subjektov vybraných do výberového súboru zo základného súboru.

V praxi nebývajú skúmané všetky charakteristiky či vlastnosti jednotlivých objektov alebo subjektov, ktorí sú predmetom záujmu konkrétneho výskumu alebo cielene zameranej štúdie. Preto v rámci ich realizácie nie je požadované merať alebo získať všetky informácie od jednotlivcov v základnom súbore. Táto skutočnosť zvýrazňuje dôležitosť určenia skupiny požadovaných meraní. Potom



Obr. 1.1: Základný súbor (vľavo) a výberový súbor (vpravo) štatistických jednotiek.

môžeme hovoriť o tzv. štatistickom základnom súbore reprezentujúcom skupinu meraní alebo záznamov nejakého štatistického znaku (veličiny), ktorá zodpovedá celému súboru jednotiek, pre ktoré sa majú odvodiť závery. Výberovým súborom zo štatistického základného súboru potom bude súbor meraní, ktoré sa skutočne zhromažďujú pre potreby danej výskumnej úlohy.

Základný súbor môže byť konečný, teda taký, ktorý vieme fyzicky uviesť či popísať. Napríklad pacienti hospitalizovaní v univerzitnej nemocnici, alebo lieky dostupné v lekárni. Avšak, v mnohých prípadoch môže byť základný súbor abstraktný (hypotetický) a môže vyplývať zo skúmaného problému. Napríklad závery odvodené z výberového súboru pacientov s určitou diagnózou v aktuálnej štúdii môžu byť použité na odvodenie záverov pre pacientov, ktorí budú mať takéto ochorenie v budúcnosti. V oboch prípadoch základný súbor vždy predstavuje cieľ danej štúdie, t. j. odvodzujeme závery o základnom súbore na základe štatistických jednotiek vybraných do výberového súboru, ktoré sú získané procesom vzorkovania.

1.3 Výber zo základného súboru

Výber štatistických jednotiek je jednou z hlavných úloh, ktorá využíva spravidla presne stanovené postupy na zabezpečenie reprezentatívnosti výberového súboru reprezentujúceho základný súbor, z ktorého je výberový súbor vybraný. Výberový súbor vybraný zo základného súboru v rámci náhodného experimentu je označovaný aj ako náhodný výberový súbor. Náhodným výberom je tak možné vybrať množstvo náhodných výberových súborov o veľkosti n zo základného súboru o veľkosti N. Výber ktorejkoľvek štatistickej jednotky

do výberového súboru je závislý na šanci respektíve pravdepodobnosti, že bude daná štatistická jednotka vybraná zo základného súboru. Vo všeobecnosti, každý prvok základného súboru by mal mať rovnakú pravdepodobnosť, že sa stane prvkom výberového súboru. U jednoduchého náhodného výberu by teda malo platiť, že o tom, či sa vybrala alebo nevybrala jednotka základného súboru rozhoduje len náhoda.

Podľa opakovateľnosti jednoduchého náhodného výberu rozoznávame:

- **náhodný výber s opakovaním** – štatistická jednotka je vyberaná vždy z rovnakého základného súboru, t. j. môže byť zaradená do výberového súboru viackrát (rôzne štatistické experimenty kedy sa vyberá napríklad jedna karta vždy z balíka všetkých kariet – po výbere sa karta vracia späť do balíka a nasleduje ďalší výber),
- **náhodný výber bez opakovania** – štatistická jednotka môže byť vo výberovom súbore len raz, t. j. pravdepodobnosť výberu jednotky závisí od predchádzajúcich výberov (rôzne klinické štúdie, kedy jeden pacient býva zaradený len do jednej skupiny, napríklad experimentálnej alebo kontrolnej).

Snahou výberu vzoriek do výberového súboru je získať čo najviac informácií pri zachovaní fixných nákladov. Okrem jednoduchého náhodného výberu sa využívajú aj alternatívne metódy. Jednou z nich je tzv. stratifikovaný výber, kedy je z hľadiska veľkosti vzorky finančne efektívnejšie prvky základného súboru rozdeliť do neprekrývajúcich sa skupín – vrstiev. Potom jednoduchý náhodný výber z každej vrstvy poskytne stratifikovanú vzorku. Inou metódou je systematický výber, pri ktorom sa namiesto vzorkovania každej jednotky naraz zvolí vhodný interval rozdelenia jednotiek základného súboru a počiatočný bod výberu sa vyberie náhodne z prvého intervalu. Každý ďalší prvok sa následne vyberie v rovnakých intervaloch.

Kapitola 2

Štatistika a bioštatistika

Aj neskúsenému čitateľovi je zrejmé, že štatistika je oveľa viac než iba vkladanie hodnôt do tabuliek textových editorov či tabuľkových kalkulátorov a ich prípadné grafické zobrazenie. Vo všeobecnom ponímaní pre nás štatistika predstavuje vedu o získavaní informácií z kvantitatívnych a kvalitatívnych údajov. Odhliadnuc od rozmanitosti štatistických metód nám štatistika ponúka možnosti odpovedí na otázky súvisiace s prípravou, realizáciou a hodnotením výskumných úloh. Vieme napríklad určiť aké údaje a koľko ich potrebujeme získať, ako ich máme organizovať a sumarizovať, ako ich vieme analyzovať a vyhodnocovať, prípadne aj to, ako môžeme určiť silu odvodených záverov, či vyhodnotiť mieru ich neistoty. To znamená, že štatistika nám ponúka metódy, ktoré umožňujú realizovať:

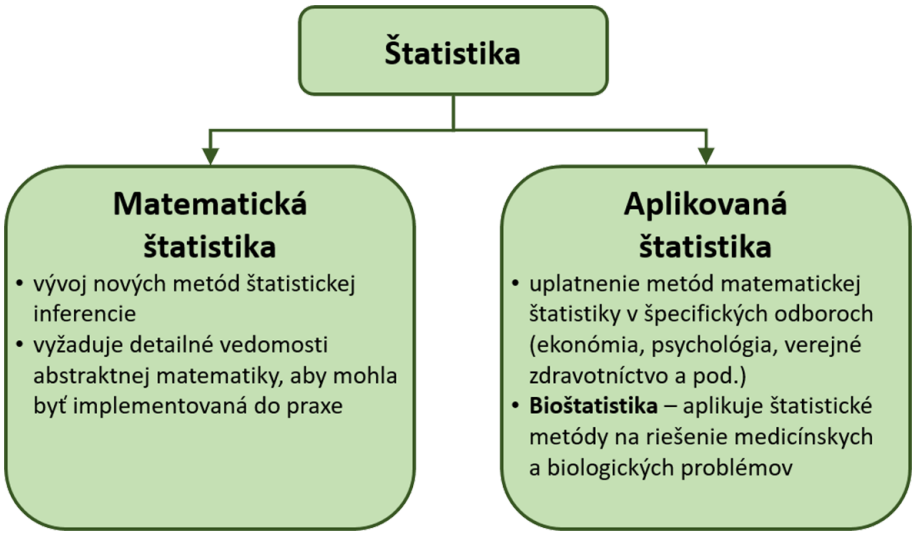
- návrh – plánovanie a realizáciu výskumných štúdií,
- popis – sumarizáciu a skúmanie údajov,
- závery (inferenciu) – predpovedanie a zovšeobecňovanie javov reprezentovaných údajmi.

Okrem tohto globálneho záberu je štatistika vedou, ktorá sa zaoberá aj neistými javmi a udalosťami. Preto je dnes štatistika aplikovaná prakticky vo všetkých vedných oblastiach. Pre cieľovú skupinu čitateľov tejto učebnice budú samozrejme zaujímavé oblasti medicíny, napríklad zamerané na štúdium účinnosti liečebných postupov, príčin vzniku rakoviny, dedičnosti ochorení, prognostických faktorov alebo potenciálnych rizík súvisiacich s konkrétnym ochorením, možnosti prežitia alebo zníženia mortality a pod.

2.1 Matematická a aplikovaná štatistika

Štatistiku, resp. prístupy využitia štatistiky je možné rozdeliť do dvoch základných skupín (obrázok 2.1). Prvú skupinu tvorí matematická štatistika, ktorá sa

zaoberá samotným vývojom a overovaním matematického aparátu štatistiky. Vyžaduje si skúsenosti a detailné vedomosti z oblasti abstraktnej matematiky, tak aby mohli byť štatistické metódy využiteľné v praxi. Druhá skupina štatistiky predstavuje tzv. aplikovanú štatistiku, ktorá uplatňuje metódy matematickej štatistiky v rôznych odboroch, ako sú napríklad ekonómia, psychológia, verejné zdravotníctvo, medicína a pod. Jednou z takýchto aplikačných oblastí je aj oblasť bioštatistiky, ktorá aplikuje štatistické metódy na riešenie biologicky a medicínsky orientovaných výskumných problémov.



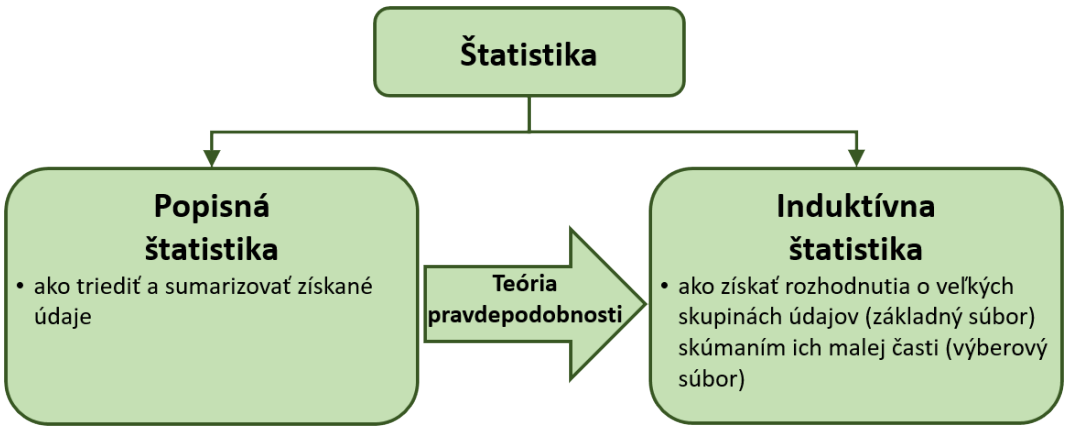
Obr. 2.1: Rozdelenie oboru štatistiky.

Bioštatistika tak predstavuje odbor s aplikáciami v mnohých oblastiach biológie vrátane laboratórnych biologických experimentov, epidemiológie, medicínskych vied, zdravotníckych vied, ostatných vied o živých organizmoch, základného biomedicínskeho výskumu, rizík spojených s výkonom povolania či dokonca príbuzných oblastí ekológie a ekonómie zdravia. Zaoberá sa návrhom štúdií, analýzou údajov, ale aj nových štatistických techník s cieľom riešenia problémov v oblastiach vied o živote.

2.2 Popisná a induktívna štatistika

Štatistiku podľa jej štatistických metód používaných na hodnotenie pozorovaných údajov rozdeľujeme do dvoch základných skupín (obrázok 2.2). Oblasť štatistiky zaoberajúca sa sumarizovaním, organizovaním a popisom údajov sa nazýva popisná štatistika, často z anglosaskej literatúry aj ako deskriptívna štatistika. Popisná štatistika teda pozostáva zo základných štatistických metód používaných na organizáciu a sumarizáciu informácií. Oblasť štatistiky zaobera-

júcu sa používaním údajov výberového súboru na odvodenie záverov o údajoch základného súboru označujeme ako induktívnu štatistiku, niekedy z anglosaskej literatúry aj ako inferenčnú štatistiku. Induktívna štatistika pozostáva z metód na vyvodzovanie a meranie spoľahlivosti záverov o základnom súbore na základe informácií získaných z výberového súboru.



Obr. 2.2: Rozdelenie metód štatistiky.

Popisná štatistika zahŕňa vytváranie tabuliek, grafov a diagramov, ako aj výpočty rôznych popisných charakteristík, medzi ktoré patria napríklad priemery, miery variability, percentily a pod. Vo všeobecnosti môžeme považovať popisnú štatistiku za základ ďalšieho štatistického uvažovania, ktorým začína analýza rôznych výskumných úloh a štúdií. V predmetoch venovaných úvodu do štatistiky či bioštatistiky sa snažíme primárne pochopiť a zvládnuť práve problematiku popisnej štatistiky.

Induktívna štatistika zahŕňa metódy ako sú napríklad bodový a intervalový odhad hodnôt, či testovanie štatistických hypotéz, ktoré sú všetky založené na teórii pravdepodobnosti. Samotná teória pravdepodobnosti potom tvorí akési prepojenie alebo premostenie medzi metódami popisnej a induktívnej štatistiky, tak ako je to znázornené na obrázku 2.2.

Popisná a induktívna štatistika sú teda navzájom prepojené a nemalo by sa k ich využívaniu pristupovať individuálne. Takmer vždy je potrebné použiť metódy popisnej štatistiky na organizáciu a sumarizáciu informácií získaných z výberového súboru a až následne aplikovať metódy induktívnej štatistiky na detailnejšiu a dôkladnejšiu analýzu skúmaného problému. Navyše, prvotná popisná analýza výberového súboru často odhaľuje charakteristiky, ktoré vedú k rozhodnutiu o výbere a následnom použití vhodnej induktívnej metódy.

2.3 Štatistické veličiny

Veličina predstavuje charakteristiku, resp. premennú, ktorá môže nadobúdať u rôznych subjektov alebo objektov rôzne hodnoty. Veličiny, nielen v štatistickom skúmaní, je možné rozdeliť do dvoch základných skupín, ktoré tvoria:

- **kvalitatívne veličiny** – premenné, ktoré predpokladajú nečíselné hodnoty, t. j. hodnoty, ktoré popisujeme slovne. Hodnoty kvalitatívnych veličín môžu vyplývať z atribútov označujúcich kategórie (napríklad krvná skupina, národnosť, úroveň vzdelania a pod.),
- **kvantitatívne veličiny** – premenné, ktoré predpokladajú číselné hodnoty, t. j. hodnoty, ktorými sú reálne čísla. Hodnoty kvantitatívnych veličín je možné získavať meraním (napríklad teplota, výška, hmotnosť a pod.), alebo spočítaním (napríklad počet pacientov, počet udalostí a pod.).

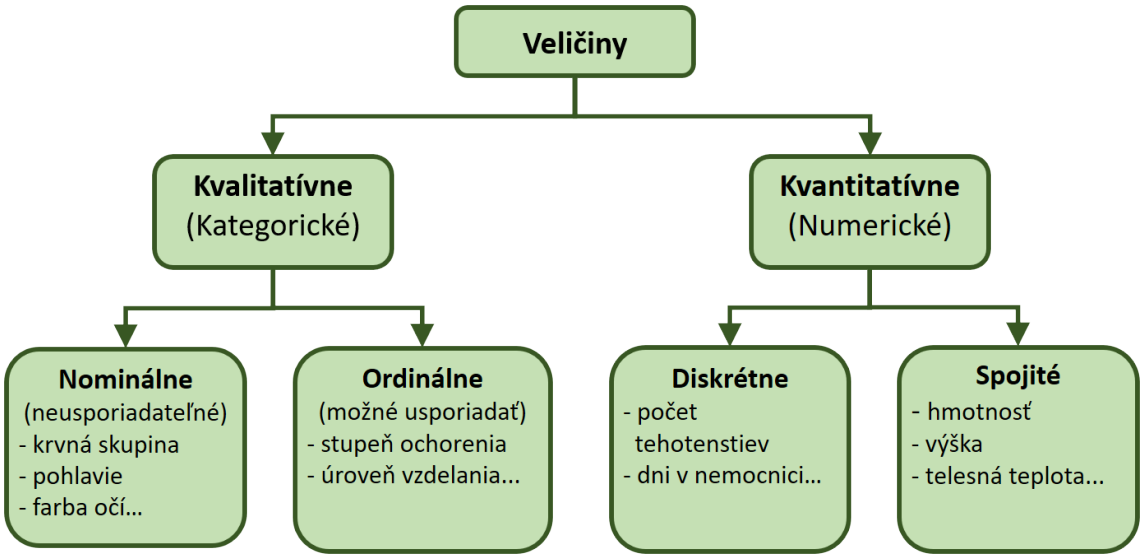
Kvalitatívne veličiny ďalej rozdeľujeme na:

- **nominálne veličiny** – kategórie sa vzájomne vylučujú a sú neusporadúvané, t. j. na poradí možných hodnôt nezáleží, keďže tieto sú len názvami (napríklad pohlavie, rodinný stav, národnosť, farba očí, krvná skupina a pod.). Nominálne veličiny tvoria najnižšiu úroveň škály merania umožňujúcu identifikáciu objektu alebo subjektu, pričom tu neexistuje významné poradie, rozdiel medzi hodnotami či významovo definovaná nula,
- **ordinálne veličiny** – kategórie sa vzájomne vylučujú a sú usporiadané, t. j. poradie hodnôt je dôležité (napríklad úroveň vzdelania, štádium rakoviny prsníka, stupeň daného ochorenia a pod.). Hodnoty ordinálnych veličín je možné porovnávať a určiť, ktoré sú väčšie, menšie alebo rovnaké, teda aj definovať ich poradie.

Kvantitatívne veličine je možné ďalej rozdeliť na:

- **diskrétne veličiny** – premenné, ktoré predpokladajú konečný alebo spočítateľný počet možných hodnôt, zvyčajne získané počítaním (napríklad počet tehotenstiev, počet infarktov myokardu, počet detí v rodine, počet obéznych pacientov, počet dní hospitalizácie a pod.),
- **spojité veličiny** – premenné, ktoré predpokladajú teoreticky nekonečný počet možných hodnôt, obvykle získaných meraním (napríklad hladina cholesterolu, hladina cukru v krvi, hmotnosť, výška, teplota a pod.).

Rozdelenie štatistických veličín je zobrazené na obrázku 2.3.



Obr. 2.3: Delenie štatistických veličín.

Kvantitatívne veličiny, tak diskrétne, ako aj spojité, je možné definovať na intervalovej alebo pomerovej škále. V prípade, že je možné zmysluplne porovnať rozdiely medzi hodnotami (meraniami) veličiny, t. j. určiť o koľko je jedna hodnota väčšia alebo menšia ako iná nameraná hodnota, potom takáto kvantitatívna veličina je označovaná ako **intervalová veličina**. Intervalové veličiny majú definovanú jednotku merania, avšak nemajú prirodzenú absolútnu nulu. Ak je možné porovnávať rozdiely medzi meraniami premennej a zároveň aj pomery meraní, potom je kvantitatívna veličina označovaná ako **pomerová veličina**. Aby bol pomer meraní zmysluplný, musí mať pomerová veličina zmysluplný absolútny nulový bod (absolútny začiatok stupnice, pod ktorý hodnota veličiny nemôže klesnúť), t. j. pomerová škála je intervalovou škálou s absolútnym nulovým bodom (skutočnou absenciou meranej veličiny). Napríklad, ak je prvý pacient vysoký 200 cm a druhý 160 cm (v tomto prípade nemôže byť výška pacienta záporná), potom vieme určiť, že prvý pacient je 1,25 krát vyšší ako ten druhý.

Kapitola 3

Charakteristiky popisnej štatistiky

Súbory získavaných a nespracovaných údajov vo všeobecnosti pozostávajú z veľkého počtu meraní alebo pozorovaní, ktoré sú často veľmi početné, a teda aj komplikované na to, aby sa jednoduchým vizuálnym porovnaním dali pochopiť. Úlohou popisnej štatistiky je preto z týchto súborov údajov, ktoré sú inak neprehľadné a spravidla individuálne pre nezainteresované osoby nič nehovoriace, vytvoriť organizované prehľady a súhrny. Popisná štatistika teda poskytuje spôsoby, ako kondenzovať a organizovať získané informácie do súboru popisných charakteristík a vizuálnych výstupov, tak aby bolo možné pochopiť a interpretovať význam komplexných údajov.

Údaje sú sumarizované spravidla pomocou jednotlivých hodnôt, označovaných ako popisné charakteristiky. Hodnoty popisných charakteristík je možné vypočítavať z:

- údajov výberového súboru (vzorky)
- údajov základného súboru (populácie)

Na základe zdrojového súboru hodnôt, potom popisné charakteristiky vypočítané z údajov výberového súboru označujeme aj ako štatistiky a popisné charakteristiky získané z údajov základného súboru označujeme aj ako parametre, t. j. sú sumarizované numerickými parametrami. Hodnoty základného súboru o veľkosti N budeme označovať veľkými písmenami ako $X_1, X_2, \ldots, X_N$. Hodnoty výberového súboru o veľkosti n budeme označovať malými písmenami ako $x_1, x_2, \ldots, x_n$ a sú známe, keďže sú získavané pozorovaním alebo meraním v rámci riešenej výskumnej úlohy.

Parametre základného súboru spravidla nie sú známe a na ich odvodenie sú používané štatistiky výberového súboru. To znamená, že štatistika popisuje charakteristiku výberového súboru, ktorú je možné následne použiť na odvodenie záverov o neznámych parametroch. Primárnym cieľom väčšiny výskumne

orientovaných štúdií sú parametre základného súboru a nie štatistiky vypočítané z náhodného výberového súboru. Výberový súbor a štatistiky, ktoré ho popisujú, sú dôležité len pokiaľ poskytujú informácie o neznámych parametroch základného súboru.

Aby bolo možné špecifikovať informácie o kvantitatívnej alebo kvalitatívnej charakteristike štatistického súboru hodnôt je potrebné poznať odpovede na nasledovné základné otázky:

1. Kde sa nachádza?
2. Ako veľmi sa mení?
3. Aké je zastúpenie jednotlivých údajov?

Význam popisnej štatistiky je týmto zrejmý a takmer samovysvetľujúci. Základné štatistické charakteristiky uvádzané v popisnej štatistike je preto možné rozdeliť do troch hlavných skupín, ktorými sú:

- **Charakteristiky polohy**, niekedy označované aj ako charakteristiky centrálnej tendencie, ktoré charakterizujú úroveň hodnôt štatistického súboru, t. j. určujú polohu, kde sa daná charakteristika v súbore hodnôt nachádza. Patria sem rôzne priemery, ale aj medián a modus.
- **Charakteristiky variability**, niekedy tiež uvádzané ako charakteristiky rozptýlenosti, ktoré charakterizujú mieru rozptýlenia hodnôt štatistického súboru, t. j. informujú nás o rozptýlení jednotlivých hodnôt okolo ich stredných hodnôt. Patria sem napríklad variačné rozpätie, kvartilové rozpätie, kvantilové rozpätie, priemerná odchýlka, rozptyl, smerodajná odchýlka či variačný koeficient.
- **Charakteristiky tvaru**, ktoré kvantitatívne charakterizujú tvar rozdelenia hodnôt štatistického súboru. Patria sem koeficienty šikmosti a špicatosti.

Pri bežnom spracovaní súborov údajov sa často uvádzajú len charakteristiky z prvých dvoch skupín, avšak z pohľadu komplexného popisu údajov i získania informácií o celkovom rozložení údajov v súbore hodnôt a následnom výbere vhodných metód testovania štatistických hypotéz by sme nemali zabúdať ani na charakteristiky z tretej skupiny.

Charakteristiky popisnej štatistiky a ich hodnoty sú užitočné najmä v situáciách, kedy chceme charakterizovať skupinu údajov pomocou jednej hodnoty, resp. jedným číslom. Takáto hodnota býva veľmi užitočná pri prezentácii výsledkov výskumu, ale taktiež pri ich porovnávaní (napríklad údaje pacientov

z rôznych regiónov, údaje získané v rôznych obdobiach, údaje z rôznych výskumov a pod.). Jediná hodnota pritom umožňuje veľmi pohodlne reprezentovať celý súbor údajov tak malej, strednej, ako aj veľkej výberovej skupiny. Mohlo by sa zdať, že popisná štatistika ponúka ideálne možnosti ako reprezentovať dôležité vlastnosti súboru údajov iba pomocou jednej hodnoty alebo jedinej charakteristiky. Opak je však pravdou, a v skutočnosti vždy potrebujeme použiť viac hodnôt alebo charakteristík, aby sme poskytli čo najlepší prehľad o získanom súbore údajov. Aj preto v sumárnom hodnotení prezentujeme kombináciu charakteristík popisnej štatistiky vyššie uvedených skupín, ktoré sú používané na popis rôznych vlastností obsiahnutých v súbore údajov. Charakteristiky polohy poskytujú napríklad centrálnu hodnotu, okolo ktorej sa nachádzajú všetky ostatné hodnoty v súbore údajov. Charakteristiky variability zasa kvantifikujú rozdiely medzi hodnotami a tiež popisujú ako sú údaje rozptýlené voči centrálnej hodnote alebo aký rozsah variácií je medzi hodnotami údajov. Charakteristiky tvaru sú používané na zistenie toho, či existuje symetria hodnôt od centrálnej hodnoty alebo aj toho, ako sú hodnoty koncentrované.

3.1 Charakteristiky polohy

Hodnoty skúmanej veličiny sú často koncentrované v okolí stredu údajov. Aj preto sa charakteristiky polohy niekedy označujú ako charakteristiky centrálnej tendencie. Medzi typické charakteristiky polohy patria priemerné hodnoty, modus a medián.

3.1.1 Aritmetický priemer

Aritmetický priemer je najčastejšie prezentovaným priemerom a charakteristikou polohy, pričom pri jeho prezentácii sa zvyčajne používa len pojem priemer alebo priemerná hodnota. Aritmetický priemer vyjadruje, aký objem hodnôt skúmanej štatistickej veličiny pripadá v priemere na jednu štatistickú jednotku súboru.

Aritmetický priemer základného súboru (populačný priemer) označujeme ako μ a vypočítame ho ako sumu všetkých hodnôt základného súboru $X_1, X_2, \ldots, X_N$, ktorú vydelíme rozsahom základného súboru N:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + X_2 + \cdots + X_N}{N} \tag{3.1}$$

pričom $N \to \infty$. Aritmetický priemer základného súboru μ je parameter, ktorý zvyčajne nie je známy, ale jeho hodnotu sa v štatistických analýzach, podobne ako aj ďalšie parametre základného súboru snažíme odhadnúť.

Aritmetický priemer výberového súboru (výberový priemer) $\bar{x}$ vypočítame zo známych hodnôt výberového súboru $x_1, x_2, \ldots, x_n$ s rozsahom n podľa vzťahu (3.2).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.2}$$

Výpočet aritmetického priemeru môžeme ilustrovať na nasledovnom príklade náhodného výberu.

Príklad 3.1. Pri meraní hmotnosti stredoškolákov bola náhodne vybraná skupina 30 žiakov z rôznych stredných škôl. Ich hmotnosti v kilogramoch boli zaznamenané do tabuľky 3.1.

Tabuľka 3.1: Hmotnosti náhodne vybraných stredoškolákov.

Žiak č. i	**Hmotnosť (kg)**	**Žiak č. i**	**Hmotnosť (kg)**	**Žiak č. i**	**Hmotnosť (kg)**
1	39	11	49	21	61
2	66	12	60	22	42
3	61	13	57	23	53
4	63	14	52	24	57
5	64	15	65	25	63
6	38	16	57	26	60
7	59	17	48	27	51
8	57	18	59	28	62
9	57	19	46	29	61
10	50	20	60	30	63

Keďže sa jedná o výberový súbor, stačí hodnoty hmotností stredoškolákov dosadiť do vzťahu (3.2). Potom bude aritmetický priemer výberového súboru:

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{39 + 66 + 61 + 63 + 64 + \cdots + 63}{30} = \frac{1680}{30} = 56$$

Naším záverom teda bude, že priemerná hmotnosť 30 náhodne vybraných stredoškolákov je 56 kg.

Funkcie v MS Excel

AVERAGE(číslo1; číslo2, ...)
Funkcia AVERAGE vráti priemernú hodnotu (aritmetický priemer) jej argumentov. Ako argumenty funkcie (číslo1; číslo2; ...) je možné použiť priamo hodnoty, z ktorých chceme aritmetický priemer vypočítať alebo odkazy na bunky (adresy buniek) tabuľky MS Excel, v ktorých sú hodnoty uložené.

Základné vlastnosti aritmetického priemeru sú:

- jedinečnosť – pre daný súbor hodnôt existuje len jeden aritmetický priemer,
- jednoduchosť – aritmetický priemer je zrozumiteľný a je ho možné ľahko vypočítať,
- aritmetický priemer je ovplyvnený každou hodnotou – každá hodnota v súbore hodnôt vstupuje do jeho výpočtu. To znamená, že ak súbor hodnôt obsahuje aj extrémne hodnoty, tieto významne ovplyvňujú výslednú hodnotu aritmetického priemeru (vplyv neobvyklých/odľahlých hodnôt),
- aritmetický priemer nie je možné použiť pre kvalitatívne údaje.

Vplyv odľahlých hodnôt na hodnotu aritmetického priemeru je možné pochopiť na nasledovnom príklade.

Príklad 3.2. V prieskume zameranom na kvalitu poskytovanej zdravotnej starostlivosti boli okrem iného zaznamenané aj informácie o dobe čakania na CT vyšetrenie. Päť respondentov, ktorí boli zahrnutí do prieskumu a navštívili nemocnicu A uviedlo dobu čakania 3, 1, 90, 4 a 2 dni.

Priemernú dobu čakania vieme vypočítať opäť pomocou vzťahu (3.2):

$$\bar{x} = \frac{3 + 1 + 90 + 4 + 2}{5} = \frac{100}{5} = 20$$

Hodnota výsledného aritmetického priemeru nás informuje, že priemerná doba čakania na CT vyšetrenie v nemocnici A je 20 dní. Avšak, už na prvý pohľad vidíme, že táto hodnota nekorešponduje so zastúpenými údajmi a skreslený pohľad na dobu čakania spôsobila jediná odľahlá hodnota. Aj kvôli tejto vlastnosti aritmetického priemeru je potrebné interpretovať jeho hodnoty s ohľadom na spôsob, akým boli údaje získané (viď kapitola 1), ako aj v spojení s ďalšími charakteristikami popisnej štatistiky.

Často sa taktiež stretávame s prezentáciou výsledkov rôznych štúdií, ktoré porovnávajú dve alebo aj viaceré skupiny údajov len na základe priemerných hodnôt vypočítaných z hodnôt sledovaných veličín zozbieraných od zahrnutých

štatistických jednotiek. Možné riziko takéhoto zjednodušeného využitia štatistiky si vysvetlíme na rozšírení predošlého príkladu.

Príklad 3.3. V prieskume zameranom na kvalitu poskytovanej zdravotnej starostlivosti uvedenom v príklade 3.2 boli zaznamenané aj informácie o dobe čakania na CT vyšetrenie u piatich respondentov, ktorí navštívili nemocnicu B a uviedli dobu čakania 21, 18, 22, 20 a 19 dní.

Priemernú dobu čakania na CT vyšetrenie v nemocnici B vypočítame pomocou vzťahu (3.2):

$$\bar{x} = \frac{21 + 18 + 22 + 20 + 19}{5} = \frac{100}{5} = 20$$

Priemerná doba čakania na CT vyšetrenie v nemocnici B je 20 dní a je teda rovnaká ako v nemocnici A. Mohlo by sa zdať, že na základe tejto informácie budú pacienti na CT vyšetrenie čakať rovnako. Avšak, kým v nemocnici B sú všetky zistené doby čakania blízko hodnoty 20 dní, u pacientov z nemocnice A tomu tak nie je. Ako už bolo uvedené vyššie, je potrebné overiť príčinu vzniku odľahlej hodnoty 90, ktorá mohla vzniknúť tak z objektívnych (množstvo požiadaviek, obsadené termíny a pod.), ako aj subjektívnych (nedostavenie sa na stanovený termín, omyl pri evidencii záznamu a pod.) dôvodov. Bez tejto odľahlej hodnoty by bola priemerná doba čakania oveľa menšia, a teda pre pacientov výhodnejšia. K rozhodnutiu o výbere nemocnice však nemusíme dospieť len odstraňovaním odľahlej hodnoty, ale aj pomocou ďalších charakteristík popisnej štatistiky.

Funkcie v MS Excel

SUM(číslo1; číslo2; ...)
Funkcia SUM spočíta všetky čísla v rozsahu buniek.

COUNT(hodnota1; hodnota2; ...)
Funkcia COUNT spočíta počet buniek v rozsahu, ktorý obsahuje čísla.

SUBTOTAL(číslo_funkcie; odkaz1; odkaz2; ...)
Funkcia SUBTOTAL vracia medzisúčet v zozname alebo databáze. Ak je prvý argument (číslo_funkcie) rovný 1, potom funkcia SUBTOTAL vráti aritmetický priemer hodnôt uložených v bunkách, ktorých odkazy tvoria ďalšie argumenty funkcie. Ak je prvý argument rovný 2, potom funkcia vráti počet buniek v rozsahu, ktorý obsahuje čísla. Ak je prvý argument rovný 9, potom funkcia vráti sumu hodnôt v rozsahu buniek, ktorý obsahuje čísla. Použitie funkcie SUBTOTAL má svoje uplatnenie pri filtrovaní záznamov tabuľky, kedy na rozdiel od funkcií AVERAGE, COUNT či SUM, vráti výsledok len z hodnôt uvedených v zobrazených bunkách, a teda do výpočtu sa nezahŕňajú hodnoty v skrytých bunkách.

Aritmetický priemer nie je jedinou priemernou hodnotou popisnej štatistiky, keďže existujú mnohé ďalšie typy priemerných hodnôt. Niekoľko z nich bude uvedených v nasledujúcich kapitolách.

3.1.2 Geometrický priemer

Geometrický priemer je používaný u veličín, ktorých hodnoty narastajú geometricky a vyjadruje priemernú veľkosť zmeny (napríklad pri časových radoch v ekonómii, biológii a pod.). Je tiež vhodný, ak súbor pozorovaní obsahuje extrémne hodnoty, ktoré sú výrazne väčšie ako väčšina ostatných hodnôt v súbore. Avšak je potrebné si uvedomiť, že je ho možné použiť len pre kladné čísla.

Geometrický priemer $\bar{x_g}$ výberového súboru n pozorovaní nezáporných hodnôt je definovaný ako n-tá odmocnina súčinu hodnôt všetkých pozorovaní:

$$\bar{x_g} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1.x_2.\cdots.x_n} \tag{3.3}$$

Geometrický priemer hodnôt doby čakania pacientov na CT vyšetrenie v nemocnici A z príkladu 3.2 vypočítame podľa vzťahu (3.3) takto:

$$\bar{x_g} = \sqrt[5]{3.1.90.4.2} = \sqrt[5]{2160} = 4,643984 \doteq 4,64$$

Priemerná doba čakania pacientov na CT vyšetrenie v nemocnici A, vypočítaná ako geometrický priemer hodnôt doby čakania pacientov na CT vyšetrenie je tak 4,64 dňa. V porovnaní s aritmetickým priemerom, ktorý je 20 dní sme bližšie k hodnotám, ktoré sú viac zastúpené v tomto malom výberovom súbore a bol minimalizovaný vplyv extrémnej hodnoty 90. Na druhej strane, geometrický priemer doby čakania pacientov na CT vyšetrenie v nemocnici B z príkladu 3.3 je 19,95 dní, čo je blízko aritmetického priemeru, keďže v tomto výberovom súbore nebola prítomná extrémna hodnota.

Príklad 3.4. Predpokladajme, že v rámci vstupných vyšetrení pacientov na internom oddelení boli zaznamenávané hodnoty krvného tlaku. U desiatich pacientov boli zaznamenané pozorovania systolického krvného tlaku v mmHg tak, ako sú uvedené v tabuľke 3.2.

Aritmetický priemer hodnôt systolického krvného tlaku je 125 mmHg, čo je viac ako deväť z desiatich pozorovaných hodnôt, keďže hodnota 180 mmHg posúva priemer doprava. Geometrický priemer je 123,9 mmHg a je menej ovplyvnený touto extrémnou hodnotou. Ak by sa hodnota 180 mmHg zmenila na ešte

Tabuľka 3.2: Hodnoty systolického krvného tlaku.

Pacient č. i	Systolický KT (mmHg)	Pacient č. i	Systolický KT (mmHg)
1	122	6	120
2	116	7	118
3	120	8	114
4	180	9	120
5	118	10	122

extrémnejšiu hodnotu, potom by bol aritmetický priemer ovplyvnený ešte viac, ale geometrický priemer o niečo menej. Geometrický priemer je zvyčajne blízko mediánu (viď kapitola 3.1.4).

Základné vlastnosti geometrického priemeru sú:

- jedinečnosť – pre daný súbor hodnôt existuje len jeden geometrický priemer,
- geometrický priemer je použiteľný na kladné kvantitatívne hodnoty,
- vplyv extrémnych hodnôt na výslednú hodnotu geometrického priemeru je minimálny i napriek tomu, že do jeho výpočtu sú zahrnuté všetky hodnoty,
- vynásobením všetkých hodnôt pôvodného súboru konštantným číslom sa rovnakým číslom násobí aj geometrický priemer.

Funkcie v MS Excel

GEOMEAN(číslo1; číslo2; ...)
Funkcia GEOMEAN vráti geometrický priemer poľa alebo rozsahu kladných číselných údajov.

PRODUCT(hodnota1; hodnota2; ...)
Funkcia PRODUCT vynásobí všetky čísla zadané ako jej argumenty, t. j. vráti súčin všetkých číselných hodnôt v rozsahu buniek.

SUBTOTAL(číslo_funkcie; odkaz1; odkaz2; ...)
Funkcia SUBTOTAL vracia medzisúčet v zozname alebo databáze. Ak je prvý argument (číslo_funkcie) rovný 6, potom funkcia SUBTOTAL vráti súčin všetkých číselných hodnôt uložených v bunkách, ktorých odkazy tvoria ďalšie argumenty funkcie, pričom do výpočtu sa nezahŕňajú hodnoty v skrytých bunkách.

3.1.3 Harmonický priemer

Harmonický priemer, podobne ako geometrický priemer využívame v prípade prítomnosti hodnôt v súbore, ktoré sa odlišujú od väčšiny pozorovaných hodnôt. Kým geometrický priemer je používaný na obmedzenie vplyvu extrémnych

hodnôt, harmonický priemer je používaný na riešenie vplyvu odľahlých hodnôt. Harmonický priemer sa zvyčajne používa na výpočet priemerných rýchlostí alebo iných mier či pomerov (napríklad výpočet priemernej rýchlosti ak sú vzdialenosti konštantné a čas premenlivý, priemernej rýchlosti ak sú úsekové priemerné rýchlosti rôzne a pod.). Hodnoty výberového súboru však nesmú obsahovať nulové hodnoty.

Harmonický priemer $\bar{x_h}$ výberového súboru n pozorovaní je definovaný ako podiel počtu hodnôt všetkých pozorovaní a sumy ich prevrátených hodnôt. Je to teda prevrátenou hodnotou aritmetického priemeru prevrátených hodnôt:

$$\bar{x_h} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \tag{3.4}$$

Harmonický priemer hodnôt doby čakania pacientov na CT vyšetrenie v nemocnici A z príkladu 3.2 vypočítame podľa vzťahu (3.4) takto:

$$\bar{x_h} = \frac{5}{\frac{1}{3} + \frac{1}{1} + \frac{1}{90} + \frac{1}{4} + \frac{1}{2}} = 2,387268 \doteq 2,39$$

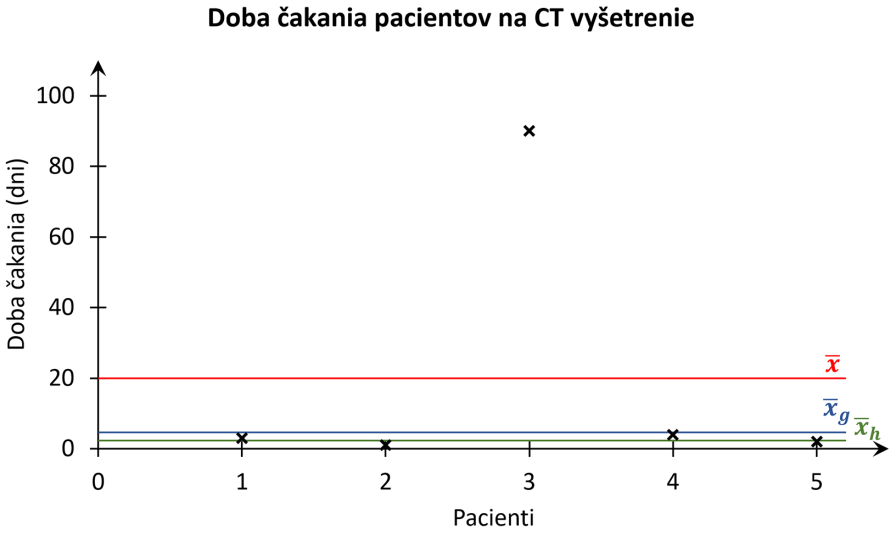
Priemerná doba čakania pacientov na CT vyšetrenie v nemocnici A, vypočítaná ako harmonický priemer hodnôt doby čakania pacientov na CT vyšetrenie je 2,39 dňa. V porovnaní s aritmetickým priemerom je harmonický priemer umiestnený uprostred najväčšieho výskytu hodnôt a odľahlá hodnota 90 má na harmonický priemer minimálny vplyv. Ak vypočítame harmonický priemer doby čakania pacientov na CT vyšetrenie v nemocnici B z príkladu 3.3, získame hodnotu 19,89 dní. Harmonický priemer je menší ako geometrický a aritmetický priemer. Grafické znázornenie hodnôt doby čakania pacientov na CT vyšetrenie v nemocnici A a úroveň jednotlivých priemerov sú uvedené na obrázku 3.1.

Základné vlastnosti harmonického priemeru sú:

- jedinečnosť – pre daný súbor hodnôt existuje len jeden harmonický priemer,
- harmonický priemer je použiteľný na kladné kvantitatívne hodnoty,
- vplyv extrémnych hodnôt na výslednú hodnotu harmonického priemeru je minimálny i napriek tomu, že do jeho výpočtu sú zahrnuté všetky hodnoty.

Funkcie v MS Excel

HARMEAN(číslo1; číslo2; ...)
Funkcia HARMEAN vráti harmonický priemer množiny kladných číselných údajov.



Obr. 3.1: Doba čakania pacientov na CT vyšetrenie v nemocnici A a úroveň aritmetického (červená), geometrického (modrá) a harmonického (zelená) priemeru.

3.1.4 Medián

Okrem rôznych priemerov, nielen tých, ktoré už boli uvedené v predošlých kapitolách, existujú aj iné charakteristiky na definovanie alebo výpočet stredu množiny hodnôt. Veľmi často používanou charakteristikou polohy je medián.

Medián $\tilde{x}$ je hodnota rozdeľujúca usporiadaný súbor hodnôt s rozsahom n (hodnoty musia byť usporiadané spravidla do neklesajúcej postupnosti) na dve rovnaké časti (napríklad menšie a väčšie čísla). Čísla v prvej časti sú menšie nanajvýš rovnaké ako medián a čísla v druhej časti hodnôt sú väčšie alebo rovnaké ako medián.

Medián je teda prostredná hodnota súboru usporiadaných hodnôt. V prípade, ak je počet pozorovaných hodnôt v súbore nepárny, určíme medián podľa vzťahu (3.5):

$$\tilde{x} = \frac{(n+1)}{2}ta\ hodnota \tag{3.5}$$

Medián pre dobu čakania pacientov na CT vyšetrenie v nemocnici A z príkladu 3.2 potom určíme ako tretiu hodnotu usporiadaných dôb čakania, keďže súbor obsahuje 5 hodnôt a prostrednou je práve tretia hodnota. Vzostupne usporiadané hodnoty sú: 1, 2, 3, 4 a 90. Mediánom doby čakania pacientov na CT vyšetrenie v nemocnici A sú 3 dni.

Podobne pre hodnoty z príkladu 3.3 (18, 19, 20, 21 a 22 dní) identifikujeme, že medián doby čakania pacientov na CT vyšetrenie v nemocnici B je rovný 20 dňom.

Ak je počet hodnôt v súbore párny, potom je mediánom priemer (aritmetický) dvoch prostredných hodnôt a určíme ho podľa vzťahu (3.6):

$$\tilde{x} = \frac{\frac{n}{2}ta\ hodnota + \frac{n+2}{2}ta\ hodnota}{2} \tag{3.6}$$

Ak sa pozrieme na príklad 3.1, z rozsahu súboru hodnôt hmotností (párny počet) je zrejmé, že uprostred sa nachádzajú dve hodnoty, t. j. 15-ta a 16-ta. Usporiadaná neklesajúca postupnosť hmotností stredoškolákov je: 38, 39, 42, 46, 48, 49, 50, 51, 52, 53, 57, 57, 57, 57, 57, 59, 59, 60, 60, 60, 61, 61, 61, 62, 63, 63, 63, 64, 65 a 66. 15-ta hodnota je rovná 57 kg a 16-ta hodnota je 59 kg. Medián vypočítame podľa vzťahu (3.6) takto:

$$\tilde{x} = \frac{57 + 59}{2} = 58$$

Potom medián hmotnosti 30 náhodne vybraných stredoškolákov sa rovná 58 kg. V tomto prípade sa ani hodnota mediánu (58 kg), ani hodnota aritmetického priemeru (56 kg) v hodnotách výberového súboru nenachádza. To poukazuje na ďalšiu vlastnosť týchto štatistických charakteristík, a síce, že ich hodnoty nemusia byť zastúpené v skupine pozorovaní.

Základné vlastnosti mediánu sú:

- jedinečnosť – pre daný súbor hodnôt existuje len jeden medián,
- jednoduchosť – medián je možné ľahko určiť (vypočítať),
- medián nie je ovplyvnený každou hodnotou – len jednou, resp. dvoma, nachádzajúcimi sa v danom súbore pozorovaní. To znamená, že extrémne hodnoty neovplyvňujú výslednú hodnotu mediánu (žiaden vplyv neobvyklých/odľahlých hodnôt),
- medián je možné vo všeobecnosti nájsť len pre kvantitatívne veličiny. V niektorých prípadoch však možno medián nájsť aj pre ordinálne kvalitatívne premenné.

Funkcie v MS Excel

MEDIAN(číslo1; číslo2; ...)
Funkcia MEDIAN vráti hodnotu v strede množiny daných číselných údajov.

3.1.5 Modus

Modus $\hat{x}$ je najčastejšie sa vyskytujúca hodnota v súbore n pozorovaní. Ak sú všetky hodnoty rôzne, alebo ak sú početnosti všetkých zastúpených hodnôt rovnaké, potom súbor pozorovaní nemá žiaden modus. Súbor pozorovaní môže obsahovať aj viac ako jeden modus.

Modus je zvyčajne vypočítavaný pre veľké súbory diskrétnych údajov, avšak je ho možné určovať pre všetky typy veličín. Vo svojej podstate nám poskytuje informáciu o hodnote, ktorá má vyššiu pravdepodobnosť výskytu pri pozorovaní základného súboru. Je to teda reprezentatívna charakteristika použiteľná pri predpovedaní hodnôt. Zjednodušene je ju možné určiť ako:

$$\hat{x} = x_{f_{max}} \tag{3.7}$$

kde, $x_{f_{max}}$ predstavuje hodnotu s najväčšou početnosťou, pre ktorú platí, že jej početnosť $f_{max} \geq 2$ a zároveň platí, že aspoň jedna početnosť ostatných hodnôt v súbore je menšia, alebo aspoň jedna pozorovaná hodnota je jedinečná, t. j. vyskytuje sa len raz.

Modus je teda charakteristikou, ktorá súvisí s vrcholom (vrcholmi) rozdelenia početností. Ak rozdelenie početností obsahuje iba jeden vrchol, potom je rozdelenie hodnôt unimodálne, ak má rozdelenie početností dva vrcholy, potom je rozdelenie hodnôt bimodálne atď.

Pre hodnoty z príkladu 3.1 zistíme, že najčastejšie vyskytujúca sa hodnota hmotnosti stredoškolákov je hodnota 57 kg, ktorá sa vo výberovom súbore vyskytuje 5 krát. Všetky ostatné hodnoty sa vyskytujú menej často. Teda modusom hmotnosti 30 náhodne vybraných stredoškolákov je 57 kg.

Ak sa pozrieme na príklad 3.1 komplexne, zistíme, že hmotnosť výberového súboru stredoškolákov vykazuje aritmetický priemer 56 kg, modus 57 kg a medián 58 kg. V niektorých prípadoch sú tieto hodnoty približne rovnaké, v iných, ako je aj tento, sú medzi nimi väčšie či menšie rozdiely. Tieto informácie budú predmetom ďalších analýz, napríklad v skúmaní normality údajov i rozhodovaní o výbere vhodnej metódy testovania štatistických hypotéz. Ak majú hodnoty súboru pozorovaní normálne rozdelenie (sú symetricky rozložené), potom aritmetický priemer, medián a modus sú rovnaké.

V príklade 3.4 nájdeme taktiež jeden modus, ktorým je hodnota systolického krvného tlaku 120 mmHg nachádzajúca sa v súbore 3 krát. V tomto prípade je hodnota 120 mmHg aj mediánom systolického krvného tlaku, t. j. medián aj modus sú rovnaké a ležia naľavo od aritmetického priemeru, ktorý je rovný hodnote 125 mmHg.

Základné vlastnosti modusu sú:

- jednoduchosť – modus je zrozumiteľný a je ho možné ľahko vypočítať,
- modus nie je ovplyvnený každou hodnotou – extrémne hodnoty neovplyvňujú jeho výslednú hodnotu (žiaden vplyv extrémnych/odľahlých hodnôt),
- modus je možné určiť pre kvantitatívne aj kvalitatívne veličiny.

Funkcie v MS Excel

MODE.SNGL(číslo1; číslo2; ...)
Funkcia MODE.SNGL vráti hodnotu, ktorá sa v poli alebo rozsahu údajov vyskytuje alebo opakuje najčastejšie.

MODE.MULT(číslo1; číslo2; ...)
Funkcia MODE.MULT vráti zvislé pole najčastejšie sa vyskytujúcich alebo opakujúcich sa hodnôt v poli alebo rozsahu údajov. MODE.MULT použijeme v prípadoch, ak sa v skupine údajov môže vyskytovať viac ako jeden modus.

Aj keď je najčastejšie prezentovanou charakteristikou polohy aritmetický priemer, ktorý je vypočítavaný zo všetkých hodnôt súboru, poukázali sme aj na prípady kedy je jeho použitie nevhodné, resp. minimálne poskytuje skreslené závery. V prípade, ak súbor hodnôt obsahuje údaje, ktorých rozdelenie je prezentované viacerými vrcholmi (napríklad bimodálne), alebo ak spracovávame kvalitatívne údaje, potom preferujeme použitie modusu. Ak je rozdelenie údajov asymetrické, potom je vhodnejšie použiť medián. Štatistické súbory sa však líšia nielen v úrovni hodnôt sledovaných veličín, ale aj ich variabilitou, ktorá charakterizuje mieru rozptýlenia hodnôt okolo charakteristík polohy.

3.2 Charakteristiky variability

Namerané alebo pozorované hodnoty získavané v rámci rôznych výskumných úloh sú zvyčajne rozptýlené v intervaloch očakávaných, prípadne sledovanej veličine zodpovedajúcich hodnôt. Určenie rozptýlenia hodnôt si vyžaduje výpočet charakteristík, ktoré túto variáciu popisujú. Medzi najdôležitejšie popisné charakteristiky variability údajov patria rozptyl a smerodajná odchýlka. Okrem týchto dvoch charakteristík však využívame aj ďalšie, ktoré nielen rozširujú popis údajov, ale sú dôležité aj z pohľadu ich analýzy a interpretácie výsledkov. Vo všeobecnosti je možné charakteristiky variability rozdeliť na charakteristiky, ktoré sú ovplyvnené:

- len niektorými hodnotami v súbore – na ich určenie/výpočet sú použité len niektoré hodnoty, napríklad variačné rozpätie, kvantilové rozpätie, kvartilové rozpätie, kvartilová odchýlka a pod.,

- všetkými hodnotami v súbore – na ich určenie/výpočet sú použité všetky hodnoty, napríklad priemerná odchýlka, rozptyl, smerodajná odchýlka, variačný koeficient a pod.

3.2.1 Variačné rozpätie

Variačné rozpätie R je najbežnejšou a najjednoduchšou charakteristikou popisu rozloženia údajov. Je určená ako rozdiel medzi najväčšou a najmenšou získanou hodnotou v súbore pozorovaní:

$$R = x_{max} - x_{min} \tag{3.8}$$

kde x_{max} reprezentuje najväčšiu (maximum) a x_{min} najmenšiu (minimum) hodnotu.

Veľká hodnota variačného rozpätia znamená, že najväčšia a najmenšia hodnota v súbore pozorovaní sa výrazne líšia. Variačné rozpätie je však veľmi výrazne ovplyvňované jednou či dvoma extrémnymi hodnotami (na oboch stranách rozloženia hodnôt). Je preto iba približnou charakteristikou variability hodnôt sledovanej veličiny a neposkytuje žiadnu informáciu o tom, ako veľmi sa hodnoty v súbore pozorovaní líšia.

Najmenšou hodnotou hmotnosti z príkladu 3.1 bola hodnota 38 kg a najväčšou hodnota 66 kg. Rozdiel medzi týmito dvoma hodnotami určuje variačné rozpätie hmotností stredoškolákov a hovorí o tom, že najťažší stredoškolák je o 28 kg ťažší ako najľahší stredoškolák vo výberovom súbore stredoškolákov.

Podobne v prípade hodnôt systolického krvného tlaku z príkladu 3.4 môžeme určiť, že variačné rozpätie hodnôt systolického krvného tlaku je 66 mmHg, ako výsledok rozdielu najväčšej hodnoty 180 mmHg a tej najmenšej v pozorovanom súbore. Tu sa však už prejavil vplyv extrémnej hodnoty, keďže všetky ostatné hodnoty sa nachádzajú v okolí mediánu 120 mmHg, a ktorých variačné rozpätie by bez tejto extrémnej hodnoty bolo 8 mmHg.

Variačné rozpätie R je vhodné skôr pre malé súbory nameraných hodnôt, keďže poskytuje len obmedzené informácie o rozptýlení hodnôt. Využíva sa napríklad pri kontrole kvality. Kvocient medzi najväčšou a najmenšou hodnotou výberového súboru je koeficient variačného rozpätia K:

$$K = \frac{x_{max}}{x_{min}} \tag{3.9}$$

Variačné rozpätie R je možné taktiež dať do súvislosti s priemernou hodnotou (aritmetický priemer) výberového súboru $\bar{x}$ a určiť relatívne variačné rozpätie:

$$R_{rel} = \frac{R}{\bar{x}} \tag{3.10}$$

Na užitočnosť variačného rozpätia v spojení s informáciou o priemernej hodnote môžeme poukázať na údajoch z príkladov 3.2 a 3.3. U oboch bol aritmetický priemer rovný hodnote 20 dní, avšak v prípade doby čakania na CT vyšetrenie v nemocnici A je variačné rozpätie 89 dní, koeficient variačného rozpätia 90 a relatívne variačné rozpätie 4,45. V nemocnici B sú to hodnoty variačného rozpätia 4 dni, koeficientu variačného rozpätia 1,22 a relatívneho variačného rozpätia 0,2. Z týchto informácií je zrejmé, že kým v prvom prípade existujú medzi údajmi značné rozdiely, v druhom sú doby čakania na CT vyšetrenie približne rovnaké.

Funkcie v MS Excel

MIN(číslo1; číslo2; ...)
Funkcia MIN vráti najnižšie číslo z množiny hodnôt, ignoruje logické hodnoty a text.

MAX(číslo1; číslo2; ...)
Funkcia MAX vráti navyššiu hodnotu z množiny hodnôt, ignoruje logické hodnoty a text.

3.2.2 Medzikvartilové rozpätie a medzikvartilová odchýlka

Variačné rozpätie R je silne ovplyvňované extrémnymi či odľahlými hodnotami. V prípadoch, kedy máme dostatočne veľké súbory analyzovaných hodnôt, je možné tento vplyv eliminovať aj odstránením krajných hodnôt z ich usporiadanej postupnosti. V tejto súvislosti sa v štatistike často uvádzajú kvartily a medzikvartilové rozpätie R_Q.

Kvartily je možné chápať ako hodnoty, ktoré rozdeľujú súbor usporiadaných údajov na štyri rovnako veľké časti. Prvý kvartil Q_1 oddeľuje najmenších 25% hodnôt od najväčších 75% hodnôt. Druhý kvartil Q_2 rozdeľuje súbor hodnôt na dve rovnaké polovice, a je teda mediánom pozorovaných hodnôt. Tretí kvartil Q_3 najmenších 75% hodnôt od najväčších 25% hodnôt. Jednotlivé kvartily usporiadanej postupnosti súboru n hodnôt určíme podľa vzťahu (3.11):

$$\begin{aligned} Q_1 &= \frac{(n+1)}{4}\ ta\ hodnota \\ Q_2 &= \frac{2\,(n+1)}{4}\ ta\ hodnota \\ Q_3 &= \frac{3\,(n+1)}{4}\ ta\ hodnota \end{aligned} \tag{3.11}$$

V závislosti od počtu hodnôt n, môžu byť kvartily identifikované ako konkrétne hodnoty z danej postupnosti, alebo ako priemery dvoch susedných hodnôt. Zjednodušene povedané, prvý kvartil Q_1 je mediánom prvej polovice menších hodnôt, druhý kvartil Q_2 je mediánom všetkých hodnôt a tretí kvartil Q_3 je mediánom druhej polovice väčších hodnôt. Ak indexy $(n-1)/4$, $2(n-1)/4$ alebo $3(n-1)/4$ nie sú prirodzené čísla, potom je možné použiť interpoláciu medzi okolitými prvkami a vypočítaný kvartil už nemusí byť jednou z hodnôt vstupného súboru údajov (interpoláciu využíva väčšina výpočtových nástrojov, napríklad aj MS Excel).

Číselný rozdiel medzi tretím a prvým kvartilom súboru určuje jeho medzikvartilové rozpätie:

$$R_Q = Q_3 - Q_1 \tag{3.12}$$

Medzikvartilové rozpätie teda špecifikuje rozsah prostredných 50% usporiadaných hodnôt, keďže sú z neho vylúčené prípady 25% najväčších a 25% najmenších hodnôt. Inými slovami je možné povedať, že je to vzdialenosť medzi 75. a 25. percentilom (viď nasledujúca kapitola 3.2.3).

Medzikvartilovú odchýlku hodnôt pozorovanej veličiny potom definujeme vzťahom:

$$Q = \frac{R_Q}{2} = \frac{Q_3 - Q_1}{2} \tag{3.13}$$

V príklade 3.1 máme výberový súbor 30 hmotností stredoškolákov. Medián rozdeľuje túto skupinu na 15 menších hodnôt a 15 väčších hodnôt (je aritmetickým priemerom 15. a 16. hodnoty). Prvý kvartil potom predstavuje 8. hodnotu (rozdeľuje menšie hodnoty na dve rovnaké polovice) a tretí kvartil 23. hodnotu (rozdeľuje väčšie hodnoty na dve rovnaké polovice). Usporiadaná neklesajúca postupnosť hmotností stredoškolákov z príkladu 3.1 je: 38, 39, 42, 46, 48, 49, 50, 51, 52, 53, 57, 57, 57, 57, 57, 59, 59, 60, 60, 60, 61, 61, 61, 62, 63, 63, 63,

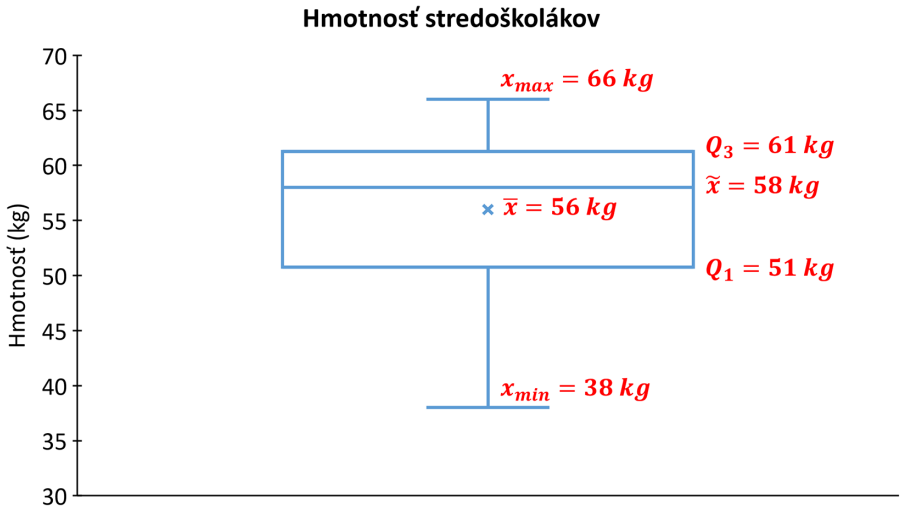
64, 65 a 66, pričom hodnoty, ktoré určujú prvý kvartil, medián a tretí kvartil sú podčiarknuté. Medzikvartilové rozpätie hmotností stredoškolákov je:

$$R_Q = 61 - 51 = 10\ kg$$

a medzikvartilová odchýlka hmotnosti stredoškolákov je:

$$Q = \frac{10}{2} = 5\ kg$$

Do výpočtu prvého Q_1 a tretieho Q_3 kvartilu je možné zahrnúť aj medián (druhý kvartil). Potom prvý kvartil bude aritmetickým priemerom 8. a 9. hodnoty a tretí kvartil aritmetickým priemerom 22. a 23. hodnoty. Jednotlivé charakteristiky popisnej štatistiky môžu byť v závislosti od charakteru výskumnej práce uvedené s komentárom v texte hodnotenia, sumarizované v prehľadnej tabuľke, alebo znázornené graficky. Grafická sumarizácia doteraz odvodených údajov k príkladu 3.1 môže byť znázornená pomocou krabicového grafu (niekedy označovaný aj ako škatuľový).



Obr. 3.2: Krabicový graf hmotnosti stredoškolákov.

Krabicový graf predstavuje obdĺžnik, ktorý ohraničuje dolný a horný kvartil hodnôt zastúpených v súbore. Obdĺžnik je vnútri predelený čiarou na úrovni mediánu (druhého kvartilu, resp. 50. percentilu). Výška obdĺžnika reprezentuje medzikvartilové rozpätie. Čiary (fúzy) nakreslené od tohoto obdĺžnika smerujú k najväčšej a najmenšej hodnote v súbore. Rozdiel medzi týmito hodnotami reprezentuje variačné rozpätie. Do krabicového grafu bola pre názornosť vložená aj hodnota priemeru (krížik uprostred obdĺžnika) a tiež popisy, ktoré reprezentujú vypočítané charakteristiky popisnej štatistiky. Krabicový graf je možné

znázorniť vertikálne, tak ako je to na obrázku 3.2, alebo aj horizontálne. Pri porovnávaní údajov z viacerých súborov sú v grafických prezentáciách zobrazované vedľa seba krabicové grafy všetkých súborov údajov.

V predchádzajúcich kapitolách už boli viackrát spomínané odľahlé, resp. extrémne hodnoty, ktoré spôsobujú skreslenie štatistických charakteristík. Odľahlé hodnoty môžu poukazovať na to, že výberový súbor nemá normálne rozdelenie, avšak nepoužívajú sa ako test hypotézy na určenie normality. Pri popisnej štatistike je preto potrebné analyzovať a vysvetliť možné dôvody prítomnosti, prípadne pôvod odľahlých hodnôt.

Jedným zo spôsobov ako definovať hranice pre odľahlé hodnoty je použitie medzikvartilového rozpätia R_Q. Potom hranice odľahlých hodnôt, horná a dolná, vieme určiť podľa vzťahu (3.14):

$$\begin{aligned} Q_d &= Q_1 - 1,5R_Q \\ Q_h &= Q_3 + 1,5R_Q \end{aligned} \tag{3.14}$$

Dolná a horná hranica odľahlých hodnôt špecifikuje oblasť, mimo ktorej je možné považovať akékoľvek hodnoty za odľahlé. Tieto hranice je možné zobraziť v krabicovom grafe (obrázok 3.2) fúzmi namiesto maximálnej a minimálnej hodnoty, pričom existujúce odľahlé hodnoty sú zobrazované ako samostatné body.

Pre hodnoty hmotnosti z príkladu 3.1 by bola hodnota dolnej hranice Q_d rovná 36 kg a hodnota hornej hranice Q_h rovná 76 kg, a keďže minimum (38 kg) aj maximum (66 kg) hodnôt hmotností stredoškolákov ležia medzi týmito hodnotami hraníc, môžeme konštatovať, že v danom výberovom súbore sa odľahlé hodnoty nenachádzajú.

Funkcie v MS Excel

QUARTILE.EXC(pole; kvart)
Funkcia QUARTILE.EXC vráti kvartil definovaný argumentom kvart z množiny údajov definovanej argumentom pole na základe hodnôt percentilov, ktoré sú väčšie ako 0 a menšie ako 1.

QUARTILE.INC(pole; kvart)
Funkcia QUARTILE.INC vráti kvartil definovaný argumentom kvart z množiny údajov definovanej argumentom pole na základe hodnôt percentilov od 0 do 1 vrátane.

3.2.3 Kvantily a kvantilové rozpätia

V predchádzajúcej kapitole bol popísaný špecifický prípad kvantilového rozdelenia údajov a následného určenia jeho rozpätia, kedy bol súbor údajov rozdelený

na štyri rovnaké časti. Pri riešení praktických úloh sa však stretávame aj s inými rozdeleniami hodnôt pomocou kvantilov, ktoré sú síce charakteristikami polohy (podobne ako minimum či maximum), ale na ich základe vieme následne definovať charakteristiky variability, medzi ktoré patria aj rozpätia i odchýlky.

Kvantily používame na získanie podrobnejšieho popisu rozdelenia hodnôt v súbore. Vo všeobecnosti je možné chápať kvantil ako hranicu, ktorá rozdeľuje súbor údajov na dve časti. V jednej časti sú hodnoty súboru, ktoré sú menšie alebo nanajvýš rovné kvantilu a v tej druhej sú hodnoty väčšie, prípadne rovné kvantilu (v závislosti od početnosti konkrétnych hodnôt nachádzajúcich sa v okolí daného kvantilu).

Najjemnejšie používané rozdelenie veľkých súborov hodnôt je rozdelenie na 100 rovnakých častí. V takom prípade kvantil označujeme ako percentil. V prípade percentilov je súbor usporiadaných údajov rozdelený na rovnaké časti, z ktorých každá obsahuje 1% hodnôt. Pre prvý percentil potom platí, že 1% hodnôt je menších alebo nanajvýš rovných ako prvý percentil a 99% hodnôt je väčších, prípadne rovných prvému percentilu. Druhý percentil potom oddeľuje 2% menších alebo nanajvýš hodnôt rovných druhému percentilu a 98% hodnôt väčších, prípadne rovných druhému percentilu. Tretí percentil oddeľuje 3% menších alebo nanajvýš hodnôt rovných tretiemu percentilu a 97% hodnôt väčších, prípadne rovných tretiemu percentilu atď. Medzipercentilové rozpätie tak obsahuje 98% hodnôt, pričom 1% hodnôt zľava a 1% hodnôt sprava je odstránených.

Iným, často využívaným rozdelením súboru hodnôt sú decily, ktoré rozdeľujú hodnoty na desať rovnakých častí. V tomto prípade platí, že prvý decil oddeľuje 10% hodnôt, ktoré sú menšie alebo nanajvýš rovnaké ako prvý decil a 90% hodnôt, ktoré sú väčšie, prípadne rovnaké ako prvý decil. Druhý decil ďalej oddeľuje 20% menších alebo nanajvýš hodnôt rovných druhému decilu a 80% hodnôt väčších, prípadne rovných druhému decilu atď. Medzidecilové rozpätie tak obsahuje 80% hodnôt, pričom 10% hodnôt zľava a 10% hodnôt sprava je odstránených.

Vo všeobecnosti, ak je súbor usporiadaných hodnôt rozdelený na α rovnakých častí, t. j. častí s rovnakou pravdepodobnosťou $p = 1/\alpha$, potom k-ty kvantil určíme podľa vzťahu (3.15):

$$Q_k = \frac{k\,(n+1)}{\alpha}\ ta\ hodnota \tag{3.15}$$

Kvantilové rozpätie, ktoré určuje vzdialenosť medzi dolným a horným kvantilom potom vypočítame ako:

$$R_Q = Q_{\alpha-1} - Q_1 \tag{3.16}$$

Aj tu platí, že používanie kvantilových rozpätí minimalizuje vplyv extrémnych, resp. odľahlých hodnôt.

Typy jednotlivých kvantilov, ich názvy a základné charakteristiky sú prehľadne uvedené v tabuľke 3.3.

Tabuľka 3.3: Typy kvantilov.

Názov	**Počet kvantilov**	**Počet častí**	**Veľkosť časti (%)**	**Veľkosť rozpätia (%)**
medián	1	2	50	-
tercily	2	3	33,33	33,33
kvartily	3	4	25	50
kvintily	4	5	20	60
sextily	5	6	16,67	66,66
septily	6	7	14,29	71,42
oktávily	7	8	12,5	75
nonily	8	9	11,11	77,78
decily	9	10	10	80
percentily	99	100	1	98

3.2.4 Priemerná odchýlka

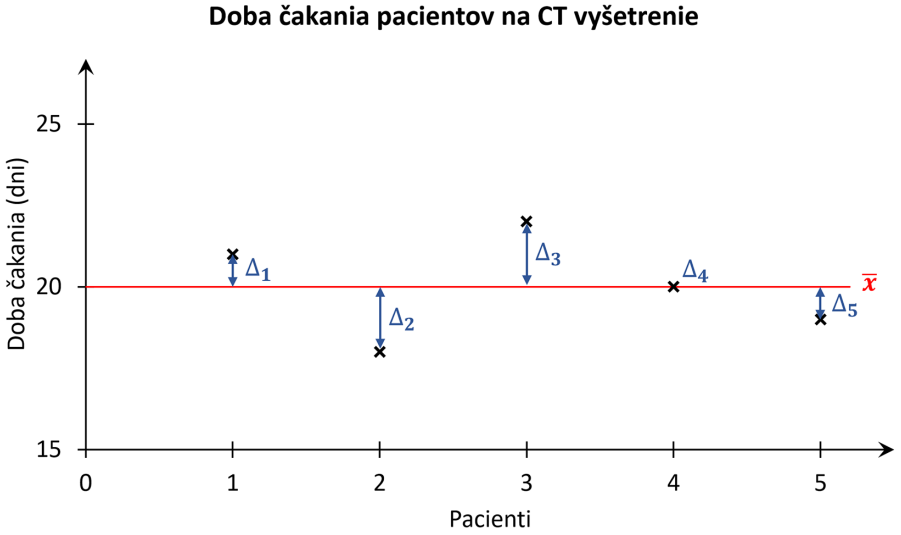
Iným prístupom ku kvantifikácii variability hodnôt v súbore, než je hľadanie hraničných hodnôt a určovanie ich rozpätí, je využitie všetkých hodnôt a hľadanie ich rozdielov od priemernej hodnoty. Tieto rozdiely sú následne spriemerované. Takouto charakteristikou je napríklad priemerná odchýlka.

Rozdielom individuálnej hodnoty x_i od priemeru $\bar{x}$ je vzdialenosť:

$$\Delta_i = x_i - \bar{x} \tag{3.17}$$

pričom ak je rozsah súboru n hodnôt, potom máme n takýchto rozdielov.

Pre hodnoty menšie ako priemer budú tieto rozdiely záporné a pre hodnoty väčšie ako je priemer budú rozdiely kladné. Z týchto rozdielov je možné vypočítať priemernú hodnotu, avšak suma rozdielov bude vždy rovná nule. Je to dané tým, že priemerná hodnota je stredom všetkých hodnôt a preto súčet všetkých záporných rozdielov bude vždy rovnaký ako súčet všetkých kladných rozdielov



Obr. 3.3: Doba čakania pacientov na CT vyšetrenie v nemocnici B a rozdiely pozorovaných hodnôt od priemeru.

s opačným znamienkom. Túto vlastnosť si vysvetlíme na hodnotách z príkladu 3.3, ktoré sú zobrazené na obrázku 3.3.

Priemerná doba čakania na CT vyšetrenie je 20 dní. Rozdiely jednotlivých hodnôt od priemeru teda sú: 1, -2, 2, 0 a -1 a ich súčet je rovný 0, rovnako ako bude súčet rozdielov hodnôt od priemeru v ktoromkoľvek inom súbore údajov. Nulový bude teda vždy aj aritmetický priemer týchto rozdielov, čo nám neumožní kvantifikovať variabilitu hodnôt súboru. Avšak, riešením je použiť priemernú odchýklu e , ktorú vypočítame ako aritmetický priemer absolútnych hodnôt všetkých odchýlok od ich priemeru:

$$e = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \tag{3.18}$$

Potom priemerná odchýlka doby čakania na CT vyšetrenie v nemocnici B je 1,2 dňa, vypočítaná podľa vzťahu (3.18):

$$e = \frac{|21-20| + |18-20| + |22-20| + |20-20| + |19-20|}{5} = \frac{6}{5} = 1,2$$

Podobne zistíme, že priemerná odchýlka doby čakania na CT vyšetrenie v nemocnici A z príkladu 3.2 je 28 dní, keďže platí, že:

$$e = \frac{|3-20| + |1-20| + |90-20| + |4-20| + |2-20|}{5} = \frac{140}{5} = 28$$

Z informácií o štatistických charakteristikách vidíme, že priemerné doby čakania na CT vyšetrenie boli u oboch skupín rovnaké, ale priemerné odchýlky

poukazujú že, doby čakania sú rozdielne a pacienti v nemocnici A môžu čakať na CT vyšetrenie výrazne dlhšie, alebo aj výrazne kratšie ako 20 dní.

Funkcie v MS Excel

ABS(číslo)
Funkcia ABS vráti absolútnu hodnotu čísla, t. j. číslo bez znamienka.

3.2.5 Rozptyl

Absolútne hodnoty rozdielov hodnôt od priemeru, použité vo výpočte priemernej odchýlky odstraňujú vplyv znamienok a zabezpečia, že ich suma nebude vždy nulová. Inou možnosťou, ako odstrániť vplyv znamienka je použiť druhé mocniny odchýlok od priemeru. Umocnením odchýlky Δ_i na druhú odstránime vplyv znamienka (všetky umocnené odchýlky sú kladné) a zároveň zvýrazníme odľahlé, resp. extrémne hodnoty v danom súbore. Odchýlky, ktorých absolútne hodnoty sú malé, t. j. menšie ako 1 umocnením budú ešte menšie a naopak, hodnoty veľkých odchýlok sa umocnením ešte zvýraznia. Aritmetický priemer štvorcov odchýlok hodnôt (umocnených na druhú) od ich aritmetického priemeru predstavuje najčastejšie používanú charakteristiku variability, ktorou je rozptyl. Rozptyl základného súboru vypočítame podľa vzťahu:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{3.19}$$

kde N je veľkosť základného súboru a μ je aritmetický priemer základného súboru.

Ak budeme považovať hodnoty z príkladu 3.1 za hodnoty základného súboru (teoreticky, inak by to mali byť hodnoty hmotností všetkých stredoškolákov), potom ich rozptyl vypočítame dosadením hodnôt do vzťahu (3.19):

$$\sigma^2 = \frac{(39-56)^2 + (66-56)^2 + \cdots + (63-56)^2}{30} = \frac{1682}{30} = 56,066\bar{6}$$

Rozptyl základného súboru hmotností stredoškolákov je 56,07 kg^2. Podobne, ak by boli hodnoty systolického krvného tlaku z príkladu 3.4, hodnotami základného súboru, potom ich rozptyl:

$$\sigma^2 = \frac{(122-125)^2 + (116-125)^2 + \cdots + (122-125)^2}{10} = \frac{3418}{10} = 341,8$$

Rozptyl základného súboru systolického krvného tlaku je 341,8 mmHg^2. Hodnoty základného súboru však väčšinou nie sú známe a preto vo väčšine štúdií

vypočítavame štatistické charakteristiky z výberového súboru a charakteristiky základného súboru odhadujeme.

Rozptyl hodnôt $x_1, x_2, \cdots, x_n$ výberového súboru o veľkosti n vypočítame podľa vzťahu (3.20):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.20}$$

kde $\bar{x}$ je priemerná hodnota a $n-1$ predstavuje počet stupňov voľnosti, ktorý uprednostňujeme pred samotným použitím n pri odhadoch či testovaní hypotéz (ktoré budú diskutované v ďalších kapitolách). Delenie číslom $n-1$ nám dáva najlepší matematický odhad rozptylu základného súboru, aj keď pre veľké vzorky je rozdiel zanedbateľný. Počet stupňov voľnosti je spojený so štatistikou či štatistikami. Napríklad číslo 1 je tu spojené s jednou štatistikou, ktorou je rozptyl. V iných prípadoch, kde bude uvažovaných viacero štatistík, bude toto číslo zodpovedať počtu štatistík.

Rozptyl výberového súboru hodnôt z príkladu 3.1 potom vypočítame dosadením hodnôt hmotností do vzťahu (3.20):

$$s^2 = \frac{(39-56)^2 + (66-56)^2 + \cdots + (63-56)^2}{29} = \frac{1682}{29} = 58$$

Rozptyl výberového súboru hmotností stredoškolákov je 58 kg^2. Podobne, pre hodnoty systolického krvného tlaku z príkladu 3.4 bude rozptyl:

$$s^2 = \frac{(122-125)^2 + (116-125)^2 + \cdots + (122-125)^2}{9} = \frac{3418}{9} = 379,7\bar{7}$$

Rozptyl výberového súboru hodnôt systolického krvného tlaku je 379,78 mmHg^2.

Čím je hodnota rozptylu väčšia, tým viac sú hodnoty distribuované od ich strednej hodnoty a naopak. Špecifický prípad by nastal, ak by boli všetky hodnoty pozorovanej veličiny rovnaké. Vtedy je rozptyl rovný 0.

Funkcie v MS Excel

VAR.P(číslo1; číslo2; ...)
Funkcia VAR.P vypočíta rozptyl základného súboru, ktorý bol zadaný ako argumenty funkcie, pričom ignoruje logické hodnoty a text.

VAR.S(číslo1; číslo2; ...)
Funkcia VAR.S vráti odhad rozptylu na základe výberového súboru, vo výpočte ignoruje logické hodnoty a text.

SUBTOTAL(číslo_funkcie; odkaz1; odkaz2; ...)
Funkcia SUBTOTAL vracia medzisúčet v zozname alebo databáze. Ak je prvý argument (číslo_funkcie) rovný 11, potom funkcia SUBTOTAL vráti rozptyl základného súboru z číselných hodnôt uložených v bunkách, ktorých odkazy tvoria ďalšie argumenty funkcie. Ak je prvý argument funkcie rovný 10, potom funkcia SUBTOTAL vráti odhad rozptylu na základe výberového súboru. Vo výpočte funkcie SUBTOTAL sa nezahŕňajú hodnoty v skrytých bunkách.

3.2.6 Smerodajná odchýlka

Pozornému čitateľovi určite neušlo, že hodnoty rozptylu sú uvádzané v štvorcoch (kvadrátoch) jednotiek danej veličiny. Napríklad rozptyl hmotnosti je v kg^2, rozptyl výšky v m^2, rozptyl krvného tlaku v $mmHg^2$ atď. Takáto prezentácia výsledkov je síce matematicky správna, avšak z pohľadu interpretácie a pochopenia súvislostí náročná.

Z tohoto dôvodu sa v prezentáciách štatistických charakteristík častejšie používa smerodajná odchýlka, ktorá je udávaná v merných jednotkách danej veličiny. Smerodajná odchýlka nám tak poskytuje informáciu o tom, ako sa v priemere líšia jednotlivé hodnoty od aritmetického priemeru, a to v oboch smeroch, t. j. v kladnom aj zápornom. Je vypočítavaná ako druhá odmocnina rozptylu. Smerodajná odchýlka základného súboru σ potom bude vypočítaná ako:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2} \tag{3.21}$$

Ak budeme považovať hodnoty z príkladu 3.1 za hodnoty základného súboru, potom ich smerodajnú odchýlku vypočítame ako odmocninu rozptylu základného súboru σ^2:

$$\sigma = \sqrt{56,07} = 7,487768 \doteq 7,49$$

Smerodajná odchýlka základného súboru hodnôt hmotností stredoškolákov je teda 7,49 kg. Podobne, ak by sme predpokladali, že hodnoty systolického

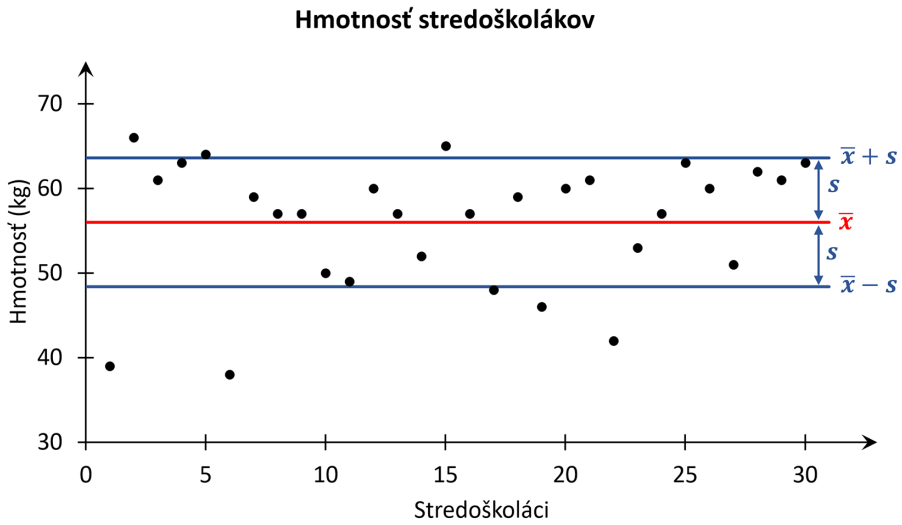
krvného tlaku z príkladu 3.4, sú hodnotami základného súboru, potom ich smerodajná odchýlka bude:

$$\sigma = \sqrt{341,8} = 18,487834 \doteq 18,49$$

t. j. smerodajná odchýlka základného súboru hodnôt systolického krvného tlaku je 18,49 mmHg. Keďže hodnoty základného súboru väčšinou nie sú známe, aj tu častejšie vypočítavame smerodajnú odchýlku z hodnôt výberového súboru. Smerodajnú odchýlku výberového súboru s vypočítame podľa vzťahu:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.22}$$

Smerodajná odchýlka výberového súboru hodnôt hmotností stredoškolákov z príkladu 3.1 potom bude $s = \sqrt{58} = 7,615773 \doteq 7,62\ kg$. Podobne, pre hodnoty systolického krvného tlaku z príkladu 3.4 bude smerodajná odchýlka výberového súboru $s = \sqrt{379,78} = 19,487888 \doteq 19,49\ mmHg$. Hodnoty smerodajnej odchýlky v jednotkách skúmanej veličiny sú takto jednoduchšie interpretovateľné a pochopiteľné. Význam smerodajnej odchýlky pochopíme aj z grafického znázornenia, kde na obrázku 3.4 sú prezentované hodnoty hmotnosti stredoškolákov z príkladu 3.1.



Obr. 3.4: Hmotnosť stredoškolákov a priemerné rozdiely od priemernej hodnoty dané hodnotou smerodajnej odchýlky.

Čierne body na obrázku 3.4 znázorňujú jednotlivé hmotnosti stredoškolákov. Červená čiara je priemernou hodnotou $\bar{x}$ (56 kg) a modré čiary určujú oblasť s výškou $2s$, ktorá vznikla pripočítaním a odpočítaním hodnoty smerodajnej

odchýlky s (7,62 kg) od priemernej hodnoty hmotnosti $\bar{x}$. Pripomíname, že modus hmotnosti $\hat{x}$ je 57 kg a medián $\tilde{x}$ je 58 kg, teda sú to hodnoty väčšie ako priemer. Tieto charakteristiky výberového súboru potvrdzuje aj vyššia koncentrácia bodov (hmotností) v hornej časti rozdelenia hodnôt, t. j. nad priemerom. Ak by sme hmotnosť vynášali na x-ovej osi, hovoríme že ležia napravo od priemernej hodnoty. Taktiež je zrejmé, že hodnoty v spodnej časti (menšie hodnoty, pri prezentácii na x-ovej osi ležia naľavo) majú väčšiu variabilitu, čo môže byť znakom asymetrie rozloženia údajov (bude diskutované v ďalších kapitolách).

Funkcie v MS Excel

STDEV.P(číslo1; číslo2; ...)
Funkcia STDEV.P vypočíta smerodajnú odchýlku základného súboru, ktorý bol zadaný ako argumenty funkcie, pričom ignoruje logické hodnoty a text.

STDEV.S(číslo1; číslo2; ...)
Funkcia STDEV.S vráti odhad smerodajnej odchýlky na základe výberového súboru, vo výpočte ignoruje logické hodnoty a text.

SQRT(číslo)
Funkcia SQRT vráti druhú odmocninu čísla.

SUBTOTAL(číslo_funkcie; odkaz1; odkaz2; ...)
Funkcia SUBTOTAL vracia medzisúčet v zozname alebo databáze. Ak je prvý argument (číslo_funkcie) rovný 8, potom funkcia SUBTOTAL vráti smerodajnú odchýlku základného súboru z číselných hodnôt uložených v bunkách, ktorých odkazy tvoria ďalšie argumenty funkcie. Ak je prvý argument funkcie rovný 7, potom funkcia SUBTOTAL vráti odhad smerodajnej odchýlky na základe výberového súboru. Vo výpočte funkcie SUBTOTAL sa nezahŕňajú hodnoty v skrytých bunkách.

3.2.7 Variačný koeficient

Rozptyl s^2 aj smerodajná odchýlka s sú v štatistických analýzach užitočnými charakteristikami variability hodnôt jednej veličiny (premennej) v danom výberovom súbore. Niekedy je však potrebné porovnať variabilitu viacerých veličín, napríklad v dvoch rôznych súboroch údajov. V takýchto prípadoch nie je možné použiť samostatný rozptyl alebo smerodajnú odchýlku. Údaje dvoch rôznych veličín totiž môžu mať buď rôzne jednotky alebo veličiny môžu mať rôzne priemery. Na porovnanie preto používame mieru relatívnej variability, ktorá nie je závislá ani od jednotiek a ani od hodnôt priemerov.

Charakteristikou variability vhodnou na porovnávanie dvoch súborov a ich veličín je variačný koeficient v, ktorý predstavuje bezrozmerné číslo a vypočítame ho ako podiel smerodajnej odchýlky a aritmetického priemeru podľa vzťahu (3.23):

$$v = \frac{s}{\bar{x}} \tag{3.23}$$

kde s je smerodajná odchýlka a $\bar{x}$ je aritmetický priemer hodnôt sledovaného výberového súboru.

Variačný koeficient možno považovať za mieru variability údajov vo vzťahu k priemeru údajov a niekedy sa vyjadruje v percentách, t. j. variabilita daná smerodajnou odchýlkou je vyjadrená v percentách aritmetického priemeru. Vtedy je vzťah (3.23) upravený takto:

$$v = \frac{s}{\bar{x}} 100\% \tag{3.24}$$

Pomocou variačného koeficientu tak vieme určiť, aj v prípade ak sú rozptyly alebo smerodajné odchýlky dvoch veličín rovnaké, či sú hodnoty viac rozptýlené v prvom alebo v druhom súbore. Záver, že dva súbory hodnôt majú rovnakú variabilitu lebo majú rovnakú smerodajnú odchýlku, odvodený len z hodnoty smerodajnej odchýlky a bez ďalších charakteristík je totiž mylný.

Pripomeňme si opäť príklady 3.2 a 3.3, kde sme sledovali dobu čakania na CT vyšetrenie v dvoch rôznych nemocniciach a porovnajme ich variabilitu pomocou variačného koeficienta. Variačné koeficienty doby čakania v nemocnici A a v nemocnici B budú:

$$v_A = \frac{39,15}{20} 100\% = 195,74\%$$

$$v_B = \frac{1,58}{20} 100\% = 7,91\%$$

Z vypočítaných hodnôt variačného koeficientu je vidieť, že variabilita doby čakania pacientov na CT vyšetrenie v nemocnici A je výrazne vyššia ako variabilita doby čakania v nemocnici B.

Príklad 3.5. Predpokladajme, že v rámci vstupných vyšetrení pacientov na internom oddelení boli zaznamenávané hodnoty krvného tlaku. U desiatich pacientov z príkladu 3.4 boli zaznamenané aj hodnoty diastolického krvného tlaku v mmHg tak, ako sú uvedené v tabuľke 3.4. Pomocou variačného koeficientu porovnajme variabilitu hodnôt systolického a diastolického krvného tlaku.

Ak vypočítame charakteristiky popisnej štatistiky pre súbor pozorovaných hodnôt diastolického tlaku zistíme, že jeho priemerná hodnota $\bar{x}$ sa rovná 82 mmHg, medián $\tilde{x}$ je 80 mmHg, modus $\hat{x}$ je 80 mmHg, rozptyl s^2 sa rovná

Tabuľka 3.4: Hodnoty diastolického krvného tlaku.

Pacient č. i	**Diastolický KT (mmHg)**	**Pacient č. i**	**Diastolický KT (mmHg)**
1	80	6	80
2	62	7	74
3	80	8	62
4	130	9	80
5	76	10	96

379,56 mmHg^2 a smerodajná odchýlka s je 19,48 mmHg. Smerodajné odchýlky diastolického a systolického krvného tlaku sú teda rovnaké.

Vypočítajme pre obe sledované veličiny variačné koeficienty podľa vzťahu (3.23). Ak variačný koeficient diastolického krvného tlaku označíme ako v_D a systolického krvného tlaku ako v_S potom:

$$v_S = \frac{19,48}{125}100\% = 15,59\%$$

$$v_D = \frac{19,48}{80}100\% = 23,76\%$$

Z vypočítaných hodnôt variačného koeficientu je vidieť, že variabilita diastolického krvného tlaku je vyššia ako variabilita systolického krvného tlaku.

3.3 Charakteristiky tvaru

Charakteristiky tvaru popisujú ako sú údaje v rámci súboru údajov distribuované. Spravidla ich asociujeme s grafickou prezentáciou, avšak vzor ich rozloženia vieme popísať aj číselne, a to pomocou koeficientov šikmosti a špicatosti. Samotný tvar rozloženia kvantitatívnych údajov je možné popísať, ak sú hodnoty logicky usporiadateľné tak, aby mohli byť premennou zobrazovanou na x-ovej osi. Tvar distribúcie neusporiadateľných kvalitatívnych údajov nie je možné popísať, keďže údaje nie sú číselné. Informácie o tvare distribúcie údajov nám môžu pomôcť pri identifikácii iných popisných štatistík. Napríklad, ktorú charakteristiku polohy je vhodné použiť na popis súboru údajov.

Pri riešení výskumných úloh a štatistickom hodnotení získaných údajov sa stretávame s prípadmi, kedy má súbor údajov a ich rozloženie:

- jeden vrchol – unimodálne rozdelenie, ktoré je možné dobre popísať šikmosťou a špicatosťou, ktoré sú tretím a štvrtým centrálnym momentom

distribučnej funkcie (distribúcia početností). Poznamenávame, že druhý centrálny moment distribučnej funkcie korešponduje s rozptylom,

- viac vrcholov – multimodálne rozdelenie, ktoré popisujeme polohou a výškou jednotlivých vrcholov.

Pri popisovaní rozloženia údajov by sme sa mali pozerať práve na zošikmenie, prípadne nadmernú špicatosť, ktoré lepšie reprezentujú extrémy súboru údajov než samotný priemer.

3.3.1 Šikmosť

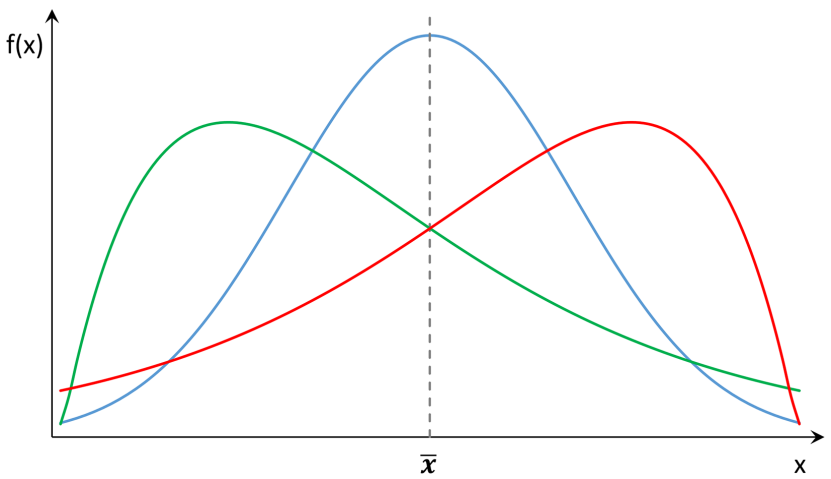
Šikmosť (*skewness*), resp. koeficient šikmosti charakterizuje smer a stupeň asymetrie rozdelenia hodnôt. Asymetriu posudzujeme voči ideálnemu symetrickému rozdeleniu, ktorým je normálne rozdelenie, pričom asymetrické rozdelenie pozorovaných údajov môže byť reprezentované krivkou, ktorej jedna strana je posunutá, či skôr natiahnutá smerom doľava alebo doprava.

Predpokladajme, že máme výberový súbor hodnôt rozsahu n, potom koeficient šikmosti vypočítame podľa vzťahu (3.25):

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \tag{3.25}$$

kde $\bar{x}$ je aritmetický priemer výberového súboru a s je smerodajná odchýlka výberového súboru.

Príklad symetrického a asymetrického rozdelenia zobrazuje obrázok 3.5.

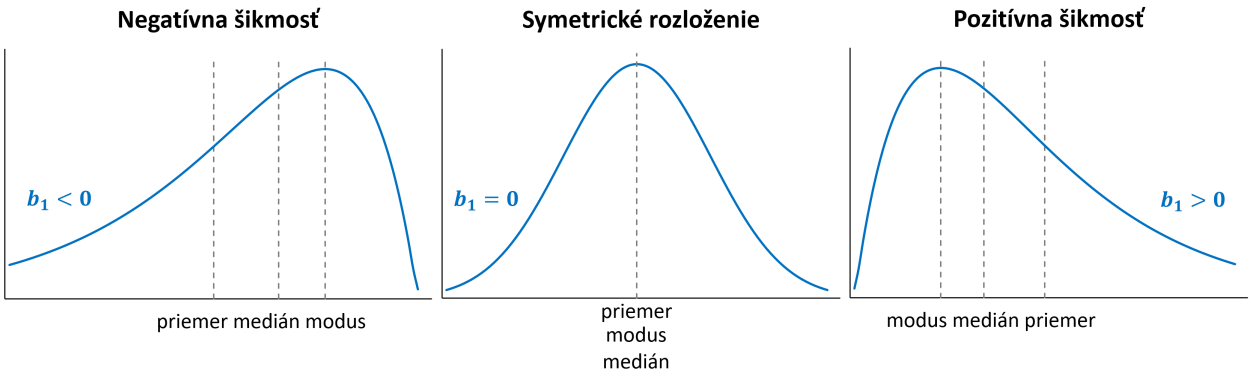


Obr. 3.5: Symetrické (modré), pravostranne asymetrické (zelené) a ľavostranne asymetrické (červené) rozdelenia.

Symetrické rozloženie, akým je aj tzv. normálne rozloženie (známe aj ako Gaussovo rozloženie) má priebeh krivky rozloženia údajov v tvare zvona a jeho koeficient šikmosti je rovný nule. Normálne rozdelenie má priemer, modus aj medián rovnaký. Symetrické rozdelenie okolo strednej hodnoty $\bar{x}$ je na obrázku 3.5 znázornené modrou čiarou. Pravá aj ľavá strana rozloženia sú rovnaké, t. j. sú zrkadlovým obrazom. Ak je koeficient šikmosti b_1 blízky nule, potom sa rozdelenie hodnôt blíži k symetrickému. U takýchto rozdelení, ktoré považujeme za symetrické, sú priemer, modus aj medián približne rovnaké.

Ak je koeficient šikmosti $b_1 < 0$, potom je rozdelenie zošikmené doľava, tzv. ľavostranná šikmosť. Ľavostrannú šikmosť označujeme aj ako negatívna šikmosť, a hodnoty na ľavej strane vykazujú väčšiu variabilitu ako hodnoty na strane pravej. Negatívna šikmosť je na obrázku 3.5 znázornená červenou čiarou. Analogicky, ak je koeficient šikmosti $b_1 > 0$, potom je rozdelenie zošikmené doprava, tzv. pravostranná šikmosť. Pravostrannú šikmosť označujeme aj ako pozitívna šikmosť a hodnoty na pravej strane vykazujú väčšiu variabilitu ako hodnoty na strane ľavej. Pozitívna šikmosť je na obrázku 3.5 znázornená zelenou čiarou.

U zošikmených rozložení má priemer $\bar{x}$ tendenciu ležať na tej istej strane rozloženia, na ktorú je rozloženie zošikmené. Ako už bolo uvedené vyššie, u symetrického rozloženia sa priemer, modus a medián zhodujú. Zvyčajné rozmiestnenie týchto charakteristík polohy (závisí však od konkrétnych údajov) u jednotlivých typov symetrického a asymetrického rozloženia je zobrazené na obrázku 3.6.



Obr. 3.6: Negatívne zošikmené rozloženie (vľavo), symetrické rozloženie (v strede) a pozitívne zošikmené rozloženie (vpravo).

Pre unimodálne rozloženia vo všeobecnosti platí, že:

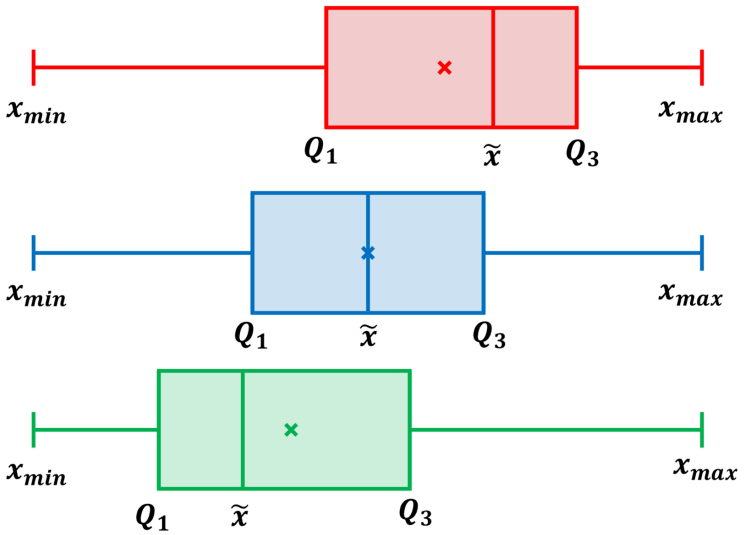
- u negatívnej šikmosti (ľavostrannej) je priemer menší ako medián,
- u pozitívnej šikmosti (pravostrannej) je priemer väčší ako medián,
- u šikmosti rovnej 0 (symetrické rozdelenie) sa priemer a medián rovnajú.

Koeficient šikmosti pre výberový súbor údajov z príkladu 3.1 vypočítame dosadením hodnôt hmotností stredoškolákov do vzťahu 3.25. Pripomenieme, že veľkosť vzorky n je 30, priemerná hodnota hmotností $\bar{x}$ je 56 kg a smerodajná odchýlka s je 7,62 kg.

$$b_1 = \frac{\frac{1}{30}\left((39-56)^3 + (66-56)^3 + \cdots + (63-56)^3\right)}{7,62^3} = -0,8802$$

Koeficient šikmosti b_1 hodnôt hmotností stredoškolákov je -0,8802 čo signalizuje, že údaje sú negatívne zošikmené, t. j. viac distribuované na ľavej strane. Na túto skutočnosť sme už poukázali aj v kapitole 3.2.2 a graficky na obrázku 3.2. Negatívny charakter rozloženia bol prezentovaný aj v kapitole 3.2.6 na obrázku 3.4, kde boli menšie hodnoty údajov (pod priemerom) rozložené vo väčšej oblasti. Aj tu je pozorované, že priemer $\bar{x}$ (56 kg) je menší ako medián $\tilde{x}$ (58 kg).

Asymetriu rozloženia, a teda aj šikmosť je možné pozorovať aj pomocou krabicových grafov. Príklad symetrického a asymetrického rozloženia v krabicových grafoch je zobrazený na obrázku 3.7.



Obr. 3.7: Krabicové grafy symetrického (modrý), pozitívne šikmého (zelený) a negatívne šikmého (červený) rozdelenia.

Z modrého krabicového grafu na obrázku 3.7, ktoré znázorňuje symetrické rozdelenie je vidieť, že medián aj priemer (označený krížikom) sú rovnaké a ležia uprostred medzikvartilového intervalu. Fúzy, smerujúce k minimálnej hodnote na ľavej strane a k maximálnej hodnote na pravej strane sú rovnako veľké (ideálne symetrické rozdelenie). Červený krabicový graf reprezentuje negatívnu šikmosť, u ktorej je priemer menší ako medián. Taktiež hodnoty na ľavej strane

sú viac rozptýlené, čo znázorňuje dlhší fúz smerujúci k minimálnej hodnote. Analogicky je možné interpretovať krabicové grafy pre pozitívnu šikmosť (na obrázku 3.7 zobrazená zelenou farbou), kedy je priemerná hodnota väčšia ako medián a dlhší fúz je na pravej strane, poukazujúci na dlhšiu vzdialenosť k maximálnej hodnote. Špecifickým prípadom môže byť krabicový graf, ktorý má priemer aj medián približne rovnaké, ale fúzy vykazujú rôzne dĺžky. Potom rozdiel v dĺžkach týchto fúzov môže indikovať prítomnosť asymetrie a jej smer.

Vyššie uvedený vzťah (3.25) pre výpočet koeficientu šikmosti výberového súboru, ktorý nezohľadňuje systematickú odchýlku (bias), je možné upraviť a použiť aj na odhad šikmosti základného súboru. Na odhad šikmosti základného súboru z údajov výberového súboru použijeme nasledujúcu koreláciu:

$$B_1 = \frac{\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3} \tag{3.26}$$

Vzťah (3.26) korešponduje s korekciou systematickej odchýlky a je používaný v mnohých výpočtových aplikáciách a tabuľkových kalkulátoroch. Rozdiely vo výsledných hodnotách vypočítaných vzťahom (3.25) a vzťahom (3.26) sú minimálne a so zvyšovaním rozsahu n klesajú.

Pre porovnanie sa vráťme k príkladu 3.1 a dosadením hodnôt hmotností stredoškolákov do vzťahu (3.26) získame hodnotu koeficientu šikmosti -0,9756, ktorá rovnako ako hodnota -0,8802, vypočítaná pomocou vzťahu (3.25) poukazuje na negatívnu šikmosť rozdelenia hodnôt hmotností stredoškolákov.

Funkcie v MS Excel

SKEW(číslo1; číslo2; ...)
Funkcia SKEW vráti hodnotu šikmosti množiny údajov.

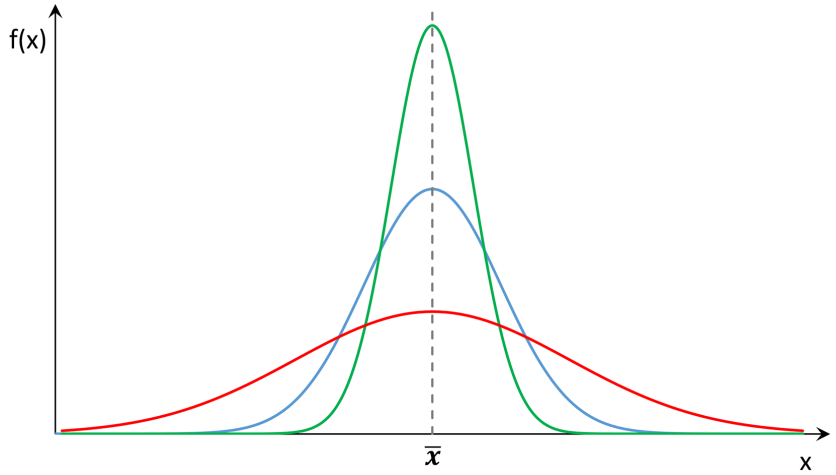
3.3.2 Špicatosť

Špicatosť (*kurtosis*), resp. koeficient špicatosti je charakteristikou, ktorá popisuje koncentráciu hodnôt sledovanej veličiny okolo ich priemernej hodnoty. Poukazuje tiež na hustotu koncových častí rozdelenia, prípadne výskyt extrémne vysokých alebo extrémne nízkych hodnôt. Na výpočet koeficientu špicatosti výberového súboru použijeme vzťah (3.27):

$$b_2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} \tag{3.27}$$

kde $\bar{x}$ je aritmetický priemer hodnôt výberového súboru a s je ich smerodajnou odchýlkou.

Príklad špicatého, normálneho a plochého rozdelenia je znázornený na obrázku 3.8.



Obr. 3.8: Normálne (modré), špicaté (zelené) a ploché (červené) rozdelenia.

Špicatosť či plochosť daného rozloženia hodnôt posudzujeme voči normálnemu rozdeleniu, ktoré je na obrázku 3.8 znázornené modrou čiarou. Tvar normálneho rozloženia označujeme aj ako *mezokurtické*. Ak výberový súbor pozostáva z väčšiny hodnôt blízkych priemeru, t. j. najväčšia koncentrácia je v okolí strednej hodnoty a len málo hodnôt je od priemeru vzdialených, potom grafické znázornenie rozloženia údajov je užšie a vyššie vo svojom vrchole ako normálne rozdelenie. Tvar špicatého rozloženia údajov je na obrázku 3.8 zobrazený zelenou čiarou a označujeme ho aj ako *leptokurtické*. Naopak, červená čiara predstavuje príklad plochého rozloženia údajov, kedy hodnoty v súbore majú väčšiu variabilitu a len málo (menej v rovnaní s normálnym rozdelením) sa ich nachádza okolo priemeru. Ploché rozloženia označujeme aj ako *platykurtické*.

Špicatosť, je rovnako ako šikmosť bezrozmernou veličinou. Pre normované normálne rozdelenie hodnôt je koeficient špicatosti b_2 vypočítaný podľa vzťahu (3.27) rovný 3. Preto platí, ak koeficient špicatosti vypočítaný podľa vzťahu (3.27) bude $b_2 = 3$ dané rozdelenie bude mať tvar normálneho rozdelenia. Ak bude koeficient špicatosti $b_2 > 3$ dané rozdelenie bude špicaté, t. j. s úzkym vrcholom. Ak bude koeficient špicatosti $b_2 < 3$, potom dané rozdelenie bude ploché, t. j. široké s väčším počtom údajov po stranách. Pri posudzovaní špicatosti je dôležité si uvedomiť, že ma význam len pre symetrické distribúcie. V prípade nesymetrických údajov je potrebné sledovať pre vytváranie záverov o analyzovaných údajoch aj iné štatistické charakteristiky.

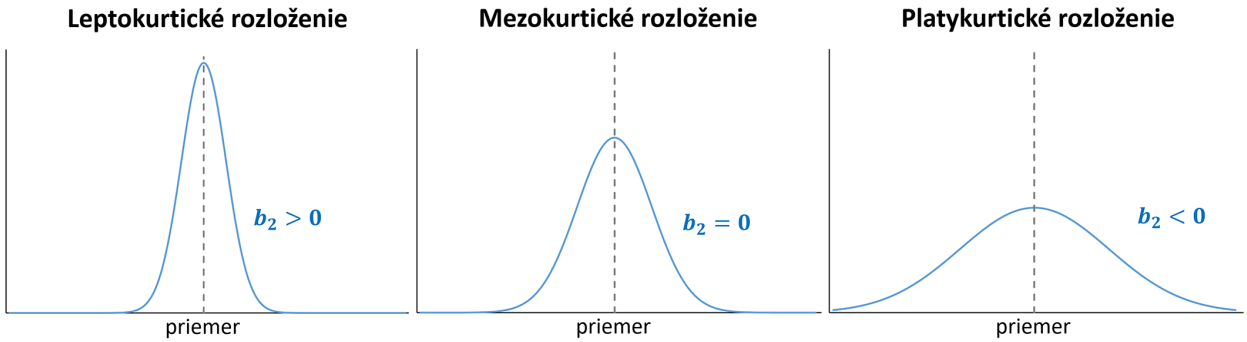
Pre jednoduchšiu interpretáciu údajov o špicatosti, resp. koeficientoch špicatosti sa v praxi, ako aj v mnohých výpočtových programoch a tabuľkových kalkulátoroch koeficient šikmosti normálneho rozdelenia znižuje na hodnotu 0 (tzv. nadmerná špicatosť, *excess kurtosis*), ktorá je ako rozhodovacia úroveň o špicatosti či plochosti prirodzenejšou, napríklad tak ako je to u šikmosti. Potom upravený vzťah (3.27) bude doplnený o túto korekciu:

$$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3 \tag{3.28}$$

a pre interpretáciu hodnôt koeficientu špicatosti platí:

- ak $b_2=0$, rozdelenie má tvar normálnej špicatosti,
- ak $b_2>0$, rozdelenie je špicaté, špicatejšie v porovnaní s normálnym,
- ak $b_2<0$, rozdelenie je ploché.

U špicatých rozložení je najviac hodnôt umiestnených okolo priemeru $\bar{x}$, preto je vrchol rozloženia vysoký. Naopak, u plochých rozložení je hodnôt v okolí priemeru menej, vrchol je nižší, po stranách je hodnôt viac a rozloženie je širšie. Príklady jednotlivých typov špicatosti unimodálnych rozložení sú zobrazené na obrázku 3.9.



Obr. 3.9: Leptokurtické rozloženie (vľavo), mezokutrické rozloženie (v strede) a platykurtické rozloženie (vpravo).

Pri riešení štatistických problémov výskumných úloh spracovávame údaje z rôzne veľkých výberových súborov a korekcia koeficientu šikmosti číslom 3 vo vzťahu (3.28) je výsledkom pozorovaní u dostatočne veľkých výberových súborov. Štatistické výpočtové aplikácie a tabuľkové kalkulátory používajú na výpočet šikmosti vzťah, ktorý zmenu veľkosti súboru zohľadňuje:

$$B_2 = \frac{n\,(n+1)}{(n-1)\,(n-2)\,(n-3)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - \frac{3\,(n-1)^2}{(n-2)\,(n-3)} \tag{3.29}$$

kde pre veľké rozsahy n výberových súborov platí, že:

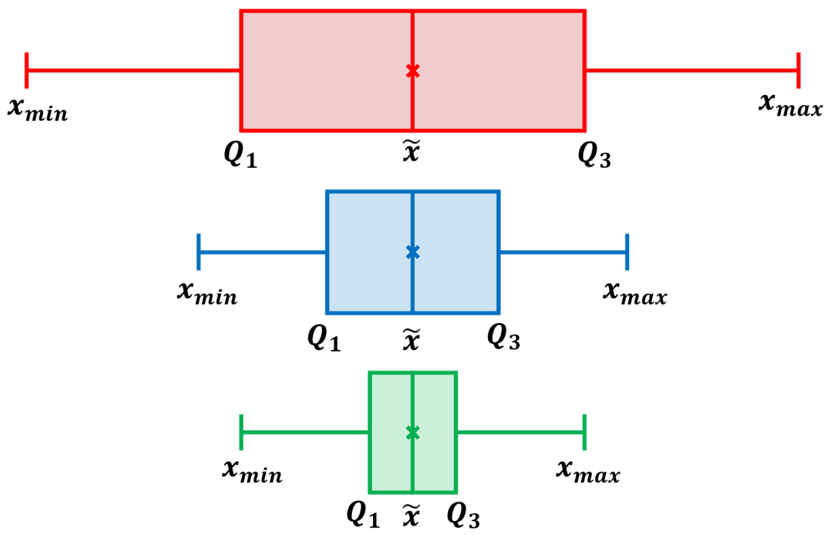
$$\frac{3\,(n-1)^2}{(n-2)\,(n-3)} \approx 3$$

Koeficient špicatosti pre hodnoty hmotností z príkladu 3.1 vypočítame podľa vzťahu (3.29):

$$B_2 = \frac{30\,(30+1)}{(30-1)\,(30-2)\,(30-3)} \frac{(39-56)^4 + (66-56)^4 + \cdots + (63-56)^4}{7,62^4} - \frac{3\,(30-1)}{(30-2)\,(30-3)} = 0,1636$$

Charakter rozloženia údajov hmotností stredoškolákov vykazuje v porovnaní s normálnym rozložením vlastnosti špicatého rozloženia s hodnotou koeficientu špicatosti 0,1636.

Špicatosť rozloženia údajov, napríklad pri porovnávaní dvoch či viacerých súborov údajov je možné pozorovať aj pomocou krabicových grafov. Príklad špicatého, normálneho a plochého rozloženia v krabicových grafoch je zobrazený na obrázku 3.10.



Obr. 3.10: Krabicové grafy normálneho (modrý), špicatého (zelený) a plochého (červený) rozdelenia.

Modrý krabicový graf na obrázku 3.10 znázorňuje normálne rozdelenie, ktoré slúži ako referencia pre určenie špicatosti rozloženia. Červený krabicový graf reprezentuje ploché rozdelenie, u ktorého je vplyvom väčšej variability údajov medzikvartilové rozpätie širšie ako medzikvartilové rozpätie normálneho rozdelenia. Variačné rozpätie hodnôt, dané minimálnou a maximálnou hodnotou je

taktiež väčšie, čo je prezentované dĺžkou fúzov po oboch stranách krabicového grafu. Analogicky je možné interpretovať krabicový graf pre špicaté rozdelenie (na obrázku 3.10 zobrazené zelenou farbou), kedy je medzikvartilové rozpätie hodnôt užšie. Rovnako je užšie aj celkové variačné rozpätie v porovnaní s normálnym rozdelením. Krabicové grafy na obrázku sú pre jednoduchšiu interpretáciu symetrické a s rovnakou strednou hodnotou. V praktických úlohách však tieto hodnoty môžu byť iné, napríklad pri porovnávaní priemernej výšky mužov a žien, porovnávaní hodnôt cholesterolu u rôznych skupín pacientov a pod.

Funkcie v MS Excel

KURT(číslo1; číslo2; ...)
Funkcia KURT vráti hodnotu špicatosti množiny údajov.

Kapitola 4

Triedenie údajov

Neoddeliteľnou súčasťou popisnej štatistiky je organizácia údajov spôsobom, ktorý nám pomáha nájsť, resp. odhaliť informácie súvisiace s predmetom daného výskumu (štúdie). Dôležité je preto získané hromadné údaje, ktoré sú často neprehľadné, vedieť jednak správne zotriediť a následne aj vhodne prezentovať, napríklad pomocou tabuliek (tabuľka početností, tabuľka relatívnych početností, tabuľka kumulatívnych početností, tabuľka kumulatívnych relatívnych početností) a grafov (histogram, polygón, koláčový graf a pod.). Cieľom je, aby takýmto triedením údajov vynikli charakteristické črty spracovávaného súboru údajov.

Triedením údajov zabezpečujeme ich usporiadanie a „zhustenie“ do skupín údajov, tzv. tried. Aby prezentácia usporiadaných údajov zodpovedala získaným údajom musíme pri ich triedení zabezpečiť požiadavky na:

- **úplnosť** – každá hodnota musí byť zahrnutá do triedenia,
- **jednoznačnosť** – každá hodnota môže byť priradená len jednej triede.

Spravidla je pri triedení usporadúvaná skupina údajov jednej skúmanej veličiny. Avšak, niekedy sú v štúdiách triedené kombinácie hodnôt viacerých veličín. Podľa toho je možné metódy triedenia rozdeliť na:

- **jednostupňové** – triedenie, v ktorom triedime hodnoty len jednej skúmanej veličiny (napríklad hodnoty cholesterolu),
- **viacstupňové** – triedenie, v rámci ktorého triedime hodnoty súčasne podľa viacerých skúmaných veličín (napríklad pri dvojstupňovom triedení podľa veku a cholesterolu, podľa krvnej skupiny a Rh faktora a pod.).

Údaje, ktoré sú zhromažďované o rôznych skúmaných veličinách vo výberových súboroch reprezentujú viac či menej presne údaje z definovaného základného súboru. Môžu byť tak kvalitatívne, ako aj kvantitatívne, t. j. dané možnými hodnotami alebo oblasťami hodnôt, ktoré môžu v rámci skúmanej veličiny

nastať. Podľa toho tiež vytvárame jednotlivé triedy, resp. určujeme hodnoty, ktoré budú tvoriť konkrétnu triedu. Charakter definovaných tried preto určuje dve hlavné skupiny triedenia údajov:

- **jednoduché** – triedenie využívané pre diskrétne hodnoty, za predpokladu, že súbor údajov obsahuje menší počet možných vyskytujúcich sa hodnôt. Každá hodnota tvorí jednu samostatnú triedu,
- **intervalové** – triedenie využívané pre spojité hodnoty, kedy súbory údajov obsahujú veľké množstvo vyskytujúcich sa hodnôt. Jedna trieda je tvorená viacerými hodnotami (intervalom hodnôt).

Proces triedenia údajov začína voľbou spôsobu ich triedenia, ktoré je závislé od typu údajov, ktoré majú byť triedené a definovaním zoznamu tried, do ktorých budú údaje zatriedené.

4.1 Jednoduché triedenie

Usporiadanie údajov pomocou jednoduchého triedenia je výpočtovo relatívne nenáročné. V mnohých prípadoch menších výberových súborov je ho možné rýchlo vytvoriť aj ručne (to však samozrejme nie je v dnešnej dobe výpočtových prostriedkov potrebné). Jednoduché triedenie je možné použiť tak u kvalitatívnych, ako aj u kvantitatívnych veličín s malým počtom odlišných hodnôt, pričom každá takáto hodnota potom tvorí samostatnú triedu. V prípade kvantitatívnych veličín sa spravidla jedná o súbory diskrétnych veličín, ktorých hodnoty nadobúdajú len celočíselné hodnoty.

U kvalitatívnych nominálnych veličín, môže byť poradie tried náhodné, t. j. nezáleží na poradí v akom budú prezentované (tabuľkovo či graficky). Príklad tried dvoch kvalitatívnych nominálnych veličín je uvedený v tabuľke 4.1

Tabuľka 4.1: Triedy dvoch kvalitatívnych nominálnych veličín.

Veličina	**Triedy**	**Veličina**	**Triedy**
krvná skupina	A B AB 0	farba očí	hnedá modrá čierna zelená sivá

Kvalitatívne veličiny je možné na uľahčenie zaznamenávania a triedenia hodnôt kódovať, napríklad priradením čísel k jednotlivým kategóriám. Textové

údaje sa tak prevedú na číselné údaje v poradovom zmysle. Napríklad možnosti rodinného stavu by sme mohli kódovať tak, že číslo 1 bude označovať stav slobodný(á), 2 ženatý/vydatá, 3 rozvedený(á) a 4 vdovec/vdova. Avšak, je potrebné si uvedomiť, že kódované kvalitatívne nominálne údaje stále zostávajú nominálnymi údajmi, ktoré nemajú charakter štandardných čísel ako ich poznáme a používame v bežnej aritmetike. To znamená, že ich nie je možné porovnávať, spočítavať či odpočítavať ako číselné hodnoty (sú len kódmi kvalitatívnych hodnôt). Tento prípad poukazuje na to, aké je vždy dôležité kontrolovať, či matematické spracovanie štatistických údajov je skutočne legitímne.

Triedy kvalitatívnych ordinálnych veličín a kvantitatívnych diskrétnych veličín by mali byť vždy prezentované v usporiadanom vzostupnom poradí. Pre každú triedu následne určujeme početnosť výskytu jej hodnoty v súbore spracovávaných údajov.

4.1.1 Početnosť

Početnosť (frekvencia) je pojem, ktorým v štatistickom spracovávaní údajov označujeme absolútnu početnosť, t. j. počet všetkých výskytov danej hodnoty v skupine údajov. Početnosť konkrétnej hodnoty vo výberovom súbore teda predstavuje informáciu o tom, koľkokrát sa táto hodnota vo výberovom súbore nachádza.

Napríklad, ak sa v súbore 100 pacientov bude nachádzať 19 pacientov s krvnou skupinou A, potom túto hodnotu nie je potrebné zapisovať 19-krát, ale použijeme početnosť krvnej skupiny A a uvedieme hodnotu 19. Podobne uvedieme hodnoty výskytu ostatných krvných skupín a namiesto prezentácie zoznamu 100 samostatných hodnôt budeme prezentovať len 4 (4 krvné skupiny). To znamená, že takto možno sumarizovať i vizualizovať rozsiahle súbory údajov jednoduchšie a prehľadnejšie.

Predpokladajme, že máme súbor hodnôt s rozsahom n, v ktorom boli zaznamenané hodnoty $x_1, x_2, \cdots, x_n$, pričom počet možností k, ktoré môžu hodnoty x_i nadobúdať je malý (napríklad hodnoty diskrétnej veličiny). Potom vieme zostaviť vzostupnú postupnosť všetkých možností $x_1^* < x_2^* < \cdots < x_k^*$, ktoré výberový súbor obsahuje, prípadne môže obsahovať, ak je známe aké možnosti sa môžu nachádzať v základnom súbore (napríklad uvedieme aj krvnú skupinu B hoci ju nikto v našom výberovom súbore nemal). Takáto postupnosť zároveň definuje triedy jednoduchého triedenia. Následne je možné spočítavať koľkokrát sa nachádza hodnota x_1^* v súbore hodnôt, koľkokrát hodnota x_2^*, koľkokrát hodnota x_3^* atď., až po poslednú možnosť, t. j. hodnotu x_k^*.

Zistené hodnoty početností n_i jednotlivých hodnôt x_i^* zapisujeme do tabuľky početností, ktorá je vlastne zoznamom hodnôt (tried) vyskytujúcich sa v skupine údajov usporiadaných vzostupne (prvý stĺpec) s ich zodpovedajúcimi absolútnymi početnosťami (druhý stĺpec), tak ako je to uvedené v tabuľke 4.2

Tabuľka 4.2: Tabuľka početností hodnôt diskrétnej veličiny.

Trieda	**Početnosť**
x_i^*	n_i
x_1^*	n_1
x_2^*	n_2
$\vdots$	$\vdots$
x_k^*	n_k

Početnosť n_i označuje koľkokrát sa hodnota x_i^* nachádza v súbore hodnôt, pričom platí, že:

$$\sum_{i=1}^{k} n_i = n \tag{4.1}$$

Inými slovami povedané, suma všetkých početností sa musí rovnať rozsahu súboru hodnôt, t. j. $n_1 + n_2 + \cdots + n_k = n$. Ak je hodnota sumy menšia ako n, potom do triedenia neboli započítané všetky hodnoty a bola porušená podmienka úplnosti triedenia údajov. Naopak, ak je hodnota sumy väčšia ako n, potom do triedenia bola(i) niektorá(é) hodnota(y) započítaná(é) viac ako raz, čím bola porušená podmienka jednoznačnosti triedenia údajov.

Príklad 4.1. Predpokladajme, že na oddelení intenzívnej starostlivosti boli zaznamenávané spoločne s ďalšími údajmi o zdravotnom stave pacienta aj informácie o stupni porúch vedomia pacientov podľa Glasgowskej škály porúch vedomia (*Glasgow Coma Scale, GCS*). V hodnoteniach slovných reakcií, ktoré môžu nadobúdať tieto možnosti: 1 - žiadna (pacient nevydá ani zvuk), 2 - nezrozumiteľná (nezrozumiteľné slová a zvuky), 3 - neprimeraná (neprimerané výrazy), 4 - zmätená (dezorientovaná) a 5 - orientovaná (zmysluplný rozhovor), boli zaznamenané nasledovné hodnoty hospitalizovaných pacientov: 4, 1, 3, 2, 3, 3, 1, 3, 1, 4, 2, 1, 3, 3, 1, 2, 2, 1, 5, 1, 5, 1, 3, 2, 1, 4, 1 a 2. Ako súbor hodnôt (vybraný z množstva iných sledovaných veličín) nám takto zapísané údaje veľa informácií bez detailnejšej analýzy neposkytujú. Popíšme ich teda pomocou tried a početností.

Ako prvú informáciu, ktorú vieme určiť spočítaním hodnôt je veľkosť (rozsah) súboru. V tomto prípade je to zoznam 28 hodnôt získaných od 28 pacientov. Rovnako vieme identifikovať jednotlivé triedy, keďže na určenie slovnej reakcie máme daných len päť možností (diskrétnych): 1, 2, 3, 4 a 5 (prípadne slovný popis, poradie je dôležité). Z tohoto dôvodu môžeme použiť jednoduché triedenie, v ktorom každá možnosť slovnej reakcie bude predstavovať jednu samostatnú triedu. Získané hodnoty je možné zoradiť do vzostupnej postupnosti, v ktorej budú všetky rovnaké hodnoty vedľa seba: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5. Zostrojíme teda tabuľku početností, do ktorej spočítaním zapíšeme počty výskytov jednotlivých tried.

Tabuľka 4.3: Tabuľka početností hodnôt slovných reakcií podľa GCS.

Slovná reakcia (x_i^*)	**Početnosť** (n_i)
1 - žiadna	10
2 - nezrozumiteľná	6
3 - neprimeraná	7
4 - zmätená	3
5 - orientovaná	2
Spolu	28

Z tabuľky početností je vidieť, že na oddelení intenzívnej starostlivosti je najviac pacientov (10) takých, ktorí nemajú žiadnu slovnú reakciu. 6 pacientov má nezrozumiteľnú reakciu, 7 neprimeranú, 3 zmätenú a 2 orientovanú. Súčet týchto početností je rovný 28, čo sa zhoduje s veľkosťou súboru a uisťuje nás v tom, že sme všetky hodnoty zahrnuli do triedenia.

Funkcie v MS Excel

COUNTIF(rozsah; kritériá)
Funkcia COUNTIF spočíta bunky v danom rozsahu, ktoré spĺňajú danú podmienku.

FREQUENCY(údajové_pole; binárne_pole)
Funkcia FREQUENCY vytvorí tabuľku početností z hodnôt v údajovom poli podľa tried uvedených v binárnom poli, t. j. zvislé pole čísel s počtami výskytov hodnôt v jednotlivých rozsahoch.

4.1.2 Relatívna početnosť

Pri analýzach individuálnych súborov, alebo pri porovnávaní viacerých súborov hodnôt nás často zaujíma v akom pomere sú hodnoty danej triedy zastúpené voči ostatným hodnotám. Túto informáciu získavame z relatívnej početnosti,

ktorá predstavuje podiel výskytu (početností tried) n_i jednotlivých hodnôt x_i^* a celkového rozsahu súboru n vypočítanej podľa vzťahu (4.2):

$$p_i = \frac{n_i}{n} \tag{4.2}$$

pričom platí, že $0 \leq p_i \leq 1$. Ak $p_i = 0$, potom sa hodnota (trieda) x_i^* v súbore všetkých hodnôt nenachádza, a ak je $p_i = 1$, potom súbor všetkých hodnôt obsahuje len hodnoty x_i^*.

Z hodnôt p_i vypočítaných pre jednotlivé triedy môžeme následne vytvoriť samostatnú tabuľku relatívnych početností, rovnako ako v prípade (absolútnych) početností (tabuľka 4.2). Kvôli komplexnému prehľadu sa však častejšie relatívne početnosti prezentujú v spoločnej tabuľke s početnosťami, tak ako je to uvedené v tabuľke 4.4.

Tabuľka 4.4: Tabuľka početností doplnená o relatívne početnosti.

Trieda x_i^*	**Početnosť** n_i	**Relatívna početnosť** p_i
x_1^*	n_1	p_1
x_2^*	n_2	p_2
$\vdots$	$\vdots$	
x_k^*	n_k	p_k

Pre relatívne početnosti platí:

$$\sum_{i=1}^{k} p_i = 1 \tag{4.3}$$

Vzťah (4.3) hovorí, že suma všetkých relatívnych početností musí byť vždy rovná hodnote 1, t. j. $p_1 + p_2 + \cdots + p_k = 1$. Ak je výsledok sumy relatívnych početností menší ako 1, alebo väčší ako 1, potom sme pri výpočte hodnôt relatívnych početností, alebo hodnôt početností, z ktorých boli relatívne početnosti odvodené, urobili chybu a je potrebné všetky hodnoty skontrolovať, prípadne prepočítať znova.

Hodnoty relatívnych početností je možné vyjadriť aj v percentách veľkosti súboru, a to tak, že ich vynásobíme hodnotou 100. Relatívne početnosti, ako aj relatívne početnosti v percentách poskytujú najdôležitejšie informácie o vzore údajov. Pri prezentácii výsledkov je tiež potrebné uvádzať veľkosť súboru, ktorá slúži ako ukazovateľ hodnovernosti relatívnych početností. Súčet všetkých relatívnych početností v percentách musí byť vždy rovný 100%.

Pre údaje z príkladu 4.1 vypočítame relatívne početnosti jednotlivých tried tak, že hodnoty početností vydelíme číslom 28, keďže všetkých hodnôt zahrnutých do triedenia údajov bolo 28. Výsledná tabuľka s relatívnymi početnosťami je uvedená ako tabuľka 4.5.

Tabuľka 4.5: Tabuľka absolútnych a relatívnych početností hodnôt slovných reakcií pacientov podľa GCS.

Slovná reakcia (x_i^*)	**Početnosť** (n_i)	**Relatívna početnosť** (p_i)
1 - žiadna	10	0,3572
2 - nezrozumiteľná	6	0,2143
3 - neprimeraná	7	0,2500
4 - zmätená	3	0,1071
5 - orientovaná	2	0,0714
Spolu	28	1

Z hodnôt relatívnych početností je možné jednoducho vyčítať, aký je podiel pacientov v jednotlivých skupinách slovných reakcií na oddelení intenzívnej starostlivosti. Vyjadrením v percentách (vynásobením hodnotou 100) vieme, že na oddelení je 35,72% pacientov, ktorí nemajú žiadnu slovnú reakciu, 21,43% pacientov má nezrozumiteľnú reakciu, 25% neprimeranú, 10,71% zmätenú a 7,14% orientovanú. Hodnoty v % je možné zaokrúhliť na celé čísla.

4.1.3 Kumulatívne početnosti

Pri hodnotení údajov je možné tiež nájsť početnosť do určitej hodnoty (triedy), ktorá sa nazýva kumulatívna početnosť. Kumulatívnu početnosť získavame sčítaním početností všetkých tried až po konkrétnu triedu. V prípade kvalitatívnych premenných majú kumulatívne početnosti zmysel len pre ordinálne premenné, nie pre nominálne premenné.

Kumulatívnu početnosť N_i triedy i vypočítame podľa vzťahu (4.4):

$$N_i = \sum_{j=1}^{i} n_j \tag{4.4}$$

kde i reprezentuje poradie triedy, pričom $i = 1, 2, \cdots, k$.

Kumulatívna početnosť prvej triedy je tak rovná početnosti prvej triedy. Kumulatívna početnosť druhej triedy je rovná súčtu početností prvej a druhej

triedy, kumulatívna početnosť tretej triedy je rovná súčtu početností prvej až tretej triedy, atď. Kumulatívna početnosť teda udáva aká časť súboru je menšia alebo rovná ako príslušná hodnota.

Podobne ako početnosť je možné kumulovať aj relatívnu početnosť. Kumulatívnu relatívnu početnosť získavame sčítaním relatívnych početností všetkých tried až po konkrétnu triedu.

Kumulatívnu relatívnu početnosť P_i triedy i vypočítame podľa vzťahu (4.5):

$$P_i = \sum_{j=1}^{i} p_j \tag{4.5}$$

kde i reprezentuje poradie danej triedy z rozsahu $i = 1, 2, \cdots, k$.

Kumulatívna relatívna početnosť prvej triedy sa preto rovná relatívnej početnosti prvej triedy. Kumulatívna relatívna početnosť druhej triedy je rovná súčtu relatívnych početností prvej a druhej triedy, kumulatívna relatívna početnosť tretej triedy je rovná súčtu relatívnych početností prvej, druhej a tretej triedy, atď.

Kumulatívne početnosti a relatívne kumulatívne početnosti bývajú súčasťou prezentácie triedených údajov v spoločnej tabuľke s početnosťami a relatívnymi početnosťami, tak ako je to uvedené v tabuľke 4.6.

Tabuľka 4.6: Tabuľka početností doplnená o relatívne a kumulatívne početnosti.

Trieda x_i^*	**Početnosť** n_i	**Relatívna početnosť** p_i	**Kumulatívna početnosť** N_i	**Kumulatívna relatívna početnosť** P_i
x_1^*	n_1	p_1	n_1	p_1
x_2^*	n_2	p_2	$n_1 + n_2$	$p_1 + p_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_k^*	n_k	p_k	$n_1 + \cdots + n_k$	$p_1 + \cdots + p_k$

Pre kumulatívne početnosti platí, že ich posledná hodnota v poslednej triede je rovná rozsahu súboru n, kedy boli akumulované všetky hodnoty súboru. Hodnota kumulatívnej početnosti rovná hodnote n sa môže nachádzať aj v predošlých triedach, ale len v prípade, ak v nasledujúcich triedach budú početnosti nulové, t. j. hodnoty patriace do nasledujúcich tried sa v súbore hodnôt nenachádzajú. Analogicky, poslednou hodnotou v poslednej triede kumulatívnych relatívnych početností je hodnota 1, ktorá značí, že do tejto a predošlých tried boli zaradené všetky hodnoty súboru. Aj tu sa hodnota kumulatívnej relatívnej

početnosti 1 môže vyskytnúť v predošlých triedach, a to v prípadoch, kedy do nasledujúcich tried nebola priradená žiadna hodnota. V ďalších triedach teda bude pripočítavaná už len hodnota 0, ktorá kumulatívnu relatívnu početnosť už nebude zvyšovať.

Kumulatívne početnosti a kumulatívne relatívne početnosti údajov z príkladu 4.1 sú uvedené v doplnenej tabuľke početností 4.7.

Tabuľka 4.7: Tabuľka absolútnych, relatívnych a kumulatívnych početností hodnôt slovných reakcií pacientov podľa GCS.

Slovná reakcia (x_i^*)	**Početnosť** (n_i)	**Relatívna početnosť** (p_i)	**Kumulatívna početnosť** (N_i)	**Kumulatívna relatívna početnosť** (P_i)
1 - žiadna	10	0,3572	10	0,3572
2 - nezrozumiteľná	6	0,2143	16	0,5715
3 - neprimeraná	7	0,2500	23	0,8215
4 - zmätená	3	0,1071	26	0,9286
5 - orientovaná	2	0,0714	28	1
Spolu	28	1	-	-

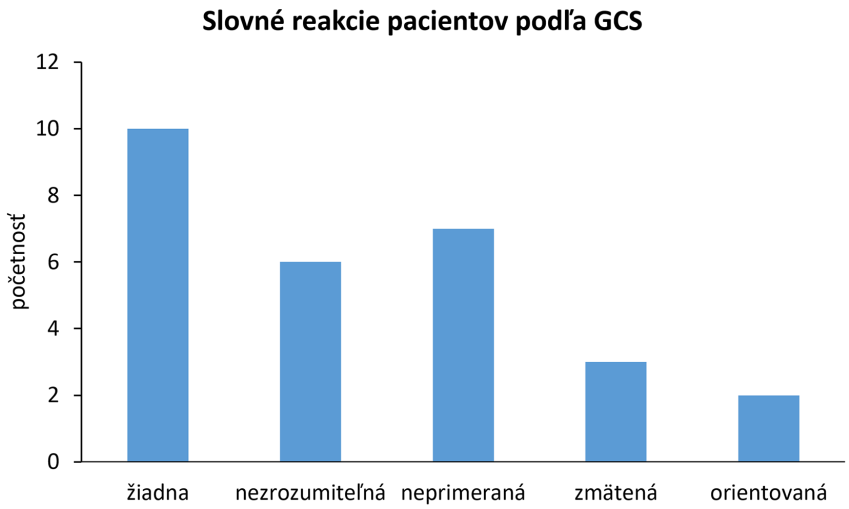
Z kumulatívnych hodnôt je teda možné hodnotiť, že na oddelení je 10 pacientov, ktorí nemajú žiadnu slovnú reakciu, čo predstavuje 35,72%. Prípadne, že na oddelení je 16 pacientov, ktorí nemajú žiadnu slovnú reakciu alebo majú nezrozumiteľnú reakciu, čo predstavuje 57,15%. Tiež, že na oddelení je 23 pacientov, ktorí nemajú žiadnu slovnú reakciu, majú nezrozumiteľnú reakciu alebo neprimeranú reakciu, čo predstavuje 82,15% pacientov, atď.

4.1.4 Grafická prezentácia početností

Vo všeobecnosti je prezentácia údajov pomocou grafov najlepším prostriedkom pre pochopenie charakteru údajov v súbore a tiež súvislostí medzi prezentovanými údajmi, keďže v malom vizuálnom priestore je spravidla zhrnutých viacero najdôležitejších štatistických charakteristík. Grafické zobrazenie je teda v porovnaní so záznamami získaných hodnôt alebo tabuľkami, v ktorých sú tieto hodnoty zozbierané nielen prehľadné, ale aj názorné. Navyše, ak je použitá vhodná grafická prezentácia, môže byť použitá aj samostatne. Existuje veľa typov grafov, avšak na prezentáciu údajov je potrebné vybrať taký, ktorý odpovedá charakteru údajov. Najčastejšie sa využíva pravouhlá súradnicová sústava xy, t. j. hodnoty sú prezentované ako xy závisloť, kedy na x-ovú os (vodorovnú) vynášame hodnoty sledovanej veličiny (triedy, resp. hodnoty v triedach) a na y-

ovú os (zvislú) príslušné hodnoty početností alebo relatívnych početností. Údaje kvalitatívnych alebo kvantitatívnych diskrétnych veličín, tak ako ich triedime v rámci tejto kapitoly, sú najčastejšie prezentované stĺpcovým alebo kruhovým (koláčovým) grafom. Vhodné bývajú aj polygón a bodový diagram.

Stĺpcový graf predstavuje sériu oddelených stĺpcov, ktorých počet sa rovná počtu tried zotriedeného súboru údajov a ich výška zodpovedá hodnote početnosti alebo relatívnej početnosti. Stĺpce majú rovnakú šírku a sú od seba oddelené medzerou. Ak použijeme ako zdroj údajov stĺpcového grafu hodnoty početností z tabuľky 4.7, potom výsledkom grafickej prezentácie slovných reakcií pacientov bude graf zobrazený na obrázku 4.1.

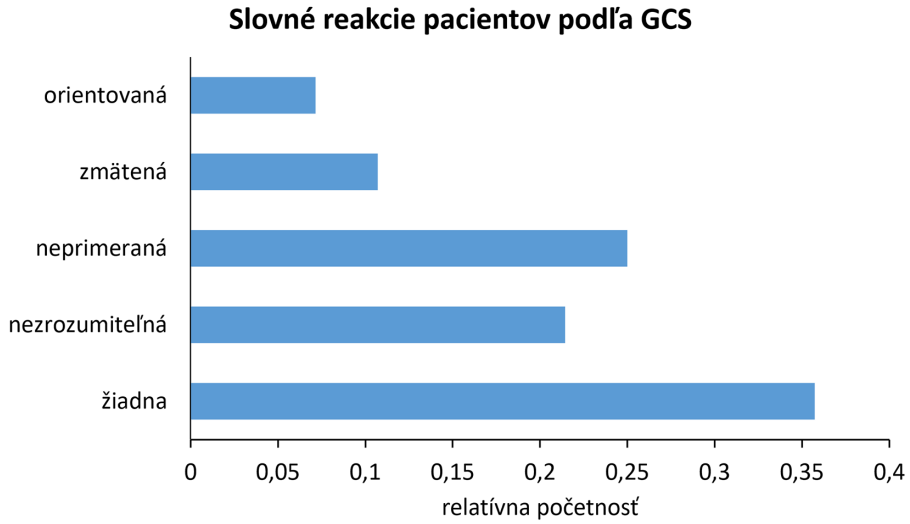


Obr. 4.1: Stĺpcový graf početností slovných reakcií pacientov.

Stĺpcový graf nám poskytuje veľa, pre interpretáciu a hodnotenie údajov užitočných informácií, a to aj v prípadoch kedy nemáme dostupné zdrojové údaje. Štatistické charakteristiky, ktoré vieme z takejto grafickej prezentácie odčítať, závisia od typu prezentovaných údajov a toho, či reprezentujú kvalitatívne alebo kvantitatívne veličiny. Z grafu 4.1 je čitateľovi zrejmé, že prezentujeme kvalitatívne ordinálne údaje rozdelené do piatich tried. Spočítaním výšok jednotlivých stĺpcov zisťujeme, že veľkosť súboru je 28. Modusom je hodnota slovnej reakcie „žiadna“, ktorá sa vyskytuje najčastejšie (najvyšší stĺpec). Mediánom hodnôt nemusí byť zákonite prostredný stĺpec, keďže stĺpce nereprezentujú hodnoty, ale ich početnosti. Ak vieme, že máme 28 hodnôt, potom prostrednú hodnotu (medián) definujú 14. a 15. hodnota. V prvom stĺpci je desať hodnôt, preto sa tu medián nenachádza. V druhom stĺpci je 6 hodnôt, čo spolu s prvým stĺpcom predstavuje 16 hodnôt súboru. Môžeme teda povedať, že 14. a 15. hodnota sa nachádza v druhom stĺpci, a teda mediánom rozdelenia je slovná reakcia „nezrozumiteľná“. Z pohľadu symetrie rozloženia vidíme

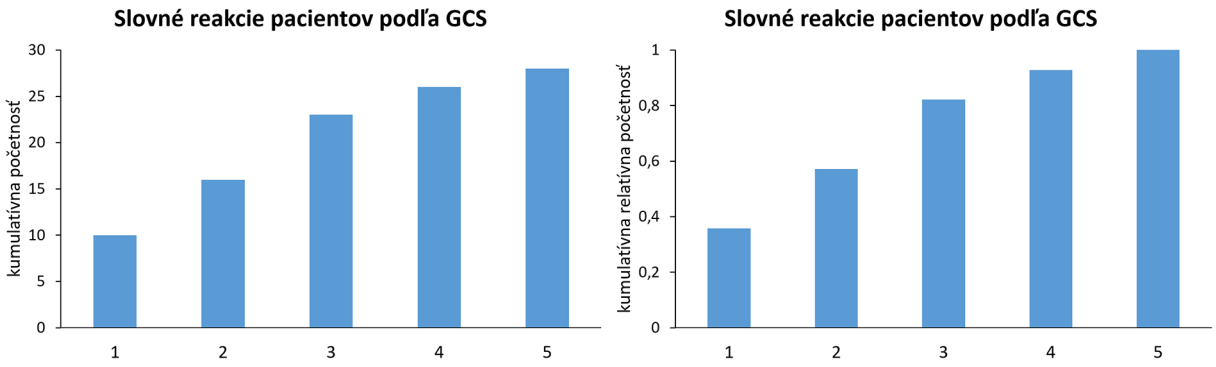
pozitívnu šikmosť. V prípade kvantitatívnych údajov by sme ďalej vedeli určiť minimálnu a maximálnu hodnotu, variačné rozpätie, priemer hodnôt podľa váženého priemeru atď.

Grafickou prezentáciou relatívnych početností získavame identickú interpretáciu údajov, avšak v inej škále. Stĺpcové grafy je možné prezentovať vertikálne alebo horizontálne. Graf na obrázku 4.2 prezentuje tie isté informácie s hodnotami relatívnych početností a stĺpcami umiestnenými horizontálne, t. j. triedy údajov sú na vertikálnej osi a relatívne početnosti na horizontálnej osi.



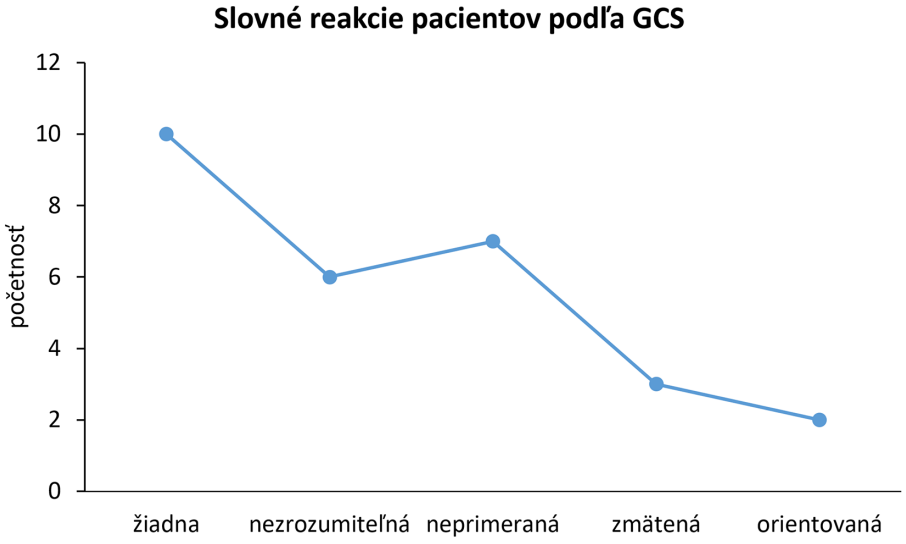
Obr. 4.2: Stĺpcový graf relatívnych početností slovných reakcií pacientov.

V prípadoch, kedy si to charakter údajov vyžaduje a pre posudzovanie výsledkov je dôležitá informácia o akumulovaných údajoch, do stĺpcových grafov vynášame hodnoty kumulatívnych početností alebo relatívnych kumulatívnych početností. Príkladom takéhoto zobrazenia je obrázok 4.3, ktorý taktiež prezentuje informácie z príkladu 4.1.



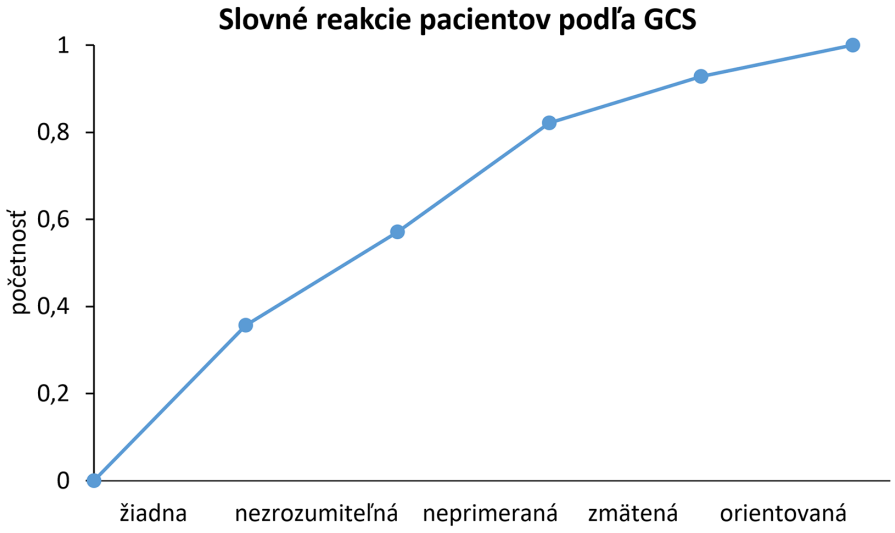
Obr. 4.3: Stĺpcový graf kumulatívnych početností (vľavo) a kumulatívnych relatívnych početností (vpravo) slovných reakcií pacientov.

Polygón početností a relatívnych početností predstavuje ďalší graf, ktorý je vhodný pre zobrazovanie údajov z jednoduchého triedenia. Je to graf, v ktorom úsečkami spájame vrcholy stĺpcového grafu, t. j. hodnoty početností alebo hodnoty relatívnych početností. Obrázok 4.4 prezentuje polygón početností.



Obr. 4.4: Polygón početností slovných reakcií pacientov.

Z pohľadu riešení úloh induktívnej štatistiky je zaujímavý aj polygón kumulatívnych početností a kumulatívnych relatívnych početností, z ktorého vieme identifikovať napríklad pravdepodobnosti pre výpočet testovaných charakteristík (bude diskutované v ďalších kapitolách).



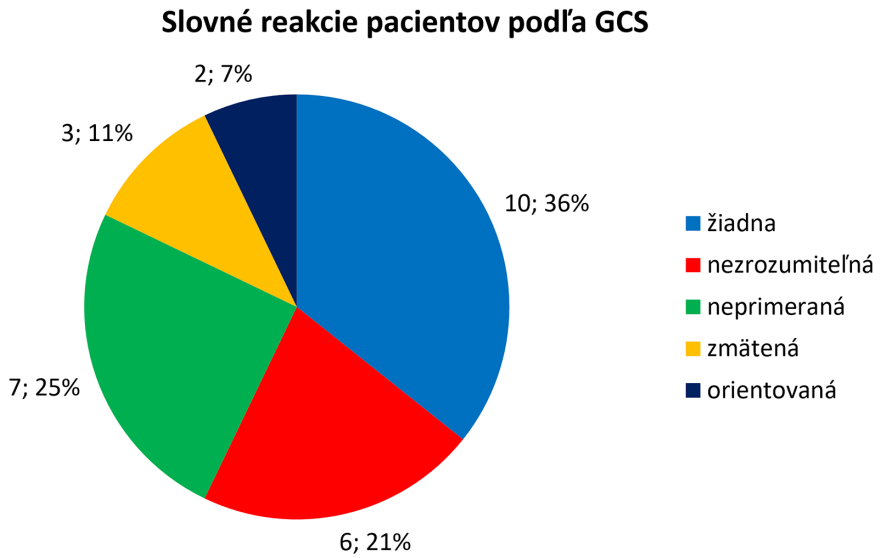
Obr. 4.5: Polygón kumulatívnych relatívnych početností slovných reakcií pacientov.

Polygón kumulatívnych relatívnych početností vykresľujeme so začiatkom v nule, napríklad tak ako je to znázornené na obrázku 4.5. Tvar kumulovaného

rozloženia údajov vynikne s väčším počtom definovaných tried a pre kvantitatívne veličiny. Pri kumulatívnych charakteristikách je dôležité poznamenať, že sa jedná vždy o neklesajúce charakteristiky, tak ako sú definované aj distribučné funkcie rôznych rozdelení.

Kruhový graf (koláčový) je často využívaným grafom, v ktorom je možné identifikovať pomerové charakteristiky zastúpených hodnôt (možností) sledovanej veličiny. Základom kruhového grafu je kruh, ktorý je delený na výseče pripadajúce jednotlivým triedam údajov. Zastúpenie hodnôt v percentách je definované tak, že na 1% údajov pripadá výseč s uhlom 3,6°(100% = 360°).

Kruhový graf umožňuje na rovnako veľkom priestore prehľadné zobrazovanie tak malých, ako aj veľkých súborov. Rozhodujúci je počet tried, do ktorých sú údaje z týchto súborov zatriedené. Veľkosť výseče zároveň umožňuje identifikovať zastúpenie danej triedy v celom súbore údajov. Príklad kruhového grafu je zobrazený na obrázku 4.6.



Obr. 4.6: Kruhový graf početností slovných reakcií pacientov.

4.2 Intervalové triedenie

Jednoduché triedenie je vhodné iba v prípadoch, kedy máme v súboroch iba malý počet možností, ktoré môžu hodnoty sledovanej veličiny nadobúdať. Avšak, v prípadoch spojitých kvantitatívnych veličín a diskrétnych kvantitatívnych veličín s veľkým počtom možností ho nepoužívame. Rezultovalo by totiž do veľkého a neprehľadného počtu tried a ich veľmi nízkych alebo aj nulových

početností. Aby triedenie takýchto spojitých hodnôt bolo zmysluplné, využívame intervalové triedenie.

Všeobecná postupnosť krokov zoskupovania spojitých kvantitatívnych veličín do intervalových tried je nasledovná:

- nájdeme najmenšiu a najväčšiu hodnotu v súbore údajov,
- stanovíme intervaly rovnakej dĺžky, tak aby pokrývali rozsah hodnôt medzi minimom a maximom bez prekrývania,
- spočítame počet hodnôt v súbore, ktoré patria do jednotlivých triednych intervalov (početnosť v danej triede),
- vypočítame relatívne a kumulatívne početnosti jednotlivých tried.

4.2.1 Určovanie intervalov

Z vyššie uvedených informácií vyplýva, že hlavný rozdiel medzi jednoduchým a intervalovým triedením spočíva definovaní skupín hodnôt, podľa ktorých budú údaje zoskupované. Kým u jednoduchého triedenia každá hodnota predstavovala jednu samostatnú triedu (skupinu), u intervalového triedenia tvorí jednu triedu viacero po sebe nasledujúcich hodnôt spadajúcich do definovaného intervalu. Až po definovaní intervalov triedenia pokrývajúcich celý rozsah hodnôt súboru bude možné vytvoriť tabuľku početností.

Predpokladajme, že máme súbor hodnôt spojitej veličiny s rozsahom n. V danom súbore vieme určiť minimálnu a maximálnu hodnotu a z týchto hodnôt vypočítame variačné rozpätie $R = x_{max} - x_{min}$. Variačné rozpätie je teraz potrebné rozdeliť na niekoľko rovnako veľkých intervalov. Ich počet však závisí od rozsahu súboru, keďže pre relatívne malé súbory nám bude postačovať menej tried a naopak pre väčšie súbory bude žiadúce mať väčší počet tried. Vo všeobecnosti je bežné zoskupovať údaje do 5 až 15 triednych intervalov. Avšak, pre veľmi veľké súbory môže byť počet intervalov aj väčší ako 15.

Počet odporúčaných tried k pre konkrétny rozsah súboru n určíme podľa Sturgesovho pravidla (4.6):

$$k \doteq 1 + 3,322 log_{10} n \tag{4.6}$$

pričom výslednú hodnotu zaokrúhľujeme na celé číslo. Napríklad, ak by bola veľkosť súboru 30, potom by bol odporúčaný počet tried:

$$k = 1 + 3,322 log_{10} 30 = 5,9069 \doteq 6$$

t. j. rozsah hodnôt súboru by sme rozdelili na 6 rovnakých intervalov.

Šírku jednotlivých tried (intervalov) h určíme vydelením variačného rozpätia súboru R a vypočítaného počtu intervalov k. Potom každý z k intervalov bude mať rovnakú šírku:

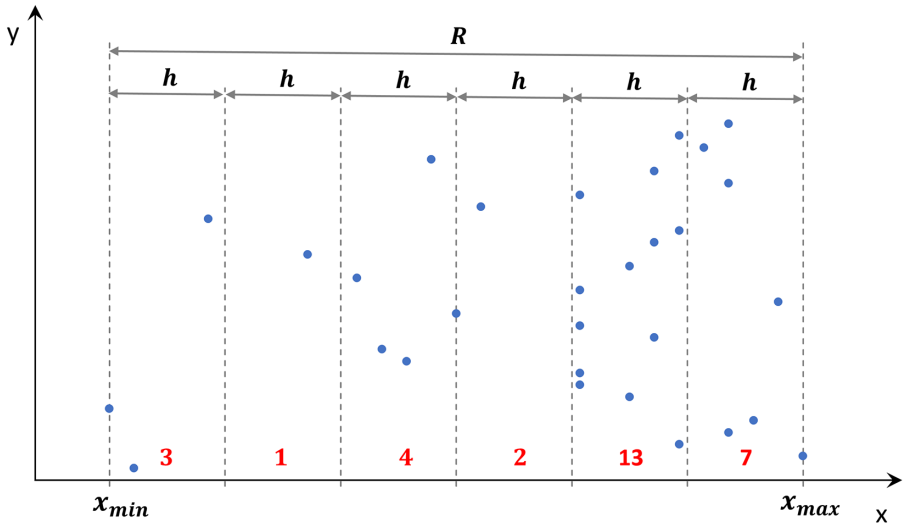
$$h \cong \frac{R}{k} \cong \frac{x_{max} - x_{min}}{k} \tag{4.7}$$

Hranice jednotlivých intervalov získavame postupným pripočítavaním šírky intervalu h k minimálnej hodnote x_{min} až kým nedosiahneme maximálnu hodnotu x_{max}. Prvý interval teda bude obsahovať minimálnu hodnotu a posledný maximálnu, čím zabezpečíme, že všetky hodnoty súboru spadnú do niektorého z intervalov. Potom hranice intervalov budú definované takto:

$$\langle x_{min}; x_{min} + h\rangle \quad (x_{min} + h; x_{min} + 2h\rangle \cdots (x_{min} + (k-1)h; x_{min} + kh\rangle$$

pričom $x_{min} + kh = x_{max}$.

Príklad rozdelenia variačného rozpätia R na 6 rovnakých intervalov (k) so šírkou h je znázornený na obrázku 4.7.



Obr. 4.7: Rozdelenie hodnôt do triednych intervalov s rovnakou šírkou.

Modré body na obrázku 4.7 reprezentujú hmotnosti 30 stredoškolákov z príkladu 3.1, pričom na x-ovej osi je znázornená hmotnosť v kg a na y-osi sú poradové čísla stredoškolákov, tak ako boli zahrnutí do výberového súboru. Rozdiel medzi maximálnou (66 kg) a minimálnou (38 kg) hmotnosťou bol 28 kg, ktorý pri rozdelení na 6 intervalov predstavuje šírku jedného intervalu o veľkosti cca 4,6667 kg. Začínajúc od hodnoty 38 kg boli horné hranice triednych intervalov tvorené hodnotami po zaokrúhlení na dve desatinné miesta: 42,67; 47,34; 52,01;

56,68; 61,35 a 66,02 kg. Červené čísla znázorňujú početnosti hmotností stredoškolákov v danom intervale. Z pohľadu symetrie rozloženia je aj tu vidieť, že údaje sú negatívne zošikmené.

Je potrebné zdôrazniť, že stanovanie počtu tried závisí jednak od veľkosti súboru, ale aj od typu triedených údajov. Niekedy sú spracovávané údaje, u ktorých je potrebné poznať len to či sa nachádzajú v intervale normálnych (štandardných) hodnôt, alebo sú menšie či väčšie (patologické) a podľa toho zabezpečiť liečbu pacientov. V tomto prípade by nám mohli postačovať tri intervaly. Inokedy sú spracovávané údaje, ktorých triedenie bude prirodzenejšie v násobkoch celých čísel. Napríklad pre hmotnosti pacientov, kedy namiesto vypočítaných desatinných čísel hraníc intervalov použijeme celé hodnoty ako 40 kg, 45 kg, 50 kg, 55 kg, 60 kg, 65 kg atď.

Pri stanovovaní hraníc triednych intervalov teda zvyčajne používame rovnakú šírku intervalov h, pričom krajné intervaly môžu byť otvorené a taktiež používame odporúčaný počet tried intervalov k. Potom hranice intervalov môžeme upraviť takto:

$$(-\infty; x_{min} + h\rangle \quad (x_{min} + h; x_{min} + 2h\rangle \quad \cdots \quad (x_{min} + (k-1)h; -\infty)$$

V prípade malých a stredne veľkých súborov údajov (do 100) je možné výpočet odporúčaného počtu tried k zjednodušiť a použiť vzťah (4.8):

$$k \cong \sqrt{n} \tag{4.8}$$

Takto vypočítané počty intervalov sú približne rovnaké ako tie vypočítané podľa Sturgesovho pravidla (4.6). Pre veľké súbory by však toto zjednodušenie bolo nepoužiteľné, keďže generuje výrazne väčšie počty tried, čo býva nepraktické a neprehľadné.

Funkcie v MS Excel

LOG(číslo; [základ])
Funkcia LOG vráti logaritmus čísla pri určenom základe.

LOG10(číslo)
Funkcia LOG10 vráti dekadický logaritmus čísla.

4.2.2 Početnosti spojitých veličín

Po zadefinovaní hodnôt triednych intervalov pokračujeme určovaním početností, relatívnych početností, kumulatívnych početností a kumulatívnych relatívnych početností, ktoré je analogické k postupu použitom v jednoduchom

triedení. Nehľadáme tu však jednu konkrétnu hodnotu, ale hľadáme všetky hodnoty, ktoré spadajú do príslušného intervalu hodnôt (triedy) a spočítame ich, aby sme dostali početnosť v danej triede, napríklad tak, ako to bolo naznačené na obrázku 4.7, a následne aj relatívne a kumulatívne početnosti. Kompletná tabuľka početností rozsiahlych súborov diskrétnych alebo spojitých kvantitatívnych veličín podľa intervalového triedenia je uvedená ako tabuľka 4.8.

Tabuľka 4.8: Tabuľka početností intervalového triedenia údajov.

Triedny interval $(x_i; x_{i+1})$	**Početnosť** n_i	**Relatívna početnosť** p_i	**Kumulatívna početnosť** N_i	**Kumulatívna relatívna početnosť** P_i
$\langle x_{min}; x_{min}+h\rangle$	n_1	p_1	n_1	p_1
$(x_{min}+h; x_{min}+2h\rangle$	n_2	p_2	n_1+n_2	p_1+p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$(x_{min}+(k-1)h; x_{min}+kh\rangle$	n_k	p_k	$n_1+\cdots+n_k$	$p_1+\cdots+p_k$

Ako už bolo uvedené vyššie, dolnú hranicu prvého intervalu a hornú hranicu posledného intervalu môžeme pri spracovaní výpočtovými prostriedkami nechať otvorenú. Taktiež je potrebné poznamenať, že v aplikáciách určených na štatistické spracovanie údajov postačuje definovať zoznam horných hraníc jednotlivých intervalov. Takto sa do prvého intervalu zahrnú všetky hodnoty menšie alebo rovné hodnote prvej hranice a do poslednej triedy všetky väčšie hodnoty ako horná hranica predposlednej triedy.

Príklad 4.2. Predpokladajme, že v biochemickom laboratóriu boli v rámci žiadaniek na biochemické vyšetrenie okrem iných parametrov analyzované aj údaje bilirubínu. Zaznamenané hodnoty bilirubínu u sledovaného súboru pacientov boli: 16,2; 19,7; 16,4; 17,7; 13,7; 17,6; 6,7; 14,4; 16,4; 15,1; 15,6; 16,5; 17,2; 12,5; 12,2; 15,9; 16,8; 7,6; 17,8; 18,5; 20,7; 14,1; 18,7; 20,6; 16,8; 16; 14,4; 17,9; 14,8; 17,6; 9,2; 14,1; 14,4; 19,3; 12,2; 13,8; 17; 18,7; 17,6; 14,3; 19,6; 13,3; 16,4; 12,4; 16,3; 17,5; 13,6; 17,1; 13,2 a 15,4 $\mu mol/l$. Zostavme tabuľku početností pre súbor získaných hodnôt.

Bilirubín predstavuje kvantitatívnu spojitú veličinu, ktorá môže nadobúdať veľmi veľké množstvo hodnôt, teoreticky neobmedzené. Najmenšou zaznamenanou hodnotou bilirubínu x_{min} v tomto výberovom súbore bola hodnota 6,7 $\mu mol/l$. Najväčšou zaznamenanou hodnotou x_{max} bola hodnota 20,7 $\mu mol/l$. Variačné rozpätie R hodnôt bilirubínu je teda 20,7-6,7=14 $\mu mol/l$. Spočítaním všetkých hodnôt zistíme, že veľkosť výberového súboru n je 50.

Potom podľa Sturgesovho pravidla vypočítame odporúčaný počet intervalov k bude $1+3{,}322\log_{10}50=6{,}6439$, čo po zaokrúhlení smerom nahor predstavuje 7 intervalov. Šírka jedného intervalu h preto bude $14/7=2$ $\mu mol/l$, t. j. hodnoty bilirubínu budeme triediť do siedmich intervalov so šírkou 2 $\mu mol/l$, počínajúc hodnotou 6,7 $\mu mol/l$ vrátane. Podľa tohoto výpočtu budeme mať definované intervaly triedenia hodnôt bilirubínu takto: $\langle 6,7; 8,7\rangle$, $(8,7; 10,7\rangle$, $(10,7; 12,7\rangle$, $(12,7; 14,7\rangle$, $(14,7; 16,7\rangle$, $(16,7; 18,7\rangle$, $(18,7; 20,7\rangle$. V závislosti od charakteru údajov a potrieb ich prezentácie, by sme mohli okrem hraníc intervalov určiť aj ich stredy. V tomto prípade by to boli hodnoty: 7,7; 9,7; 11,7; 13,7; 15,7; 17,7 a 19,7 $\mu mol/l$.

Nakoniec zostrojíme tabuľku početností intervalového rozdelenia početností hodnôt bilirubínu sledovaných pacientov (tabuľka 4.9).

Tabuľka 4.9: Tabuľka početností hodnôt bilirubínu.

Triedny interval $(x_i; x_{i+1})$	**Početnosť** n_i	**Relatívna početnosť** p_i	**Kumulatívna početnosť** N_i	**Kumulatívna relatívna početnosť** P_i
$\langle 6,7; 8,7\rangle$	2	0,04	2	0,04
$(8,7; 10,7\rangle$	1	0,02	3	0,06
$(10,7; 12,7\rangle$	4	0,08	7	0,14
$(12,7; 14,7\rangle$	11	0,22	18	0,36
$(14,7; 16,7\rangle$	12	0,24	30	0,60
$(16,7; 18,7\rangle$	15	0,30	45	0,90
$(18,7; 20,7\rangle$	5	0,10	50	1
Spolu	50	1	-	-

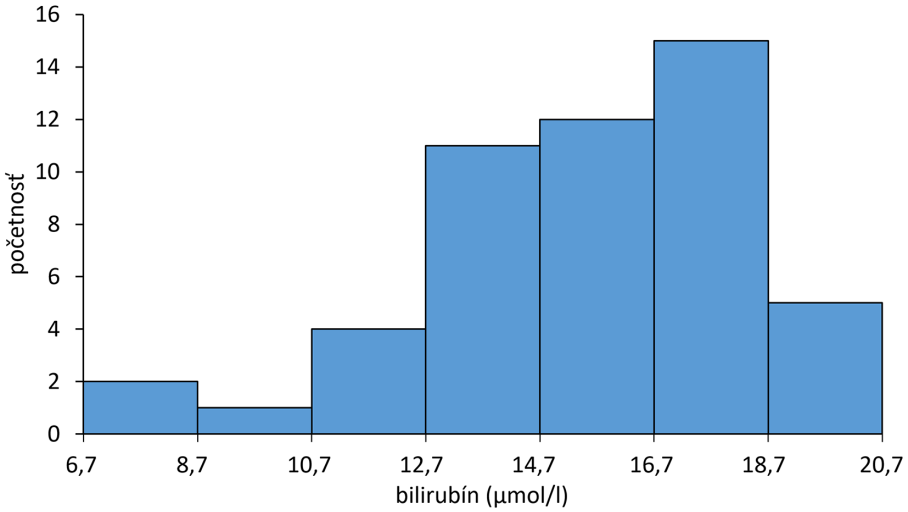
Z tabuľky 4.9 vidíme nielen to, koľko hodnôt máme vo výberovom súbore, ale aj to, ktoré hodnoty sú zastúpené najčastejšie, aké je ich rozloženie naprieč variačným rozpätím, aké zastúpenie majú jednotlivé intervaly hodnôt v rámci celého súboru a mnohé ďalšie charakteristiky, ako to už bolo diskutované vyššie.

4.2.3 Grafická prezentácia

Pri intervalovom triedení údajov používame na ich grafickú prezentáciu najčastejšie histogram, prípadne polygón. V niektorých prípadoch je využívaný taktiež diagram stoniek a listov, či iné typy grafov.

Histogram je optimálnym spôsobom grafického zobrazenia spojitých kvantitatívnych veličín bez ohľadu na rozsah hodnôt analyzovaných súborov. Je obdobou stĺpcového grafu, avšak stĺpce histogramu sú spojené (nie sú medzi

nimi medzery), keďže spojité hodnoty na seba nadväzujú. Na x-ovej osi sú zobrazované hranice jednotlivých tried, takže šírka x-ovej osi zodpovedá variačnému rozpätiu súboru zobrazovaných údajov. Na y-ovej osi sú zobrazované početnosti, prípadne aj relatívne alebo kumulatívne početnosti. Histogram tak zobrazuje rozdelenie početností, a preto sú z neho základné charakteristiky rozdelenia údajov dobre čitateľné. Histogram je považovaný za veľmi užitočný graf pre svoju samovysvetľujúcu vizualizáciu údajov. Na obrázku 4.8 je znázornený histogram početností údajov z príkladu 4.2.



Obr. 4.8: Histogram početností hodnôt bilirubínu.

Z histogramu početností hodnôt bilirubínu je možné okrem vyššie uvedených informácií identifikovať viaceré popisné charakteristiky. Z tvaru histogramu je vidieť, že rozloženie údajov nie je symetrické a šikmosť má negatívy charakter. Spočítaním výšok histogramu zisťujeme rozsah súboru, ktorý je 50. Medián rozdelenia je teda daný 25. a 26. hodnotou. Kumulovaním početností (výšok stĺpcou) zistíme, že medián leží v intervale 14,7 až 16,7 $\mu mol/l$, keďže tento interval obsahuje hodnoty od 19 do 30. Z rozdelenia početností vieme taktiež odhadnúť priemer hodnôt, a to pomocou vzťahu (4.9):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i p_i = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_k p_k}{n} \tag{4.9}$$

kde k je počet intervalov triedenia, x_i sú stredné hodnoty intervalov sledovanej veličiny (u jednoduchého triedenia je to hodnota triedy) a p_i sú početnosti hodnôt v príslušných intervaloch. Početnosti vystupujú do výpočtu priemeru ako váhy stredných hodnôt intervalov, a preto takto vypočítaný aritmetický priemer nazývame aj ako vážený aritmetický priemer.

Potom priemerná hodnota bilirubínu určená na základe rozdelenia početností bude:

$$\bar{x} = \frac{7,7 \times 2 + 9,7 \times 1 + \cdots + 19,7 \times 5}{50} = \frac{775}{50} = 15,5$$

Priemerná hodnota vypočítaná z hodnôt zdrojového súboru je 15,63 $\mu mol/l$. Vidíme teda, že aritmetický priemer vypočítaný z rozdelenia početností je takmer totožný. U jednoduchého triedenia by boli hodnoty aritmetických priemerov vypočítané oboma spôsobmi rovnaké.

Podobne by sme z informácií o rozdelení početnosti vedeli určiť rozptyl alebo smerodajnú odchýlku daného výberového súboru. Rozptyl hodnôt vypočítame podľa vzťahu (4.10):

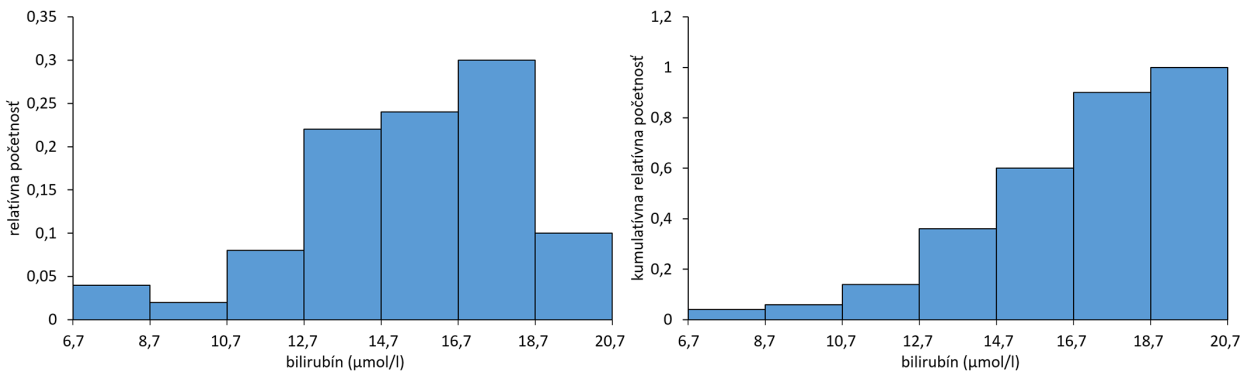
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} p_i \left(x_i - \bar{x}\right)^2 \tag{4.10}$$

a smerodajná odchýlka s bude potom odmocninou rozptylu. Rozptyl vypočítaný z informácií odčítaných z histogramu početností hodnôt bilirubínu bude:

$$s^2 = \frac{2\left(7,7 - 15,5\right)^2 + 1\left(9,7 - 15,5\right)^2 + \cdots + 5\left(19,7 - 15,5\right)^2}{49} = \frac{410}{49} = 8,37$$

Rozptyl hodnôt bilirubínu určený z rozdelenia početností je 8,37 $\mu mol^2/l^2$ a smerodajná odchýlka 2,89 $\mu mol/l$. Pre porovnanie, rozptyl vypočítaný zo zdrojových údajov súboru je 8,78 $\mu mol^2/l^2$ a smerodajná odchýlka 2,96 $\mu mol/l$.

Podobne ako u jednoduchého triedenia je často graficky zobrazovaná aj informácia o rozdelení relatívnych či kumulatívnych početností, napríklad tak ako je to znázornené na obrázku 4.9.

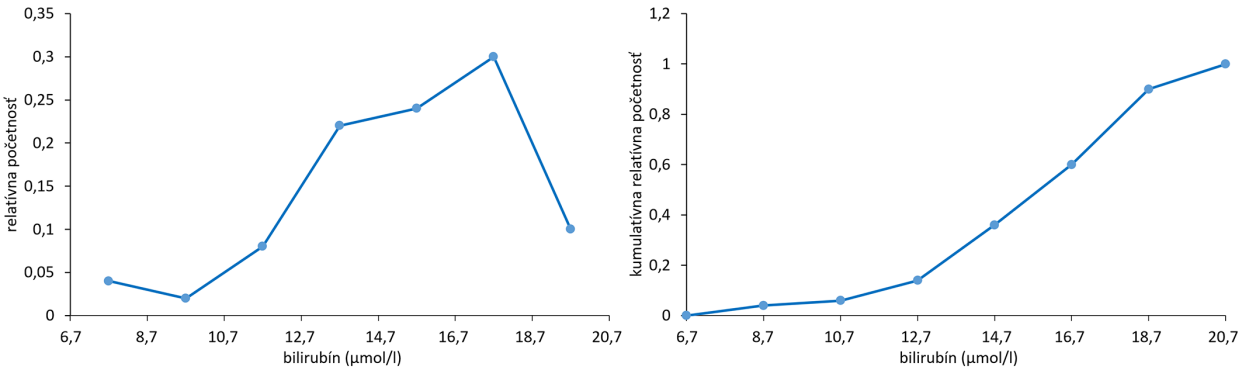


Obr. 4.9: Histogram relatívnych a kumulatívnych relatívnych početností hodnôt bilirubínu.

Pri zostavovaní histogramu je dôležité zachovať rovnakú šírku jednotlivých intervalov, aby prezentované početnosti, resp. relatívne početnosti zodpovedali

skutočným údajom. Ak by sme totiž používali rôzne šírky intervalov triedenia údajov, ale stĺpce histogramu by boli prezentované s rovnakou šírkou, boli by tvar i prezentácia rozdelenia údajov a ich zastúpenia v súbore skreslené. V takom prípade je potrebné relatívnu početnosť vydeliť šírkou daného intervalu a do histogramu vynášať výsledné hodnoty tohto pomeru.

Polygón kvantitatívnych spojitých veličín je tvorený početnosťami, resp. relatívnymi početnosťami korešpondujúcimi so stredmi triednych intervalov, tak ako je to zobrazené na obrázku 4.10.



Obr. 4.10: Polygón relatívnych a kumulatívnych relatívnych početností hodnôt bilirubínu.

Polygón rozdelenia početností/relatívnych početností by sa mal na oboch stranách predĺžiť k osi x tak, že sa spojí so strednými hodnotami ďalšieho „odľahlého“ intervalu nulovej početnosti. Takýto polygón nazývame uzavretým polygónom. Ak tieto odľahlé intervaly nulovej početnosti nezakreslíme, potom polygón nazývame otvoreným, tak ako je to na obrázku 4.10 vľavo. U kumulatívnych početností spojitých veličín vykresľujeme polygón so začiatkom v nule na dolnej hranici prvého intervalu a pokračujeme spájaním horných hraníc jednotlivých triednych intervalov až k poslednému intervalu. Jednou z výhod polygónu kumulatívnych početností je, že podiel (alebo percento) pozorovaní nižších ako špecifikovaná hodnota možno ľahko odčítať z grafu. Napríklad z obrázku 4.10 vpravo je vidieť, že 50% pacientov má hodnotu bilirubínu 16,25 $\mu mol/l$, čo je mediánom súboru. Polygóny kumulatívnych početností je možné použiť na účely porovnávania, a to prekrytím dvoch alebo viacerých polygónov.

Diagram stonky a listov je užitočný spôsob kombinovanej prezentácie zdrojových údajov a ich charakteristík. Štandardná organizácia údajov totiž zahŕňa najmä usporiadanie údajov a zostavenie tabuľky početností, pričom pôvodné údaje sú zoskupené a prezentované tak, že nie je možné skúmať charakteristiky s použitím individuálnych údajov.

V tejto súvislosti je diagram stonky a listov (*stem-and-leaf plot*) vhodnou alternatívou, pretože sú v ňom viditeľné zdrojové údaje a zároveň sú graficky usporiadané. Grafické znázornenie diagramu stonky a listov umožňuje prieskumnú analýzu údajov. Zobrazenie a účel diagramu stonky a listov je pritom obdobné ako u histogramu, ktorý je užitočný pre kvantitatívne súbory údajov. Poskytuje prehľadné informácie o rozsahu súboru, distribúcii a koncentrácii údajov, odhaľuje prítomnosť alebo neprítomnosť symetrie a pod. Hlavnou výhodou diagramu stonky a listov je, že zachováva pôvodný súbor údajov a možno ho vytvárať a dopĺňať už počas zberu údajov. Predstavuje teda výkonnú, ale pritom jednoduchú techniku na skúmanie údajov.

Pri zostavovaní diagramu stonky a listov rozdelíme jednotlivé hodnoty na dve časti, stonku a list. Stonku predstavuje jedna alebo viac počiatočných číslic hodnoty a listovú časť predstavujú zostávajúce číslice. Pri zobrazovaní stonka začína najmenšou hodnotou a končí najväčšou, pričom hodnoty sú zobrazené v usporiadanom stĺpci. Do stonky je potrebné zahrnúť všetky hodnoty v postupnosti, a to aj v prípade, že nie sú v sledovanom súbore. Listy diagramu sú umiestnené na pravej strane zodpovedajúcich stoniek v radoch. Ak je v listoch viac ako jedna číslica, berie sa do úvahy prvá číslica a zostávajúce číslice sa ignorujú. Ak sú v pozorovaných údajoch desatinné miesta, potom sa desatinné miesta v diagrame vynechajú. Medzi stonkou a listami je možné zakresliť zvislú čiaru.

Na obrázku 4.11 je uvedený príklad diagramu stonky a listov hodnôt hmotností stredoškolákov z príkladu 3.1.

3	**89**
4	**2689**
5	**01237777799**
6	**0001112333456**

Obr. 4.11: Diagram stonky a listov hmotností stredoškolákov.

Z vyššie uvedeného diagramu stonky a listov, môžeme vidieť napríklad, že najmenšia hodnota hmotnosti je 38 kg, najväčšia 66 kg, medián 58 kg, modus 57 kg, variačné rozpätie 28 kg, početnosť najvyššia v dekáde hmotností 60-kilogramových stredoškolákov.

Kapitola 5

Úvod do teórie pravdepodobnosti

V kapitole 2.2 sme naznačili, že teória pravdepodobnosti predstavuje premostenie medzi popisnou a induktívnou štatistikou. Teória pravdepodobnosti je odvetvím matematiky, ktorá nám poskytuje základ pre odvodenie štatistických záverov induktívnej štatistiky. Musíme však zdôrazniť, že sa jedná o pomerne rozsiahlu problematiku, ktorá nie je hlavným predmetom tejto učebnice, a preto uvedieme len jej základné koncepty a najdôležitejšie princípy potrebné pre pochopenie tém štatistickej indukcie, tak ako budú prezentované v ďalších kapitolách tejto učebnice.

Termín pravdepodobnosť je v medicíne používaný pomerne často. Môžeme napríklad povedať, že pacient má 70% šancu na uzdravenie, šancu 50 na 50 prežiť určitú operáciu, na 95% konkrétnu chorobu a pod. Ako je vidieť z týchto príkladov, väčšinou vyjadrujeme pravdepodobnosti v percentách. V matematickom ponímaní pravdepodobnosti je však vhodnejšie vyjadrovať pravdepodobnosti ako pomery (resp. zlomky). Percentá potom získame vynásobením hodnoty pravdepodobnosti číslom 100. Z toho je zrejmé, že pravdepodobnosť výskytu nejakého javu alebo udalosti označujeme číslom medzi nulou a jednotkou, pričom daný jav alebo udalosť je tým viac pravdepodobnejšia, čím je toto číslo väčšie, t. j. bližšie k jednotke. Pravdepodobnosť, že jav alebo udalosť určite nastane, má hodnotu jedna. Naopak, javy alebo udalosti, ktoré nemôžu nastať, majú pravdepodobnosť rovnú nule.

Z pohľadu bioštatistiky by bolo legitímne pýtať sa, napríklad či zlepšenie hodnôt biomarkerov, zmena zdravotného stavu, vznik určitého ochorenia a pod. sa mohli objaviť len náhodou, alebo či sú dôsledkom nejakých vonkajších vplyvov, napríklad liečby, životného štýlu, prostredia a pod. Predstavme si situáciu, že sedem z desiatich pacientov trpiacich nejakým ochorením sa po absolvovaní liečby úspešne vylieči. Je pravdepodobné, že dôjde k takejto miere vyliečenia, ak pacienti liečbu nepodstúpia? Alebo, je to skutočný dôkaz, že poskytnutá liečba

je účinná? Prípadne pri riešení iných problémov zisťujeme aká je pravdepodobnosť, že pacient po transplantácii srdca bude žiť tri roky? Aká je pravdepodobnosť, že pacient bude reagovať na stanovenú liečbu? Aká je pravdepodobnosť, že pacient trpiaci bolesťami žalúdka má vred? Aj na takéto otázky sa pokúšame nájsť odpovede s využitím pojmov a zákonitostí platných v teórii pravdepodobnosti. Je však potrebné zdôrazniť aj to, že štatistické analýzy zriedkavo vedú k jednoznačnej odpovedi, a preto by mal byť do odpovedí zahrnutý aj určitý stupeň neistoty.

5.1 Východiská a základné pojmy teórie pravdepodobnosti

Teória pravdepodobnosti vo svojej všeobecnej podstate skúma zákonitosti, u ktorých sa vyskytujú prvky náhodnosti. To znamená, že rieši problémy spojené s úlohami, v ktorých nejaká udalosť môže, ale nemusí nastať, a to aj pri zachovaní rovnakých vstupných podmienok. Podstata teórie pravdepodobnosti je zvyčajne vysvetľovaná na príkladoch ako sú hod mincou, hod kockou, výber hracej karty, narodenie dieťaťa istého pohlavia, výroba súčiastok a pod., pričom sa zdôrazňuje, že výsledok závisí na náhode. Jednotlivé činnosti z týchto príkladov vedú k rôznym výsledkom, ktorých počet môže byť konečný, napríklad u diskrétnych veličín, alebo teoreticky nekonečný, resp. s veľmi veľkým počtom možných výsledkov, tak ako je to u spojitých veličín. Hod mincou sa môže skončiť jednou z dvoch možností, hod jednou kockou jednou zo šiestich možností, výber karty z celého balíka sedmových kariet jednou z tridsaťdva možností a pod.

Ak nie je výsledok nejakej činnosti jednoznačne určený pri zabezpečení rovnakých podmienok a nie je ho možné jednoznačne predpovedať, potom takúto činnosť označujeme ako náhodný pokus. Náhodný pokus je možné za rovnakých podmienok (teoreticky neobmedzene) opakovať, čo definuje hromadnosť javov, t. j. vlastnosť, ktorá je pre teóriu pravdepodobnosti základom. Pri realizácii náhodných pokusov (experimentov) teda budeme opakovane hádzať mincou alebo kockou, ťahať kartu z celého balíka kariet, pozorovať počty a pohlavia narodených detí a pod.

Ak vieme, že výsledok náhodného pokusu ja závislý na náhode, a že pokusy je možné za rovnakých podmienok opakovať, potom môžeme definovať pojem náhodná veličina. Náhodná veličina predstavuje premennú, ktorá môže nadobúdať rôzne hodnoty v závislosti na náhode. Ak náhodnú veličinu (pokus,

premennú) označíme ako X, potom jej hodnoty (výsledky, udalosti) budú x_i, kde $i = 1, 2, 3, \ldots, n$:

$$X = \{x_1, x_2, x_3, \ldots, x_n\} \tag{5.1}$$

kde x_i predstavujú elementárne javy, ktoré sa navzájom vylučujú, ale pri realizácii náhodného pokusu jeden musí vždy nastať.

V závislosti od nadobúdaných hodnôt môžu byť náhodné veličiny diskrétne (počet elementárnych javov je obmedzený) alebo spojité (elementárne javy môžu nadobúdať ľubovoľné hodnoty).

Náhodný jav potom bude predstavovať ľubovoľné tvrdenie o výsledku náhodného pokusu a je reprezentovaný množinou možných elementárnych javov, napríklad pri hode mincou padne znak, pri hode kockou padne 6, narodí sa chlapec, výrobok bude nepodarok, v čakárni bude viac ako 10 pacientov a pod. V špecifických prípadoch sa môžeme stretnúť s istým alebo aj nemožným javom. Istý jav je náhodný jav, ktorý nastane pri každom pokuse. Nemožný jav je zasa náhodný jav, ktorý nenastane pri žiadnom pokuse (nenastane nikdy).

Podstatu elementárnych javov a náhodných javov si vysvetlíme na nasledovnom príklade. Predpokladajme, že náhodný pokus spočíva v jednom hode šesťstennou hracou kockou so stenami očíslovanými od 1 do 6. Potom náhodná veličina určuje množinu šiestich elementárnych javov, ktoré môžu nastať $X = \{1, 2, 3, 4, 5, 6\}$. Predpokladajme, náhodným javom, označme ho napríklad A, je ak padne jednotka. Potom náhodný jav vymedzuje len jednu elementárnu možnosť a platí, že $A = \{1\}$. Alebo, náhodný jav B nastane, ak padne párne číslo, potom $B = \{2, 4, 6\}$. Prípadne, náhodný jav C nastane, ak padne číslo menšie ako 4, potom $C = \{1, 2, 3\}$ a pod. Ak nie sú definované žiadne obmedzenia náhodných javov, potom by sme vedeli definovať 2^6=64 rôznych náhodných javov, t.j. všetky podmnožiny náhodnej veličiny X.

Ak poznáme množinu všetkých možných výsledkov náhodnej veličiny, potom je možné vyjadriť šancu, že náhodný pokus realizovaný v rámci danej náhodnej veličiny skončí určitým výsledkom. Ak predpokladáme, že udalosť x_1 je jednou z možností, ktoré môžu nastať v náhodnom pokuse náhodnej veličiny X, potom pri viacnásobnom opakovaní pokusu za rovnakých podmienok udalosť x_1 v niektorých pokusoch nastane a v iných nie.

Z popisnej štatistiky a triedenia údajov (viď kapitola 4.1) môžeme analogicky určiť, že ak nastala udalosť x_1 v n pokusoch k–krát, potom číslo k nazývame absolútnou početnosťou udalosti x_1 a podiel k/n nazývame relatívnou početnosťou udalosti x_1.

Ak relatívne početnosti sledovanej udalosti u opakovaných náhodných pokusov kolíšu okolo určitej hodnoty, potom odchýlky od tejto hodnoty budú tým menšie, čím viac pokusov sa uskutoční. Takúto vlastnosť relatívnych početností označujeme ako stabilita relatívnych početností, resp. štatistická stabilita. Základným predpokladom je, že ak sa udalosť vyskytla v určitom pomere prípadov v minulosti, vyskytne sa približne v rovnakom pomere prípadov aj v budúcnosti. Tento predpoklad možno oprieť o veľmi všeobecné tvrdenie nazývané aj ako pravidlo veľkých čísel, ktoré hovorí, že so zvyšovaním počtu pokusov sa relatívne početnosti výsledkov viac a viac približujú k teoretickej (alebo skutočnej) hodnote pravdepodobnosti.

Hodnotu, okolo ktorej kolíše relatívna početnosť sledovanej udalosti x_1 považujeme za pravdepodobnosť tejto náhodnej udalosti a označujeme $P(x_1)$, pričom platí, že:

$$P\left(x_1\right) = \frac{k}{n}; n \to \infty \tag{5.2}$$

Definíciu pravdepodobnosti konkrétneho výsledku teda môžeme chápať ako podiel prípadov, kedy by sa tento výsledok mal vyskytnúť v dlhodobom slede opakovaných pozorovaní. Priradenie pravdepodobnosti jednotlivým elementárnym javom na základe relatívnych početností reálne vykonaných experimentov (pokusov) je v medicínskych úlohách najčastejšie používaným štatistickým prístupom. Okrem tohto prístupu by bolo možné pravdepodobnosti javom prideľovať na základe intuície vychádzajúcej z predošlých skúseností (zaťažené istou mierou subjektívnosti), alebo na základe priradenia rovnakých pravdepodobností všetkým javom (rovnomerné rozdelenie).

Príklad 5.1. Predpokladajme, že farba ľudských očí je daná párom génov (jeden od otca a jeden od matky), ktorý označujeme genotyp. Hnedá farba očí (H) je dominantná nad modrou farbou očí (M). Preto v genotype HM, pozostávajúcom z jedného hnedého génu H a jedného modrého génu M, dominuje hnedý gén. Osoba s genotypom HM má hnedé oči. Máme určiť pravdepodobnosť modrej farby očí dieťaťa, ak obaja rodičia majú hnedé oči a genotyp HM. Aká je pravdepodobnosť, že dieťa bude mať hnedé oči?

Pre vyriešenie tejto úlohy sa potrebujeme pozrieť na všetky možnosti, ktoré môžu nastať pri určení genotypu farby očí dieťaťa. Tieto možnosti vynesieme do tabuľky 5.1. Podľa teórie genetiky máme dostupné štyri možné genotypy pre dieťa, ktoré sú rovnako pravdepodobné. Preto môžeme na výpočet pravdepodobnosti javu, ktorým je daná farba očí, použiť pomer počtu prípadov, kedy nastane táto farba očí ku všetkým možnostiam.

Tabuľka 5.1: Genotypy farby očí dieťaťa.

Otec	**Matka**	
	H	M
H	HH	HM
M	MH	MM

Modré oči sa môžu vyskytnúť iba s genotypom MM, takže existuje iba jeden výsledok priaznivý pre modré oči, potom P(M)=1/4=0,25. Pravdepodobnosť, že dieťa bude mať modré oči je 25%.

Podobne, hnedé oči budú definované ostávajúcimi troma elementárnymi javmi (možnosťami) a preto P(H)=3/4=0,75. Pravdepodobnosť, že dieťa bude mať hnedé oči je 75%.

5.2 Elementárne vlastnosti pravdepodobnosti

Všeobecný matematický prístup k riešeniu problematiky pravdepodobnosti vychádza z troch elementárnych vlastností, z ktorých je pomocou matematickej logiky ďalej konštruovaný celý systém teórie pravdepodobnosti. Tieto tri vlastnosti pravdepodobnosti sú nasledovné.

Predpokladajme, že máme náhodnú veličinu s n navzájom sa vylučujúcimi udalosťami $A_1, A_2, \ldots, A_n$. Potom pre ktorúkoľvek udalosť A_i existuje pravdepodobnosť $P(A_i)$ daná nezáporným číslom a platí, že:

$$0 \leq P\left(A_i\right) \leq 1 \tag{5.3}$$

Inými slovami povedané, všetky udalosti musia mať pravdepodobnosť väčšiu alebo rovnú nule, ale nanajvýš rovnú jednotke. Je pochopiteľné, že pravdepodobnosť nemôže byť negatívna. Kľúčovým konceptom vo vyjadrení tejto vlastnosti je podstata vzájomne sa vylučujúcich udalostí, t. j. takých, ktoré nemôžu nastať súčasne.

Druhou elementárnou vlastnosťou je, že suma pravdepodobností všetkých navzájom sa vylučujúcich udalostí $A_1, A_2, \ldots, A_n$ je rovná jednej:

$$P\left(A_1\right) + P\left(A_2\right) + \ldots + P\left(A_n\right) = 1 \tag{5.4}$$

Toto predstavuje vlastnosť úplnosti a vzťahuje sa na skutočnosť, že pri pozorovaní procesov založených na pravdepodobnostiach musíme počítať so všetkými možnými udalosťami, a keď sa vezmú všetky dohromady, ich celková pravdepodobnosť je 1. Samozrejme, platí požiadavka, že máme kompletný zoznam

možných udalostí, ktoré sa neprekrývajú a žiadne dve z nich sa teda nemôžu vyskytnúť súčasne.

A do tretice, predpokladajme, že máme akékoľvek dve vzájomne sa vylučujúce udalosti A_i a A_j. Potom pravdepodobnosť výskytu A_i alebo A_j sa rovná súčtu ich jednotlivých pravdepodobností:

$$P(A_i + A_j) = P(A_i) + P(A_j) \tag{5.5}$$

Ak by sa A_i a A_j navzájom nevylučovali, potom by táto vlastnosť neplatila, je potrebné hľadať ich spoločné prvky a použiť iné operácie s udalosťami a ich pravdepodobnosťami.

5.3 Vybrané operácie nad pravdepodobnosťami

Náhodné javy (udalosti) predstavujú podmnožiny súboru všetkých možností náhodnej veličiny, a preto aj operácie medzi nimi odpovedajú množinovým operáciám. Výsledky operácií s náhodnými javmi môžu byť teda aplikované na operácie s ich pravdepodobnosťami.

Ak máme udalosť A, potom jej doplnkom je podmnožina všetkých ostatných udalostí, ktoré môžu nastať. Teda platí, že ak nastane doplnok udalosti A, potom nenastane udalosť A. Takúto doplnkovú udalosť označujeme ako $\bar{A}$ a pre ich pravdepodobnosti platí:

$$P(A) + P\left(\bar{A}\right) = 1 \tag{5.6}$$

Ak máme dve ľubovoľné udalosti A a B, potom udalosť $A \cup B$ interpretujeme tak, že nastala aspoň jedna z týchto dvoch udalostí a pre pravdepodobnosť $P(A \cup B)$ platí:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.7}$$

kde udalosť $A \cap B$, znamená, že udalosti A a B nastali súčasne. Ak sú udalosti A a B nezlučiteľné, teda nemôžu nastať súčasne, potom:

$$P(A \cap B) = 0 \tag{5.8}$$

Ak sú udalosti A a B navzájom nezávislé a môžu nastať súčasne, potom:

$$P(A \cap B) = P(A)\,P(B) \tag{5.9}$$

Situácia, kedy nastane udalosť A a nenastane udalosť B je označovaná ako $A - B$, pričom ak sú obe udalosti nezlučiteľné, potom platí:

$$P(A - B) = P(A) \tag{5.10}$$

Ak je udalosť A podmnožinou udalosti B, označujeme to ako $A \subset B$. Pre $P(B - A)$ potom platí:

$$P(B - A) = P(B) - P(A) \tag{5.11}$$

Často sa stretávame s prípadmi tzv. podmienenej pravdepodobnosti, t. j. napríklad kedy nastala udalosť A za podmienky, že nastala udalosť B. Označíme to ako A/B a pre pravdepodobnosť platí:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{5.12}$$

Príklad 5.2. Predpokladajme, že u hospitalizovaných pacientov, ktorí mali podstúpiť operačný zákrok, boli spoločne s ostatnými údajmi v rámci predoperačných vyšetrení zaznamenávané aj informácie o krvnej skupine a zo zaznamenaných údajov bola zostrojená tabuľka početností krvných skupín. Určte pravdepodobnosti nasledujúcich udalostí (javov): A_1 - pacient má krvnú skupinu A, A_2 - pacient má krvnú skupinu AB, A_3 - pacient má krvnú skupinu B, A_4 - pacient má krvnú skupinu 0, A_5 - pacient má krvnú skupinu A a AB, A_6 - pacient má krvnú skupinu A alebo 0, A_7 - pacient nemá krvnú skupinu B.

Tabuľka 5.2: Tabuľka početností krvných skupín hospitalizovaných pacientov.

Krvná skupina	**Početnosť**
A	90
AB	80
B	20
0	10

Pri riešení tohoto príkladu môžeme vychádzať z využitia relatívnych početností na odvodenie pravdepodobností. Hodnoty sa navzájom vylučujú (jeden pacient nemôže mať dve krvné skupiny) a spočítaním početností zisťujeme, že do súboru bolo zahrnutých 200 hodnôt (pacientov). Pravdepodobnosti jednotlivých udalostí potom určíme nasledovne:

$$P(A_1) = \frac{90}{200} = 0,45$$

$$P(A_2) = \frac{80}{200} = 0,4$$

$$P(A_3) = \frac{20}{200} = 0,1$$

$$P(A_4) = \frac{10}{200} = 0,05$$
$$P(A_5) = P(A_1 \cap A_2) = 0$$
$$P(A_6) = P(A_1 \cup A_4) = \frac{90}{200} + \frac{10}{200} = 0,5$$
$$P(A_7) = P\left(\bar{A}_3\right) = 1 - P(A_3) = 1 - \frac{20}{200} = 1 - 0,1 = 0,9$$

Jednotlivé pravdepodobnosti vyjadrené v percentách získame, ak výsledné hodnoty pravdepodobnosti vynásobíme hodnotou 100. Pripomenieme, že udalosť A_5 je rovná nule preto, lebo sa jedná o javy, ktoré nemôžu nastať súčasne.

Kapitola 6

Rozdelenia pravdepodobnosti

Schopnosť odvodiť pravdepodobnosti náhodných veličín nám umožní nielen pozorovať pravdepodobnosti individuálnych javov, tak ako to bolo naznačené v predchádzajúcich úvahách, ale umožňuje nám tiež definovať a pochopiť charakteristiky ich rozdelenia. Rozdelenia pravdepodobnosti náhodných veličín majú svoju nezastupiteľnú úlohu v štatistických analýzach, keďže umožňujú vykonávanie objektívnych rozhodnutí na základe výberových súborov. Rozdelenia pravdepodobnosti sa používajú na výpočet teoretickej pravdepodobnosti výskytu rôznych hodnôt a sú teda teoretickým ekvivalentom empirických rozdelení relatívnych početností.

Napríklad, ak poznáme priemernú hodnotu a smerodajnú odchýlku výšky dospelých mužov, môžeme vypočítať pravdepodobnosť, že budú vyšší ako 180 cm, ak predpokladáme, že rozdelenie výšky v základnom súbore je rovnaké ako príslušné rozdelenie pravdepodobnosti. V nasledujúcich podkapitolách si preto uvedieme najčastejšie používané rozdelenia pravdepodobnosti diskrétnych a spojitých veličín.

6.1 Rozdelenia pravdepodobnosti diskrétnych veličín

Diskrétne náhodné veličiny nadobúdajú konečné (spočítateľné) množstvo hodnôt (spravidla celočíselných). Príkladmi môžu byť počet narodených dievčat, počet vyliečených pacientov, počet dní hospitalizácie v nemocnici, počet úmrtí a pod. Medzi najznámejšie teoretické diskrétne rozdelenia pravdepodobnosti patria napríklad diskrétne rovnomerné rozdelenie, alternatívne rozdelenie, binomické rozdelenie, Poissonovo rozdelenie, hypergeometrické rozdelenie či logaritmické rozdelenie.

Rozdelenie pravdepodobnosti diskrétnej náhodnej veličiny predstavuje pravdepodobnosti priradené každej hodnote diskrétnej náhodnej veličiny. Relatívna

početnosť získaná z údajov výberového súboru tak aproximuje rozdelenie pravdepodobnosti pre veľké súbory. Každá pravdepodobnosť priradená konkrétnej hodnote je číslo medzi 0 a 1 a súčet pravdepodobností všetkých možných hodnôt diskrétnej náhodnej veličiny sa rovná 1. Ak hodnoty $x_i;\ i = 1, 2, \ldots, n$ označujú možný výsledok náhodnej veličiny X a ak $P(X = x_i) = P(x_i) = p_i$ označuje pravdepodobnosť tohto výsledku, potom:

$$0 \leq p_i \leq 1 \quad \text{a} \quad \sum_{i=1}^{n} p_i = 1 \tag{6.1}$$

Rozdelenie pravdepodobnosti diskrétnej náhodnej veličiny definujeme pravdepodobnostnou tabuľkou diskrétnej náhodnej veličiny (tabuľka 6.1).

Tabuľka 6.1: Pravdepodobnostná tabuľka diskrétnej náhodnej veličiny.

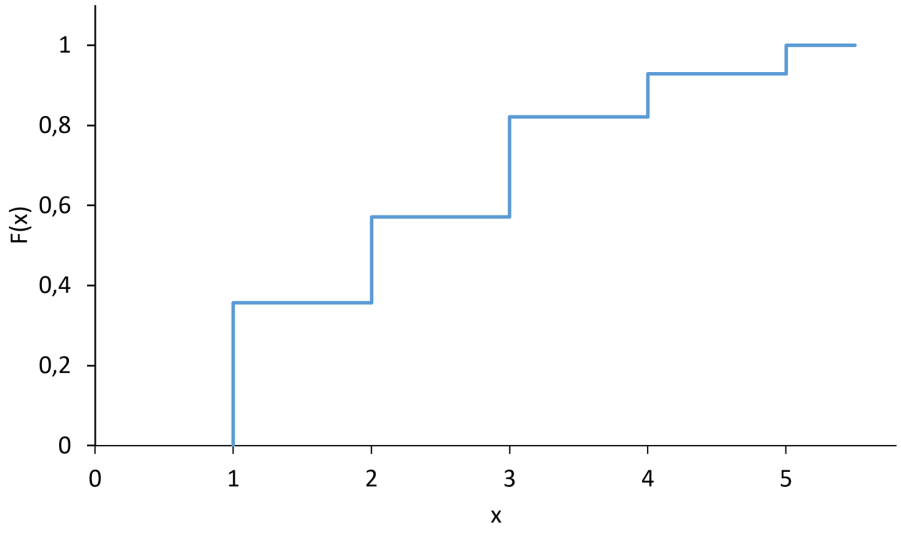
x_i	x_1	x_2	x_3	$\cdots$	x_n
p_i	p_1	p_2	p_3	$\cdots$	p_n

Alternatívne môžeme toto rozdelenie pravdepodobnosti diskrétnej náhodnej veličiny prezentovať vo forme stĺpcového grafu. Grafickou prezentáciou a spojením bodov $[x_i; p_i]$ dostaneme polygón rozdelenia pravdepodobnosti.

Rozdelenie pravdepodobnosti diskrétnej náhodnej veličiny definujeme taktiež pomocou distribučnej funkcie. Kumulatívne rozdelenie pravdepodobnosti pre diskrétnu náhodnú veličinu, ktorej rozdelenie pravdepodobnosti je uvedené v tabuľke 6.1, možno získať postupným sčítaním pravdepodobností p_i. Kumulatívnu pravdepodobnosť pre x_i potom zapíšeme ako $F(x_i) = P(X \leq x_i)$. Hodnota distribučnej funkcie rozdelenia pravdepodobnosti diskrétnej náhodnej veličiny udáva pravdepodobnosť, že náhodná veličina X nadobudne hodnoty menšie alebo rovné špecifikovanej hodnote x_i. Distribučnú funkciu rozdelenia pravdepodobnosti diskrétnej náhodnej veličiny potom zapíšeme v tvare (6.2):

$$F(x) = P(X \leq x), x \in R \tag{6.2}$$

Príklad grafickej prezentácie distribučnej funkcie rozdelenia pravdepodobnosti diskrétnej náhodnej veličiny je znázornený na obrázku 6.1 (prezentuje údaje z príkladu 4.1). Graf funkcie $F(x)$ pozostáva výlučne z horizontálnych čiar, avšak zvislé (vertikálne) čiary dodávajú grafu súvislý vzhľad. Nahliadnutím do takejto grafickej prezentácie kumulatívneho rozdelenia pravdepodobnosti môžeme rýchlo identifikovať pravdepodobnosti jednotlivých hodnôt, ale aj skupín hodnôt diskrétnej náhodnej veličiny.



Obr. 6.1: Distribučná funkcia rozdelenia pravdepodobnosti diskrétnej náhodnej veličiny.

Rozdelenie pravdepodobnosti reprezentujúce hodnoty základného súboru je možné použiť aj na odvodenie priemernej hodnoty či rozptylu hodnôt náhodnej veličiny. Priemer základného súboru μ je definovaný ako očakávaná hodnota diskrétnej náhodnej veličiny X, ktorú vypočítame pomocou vzťahu (6.3):

$$\mu = \sum_{i=1}^{n} x_i p_i \tag{6.3}$$

Podobne je možné vypočítať rozptyl základného súboru σ^2 diskrétnej náhodnej veličiny X, a to podľa vzťahu (6.4):

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p_i \tag{6.4}$$

z ktorého druhou odmocninou získame smerodajnú odchýlku σ.

6.1.1 Diskrétne rovnomerné rozdelenie

Diskrétne rovnomerné rozdelenie $R(n)$ je veľmi jednoduchým teoretickým rozdelením diskrétnej náhodnej veličiny. Predpokladá, že pravdepodobnosť výskytu ktorejkoľvek hodnoty x_i diskrétnej náhodnej veličiny X je rovnaká:

$$P(x_1) = P(x_2) = \cdots = P(x_n) = \frac{1}{n} \tag{6.5}$$

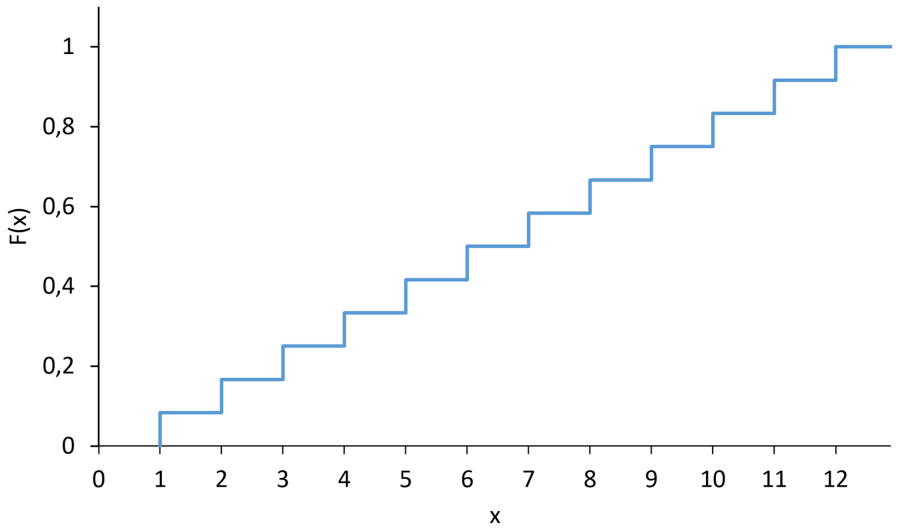
kde n je rozsah diskrétnej náhodnej veličiny X a zároveň jediným parametrom rovnomerného rozdelenia $R(n)$.

Rovnomerné diskrétne rozdelenie $R(n)$ potom definuje pravdepodobnostnú tabuľku diskrétnej náhodnej veličiny X s rovnakými pravdepodobnosťami pre všetky hodnoty, tak ako je to uvedené v tabuľke 6.2.

Tabuľka 6.2: Pravdepodobnostná tabuľka diskrétneho rovnomerného rozdelenia.

X	x_1	x_2	x_3	$\cdots$	x_n
p_i	$1/n$	$1/n$	$1/n$	$\cdots$	$1/n$

V prípade, že máme diskrétnu náhodnú veličinu, ktorej hodnoty môžu nadobúdať celočíselné hodnoty 1 až 12, potom pravdepodobnosť pre ktorúkoľvek z týchto hodnôt bude $p_i = 1/12 = 0{,}8333$. Pre distribučnú funkciu platí vzťah (6.2) a jej grafické znázornenie je zobrazené na obrázku 6.2.



Obr. 6.2: Distribučná funkcia diskrétnej náhodnej veličiny s dvanástimi možnosťami.

Diskrétne rovnomerné rozdelenie $R(n)$ používame pri teoretických úvahách o rozložení diskrétnych hodnôt, ak nie sú známe empirické hodnoty relatívnych početností, alebo ak chceme otestovať, či získané údaje majú rovnomerné rozdelenie. Napríklad, pri posudzovaní výskytu nejakého ochorenia počas roka, ak predpokladáme, že výskyt ochorenia je v každom mesiaci rovnaký, t. j. nejedná sa o sezónne ochorenie.

6.1.2 Alternatívne rozdelenie

Alternatívne rozdelenie $A(p)$ je špecifickým teoretickým rozdelením diskrétnej náhodnej veličiny, kedy skúmaný jav môže nastať alebo nenastane (dve možnosti). Napríklad, pozorujeme vývoj zdravotného stavu pacienta a hodnotíme,

či ochorenie vzniklo alebo nie. Prípadne sledujeme pohlavie narodeného dieťaťa, reakciu na podaný liek, dostupnosť lieku v lekárni a pod.

U alternatívneho rozdelenia $A\,(p)$ náhodná premenná X s pravdepodobnosťou p nadobúda hodnotu 1 ak jav nastal a s pravdepodobnosťou $q = 1 - p$ hodnotu 0 ak jav nenastal. Pravdepodobnosť p je jediným parametrom alternatívneho rozdelenia diskrétnej náhodnej veličiny.

Pravdepodobnostná tabuľka má teda v prípade alternatívnej náhodnej veličiny X len dve možnosti (neznamená to automaticky rovnaké pravdepodobnosti 0,5 a 0,5), tak ako je to uvedené v tabuľke 6.3.

Tabuľka 6.3: Pravdepodobnostná tabuľka alternatívneho rozdelenia.

X	0	1
p_i	$1-p$	p

Distribučná funkcia $F\,(x)$ predstavuje len jednu úroveň, t. j. hodnotu jedna, a to ak nastala sledovaná udalosť.

6.1.3 Binomické rozdelenie

Najjednoduchšie rozdelenie pravdepodobnosti pre diskrétne údaje vzniká, ak existujú len dve možnosti, napríklad tak, ako to bolo v predošlom prípade alternatívneho rozdelenia. Avšak, v analýzach údajov sa stretávame častejšie s prípadmi, kedy sa alternatívne pokusy opakujú a hľadajú sa rozdelenia kombinácie opakovaných pokusov. Pri takejto úvahe sa používa binomické rozdelenie $B\,(n, p)$, ktoré uvažuje s dvoma vzájomne sa vylučujúcimi možnosťami, teda definuje len pravdepodobnosť p kedy udalosť nastala alebo nenastala (pravdepodobnosť $q = 1 - p$), pričom sleduje koľkokrát daná udalosť nastala v n pokusoch. Binomické rozdelenie predpokladá, že pravdepodobnosť p je stanovená pre všetky pokusy (nemení sa), pričom jednotlivé pokusy sú nezávislé, t. j. výsledok jedného pokusu neovplyvňuje výsledky ostatných pokusov.

Predpokladajme, že máme binomickú náhodnú veličinu X, ktorá reprezentuje počty prípadov, kedy nastala sledovaná udalosť v n pokusoch. Môže teda nadobúdať iba hodnoty $0, 1, 2, \ldots, n$, pričom 0 znamená, že udalosť v n pokusoch nenastala, 1 znamená, že nastala v n pokusoch raz atď. Ak je pravdepodobnosť, že udalosť nastane p veľká, potom aj pravdepodobnosti pre počty udalostí veličiny X budú mať tendenciu byť veľké. Naopak, ak je pravdepodobnosť, že udalosť nastane p malá, aj pravdepodobnosti pre počty udalostí veličiny X budú mať tendenciu byť malé.

Ďalej predpokladajme, že máme hodnotu k, pre ktorú platí, že je to celé číslo z rozsahu od 0 do n vrátane, a predstavuje konkrétny počet prípadov kedy nastala sledovaná udalosť v n pokusoch. Potom výpočet pravdepodobnosti binomickej náhodnej veličiny X pre hodnotu k vypočítame podľa vzťahu (6.6):

$$p_k = P(X = k) = \binom{n}{k} p^k q^{n-k} \tag{6.6}$$

kde $p \in (0, 1)$ a $q = 1 - p$. Binomický koeficient $\binom{n}{k}$ je koeficient, ktorý určuje počet rôznych spôsobov, ktorými možno vybrať k objektov z n objektov. Napríklad, ak máme piatich pacientov (n) a dvaja z nich (k) majú krvnú skupinu A, tak to môže byť pacient 1 a 2, alebo 1 a 3, alebo 1 a 4, alebo 1 a 5, alebo 2 a 3 atď., celkom 10 možností. Vypočítame ho podľa vzťahu (6.7):

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{6.7}$$

kde $n!$, resp. $k!$ je n-faktoriál, resp. k-faktoriál, ktoré počítame pre kladné čísla ako súčin všetkých celých čísel od 1 do n, resp. do k. Napríklad, 5!=1.2.3.4.5=120. Špecifickým prípadom je 0, kedy 0!=1.

Pravdepodobnostná tabuľka binomického rozdelenia diskrétnej náhodnej veličiny $B(n, p)$ potom definuje pravdepodobnosti pre všetky hodnoty počtov možného výskytu sledovanej udalosti, tak ako je to uvedené v tabuľke 6.4.

Tabuľka 6.4: Pravdepodobnostná tabuľka binomického rozdelenia.

x_i	0	1	2	$\cdots$	n
p_i	p_0	p_1	p_2	$\cdots$	p_n

Ak má náhodná veličina binomické rozdelenie X, potom je možné vypočítať očakávanú hodnotu jej priemeru a rozptylu. Očakávaná hodnota priemeru je:

$$\mu = np \tag{6.8}$$

Podobne, s využitím informácie o počte pokusov a pravdepodobnosti bude hodnota rozptylu binomickej veličiny X vypočítaná ako:

$$\sigma^2 = npq \tag{6.9}$$

Príklad 6.1. Predpokladajme, že pravdepodobnosť, že dospelý človek bude trpieť diabetom je 0,12. Predpokladajme tiež, že náhodne vyberieme 8 osôb

(náhodný výber zo základného súboru). Zostrojme tabuľku početností binomického rozdelenia náhodnej veličiny reprezentujúcu počet osôb s diabetom vo výberovom súbore.

Počet osôb v súbore je 8 (počet pokusov), a preto hodnoty veličiny X budú x=0, 1, 2, 3, 4, 5, 6, 7 a 8, t. j. diabetom nebude trpieť nikto, alebo jeden, dvaja atď. pacienti. Pravdepodobnosť udalosti p=0,12 (pacient bude mať diabetes). Potom pravdepodobnosť q=0,88 (nenastane sledovaná udalosť). Pokusy sú nezávislé, keďže výsledok jednej osoby neovplyvňuje výsledok iných osôb a výber bol náhodný.

Pravdepodobnosti jednotlivých hodnôt náhodnej premennej X určíme podľa vzťahu (6.6). Pre možnosť, že nikto nebude mať diabetes je:

$$p_0 = \binom{8}{0} 0,12^0.0,88^8 = \frac{8!}{0!.8!} 0,12^0.0,88^8 = 0,359634525$$

Pravdepodobnosť, že jedna z vybraných osôb bude trpieť diabetom je:

$$p_1 = \binom{8}{1} 0,12^1.0,88^7 = \frac{8!}{1!.7!} 0,12^1.0,88^7 = 0,392328573$$

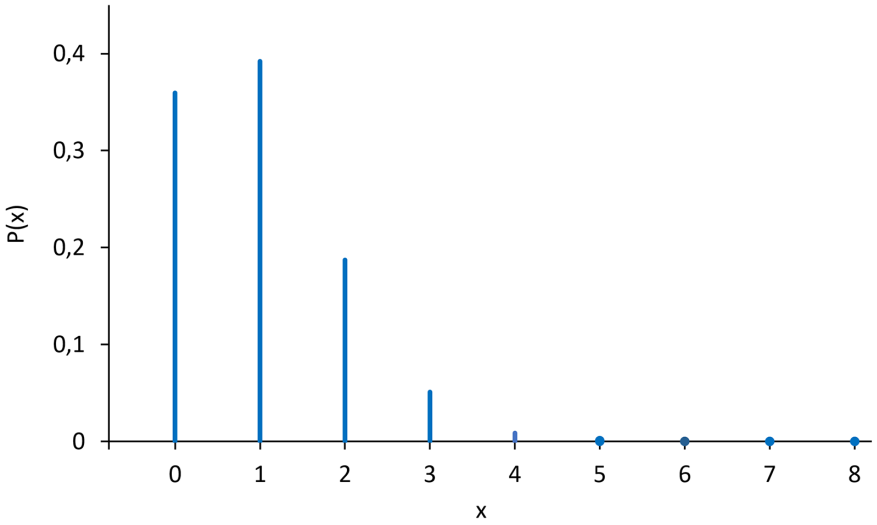
Podobne vypočítame všetky ostatné pravdepodobnosti. Z vypočítaných hodnôt zostrojíme tabuľku pravdepodobností binomickej veličiny. Údaje sú uvedené v tabuľke 6.5.

Tabuľka 6.5: Hodnoty pravdepodobnosti výskytu diabetu v skupine 8 osôb.

x_i	p_i
0	0,359634525
1	0,392328573
2	0,187247728
3	0,051067562
4	0,008704698
5	0,000949603
6	0,000064745
7	0,000005225
8	0,000000042

Z tabuľky pravdepodobností vidíme, že pravdepodobnosť, že nikto nebude mať diabetes je 36%. Pravdepodobnosť, že jedna z ôsmich vybraných osôb bude mať diabetes je 39%. Pravdepodobnosť, že dve z ôsmich vybraných osôb budú mať diabetes je 19% atď. Priemerná hodnota očakávaného počtu osôb s diabetom z 8 je np=8.0,12=0,96. Rozptyl vypočítaný ako npq=8.0,12.0,88=0,8448, čo predstavuje smerodajnú odchýlku (druhá odmocnina) 0,9191.

Rozdelenie pravdepodobnosti náhodnej diskrétnej veličiny X s binomickým rozdelením je možné znázorniť aj graficky, tak ako je to na obrázku 6.3.



Obr. 6.3: Binomické pravdepodobnosti výskytu diabetu.

Funkcie v MS Excel

BINOM.DIST(číslo_s; pokusy; pravdepodobnosť_s; kumulatívne)
Funkcia BINOM.DIST vráti hodnotu binomického rozdelenia pravdepodobnosti jednotlivých veličín.

BINOM.INV(pokusy; pravdepodobnosť_s; alfa)
Funkcia BINOM.INV vráti najmenšiu hodnotu, pre ktorú má distribučná funkcia binomického rozdelenia hodnotu väčšiu alebo rovnajúcu sa hodnote kritéria.

6.1.4 Poissonovo rozdelenie

Poissonovo rozdelenie $P(\lambda)$ je rozdelením diskrétnych náhodných veličín, ktoré je veľmi často využívané v biológii a medicíne ako teoretický pravdepodobnostný model. Možno ho použiť na modelovanie počtu udalostí, ktoré nastanú počas nejakého časového obdobia, prípadne v nejakom priestore, ale kde platí, že každá udalosť nastáva nezávisle a náhodne. Môžu to byť napríklad udalosti ako denný počet nových prípadov rakoviny prsníka, počet vykonaných operačných zákrokov za týždeň, počet abnormálnych buniek v pevnej oblasti histologických preparátov zo série biopsií pečene, počet pozorovaných bakteriálnych kolónií na Petriho misku v mikrobiologickej štúdii, zaznamenané rádioaktívne prípady za jednotku času a pod.

Ak je x konkrétna hodnota diskrétnej náhodnej premennej X s Poissonovým rozdelením, ktorá predstavuje počet výskytov nejakej náhodnej udalosti

v časovom období alebo v priestore, potom pravdepodobnosť, že táto hodnota x nastane, je daná vzťahom (6.10):

$$p_x = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{6.10}$$

kde λ je parameter tvaru, ktorý udáva priemernú hodnotu náhodnej veličiny v danom časovom intervale alebo priestore (je očakávanou hodnotou), e je tzv. prirodzené číslo (2,71828182845904) a x je celé nezáporné číslo, ktoré nadobúda hodnoty 0, 1, 2, 3, . . . , n.

Poissonovo rozdelenie možno tiež reprezentovať ako obmedzený prípad binomického rozdelenia a je niekedy nazývané aj rozdelením zriedkavých udalostí. Predpokladajme, že máme binomické rozdelenie, ale počet pokusov n je veľmi veľký a pravdepodobnosť konkrétnej udalosti pri každom pokuse sa blíži k nule. Potom, výsledné rozdelenie, za predpokladu, že priemerný počet udalostí očakávaných v n pokusoch je konečné, je Poissonovo rozdelenie. Poissonovo rozdelenie tiež predpokladá nezávislé udalosti, z ktorých každá má rovnakú pravdepodobnosť výskytu, no okrem toho predpokladá, že celkový počet takýchto udalostí môže byť (hoci s veľmi malou pravdepodobnosťou) neobmedzene veľký.

Distribučnú funkciu Poissonovho rozdelenia môžeme zapísať v tvare podľa vzťahu (6.11):

$$F(x) = \sum_{i=0}^{x} \frac{\lambda^i e^{-\lambda}}{i!} \tag{6.11}$$

Keďže platí, že pre každé x je $p_x \geq 0$ a $\sum p_x = 1$ Poissonovo rozdelenie vyhovuje požiadavkám pre rozdelenie pravdepodobnosti.

Ak je náhodná veličina X náhodnou veličinou s Poissonovým rozdelením, potom jej priemer a rozptyl sú rovnaké, keďže očakávaná hodnota veličiny X je $\mu = \lambda$ a rozptyl veličiny X je $\sigma^2 = \lambda$.

V prípadoch keď je priemerná hodnota prípadov náhodnej veličiny malá, potom Poissonovo rozdelenie je veľmi asymetrické. So zvyšovaním priemernej hodnoty sa však tvar rozloženia mení a stáva sa takmer symetrickým. Poissonovo rozdelenie nemá žiadnu teoretickú maximálnu hodnotu, ale pravdepodobnosti veľmi rýchlo klesajú k nule.

Príklad 6.2. Predpokladajme, že denný počet nových registrácií rakoviny prsníka môže byť v priemere 2,4, avšak v niektorý deň sa nemusia vyskytnúť žiadne nové prípady, alebo inokedy sa ich môže zasa vyskytnúť niekoľko (teoreticky neobmedzene veľa). Ak predpokladáme, že platia podmienky pre Po-

issonovo rozdelenie, potom môžeme vypočítať pravdepodobnosť ľubovoľného počtu nových prípadov za jeden deň.

Riešením problému bude určiť jednotlivé pravdepodobnosti podľa vzťahu (6.10), a to tak, že budeme predpokladať, že za jeden deň sa nevyskytne žiadna nová registrácia rakoviny prsníka P(0), potom že sa vyskytne jeden prípad P(1), dva prípady P(2), tri prípady P(3) atď. Vychádzame z predpokladu, že priemerný počet prípadov na deň λ je 2,4. Jednotlivé pravdepodobnosti zaokrúhlime na štyri desatinné miesta.

$$P(0) = \frac{2,4^0 e^{-2,4}}{0!} = 0,0907 \quad \text{a} \quad P(1) = \frac{2,4^1 e^{-2,4}}{1!} = 0,2177$$

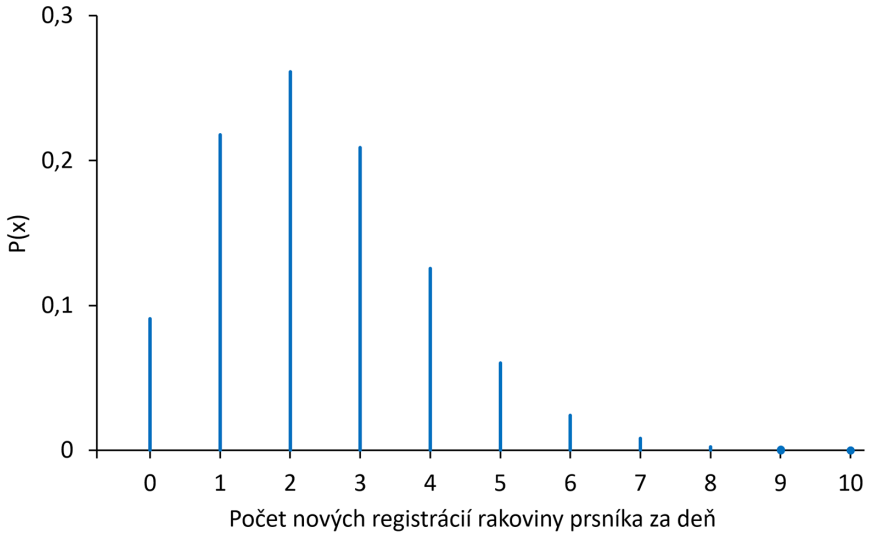
$$P(2) = \frac{2,4^2 e^{-2,4}}{2!} = 0,2613 \quad \text{a} \quad P(3) = \frac{2,4^3 e^{-2,4}}{3!} = 0,2090$$

$$P(4) = \frac{2,4^4 e^{-2,4}}{4!} = 0,1254 \quad \text{a} \quad P(5) = \frac{2,4^5 e^{-2,4}}{5!} = 0,0602$$

$$P(6) = \frac{2,4^6 e^{-2,4}}{6!} = 0,0241 \quad \text{a} \quad P(7) = \frac{2,4^7 e^{-2,4}}{7!} = 0,0083$$

$$P(8) = \frac{2,4^8 e^{-2,4}}{8!} = 0,0024 \quad \text{a} \quad P(9) = \frac{2,4^9 e^{-2,4}}{9!} = 0,0007$$

Vo výpočte pravdepodobností je možné pokračovať, ale ako vidíme tieto budú už len klesať a stále viac sa približovať k nule. Grafický priebeh rozloženia vypočítaných pravdepodobností je znázornený na obrázku 6.4.



Obr. 6.4: Poissonovo rozdelenie pravdepodobnosti pre denné počty nových registrácií rakoviny prsníka.

Funkcie v MS Excel

EXP(číslo)
Funkcia EXP vráti *e* (prirodzené číslo, základ prirodzeného logaritmu) umocnené na zadané číslo.

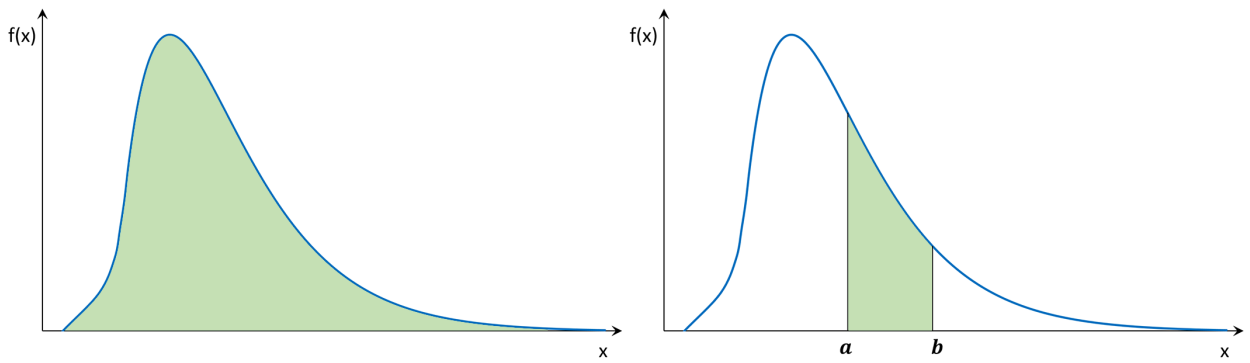
POISSON.DIST(x; stred; kumulatívne)
Funkcia POISSON.DIST vráti hodnoty Poissonovho rozdelenia.

6.2 Rozdelenia pravdepodobnosti spojitých veličín

Spojité náhodné veličiny nadobúdajú ľubovoľné hodnoty z ohraničeného alebo neohraničeného intervalu. Príkladmi môžu byť hmotnosť, výška, cholesterol, čas od nástupu choroby po vyliečenie, čas v čakárni a pod. Na rozdiel od diskrétnych náhodných veličín, u spojitých náhodných veličín nedefinujeme tabuľku pravdepodobností. Pri intervalovom triedení síce vieme pre jednotlivé intervaly hodnôt stanoviť relatívne početnosti, avšak v danom intervale sú hodnoty spojitých veličín teoreticky nekonečné (nespočítateľné).

Pre každú spojitú náhodnú veličinu X existuje jej funkcia $f(x)$. Funkciu $f(x)$ nazývame funkcia hustoty rozdelenia pravdepodobnosti spojitej náhodnej veličiny (funkcia hustoty), ak pre ňu platí, že:

- celková plocha pod krivkou $f(x)$ sa rovná 1 (obrázok 6.5 vľavo),
- pravdepodobnosť, že hodnota spojitej náhodnej veličiny X leží medzi bodmi a a b sa rovná ploche pod krivkou $f(x)$, ktorá je ohraničená bodmi a a b (obrázok 6.5 vpravo),
- pravdepodobnosť intervalovej udalosti je daná plochou pod krivkou $f(x)$, ktorá zodpovedá tomuto intervalu.



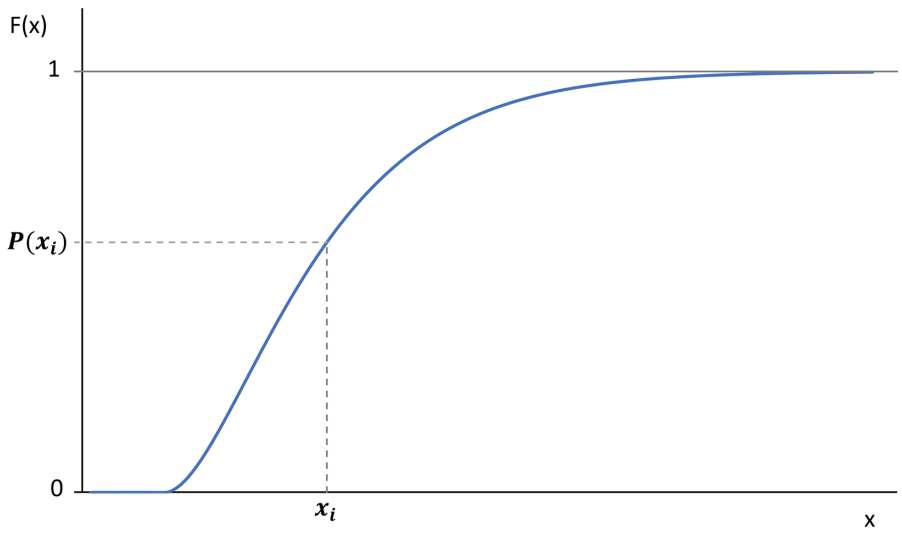
Obr. 6.5: Celková plocha (vľavo) a ohraničená plocha (vpravo) funkcie hustoty rozdelenia pravdepodobnosti spojitej náhodnej veličiny.

Pre celkovú plochu pod funkciou hustoty $f(x)$ teda platí $\int_{-\infty}^{\infty} f(x)\,dx = 1$ a pre ohraničenú plochu platí $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx = P(b) - P(a)$.

Distribučná funkcia $F(x)$ spojitej náhodnej veličiny X pre hodnoty $x \in R$ je definovaná ako:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx \tag{6.12}$$

Distribučná funkcia spojitej náhodnej veličiny $F(x)$ pre hodnotu x_i nám udáva aká je pravdepodobnosť výskytu hodnôt menších alebo rovných ako hodnota x_i. Jej grafický priebeh (obrázok 6.6) je plynulý (nie skokový ako u diskrétnych veličín), neklesajúci, začína v nule a končí v jednotke (pokryté sú všetky hodnoty).



Obr. 6.6: Distribučná funkcia rozdelenia pravdepodobnosti spojitej náhodnej veličiny.

Medzi najznámejšie spojité rozdelenia pravdepodobnosti patria normálne rozdelenie, Studentovo (t) rozdelenie, Fisherove (F) rozdelenie či Chí kvadrát (χ^2) rozdelenie.

6.2.1 Normálne rozdelenie

Normálne rozdelenie $N(\mu, \sigma^2)$ je v štatistickom ponímaní najčastejšie zmieňovaným, používaným a aj najdôležitejším rozdelením pravdepodobnosti spojitých náhodných veličín. Preto je pochopenie jeho podstaty a účelu veľmi dôležité. Je tiež známe ako Gaussovo rozdelenie.

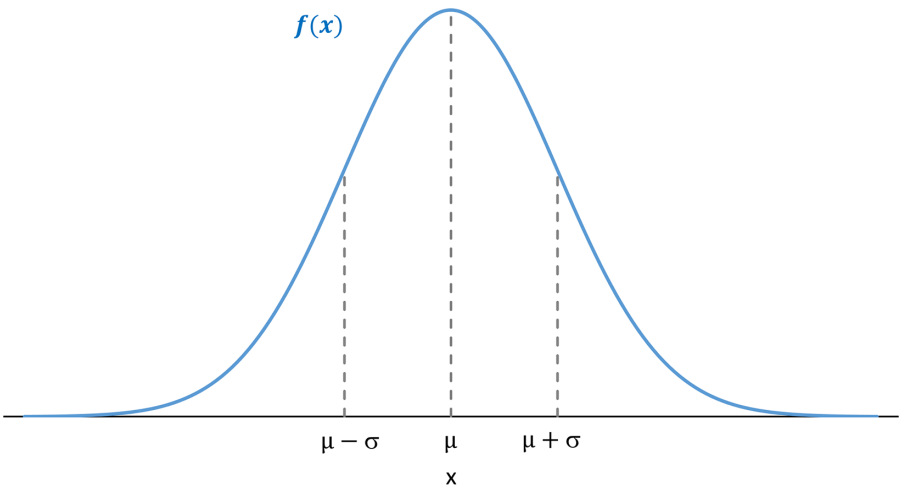
Normálne rozdelenie $N\left(\mu, \sigma^2\right)$ je rozdelenie pravdepodobnosti, ktoré je unimodálne, symetrické a je charakterizované svojou strednou hodnotou (priemerom) μ a rozptylom σ^2. Vo všeobecnosti o spojitej náhodnej veličine, ktorej krivka hustoty rozdelenia je v tvare zvona hovoríme, že má normálne rozdelenie. Normálne rozdelenie je dôležité aj preto, že dobre aproximuje distribúcie mnohých veličín. Histogramy údajov výberových súborov často reprezentujú približne zvonovitý tvar. V takýchto prípadoch hovoríme, že veličina je približne normálne rozdelená. Avšak, hlavným dôvodom prečo je normálne rozdelenie tak dôležité je, že väčšina induktívnych štatistických metód využíva vlastnosti normálneho rozdelenia, aj keď údaje výberového súboru nemajú zvonovitý tvar.

Spojitá náhodná premenná X má normálne rozdelenie pravdepodobnosti s parametrami μ a σ^2 ak jej funkcia hustoty $f\left(x\right)$ má tvar:

$$f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.13}$$

kde $x \in R$, e je tzv. prirodzené číslo (2,71828182845904) a π je matematická konštanta pí (3,14159265358979).

Grafická prezentácia funkcie hustoty normálneho rozdelenia je znázornená na obrázku 6.7

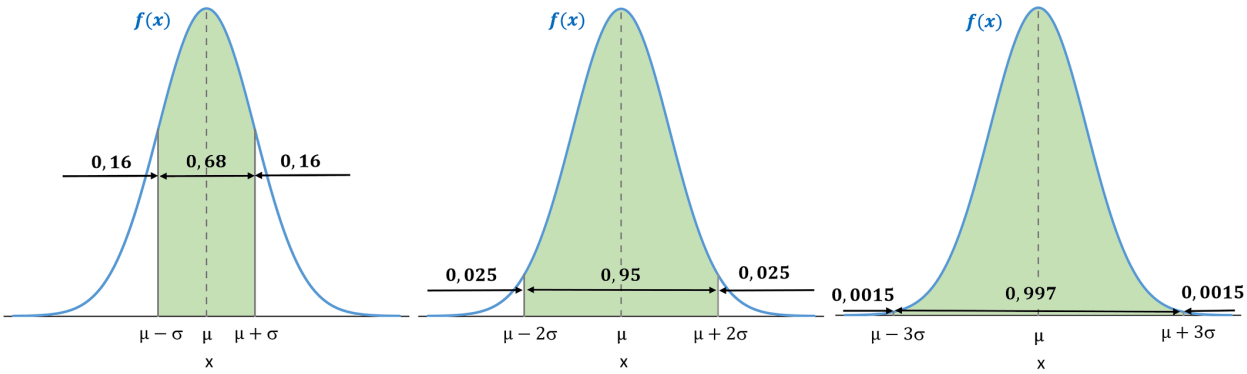


Obr. 6.7: Funkcia hustoty f(x) normálneho rozdelenia.

Funkcia hustoty $f\left(x\right)$ normálneho rozdelenia nadobúda maximum v bode $x = \mu$ (stredná hodnota), okolo ktorého je krivka symetrická. Pre funkciu $f\left(x\right)$ normálneho rozdelenia platí, že priemer = modus = medián, t. j ich hodnoty sú rovnaké. Smerodajná odchýlka σ určuje šírku krivky, ktorá je celkom blízko osi x na úrovni $x = \mu - 3\sigma$ na ľavej strane a na úrovni $x = \mu + 3\sigma$ na pravej

strane. Po okrajoch má krivka tendenciu vyrovnávať sa a približovať sa k horizontálnej osi x, avšak v matematickej teórii sa nikdy osi x nedotýka. Obrázok 6.7 zobrazuje aj hodnoty $x = \mu - \sigma$ a $x = \mu + \sigma$, ktoré na krivke normálneho rozdelenia definujú inflexné body, t.j. miesta, v ktorých dochádza k prechodom medzi konkávnym (stred) a konvexným tvarom (na ľavej a na pravej strane) krivky. Šikmosť (koeficient šikmosti) je pre funkciu hustoty normálneho rozdelenia rovný 0 a špicatosť (koeficient špicatosti) je rovný hodnote 3 (0 v prípade ak je použitá korekcia, viď kapitola 3.3.2).

Normálne rozdelenie má aj ďalšie dôležité vlastnosti. Rovnako ako pri akejkoľvek inej funkcii hustoty je celková plocha pod normálnou krivkou rovná 1. Pravdepodobnosť, že náhodná veličina s normálnym rozdelením leží v rozsahu jednej smerodajnej odchýlky od priemeru, t. j. že leží medzi $\mu - \sigma$ a $\mu + \sigma$ (obrázok 6.8 vľavo) je približne 68%. Z toho vyplýva, že je 16% pravdepodobnosť, že bude ležať pod hodnotou $\mu - \sigma$ a 16% pravdepodobnosť, že bude ležať nad hodnotou $\mu + \sigma$ (pretože 16%+68%+16%=100%).



Obr. 6.8: Oblasti pod špecifikovaným rozsahom funkcie hustoty normálneho rozdelenia.

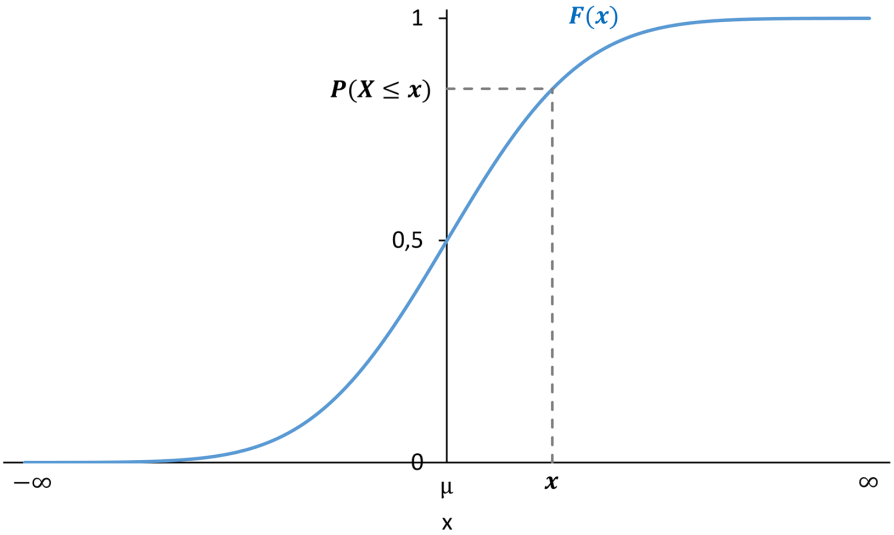
Pravdepodobnosť, že náhodná veličina s normálnym rozdelením leží v rozsahu dvoch smerodajných odchýlok od priemeru, t. j. že leží medzi $\mu - 2\sigma$ a $\mu + 2\sigma$ (obrázok 6.8 v strede) je približne 95%. Z toho vyplýva, že je 2,5% pravdepodobnosť, že bude ležať pod hodnotou $\mu - 2\sigma$ a 2,5% pravdepodobnosť, že bude ležať nad hodnotou a $\mu + 2\sigma$ (pretože 2,5%+95%+2,5%=100%). A tiež pravdepodobnosť, že náhodná veličina s normálnym rozdelením leží v rozsahu troch smerodajných odchýlok od priemeru, t. j. že leží medzi $\mu - 3\sigma$ a $\mu + 3\sigma$ (obrázok 6.8 vpravo) je približne 99,7%. Z toho vyplýva, že je 0,15% pravdepodobnosť, že bude ležať pod hodnotou $\mu - 3\sigma$ a 0,15% pravdepodobnosť, že bude ležať nad hodnotou a $\mu + 3\sigma$ (pretože 0,15%+99,7%+0,15%=100%).

Pre každú funkciu hustoty rozdelenia pravdepodobnosti $f(x)$ existuje práve jedna zodpovedajúca kumulatívna distribučná funkcia $F(x)$ a naopak. Obe po-

skytujú tie isté informácie o rozdelení náhodnej veličiny, avšak iným spôsobom. Zatiaľ čo funkcia hustoty dáva hustotu konkrétnej hodnoty x z X (je analogická s histogramom alebo polygónom početností), distribučná funkcia (kumulatívna) dáva pravdepodobnosť, že X je menšie alebo rovné konkrétnej hodnote x (je analogická kumulatívnemu grafu). Distribučná funkcia normálneho rozdelenia $F(x)$ je integrálom funkcie hustoty, a teda daná vzťahom (6.14):

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{6.14}$$

Grafická prezentácia distribučnej funkcie normálneho rozdelenia $F(x)$ je na obrázku 6.9.

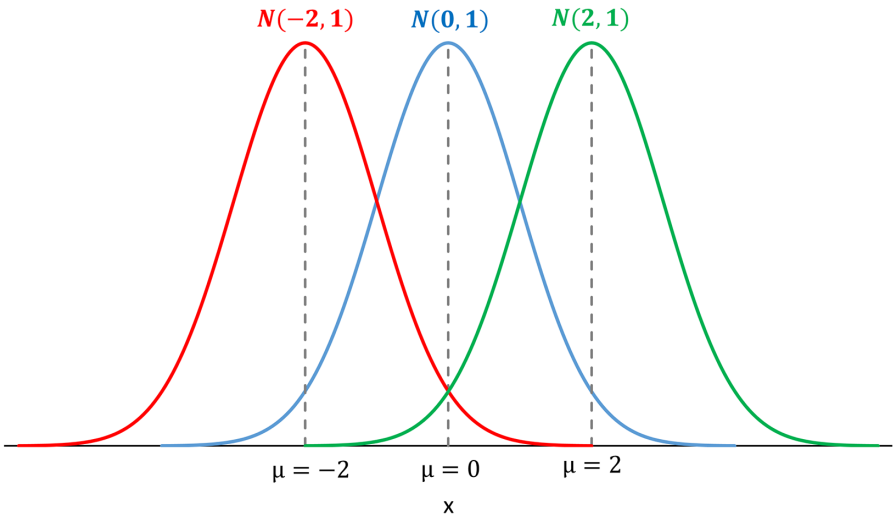


Obr. 6.9: Distribučná funkcia normálneho rozdelenia.

Pre akúkoľvek hodnotu x z rozsahu $-\infty$ až ∞ dáva distribučná funkcia $F(x)$ hodnotu pravdepodobnosti, že spojitá veličina normálneho rozdelenia X bude nadobúdať akékoľvek hodnoty od $-\infty$ do x (obrázok 6.9). Hoci riešenie v jednoduchej uzavretej forme pre integrál neexistuje, môžeme použiť numerické metódy výpočtu na aproximáciu požadovaných oblastí. Našťastie sa takýmito výpočtovými operáciami nemusíme zaoberať, keďže na tieto účely existujú štatistické tabuľky, v ktorých vieme nájsť výsledky akejkoľvek integrácie, ktorá by nás mohla zaujímať. Štatistické tabuľky (aj pre iné rozdelenia) uvádzajú hodnoty pre tzv. štandardné, resp. normované rozdelenia.

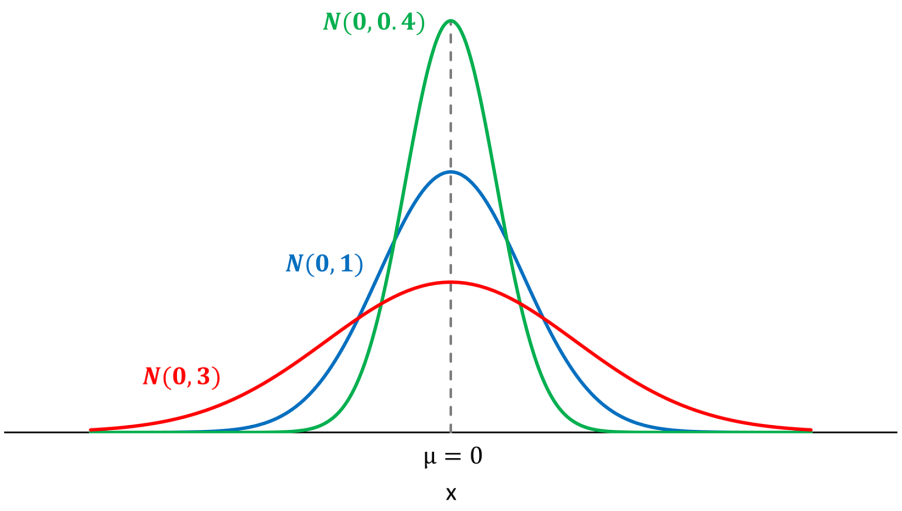
Špecifickým prípadom normálneho rozdelenia je, ak stredná hodnota $\mu = 0$ a rozptyl $\sigma^2 = 1$. Takéto rozdelenie nazývame štandardné (alebo normované) normálne rozdelenie a označujeme ho $N(0, 1)$.

Vplyv zmeny strednej hodnoty μ je možné vidieť na obrázku 6.10. Ak sa hodnota μ zvyšuje, krivka funkcie hustoty sa posúva doprava a naopak, ak sa znižuje, potom sa krivka rozdelenia posúva doľava.



Obr. 6.10: Posun funkcie hustoty pri zmene strednej hodnoty.

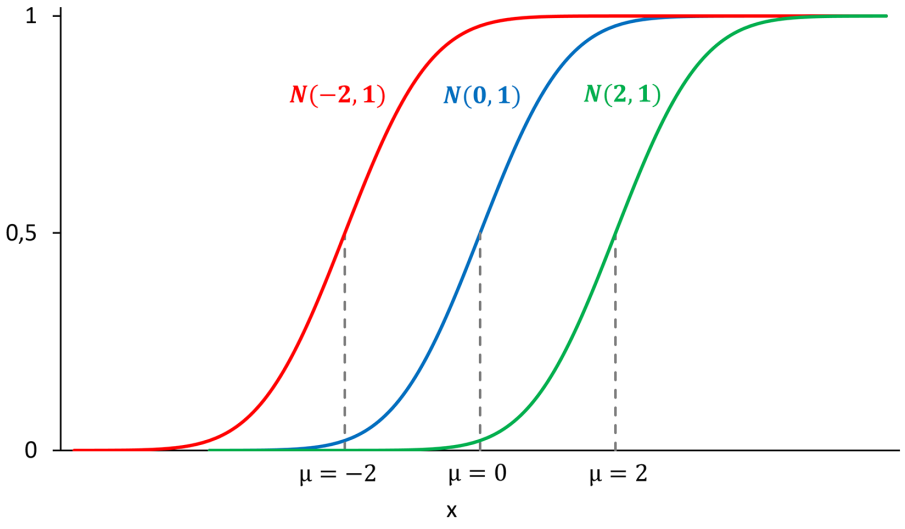
Vplyv zmeny rozptylu σ^2 je možné vidieť na obrázku 6.11. Ak je rozptyl (alebo smerodajná odchýlka ako jeho odmocnina) veľký, krivka funkcie hustoty bude viac roztiahnutá a v strede bude klesať (hodnoty sú viac rozptýlené). Ak sa hodnota rozptylu zmenšuje, potom sa aj krivka funkcie hustoty bude zužovať a v strede bude narastať (väčšie zastúpenie hodnôt okolo strednej hodnoty).



Obr. 6.11: Zmena funkcie hustoty pri zmene rozptylu.

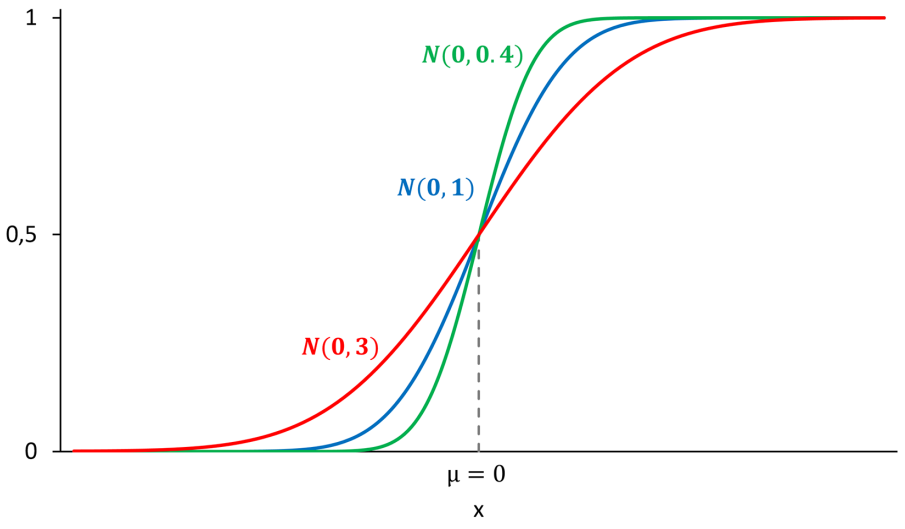
Podobne je možné uviesť, že meny strednej hodnoty alebo rozptylu sa prejavia aj zmenou distribučnej funkcie. Vplyv zmeny strednej hodnoty μ je možné

vidieť na obrázku 6.12. Ak sa hodnota μ zvyšuje, krivka distribučnej funkcie sa posúva doprava a naopak, ak sa znižuje, potom sa krivka posúva doľava.



Obr. 6.12: Posun distribučnej funkcie pri zmene strednej hodnoty.

Vplyv zmeny rozptylu σ^2 je možné vidieť na obrázku 6.13. Ak je rozptyl (alebo smerodajná odchýlka ako jeho odmocnina) veľký, krivka distribučnej funkcie bude viac naklonená a bude rásť pomalšie so zmenou x. Ak sa hodnota rozptylu zmenšuje, potom sa aj okraje krivky distribučnej funkcie priblížia k strednej hodnote a distribučná funkcia bude rásť strmšie.



Obr. 6.13: Zmena distribučnej funkcie pri zmene rozptylu.

Spojitú náhodnú veličinu so štandardným normálnym rozdelením $N(0,1)$ označujeme ako Z, pre ktorú platí:

$$Z = \frac{X - \mu}{\sigma} \tag{6.15}$$

Potom, keďže $\mu = 0$ a $\sigma^2 = 1$, funkcia hustoty rozdelenia pravdepodobnosti náhodnej spojitej veličiny Z je daná funkciou:

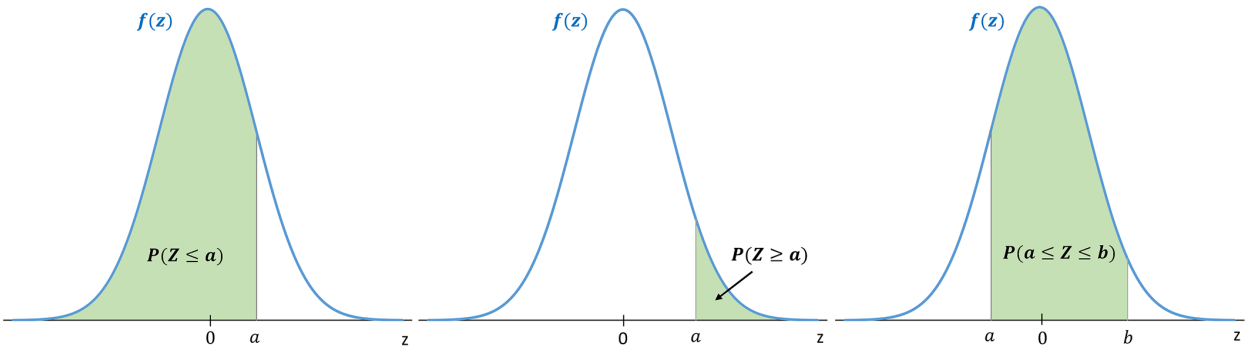
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{6.16}$$

definovanou pre hodnoty z na intervale $-\infty < z < \infty$.

Štandardné normálne rozdelenie $N(0,1)$ má v štatistike veľmi dôležitú úlohu, pretože plocha pod akoukoľvek normálnou krivkou sa dá vypočítať z pravdepodobnosti štandardného normálneho rozdelenia. Berieme pritom do úvahy, že celková plocha pod krivkou funkcie hustoty je rovná 1 a tiež, že priemer aj rozptyl štandardného normálneho rozdelenia sú známe, t. j. rozdelenie je úplne definované.

Hodnota z pre hodnotu náhodnej veličiny predstavuje počet smerodajných odchýlok, ktorých hodnota bude pod $-z$ alebo nad z, ktoré je v prípade štandardného normálneho rozdelenia 0. Napríklad, z-transformácia pre $z = 1$ znamená, že hodnota x použitá v transformácii je o 1 smerodajnú odchýlku nad 0 a hodnota $z = -1$ znamená, že hodnota x je o 1 smerodajnú odchýlku pod 0.

Ak chceme určiť pravdepodobnosť, že Z nadobudne hodnotu medzi ľubovoľnými dvoma bodmi na osi z (horizontálna os štandardného normálneho rozdelenia), musíme nájsť oblasť pod krivkou normálneho rozdelenia ohraničenú týmito dvoma bodmi. Ako už bolo uvedené vyššie, hodnoty z používame zo štatistických tabuliek. Pravdepodobnosť $P(Z \leq a)$ je potom daná oblasťou pod krivkou od a doľava (obrázok 6.14 vľavo), t. j. pre $z = a$ nájdeme odpovedajúcu hodnotu P v štatistických tabuľkách Z rozdelenia.



Obr. 6.14: Oblasti pod krivkou štandardného normálneho rozdelenia.

Podobne, pravdepodobnosť $P(Z \geq a)$ (obrázok 6.14 v strede) nájdeme ako $P(Z \geq a) = 1 - P(Z \leq a)$, kde $P(Z \leq a)$ opäť nájdeme v štatistických tabuľkách. Alebo iný ohraničený prípad $P(a \leq Z \leq b)$ (obrázok 6.14 vpravo) určíme ako $P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$.

Príklad 6.3. Predpokladajme, že hmotnosť žien s onkologickými ochoreniami má normálne rozdelenie s priemernou hodnotou 53 kg a rozptylom 16 kg^2. Určme pravdepodobnosť, že náhodne vybraná žena s onkologickým ochorením má hmotnosť nižšiu ako 48 kg. Určme tiež aké je percento žien s hmotnosťou nižšou ako 48 kg. V základnom súbore 12 000 žien s onkologickými ochoreniami, koľko očakávame, že bude mať hmotnosť nižšiu ako 48 kg?

Spojitou náhodnou veličinou X je hmotnosť žien s onkologickými ochoreniami. Priemerná hodnota hmotnosti žien s onkologickými ochoreniami je 53 kg a smerodajná odchýlka σ=4 kg (druhá odmocnina rozptylu). Potom, pravdepodobnosť P(X≤48) bude:

$$\begin{aligned} P(X \leq 48) &= P\left(Z \leq \tfrac{x-\mu}{\sigma}\right) = P\left(Z \leq \tfrac{48-53}{4}\right) = \\ &= P(Z \leq -1,25) = 0,1056 \end{aligned}$$

Hodnotu pravdepodobnosti sme našli v štatistických tabuľkách pre Z rozdelenie, tak ako je to zobrazené na obrázku 6.15.

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
⋮											
−1.50	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668	−1.50
−1.40	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808	−1.40
−1.30	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968	−1.30
−1.20	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151	−1.20
−1.10	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357	−1.10
⋮											

Obr. 6.15: Výber záporných tabuľkových hodnôt Z štandardného normálneho rozdelenia.

V prípade, ak budeme mať k dispozícii tabuľky kladné hodnoty, potom P(X≤ -z) nájdeme ako P(X≤ -z)=1-P(X≤z), tak ako je to na obrázku 6.16.

Pravdepodobnosť, že náhodne vybraná žena s onkologickým ochorením má hmotnosť nižšiu ako 48 kg je 0,1056 čo predstavuje tiež 10,56% žien s hmotnosťou nižšou ako 48 kg. V základnom súbore 12 000 žien s onkologickými ochoreniami, očakávame, že 12000.0,1056=1267,2≐1267 z nich bude mať hmotnosť nižšiu ako 48 kg.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
⋮											
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	1.00
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	1.10
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015	1.20
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	1.30
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	1.40
⋮											

Obr. 6.16: Výber kladných tabuľkových hodnôt Z štandardného normálneho rozdelenia.

Funkcie v MS Excel

PI()
Funkcia PI nemá argumenty a vráti číslo 3,14159265358979, matematickú konštantu pi, s presnosťou na 15 číslic.

NORM.DIST(x; stred; smerodajná_odch; kumulatívne)
Funkcia NORM.DIST vráti hodnotu normálneho rozdelenia pre zadanú strednú hodnotu a smerodajnú odchýlku.

NORM.INV(pravdepodobnosť; stred; smerodajná_odch)
Funkcia NORM.INV vráti inverznú funkciu k distribučnej funkcii normálneho rozdelenia pre zadanú strednú hodnotu a smerodajnú odchýlku.

6.2.2 t rozdelenie

V predchádzajúcej kapitole sme poukázali na vlastnosti normálneho rozdelenia, vplyv strednej hodnoty a rozptylu na jeho rozdelenie a niektoré úvahy o jeho využití. Normálne rozdelenie predpokladá dostatočne veľké súbory údajov, no nie vždy pracujeme so súbormi, ktoré sú veľké (napríklad viac ako 100). Pri analýze spojitých údajov najčastejšie pracujeme s priemernou hodnotou, ktorá však môže byť pri malých súboroch značne skreslená. Navyše, často nie je známy rozptyl základného súboru. V takýchto prípadoch je vhodné použiť t rozdelenie, ktoré je označované aj ako Studentovo rozdelenie.

Studentovo t-rozdelenie predstavuje rozdelenie spojitej náhodnej veličiny podobnej veličine Z s normálnym rozdelením. Obe rozdelenia, t aj Z predpokladajú, že súbor hodnôt $x_1, x_2, \ldots, x_n$ je náhodný výberový súbor o veľkosti n z normálneho rozdelenia s priemerom μ a rozptylom σ^2. Avšak, kým u normálneho rozdelenia $N\left(\mu, \sigma^2\right)$ je rozptyl σ^2 známy a vieme určiť Z, u t-rozdelenia známy nie je, a preto je ho potrebné odhadnúť. Ak je náhodný výberový súbor o veľkosti n vybraný zo základného súboru s priemerom μ a rozptylom

σ^2, potom má priemer výberového súboru normálne rozdelenie s priemerom μ a rozptylom σ^2/n. Pre náhodnú veličinu Z so štandardným normálnym rozdelením potom platí:

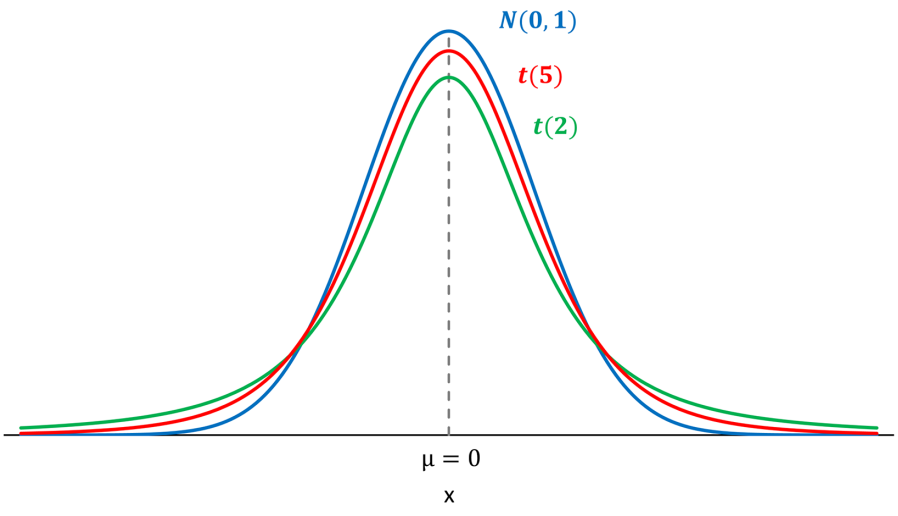
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{6.17}$$

V prípadoch, kedy rozptyl základného súboru σ^2 nie je známy, nahradíme ho rozptylom výberového súboru s^2 a definujeme novú štatistiku označovanú ako T alebo Studentovo T:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \tag{6.18}$$

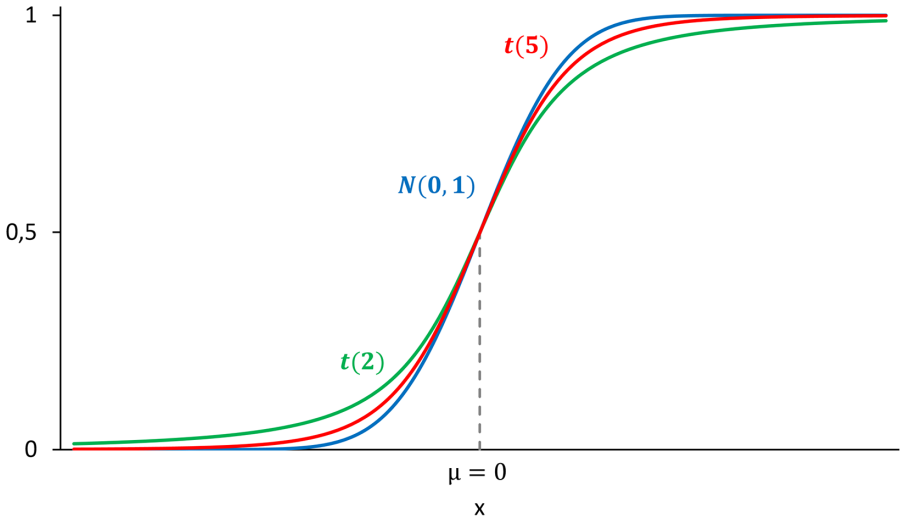
Ak $x_1, x_2, \ldots, x_n$ je náhodný výberový súbor o veľkosti n z normálneho rozdelenia a náhodné veličiny sú nezávislé s priemerom μ a rozptylom σ^2, potom hovoríme, že štatistika T má t-rozdelenie s $v = (n-1)$ stupňami voľnosti.

t-rozdelenie je spojitým symetrickým rozdelením, pričom hodnota náhodnej veličiny T sa pohybuje od $-\infty$ do ∞, priemer T je 0 a rozptyl je $v/(v-2)$. Tvar t-rozdelenia je podobný tvaru štandardného normálneho rozdelenia, pričom so zvyšovaním n sa približuje k štandardnému normálnemu rozdeleniu. Na obrázku 6.17 sú znázornené krivky funkcie hustoty t-rozdelenia pre rôzne stupne voľnosti a štandardného normálneho rozdelenia.



Obr. 6.17: Funkcie hustoty t-rozdelenia v porovnaní so štandardným normálnym rozdelením.

Obdobne sú porovnané grafické priebehy kriviek distribučných funkcií t-rozdelenia pre rôzne stupne voľnosti a štandardného normálneho rozdelenia, a to na obrázku 6.18.



Obr. 6.18: Distribučné funkcie t-rozdelenia v porovnaní so štandardným normálnym rozdelením.

Studentovo t-rozdelenie, podobne ako štandardné normálne rozdelenie, má svoje hodnoty uvedené v štatistických tabuľkách a je teda možné odvodiť jednotlivé pravdepodobnosti oblastí pod krivkou funkcie hustoty.

Príklad 6.4. Predpokladajme, že chceme porovnať priemerný príjem potravy pre konkrétnu skupinu jednotlivcov s odporúčaným denným príjmom. Vo výberovom súbore 11 zdravých žien vo veku 25-35 rokov bol počas desiatich dní pozorovaný priemerný denný energetický príjem. Ich priemerný denný príjem bol 6753,6 kJ a smerodajná odchýlka 1142,1 kJ, pričom táto malá skupina hodnôt nevykazovala zjavnú šikmosť a je ju možné považovať za približne normálne rozdelenú. Čo môžeme povedať o energetickom príjme týchto žien v porovnaní s odporúčaným denným príjmom 7725 kJ?

Priemerná hodnota denného príjmu potravy je 6753,6 kJ, smerodajná odchýlka 1142,1 kJ a hypotetická hodnota príjmu potravy je odporúčaný denný príjem 7725 kJ. Môžeme teda vypočítať hodnotu T Studentovho t-rozdelenia ako:

$$T = \frac{6753,6 - 7725}{\frac{1142,1}{\sqrt{11}}} = -2,821$$

V štatistických tabuľkách pre Studentovo t-rozdelenie nájdeme hodnotu P korešpondujúcu so získanou hodnotou T a počtom stupňov voľnosti v=10 (n-1), tak ako je to na obrázku 6.19. Znamienko mínus môžeme ignorovať pre

t obojstranný test a hľadať najväčšiu tabuľkovú hodnotu pod našou získanou hodnotou T s použitím 10 stupňov voľnosti. Z tabuľky dostávame P<0,02, takže príjem potravy žien vo výberovom súbore bol výrazne nižší ako odporúčaná hladina pri použití obvyklého kritéria P<0,05 (obvyklá hladina testovania).

			(P)			
d.f.	0.2	0.1	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
⋮						
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
⋮						

Obr. 6.19: Výber tabuľkových hodnôt Studentovho t-obojstranného rozdelenia.

Neskôr si ukážeme, ako pri použití tabuľky t-rozdelenia musíme brať do úvahy koeficient spoľahlivosti aj stupne voľnosti.

Funkcie v MS Excel

T.DIST(x; st_voľnosti; kumulatívne)
Funkcia T.DIST vráti ľavostranné Studentovho t-rozdelenie.

T.INV(pravdepodobnosť; st_voľnosti)
Funkcia T.INV vráti inverznú funkciu ľavostranného Studentovho t-rozdelenia.

6.2.3 Chí kvadrát rozdelenie

Nie všetky súbory údajov majú vždy symetrické rozdelenie, ale často sa stretávame aj s rozdeleniami nesymetrickými. Príkladom nesymetrického rozdelenia je aj χ^2 rozdelenie (chí kvadrát rozdelenie). Definíciu χ^2 rozdelenia je možné vysvetliť na príklade spojitej náhodnej veličiny X. Ak má náhodná veličina X štandardné normálne rozdelenie, potom X^2 má χ^2 rozdelenie. Keďže akákoľvek hodnota umocnená na druhú je vždy kladné číslo, potom aj χ^2 môže mať len kladné hodnoty. Rozloženie χ^2 je značne zošikmené, t. j. nie je symetrické, napríklad tak ako sme to videli u normálneho alebo u t rozdelenia. Takéto rozdelenie X^2 má jeden stupeň voľnosti a je najjednoduchším príkladom skupiny χ^2 rozdelení. Mohli by sme teda zapísať:

$$\chi^2_{(1)} = z^2 = \left(\frac{X-\mu}{\sigma}\right)^2 \tag{6.19}$$

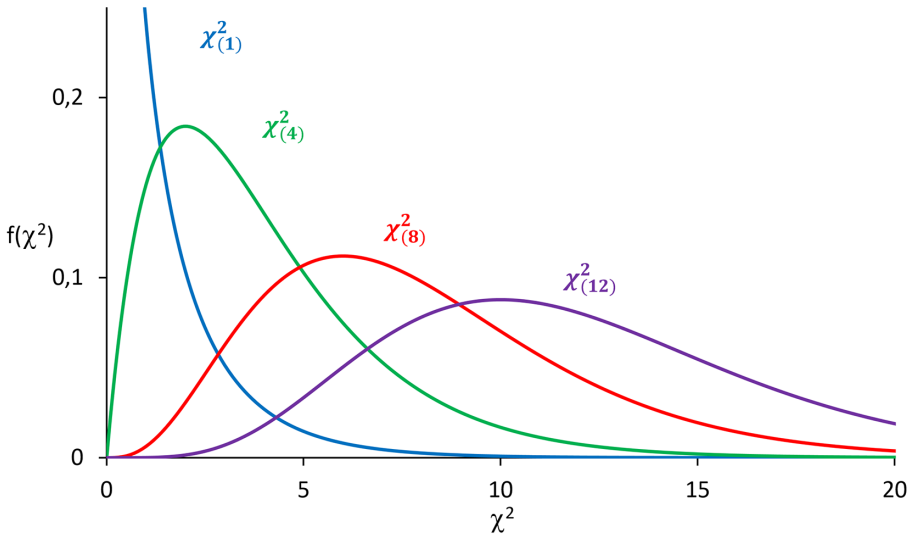
V prípade kedy máme viacero nezávislých veličín, napríklad $X_1, X_2, \ldots, X_n$, z ktorých každá má štandardné normálne rozdelenie, potom súčet druhých mocnín všetkých náhodných veličín X_i má χ^2 rozdelenie s n stupňami voľnosti:

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2 \tag{6.20}$$

Hodnoty x jednotlivých výberových súborov môžeme transformovať na štandardnú normálnu veličinu z a umocniť na druhú. Potom pre χ^2 rozdelenie platí:

$$\chi^2_{(n)} = \left(\frac{X_1 - \mu}{\sigma}\right)^2 + \left(\frac{X_2 - \mu}{\sigma}\right)^2 + \ldots + \left(\frac{X_n - \mu}{\sigma}\right)^2 \tag{6.21}$$

Na obrázku 6.20 sú znázornené teoretické χ^2 rozdelenia s rôznymi počtami stupňov voľnosti.

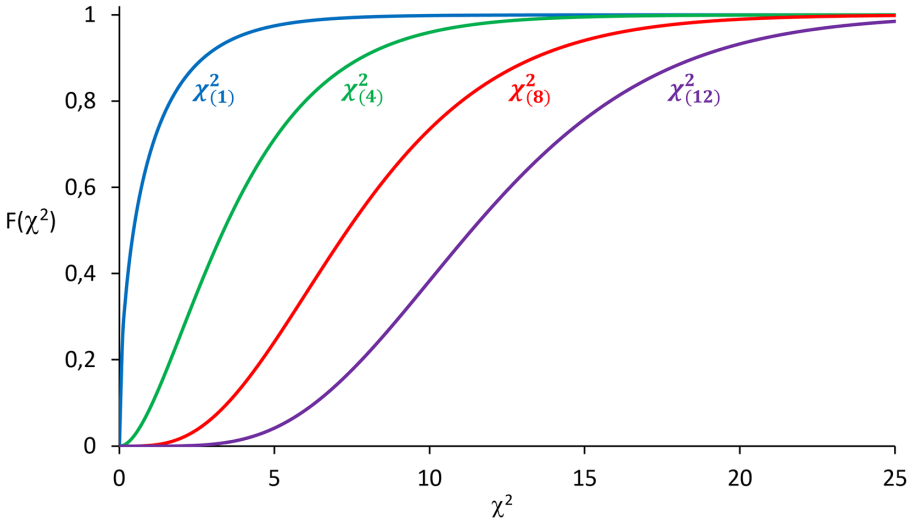


Obr. 6.20: Funkcie hustoty χ^2 rozdelenia pre rôzne počty stupňov voľnosti.

Pripomíname, že u štandardného normálneho rozdelenia Z leží približne 68% rozdelenia medzi hodnotami -1 a +1 (priemer je rovný 0 a smerodajná odchýlka 1). Analogicky, 68% χ^2 rozdelenia s jedným stupňom voľnosti (Z^2 rozdelenie) leží medzi hodnotami 0 a +1. Zvyšných 32% χ^2 rozdelenia leží medzi +1 a ∞. Preto je funkcia hustoty χ^2 rozdelenia pozitívne zošikmená.

Tvar funkcie hustoty χ^2 rozdelenia sa so zmenou počtu stupňov voľnosti mení. Čím je počet stupňov voľnosti väčší, tým viac sa šikmosť rozdelenia zmenšuje a približuje sa k normálnemu rozdeleniu. Priemer μ je rovný počtu stupňov voľnosti a rozptyl σ^2 je dvojnásobkom počtu stupňov voľnosti. Vychádzajúc z definície χ^2 rozdelenia je zrejmé, že súčet dvoch alebo viacerých nezávislých veličín s χ^2 rozdelením bude mať tiež χ^2 rozdelenie.

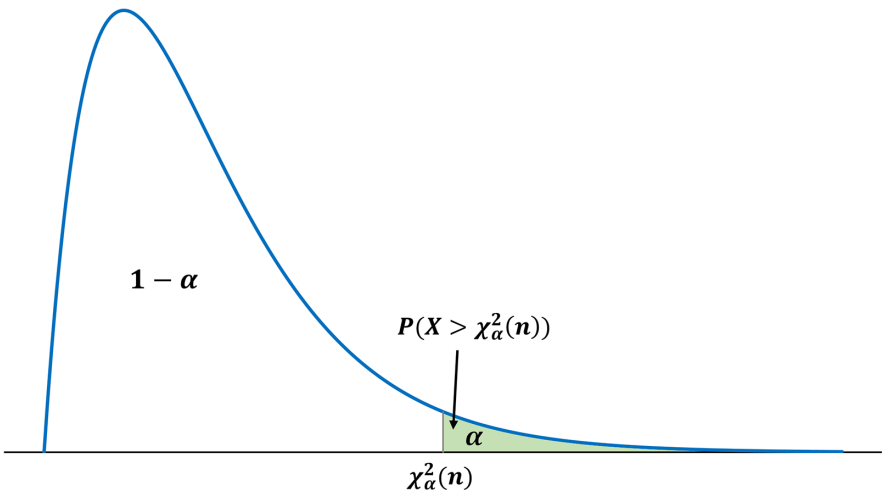
Na obrázku 6.21 sú znázornené rôzne distribučné funkcie χ^2 rozdelenia, z ktorých rovnako ako pri iných rozdeleniach vieme určiť pravdepodobnosť, že sledovaná hodnota náhodnej veličiny bude rovná alebo menšia ako daná hodnota.



Obr. 6.21: Distribučné funkcie χ^2 rozdelenia pre rôzne počty stupňov voľnosti.

Vo väčšine prípadov použitia χ^2 rozdelenia nás zaujímajú jeho kritické hodnoty (viď testovanie hypotéz), ktoré predstavujú takú hodnotu náhodnej veličiny X, ktorá ohraničuje oblasť pod krivkou funkcie hustoty napravo od tejto hodnoty, pričom veľkosť oblasti je rovná hodnote α (obrázok 6.22). Hodnota kvantilu $\chi^2_\alpha(n)$ teda pre tento prípad špecifikuje pravdepodobnosť:

$$P\left(X > \chi^2_\alpha(n)\right) = \alpha \tag{6.22}$$



Obr. 6.22: χ^2 rozdelenie a ohraničená oblasť α.

Hodnoty kvantilov χ^2 rozdelenia bývajú taktiež dostupné v štatistických tabuľkách, ktoré sú špecifikované pre rôzne hodnoty n a α. Predpokladajme, že máme náhodnú veličinu s χ^2 rozdelením, ktorá má 6 stupňov voľnosti a α je 0,05. Potom, ak máme pokryť oblasť o veľkosti 0,05 na pravej strane, použijeme na nájdenie hodnoty kvantilu $\chi^2_\alpha(n)$ štatistické tabuľky pravostranného χ^2 rozdelenia (je to vlastne invertovaná tabuľka χ^2 rozdelenia, kedy α=0 začína napravo a smerom doľava sa zväčšuje až kým nepokryje celú plochu pod krivkou χ^2 rozdelenia, t. j. α=1). Ak máme len štatistické tabuľky χ^2 rozdelenia (α=0 začína naľavo), potom ak chceme pokryť oblasť o veľkosti 0,05 na pravej strane, všetko čo je na ľavej strane od hodnoty kvantilu $\chi^2_\alpha(n)$ sa rovná 1-α=1-0,05=0,95. Zo štatistických tabuliek pre χ^2 rozdelenie (obrázok 6.23) nájdeme hodnotu kvantilu $\chi^2_\alpha(n)$, ktorá odpovedá n=6 a α=0,95. Ako vidíme hodnotou kvantilu $\chi^2_{0,95}(6)$ je 12,592, čo znamená, že 5% hodnôt náhodnej veličiny s χ^2 rozdelením a 6 stupňami leží napravo od tejto hodnoty.

d.f.	$\chi^2_{.005}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.0000393	.000982	.00393	2.706	3.841	5.024	6.635	7.879
2	.0100	.0506	.103	4.605	5.991	7.378	9.210	10.597
3	.0717	.216	.352	6.251	7.815	9.348	11.345	12.838
4	.207	.484	.711	7.779	9.488	11.143	13.277	14.860
5	.412	.831	1.145	9.236	11.070	12.832	15.086	16.750
6	.676	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	.989	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	13.362	15.507	17.535	20.090	21.955
⋮								

Obr. 6.23: Výber tabuľkových hodnôt χ^2 rozdelenia.

Funkcie v MS Excel

CHISQ.DIST(x; st_voľnosti; kumulatívne)
Funkcia CHISQ.DIST vráti ľavostrannú pravdepodobnosť χ^2 rozdelenia.

CHISQ.DIST.RT(x; st_voľnosti)
Funkcia CHISQ.DIST.RT vráti pravostrannú pravdepodobnosť χ^2 rozdelenia.

6.2.4 F rozdelenie

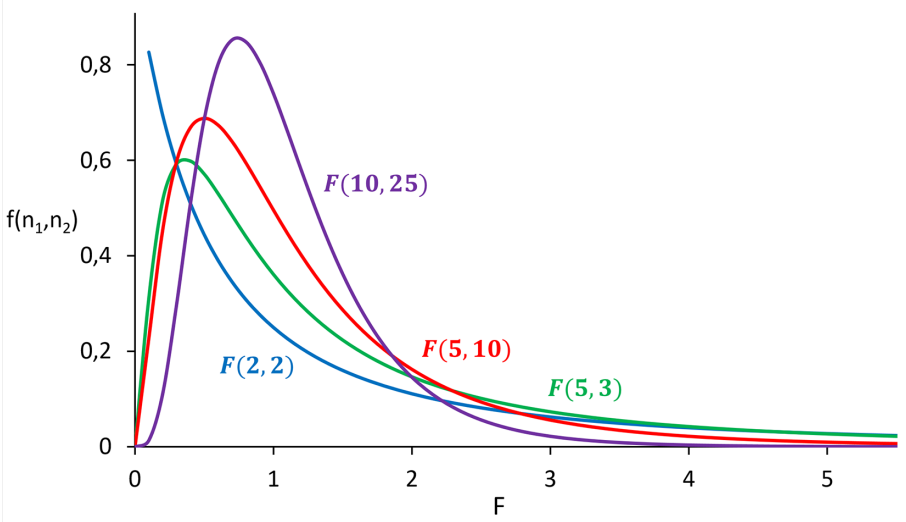
F rozdelenie alebo Fisherovo rozdelenie je ďalším dôležitým a často využívaným rozdelením pri riešení štatistických úloh. F rozdelenie predpokladá vytvorenie spojitej náhodnej veličiny F pomocou dvoch náhodných a nezávislých veličín χ^2_1 a χ^2_2 s χ^2 rozdelením a n_1, resp. n_2 stupňami voľnosti. Potom náhodná veličina F je daná ako ich podiel v tvare:

$$F = \frac{\frac{\chi_1^2}{n_1}}{\frac{\chi_2^2}{n_2}} \tag{6.23}$$

a jej rozdelenie sa nazýva F rozdelenie s n_1 a n_2 stupňami voľnosti.

F rozdelenie je teda charakterizované dvoma parametrami, ktorými sú stupne voľnosti dvoch veličín, pričom na základe týchto dvoch hodnôt je možné nájsť pravdepodobnosti F rozdelenia. Vo všeobecnosti F rozdelenie využívame na porovnávanie rozptylov výberových súborov so širokým spektrom možností na porovnanie dvoch alebo viacerých súborov údajov.

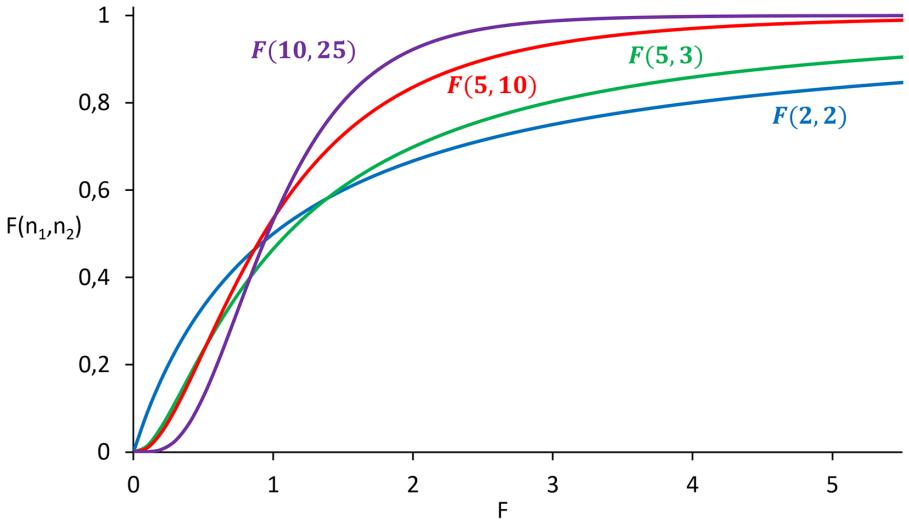
Funkcia hustoty F rozdelenia nemá jednoduchý matematický vzorec, ale väčšina štatistických výpočtových nástrojov dokáže tabuľky rozdelenia vypočítať. Hodnoty F rozdelenia sú v tabuľkách indexované stupňami voľnosti súvisiacimi s rozptylom použitým v menovateli a rozptylom použitým v čitateli. Na obrázku 6.24 sú zobrazené F rozdelenia pre niekoľko kombinácií rôznych stupňov voľnosti čitateľa a menovateľa F rozdelenia.



Obr. 6.24: Funkcie hustoty F rozdelenia pre rôzne kombinácie stupňov voľnosti.

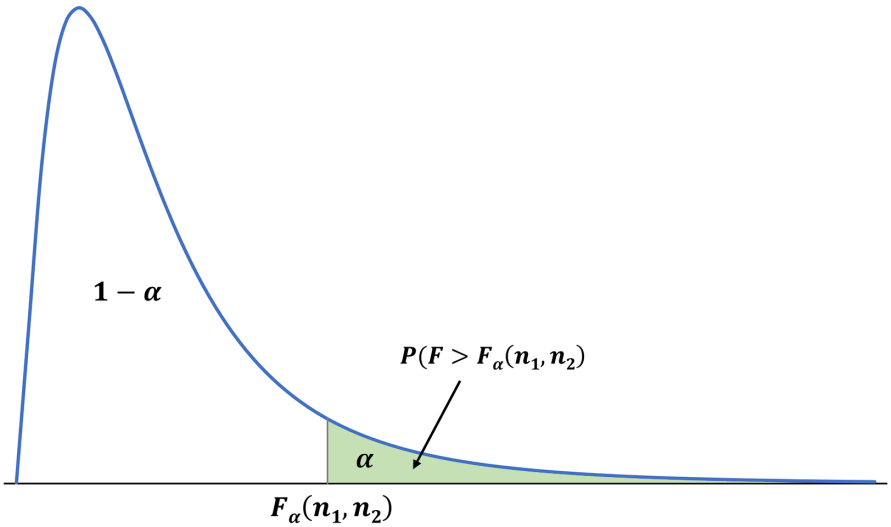
Keďže F rozdelenie využíva ako premenné veličiny s χ^2 rozdelením, má aj podobné vlastnosti ako χ^2 rozdelenie. Hodnoty F rozdelenia sú definované na intervale od 0 do ∞, t. j. pre kladné hodnoty. Tvar F rozdelenia má pravostrannú šikmosť, ktorá sa so zvyšovaním počtov stupňov voľnosti zmenšuje.

Na obrázku 6.25 sú znázornené distribučné funkcie F rozdelenia s rovnakými kombináciami stupňov voľnosti ako v prípade predošlého obrázku s funkciami hustoty F rozdelenia.



Obr. 6.25: Distribučné funkcie F rozdelenia pre rôzne kombinácie stupňov voľnosti.

F rozdelenie sa prevažne používa na určenie kritických oblastí pri testovaní štatistických hypotéz a tiež pri určovaní intervalov spoľahlivosti (viď ďalšie kapitoly). Ako príklady možno uviesť analýzu rozptylu a F test na určenie, či sú rozptyly dvoch základných súborov rovnaké.



Obr. 6.26: F rozdelenie a ohraničená oblasť α.

Ak potrebujeme vymedziť kritickú oblasť F rozdelenia s počtom stupňov voľnosti n_1 a n_2, potom potrebujeme určiť jeho hodnotu kvantilu, ktorá oddeľuje hodnoty rozdelenia napravo, v strede alebo naľavo (v závislosti od konkrétnej testovanej hypotézy). Zoberme si identický prípad ako v prípade χ^2 rozdelenia a predpokladajme, že nás zaujíma získanie hodnoty kvantilu náhodnej spojitej premennej F, ktorá pokrýva oblasť rovnajúcu sa α napravo (tak ako je to zo-

brazené na obrázku 6.26). Takúto hodnotu kvantilu označíme $F_\alpha(n_1, n_2)$ a pre pravdepodobnosť oblasti napravo platí, že:

$$P(F > F_\alpha(n_1, n_2)) = \alpha \tag{6.24}$$

Hodnoty kvantilov F rozdelenia opäť nájdeme v štatistických tabuľkách. Predpokladajme, že náhodná veličina F má F rozdelenie s počtom stupňov voľnosti $n_1 = 6$ a počtom stupňov voľnosti $n_2 = 8$, t. j. $F(6, 8)$. Oblasť napravo od hodnoty kvantilu má byť $\alpha = 0,05$. Keďže kritická oblasť je na pravej strane, potrebujeme použiť štatistické tabuľky pre pravostranné F rozdelenie (hodnota α sa bude zvyšovať sprava doľava). Väčšinou máme len tabuľky pre štandardné F rozdelenie (obrázok 6.27), t. j. na hľadanie hodnoty kvantilu použijeme tie, ktoré platia pre $1 - \alpha = 1 - 0,05 = 0,95$.

$F_{.95}$

n_2 \ n_1	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	13.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
⋮									
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
⋮									

Obr. 6.27: Výber tabuľkových hodnôt F rozdelenia pre $\alpha = 0,95$.

Hodnota náhodnej premennej F, ktorá vymedzuje oblasť rovnajúcu sa $\alpha = 0,05$ napravo od tejto hodnoty, je hodnota kvantilu $F_{0,05}(6, 8) = 3,58$. Poznamenajme, že pre takéto prípady je vždy potrebné uviesť stranu vymedzenia hodnôt (napravo), inak môže byť hodnota 0,05 považovaná za oblasť vymedzenú vľavo.

Funkcie v MS Excel

F.DIST(x; st_voľnosti1; st_voľnosti2; kumulatívne)
Funkcia F.DIST vráti ľavostrannú pravdepodobnosť F rozdelenia pre dve skupiny údajov.

F.DIST.RT(x; st_voľnosti1; st_voľnosti2)
Funkcia F.DIST.RT vráti pravostrannú pravdepodobnosť F rozdelenia pre dve skupiny údajov.

Kapitola 7

Štatistické odhady

Štatistické odhady, alebo skrátene odhady, sú jednou z dvoch základných oblastí induktívnej štatistiky (tou druhou je testovanie štatistických hypotéz). Pričom vieme, že induktívna štatistika predstavuje spôsob, akým sa môžme dopracovať k záveru o základnom súbore na základe informácií, ktoré sú obsiahnuté vo výberovom súbore vybranom z tohoto základného súboru.

Výberový súbor je vyberaný zo základného súboru náhodne, a preto náhodný výberový súbor o veľkosti n je závislý od metódy výberu vzorky s vopred stanovenou pravdepodobnosťou. V predchádzajúcich úvahách sme ukázali, že výberové rozdelenie poskytuje rozdelenie pravdepodobnosti sledovanej štatistiky, ktoré zovšeobecňujeme pre základný súbor. Keďže výberové rozdelenie je založené na všetkých možných vzorkách zo základného súboru o veľkosti N, potom očakávanú hodnotu, napríklad rozptyl či ďalšie charakteristiky základného súboru možno teoreticky získať zo štatistík náhodného výberového súboru.

Odhady ako také, sa vzťahujú na metódy zisťovania hodnoty štatistiky výberového súboru zodpovedajúcej hodnote základného súboru tak, aby táto štatistika spĺňala vlastnosti reprezentujúce parameter. Na druhej strane, akonáhle získame hodnotu pre odhad zodpovedajúcej hodnoty základného súboru, je potrebné vedieť, či sa hodnota výberového súboru použitá ako odhad hodnoty základného súboru blíži skutočnej hodnote. Tento problém je v štatistike veľmi dôležitý, a preto existujú rôzne techniky používané na rozhodovanie o zovšeobecnení odhadov parametrov základného súboru. Keďže odhady sú získané z náhodne vybraných výberových súborov, je pravdepodobné, že hodnoty, ktoré považujeme za charakteristiky základného súboru sa môžu líšiť od skutočnej hodnoty. To znamená, že hoci je výberový súbor vybraný zo základného súboru, tak odhad sa nemusí presne rovnať zodpovedajúcemu parametru. Aj preto je potrebné poznať postupy, ktoré nám umožnia vysloviť závery o hodnotách základného súboru.

Využitie odhadov v oblasti medicíny nie je potrebné nejak zvlášť zdôrazňovať. Mnohé cieľové skupiny, aj keď sú konečné, sú tak veľké, že realizácia 100-percentného pokrytia (celý základný súbor) by bola z hľadiska nákladov takmer nemožná. Ale sú tu aj ďalšie faktory. Napríklad, lekára môže zaujímať aká časť určitého typu pacientov, liečených konkrétnym liekom trpí nežiaducimi vedľajšími účinkami. Je jasné, že predstava o základnom súbore pozostáva zo všetkých osôb, ktoré kedy boli alebo niekedy budú liečené týmto liekom. Čakanie na odvodenie záverov, kým nebudú získané informácie od všetkých pacientov, by mohlo mať nepriaznivý vplyv na pacientov, alebo aj na samotnú klinickú prax lekára.

Predmetom štatistických odhadov sú najčastejšie priemer základného súboru a rozptyl základného súboru. Odhadovať tiež môžme rozdiely medzi dvoma priemermi, rozdiely dvoch rozptylov, či pomer dvoch rozptylov.

Vo všeobecnosti je možné metódy odhadov parametrov základného súboru zo štatistík výberového súboru v induktívnej štatistike rozdeliť do dvoch skupín, ktoré definujú:

- **bodový odhad** – odhad pozostávajúci z jedného čísla, vypočítaného z výberového súboru, ktoré predstavuje najlepší odhad neznámeho parametra,
- **intervalový odhad** – odhad pozostávajúci z rozsahu čísel okolo bodového odhadu, do ktorého predpokladáme, že daný parameter spadá.

7.1 Bodový odhad

Bodový odhad (proces) neznámeho parametra základného súboru je štatistika, ktorá odhaduje hodnotu tohto parametra. Bodovým odhadom (hodnota) parametra je hodnota štatistiky, ktorá sa používa na odhad parametra.

Zmyslom bodového odhadu je vypočítať z údajov výberového súboru jediné číslo, ktoré je pravdepodobne najbližšie k neznámej hodnote parametra základného súboru. Predpokladáme, že dostupné informácie sú vo forme náhodného výberového súboru zo základného súboru, pričom zo základného súboru je možné vybrať rôzne výberové súbory $X_1, X_2, \ldots, X_n$ o veľkosti n. Cieľom je sformulovať štatistiku tak, aby sa jej hodnota vypočítaná z údajov výberového súboru čo najviac priblížila hodnote parametra základného súboru.

Predpokladajme teda, že náhodným výberom zo základného súboru získame n-ticu navzájom nezávislých náhodných premenných $X_1, X_2, \ldots, X_n$, z ktorých je potom možné vypočítať napríklad nasledovné najčastejšie používané výberové charakteristiky.

Výberový priemer $\bar{X}$ vypočítaný ako:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.1}$$

a ktorý sa používa ako odhad priemeru základného súboru.

Výberový rozptyl S^2 vypočítaný ako:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{7.2}$$

a ktorý sa používa ako odhad rozptylu základného súboru.

Výberová smerodajnú odchýlka S vypočítaná ako:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \tag{7.3}$$

a ktorá sa používa ako odhad smerodajnej odchýlky základného súboru.

Pri praktickej realizácii experimentálnych úloh máme spravidla dostupný len jeden výberový súbor zo základného súboru. Potom, napríklad na odhad priemernej hodnoty základného súboru μ (vzťah (3.1)) je pravdepodobne najlepším bodovým odhadom priemer výberového súboru $\bar{x}$ (vzťah (3.2)). Preto, ak máme pozorované hodnoty $x_1, x_2, \ldots, x_n$ náhodnej veličiny X, ktorá má pravdepodobnostné rozdelenie so strednou hodnotou μ a rozptylom σ^2, potom z týchto hodnôt môžeme vypočítať pozorovanú hodnotu výberového priemeru $\bar{x}$, ktorý nazývame bodovým odhadom priemeru základného súboru μ.

Podobne, pravdepodobne najlepším bodovým odhadom rozptylu základného súboru σ^2 (vzťah (3.19)) je rozptyl výberového súboru s^2 (vzťah (3.20)) a najlepším bodovým odhadom smerodajnej odchýlky základného súboru σ (vzťah (3.21)) je smerodajná odchýlka výberového súboru s (vzťah (3.22)).

Kvalitu bodového odhadu určuje jeho:

- neskreslenosť – stredná hodnota odhadu je rovná skutočnej hodnote odhadovaného parametra, t. j. požadujeme, aby stredná hodnota chyby odhadu bola rovná nule,
- konzistentnosť – s rastúcim rozsahom n náhodného výberu rastie aj pravdepodobnosť, že sa bude hodnota bodového odhadu veľmi málo líšiť od skutočnej hodnoty odhadovaného parametra,
- výdatnosť – zo všetkých charakteristík, ktoré poskytujú neskreslený bodový odhad parametra má najmenší rozptyl.

7.2 Intervalový odhad

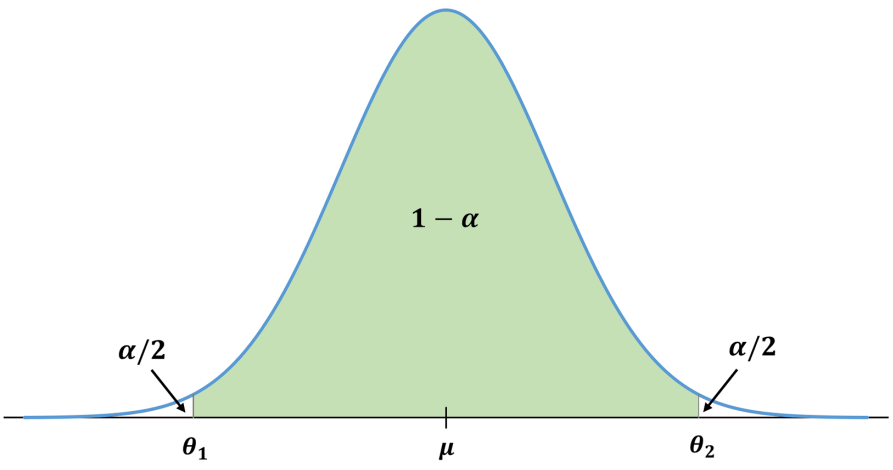
Z bodového odhadu diskutovaného v predchádzajúcej kapitole vieme, že parameter základného súboru predstavuje jedna hodnota. Použitie bodového odhadu na zovšeobecnenie výsledku z výberového súboru na základný súbor je relatívne jednoduché a pohodlné, ale neposkytuje niektoré dôležité informácie, napríklad tie, ktoré sa týkajú veľkosti chyby spojenej s bodovým odhadom. Inými slovami povedané, ak je bodový odhad sprevádzaný chybou spojenou s bodovým odhadom, potom je dôležitá aj spoľahlivosť odhadu a potom k takémuto bodovému odhadu potrebujeme priradiť aj mieru jeho spoľahlivosti. Miera spoľahlivosti odhadu je zohľadnená v intervalovom odhade, vyjadrenom takým intervalom (intervalom spoľahlivosti), ktorý pokrýva hľadaný parameter s vopred definovanou (dostatočne veľkou) presnosťou.

Intervalový odhad sledovanej veličiny, označme ju napríklad Θ, predstavuje ohraničený interval $(\theta_1; \theta_2)$, ktorý obsahuje skutočnú hodnotu Θ s pravdepodobnosťou $1 - \alpha$, a teda platí, že:

$$P\left(\theta_1 < \Theta < \theta_2\right) = 1 - \alpha \tag{7.4}$$

kde $1 - \alpha$ sa nazýva **hladina spoľahlivosti**, ktorá určuje spoľahlivosť odhadu, θ_1 je dolná hranica intervalu spoľahlivosti a θ_2 je horná hranica intervalu spoľahlivosti.

Interval spoľahlivosti $(\theta_1; \theta_2)$ označujeme aj ako obojstranný interval spoľahlivosti, keďže ohraničuje oblasť pokrytia hodnôt z oboch strán rozdelenia sledovanej veličiny. Oblasť obojstranného intervalu spoľahlivosti je znázornená na obrázku 7.1.



Obr. 7.1: Obojstranný interval spoľahlivosti.

Pravdepodobnosť α predstavuje **hladinu významnosti**, ktorá určuje tzv. riziko odhadu, t. j. pravdepodobnosť, že interval spoľahlivosti neobsahuje skutočnú hodnotu. V mnohých štúdiách, vrátane tých medicínskych sa hladina významnosti α volí spravidla 0,05 alebo 0,01. Pre hodnoty mimo intervalu spoľahlivosti platí:

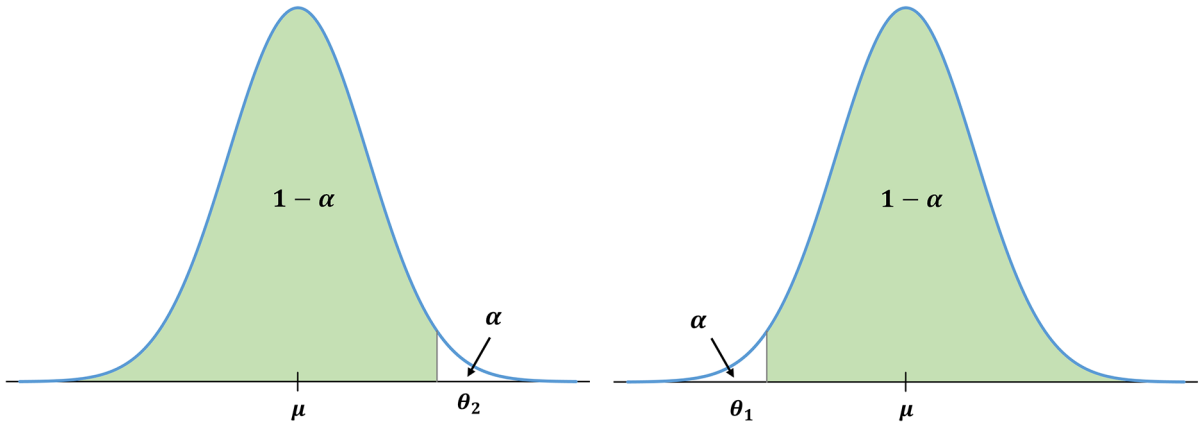
$$P(\Theta \leq \theta_1) = P(\Theta \geq \theta_2) = \frac{\alpha}{2} \tag{7.5}$$

Interval spoľahlivosti je možné vyjadriť aj v percentách, a to tak, že hladinu spoľahlivosti vynásobíme číslom 100, t. j. $(1-\alpha)\,100\%$.

Podobne môžeme definovať aj jednostranný interval spoľahlivosti, ktorý ohraničuje oblasť pokrytia hodnôt odhadovaného parametra len z jednej strany. Horný, resp. pravostranný interval spoľahlivosti potom vymedzuje oblasť, pre ktorú platí:

$$P(\Theta < \theta_2) = 1 - \alpha \tag{7.6}$$

kde θ_2 je horná hranica jednostranného intervalu spoľahlivosti, ktorý zapíšeme ako $(-\infty; \theta_2)$ a pokrýva oblasť $1-\alpha$ náhodnej veličiny Θ tak, ako je to znázornené na obrázku 7.2 vľavo.



Obr. 7.2: Pravostranný (vľavo) a ľavostranný (vpravo) interval spoľahlivosti.

Pre hodnoty mimo horného intervalu spoľahlivosti platí:

$$P(\Theta \geq \theta_2) = \alpha \tag{7.7}$$

Ľavostranný (dolný) interval spoľahlivosti zasa vymedzuje oblasť, pre ktorú platí:

$$P(\Theta > \theta_1) = 1 - \alpha \tag{7.8}$$

kde θ_1 je dolná hranica jednostranného intervalu spoľahlivosti, ktorý zapíšeme ako $(\theta_1; \infty)$ a pokrýva oblasť $1-\alpha$ náhodnej veličiny Θ tak, ako je to znázornené na obrázku 7.2 vpravo.

Pre hodnoty mimo dolného intervalu spoľahlivosti platí:

$$P(\Theta \leq \theta_1) = \alpha \tag{7.9}$$

Je dôležité si uvedomiť, že θ_1 a θ_2 sú premenné náhodnej veličiny, a tomu má odpovedať aj interpretácia intervalu spoľahlivosti. Keďže θ, ako hodnota náhodnej veličiny Θ zostáva konštantná, je parametrom, a môžeme ho interpretovať iba z hľadiska toho, či je parameter θ zahrnutý v intervale spoľahlivosti alebo nie. Navyše, interval vytvorený zo samostatného výberového súboru nemôže byť použitý na deklarovanie pravdepodobnosti, ale musí sa interpretovať s ohľadom na vytvorenie podobných intervalov spoľahlivosti pre každý výberový súbor opakovaný veľakrát s rovnakou veľkosťou a z rovnakého základného súboru. Inak povedané, výberové rozdelenie sledovanej veličiny nám poskytuje možnosti pre odvodenie a interpretáciu intervalu spoľahlivosti.

7.2.1 Intervalový odhad strednej hodnoty

Intervalový odhad strednej hodnoty (priemeru) základného súboru μ predstavuje interval spoľahlivosti $(\theta_1; \theta_2)$, ktorý obsahuje skutočnú hodnotu μ s pravdepodobnosťou $1 - \alpha$. Predpokladáme, že hranice intervalu spoľahlivosti sú premenné, a teda pre rôzne výberové súbory predstavujú rôzne hodnoty.

Pri intervalovom odhade strednej hodnoty základného súboru môžu vo všeobecnosti nastať dva prípady:

- poznáme rozptyl základného súboru – teoretické východisko
- nepoznáme rozptyl základného súboru – častejší prípad

7.2.1.1 Riešenie ak poznáme rozptyl základného súboru

Predpokladajme, že náhodná veličina X reprezentujúca základný súbor má normálne rozdelenie $N(\mu, \sigma^2)$ so strednou hodnotou μ a rozptylom σ^2 (alebo smerodajnou odchýlkou σ), ktorý poznáme. Ďalej predpokladajme, že $X_1, X_2, \ldots, X_n$ predstavuje náhodný výber rozsahu n z X (séria náhodných výberových súborov zo základného súboru). Výberovým odhadom strednej hodnoty μ je potom výberový priemer $\bar{X}$ (vzťah (7.1)), ktorý má normálne rozdelenie $N\left(\mu, \frac{\sigma^2}{n}\right)$. Tiež vieme, že pre štandardné normálne rozdelenie $N(0, 1)$

máme definovanú náhodnú premennú Z (vzťah (6.15)). Pre náhodný výber platí, že náhodná premenná Z bude daná vzťahom:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{7.10}$$

Potom, pravdepodobnosť $1 - \alpha$, ktorá predstavuje interval spoľahlivosti, môže byť získaná z nasledovného intervalu:

$$P\left(z_{\alpha/2} < Z < z_{1-\alpha/2}\right) = 1 - \alpha \tag{7.11}$$

kde $z_{\alpha/2}$ a $z_{1-\alpha/2}$ sú hodnoty kvantilov náhodnej veličiny Z so štandardným normálnym rozdelením $N(0, 1)$ a získame ich zo štatistických tabuliek pre príslušnú hodnotu α.

Úpravou vzťahu (7.11) dostaneme pre odhad strednej hodnoty μ vzťah:

$$P\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{7.12}$$

kde $\bar{x}$ predstavuje priemernú hodnotu (strednú) vypočítanú z konkrétneho výberového súboru s rozsahom n. Hodnota $\frac{\sigma}{\sqrt{n}}$ nám určuje strednú (štandardnú) chybu priemeru (označujeme SE).

Keďže štandardné normálne rozdelenie má strednú hodnotu rovnú nule, potom platí, že hodnoty kvantilov na opačných stranách rozdelenia náhodnej veličiny majú rovnaké hodnoty, ale s opačným znamienkom. Platí teda, že $z_{\alpha/2} = -z_{1-\alpha/2}$. Obojstranný interval spoľahlivosti potom pri známom rozptyle σ^2, bez ohľadu na veľkosť výberového súboru, určujeme podľa dolnej a hornej hranice takto:

$$\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{7.13}$$

Analogicky môžme odvodiť hranice jednostranných intervalov spoľahlivosti. V prípade dolnej hranice pri ľavostrannom intervale spoľahlivosti, hľadáme hodnotu kvantilu štandardného normálneho rozdelenia na úrovni α (pozor, nedelíme číslom 2, keďže všetky možnosti ležia na jednej strane). Ľavostranný interval spoľahlivosti je potom daný vzťahom:

$$\left(\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \infty\right) \tag{7.14}$$

Pre hornú hranicu pravostranného intervalu spoľahlivosti hľadáme úroveň, ktorá vylučuje oblasť rozdelenia o veľkosti α na pravej strane rozdelenia hodnôt

náhodnej veličiny, preto pravostranný interval spoľahlivosti budeme určovať nasledovným vzťahom:

$$\left(-\infty; \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) \tag{7.15}$$

Príklad 7.1. Náhodným výberom bolo vybraných 100 novorodencov. Ich priemerná hmotnosť je 3,45 kg. Na základe dlhodobých pozorovaní bolo zistené, že smerodajná odchýlka hmotnosti σ je 0,45 kg. 95-percentným intervalom spoľahlivosti odhadnime priemernú hmotnosť novorodenca.

Zo zadaných údajov je zrejmé, že veľkosť výberového súboru je n=100, priemerná hodnota hmotnosti $\bar{x}$=3,45 kg a smerodajná odchýlka základného súboru je σ=0,45 kg. Odhad intervalu spoľahlivosti máme urobiť na hladine významnosti α=1-0,95=0,05 (5% riziko chyby odhadu pre 95% interval spoľahlivosti). Pre odhad obojstranného intervalu spoľahlivosti potrebujeme zistiť hodnoty kvantilov štandardného normálneho rozdelenia. Nájdeme ich v štatistických tabuľkách (obrázok 7.3).

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
−3.70	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.70
⋮											
−2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	−2.00
−1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	−1.90
−1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	−1.80
⋮											

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.10
⋮											
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706	1.80
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767	1.90
2.00	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817	2.00
⋮											

Obr. 7.3: Výber tabuľkových hodnôt štandardného normálneho rozdelenia.

Pri hľadaní hodnoty kvantilu štandardného normálneho rozdelenia postupujeme tak, že pre známu hodnotu pravdepodobnosti hľadáme skóre, ktorému táto pravdepodobnosť patrí. Ak je hodnota α=0,05, potom hľadáme pre obojstranný interval spoľahlivosti hodnoty kvantilov $z_{\alpha/2}$ a $z_{1-\alpha/2}$. Pre $\alpha/2$=0,05/2=0,025 nachádzame v tabuľke na obrázku 7.3 hore, že $z_{0,025}$=(-1,90)+(-0,06)=-1,96.

Podobne pre $1 - \alpha/2$=1-0,05/2=0,975 nachádzame v tabuľke na obrázku 7.3 dole, že $z_{0,975}$=1,90+0,06=1,96. Vidíme, že obe hodnoty kvantilov sú rovnaké (s opačným znamienkom, viď úvaha vyššie), čo je dané tým, že štandardné normálne rozdelenie je symetrické so stredom v nule.

Dolnú hranicu intervalu spoľahlivosti potom vypočítame ako:

$$\theta_1 = 3,45 - 1,96\frac{0,45}{\sqrt{100}} \doteq 3,36$$

a horná hranica bude:

$$\theta_2 = 3,45 + 1,96\frac{0,45}{\sqrt{100}} \doteq 3,54$$

95% interval spoľahlivosti priemernej hmotnosti novorodencov je (3,36; 3,54) kg. Inak povedané 3,36 až 3,54 kg je interval, v ktorom s 95% pravdepodobnosťou leží priemerná hodnota hmotnosti novorodencov.

Funkcie v MS Excel

NORM.S.DIST(z; kumulatívne)
Funkcia NORM.S.DIST vráti hodnotu štandardného normálneho rozdelenia (stredná hodnota je rovná 0 a smerodajná odchýlka 1).

NORM.S.INV(pravdepodobnosť)
Funkcia NORM.S.INV vráti inverznú hodnotu kumulatívnej funkcie štandardného normálneho rozdelenia (stredná hodnota je rovná 0 a smerodajná odchýlka 1).

7.2.1.2 Riešenie ak nepoznáme rozptyl základného súboru

V mnohých experimentálnych a výskumných úlohách pracujeme len s údajmi výberového súboru a rozptyl základného súboru σ^2 nepoznáme. Pre určenie intervalu spoľahlivosti potom musíme brať do úvahy veľkosť výberového súboru n, pričom môžu nastať dva prípady:

- výberový súbor je dostatočne veľký – $n > 30$
- výberový súbor je malý – $n \leq 30$

A) Ak je výberový súbor dostatočne veľký, potom teorém o centrálnom rozložení údajov predpokladá, že platí definícia pravdepodobnosti daná vzťahom (7.12), a to bez zásadného vplyvu rozdelenia základného súboru. Tiež predpokladáme, že v prípade veľkého n, nahradenie strednej chyby priemeru $\frac{\sigma}{\sqrt{n}}$ odhadom $\frac{s}{\sqrt{n}}$ výrazne neovplyvní vyššie uvedený výrok o pravdepodobnosti.

Preto platí, ak je n veľké a σ^2 nie je známe, potom $100\,(1-\alpha)\,\%$-tný interval spoľahlivosti pre strednú hodnotu μ je daný ako:

$$\left(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right) \tag{7.16}$$

kde s je smerodajná odchýlka výberového súboru, teda smerodajná odchýlka vypočítaná z údajov konkrétneho výberového súboru.

Pre jednostranné intervaly spoľahlivosti platia rovnaké predpoklady, a preto ľavostranný interval spoľahlivosti je daný vzťahom:

$$\left(\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}}; \infty\right) \tag{7.17}$$

Podobne pravostranný interval spoľahlivosti budeme určovať vzťahom:

$$\left(-\infty; \bar{x} + z_{1-\alpha}\frac{s}{\sqrt{n}}\right) \tag{7.18}$$

Príklad 7.2. Predpokladajme, že počas preventívnych prehliadok u pediatra bola zistená výška 250 náhodne vybraných 10 ročných chlapcov z populácie všetkých 10 ročných chlapcov. Ich priemerná výška bola 136,2 cm a rozptyl 41,56 cm^2. 99-percentným intervalom spoľahlivosti odhadnime priemernú výšku 10 ročných chlapcov.

Z uvedených informácií vieme, že máme dostupné len údaje výberového súboru, ktorý je dostatočne veľký, pričom n=250, priemerná hodnota výšky $\bar{x}$=136,2 cm a smerodajná odchýlka $s = \sqrt{41,56} \doteq 6{,}45$ cm. Hladina významnosti $\alpha = 1-0,99 = 0,01$. Hodnoty kvantilov potrebujeme nájsť pre $z_{0,005}$ alebo $z_{0,995}$. V štatistických tabuľkách štandardného normálneho rozdelenia (obrázok 7.4) nájdeme pre hornú hranicu len hodnoty 0,9949 a 0,9951. Keďže hodnota 0,995 leží uprostred, potom $z_{0,995}$=2,575 (môže byť zaokrúhlená na 2,58).

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
⋮											
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	2.40
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	2.50
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	2.60
⋮											

Obr. 7.4: Výber tabuľkových hodnôt štandardného normálneho rozdelenia.

Dolnú hranicu intervalu spoľahlivosti potom vypočítame ako:

$$\theta_1 = 136,2 - 2,575\frac{6,45}{\sqrt{250}} \doteq 135,2$$

a horná hranica bude:

$$\theta_2 = 136,2 + 2,575\frac{6,45}{\sqrt{250}} \doteq 137,3$$

Z výberového súboru náhodne vybraných 250 desaťročných chlapcov sme odhadli, že 99% interval spoľahlivosti priemernej výšky 10 ročných chlapcov je (135,2; 137,3) cm.

B) Ak je výberový súbor malý, potom už vyššie uvedené predpoklady neplatia a nemôžeme použiť výberové charakteristiky štandardného normálneho rozdelenia Z. Namiesto nich použijeme štatistiku Studentovho t-rozdelenia, ktoré je vhodné práve pre malé súbory (viď kapitola 6.2.2). Táto štatistika má tvar:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{7.19}$$

kde $\bar{X}$ a S sú charakteristiky výberového rozdelenia.

Pravdepodobnosť $1 - \alpha$ vieme odvodiť podľa:

$$P\left(t_{(n-1),\alpha/2} < T < t_{(n-1),1-\alpha/2}\right) = 1 - \alpha \tag{7.20}$$

kde $t_{(n-1),\alpha/2}$ a $t_{(n-1),1-\alpha/2}$ sú hodnoty kvantilov náhodnej veličiny T s t-rozdelením a získame ich zo štatistických tabuliek pre príslušnú hodnotu stupňov voľnosti $n - 1$ a α. Keďže aj Studentovo t-rozdelenie je symetrické okolo nuly, budú hodnoty kvantilov rovnaké, ale s opačným znamienkom.

Úpravou vzťahu (7.20) získame hranice intervalu spoľahlivosti pre priemer základného súboru μ, ak rozptyl σ^2 nepoznáme a výberový súbor je malý, takto:

$$\left(\bar{x} - t_{(n-1),1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{(n-1),1-\alpha/2}\frac{s}{\sqrt{n}}\right) \tag{7.21}$$

kde priemer $\bar{x}$ a smerodajná odchýlka s sú štatistiky vypočítané z konkrétneho výberového súboru s rozsahom n.

Pre jednostranné intervaly spoľahlivosti strednej hodnoty základného súboru μ odhadované z malých súborov, ak nepoznáme rozptyl základného súboru

σ^2 použijeme tiež predpoklady založené na Studentovom t-rozdelení. Potom ľavostranný interval spoľahlivosti bude daný vzťahom:

$$\left(\bar{x} - t_{(n-1),1-\alpha}\frac{s}{\sqrt{n}}; \infty\right) \tag{7.22}$$

Podobne pravostranný interval spoľahlivosti budeme určovať vzťahom:

$$\left(-\infty; \bar{x} + t_{(n-1),1-\alpha}\frac{s}{\sqrt{n}}\right) \tag{7.23}$$

Príklad 7.3. V príklade 3.1 sme zisťovali priemernú hmotnosť stredoškolákov. Vráťme sa teraz k údajom z tohto príkladu a 95-percentným intervalom spoľahlivosti odhadnime priemernú hmotnosť stredoškolákov.

Z dostupných informácií vieme, že priemerná hodnota hmotnosti $\bar{x}$=56 kg, smerodajná odchýlka s=7,62 kg a veľkosť súboru n=30 je malá. Hodnoty kvantilov hľadáme pre počet stupňov voľnosti $(n-1)$=30-1=29 a hladina významnosti α=0,05. Potrebnú hodnotu kvantilu $t_{(29),0,975}$ nájdeme v štatistických tabuľkách (obrázok 7.5), a má hodnotu 2,0452.

d.f.	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
⋮					
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
⋮					

Obr. 7.5: Výber tabuľkových hodnôt Studentovho t-rozdelenia.

Dolnú hranicu intervalu spoľahlivosti potom vypočítame ako:

$$\theta_1 = 56 - 2,0452\frac{7,62}{\sqrt{30}} \doteq 53,15$$

a horná hranica bude:

$$\theta_2 = 56 + 2,0452\frac{7,62}{\sqrt{30}} \doteq 58,85$$

95% interval spoľahlivosti priemernej hodnoty hmotnosti stredoškolákov je (53,15;58,85) kg.

Funkcie v MS Excel

T.DIST.2T(x; st_voľnosti; kumulatívne)
Funkcia T.DIST.2T vráti hodnotu obojstranného Studentovho *t*-rozdelenia.

T.INV.2T(pravdepodobnosť; st_voľnosti)
Funkcia T.INV.2T vráti inverznú hodnotu obojstranného Studentovho *t*-rozdelenia.

7.2.2 Intervalový odhad rozptylu

Úspešnosť vytvorenia intervalu spoľahlivosti pre rozptyl základného súboru σ^2 závisí od schopnosti nájsť vhodné výberové rozdelenie. Avšak, intervaly spoľahlivosti pre σ^2 sú zvyčajne založené na rozdelení vzoriek $(n-1)\,s^2/\sigma^2$. Ak sa výberové súbory veľkosti n vyberajú zo základného súboru s normálnym rozdelením, potom táto charakteristika má rozdelenie známe ako χ^2 (chí kvadrát) rozdelenie s $n-1$ stupňami voľnosti:

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma^2} \tag{7.24}$$

Pri určovaní intervalu spoľahlivosti rozptylu σ^2, potrebujeme podobne ako pri odhade intervalu spoľahlivosti pre μ, definovať oblasť $1-\alpha$. Túto oblasť pre χ^2 rozdelenie s $n-1$ stupňami voľnosti vyberieme tak, aby jeho hodnoty (kvantily $\chi_{(n-1),\alpha/2}$ a $\chi_{(n-1),1-\alpha/2}$) vymedzovali oblasť $\alpha/2$ najmenších hodnôt naľavo a tiež oblasť $\alpha/2$ najväčších hodnôt napravo. Potom pre pravdepodobnosť oblasti $1-\alpha$ platí:

$$P\left(\chi^2_{(n-1),\alpha/2} < \chi^2 < \chi^2_{(n-1),1-\alpha/2}\right) = 1-\alpha \tag{7.25}$$

Po dosadení vzťahu (7.24) do nerovnosti vo vzťahu (7.24) dostávame:

$$\chi^2_{(n-1),\alpha/2} < \frac{(n-1)\,s^2}{\sigma^2} < \chi^2_{(n-1),1-\alpha/2} \tag{7.26}$$

Vzťah (7.26) upravíme tak, aby nám v strede zostal len samotný rozptyl σ^2, t. j. najprv všetky zložky nerovnosti vydelíme hodnotou $(n-1)\,s^2$, potom nerovnosť prevrátime (vymeníme čitateľov a menovateľov) a dostávame:

$$\frac{(n-1)\,s^2}{\chi^2_{(n-1),1-\alpha/2}} < \sigma^2 < \frac{(n-1)\,s^2}{\chi^2_{(n-1),\alpha/2}} \tag{7.27}$$

Upozorňujeme, že pri prevrátení nerovnosti (7.26) sa vymenili znamienka nerovnosti. Ľavá a pravá strana nerovnosti vo vzťahu (7.27) reprezentujú dolnú a hornú hranicu intervalu spoľahlivosti pre rozptyl σ^2, ktorý definujeme nasledovným výrazom:

$$\left(\frac{(n-1)\,s^2}{\chi^2_{(n-1),1-\alpha/2}}; \frac{(n-1)\,s^2}{\chi^2_{(n-1),\alpha/2}} \right) \tag{7.28}$$

Kým intervaly spoľahlivosti pre priemernú hodnotu základného súboru boli symetrické, intervaly spoľahlivosti pre rozptyl základného súboru nie sú symetrické. Je to dané tým, že χ^2 rozdelenie je nesymetrické.

Z intervalu spoľahlivosti pre rozptyl zakladaného súboru vieme odmocninou dolnej a hornej hranice odvodiť interval spoľahlivosti pre smerodajnú odchýlku zakladaného súboru:

$$\left(\sqrt{\frac{(n-1)\,s^2}{\chi^2_{(n-1),1-\alpha/2}}}; \sqrt{\frac{(n-1)\,s^2}{\chi^2_{(n-1),\alpha/2}}} \right) \tag{7.29}$$

Jednostranné intervaly spoľahlivosti pre rozptyl základného súboru vytvoríme vymedzením hornej, resp. dolnej hranice intervalu $1-\alpha$ tak, aby hladina významnosti α, definovala len oblasť rizika odhadu naľavo, resp. napravo. Ľavostranný interval spoľahlivosti pre rozptyl základného súboru potom bude daný výrazom:

$$\left(\frac{(n-1)\,s^2}{\chi^2_{(n-1),1-\alpha}}; \infty \right) \tag{7.30}$$

a pravostranný interval spoľahlivosti pre rozptyl základného súboru bude:

$$\left(-\infty; \frac{(n-1)\,s^2}{\chi^2_{(n-1),\alpha}} \right) \tag{7.31}$$

Analogicky, aj tu získame jednostranné intervaly spoľahlivosti (pravostranný a ľavostranný) pre smerodajnú odchýlku základného súboru druhou odmocninou hodnôt rozptylu určujúcich hranice jednostranných intervalov spoľahlivosti pre rozptyl základného súboru.

Príklad 7.4. Predpokladajme, že v rámci vstupných vyšetrení pacientov na internom oddelení boli sledované, okrem iného, aj hodnoty celkového cholesterolu. Zaznamenané boli hodnoty dvadsiatich náhodne vybraných pacientov a je potrebné odhadnúť 95% interval spoľahlivosti pre rozptyl a smerodajnú odchýlku celkového cholesterolu základného súboru. Zaznamenané údaje sú uvedené v tabuľke 7.1.

Tabuľka 7.1: Hodnoty celkového cholesterolu.

Pacient č. i	**Cholesterol** ***mmol/l***	**Pacient č. i**	**Cholesterol** ***mmol/l***
1	5,00	11	3,37
2	6,50	12	3,39
3	4,07	13	4,19
4	3,23	14	5,19
5	3,65	15	4,00
6	4,86	16	5,13
7	3,74	17	2,99
8	3,49	18	4,31
9	4,11	19	4,91
10	3,34	20	3,94

Z dostupných informácií vieme, že hladina spoľahlivosti $1-\alpha$=0,95 a hladina významnosti α=0,05. Ďalej vieme, že náhodný výberový súbor má veľkosť n=20. Ostatné údaje pre výpočet intervalu spoľahlivosti musíme vypočítať.

Rozptyl výberového súboru s^2 vypočítame podľa vzťahu (3.20), do ktorého potrebujeme dosadiť priemernú hodnotu $\bar{x}$. Priemernú hodnotu celkového cholesterolu vypočítame podľa vzťahu (3.2):

$$\bar{x} = \frac{5 + 6,5 + 4,07 + \cdots + 3,94}{20} = \frac{83,41}{20} \doteq 4,17$$

potom rozptyl výberového súboru sa bude rovnať:

$$s^2 = \frac{(5 - 4,17)^2 + (6,5 - 4,17)^2 + \cdots + (3,94 - 4,17)^2}{19} = \frac{14,3499}{19} \doteq 0,755$$

Hodnoty kvantilov χ^2 rozdelenia nájdeme v štatistických tabuľkách, pričom platí, že $n-1$=19, $\alpha/2$=0,025 a $1-\alpha/2$=0,975, tak ako je to zobrazené na obrázku 7.6.

Potom dolná hranica intervalu spoľahlivosti pre rozptyl μ bude:

$$\theta_1 = \frac{(20-1)\,4,17^2}{32,852} \doteq 0,4368$$

a horná hranica obdobne:

$$\theta_2 = \frac{(20-1)\,4,17^2}{8,907} \doteq 1,6112$$

95% interval spoľahlivosti pre rozptyl celkového cholesterolu je (0,4368; 1,6112) $\mathrm{mmol}^2/\mathrm{l}^2$.

d.f.	$\chi^2_{.005}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.0000393	.000982	.00393	2.706	3.841	5.024	6.635	7.879
2	.0100	.0506	.103	4.605	5.991	7.378	9.210	10.597
3	.0717	.216	.352	6.251	7.815	9.348	11.345	12.838
⋮								
17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401
⋮								

Obr. 7.6: Výber tabuľkových hodnôt Chí kvadrát rozdelenia.

Druhou odmocninou hodnôt rozptylu získavame smerodajnú odchýlku:

$$\left(\sqrt{0,4368}; \sqrt{1,6112}\right) = (0,6609; 1,2693)$$

a teda 95% interval spoľahlivosti pre smerodajnú odchýlku celkového cholesterolu je (0,6609; 1,2693) mmol/l.

Funkcie v MS Excel

CHISQ.INV(pravdepodobnosť; st_voľnosti)
Funkcia CHISQ.INV vráti inverznú hodnotu ľavostrannej pravdepodobnosti Chí-kvadrát rozdelenia.

CHISQ.INV.RT(pravdepodobnosť; st_voľnosti)
Funkcia CHISQ.INV.RT vráti inverznú hodnotu pravostrannej pravdepodobnosti Chí-kvadrát rozdelenia.

Kapitola 8

Testovanie štatistických hypotéz

V predchádzajúcej kapitole sme sa venovali základným odhadom neznámych parametrov základného súboru z náhodných výberových súborov. Metódy odhadu nám teda zabezpečia, že s použitím údajov náhodného výberového súboru vieme získať dobré odhady neznámych parametrov reprezentujúcich zodpovedajúce hodnoty základného súboru. Po získaní odhadu býva ďalším krokom štatistického uvažovania potvrdenie toho, či získaný odhad adekvátne reprezentuje hodnotu základného súboru. Táto otázka je legitímna, keďže existuje veľa možností pri výbere náhodného súboru reprezentujúceho základný súbor, a teda aj veľa možných odhadov. Je preto náročné jednoznačne potvrdiť, že odhad z jedného výberového súboru možno považovať za hodnotu predstavujúcu parameter základného súboru. Týmto súvislostiam nám pomáha porozumieť druhá oblasť induktívnej štatistiky, ktorou je testovanie štatistických hypotéz. Testovanie hypotéz sa zaoberá zodpovedaním výskumných otázok o neznámych parametroch základného súboru, resp. potvrdením alebo vyvrátením niektorých dohadov alebo tvrdení o neznámych parametroch. Proces, ktorý zahŕňa rozhodovanie o parametri na základe náhodného výberového súboru označujeme ako test hypotézy.

8.1 Hypotézy

Vo výskumných úlohách a experimentoch sú riešené problémy transformované do hypotéz, ktorých platnosť je potrebné overiť a na základe výsledkov overovania zamietnuť alebo nezamietnuť. Vyslovujeme teda predpoklady o neznámych parametroch základného súboru a ich platnosť overujeme štatistickými postupmi.

Štatistická hypotéza predstavuje tvrdenie o jednom alebo o viacerých základných súboroch a týka sa rozdelenia pravdepodobnosti pozorovanej veličiny,

prípadne jej parametrov. Napríklad výskumníci môžu predpokladať, že celkový cholesterol je v priemere rovnaký u mužov aj u žien; vedenie nemocnice môže predpokladať, že priemerná dĺžka hospitalizácie pacientov je 6 dní; praktický lekár môže predpokladať, že dve rôzne liečebné metódy používané pri bolestiach hlavy sú rovnako účinné; zdravotná sestra môže predpokladať, že konkrétny vzdelávací program povedie k zlepšeniu komunikácie medzi sestrou a pacientom; lekár internista môže predpokladať, že určitý liek bude účinný v 90 percentách prípadov, na ktoré sa používa a pod. Pomocou testovania hypotéz rozhodneme, či takéto tvrdenia sú alebo nie sú v súlade s dostupnými údajmi.

V induktívnej štatistike formulujeme dva typy hypotéz o neznámom parametri, ktoré si navzájom odporujú (navzájom sa vylučujú), a ktoré by mali byť vždy explicitne uvedené. Sú to:

- **nulová hypotéza**, ktorú označujeme ako H_0 a
- **alternatívna hypotéza**, ktorú označujeme ako H_1.

Nulová hypotéza H_0 je hypotéza, ktorej platnosť overujeme (testujeme). Zvyčajne predstavuje tvrdenie, že parameter má hodnotu zodpovedajúcu, v zmysle riešeného problému, žiadnemu účinku, resp. žiadnej zmene. Nulovú hypotézu môžeme zamietnuť alebo nezamietnuť, a to na základe dôkazov z údajov výberového súboru, ako aj zo základného výberového rozdelenia štatistiky, použitého na odhad neznámeho parametra. Ak nie je nulová hypotéza zamietnutá, hovoríme, že údaje, na ktorých je test založený, neposkytujú dostatočné dôkazy na jej zamietnutie. Ak testovací postup vedie k zamietnutiu, hovoríme, že dostupné údaje nie sú v súlade s nulovou hypotézou, ale podporujú nejakú inú hypotézu. Zamietnutie nulovej hypotézy vedie k alternatívnej hypotéze.

Alternatívna hypotéza H_1 je negáciou nulovej hypotézy a prijímame ju, ak neplatí nulová hypotéza. Alternatívna hypotéza je teda hypotézou, ktorá je v rozpore s nulovou hypotézou a uvádza, že parameter spadá do nejakej alternatívnej oblasti, resp. množiny hodnôt, v porovnaní s tým čo špecifikuje nulová hypotéza. Vo všeobecnosti sa alternatívna hypotéza rovná výskumnej hypotéze, t. j. domnienke alebo predpokladu, ktorý motivuje samotnú realizáciu výskumu (napríklad, ak výskum tvrdí nejakú zmenu).

Zatiaľ čo prijatie nulovej hypotézy znamená, že nie je veľa dôkazov na zamietnutie tvrdenia v nulovej hypotéze, alternatívna hypotéza znamená, že ak nemôžeme prijať nulovú hypotézu alebo presnejšie, ak nemáme dostatok dôkazov v prospech nulovej hypotézy, alebo na podporu nulovej hypotézy, potom môžeme prijať alternatívne tvrdenie dané alternatívnou hypotézou.

8.2 Stanovenie štatistických hypotéz

Pri stanovovaní štatistických hypotéz používame znamienka rovnosti (ako sú znamienka $=$, $\leq$ alebo $\geq$) vždy v nulovej hypotéze H_0, t. j. nulová hypotéza musí tvrdiť, že parameter, ktorý je predmetom záujmu sa rovná špecifikovanej hodnote. Potom alternatívna hypotéza H_1 tvrdí, že daný parameter je menší, väčší alebo sa jednoducho nerovná špecifikovanej hodnote. Štatistický test podľa toho kategorizujeme ako ľavostranný, pravostranný alebo obojstranný, a to v závislosti od alternatívnej hypotézy. Štatistický test bude:

- **ľavostranný** – ak H_1 tvrdí, že parameter je menší ako požadovaná hodnota,
- **pravostranný** – ak H_1 tvrdí, že parameter je väčší ako požadovaná hodnota,
- **obojstranný** – ak H_1 tvrdí, že parameter sa líši od hodnoty uvedenej v H_0, resp. sa jej nerovná.

Predpokladajme, že máme základný súbor s neznámym parametrom θ a chceme otestovať (potvrdiť alebo vyvrátiť) nejaké hypotézy o θ. Napríklad, chceme odpovedať na otázku, či môžeme dospieť k záveru, že určitý parameter základného súboru θ nie je rovný konkrétnej hodnote θ_0? Potom nulová hypotéza je:

$$H_0 : \ \theta = \theta_0 \tag{8.1}$$

a vyjadruje predpoklad o zhode parametra θ základného súboru so známou konštantou θ_0. Alternatívna hypotéza potom hovorí, že:

$$H_1 : \ \theta \neq \theta_0 \tag{8.2}$$

a popiera platnosť H_0 bez špecifikovania hodnoty parametra θ.

V prípade pravostrannej alternatívy by sme mohli chcieť vedieť, či môžeme dospieť k záveru, že určitý parameter základného súboru θ je väčší ako konkrétna hodnota θ_0? Potom nulová a alternatívna hypotéza sú:

$$H_0 : \ \theta \leq \theta_0 \ \text{ a } \ H_1 : \ \theta > \theta_0 \tag{8.3}$$

Alebo, v prípade ľavostrannej alternatívnej hypotézy by sme mohli chcieť vedieť, či môžeme dospieť k záveru, že určitý parameter základného súboru θ je menší ako konkrétna hodnota θ_0? Potom nulová a alternatívna hypotéza sú:

$$H_0 : \ \theta \geq \theta_0 \ \text{ a } \ H_1 : \ \theta < \theta_0 \tag{8.4}$$

Predpokladajme tiež, že máme základný súbor s neznámym parametrom θ_1 a základný súbor s neznámym parametrom θ_2 a chceme odpovedať na otázku, či môžeme dospieť k záveru, že určitý parameter θ_1 prvého základného súboru nie je rovný určitému parametru θ_2 druhého základného súboru? Potom nulová hypotéza je:

$$H_0 : \ \theta_1 = \theta_2 \tag{8.5}$$

a vyjadruje predpoklad o zhode parametra θ_1 s parametrom θ_2. Alternatívna hypotéza potom hovorí, že:

$$H_1 : \ \theta_1 \neq \theta_2 \tag{8.6}$$

a popiera platnosť H_0 bez špecifikovania hodnôt parametrov θ_1 a θ_2.

V prípade pravostrannej alternatívy by sme mohli chcieť vedieť, či môžeme dospieť k záveru, že určitý parameter θ_1 prvého základného súboru je väčší ako určitý parameter θ_2 druhého základného súboru? Potom nulová a alternatívna hypotéza sú:

$$H_0 : \ \theta_1 \leq \theta_2 \ \text{ a } \ H_1 : \ \theta_1 > \theta_2 \tag{8.7}$$

Alebo, v prípade ľavostrannej alternatívnej hypotézy by sme mohli chcieť vedieť, či môžeme dospieť k záveru, že určitý parameter θ_1 prvého základného súboru je menší ako určitý parameter θ_2 druhého základného súboru? Potom nulová a alternatívna hypotéza sú:

$$H_0 : \ \theta_1 \geq \theta_2 \ \text{ a } \ H_1 : \ \theta_1 < \theta_2 \tag{8.8}$$

8.3 Chyby v štatistickom rozhodovaní

Pri testovaní štatistických hypotéz robíme závery na základe údajov z výberových súborov a môžu nastať situácie, kedy sa dopúšťame aj nesprávnych rozhodnutí, t. j. rozhodujeme na základe realizácie náhodného výberového súboru a preto nie je možné „zaručiť“ bezchybné rozhodnutie. Aby boli testy hypotéz vierohodné a použiteľné v analýzach údajov výskumných úloh, musia byť navrhnuté tak, aby minimalizovali možné chyby v rozhodovaní. Zvyčajne však nevieme, či došlo ku chybe, a preto môžeme hovoriť len o pravdepodobnosti chyby. Možnosti rozhodnutí pri testovaní hypotéz sú uvedené v tabuľke 8.1.

V rozhodovacom procese o nezamietnutí alebo zamietnutí nulovej hypotézy existujú dva druhy chýb. Ak zamietneme nulovú hypotézu, keď je pravdivá,

Tabuľka 8.1: Možnosti rozhodnutí pri testovaní hypotéz.

		Skutočnosť	
		H_0 platí	H_0 neplatí
Rozhodnutie	H_0 nezamietame	správne rozhodnutie $(1-\alpha)$ (hladina spoľahlivosti)	chyba 2. druhu β pravdepodobnosť chyby 2. druhu
	H_0 zamietame	chyba 1. druhu α pravdepodobnosť chyby 1. druhu (hladina významnosti)	správne rozhodnutie $(1-\beta)$ (sila testu)

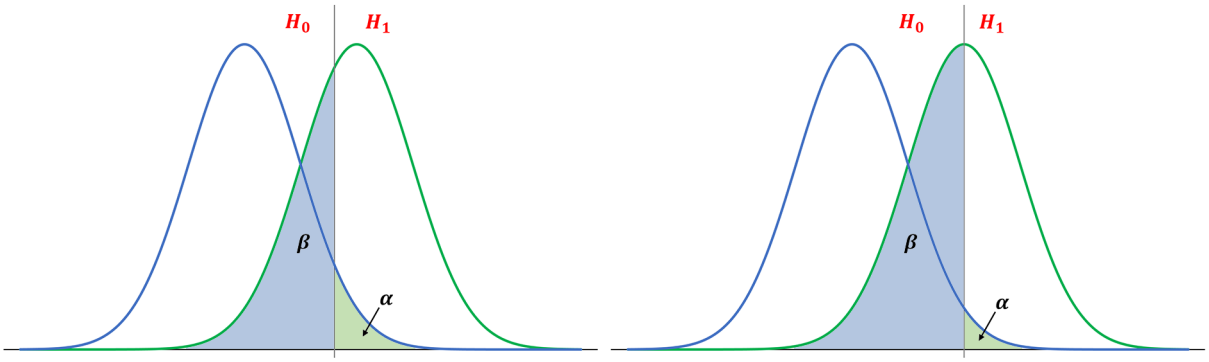
nazývame túto chybu chybou 1. druhu. Pravdepodobnosť chyby 1. druhu označujeme ako α a nazývame ju aj hladina významnosti. Druhý typ chyby súvisí s rozhodnutím, kedy nezamietneme nulovú hypotézu aj keď v skutočnosti neplatí. Pravdepodobnosť chyby 2. druhu označujeme ako β.

Správne rozhodnutia nastanú, ak je nulová hypotéza pravdivá a nezamietame ju a ak nulová hypotéza neplatí a zamietame ju. Pravdepodobnosť, že nezamietame nulovú hypotézu, ktorá v skutočnosti platí označujeme $1-\alpha$ a hovoríme o hladine spoľahlivosti. Pravdepodobnosť, že pri rozhodovaní nenastala chyba 2. druhu označujeme ako pravdepodobnosť $1-\beta$ a nazývame ju sila testu. Pravdepodobnosť $1-\beta$ teda vyjadruje s akou pravdepodobnosťou zamietneme nulovú hypotézu, ak v skutočnosti platí alternatívna hypotéza.

Tu je potrebné pripomenúť, že príslušná hodnota základného súboru nie je vo väčšine prípadov známa, ale môžeme uvažovať o hypotetickej hodnote, na základe ktorej sa test hypotézy vykonáva. Takže, ak tvrdíme, že nulová hypotéza je pravdivá alebo nepravdivá, je to iba hypotetické tvrdenie o hodnote základného súboru, keďže môžu existovať iné hypotetické hodnoty nulovej hypotézy, ktoré by bolo možné testovať s rovnakou dôležitosťou. Keďže sa test vykonáva v hypotetickej situácii o základnom súbore a v relatívnom zmysle, pravdepodobnostné výroky vedúce ku konkrétnym rozhodnutiam o hodnote základného súboru reflektujú iba jednu z mnohých hypotetických situácií.

V procese testovania štatistických hypotéz sa snažíme použiť test, pri ktorom sa maximalizuje pravdepodobnosť správnych rozhodnutí, a teda minimalizuje pravdepodobnosť chyby 1. druhu aj chyby 2. druhu. Avšak, ak budeme pre stanovenú veľkosť výberového súboru znižovať pravdepodobnosť chyby 1. druhu, neznamená to, že sa bude zmenšovať aj chyba 2. druhu. Naopak, tá sa môže

ešte zvyšovať. Príklad vplyvu zmeny hladiny významnosti α na β je znázornený na obrázku 8.1.



Obr. 8.1: Vzťah chyby 1. druhu a chyby 2. druhu.

Na obrázku 8.1 sú funkcie hustoty testového kritéria pri platnosti hypotézy H_0 a pri platnosti hypotézy H_1. Vľavo je situácia, kedy sa snažíme znížiť chybu 1. druhu α, no vidíme, že chyba 2. druhu β sa zvyšuje.

Pravdepodobnosť chýb 1. a 2. druhu však môžme znižovať zvyšovaním veľkosti výberového súboru. V praktickej realizácii hodnotu chyby 1. druhu volíme, zvyčajne ako α=0,05 (prípadne 0,01 alebo 0,1), a je teda pevnou hodnotou, ako sme to videli v predošlej kapitole pri určovaní intervalu spoľahlivosti.

8.4 Testovacia štatistika

Na testovanie hypotéz potrebujeme testovaciu štatistiku, t. j. štatistiku, ktorá poskytuje výberové rozdelenie vhodné pre definovanie pravdepodobnosti prijatia alebo zamietnutia nulovej hypotézy. Testovacia štatistika je štatistika definovaná ako funkcia premenných náhodného výberového súboru a základného súboru. Testovaciu štatistiku je možné tiež považovať za náhodnú veličinu, pričom jej výberové rozdelenie je nezávislé od parametrov, pre ktoré sa test hypotézy vykonáva.

Všeobecný tvar testovacej štatistiky možno vyjadriť vzťahom medzi pozorovanou hodnotou sledovanej veličiny a očakávanou hodnotou, ktorá platí pre nulovú hypotézu. Pre testovaciu štatistiku je teda možné zapísať:

$$\text{Testovacia štatistika} = \frac{\text{Pozorovaná hodnota} - \text{Hypotetická hodnota}}{\text{Stredná chyba pozorovanej hodnoty}} \tag{8.9}$$

Výsledok testovacej štatistiky (vypočítaná hodnota) sa následne porovná s rozdelením, ktoré očakávame, ak je nulová hypotéza pravdivá. Napríklad so

štandardným normálnym rozdelením, ktoré má priemernú hodnotu rovnú nule a rozptyl rovný jednej.

Testovaciu štatistiku je teda možné vypočítať z údajov výberového súboru. Je však potrebné si uvedomiť, že existuje veľa možných výsledkov, keďže existuje veľa možných výberových súborov zo základného súboru, t. j. konkrétna hodnota závisí od konkrétneho výberového súboru. Príkladom testovacej štatistiky je napríklad z štatistika:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{8.10}$$

kde μ_0 je hypotetickou hodnotou priemeru základného súboru.

V mnohých prípadoch je hypotetická hodnota rovná nule, takže testovacia štatistika sa stáva pomerom pozorovanej veličiny k jej strednej (štandardnej) chybe. Metóda vyhodnocovania veľkosti sledovanej veličiny, ako násobku jej strednej chyby je aplikovaná v mnohých metódach štatistickej analýzy. Avšak, využívané sú aj iné metódy a formy testovacích štatistík.

8.5 Oblasti rozhodovania

Hodnota testovacej štatistiky rozhoduje o zamietnutí alebo nezamietnutí nulovej hypotézy. Testovacia štatistika môže vo všeobecnosti nadobúdať hodnoty z reálnej množiny hodnôt, ktorú rozdelíme na dve podmnožiny, t. j. dve navzájom sa neprekrývajúce oblasti:

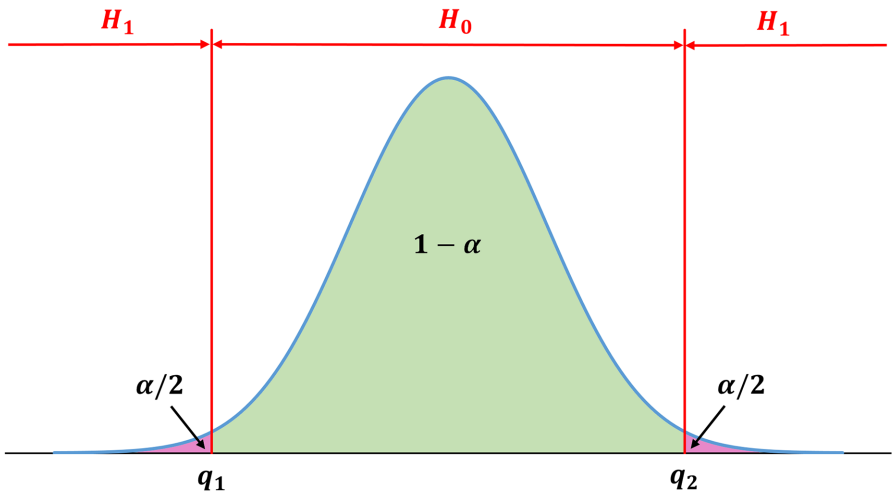
- **kritickú oblasť** W_α – oblasť zamietnutia nulovej hypotézy H_0 a
- **doplnkovú oblasť** W_0 – oblasť nezamietnutia nulovej hypotézy H_0

Hranice, ktoré oddeľujú tieto oblasti označujeme ako kritické hodnoty a určujeme ich ako kvantity rozdelenia použitej testovacej štatistiky, označme ju napríklad Q, pre zvolenú chybu 1. druhu, teda hladinu významnosti α. Kritická a doplnková oblasť sú definované na základe testovacej štatistiky, ako aj na základe nulovej a alternatívnej hypotézy, t. j. podľa toho či je alternatívna hypotéza obojstranná alebo jednostranná. O zamietnutí alebo nezamietnutí testovanej hypotézy potom rozhodujeme na základe nameraných hodnôt, z ktorých vypočítame hodnotu testovacej štatistiky Q. Ak jej hodnota padne do kritickej oblasti W_α, potom zamietame testovanú hypotézu H_0 a prijímame alternatívnu hypotézu H_1. Ak hodnota testovacej štatistiky Q padne do doplnkovej oblasti W_0, potom nulovú hypotézu H_0 nezamietame.

Ak je test hypotézy obojstranný, potom je kritická oblasť definovaná tak na ľavej, ako aj na pravej strane výberového rozdelenia testovacej štatistiky. Pre obojstranný test definujeme dve kritické hodnoty q_1 a q_2. Potom kritická oblasť a doplnková oblasť sú dané:

$$\begin{aligned} W_\alpha &: \ (-\infty; q_1\rangle \cup \langle q_2; \infty) \\ W_0 &: \ (q_1; q_2) \end{aligned} \tag{8.11}$$

Kritické hodnoty testovacej štatistiky Q (obrázok 8.2) určíme na ľavej strane jej rozdelenia pre oblasť pokrytia $\alpha/2$ a na pravej strane rozdelenia pre oblasť pokrytia $1 - \alpha/2$. Kritické hodnoty môžme nájsť v štatistických tabuľkách pre rozdelenie danej testovacej štatistiky.



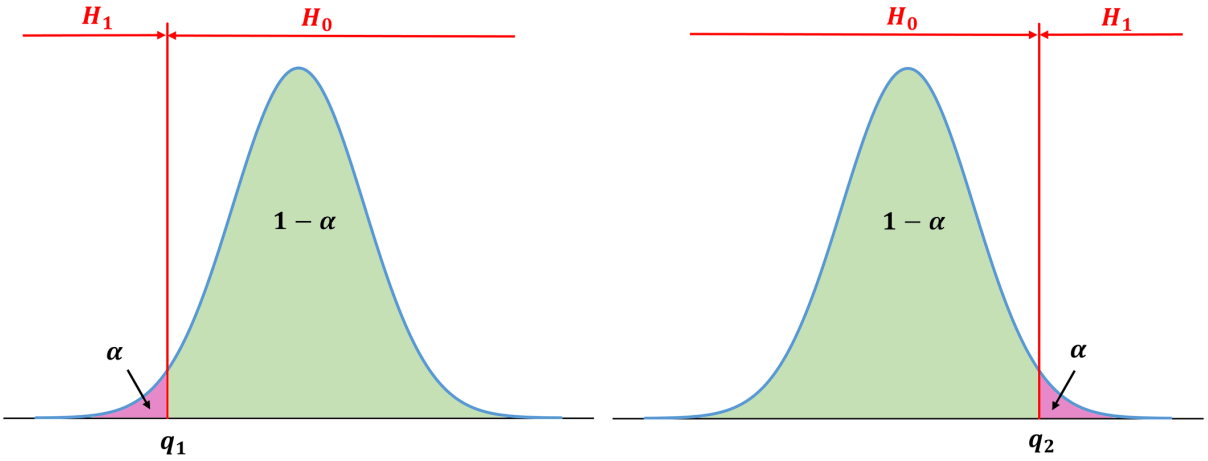
Obr. 8.2: Oblasti zamietnutia a nezamietnutia nulovej hypotézy pre obojstrannú alternatívnu hypotézu.

Ak je test hypotézy jednostranný, potom na základe alternatívnej hypotézy môžeme definovať kritickú oblasť buď na ľavej alebo na pravej strane výberového rozdelenia testovacej štatistiky Q (obrázok 8.3).

Pre jednostranný test hypotézy je teda definovaná jedna kritická hodnota, ktorá vymedzuje oblasť o veľkosti pravdepodobnosti α, a to na ľavej alebo na pravej strane rozdelenia testovacej štatistiky pre ľavostranný, resp. pravostranný test.

Pre ľavostrannú alternatívnu hypotézu sú kritická oblasť a doplnková oblasť dané nasledovnými dvoma intervalmi:

$$\begin{aligned} W_\alpha &: \ (-\infty; q_1\rangle \\ W_0 &: \ (q_1; \infty) \end{aligned} \tag{8.12}$$



Obr. 8.3: Oblasti zamietnutia a nezamietnutia nulovej hypotézy pre ľavostrannú (vľavo) a pravostrannú (vpravo) alternatívnu hypotézu.

Pre pravostrannú alternatívnu hypotézu sú kritická oblasť a doplnková oblasť dané nasledovnými dvoma intervalmi:

$$\begin{aligned} W_\alpha &: \ \langle q_2; \infty) \\ W_0 &: \ (-\infty; q_2) \end{aligned} \tag{8.13}$$

Hodnoty testovacej štatistiky tvoriace kritickú oblasť sú tie, pri ktorých je menej pravdepodobné, že sa vyskytnú, ak je nulová hypotéza pravdivá. Na druhej strane, hodnoty tvoriace doplnkovú oblasť sa vyskytnú s väčšou pravdepodobnosťou, ak je nulová hypotéza pravdivá. Takzvané rozhodovacie pravidlo teda hovorí, že máme zamietnuť nulovú hypotézu, ak hodnota testovacej štatistiky, ktorú vypočítame z výberového súboru, je jednou z hodnôt v kritickej oblasti a nezamietnuť nulovú hypotézu, ak je vypočítaná hodnota testovacej štatistiky jednou z hodnôt v doplnkovej oblasti.

Pripomíname, že je vhodnejšie tvrdiť, že nulová hypotéza nie je zamietnutá, ako tvrdiť, že nulová hypotéza je prijatá. Vyhneme sa tak prípadným pochybnostiam, keďže sme sa mohli dopustiť chyby 2. druhu, ktorej pravdepodobnosť môže byť často vysoká.

8.6 p hodnota

Hodnota p je číslo, ktoré nám hovorí, ako nezvyčajné sú výsledky konkrétneho výberového súboru, ak platí nulová hypotéza. Hodnota p označujúca, že výsledky výberového súboru pravdepodobne nenastali, nám poskytuje opodstatnenie pre pochybnosti o pravdivosti nulovej hypotézy.

Mohli by sme teda definovať, že hodnota p je pravdepodobnosť, že vypočítaná hodnota testovacej štatistiky je aspoň taká extrémna ako špecifikovaná hodnota testovacej štatistiky, keď je pravdivá nulová hypotéza. Hodnota p je teda najmenšia hodnota α, pre ktorú môžeme zamietnuť nulovú hypotézu.

Často sa význam hodnoty p a hladiny významnosti α nesprávne zamieňajú a interpretujú. Hovoríme, že výsledky sú štatisticky významné, ak je p hodnota nižšia ako hladina významnosti α, ktorá je pevne stanovená na začiatku analýzy, zvyčajne na hodnotu 0,05, resp. 5%. Na druhej strane, p hodnota je odvodená s ohľadom na test hypotézy, v ktorom sa testovacia štatistika vypočítava z výsledkov daného súboru údajov a za predpokladu, že je nulová hypotéza pravdivá. Hodnoty p môžu naznačovať, do akej miery sú údaje výberového súboru nekompatibilné so špecifikovaným štatistickým modelom, avšak neurčujú pravdepodobnosť, že študovaná hypotéza je pravdivá, či pravdepodobnosť, že výber údajov bol získaný len na základe náhody.

Predpokladajme, že máme testovaciu štatistiku Z výberového rozdelenia na testovanie nulovej hypotézy $H_0 : \mu = \mu_0$ oproti alternatívnej hypotéze $H_1 : \mu \neq \mu_0$ pre náhodný výberový súbor o veľkosti n, ktorý je výberom zo základného súboru s normálnym rozdelením $N\left(\mu, sigma^2\right)$. Hodnotu testovacej štatistiky z pre konkrétny výberový súbor a danú nulovú hypotézu je možné získať podľa vzťahu (8.10). Potom môžeme definovať p hodnotu ako pravdepodobnosť podľa špecifikovanej štatistickej hypotézy, kedy by testovacia štatistika bola rovná alebo extrémnejšia ako jej pozorovaná hodnota:

$$p = P\left(Z > |z|\right) + P\left(Z < -|z|\right) \tag{8.14}$$

pričom absolútnu hodnotu z používame na definovanie okrajových plôch pozorovanej hodnoty (ľavá a pravá strana rozdelenia), pre prípady negatívnej alebo pozitívnej hodnoty testovacej štatistiky, tak ako je to zobrazené na obrázku 8.4.

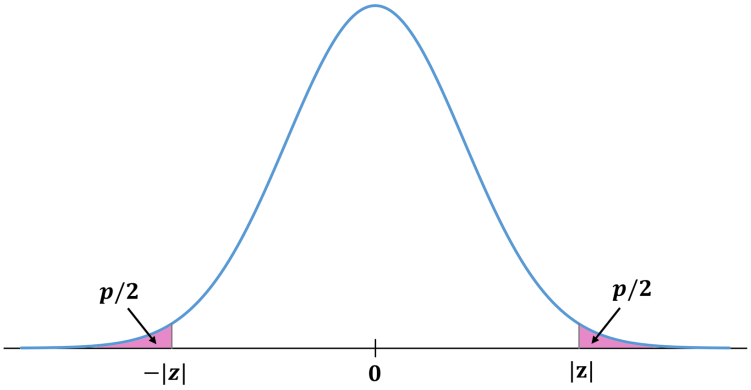
Pri obojstrannej alternatívnej hypotéze je možné extrémnejšie hodnoty testovacej štatistiky získať z ľavej alebo pravej strany výberového rozdelenia. Pre kladnú hodnotu z môžme vzťah (8.14) upraviť takto:

$$p = P\left(Z > z\right) + P\left(Z < -z\right) \tag{8.15}$$

pretože štandardné normálne rozdelenie je symetrické a krajné plochy rozdelenia sú väčšie ako z a menšie ako $-z$.

Podobne, vieme určiť p hodnotu pre jednostranné testy hypotézy. V prípade pravostrannej alternatívnej hypotézy bude platiť, že:

$$p = P\left(Z > z\right) = 1 - P\left(Z < z\right) = 1 - F\left(z\right) \tag{8.16}$$



Obr. 8.4: Funkcia hustoty testovacej štatistiky so štandardným normálnym rozdelením pre obojstranný test hypotézy.

Pre ľavostrannú alternatívnu hypotézu:

$$p = P\left(Z < z\right) = F\left(z\right) \tag{8.17}$$

kde $F\left(z\right)$ je hodnota distribučnej funkcie štandardného normálneho rozdelenia pre $Z = z$.

Predpokladajme, že hladina významnosti α bola špecifikovaná ako hodnota 0,05. Potom, na zistenie, či sú výsledky testov hypotéz štatisticky významné, možno použiť jednu z nasledujúcich metód.

- Vypočítame testovaciu štatistiku a porovnáme ju s kritickou hodnotou testovacieho kritéria na úrovni α=0,05. Ak testovacia štatistika padne do kritickej oblasti, potom H_0 zamietame a výsledok považujeme za štatisticky významný ($p \leq 0,05$). V opačnom prípade H_0 nezamietame a výsledok považujeme za štatisticky nevýznamný ($p > 0,05$). Tento prístup bol diskutovaný v predošlej kapitole 8.5, a niekedy ho označujeme aj ako metóda kritických hodnôt.
- Vypočítame presnú p hodnotu, a ak $p \leq 0,05$, potom H_0 zamietneme a výsledky považujeme za štatisticky významné. V opačnom prípade, ak $p > 0,05$, potom nezamietame H_0 a výsledky považujeme za štatisticky nevýznamné. Tento prístup označujeme aj ako metóda p hodnoty.

8.7 Všeobecný postup testovania hypotéz

Všeobecný postup testovania štatistických hypotéz je možné zhrnúť do nasledujúcich krokov:

1. Sformulujeme nulovú hypotézu H_0, t. j. spravidla tvrdenie, ktoré chceme vyvrátiť.
2. Sformulujeme alternatívnu hypotézu H_1, t. j. spravidla tvrdenie, ktoré chceme výskumom dokázať.
3. Zvolíme hladinu významnosti α, na ktorej budeme nulovú hypotézu testovať. Zvyčajne ju volíme na úrovni 0,05. Niekedy 0,01 alebo 0,1.
4. Z údajov výberového súboru vypočítame hodnotu testovacej štatistiky (testovacieho kritéria, podľa testovaného problému). Pri výbere vhodného typu testu zohľadňujeme napríklad rozdelenie pravdepodobnosti náhodnej veličiny, z ktorej je vyberaný výberový súbor, dostupnosť informácie o rozptyle základného súboru, či veľkosť výberového súboru.
5. Určíme kritickú oblasť W_α, t. j. určíme kritické hodnoty, ktoré vymedzujú oblasť zamietnutia, resp. nezamietnutia nulovej hypotézy (napríklad vyhľadaním v štatistických tabuľkách).
6. Rozhodneme o zamietnutí, resp. nezamietnutí nulovej hypotézy H_0 na zvolenej hladine významnosti α.

Nezamietnutie nulovej hypotézy neznamená, že hypotéza je správna. Preveruje sa či je β dostatočne malé číslo. Prakticky sa však vo väčšine prípadov postupuje tak, že v prípade nezamietnutia nulovej hypotézy, sa táto hypotéza považuje za správnu.

Je potrebné zdôrazniť, že v praxi sa stretávame tak so súbormi, ktoré majú normálne rozdelenie údajov, ako aj so súbormi, ktoré sú malé v porovnaní s veľkosťou základného súboru, alebo majú rozdelenie, ktoré je zošikmené, a u ktorých neplatí predpoklad normality. Preto je dôležité, aby sa pred analýzou údajov vykonali testy normality. Navyše, pri testovaní dvoch výberových súborov existuje predpoklad, že ich odchýlky sú rovnaké, ktorý je tiež potrebné otestovať. Nakoniec je potrebné poznamenať, že testy hypotéz, rovnako ako intervaly spoľahlivosti, sú relatívne citlivé na veľkosť testovaných výberových súborov a pri interpretácii výsledkov zahŕňajúcich veľmi malé výberové súbory je potrebné postupovať opatrne.

Vo všeobecnosti je možné metódy testovania rozdeliť na:

- **parametrické** – ich použitie je viazané na určité predpoklady o rozdelení, resp. o parametroch základných súborov (napríklad o normálnom rozdelení, či o rovnakej variabilite),
- **neparametrické** – používajú sa, ak nie sú splnené požadované predpoklady pre použitie parametrického testu. Nevyžadujú výpočet parametrov štatistického súboru ani znalosť funkcie rozdelenia základného súboru.

Kapitola 9

Overovanie hypotéz o normálnom rozdelení

Ako už bolo naznačené v predchádzajúcich kapitolách, jedným z predpokladov pre použitie parametrických metód testovania štatistických hypotéz je, aby boli sledované veličiny základného súboru s normálnym rozdelením. Teda predpokladáme, že aj výberový súbor je náhodným výberom zo základného súboru s normálnym rozdelením. Zanedbanie predpokladu o type rozdelenia môže v praxi viesť k zavádzajúcim, resp. nesprávne interpretovaným výsledkom.

Na overovanie normálneho rozdelenia údajov (alebo aspoň ich symetrickosti) je možné použiť viacero prístupov:

- vizuálne hodnotenie grafickej prezentácie rozdelenia údajov výberového súboru, napríklad histogramu,
- analýza charakteristík popisnej štatistiky, najmä koeficientov šikmosti a špicatosti,
- testy hypotéz o normálnom rozdelení, napríklad Chí-kvadrát test dobrej zhody, test podľa Shapira-Wilka, D'Agostinov test, Kolmogorov-Smironov test a pod.

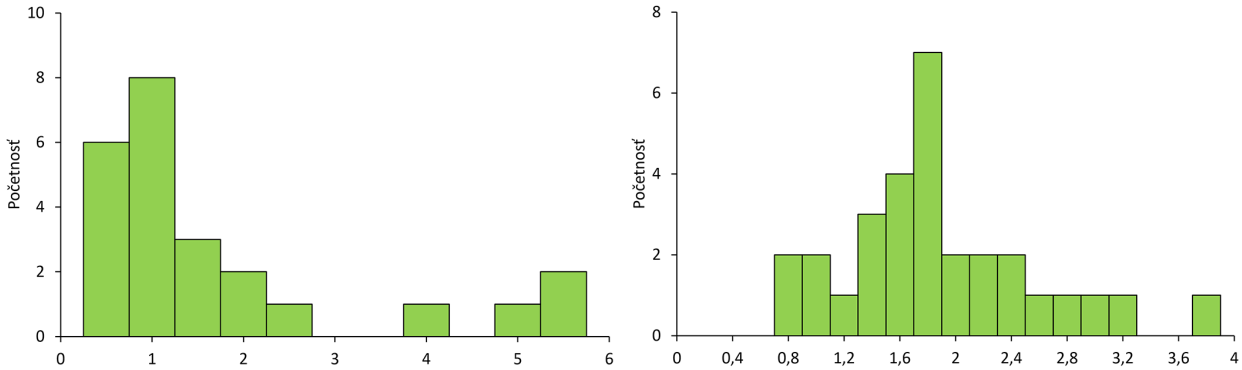
V niektorých prípadoch, kedy nie je rozdelenie údajov symetrické, je možné použiť vhodnú transformáciu údajov tak, aby boli transformované údaje symetrické. Následne je možné transformované údaje analyzovať štatistickými metódami určenými pre súbory s normálnym rozdelením.

9.1 Vizuálne hodnotenie

Prvým riešením, ako odhadnúť, či výberový súbor, ktorý sme získali v rámci riešenia nejakej výskumnej úlohy alebo experimentu, je náhodným výberom zo

základného súboru s normálnym rozdelením, je pozrieť si grafickú prezentáciu jeho hodnôt. Výraz „odhadnúť“ tu chápeme v zmysle subjektívneho hodnotenia, keďže sa opierame prevažne o skúsenosti pozorovateľa.

Jednoduchou a zároveň názornou reprezentatívnou metódou vizualizácie súboru údajov je histogram (viď kapitola 4.2.3), ktorý zobrazuje zastúpenie hodnôt v definovaných intervaloch formou početností, resp. relatívnych početností.

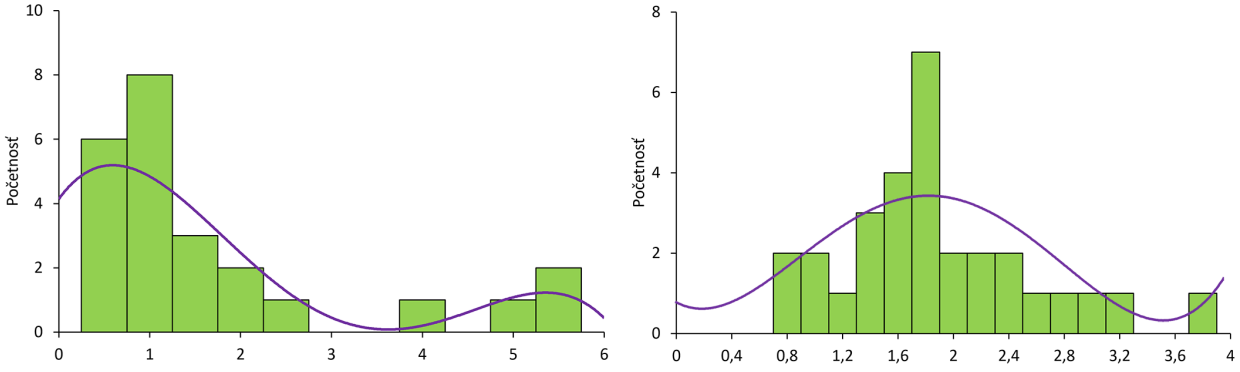


Obr. 9.1: Histogramy dvoch náhodných výberových súborov.

Na obrázku 9.1 sú znázornené príklady histogramov dvoch náhodných výberových súborov. Vizuálnym hodnotením sa snažíme zistiť napríklad to, či histogram má svoj stred, v ktorého okolí by malo byť najväčšie zastúpenie hodnôt a smerom k okrajom na oboch stranách by malo byť hodnôt menej. Tiež si všímame, či má histogram jeden, alebo viacero vrcholov. Prípadne to, či nie sú údaje približne rovnako zastúpené v rámci celého rozsahu hodnôt.

Na obrázku 9.1 vľavo, je najviac hodnôt na začiatku rozdelenia, kde je aj vrchol histogramu, a smerom k väčším hodnotám početnosti klesajú. V strede rozloženia dokonca nie sú žiadne hodnoty sledovanej premennej zastúpené vo výberovom súbore a potom ich je niekoľko na pravej strane histogramu. V tomto prípade, môžme priamo tvrdiť, že daný výberový súbor nepredstavuje súbor hodnôt s normálnym rozdelením.

Iné rozloženie reprezentuje histogram na obrázku 9.1 vpravo. Histogram má jeden vrchol umiestnený približne v strede rozdelenia výberového súboru. Smerom k okrajom histogramu početnosti hodnôt klesajú, nie však symetricky, keďže na ľavej strane klesajú rýchlejšie, kým na pravej strane klesajú voľnejšie na o niečo širšom intervale. To dáva predpoklad, že rozloženie bude mierne pravostranne zošikmené. V porovnaní s histogramom na obrázku 9.1 vľavo, však už nemôžeme vylúčiť, že (s istou pravdepodobnosťou) sa dané rozloženie blíži k normálnemu rozdeleniu. Pri vizualizácii si môžeme pomôcť krivkou, aproximujúcou priebeh rozloženia, napríklad tak ako je to na obrázku 9.2.



Obr. 9.2: Histogramy dvoch náhodných výberových súborov s krivkami aproximujúcimi tvar rozloženia.

Aproximáciou síce nezískavame potvrdenie, že údaje sú normálne rozdelené, ale vieme posúdiť, ktorý zo súborov podrobíme prípadnému testu normality v ďalšej analýze, a ktorý nie.

Histogram je teda vhodný spôsob, ako rýchlo vizualizovať rozdelenie jednej premennej. Okrem histogramu by sme na grafickú vizualizáciu mohli použiť napríklad polygón, krabicový graf, graf závislosti teoretických a empirických kvantilov a pod.

9.2 Porovnanie charakteristík popisnej štatistiky

Objektívnejšou metódou, ako je vizuálne pozorovanie grafickej prezentácie rozloženia údajov výberového súboru, je vypočítanie charakteristík popisnej štatistiky. Typickým príkladom sú charakteristiky tvaru, tak ako boli diskutované v kapitole 3.3.

Analýza koeficientov šikmosti a špicatosti môže byť využitá na prvotné posúdenie normality a symetrickosti rozdelenia skupiny údajov (výberového súboru, ktorý máme k dispozícii). Vychádzame z predpokladu, že u štandardného normálneho rozdelenia je šikmosť a špicatosť rovná nule (pozor, špicatosť nemusí byť znížená na nulu v niektorých výpočtových nástrojoch a potom bude rovná hodnote 3). Preto by aj hodnoty týchto charakteristík vypočítané z pozorovaného výberového súboru, mali byť blízke nule. Navyše, v prípade ak je absolútna hodnota šikmosti viac ako dvojnásobkom štandardnej chyby strednej hodnoty, potom to naznačuje, že údaje nie sú symetrické, a teda nemajú charakter normálneho rozdelenia.

Predpokladajme, že v rámci riešenia výskumnej úlohy bola sledovaná spojitá veličina Θ a zo základného súboru všetkých možných subjektov, boli náhodne a nezávisle na sebe vybrané dva výberové súbory, t. j. výberový súbor

A a výberový súbor B. Hodnoty sledovanej veličiny boli zaznamenané a popísané štatistickými charakteristikami. Vybrané charakteristiky popisujúce tieto výberové súbory sú uvedené v tabuľke 9.1.

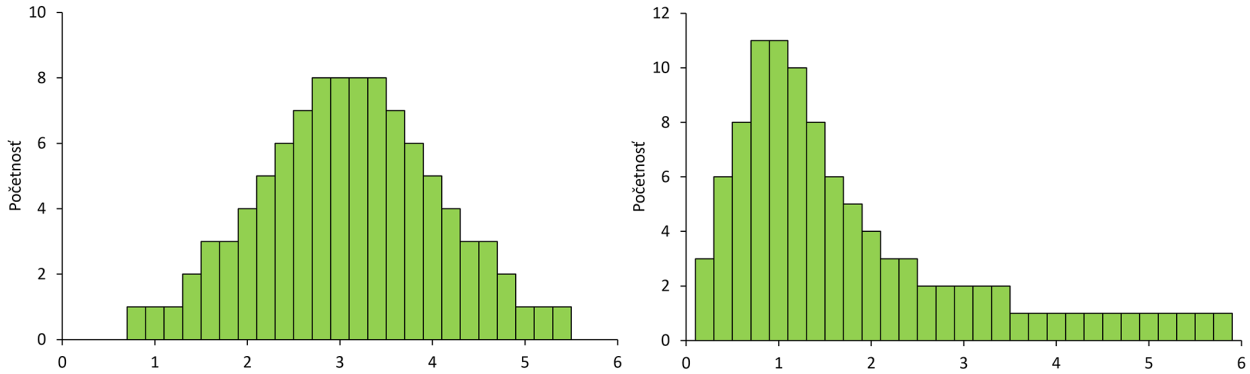
Tabuľka 9.1: Charakteristiky sledovanej veličiny Θ vypočítané z výberového súboru A a výberového súboru B.

Charakteristika	**Výberový súbor A**	**Výberový súbor B**
n	98	100
$\bar{x}$	3,1	1,776
$\tilde{x}$	3,1	1,4
s	0,9663	1,3528
SE	0,0976	0,1353
b_1	0,0000	1,3481
b_2	-0,2777	1,1622

Výberové súbory A a B sú stredne veľké, keďže ich veľkosť je 98, resp. 100 subjektov (hodnôt). Koeficient šikmosti b_1 výberového súboru A sa rovná 0, čo nám hovorí, že údaje sú symetricky rozdelené. Koeficient špicatosti b_2 je -0,2777 a informuje nás, že rozdelenie má v porovnaní s normálnym rozdelením nižší vrchol, keďže je hodnota záporná, a je len mierne roztiahnuté do strán, keďže hodnota je blízko nuly. Štandardná chyba strednej hodnoty SE má hodnotu 0,0976, ale keďže b_1 je nulové, jedná sa o identické rozloženie na pravej aj na ľavej strane voči strednej hodnote rozdelenia. Priemerná hodnota sledovanej veličiny sa v prípade výberového súboru A rovná 3,1, a je totožná s hodnotou mediánu, tak ako to býva u normálneho rozdelenia. Na základe týchto charakteristík je možné povedať, že rozdelenie hodnôt výberového súboru A vykazuje charakter normálneho rozdelenia.

V prípade výberového súboru B je koeficient šikmosti b_1 rovný 1,3481, čo nám hovorí, že rozdelenie je zošikmené na pravú stranu a údaje nie sú symetricky rozdelené. Koeficient špicatosti b_2 je 1,1622 a informuje nás, že rozdelenie má v porovnaní s normálnym rozdelením užší a vyšší vrchol, keďže hodnota b_2 je kladná. Štandardná chyba strednej hodnoty SE má hodnotu 0,1353, a keďže šikmosť b_1 je takmer desať krát väčšia, nemôžeme považovať rozdelenie výberového súboru B za rozdelenie s normálnym rozdelením hodnôt. Poukazuje na to čiastočne aj rozdiel priemernej hodnoty sledovanej veličiny, ktorý sa rovná hodnote 1,776 a je napravo od mediánu, ktorý má hodnotu 1,4. Na základe týchto charakteristík je možné povedať, že rozdelenie hodnôt výberového súboru B je zošikmené a nemá charakter normálneho rozdelenia.

Údaje oboch výberových súborov boli taktiež zotriedené a prezentované histogramom početností, tak ako je to znázornené na obrázku obrázku 9.3, pričom histogram výberového súboru A je zobrazený vľavo a histogram výberového súboru B je zobrazený vpravo.



Obr. 9.3: Histogramy náhodných výberových súborov A (vľavo) a B (vpravo).

Z histogramov rozdelenia početností je vidieť, že výberový súbor A je symetrický s identickou ľavou a pravou stranou, kým výberový súbor B je zošikmený napravo, tak ako na to poukazovali charakteristiky popisnej štatistiky.

9.3 Testy hypotéz o normálnom rozdelení

Predpoklad o type rozdelenia náhodnej veličiny možno otestovať aj objektívnymi metódami založenými na štatistickom testovaní. Medzi najpoužívanejšie testy hypotézy o normálnom rozdelení základného súboru patria:

- χ^2 **test dobrej zhody**, ktorý porovnáva teoretické a empirické početnosti, pričom súbory by mali byť stredné a veľké,
- **Shapiro-Wilkov test normality**, ktorý porovnáva teoretické a empirické kvantily, a je vhodný pre malé súbory,
- **D´Agostinov test**, ktorý porovnáva teoretické a empirické kvantily stredne veľkých súborov,
- **Kolmogorov-Smirnovov test**, ktorý využíva porovnanie teoretickej a empirickej distribučnej funkcie,
- ale aj **ďalšie testy normality** ako napríklad Anderson-Darlingov test, či Martinez-Iglewiczov test.

9.3.1 Chí kvadrát test dobrej zhody

χ^2 test dobrej zhody je vhodným testom, keď potrebujeme rozhodnúť, či pozorované rozloženie početností je alebo nie je kompatibilné s nejakým vopred vytvoreným alebo predpokladaným rozdelením. Môžeme teda určiť aj to, či hodnoty výberového súboru nejakej náhodnej premennej vyhovujú hypotéze, že boli získané zo základného súboru, ktorý má normálne rozdelenie.

Postup vedúci k dosiahnutiu rozhodnutia spočíva v zatriedení hodnôt výberového súboru do vzájomne sa vylučujúcich tried alebo triednych intervalov a určení ich početností. Potom tieto hodnoty porovnáme s očakávanými početnosťami, ktoré by malo mať normálne rozdelenie pre definované triedy alebo triedne intervaly. Ak sú rozdiely medzi početnosťami výberového súboru a očakávanými početnosťami malé, také, že by mohli vzniknúť náhodne, potom môžme dospieť k záveru, že výberový súbor môže byť výberom z normálneho rozdelenia.

Predpokladajme, že máme náhodnú veličinu X, ktorej hodnoty vo výberovom súbore sú $x_1, x_2, \ldots, x_n$. Sformulujeme nulovú a alternatívnu hypotézu tak, že predpokladáme:

H_0 : náhodná veličina má normálne rozdelenie
H_1 : náhodná veličina nemá normálne rozdelenie

Hodnoty náhodného výberového súboru s rozsahom n zatriedime do r tried (triednych intervalov) a ich početnosti porovnáme s teoretickými početnosťami, pričom testovacou štatistikou je náhodná premenná:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \tag{9.1}$$

kde n_i je empirická početnosť i-tej triedy, p_i je pravdepodobnosť, že X nadobudne hodnotu z i-tej triedy, teda ak platí nulová hypotéza a np_i je teoretická (očakávaná) početnosť i-tej triedy.

χ^2 má pre veľké súbory, teoreticky $n \to \infty$, rozdelenie s $(r - s - 1)$ stupňami voľnosti, kde s je počet parametrov, ktoré je potrebné odhadnúť pomocou výberových údajov. Pre normálne rozdelenie odhadujeme dva parametre, t. j. μ a σ^2, teda $s = 2$.

Pre kritickú oblasť, resp. oblasť zamietnutia H_0 platí:

$$\chi^2 > \chi^2_{1-\alpha}\,(r - s - 1) \tag{9.2}$$

Kritickú hodnotu pre príslušné α a počet stupňov voľnosti nájdeme v štatistických tabuľkách. Triedy, resp. triedne intervaly treba voliť tak, aby pre všetky $i = 1, 2, \ldots, r$ platilo $np_i \geq 5$. χ^2 test dobrej zhody teda nie je vhodný pre malé súbory.

Príklad 9.1. Náhodným výberom bolo vybraných 100 pacientov, u ktorých boli okrem iných údajov zaznamenané aj hodnoty bilirubínu. Je potrebné overiť, či je súbor hodnôt bilirubínu náhodným výberom z normálneho rozdelenia. Získané hodnoty sú: 17,1; 11,1; 15,9; 14,8; 12,2; 13,2; 15,5; 16,4; 19,2; 16,8; 13,2; 9,2; 16,8; 8,1; 20,6; 19,7; 13,7; 11,3; 17,6; 15,9; 13,3; 10,7; 16,3; 14,7; 18,5; 18,4; 17,7; 13,4; 15,0; 16,8; 17,2; 18,4; 19,6; 17,7; 14,4; 15,2; 13,8; 19,3; 14,6; 17,5; 13,3; 15,8; 12,3; 18,7; 14,5; 17,9; 16,5; 10,7; 14,5; 16,0; 15,1; 19,1; 14,1; 13,0; 13,6; 6,8; 12,2; 17,5; 16,7; 12,4; 11,6; 13,6; 9,2; 14,1; 16,4; 20,3; 7,6; 14,1; 18,5; 17,2; 14,4; 16,2; 17,0; 14,4; 15,6; 15,4; 15,2; 17,6; 20,7; 14,0; 18,7; 15,2; 14,2; 16,5; 13,2; 15,4; 17,6; 17,8; 15,9; 15,8; 4,7; 16,5; 12,5; 5,4; 15,8; 16,0; 14,3; 16,8; 15,6 a 16,4 $\mu mol/l$.

Našou nulovou hypotézou H_0 je, že výberový súbor hodnôt bilirubínu je výberom z normálneho rozdelenia a alternatívnou hypotézou H_1 je, že výberový súbor hodnôt bilirubínu nie je výberom z normálneho rozdelenia. Keďže hladina významnosti α nie je určená, použijeme hodnotu 0,05. Na rozhodnutie použijeme testovaciu štatistiku (9.1).

Postup výpočtu testovacej štatistiky χ^2 testu dobrej zhody budeme kvôli prehľadnosti zapisovať do tabuľky 9.2.

Tabuľka 9.2: Postup výpočtu χ^2 testovacej štatistiky.

Bilirubín	n_i	$F_0(x_i)$	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
≤ 10	7	0,0517	0,0517	5,18	0,6415
$(10; 12\rangle$	5	0,1615	0,1098	10,98	3,2547
$(12; 14\rangle$	17	0,3638	0,2022	20,22	0,5139
$(14; 16\rangle$	31	0,6146	0,2508	25,08	1,3951
$(16; 18\rangle$	26	0,8241	0,2095	20,95	1,2180
$(18; 20\rangle$	11	0,9419	0,1178	11,78	0,0514
> 20	3	0,9865	0,0446	4,46	0,4763

Početnosti získaných hodnôt bilirubínu určíme podľa intervalového triedenia diskutovaného v kapitole 4.2.2. Vzhľadom na veľkosť súboru a rozpätie hodnôt definujeme r=7 intervalov s otvorenou dolnou a hornou hranicou, tak ako je to

uvedené v prvom stĺpci tabuľky 9.2. Do ďalšieho stĺpca spočítame početnosti hodnôt bilirubínu n_i, spadajúce do príslušnej triedy.

Pre určenie pravdepodobností p_i teoretických početností normálneho rozdelenia, použijeme distribučnú funkciu F_0 štandardného normálneho rozdelenia $N(0,1)$, kde pravdepodobnosť, že hodnota bilirubínu bude rovná alebo menšia ako x_i je daná ako $P(F_0(x_i))$, a nájdeme ju v štatistických tabuľkách, tak ako to bolo vysvetlené v kapitole 6.2.1.

Pravdepodobnosti $P(F_0(x_i))$ hľadáme pre všetky horné hranice triednych intervalov bilirubínu tak, že $F_0(x_i) = (x_i - \mu)/\sigma$, kde μ je odhad priemernej hodnoty bilirubínu vypočítaný z výberového súboru a rovná sa 15,1 $\mu mol/l$ a podobne odhad smerodajnej odchýlky σ je 3,1262 $\mu mol/l$. Hodnoty distribučnej funkcie sme zapísali do tretieho stĺpca tabuľky.

Keďže nás zaujímajú pravdepodobnosti triednych intervalov a nie pravdepodobnosti všetkých predošlých hodnôt, potom do štvrtého stĺpca tabuľky zapíšeme pravdepodobnosti $p_i = F_0(x_i) - F_0(x_{i-1})$, ktoré reprezentujú pravdepodobnosti normálneho rozdelenia bilirubínu v jednotlivých triedach pri danej strednej hodnote a smerodajnej odchýlke.

Teoretické (očakávane) početnosti podľa normálneho rozdelenia vypočítame vynásobením počtu hodnôt, ktoré chceme zatriediť do jednotlivých triednych intervalov príslušnými pravdepodobnosťami danej triedy. Počet hodnôt n=100, keďže chceme porovnať rovnako veľké skupiny. Teoretické početnosti sú uvedené v stĺpci np_i.

Nakoniec vypočítame zložky sumy testovacieho kritéria χ^2, ktoré sú uvedené v poslednom stĺpci tabuľky 9.2. Ich spočítaním dostávame výsledok testovacej štatistiky χ^2=7,551.

Posledným krokom testovania hypotézy o normálnom rozdelení je rozhodnúť, či nulovú hypotézu zamietame alebo nezamietame, t. j. rozhodneme podľa toho, či je výsledná hodnota χ^2=7,551 v kritickej oblasti, alebo nie. Kritická hodnota $\chi^2_{1-\alpha}(4)$, keďže počet stupňov voľnosti je r-s-1=7-2-1=4, nájdená v štatistických tabuľkách je 9,488. Podľa definície kritickej oblasti (9.2) je vypočítaná testovacia štatistika menšia ako kritická hodnota a tak nulovú hypotézu nezamietame a tvrdíme, že hodnoty bilirubínu sú výberom z normálneho rozdelenia.

9.3.2 Shapiro-Wilkov test

Shapiro-Wilkov test je používaný väčšinou u malých výberových súborov, kde sa veľkosť n pohybuje často na úrovni medzi 3 až 50. Publikované sú však aj prístupy pre väčšie rozsahy súborov.

Nulovú a alternatívnu hypotézu formulujeme s predpokladom, že:

H_0 : náhodná veličina má normálne rozdelenie
H_1 : náhodná veličina nemá normálne rozdelenie

Realizácia Shapiro-Wilkovho testu spočíva v usporiadaní hodnôt $x_1, x_2, \ldots, x_n$ náhodnej premennej X podľa veľkosti tak, aby vznikla neklesajúca postupnosť $x_1^* \leq x_2^* \leq \ldots \leq x_n^*$.

Potom testovacou štatistikou je náhodná premenná:

$$W = \frac{\left(\sum_{i=1}^{m} a_i^{(n)} \left(x_{n-i+1}^* - x_i^*\right)\right)^2}{\sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2} \tag{9.3}$$

kde $m = n/2$ pre párnu hodnotu veľkosti výberového súboru n, $m = (n-1)/2$ pre nepárnu hodnotu veľkosti výberového súboru n a určuje počet možných rozdielov, ktoré vzniknú vypočítaním rozdielu medzi najväčšou hodnotou vo výberovom súbore a najmenšou hodnotou vo výberovom súbore, druhou najväčšou hodnotou a druhou najmenšou hodnotou, treťou najväčšou hodnotou a treťou najmenšou hodnotou atď., až kým sa nedosiahne stred usporiadaných hodnôt výberového súboru. $a_i^{(n)}$ sú koeficienty Shapiro-Wilkovho testu pre danú veľkosť výberového súboru n, a sú uvádzané v štatistických tabuľkách.

Pre oblasť zamietnutia (kritickú oblasť) H_0 platí, že:

$$W \leq W\left(n, \alpha\right) \tag{9.4}$$

kde $W\left(n, \alpha\right)$ sú kritické hodnoty Shapirovho-Wilkovho testu a taktiež sú uvádzané v štatistických tabuľkách.

Príklad 9.2. Náhodným výberom bola vybraná skupina desaťročných dievčat, u ktorých boli zaznamenané hodnoty výšky. Je potrebné overiť, či je výberový súbor hodnôt výšky desaťročných dievčat náhodným výberom z normálneho rozdelenia. Získané hodnoty výšky desaťročných dievčat sú: 130; 127; 136; 139; 147; 142; 138; 140; 139; 141; 139; 133; 151 a 136 cm.

Z dostupných informácií je zrejmé, že výberový súbor je malý, pričom n=14 a $\bar{x}$=138,43 cm. V tomto prípade nie je možné použiť χ^2 test dobrej zhody, a preto overovanie normality rozdelenia hodnôt výšky desaťročných dievčat vykonáme pomocou Shapiro-Wilkovho testu.

Nulovou hypotézou H_0 je, že súbor hodnôt výšok desaťročných dievčat je výberom z normálneho rozdelenia a alternatívnou hypotézou H_1 je, že súbor hodnôt výšok desaťročných dievčat nie je výberom z normálneho rozdelenia.

Keďže hladina významnosti α nie je určená, použijeme štandardne aplikovanú hodnotu α=0,05. Na rozhodnutie o zamietnutí alebo nezamietnutí nulovej hypotézy použijeme testovaciu štatistiku (9.3), pričom postup výpočtu budeme zapisovať kvôli názornosti aj do tabuľky 9.3.

Tabuľka 9.3: Postup výpočtu testovacej štatistiky Shapiro-Wilkovho testu.

x_i	x_i^*	$a_i^{(14)}$	$\left(x_{n-i+1}^* - x_i^*\right)$	$a_i^{(14)}\left(x_{n-i+1}^* - x_i^*\right)$	$\left(x_i^* - \bar{x}\right)^2$
130	127	0,5251	24	12,6024	130,6122
127	130	0,3318	17	5,6406	71,0408
136	133	0,2460	9	2,2140	29,4694
139	136	0,1802	5	0,9010	5,8980
147	136	0,1240	4	0,4960	5,8980
142	138	0,0727	1	0,0727	0,1837
138	139	0,0240	0	0	0,3265
140	139				0,3265
139	139				0,3265
141	140				2,4694
139	141				6,6122
133	142				12,7551
151	147				73,4694
136	151				158,0408
			Suma	21,9267	497,4286

Súbor všetkých zaznamenaných hodnôt výšky x_i usporiadame do neklesajúcej postupnosti x_i^*, t. j. od najmenšej po najväčšiu hodnotu výšky, tak ako je to uvedené v druhom stĺpci tabuľky 9.3.

Počet koeficientov $a_i^{(n)}$ Shapiro-Wilkovho testu m určíme ako počet všetkých rozdielov medzi krajnými hodnotami výberového súboru. Keďže n=14, tak m=14/2=7, t. j. budeme mať 7 rozdielov násobených siedmimi koeficientmi $a_i^{(14)}$, ktoré vyhľadáme v štatistických tabuľkách, tak ako je to zobrazené na obrázku 9.4. Hodnoty zapíšeme v tabuľke 9.3 do tretieho stĺpca.

Rozdiely medzi väčšími a menšími hodnotami výberového súboru vypočítame z usporiadanej postupnosti hodnôt, pričom postupne odpočítavame krajné menšie hodnoty od krajných väčších hodnôt smerom ku stredu usporiadaných hodnôt (k mediánu). Prvý rozdiel bude 151-127=24, druhý 147-130=17 atď. Vypočítané rozdiely sú zapísané v štvrtom stĺpci tabuľky 9.3.

i \ n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	-	-	0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8	-	-	-	-	0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9	-	-	-	-	-	-	0.0000	0.0163	0.0303	0.0422
10	-	-	-	-	-	-	-	-	0.0000	0.0140

Obr. 9.4: Výber tabuľkových hodnôt koeficientov Shapiro-Wilkovho testu.

Do piateho stĺpca tabuľky 9.3 boli zapísané všetky zložky sumy čitateľa testovacej štatistiky W, ktoré vznikli vynásobením koeficientov $a_i^{(14)}$ Shapiro-Wilkovho testu a rozdielov medzi krajnými hodnotami usporiadanej postupnosti hodnôt výberového súboru.

Jednotlivé zložky sumy menovateľa testovacej štatistiky W, získame druhou mocninou rozdielu medzi hodnotou výberového súboru a priemerom všetkých hodnôt. Prvá zložka tejto sumy bude vypočítaná ako $(127 - 138,43)^2 = 130,6122$, druhá $(130 - 138,43)^2 = 71,0408$ atď., tak ako je to zapísané v poslednom stĺpci tabuľky 9.3.

Testovacia štatistika Shapiro-Wilkovho testu W sa potom rovná podielu druhej mocniny sumy zložiek čitateľa a sumy zložiek menovateľa. Potom je výsledkom testovacej štatistiky $W = 21,92676^2/497,4286 \doteq 0,967$. Nulovú hypotézu zamietame, ak je testovacia štatistika menšia ako kritická hodnota $W(n, \alpha)$, ktorá je pre n=14 a α=0,05 rovná hodnote 0,874, tak ako to môžme nájsť v štatistických tabuľkách (obrázok 9.5).

n	$W_{0.01}$	$W_{0.02}$	$W_{0.05}$	$W_{0.10}$	$W_{0.50}$
3	0.753	0.756	0.767	0.789	0.959
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
⋮					
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
⋮					

Obr. 9.5: Výber tabuľkových hodnôt kvantilov Shapiro-Wilkovho testu.

Hodnota testovacej štatistiky nespadá do kritickej oblasti a preto nulovú hypotézu nezamietame. Záverom je, že výberový súbor hodnôt výšky desaťročných dievčat je výberom z normálneho rozdelenia.

9.3.3 D´Agostinov test

Ďalším testom na overenie normality údajov je D´Agostinov test, ktorý je v svojej základnej podstate integráciou testov šikmosti a špicatosti.

Test šikmosti vychádza z predpokladu, že normálne rozdelenie má šikmosť rovnú nule. Potom samotný test šikmosti zisťuje, či je šikmosť údajov štatisticky odlišná od nuly. Test je založený na skutočnosti, že ak sú údaje normálne rozdelené, potom testovacia štatistika, označme ju z_s, má štandardné normálne rozdelenie a platí:

$$z_s = \frac{b_1}{SE_s} \tag{9.5}$$

kde b_1 je šikmosť (koeficient šikmosti, pozri vzťah (3.25)) údajov výberového súboru a SE_s je štandardná chyba, ktorá je daná vzťahom:

$$SE_s = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{9.6}$$

kde n je veľkosť výberového súboru.

Šikmosť nie je štatisticky významne odlišná ako 0 (H_0), ak hodnota testovacej štatistiky spadá do doplnkovej oblasti (interval spoľahlivosti):

$$\left(b_1 - z_{1-\alpha/2}SE_s; b_1 + z_{1-\alpha/2}SE_s\right) \tag{9.7}$$

kde $z_{1-\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia.

Obdobne, test špicatosti vychádza z predpokladu, že normálne rozdelenie má šikmosť rovnú nule (po korekcii -3, ako to bolo naznačené v kapitole 3.3.2). Potom samotný test špicatosti zisťuje, či je špicatosť údajov štatisticky odlišná od nuly. Test je založený na skutočnosti, že ak sú údaje normálne rozdelené, potom testovacia štatistika, označme ju z_k, má štandardné normálne rozdelenie a platí:

$$z_k = \frac{b_2}{SE_k} \tag{9.8}$$

kde b_2 je špicatosť (koeficient špicatosti, pozri vzťah (3.27)) údajov výberového súboru a SE_k je štandardná chyba, ktorá je daná vzťahom:

$$SE_k = 2(n-1)\sqrt{\frac{6n}{(n-2)(n-3)(n+3)(n+5)}} \tag{9.9}$$

kde n je veľkosť výberového súboru.

Špicatosť nie je štatisticky významne odlišná ako 0 (H_0), ak hodnota testovacej štatistiky spadá do doplnkovej oblasti (interval spoľahlivosti):

$$\left(b_2 - z_{1-\alpha/2} SE_k; b_2 + z_{1-\alpha/2} SE_k\right) \tag{9.10}$$

kde $z_{1-\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia.

Samotný D'Agostinov test je potom založený na skutočnosti, že pri normálnom rozložení údajov má testovacia štatistika χ^2 rozdelenie s dvoma stupňami voľnosti:

$$\chi^2 = z_s^2 + z_k^2 \tag{9.11}$$

Test je vhodný pre stredne veľké súbory a nulovú hypotézu o normálnom rozdelení súboru hodnôt zamietame, ak:

$$\chi^2 > \chi^2_{1-\alpha}(2) \tag{9.12}$$

pričom, podobne ako v ostatných prípadoch teoretických rozdelení, kvantily χ^2 rozdelenia máme dostupné v štatistických tabuľkách.

9.3.4 Kolmogorov-Smirnovov test

Kolmogorov-Smirnovov test je ďalším vhodným testom, ak potrebujeme vedieť ako rozdelenie hodnôt vo výberovom súbore zodpovedá určitému teoretickému rozdeleniu. Kolmogorov-Smirnovov test dobrej zhody je vhodnou alternatívou χ^2 testu dobrej zhody (viď. kapitola 9.3.1).

Podstata Kolmogorov-Smirnovovho testu dobrej zhody spočíva v porovnaní distribučnej funkcie výberového súboru $F_v(x)$ s nejakou teoretickou distribučnou funkciou. Pri testovaní normality rozdelenia to bude distribučná funkcia normálneho rozdelenia $F(x)$.

V prípade, ak existuje úzka zhoda medzi distribučnými funkciami výberového súboru a normálneho rozdelenia, potom nulovú hypotézu, že výberový súbor je výberom zo základného súboru s distribučnou funkciou normálneho rozdelenia nezamietame. Naopak, ak existujú veľmi veľké rozdiely medzi distribučnými funkciami, nulovú hypotézu zamietneme.

Pripomíname, že distribučná funkcia je neklesajúcou funkciou definovanou na intervale od $-\infty$ do ∞ a hodnotami od 0 do 1, ktoré definujú pravdepodobnosť výskytu konkrétnej hodnoty a všetkých menších hodnôt ako je táto konkrétna hodnota (viď. napríklad obrázok 6.9).

Predpokladajme, že máme náhodný výberový súbor X s hodnotami $x_1, x_2, \ldots, x_n$. Ak predpokladáme porovnávanie distribučných funkcií, môžeme definovať hypotézy takto:

$$\begin{aligned} H_0: &\ F(x) = F_v(x) \quad \text{pre všetky hodnoty } x \\ H_1: &\ F(x) \neq F_v(x) \quad \text{pre aspoň jednu hodnotu } x \end{aligned}$$

Rozdiel medzi distribučnou funkciou normálneho rozdelenia $F(x)$ a distribučnou funkciou výberového súboru $F_v(x)$ kvantifikujeme pomocou testovacej štatistiky D, ktorá predstavuje najväčšiu vertikálnu vzdialenosť medzi $F(x)$ a $F_v(x)$. Testovaciu štatistiku zapíšeme:

$$D = \sup_x |F(x) - F_v(x)| \tag{9.13}$$

t. j. D sa rovná najväčšej hodnote zo všetkých absolútnych hodnôt rozdielov medzi $F(x)$ a $F_v(x)$.

Nulovú hypotézu zamietneme ak:

$$D > D_{n,\alpha} \tag{9.14}$$

kde $D_{n,\alpha}$ je kritická hodnota, ktorú nájdeme v štatistických tabuľkách pre veľkosť vzorky n a hladinu významnosti α.

Príklad 9.3. Náhodným výberom bola vybraná skupina pacientov, u ktorých bola sledovaná úroveň glukózy v krvi nalačno. Chceme overiť, či môžeme dospieť k záveru, že tieto údaje pochádzajú z normálneho rozdelenia základného súboru. Zaznamenané údaje pacientov sú: 75; 84; 80; 77; 68; 87; 92; 77; 92; 86; 78; 76; 80; 81; 72; 77; 92; 80; 80; 77; 77; 92; 68; 87; 84; 75; 78; 80; 80; 77; 72; 81; 76; 78; 81 a 86 mg/100ml.

Budeme testovať nulovú hypotézu, ktorá hovorí, že distribučná funkcia glukózy $F_g(x)$ sa rovná distribučnej funkcii normálneho rozdelenia $F(x)$ pre všetky x voči alternatívnej hypotéze, ktorá hovorí, že distribučná funkcia glukózy $F_g(x)$ sa nerovná distribučnej funkcii normálneho rozdelenia $F(x)$ aspoň pre jednu hodnotu x. Hladinu významnosti budeme uvažovať α=0,05.

Najprv vypočítame hodnoty distribučnej funkcie $F_g(x)$ pre hodnoty glukózy pacientov z výberového súboru, ktorého rozsah je n=36 (zistili sme ako počet hodnôt výberového súboru). Výpočet hodnôt distribučnej funkcie $F_g(x)$ spoločne s postupom výpočtu Kolmogorov-Smirnovovho testu pre názornosť zapíšeme do tabuľky 9.4.

Tabuľka 9.4: Výpočet hodnôt distribučnej funkcie glukózy pacientov z výberového súboru a distribučnej funkcie normálneho rozdelenia.

Glukóza ($mg/100ml$)	n_i	N_i	$F_g(x)$	z_i	$F_{(x)}$	D_i
68	2	2	0,0556	-1,9494	0,0256	0,0300
72	2	4	0,1111	-1,3041	0,0961	0,0150
75	2	6	0,1667	-0,8201	0,2061	0,0394
76	2	8	0,2222	-0,6588	0,2550	0,0328
77	6	14	0,3889	-0,4974	0,3094	0,0795
78	3	17	0,4722	-0,3361	0,3684	0,1038
80	6	23	0,6389	-0,0134	0,4946	0,1443
81	3	26	0,7222	0,1479	0,5588	**0,1634**
84	2	28	0,7778	0,6319	0,7363	0,0415
86	2	30	0,8333	0,9545	0,8301	0,0032
87	2	32	0,8889	1,1159	0,8678	0,0211
92	4	36	1,0000	1,9225	0,9727	0,0273

.

Prvú časť tabuľky 9.4 predstavuje tabuľka početností pre všetky jedinečné hodnoty glukózy v krvy pacientov výberového súboru (neboli vytvárané triedne intervaly), pričom absolútna početnosť n_i jednotlivých hodnôt bola zaznamenaná do druhého stĺpca tabuľky.

Distribučná funkcia je kumulatívnou funkciou, a preto sme do ďalšieho stĺpca tabuľky vypočítali kumulatívne absolútne početnosti N_i. Hodnoty distribučnej funkcie $F_g(x)$ pre hodnoty glukózy pacientov z výberového súboru boli získané vydelením hodnoty kumulatívnej absolútnej početnosti počtom všetkých hodnôt: $F_g(x_i) = N_i/n$. Napríklad, pre hodnotu glukózy 68 mg/100ml je $F_g(68)$=2/36=0,0556, pre $F_g(72)$=4/36=0,1111, pre $F_g(75)$=6/36=0,1667 atď. Výsledné hodnoty sú uvedené v štvrtom stĺpci tabuľky 9.4.

Na výpočet hodnôt distribučnej funkcie normálneho rozdelenia potrebujeme najprv transformovať získané hodnoty glukózy na hodnoty štandardného normálneho rozdelenia z, s ktorým náš výberový súbor porovnávame. Vypočítame ich podľa vzťahu:

$$z = \frac{x - \mu}{\sigma}$$

Na odhad μ použijeme priemernú hodnotu výberového súboru, ktorou je $\bar{x}$=80,08 mg/100ml. Podobne vypočítame smerodajnú odchýlku výberového súboru, ktorú použijeme ako odhad σ. Z hodnôt glukózy pacientov vo výberovom súbore dostávame smerodajnú odchýlku rovnú 6,1985 mg/100ml.

Potom z pre hodnotu glukózy v krvi rovnú 68 mg/100ml bude $(68 - 80,08)/6,1985 = -1,9494$, pre hodnotu glukózy v krvi rovnú 72 mg/100ml bude $(72 - 80,08)/6,1985 = -1,3041$ atď. Hodnoty štandardného normálneho rozdelenia pre glukózu so strednou hodnotou 80,08 mg/100ml a smerodajnú odchýlku 6,1985 mg/100ml sú uvedené v piatom stĺpci tabuľky 9.4.

Ďalej v štatistických tabuľkách pre štandardné normálne rozdelenie nájdeme hodnotu pravdepodobnosti $P(-\infty \leq z)$, ktorá pokrýva hodnotu z a všetky menšie, t. j. hodnotu distribučnej funkcie normálneho rozdelenia $F(x)$. V tabuľkách sú hodnoty z uvádzané s dvoma desatinnými miestami a pre príslušné α. Ako hodnotu distribučnej funkcie $F(x)$ vyberieme hodnotu, ktorá zodpovedá z najbližšiemu k hľadanej hodnote z_i z tabuľky 9.4. Prípadne je možné hodnotu pravdepodobnosti prepočítať podľa dvoch hodnôt, ktoré vymedzujú oblasť pre ďalšie desatinné miesta hodnoty z_i.

Obrázok 9.6 zobrazuje výber tabuľky hodnôt distribučnej funkcie normálneho rozdelenia a spôsob určenia prvých štyroch hodnôt hľadanej distribučnej funkcie $F(x)$.

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
⋮											
−2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	−2.00
−1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	−1.90
−1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	−1.80
⋮											
−1.40	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808	−1.40
−1.30	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968	−1.30
−1.20	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151	−1.20
⋮											
−0.80	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119	−0.80
−0.70	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420	−0.70
−0.60	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743	−0.60
−0.50	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085	−0.50
⋮											

Obr. 9.6: Výber tabuľkových hodnôt štandardného normálneho rozdelenia.

Teraz je možné vypočítať testovaciu štatistiku Kolmogorov-Smirnovovho testu, ktorá je maximom z rozdielov medzi distribučnými funkciami výberového súboru a normálneho rozdelenia $D_i = \left|F_g(x) - F_(x)\right|$. Absolútne hodnoty týchto rozdielov sú uvedené v poslednom stĺpci tabuľky 9.4.

Na základe vypočítaných rozdielov medzi distribučnými funkciami pre hodnoty glukózy je testovacia štatistika Kolmogorov-Smirnovovho testu rovná hodnote 0,1634 (v tabuľke 9.4 je zvýraznená tučným písmom).

Kritickú hodnotu Kolmogorov-Smirnovovho testu nájdeme v štatistických tabuľkách pre príslušné n a hladinu významnosti α, napríklad tak ako je to zobrazené na je obrázku 9.7.

One-Sided Test $p = .90$	**.95**	**.975**	**.99**	**.995**
Two-Sided Test $p = .80$	**.90**	**.95**	**.98**	**.99**
$n = 1$	.900	.950	.975	.990	.995
2	.684	.776	.842	.900	.929
3	.565	.636	.708	.785	.829
⋮					
35	.177	.202	.224	.251	.269
36	.174	.199	.221	.247	.265
37	.172	.196	.218	.244	.262
38	.170	.194	.215	.241	.258
39	.168	.191	.213	.238	.255
40	.165	.189	.210	.235	.252
Approximation for $n > 40$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Obr. 9.7: Výber tabuľkových hodnôt kvantilov Kolmogorov-Smirnovovej štatistiky.

Hodnota testovacej štatistiky Kolmogorov-Smirnovovho testu je menšia ako kritická hodnota $D_{36,0,05}$, keďže $0{,}1634 < 0{,}221$. Nulovú hypotézu teda nezamietame a môžeme povedať, že hodnoty glukózy v krvy pacientov výberového súboru sú výberom z normálneho rozdelenia.

Kapitola 10

Parametrické testy

V tejto kapitole si uvedieme niektoré najčastejšie používané prístupy testovania štatistických hypotéz, ak sú splnené predpoklady o výbere z normálneho rozdelenia. Hoci teória podmienok závisí od základného súboru s normálnym rozdelením, bežne sa tieto prístupy používajú aj v prípadoch, kedy sú relevantné základné súbory len približne normálne rozdelené. Toto je však akceptovateľné, len ak nie je odchýlka od normálneho rozdelenia významná.

Vo všeobecnosti je možné najčastejšie používané testy hypotéz rozdeliť do nasledovných skupín:

- **testy hypotéz o strednej hodnote**
  - testy zhody strednej hodnoty so známou konštantou
  - testy hypotéz o zhode dvoch stredných hodnôt nezávislých súborov
  - testy hypotéz o zhode dvoch stredných hodnôt závislých súborov (údaje tvoria párové pozorovania, dvojice)
- **testy hypotéz o rozptyloch**
  - testy zhody rozptylu so známou konštantou
  - testy zhody dvoch rozptylov

10.1 Testy hypotéz o strednej hodnote

10.1.1 Testy zhody strednej hodnoty základného súboru

Predpokladajme, že máme náhodný výberový súbor X, ktorého hodnoty sú $x_1, x_2, \ldots, x_n$, a sú výberom zo základného súboru s normálnym rozdelením $N\left(\mu, \sigma^2\right)$, pričom strednú hodnotu (priemer) základného súboru μ nepoznáme.

Potom môžu nastať prípady, kedy rozptyl základného súboru poznáme, napríklad z predošlých rozsiahlych experimentov či štúdií, alebo, čo je častejším prípadom, rozptyl základného súboru nepoznáme. Taktiež rozlišujeme situácie, kedy je náhodný výberový súbor malý, t. j. $n \leq 30$, alebo je dostatočne veľký, $n > 30$.

10.1.1.1 Rozptyl základného súboru je známy

Predpokladajme, že máme náhodnú veličinu X, ktorej rozptyl základného súboru σ^2 je známy. Na základe náhodného výberového súboru chceme overiť, či sa neznáma stredná hodnota základného súboru μ rovná alebo nerovná známej hodnote μ_0 (konštantnej hodnote). Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ sa rovná tejto konštante μ_0. Alternatívou bude, že sa tejto konštante μ_0 nerovná. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu = \mu_0 \\ H_1 &: \ \mu \neq \mu_0 \end{aligned}$$

Testovacou štatistikou je Z štatistika štandardného normálneho rozdelenia:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{10.1}$$

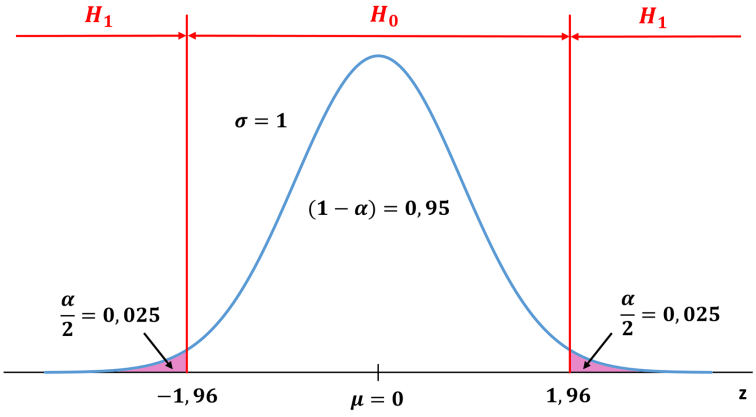
so štandardným normálnym rozdelením pravdepodobnosti $N(0,1)$.

Kritickú oblasť W_α, do ktorej, ak spadne hodnota testovacej štatistiky Z, tak zamietame nulovú hypotézu H_0, a doplnkovú oblasť W_0, do ktorej, ak spadne hodnota testovacej štatistiky Z, tak nezamietame nulovú hypotézu H_0, určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ \left(-\infty; z_{\alpha/2}\right\rangle \cup \left\langle z_{1-\alpha/2}; \infty\right) \\ W_0 &: \ \left(z_{\alpha/2}; z_{1-\alpha/2}\right) \end{aligned} \tag{10.2}$$

kde $z_{\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia a $z_{1-\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom (zaokrúhlené na dve desatinné miesta) je kritická hodnota $z_{0,025}$=-1,96 a kritická hodnota $z_{0,975}$=1,96 (pozri štatistické tabuľky štandardného normálneho rozdelenia). Potom oblasti pre rozhodnutie o prijatí alebo



Obr. 10.1: Oblasti prijatia alebo neprijatia nulovej hypotézy obojstranného Z testu na hladine významnosti 0,05.

neprijatí nulovej hypotézy H_0 môžeme znázorniť graficky, napríklad tak ako je to na obrázku 10.1.

Predpokladajme, že máme inú situáciu, kedy na základe náhodného výberového súboru chceme overiť, či je neznáma stredná hodnota základného súboru μ menšia ako známa hodnota μ_0. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ je rovná alebo väčšia ako μ_0 a alternatívna hypotéza hovorí, že je menšia ako μ_0. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu \geq \mu_0 \\ H_1 &: \ \mu < \mu_0 \end{aligned}$$

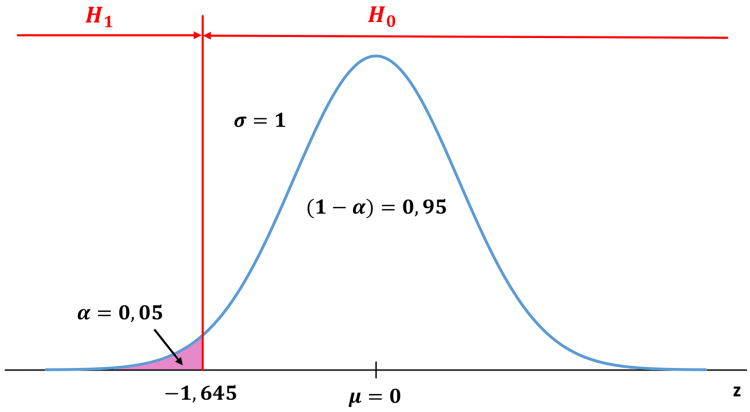
Testovacou štatistikou je opäť Z štatistika daná vzťahom (10.1). Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ (-\infty; z_\alpha\rangle \\ W_0 &: \ (z_\alpha; \infty) \end{aligned} \tag{10.3}$$

kde z_α je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,05}$=-1,645. Oblasti pre rozhodnutie o prijatí alebo neprijatí nulovej hypotézy H_0 môžeme taktiež znázorniť graficky, napríklad tak ako je to na obrázku 10.2.

Obdobne predpokladajme, že máme situáciu, kedy na základe náhodného výberového súboru chceme overiť, či je neznáma stredná hodnota základného súboru μ väčšia ako známa hodnota μ_0. Potom testujeme nulovú hypotézu,

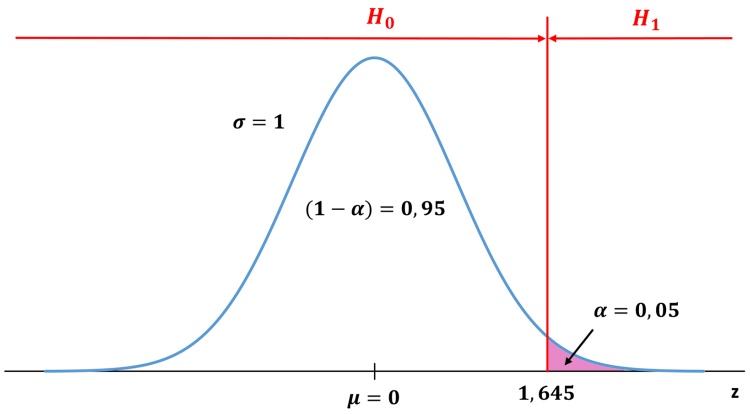


Obr. 10.2: Oblasti prijatia alebo neprijatia nulovej hypotézy ľavostranného Z testu na hladine významnosti 0,05.

ktorá hovorí, že stredná hodnota základného súboru μ je rovná alebo menšia ako μ_0 a alternatívna hypotéza hovorí, že je väčšia ako μ_0. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$H_0: \ \mu \leq \mu_0$$
$$H_1: \ \mu > \mu_0$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť Z štatistiku danú vzťahom (10.1). Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.



Obr. 10.3: Oblasti prijatia alebo neprijatia nulovej hypotézy pravostranného Z testu na hladine významnosti 0,05.

$$W_\alpha: \ \langle z_{1-\alpha}; \infty)$$
$$W_0: \ (-\infty; z_{1-\alpha}) \tag{10.4}$$

kde $z_{1-\alpha}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,95}$=1,645. Oblasti pre rozhodnutie o prijatí alebo neprijatí nulovej hypotézy H_0 môžeme taktiež znázorniť graficky, napríklad tak ako je to na obrázku 10.3.

10.1.1.2 Rozptyl základného súboru nie je známy a výberový súbor je veľký

Predpokladajme, že máme náhodnú veličinu X, ktorej rozptyl základného súboru σ^2 nie je známy. Na základe náhodného výberového súboru chceme overiť, či sa neznáma stredná hodnota základného súboru μ rovná alebo nerovná známej hodnote μ_0. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ sa rovná tejto konštante μ_0. Alternatívou bude, že sa tejto konštante μ_0 nerovná. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu = \mu_0 \\ H_1 &: \ \mu \neq \mu_0 \end{aligned}$$

Testovacou štatistikou je Z štatistika štandardného normálneho rozdelenia:

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{10.5}$$

so štandardným normálnym rozdelením pravdepodobnosti $N(0,1)$, kde s je smerodajná odchýlka vypočítaná z hodnôt výberového súboru.

Kritickú oblasť W_α určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ \left(-\infty; z_{\alpha/2}\right\rangle \cup \left\langle z_{1-\alpha/2}; \infty\right) \\ W_0 &: \ \left(z_{\alpha/2}; z_{1-\alpha/2}\right) \end{aligned} \tag{10.6}$$

kde $z_{\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia a $z_{1-\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,025}$=-1,96 a kritická hodnota $z_{0,975}$=1,96, a ak je α=0,01, potom je kritická hodnota $z_{0,005}$=-2,576 a kritická hodnota $z_{0,995}$=2,576.

V prípade, že na základe náhodného výberového súboru chceme overiť, či je neznáma stredná hodnota základného súboru μ menšia ako známa hodnota μ_0,

potom sú hypotézy identické, ako v prípade testu hypotézy strednej hodnoty základného súboru so známym rozptylom:

$$\begin{aligned} H_0 &: \ \mu \geq \mu_0 \\ H_1 &: \ \mu < \mu_0 \end{aligned}$$

Testovacou štatistikou je Z štatistika daná vzťahom (10.5), a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ (-\infty; z_\alpha\rangle \\ W_0 &: \ (z_\alpha; \infty) \end{aligned} \tag{10.7}$$

kde z_α je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,05}$=-1,645, a ak je α=0,01, potom je kritická hodnota $z_{0,01}$=-2,326.

Analogicky, ak chceme overiť, či je neznáma stredná hodnota základného súboru μ väčšia ako známa hodnota μ_0, potom sú hypotézy:

$$\begin{aligned} H_0 &: \ \mu \leq \mu_0 \\ H_1 &: \ \mu > \mu_0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť Z štatistiku danú vzťahom (10.5) a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \langle z_{1-\alpha}; \infty) \\ W_0 &: \ (-\infty; z_{1-\alpha}) \end{aligned} \tag{10.8}$$

kde $z_{1-\alpha}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,95}$=1,645, a ak je α=0,01, potom je kritická hodnota $z_{0,99}$=2,326.

10.1.1.3 Rozptyl základného súboru nie je známy a výberový súbor je malý

Predpokladajme, že máme náhodnú veličinu X, ktorej rozptyl základného súboru σ^2 nie je známy a výberový súbor je malý, t. j. $n \leq 30$. Na základe náhodného výberového súboru chceme overiť, či sa neznáma stredná hodnota základného súboru μ rovná alebo nerovná známej hodnote μ_0. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ

sa rovná tejto konštante μ_0. Alternatívou bude, že sa tejto konštante μ_0 nerovná. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu = \mu_0 \\ H_1 &: \ \mu \neq \mu_0 \end{aligned}$$

Testovacou štatistikou je T štatistika so Studentovým t rozdelením a $n-1$ stupňami voľnosti, ktoré je vhodné pre malé súbory (viď kapitola 6.2.2):

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{10.9}$$

kde s je smerodajná odchýlka vypočítaná z hodnôt výberového súboru.

Kritickú oblasť W_α určíme pomocou kritických hodnôt Studentovho t rozdelenia s $n-1$ stupňami voľnosti na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha/2,n-1}\right\rangle \cup \left\langle t_{1-\alpha/2,n-1}; \infty\right) \\ W_0 &: \ \left(t_{\alpha/2,n-1}; t_{1-\alpha/2,n-1}\right) \end{aligned} \tag{10.10}$$

kde $t_{\alpha/2,n-1}$ je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia a $t_{1-\alpha/2,n-1}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia pre $n-1$ stupňov voľnosti.

Ak je α=0,05 a n napríklad 5, potom je kritická hodnota $t_{0,025,4}$=-2,776 a kritická hodnota $t_{0,975,4}$=2,776, a ak je α=0,01, potom je kritická hodnota $t_{0,005,4}$=-4,604 a kritická hodnota $t_{0,995,4}$=4,604.

V prípade, že na základe náhodného výberového súboru chceme overiť, či je neznáma stredná hodnota základného súboru μ menšia ako známa hodnota μ_0, potom sú hypotézy dané ako:

$$\begin{aligned} H_0 &: \ \mu \geq \mu_0 \\ H_1 &: \ \mu < \mu_0 \end{aligned}$$

Testovacou štatistikou je T štatistika daná vzťahom (10.9), a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha,n-1}\right\rangle \\ W_0 &: \ \left(t_{\alpha,n-1}; \infty\right) \end{aligned} \tag{10.11}$$

kde $t_{\alpha,n-1}$ je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia pre $n-1$ stupňov voľnosti.

Ak je α=0,05 a n napríklad 5, potom je kritická hodnota $t_{0,05,4}$=-2,132, a ak je α=0,01, potom je kritická hodnota $t_{0,01,4}$=-3,747.

Analogicky, ak chceme overiť, či je neznáma stredná hodnota základného súboru μ väčšia ako známa hodnota μ_0, potom sú hypotézy:

$$\begin{aligned} H_0 &: \ \mu \leq \mu_0 \\ H_1 &: \ \mu > \mu_0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť T štatistiku danú vzťahom (10.9) a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \langle t_{1-\alpha,n-1}; \infty) \\ W_0 &: \ (-\infty; t_{1-\alpha,n-1}) \end{aligned} \tag{10.12}$$

kde $t_{1-\alpha,n-1}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia pre $n-1$ stupňov voľnosti.

Ak je α=0,05 a n napríklad 5, potom je kritická hodnota $t_{0,95,4}$=2,132, a ak je α=0,01, potom je kritická hodnota $t_{0,99,4}$=3,747.

10.1.2 Testy hypotéz o zhode dvoch stredných hodnôt nezávislých súborov

Testovanie hypotéz, ktoré sa zaoberá rozdielom medzi strednými hodnotami dvoch základných súborov je najčastejšie používané na určenie toho, či je alebo nie je možné dospieť k záveru, že sa tieto stredné hodnoty nerovnajú. Predpokladajme, že máme náhodný výberový súbor X_1, ktorého hodnoty sú $x_{11}, x_{12}, \ldots, x_{1n_1}$ a náhodný výberový súbor X_2, ktorého hodnoty sú $x_{21}, x_{22}, \ldots, x_{2n_2}$, ktoré sú výberom zo základných súborov s normálnym rozdelením $N\left(\mu, \sigma^2\right)$, pričom strednú hodnotu μ_1 základného súboru X_1 ani strednú hodnotu μ_2 základného súboru X_2 nepoznáme. Potom môžu nastať prípady, kedy rozptyly základných súborov poznáme alebo nepoznáme, a taktiež kedy sú náhodné výberové súbory malé alebo veľké.

10.1.2.1 Rozptyly základných súborov sú známe

Predpokladajme, že máme náhodné veličiny X_1 a X_2, ktorých rozptyly základného súboru σ_1^2 a σ_2^2 sú známe. Na základe dvoch náhodných výberových súborov chceme overiť, či sa neznáma stredná hodnota základného súboru μ_1 rovná alebo nerovná neznámej strednej hodnote základného súboru μ_2. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ_1 sa rovná strednej hodnote základného súboru μ_2. Alternatívou bude,

že sa tieto dve stredné hodnoty nerovnajú. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 = 0 \\ H_1 &: \ \mu_1 \neq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Testovacou štatistikou je Z štatistika štandardného normálneho rozdelenia:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{10.13}$$

so štandardným normálnym rozdelením pravdepodobnosti $N(0,1)$, kde n_1 a n_2 sú veľkosti výberových súborov.

Kritickú oblasť W_α, do ktorej, ak spadne hodnota testovacej štatistiky Z, tak zamietame nulovú hypotézu H_0, a doplnkovú oblasť W_0, do ktorej, ak spadne hodnota testovacej štatistiky Z, tak nezamietame nulovú hypotézu H_0, určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ \left(-\infty; z_{\alpha/2}\right\rangle \cup \left\langle z_{1-\alpha/2}; \infty\right) \\ W_0 &: \ \left(z_{\alpha/2}; z_{1-\alpha/2}\right) \end{aligned} \tag{10.14}$$

kde $z_{\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia a $z_{1-\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,025}$=-1,96 a kritická hodnota $z_{0,975}$=1,96, a ak je α=0,01, potom je kritická hodnota $z_{0,005}$=-2,576 a kritická hodnota $z_{0,995}$=2,576.

Predpokladajme, že máme inú situáciu, kedy na základe dvoch náhodných výberových súborov chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 menšia ako neznáma hodnota základného súboru μ_2. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ_1 je rovná alebo väčšia ako stredná hodnota základného súboru μ_2 a alternatívna hypotéza hovorí, že je menšia ako μ_2. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu_1 \geq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \geq 0 \\ H_1 &: \ \mu_1 < \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 < 0 \end{aligned}$$

Testovacou štatistikou je Z štatistika daná vzťahom (10.13). Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt štandardného

normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ (-\infty; z_\alpha\rangle \\ W_0 &: \ (z_\alpha; \infty) \end{aligned} \tag{10.15}$$

kde z_α je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,05}$=-1,645, a ak je α=0,01, potom je kritická hodnota $z_{0,01}$=-2,326.

Obdobne predpokladajme, že máme situáciu, kedy na základe dvoch náhodných výberových súborov chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 väčšia ako neznáma stredná hodnota základného súboru μ_2. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ_1 je rovná alebo menšia ako stredná hodnota základného súboru μ_2 a alternatívna hypotéza hovorí, že je väčšia ako μ_2. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu_1 \leq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \leq 0 \\ H_1 &: \ \mu_1 > \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 > 0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť Z štatistiku danú vzťahom (10.13). Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha &: \ \langle z_{1-\alpha}; \infty) \\ W_0 &: \ (-\infty; z_{1-\alpha}) \end{aligned} \tag{10.16}$$

kde $z_{1-\alpha}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,95}$=1,645, a ak je α=0,01, potom je kritická hodnota $z_{0,99}$=2,326.

10.1.2.2 Rozptyly základných súborov nie sú známe a výberové súbory sú veľké

Predpokladajme, že máme náhodné veličiny X_1 a X_2, ktorých rozptyly základného súboru σ_1^2 a σ_2^2 nie sú známe. Na základe dvoch náhodných výberových súborov, ktoré sú oba dostatočne veľké (>30), chceme overiť, či sa neznáma stredná hodnota základného súboru μ_1 rovná alebo nerovná neznámej strednej hodnote základného súboru μ_2. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ_1 sa rovná strednej hodnote

základného súboru μ_2. Alternatívou bude, že sa tieto dve stredné hodnoty nerovnajú. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0: &\ \mu_1 = \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 = 0 \\ H_1: &\ \mu_1 \neq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Testovacou štatistikou je Z štatistika štandardného normálneho rozdelenia:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.17}$$

so štandardným normálnym rozdelením pravdepodobnosti $N(0,1)$, kde s_1^2 a s_2^2 sú rozptyly prvého, resp. druhého výberového súboru.

Kritickú oblasť W_α určíme pomocou kritických hodnôt štandardného normálneho rozdelenia $N(0,1)$ na zvolenej hladine významnosti α.

$$\begin{aligned} W_\alpha: &\ \left(-\infty; z_{\alpha/2}\right\rangle \cup \left\langle z_{1-\alpha/2}; \infty\right) \\ W_0: &\ \left(z_{\alpha/2}; z_{1-\alpha/2}\right) \end{aligned} \tag{10.18}$$

kde $z_{\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia a $z_{1-\alpha/2}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,025}$=-1,96 a kritická hodnota $z_{0,975}$=1,96, a ak je α=0,01, potom je kritická hodnota $z_{0,005}$=-2,576 a kritická hodnota $z_{0,995}$=2,576.

V prípade, ak na základe dvoch náhodných výberových súborov chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 menšia ako neznáma stredná hodnota základného súboru μ_2, potom hypotézy sú:

$$\begin{aligned} H_0: &\ \mu_1 \geq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \geq 0 \\ H_1: &\ \mu_1 < \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 < 0 \end{aligned}$$

Testovacou štatistikou je Z štatistika daná vzťahom (10.17), a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha: &\ \left(-\infty; z_\alpha\right\rangle \\ W_0: &\ \left(z_\alpha; \infty\right) \end{aligned} \tag{10.19}$$

kde z_α je kritická hodnota štandardného normálneho rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,05}$=-1,645, a ak je α=0,01, potom je kritická hodnota $z_{0,01}$=-2,326.

Analogicky, ak chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 väčšia ako neznáma stredná hodnota základného súboru μ_2, potom sú hypotézy:

$$\begin{aligned} H_0: &\ \mu_1 \leq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \leq 0 \\ H_1: &\ \mu_1 > \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 > 0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť Z štatistiku danú vzťahom (10.17) a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha: &\ \langle z_{1-\alpha}; \infty) \\ W_0: &\ (-\infty; z_{1-\alpha}) \end{aligned} \tag{10.20}$$

kde $z_{1-\alpha}$ je kritická hodnota štandardného normálneho rozdelenia na pravej strane rozdelenia.

Ak je α=0,05, potom je kritická hodnota $z_{0,95}$=1,645, a ak je α=0,01, potom je kritická hodnota $z_{0,99}$=2,326.

10.1.2.3 Rozptyly základných súborov nie sú známe, ale predpokladáme, že sú rovnaké a výberové súbory sú malé

Predpokladajme, že máme náhodné veličiny X_1 a X_2, ktorých rozptyly základného súboru σ_1^2 a σ_2^2 nie sú známe, ale predpokladáme, že sú rovnaké, t. j. $\sigma_1^2 = \sigma_2^2$. Na základe dvoch náhodných výberových súborov, z ktorých je aspoň jeden malý (≤ 30), chceme overiť, či sa neznáma stredná hodnota základného súboru μ_1 rovná alebo nerovná neznámej strednej hodnote základného súboru μ_2. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ_1 sa rovná strednej hodnote základného súboru μ_2. Alternatívou bude, že sa tieto dve stredné hodnoty nerovnajú. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0: &\ \mu_1 = \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 = 0 \\ H_1: &\ \mu_1 \neq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Testovacou štatistikou je T štatistika Studentovho t rozdelenia:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10.21}$$

kde S_p^2 predstavuje združený výberový rozptyl, ktorý vypočítame podľa vzťahu:

$$S_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \tag{10.22}$$

z rozptylov s_1^2 a s_2^2 a veľkostí n_1 a n_2 výberových súborov.

Kritickú oblasť W_α určíme pomocou kritických hodnôt Studentovho t rozdelenia na zvolenej hladine významnosti α a s počtom stupňov voľnosti n_1+n_2-2.

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha/2,n_1+n_2-2}\right\rangle \cup \left\langle t_{1-\alpha/2,n_1+n_2-2}; \infty\right) \\ W_0 &: \ \left(t_{\alpha/2,n_1+n_2-2}; t_{1-\alpha/2,n_1+n_2-2}\right) \end{aligned} \tag{10.23}$$

kde $t_{\alpha/2,n_1+n_2-2}$ je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia a $t_{1-\alpha/2,n_1+n_2-2}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 19, potom je kritická hodnota $t_{0,025,19}$=-2,093 a kritická hodnota $t_{0,975,19}$=2,093.

V prípade, ak na základe dvoch náhodných výberových súborov chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 menšia ako neznáma stredná hodnota základného súboru μ_2, potom hypotézy sú:

$$\begin{aligned} H_0 &: \ \mu_1 \geq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \geq 0 \\ H_1 &: \ \mu_1 < \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 < 0 \end{aligned}$$

Testovacou štatistikou je T štatistika daná vzťahom (10.22), a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha,n_1+n_2-2}\right\rangle \\ W_0 &: \ \left(t_{\alpha,n_1+n_2-2}; \infty\right) \end{aligned} \tag{10.24}$$

kde t_{α,n_1+n_2-2} je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 19, potom je kritická hodnota $t_{0,05,19}$=-1,729, a ak je α=0,01, potom je kritická hodnota $t_{0,01,19}$=-2,539.

Analogicky, ak chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 väčšia ako neznáma stredná hodnota základného súboru μ_2, potom sú hypotézy:

$$\begin{aligned} H_0 &: \ \mu_1 \leq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \leq 0 \\ H_1 &: \ \mu_1 > \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 > 0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť T štatistiku danú vzťahom (10.22) a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \langle t_{1-\alpha,n_1+n_2-2}; \infty) \\ W_0 &: \ (-\infty; t_{1-\alpha,n_1+n_2-2}) \end{aligned} \tag{10.25}$$

kde $t_{1-\alpha,n_1+n_2-2}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 19, potom je kritická hodnota $t_{0,95,19}$=1,729, a ak je α=0,01, potom je kritická hodnota $t_{0,99,19}$=2,539.

10.1.2.4 Rozptyly základných súborov nie sú známe, nie sú rovnaké a výberové súbory sú malé

Predpokladajme, že máme dve nezávislé náhodné veličiny X_1 a X_2, ktorých rozptyly základného súboru σ_1^2 a σ_2^2 nie sú známe, a nie sú rovnaké (nezávislé súbory hodnôt), t. j. $\sigma_1^2 \neq \sigma_2^2$. Na základe dvoch nezávislých náhodných výberových súborov, z ktorých je aspoň jeden malý (≤ 30), chceme overiť, či sa neznáma stredná hodnota základného súboru μ_1 rovná alebo nerovná neznámej strednej hodnote základného súboru μ_2. Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota základného súboru μ_1 sa rovná strednej hodnote základného súboru μ_2. Alternatívou bude, že sa tieto dve stredné hodnoty nerovnajú. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 = 0 \\ H_1 &: \ \mu_1 \neq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Testovacou štatistikou je T štatistika Studentovho t rozdelenia:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.26}$$

kde s_1^2 a s_2^2 sú rozptyly dvoch nezávislých výberových súborov s veľkosťou n_1, resp. n_2.

Kritickú oblasť W_α určíme pomocou kritických hodnôt Studentovho t rozdelenia, ktoré na zvolenej hladine významnosti α výpočítame podľa vzťahu:

$$t_\alpha = \frac{t_{\alpha,n_1}\frac{s_1^2}{n_1} + t_{\alpha,n_2}\frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.27}$$

kde t_{α,n_1} je kritická hodnota Studentovho t rozdelenia pre hľadanú hodnotu α a počet stupňov voľnosti $n_1 - 1$ a t_{α,n_2} je kritická hodnota Studentovho t rozdelenia pre hľadanú hodnotu α a počet stupňov voľnosti $n_2 - 1$.

Pre kritickú oblasť W_α a doplnkovú oblasť W_0 potom platí:

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha/2}\right\rangle \cup \left\langle t_{1-\alpha/2}; \infty\right) \\ W_0 &: \ \left(t_{\alpha/2}; t_{1-\alpha/2}\right) \end{aligned} \tag{10.28}$$

kde $t_{\alpha/2}$ je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia vypočítaná podľa vzťahu (10.27) a $t_{1-\alpha/2}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia vypočítaná podľa vzťahu (10.27).

V prípade, ak na základe dvoch nezávislých náhodných výberových súborov chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 menšia ako neznáma stredná hodnota základného súboru μ_2, potom hypotézy sú:

$$\begin{aligned} H_0 &: \ \mu_1 \geq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \geq 0 \\ H_1 &: \ \mu_1 < \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 < 0 \end{aligned}$$

Testovacou štatistikou je T štatistika daná vzťahom (10.26), a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_\alpha\right\rangle \\ W_0 &: \ \left(t_\alpha; \infty\right) \end{aligned} \tag{10.29}$$

kde t_α je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia vypočítaná podľa vzťahu (10.27).

Analogicky, ak chceme overiť, či je neznáma stredná hodnota základného súboru μ_1 väčšia ako neznáma stredná hodnota základného súboru μ_2, potom sú hypotézy:

$$\begin{aligned} H_0 &: \ \mu_1 \leq \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 \leq 0 \\ H_1 &: \ \mu_1 > \mu_2 \ \text{ alebo } \ \mu_1 - \mu_2 > 0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť T štatistiku danú vzťahom (10.26) a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \left\langle t_{1-\alpha}; \infty\right) \\ W_0 &: \ \left(-\infty; t_{1-\alpha}\right) \end{aligned} \tag{10.30}$$

kde $t_{1-\alpha}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia vypočítaná podľa vzťahu (10.27).

10.1.3 Test hypotézy o zhode dvoch stredných hodnôt závislých súborov

V predchádzajúcej kapitole sme si priblížili niektoré možnosti porovnávania stredných hodnôt dvoch základných súborov, pričom sme predpokladali, že súbory sú nezávislé. V medicínskych úlohách sa však často stretávame aj s potrebou porovnávania závislých súborov, napríklad pri hodnotení účinku poskytovanej liečby, kedy pozorujeme hodnoty pacientov pred liečbou a po liečbe. Test hypotézy využívajúci navzájom súvisiace pozorovania označujeme aj ako párový test, prípadne test párových porovnaní. Zmyslom párových testov je zabezpečiť získanie vzoriek, z ktorých by bolo možné pozorovať vplyv skúmanej veličiny alebo veličín, a prípadne minimalizovať vonkajšie vplyvy, ktoré so skúmanou veličinou nemusia súvisieť.

Predpokladajme, že máme výberový súbor štatistických jednotiek, u ktorých pozorujeme náhodné veličiny X_1 a X_2. Realizáciou veličiny X_1 je súbor hodnôt $x_{11}, x_{12}, \ldots, x_{1n}$ a realizáciou veličiny X_2 je súbor hodnôt $x_{21}, x_{22}, \ldots, x_{2n}$, t. j. od každej štatistickej jednotky (napríklad pacient) sme získali dve hodnoty (napríklad pred liečbou a po liečbe). Oba súbory hodnôt teda musia byť rovnako veľké. Namiesto vykonávania analýzy s jednotlivými pozorovaniami a porovnávania strednej hodnoty μ_1 prvého základného súboru so strednou hodnotou μ_2 druhého základného súboru, použijeme ako pozorovanú veličinu množinu n rozdielov medzi pármi pozorovaní a vykonáme takzvaný párový t-test.

Potom testujeme nulovú hypotézu, ktorá hovorí, že stredná hodnota párových rozdielov základného súboru μ_D je rovná nule (medzi pármi pozorovaní nie sú rozdiely). Alternatívou bude, že sa stredná hodnota rozdielov nerovná nule. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \mu_D = 0 \\ H_1 &: \ \mu_D \neq 0 \end{aligned}$$

kde $\mu_D = \mu_1 - \mu_2$.

Testovacou štatistikou je T štatistika Studentovho t rozdelenia s $n-1$ stupňami voľnosti:

$$T = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} \tag{10.31}$$

kde $\bar{D}$ je aritmetický priemer rozdielov, S_D je smerodajná odchýlka rozdielov a n je počet rozdielov párového pozorovania.

Kritickú oblasť W_α určíme pomocou kritických hodnôt Studentovho t rozdelenia pre zvolenú hladinu významnosti α a $n-1$ stupňov voľnosti. Pre kritickú oblasť W_α a doplnkovú oblasť W_0 potom platí:

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha/2,n-1}\right\rangle \cup \left\langle t_{1-\alpha/2,n-1}; \infty\right) \\ W_0 &: \ \left(t_{\alpha/2,n-1}; t_{1-\alpha/2,n-1}\right) \end{aligned} \tag{10.32}$$

kde $t_{\alpha/2,n-1}$ je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia a $t_{1-\alpha/2,n-1}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 12, potom je kritická hodnota $t_{0,025,12}$=-2,179 a kritická hodnota $t_{0,975,12}$=2,179. Ak je α=0,01, potom je kritická hodnota $t_{0,005,12}$=-3,055 a kritická hodnota $t_{0,995,12}$=3,055.

V prípade, ak na základe párových rozdielov chceme overiť, či je neznáma stredná hodnota rozdielov základného súboru μ_D menšia ako nula, potom hypotézy sú:

$$\begin{aligned} H_0 &: \ \mu_D \geq 0 \\ H_1 &: \ \mu_D < 0 \end{aligned}$$

Testovacou štatistikou je T štatistika daná vzťahom (10.31), a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \left(-\infty; t_{\alpha,n-1}\right\rangle \\ W_0 &: \ \left(t_{\alpha,n-1}; \infty\right) \end{aligned} \tag{10.33}$$

kde $t_{\alpha,n-1}$ je kritická hodnota Studentovho t rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 12, potom je kritická hodnota $t_{0,05,12}$=-1,782, a ak je α=0,01, potom je kritická hodnota $t_{0,005,12}$=-2,681.

Analogicky, ak chceme na základe párových rozdielov overiť, či je neznáma stredná hodnota rozdielov základného súboru μ_D väčšia ako nula, potom sú hypotézy:

$$\begin{aligned} H_0 &: \ \mu_D \leq 0 \\ H_1 &: \ \mu_D > 0 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť T štatistiku danú vzťahom (10.31) a pre kritickú oblasť W_α a doplnkovú oblasť W_0 platí:

$$\begin{aligned} W_\alpha &: \ \left\langle t_{1-\alpha,n-1}; \infty\right) \\ W_0 &: \ \left(-\infty; t_{1-\alpha,n-1}\right) \end{aligned} \tag{10.34}$$

kde $t_{1-\alpha,n-1}$ je kritická hodnota Studentovho t rozdelenia na pravej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 12, potom je kritická hodnota $t_{0,95,12}$=1,782, a ak je α=0,01, potom je kritická hodnota $t_{0,995,12}$=2,681.

10.2 Testy hypotéz o rozptyloch

10.2.1 Test hypotézy o rozptyle základného súboru

Predpokladajme, že máme náhodný výberový súbor X, ktorého hodnoty sú $x_1, x_2, \ldots, x_n$, a sú výberom zo základného súboru s normálnym rozdelením $N\left(\mu, \sigma^2\right)$. Na základe náhodného výberového súboru chceme overiť, či sa neznáma hodnota rozptylu základného súboru σ^2 rovná alebo nerovná známej hodnote rozptylu σ_0^2. Potom testujeme nulovú hypotézu, ktorá hovorí, že rozptyl základného súboru σ^2 sa rovná tejto konštante σ_0^2. Alternatívou bude, že sa tejto konštante σ_0^2 nerovná. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \sigma^2 = \sigma_0^2 \\ H_1 &: \ \sigma^2 \neq \sigma_0^2 \end{aligned}$$

Testovacou štatistikou bude χ^2 štatistika s $n-1$ stupňami voľnosti a s χ^2 rozdelením pravdepodobnosti:

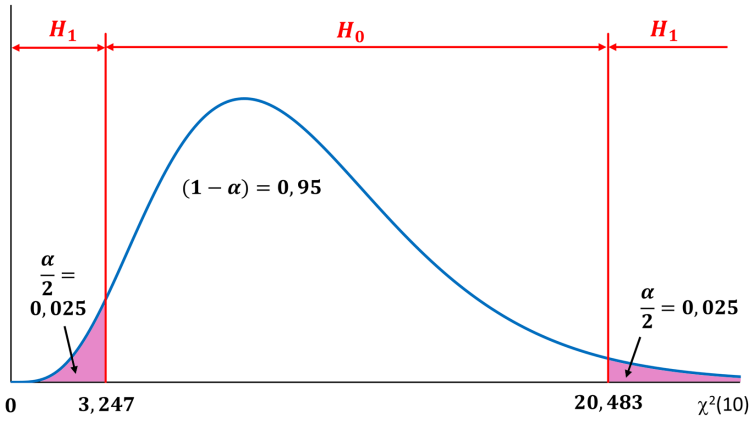
$$\chi^2 = \frac{(n-1)\, S^2}{\sigma_0^2} \tag{10.35}$$

Kritickú oblasť W_α, do ktorej, ak spadne hodnota testovacej štatistiky χ^2, tak zamietame nulovú hypotézu H_0, a doplnkovú oblasť W_0, do ktorej, ak spadne hodnota testovacej štatistiky χ^2, tak nezamietame nulovú hypotézu H_0, určíme pomocou kritických hodnôt χ^2 rozdelenia na zvolenej hladine významnosti α a s počtom stupňov voľnosti $n-1$.

$$\begin{aligned} W_\alpha &: \ \left(0; \chi^2_{\alpha/2}(n-1)\right\rangle \cup \left\langle \chi^2_{1-\alpha/2}(n-1); \infty\right) \\ W_0 &: \ \left(\chi^2_{\alpha/2}(n-1); \chi^2_{1-\alpha/2}(n-1)\right) \end{aligned} \tag{10.36}$$

kde $\chi^2_{\alpha/2}(n-1)$ je kritická hodnota χ^2 rozdelenia na ľavej strane rozdelenia a $\chi^2_{1-\alpha/2}(n-1)$ je kritická hodnota χ^2 rozdelenia na pravej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 10, potom je kritická hodnota $\chi^2_{0,025}(10)$=3,247 a kritická hodnota $\chi^2_{0,975}(10)$=20,483.



Obr. 10.4: Oblasti prijatia alebo neprijatia nulovej hypotézy obojstranného χ^2 testu na hladine významnosti 0,05 a s počtom stupňov voľnosti 10.

Oblasti pre rozhodnutie o prijatí alebo neprijatí nulovej hypotézy H_0 môžeme znázorniť graficky, napríklad tak ako je to na obrázku 10.4.

Predpokladajme, že máme inú situáciu, kedy na základe náhodného výberového súboru chceme overiť, či je neznáma hodnota rozptylu základného súboru σ^2 menšia ako známa hodnota σ_0^2. Potom testujeme nulovú hypotézu, ktorá hovorí, že rozptyl základného súboru σ^2 je rovný alebo väčší ako σ_0^2 a alternatívna hypotéza hovorí, že je menší ako σ_0^2. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \sigma^2 \geq \sigma_0^2 \\ H_1 &: \ \sigma^2 < \sigma_0^2 \end{aligned}$$

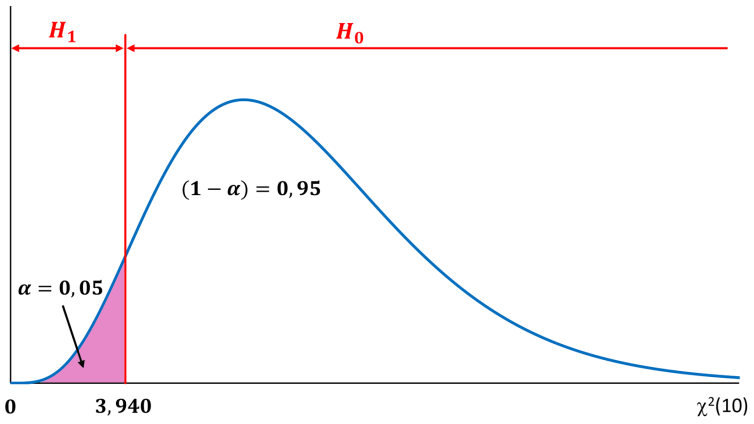
Testovacou štatistikou je opäť χ^2 štatistika daná vzťahom (10.35). Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt χ^2 rozdelenia na zvolenej hladine významnosti α a s počtom stupňov voľnosti $n-1$.

$$\begin{aligned} W_\alpha &: \ \left(0; \chi_\alpha^2(n-1)\right\rangle \\ W_0 &: \ \left(\chi_\alpha^2(n-1); \infty\right) \end{aligned} \tag{10.37}$$

kde $\chi_\alpha^2(n-1)$ je kritická hodnota χ^2 rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 10, potom je kritická hodnota $\chi_{0,05}^2(10)$=3,940. Oblasti pre rozhodnutie o prijatí alebo neprijatí nulovej hypotézy H_0 môžeme taktiež znázorniť graficky, napríklad tak ako je to na obrázku 10.5.

Obdobne predpokladajme, že máme situáciu, kedy na základe náhodného výberového súboru chceme overiť, či je neznámy rozptyl základného súboru σ^2



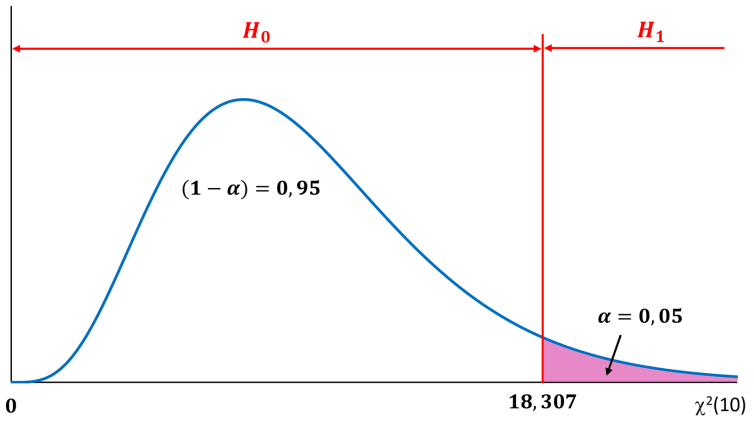
Obr. 10.5: Oblasti prijatia alebo neprijatia nulovej hypotézy ľavostranného χ^2 testu na hladine významnosti 0,05 a s počtom stupňov voľnosti 10.

väčší ako známa hodnota rozptylu σ_0^2. Potom testujeme nulovú hypotézu, ktorá hovorí, že rozptyl základného súboru σ^2 je rovný alebo menší ako σ_0^2 a alternatívna hypotéza hovorí, že je väčší ako σ_0^2. Nulovú a alternatívnu hypotézu potom zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \sigma^2 \leq \sigma_0^2 \\ H_1 &: \ \sigma^2 > \sigma_0^2 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť χ^2 štatistiku danú vzťahom (10.35).

Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt χ^2 rozdelenia na hladine významnosti α a s počtom stupňov voľnosti $n-1$.



Obr. 10.6: Oblasti prijatia alebo neprijatia nulovej hypotézy pravostranného χ^2 testu na hladine významnosti 0,05 a s počtom stupňov voľnosti 10.

$$\begin{aligned} W_\alpha &: \ \left\langle \chi^2_{1-\alpha}(n-1); \infty \right) \\ W_0 &: \ \left(0; \chi^2_{1-\alpha}(n-1)\right) \end{aligned} \tag{10.38}$$

kde $\chi^2_{1-\alpha}(n-1)$ je kritická hodnota χ^2 rozdelenia na pravej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 10, potom je kritická hodnota $\chi^2_{0,95}$=18,307. Oblasti pre rozhodnutie o prijatí alebo neprijatí nulovej hypotézy H_0 môžeme taktiež znázorniť graficky, napríklad tak ako je to na obrázku 10.6.

10.2.2 Test hypotézy o zhode rozptylov dvoch základných súborov

V prípadoch testovania hypotéz o stredných hodnotách, napríklad aj tých, ktoré boli uvedené v predchádzajúcich kapitolách, sa objavujú predpoklady o zhode rozptylov základných súborov. Ak rozptyly základných súborov nie sú známe, potom je možné tento predpoklad overiť porovnaním rozptylov výberových súborov získaných zo základných súborov.

Predpokladajme, že máme dva nezávislé výberové súbory X_1 a X_2 s rozsahom n_1 a n_2, ktoré sú náhodným výberom zo základných súborov s normálnym rozdelením $N\left(\mu_1, \sigma_1^2\right)$, resp. $N\left(\mu_2, \sigma_2^2\right)$, pričom rozptyly základných súborov σ_1^2 a σ_2^2 nepoznáme. Taktiež predpokladajme, že na základe náhodných výberových súborov chceme overiť, či existuje rozdiel medzi rozptylom prvého a rozptylom druhého základného súboru. Napríklad, ak sú priemerné hodnoty dvoch súborov rovnaké, a chceme vedieť, v ktorom súbore sa hodnoty viac menia.

Potom, testovanie týkajúce sa porovnania dvoch rozptylov základného súboru je zvyčajne založené na zisťovaní pomeru týchto rozptylov. Preto, keď testujeme hypotézu, že rozptyly dvoch základných súborov sú rovnaké, v skutočnosti testujeme hypotézu, že ich pomer je rovný 1. Sformulujme teda nulovú a alternatívnu hypotézu nasledovne:

$$\begin{aligned} H_0 &: \ \sigma_1^2 = \sigma_2^2 \\ H_1 &: \ \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

Testovacou štatistikou na overenie nulovej hypotézy bude F štatistika s Fisherovým rozdelením pravdepodobnosti a s $n_1 - 1$ a $n_2 - 1$ stupňami voľnosti :

$$F = \frac{S_1^2}{S_2^2} \tag{10.39}$$

kde ako realizáciu výberového rozdelenia použijeme rozptyly s_1^2 a s_2^2 vypočítané z individuálnych výberových súborov. Do čitateľa testovacej štatistiky F umiestňujeme väčšiu hodnotu rozptylu a do menovateľa tú menšiu. Výsledná hodnota tak bude vždy väčšia alebo nanajvýš rovná jednej (ako pravostranný

test). Čím viac sa tento pomer bude odchyľovať od 1, tým silnejší bude dôkaz o nerovnosti rozptylov základných súborov.

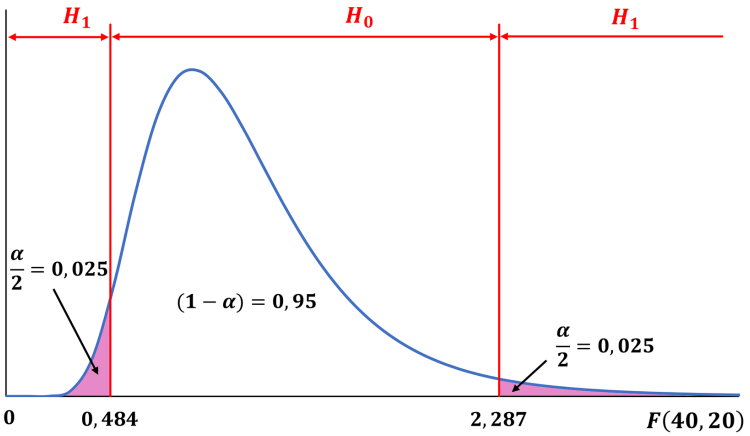
Kritickú oblasť W_α, do ktorej ak spadne hodnota testovacej štatistiky F, tak zamietame nulovú hypotézu H_0, a doplnkovú oblasť W_0, do ktorej ak spadne hodnota testovacej štatistiky F, tak nezamietame nulovú hypotézu H_0, určíme pomocou kritických hodnôt F rozdelenia na zvolenej hladine významnosti α a s počtom stupňov voľnosti $n_1 - 1$ a $n_2 - 1$.

$$\begin{array}{ll} W_\alpha: & \big(0; F_{\alpha/2}(n_1-1, n_2-1)\big\rangle \cup \big\langle F_{1-\alpha/2}(n_1-1, n_2-1); \infty\big) \\ W_0: & \big(F_{\alpha/2}(n_1-1, n_2-1); F_{1-\alpha/2}(n_1-1, n_2-1)\big) \end{array} \tag{10.40}$$

kde $F_{\alpha/2}(n_1 - 1, n_2 - 1)$ je kritická hodnota F rozdelenia na ľavej strane rozdelenia a $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ je kritická hodnota F rozdelenia na pravej strane rozdelenia. Prakticky porovnávame výsledok testovacej štatistiky F vždy s kritickou hodnotou na pravej strane, keďže do čitateľa sme umiestnili väčšiu z porovnávaných hodnôt rozptylov.

Ak je α=0,05 a počet stupňov voľnosti napríklad 40 a 20, potom je kritická hodnota $F_{0,025}(40, 20)$=0,484 a kritická hodnota $F_{0,975}(40, 20)$=2,287.

Oblasti pre rozhodnutie o prijatí alebo neprijatí nulovej hypotézy H_0 môžeme znázorniť graficky, napríklad tak ako je to na obrázku 10.7.



Obr. 10.7: Oblasti prijatia alebo neprijatia nulovej hypotézy obojstranného F testu na hladine významnosti 0,05 a s počtom stupňov voľnosti 40 a 20.

Predpokladajme, že máme inú situáciu, kedy na základe dvoch náhodných výberových súborov chceme overiť, či je neznáma hodnota rozptylu základného súboru σ_1^2 menšia ako neznáma hodnota rozptylu základného súboru σ_2^2. Potom nulovú a alternatívnu hypotézu zapíšeme nasledovne:

$$\begin{array}{ll} H_0: & \sigma_1^2 \geq \sigma_2^2 \\ H_1: & \sigma_1^2 < \sigma_2^2 \end{array}$$

Testovacou štatistikou je F štatistika daná vzťahom (10.39) a kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt F rozdelenia na zvolenej hladine významnosti α a s počtom stupňov voľnosti $n_1 - 1$ a $n_2 - 1$.

$$\begin{aligned} W_\alpha &: \ (0; F_\alpha(n_1 - 1, n_2 - 1)\rangle \\ W_0 &: \ (F_\alpha(n_1 - 1, n_2 - 1); \infty) \end{aligned} \tag{10.41}$$

kde $F_\alpha(n_1-1, n_2-1)$ je kritická hodnota F rozdelenia na ľavej strane rozdelenia.

Ak je α=0,05 a počet stupňov voľnosti napríklad 40 a 10, potom je kritická hodnota $F_{0,05}(40, 20)$=0,544.

Obdobne predpokladajme, že máme situáciu, kedy na základe dvoch náhodných výberových súborov chceme overiť, či je neznámy rozptyl základného súboru σ_1^2 väčší ako neznámy rozptyl základného súboru σ_2^2. Potom nulovú a alternatívnu hypotézu zapíšeme nasledovne:

$$\begin{aligned} H_0 &: \ \sigma_1^2 \leq \sigma_2^2 \\ H_1 &: \ \sigma_1^2 > \sigma_2^2 \end{aligned}$$

Testovacia štatistika sa nemení, používame pre dané podmienky opäť F štatistiku danú vzťahom (10.39).

Kritickú oblasť W_α a doplnkovú oblasť W_0 určíme pomocou kritických hodnôt F rozdelenia na hladine významnosti α a s počtom stupňov voľnosti $n_1 - 1$ a $n_2 - 1$.

$$\begin{aligned} W_\alpha &: \ \langle F_{1-\alpha}(n_1 - 1, n_2 - 1); \infty) \\ W_0 &: \ (0; F_{1-\alpha}(n_1 - 1, n_2 - 1)) \end{aligned} \tag{10.42}$$

kde $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ je kritická hodnota F rozdelenia na pravej strane rozdelenia.

10.3 Analýza rozptylu (ANOVA)

Analýza rozptylu (*analysis of variance*), predstavuje metódu rozdelenia celkového rozptylu vypočítaného zo súboru údajov na zložky, z ktorých každá predstavuje množstvo celkového rozptylu, ktorý možno pripísať konkrétnemu zdroju variácie. Výsledky takéhoto rozdelenia je potom možné použiť na odhad a testovanie hypotéz o rozptyloch a stredných hodnotách základných súborov. Najčastejšie sa v rámci metód analýzy rozptylu zameriavame na testovanie hypotéz o stredných hodnotách, ak máme viac ako dva základné súbory alebo dve a viac náhodných veličín, pričom tieto metódy sú často aplikované v medicínskych štúdiách a experimentoch.

Analýza rozptylu sú spoločne s lineárnou regresiou najbežnejšie používané analytické nástroje, ktorých koncepčným základom sú štatistické modely poskytujúce užitočné reprezentácie vzťahov medzi viacerými veličinami súčasne. Štatistický model pritom chápeme ako matematické vyjadrenie vzťahov medzi veličinami. Napríklad, štatistický model môžeme použiť na opis toho, ako náhodné veličiny navzájom súvisia v kontexte, v ktorom môže byť hodnota jednej výslednej veličiny, modelovaná ako funkcia jednej alebo viacerých vonkajších veličín. Týmto spôsobom nás zaujíma, do akej miery variabilitu výsledkov možno vysvetliť náhodnými veličinami, ktoré boli pozorované alebo kontrolované v rámci realizácie experimentu.

Predpokladmi pre použitie analýzy rozptylu budú, že chceme overiť zhodu, resp. porovnať stredné hodnoty viacerých základných súborov. Avšak, musíme si uvedomiť, že všeobecnými podmienkami pre použitie metód analýzy rozptylu je, aby všetky porovnávané výberové súbory pochádzali zo základných súborov s normálnym rozdelením, výberové súbory boli nezávislé a rozptyly všetkých základných súborov sa rovnali. V prípade, že porovnávané súbory nie sú výberom z normálneho rozdelenia, potom je možné použiť neparametrické alternatívy analýzy rozptylu.

Ako sme už naznačili vyššie, analýzou rozptylu porovnávame stredné hodnoty viac ako dvoch základných súborov, ktoré spravidla vzniknú rozdelením jedného základného súboru podľa hodnôt nejakého faktora. Napríklad, súbor pacientov, u ktorých sledujeme účinok liečby, môžeme rozdeliť podľa faktora, ktorým je typ podaného lieku (liek A, liek B, liek C atď.). Konkrétne lieky sú teda hodnoty, presnejšie úrovne daného faktora. Takto môžeme overiť, či je účinok rôznych liekov porovnateľný. Často môžeme testovať, či hodnota sledovanej veličiny závisí od úrovní viacerých faktorov. Mohli by to byť vplyv lieku, pohlavie pacientov, fajčiar či nefajčiar a pod. Potom, v závislosti od testovania vplyv úrovní jedného alebo viacerých faktorov na variabilitu hodnôt analyzovanej veličiny hovoríme o jednofaktorovej alebo o viacfaktorovej analýze rozptylu.

10.3.1 Jednofaktorová analýza rozptylu

Jednofaktorová analýza rozptylu reprezentuje najjednoduchší prípad analýzy rozptylu, kedy skúmame jeden zdroj variácie (zmien sledovanej veličiny). Vo všeobecnosti sa jedná o rozšírenie postupu t testu (viď napríklad kapitola 10.1.2.4) s dvoma nezávislými výberovými súbormi na tri alebo viac výberových súborov. Pri praktickej realizácii teda chceme použiť jednofaktorovú analýzu rozptylu na testovanie nulovej hypotézy, že tri alebo viaceré metódy

liečby sú rovnako účinné. Samotný výskum je potom navrhnutý a realizovaný tak, že pacientom je náhodným spôsobom priradený jeden zo sledovaných typov liečby, ktoré sú predmetom štúdie. Hodnoty získavané z jednotlivých výberových skupín takto postavenej randomizovanej štúdie by sme mohli zapísať do tabuľky, napríklad tak ako je to uvedené v tabuľke 10.1.

Tabuľka 10.1: Tabuľka hodnôt náhodných výberových súborov.

	Výberový súbor (liečba)					
	1	2	3	$\cdots$	k	
	x_{11}	x_{21}	x_{31}	$\cdots$	x_{k1}	
	x_{12}	x_{22}	x_{32}	$\cdots$	x_{k2}	
	x_{13}	x_{23}	x_{33}	$\cdots$	x_{k3}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	x_{1n_1}	x_{2n_2}	x_{3n_3}	$\cdots$	x_{kn_k}	
Suma	$\sum_{i=1}^{n_1} x_{1i}$	$\sum_{i=1}^{n_2} x_{2i}$	$\sum_{i=1}^{n_3} x_{3i}$	$\cdots$	$\sum_{i=1}^{n_k} x_{ki}$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ji}$
Priemer	$\bar{x_1}$	$\bar{x_2}$	$\bar{x_3}$	$\cdots$	$\bar{x_k}$	$\bar{x}$

To znamená, že pozorovanú veličinu X môžeme sledovať u celkového počtu n subjektov alebo objektov, rozdelených do k náhodných nezávislých výberových súborov zo základného súboru, pričom platí, že $n = n_1 + n_2 + \ldots + n_k$. Pre každý výberový súbor vieme vypočíta samostatne sumu aj priemernú hodnotu sledovanej veličiny. Taktiež vieme vypočítať celkovú sumu aj celkovú priemernú hodnotu zo všetkých hodnôt vo všetkých skupinách.

Testovať budeme nulovú hypotézu, ktorá predpokladá, že všetky stredné hodnoty (priemery) sú rovnaké, napríklad v rôznych sledovaných liečbach nie sú rozdiely. Alternatívnou hypotézou bude, že stredná hodnota aspoň jedného súboru nie je rovnaká. Potom nulovú a alternatívnu hypotézu zapíšeme:

$$H_0: \ \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \ \text{, kde} \ \ k > 2$$
$$H_1: \ \text{nie všetky } \mu_j \text{ sú rovnaké}$$

Prvým krokom pri výpočte testovacej štatistiky jednofaktorovej analýzy rozptylu je rozdelenie celkovej variability prítomnej v pozorovaných údajoch na jej základné zložky, z ktorých každá súvisí s identifikovateľným zdrojom (v našom prípade výberovým súborom). Pojem variabilita používaný v tomto kontexte budeme vzťahovať k sume druhých mocnín odchýlok pozorovaní od ich priemernej hodnoty. Potm celková variabilita bude predstavovať súčet druhých mocnín odchýlok jednotlivých pozorovaní od celkového priemeru všetkých pozorovaní a je vypočítaná ako:

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \bar{x})^2 \tag{10.43}$$

Celková variabilita výberových údajov SST predstavuje súčet variability v rámci skupín (výberových súborov), t. j. vnútroúrovňovej variability a variability medzi skupinami, t. j. medziúrovňovej variability.

Vnútroúrovňovú variabilitu určíme ako súčet všetkých súčtov štvorcových odchýlok jednotlivých pozorovaní od ich priemeru v rámci danej skupiny. Výpočet vnútroúrovňovej variability vypočítame podľa vzťahu:

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \bar{x_j})^2 \tag{10.44}$$

Medziúrovňovú variabilitu vypočítame tak, že pre každú skupinu údajov vypočítame štvorcovú odchýlku priemeru skupiny od celkového priemeru a výsledok vynásobíme veľkosťou danej skupiny. Potom tieto výsledky spočítame a získame hodnotu medziúrovňovej variability:

$$SSA = \sum_{j=1}^{k} n_j (\bar{x_j} - \bar{x})^2 \tag{10.45}$$

Z vnútroúrovňovej a medziúrovňovej variability je ďalej možné získať dva odhady rozptylu základného súboru σ^2, ktoré sú reprezentované priemernou hodnotou danej variability. Ak platia predpoklady analýzy rozptylu uvedené vyššie, potom vydelením variability príslušným počtom stupňov voľnosti získavame priemerné variability ako nezávislé a neskreslené odhady rozptylu. Pre vnútroúrovňovú priemernú hodnotu variability bude platiť vzťah:

$$MSW = \frac{SSW}{n - k} \tag{10.46}$$

kde n je počet všetkých hodnôt a k je počet skupín (výberových súborov), ktoré sú analyzované.

Pre medziúrovňovú priemernú hodnotu variability platí vzťah:

$$MSA = \frac{SSA}{k - 1} \tag{10.47}$$

Ak sú tieto dva odhady rozptylov (vnútroúrovňový a medziúrovňový) približne rovnaké, potom sa dá očakávať, že bude platiť nulová hypotéza. A naopak, ak bude medziúrovňová stredná hodnota variability väčšia ako tá vnútroúrovňová, potom bude pravdepodobne platiť alternatívna hypotéza. Testovacou

štatistikou analýzy rozptylu potom bude F štatistika, ktorá je pomerom medziúrovňovej a vnútroúrovňovej priemernej hodnoty variability:

$$F = \frac{MSA}{MSW} \tag{10.48}$$

Po vypočítaní testovacej štatistiky určíme kritickú hodnotu F rozdelenia, ktorá rozhodne o zamietnutí alebo nezamietnutí nulovej hypotézy. Kritickú hodnotu určíme pre zvolenú hladinu významnosti α a počty stupňov voľnosti medzi skupinami $(k-1)$ a v rámci skupín $(n-k)$. Ak bude vypočítaná hodnota rovná alebo väčšia ako kritická hodnota, potom zamietame nulovú hypotézu. Ak bude vypočítaná hodnota menšia ako kritická hodnota, potom nezamietame nulovú hypotézu. Oblasť zamietnutia H_0 je teda definovaná takto:

$$F \geq F_{1-\alpha}(k-1, n-k) \tag{10.49}$$

Ak na základe testovacej štatistiky nezamietneme nulovú hypotézu, záverom je, že stredné hodnoty základného súboru sa navzájom výrazne nelíšia. Ak zamietame nulovú hypotézu, potom záverom je, že nie všetky stredné hodnoty základného súboru sú rovnaké.

Zamietnutie nulovej hypotézy nám nič nehovorí o tom, ktoré stredné hodnoty sa líšia. Preto vyvstáva otázka, ktoré dvojice stredných hodnôt sú odlišné a ako to overiť? Je teda potrebné vykonať test významnosti na každej jednej dvojici sledovaných výberových súborov. Na tieto účely sa v praxi bežne používa niekoľko viacnásobných porovnávacích postupov. Medzi najčastejšie využívané patria Tukeyho test, Duncanov test či Bonferroniho test.

10.3.1.1 Tukeyho test

Tukeyho postup viacnásobného porovnávania sa často používa na testovanie nulovej hypotézy, že všetky možné dvojice stredných hodnôt výberových súborov (skupín) sú rovnaké, za predpokladu, že sú všetky pozorované výberové súbory rovnako veľké.

Tukeyho test využíva jedinú hodnotu, s ktorou sú porovnávané všetky rozdiely v stredných hodnotách dvojíc výberových súborov. Túto hodnotu vypočítame podľa vzťahu:

$$T_k = q_{\alpha,k,n-k}\sqrt{\frac{MSW}{n_i}} \tag{10.50}$$

kde α je zvolená hladina významnosti, k je počet stredných hodnôt, resp. počet skupín, n je celkový počet pozorovaní v experimente (počet všetkých hodnôt

spolu), n_i je počet pozorovaní v jednej skupine (všetky skupiny sú rovnako veľké), MSW je vnútroúrovňová priemerná hodnota variability analýzy rozptylu (ANOVA) a q je tabuľková hodnota percentilu pre dané α, n a $n-k$, ktorú získame zo štatistických tabuliek.

Potom všetky absolútne hodnoty rozdielov medzi pármi stredných hodnôt výberových súborov porovnávame s hodnotou T_k. každý rozdiel, ktorý presahuje hodnotu T_k, vyhodnocujeme ako významný, t. j. konštatujeme, že významný je rozdiel medzi tými dvoma výberovými súbormi, u ktorých rozdiel stredných hodnôt v absolútnom vyjadrení prevýšil hodnotu T_k.

Predošlý postup je použiteľný pre rovnaké veľkosti výberových súborov. Ak je realizovaný experiment, v ktorom je veľkosť jednotlivých výberových súborov rôzna, potom je možné Tukeyho test upraviť tak, aby tieto rozdiely boli zohľadnené. Rozšírený Tukeyho test, ktorý umožňuje takého porovnávanie označujeme aj ako Tukeyho-Kramerov test.

Hodnotu, s ktorou porovnávame rozdiely medzi strednými hodnotami výberových súborov potom vypočítavame podľa vzťahu:

$$T_k^* = q_{\alpha,k,n-k}\sqrt{\frac{MSW}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{10.51}$$

kde n_i a n_j sú veľkosti výberových súborov, ktoré porovnávame.

Aj tu platí, že akákoľvek absolútna hodnota rozdielu medzi dvoma strednými hodnotami výberových súborov, ktorá je väčšia ako T_k^* je považovaná za významnú.

10.3.1.2 Duncanov test

Duncanov test je založený na porovnávaní rozdielov stredných hodnôt jednotlivých výberových súborov s najmenším významným rozsahom.

Najmenší významný rozsah pre podskupinu p priemerov výberových súborov vypočítame podľa vzťahu:

$$R_p = r_p\sqrt{\frac{MSW}{n_i}} \tag{10.52}$$

kde r_p je kritická hodnota Duncanovho testu, ktorá závisí od stupňov voľnosti $n-k$ a počtu priemerov p v porovnávanej podskupine priemerov. Kritické hodnoty potrebné pre jednotlivé porovnania nájdeme v štatistických tabuľkách. MSW je vnútroúrovňová priemerná hodnota variability analýzy rozptylu

(ANOVA) a n_i je veľkosť výberových súborov (za predpokladu, že sú rovnaké. Ak veľkosti vzoriek nie sú rovnaké, potom sa n_i nahradí harmonickým priemerom veľkostí vzoriek.

Porovnávanie vykonávame tak, že všetky stredné hodnoty výberových súborov usporiadame podľa veľkosti a porovnávanie začíname medzi najväčšou a najmenšou strednou hodnotou. Rozsah (počet) stredných hodnôt p bude označovať počet všetkých stredných hodnôt od najväčšieho po najmenší (ak porovnávame 7 výberových súborov, potom pri takom porovnaní bude p=7). Ak bude rozdiel medzi najväčšou a najmenšou strednou hodnotou väčší ako najmenší významný rozsah R_p (v tomto prípade R_7), potom je rozdiel významný a výberové súbory, ktorým patria stredné hodnoty sú odlišné. Ďalej pokračujeme v porovnávaní medzi hodnotou druhého najväčšieho rozdielu stredných hodnôt a súvisiacim najmenším významným rozsahom R_p. Opäť, ak bude tento rozdiel väčší, potom je významný a aj tieto výberové súbory, ktorým tieto stredné hodnoty patria sa líšia. Takto sa pokračuje v porovnávaní ďalších rozdielov medzi strednými hodnotami párov výberových súborov s príslušným najmenším významným rozsahom a identifikuje sa, ktoré rozdiely sú, a ktoré nie sú významné.

10.3.1.3 Bonferroniho test

Bonferroniho test, resp. Bonferroniho metóda, predstavuje ďalší veľmi často používaný viacnásobný spôsob porovnávania stredných hodnôt výberových súborov. Rovnako ako pri iných metódach, napríklad Tukeyho teste, sa Bonferroniho metóda snaží zachovať celkovú hladinu významnosti α pre súčet všetkých párových testov. Metóda vychádza z predpokladu, že pravdepodobnosť výskytu jednej alebo viacerých skupín udalostí (napríklad A_1, A_2 atď.), ktoré sa vyskytnú, je menšia alebo rovná súčtu pravdepodobností. To znamená, že Bonferroniho nerovnosť vyjadruje:

$$P\left(A_1 \cup A_2 \cup \ldots \cup A_n\right) \leq \sum_{i=1}^{n} P\left(A_i\right) \tag{10.53}$$

Ak takéto udalosti nie sú disjunktné, potom sa časť pravdepodobnosti počíta dva alebo aj viackrát. V takomto prípade existuje vo vyššie uvedenom vzťahu nerovnosť.

Pri Bonferroniho metóde teda jednoducho vydelíme požadovanú hladinu významnosti počtom jednotlivých párov, u ktorých testujeme zhodu ich stredných hodnôt. To znamená, že namiesto testovania na hladine významnosti α, testu-

jeme na hladine významnosti α/k, kde k je počet párových porovnaní. Súčet všetkých členov α/k teda nemôže prekročiť stanovenú hodnotu α.

Napríklad, ak máme tri výberové súbory označené ako A_1, A_2 a A_3, potom existuje k=3 párových porovnaní. Sú to $\mu_{A_1} = \mu_{A_2}$, $\mu_{A_1} = \mu_{A_3}$ a $\mu_{A_2} = \mu_{A_3}$. Ak zvolíme hladinu významnosti α=0,05, potom by sme mali realizovať porovnania a použiť tzv. Bonferroniho korigovanú hladinu významnosti $\alpha/3 \doteq 0,017$. Preto hodnota p nesmie byť väčšia ako 0,017, aby sme zamietli nulovú hypotézu a dospeli k záveru, že sa dve stredné hodnoty líšia. Väčšina výpočtových aplikácií používa na porovnávanie stredných hodnôt párov výberových súborov práve Bonferroniho metódu.

10.3.2 Dvojfaktorová analýza rozptylu

Dvojfaktorová analýza rozptylu predstavuje spôsob analýzy údajov z náhodného výberu, kde jednotlivé pozorovania môžeme kategorizovať na základe dvoch kritérií. Napríklad sledovanou veličinou môže byť krvný tlak a klasifikačné veličiny typ liečby a pohlavie. Pozorované údaje je možné prehľadne usporiadať vo forme tabuľky, napríklad tak ako je to uvedené v tabuľke 10.2.

Tabuľka 10.2: Tabuľka hodnôt náhodných výberových súborov kategorizoaných podľa dvoch kritérií.

Faktor 1	**Faktor 2**						
	1	2	3	$\cdots$	k	Suma	Priemer
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1k}	$\sum_{i=1}^{k} x_{1i}$	$\bar{x_{1.}}$
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2k}	$\sum_{i=1}^{k} x_{2i}$	$\bar{x_{2.}}$
3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3k}	$\sum_{i=1}^{k} x_{3i}$	$\bar{x_{3.}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$	$\cdots$	$x_{n_k k}$	$\sum_{i=1}^{k} x_{li}$	$\bar{x_{l.}}$
Suma	$\sum_{i=1}^{n_1} x_{i1}$	$\sum_{i=1}^{n_2} x_{i2}$	$\sum_{i=1}^{n_3} x_{i3}$	$\cdots$	$\sum_{i=1}^{n_k} x_{ik}$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ji}$	
Priemer	$\bar{x_{.1}}$	$\bar{x_{.2}}$	$\bar{x_{.3}}$	$\cdots$	$\bar{x_{.k}}$		$\bar{x_{..}}$

Matematický model pre realizáciu dvojfaktorovej analýzy rozptylu je možné zapísať v tvare:

$$x_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij} \tag{10.54}$$

pre $i = 1, 2, \ldots, l$ a $j = 1, 2, \ldots, k$. x_{ij} je typická hodnota zo základného súboru, μ je neznáma konštanta, β_i reprezentuje vplyv úrovne i faktora 1, τ_j reprezentuje vplyv úrovne j faktora 2 a ϵ_{ij} je reziduálna zložka reprezentujúca všetky zdroje variability, ktoré nepochádzajú od faktora 1 a 2. Zároveň predpokladáme, že platí:

$$\sum_{j=1}^{k} \tau_j = \sum_{i=1}^{l} \beta_i = 0 \tag{10.55}$$

Nulové a alternatívne hypotézy, ktoré je možné testovať budú:

$$\begin{aligned} H_0 &: \ \tau_j = 0, \text{ kde} \qquad j = 1, 2, \ldots, k \\ H_1 &: \ \text{nie všetky } \tau_j = 0 \end{aligned}$$

a

$$\begin{aligned} H_0 &: \ \beta_i = 0, \text{ kde} \qquad i = 1, 2, \ldots, l \\ H_1 &: \ \text{nie všetky } \beta_i = 0 \end{aligned}$$

a teda nulové hypotézy predpokladajú, že všetky úrovne faktora 1, resp. 2 majú rovnaký vplyv.

Obdobne ako pri jednofaktorovej analýze rozptylu, aj tu je možné zapísať, že celková variabilita (súčet štvorcov odchýlok od priemernej hodnoty) je rovná súčtu troch zložiek, t. j. vnútroúrovňovej variabilite faktora 1, vnútroúrovňovej variabilite faktora 2 a medziúrovňovej variabilite.

Vnútroúrovňovú variabilitu faktora 1 určíme ako súčet všetkých súčtov štvorcových odchýlok priemerov jednotlivých pozorovaní v rámci úrovní faktora 1 od celkového priemeru. Výpočet vnútroúrovňovej variability faktora 1 vypočítame podľa vzťahu:

$$SSW1 = \sum_{j=1}^{k} \sum_{i=1}^{l} \left(\bar{x_{i.}} - \bar{x_{..}}\right)^2 \tag{10.56}$$

Vnútroúrovňovú variabilitu faktora 2 určíme ako súčet všetkých súčtov štvorcových odchýlok priemerov jednotlivých pozorovaní v rámci úrovní faktora 2 od celkového priemeru. Výpočet vnútroúrovňovej variability faktora 2 vypočítame podľa vzťahu:

$$SSW2 = \sum_{j=1}^{k} \sum_{i=1}^{l} \left(\bar{x_{.j}} - \bar{x_{..}}\right)^2 \tag{10.57}$$

Celkovú variabilitu vypočítame tak, že pre všetky hodnoty vypočítame štvorcovú odchýlku od celkového priemeru:

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{l} (x_{ij} - \bar{x}_{..})^2 \tag{10.58}$$

Potom medziúrovňovú variabilitu vypočítame tak, že od celkovej variability odpočítame vnútroúrovňové:

$$SSA = SST - SSW1 - SSW2 \tag{10.59}$$

Jednotlivé zložky celkovej variability potom môžeme použiť na odhad rozptylov. Pre vnútroúrovňovú priemernú hodnotu variability úrovné faktora 1 bude platiť vzťah:

$$MSW1 = \frac{SSW1}{l-1} \tag{10.60}$$

kde $l-1$ je počet stupňov voľnosti úrovní faktora 1.

Pre vnútroúrovňovú priemernú hodnotu variability úrovné faktora 1 bude platiť vzťah:

$$MSW2 = \frac{SSW2}{k-1} \tag{10.61}$$

kde $k-1$ je počet stupňov voľnosti úrovní faktora 2.

Pre medziúrovňovú priemernú hodnotu variability platí vzťah:

$$MSA = \frac{SSA}{(l-1)(k-1)} \tag{10.62}$$

Pre vyhodnotenie vplyvu faktora 1 použijeme testovaciu štatistiku F, ktorá je pomerom medziúrovňovej a vnútroúrovňovej priemernej hodnoty variability faktora 1:

$$F = \frac{MSA}{MSW1} \tag{10.63}$$

Po vypočítaní testovacej štatistiky určíme kritickú hodnotu F rozdelenia, ktorá rozhodne o zamietnutí alebo nezamietnutí nulovej hypotézy. Kritickú hodnotu určíme pre zvolenú hladinu významnosti α a počty stupňov voľnosti medzi úrovňami faktora 1 $(l-1)$ a v rámci úrovní oboch faktorov $(l-1)(k-1)$. Ak bude vypočítaná hodnota rovná alebo väčšia ako kritická hodnota, potom zamietame nulovú hypotézu. Ak bude vypočítaná hodnota menšia ako kritická

hodnota, potom nezamietame nulovú hypotézu. Oblasť zamietnutia pre vplyv faktora 1 H_0 je teda definovaná takto:

$$F \geq F_{1-\alpha}(l-1,(l-1)(k-1)) \tag{10.64}$$

Ak na základe testovacej štatistiky nezamietneme nulovú hypotézu, záverom je, že faktor 1 nemá vplyv na hodnoty pozorovanej veličiny. Ak zamietame nulovú hypotézu, potom záverom je, že faktor 1 má vplyv na hodnoty pozorovanej veličiny.

Pre vyhodnotenie vplyvu faktora 2 použijeme testovaciu štatistiku F, ktorá je pomerom medziúrovňovej a vnútroúrovňovej priemernej hodnoty variability faktora 2:

$$F = \frac{MSA}{MSW2} \tag{10.65}$$

Po vypočítaní testovacej štatistiky určíme kritickú hodnotu F rozdelenia, ktorá rozhodne o zamietnutí alebo nezamietnutí nulovej hypotézy. Kritickú hodnotu určíme pre zvolenú hladinu významnosti α a počty stupňov voľnosti medzi úrovňami faktora 2 $(k-1)$ a v rámci úrovní oboch faktorov $(l-1)(k-1)$. Ak bude vypočítaná hodnota rovná alebo väčšia ako kritická hodnota, potom zamietame nulovú hypotézu. Ak bude vypočítaná hodnota menšia ako kritická hodnota, potom nezamietame nulovú hypotézu. Oblasť zamietnutia pre vplyv faktora 2 H_0 je teda definovaná takto:

$$F \geq F_{1-\alpha}(k-1,(l-1)(k-1)) \tag{10.66}$$

Ak na základe testovacej štatistiky nezamietneme nulovú hypotézu, záverom je, že faktor 2 nemá vplyv na hodnoty pozorovanej veličiny. Ak zamietame nulovú hypotézu, potom záverom je, že faktor 2 má vplyv na hodnoty pozorovanej veličiny.

V prípade, že testovanú hypotézu faktora 1 alebo faktora 2 zamietame, budeme pokračujeme testovaním štatistickej významnosti vplyvov faktora 1 alebo 2, napríklad Tukeyho, Duncanovým či ďalšími post hoc testami.

Kapitola 11

Neparametrické testy

V predchádzajúcich kapitolách sme naznačili, že nie vždy sú splnené podmienky a predpoklady o rozložení základného súboru, a vtedy je potrebné použiť iné štatistické postupy. Takzvané neparametrické metódy poskytujú riešenia, ako analyzovať údaje, ak sú porušené základné predpoklady testov tradičných hypotéz, alebo ak potrebujeme vykonať test bez vytvárania predpokladov o výberovom súbore.

Kým v prípade parametrických metód boli postupy zamerané na odhad alebo testovanie hypotézy o jednom alebo viacerých parametroch základného súboru, pričom bolo potrebné poznať funkčnú formu základného súboru pre odvodenie záveru, u neparametrických postupov sa buď nezameriavame na parametre základného súboru, alebo tieto postupy nezávisia od znalostí o výberovom súbore zo základného súboru.

Neparametrické metódy štatistiky teda umožňujú testovanie hypotéz, ktoré nie sú tvrdeniami o hodnotách parametrov základného súboru (niektoré z χ^2 testov dobrej zhody a testy nezávislosti sú príkladmi testov, ktoré majú túto vlastnosť). Neparametrické testy je možné použiť, ak forma výberového súboru základného súboru nie je známa, pričom samotné neparametrické procedúry bývajú výpočtovo jednoduchšie a následne rýchlejšie aplikované. Čas by však nemal byť určujúcim kritériom pre výber neparametrického testu. Navyše, samotnú realizáciu výpočtov vykonávame pomocou rôznych výpočtových nástrojov, kedy úvahy o rýchlosti výpočtu majú byť nahradené znalosťami o rôznych metódach a výberom tej správnej metódy na odvodenie záverov o pozorovaných údajoch. Ďalšou vlastnosťou neparametrických postupov je, že môžu byť použité, keď analyzované údaje pozostávajú iba z klasifikácií, t. j. údaje nemusia byť založené na dostatočne silnej meracej škále, aby umožnili vykonávanie aritmetických operácií. Pri neparametrických metódach si však musíme uvedomiť aj to, že ak ich použijeme na údaje, ktoré je možné spracovať pomocou

parametrických postupov, vedie to k strate sily testu, resp. plytvaniu údajov a tiež to, že pre veľké výberové súbory môže byť aplikácia niektorých neparametrických testov pracná. V nasledujúcich kapitolách si priblížime niektoré najčastejšie aplikované metódy z tejto kategórie induktívnej štatistiky.

11.1 Wilcoxonov test

Wilcoxonov test je testom, ktorého použitie môžme aplikovať v situáciách, kedy sme potrebovali otestovať nulovú hypotézu o základnom súbore, ale neboli splnené podmienky pre použitie Z alebo t testu. Napríklad, ak máme malý výberový súbor, alebo párové výberové súbory, ktoré nemajú normálne rozdelenie a predpoklad o centrálnom rozložení údajov neplatí, potom vhodným neparametrickým postupom môže byť použitie hodnôt rozdielov medzi pozorovaniami a hypotetický parameter polohy, tak ako to definuje Wilcoxonov znamienkovo-poradový test.

Wilcoxonov test predpokladá, že výberový súbor je náhodný, ktorého sledovaná veličina je spojitá, a základný súbor je symetricky rozdelený okolo svojej strednej hodnoty μ. Potom je možné definovať nasledujúcu nulovú a alternatívnu hypotézu, ktoré testujeme na nejakej neznámej strednej hodnote základného súboru μ_0.

$$\begin{aligned} H_0 &: \ \mu = \mu_0 \\ H_1 &: \ \mu \neq \mu_0 \end{aligned}$$

Pre vykonanie Wilcoxonovho testu vypočítame všetky rozdiely pozorovaných hodnôt od hypotetickej strednej hodnoty μ_0:

$$d_i = x_i - \mu_0 \tag{11.1}$$

kde x_i reprezentuje pozorované hodnoty výberového súboru.

Následne priradíme jednotlivým hodnotám rozdielov d_i, zoradených do postupnosti od najmenšieho rozdielu po najväčší, bez ohľadu na znamienko rozdielu, poradové číslo. Napríklad, najmenší rozdiel v absolútnom vyjadrení dostane číslo 1, druhý najmenší rozdiel v absolútnom vyjadrení dostane číslo 2 atď. Ak sú dve alebo viaceré hodnoty $|d_i|$ rovnaké, potom priradíme každej takejto hodnote priemer pozícií ich poradia. Napríklad, ak tretí, štvrtý a piaty rozdiel sú rovnaké, potom každému z týchto rozdielov priradíme priemernú hodnotu poradia: (3+4+5)/3=4. Po priradení poradových čísel jednotlivým rozdielom, týmto poradovým číslam priradíme znamienko daného rozdielu, t. j. ak bol rozdiel záporný, aj poradové číslo bude záporné, a ak bol rozdiel kladný, potom aj

poradové číslo bude kladné. Pre názornosť je možné tieto informácie zapísať do tabuľky, napríklad tak ako je to naznačené v tabuľke 11.1.

Tabuľka 11.1: Tabuľka výpočtov pre Wilcoxonov test.

x_i	$d_i = x_i - \mu_0$	Poradie $\lvert d_i \rvert$	Poradie $\lvert d_i \rvert$ so znamienkom
x_1	$d_1 = x_1 - \mu_0$	r_1^*	sr_1^*
x_2	$d_2 = x_2 - \mu_0$	r_2^*	sr_2^*
x_3	$d_3 = x_3 - \mu_0$	r_3^*	sr_3^*
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	$d_n = x_n - \mu_0$	r_n^*	sr_n^*

Hodnota r_i^* v tabuľke 11.1 predstavuje poradové číslo absolútnej hodnoty rozdielu d_i a sr_i^* je poradové číslo rozdielu d_i, ktorému je priradené znamienko tohto rozdielu.

Testovacou štatistikou Wilcoxonovho testu bude:

$$T_W = min\left(\sum_{i=1}^{n} sr_i^*(+), \sum_{i=1}^{n} sr_i^*(-)\right) \tag{11.2}$$

kde $\sum_{i=1}^{n} sr_i^*(+)$ predstavuje súčet poradových čísiel, ktoré sú priradené kladným hodnotám rozdielov d_i a $\sum_{i=1}^{n} sr_i^*(-)$ predstavuje súčet poradových čísiel (bez znamienka mínus), ktoré sú priradené záporným hodnotám rozdielov d_i.

Ak je nulová hypotéza pravdivá, teda, ak sa skutočná stredná hodnota základného súboru rovná predpokladanej strednej hodnote μ_0, a ak sú splnené predpoklady pre použitie Wilcoxonovho testu, potom pravdepodobnosť pozorovania kladného rozdielu $d_i = x_i - \mu_0$ danej veľkosti sa rovná pravdepodobnosti pozorovania záporného rozdielu rovnakej veľkosti. Pri opakovaných náhodných výberoch sa očakávaná hodnota $\sum_{i=1}^{n} sr_i^*(+)$ rovná očakávanej hodnote $\sum_{i=1}^{n} sr_i^*(-)$, pričom neočakávame veľký rozdiel v hodnotách týchto súm. Na druhej strane, dostatočne malá hodnota $\sum_{i=1}^{n} sr_i^*(+)$ alebo dostatočne malá hodnota $\sum_{i=1}^{n} sr_i^*(-)$ spôsobí zamietnutie nulovej hypotézy. Na rozhodnutie o zamietnutí alebo nezamietnutí nulovej hypotézy sa použije menšia hodnota z týchto dvoch súm.

Kritické hodnoty Wilcoxonovej testovacej štatistiky sú tabelované hodnoty, ktoré sú uvedené v štatistických tabuľkách. Nulovú hypotézu zamietneme na hladine významnosti α, ak je hodnota testovacej štatistiky T_W menšia alebo rovná ktirickej hodnote T pre dané n a vopred známu úroveň $\alpha/2$. Nulovú hypotézu teda zamietame, ak platí:

$$T_W \leq T_{n,\alpha/2} \tag{11.3}$$

Analogické je použitie Wilcoxonovho testu pre jednostranné alternatívne hypotézy, ktoré je možné definovať ako:

$$\begin{aligned} H_0 &: \ \mu \geq \mu_0 \\ H_1 &: \ \mu < \mu_0 \end{aligned}$$

alebo:

$$\begin{aligned} H_0 &: \ \mu \leq \mu_0 \\ H_1 &: \ \mu > \mu_0 \end{aligned}$$

Keď platí nulová hypotéza s ľavostrannou alternatívou, potom predpokladáme, že výberový súbor bude mať veľkú hodnotu $\sum_{i=1}^{n} sr_i^*(+)$. Preto, keď jednostranná alternatívna hypotéza uvádza, že skutočná stredná hodnota základného súboru je menšia ako predpokladaná hodnota, potom dostatočne malá hodnota $\sum_{i=1}^{n} sr_i^*(+)$ spôsobí zamietnutie nulovej hypotézy, a teda $\sum_{i=1}^{n} sr_i^*(+)$ je testovacia štatistika. Nulovú hypotézu zamietame, ak platí:

$$T_{W(+)} \leq T_{n,\alpha} \tag{11.4}$$

Podobne, ak platí nulová hypotéza s pravostrannou alternatívou, potom predpokladáme, že výberový súbor bude mať veľkú hodnotu $\sum_{i=1}^{n} sr_i^*(-)$. Preto, keď jednostranná alternatívna hypotéza uvádza, že skutočná stredná hodnota základného súboru je väčšia ako predpokladaná hodnota, potom dostatočne malá hodnota $\sum_{i=1}^{n} sr_i^*(-)$ spôsobí zamietnutie nulovej hypotézy, a teda $\sum_{i=1}^{n} sr_i^*(-)$ je testovacia štatistika. Nulovú hypotézu zamietame, ak platí:

$$T_{W(-)} \leq T_{n,\alpha} \tag{11.5}$$

Wilcoxonov test je možné aplikovať aj na súbory s párovými údajmi. Aj tu volíme túto metódu v prípadoch, kedy nie je vhodné použiť párový t test. Pri

párovom Wilcoxonovom teste vypočítavame rozdiely medzi pármi pozorovaní. Potom hypotézy formulujeme v tvare:

$$H_0: \ \mu_d = 0$$
$$H_1: \ \mu_d \neq 0$$

kde μ_d je priemerná hodnota rozdielov medzi pármi jednotlivých pozorovaní. Postup testovania je potom analogický s postupom uvedeným vyššie.

11.2 Mann-Whitneyho test

Mann-Whitneyho test je neparametrickým postupom, ktorý možno často použiť namiesto testu mediánu, resp. ako neparametrickú alternatívu k t testu pre nezávislé výberové súbory na porovnávanie priemerov dvoch základných súborov. Tento test je označovaný ak ako Mann–Whitney–Wilcoxonov test, keďže je založený na označovaní poradia jednotlivých hodnôt. Predpokladom pre použitie Mann-Whitneyho testu je, aby náhodné výberové súbory boli nezávislé, náhodne vybrané z ich príslušných základných súborov a sledovaná veličina by mala byť spojitá.

Ak sú predpoklady pre použitie Mann-Whitneyho testu splnené, potom je možné testovať nulovú hypotézu o zhode stredných hodnôt (alebo mediánov) dvoch výberových súborov voči obojstrannej, ľavostrannej alebo pravostrannej alternatíve. Obojstranný test bude uvažovať o hypotézach:

$$H_0: \ \mu_1 = \mu_2$$
$$H_1: \ \mu_1 \neq \mu_2$$

Na výpočet testovacej štatistiky priradíme všetkým hodnotám z oboch výberových súborov, zoradených podľa veľkosti od najmenšej hodnoty po najväčšiu, korešpondujúce poradové číslo pozorovania. Zachovávame si pritom informáciu o príslušnosti k danému súboru tak, že hodnoty sú evidované pod hlavičkou toho výberového súboru, ktorému patria. Ak sú hodnoty niektorých pozorovaní rovnaké, potom im všetkým prislúcha rovnaká poradová hodnota, ktorá je priemerom hodnôt poradí pre tieto hodnoty. Fiktívny príklad, takéhoto usporiadania je uvedený v tabuľke 11.2.

Testovacou štatistikou Mann-Whitneyho testu bude:

$$T_M = min\,(T_{M1}, T_{M2}) \tag{11.6}$$

t. j. menšia z dvoch vypočítaných charakteristík T_{M1} a T_{M2} porovnávaných výberových súborov, pričom T_{M1} je charakteristika prvého výberového súboru

Tabuľka 11.2: Priradenie poradových hodnôt meraniam dvoch nezávislých súborov.

Hodnoty súboru A	**Poradie**	**Hodnoty súboru B**	**Poradie**
10,3	1		
10,4	2		
10,6	3,5		
10,6	3,5		
		10,7	5
		10,8	6
10,9	7		
		11,3	8
11,8	9		
		12,0	10
⋮	⋮	⋮	⋮

a T_{M2} je charakteristika druhého výberového súboru, ktoré vypočítame podľa vzťahu Mann-Whitneyho testovacej štatistiky:

$$T_{M1} = S_1 - \frac{n_1 (n_1 + 1)}{2} \tag{11.7}$$

a

$$T_{M2} = S_2 - \frac{n_2 (n_2 + 1)}{2} \tag{11.8}$$

kde n_1 je počet pozorovaní prvého súboru, n_2 je počet pozorovaní druhého súboru, S_1 je súčet poradí priradených pozorovaniam prvého výberového súboru a S_2 je súčet poradí priradených pozorovaniam druhého výberového súboru.

Vypočítanú hodnotu Mann-Whitneyho testovacej štatistiky porovnávame s kritickou hodnotou, ktorú nájdeme v štatistických tabuľkách pre zvolenú úroveň α a veľkosti výberových súborov. Ak je vypočítaná hodnota testovacej štatistiky T_M menšia ako kritická hodnota na úrovni $\alpha/2$, potom nulovú hypotézu zamietame.

Pre jednostranné alternatívy to budú hypotézy nasledovné:

$$\begin{aligned} H_0 &: \ \mu_1 \geq \mu_2 \\ H_1 &: \ \mu_1 < \mu_2 \end{aligned}$$

alebo:

$$\begin{aligned} H_0 &: \ \mu_1 \leq \mu_2 \\ H_1 &: \ \mu_1 > \mu_2 \end{aligned}$$

Pre jednostranné alternatívy nulovú hypotézu zamietame, ak je vypočítaná hodnota testovacej štatistiky T_M menšia ako kritická hodnota na úrovni α.

Ak je niektorý z výberových súborov väčší ako 20, potom už nevieme použiť kritické hodnoty Mann-Whitneyho testu a testovaciu štatistiku nahrádzame štatistikou z:

$$z = \frac{T - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}} \tag{11.9}$$

ktorá má pri platnosti testovanej hypotézy normálne normované rozdelenie $N(0, 1)$. Testovanú hypotézu potom zamietame alebo nezamietame na základe kritických hodnôt z rozdelenia.

11.3 Kruskal-Wallisov test

Kruskal-Wallisov test je možné používať v prípadoch, kedy údaje dostupné na analýzu nespĺňajú predpoklady pre jednofaktorovú analýzu rozptylu pri porovnávaní viac ako dvoch výberových súborov.

Predpokladajme, že máme k výberových súborov pochádzajúcich zo základných súborov so spojitou veličinou a s rozsahom $n_1, n_2, \ldots, n_k$, u ktorých chceme overiť zhodu stredných hodnôt. Potom nulovú a alternatívnu hypotézu môžeme zapísať v tvare:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1 &: \ \text{nie všetky } \mu_i \text{ sú rovnaké} \end{aligned}$$

Overovanie platnosti nulovej hypotézy budeme vykonávať pomocou Kruskal-Wallisovho testu, ktorého princípom je, že hodnoty všetkých súborov zoradíme od najmenšej po najväčšiu a každej hodnote potom priradíme poradie od čísla 1 až po n, pričom platí, že $n = n_1 + n_2 + \ldots + n_k$. Poradie 1 priradíme najmenšej hodnote v celkovom súbore a poradie n priradíme najväčšej hodnote, t. j. poslednej hodnote v súbore všetkých hodnôt. V prípade, ak sa v celkovom súbore vyskytne viacero rovnakých hodnôt, potom sa všetkým týmto hodnotám priradí priemerná hodnota poradí, ktoré im prislúchajú. Následne vieme pre všetky výberové súbory spočítať sumu priradených poradí, čím dostávame k poradových súm. Príklad takéhoto usporiadania je uvedený v tabuľke 11.3.

Tabuľka 11.3: Priradenie poradových hodnôt meraniam troch nezávislých súborov.

Súbor A	**Poradie A**	**Súbor B**	**Poradie B**	**Súbor C**	**Poradie C**
1,42	7	1,23	3	1,62	10
1,13	1	1,63	11	1,41	6
1,59	9	1,93	15	1,74	12
1,80	13	1,27	5	1,58	8
1,24	4	1,15	2	1,87	14
	T_1=34		T_2=36		T_3=50

Testovacou štatistikou Kruskal-Wallisovho testu bude:

$$K = \frac{12}{n\,(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3\,(n+1) \tag{11.10}$$

kde n je počet výberových súborov, n_i je počet hodnôt v i-tom výberovom súbore, n je počet všetkých hodnôt vo výberových súboroch a T_i je súčet poradí priradených hodnotám i-teho výberového súboru.

Rozhodovanie o zamietnutí alebo nezamietnutí nulovej hypotézy vykonávame na základe porovnania testovacej štatistiky a kritickej hodnoty. Ak sú porovnávané tri výberové súbory a v každom výberovom súbore je najviac 5 hodnôt, potom kritickú hodnotu pre zvolné α nájdeme v štatistických tabuľkách pre Kruskal-Wallisov test. Nulovú hypotézu zamietneme ak platí:

$$K \geq K_\alpha \tag{11.11}$$

Avšak, ak je v jednom alebo vo viacerých výberových súboroch viac ako päť hodnôt, prípadne ak porovnávame viac výberových súborov ako tri, potom kritickú hodnotu používame z χ^2 rozdelenia s $k-1$ stupňami voľnosti. V takom prípade nulovú hypotézu zamietame ak:

$$K \geq \chi^2_{1-\alpha,k-1} \tag{11.12}$$

Ak existujú v údajoch viaceré rovnaké hodnoty, potom je možné testovaciu štatistiku upraviť tak, že ju vydelíme korekciou zohľadňujúcou opakujúce sa hodnoty. Korigovaná testovacia štatistika Kruskal-Wallisovho testu potom bude vypočítaná takto:

$$K_t = \frac{K}{1 - \dfrac{\sum T}{n^3 - n}} \tag{11.13}$$

kde $T = t^3 - t$ a t označuje počet zhodných pozorovaní v skupine zhodných hodnôt. V hodnotách uvedených v tabuľke 11.3 nie sú žiadne skupiny zhodných hodnôt, ale vo všeobecnosti sa môžu vyskytnúť. Ak ich bude viac, potom bude aj viac hodnôt T vo vzťahu (11.13). Efekt tejto korekcie je však zvyčajne zanedbateľný.

Rovnako ako u jednofaktorovej analýzy rozptylu, aj tu v prípade, že záverom bude zamietnutie nulovej hypotézy nevieme, ktoré zo súborov sa líšia, resp. nepochádzajú z toho istého rozdelenia. Potom je potrebné vykonať porovnávanie všetkých párov výberových súborov.

11.4 Friedmanov test

V predchádzajúcej kapitole sme poukázali na možnosť využitia neparametrickej alternatívy k jednofaktorovej analýze rozptylu. V niektorých štúdiách môžeme obdobným spôsobom potrebovať neparametrickú alternatívu k dvojfaktorovej analýze rozptylu. Pre takéto prípady sa najčastejšie používa Friedmanov test, ktorý je vhodný, ak sú údaje získavané aspoň na ordinálnej škále a možno ich zmysluplne usporiadať do dvojfaktorovej klasifikácie.

Predpokladajme, že máme štúdiu, v ktorej porovnávame hodnoty získanej veličiny ovplyvnenej dvoma faktormi. Mohlo by to byť napríklad skúmanie vplyvu liečby na znižovanie hodnoty celkového cholestorolu (niekoľko rôznych liekov), ktoré sú pacientom aplikované v rôznych blokoch (rôzne dávky, obdobia podávania lieku a pod.). Chceme rozhodnúť, či liečebné postupy majú rovnaký účinok. Ak nemajú rovnaký účinok, potom je potrebné určiť, ktoré liečebné postupy sa od seba líšia.

Nulovú a alternatívnu hypotézu by sme mohli definovať takto:

H_0 : všetky liečebné postupy majú rovnaký účinok
H_1 : aspoň jeden liečebný postup sa líši od ostatných liečebných postupov

Pomocou Friedmanovho testu môžeme určiť, či je rozumné predpokladať, že stĺpce poradí hodnôt (liečba) boli vybrané z rovnakého základného súboru. Ak je nulová hypotéza pravdivá, potom očakávame, že pozorované rozdelenie poradí v ktoromkoľvek stĺpci bude výsledkom náhodných faktorov, a preto očakávame, že hodnoty sa budú v každom stĺpci vyskytovať s približne rovnakou početnosťou. Na druhej strane, ak nulová hypotéza nie je pravdivá (liečebné postupy nie sú rovnako účinné), očakávame prevahu relatívne vysokých (alebo nízkych) poradí aspoň v jednom stĺpci. Táto podmienka by mala byť premietnutá do súčtu poradí hodnôt. Friedmanov test nám určí, či sú alebo nie sú

pozorované súčty poradí hodnôt také rozdielne, že nie je pravdepodobné, že by boli výsledkom náhody, za predpokladu, že platí nulová hypotéza.

Údajom v jednotlivých blokoch (riadkoch) priradíme odpovedajúce poradie od 1 do k, kde k je počet výberových súborov (stĺpcov) a potom poradia v každom stĺpci spočítame (T_i), napríklad tak ako v tabuľke 11.4.

Tabuľka 11.4: Priradenie poradových hodnôt podľa Friedmanovho testu.

Blok	**Liečba A**	**Liečba B**	**Liečba C**	**Poradie A**	**Poradie B**	**Poradie C**
1	15,38	14,97	14,51	3	2	1
2	15,13	14,29	14,86	3	1	2
3	14,95	15,14	14,88	2	3	1
4	14,58	14,05	14,72	2	1	3
5	14,74	14,89	14,65	2	3	1
6	15,63	14,56	14,93	3	1	2
7	14,42	15,04	14,71	1	3	2
				T_1=16	T_2=14	T_3=12

Testovacia štatistika Friedmanovho testu bude daná vzťahom:

$$\chi_r^2 = \frac{12}{nk\,(k+1)} \sum_{i=1}^{k} (T_i)^2 - 3n\,(k+1) \tag{11.14}$$

kde n je počet riadkov (blokov) a k je počet stĺpcov (liečba).

Kritické hodnoty potrebné pre vyhodnotenie Friedmanovho testu a rozhodnutie o zamietnutí alebo nezamietnutí nulovej hypotézy nájdeme v štatistických tabuľkách pre dané n a k. Nulovú hypotézu zamietame, ak je hodnota testovacej štatistiky väčšia alebo rovná ako kritická hodnota:

$$\chi_r^2 \geq \chi_{1-\alpha,k-1}^2 \tag{11.15}$$

Kritickú hodnotu χ^2 rozdelenia používame, ak výberové súbory majú rozsah stredne veľký a veľký. V prípade malých hodnôt n a k (dvadsať a menej), porovnávame testovaciu štatistiku s kritickou hodnotou Friedmanovho testu.

Literatúra

[1] Altman D.G.: *Practical statistics for medical research*, Chapman and Hall, 1991, ISBN 0-412-27630-5.

[2] Anděl J.: *Základy matematické statistiky*, Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Praha, 2005, ISBN 80-86732-40-1.

[3] Ataharul Islam M., Al-Shiha A.: *Foundations of Biostatistics*, Springer, 2020, ISBN 978-981-10-8626-7.

[4] Bekker A., Chen D.G., Ferreira J.T.: *Computational and Methodological Statistics and Biostatistics*, Springer, 2020, ISBN 978-3-030-42195-3.

[5] Berk R.A.: *Statistical Learning from a Regression Perspective*, Springer, 2020, ISBN 978-3-030-40188-7.

[6] Brase Ch.H., Brase C.P.: *Understandable Statistics: Concepts and Methods*, Houghton Mifflin Company, Boston, 2009, ISBN 978-0-618-94989-2.

[7] Cleophas T.J., Zwinderman A.H.: *Statistics Applied to Clinical Studies*, Springer, 2012, ISBN 978-94-007-2862-2.

[8] Cleophas T.J., Zwinderman A.H.: *Regression Analysis in Medical Research for Starters and 2nd Levelers*, Springer, 2021, ISBN 978-3-030-61393-8.

[9] Culliford D.: *Applied Statistical Considerations for Clinical Researchers*, Blackwell Science, 2010, ISBN 0-632-04763-1.

[10] Daly L.E., Bourke G.J.: *Interpretation and Uses of Medical Statistics*, John Wiley & Sons, Inc., 2014, ISBN 978-1-118-30279-8.

[11] Daniel W.W., Cross Ch.L.: *Biostatistics: A Foundation for Analysis in the Health Sciences*, John Wiley & Sons, Inc., 2014, ISBN 978-1-118-30279-8.

[12] Elston R.C., Johnson W.D.: *Basic Biostatistics for Geneticists and Epidemiologists: A Practical Approach*, John Wiley & Sons, Inc., 2008, ISBN 978-0-470-02489-8.

[13] Feinstein A.R.: *Principles of Medical Statistics*, Chapman & Hall/CRC, 2002, ISBN 1-58488-216-6.

[14] Ghosh I., Balakrishnan N., Ng H.K.T.: *Advances in Statistics: Theory and Applications*, Springer, 2021, ISBN 978-3-030-62899-4.

[15] Herzog M.H., Francis G., Clarke A.: *Understanding Statistics and Experimental Design, How to Not Lie with Statistics*, Springer, 2019, ISBN 978-3-030-03498-6.

[16] James G., Witten D., Hastie T., Tibshiran R.: *An Introduction to Statistical Learning: with Applications in R*, Springer, 2017, ISBN 978-1-4614-7137-0.

[17] Jurečková M., Molnárová I.: *Štatistika s Excelom*, Akadémia ozbrojených síl gen. M.R. Štefánika v Liptovskom Mikuláši, 2005, ISBN 80-8040-257-4.

[18] Kalina M., Bacigál T., Schiesslová A.: *Základy pravdepodobnosti a matematickej štatistiky*, Slovenská technická univerzita v Bratislave, 2010, ISBN 978-80-227-3273-4.

[19] Kaltenbach H.M.: *Statistical Design and Analysis of Biological Experiments*, Springer, 2021, ISBN 978-3-030-69640-5.

[20] Kauermann G., Kuchenhoff H., Heumann Ch.: *Statistical Foundations, Reasoning and Inference: For Science and Data Science*, Springer, 2021, ISBN 978-3-030-69826-3.

[21] Lee H.: *Foundations of Applied Statistical Methods*, Springer, 2014, ISBN 978-3-319-02401-1.

[22] Linebach J.A., Tesch B.P., Kovacsiss L.M.: *Nonparametric Statistics for Applied Research*, Springer, 2014, ISBN 978-1-4614-9040-1.

[23] Logan M.: *Biostatistical Design and Analysis Using R: A Practical Guide*, John Wiley & Sons, Inc, 2010, ISBN 978-1-4443-3524-8.

[24] MacFarland T.W., Yates J.M.: *Using R for Biostatistics*, Springer, 2021, ISBN 978-3-030-62403-3.

[25] Machin D., Campbell M.J., Walters S.J.: *Medical Statistics: Fourth Edition, A Textbook for the Health Sciences*, John Wiley & Sons Ltd, 2007, ISBN 978-0-470-02519-2.

[26] Markechová D., Tirpáková A., Stehlíková B.: *Základy štatistiky pre pedagógov*, School Science Review, December 2001, 83(303), 23-34.

[27] Millar N.: *Biology statistics made simple using Excel*, Univerzita Konštantína Filozofa v Nitre, 2011, ISBN 978-80-8094-899-3.

[28] Norušis M.J.: *Basics of Biostatistics*, Creative Commons, 2013, ISBN 978-80-8094-899-3.

[29] Pecáková I.: *Statistika v terénních průzkumech*, Professional Publishing PBtisk Příbram, 2011, ISBN 978-80-7431-039-3.

[30] Procházka B.: *Stručná biostatistika pro lékaře*, Univerzita Karlova v Praze, Karolinum, 2015, ISBN 978-80-246-2783-0.

[31] Quirk T.J., Cummings S.H.: *Excel 2019 for Health Services Management Statistics: A Guide to Solving Practical Problems*, Springer, 2021, ISBN 978-3-030-57827-5.

[32] Quirk T.J., Quirk M.H., Horton H.F.: *Excel 2019 for Biological and Life Sciences Statistics: A Guide to Solving Practical Problems*, Springer, 2020, ISBN 978-3-030-39280-2.

[33] Quirk T.J., Quirk M.H., Horton H.F.: *Excel 2019 for Physical Sciences Statistics: A Guide to Solving Practical Problems*, Springer, 2021, ISBN 978-3-030-63237-3.

[34] Radermacher W.J.: *Official Statistics 4.0: Verified Facts for People in the 21st Century*, Springer, 2020, ISBN 978-0-538-73349-6.

[35] Rosner B.: *Fundamentals of Biostatistics: Seventh Edition*, Brooks/Cole, Cengage Learning, 2010, ISBN 978-3-030-31491-0.

[36] Saiz A.Z., González C.Q., Gil L.H., Ruiz D.M.: *An Introduction to Data Analysis in R: Hands-on Coding, Data Mining, Visualization and Statistics from Scratch*, Springer, 2020, ISBN 978-3-030-48996-0.

[37] Schiefer H., Schiefer F.: *Statistics for Engineers: An Introduction with Examples from Practice*, Springer, 2021, ISBN 978-3-658-32396-7.

[38] Simera I., Moher D., Hoey J., Schulz K.F., Altman D.G.: *A catalogue of reporting guidelines for health research*, European Journal of Clinical Investigation Vol 40, 2009, 35-53.

[39] van Belle G., Fisher L.D., Heagerty P.J., Lumley T.: *Biostatistics: A Methodology for the Health Sciences*, Routledge, 2009, ISBN 978–0–203–88786–8.

[40] Van Blerkom M.L.: *Measurement and Statistics for Teachers*, John Wiley & Sons, Inc., 2004, ISBN 0-471-03185-2.

[41] Zhao Y., Chen D.G.: *Modern Statistical Methods for Health Research*, Springer, 2021, ISBN 978-3-030-72436-8.

[42] Zvára K.: *Biostatistika*, Univerzita Karlova v Praze, Karolinum, Praha, 2008, ISBN 978-80-246-0739-9.

[43] Zvára K., Štěpán J.: *Pravdepodobnost a matematická statistika*, matfyzpress, Praha, 2006, ISBN 80-85863-93-6.

Register

Základy (bio)štatistiky pre medikov
Vysokoškolská učebnica
Autor: doc. Ing. Jaroslav Majerník, PhD.

Vydavateľ: Univerzita Pavla Jozefa Šafárika v Košiciach
Vydavateľstvo ŠafárikPress

Rok vydania: 2021
Náklad: 200 ks
Rozsah strán: 218
Rozsah: 10,9 AH
Vydanie: prvé
Tlač: EQUILIBRIA, s.r.o.

ISBN 978-80-574-0066-0