EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Biostatistika

Tomáš Pavlík, Ladislav Dušek

Leden 2012

Příprava a vydání této publikace byly podporovány projektem ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“ a státním rozpočtem České republiky.

Předmluva

Publikace *Biostatistika* je součástí série učebních textů vzniklých v rámci řešení projektu ESF č. CZ.1.07/2.2.00/07.0318 „VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE“, který je zaměřen na zkvalitnění a rozšíření výuky klíčových předmětů studijního oboru Matematická biologie akreditovaného pod Přírodovědeckou fakultou Masarykovy univerzity. Cílovými čtenáři jsou tak studenti matematické biologie, kterým chceme tímto učebním textem podat srozumitelný přehled základů biostatistiky v kontextu hodnocení biologických a klinických dat. Vzhledem k rozsahu skript však bylo nutné vybrat pouze klíčové metody, bez nichž si nelze zpracování dat vůbec přestavit, a na řadu používaných metod se tak v tomto textu vůbec nedostalo. Skripta proto neslouží jako náhrada přednášek, slouží pouze jako jejich doplnění.

Tato publikace nemá v žádném případě ambici nahradit jakýkoliv stávající učební text o biostatistice či statistických metodách. Pro studenty i další čtenáře je totiž vždy lepší čerpat z více zdrojů a udělat si o dané problematice komplexní obrázek. V biostatistice toto platí dvojnásob, neboť co autor, to jiný úhel pohledu a jiná zkušenost s praktickými aplikacemi.

Naším cílem bylo podat biostatistické metody korektně nejen z teoretického, ale i z praktického hlediska. Proto je kromě teoretického zázemí jednotlivých statistických metod text zaměřen také na praktickou stránku hodnocení dat, a to zejména na nástrahy, které mohou kohokoliv při zpracování konkrétních datových souborů potkat, ať už při výpočtech nebo interpretaci výsledků. Doufáme tedy, že publikace *Biostatistika* poslouží studentům nejen jako příprava ke složení zkoušky, ale také jako referenční text pro statistické zpracování dat v rámci bakalářských a diplomových prací.

V Brně 24. 12. 2011

Tomáš Pavlík

Ladislav Dušek

ISBN 978-80-7204-782-6

1 Úvod do biostatistiky

Biostatistika je vědní obor na pomezí matematické statistiky a věd o živých systémech. Jednoduše ji lze charakterizovat jako aplikaci a vývoj statistických metod pro řešení biologických a klinických problémů. Jinak řečeno, naší snahou je získání užitečné informace z pozorovaných dat. Tou může být prostý popis stavu sledovaného souboru, identifikace faktorů ovlivňujících jeho chování, nebo rozhodnutí o nějaké jeho neznámé charakteristice. Získaná informace pak nemusí mít vůbec žádný důsledek a může sloužit pouze pro informaci hodnotitele, nebo naopak, může vést k výrazné změně lidské činnosti, např. ke změně metodických a léčebných postupů nebo klinických doporučení. Příkladem může být hodnocení účinnosti a bezpečnosti léčivých přípravků v klinických studiích [33].

Biostatistika má zásadní postavení v dnešní vědě a výzkumu, kdy řada vědeckých časopisů nepřijme k publikaci experimentální výsledky bez jejich statistického zpracování s použitím metod, které lze považovat za standardní pro danou vědeckou oblast. V hodnocení biomedicínských dat je tento fenomén patrný zřejmě nejvíce, význam aplikace statistických metod v této oblasti lze dokumentovat tím, že jej prestižní americký časopis *The New England Journal of Medicine* zařadil v roce 2000 mezi 11 nejvýznamnějších událostí, které ovlivnily medicínu 20. století [18].

Je však použití statistických metod pro získávání informací nezbytné? Není možné se bez nich obejít? Odpověď je jednoznačná, není možné se bez nich obejít. Důvodem je tendence lidského uvažování dělat jasné závěry z nejasných podkladů, přičemž tento fenomén byl pozorován již u osmiměsíčních dětí [35]. Jinými slovy, lidský mozek má schopnost dělat ukvapené závěry, z čehož plyne, že v mnoha oblastech lidské činnosti, biologii a medicínu nevyjímaje, nelze při hodnocení výsledků experimentů či pokusů spoléhat výhradně na selský rozum. Vědci a odborníci tak využívají biostatistiku, respektive statistiku a její metodické zázemí k tomu, aby se při hodnocení experimentů na základě limitovaných dat a údajů nedopouštěli nesprávných interpretací a závěrů [17].

Jak již bylo naznačeno, biostatistika primárně vychází ze statistiky, jejich hranice však nejsou ostré. Biostatistika je navíc často zaměňována s analýzou dat, se kterou může mít společný cíl a někdy i metodiku. Rozdíly mezi těmito třemi oblastmi lze shrnout následovně:

- ***Statistika*** je primárně zaměřena na teoretické aspekty, respektive na vývoj metod a algoritmů. Nicméně i vývoj ve statistice byl a je motivován reálnými problémy, cílem je však zejména jejich adekvátní teoretické řešení.

- ***Biostatistika*** představuje propojení znalosti statistických metod a dané problematiky v řešení biologických a klinických úloh. Biostatistika také zahrnuje metodický vývoj, nicméně vždy je primárně orientována na řešení konkrétního biologického a medicínského problému, je tedy zaměřena převážně prakticky.

- ***Analýza dat*** je velmi obecná oblast, která nemusí být nutně spojována se statistickými metodami a která prostupuje různými odvětvími. Zahrnuje komplexní postupy pro získávání informací z dat, včetně jejich zpracování a přípravy, tedy čištění dat, analýzu odlehlých pozorování a kódování dat. Metody analýzy dat mohou i nemusí mít matematický základ, často se např. setkáváme v analýze dat s metodami a algoritmy dolování dat.

1.1 Cíl biostatistiky a základní pojmy

Hlavním cílem biostatistiky je získání informace o tzv. ***cílové populaci*** (***základním souboru***), jejíž prvky jsou nejčastěji dány vymezením společných vlastností. Příkladem může být populace pacientek s karcinomem prsu, populace mužů starších 60 let nebo populace motýlů druhu *Papilio machaon*. Na druhou stranu, prvky cílové populace mohou být dány i výčtem, můžeme např. studovat populaci zdravotnických zařízení v ČR, populaci studentů oboru Matematická biologie nebo populaci zaměstnanců Institutu biostatistiky a analýz Masarykovy univerzity. Ve většině případů je však zjišťování sledovaných charakteristik u všech subjektů cílové populace nereálné a my jsme v našem bádání omezeni pouze na část cílové populace, tzv. ***výběr z cílové populace*** (***experimentální vzorek***) [38].

Experimentální vzorek představuje podsoubor cílové populace zahrnutý v naší studii nebo experimentu. Jinak řečeno, je to skupina subjektů, kterou máme k dispozici a která představuje ***pozorování*** cílové populace. Sledované vlastnosti experimentálního vzorku pro hodnocení převedeme na číselné vyjádření (***data***). Ta jsou dále předmětem našeho zájmu, nicméně to, jak budeme dále postupovat při jejich hodnocení, do značné míry závisí na účelu studie nebo experimentu. Obecně se dá říci, že předpokládáme určité pravděpodobnostní chování (***model***) studované cílové populace a tím i experimentálního vzorku. Konkrétní problém následně vyjádříme v našem modelu jako hypotézu, jejíž platnost vyhodnotíme na základě vybraného modelu a pozorovaných dat.

Charakteristika sledovaná u cílové populace se nejčastěji označuje jako ***znak***. Jinak řečeno, znaky odpovídají sledovaným vlastnostem subjektů cílové populace. Dle povahy popisu jednotlivých variant daného znaku dělíme znaky na ***kvalitativní*** a ***kvantitativní***. Můžeme-li jednotlivé varianty vyjádřit slovně, mluvíme o znaku kvalitativním. Naopak, můžeme-li varianty vyjádřit číslem, mluvíme o znaku kvantitativním. Slovní vyjádření jsou pro jakékoliv matematické zpracování nevhodná, proto i varianty kvalitativního znaku převádíme na čísla (na rozdíl od kvantitativních znaků jsou však tato čísla pouze pomocná a nemají většinou žádnou interpretaci). Číselnou reprezentaci daného znaku, která je nezbytná pro statistické zpracování, pak nazýváme ***veličinou***. Vzhledem k tomu, že ve statistice a tudíž i v biostatistice reprezentujeme skutečnost matematickým modelem, ve kterém hraje roli náhoda, představuje měření daného znaku u jednoho prvku experimentálního vzorku výsledek náhodného pokusu. V tomto případě nazýváme číselnou reprezentaci daného znaku ***náhodnou veličinou***. Konkrétní číselný výsledek náhodného pokusu, tedy pozorovanou hodnotu náhodné veličiny u *i*-tého prvku experimentálního vzorku, pak označujeme jako ***realizaci náhodné veličiny***. Náhodná veličina má v biostatistice klíčové postavení, neboť je základním konceptem všech biostatistických úloh, v detailu se náhodné veličině věnuje kapitola 3.

1.2 Typy biostatistických úloh

Existuje několik typů biostatistických úloh, čtyři z nich jsou však základní:

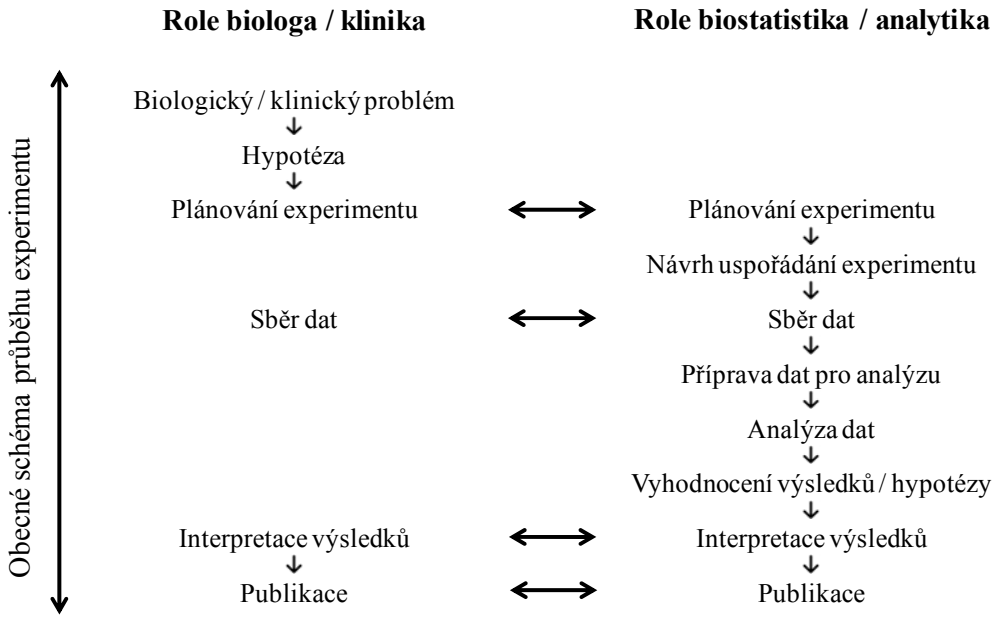
- ***Popis cílové populace*** – Popisem myslíme sumarizaci sledovaných znaků (veličin) cílové populace. Jde o grafické a početní techniky vedoucí k vyjádření informace z dat v srozumitelné, korektní a rozsahem akceptovatelné podobě. Přesněji řečeno, často nepřehledné záznamy o jednotlivých subjektech hodnocení (primární data) jsou nahrazeny vypočítanými hodnotami, které nazýváme ***sumární statistiky***. Ty představují ***odhady parametrů*** modelu cílové populace. Popis musí pravdivě odpovídat primárním datům bez ztráty podstatné informace. Přínos popisné analýzy je ale podmíněn adekvátně zvolenou sumarizací, špatná volba sumární statistiky může znehodnotit celou práci.

- ***Srovnání skupin*** – Na rozdíl od popisné statistiky, u srovnávacích postupů většinou vycházíme z nějaké hypotézy nebo předpokladu o sledovaném znaku (veličině), který měřením a následným testováním ověřujeme. Jinak řečeno, testování hypotéz o sledovaných veličinách se zabývá rozhodováním o platnosti stanovených hypotéz na základě pozorovaných dat. Platnost hypotéz ověřujeme pomocí ***statistického testu*** – rozhodovacího pravidla, které každému náhodnému výběru přiřadí právě jedno ze dvou možných rozhodnutí – hypotézu nezamítáme nebo hypotézu zamítáme.

- ***Regresní analýza*** – Velmi často zaznamenáváme u sledovaných subjektů více znaků zároveň s tím, že nás zajímá, jestli mezi nimi existuje nějaký vztah. Regresní metody slouží k modelování a kvantifikaci tohoto vztahu. Hlavním cílem je vysvětlit pozorovanou variabilitu ve sledovaných znacích a odhalení případné společné tendence ve výskytu jednotlivých hodnot těchto znaků. Klíčovou roli v regresní analýze hrají stochastické modely, jejichž nejjednodušším příkladem je korelační analýza.

- ***Predikce a klasifikace*** – Cílem prediktivního modelování a klasifikačních algoritmů je předpovědět neznámé hodnoty, které jsou v případě prediktivního modelování většinou kvantitativního charakteru, zatímco v případě klasifikace jsou to většinou kategorie. Hlavní pointa je stejná jako v případě regresního modelování, tedy primárně je třeba modelovat pozorovanou variabilitu v datovém souboru. Výsledkem je ale vytvoření rozhodovacího pravidla, které lze následně po zadání vstupních hodnot použít pro předpověď. Pro úplnost je nutno dodat, že problematika klasifikace nemusí být vůbec spojena s použitím statistických metod.

Biostatistika se však netýká pouze závěrečné fáze zpracování nebo modelování dat, obecně lze říci, že se biostatistik nebo analytik dat účastní téměř všech fází experimentu, ať už na nich pracuje sám nebo ve spolupráci s biologem či klinikem [2]. Roli biostatistiky v průběhu celého experimentu blíže sumarizuje obrázek 1.1.



Obr. 1.1 Využití biostatistiky v průběhu biologického/klinického experimentu.

1.3 Příklady biostatistických úloh

Konkrétní příklady použití biostatistiky v medicíně a biologii jsou následující:

- ***Modelování demografické struktury obyvatelstva***. Sledování vývoje počtu obyvatel České republiky a jejich věkové struktury je zásadní pro jakékoliv plánování v rámci státu. Při modelování se většinou pracuje s různými scénáři, které představují různé možnosti vývoje s ohledem na migraci, porodnost, stav zdravotní péče, apod.

- ***Identifikace vlivu genetických a environmentálních rizikových faktorů na vznik různých onemocnění – astma, diabetes, hypertenze***. Již celá staletí lidstvo zajímá, které charakteristiky člověka a jeho chování mohou ovlivnit vznik a vývoj závažných onemocnění, a ani v 21. století není tato otázka uspokojivě zodpovězena. Vedle samotného vlivu sledovaných faktorů na výskyt onemocnění můžeme zároveň studovat, jak spolu jednotlivé vysvětlující faktory interagují a vzájemně se ovlivňují.

- ***Hodnocení úspěšnosti programů prevence nádorových onemocnění***. V rámci hodnocení úspěšnosti včasné detekce nádorových onemocnění lze studovat tzv. indikátory kvality, což jsou měřitelné výstupy kvality, a to na celonárodní úrovni, na regionální úrovni i na úrovni jednotlivých zdravotnických zařízení. Tyto indikátory lze zpětně použít pro zlepšování výkonnosti zdravotního systému a potažmo i péče o těžce nemocné.

- ***Identifikace podskupin pacientů s leukémií na základě genetických dat***. Hemato-onkologické diagnózy představují heterogenní skupinu onemocnění s velmi různou prognózou pro jednotlivé pacienty. Genetická data mohou být vodítkem k lepší klasifikaci a diagnostice těchto onemocnění, která umožní jejich rychlejší a účinnější léčbu.

- ***Prostorové modelování koncentrací škodlivých látek***. Produkce odpadů v jakékoliv formě je závažným problémem současného světa. S tím souvisí i fakt, že se do prostředí dostává čím dál více nebezpečných látek. Monitoring koncentrací těchto látek a případné prostorové modelování jejich šíření v prostředí (např. půdě, vodě, vzduchu) je velmi prospěšné, neboť tyto látky mohou zásadním způsobem ovlivňovat vývoj zdravotního stavu lidí i zvířat.

- ***Prediktivní modelování potenciálního rozšíření biologických společenstev***. Poznání zákonitostí, které ovlivňují šíření rostlinných druhů v přírodě, nám může v budoucnu sloužit k rozpoznávání a modelování změn v prostředí způsobených činností člověka a následnému hodnocení ekologického stavu daného prostředí.

- ***Definice indikačních taxonů a jejich vztah k parametrům prostředí***. Tzv. indikační druhy rostlin nebo živočichů hrají významnou roli v hodnocení kvality životního prostředí, neboť jsou zvláště citlivé na změnu určité charakteristiky prostředí, ve kterém žijí. Pro využití indikačních druhů směrem k monitoringu kvality životního prostředí je však nejprve nutné pro jednotlivé typy prostředí tyto indikační druhy korektně identifikovat a validovat, což nelze bez použití biostatistických metod.

- ***Analýza vztahu dávka-odpověď mezi koncentrací toxické látky, např. pesticidu a reakcí biologických receptorů***. Hodnocení závislosti mezi dávkou určité chemické látky a odpovědí daného biologického receptoru zasahuje prakticky všechny oblasti biologického výzkumu a představuje téměř univerzální přírodovědný problém. Je logické, že nejintenzivněji je tato problematika rozvíjena v rámci oboru toxikologie, kde lze z testů toxicity a karcinogeneze odvozovat parametry biologické účinnosti látek.

Ovšem i metody biostatistiky mohou být použity nekorektně a vést k nesprávným závěrům, příkladem jsou výsledky německé studie Hemkense a kol., kteří se na základě dat největší německé zdravotní pojišťovny věnovali riziku vzniku nádorových onemocnění a celkové mortalitě pacientů léčených humánním inzulinem a inzulinovými analogy [11]. V této studii bylo publikováno vyšší riziko vzniku zhoubného nádoru při užívání inzulinu glargin při srovnání s adekvátní dávkou humánního inzulinu. Tato studie však obsahovala ze statistického hlediska řadu nekorektních kroků a nesprávných postupů, které naprosto znemožnily jakoukoliv interpretaci získaných výsledků. Prvním problémem bylo, že se jednalo o tzv. ***observační studii***, což je epidemiologická studie bez náhodného přiřazování sledovaných subjektů do srovnávaných skupin. Problém byl právě v nenáhodném přiřazování subjektů, které nemůže zaručit stejné zastoupení jejich charakteristik v jednotlivých sledovaných skupinách. Výsledky studie tak byly ovlivněny různým zastoupením pacientů s diabetem 1. a 2. typu a pacientů s různou hmotností v jednotlivých skupinách. Dalším špatným krokem byla použitá adjustace na dávkování, která neodpovídá statistickým standardům. Autoři studie adjustovali výsledky na průměrnou dávku zjištěnou v průběhu sledování. Z hlediska statistiky je však nepřijatelné adjustovat model na informaci, která je získána až v průběhu sledování, neboť při tomto kroku můžeme zaměňovat příčinu s důsledkem.

1.4 Klíčové pojmy biostatistiky

Každá oblast biostatistiky má svá specifika a aspekty, kterým je nutné se při hodnocení dat věnovat. Některé z těchto aspektů jsou však společné pro všechny oblasti biostatistiky, a vlastně i analýzy dat obecně, a jsou naprosto klíčové pro korektní hodnocení a interpretaci výsledků.

1.4.1 Zkreslení výsledků

Jak již bylo řečeno, hlavním cílem biostatistiky je získat užitečnou informaci na základě dostupných dat, která pravdivě popisuje skutečný stav cílové populace. Snažíme se tak vyhnout ***zkreslení výsledků*** (*biased results*), které by tento pravdivý popis jakkoliv změnilo. Jinak řečeno, snažíme se vyhnout zkreslení hodnot sledované náhodné veličiny veličinami, které nejsou cílem studie. Příkladem zavádějícího srovnání může být srovnání pětiletého celkového přežití dosaženého pro konkrétní onkologickou diagnózu v jednotlivých krajích ČR. Ty se totiž mohou natolik lišit ve věkové struktuře onkologických pacientů, že výpočet celkového přežití bez zohlednění vlivu této věkové struktury může vést k zavádějícím závěrům.

Použití statistických metod nikdy nedává stoprocentní jistotu, že nějaká zjištěná skutečnost opravdu platí, neboť musíme počítat s vlivem náhody a tedy i pravděpodobností chybného úsudku. Tento fakt nelze ovlivnit, nicméně naší úlohou je použít metody pro odstranění vlivů, které by zkreslily výsledky a nebyly přitom náhodné. Příkladem může být hodnocení vlivu dvou typů léčby na mortalitu sledovaného onemocnění, která je však zásadně ovlivněna také stadiem neboli tíží onemocnění. Ve chvíli, kdy by oba soubory s různým typem léčby neměly stejné zastoupení stadií, nemohli bychom korektně rozhodnout, jestli pozorované rozdíly v mortalitě jsou dány rozdíly v léčbě nebo různým stadiem onemocnění u sledovaných pacientů.

Jiným příkladem, kdy může dojít ke zkreslení výsledků, jsou již dříve zmiňované epidemiologické studie [7]. Uvažujme studii, kdy sledujeme vztah nošení zapalovače a výskytu rakoviny plic. Bez znalosti vlivu kouření a toho, že zapalovače nosí zejména kuřáci,

bychom se mohli mylně domnívat, že nošení zapalovače způsobuje rakovinu plic. Nicméně je jasné, že pravým důvodem je kouření a nošení zapalovače s výskytem rakoviny plic pouze koreluje. Nošení zapalovače z tohoto příkladu označujeme jako tzv. ***zavádějící veličinu*** neboli ***zavádějící faktor*** (*confounding variable, confounder*), který představuje nepravou příčinu sledovaného výsledku (výskyt rakoviny plic), kdy korelace se sledovaným výsledkem je dána vztahem k pravé příčině sledovaného výsledku (kouření).

1.4.2 Reprezentativnost

Reprezentativnost experimentálního vzorku, respektive fakt, že vybraný experimentální vzorek musí svými charakteristikami odpovídat cílové populaci, je dalším klíčovým aspektem biostatistiky, který podmiňuje možnost zobecnění výsledků na celou cílovou populaci. Nebude-li totiž zkoumaný vzorek reprezentativní vzhledem k cílové populaci, zobecnění výsledků získaných statistickým zpracování dat vzorku na cílovou populaci může být nesprávné a jejich interpretace nekorektní, zkreslená.

Charakter cílové populace je důležité si uvědomit i při interpretaci cizích, respektive publikovaných výsledků. A to právě z důvodu, abychom zobecňovali výsledky pouze na populaci, na které těchto výsledků bylo dosaženo. Je-li například sledovaná léčba účinná z hlediska snížení rizika celkové mortality u kardiologických pacientů s normální funkcí ledvin, nelze účinnost této léčby jednoduše předpokládat i u skupiny pacientů se stejným kardiologickým problémem a dysfunkcí ledvin. Na druhou stranu, neúčinnost léčby na jednom souboru jedinců nebo vzorků neznamená neúčinnost také u souboru, který v dané studii nebyl uvažován.

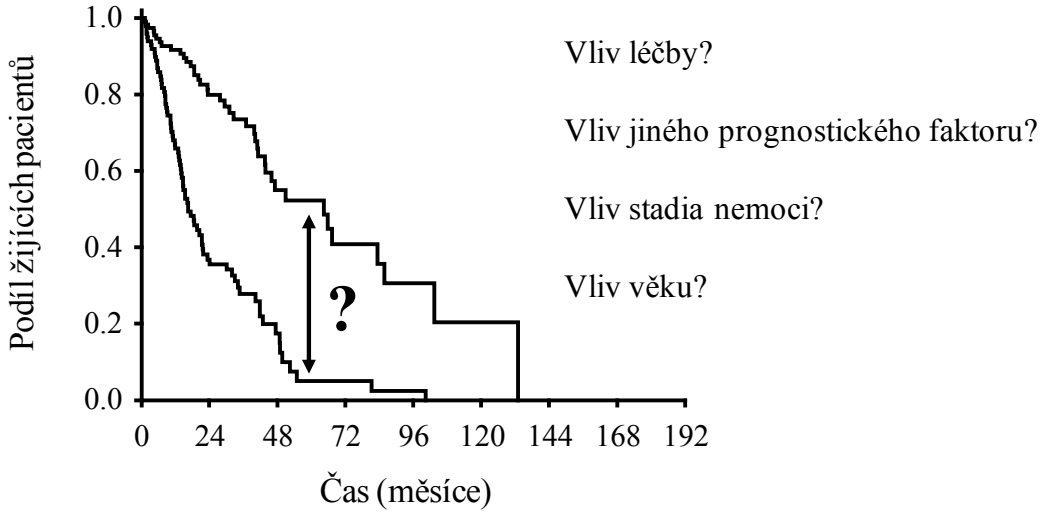
Klasickým příkladem reprezentativnosti je odhad střední výšky české dospělé populace. Aby byl odhad kvalitní, je třeba oslovit a změřit reprezentativní vzorek české populace nad 18 let, což znamená, že by naším vzorkem (myšleno ve významu pouze) jistě neměla být mužská basketbalová reprezentace (tito jedinci budou zřejmě výsledný odhad nadhodnocovat), ženská reprezentace ve sportovní gymnastice (tito jedinci budou odhad naopak spíše podhodnocovat), nebo dospělí návštěvníci akvaparku. V posledním případě je problém zejména v tom, že dospělí návštěvníci akvaparku s velkou pravděpodobností nebudou svojí věkovou strukturou odpovídat celé české populaci, což je vzhledem k reprezentativnosti problém, neboť je známo, že se v České republice výška postavy v průběhu času zvyšuje.

1.4.3 Srovnatelnost

V úlohách, kde je naším cílem srovnání dvou a více skupin je nutné zajistit jejich vzájemnou ***srovnatelnost***, neboť korektní výsledky lze získat pouze při srovnávání srovnatelného (tedy srovnávání jablek s jablky a ne jablek s hruškami). V nejpřísněji kontrolovaném medicínském výzkumu, klinických studiích, je srovnatelnost do značné míry (nikdy nelze říci, že stoprocentně) zajištěna pomocí tzv. ***randomizace***. U studií bez randomizace je nutné se tématu srovnatelnosti skupin věnovat, protože i malý nepoměr mezi srovnávanými skupinami, zvláště týká-li se veličiny spojené s výsledkem experimentu (např. věku), může vést ke zkreslení výsledků (tato problematika souvisí se zavádějícím faktorem – viz výše). Ve chvíli, kdy nemáme k dispozici randomizaci a víme, že naše skupiny nejsou v nějakém ohledu plně srovnatelné, je třeba použít metody adjustace, případně rozdělit soubor na podskupiny a srovnávat výsledky experimentu v rámci podskupin.

Význam srovnatelnosti sledovaných skupin lze demonstrovat na obrázku 1.2, který zobrazuje vývoj pravděpodobnosti přežití dvou skupin pacientů s nádorem trávicího traktu. Řekněme, že skupiny jsou dány typem léčby, kterou pacienti podstoupili. Ve chvíli, kdy obě

skupiny budou srovnatelné z hlediska všech charakteristik souvisejících s délkou přežití, můžeme říci, že pozorovaný rozdíl v přežití je zřejmě dán rozdílnou léčbou obou skupin. Pokud však obě skupiny srovnatelné nejsou, nelze jednoduše říci, že je pozorovaný rozdíl dán rozdílnou léčbou, neboť může být současně ovlivněn rozdílným zastoupením pohlaví, věkových kategorií nebo různou tíží onemocnění pacientů ve srovnávaných skupinách.

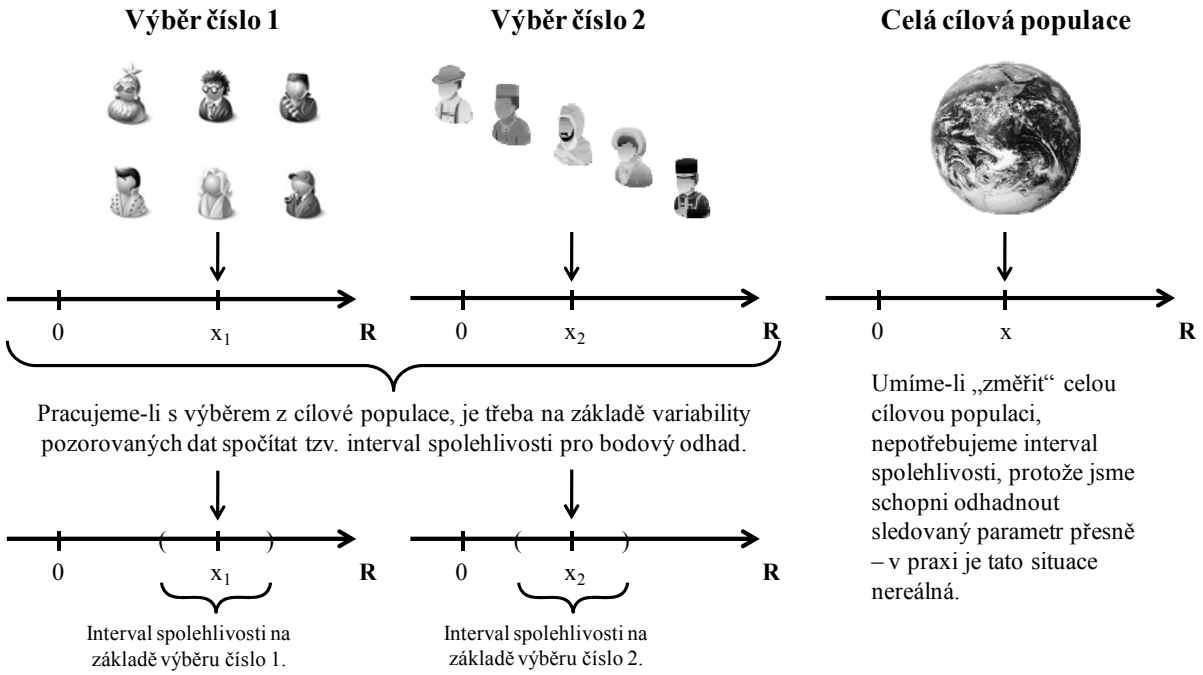


Obr. 1.2 Rozdíl ve vývoji pravděpodobnosti přežití dvou skupin pacientů s nádorem v čase.

1.4.4 Spolehlivost

Ve většině studií nás zajímá kvantifikace sledovaného znaku, respektive náhodné veličiny, ve formě jednoho čísla, tzv. ***bodového odhadu***. Bodový odhad je však sám o sobě nedostatečný, neboť nepostihuje variabilitu pozorovaných dat. Příkladem mohou být dva odhady průměrné výšky nějaké populace, jeden naměřený na 10 jedincích, druhý naměřený na 1000 jedincích. Je zřejmé, že druhý odhad bude přesnější než ten první, jinými slovy, bude méně zatížen variabilitou sledované veličiny, tedy výšky. Bodový odhad je tedy nutné doplnit měřítkem jeho kvality, respektive ***spolehlivosti***. Tím je většinou tzv. ***intervalový odhad***, který představuje rozsah hodnot (interval), který se zvolenou spolehlivostí (pravděpodobností) pokrývá neznámý parametr, který se snažíme odhadnout bodovým odhadem.

Umíme-li kvantifikovat sledovaný znak na celé cílové populaci, nepotřebujeme interval spolehlivosti, protože jsme schopni odhadnout neznámou veličinu úplně přesně. V praxi je však tato situace nereálná a doplňování intervalu spolehlivosti k bodovým odhadům by mělo být jak v případě medicínských, tak biologických analýz standardem. Význam intervalu spolehlivosti jako měřítka přesnosti bodového odhadu ilustruje obrázek 1.3, detailně se problematice intervalových odhadů věnuje kapitola 5.

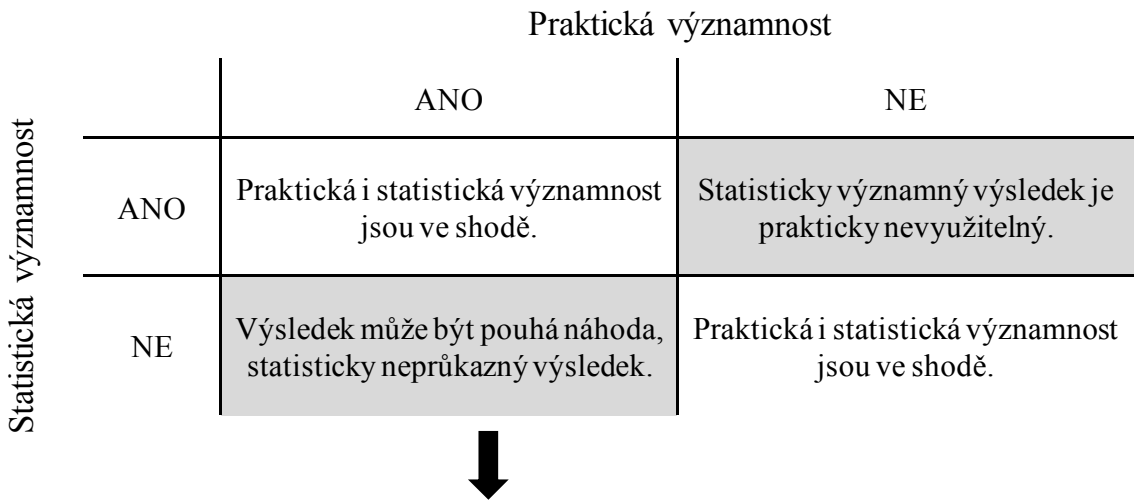


Obr. 1.3 Ilustrace významu intervalu spolehlivosti jako měřítka přesnosti bodového odhadu.

1.4.5 Významnost

Statistika i biostatistika umí na základě pravděpodobnosti ohodnotit výsledek experimentu, tedy umí rozhodnout, zda pozorovaný rozdíl mezi dvěma skupinami vznikl náhodou. Tzv. ***statistická významnost*** je však někdy mylně zaměňována s pravdou. Ve skutečnosti statistická významnost pouze indikuje, že pozorovaný rozdíl není na základě výběrových souborů náhodný (ve smyslu námi vybraného modelu skutečnosti). Zvláště při interpretaci výsledků je třeba si uvědomit, že statistická významnost je pouze jednou stranou mince a nemusí jednoduše znamenat příčinný vztah. Stejně důležitá jako statistická významnost je i tzv. ***praktická významnost***, tedy významnost z hlediska experimentátora (např. lékaře nebo biologa), který má odborné znalosti v dané oblasti vědy, případně čerpá z informací dostupných z literatury. Experimentátor musí na základě pozorovaného efektu vedle statistické významnosti posoudit i jeho věcný význam, tedy zhodnotit, zda je biologicky/klinicky podstatný [2].

Důvodem těchto úvah je fakt, že statistická významnost souvisí s velikostí experimentálního vzorku a lze ji pomocí ní ovlivnit. Vezměme si jako příklad srovnání průměrné výšky lidské postavy mezi dvěma populacemi, například Čechy a Slováky. Srovnáním bodových odhadů průměru můžeme dostat numerický rozdíl např. 0,5 cm, který při znalosti rozsahu hodnot jistě nikdo neoznačí za biologicky podstatný. Přesto lze i tak malý rozdíl při velkém vzorku vyhodnotit jako statisticky významný. Na druhou stranu je třeba říci, že statisticky nevýznamný výsledek nemusí nutně znamenat, že pozorovaný rozdíl ve skutečnosti neexistuje. Opět to může být způsobeno velikostí vzorku, tentokrát ale jeho nedostatečnou velikostí (nedostatečnou informací v pozorovaných datech). Souvislost statistické a praktické významnosti výsledků experimentu ilustruje obrázek 1.4.



Obr. 1.4 Souvislost statistické a praktické významnosti výsledků experimentu.

1.5 Shrnutí

Biostatistika je aplikovaná věda vycházející ze statistiky, která má za cíl získání užitečné informace z pozorovaných dat. V popředí našeho zájmu je popis a vysvětlení pozorované variability studovaných subjektů ve znaku, který nás zajímá a který lze dle povahy jednotlivých variant označit jako kvalitativní nebo kvantitativní. Můžeme-li jednotlivé varianty vyjádřit slovně, mluvíme o znaku kvalitativním, můžeme-li varianty vyjádřit číslem, mluvíme o znaku kvantitativním. Sledovaný znak matematicky reprezentujeme jako náhodnou veličinu, jejíž realizaci získáváme pozorováním konkrétních hodnot této veličiny na výběrovém souboru. Pod pojem biostatistika lze zařadit metodické postupy pro řešení mnoha úloh, nejčastěji jedné z následujících: popis vlastností cílové populace spolu s grafickým znázorněním pozorovaných hodnot, srovnání sledované vlastnosti v rámci dvou nebo více experimentálních skupin, analýzu vztahů náhodných veličin pomocí stochastických modelů nebo vytvoření rozhodovacího pravidla pro predikci či klasifikaci neznámých hodnot. Každá z těchto úloh má svá specifika, kterým je nutné se při hodnocení věnovat, naším cílem totiž není bezhlavé použití biostatistických metod, ale korektní použití biostatistických metod. Jak při hodnocení experimentálních dat, tak při čtení publikovaných výsledků je vždy nutné klást si tyto otázky: „Na jakých subjektech/objektech byl experiment prováděn a jsou tyto subjekty reprezentativní vzhledem k cílové populaci?“, „Jsou srovnávané skupiny subjektů skutečně srovnatelné?“, „Jaká je variabilita pozorovaného efektu, respektive rozdílu?“, „Je statisticky významný výsledek využitelný i z praktického hlediska?“ a „Čím mohou být výsledky experimentu zkresleny?“.

2 Data, jejich popis a vizualizace

V kapitole 1 jsme definovali data jako číselný nebo slovní záznam vlastností našeho pozorovaného souboru, který reprezentuje námi studovanou cílovou populaci. Jednoduše řečeno, data vznikají záznamem skutečnosti, kterou chceme dále analyzovat. Jakýkoliv záznam skutečnosti musí být smysluplný a promyšlený, tedy musíme vědět, co a proč měříme. Nemá totiž smysl měřit a zaznamenávat něco, co buď v cílové populaci nevykazuje žádné rozdíly mezi sledovanými jedinci (např. počet srdcí u mužů), nebo nijak nesouvisí s tím, co se snažíme o cílové populaci zjistit (např. počet rukou u pacientů s rakovinou). Smysluplnost a promyšlenost souvisí s adekvátním plánováním experimentu, který u řady problémů do značné míry ovlivňuje kvalitu výsledků a možnosti interpretace celého experimentu.

Je třeba si také uvědomit, že záznam skutečnosti nikdy není dokonalý a data tedy mohou mít v různých oblastech různou kvalitu. Variabilitu pozorovanou v datech lze rozdělit na dvě složky, informaci a chybu měření. S obojím se lze vypořádat s pomocí statistických metod, obecně však platí, že chybu danou experimentem (samotným měřením hodnot) se snažíme ještě před začátkem experimentu minimalizovat.

2.1 Typy dat

Data reprezentují sledované veličiny, respektive znaky, a proto i typy dat odpovídají typům veličin. ***Kvalitativní*** **(*kategoriální*)** data lze řadit do kategorií, ale nelze je kvantifikovat, respektive jednotlivým kategoriím lze přiřadit číselné kódy, které však nemají logickou souvislost s úrovní sledovaného znaku. Jako příklad můžeme uvést pohlaví, přítomnost viru HIV v krvi, užívání drog nebo barvu vlasů. Naopak, ***kvantitativní*** **(*numerická*)** data můžeme charakterizovat číselnou hodnotou.

Kvalitativní data lze dále dělit do následujících skupin:

- ***Binární data*** mohou nabývat pouze dvou hodnot. Většinou jsou to data typu ano/ne. Příkladem binárních dat je např. přítomnost diabetu (osoba s diabetem / osoba bez diabetu), pohlaví (muž/žena), stav (ženatý/svobodný). Číselně se obvykle kódují pomocí číslic 0 (ne) a 1 (ano).

- ***Nominální data*** obsahují více kategorií, které nelze vzájemně seřadit (neexistuje u nich přirozené pořadí jednotlivých hodnot) a u nichž nemá smysl ptát se na relaci větší/menší. Příkladem nominálních dat je např. krevní skupina (A/B/AB/0), stát EU (Belgie/…/Česká republika/…/Velká Británie), stav (ženatý/svobodný/rozvedený/vdovec).

- ***Ordinální data*** také obsahují více kategorií, na rozdíl od nominálních dat je však lze vzájemně seřadit. U ordinálních dat má smysl ptát se na relaci větší/menší. Příkladem ordinálních dat je např. stupeň bolesti (mírná/střední/velká/nesnesitelná), spotřeba cigaret (nekuřák / ex-kuřák / občasný kuřák / pravidelný kuřák), stadium maligního onemocnění (I/II/III/IV).

Kvantitativní data lze také dále dělit:

- ***Spojitá data*** mohou nabývat jakýchkoliv hodnot v určitém rozmezí (intervalu). Příkladem spojitých dat je výška a hmotnost osob, délka časového období od narození do smrti, velikost nádoru nebo teplota.

- ***Diskrétní data*** mohou nabývat pouze spočetně mnoha hodnot. Při číselné reprezentaci jsou taková data na reálné ose zobrazena pomocí izolovaných bodů. Příkladem diskrétních dat je počet krevních buněk v 1 ml krve, počet králíků v králíkárně, počet hospitalizací pro srdeční slabost, počet krvácivých epizod za rok u pacienta s hemofilií nebo počet dětí v rodině.

Kvantitativní data můžeme rozlišovat také dle toho, jestli je měříme na intervalové nebo poměrové stupnici. V případě ***intervalové stupnice*** se při srovnání jakýchkoliv dvou hodnot můžeme ptát na otázku, o kolik jednotek se tyto dvě hodnoty liší. Můžeme se tedy ptát na rozdíl, nikoliv ale na podíl dvou hodnot, a to z toho důvodu, že u intervalové stupnice je nulová hodnota na místě daném konvencí, které nemusí vyjadřovat absenci daného znaku. Nelze se tedy ptát, kolikrát je jedna hodnota větší než druhá. Typickým příkladem je teplota měřená ve stupních Celsia, kde se můžeme ptát, o kolik stupňů je dnes tepleji než bylo včera, ale nemá smysl se ptát na to, kolikrát je dnes tepleji než včera (nula stupňů Celsia není počátek stupnice, připouštíme zde i záporné hodnoty). ***Poměrová stupnice*** má nulovou hodnotu na místě, které odpovídá nepřítomnosti sledovaného znaku, a umožňuje nám tak se ptát i na otázku, kolikrát je jedna hodnota větší než druhá. Kromě podílu se samozřejmě můžeme u poměrové stupnice ptát i na rozdíl dvou hodnot. Příkladem poměrových dat jsou již zmiňované výška a váha osob, velikost nádoru nebo počet krevních buněk v 1 ml krve.

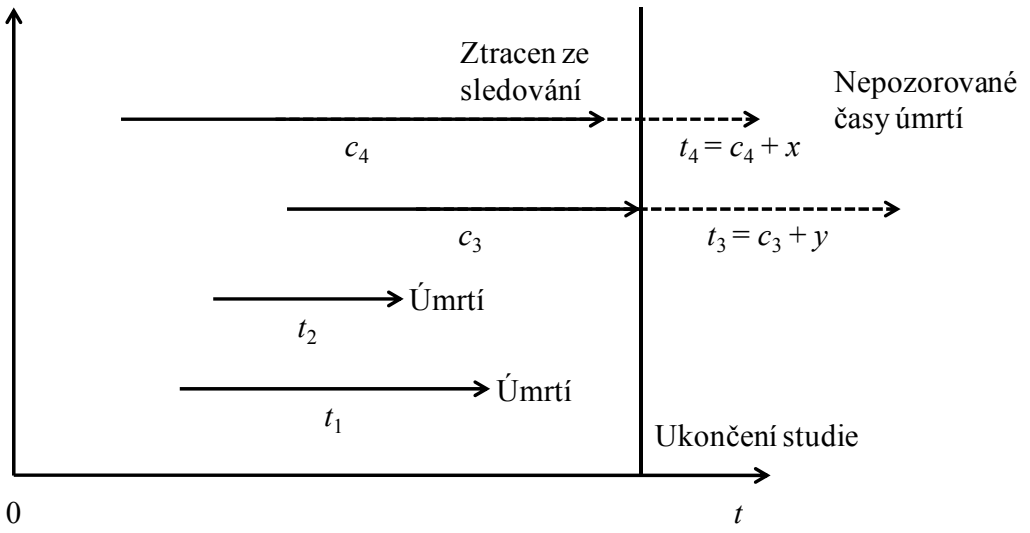
Data lze samozřejmě pro analýzu převádět (zjednodušovat) ze spojitých na diskrétní, případně ordinální. Je to výhodné zejména kvůli lepší interpretaci výsledků, ale také kvůli snazší práci s daty a jejich jednodušší analýze. Je třeba si však uvědomit, že agregace kvantitativních dat do kategorií (např. kategorizace věku do desetiletých věkových skupin) znamená ztrátu části informace uložené v datech, kterou nejsme schopni bez primárních dat zpětně rekonstruovat, a která v případě testování hypotéz vede většinou ke snížení schopnosti testu rozhodnout o platnosti nebo neplatnosti studované hypotézy.

2.1.1 Data cenzorovaná

Cenzorovaná data charakterizují experimenty, kde sledujeme čas do výskytu předem definované události [4, 15]. V průběhu sledování však událost nemusí nastat u všech subjektů. Subjekty pak nelze vinit z toho, že jsme u nich nebyli schopni danou událost pozorovat a už vůbec je nelze z hodnocení vyloučit. O čase sledování takového subjektu pak mluvíme jako o cenzorovaném. Toto označení indikuje, že sledování bylo ukončeno dříve, než u subjektu došlo k definované události. Nevíme tedy, kdy a jestli vůbec daná událost u subjektu nastala, víme pouze, že nenastala před ukončením sledování. Ukázka časového sledování čtyř osob, kdy u dvou z nich je pozorovaný čas do úmrtí cenzorován, je znázorněna na obrázku 2.1.

Abychom byli úplně přesní, výše popsaný princip cenzorování se označuje jako tzv. cenzorování zprava a odpovídá nejčastější situaci, ke které dochází v klinickém výzkumu. Kromě něj existují ještě dva méně časté typy cenzorování, a to cenzorování zleva a intervalové cenzorování. O cenzorování zleva mluvíme v případě, že sledovaná událost nastala před určitým časovým bodem, ale nevíme, kdy přesně (budeme-li sledovat věk při prvním požití drog u studentů středních škol, je možné, že někteří studenti je zkusili již na základní škole a přesný věk již nelze stanovit). Pozorovaná doba bez události je tedy větší než skutečná doba bez události. Intervalové cenzorování označuje případ, kdy sledovaná událost nastane mezi dvěma danými časy, které jsme schopni pozorovat, tedy v nějakém časovém intervalu. Příkladem takového intervalu může být doba mezi dvěma návštěvami u lékaře,

mezi nimiž se může u osoby manifestovat nějaké onemocnění, které lze ale odhalit pouze lékařským vyšetřením. Datum diagnózy pak bude shodné s datem návštěvy lékaře, i když je jasné, že onemocnění se u dané osoby vyskytlo již dříve.



Obr. 2.1 Rozdíl mezi kompletním časem do sledované události (t_1, t_2) a cenzorovaným časem (c_3, c_4).

2.1.2 Další typy dat, data odvozená

Kromě již zmíněných datových typů existuje ještě řada dalších, které jsou také běžné v biologickém a klinickém výzkumu. Za zmínku stojí následující typy dat:

- ***Podílem*** (*ratio*) je reprezentována řada indexů. Nejjednodušším příkladem je tzv. *body mass index* (BMI), který je dán vzájemným poměrem mezi tělesnou hmotností v kilogramech a výškou v metrech na druhou.

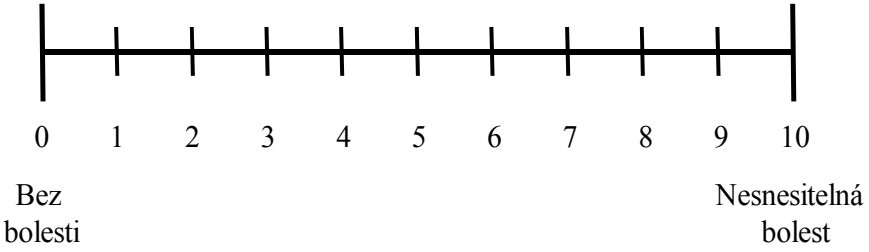
- ***Procento*** (*percentage*) může vyjadřovat jak relativní četnost výskytu sledované události tak třeba zlepšení určité charakteristiky (veličiny). Sledujeme-li např. vývoj ejekční frakce levé srdeční komory u pacientů s prodělaným srdečním selháním v čase, je výhodné kromě absolutní změny sledovat i procentuální zlepšení. Formálně se jedná o podíl aktuální a referenční hodnoty, která reprezentuje 100 %.

- ***Míra pravděpodobnosti*** (*rate*) se týká především výskytu různých onemocnění, kdy počet nových pacientů za dané období (délka trvání studie) je vztažen na celkový počet zaznamenaných osobo-roků. Příkladem tohoto typu dat je roční incidence nádorových onemocnění u dospělých osob žijících v ČR.

- ***Pořadí*** (*rank*) někdy nahrazuje absolutní hodnoty, které nejsme schopni přesně zaznamenat. Z hlediska informační hodnoty se sice jedná o ztrátu určitého množství informace, nicméně i pořadí lze v biostatistice využít – řada metod je založena na práci s pořadími pozorovaných hodnot, zejména neparametrické testy jako např. Wilcoxonův test a Mannův-Whitneyho test (tyto testy jsou uvedeny v kapitole 7).

- ***Skóre*** (*score*) představují uměle vytvořené hodnoty charakterizující nejčastěji určitý stav sledovaného subjektu, který nelze jednoduše změřit jako číselnou hodnotu. Příkladem jsou indexy vyjadřující kvalitu života.

- ***Vizuální škála*** (*visual scale*) také většinou souvisí s hodnocením kvality života, neboť pacienti často hodnotí svoje obtíže na škále, která má formu úsečky, kde jeden konec úsečky představuje minimální a druhý naopak maximální obtíže spojené s daným onemocněním. Příklad vizuální škály je uveden na obrázku 2.2.



Obr. 2.2 Příklad vizuální škály pro hodnocení stupně bolesti.

Je třeba poznamenat, že vyjádření výsledků v relativní formě (procento) má často příjemnou interpretaci, ale může být zavádějící. Uvažujme následující příklad srovnání účinnosti léčiv A a B ve smyslu prevence cévní mozkové příhody (CMP) u dvou skupin osob, respektive ve dvou různých studiích. Ve studii 1 byl sledovaný výskyt CMP ve skupině A 12 %, ve skupině B pak 20 %. Vypočteme-li relativní změnu v účinnosti obou preparátů, pak dostaneme číslo 40 %, absolutní změna je pak rovna 8 %. Ve studii 2 byl sledovaný výskyt CMP ve skupině A pouze 0,9 %, ve skupině B pak 1,5 %. Z toho vyplývá, že ve druhé studii byla relativní změna v účinnosti opět 40 %, nicméně absolutní změna je v tomto případě pouze 0,6 %. Výsledkem je rozdílný přínos léčby A při stejné relativní účinnosti a je zřejmé, že efekt léčby A je mnohem výraznější ve studii 1. Relativní vyjádření účinnosti by proto mělo vždy být doprovázeno i absolutním vyjádřením účinnosti. Interpretační problém je v rozdílném významu procenta v obou typech srovnání: absolutní změna je vyjádřena vzhledem k původním referenčním hodnotám (počet osob ve studii), kdežto relativní změna je vyjádřena vzhledem k jedné ze dvou účinností (*de facto* tak počítáme procenta z procent).

Dalším důležitým aspektem je v souvislosti s typy dat míra subjektivity, kterou je ten či onen datový typ ovlivněn, respektive do jaké míry je měření dané veličiny ovlivnitelné subjektivními faktory. Klasickým příkladem je rozdíl mezi měřením tlaku krve a bolestivosti rány po operaci. Zatímco měření tlaku krve je standardní procedura, hodnocení bolestivosti rány je do značné míry závislé na povaze pacienta a jeho psychickém stavu. Lze tedy očekávat, že měření bolestivosti bude zatíženo subjektivitou v daleko větší míře než zmíněné měření tlaku. Obecně lze říci, že pozorované hodnoty veličin ovlivněných subjektivitou mají větší variabilitu než u veličin neovlivněných subjektivitou a lze je tedy i obtížněji hodnotit.

2.2 Význam popisu a vizualizace dat

Metody popisné statistiky mají za cíl sumarizovat pozorované hodnoty tak, abychom mohli lépe a jednodušeji pracovat s informací v nich uloženou. Na základě pozorovaných hodnot chceme vypočítat hodnoty označované jako ***statistiky*** (formálně se jako statistika označuje náhodná veličina, která je funkcí náhodného výběru), které budou pozorovaná data dále zastupovat při prezentaci, testování apod. Jinak řečeno, tyto zástupné hodnoty slouží k „uložení“ informace obsažené v datech, neboť použití všech pozorovaných hodnot je nepraktické a často vlastně i nemožné.

Cílem vizualizace dat je především pozorovaná data graficky zpřehlednit a poskytnout uživateli na minimální ploše maximum informací. Vizualizace dat je nezbytná bez ohledu na to, jak pokročilé metody chceme následně při zpracování dat použít, vizualizace nám totiž dává informaci nejen o charakteru dat, případně procesu, při kterém vznikla, ale i o problematických záznamech a chybných hodnotách. Bez adekvátní vizualizace můžeme v datech nechat pozorování, která negativně ovlivní další hodnocení a znemožní správnou interpretaci výsledků. Identifikaci problematických hodnot se věnuje část 2.3.

Výstupy popisného zpracování dat jsou obecně neformální, jde pouze o shrnutí pozorovaného a ne o formální testování hypotéz s použitím statistických testů. Vztahují se pouze na vybraný soubor dat, respektive na experimentální vzorek, a zobecnění získaných informací na celou cílovou populaci je tak opět podmíněno jeho reprezentativností vůči cílové populaci. Výstupy popisné statistiky často slouží jako podklad pro stanovení hypotéz, které jsou následně ověřeny dalšími experimenty.

2.2.1 Popis a vizualizace kvalitativních dat

Označme $x_1,\ldots, x_n$ zaznamenané hodnoty sledovaného znaku u výběrového souboru n subjektů. U kvalitativních dat předpokládáme opakování pozorování jednotlivých hodnot daného znaku, proto je logické sumarizovat tato data pomocí tabulky s četnostmi možných hodnot (***tabulka četností***). Označíme-li $y_1,\ldots, y_m$ možné hodnoty sledovaného znaku, ***pozorovanou (absolutní) četnost*** odpovídající variantě znaku y_j budeme označovat jako n_j. Pro lepší orientaci a možnost srovnání je vhodné doplnit pozorovanou četnost i ***relativní četností***, která má pro variantu znaku y_j tvar n_j / n.

Příklad 2.1. Sledujeme přítomnost diabetu u pacientů zdravotnického zařízení za období jednoho roku s tím, že rozlišujeme pacienta bez diabetu a pacienty s diabetem 1. nebo 2. typu ($m = 3$). Celkem bylo pozorováno $n = 687$ pacientů, sumarizaci výsledků uvádí tabulka 2.1.

Tabulka 2.1 *Počty pacientů ve zdravotnickém zařízení dle přítomnosti diabetu*

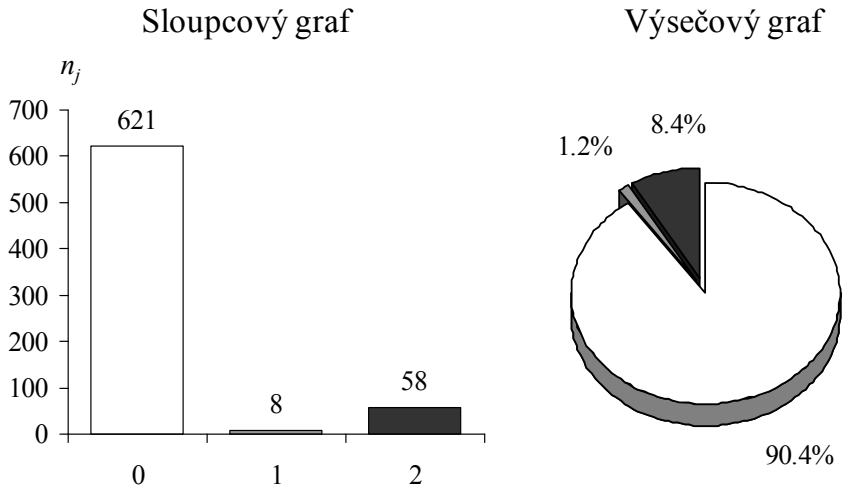
Přítomnost diabetu	y_j	n_j	n_j / n	n_j / n (%)
Bez diabetu	0	621	0,904	90,4 %
Diabetes 1. typu	1	8	0,084	1,2 %
Diabetes 2. typu	2	58	0,012	8,4 %
Celkem		687	1	100 %

Vzhledem k tomu, že kvalitativní data často nelze seřadit dle velikosti, používá se jako frekvenční charakteristika těchto dat tzv. ***mód***, což je varianta znaku s největší četností. V příkladu 2.1 je modální hodnotou pacient bez diabetu. Vypovídací hodnota módu jako reprezentanta pozorovaných dat závisí především na počtu kategorií sledovaného znaku a vyrovnanosti jejich četností. Někdy může být mód opravdu typickou hodnotou, jindy mohou být četnosti jednotlivých variant znaku tak vyrovnané, že to spíše indikuje neexistenci typické hodnoty pro daný znak.

V případě nízkých pozorovaných četností některých kategorií je často vhodné tyto kategorie sloučit a dále pracovat již pouze se sloučenými kategoriemi. Slučovat by se však měly pouze sousední kategorie a ještě pouze v případě, kdy jejich sloučení zachovává data interpretovatelnými.

Pro vizualizaci kvalitativních dat se nejčastěji používají ***sloupcový*** a ***výsečový*** (***koláčový***) ***graf***, kde výška sloupců (šířka je pro všechny sloupce stejná), respektive plocha výsečí, pro jednotlivé varianty je úměrná jejich četnosti. U koláčového grafu jeho plocha představuje 100 %, proto je vhodný pro vizualizaci relativních četností, ve sloupcovém grafu můžeme zobrazit

obojí, jak absolutní, tak relativní četnosti. Příklad sloupcového a koláčového grafu s absolutními a relativními četnostmi z tabulky 2.1 je uveden na obrázku 2.3.



Obr. 2.3 Příklad sloupcového a výsečového grafu na datech z tabulky 2.1.

2.2.2 Popis a vizualizace kvantitativních dat

Opět označme pozorované hodnoty sledovaného znaku u n subjektů výběrového souboru jako $x_1,\ldots, x_n$. Na rozdíl od kvalitativních dat dochází u kvantitativních dat k opakování pozorování jednotlivých hodnot daného znaku zřídka a tabulku četností, tak jak byla definována výše, nelze pro popis dat použít. Pro použití tabulky četností je třeba nejprve seskupit pozorované hodnoty do m disjunktních, vyčerpávajících a hlavně smysluplných intervalů, které pak v tabulce četností nahrazují kategorie kvalitativního znaku. Znázornění tabulky četností je pak stejné jako v předchozím případě, pro přehlednost je v ní vhodné uvádět i šířku zvolených intervalů (šířku j-tého intervalu budeme značit d_j), zejména kvůli srovnatelnosti výsledků.

Příklad 2.2. Uvažujme věk n = 6500 pacientek s karcinomem prsu, který chceme sumarizovat v následujících věkových intervalech: 0–39 let, 40–49 let, 50–59 let, 60–69 let, 70 a více let. Sumarizaci zvolených intervalů, jejich absolutních četností, n_j, i relativních četností, n_j / n, ukazuje tabulka 2.2.

Tabulka 2.2 *Věková struktura souboru n = 6500 pacientek s karcinomem prsu*

Věkový interval	d_j	n_j	n_j / n	n_j / n (%)
0–39 let	40	231	0,036	3,6 %
40–49 let	10	747	0,115	11,5 %
50–59 let	10	1559	0,240	24,0 %
60–69 let	10	1894	0,291	29,1 %
70 a více let	20	2069	0,318	31,8 %
Celkem	90	6500	1	100 %

Míry polohy

I když nám frekvenční tabulka zpřehledňuje pozorované hodnoty a umožňuje zjistit, kterých hodnot je v našem souboru více a kterých naopak méně, je vhodné ji vždy doplnit statistikou, která shrnuje soubor dat jedním číslem a představuje „typickou hodnotu“, kolem které mají ostatní pozorované hodnoty tendenci kolísat. Nejčastěji jsou jako charakteristiky polohy používány statistiky ***průměr*** a ***medián***. Průměr (také ***aritmetický průměr*** či ***výběrový průměr***) lze jednoduše spočítat jako součet pozorovaných hodnot dělený jejich počtem:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{2.1}$$

Použití průměru jako sumarizace n pozorovaných hodnot se učí už na základní škole, zmínka o jeho používání je již z konce 17. století. Byl navržen bez ohledu na jakoukoliv souvislost s teorií pravděpodobnosti jako hodnota, označme ji a, která má následující vlastnosti:

1. Hodnota a minimalizuje reziduální součet čtverců, tedy součet čtverců rozdílů (odchylek, reziduí) pozorovaných hodnot od hodnoty a:

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2 \,. \tag{2.2}$$

2. Součet reziduí vzhledem k hodnotě a je nula, tedy kladná i záporná rezidua jsou v rovnováze:

$$\sum_{i=1}^{n}(x_i - a) = 0 \,. \tag{2.3}$$

Abychom mohli definovat medián, je třeba kromě neuspořádaného výběrového souboru $x_1,\ldots, x_n$ uvažovat i jeho uspořádanou variantu $x_{(1)} \le x_{(2)} \ldots \le x_{(n)}$, kde $x_{(1)}$ značí minimální pozorovanou hodnotu a $x_{(n)}$ značí maximální pozorovanou hodnotu. Medián pak definujeme následovně:

$$\begin{aligned} \tilde{x} &= x_{((n+1)/2)} \quad \text{je-li } n \text{ liché} \\ \tilde{x} &= \tfrac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) \quad \text{je-li } n \text{ sudé} \end{aligned} \tag{2.4}$$

Z výše uvedeného je vidět, že zatímco průměr je vypočten ze všech pozorovaných hodnot a všechny hodnoty souboru se tak podílejí na jeho výsledné číselné realizaci, medián je prostřední pozorovaná hodnota, která dělí celý soubor na dvě poloviny, tedy polovina souboru je menší než medián a naopak polovina souboru je větší než medián. S těmito vlastnostmi obou statistik jsou spojeny jejich výhody i nevýhody.

Chceme-li, aby naše vypočtená statistika byla dobrým odhadem frekvenčního středu dat, je medián vždy dobrou volbou. Průměr je v tomto případě dobrou volbou pouze tehdy, když jsou naše data symetrická a neobsahují odlehlé či nesprávné hodnoty. V případě asymetrických dat nebo přítomnosti odlehlých hodnot má totiž průměr tendenci se těmto

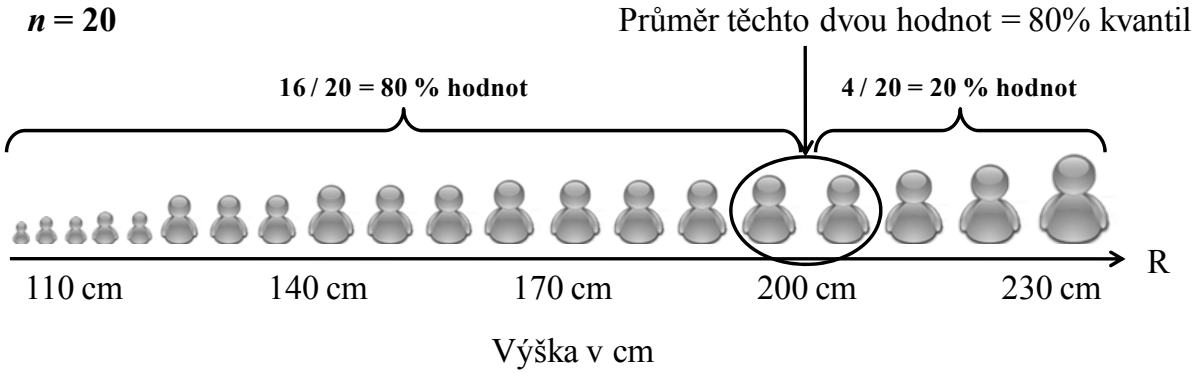
„netypickým“ hodnotám přizpůsobovat, což ho jako odhad frekvenčního středu dat diskvalifikuje. Typickým příkladem pro vysvětlení této vlastnosti průměru je výpočet průměrného platu v České republice. Je totiž zřejmé, že průměrný plat není dobrým odhadem středního výdělku české populace, neboť jeho hodnota je značně ovlivněna malou skupinou lidí s velmi vysokými příjmy. Medián je na druhou stranu dobrým odhadem středního výdělku české populace, protože jednoznačně určuje frekvenční střed dosahovaných příjmů. Problematickou situací pro obě míry, tedy průměr i medián, jsou data se dvěma (případně více) frekvenčními středy, kde může být zavádějící použití obou měr. V tomto případě by mělo primárně dojít k analýze toho, co způsobuje toto chování, a případně by mělo dojít k adekvátnímu rozdělení souboru (může se nám např. stát, že máme ve výběrovém souboru dvě homogenní skupiny, které se však ve sledovaném znaku vzájemně liší).

Jako míry polohy lze použít i minimální (hodnota $x_{(1)}$) a maximální (hodnota $x_{(n)}$) pozorované hodnoty, které nám také dávají obraz o tom, kde se námi sledovaná náhodná veličina X pohybuje na reálné ose. S uspořádanou variantou výběrového souboru, tedy s hodnotami $x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)}$ souvisí i další důležitý pojem statistiky a analýzy dat, a to pojem ***kvantil***. Ve statistice je kvantil definován pomocí kvantilové funkce (více o kvantilech v kapitole 3), laicky lze kvantil definovat jako číslo na reálné ose, které rozděluje pozorované hodnoty na dvě části dle pravděpodobnosti. Jinak řečeno, tzv. p% kvantil (p-procentní kvantil) rozděluje data na p procent hodnot a $(100 - p)$ procent hodnot, kdy p procent hodnot je menších (nebo rovno) než p% kvantil a naopak $(100 - p)$ procent hodnot je větších (nebo rovno) než p% kvantil. Mluvíme-li o p% kvantilu pozorovaných hodnot, je třeba si uvědomit, že se vždy jedná o jednu z naměřených hodnot, tedy jednu z hodnot $x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)}$, případně o průměr dvou takových sousedních hodnot. Označíme-li p% kvantil jako $x_{p/100}$, můžeme ho mezi seřazenými hodnotami najít následovně:

$$\begin{aligned} x_{p/100} &= x_{(k)} \quad \text{pro } np/100 \text{ neceločíselné, pak } k = \lceil np/100 \rceil; \\ x_{p/100} &= \tfrac{1}{2}(x_{(k)} + x_{(k+1)}) \quad \text{pro } np/100 \text{ celočíselné, pak } k = np/100; \end{aligned} \tag{2.5}$$

přitom $\lceil np/100 \rceil$ představuje horní celou část čísla $np/100$.

Příklad nalezení 80% kvantilu hodnot výšky v souboru 20 osob ukazuje obrázek 2.4. Významnými kvantily jsou již zmíněné minimální (0% kvantil) a maximální (100% kvantil) pozorovaná hodnota a medián (50% kvantil), kromě nich jsou ještě používány hodnoty 25% a 75% kvantilu, které se standardně nazývají ***dolní*** a ***horní kvartil***.



Obr. 2.4 Příklad nalezení 80% kvantilu hodnot výšky v souboru 20 osob.

Míry variability

Výpočet míry polohy jako „typického“ pozorování je nezbytné doplnit také informací o tom, jak jsou kolem této hodnoty rozložena ostatní pozorování, což znamená doplnit míru polohy tzv. ***mírou variability***. Důvod je zřejmý, je třeba od sebe odlišit dva znaky, které nabývají stejné průměrné hodnoty (např. 50), ale zásadně se liší ve spektru hodnot, jež tento znak může nabývat. Ve chvíli, kdy první znak může nabývat např. hodnot od 0 do 100 a druhý od 40 do 60, je jasné, že první znak vykazuje větší variabilitu než znak druhý, což bychom nebyli z pouhé znalosti průměru schopni zjistit. Jak již bylo naznačeno, nejjednodušší charakteristikou variability pozorovaných dat je ***rozsah hodnot*** (***rozpětí***), který je dán minimální a maximální pozorovanou hodnotou. Nevýhodou prezentování rozsahu pozorovaných hodnot je jeho náchylnost k netypickým, odlehlým, případně chybným hodnotám. Tento fakt lze na druhou stranu využít právě pro identifikaci problematických hodnot a čištění datového souboru ještě před začátkem jakéhokoliv statistického zpracování.

Další mírou variability, která není téměř vůbec náchylná na odlehlá pozorování je tzv. ***kvantilové rozpětí***, což je interval definovaný hodnotami p% kvantilu a (100 – p)% kvantilu. Speciálním případem kvantilového rozpětí je tzv. ***kvartilové rozpětí*** (*interquartile range, IQR*), které je dáno dolním a horním kvartilem a které pokrývá 50 % pozorovaných hodnot:

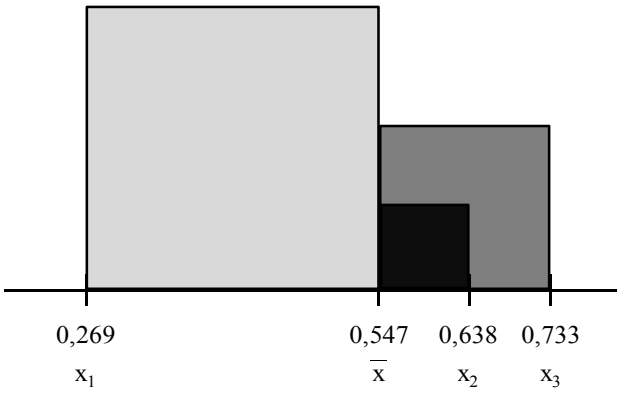
$$IQR = x_{0,75} - x_{0,25} \,. \tag{2.6}$$

Rozsah hodnot i kvantilové rozpětí nám sice dávají informaci o variabilitě, ale v obou případech se jedná o charakteristiky vypočtené na základě dvou pozorování, které nezohledňují polohu „typického“ pozorování, např. průměru nebo mediánu. Fluktuaci pozorovaných hodnot kolem průměru odráží ***výběrový rozptyl*** (*sample variance*), značíme ho s^2, a je definován jako průměrný kvadrát odchylky pozorovaných hodnot od hodnoty průměru:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right), \tag{2.7}$$

Je třeba poznamenat, že ve jmenovateli vzorce (2.7) pro výběrový rozptyl je výraz n – 1 a nikoliv n. Jedná se o výpočetní korekci, která má zamezit podhodnocení výběrového rozptylu u malých výběrových souborů a která je známa pod označením Besselova korekce.

Ukázka kvadrátů odchylek tří pozorovaných hodnot od jejich průměru je zobrazena na obrázku 2.5, ze kterého je evidentní, že výběrový rozptyl trpí stejnou nevýhodou jako průměr, a to citlivostí na odlehlé a chybné hodnoty, která je ještě zvýrazněna druhou mocninou. Výběrový rozptyl má navíc interpretační nevýhodu v tom, že nemá stejné jednotky jako pozorované hodnoty a jejich průměr, a proto se častěji jako míra variability používá jeho odmocnina, tzv. ***výběrová směrodatná odchylka***, kterou značíme s.



Obr. 2.5 Kvadráty odchylek tří pozorovaných hodnot od průměru.

Příklad 2.3. Naším cílem je vypočítat průměr, medián a výběrovou směrodatnou odchylku hladiny cholesterolu vybrané populace mužů (n = 22). Naměřené hodnoty jsou uvedeny v mmol/l a jsou dány v tabulce 2.3.

Tabulka 2.3 *Hodnoty cholesterolu vybrané populace mužů (mmol/l).*

6.2	7.6	6.3	9.1	4.2	5.8	5.65	6.3	8.6	6.0	6.2
6.7	4.6	6.25	6.4	4.04	6.3	9.1	6.3	5.2	6.4	5.75

Výpočet požadovaných statistik (v mmol/l) je pak následující:

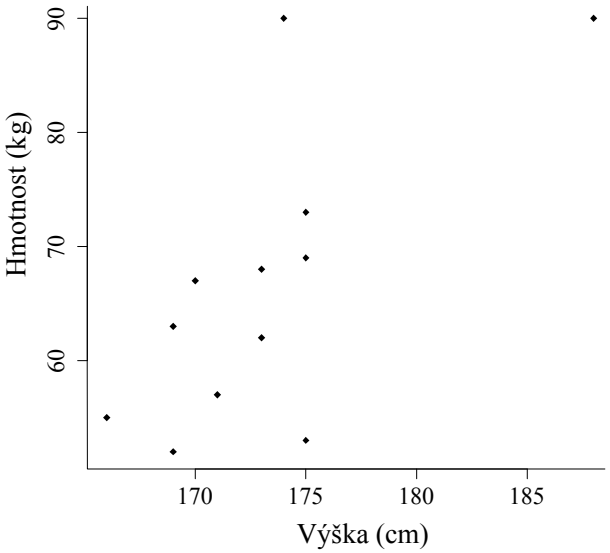
$$\text{Průměr:}\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{22}\sum_{i=1}^{22} x_i = \frac{1}{22}138{,}99 = 6{,}318\,,$$

$$\text{Medián:}\quad \tilde{x} = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) = \frac{1}{2}(6{,}25 + 6{,}3) = 6{,}275\,, \tag{2.8}$$

$$\text{Směrodatná odchylka:}\quad s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} = \sqrt{\frac{1}{21}\left(915{,}639 - 878{,}101\right)} = 1{,}337\,.$$

Bodový graf

Bodový graf (*scatter plot*) je grafický nástroj pro vizualizaci kvantitativních dat zobrazující každou měřenou hodnotu jako bod plochy. Lze ho použít na vizualizaci naměřených hodnot v několika kategoriích (od jedné až po mnoho), ale jeho největší přínos je zejména ve vizualizaci vzájemného vztahu dvou veličin spojitého typu, kdy hodnoty jedné veličiny jsou zobrazeny na ose x a hodnoty druhé veličiny jsou zobrazeny na ose y. Příklad bodového grafu je uveden na obrázku 2.6, kde vidíme bodové znázornění hodnot výšky a hmotnosti studentů 2. ročníku Matematické biologie.



Obr. 2.6 Bodový graf hodnot výšky a hmotnosti studentů Matematické biologie.

Histogram

Neocenitelným a možná nejpoužívanějším grafickým nástrojem pro vizualizaci poměrových a intervalových dat je tzv. ***histogram***. Histogram vzhledem připomíná sloupcový graf, ale na rozdíl od sloupcového grafu každý sloupec v histogramu odráží absolutní nebo relativní četnost na jednotku sledované veličiny na vodorovné ose. Naproti tomu sloupcový graf znázorňuje kvalitativní data a jako takový s žádnými jednotkami na vodorovné ose nepracuje; na kvantitativní data jej lze použít až po jejich kategorizaci (agregaci do intervalů).

Máme-li n hodnot sledované veličiny u výběrového souboru: $x_1,\ldots, x_n$, je třeba je pro vytvoření histogramu nejdříve seřadit dle velikosti a rozdělit do m vzájemně disjunktních intervalů, které vytvoříme na vodorovné ose. Šířku j-tého intervalu označíme jako d_j a počet pozorovaných hodnot, které padly do j-tého intervalu, označíme symbolem n_j. Výšku sloupců histogramu pro j-tý interval pak můžeme vyjádřit pomocí relativní četnosti jako

$$f(j) = \frac{n_j / n}{d_j} \ , \tag{2.9}$$

nebo pomocí absolutní četnosti jako

$$f^*(j) = \frac{n_j}{d_j} \ . \tag{2.10}$$

Příklad 2.4. Vraťme se k datům o věku 6500 pacientek s karcinomem prsu z příkladu 2.2 a sestrojme histogram s použitím věkových kategorií: 0–39 let, 40–49 let, 50–59 let, 60–69 let, 70 a více let. Pozorované absolutní a relativní četnosti i hodnoty $f(j)$ a $f^*(j)$ nezbytné pro sestrojení histogramu sumarizuje tabulka 2.4, histogramy pro absolutní a relativní četnost s použitím dat z tabulky 2.4 ukazuje obrázek 2.7.

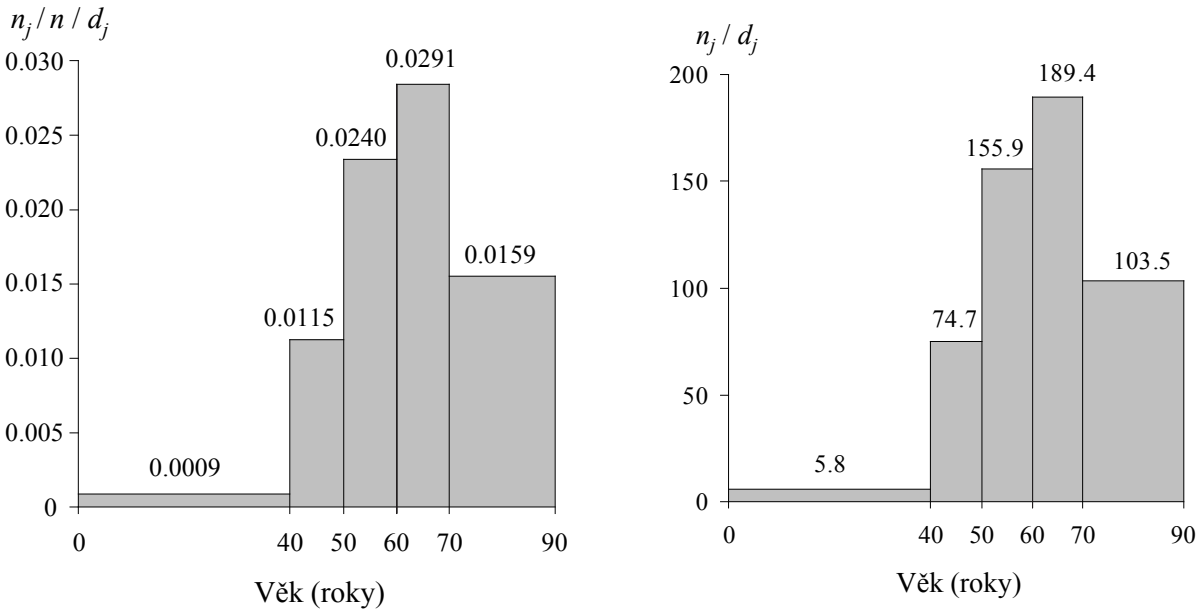
Tabulka 2.4 *Věková struktura souboru n = 6500 pacientek s karcinomem prsu*

Věkový interval	d_j	n_j	n_j / n	n_j / d_j	$n_j / n / d_j$
0–39 let	40	231	0,036	5,8	0,0009
40–49 let	10	747	0,115	74,7	0,0115
50–59 let	10	1559	0,240	155,9	0,0240
60–69 let	10	1894	0,291	189,4	0,0291
70 a více let	20	2069	0,318	103,5	0,0159
Celkem	90	6500	1	-	-

Přepočet absolutních a relativních četností na šířku intervalu vypadá na první pohled zbytečně, nicméně důvody pro tento výpočet jsou dva:

1. Přepočet na šířku intervalu zajišťuje zároveň jejich srovnatelnost vzhledem k absolutním i relativním četnostem. Příkladem může být srovnání četností věkových kategorií 60–69 let a 70 a více let v tabulce 2.4. Z hlediska absolutní i relativní četnosti nestandardizované na šířku intervalu se zdá být věkový interval 70 a více let četnější než interval 60–69 let, je to ale dáno tím, že zahrnuje širší věkové spektrum. Po standardizaci na šířku intervalu je vidět, že četnější je naopak věková kategorie 60–69 let.

2. Celková plocha histogramu pro absolutní četnost je rovna celkové velikosti výběru, zatímco celková plocha histogramu pro relativní četnost je rovna 1. Tato skutečnost má těsnou vazbu na základní popis pravděpodobnostního chování náhodné veličiny, kterým je ***hustota rozdělení pravděpodobnosti***. Ta je definována spolu s dalšími charakteristikami náhodné veličiny v kapitole 3, nicméně je třeba poznamenat, že histogram lze chápat jako odhad tvaru hustoty pravděpodobnosti. Jinými slovy, je to grafická vizualizace pravděpodobnostního chování kvantitativních (zejména spojitých) dat.



Obr. 2.7 Histogram pro relativní četnost (vlevo) a absolutní četnost (vpravo) z příkladu 2.4.

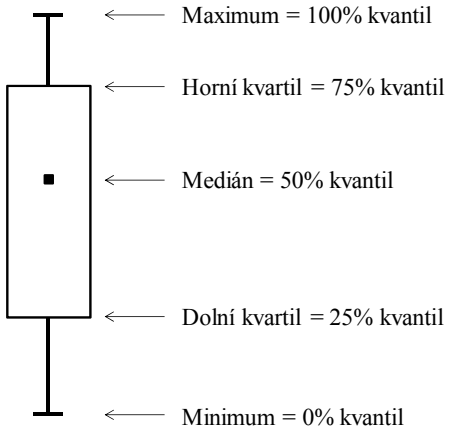
Na druhou stranu, v dnešním statistickém software je histogram zřídka vyjadřován pomocí výrazů 2.9 a 2.10. Daleko častěji se jedná o prosté absolutní nebo relativní počty pozorování v daném intervalu, které jsou výhodné zejména kvůli snadné čitelnosti

a interpretaci. Abychom však byli schopni adekvátní interpretace, je důležité, aby intervaly měly stejnou šířku, a to z důvodu srovnatelnosti, který byl popsán výše.

Dalším důležitým aspektem tvorby histogramu je počet jeho intervalů, neboť ten v zásadě rozhoduje o tom, jak bude histogram vypadat. Při malém počtu intervalů může být charakter dat maskován, zatímco při velkém počtu intervalů zase můžeme pozorovat velkou variabilitu v četnostech jednotlivých intervalů. Jak tedy volit počet intervalů? Nejčastěji jsou používány dvě jednoduché metody, kdy v prvním případě volíme počet intervalů (m) roven odmocnině z celkového počtu pozorování, tedy $m = \sqrt{n}$, v druhém případě pak podle tzv. ***Sturgesova pravidla*** volíme počet intervalů jako 1 + logaritmus o základu dva z celkového počtu pozorování, tedy $m = 1 + \log_2(n)$.

Krabicový graf

Dalším nástrojem pro vizualizaci kvantitativních dat je tzv. ***krabicový graf*** (*box plot*), což je, jak název napovídá, graf ve tvaru obdélníku doplněný tzv. ***fousky*** (*whiskers*). Jednotlivé prvky krabicového grafu nejčastěji odpovídají významným kvantilům vypočteným na základě pozorovaných dat. Uvnitř obdélníkového tvaru je naznačena pozice mediánu (50% kvantilu) a obdélník samotný značí polohu dolního a horního kvartilu, tedy 25% a 75% kvantilu. Tyto dva kvantily odpovídají ***kvartilovému rozpětí***, které ohraničuje 50% pozorovaných hodnot. Fousky dosahující za hranice obdélníkového tvaru pak signalizují polohu hodnot více vzdálených od mediánu, nejčastěji odpovídají 5% kvantilu (spodní fousek) a 95% kvantilu (horní fousek), případně minimu a maximu pozorovaných hodnot.



Obr. 2.8 Příklad krabicového grafu s vyznačením významných kvantilů pozorovaných dat.

2.3 Identifikace odlehlých hodnot

Zásadní vliv odlehlých hodnot na popisné statistiky a tedy i nezbytnost jejich identifikace lze nejlépe ilustrovat příkladem.

Příklad 2.5. Uvažujme data z příkladu 2.3, v nichž zaměníme jednu jedinou správnou hodnotu za hodnotu odlehlou (a to tak, že pouze vynecháme desetinnou čárku). Data s odlehlou hodnotou jsou dána v tabulce 2.5, odlehlá hodnota je zobrazena tučně.

Tabulka 2.5 *Hodnoty cholesterolu vybrané populace mužů s odlehlou hodnotou.*

6.2	7.6	6.3	9.1	4.2	5.8	5.65	6.3	8.6	6.0	6.2
6.7	4.6	6.25	6.4	4.04	6.3	9.1	6.3	5.2	**64**	5.75

Výpočet popisných statistik je uveden v tabulce 2.6. Srovnáme-li výsledky výpočtů na datech s a bez odlehlé hodnoty, je vidět, že odlehlá hodnota velmi výrazně ovlivňuje jak hodnotu průměru, tak výběrové směrodatné odchylky, které již vůbec neodrážejí původně naměřené hodnoty hladiny cholesterolu. Jinými slovy, průměr ovlivněný odlehlou hodnotou nelze považovat za adekvátní míru střední tendence těchto dat a výběrovou směrodatnou odchylku ovlivněnou odlehlou hodnotou nelze považovat za adekvátní míru jejich variability. Na druhou stranu, hodnota mediánu se vlivem odlehlé hodnoty nemění, neboť odlehlá hodnota nemění frekvenční střed dat.

Tabulka 2.6 *Popisné statistiky vypočtené na datech s a bez odlehlé hodnoty (v mmol/l).*

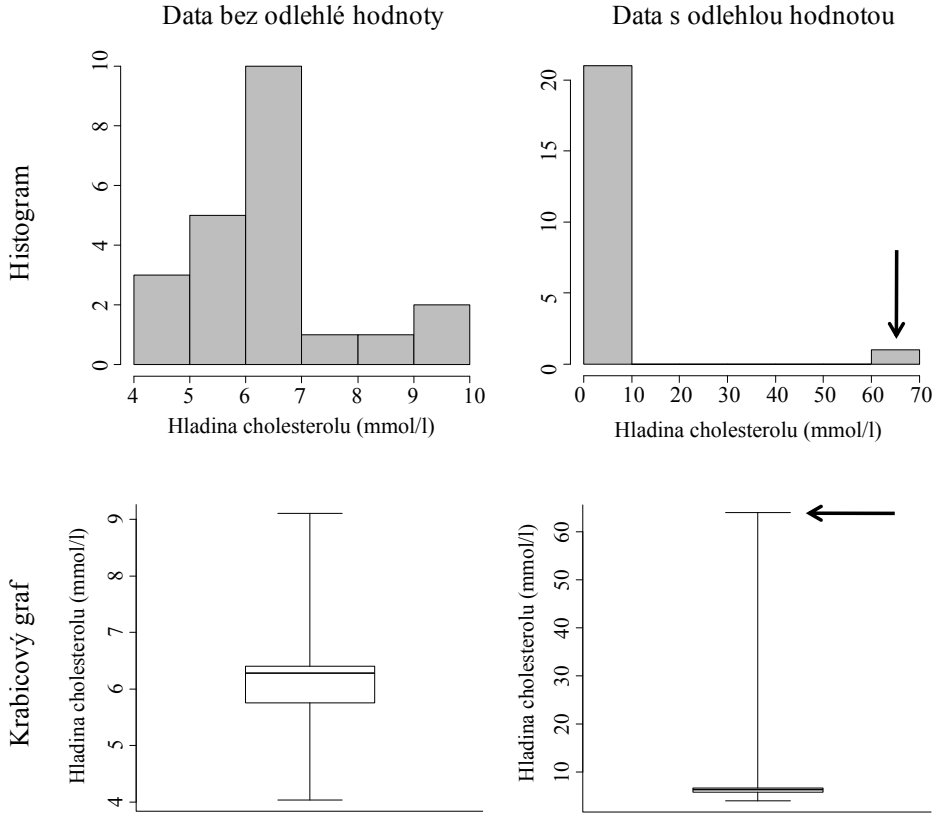
Statistika	Výpočet na datech bez odlehlé hodnoty	Výpočet na datech s odlehlou hodnotou
Průměr:	$\bar{x} = 6{,}318$	$\bar{x} = 8{,}936$
Medián:	$\tilde{x} = 6{,}275$	$\tilde{x} = 6{,}275$
Směrodatná odchylka:	$s = 1{,}337$	$s = 12{,}371$

Jak je vidět z příkladu 2.5, chybné hodnoty nebo také odlehlá pozorování mohou zásadním způsobem ovlivnit výsledky sumarizace dat, což může vést k mylné interpretaci a závěrům. Stejně tomu je i v případě pokročilejších statistických metod a modelů, kde je však naše schopnost odhalení odlehlé hodnoty na základě výsledků řádově horší než u jednoduché sumarizace. Je tak zřejmé, že problému odlehlých pozorování je nutné se věnovat ještě před zahájením jakýchkoliv výpočtů. Definice extrémních (odlehlých) hodnot není jednoduchá, neboť obor možných hodnot náhodné veličiny vždy závisí na konkrétním problému, který řešíme (v případě klinických dat je většinou dán rozmezím fyziologických hodnot). Někteří autoři definují odlehlou hodnotu jako hodnotu, která leží několikanásobek (tří, pěti, sedminásobek) výběrové směrodatné odchylky, respektive kvartilového rozpětí (často jedna a půl nebo třínásobek *IQR*), od průměru, respektive mediánu. Toto pravidlo však nelze brát striktně, neboť skutečnost, které hodnoty jsou či nejsou možné, by měl definovat hlavně zadavatel analýzy (expert na danou problematiku). Mnohem lepší je volná definice odlehlé hodnoty, která ji definuje jako netypické pozorování, které nezapadá do pravděpodobnostního chování souboru dat.

Ideálními nástroji pro identifikaci odlehlých hodnot jsou zejména výše uvedené grafy, které většinou jednoznačně odhalí problematickou hodnotu jako nezvykle vzdálenou od ostatních pozorovaných hodnot. Zajímá-li nás jedna náhodná veličina, je na místě použít histogram a krabicový graf, v případě hodnocení vztahu dvou náhodných veličin je vhodný pro identifikaci odlehlé hodnoty bodový graf. Identifikaci odlehlé hodnoty z příkladu 2.5 pomocí histogramu a krabicového grafu ukazuje obrázek 2.9.

Popisné statistiky jsou další pomůckou pro odhalování problematických hodnot, sumarizace minimálních a maximálních pozorovaných hodnot, případně 5% a 95% kvantilů, nám vždy jasně ukáže, v jakém rozsahu hodnot se náš soubor pohybuje. Na přítomnost či nepřítomnost odlehlých hodnot ukazuje i srovnání průměru a mediánu. Ve chvíli, kdy nám obě hodnoty vycházejí číselně podobně, můžeme usuzovat na nepřítomnost odlehlých hodnot, zatímco ve chvíli, kdy se hodnota průměru liší od hodnoty mediánu, svědčí to o přítomnosti odlehlých hodnot.

Je zřejmé, že zejména na větších datových souborech se nelze v identifikaci odlehlých hodnot obejít bez vizualizace a popisných statistik. Stejně tak se ale nelze obejít bez znalosti daného problému, která nám pomáhá se orientovat v tom, jaký je vůbec obor možných hodnot sledované náhodné veličiny.



Obr. 2.9 Identifikace odlehlé hodnoty pomocí histogramu a krabicového grafu.

2.4 Shrnutí

I když dle povahy rozdělujeme znaky v zásadě pouze na dvě skupiny, kvalitativní a kvantitativní, datových typů, se kterými potom při zpracování pracujeme, je více. Kromě jejich definice a toho, jak s daným datovým typem zacházet, je třeba si uvědomovat i jejich informační hodnotu, která souvisí jak s variabilitou měření, tak případně s jeho subjektivitou. Příkladem, který byl již uveden výše, je, že zatímco měření tlaku krve lze při použití adekvátního nástroje považovat za přesné, hodnocení bolestivosti rány je do značné míry závislé na povaze pacienta a jeho psychickém stavu.

Prvním krokem jakékoliv analýzy bez ohledu na typ dat by měla být sumarizace a vizualizace pozorovaných hodnot. Obojí používáme proto, že chceme do dat vidět. Nikdo pouhým pohledem neodečte potřebné informace ze souboru 1000 pacientů. Potřebujeme grafy a popisné statistiky, abychom mohli o datech vůbec komunikovat a přemýšlet. Vizualizace a výpočet popisných statistik by se tedy nikdy neměly podceňovat, i když jsou to kroky zdánlivě primitivní a zbytečné. S jejich pomocí totiž můžeme velmi snadno odhalit odlehlá pozorování a nesprávné hodnoty v datech, např. číselné překlepy v desetinných místech. Máme-li uprostřed datového souboru místo správné položky 1,21 zapsáno 121, je jistě lepší na to přijít na počátku zpracování dat než na jeho konci.

3 Náhodná veličina a její rozdělení pravděpodobnosti

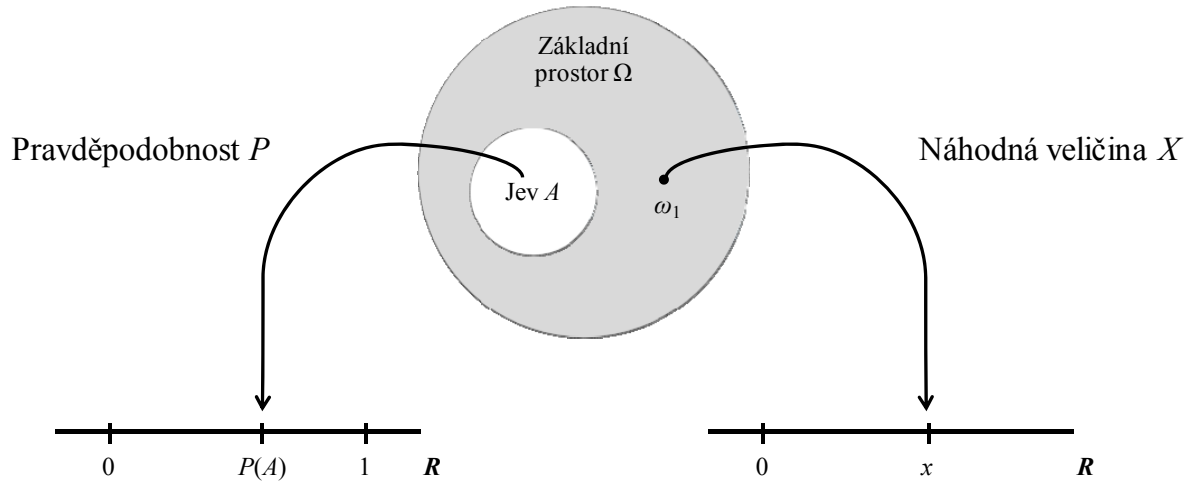
Náhodná veličina je základním konceptem matematické statistiky, která nám umožňuje pracovat pomocí statistické metodiky se znaky, které jsou v biologii, medicíně i dalších vědách předmětem našeho zájmu.

Označme Ω množinu všech možných výsledků náhodného pokusu (Ω reprezentuje základní soubor), a ω jednotlivé elementární jevy (ω_i reprezentuje i-tý prvek základního souboru). Náhodná veličina představuje číselné vyjádření výsledku náhodného pokusu, matematicky řečeno je to funkce, která každému elementárnímu jevu ω_i z Ω přiřadí hodnotu $X(\omega_i)$ z množiny možných hodnot (ta je podmnožinou množiny reálných čísel, $\boldsymbol{R}$). Matematicky zapsáno, je náhodná veličina definována jako následující funkce:

$$X : \Omega \rightarrow \boldsymbol{R} \ . \tag{3.1}$$

Formální definice náhodné veličiny je složitější (vyžaduje např. měřitelnost této funkce) a přesahuje rámec těchto skript. Úplnou definici náhodné veličiny lze nalézt např. v knize Matematická statistika od Jiřího Anděla [3].

Celý základní soubor Ω často není znám (množina Ω může být i nekonečná) a nejsme tak schopni ho popsat. Výhodou náhodné veličiny X je, že převádí základní prostor na čísla a teprve na jejich základě usuzujeme na vlastnosti Ω. Schematicky je vztah množiny všech možných výsledků náhodného pokusu, pravděpodobnosti a náhodné veličiny znázorněn na obrázku 3.1. Náhodné veličiny je zvykem označovat velkými písmeny z konce abecedy, např. X, Y, Z, jejich číselné realizace pak odpovídajícími malými písmeny, např. x, y, z.



Obr. 3.1 Schematické vyjádření konceptu náhodné veličiny.

Pravděpodobnostní chování náhodné veličiny, tedy přiřazení pravděpodobnosti každému možnému výsledku náhodné veličiny, jednoznačně popisuje tzv. ***rozdělení pravděpodobnosti***, což je předpis daný buď jako funkce zadaná analyticky, nebo jako výčet možností a jim příslušných pravděpodobností. Druhou možnost lze ilustrovat jednoduchým příkladem v podobě sledování skutečnosti, zda při hodu kostkou padne číslo 6. Náhodná veličina X pak nabývá hodnot 1 (číslo 6 padlo, pravděpodobnost je rovna 1/6) nebo 0 (číslo 6 nepadlo, pravděpodobnost je rovna 5/6). Je tedy zřejmé, že náhodná veličina se netýká pouze

kvantitativních znaků, neboť číselné vyjádření výsledku náhodného pokusu může popisovat i pohlaví.

Rozdělení pravděpodobnosti představuje model chování náhodné veličiny v cílové populaci. Pomocí vzorku (naměřených pozorování) se ptáme, jestli je model správný a jaké jsou jeho charakteristiky. Rozdělení pravděpodobnosti náhodné veličiny lze jednoznačně popsat pomocí tzv. ***distribuční funkce*** (*cumulative distribution function*), kterou standardně značíme $F(x)$. Distribuční funkce vyjadřuje pravděpodobnost, že číselná realizace náhodné veličiny X nepřekročí na reálné ose danou hodnotu x, což lze zapsat jako

$$F(x) = P(X \le x) = P(\omega_i \in \Omega : X(\omega_i) \le x). \tag{3.2}$$

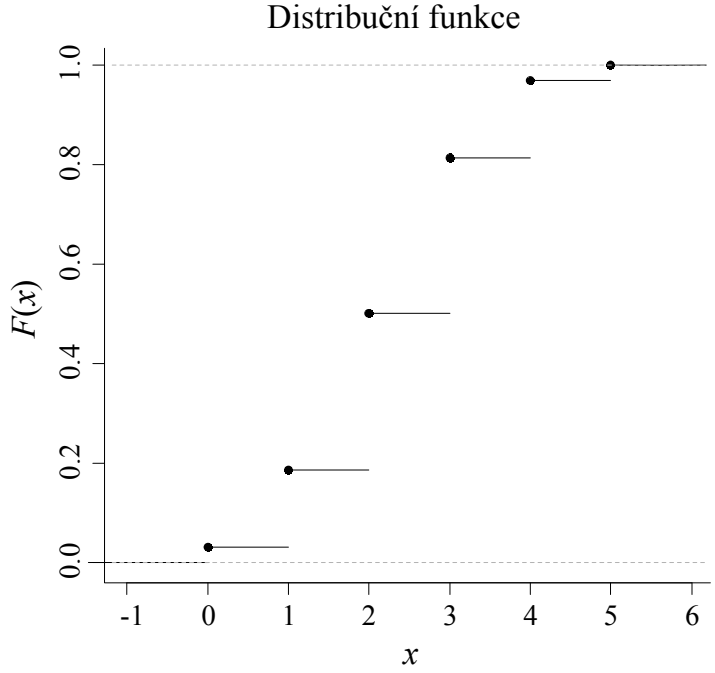
Distribuční funkce má několik vlastností, které plynou z toho, že je definována jako pravděpodobnost:

1. $F(x)$ je neklesající a zprava spojitá.
2. Platí, že $0 \le F(x) \le 1$.
3. Platí, že $F(x) \to 0$ pro $x \to -\infty$ a $F(x) \to 1$ pro $x \to \infty$.

Příklad 3.1. Uvažujme 5 hodů mincí. Náhodná veličina X představuje počet líců a může nabývat pouze hodnot z množiny $\{0, 1, 2, 3, 4, 5\}$. Pro úplnost dodejme, že množina Ω je v tomto případě množina všech uspořádaných pětic složených z nul a jedniček reprezentujících ruby, respektive líce. Pravděpodobnosti jednotlivých realizací náhodné veličiny X lze spočítat jednoduše pomocí kombinatoriky: $P(X = 0) = 1/32$, $P(X = 1) = 5/32$, $P(X = 2) = 10/32$, $P(X = 3) = 10/32$, $P(X = 4) = 5/32$, $P(X = 5) = 1/32$. Distribuční funkce náhodné veličiny X je pak schodovitá funkce daná tabulkou 3.1. Graficky je distribuční funkce náhodné veličiny X znázorněna na obrázku 3.2.

Tabulka 3.1 *Hodnoty distribuční funkce náhodné veličiny X udávají počet líců v pěti hodech mincí.*

x	$x < 0$	$x \in \langle 0,1)$	$x \in \langle 1,2)$	$x \in \langle 2,3)$	$x \in \langle 3,4)$	$x \in \langle 4,5)$	$x \ge 5$
$F(x)$	0	1/32	6/32	16/32	26/32	31/32	1



Obr. 3.2 Distribuční funkce náhodné veličiny X udávají počet líců v pěti hodech mincí.

Distribuční funkce je teoretický předpis, který sice definuje pravděpodobnostní model pro náhodnou veličinu X, ale v řadě případů neznáme jeho přesné vyjádření. Jejím výběrovým ekvivalentem, který kumulativním způsobem popisuje pravděpodobnostní chování pozorovaných hodnot je tzv. ***výběrová*** (***empirická***) ***distribuční funkce***, $F_n(x)$, která je definována následovně:

$$F_n(x) = \frac{\#(x_i \le x)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)\,, \tag{3.3}$$

kde symbol # vyjadřuje počet a I je indikátorová funkce nabývající hodnoty 1, když je podmínka v argumentu funkce splněna, a hodnoty 0, pokud podmínka v závorce splněna není.

Výběrová distribuční funkce je při splnění předpokladu reprezentativnosti experimentálního vzorku odhadem teoretické distribuční funkce, což znamená, že z jejích hodnot a grafického znázornění můžeme usuzovat na vlastnosti teoretické distribuční funkce. Distribuční funkce jednoznačně přiřazuje každému číslu x na reálné ose pravděpodobnost, když odpovídá na otázku, s jakou pravděpodobností náhodná veličina X právě toto x nepřekročí. Často nás zajímá ale i opačná úvaha, tedy odpověď na otázku, jaké číslo x na reálné ose nepřekročí náhodná veličina X s určitou pravděpodobností (označme ji p), což může být např. číslo p = 0,8, 0,9 nebo 0,95. Odpověď na tuto otázku dává tzv. ***kvantilová funkce***, což je funkce inverzní k distribuční funkci, jejímž výsledkem není pravděpodobnost, ale právě číslo na reálné ose, které této pravděpodobnosti p odpovídá. Rozdíl mezi distribuční funkcí a kvantilovou funkcí ukazují vztahy 3.4 a 3.5.

Distribuční funkce: $$F(x_p) = P(X \le x_p) = p \tag{3.4}$$

Kvantilová funkce: $$x_p = F^{-1}(P(X \le x_p)) = F^{-1}(p) \qquad (3.5)$$

Kvantilová funkce úzce souvisí s pojmem kvantil, který byl vysvětlen v kapitole 2, ale zatímco tam byl kvantil zaveden jako jedna z pozorovaných hodnot s určitou vlastností (*p*-procentní kvantil rozděluje data na *p* procent hodnot a (100 – *p*) procent hodnot), zde se jedná o teoretickou funkci, která je charakteristikou rozdělení náhodné veličiny *X*.

3.1 Spojité a diskrétní náhodné veličiny

Náhodné veličiny dělíme dle množiny hodnot, kterých mohou nabývat, na spojité a diskrétní. ***Diskrétní náhodná veličina*** může nabýt nejvýše spočetně mnoha hodnot (představovaných izolovanými body na reálné ose), zatímco ***spojitá náhodná veličina*** může nabýt všech hodnot v určitém intervalu (tedy může nabýt nespočetně mnoha hodnot). Medicínským příkladem spojité náhodné veličiny je např. výška osoby, váha osoby, krevní tlak pacienta, koncentrace glukózy v krvi (glykémie) nebo čas do výskytu sledované události; příkladem z oblasti biologie pak biomasa na m^2, listová plocha, pH, koncentrace toxických látek ve vodě nebo v ovzduší apod. Jako příklad diskrétní náhodné veličiny z oblasti medicíny můžeme uvést počet krvácivých epizod u pacienta za rok, počet opakovaných hospitalizací, počet dní po operaci do odeznění bolesti; z oblasti biologie pak např. počet zvířecích druhů na jednotku plochy nebo objemu, počet bakteriálních kolonií na experimentální misku, apod. Diskrétní náhodná veličina má distribuční funkci schodovitého tvaru, zatímco spojitá náhodná veličina má spojitou distribuční funkci.

Zatímco distribuční funkce popisující rozdělení pravděpodobnosti náhodné veličiny kumulativním způsobem je charakteristika společná pro spojité i diskrétní náhodné veličiny, ve chvíli, kdy chceme popsat rozdělení pravděpodobnosti pro jednotlivé hodnoty, respektive intervaly na reálné ose, musíme definovat příslušnou funkci zvlášť pro spojité a zvlášť pro diskrétní náhodné veličiny. Spojitou náhodnou veličinu *X* s distribuční funkcí *F*(*x*) charakterizuje tzv. ***hustota pravděpodobnosti*** (*probability density function*), což je funkce *f*(*x*) taková, že platí:

$$F(x) = \int_{-\infty}^{x} f(x)dt \,. \qquad (3.6)$$

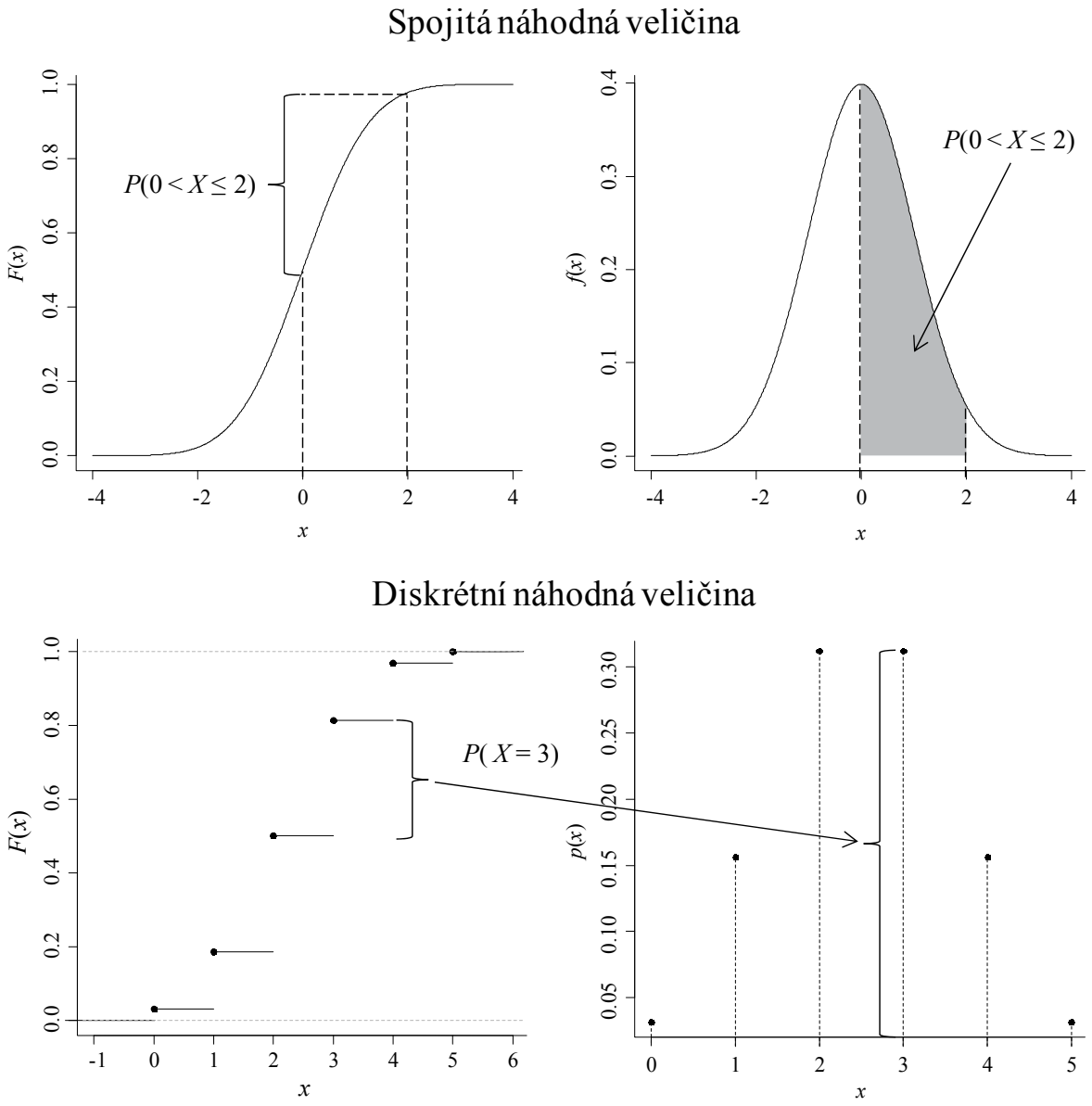
To znamená, že distribuční funkce spojité náhodné veličiny geometricky znamená plochu pod grafem hustoty pravděpodobnosti *f*(*x*), viz obrázek 3.3. Jako přímý důsledek lze hustotu pravděpodobnosti získat derivací distribuční funkce, tedy $f(x) = \mathrm{d}F(x)/\mathrm{d}x$.

Diskrétní náhodnou veličinu *X* s distribuční funkcí *F*(*x*) charakterizuje tzv. ***pravděpodobnostní funkce*** (*probability mass function*), což je funkce *p*(*x*) taková, že platí:

$$F(x) = \sum_{t \le x} p(t) = \sum_{t \le x} P(X = t) \,. \qquad (3.7)$$

Pravděpodobnostní funkce vyjadřuje vztah $p(x) = P(X = x)$, což geometricky odráží skutečnost, že hodnota *p*(*x*) je rovna výšce skoku (hrany schodu) v bodě *x* na grafu schodovité distribuční funkce (viz obrázek 3.3).

Ze vztahů 3.6 a 3.7 je vidět, že distribuční funkce a hustota, respektive distribuční funkce a pravděpodobnostní funkce, jsou navzájem ekvivalentní, tedy známe-li jednu, můžeme dopočítat druhou.



Obr. 3.3 Vztah distribuční funkce náhodné veličiny s hustotou a pravděpodobnostní funkcí.

3.2 Charakteristiky náhodných veličin

Výše definovaný popis pravděpodobnostního chování náhodné veličiny pomocí distribuční funkce, hustoty a pravděpodobnostní funkce je sice úplný, ale trochu složitý a velmi nepraktický. Často se tak pro popis jednotlivých rozdělení pravděpodobnosti používají číselné charakteristiky, které shrnují vlastnosti rozdělení pravděpodobnosti do jednoho čísla, které je snadno interpretovatelné a lze s ním pracovat jednodušeji než s funkčním vyjádřením. Dvě nejznámější a nejpoužívanější charakteristiky, které odráží vlastnosti rozdělení pravděpodobnosti náhodné veličiny, jsou ***střední hodnota*** (*mean value*) a ***rozptyl*** (*dispersion, variance*). Střední hodnota náhodné veličiny X, značíme ji $E(X)$, je mírou polohy a popisuje

tak oblast reálné osy, kde má náhodná veličina X „tendenci“ se realizovat, zatímco rozptyl náhodné veličiny X, značíme ho $D(X)$, je mírou variability, který ukazuje, jak moc jednotlivé možné hodnoty náhodné veličiny X kolísají kolem její střední hodnoty.

Vzhledem k tomu, že střední hodnota i rozptyl charakterizují rozdělení pravděpodobnosti, není překvapivé, že jsou definovány pomocí odpovídajících funkcí, tedy střední hodnota spojité náhodné veličiny X s hustotou $f(x)$ je definována jako integrál

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \,, \tag{3.8}$$

zatímco střední hodnota diskrétní náhodné veličiny X s pravděpodobnostní funkcí $p(x)$ a oborem možných hodnot R je definována jako suma

$$E(X) = \sum_{x \in R} x p(x) \,. \tag{3.9}$$

Výraz pro výpočet střední hodnoty může vypadat složitě, ale nejedná se o nic jiného než o formu váženého průměru, kde jednotlivé možné hodnoty, x, jsou váženy jejich pravděpodobností výskytu, $p(x)$. Jinak řečeno, reálné hodnoty s větší pravděpodobností výskytu v rámci realizace náhodné veličiny X mají větší vliv na její výslednou střední hodnotu než hodnoty s menší pravděpodobností výskytu.

Rozptyl náhodné veličiny X, $D(X)$, je definován stejně pro spojitou i diskrétní náhodnou veličinu, a to jako střední hodnota kvadrátu odchylky náhodné veličiny od její střední hodnoty:

$$D(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2 \,, \tag{3.10}$$

kde výraz $E(X^2)$ představuje střední hodnotu transformované náhodné veličiny X^2 [3]. Stejně jako v případě výběrového rozptylu není ani rozptyl náhodné veličiny v týchž jednotkách jako střední hodnota a hodnoty náhodné veličiny, a proto se jako míra variability používá spíše jeho odmocnina, tzv. ***směrodatná odchylka*** (*standard deviation*) náhodné veličiny, kterou značíme $SD(X)$:

$$SD(X) = \sqrt{D(X)} \,. \tag{3.11}$$

Střední hodnota a rozptyl náhodné veličiny představují teoretický ekvivalent (ve smyslu pravděpodobnosti) popisných ukazatelů, které nás zajímaly u pozorovaných dat, tedy střední hodnota, $E(X)$, je teoretickým ekvivalentem průměru a rozptyl, $D(X)$, je teoretickým ekvivalentem výběrového rozptylu (viz část 2.2.2). Střední hodnota a rozptyl náhodné veličiny X představují klíčové parametry jejího rozdělení pravděpodobnosti a při statistickém zpracování dat jsou většinou hlavním předmětem našeho zájmu. U spojitých náhodných veličin mají výše definované charakteristiky většinou jasnou interpretaci, v případě diskrétních náhodných veličin však mohou být i lehce zavádějící, neboť diskrétní náhodná veličina vůbec nemusí nabývat své střední hodnoty. Jako příklad lze uvést náhodnou veličinu X, která nabývá hodnot -1 a 1, obou s pravděpodobností 0,5. Je zřejmé, že její střední hodnota je 0, což je ale hodnota, které tato náhodná veličina nikdy nemůže nabývat.

3.3 Shrnutí

V této kapitole jsme definovali náhodnou veličinu jako základní myšlenku statistiky, která nám umožňuje práci s pozorovanými hodnotami sledovaných znaků. Chování náhodné veličiny z hlediska pravděpodobnosti popisujeme pomocí jejího rozdělení pravděpodobnosti, které představuje model chování náhodné veličiny v cílové populaci. Rozdělení náhodné veličiny je nejčastěji charakterizováno pomocí distribuční funkce nebo hustoty pravděpodobnosti v případě spojité náhodné veličiny, respektive pravděpodobnostní funkce v případě diskrétní náhodné veličiny. V praxi nás však nejvíce zajímají číselné charakteristiky náhodných veličin, střední hodnota a rozptyl, které jsou při statistickém zpracování dat většinou hlavním předmětem našeho zájmu.

Tyto funkce a číselné charakteristiky lze použít jak pro popis vlastností cílové populace, tak pro predikci budoucího chování náhodné veličiny. Na základě pozorovaných dat jsme totiž schopni pomocí dostupných nástrojů (histogram, box plot, popisné statistiky) usuzovat na charakter rozdělení pravděpodobnosti sledované veličiny a jsme schopni otestovat míru shody pozorovaných hodnot s teoretickým rozdělením. Za předpokladu shody se můžeme následně ptát, jaká je pravděpodobnost, že se sledovaná náhodná veličina realizuje v nějakém konkrétním intervalu hodnot. Dále můžeme např. provádět srovnání sledované vlastnosti v rámci jedné nebo více cílových populací. Na základě pozorovaných dat a našich předpokladů o teoretickém pravděpodobnostním modelu (hypotéz) jsme totiž schopni pomocí statistických testů srovnávat charakteristiky dané náhodné veličiny v rámci jedné nebo více zkoumaných skupin subjektů/objektů.

4 Vybraná rozdělení pravděpodobnosti

Znalost rozdělení pravděpodobnosti, kterým se řídí námi studovaná náhodná veličina, není nezbytná. Existují totiž statistické metody, standardně jsou označovány jako ***neparametrické metody***, které nevyžadují specifikaci konkrétního rozdělení pozorovaných hodnot. Na druhou stranu, tato znalost, kterou vyžadují tzv. ***parametrické metody***, je vždy výhodná, neboť použití parametrických metod je většinou jednodušší a při korektně specifikovaném rozdělení i přesnější. Podmínka korektní specifikace je však nesmírně důležitá. Pokud totiž předpokládáme pravděpodobnostní chování studované cílové populace dle určitého rozdělení, ale ve skutečnosti tento předpoklad splněn není, je špatně specifikace celého statistického modelu, což vede k zavádějícím výsledkům a neinterpretovatelným závěrům.

Neparametrické metody získaly v biologické a klinické komunitě na popularitě zejména díky rozvoji moderních genetických a molekulárně-biologických metod zaměřených na monitorování hladin a intenzit nejrůznějšího původu. V těchto případech je znalost konkrétních rozdělení velmi omezená, což nahrává použití neparametrických metod. I přes rostoucí popularitu neparametrické statistiky však parametrické metody zůstávají dominantní oblastí statistiky, z čehož samozřejmě vyplývá i význam jednotlivých rozdělení. V další části této kapitoly budou představena rozdělení pravděpodobnosti, která hrají důležitou roli jak v praktické, tak v teoretické biostatistice [3, 37, 38].

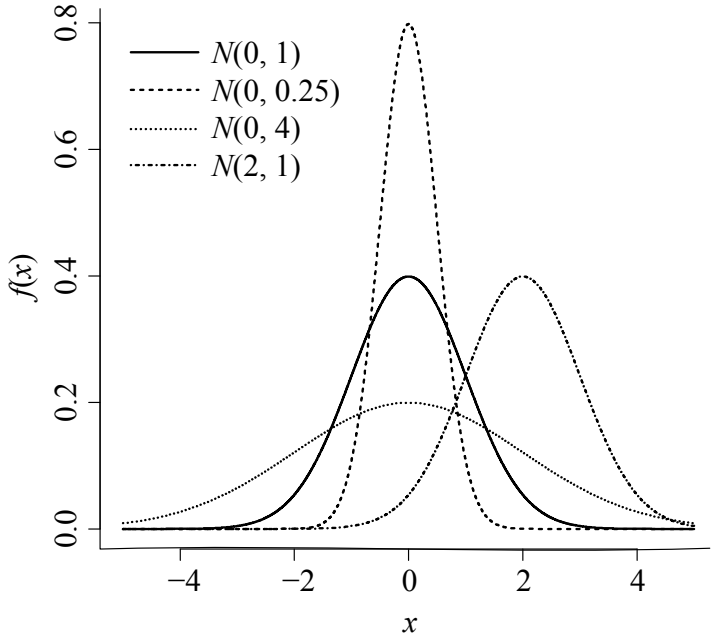
4.1 Normální rozdělení

Normální rozdělení je spojité rozdělení pravděpodobnosti, které popisuje celou řadu veličin, jejichž hodnoty se symetricky shlukují kolem střední hodnoty a vytvářejí tak charakteristický tvar hustoty pravděpodobnosti, která je známá také pod pojmem ***Gaussova křivka***. Tento zvonovitý tvar souvisí s faktem, že variabilita normálního rozdělení kolem jeho střední hodnoty je dána aditivním vlivem mnoha tzv. „slabě působících“ faktorů, což znamená, že se s normálním rozdělením setkáváme u řady biologických a klinických znaků, např. výšky člověka, délky končetin a kostí, maximální dosažené rychlosti ještěrky, apod. Je třeba poznamenat, že označení normální rozdělení neznamená, že by toto rozdělení pravděpodobnosti bylo v přírodě normálnější než rozdělení jiná, na rozdíl od ostatních rozdělení má však normální rozdělení stěžejní význam v teoretické statistice.

Normální rozdělení pravděpodobnosti je zcela popsáno dvěma parametry, které jsou standardně označovány jako μ a σ^2, kdy první z nich představuje střední hodnotu normálního rozdělení a druhý představuje rozptyl normálního rozdělení. Fakt, že náhodná veličina X má normální rozdělení pravděpodobnosti se střední hodnotou μ a rozptylem σ^2, zapisujeme jako $X \sim N(\mu,\sigma^2)$. Hustota náhodné veličiny X pak má následující tvar:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{4.1}$$

Ukázky hustot náhodných veličin s normálním rozdělením pro různé hodnoty parametrů μ a σ^2 jsou uvedeny na obrázku 4.1.



Obr. 4.1 Ukázky hustot náhodných veličin s normálním rozdělením.

Jak již bylo uvedeno, normální rozdělení pravděpodobnosti je pro biostatistiku důležité, neboť je klíčovým předpokladem řady základních testů a modelů. Rozhodneme-li se použít pro zpracování dat metodu založenou na předpokladu, že data pocházejí z normálního rozdělení, je ověření tohoto předpokladu stejně důležité jako výběr samotného testu. Pro ověření normality existuje řada testů a grafických metod, některé z nich budou rozebrány v kapitole 8 v souvislosti s analýzou rozptylu, což je právě jedna z metod postavených na předpokladu normálního rozdělení. Nicméně stoprocentně ověřit to, zda se sledovaná veličina (znak) chová dle normálního rozdělení je prakticky nemožné, vždy jsme totiž limitováni množstvím a kvalitou pozorovaných dat. Uvažujme soubor 22 pozorování z příkladu 2.3, která jsou zobrazena pomocí histogramu a krabicového grafu na obrázku 2.9. Stěží lze rozhodnout, zda se jedná nebo nejedná o hodnoty pocházející z normálního rozdělení; na takto malém datovém souboru je jakákoliv vizualizace i testování velmi obtížné. V praxi tak normalitu vlastně nepotvrzujeme, spíše připouštíme, že se pozorované hodnoty neodchylují až příliš.

Vlastností normálního rozdělení s velkým praktickým významem je, že jsme u něj schopni vyčíslit procento pozorování, která by se měla realizovat v rozmezí ± x-násobku směrodatné odchylky σ od střední hodnoty μ. Zmíněná procenta pro jedno, dvou a třínásobek směrodatné odchylky uvádí tabulka 4.1.

Tabulka 4.1 *Pravděpodobnost realizace normální náhodné veličiny v rozmezí ± x-násobku směrodatné odchylky σ od střední hodnoty μ.*

Interval	Pravděpodobnost realizace uvnitř intervalu	Pravděpodobnost realizace vně intervalu
$\mu \pm 1\sigma$	0,683	0,317
$\mu \pm 2\sigma$	0,954	0,046
$\mu \pm 3\sigma$	0,997	0,003

Z tabulky 4.1 vyplývá, že pravděpodobnost realizace náhodné veličiny s normálním rozdělením v intervalu $\mu \pm 2\sigma$ je lehce přes 95%. Jinými slovy, zhruba 95% pozorování náhodné veličiny s normálním rozdělením by se mělo realizovat v rozmezí $\mu - 2\sigma$ a $\mu + 2\sigma$. Uvážíme-li dokonce interval $\mu \pm 3\sigma$, pak je pravděpodobnost realizace náhodné veličiny uvnitř tohoto intervalu dokonce více než 99,5%. Tato jednoduchá poučka je někdy označována jako pravidlo ± 3 sigma.

Další důležitou vlastností normálního rozdělení, která má praktické využití, je, že při sčítání dvou a více náhodných veličin se zachovává normalita. Jinými slovy, pro nezávislé náhodné veličiny X a Y platí, že i jejich součet, tedy náhodná veličina $Z = X + Y$, má normální rozdělení pravděpodobnosti.

4.2 Standardizované normální rozdělení

Mezi výhodné vlastnosti normálního rozdělení patří zachování normality při změně měřítka osy, na které měříme jednotky náhodné veličiny X. Jinými slovy, pokud veličinu X s rozdělením $N(\mu,\sigma^2)$ transformujeme podle vztahu $Y = a + bX$, pak platí, že náhodná veličina Y má rozdělení pravděpodobnosti $N(a + b\mu, b^2\sigma^2)$. S využitím této vlastnosti jsme vždy schopni transformovat náhodnou veličinu X s rozdělením $N(\mu,\sigma^2)$ na náhodnou veličinu Z s rozdělením $N(0,1)$, tedy s normálním rozdělením s nulovou střední hodnotou a jednotkovým rozptylem. Platí

$$X \sim N(\mu,\sigma^2) \rightarrow Z = \frac{X-\mu}{\sqrt{\sigma^2}} \sim N(0,1)\,. \tag{4.2}$$

Toto rozdělení má ve statistice výsadní postavení a označuje se jako ***standardizované normální rozdělení*** (***normované normální rozdělení***). Jeho hustota pak má následující tvar:

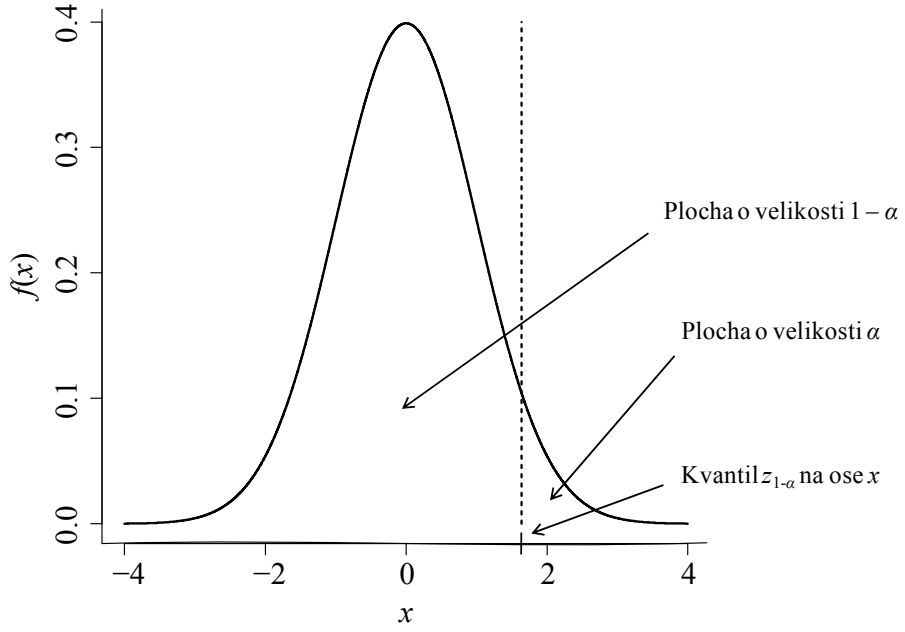
$$f(x;0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,. \tag{4.3}$$

Výhoda je, že všechny hodnoty distribuční i kvantilové funkce jsou tabelovány a obsaženy v dostupných softwarech (kvantily standardizovaného normálního rozdělení se označují jako z). Můžeme tak jednoduše kvantifikovat pravděpodobnost, s jakou se náhodná veličina Z se standardizovaným normálním rozdělením realizuje nad určitou hodnotou z (případně pod ní, nebo mezi dvěma danými hodnotami). Obecně lze plochu pod hustotou rozdělit pomocí kvantilu na dvě části, např. pomocí 100(1 – α)procentního kvantilu, označme ho $z_{1-\alpha}$, na část s plochou $1 - \alpha$ a na část s plochou α (viz obrázek 4.2). Toto dělení samozřejmě odpovídá pravděpodobnosti, tedy náhodná veličina Z se realizuje číslem menším než $z_{1-\alpha}$ s pravděpodobností $1 - \alpha$ a číslem větším než $z_{1-\alpha}$ s pravděpodobností α.

Příklad 4.1. Při populačním epidemiologickém průzkumu se zjistilo, že průměrný objem prostaty u mužů (veličina X) je 52,73 ml se směrodatnou odchylkou rovnou 13,12 ml. Předpokládáme, že objem prostaty se řídí normálním rozdělením, za hodnoty parametrů μ a σ^2 bereme populační odhady. Zajímá nás, jaká je pravděpodobnost, že objem prostaty u muže bude větší než 80 ml. Abychom zjistili, jaká pravděpodobnost přísluší hodnotě 80 ml jako kvantilu rozdělení náhodné veličiny X, provedeme standardizaci a zjistíme příslušnou pravděpodobnost na základě kvantilu standardizované normální veličiny Z. Výpočet hodnoty veličiny Z je následující:

$$Z = \frac{X - \mu}{\sigma} = \frac{80 - 52{,}73}{13{,}12} = 2{,}08\,. \tag{4.4}$$

Víme, že hodnota 2,08 představuje 100(1 – α)procentní kvantil, $z_{1-\alpha}$, standardizované normální veličiny Z, k ní odpovídající hladinu α zjistíme z tabulek hodnot kvantilové funkce. Lze zjistit, že pravděpodobnost výskytu hodnoty větší než 2,08 je pro standardizovanou normální veličinu rovna 0,0188, což tedy znamená, že pravděpodobnost výskytu prostaty s objemem větším než 80 ml je rovna přibližně 2%.



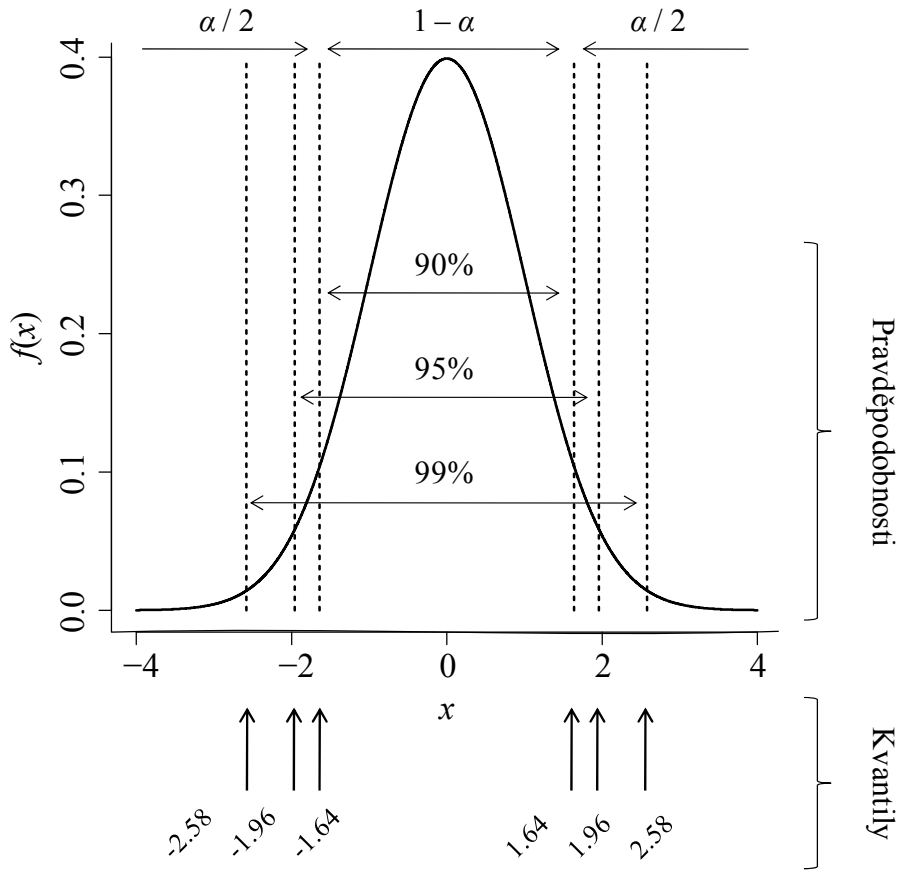
Obr. 4.2 Plochy pod hustotou pravděpodobnosti příslušné kvantilu $z_{1-\alpha}$.

Oblast, kde se náhodná veličina Z se standardizovaným normálním rozdělením realizuje s pravděpodobností 1 – α lze vyjádřit pomocí její distribuční funkce (ta vyjadřuje pravděpodobnost, že číselná realizace náhodné veličiny nepřekročí na reálné ose danou hodnotu) a příslušných kvantilů následujícím způsobem:

$$1-\alpha = 1-\tfrac{\alpha}{2}-\tfrac{\alpha}{2} = \left(1-\tfrac{\alpha}{2}\right)-\tfrac{\alpha}{2} = F_{N(0,1)}(z_{1-\alpha/2}) - F_{N(0,1)}(z_{\alpha/2}) = P(z_{\alpha/2} \le Z \le z_{1-\alpha/2})\,. \tag{4.5}$$

Jinými slovy, oblast realizace náhodné veličiny Z s rozdělením $N(0,1)$ odpovídající pravděpodobnosti 1 – α lze vymezit pomocí jejích kvantilů, jmenovitě pomocí 100(α/2)procentního kvantilu, $z_{\alpha/2}$, a 100(1 – α/2)procentního kvantilu, $z_{1-\alpha/2}$. Vzhledem k symetrii hustoty standardizovaného normálního rozdělení jsou vždy tyto dva kvantily identické až na znaménko, tedy platí $z_{1-\alpha/2} = -z_{\alpha/2}$.

Klíčové kvantily standardizovaného normálního rozdělení uvádí obrázek 4.3, ze kterého vyplývá, že náhodná veličina s rozdělením $N(0,1)$ se s pravděpodobností 90% realizuje mezi hodnotou -1,64 a hodnotou 1,64, s pravděpodobností 95% mezi hodnotami -1,96 a 1,96 a s pravděpodobností 99% nepřekročí v absolutní hodnotě číslo 2,58.



Obr. 4.3 Klíčové kvantily standardizovaného normálního rozdělení pravděpodobnosti.

Vymezení oblasti, kde se náhodná veličina realizuje s určitou pravděpodobností je platné pro všechna rozdělení pravděpodobnosti, nejen pro standardizované normální (i když u rozdělení $N(0,1)$ se vzhledem k jeho symetrii významné kvantily dobře pamatují). Tento fakt je velmi důležitý zejména v testování hypotéz (viz kapitola 6), kde na základě toho, v jaké oblasti se realizuje hodnota testové statistiky (náhodné veličiny s daným rozdělením pravděpodobnosti), rozhodujeme o platnosti nebo neplatnosti sledované hypotézy.

4.3 Další rozdělení pravděpodobnosti

V další části této kapitoly budou představena rozdělení pravděpodobnosti, která jsou buď důležitá pro pochopení dále uvedené statistické metodiky (zejména pro proces testování hypotéz) nebo jsou to rozdělení často charakterizující znaky a jevy v přírodě.

4.3.1 Rovnoměrně spojité rozdělení – *Rs*(*a*,*b*)

Rovnoměrně spojité rozdělení je základní spojité rozdělení pravděpodobnosti, pro které platí, že jeho hustota pravděpodobnosti je na intervalu (a, b) konstantní a mimo tento interval nulová. Pro x z intervalu (a, b), kde $a < b$, má hustota pravděpodobnosti rovnoměrně spojitého rozdělení tvar

$$f(x) = \frac{1}{b-a}. \tag{4.6}$$

Lze ukázat (laskavému čtenáři necháváme jako cvičení), že střední hodnota náhodné veličiny s rovnoměrně spojitým rozdělením, $E(X)$, je rovna $(a + b)/2$, její teoretický rozptyl, $D(X)$, je pak roven $(b - a)^2/12$.

4.3.2 Chí-kvadrát rozdělení – $\chi^2(k)$

Chí-kvadrát rozdělení je spojité rozdělení pravděpodobnosti s velkým významem v teoretické statistice. Využíváme ho při konstrukci intervalu spolehlivosti pro rozptyl náhodné veličiny (viz kapitola 5) a je to modelové rozdělení pravděpodobnosti testové statistiky při testování hypotéz o nezávislosti kvalitativních dat a testech dobré shody (viz kapitola 9). Chí-kvadrát rozdělení vzniká jako součet druhých mocnin k nezávislých náhodných veličin se standardizovaným normálním rozdělením, $N(0,1)$, jedná se tedy o rozdělení odvozené z normálního. Platí následující:

$$Z_i \sim N(0,1) \rightarrow K = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k) \tag{4.7}$$

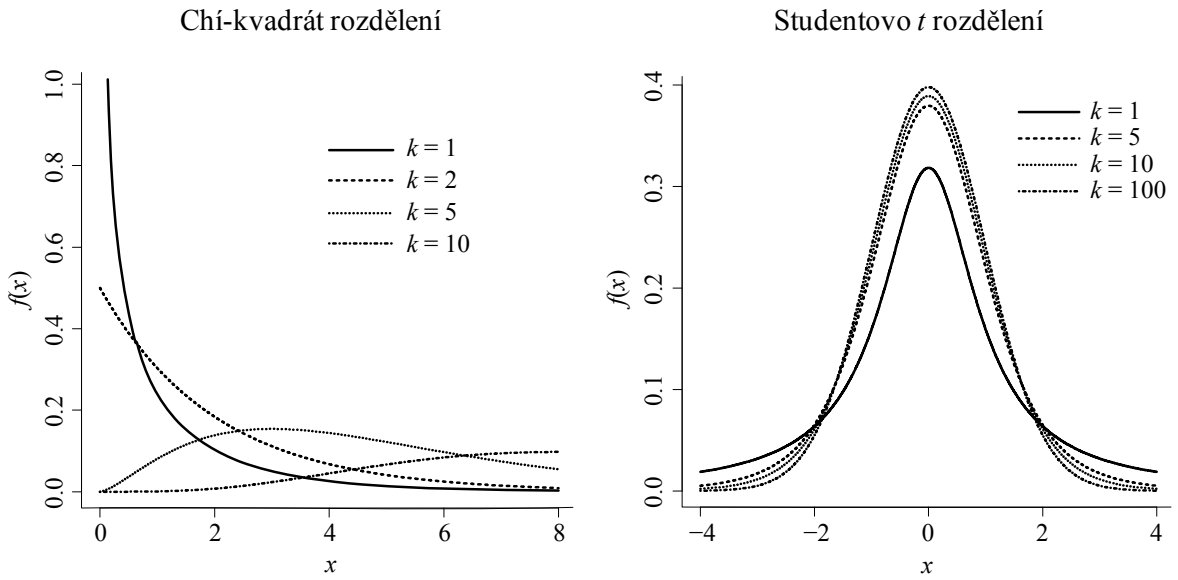
Konstanta k, která je jediným parametrem tohoto rozdělení, je standardně nazývána ***počet stupňů volnosti***. Ze vztahu (4.7) je zřejmé, že náhodná veličina K nabývá pouze nezáporných hodnot. Hustoty chí-kvadrát rozdělení pro čtyři různé hodnoty parametru k jsou zobrazeny na obrázku 4.4 vlevo.

4.3.3 Studentovo *t* rozdělení – *t*(*k*)

Studentovo t rozdělení je také spojité rozdělení pravděpodobnosti, které stejně jako v předchozím případě nachází spíše uplatnění v teoretické statistice než v přírodě [29]. Toto rozdělení charakterizuje rozdělení pravděpodobnosti průměru jako odhadu střední hodnoty veličiny s normálním rozdělením v případě, že neznáme přesnou hodnotu rozptylu (což je v praktickém životě téměř vždy). Studentovo t rozdělení vzniká jako podíl dvou nezávislých náhodných veličin, jedné s rozdělením $N(0,1)$ a druhé s rozdělením $\chi^2(k)$. Platí tedy:

$$Z \sim N(0,1), K \sim \chi^2(k) \rightarrow T = \frac{Z}{\sqrt{K/k}} \sim t(k) \tag{4.8}$$

Parametrem Studentova t rozdělení je opět počet stupňů volnosti k, který přebírá od rozdělení chí-kvadrát. Studentovo rozdělení lze také chápat jako aproximaci standardizovaného normálního rozdělení pro malé výběrové soubory (tomu odpovídá malá hodnota parametru k), s rostoucí velikostí souboru (s rostoucím parametrem k) se hustota Studentova rozdělení (a tedy i kvantily) přibližuje hustotě normálního rozdělení. Srovnáním obrázku 4.4 vpravo s obrázkem 4.3 lze zjistit, že již pro $k = 100$ je hustota Studentova t rozdělení téměř shodná s hustotou standardizovaného normálního rozdělení.



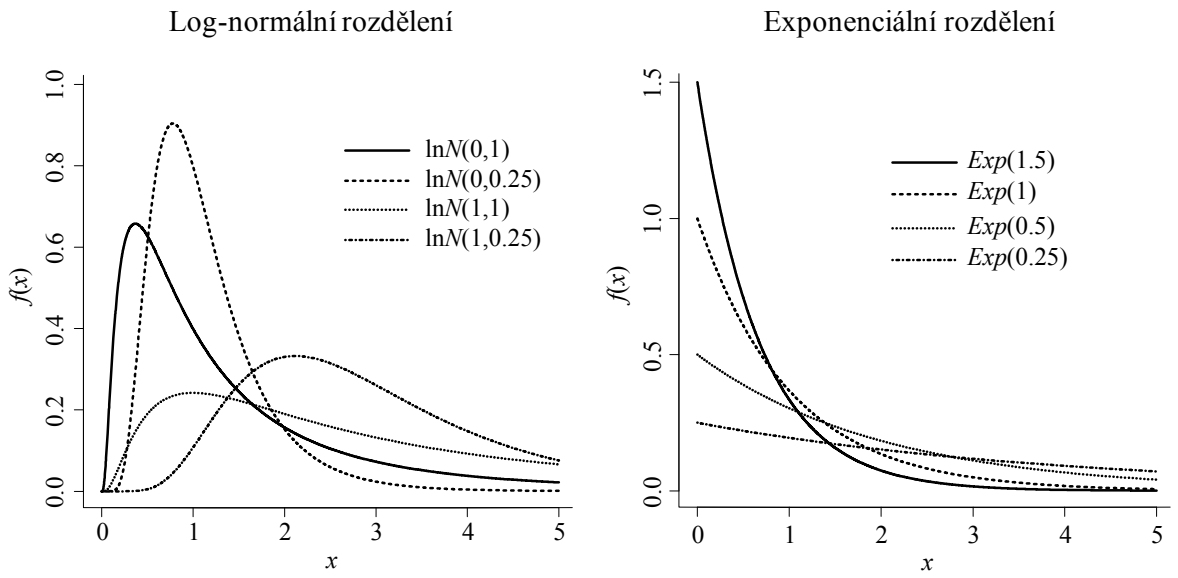
Obr. 4.4 Ukázky hustot náhodných veličin s chí-kvadrát rozdělením a Studentovým t rozdělením.

4.3.4 Logaritmicko-normální rozdělení – $lnN(\mu,\sigma^2)$

S ***logaritmicko-normálním*** (***log-normálním***) ***rozdělením*** se na rozdíl od dvou předchozích rozdělení můžeme relativně často setkat v přírodě (respektive jak v biologii, tak v medicíně). Logaritmicko-normální rozdělení má např. tělesná hmotnost, délka inkubační doby infekčního onemocnění nebo abundance živočišných druhů. V neposlední řadě toto rozdělení charakterizuje i řadu krevních parametrů (např. počet krevních buněk v daném objemu, sérový bilirubin u pacientů s cirhózou). Náhodná veličina X má logaritmicko-normální rozdělení právě tehdy, když veličina $Y = \ln(X)$ má normální rozdělení (výraz ln zde zastupuje přirozený logaritmus). A to samé platí i naopak, když veličina Y má normální rozdělení, pak náhodná veličina $X = \exp(Y)$ má rozdělení logaritmicko-normální. Hustota je dána vztahem

$$f(x;\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\ln x-\mu)^2/2\sigma^2}, \qquad (4.9)$$

kde parametry μ a σ^2 mají význam střední hodnoty a rozptylu normálního rozdělení odpovídající náhodné veličiny $Y = \ln(X)$. Ukázky hustot logaritmicko-normálního rozdělení pro čtyři různé kombinace parametrů μ a σ^2 jsou zobrazeny na obrázku 4.5 vlevo. Logaritmicko-normální náhodná veličina X opět nabývá pouze nezáporných hodnot, platí tedy $f(x) = 0$ pro $x < 0$.



Obr. 4.5 Ukázky hustot náhodných veličin s log-normálním a exponenciálním rozdělením.

4.3.5 Exponenciální rozdělení – *Exp*(λ)

Exponenciální rozdělení je spojité rozdělení pravděpodobnosti, které popisuje délky časových intervalů mezi jednotlivými událostmi tzv. ***Poissonova procesu***, což znamená, že popisuje délku časových intervalů mezi jednotlivými událostmi, když se tyto události vyskytují vzájemně nezávisle a s konstantní intenzitou (tu popisuje jediný parametr tohoto rozdělení λ). Hustota je dána vztahem

$$f(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0\,. \tag{4.10}$$

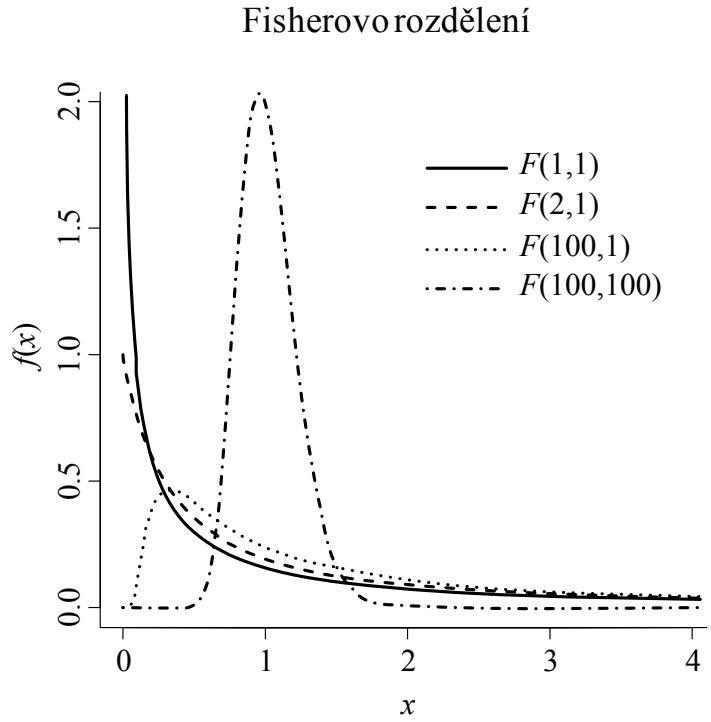
Hustoty exponenciálního rozdělení pro čtyři různé hodnoty parametru λ jsou zobrazeny na obrázku 4.5 vpravo. Exponenciální rozdělení má význam v analýze přežití, neboť je to nejjednodušší modelové rozdělení pro délku doby do výskytu sledované události, jeho jednoduchost je právě v konstantní intenzitě procesu, což přeneseně znamená, že systém nemá paměť a že doba od začátku sledování neovlivňuje intenzitu procesu. Zobecněními exponenciálního rozdělení (která umožňují časově závislou intenzitu procesu) jsou další rozdělení používaná zejména v analýze přežití, jmenovitě ***Weibullovo rozdělení*** a ***gamma rozdělení*** [4].

4.3.6 Fisherovo *F* rozdělení – $F(k_1,k_2)$

Fisherovo F rozdělení pravděpodobnosti je opět rozdělení s velkým využitím v teoretické statistice, kde vzniká jako podíl dvou chí-kvadrát rozdělení. Máme-li tedy dvě nezávislé náhodné veličiny, K_1 a K_2, s chí-kvadrát rozdělením se stupni volnosti k_1 a k_2, tedy platí $K_1 \sim \chi^2(k_1)$ a $K_2 \sim \chi^2(k_2)$, pak náhodná veličina F definovaná pomocí vztahu

$$F = \frac{K_1 / k_1}{K_2 / k_2} \tag{4.11}$$

má Fisherovo F rozdělení s parametry k_1 a k_2, které se zde, stejně jako v případě chí-kvadrát rozdělení, nazývají stupně volnosti. F rozdělení je používáno pro sestrojení $100(1 - \alpha)\%$ intervalu spolehlivosti (viz kapitola 5) pro podíl dvou rozptylů normálního rozdělení, σ_1^2 a σ_2^2, navíc je to také modelové rozdělení testové statistiky, kterou používáme pro ověření hypotézy o rovnosti dvou rozptylů (více v kapitole 7). V neposlední řadě má F rozdělení také významné místo v testování hypotézy o rovnosti středních hodnot veličiny X u více než dvou výběrových souborů. Použití F rozdělení v tzv. ***analýze rozptylu*** se věnuje kapitola 8. Hustoty Fisherova F rozdělení pro čtyři různé kombinace parametrů k_1 a k_2 jsou zobrazeny na obrázku 4.6.



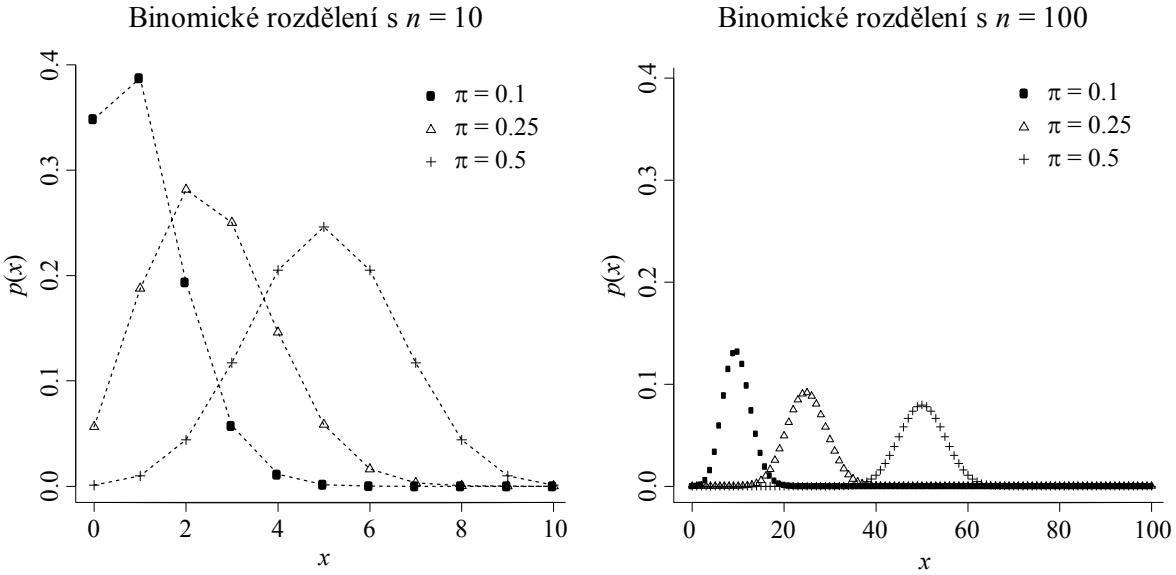
Obr. 4.6 Ukázky hustot náhodných veličin s Fisherovým F rozdělením.

4.3.7 Binomické rozdělení – *Bi*(n,π)

Prvním příkladem diskrétního rozdělení je ***binomické rozdělení***, které je základním modelem pro diskrétní veličiny v biologii i medicíně. Popisuje totiž počet výskytů sledovaného znaku nebo události (ve formě ano/ne, nastala/nenastala) v sérii n nezávislých experimentů, kdy v každém experimentu máme stejnou pravděpodobnost výskytu daného znaku (události), označenou π. Příklad náhodné veličiny s binomickým rozdělením byl uveden již v příkladu 3.1, kde jsme studovali počet líců v sérii pěti hodů mincí. Zde byl počet nezávislých experimentů $n = 5$ a pravděpodobnost výskytu dané události (padnutí líce) $\pi = 0{,}5$. Jiným příkladem může být sledování výskytu nežádoucích účinků u léčených pacientů nebo sledování přítomnosti biologického druhu na dané lokalitě. Pravděpodobnostní funkce binomického rozdělení s parametry n a π má tvar

$$p(x;n,\pi) = P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, x = 0, 1, 2, \ldots, n\,. \tag{4.12}$$

Pravděpodobnostní funkce binomické náhodné veličiny pro experiment s 10 a 100 opakováními vzhledem ke třem různým možnostem parametru π jsou uvedeny na obrázku 4.7 (na obrázku vlevo jsou body jednotlivých funkcí pro přehlednost spojeny tečkovanou čárou).



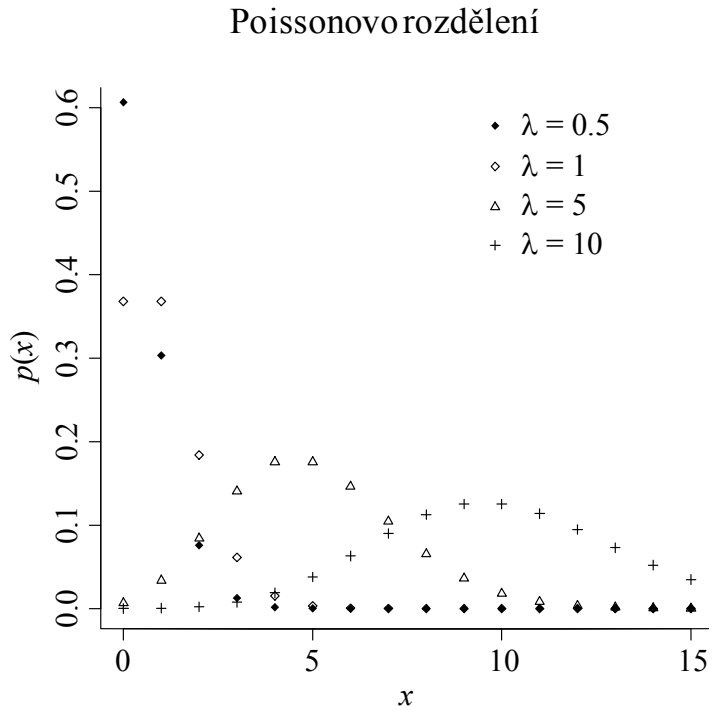
Obr. 4.7 Pravděpodobnostní funkce náhodné veličiny s binomickým rozdělením pro n=10 a n=100.

4.3.8 Poissonovo rozdělení – *Po*(λ)

Druhým příkladem diskrétního rozdělení pravděpodobnosti je ***Poissonovo rozdělení***, které popisuje počet výskytů sledovaného znaku nebo události na danou jednotku času, plochy, případně objemu s tím, že se tyto události vyskytují vzájemně nezávisle a s konstantní intenzitou (tu popisuje jediný parametr tohoto rozdělení, intenzita λ, známá již z popisu exponenciálního rozdělení). Poissonovo rozdělení lze chápat jako limitní případ binomického rozdělení, které přechází v Poissonovo při rostoucím počtu opakování experimentu n (dle definice $n \to \infty$) a klesající pravděpodobnosti výskytu jednotlivé události π (dle definice $\pi \to 0$), přičemž součin $n\pi$ přechází v intenzitu λ. Ve chvíli, kdy můžeme odlišit jednotlivá opakování experimentu (např. hody kostkou), není důvod pracovat s jiným rozdělením, než je binomické. Nicméně, úplně jiná situace nastává ve chvíli, kdy počet jednotlivých experimentů začne být neměřitelný ($n \to \infty$). Jedinou možností v takové situaci je práce s jinak definovanými experimentálními jednotkami, tedy např. s časovými intervaly, plochou nebo objemovými intervaly, ve kterých provádíme sledování (jeden kalendářní rok, výrobní směna, směna v nemocnici, letní sezona, apod.). V takové situaci binomické rozdělení již použít nelze. Jako příklady veličin s Poissonovým rozdělením lze uvést počet komplikací během určitého časového intervalu po operaci, počet žížal vyskytujících se na 1 m^2 pole, počet krvinek v poli mikroskopu nebo průměrný počet mutací bakterií měřený klasickým výsevem kolonií na jednu Petriho misku. Pro úplnost uvádíme pravděpodobnostní funkci Poissonova rozdělení s parametrem λ:

$$p(x;\lambda) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, \ldots \tag{4.13}$$

Ukázky pravděpodobnostní funkce Poissonova rozdělení pro čtyři různé hodnoty parametru λ jsou zobrazeny na obrázku 4.8.



Obr. 4.8 Ukázky pravděpodobnostní funkce náhodných veličin s Poissonovým rozdělením.

4.4 Shrnutí

V této kapitole jsme definovali základní spojitá a diskrétní rozdělení pravděpodobnosti používaná v praktické i teoretické biostatistice. Tato rozdělení popisují pravděpodobnostní chování řady znaků v přírodě i medicíně a jejich znalost je tak důležitá pro pochopení základů biostatistiky. V neposlední řadě je také důležitý teoretický význam standardizovaného (normovaného) normálního rozdělení, případně chí-kvadrát rozdělení, Studentova *t* rozdělení a Fisherova *F* rozdělení, na kterých jsou založeny základní statistické testy. Mezi rozděleními jsou zásadní rozdíly, např. normální veličina se vždy vyznačuje symetrickým rozdělení hodnot, zatímco u logaritmicko-normální veličiny je rozdělení hodnot vždy asymetrické. Vždy je třeba se zamýšlet např. nad tím, jaké problémy můžeme řešit pomocí binomického rozdělení, a kdy na náš problém lépe sedí pravděpodobnostní model Poissonova rozdělení.

Často lze na konkrétní rozdělení usuzovat již ze samotné podstaty studované problematiky, v případě nejasností nám mohou pomoci jak grafické vizualizační nástroje (histogram a krabicový graf), tak i popisné statistiky (kvantily, průměr, medián, výběrová směrodatná odchylka).

5 Bodové a intervalové odhady

V předchozích kapitolách jsme popsali pravděpodobnostní chování náhodné veličiny jako funkci zahrnující jeden či více parametrů. Abychom byli schopni popisovat, testovat a jinak rozhodovat o vlastnostech náhodných veličin, je nezbytné tyto parametry odhadnout (v drtivé většině biologických a klinických úloh jsou totiž konkrétní hodnoty těchto parametrů neznámé). Naším cílem je tedy sestrojit na základě pozorování náhodné veličiny X statistiku (vymyslet jakoukoliv smysluplnou transformaci dat), která by poskytla jejich nejlepší možný odhad. Zásadním předpokladem praktického použití odhadů samozřejmě je, že pozorované hodnoty nesou informaci o neznámých parametrech, což znamená, že námi předpokládaný model pravděpodobnostního chování dané náhodné veličiny je správný.

V této kapitole se krátce zmíníme o dvou hlavních typech bodových odhadů, a to o nestranných odhadech a maximálně věrohodných odhadech, a budeme se věnovat použití průměru a mediánu jako odhadů charakteristik symetrických a asymetrických dat. Hlavním cílem této kapitoly je však zavedení intervalových odhadů, intervalů spolehlivosti, které jsou jedním ze základních kamenů biostatistiky s naprosto klíčovým významem pro praxi.

5.1 Nestranné odhady

Existuje řada postupů pro nalezení bodového odhadu neznámých parametrů nebo charakteristik rozdělení pravděpodobnosti, které se liší jak svojí filozofií (např. Bayesovské odhady nehledají jednu hodnotu parametru, ale celé rozdělení pravděpodobnosti, neboť chápou parametr rozdělení jako náhodnou veličinu [34]), tak definicí kritéria optimálních vlastností odhadu. V teoretické statistice má významné místo ***metoda založená na Raově-Blackwellově větě***, která slouží k nalezení nestranného odhadu s nejmenší variabilitou (Raova-Blackwellova věta představuje složitější téma, které přesahuje rámec těchto skript, více lze nalézt v [13]). Označme řeckým θ odhadovaný parametr daného rozdělení pravděpodobnosti (konkrétní řecké symboly pro jednotlivá rozdělení jsou uvedena v kapitole 4). ***Nestranný odhad*** parametru θ je pak definován jako odhad, jehož střední hodnota je rovna θ a to pro každou hodnotu, které může tento parametr ze své definice nabývat. Nestrannost odhadu je celkem logickým omezením, které nám říká, že tento odhad má vzhledem ke střední hodnotě nulové vychýlení (v úvodní kapitole bylo zmíněno, že se chceme vyvarovat zkreslení výsledků (*biased results*), přičemž nestrannost je ve své podstatě to samé, neboť náš odhad chceme oprostit od systematické chyby). Jako příklad nestranného odhadu lze uvést výběrový průměr jako odhad střední hodnoty (parametru μ) normálního rozdělení. Mějme náhodný výběr $X_1, ..., X_n$, s tím, že $X_i \sim N(\mu,\sigma^2)$. Pak platí

$$E(\bar{X}) = E(\tfrac{1}{n}\sum_{i=1}^{n} X_i) = \tfrac{1}{n}\sum_{i=1}^{n} EX_i = \mu \text{ pro každé } \mu \in \boldsymbol{R}. \tag{5.1}$$

Stejně tak lze ukázat, že výběrový průměr je nestranným odhadem střední hodnoty (parametru λ) Poissonova rozdělení. Mějme náhodný výběr $X_1, ..., X_n$, kde $X_i \sim Po(\lambda)$. Pak platí

$$E(\bar{X}) = E(\tfrac{1}{n}\sum_{i=1}^{n} X_i) = \tfrac{1}{n}\sum_{i=1}^{n} EX_i = \lambda \text{ pro každé } \lambda \in \boldsymbol{R}, \lambda > 0. \tag{5.2}$$

Nestranných odhadů pro odhad parametru θ může být více, pro nás je z praktického hlediska nejvýhodnější ten, který má ze všech nestranných odhadů nejmenší rozptyl (variabilitu). Ten je pak označován jako ***nejlepší nestranný odhad***. Problematika nestranných odhadů není úplně intuitivní, proto zde uvádíme na toto téma příklad.

Příklad 5.1. Uvažujme náhodnou veličinu X, která představuje časovou délku návštěvy u praktického lékaře, u níž předpokládáme rovnoměrně spojité rozdělení pravděpodobnosti na intervalu $[0,\theta]$, tedy $X \sim Rs(0,\theta)$, kde θ je neznámý parametr. Naším cílem je odhad parametru θ, tedy odhad maximální doby, kterou je možno strávit v ambulanci (motivací může být např. optimalizace počtu praktických lékařů v daném regionu). Uvažujme náhodný výběr $X_1,\ldots, X_n$ i jeho uspořádanou variantu $X_{(1)} \le X_{(2)} \ldots \le X_{(n)}$, kde $X_{(1)}$ značí minimální hodnotu v náhodném výběru a naopak $X_{(n)}$ hodnotu maximální. Pak máme dva zajímavé a relativně intuitivní odhady parametru θ:

Odhad T_1: $$T_1 = 2\bar{X} = \frac{2}{n}\sum_{i=1}^{n} X_i \qquad (5.3)$$

Odhad T_2: $$T_2 = \frac{n+1}{n} X_{(n)} = \frac{n+1}{n}\max(X_i) \qquad (5.4)$$

První odhad je dvojnásobek výběrového průměru, druhý odhad je pak $(n+1)/n$ násobkem maximální hodnoty pozorované v náhodném výběru. K ověření charakteristik odhadů T_1 a T_2 potřebujeme nejprve získat charakteristiky výběrového průměru a maximální hodnoty. Ty lze odvodit s použitím pravidel pro počítání se středními hodnotami a rozptyly jako následující (přenecháváme čtenáři jako cvičení na výpočet střední hodnoty a rozptylu transformované náhodné veličiny)

$$E(\bar{X}) = \tfrac{1}{n}\sum_{i=1} E(X_i) = \theta/2 \qquad D(\bar{X}) = \frac{1}{12n}\theta^2 \qquad (5.5)$$

$$E(X_{(n)}) = E(\max(X_i)) = \frac{n}{n+1}\theta \qquad D(X_{(n)}) = \frac{n\theta^2}{(n+1)^2(n+2)} \qquad (5.6)$$

S pomocí výrazů (5.5) a (5.6) pak odvodíme charakteristiky odhadů T_1 a T_2:

$$ET_1 = E(2\bar{X}) = 2(\theta/2) = \theta \qquad D(T_1) = \frac{1}{3n}\theta^2 \qquad (5.7)$$

$$ET_2 = E(\tfrac{n+1}{n} X_{(n)}) = \frac{n+1}{n}\frac{n}{n+1}\theta = \theta \qquad D(T_2) = \frac{1}{n(n+2)}\theta^2 \qquad (5.8)$$

Jak je vidět z výrazů (5.7) a (5.8), oba odhady, T_1 i T_2, jsou nestrannými odhady parametru θ, neboť jejich střední hodnota je rovna θ. Z hlediska jejich rozptylu (variability odhadu) je však lepším odhadem T_2, jehož rozptyl s rostoucím n rychleji klesá k 0 (ve jmenovateli má kvadrát velikosti vzorku). Závěrem lze tedy říci, že pro odhad parametru θ je vhodnější použít odhad, který je $(n+1)/n$ násobkem maximální pozorované hodnoty.

5.2 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti je důležitým nástrojem biostatistiky, který je používán pro jednoduché odhady, jako je např. odhad směrodatné odchylky normální náhodné veličiny, i velmi netriviální odhady v nelineárních modelech s daty z jiného než normálního rozdělení pravděpodobnosti [13]. Principem metody maximální věrohodnosti je najít odhad parametru θ (případně vektoru parametrů), který maximalizuje pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení pravděpodobnosti. Jinými slovy se snažíme najít takovou hodnotu θ, pro niž je pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení, maximální. Odhad se tedy snaží maximálně přizpůsobit pozorovaným datům, což je logické, když připouštíme, že data představují jediný zdroj informací o našem neznámém parametru. Opět ale platí, že celý úspěch maximálně věrohodného odhadu je závislý na korektní specifikaci pravděpodobnostního chování, tedy volbě konkrétního rozdělení pravděpodobnosti.

Uvažujme náhodný výběr $X_1,\ldots, X_n$, tedy n nezávislých náhodných veličin se stejným rozdělením pravděpodobnosti s hustotou $f(x,\boldsymbol{\theta})$, kde $\boldsymbol{\theta}$ představuje vektor neznámých parametrů. Sdružená hustota, případně pravděpodobnostní funkce, odpovídající n realizacím náhodné veličiny X, tedy hodnotám $x_1, x_2,\ldots, x_n$, pak má tvar:

$$f(x_1,\ldots,x_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}). \tag{5.9}$$

Za předpokladu, že známe $\boldsymbol{\theta}$, vyjadřuje větší hodnota sdružené hustoty větší shodu pozorovaných hodnot s předpokládaným rozdělením s hustotou $f(x,\boldsymbol{\theta})$. Hlavní myšlenkou metody maximální věrohodnosti je dívat se na sdruženou hustotu nikoliv jako na funkci x_1, $x_2,\ldots, x_n$, ale jako na funkci vektoru $\boldsymbol{\theta}$ (při pevně daných x_1, $x_2,\ldots, x_n$), a vybrat ze všech možných hodnot $\boldsymbol{\theta}$ takové, aby výraz (5.9) nabýval svého maxima. Pro tento účel zavádíme tzv. ***funkci věrohodnosti*** (*likelihood function*) ve tvaru

$$L(\boldsymbol{\theta} \mid x_1,\ldots,x_n) = f(x_1,\ldots,x_n \mid \boldsymbol{\theta}), \tag{5.10}$$

což je vyjádření shodné se sdruženou hustotou, kde ovšem jako závisle proměnná vystupuje vektor neznámých parametrů $\boldsymbol{\theta}$. Maximálně věrohodný odhad vektoru $\boldsymbol{\theta}$ pak značíme $\hat{\boldsymbol{\theta}}_{MLE}$, a je to číselný vektor, který maximalizuje funkci věrohodnosti, tedy

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\theta\in\Theta} L(\boldsymbol{\theta} \mid x_1,\ldots,x_n), \tag{5.11}$$

kde výraz Θ symbolizuje parametrický prostor, tedy prostor všech možných hodnot vektoru $\boldsymbol{\theta}$. Často je pro nás výhodnější maximalizovat logaritmus funkce věrohodnosti (řada rozdělení pravděpodobnosti má hustotu vyjádřenou pomocí exponenciály a přirozený logaritmus je tudíž výhodný pro zjednodušení součinu). Tato tzv. ***logaritmická věrohodnostní funkce*** (*log-likelihood function*) má tvar

$$l(\boldsymbol{\theta} \mid x_1,\ldots,x_n) = \ln L(\boldsymbol{\theta} \mid x_1,\ldots,x_n) = \ln \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta}). \tag{5.12}$$

Tuto operaci si můžeme dovolit, protože přirozený logaritmus je funkce, která zachovává extrémy (je monotónní). Je-li funkce věrohodnosti diferencovatelná, lze najít maximálně věrohodný odhad jako stacionární bod funkce L nebo l, tedy řešení systému rovnic, kdy první derivace funkce věrohodnosti (nebo jejího logaritmu) podle parametrů položíme rovny 0. Následně bychom měli ověřit, zda jsme opravdu nalezli maximum, např. pomocí druhých derivací.

Příklad 5.2. Najděme maximálně věrohodný odhad parametru λ Poissonova rozdělení. Uvažujme n nezávislých pozorování, x_1, x_2,…, x_n, z Poissonova rozdělení, funkce věrohodnosti pak má tvar

$$L(\lambda \mid x_1,\ldots,x_n) = p(x_1,\ldots,x_n \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\Sigma_i x_i}}{\prod_i x_i!}. \tag{5.13}$$

Logaritmus funkce věrohodnosti lze pomocí jednoduchých pravidel pro počítání s logaritmy vyjádřit jako

$$\ln L(\lambda \mid x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i \ln\lambda - n\lambda - \ln(\prod_{i=1}^{n} x_i!), \tag{5.14}$$

derivace logaritmu funkce věrohodnosti podle λ (která se pro dosažení maxima má rovnat nule) pak vypadá následovně:

$$\frac{d \ln L}{d\lambda} = \sum_{i=1}^{n} x_i / \lambda - n = 0. \tag{5.15}$$

Jednoduchou úpravou dostáváme, že maximálně věrohodným odhadem parametru λ Poissonova rozdělení je průměrný počet pozorovaných událostí v n opakováních experimentu (že je průměr opravdu maximem lze ověřit pomocí druhých derivací), tedy

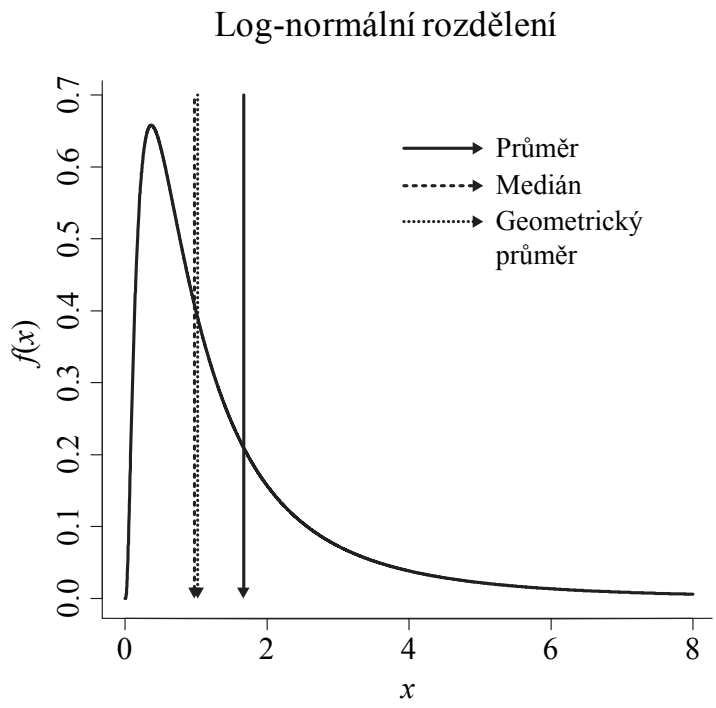
$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{5.16}$$

Obdobně bychom mohli pro normální rozdělení ověřit, že průměr je maximálně věrohodným odhadem parametru μ.

5.3 Srovnání průměru a mediánu

V kapitole věnované odhadům nesmí chybět rozvaha nad použitím průměru a mediánu jako bodových odhadů neznámých parametrů (oba tyto odhady byly definovány v kapitole 2). Vhodnost jejich použití totiž není dána pouze symetrií, respektive asymetrií pozorovaných hodnot, ale také účelem studie. Platí sice, že průměr je dobrou charakteristikou frekvenčního středu (dobrým odhadem střední hodnoty) tehdy, když jsou naše data symetrická a neobsahují odlehlé či nesprávné hodnoty, to však neznamená, že ho např. v případě asymetrických dat nelze nikdy použít.

Ideálním příkladem pro vysvětlení je právě veličina s asymetrickým logaritmicko-normálním rozdělením pravděpodobnosti. Chceme-li charakterizovat logaritmicko-normální rozdělení z hlediska střední hodnoty, je použití průměru opravdu nevhodné, neboť v případě těchto dat má průměr tendenci se přizpůsobovat vysokým hodnotám, které jsou pozorovány s malou četností. To ho jako odhad frekvenčního středu dat diskvalifikuje. Nejvhodnějším odhadem je tzv. ***geometrický průměr***, což není nic jiného než průměr spočítaný na normalizovaných hodnotách, tedy na hodnotách po transformaci $y = \ln(x)$, případně medián. Srovnání průměru, geometrického průměru a mediánu u dat pocházejících z logaritmicko-normálního rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$ je uvedeno na obrázku 5.1, na kterém je také vidět, že hodnoty mediánu a geometrického průměru téměř splývají.



Obr. 5.1 Srovnání průměru, geometrického průměru a mediánu na datech z logaritmicko-normálního rozdělení.

Na druhou stranu, chceme-li charakterizovat logaritmicko-normální rozdělení z hlediska celkového součtu pozorovaných hodnot, může být použití průměru smysluplné. Pokud nám v dané studii jde o to charakterizovat např. spotřebu nějakého materiálu (papíru, dřeva, léků, alkoholu) nebo třeba peněz, pak aritmetický průměr popisuje z hlediska celkového součtu spotřebu lépe než výše uvedený geometrický průměr nebo medián. Motivací pro tento typ studie může být např. plánování finančních prostředků na léčbu nějakého onemocnění na další rok. Pokud bychom postupovali tak, že bychom předpokládaný počet pacientů vynásobili

hodnotou geometrického průměru nákladů na léčbu (nebo mediánu), dostali bychom objem financí, které by spotřeboval předpokládaný počet „typických“ pacientů s daným onemocněním. Tento výpočet by však neodpovídal realitě, neboť v praxi se nevyskytují pouze „typičtí“ pacienti. Odhad, který bychom dostali vynásobením předpokládaného počtu pacientů klasickým průměrem nákladů na léčbu, by byl v tomto případě vhodnější, neboť počítá právě i s „netypickými“ pacienty (jinými slovy s pacienty, jejichž náklady na léčbu jsou z nějakého důvodu vyšší než u ostatních). Důvod pro to, že průměr dobře charakterizuje celkový součet pozorovaných hodnot je prostý a vychází z jeho definice

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \sum_{i=1}^{n} x_i = n\bar{x}\,. \tag{5.17}$$

Naopak ze znalosti mediánu a počtu pozorování nejsme schopni celkový součet pozorovaných hodnot zrekonstruovat.

Závěrem tedy nelze říci, že by jedna ze sumárních statistik byla lepší než druhá, určitě je na místě používat při výpočtech jak průměr, tak medián (a samozřejmě i geometrický průměr), nicméně je vždy třeba se zamyslet nad účelem použití této sumární statistiky a nad charakterem dat, která chceme sumarizovat. Na prvním místě zpracování dat by vždy měla být identifikace chybných a jinak „nevěrohodných“ pozorování, hned v závěsu bychom se pak měli věnovat identifikaci rozdělení, z něhož data pochází (ověřit předpoklad normality hodnot, nebo alespoň jejich symetrie, případně ověření Poissonova rozdělení). Nakonec je však jistě nejlepší radou výpočet obou hodnot, tedy průměru i mediánu, které jsou spolu s různými mírami variability cenným zdrojem informací o vlastnostech sledované náhodné veličiny.

5.4 Teoretické pozadí intervalových odhadů

Výpočet bodového odhadu neznámého parametru rozdělení pravděpodobnosti nebo nějaké jeho funkce je bez diskuze prvním krokem ve statistickém zpracování dat. Představme si však situaci, kdy dva různí lidé budou sledovat stejný znak, respektive měřit stejnou náhodnou veličinu. Vzhledem k variabilitě měření a faktu, že oba výzkumníci budou mít jistě rozdílné výběrové soubory, lze předpokládat, že oba při měření dané veličiny dojdou ke dvěma různým bodovým odhadům. Na místě je pak otázka, které z těchto dvou čísel je lepší, přesnější, správnější? Bez další znalosti nejsme schopni na tuto otázku korektně odpovědět, může nám však napovědět skutečnost, že první z výzkumníků měl soubor např. 1000 pacientů, zatímco druhý z nich měl soubor pouze 10 pacientů. Jistě bychom se v tomto případě přiklonili spíš k prvnímu odhadu, neboť instinktivně tušíme, že odhad založený na větším množství pacientů (informace) lze považovat za lepší (přesnější).

Bodový odhad je tak sám o sobě nedostatečný pro popis parametru rozdělení pravděpodobnosti náhodné veličiny, neboť nevíme nic o jeho přesnosti (spolehlivosti). Jinak řečeno, nemáme nijak pravděpodobnostně vyjádřeno, jak je tento bodový odhad ve skutečnosti vzdálen od skutečné hodnoty neznámého parametru. Samozřejmě jsme-li v situaci, kdy jsme schopni měřením postihnout celou cílovou populaci, nepotřebujeme žádné vyjádření spolehlivosti, protože jsme schopni odhadnout sledovaný parametr přesně. V praxi je však tato situace nereálná.

5.4.1 Vlastnosti výběrového průměru

Nejen průměr, ale jakákoliv statistika je jako transformace náhodných veličin také náhodnou veličinou a má tudíž i vlastní rozdělení pravděpodobnosti. Vzhledem k tomu, že jednotlivé realizace náhodné veličiny X vykazují variabilitu (popsanou směrodatnou odchylkou, $SD(X)$), pak i jednotlivé realizace statistiky nad různými náhodnými výběry vykazují variabilitu, která je úměrná $SD(X)$.

Co se týče výběrového průměru, má tento odhad dvě zajímavé vlastnosti, které jsou stěžejní nejen pro konstrukci intervalů spolehlivosti, ale i pro řadu dalších biostatistických úloh:

1. Rozdělení pravděpodobnosti výběrového průměru má tím menší rozptyl (variabilitu), čím více pozorování je v průměru zahrnuto, tedy čím větší je výběrový soubor (větší n). Jinými slovy, máme-li více informací, jsme schopni odhadovat s větší přesností. Tato vlastnost průměru plyne z vlastností rozptylu transformované náhodné veličiny.

2. Rozdělení pravděpodobnosti výběrového průměru se s rostoucí velikostí souboru (rostoucím n) přestává podobat rozdělení původní náhodné veličiny X a začíná se podobat rozdělení normálnímu. Tato vlastnost plyne z centrální limitní věty, která je klíčovým tvrzením teoretické statistiky.

Nejprve se věnujme rozptylu průměru jako transformované náhodné veličiny. Mějme posloupnost $X_1, \ldots, X_n$ nezávislých náhodných veličin se stejným rozdělením pravděpodobnosti, které má konečnou střední hodnotu μ a rozptyl σ^2. Pak z pravidel pro výpočet rozptylu platí, že rozptyl výběrového průměru má tvar

$$D(\bar{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \qquad (5.18)$$

Pro praktické počítání je třeba pracovat se stejnými jednotkami, jako má původní náhodná veličina, což znamená vyjádřit i směrodatnou odchylku výběrového průměru:

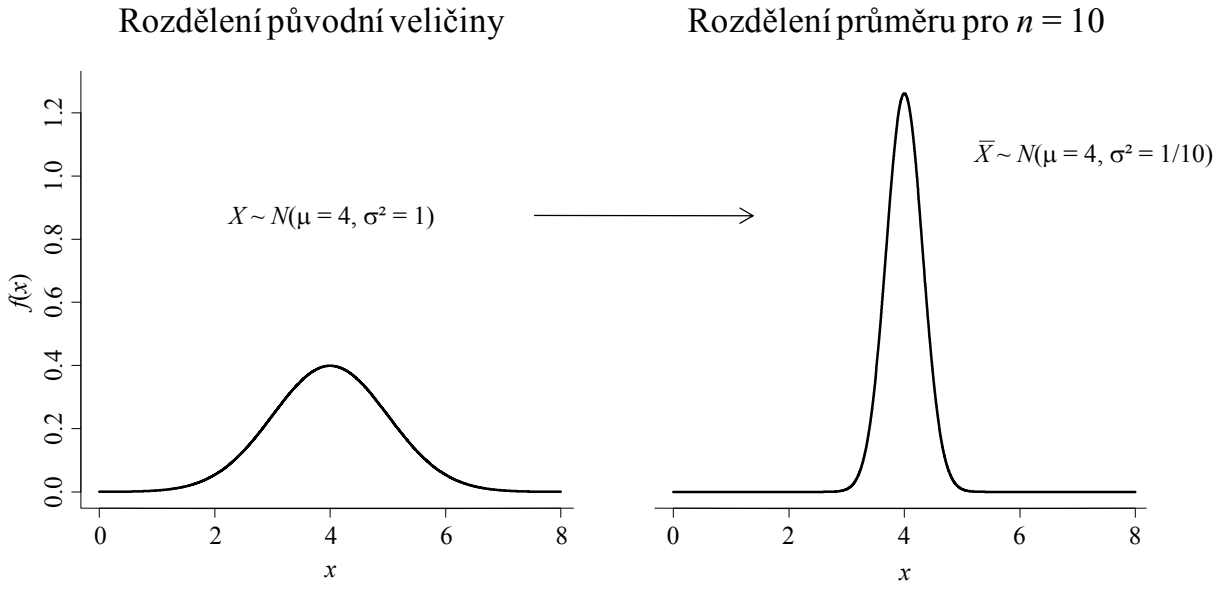
$$SD(\bar{X}) = \sqrt{D(\bar{X})} = \frac{\sigma}{\sqrt{n}}. \qquad (5.19)$$

Výraz (5.19), tedy směrodatná odchylka výběrového průměru, se nejčastěji označuje pojmem ***standardní chyba*** (*standard error*), zkráceně značeno *SE*. Platí tedy

$$SE(\bar{X}) = SD(\bar{X}) = \sigma/\sqrt{n} = SD(X)/\sqrt{n}. \qquad (5.20)$$

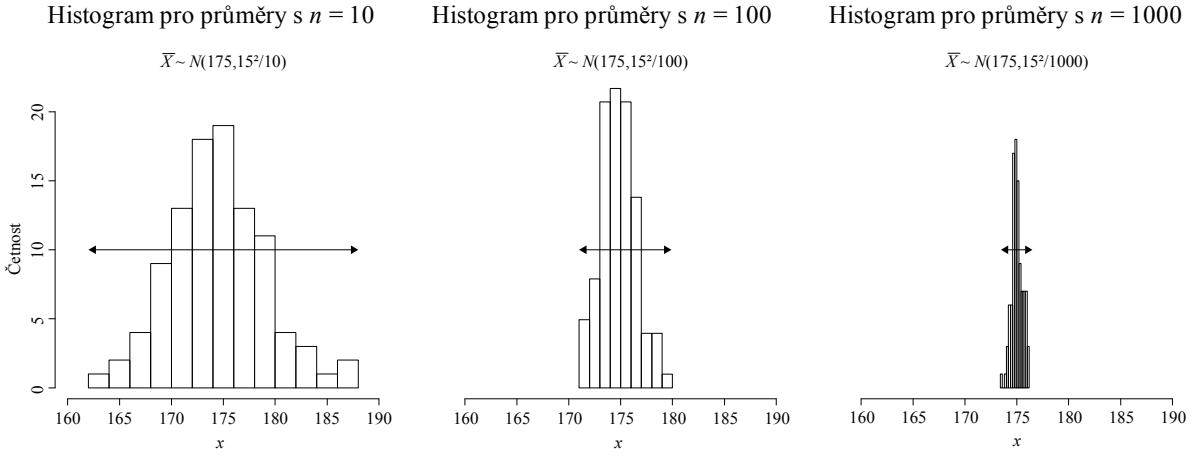
Je velmi důležité si uvědomit, že směrodatná odchylka náhodné veličiny, tedy $SD(X)$, je odrazem variability náhodné veličiny ve sledované populaci a souvisí tak s variabilitou biologického procesu (nelze ji tudíž ovlivnit). Na druhou stranu, směrodatná odchylka výběrového průměru, tedy standardní chyba $SE(\bar{X})$, je odrazem přesnosti výběrového průměru jako odhadu střední hodnoty náhodné veličiny a jako taková souvisí nejen s variabilitou biologického procesu, ale zejména s velikostí vzorku, která hodnotu standardní chyby ovlivňuje zásadním způsobem. Rozdíl mezi rozdělením pravděpodobnosti náhodné veličiny a výběrového průměru pro velikost výběru $n = 10$ je uveden na obrázku 5.2.

Z obrázku je vidět i to, že zatímco realizace náhodné veličiny z rozdělení $N(4,1)$ v blízkosti čísla 5 je očekávatelná, realizace průměru deseti pozorování této veličiny v blízkosti čísla 5 je již velmi málo pravděpodobná.



Obr. 5.2 Srovnání hustoty rozdělení původní veličiny a výběrového průměru pro n=10.

Příklad 5.3. Předpokládejme, že výška člověka je náhodná veličina, označme ji X, pocházející z normálního rozdělení pravděpodobnosti se střední hodnotou 175 cm a směrodatnou odchylkou 15 cm, tedy že platí $X \sim N(175,15^2)$. Pomocí pravidla ± 3 sigma si lze ověřit, že náhodná veličina se (z 99 %) realizuje zhruba v rozsahu hodnot od 120 cm do 220 cm. Zajímá nás, jak se realizují průměry jako náhodné veličiny pro měnící se velikost výběru, tedy průměry pro výběry o velikosti $n = 10$, $n = 100$ a $n = 1000$ pozorování. Výsledky jsou zobrazeny pomocí histogramů na obrázku 5.3.



Obr. 5.3 Histogramy realizací výběrového průměru pro různé počty pozorování.

5.4.2 Centrální limitní věta

Centrální limitní věta je klíčové matematické tvrzení, které popisuje pravděpodobnostní chování výběrového průměru pro velké vzorky a umožňuje tak sestrojení intervalových odhadů, a to nejen pro normálně rozdělené náhodné veličiny.

Opět mějme posloupnost $X_1, \ldots, X_n$ nezávislých, stejně rozdělených náhodných veličin, které mají konečnou střední hodnotu μ a rozptyl σ^2. Zjednodušeně řečeno, dle centrální limitní věty pak platí, že pro $n \to \infty$ má výběrový průměr

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.21}$$

normální rozdělení se střední hodnotou μ a rozptylem σ^2/n (rozdělení výběrového průměru ***konverguje*** k normálnímu rozdělení tzv. ***v distribuci***, což matematicky popisuje níže uvedený vztah 5.23). Průměr je zde záměrně zapsán pomocí velkého písmene X, abychom zdůraznili, že se jedná o náhodnou veličinu. Toto tvrzení je ekvivalentní s tvrzením, že za výše uvedených podmínek má náhodná veličina

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \tag{5.22}$$

přibližně standardizované normální rozdělení pravděpodobnosti, $N(0,1)$, tedy že platí

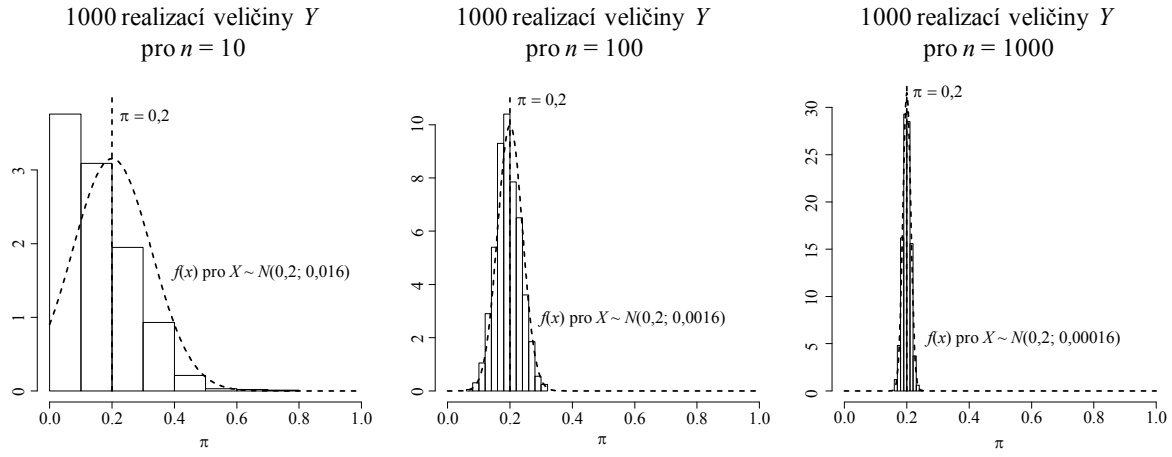
$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le x\right) = \lim_{n\to\infty} \mathrm{P}(Z \le x) = F_{N(0,1)}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du\,. \tag{5.23}$$

Zjednodušeně lze centrální limitní větu interpretovat tak, že pokud je rozdělení pravděpodobnosti náhodné veličiny X normální, pak je i rozdělení průměru pozorovaných hodnot normální (a to i pro $n = 1$). Pokud však rozdělení pravděpodobnosti náhodné veličiny X normální není, pak je rozdělení průměru pozorovaných hodnot přibližně normální, když n je dostatečně velké (matematicky řečeno, pro n jdoucí do nekonečna). Slovní obrat „dostatečně velké n“ je samozřejmě problematický, neboť každý si pod ním může představit něco jiného. Nicméně velikost souboru pro výpočet průměru by neměla být menší než 30 v případě rozdělení pravděpodobnosti podobných normálnímu a menší než 100 pro rozdělení, která nejsou podobná normálnímu.

Centrální limitní věta funguje dokonce i tehdy, když rozdělení původní náhodné veličiny není spojité, ale diskrétní. Jednoduchým příkladem je binomická náhodná veličina X, která je definována jako součet n nul a jedniček (úspěchů a neúspěchů). Pokud tuto veličinu transformujeme na $Y = X / n$, již dostáváme průměr, což znamená, že při dostatečném n můžeme s veličinou Y pracovat jako s veličinou s normálním rozdělením.

Příklad 5.4. Předpokládejme, že skutečný podíl dospělých s hypertenzí je v České republice roven 0,2. Definujeme náhodnou veličinu X, která bude udávat počet hypertoniků v souboru osob o velikosti n (tedy náhodná veličina X má binomické rozdělení pravděpodobnosti s parametry n a $\pi = 0{,}2$). Dále zavedeme náhodnou veličinu Y jako X / n. Zajímá nás, jak se chová tisíc realizací náhodné veličiny Y (zobrazeno pomocí histogramu) pro náhodné výběry o velikosti $n = 10$, $n = 100$ a $n = 1000$. Výsledky ukazuje obrázek 5.4, kde je vidět, jak se zvyšujícím se počtem pozorování ve výběru histogram dobře kopíruje

hustotu teoretického normálního rozdělení s příslušnými parametry odvozenými z parametrů binomického rozdělení veličiny Y ($\mu = \pi = 0{,}2$; $\sigma^2 = \pi(1-\pi)/n$).



Obr. 5.4 Shoda empirického rozdělení realizací veličiny Y s příslušným normálním rozdělením.

5.5 Intervalové odhady

Pro každé rozdělení pravděpodobnosti lze omezit oblast, kde se náhodná veličina s tímto rozdělením realizuje s pravděpodobností $1 - \alpha$, pomocí jejích kvantilů, tedy čísel na reálné ose. Tento fakt představuje teoretický základ konstrukce intervalových odhadů a nejlépe se demonstruje na příkladu výběrového průměru a normálního rozdělení. Jak plyne z centrální limitní věty, rozdělení pravděpodobnosti výběrového průměru lze při dostatečné velikosti souboru aproximovat normálním rozdělením a je tak možno pracovat s dobře dostupnými (tabelovanými) kvantily normálního rozdělení. Provedeme-li navíc standardizaci výběrového průměru na veličinu Z (veličina Z má tedy potom standardizované normální rozdělení), je oblast, kde se náhodná veličina Z realizuje s pravděpodobností $1 - \alpha$, vyjádřena pomocí následujícího vztahu:

$$1-\alpha = 1-\tfrac{\alpha}{2}-\tfrac{\alpha}{2} = \left(1-\tfrac{\alpha}{2}\right)-\tfrac{\alpha}{2} = F_{N(0,1)}(z_{1-\alpha/2}) - F_{N(0,1)}(z_{\alpha/2}) = P(z_{\alpha/2} \le Z \le z_{1-\alpha/2}), \qquad (5.24)$$

kde $z_{\alpha/2}$ a $z_{1-\alpha/2}$ jsou hodnoty $100(\alpha/2)$procentního, respektive $100(1 - \alpha/2)$ procentního kvantilu standardizovaného normálního rozdělení. Vztah (5.24) je shodný se vztahem (4.5), což není nic překvapujícího, neboť princip konstrukce intervalového odhadu pro odhad střední hodnoty normálního rozdělení (μ), respektive pro výběrový průměr, je shodný s teoretickým pozadím pravidla ± 3 sigma. Celá podstata konstrukce intervalu spolehlivosti spočívá v tom, že za náhodnou veličinu Z dosadíme její definiční vzorec a výraz upravíme tak, abychom mezi matematickými znaménky větší nebo rovno osamostatnili odhadovaný parametr (v případě výběrového průměru parametr μ).

5.5.1 Konstrukce intervalů spolehlivosti pro parametry normálního rozdělení

V této části kapitoly odvodíme intervaly spolehlivosti jak pro oba parametry normálního rozdělení, μ a σ^2, tak pro střední hodnotu rozdílu dvou náhodných veličin X a Y.

Konstrukce 100(1 – α)% intervalu spolehlivosti pro parametr μ

Mějme náhodný výběr $X_1, ..., X_n$ z normálního rozdělení pravděpodobnosti, tedy předpokládejme, že platí $X_i \sim N(\mu,\sigma^2)$, $i = 1, ..., n$. Nejprve budeme uvažovat situaci, kdy hodnotu σ^2 známe. Úpravou vztahu 5.24 dosazením za Z s tím, že platí $z_{1-\alpha/2} = -z_{\alpha/2}$, dostaneme

$$1-\alpha = P(z_{\alpha/2} \le Z \le z_{1-\alpha/2}) = P(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}) = P(-z_{1-\alpha/2} \le \tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}). \qquad (5.25)$$

Naším cílem je interval spolehlivosti pro μ, upravíme tedy vzorec tak, abychom μ mezi matematickými znaménky větší nebo rovno osamostatnili. Úpravou dostáváme

$$1-\alpha = P(-\tfrac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \le \overline{X}-\mu \le \tfrac{\sigma}{\sqrt{n}} z_{1-\alpha/2}) = P(\overline{X}-\tfrac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \le \mu \le \overline{X}+\tfrac{\sigma}{\sqrt{n}} z_{1-\alpha/2}). \qquad (5.26)$$

Vidíme, že jsme s pomocí pravděpodobnosti a známých kvantit vypočítali dolní a horní mez, které zdola a shora omezují neznámý parametr μ. Správně bychom řekli, že 100(1 – α)% interval spolehlivosti představuje oblast, která s pravděpodobností $1 - \alpha$ pokrývá neznámý parametr μ. 100(1 – α)% interval spolehlivosti pro parametr μ má tedy tvar

$$(\overline{X}-\tfrac{\sigma}{\sqrt{n}} z_{1-\alpha/2}; \overline{X}+\tfrac{\sigma}{\sqrt{n}} z_{1-\alpha/2}). \qquad (5.27)$$

Výraz $\sigma/\sqrt{n}$ jsme již dříve definovali jako standardní chybu výběrového průměru (viz vzorec (5.20)), proto 100(1 – α)% interval spolehlivosti pro parametr μ můžeme ještě vyjádřit v alternativní formě jako

$$100(1-\alpha)\% \text{ interval spolehlivosti pro } \mu = \left(\overline{X} - z_{1-\alpha/2} SE(\overline{X}); \overline{X} + z_{1-\alpha/2} SE(\overline{X})\right). \qquad (5.28)$$

Výše uvedené jsme odvozovali za předpokladu, že známe přesnou hodnotu parametru σ^2, což je však z praktického hlediska značně omezující (reálným příkladem však může být intervalový odhad střední hodnoty pro data měřená přístrojem s kalibrovanou a tudíž známou přesností). Ve chvíli, kdy neznáme hodnotu parametru σ^2, musíme pro konstrukci intervalu spolehlivosti použít jinou statistiku než Z, s jiným rozdělením pravděpodobnosti. Logické by bylo místo směrodatné odchylky, σ, použít výběrovou směrodatnou odchylku, s, nicméně tato náhrada není úplně jednoduchá, nejedná se o pouhé dosazení s za σ. Pomůžeme si vztahem (4.7), který definuje statistiku se Studentovým t rozdělením. Nejprve pomocí s^2 vytvoříme pomocnou statistiku K s chí-kvadrát rozdělením pravděpodobnosti (s $n - 1$ stupni volnosti):

$$K = \frac{n-1}{\sigma^2} s^2 \sim \chi^2(n-1). \qquad (5.29)$$

Tuto statistiku pak spolu se statistikou Z se standardizovaným normálním rozdělením použijeme pro vytvoření statistiky T se Studentovým t rozdělením:

$$T = \frac{Z}{\sqrt{K/(n-1)}} = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{(n-1)s^2/(n-1)\sigma^2}} = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1). \quad (5.30)$$

Vidíme, že statistika T vypadá stejně jako statistika Z, jen namísto směrodatné odchylky, σ, obsahuje výběrovou směrodatnou odchylku, s. To je přesně to, čeho jsme chtěli dosáhnout. Je však důležité si uvědomit, že statistika T má jiné rozdělení pravděpodobnosti než Z, tedy i jinou kvantilovou funkci. V souladu s (5.25) pro statistiku T platí

$$1-\alpha = P(t_{\alpha/2}(n-1) \le T \le t_{1-\alpha/2}(n-1)) = \ldots = P(-t_{1-\alpha/2}(n-1) \le \tfrac{\bar{X}-\mu}{s/\sqrt{n}} \le t_{1-\alpha/2}(n-1)), \quad (5.31)$$

kde $t_{\alpha/2}(n-1)$ a $t_{1-\alpha/2}(n-1)$ jsou $100(\alpha/2)\%$, respektive $100(1-\alpha/2)\%$, kvantily Studentova t rozdělení s $n-1$ stupni volnosti. Stejnými úpravami jako v případě intervalu spolehlivosti pro parametr μ při známém σ^2 dostaneme $100(1-\alpha)\%$ interval spolehlivosti pro parametr μ při neznámém σ^2 ve tvaru

$$(\bar{X} - \tfrac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1); \bar{X} + \tfrac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)). \quad (5.32)$$

Příklad 5.5. Chceme sestrojit 95% interval spolehlivosti pro střední hodnotu systolického tlaku studentů vysokých škol. Na vzorku $n = 100$ náhodně vybraných studentech byl výběrový průměr systolického tlaku roven hodnotě 123,4 mm Hg s výběrovou směrodatnou odchylkou $s = 14{,}0$ mm Hg. Kromě těchto hodnot je třeba k výpočtu 95% intervalu spolehlivosti ještě hodnota kvantilu t rozdělení příslušného hladině $\alpha = 0{,}05$ a $n - 1 = 99$ stupňům volnosti. V tabulkách nebo příslušném software najdeme, že $t_{0{,}975}(99) = 1{,}98$. Dosazením do vzorce (5.32) získáme

$$(\bar{X} - \tfrac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1); \bar{X} + \tfrac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)) = (123{,}4 - \tfrac{14{,}0}{\sqrt{100}} 1{,}98; 123{,}4 + \tfrac{14{,}0}{\sqrt{100}} 1{,}98), \quad (5.33)$$

což znamená, že 95% intervalem spolehlivosti pro střední hodnotu systolického tlaku studentů vysokých škol je interval (120,6 mm Hg; 126,2 mm Hg). Můžeme tedy říci, že s pravděpodobností 95 % interval (120,6 mm Hg; 126,2 mm Hg) pokrývá neznámou střední hodnotu systolického tlaku studentů vysokých škol.

Konstrukce $100(1-\alpha)\%$ intervalu spolehlivosti pro rozdíl parametrů μ_1 a μ_2

Velmi často nás zajímá odhad střední hodnoty sledované veličiny u dvou skupin subjektů, kdy tím, o co nám jde nejvíce, je rozdíl těchto dvou středních hodnot. Snažíme se totiž zjistit, jestli se sledovaný znak chová stejně u jedné i u druhé skupiny. Tuto situaci reprezentujeme dvěma navzájem nezávislými náhodnými veličinami, X_1 a X_2, u kterých předpokládáme normální rozdělení pravděpodobnosti, potažmo pak dvěma náhodnými výběry, $X_{11}, \ldots, X_{1n1}$, kde $X_{1i} \sim N(\mu_1, \sigma_1^2)$, a $X_{21}, \ldots, X_{2n2}$, kde $X_{2j} \sim N(\mu_2, \sigma_2^2)$. Z vlastností normálního rozdělení (viz kapitola 4) plyne, že i rozdíl průměrů náhodných veličin X_1 a X_2 má normální rozdělení pravděpodobnosti s tím, že platí

$$X_1 - X_2 \sim N\left(\mu_1 - \mu_2, \tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}\right). \tag{5.34}$$

V případě, že známe hodnoty parametrů σ_1 a σ_2, provedeme standardizaci náhodné veličiny $X_1 - X_2$ na veličinu Z a následně odvodíme 100(1 – α)% interval spolehlivosti naprosto stejným postupem jako při odvození intervalu spolehlivosti pro jeden parametr μ. Výsledný 100(1 – α)% interval spolehlivosti pro rozdíl středních hodnot náhodných veličin X_1 a X_2 má tvar

$$\left(X_1 - X_2 - z_{1-\alpha/2}\sqrt{\tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}}; X_1 - X_2 + z_{1-\alpha/2}\sqrt{\tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}}\right). \tag{5.35}$$

V případě, že neznáme hodnoty parametrů σ_1 a σ_2, si opět musíme pomoci statistikami, které mají chí-kvadrát rozdělení pravděpodobnosti a které nám pomohou se zbavit neznámých σ_1 a σ_2. Obdobně jako ve vztahu (5.29) tedy definujeme statistiky K_1 a K_2, které spolu se statistikou Z převedeme na statistiku T. Ta má opět Studentovo t rozdělení. Parametr t rozdělení, tedy počet stupňů volnosti, je však v obecném případě, kdy $\sigma_1 \neq \sigma_2$, dán vztahem

$$\nu = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}. \tag{5.36}$$

Budeme-li předpokládat rovnost obou směrodatných odchylek, tedy $\sigma_1 = \sigma_2$, je $\nu = n_1 + n_2 - 2$. Odpovídajícími úpravami dostaneme 100(1 – α)% interval spolehlivosti pro rozdíl středních hodnot náhodných veličin X_1 a X_2 při neznámých hodnotách parametrů σ_1 a σ_2 ve tvaru

$$\left(\overline{X}_1 - \overline{X}_2 - t_{1-\alpha/2}(\nu)\sqrt{\tfrac{s_1^2}{n_1} + \tfrac{s_2^2}{n_2}}; \overline{X}_1 - \overline{X}_2 + t_{1-\alpha/2}(\nu)\sqrt{\tfrac{s_1^2}{n_1} + \tfrac{s_2^2}{n_2}}\right). \tag{5.37}$$

Příklad 5.6. V průběhu experimentu sledujeme vliv typu chlazení okolních struktur (skupina 1 – žádné, skupina 2 – průplach vodou) na největší rozměr poškození tkáně slinivky břišní. Zajímá nás rozdíl v efektu obou typů chlazení a jeho 95% interval spolehlivosti. Popisné statistiky naměřené na obou vzorcích jsou dány v tabulce 5.1.

Tabulka 5.1 *Popisné statistiky poškození tkáně slinivky břišní u souborů 1 a 2.*

Skupina	Počet	Výběrový průměr	Výběrová směrodatná odchylka	Směrodatná chyba
1	$n_1 = 18$	$\bar{x}_1 = 25{,}1$ mm	$s_1 = 0{,}8$	$SE_1 = 0{,}8/\sqrt{18} = 0{,}19$ mm
2	$n_2 = 17$	$\bar{x}_2 = 21{,}8$ mm	$s_2 = 2{,}4$	$SE_2 = 2{,}4/\sqrt{17} = 0{,}58$ mm

S použitím příslušného kvantilu Studentova t rozdělení, $t_{0{,}975}(\nu = 19{,}33) = 2{,}09$, a dosazením hodnot z tabulky 5.1 do vzorce (5.37) dostáváme

$$\left(25{,}1-21{,}8-2{,}09\sqrt{\tfrac{0{,}8^2}{18}+\tfrac{2{,}4^2}{17}};\ 25{,}1-21{,}8+2{,}09\sqrt{\tfrac{0{,}8^2}{18}+\tfrac{2{,}4^2}{17}}\right), \tag{5.38}$$

což znamená, že 95% interval spolehlivosti pro rozdíl středních hodnot poškození tkáně slinivky břišní u skupin 1 a 2 je následující

$$0{,}95 = P(2{,}0 \le \mu_1 - \mu_2 \le 4{,}6)\,. \tag{5.39}$$

Konstrukce 100(1 – α)% intervalu spolehlivosti pro medián

Stejně jako v případě průměru jako odhadu střední hodnoty je vhodné i medián doplnit 100(1 – α)% intervalem spolehlivosti. Jednoduchým postupem konstrukce intervalu spolehlivosti je výpočet s použitím binomického rozdělení pravděpodobnosti, který je použitelný pro sestrojení intervalu spolehlivosti pro jakýkoliv kvantil (označme jej x_π) odpovídající zvolené pravděpodobnosti π. Principem tohoto postupu je odhad odpovídajícího intervalu spolehlivosti pro střední hodnotu binomické náhodné veličiny s parametry n a π (v případě mediánu π = 0,5, n je samozřejmě velikost datového souboru) pomocí aproximace na normální rozdělení (detailní popis je uveden v kapitole 9). Tento „pomocný“ interval spolehlivosti nám pak definuje pořadí čísel tvořících spodní a horní hranici intervalu spolehlivosti pro medián v rámci pozorovaných hodnot. S pomocí této aproximace a znalostí, že střední hodnota binomické náhodné veličiny je rovna $n\pi$ a její rozptyl je roven $n\pi(1 - \pi)$, můžeme pořadí hodnot tvořících hranice 100(1 – α)% intervalu spolehlivosti (IS) pro medián vyjádřit následovně:

Pořadí spodní hranice 100(1 – α)% IS: $$i = n\pi - z_{1-\alpha/2}\sqrt{n\pi(1-\pi)} \tag{5.40}$$

Pořadí horní hranice 100(1 – α)% IS: $$j = n\pi + z_{1-\alpha/2}\sqrt{n\pi(1-\pi)} \tag{5.41}$$

Symbol $z_{1-\alpha/2}$ představuje 100(1 – α/2)% kvantil standardizovaného normálního rozdělení (pro konstrukci 95% intervalu spolehlivosti je $z_{0,975}$ = 1,96).

Ke konstrukci intervalu spolehlivosti pro medián s pomocí aproximace na binomické a potažmo normální rozdělení je nutné poznamenat dvě věci. Zaprvé, tuto aproximaci nelze použít paušálně, výpočet je vhodný pouze pro soubory s dostatečným rozsahem n. Jinými slovy, podmínkou dobré aproximace normálním rozdělením je hodnota součinu $n\pi(1 - \pi)$ větší než 5, nebo ještě lépe hodnota součinu $n\pi(1 - \pi)$ větší než 10. Zadruhé, na rozdíl od výpočtu intervalu spolehlivosti pro průměr nemusí být interval spolehlivosti pro medián symetrický, protože je tvořen dvěma hodnotami z pozorovaných dat.

Příklad 5.7. Uvažujme soubor n = 200 jedinců vybraných náhodně z obecné populace, u kterých sledujeme hladinu cholesterolu v krvi. Cílem je odhad střední hodnoty pomocí mediánu a jeho doplnění 95% intervalem spolehlivosti. Medián je ve skutečnosti 50% kvantil, což znamená, že π = 0,5. Co se týče pozorovaných hodnot, medián je vzhledem k sudému počtu pozorování průměrem z pozorovaných hodnot na 100. pozici a 101. pozici, tedy hodnot $X_{(100)}$ a $X_{(101)}$:

$$\text{Medián (v mmol/l)} = \frac{X_{(100)} + X_{(101)}}{2} = \frac{3{,}85 + 3{,}92}{2} = 3{,}89\,. \tag{5.42}$$

Výpočet pořadí spodní a horní hranice 95% intervalu spolehlivosti (pro $\alpha = 5\%$) pro medián je následující ($z_{0{,}975} = 1{,}96$):

$$\text{Pořadí spodní hranice 95\% IS:} \qquad i = 200 \times 0{,}5 - 1{,}96\sqrt{200 \times 0{,}5(1-0{,}5)} = 86{,}1 \tag{5.43}$$

$$\text{Pořadí horní hranice 95\% IS:} \qquad j = 200 \times 0{,}5 + 1{,}96\sqrt{200 \times 0{,}5(1-0{,}5)} = 113{,}9 \tag{5.44}$$

Spodní hranicí 95% intervalu spolehlivosti pro medián bude pozorovaná hodnota na 86. pozici a horní hranicí bude pozorovaná hodnota na 114. pozici. Výsledný 95% interval spolehlivosti pro medián bude tvořen hodnotami $X_{(86)}$ a $X_{(114)}$, tedy hodnotami $X_{(86)}$ = 3,57 mmol/l a $X_{(114)}$ = 4,12 mmol/l.

Konstrukce 100(1 – α)% intervalu spolehlivosti pro parametr σ^2

Opět předpokládejme náhodný výběr $X_1, ..., X_n$ z normálního rozdělení pravděpodobnosti, tedy $X_i \sim N(\mu,\sigma^2)$, $i = 1, ..., n$. Pro konstrukci 100(1 – α)% intervalu spolehlivosti pro parametr σ^2 použijeme statistiku K definovanou vztahem (5.29), která se řídí chí-kvadrát rozdělením. Pro statistiku K tedy platí

$$1-\alpha = P\left(\chi^2_{\alpha/2}(n-1) \le K \le \chi^2_{1-\alpha/2}(n-1)\right) = P\left(\chi^2_{\alpha/2}(n-1) \le \tfrac{n-1}{\sigma^2}s^2 \le \chi^2_{1-\alpha/2}(n-1)\right), \tag{5.45}$$

kde $\chi^2_{\alpha/2}(n-1)$ je 100(α/2)procentní kvantil chí-kvadrát rozdělení s n – 1 stupni volnosti. S pomocí standardních matematických operací vzorec upravíme tak, abychom parametr σ^2 mezi matematickými znaménky větší nebo rovno osamostatnili. Dostáváme tedy

$$1-\alpha = P\left(\frac{\chi^2_{\alpha/2}(n-1)}{(n-1)s^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{1-\alpha/2}(n-1)}{(n-1)s^2}\right) = P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right). \tag{5.46}$$

100(1 – α)% interval spolehlivosti pro parametr σ^2 má tedy tvar

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right). \tag{5.47}$$

Konstrukce 100(1 – α)% intervalu spolehlivosti pro podíl parametrů σ_1^2 a σ_2^2

Z praktického hlediska je užitečné uvažovat i o intervalu spolehlivosti pro podíl parametrů σ_1^2 a σ_2^2, který nám může posloužit k získání informace o tom, zda dva náhodné výběry z normálního rozdělení pravděpodobnosti vykazují podobnou variabilitu či nikoliv. Pokud interval spolehlivosti pro podíl parametrů σ_1^2 a σ_2^2 obsahuje číslo 1, tento fakt indikuje

podobnou variabilitu obou souborů, pokud ale interval spolehlivosti číslo 1 neobsahuje, budou zřejmě oba parametry, σ_1^2 a σ_2^2, rozdílné.

Uvažujme opět dva náhodné výběry, $X_{11}, \ldots, X_{1n1}$, kde $X_{1i} \sim N(\mu_1,\sigma_1^2)$, a $X_{21}, \ldots, X_{2n2}$, kde $X_{2j} \sim N(\mu_2,\sigma_2^2)$. Pro sestrojení 100(1 – α)% intervalu spolehlivosti pro podíl parametrů σ_1^2 a σ_2^2 využíváme statistiku F s Fisherovým F rozdělením pravděpodobnosti (s $n_1 - 1$ a $n_2 - 1$ stupni volnosti) definovanou jako

$$F = \frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1). \tag{5.48}$$

Obdobně jako v (5.24) a (5.45) pro statistiku F platí

$$1 - \alpha = P(F_{\alpha/2}(n_1 - 1, n_2 - 1) \le F \le F_{1-\alpha/2}(n_1 - 1, n_2 - 1)), \tag{5.49}$$

z čehož po dosazení za F a s pomocí jednoduchých úprav dostáváme 100(1 – α)% interval spolehlivosti pro podíl parametrů σ_1^2 a σ_2^2, jmenovitě pro podíl σ_2^2/σ_1^2 ve tvaru

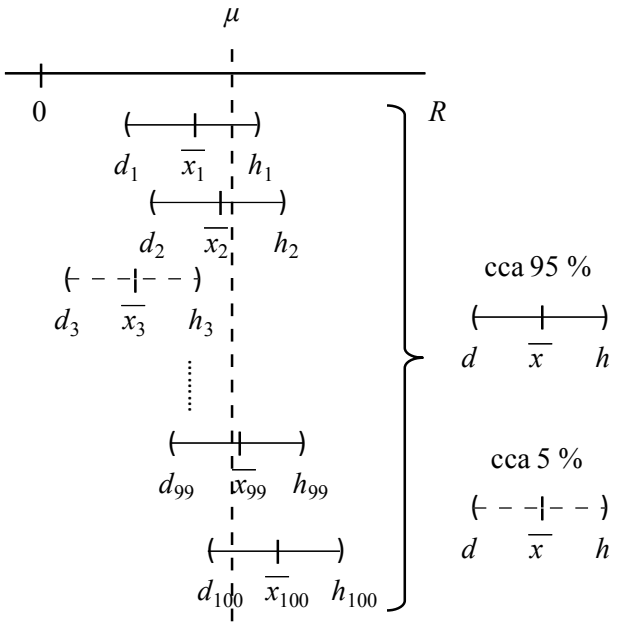
$$\left(\frac{s_2^2}{s_1^2}F_{\alpha/2}(n_1 - 1, n_2 - 1); \frac{s_2^2}{s_1^2}F_{1-\alpha/2}(n_1 - 1, n_2 - 1)\right), \tag{5.50}$$

kde $F_{\alpha/2}(n_1 - 1, n_2 - 1)$ a $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ jsou 100(α/2)% a 100(1 – α/2)% kvantily Fisherova F rozdělení, které lze najít v tabulkách nebo specializovaném software.

5.5.2 Interpretace intervalu spolehlivosti

V případě frekventistického pojetí statistiky, které je náplní těchto skript, předpokládáme, že poloha neznámého parametru je konstantní (předpokládáme tedy, že neznámý parametr není sám o sobě náhodnou veličinou). Za tohoto předpokladu má 100(1 – α)% interval spolehlivosti následující interpretaci: pokud bychom opakovali experiment za stejných podmínek, respektive opakovaně vybírali skupiny subjektů o stejné velikosti (stejném n), počítali výběrovou statistiku a sestrojovali k ní 100(1 – α)% interval spolehlivosti pro sledovaný parametr θ, pak 100(1 – α) % těchto intervalů neznámý parametr θ pokrývá (obsahuje) a 100α % ho nepokrývá (neobsahuje). Jinak řečeno, 100(1 – α)% interval spolehlivosti pokrývá neznámý parametr θ s rizikem α.

Budeme-li uvažovat např. 95% interval spolehlivosti sestrojený kolem výběrové statistiky, pak při provedení 100 experimentů za stejných podmínek a se stejnou velikostí vzorku (n) by měl alespoň v 95 případech tento interval spolehlivosti pokrývat neznámý parametr θ. Tato situace (95% interval spolehlivosti, 100 experimentů) je schematicky znázorněna na obrázku 5.5.



Obr. 5.5 Ilustrace výsledných 95% intervalů spolehlivosti při provedení 100 experimentů za stejných podmínek a se stejnou velikostí vzorku.

5.5.3 Šířka intervalu spolehlivosti

Intervaly spolehlivosti konstruované na základě různých studií budou mít jistě různou šířku. Jak již bylo uvedeno v kapitole 1, je rozdíl mezi odhadem průměrné výšky určité populace na základě 10 měření a odhadem na základě 1000 měření. Stejně tak lze očekávat různou šířku intervalu spolehlivosti ve chvíli, kdy sledujeme náhodnou veličinu (znak) s velkou variabilitou (rozptylem), a ve chvíli, kdy se zabýváme náhodnou veličinou s malou variabilitou. Šířku intervalu spolehlivosti ovlivňují tři charakteristiky, které se vyskytují ve výpočetních vzorcích. Jedná se o následující:

- ***Velikost experimentálního vzorku*** (souboru, výběru) – s rostoucí velikostí vzorku je interval spolehlivosti užší, což znamená, že náš bodový odhad je přesnější. Je to dáno tím, že máme k dispozici větší množství informace o neznámém parametru. Velikost vzorku ovlivňuje také hodnoty příslušných kvantilů rozdělení pravděpodobnosti, s rostoucím n se např. kvantily Studentova t rozdělení blíží kvantilům standardizovaného normálního rozdělení.

- ***Variabilita (rozptyl) náhodné veličiny*** – s rostoucím rozptylem sledované náhodné veličiny očekáváme i větší variabilitu bodového odhadu, což se samozřejmě odrazí v širším intervalu spolehlivosti. Příkladem je rozdělení výběrového průměru jako náhodné veličiny, které přímo závisí na rozptylu původní náhodné veličiny.

- ***Požadovaná spolehlivost intervalu*** – s rostoucí spolehlivostí, kterou požadujeme od konstruovaného intervalu, je tento interval širší, neboť požadujeme větší jistotu, že náš interval skutečně pokrývá hodnotu neznámého parametru. Stačí-li nám menší spolehlivost, pak bude interval spolehlivosti užší. Standardně je používán 95% interval spolehlivosti (odpovídající riziku α = 5 %), ale v literatuře se můžeme také setkat s 90% intervalem spolehlivosti (spokojíme-li se s rizikem α = 10 %) anebo 99% intervalem spolehlivosti (požadujeme-li naopak vysokou spolehlivost, tedy α = 1 %).

5.5.4 Neparametrické metody pro konstrukci intervalů spolehlivosti

Pro konstrukci intervalu spolehlivosti, tedy např. pro odhad variability výběrového průměru jako odhadu střední hodnoty normálního rozdělení, lze použít i neparametrické metody. Dva nejjednodušší postupy jsou založeny na metodách bootstrap a jackknife. ***Metoda bootstrap*** [6, 26] je založena na principu opakovaného vzorkování pozorovaných hodnot s vracením, kdy pro vytvoření „nového“ výběrového souboru (tzv. ***bootstrap vzorku***) může být každá hodnota z původních dat použita více než jednou, právě jednou anebo vůbec. Celý postup opakovaného vzorkování je proveden tak, aby došlo k zachování celkové velikosti souboru, n, případně i velikosti jednotlivých sledovaných skupin. Následně je na základě hodnot vybraných do bootstrap vzorku vypočten výběrový průměr. Vytvoříme-li takto např. 1000 bootstrap vzorků, můžeme následně z hodnot odpovídajících výběrových průměrů (těch je samozřejmě také 1000) sestrojit interval spolehlivosti pro původní výběrový průměr pomocí vybraných kvantilů (např. s pomocí 2,5% a 97,5% kvantilu sestrojit 95% interval spolehlivosti). ***Metoda jackknife*** [24, 26] pracuje také na principu opakovaného výpočtu sledované statistiky, ale s tím rozdílem, že pro výpočet opakovaně upravujeme původní datový soubor vynecháním vždy právě jednoho pozorování. Můžeme-li považovat původní naměřené hodnoty za reprezentativní vzorek z cílové populace, může nám tento postup stejně jako v případě metody bootstrap poskytnout představu o rozsahu hodnot, ve kterých se daná statistika může realizovat.

5.6 Shrnutí

V této kapitole jsme uvedli problematiku odhadů, a to jak bodových, kdy je naším cílem jedna číselná hodnota, tak intervalových, v jejichž případě chceme na pravděpodobnostní bázi postihnout neznámý parametr celým intervalem hodnot. Významnou úlohu v biostatistice hrají odhady pomocí metody maximální věrohodnosti, což je obecný koncept, který našel velké uplatnění v řešení mnoha biologických a medicínských problémů. Metoda maximální věrohodnosti však pracuje s jedním velmi silným předpokladem, na který se nesmí zapomínat, a tím je předpoklad konkrétního rozdělení pravděpodobnosti. Tento předpoklad je třeba vždy ověřit, metodám pro testování shody teoretického rozdělení s výběrovým rozdělením pozorovaných hodnot se věnuje část kapitoly 9.

Princip konstrukce intervalových odhadů je z velké části založen na centrální limitní větě, která popisuje pravděpodobnostní chování, rozdělení pravděpodobnosti, výběrového průměru pro soubory s velkým n. Toto rozdělení konverguje k normálnímu, a to i tehdy, když původní rozdělení náhodné veličiny není normální či dokonce spojité. Kromě konkrétních postupů pro konstrukci intervalů spolehlivosti pro parametry μ a σ^2 normálního rozdělení a intervalu spolehlivosti pro medián jsou v této kapitole diskutovány i jednotlivé charakteristiky, které zásadním způsobem ovlivňují šířku intervalu spolehlivosti.

Nakonec je třeba poznamenat, že nepřesnost bodového odhadu, kterou popisuje interval spolehlivosti, počítá pouze s variabilitou danou náhodnou veličinou, což znamená, že nepočítá se zdroji systematického zkreslení. Interval spolehlivosti tak v žádném případě nemůže postihnout systematické zkreslení dané starým měřidlem (zde mluvíme o tzv. *technical bias*), např. při měření krevního tlaku. Jiným příkladem systematického zkreslení je nereprezentativnost výběrového souboru, což může být dáno tím, že se např. do klinické studie přihlásí pouze určitá, selektovaná skupina osob (zde mluvíme o tzv. *selection bias*).

6 Úvod do testování hypotéz

V předchozí kapitole jsme se věnovali bodovým a intervalovým odhadům, které používáme k popisu jednotlivých charakteristik a parametrů náhodných veličin a jejich rozdělení pravděpodobnosti. Pokud se chceme posunout od pouhého popisu ke srovnávacím analýzám, musíme se v biostatistice přesunout k problematice ***testování hypotéz***. Pomocí statistických testů jsme schopni realizovat následující úlohy:

- Srovnat výběrovou charakteristiku jako odhad neznámého parametru θ s předpokládanou hodnotou, srovnat výběrové charakteristiky dvou náhodných výběrů mezi sebou, nebo případně vzájemně srovnat výběrové charakteristiky více náhodných výběrů.
- Hodnotit změnu v hodnotách sledované veličiny vzhledem k nějakému vnějšímu zásahu.
- Rozhodnout o nezávislosti dvou náhodných veličin.
- Rozhodnout o charakteru rozdělení pravděpodobnosti náhodné veličiny.

Klíčovou úlohu v testování hypotéz hrají samozřejmě hypotézy, což není nic jiného než tvrzení, které lze na základě pozorovaných hodnot náhodné veličiny ohodnotit ze statistického hlediska. Rozlišujeme tzv. nulovou a alternativní hypotézu. ***Nulová hypotéza*** (*null hypothesis*) je tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny (vzhledem k cílové populaci subjektů). Může být tvrzením o parametrech rozdělení nebo tvaru rozdělení pravděpodobnosti. ***Alternativní hypotéza*** (*alternative hypothesis*) je tvrzení o neznámých vlastnostech rozdělení pravděpodobnosti sledované náhodné veličiny, které popírá platnost nulové hypotézy. Vymezuje, jaká situace nastává, když nulová hypotéza neplatí. Testování hypotéz se tak zabývá rozhodováním o platnosti stanovených hypotéz na základě pozorovaných hodnot sledované náhodné veličiny. Platnost hypotéz ověřujeme pomocí ***statistického testu***, rozhodovacího pravidla, které každému náhodnému výběru (pozorovaným hodnotám náhodné veličiny) přiřadí právě jedno ze dvou možných rozhodnutí: nulovou hypotézu H_0 nezamítáme nebo naopak, nulovou hypotézu H_0 zamítáme.

Jak definovat nulovou a alternativní hypotézu ukážeme na třech klinických otázkách:

1. Urychluje použití antibiotika ve srovnání s použitím běžné dezinfekce hojení rány? Označme střední dobu hojení s antibiotiky symbolem θ_1 a střední dobu hojení bez antibiotik symbolem θ_2. Pak

Nulová hypotéza má tvar $$H_0 : \theta_1 = \theta_2 \qquad (6.1)$$

Alternativní hypotéza má tvar $$H_1 : \theta_1 < \theta_2 \qquad (6.2)$$

2. Je průměrný systolický tlak mužů nad 70 let stejný jako průměrný systolický tlak celé mužské populace? Označme střední systolický tlak mužů nad 70 let symbolem θ_1 a populační hodnotu systolického tlaku (konstantu) symbolem θ_0. Pak

Nulová hypotéza má tvar $$H_0 : \theta_1 = \theta_0 \qquad (6.3)$$

Alternativní hypotéza má tvar $H_1 : \theta_1 \neq \theta_0$ (6.4)

3. Liší se jednotlivé typy nádorového onemocnění krve v aktivitě vybraného genu? Označme střední hodnoty aktivity genu *g* u jednotlivých typů leukémie (pro zjednodušení uvažujme skupiny AML, ALL, CML, CLL) symboly $\theta^g_{AML}, \theta^g_{ALL}, \theta^g_{CML}, \theta^g_{CLL}$. Pak

Nulová hypotéza má tvar $H_0 : \theta^g_{AML} = \theta^g_{ALL} = \theta^g_{CML} = \theta^g_{CLL}$ (6.5)

Alternativní hypotéza má tvar H_1 : nejméně jedno θ^g je odlišné od ostatních (6.6)

Z uvedených příkladů si lze všimnout, že nulová hypotéza je vždy postavena jako nepřítomnost rozdílu mezi sledovanými skupinami (body 2 a 3), respektive nepřítomnost efektu léčby (bod 1). Jinak řečeno, nulová hypotéza odráží fakt, že se něco nestalo nebo neprojevilo, a je tedy stanovena jako opak toho, co chceme experimentem prokázat. Důvodem, proč nulovou hypotézu formulujeme právě takto, je skutečnost, že ji chceme pomocí pozorovaných hodnot vyvrátit. Pro zamítnutí platnosti nulové hypotézy nám totiž stačí najít jeden příklad, kdy nulová hypotéza neplatí (tím příkladem má být náš náhodný výběr, naše pozorovaná data). Zamítnutí jakékoliv hypotézy je vždy jednodušší než její potvrzení. S tím souvisí i terminologie v případě, že se nám nepodaří nulovou hypotézu vyvrátit, kdy hovoříme o případném nezamítnutí nulové hypotézy a nikoliv o přijetí nulové hypotézy.

Označme symbolem θ parametr, který nás zajímá (např. střední hodnotu sledované náhodné veličiny), a symbolem θ_0 hodnotu, se kterou chceme neznámý parametr srovnat (θ_0 může být konstanta nebo hodnota jiného neznámého parametru). Pak můžeme obě hypotézy obecně zapsat ve tvaru:

Nulová hypotéza má tvar $H_0 : \theta = \theta_0$ (6.7)

Alternativní hypotéza má jeden z tvarů $H_1 : \theta \neq \theta_0$ / $H_1 : \theta < \theta_0$ / $H_1 : \theta > \theta_0$ (6.8)

Tabulka 6.1 *Možné výsledku rozhodovacího procesu při testování statistických hypotéz.*

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítáme	správné přijetí platné nulové hypotézy	chyba II. druhu
H_0 zamítáme	chyba I. druhu	správné zamítnutí neplatné nulové hypotézy

V případě jakéhokoliv rozhodování se můžeme mýlit, a to samé platí i o testování hypotéz. Vzhledem k nulové hypotéze existují čtyři možnosti výsledku rozhodovacího procesu, které ukazuje tabulka 6.1. Dva z těchto možných výsledků, které znamenají chybný úsudek, jsou standardně označovány jako chyba I. druhu a chyba II. druhu.

Chybou I. druhu (*type I error*) označujeme falešně pozitivní závěr testu, kdy na základě výsledku testu zamítneme nulovou hypotézu, která ale ve skutečnosti platí (tedy mezi sledovanými skupinami ve skutečnosti není rozdíl, ale náš závěr na základě dat je opačný). A obdobně, ***chybou II. druhu*** (*type II error*) nazýváme zase falešně negativní závěr testu, kdy na základě výsledku testu nezamítneme nulovou hypotézu, která ale ve skutečnosti neplatí (tedy rozdíl mezi skupinami skutečně existuje, ale my ho nejsme schopni na základě dat statisticky prokázat). Příslušným výsledkům rozhodovacího procesu z tabulky 6.1 odpovídají pravděpodobnosti jejich nastání, které mají opět standardní označení, tentokrát uvedené v tabulce 6.2. Pravděpodobnost chyby I. druhu se značí α (odpovídá riziku získání falešně pozitivního výsledku), zatímco pravděpodobnost chyby II. druhu se značí β (odpovídá riziku získání falešně negativního výsledku). Při jakémkoliv testování tak máme nenulovou pravděpodobnost, že se v závěru testu mýlíme a deklarujeme opak skutečnosti. Kromě pravděpodobnosti toho, že při testování na základě dat dojdeme k chybnému závěru, je důležité vnímat i pravděpodobnost toho, že k chybnému rozhodnutí nedojde. Tedy v případě platné nulové hypotézy máme pravděpodobnost $1 - \alpha$, že tuto hypotézu nezamítneme, a v případě neplatné nulové hypotézy máme pravděpodobnost $1 - \beta$, že tuto skutečnost rozpoznáme, zamítneme H_0 a přikloníme se k alternativní hypotéze. Pravděpodobnost $1 - \beta$ se nazývá ***síla testu*** (*power of the test*) a spolu s pravděpodobností chyby I. druhu (α) je to klíčová charakteristika každého statistického testu.

Tabulka 6.2 *Možné výsledku rozhodovacího procesu a jejich příslušné pravděpodobnosti.*

Rozhodnutí	Skutečnost	
	H_0 platí	H_0 neplatí
H_0 nezamítáme	správné rozhodnutí: $P = 1 - \alpha$	chyba II. druhu: $P = \beta$
H_0 zamítáme	chyba I. druhu: $P = \alpha$	správné rozhodnutí: $P = 1 - \beta$

Testování hypotéz lze chápat i jako analogii se soudním procesem. Fakt, že nulová hypotéza odráží nepřítomnost nějakého rozdílu nebo efektu přeneseně znamená, že ctíme presumpci neviny, tedy vycházíme z toho, že obžalovaný nic neudělal (nulová hypotéza platí). Následně požadujeme důkazy pro prokázání viny, tedy důkazy pro to, že definovaný skutek, rozdíl nebo efekt skutečně existuje. Těmito důkazy není samozřejmě nic jiného než pozorované hodnoty (realizace) náhodné veličiny. Jinými slovy, na základě pozorovaných dat chceme ukázat, že nulová hypotéza neplatí.

Na analogii se soudním procesem lze demonstrovat i skutečnost, že v případě statistického testu nelze minimalizovat pravděpodobnost obou chyb (I. a II. druhu) zároveň, neboť jsou vzájemně provázané. Když nám totiž bude stačit pro usvědčení (zamítnutí hypotézy) málo důkazů, zvýší se sice procento odsouzených, kteří jsou skutečně vinni (tedy procento správně zamítnutých neplatných nulových hypotéz), ale zároveň se zvýší procento odsouzených, kteří jsou nevinní (zvýší se zastoupení chyb I. druhu). A naopak, budeme-li požadovat pro odsouzení hodně důkazů, zvýší se sice procento nevinných, kteří budou osvobozeni (tedy procento správně nezamítnutých platných nulových hypotéz), ale zároveň se

zvýší i procento viníků, kteří budou osvobozeni a nebudou potrestáni (zvýší se zastoupení chyb II. druhu).

V testování hypotéz je za důležitější považována kontrola falešně pozitivního výsledku, tedy chyby I. druhu, proto si při testování musíme nejdříve stanovit maximální možnou pravděpodobnost chyby I. druhu, kterou jsme ještě ochotni podstoupit (musíme si stanovit maximální pravděpodobnost, s jakou riskujeme falešně pozitivní výsledek). S touto hodnotou α, kterou nazýváme ***hladina významnosti testu*** (*level of significance*), pak dále pracujeme jako s pevně danou a následně k ní volíme test, který má minimální pravděpodobnost chyby II. druhu, β, tedy maximální sílu testu, $1 - \beta$. Za standardní hladiny významnosti testu jsou přijímány hodnoty $\alpha = 0{,}05$, tedy 5 %, nebo $\alpha = 0{,}01$, tedy 1 %, lze však zvolit i hladinu jinou, přísnější i méně přísnou.

6.1 Statistický test

Testování hypotéz probíhá na základě pozorovaných hodnot náhodné veličiny (dat) a statistického testu, který odpovídá testované nulové hypotéze a který nám umožní ověřit její platnost. Statistický test je reprezentován tzv. ***testovou statistikou*** (*test statistic*), což je transformace pozorovaných hodnot (náhodného výběru) pocházejících z určitého rozdělení pravděpodobnosti. To znamená, že sama testová statistika je také náhodnou veličinou a má nějaké rozdělení pravděpodobnosti. Rozdělení pravděpodobnosti testové statistiky za platnosti nulové hypotézy, H_0, lze najít v anglické literatuře pod pojmem *null distribution*.

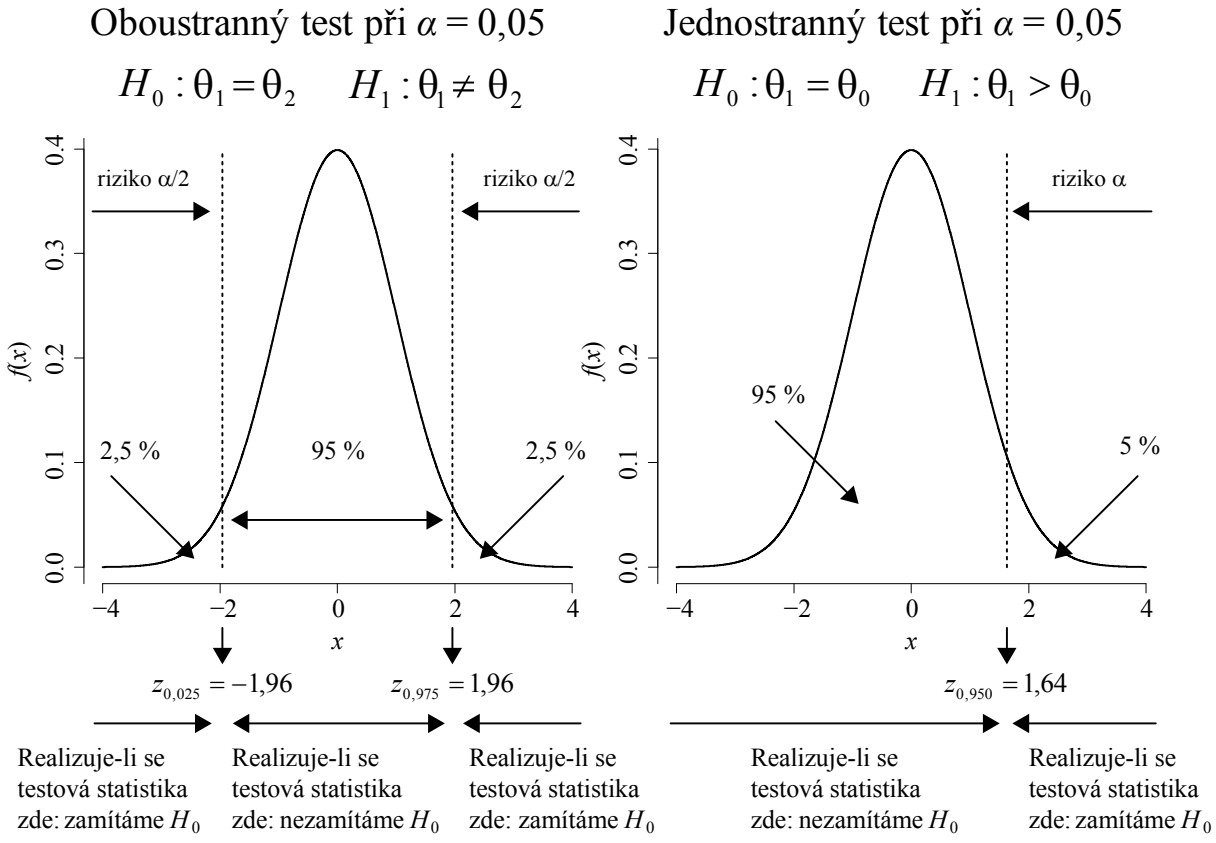
Provedení testu pak probíhá následujícím způsobem: na základě dat vypočítáme hodnotu testové statistiky, kterou srovnáme s kvantilem, často označovaným jako tzv. ***kritická hodnota***, jejího rozdělení pravděpodobnosti odpovídajícím zvolené hladině významnosti testu α. Pohybuje-li se hodnota realizace testové statistiky v rozmezí běžných hodnot daných rozdělením pravděpodobnosti testové statistiky za platnosti nulové hypotézy, H_0, tedy hodnota realizace nepřekračuje kritickou hodnotu, pak nulovou hypotézu nezamítáme. Naopak, představuje-li hodnota realizace testové statistiky extrémnější (méně pravděpodobnou) hodnotu v rámci rozdělení pravděpodobnosti odpovídajícího nulové hypotéze, než je kritická hodnota (kvantil rozdělení) odpovídající zvolenému riziku α, pak nulovou hypotézu zamítáme. Jinými slovy hodně pravděpodobné nebo běžné hodnoty realizace testové statistiky v rámci rozdělení pravděpodobnosti testové statistiky za platnosti nulové hypotézy potvrzují platnost statistické hypotézy, zatímco málo pravděpodobné až extrémní hodnoty realizace testové statistiky do tohoto rozdělení zřejmě nepatří, což naznačuje neplatnost nulové hypotézy. V souvislosti se zvolenou alternativní hypotézou riziko špatného rozhodnutí, které podstupujeme, buď rovnoměrně rozdělujeme na obě extrémní varianty výsledku (extrémně nízké i vysoké hodnoty testové statistiky) a jedná se tak o tzv. ***oboustranný test***, nebo uvažujeme pouze jednu extrémní variantu výsledku (buď extrémně nízké, nebo extrémně vysoké hodnoty testové statistiky) a jedná se tak o tzv. ***jednostranný test***. Ukázka kritických hodnot pro případ, kdy uvažujeme testovou statistiku se standardizovaným normálním rozdělením, hladinu významnosti $\alpha = 0{,}05$ a oboustrannou i jednostrannou alternativu, je uvedena na obrázku 6.1. Zde jsou pro oboustrannou alternativu kritickými hodnotami kvantily $z_{\alpha/2}$ a $z_{1-\alpha/2}$, tedy kvantily $z_{0{,}025}$ a $z_{0{,}975}$ (čísla -1,96 a 1,96), rozdělení $N(0,1)$, zatímco pro jednostrannou alternativu je kritickou hodnotou kvantil $z_{1-\alpha}$, tedy kvantil $z_{0{,}95}$ (číslo 1,64).

Fakticky realizace testové statistiky v oblasti málo pravděpodobných hodnot rozdělení pravděpodobnosti za platnosti nulové hypotézy znamená, že nastala jedna ze dvou situací:

1. H_0 platí a my jsme pozorovali málo pravděpodobný jev

2. H_0 neplatí

Pozorování málo pravděpodobných jevů máme ošetřeno rizikem α (pravděpodobností chyby I. druhu), jinými slovy málo pravděpodobné jevy jsou součástí našeho rizika, proto se v takovém případě kloníme k druhé možnosti a zamítáme H_0. Zamítáme-li nulovou hypotézu, je vždy nutné tuto informaci doplnit právě hodnotou α, tedy informací, na jaké hladině významnosti jsme test prováděli.



Obr. 6.1 Znázornění kritických hodnot pro oboustranný a jednostranný test vzhledem k riziku α.

Příklad 6.1. Při populačním epidemiologickém průzkumu bylo zjištěno, že průměrný objem prostaty u mužů je 32,73 ml (s výběrovou směrodatnou odchylkou s = 18,12 ml). Na hladině významnosti testu α = 0,05 chceme ověřit, jestli se objem prostaty u mužů nad 70 let liší od celé populace. Máme náhodný výběr o velikosti n = 100, kde byl naměřen výběrový průměr objemu prostaty 36,60 ml. Označme objem prostaty u mužů nad 70 let jako náhodnou veličinu X, střední hodnotu této veličiny symbolem μ a předpokládejme, že nemáme apriorní znalost toho, zda starší muži mají prostatu spíše větší nebo menší než mužská populace jako celek. Nulová hypotéza a příslušná oboustranná alternativní hypotéza pak mají následující tvar:

$$H_0 : \mu = 32{,}73\,, \qquad H_1 : \mu \neq 32{,}73\,. \tag{6.9}$$

Předpokládejme, že jsme v situaci, kdy víme, že výběrová směrodatná odchylka, s, zjištěná v populační studii odpovídá skutečné směrodatné odchylce σ. Za platnosti nulové hypotézy pak platí, že

$$X \sim N\left(\mu = 32{,}73, \frac{\sigma^2}{n} = \frac{18{,}12^2}{100} = 3{,}28\right). \tag{6.10}$$

Dále z centrální limitní věty víme, že platí-li (6.10), platí i následující:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - 32{,}73}{1{,}812} \sim N(0,1). \tag{6.11}$$

Pokud tedy výběrový průměr náhodné veličiny X patří do rozdělení $N(32{,}73;3{,}28)$, neměla by realizace statistiky Z být vzhledem ke standardizovanému normálnímu rozdělení nijak extrémní. Na základě pozorovaných hodnot vypočteme realizaci testové statistiky Z jako

$$z = \frac{36{,}60 - 32{,}73}{18{,}12 / \sqrt{100}} = \frac{3{,}87}{1{,}812} = 2{,}14. \tag{6.12}$$

Nyní je otázkou, můžeme zamítnout nulovou hypotézu na hladině významnosti testu $\alpha = 0{,}05$ nebo ne? Uvážíme-li zvolené riziko $\alpha = 0{,}05$, pak by se dle vztahu (5.24) měla realizace testové statistiky Z v 95 % případů pohybovat mezi kvantily $z_{\alpha/2}$ a $z_{1-\alpha/2}$, tedy hodnotami -1,96 a 1,96 (viz také obrázek 6.1 vlevo). V ideálním případě (z hlediska nulové hypotézy), pokud bychom dospěli u mužů starších 70 let ke stejnému výběrovému průměru jako v případě populační studie, by hodnota testové statistiky byla rovna nule, což je samozřejmě číslo mezi hodnotami -1,96 a 1,96. V našem případě je ale platí

$$z = 2{,}14 > 1{,}96 = z_{0{,}975} = z_{1-\alpha/2}, \tag{6.13}$$

a číslo 2,14 tak představuje extrémnější (méně pravděpodobnou) hodnotu v rámci rozdělení pravděpodobnosti odpovídajícího nulové hypotéze, než je kritická hodnota, což naznačuje neplatnost nulové hypotézy. Na hladině významnosti $\alpha = 0{,}05$ tak zamítáme nulovou hypotézu o rovnosti objemu prostaty u mužů nad 70 let populační hodnotě 32,73 ml, protože výsledná hodnota testové statistiky je větší než příslušný kvantil (kritická hodnota) standardizovaného normálního rozdělení $N(0,1)$.

6.2 *P*-hodnota a její interpretace

Místo porovnání hodnoty testovacího kritéria s kritickými hodnotami lze pro rozhodování o platnosti či neplatnosti nulové hypotézy použít i tzv. ***p-hodnotu*** (*p-value*). *P*-hodnota vyjadřuje pravděpodobnost za platnosti H_0, s níž bychom, vzhledem k jednostrannosti nebo oboustrannosti testu získali stejnou nebo extrémnější (ještě méně pravděpodobnou) hodnotu testové statistiky. Formálně lze *p*-hodnotu definovat i jako nejmenší hladinu významnosti testu, při níž na daných datech ještě zamítneme nulovou hypotézu. Platí tedy, že čím nižší *p*-hodnota testu je, tím menší nám tento test indikuje pravděpodobnost, že platí nulová hypotéza. Jinak řečeno, vyjde-li nám při vyhodnocení statistického testu *p*-hodnota „blízká nule“ (standardně jsou opět přijímány dvě hranice: 5 % a 1 %), znamená to, že naše nulová hypotéza má velmi malou oporu v pozorovaných datech a můžeme ji zamítnout.

Rozhodování o platnosti či neplatnosti nulové hypotézy tedy probíhá tak, že výslednou *p*-hodnotu testu srovnáme se zvolenou hladinou významnosti α s tím, že nulová hypotéza je zamítána ve chvíli, kdy *p*-hodnota testu klesne pod tuto hladinu. Dá se tedy říci, že ve chvíli, kdy riziko falešně pozitivního výsledku v souvislosti se zamítnutím nulové hypotézy klesne pod vybranou hladinu (např. 5 % nebo 1 %), pak ji zamítáme. Je-li tedy např. *p*-hodnota menší než 0,05, nulovou hypotézu zamítáme a hovoříme o statisticky významném výsledku na hladině významnosti $\alpha = 0{,}05$. Rozhodujeme-li o platnosti nulové hypotézy pomocí *p*-hodnoty, lze *p*-hodnotu chápat jako číselný indikátor platnosti nebo neplatnosti nulové hypotézy vyjádřený na pravděpodobnostní škále. A jako každý indikátor, může i *p*-hodnota indikovat špatný výsledek, neboť si stále musíme uvědomovat, že nám hrozí jak chyba I. druhu, tak chyba II. druhu.

Příklad 6.2. Vraťme se k příkladu 6.1, kde jsme získali výslednou hodnotu testové statistiky $z = 2{,}14$. Otázkou je, jaká jí odpovídá *p*-hodnota? Důležité je si uvědomit, že máme oboustrannou alternativní hypotézu, což znamená, že extrémnější (méně pravděpodobné) hodnoty testové statistiky v rámci rozdělení pravděpodobnosti odpovídajícího nulové hypotéze jsou jak hodnoty vyšší než 2,14, tak hodnoty nižší než -2,14. Do pravděpodobnosti, kterou *p*-hodnota představuje tak musíme načíst jak pravděpodobnost výskytu vysokých hodnot testové statistiky, tak pravděpodobnost výskytu nízkých hodnot testové statistiky. Výslednou *p*-hodnotu pro oboustrannou alternativu lze tedy vyjádřit následovně

$$p = 2 * (1 - P(Z \leq z)), \tag{6.14}$$

kde z je pozorovaná hodnota testové statistiky a $P(Z \leq z)$ označuje hodnotu distribuční funkce standardizovaného normálního rozdělení v bodě z. Výpočet *p*-hodnoty pro hodnotu testové statistiky $z = 2{,}14$ z příkladu 6.1 je

$$p = 2 * (1 - P(Z \leq 2{,}14)) = 2 * (1 - 0{,}984) = 0{,}032. \tag{6.15}$$

Výsledná hodnota 0,032 je menší než zvolené α a opět tudíž můžeme říci, že na hladině významnosti $\alpha = 0{,}05$ zamítáme nulovou hypotézu o rovnosti objemu prostaty u mužů nad 70 let populační hodnotě 32,73 ml.

Na testování hypotéz má zásadní vliv velikost výběrového souboru, jinými slovy, množství informace, na jejímž základě rozhodujeme o platnosti nulové hypotézy. Nejlépe lze tento fakt opět ilustrovat příkladem.

Příklad 6.3. Opět se vrátíme k příkladu 6.1, ale budeme uvažovat výběrový soubor mužů starších 70 let o velikosti pouze $n = 10$ jedinců (ostatní charakteristiky zůstanou beze změny). Hypotézy uvedené v (6.9) taktéž zůstávají stejné. Vzhledem k $n = 10$ ale víme, že rozdělení výběrového průměru musí být nutně jiné (opět předpokládáme znalost σ^2), a to

$$X \sim N\left(\mu = 32{,}73, \frac{\sigma^2}{n} = \frac{18{,}12^2}{10} = 32{,}8\right). \tag{6.16}$$

Když na základě pozorovaných hodnot vypočteme realizaci testové statistiky Z jako

$$z = \frac{36{,}60 - 32{,}73}{18{,}12/\sqrt{10}} = \frac{3{,}87}{5{,}73} = 0{,}68, \tag{6.17}$$

a srovnáme ji s příslušným kvantilem:

$$z = 0{,}68 < 1{,}96 = z_{0{,}975} = z_{1-\alpha/2}\,, \tag{6.18}$$

pak v tomto případě nulovou hypotézu o rovnosti objemu prostaty u mužů nad 70 let populační hodnotě 32,73 ml nezamítáme. Důležité je si uvědomit, že rozdíl v objemu prostaty pozorovaný mezi populační hodnotou a muži staršími 70 let se nezměnil, jediné, co se změnilo, je množství informace, ze které čerpáme, tedy velikost výběrového souboru.

6.3 Poznámky k testování hypotéz

Problematika testování hypotéz je velmi široká a v rozsahu těchto skript ji nelze obsáhnout. Některé aspekty testování jsou však natolik významné, že je nemůžeme nechat alespoň bez stručné zmínky. Těmito aspekty jsou: souvislost testování hypotéz s konstrukcí intervalu spolehlivosti, vztah statistické a praktické významnosti dosaženého výsledku, faktory ovlivňující sílu testu a problematika násobného testování hypotéz.

6.3.1 Spojitost testování hypotéz s intervaly spolehlivosti

Spojitost testování hypotéz s intervaly spolehlivosti lze opět nejlépe demonstrovat na příkladu s objemem mužské prostaty (příklad 6.1), kde jsme na základě výběrového souboru o velikosti n = 100 zamítli nulovou hypotézu $H_0 : \mu = 32{,}73$ proti $H_1 : \mu \neq 32{,}73$. Vypočtěme 95% interval spolehlivosti pro μ (tedy interval spolehlivosti s α = 0,05). Vycházíme ze statistiky Z, následujících charakteristik

$$Z = \tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\,, \qquad \bar{X} = 36{,}60\,, \qquad \tfrac{\sigma}{\sqrt{n}} = \tfrac{18{,}12}{\sqrt{100}} = 1{,}812\,, \qquad z_{0{,}975} = 1{,}96\,, \tag{6.19}$$

a vzorce

$$1 - 0{,}05 = P(-z_{0{,}975} \le \tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{0{,}975}) = P(\bar{X} - \tfrac{\sigma}{\sqrt{n}} z_{0{,}975} \le \mu \le \bar{X} + \tfrac{\sigma}{\sqrt{n}} z_{0{,}975})\,. \tag{6.20}$$

Výsledkem je pak 95% interval spolehlivosti (33,05; 40,15). Tento interval neobsahuje nulovou hypotézu, jinými slovy, tento interval neobsahuje předpokládanou hodnotu 32,73 ml. Fakt, že výsledný 95% interval spolehlivosti neobsahuje hodnotu neznámého parametru stanovenou v H_0, znamená, že můžeme H_0 zamítnout. Opět platí, že podstupujeme riziko α = 0,05, že se mýlíme, tedy že jsme naším 95% intervalem spolehlivosti nepokryli hodnotu neznámého parametru μ.

Testování hypotéz a intervaly spolehlivosti jsou velmi často výpočetně ekvivalentní, nicméně oba tyto přístupy sledují jiný cíl. Konstrukce intervalů spolehlivosti má za cíl charakterizovat přesnost bodového odhadu neznámého parametru, zatímco test nulové hypotézy se zaměřuje na hodnocení platnosti pravděpodobnostního modelu, který popisuje chování náhodné veličiny.

Každopádně z praktického hlediska je podstatné, aby v každé studii byla vždy vedle výsledku testu (rozhodnutí o platnosti H_0) publikována i velikost dosaženého rozdílu (efektu)

s příslušným intervalem spolehlivosti. Ze samotné *p*-hodnoty zvoleného testu nebo rozhodnutí zamítáme H_0 / nezamítáme H_0 totiž není zřejmé, v jakých mezích se pozorovaná velikost rozdílu (účinku) pohybuje.

6.3.2 Statistická a praktická významnost

Rozhodnutí o zamítnutí/nezamítnutí nulové hypotézy je vlastně rozhodnutím o tzv. ***statistické významnosti*** rozdílu dvou nebo více výběrových souborů ve sledované náhodné veličině, případně rozdílu jednoho souboru od předem dané konstantní hodnoty. Velmi často se však při testování stává, že je zanedbána praktická interpretace dosaženého výsledku, např. rozdílu v délkách nebo efektu léčby. Praktická využitelnost pozorovaných hodnot odpovídá naopak tzv. ***věcné*** (***praktické***, ***klinické***, ***biologické***) ***významnosti*** výsledku, která ale nemusí vždy odpovídat významnosti statistické. Vzhledem k tomu, že testování statistických hypotéz vždy provádíme kvůli možnosti zobecnění z náhodného výběru na celou populaci, je ověření interpretační hodnoty výsledku minimálně stejně významné jako vlastní výpočet testu. Statistická významnost totiž nutně nemusí znamenat existenci příčinného vztahu, respektive dosažení malé výsledné *p*-hodnoty nemusí znamenat dosažení velkého rozdílu ve sledované náhodné veličině, např. efektu léčby. Statistická významnost pouze indikuje, že pozorovaný rozdíl není ve smyslu stanovené hypotézy náhodný, což ale, jak bylo vidět v příkladu 6.3, lze ovlivnit velikostí vzorku.

Absolutní velikost efektu je při srovnání sledovaných skupin subjektů měřitelná například jako rozdíl výběrových průměrů. Pro stanovení, jaký dosažený efekt je ale zároveň i věcně podstatný, neexistuje žádné univerzální pravidlo, neboť vše závisí na konkrétní situaci, měřené veličině a cílech výzkumu. V jedné situaci může být za podstatný považován efekt, který v jiném kontextu podstatný není. Nastavení vždy musí provádět člověk znalý věci, který čerpá ze znalosti problému nebo z informací dostupných z literatury.

Příklad 6.4. Předpokládejme, že standardní léčba vysokého tlaku (hypertenze) snižuje systolický tlak (TKs) v průměru o 20 milimetrů rtuťového sloupce (mm Hg). Tuto hodnotu budeme pro jednoduchost nyní považovat za střední hodnotu v populaci. Z klinického hlediska by věcně významné bylo zvýšení účinnosti o dalších 10 mm Hg (klinicky významné zlepšení účinku). Za významné tedy považujeme snížení TKs novou léčbou o 30 mm Hg. Jak může nová léčba hypertenze dopadnout z hlediska statistické a praktické významnosti sumarizuje tabulka 6.3.

Interpretace výsledků je následující. Možnosti a), b) i e) nesplňují měřítko statistické významnosti, neboť 95% interval spolehlivosti obsahuje (připouští) populační hodnotu, kterou je snížení systolického tlaku o 20 milimetrů rtuťového sloupce. Možnost e) navíc nesplňuje ani měřítko praktické významnosti, neboť ani bodový odhad účinku (22,9 mm Hg) ani horní hranice intervalu spolehlivosti (27,5 mm Hg) nepřekračují hranici pro klinickou významnost, kterou je snížení TKs o 30 mm Hg. Možnosti a) a b) splňují měřítko praktické významnosti pouze možná, neboť interval spolehlivosti připouští hodnoty účinku jak pod 30 mm Hg, tak nad 30 mm Hg. Statisticky významné výsledky představují možnosti c), d), a f), kde vidíme, že interval spolehlivosti připouští pouze snížení systolického tlaku o více než 20 milimetrů rtuťového sloupce, což znamená, že nová léčba hypertenze účinkuje lépe než standardní léčba. Nicméně hranici stanovenou pro praktickou významnost překračuje pouze možnost d), kde jak bodový odhad účinku (36,2 mm Hg), tak spodní hranice 95% intervalu spolehlivosti (32,1 mm Hg) jsou větší než 30 mm Hg.

Tabulka 6.3 *Možné výsledky klinického experimentu a jejich významnost.*

Možnost	Statistická vs. klinická významnost	
a)	V průměru došlo ke snížení TKs o 24,7 mm Hg, ale byla pozorována taková variabilita v účinku, že 95% interval spolehlivosti pro výběrový průměr byl (16,5; 32,9).	Statistická významnost: ne Praktická významnost: možná
b)	V průměru došlo ke snížení TKs o 30,1 mm Hg, ale byla pozorována taková variabilita v účinku, že 95% interval spolehlivosti pro výběrový průměr byl (19,6; 40,6).	Statistická významnost: ne Praktická významnost: možná
c)	V průměru došlo ke snížení TKs o 31,5 mm Hg, ale byla pozorována taková variabilita v účinku, že 95% interval spolehlivosti pro výběrový průměr byl (26,0; 37,0).	Statistická významnost: ano Praktická významnost: možná
d)	**V průměru došlo ke snížení TKs o 36,2 mm Hg a byla pozorována taková variabilita v účinku, že 95% interval spolehlivosti pro výběrový průměr byl (32,1; 39,3).**	**Statistická významnost: ano Praktická významnost: ano**
e)	V průměru došlo ke snížení TKs o 22,9 mm Hg, ale byla pozorována taková variabilita v účinku, že 95% interval spolehlivosti pro výběrový průměr byl (18,3; 27,5).	Statistická významnost: ne Praktická významnost: ne
f)	V průměru došlo ke snížení TKs o 25,1 mm Hg, ale byla pozorována taková variabilita v účinku, že 95% interval spolehlivosti pro výběrový průměr byl (21,6; 28,6).	Statistická významnost: ano Praktická významnost: ne

6.3.3 Faktory ovlivňující sílu testu

Síla testu byla definována v úvodu této kapitoly jako pravděpodobnost, že zamítneme H_0 ve chvíli, kdy H_0 opravdu neplatí. Jedná se tedy o správné rozhodnutí a jeho pravděpodobnost se standardně označuje jako $1 - \beta$ (doplněk k pravděpodobnosti chyby II. druhu). Vzhledem k tomu, že je pro nás v testování hypotéz důležitější pravděpodobnost chyby I. druhu (α), snažíme se sílu testu optimalizovat (ideálně maximalizovat) při současném zachování hladiny α. Optimalizace síly testu je hlavním cílem odhadu velikosti experimentálního vzorku před provedením studie, kdy se snažíme zjistit, kolik je třeba experimentálních subjektů (pozorování) k tomu, aby měl výsledný test dostatečnou sílu k zamítnutí nulové hypotézy, bude-li tato hypotéza skutečně neplatná. Ptát se dopředu na velikost výběrového souboru má skutečně smysl, neboť se chceme vyvarovat situace, kdy pro zamítnutí neplatné nulové hypotézy nemáme dostatečné množství informace. Nezamítnutí nulové hypotézy by totiž nemělo automaticky znamenat její přijetí, v řadě případů se totiž jedná pouze o situaci, kdy nejsme schopni na základě pozorovaných hodnot neplatnost nulové hypotézy prokázat. Optimalizovat sílu testu a velikost vzorku před začátkem experimentu však není triviální, tento proces je spojen s řadou faktorů, které nelze ovlivnit: např. biologické limity (nelze zařadit do studie více pacientů, než kolik jich onemocní v určitém území za daný čas), nebo finanční limity (jakýkoliv experiment stojí peníze a jejich zdroje jsou vždy omezeny).

Faktory ovlivňující sílu testu jsou následující:

- ***Velikost výběrového souboru*** (velikost vzorku): čím více pozorování náhodné veličiny máme k dispozici, tedy čím více máme informace o platnosti nulové hypotézy, tím větší má test sílu. Růst síly testu s velikostí souboru však není lineární, konkrétní podoba tohoto vztahu závisí na konkrétním použitém testu. Efekt rostoucí velikosti souboru je opět stejný jako u intervalů spolehlivosti, čím více máme pozorování, tím je naše schopnost identifikovat skutečnou hodnotu (skutečnost zda platí nulová hypotéza) lepší.

- ***Velikost pozorovaného rozdílu*** (efektu, účinku): velikost rozdílu ve sledované veličině také ovlivňuje sílu testu. Pro statistický test je vždy jednodušší identifikovat jako statisticky významný velký rozdíl (např. rozdíl ve výšce mužů a žen) a naopak, je těžší prokázat jako statisticky významný malý rozdíl (např. rozdíl ve výšce populací Čechů a Slováků).

- ***Variabilita dat*** reprezentovaná rozptylem náhodné veličiny: větší rozptyl sledované náhodné veličiny zvyšuje variabilitu odhadu neznámého parametru, čímž ztěžuje i rozhodnutí o platnosti nulové hypotézy. Čím více jsou pozorované hodnoty variabilní, tím více jich bude potřeba pro přesný odhad skutečného rozdílu mezi skupinami.

- ***Hladina významnosti testu***: standardně testujeme nulovou hypotézu na hladině významnosti $\alpha = 0{,}05$. Snížíme-li hladinu významnosti, tedy zvolíme-li např. hladinu $\alpha = 0{,}01$, zamítnout nulovou hypotézu bude obtížnější, což znamená, že se sníží síla testu. Naopak zvýšení hladiny významnosti (což je ale spojeno s vyšším rizikem získání falešně pozitivního výsledku) znamená zvýšení síly testu.

6.3.4 Problém násobného testování hypotéz

V klinickém výzkumu se často setkáváme se situací, kdy potřebujeme testovat více hypotéz zároveň. Nemusí to nutně znamenat hodnocení různých výběrových souborů nebo náhodných veličin, ale např. i hodnocení stejné veličiny v rámci různých podskupin celého výběrového souboru. Když např. sledujeme rozdíl v nějaké veličině u souboru pacientů se skupinami A, B, C a D, a zjistíme, že se v celkovém pohledu sledované skupiny liší, je samozřejmě z jakéhokoliv hlediska zajímavé podívat se na tento rozdíl i mezi jednotlivými podskupinami, tedy podívat se, jak se liší skupina A od B, B od C, apod. Tento fenomén však v praxi vede k tzv. ***problému násobného testování hypotéz*** [23]. Ten spočívá v tom, že s narůstajícím počtem testovaných hypotéz nám roste také pravděpodobnost získání falešně pozitivního výsledku, tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje.

Můžeme si představit modelovou situaci, kdy provedeme zároveň 60 testů, což v době běžného provádění biochemických a genetických experimentů není zase tolik. Použijeme-li standardní hladinu významnosti $\alpha = 0{,}05$, máme pro každý test 5% riziko získání falešně pozitivního výsledku. Vynásobíme-li 60 a 0,05, vyjde nám, že zhruba u 3 testů bychom měli dospět k falešně statisticky významnému závěru. V případě genomických analýz, kde jsou často různé testy pouze formou exploratorní a popisné analýzy, nemusí být přítomnost falešně pozitivních výsledků fatální, v klinické praxi to však může vést k zavádějícím výsledkům a mylným interpretacím. Z tohoto důvodu je nutné při násobném statistickém testování uvažovat tzv. ***korekční procedury***, které by měly brát v úvahu celkový počet provedených testů.

Nejznámější korekční procedurou pro násobné testování hypotéz je ***Bonferroniho procedura*** [10], která zamítá nulovou hypotézu ve chvíli, kdy je její p-hodnota menší nebo rovna hodnotě α/m, kde α je zvolená hladina významnosti testu (obvykle 0,05 nebo 0,01), a m je počet zároveň provedených testů. Použití Bonferroniho procedury je poměrně konzervativní, což znamená, že je při jejím použití relativně obtížné dosáhnout statistické významnosti (zvláště když je počet provedených testů větší než 10). Korekčních procedur však existuje celá řada, z metod pro parametrické testy lze jmenovat např. Scheffého metodu či Tukeyho metodu, pro neparametrické testy pak např. metodu dle Steela a Dwasse [5, 28].

6.4 Shrnutí

Kapitola 6 se zabývá rozhodováním o platnosti statistických hypotéz na základě vybraného modelu a pozorovaných dat. Statistické hypotézy nejsou nic jiného než tvrzení, které lze na základě pozorovaných hodnot pomocí statistických metod ohodnotit. Rozlišujeme nulovou a alternativní hypotézu, kdy nulová hypotéza je tvrzení, které je vždy postaveno jako nepřítomnost rozdílu mezi sledovanými skupinami. Jinak řečeno, nulová hypotéza odráží fakt, že se něco nestalo nebo neprojevilo, a je tedy stanovena jako opak toho, co chceme experimentem prokázat. Alternativní hypotéza je pak tvrzení, které popírá platnost nulové hypotézy. Nulovou hypotézu ověřujeme pomocí statistického testu, kdy na základě pozorovaných dat počítáme realizaci testové statistiky, která má za platnosti nulové hypotézy známé rozdělení pravděpodobnosti. Rozhodování o přijetí nebo zamítnutí nulové hypotézy je spojeno s dvěma typy chyb. Ty jsou standardně označovány jako chyba I. druhu (její pravděpodobnost značíme jako α) a chyba II. druhu (její pravděpodobnost značíme jako β). Pravděpodobnost chyby I. druhu souvisí s falešně pozitivním závěrem testu, kdy na základě výsledku testu zamítneme nulovou hypotézu, která ale ve skutečnosti platí. Podobně, pravděpodobnost chyby II. druhu souvisí zase s falešně negativním závěrem testu, kdy na základě výsledku testu nezamítneme nulovou hypotézu, která ale ve skutečnosti neplatí. V biostatistice je za důležitější považována chyba I. druhu, kterou se snažíme omezit na přijatelné minimum. Jako standardní hranice, které potom představují riziko falešné pozitivity podstoupené při testování, jsou přijímány hladiny 5 % nebo 1 %. Ať již provádíme test pomocí kritického kvantilu nebo pomocí *p*-hodnoty, zjednodušeně můžeme říci, že ve chvíli, kdy riziko falešně pozitivního výsledku v souvislosti se zamítnutím nulové hypotézy klesne pod vybranou hladinu (5 % nebo 1 %), pak ji zamítáme. Zároveň je však třeba si uvědomit, že při jakémkoliv testování máme nenulovou pravděpodobnost, že se v závěru testu mýlíme a deklarujeme opak skutečnosti.

Testování hypotéz si lze z laického hlediska představit jako číselný indikátor platnosti nebo neplatnosti nulové hypotézy, který můžeme vyjádřit na pravděpodobnostní škále pomocí *p*-hodnoty. A jako každý indikátor, může i *p*-hodnota dát špatný výsledek, což znamená, že při posouzení pravdivosti testované hypotézy můžeme dojít ke špatnému závěru. Právě proto nesmí být nikdy stanovená hladina významnosti testu (ať už máme $\alpha = 0{,}05$, $\alpha = 0{,}01$ nebo $\alpha = 0{,}10$) slepě brána jako hranice pro existenci/neexistenci testovaného efektu. Neexistuje totiž jasná hranice pro praktickou významnost či nevýznamnost a velmi často je pouze malý rozdíl mezi *p*-hodnotou rovnou 0,04 a *p*-hodnotou rovnou 0,06. Navíc malá *p*-hodnota nemusí nutně znamenat velký efekt. Hodnota testové statistiky a odpovídající *p*-hodnota totiž může být ovlivněna velkou velikostí vzorku a malou variabilitou pozorovaných dat.

7 Testování hypotéz o kvantitativních proměnných

V této kapitole se budeme věnovat testování hypotéz o spojitých náhodných veličinách, tedy obecně těch, které mohou nabývat jakýchkoliv hodnot v určitém intervalu. Klasickými příklady jsou výška postavy, hmotnost jedince, nebo časové a teplotní měření. Použití uvedených testů lze zvažovat i v testování hypotéz o diskrétních náhodných veličinách, ale pouze v případě, že je to odůvodnitelné velkým počtem hodnot, kterých může daná veličina nabývat. Příkladem může být počet červených krvinek v 1 ml krve, což samozřejmě není spojitá náhodná veličina, ale počet možných hodnot je natolik velký, že nás opravňuje použít pro testování hypotéz testy pro spojité veličiny.

Z hlediska předpokladů, které jednotlivé testy kladou na testovanou náhodnou veličinu, lze testy hypotéz rozdělit na dvě velké skupiny, a to tzv. ***parametrické testy*** a ***neparametrické testy***. Parametrické testy, jak již samotné označení napovídá, se zabývají testováním tvrzení o neznámých parametrech rozdělení pravděpodobnosti, kterým se uvažovaná náhodná veličina řídí. Co se týče předpokladů, parametrické testy jsou obecně náročnější než neparametrické, neboť vyžadují minimálně specifikaci daného rozdělení pravděpodobnosti (specifikace rozdělení pravděpodobnosti jako modelu podstaty chování náhodné veličiny je samo o sobě velmi silný předpoklad). Neparametrické testy jsou naopak nezávislé nebo téměř nezávislé na konkrétních rozděleních pravděpodobnosti a vyžadují slabší předpoklady, nevyžadují např. normalitu rozdělení pravděpodobnosti, ale pouze jeho symetrii. Na druhou stranu mají neparametrické testy menší sílu, což je následně nutné kompenzovat větší velikostí vzorku. Obecně lze ale říci, že neparametrické testy jsou „bezpečnější“ než testy parametrické, neboť testování hypotézy v případě chybně určeného rozdělení pravděpodobnosti parametrické testové statistiky může vést k mylným závěrům z důvodu nerelevantní p-hodnoty, respektive p-hodnoty stanovené chybnou úvahou.

Formálně lze popsat postup při statistickém testování takto:

1. Formulujeme nulovou hypotézu H_0, která předpokládá neexistenci nějakého rozdílu, např. neexistenci efektu léčby. Zároveň zvolíme hladinu významnosti testu α, která představuje pravděpodobnost získání falešně pozitivního výsledku.

2. Formulujeme alternativní hypotézu H_1, která u parametrických testů může být jak oboustranná, tak jednostranná.

3. Zvolíme adekvátní testovou statistiku T jako kritérium pro rozhodnutí o nulové hypotéze. Testovou statistiku volíme tak, abychom znali rozdělení pravděpodobnosti této statistiky při platnosti nulové hypotézy a byli tak schopni posoudit, jak extrémní je výsledná hodnota testové statistiky v rámci tohoto rozdělení.

4. Vypočítáme hodnotu testové statistiky T na základě n realizací náhodné veličiny X, tedy na základě pozorovaných hodnot $x_1, x_2, \dots, x_n$.

5. Na základě rozdělení pravděpodobnosti testové statistiky T a zvolené hladiny významnosti určíme kritický obor neboli obor hodnot, v němž zamítáme H_0 a přikláníme se k platnosti H_1.

6. Zjistíme, zda hodnota testové statistiky T leží v oboru kritických hodnot. Pokud ano, zamítáme nulovou hypotézu a přikláníme se k platnosti alternativní hypotézy, pokud ne,

nezamítáme nulovou hypotézu. Alternativně tomuto postupu můžeme zjistit *p*-hodnotu výsledku a srovnat ji se zvolenou hladinou významnosti testu.

7.1 Testy o parametrech jednoho rozdělení

Pointou testů o parametrech jednoho rozdělení pravděpodobnosti je srovnání výběrové charakteristiky vypočtené na základě pozorovaných dat (ty reprezentují odhad sledované charakteristiky náhodné veličiny) s předem danou hodnotou (konstantou). Testujeme tak, zda se výběrové charakteristiky našeho datového souboru shodují s předpokládanou hodnotou.

Základní testy o parametrech jednoho rozdělení jsou následující:

- Test o střední hodnotě při známém rozptylu (*z-test pro jeden výběr*)
- Test o střední hodnotě při neznámém rozptylu (*t-test pro jeden výběr*)
- Neparametrický test pro jeden výběr (*Wilcoxonův test*)
- Test o rozdílu párových (závislých) pozorování (*párový t-test*)
- Test o rozptylu normálního rozdělení

Kromě samotného rozhodnutí o platnosti nulové hypotézy je pro korektní popis provedeného experimentu a pozorovaných dat nezbytné vypočítat i adekvátní interval spolehlivosti pro výběrovou charakteristiku sumarizující sledovanou náhodnou veličinu (výběrový průměr nebo výběrový rozptyl). Jen tak jsme si totiž schopni udělat komplexní obrázek jak o dosažené statistické významnosti, tak o dosažené praktické významnosti.

7.1.1 Test o střední hodnotě při známém rozptylu (*z*-test pro jeden výběr)

Cílem *z*-testu pro jeden výběr je testovat hypotézu, zda data náhodného výběru pochází z rozdělení se stejnou střední hodnotou, jako je předpokládaná hodnota μ_0 (konstanta). Vycházíme z realizace náhodného výběru o rozsahu n: $x_1, x_2, \ldots, x_n$, o kterém předpokládáme, že pochází z normálního rozdělení. Předpokládáme tedy, že platí $X_i \sim N(\mu,\sigma^2)$. Dále předpokládáme, že známe hodnotu parametru σ^2. Oba tyto předpoklady jsou velmi silné (a v biologické či klinické praxi téměř nereálné), neboť to znamená, že téměř přesně známe pravděpodobnostní chování náhodné veličiny X. Nulová hypotéza a příslušné alternativní hypotézy (oboustranná a jednostranné) pak mají následující tvar

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0 \qquad H_1 : \mu > \mu_0 \qquad H_1 : \mu < \mu_0 \qquad (7.1)$$

Výběrovou charakteristikou, která hraje v *z*-testu hlavní roli je samozřejmě výběrový průměr. Víme totiž, že za platnosti H_0 má výběrový průměr také normální rozdělení, což znamená, že platí

$$X \sim N(\mu_0, \sigma^2/n). \qquad (7.2)$$

Z toho plyne, že testová statistika *Z*, kterou dostaneme z výběrového průměru standardizací, má standardizované normální rozdělení:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)\,. \tag{7.3}$$

Pokud nulová hypotéza platí, statistika *Z* se bude realizovat v hodnotách běžných pro rozdělení $N(0,1)$, a naopak, neplatí-li nulová hypotéza, statistika *Z* se bude realizovat v hodnotách, které nejsou pro standardizované normální rozdělení běžné. Nulovou hypotézu tak zamítáme na hladině významnosti α ve chvíli, kdy výsledná hodnota statistiky *Z* je větší (nebo menší, v závislosti na předem zvolené alternativě) než příslušný kvantil (kritická hodnota) rozdělení $N(0,1)$. Co znamená realizace statistiky v běžných hodnotách, bylo blíže rozebráno v kapitole 6, v případě oboustranného testu na hladině významnosti α by se měla testová statistika *Z* pohybovat mezi kvantily $z_{\alpha/2}$ a $z_{1-\alpha/2}$, což pro $\alpha = 0{,}05$ jsou hodnoty -1,96 a 1,96. Bude-li se statistika *Z* realizovat mimo tento interval, zamítáme nulovou hypotézu (na hladině významnosti $\alpha = 0{,}05$, samozřejmě). Vzhledem k symetrii kvantilů rozdělení $N(0,1)$ lze pravidlo pro zamítnutí H_0 pro oboustrannou alternativu u *z*-testu zjednodušit na vyjádření, kdy absolutní hodnota statistiky *Z* překročí hodnotu kvantilu $z_{1-\alpha/2}$, tedy $|Z| > z_{1-\alpha/2}$. Souhrnně jsou pravidla pro zamítnutí nulové hypotézy pro *z*-test pro jeden výběr dle zvolené alternativy uvedena v tabulce 7.1. Příklad na výpočet *z*-testu pro jeden výběr byl uveden v kapitole 6.

Tabulka 7.1 *Pravidla pro zamítnutí H_0 pro z-test pro jeden výběr dle zvolené alternativy.*

Alternativa	$H_1 : \mu \neq \mu_0$	→ Zamítáme H_0, když	$\lvert Z \rvert > z_{1-\alpha/2}$
Alternativa	$H_1 : \mu > \mu_0$	→ Zamítáme H_0, když	$Z > z_{1-\alpha}$
Alternativa	$H_1 : \mu < \mu_0$	→ Zamítáme H_0, když	$Z < z_{\alpha}$

7.1.2 Test o střední hodnotě při neznámém rozptylu (*t*-test pro jeden výběr)

Cíl *t*-testu pro jeden výběr [9, 36, 37] je stejný jako u *z*-testu, tedy také chceme testovat hypotézu, zda data náhodného výběru pochází z rozdělení se stejnou střední hodnotou jako je předpokládaná konstanta μ_0. Stejný je i předpoklad, že data pochází z normálního rozdělení, tedy že platí $X_i \sim N(\mu,\sigma^2)$. Rozdíl mezi oběma testy je v tom, že u *t*-testu pro jeden výběr nepředpokládáme znalost parametru σ, což znamená, že pro testování nemůžeme jednoduše použít výše uvedenou statistiku *Z*. Abychom se zbavili nutnosti specifikovat parametr σ, je třeba definovat statistiku *K* tak, že

$$K = \frac{n-1}{\sigma^2} s^2\,. \tag{7.4}$$

Statistika *K* má chí-kvadrát rozdělení pravděpodobnosti s $(n - 1)$ stupni volnosti, tedy $K \sim \chi^2(n-1)$. Statistiky *Z* a *K* použijeme ke konstrukci statistiky *T*:

$$T = \frac{Z}{\sqrt{K/(n-1)}} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}. \tag{7.5}$$

Statistika T již neobsahuje neznámý parametr σ, který je nahrazen jeho výběrovým odhadem ve formě výběrové směrodatné odchylky, s. Lze ukázat, že statistika T má Studentovo t rozdělení pravděpodobnosti s $(n - 1)$ stupni volnosti, tedy $T \sim t(n-1)$. Pravidla pro zamítnutí nulové hypotézy na základě výsledné hodnoty statistiky T (dle zvolené alternativy a hladiny významnosti testu) jsou pro t-test pro jeden výběr obdobná jako pro z-test pro jeden výběr pouze s tím rozdílem, že jako kritické hodnoty používáme příslušné kvantily Studentova t rozdělení s parametrem $(n - 1)$. Pravidla pro zamítnutí nulové hypotézy platná pro t-test pro jeden výběr dle zvolené alternativy jsou uvedena v tabulce 7.2.

Tabulka 7.2 *Pravidla pro zamítnutí H_0 pro t-test pro jeden výběr dle zvolené alternativy.*

Alternativa	$H_1 : \mu \neq \mu_0$	→ Zamítáme H_0, když	$\lvert T \rvert > t_{1-\alpha/2}^{(n-1)}$
Alternativa	$H_1 : \mu > \mu_0$	→ Zamítáme H_0, když	$T > t_{1-\alpha}^{(n-1)}$
Alternativa	$H_1 : \mu < \mu_0$	→ Zamítáme H_0, když	$T < t_{\alpha}^{(n-1)}$

Příklad 7.1. Pomocí t-testu pro jeden výběr chceme srovnat průměrný denní energetický příjem skupiny 11 žen ve věku 22 – 30 let s doporučenou populační hodnotou, kterou je 7725 kJ (hodnoty převzaty z [2]). Pozorovaný průměrný energetický příjem skupiny 11 žen byl 6753,6 kJ se směrodatnou odchylkou s = 1142,1 kJ. Předpokládejme, že nemáme představu o stravovacích návycích mladých žen, proto zvolíme oboustrannou alternativu. Nulová hypotéza, H_0, a jí příslušná oboustranná alternativa, H_1, pak mají tvar

$$H_0 : \mu = \mu_0 = 7725, \qquad H_1 : \mu \neq \mu_0 = 7725. \tag{7.6}$$

K ověření nulové hypotézy použijeme testovou statistiku T, která je dána vztahem (7.5). Výpočet realizace testové statistiky T tedy znamená dosazení výběrových charakteristik do (7.5) a je následující:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{6753{,}6 - 7725}{1142{,}1/\sqrt{11}} = -2{,}821. \tag{7.7}$$

Vzhledem k tomu, že alternativní hypotéza je oboustrannou alternativou, pro rozhodnutí o platnosti H_0 je třeba srovnat absolutní hodnotu realizace testové statistiky, tedy číslo 2,821, se 100(1 – α/2)procentním kvantilem t rozdělení s $n - 1$ (tedy 10) stupni volnosti, což je hodnota 2,228. V souladu s tabulkou 7.2 platí, že

$$\lvert t \rvert = 2{,}821 > 2{,}228 = t_{0{,}975}^{10} = t_{1-\alpha/2}^{n-1}, \tag{7.8}$$

a tedy zamítáme H_0 na hladině významnosti $\alpha = 0{,}05$. Jinými slovy, na hladině významnosti $\alpha = 0{,}05$ můžeme říci, že sledovaná skupina žen měla statisticky významně odlišný (nižší) denní energetický příjem, než je doporučená hodnota 7725 kJ.

7.1.3 Neparametrický test pro jeden výběr (Wilcoxonův test)

Oba předchozí testy o střední hodnotě, *z*-test i *t*-test, jsou parametrické testy vyžadující předpoklad normality dat, který se následně odráží v nulové i alternativní hypotéze. Tento předpoklad je však velmi silný a v praxi často není splněn. V řadě případů, spojených zejména s malou velikostí výběrového souboru, dokonce ani nejsme schopni normalitu dat korektně ověřit. Neparametrickou alternativou *z*-testu a *t*-testu pro jeden výběr je Wilcoxonův test [31, 36], který není testem o střední hodnotě, ale testem o mediánu, a jeho jediným předpokladem je symetrie rozdělení náhodné veličiny X, z něhož pochází náhodný výběr. Nulová hypotéza Wilcoxonova testu se týká mediánu rozdělení sledované náhodné veličiny a spolu s oboustrannou alternativou ji lze zapsat jako

$$H_0 : \tilde{x} = x_0 \qquad H_1 : \tilde{x} \neq x_0 \tag{7.9}$$

Princip Wilcoxonova testu je velmi jednoduchý, test v podstatě hodnotí, zda je přibližně polovina hodnot $x_1, x_2, \ldots, x_n$ menších než předpokládaná hodnota x_0 a přibližně polovina hodnot $x_1, x_2, \ldots, x_n$ větších než tato konstanta s tím, že předpokládá obdobné kolísání hodnot nalevo i napravo od mediánu (předpoklad symetrie). Při samotném výpočtu Wilcoxonův test převádí pozorované hodnoty $x_1, x_2, \ldots, x_n$ na diference od x_0, tedy na hodnoty y_i, $i = 1, \ldots, n$ definované jako

$$y_i = x_i - x_0 , \tag{7.10}$$

které jsou následně seřazeny podle velikosti absolutních hodnot od nejmenší diference po největší:

$$|y_{(1)}| < |y_{(2)}| < \ldots < |y_{(n)}| . \tag{7.11}$$

Jednotlivým diferencím y_i je potom na základě tohoto seřazení přiřazeno pořadí, označme ho jako R_i. Samotná testová statistika Wilcoxonova testu je založena pouze na těchto pořadích a je definována jako

$$\min(S^+, S^-) , \tag{7.12}$$

kde veličiny S^+ a S^- spočítáme jako součty pořadí

$$S^+ = \sum_{y_i > 0} R_i , \qquad S^- = \sum_{y_i < 0} R_i . \tag{7.13}$$

V případě, že pozorované hodnoty jsou symetricky rozděleny kolem předpokládané hodnoty x_0, bude přibližně jedna polovina diferencí kladná a druhá záporná. Navíc absolutní hodnoty kladných diferencí nebudou systematicky větší než absolutní hodnoty záporných diferencí a naopak, což ve výsledku znamená, že součet pořadí příslušný kladným diferencím bude přibližně stejný jako součet pořadí příslušný záporným diferencím. Za platnosti H_0 tak lze předpokládat, že hodnoty S^+ a S^- budou zhruba vyrovnané. Na druhou stranu, ve chvíli, kdy H_0 nebude platit, bude mezi hodnotami S^+ a S^- rozdíl, kdy jedna z těchto statistik bude malé číslo a druhá velké číslo (pojem malé a velké číslo je zde závislý na velikosti souboru).

Pro rozhodnutí o platnosti H_0 je pak testová statistika Wilcoxonova testu, $\min(S^+,S^-)$, srovnána s kritickou hodnotou příslušnou dané velikosti výběrového souboru a zvolené hladině významnosti testu α. Je-li hodnota $\min(S^+,S^-)$ menší nebo rovna kritické hodnotě, zamítáme H_0 o rovnosti mediánu sledované náhodné veličiny předpokládané hodnotě x_0 (spadne-li hodnota minima obou statistik pod určitou mez, ukazuje to na statisticky významný rozdíl mezi S^+ a S^- a tudíž i na neplatnost H_0). Pro malá n (cca do 30) lze kritickou hodnotu pro statistiku $\min(S^+,S^-)$ odpovídající zvolené hladině významnosti α najít v tabulkách, pro větší n lze rozdělení testové statistiky $\min(S^+,S^-)$ aproximovat normálním rozdělením s následující střední hodnotou a rozptylem:

$$E(\min(S^+,S^-)) = n(n+1)/4\,, \qquad D(\min(S^+,S^-)) = n(n+1)(2n+1)/24\,. \tag{7.14}$$

Jak je vidět z výpočtu, Wilcoxonův test pracuje místo pozorovaných hodnot s pořadími, což je postup robustní vůči odlehlým hodnotám, které by v případě použití z-testu nebo t-testu pro jeden výběr mohly zásadním způsobem ovlivnit hodnotu výběrového průměru. Obecně samozřejmě platí, že parametrické a neparametrické testy nemusí vycházet stejně. Důvody mohou být především nesplnění předpokladů parametrického testu nebo menší síla neparametrického testu. Na druhou stranu, je-li dobře specifikován pravděpodobnostní model a máme-li k dispozici dostatek dat, výsledky parametrických i neparametrických testů budou stejné.

Příklad 7.2. Stejně jako v příkladu 7.1 budeme srovnávat denní energetický příjem skupiny 11 žen ve věku 22 – 30 let s doporučenou hodnotou 7725 kJ s tím, že pro srovnání použijeme Wilcoxonův test. Nulová a alternativní hypotéza jsou vyjádřeny následovně

$$H_0 : \tilde{x} = 7725\,, \qquad H_1 : \tilde{x} \neq 7725\,. \tag{7.15}$$

Pozorované hodnoty, diference od referenční hodnoty 7725 kJ a příslušná pořadí jsou znázorněna v tabulce 7.3 (hodnoty převzaty z [2]). Na základě pořadí absolutních hodnot kladných a záporných diferencí vypočítáme následující hodnoty pomocných statistik a testové statistiky

$$S^+ = \sum_{y_i>0} R_i = 8\,, \qquad S^- = \sum_{y_i<0} R_i = 58\,, \qquad \min(S^+,S^-) = 8\,. \tag{7.16}$$

Výslednou hodnotu testové statistiky srovnáme s kritickou hodnotou $w_n(\alpha)$ příslušnou velikosti souboru, n = 11, a hladině významnosti testu α = 0,05, která je v tomto případě $w_{11}(0{,}05) = 10$. Vzhledem k tomu, že realizace testové statistiky, číslo 8, je menší než hodnota 10, zamítáme nulovou hypotézu o tom, že medián energetického příjmu žen ve věku 22 – 30 let je roven 7725 kJ za den.

Tabulka 7.3 *Denní energetický příjem skupiny 11 žen ve věku 22 – 30 let.*

Žena	Denní energetický příjem v kJ	Diference od hodnoty 7725 kJ	Pořadí absolutní hodnoty diference
1	5260	-2465	11
2	5470	-2255	10
3	5640	-2085	9
4	6180	-1545	8
5	6390	-1335	7
6	6515	-1210	6
7	6805	-920	4
8	7515	-210	1,5
9	7515	-210	1,5
10	8230	505	3
11	8770	1045	5

7.1.4 Test o rozdílu párových (závislých) pozorování (*párový t-test*)

Samostatným problémem v biostatistice je hodnocení párových pozorování, která jsou vzájemně závislá, respektive vázaná nějakým společným prvkem. Klasickým příkladem párových pozorování jsou hodnoty dvou po sobě jdoucích měření na stejném pacientovi, které samozřejmě nelze považovat za nezávislé, neboť jsou vázány osobou pacienta [38]. Cílem testu o rozdílu párových pozorování, párového *t*-testu, je ověřit, zda se střední hodnoty náhodných veličin X a Y liší o předem danou hodnotu d_0. Předpokládáme tedy realizaci dvourozměrného náhodného vektoru o rozsahu n s tím, že u veličin X a Y předpokládáme normální rozdělení:

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}, \qquad \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} \right). \tag{7.17}$$

Nulová hypotéza a příslušné alternativní hypotézy (oboustranná a jednostranné) pak mají následující tvar

$$H_0 : \mu_1 - \mu_2 = d_0, \qquad H_1 : \mu_1 - \mu_2 \neq d_0, \qquad H_1 : \mu_1 - \mu_2 < d_0, \qquad H_1 : \mu_1 - \mu_2 > d_0. \tag{7.18}$$

I když vlastně uvažujeme sledování dvou náhodných veličin, tak párový *t*-test patří do této kapitoly, neboť výpočetně převádíme párový problém na případ jednoho výběru. To znamená, že výpočet párového *t*-testu nepočítá s dvojicemi hodnot, ale s jejich rozdíly d_i, $i = 1, \ldots, n$ definovanými jako

$$d_i = x_i - y_i. \tag{7.19}$$

Následně testujeme, zda je průměr hodnot $d_1, d_2, \ldots, d_n$ různý od předpokládané hodnoty d_0. Za předpokladu normality diferencí d_i, tedy za předpokladu, že platí $D_i \sim N(\mu_d,\sigma^2)$, to znamená, že dále postupujeme jako při *t*-testu pro jeden výběr. Testová statistika má tvar

$$T = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}, \tag{7.20}$$

kde $\bar{d}$ značí průměr pozorovaných diferencí a s_d jejich výběrovou směrodatnou odchylku. Stejně jako v případě *t*-testu pro jeden výběr má statistika T Studentovo t rozdělení pravděpodobnosti s $n - 1$ stupni volnosti; nulovou hypotézu, H_0, proto zamítáme na hladině významnosti α, když je realizace statistiky T větší nebo menší než kritická hodnota (příslušný kvantil) Studentova rozdělení $t(n - 1)$. Pravidla pro zamítnutí nulové hypotézy pro párový *t*-test dle zvolené alternativy přehledně sumarizuje tabulka 7.4.

Tabulka 7.4 *Pravidla pro zamítnutí H_0 pro párový t-test dle zvolené alternativy.*

Alternativa	$H_1 : \mu_1 - \mu_2 \neq d_0 \leftrightarrow H_1 : \mu_d \neq d_0$	→ Zamítáme H_0, když	$\lvert T \rvert > t_{1-\alpha/2}^{(n-1)}$
Alternativa	$H_1 : \mu_1 - \mu_2 > d_0 \leftrightarrow H_1 : \mu_d > d_0$	→ Zamítáme H_0, když	$T > t_{1-\alpha}^{(n-1)}$
Alternativa	$H_1 : \mu_1 - \mu_2 < d_0 \leftrightarrow H_1 : \mu_d < d_0$	→ Zamítáme H_0, když	$T < t_{\alpha}^{(n-1)}$

7.2 Testy o parametrech dvou rozdělení

Testy o parametrech dvou rozdělení pravděpodobnosti nám umožňují srovnat výběrové odhady sledované charakteristiky náhodné veličiny ve dvou nezávislých experimentálních souborech. Testujeme tak, zda se výběrové charakteristiky ve dvou nezávislých skupinách liší nebo neliší.

Základní testy o parametrech dvou rozdělení jsou následující:

- Test o rozdílu středních hodnot dvou nezávislých výběrů při stejných rozptylech (*t-test pro dva výběry*)
- Welchova korekce pro *t*-test při nestejných rozptylech
- Test o shodnosti rozptylů dvou nezávislých výběrů – *F*-test
- Neparametrický test pro 2 výběry – Mannův-Whitneyho test

Stejně jako v případě testů pro jeden výběr by měly i v případě testů pro dva výběry být spolu s výsledkem testu, tedy rozhodnutím o platnosti nulové hypotézy a případně *p*-hodnotou, reportovány i příslušné intervaly spolehlivosti pro pozorované rozdíly, případně podíly odhadovaných parametrů rozdělení pravděpodobnosti.

7.2.1 Test o rozdílu středních hodnot dvou nezávislých výběrů při stejných rozptylech (*t-test pro dva výběry*)

Základním testem pro srovnávání středních hodnot dvou nezávislých výběrů je v biostatistice *t*-test pro dva výběry [9, 36, 37], který testuje, zda náhodné výběry pochází z rozdělení se středními hodnotami, jejichž rozdíl je daná konstanta c. Umožňuje nám tak

posoudit, zda se hodnoty náhodné veličiny v jedné populaci statisticky významně liší od hodnot této náhodné veličiny v populaci druhé. Jedná se o parametrický test, jehož hlavním předpokladem je normalita rozdělení pravděpodobnosti obou náhodných výběrů. Máme-li realizaci prvního náhodného výběru o rozsahu n_1: $x_1, x_2, \ldots, x_{n1}$, a na ní nezávislou realizaci druhého náhodného výběru o rozsahu n_2: $y_1, y_2, \ldots, y_{n2}$, předpokládáme, že jak realizace x_i, tak realizace y_j pocházejí z normálního rozdělení, tedy že platí $X_i \sim N(\mu_1,\sigma^2)$, $i = 1, \ldots, n_1$, a $Y_j \sim N(\mu_2,\sigma^2)$, $j = 1, \ldots, n_2$. Nulová hypotéza, předpokládající rozdíl mezi středními hodnotami roven c (nejčastěji volíme $c = 0$), a příslušné alternativní hypotézy (oboustranná a jednostranné) mají tvar

$$H_0 : \mu_1 - \mu_2 = c\,, \qquad H_1 : \mu_1 - \mu_2 \neq c\,, \qquad H_1 : \mu_1 - \mu_2 < c\,, \qquad H_1 : \mu_1 - \mu_2 > c\,. \tag{7.21}$$

Je důležité si uvědomit, že jsme opět v situaci, kdy neznáme skutečnou hodnotu parametru σ^2, pouze předpokládáme, že je stejná pro oba výběry. Tento neznámý parametr odhadujeme pomocí váženého průměru odhadů rozptylu (výběrových rozptylů) v jednotlivých skupinách, s_1^2 a s_2^2:

$$s_*^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\,. \tag{7.22}$$

Z vlastností normálního rozdělení pravděpodobnosti plyne, že rozdíl průměrů normálních náhodných veličin X a Y je také normální náhodná veličina. Platí tedy

$$X - Y \sim N(c, \sigma^2(\tfrac{1}{n_1} + \tfrac{1}{n_2}))\,. \tag{7.23}$$

Vzhledem k neznámému parametru σ^2 nelze použít pro testování statistiku s normálním rozdělením pravděpodobnosti, proto obdobně jako v případě t-testu pro jeden výběr i zde hraje roli testové statistiky statistika T se Studentovým t rozdělením (s $n_1 + n_2 - 2$ stupni volnosti). Pro dva výběry je statistika T definována jako

$$T = \frac{X - Y - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)\,. \tag{7.24}$$

Nulovou hypotézu opět zamítáme na hladině významnosti α ve chvíli, kdy realizace statistiky T překročí určitou hranici, kterou je kvantil Studentova rozdělení $t(n_1 + n_2 - 2)$ příslušný hladině α a zvolené alternativě. Souhrn pravidel pro zamítnutí nulové hypotézy platných pro t-test pro dva výběry dle zvolené alternativy je uveden v tabulce 7.5. Kromě pravidel pro rozhodnutí o platnosti H_0 je třeba mít na paměti, že použití t-testu pro dva výběry má dva velmi silné předpoklady, kterým bychom měli před výpočtem vždy věnovat adekvátní pozornost. Těmito předpoklady jsou

1. ***Normalita pozorovaných hodnot***, a to v rámci obou náhodných výběrů. Předpoklad normality musíme předem otestovat adekvátním testem (více v kapitole 8) nebo alespoň graficky ověřit pomocí dostupných vizualizačních nástrojů (histogram, krabicový graf).

2. ***Homogenní*** (stejný) ***rozptyl náhodné veličiny***, opět v rámci obou srovnávaných výběrů. Předpoklad homogenity rozptylu lze stejně jako normalitu testovat příslušným statistickým testem (tomuto tématu je věnována část 7.2.2 o tzv. *F*-testu), možné je i grafické ověření pomocí výše zmíněných nástrojů (histogram, krabicový graf).

Tabulka 7.5 *Pravidla pro zamítnutí H_0 pro t-test pro dva výběry dle zvolené alternativy.*

Alternativa	$H_1 : \mu_1 - \mu_2 \neq c$	→ Zamítáme H_0, když	$\lvert T \rvert > t_{1-\alpha/2}^{(n_1+n_2-2)}$
Alternativa	$H_1 : \mu_1 - \mu_2 > c$	→ Zamítáme H_0, když	$T > t_{1-\alpha}^{(n_1+n_2-2)}$
Alternativa	$H_1 : \mu_1 - \mu_2 < c$	→ Zamítáme H_0, když	$T < t_{\alpha}^{(n_1+n_2-2)}$

Příklad 7.3. Uvažujme léčbu pacientů se špatně kontrolovanou hypertenzí, pro kterou je dostupná léčba tzv. ACE inhibitory (ACE-I) a antagonisty pro angiotensin II receptor (AIIA). Účinnost léčby ACE-I u pacientů se špatně kontrolovanou hypertenzí reprezentujeme náhodnou veličinou *X*, zatímco účinnost léčby AIIA u těchto pacientů popíšeme náhodnou veličinou *Y*. Nulová hypotéza pak vyjadřuje stejný účinek obou léků (ve smyslu střední hodnoty) na snížení diastolického tlaku (TKd) těchto pacientů měřený v milimetrech rtuti po šesti měsících od zahájení léčby. Tedy

$$H_0 : \mu_1 - \mu_2 = 0, \qquad H_1 : \mu_1 - \mu_2 \neq 0. \tag{7.25}$$

U pacientů léčených ACE-I (skupina 1), respektive AIIA (skupina 2), byly pozorovány následující výběrové charakteristiky:

$$n_1 = 1926; \bar{x} = 12{,}7; s_1 = 9{,}96, \qquad n_2 = 1887; \bar{y} = 12{,}8; s_2 = 9{,}79. \tag{7.26}$$

Dále byl na základě hodnot s_1 a s_2 vypočten vážený odhad parametru σ, $s_* = 9{,}88$. Víme, že za platnosti H_0 platí $\bar{X} - \bar{Y} \sim N(0, \sigma^2(\frac{1}{1926} + \frac{1}{1887}))$, což znamená, že můžeme pro testování použít statistiku *T* definovanou v (7.24). Po dosazení získáme

$$t = \frac{\bar{x} - \bar{y} - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{12{,}7 - 12{,}8 - 0}{9{,}88 \sqrt{\frac{1}{1926} + \frac{1}{1887}}} = -0{,}31. \tag{7.27}$$

Absolutní hodnotu výsledné realizace testové statistiky srovnáme s kvantilem Studentova *t* rozdělení s 3811 stupni volnosti (vzhledem k platnosti centrální limitní věty zde již můžeme použít kvantil rozdělení $N(0,1)$). Absolutní hodnota testové statistiky je menší než hodnota kvantilu $z_{1-\alpha/2} = z_{0,975} = 1{,}96$ a tedy nulovou hypotézu nezamítáme. Závěrem tedy lze říci, že na hladině významnosti $\alpha = 0{,}05$ nelze prokázat rozdíl mezi léčbou ACE-I a AIIA vzhledem ke snížení diastolického tlaku u pacientů se špatně kontrolovanou hypertenzí.

7.2.2 Test o shodnosti (homogenitě) rozptylů dvou nezávislých výběrů (*F-test*)

Cílem *F*-testu o rovnosti dvou rozptylů je ověřit, zda dva výběrové soubory pochází z rozdělení se stejným rozptylem, což znamená ověřit, zda oba soubory vykazují přibližně stejný rozptyl sledované náhodné veličiny. Předpokladem tohoto testu je normalita pozorovaných hodnot v obou výběrových souborech, tedy předpokládáme, že platí $X_i \sim N(\mu_1,\sigma_1^2)$, $i = 1, \ldots, n_1$, a $Y_j \sim N(\mu_2,\sigma_2^2)$, $j = 1, \ldots, n_2$. Nulovou hypotézu a příslušné alternativy pak zapíšeme jako

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2 \qquad H_1 : \sigma_1^2 < \sigma_2^2 \qquad H_1 : \sigma_1^2 > \sigma_2^2 \qquad (7.28)$$

Testová statistika *F*-testu využívá pro ověření nulové hypotézy informaci uloženou ve výběrových rozptylech a má tvar

$$F = \frac{s_1^2}{s_2^2}. \qquad (7.29)$$

Tato statistika má za platnosti H_0 Fisherovo *F* rozdělení s parametry $(n_1 - 1)$ a $(n_2 - 1)$, což zapisujeme jako $F \sim F(n_1 - 1, n_2 - 1)$. Pro rozhodnutí o platnosti nulové hypotézy srovnáme hodnotu realizace statistiky *F* s kvantily *F* rozdělení příslušnými hladině významnosti testu, parametrům a zvolené alternativě. Pravidla pro zamítnutí nulové hypotézy platná pro *F*-test dle zvolené alternativy jsou uvedena v tabulce 7.6.

Tabulka 7.6 *Pravidla pro zamítnutí H_0 pro F-test dle zvolené alternativy.*

Alternativa	$H_1 : \sigma_1^2 \neq \sigma_2^2$	→ Zamítáme H_0, když	$F < F_{\alpha/2}^{(n_1-1,n_2-1)}$ nebo $F > F_{1-\alpha/2}^{(n_1-1,n_2-1)}$
Alternativa	$H_1 : \sigma_1^2 > \sigma_2^2$	→ Zamítáme H_0, když	$F > F_{1-\alpha}^{(n_1-1,n_2-1)}$
Alternativa	$H_1 : \sigma_1^2 < \sigma_2^2$	→ Zamítáme H_0, když	$F < F_{\alpha}^{(n_1-1,n_2-1)}$

Příklad 7.5. Sledujeme dvě skupiny dětí s hypotyreózou, první skupinou jsou děti s mírnými symptomy, druhá skupina jsou děti s výraznými symptomy. Chceme u těchto dvou skupin srovnat hladinu tyroxinu v séru. Před použitím testu pro dva výběry si musíme položit následující otázku: Můžeme si dovolit použít *t*-test pro dva výběry ve chvíli, kdy je jedním z jeho předpokladů homogenita rozptylů ve sledovaných skupinách? Na zodpovězení této otázky použijeme *F*-test o rovnosti dvou rozptylů na hladině významnosti $\alpha = 0{,}05$ s tím, že proti nulové hypotéze použijeme jednostrannou alternativu – předpokládáme totiž, že děti s výraznými symptomy budou vykazovat větší variabilitu v hodnotách tyroxinu v séru. Naměřené výběrové charakteristiky jsou uvedeny v tabulce 7.7.

Tabulka 7.7 *Výběrové charakteristiky skupin pacientů s hypotyreózou.*

Hladina tyroxinu v séru (nmol/l)	Mírné symptomy ($n_1 = 9$)	Výrazné symptomy ($n_2 = 7$)
Průměr	56,4	42,1
Směrodatná odchylka	14,22	37,48

Výpočet testové statistiky je následující:

$$F = \frac{s_1^2}{s_2^2} = \frac{(14{,}22)^2}{(37{,}48)^2} = 0{,}144\,. \tag{7.30}$$

V souladu s tabulkou 7.6 zamítáme H_0, když realizace statistiky F bude nižší než kvantil Fisherova F rozdělení pro $\alpha = 0{,}05$ a parametry $n_1 - 1 = 8$ a $n_2 - 1 = 6$. Vzhledem k tomu, že platí

$$F = 0{,}144 < 0{,}279 = F_{0{,}05}^{(8,6)} = F_{\alpha}^{(n_1-1,n_2-1)}, \tag{7.31}$$

zamítáme na hladině významnosti $\alpha = 0{,}05$ nulovou hypotézu o shodě rozptylů měření hladiny tyroxinu v séru u dětí s mírnými symptomy a dětí s výraznými symptomy. Obě skupiny dětí se tedy statisticky významně liší ve variabilitě hladin tyroxinu v séru.

7.2.3 Welchova korekce pro *t*-test při nestejných rozptylech

Předpoklad stejných rozptylů sledované veličiny v obou srovnávaných souborech je v praxi téměř nereálný. Proto Welch v roce 1938 [30] navrhl korekci pro výpočet statistiky T se zohledněním nestejné variability skupinových pozorování. V případě nestejných rozptylů víme, že v souladu se vztahem (7.23) platí $\bar{X} - \bar{Y} \sim N(c, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, což vede na testovou statistiku T ve tvaru

$$T = \frac{\bar{X} - \bar{Y} - c}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu)\,. \tag{7.32}$$

Ze vztahu (7.32) plyne kromě jiného výrazu pro standardní chybu měření, která nyní obsahuje obě výběrové směrodatné odchylky, i fakt, že počet stupňů volnosti Studentova t rozdělení, již není roven $n_1 + n_2 - 2$, ale je třeba ho stanovit následovně

$$\nu = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}\,. \tag{7.33}$$

Kritické hodnoty pro zamítnutí H_0 jsou pak odvozeny stejně jako v případě t-testu pro dva výběry se stejným rozptylem.

7.2.4 Neparametrický test pro dva výběry (*Mannův-Whitneyho test*)

Mannův-Whitneyho test [14, 36] je neparametrickou alternativou t-testu pro dva výběry ve chvíli, kdy není splněn některý z jeho předpokladů, respektive máme-li o platnosti některého z jeho předpokladů pochyby. Nulová hypotéza Mannova-Whitneyho testu není

zaměřena na střední hodnoty, ale místo toho předpokládáme stejné rozdělení pravděpodobnosti náhodné veličiny v obou souborech, což je slabší předpoklad než normalita dat. Nulová hypotéza se tak týká srovnatelnosti dvou distribučních funkcí, kterou zapíšeme jako

$$H_0 : F(x) = F(y), \qquad H_1 : F(x) \neq F(y). \tag{7.34}$$

Mějme realizaci prvního náhodného výběru o rozsahu n_1: x_1, x_2, … , x_{n1}, a na ní nezávislou realizaci druhého náhodného výběru o rozsahu n_2: y_1, y_2, … , y_{n2}. Pointa výpočtu Mannova-Whitneyho testu je následující: pokud pozorování x_i a y_j ($i = 1, \ldots, n_1$; $j = 1, \ldots, n_2$) pochází ze stejného rozdělení pravděpodobnosti, pak by pravděpodobnost toho, že náhodně vybraná hodnota x_i bude větší než náhodně vybraná hodnota y_j ($P(x_i > y_j)$) měla být 50 %. To je ekvivalentní tomu, že při srovnání všech dostupných dvojic x_i a y_j bude v případě cca 50 % těchto dvojic větší hodnota x_i a naopak.

Pro výpočet nejprve seřadíme všechna pozorování od nejmenšího po největší tak, jako by byly z jednoho vzorku, a přiřadíme jednotlivým hodnotám jejich pořadí. Symbolem T_1 označíme součet pořadí hodnot příslušných první skupině. Testovými statistikami pak jsou statistiky U a U', definované jako

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1, \qquad U' = n_1 n_2 - U. \tag{7.35}$$

Pro rozhodnutí o platnosti nulové hypotézy srovnáme větší z hodnot U a U' s kritickou hodnotou z tabulek (v případě oboustranného testu). Je-li kritická hodnota menší, H_0 zamítáme. Pro jednostranný test uvažujeme dle nulové hypotézy pouze buď statistiku U nebo U'. Pro výběrové soubory o velikosti $n_1 > 10$ a zároveň $n_2 > 10$ lze rozdělení pravděpodobnosti testové statistiky U aproximovat normálním rozdělením s charakteristikami

$$E(U) = n_1 n_2 / 2, \qquad D(U) = n_1 n_2 (n_1 + n_2 + 1)/12, \tag{7.36}$$

což znamená, že pro ověření nulové hypotézy lze dosadit uvedené hodnoty do statistiky Z a její realizaci srovnat s příslušným kvantilem standardizovaného normálního rozdělení $N(0,1)$.

Příklad 7.6. Opět uvažujme dvě skupiny dětí s hypotyreózou z příkladu 7.5. První skupina jsou děti s mírnými symptomy, druhá skupina jsou děti s výraznými symptomy, naším cílem je srovnat u těchto dvou skupin hladinu tyroxinu v séru. T-test pro dva výběry není pro tento účel vhodný, neboť obě skupiny vykazují různý rozptyl sledované náhodné veličiny. Seřadíme-li všechna pozorování podle velikosti a přiřadíme jednotlivým hodnotám jejich pořadí, dojdeme k tomu, že součet pořadí v první skupině, tedy hodnota statistiky T_1, je roven 84,5. Toto číslo dosadíme do vztahu (7.35) a vypočteme

$$U = 9*7 + \frac{9(9+1)}{2} - 84{,}5 = 63 + 45 - 84{,}5 = 23{,}5, \qquad U' = 9*7 - 23{,}5 = 39{,}5. \tag{7.37}$$

Jako realizace testové statistiky slouží větší z vypočtených U a U', tedy číslo 39,5, které srovnáme s kritickou hodnotou ze statistických tabulek příslušnou hladině významnosti testu α. Vzhledem k tomu, že platí

$$\max(U, U') = 39{,}5 < 51 = U^{(9,7)}_{0{,}05(2)} = U^{(n_1,n_2)}_{\alpha(1/2)}, \tag{7.38}$$

nezamítáme nulovou hypotézu o shodě distribučních funkcí, z nichž pochází měření tyroxinu v séru u dvou skupin dětí s hypotyreózou. Tento výsledek je na první pohled relativně překvapivý, nicméně je třeba si uvědomit, že oba výběrové soubory jsou velmi malé a test tak zřejmě nemá dostatečnou sílu na to, aby odhalil rozdíl v hodnotách tyroxinu mezi oběma skupinami.

7.3 Shrnutí

Kapitola 7 uvádí přehled základních parametrických a neparametrických testů pro testování hypotéz o jednom, respektive dvou výběrových souborech. Zásadním rozdílem mezi parametrickými a neparametrickými testy je nutnost předpokladu o pravděpodobnostním chování náhodné veličiny nebo veličin, které pozorujeme (zde se jedná o silný předpoklad normality dat). Normalita dat je velmi spekulativní, zejména u menších výběrových souborů, nicméně v případě parametrických testů je třeba tento předpoklad vždy ověřit, např. s pomocí grafických metod, abychom se alespoň ujistili, že normalita není zásadně porušena.

Alternativu v podobě neparametrických testů také nelze použít úplně libovolně, např. pro Wilcoxonův test pro jeden výběr bychom měli, opět alespoň graficky, ověřit přibližnou symetrii výběrového rozdělení pozorovaných hodnot. Velkou výhodou neparametrických testů je fakt, že pracují s pořadími hodnot, což znamená, že téměř úplně ignorují odlehlá pozorování a případné chybné hodnoty. Na druhou stranu trpí neparametrické metody sníženou sílou testu, tedy sníženou schopností zamítnout neplatnou nulovou hypotézu, což je nepříjemné zejména v případě menších výběrových souborů, kde je pak testování hypotéz pomocí neparametrických testů většinou bez jasného závěru, kdy např. pozorujeme prakticky zajímavý rozdíl mezi sledovanými skupinami, který však vzhledem k omezené velikosti výběru nelze prokázat jako statisticky významný. Na tomto místě je třeba znovu připomenout, že statisticky nevýznamný výsledek nemusí znamenat, že pozorovaný rozdíl ve skutečnosti neexistuje, neboť se může jednat právě pouze o nedostatečnou velikost výběrového souboru. Problematikou vyváženosti velikosti vzorku potřebné pro korektní provedení experimentu vzhledem k nulové hypotéze a výše definovaným chybám I. a II. druhu se zabývá oblast statistiky nazvaná ***plánování experimentů***, která má v biostatistice velký význam.

8 Analýza rozptylu (ANOVA)

V předchozí kapitole jsme zavedli testy pro srovnávání charakteristik jednoho výběru s danou konstantou a testy pro srovnávání charakteristik dvou výběrů. V praxi je však velmi častá i situace, kdy potřebujeme srovnávat více skupin, příkladem může být sledování plicních funkcí u pacientů s chronickou obstrukční plicní nemocí ve stadiu II, III a IV. Zajímá nás, jak se pacienti v jednotlivých stadiích liší v maximálním inspiračním tlaku, tedy maximálním tlaku, který jsou schopni vygenerovat při nádechu. Otázka tedy je, jak můžeme pro stadia II, III a IV ověřit rozdíl (respektive rovnost) v maximálním inspiračním tlaku? Máme dvě možnosti:

1. Použijeme vhodný test pro dva výběry (např. *t*-test) a otestujeme, jak se liší stadium II od stadia III, stadium II od stadia IV a stadium III od stadia IV. Jinými slovy provedeme 3 testy pro dva výběry.

2. Použijeme vhodný test pro více než dva výběry.

Zásadní problém s první možností je v násobném testování hypotéz, kdy je třeba si uvědomit, že s narůstajícím počtem testovaných hypotéz (zde třemi) roste také pravděpodobnost získání falešně pozitivního výsledku, tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje (chyba I. druhu). Pravděpodobnost získání falešně pozitivního výsledku lze v tomto případě jednoduše kvantifikovat: jestliže uvažujeme tři testy a v každém z nich 95% pravděpodobnost, že neuděláme chybu I. druhu, pak za předpokladu nezávislosti provedených testů lze celkovou pravděpodobnost, že neuděláme chybu I. druhu, vyjádřit jako $0{,}95 \times 0{,}95 \times 0{,}95 = 0{,}857$. Jinými slovy pravděpodobnost, že neuděláme chybu I. druhu, nám celkově klesla na 0,857 a tedy pravděpodobnost, že uděláme chybu I. druhu, nám celkově stoupla na 0,143. Jednoznačnou volbou pro testování hypotéz u více než dvou výběrů by tedy měl být adekvátní test pro více než dva výběry.

Základní parametrickou statistickou metodou pro testování hypotéz o středních hodnotách více než dvou skupin je tzv. ***analýza rozptylu*** (*analysis of variance*, ANOVA) [37]. Zmiňované skupiny mohou být samozřejmě dány přirozeně, např. sledujeme-li rozdíl v systolickém krevním tlaku dle desetiletých věkových kategorií, nebo uměle, např. sledujeme-li rozdíl v účinnosti několika typů léčby. Nulová hypotéza je v případě analýzy rozptylu stanovena jako rovnost středních hodnot ve všech sledovaných skupinách. Označíme-li tedy k počet srovnávaných výběrů, pak nulovou a alternativní hypotézu analýzy rozptylu vyjádříme jako

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k , \qquad H_1 : \text{nejméně jedno } \mu_i \text{ je odlišné od ostatních} . \tag{8.1}$$

Příklady problémů a jim příslušných hypotéz vhodných pro analýzu rozptylu mohou být následující:

1. Liší se účinnost dvou různých dávek léčiva A od účinnosti placeba? Označme střední hodnotu účinnosti placeba μ_p, střední hodnotu účinnosti léčiva A v dávce 1 μ_{A1} a μ_{A2} v dávce 2. Pak nulovou a alternativní hypotézu stanovíme takto

$$H_0 : \mu_P = \mu_{A_1} = \mu_{A_2}, \qquad H_1 : \text{nejméně jedno } \mu \text{ je odlišné od ostatních}. \tag{8.2}$$

2. Liší se jednotlivé typy leukémie (akutní myeloidní – AML, akutní lymfoidní – ALL, chronická myeloidní – CML a chronická lymfoidní – CLL) v aktivitě vybraných genů? Označme střední hodnotu exprese genu g u AML, ALL, CML a CLL postupně jako $\theta^g_{AML}, \theta^g_{ALL}, \theta^g_{CML}, \theta^g_{CLL}$. Pak nulovou a alternativní hypotézu stanovíme takto

$$H_0 : \theta^g_{AML} = \theta^g_{ALL} = \theta^g_{CML} = \theta^g_{CLL}, \qquad H_1 : \text{nejméně jedno } \theta^g \text{ je odlišné od ostatních}. \tag{8.3}$$

8.1 Variabilita výběrových souborů a princip výpočtu

Abychom mohli adekvátně vysvětlit princip výpočtu analýzy rozptylu, je třeba nejprve zavést značení a předpoklady, na nichž je analýza rozptylu postavena. Obecně uvažujeme k nezávislých náhodných výběrů Y_{1j}, Y_{2j}, …, Y_{kj} s rozsahy n_1, n_2, …, n_k, o nichž předpokládáme, že pochází z normálního rozdělení, tedy že pro j-té pozorování z i-tého výběru platí $Y_{ij} \sim N(\mu_i,\sigma^2)$. Jinými slovy předpokládáme normalitu hodnot a homogenitu rozptylů u všech k náhodných výběrů (parametr odpovídající rozptylu není závislý na konkrétním výběru a je tedy stejný pro všech k náhodných výběrů). Na základě výše uvedených předpokladů pak definujeme skupinové průměry pro jednotlivé výběry a celkový průměr pro všechny výběry dohromady, které uvádí tabulka 8.1.

Tabulka 8.1 *Zavedení značení k analýze rozptylu.*

	Rozsah výběru	Výběrový součet	Výběrový průměr
Výběr 1	n_1	$Y_{1\cdot} = \sum_{j=1}^{n_1} Y_{1j}$	$\bar{y}_{1\cdot} = Y_{1\cdot} / n_1$
Výběr 2	n_2	$Y_{2\cdot} = \sum_{j=1}^{n_2} Y_{2j}$	$\bar{y}_{2\cdot} = Y_{2\cdot} / n_2$
:	:	:	:
Výběr k	n_k	$Y_{k\cdot} = \sum_{j=1}^{n_k} Y_{kj}$	$\bar{y}_{k\cdot} = Y_{k\cdot} / n_k$
Všechny výběry	n	$Y_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$	$\bar{y}_{\cdot\cdot} = Y_{\cdot\cdot} / n$

Dále zavádíme tři odhady variability, které charakterizují pozorovaná data. První je tzv. ***celkový součet čtverců*** (*total sum of squares*), S_T, který odráží celkovou variabilitu ve výběrovém souboru. Celkový součet čtverců je definován pomocí kvadrátů rozdílů pozorovaných hodnot od celkového průměru následovně:

$$S_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{\cdot\cdot})^2 . \tag{8.4}$$

Celkový součet čtverců je jakožto funkce pozorovaných hodnot statistikou, která má svoje rozdělení pravděpodobnosti. Lze ukázat, že za platnosti nulové hypotézy má statistika S_T chí-kvadrát rozdělení s počtem stupňů volnosti, který se označuje jako df_T a je roven $n - 1$.

Další formou variability je tzv. ***skupinový součet čtverců*** (*group sum of squares*), S_A, který odráží variabilitu mezi skupinami, respektive skupinovými průměry. Jinými slovy, skupinový součet čtverců popisuje variabilitu příslušnou vlivu sledované vysvětlující proměnné. Lze ho spočítat pomocí součtu kvadrátů rozdílů výběrových průměrů od celkového průměru. Statistiku S_A definujeme takto:

$$S_A = \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 . \tag{8.5}$$

Stejně jako v případě S_T, má i statistika S_A chí-kvadrát rozdělení pravděpodobnosti, tentokrát ale se stupni volnosti $df_A = k - 1$.

Třetí statistikou odrážející variabilitu pozorovaných dat je tzv. ***reziduální součet čtverců*** (*residual sum of squares*), S_e, odpovídající variabilitě v rámci skupin. Spočítáme ho tak, že přes všechny výběry a pozorování sečteme kvadráty rozdílů pozorovaných hodnot od příslušných skupinových průměrů, což lze zapsat takto:

$$S_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{y}_{i\cdot})^2 , \tag{8.6}$$

Pro statistiku S_e lze ukázat, že platí $S_e \sim \chi^2(n - k)$.

Příklad 8.1. Tabulka 8.2 obsahuje na fiktivních datech příklad výpočtu jednotlivých součtů čtverců. V příkladu předpokládáme tři výběrové soubory, přičemž každý z nich obsahuje tři pozorované hodnoty.

Tabulka 8.2 *Fiktivní datový soubor se třemi srovnávanými skupinami.*

Léčba	Pozorovaná hodnota	Skupinový průměr	Skupinový průměr *mínus* celkový průměr	Pozorovaná hodnota *mínus* skupinový průměr	Pozorovaná hodnota *mínus* celkový průměr
A	10	12	-4	-2	-6
A	12	12	-4	0	-4
A	14	12	-4	2	-2
B	19	20	4	-1	3
B	20	20	4	0	4
B	21	20	4	1	5
C	14	16	0	-2	-2
C	16	16	0	0	0
C	18	16	0	2	2
	Celkový průměr = 16		Součet čtverců = 96	Součet čtverců = 18	Součet čtverců = 114

V tabulce 8.2 si lze všimnout, že reziduální součet čtverců a skupinový součet čtverců dávají po sečtení dohromady celkový součet čtverců. Toto není náhoda, skutečně lze ukázat, že platí

$$S_T = S_e + S_A\,, \tag{8.7}$$

což znamená, že celková variabilita pozorovaných hodnot se dá rozložit na variabilitu v rámci skupin a variabilitu mezi skupinami:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2\,. \tag{8.8}$$

Stejný vztah jako (8.7) platí i pro stupně volnosti příslušné statistikám S_T, S_A a S_e.

Výpočet analýzy rozptylu je založen na srovnání skupinového a reziduálního součtu čtverců, jinak řečeno ANOVA srovnává pozorovanou variabilitu (rozptyl hodnot) mezi výběry s pozorovanou variabilitou (rozptylem hodnot) uvnitř výběrových souborů. Za předpokladu, že hodnoty všech k srovnávaných výběrů pocházejí z normálního rozdělení se stejným rozptylem, σ^2, představuje výraz

$$\frac{S_e}{df_e} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{y}_{i.})^2}{n-k} \tag{8.9}$$

výběrový odhad tohoto neznámého parametru. Tento podíl odpovídá průměrnému kvadrátu rozdílů pozorovaných hodnot od příslušných skupinových průměrů. Navíc, za platnosti nulové hypotézy představuje i výraz

$$\frac{S_A}{df_A} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{k-1} \tag{8.10}$$

výběrový odhad σ^2. Tento podíl odpovídá průměrnému kvadrátu rozdílů výběrových průměrů od celkového průměru. Platí-li tedy nulová hypotéza, výraz (8.10), vycházející z výběrových průměrů, bude zhruba stejný jako výraz (8.9), vycházející z pozorovaných hodnot. Naopak, neplatí-li nulová hypotéza, lze očekávat, že výraz (8.10) bude větší než výraz (8.9), neboť lze očekávat velkou variabilitu mezi výběrovými průměry (homogenita rozptylů uvnitř výběrů je základním předpokladem analýzy rozptylu). Testovou statistikou v analýze rozptylu je statistika F, která je podílem výrazů (8.10) a (8.9) a která má za platnosti H_0 Fisherovo F rozdělení s parametry $k-1$ a $n-k$. Tedy

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{y}_{i.})^2}{n-k}} = \frac{S_A/df_A}{S_e/df_e} = \frac{MS_A}{MS_e} \sim F(k-1,n-k)\,. \tag{8.11}$$

V případě, že neplatí nulová hypotéza, bude čitatel statistiky F větší než její jmenovatel a výsledná hodnota statistiky F tak bude větší než 1. Hranici pro zamítnutí nulové hypotézy ale opět představuje kvantil (kritická hodnota) rozdělení $F(k-1, n-k)$ příslušný zvolené hladině

významnosti testu α. Případně nulovou hypotézu zamítneme/nezamítneme na základě srovnání výsledné p-hodnoty testu se zvolenou hladinou významnosti testu α. Výsledné výpočty jsou standardně zaznamenávány do tzv. ***tabulky analýzy rozptylu***, kterou pro data z příkladu 8.1 představuje tabulka 8.3 (předpokládejme test na hladině významnosti $\alpha = 0{,}05$). Z této tabulky je vidět, že zamítáme nulovou hypotézu o tom, že pozorované hodnoty pocházejí z normálního rozdělení se stejnou střední hodnotou, neboť při srovnání výsledné p-hodnoty testu se zvolenou hladinou významnosti platí, že $0{,}004 < 0{,}05$. Pokud bychom chtěli rozhodnout o platnosti H_0 pomocí srovnání výsledné hodnoty statistiky F ($F = 16$) s kritickou hodnotou, pak příslušný kvantil F rozdělení je $F_{1-\alpha}^{(k-1,n-k)} = F_{0,95}^{(2,6)} = 5{,}14$. Přitom platí $16 > 5{,}14$, což je v souladu se závěrem pomocí výsledné p-hodnoty.

Tabulka 8.3 *Sumarizace výsledků analýzy rozptylu pro fiktivní data z příkladu 8.1.*

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	Statistika F	p-hodnota
Mezi skupinami	$S_A = 96$	$df_A = k - 1 = 2$	$MS_A = 48$	$F = 16$	0,004
Uvnitř skupin	$S_e = 18$	$df_e = n - k = 6$	$MS_e = 3$		
Celkem	$S_T = 114$	$df_T = n - 1 = 8$			

8.2 Předpoklady analýzy rozptylu a jejich ověření

Analýza rozptylu má stejně jako většina dalších statistických metod svoje předpoklady, bez jejichž splnění nelze na její výsledky spoléhat, respektive, bez jejichž splnění bychom tuto metodu vůbec neměli na dané hodnoty použít. Předpoklady analýzy rozptylu jsou následující:

1. ***Nezávislost pozorovaných hodnot***. Tento předpoklad často bereme za automatický, nicméně automatický není a vždy je třeba se zamyslet nad původem jednotlivých pozorování, zda jsou či nejsou vzájemně nezávislá.

2. ***Normalita hodnot jednotlivých náhodných výběrů***. Tento předpoklad je nutno korektně ověřit, buď pomocí příslušného testu, nebo alespoň pomocí grafických metod (histogramu, krabicového grafu).

3. ***Stejný rozptyl hodnot ve všech srovnávaných skupinách***. Pro ověření tohoto předpokladu platí to samé, co platí v případě ověření normality. Opět musíme buď použít adekvátní test (např. F-test uvedený v kapitole 7), nebo si pozorované hodnoty alespoň zobrazit pomocí histogramu či krabicového grafu.

8.2.1 Hodnocení normality pozorovaných hodnot

Hodnocení normality pozorovaných hodnot je klíčovým postupem v biostatistice, neboť náhodný výběr z normálního rozdělení je kromě analýzy rozptylu předpokladem i řady dalších základních testů a modelů. Zamítnutí normality rozdělení pozorovaných hodnot však nemusí znamenat povolení nebo zamítnutí použití příslušného testu, ale může např. indikovat odlehlé a nelogické hodnoty v datovém souboru. Navíc, pokud o sledované náhodné veličině prokazatelně víme, že se v cílové populaci chová dle normálního rozdělení (např. výška lidské postavy), ale v našem výběrovém souboru normální rozdělení nepotvrdíme, pak s naším

náhodným výběrem zřejmě není něco v pořádku, např. není reprezentativní ke sledované cílové populaci.

Posouzení, zda pozorované hodnoty pochází z normálního rozdělení pravděpodobnosti, není vůbec jednoduché a statistické testy nemusí být nutně nejlepším nástrojem. Vždy je důležité pozorované hodnoty zobrazit pomocí dostupných grafických nástrojů. Základní nástroje pro hodnocení normality pozorovaných dat jsou následující:

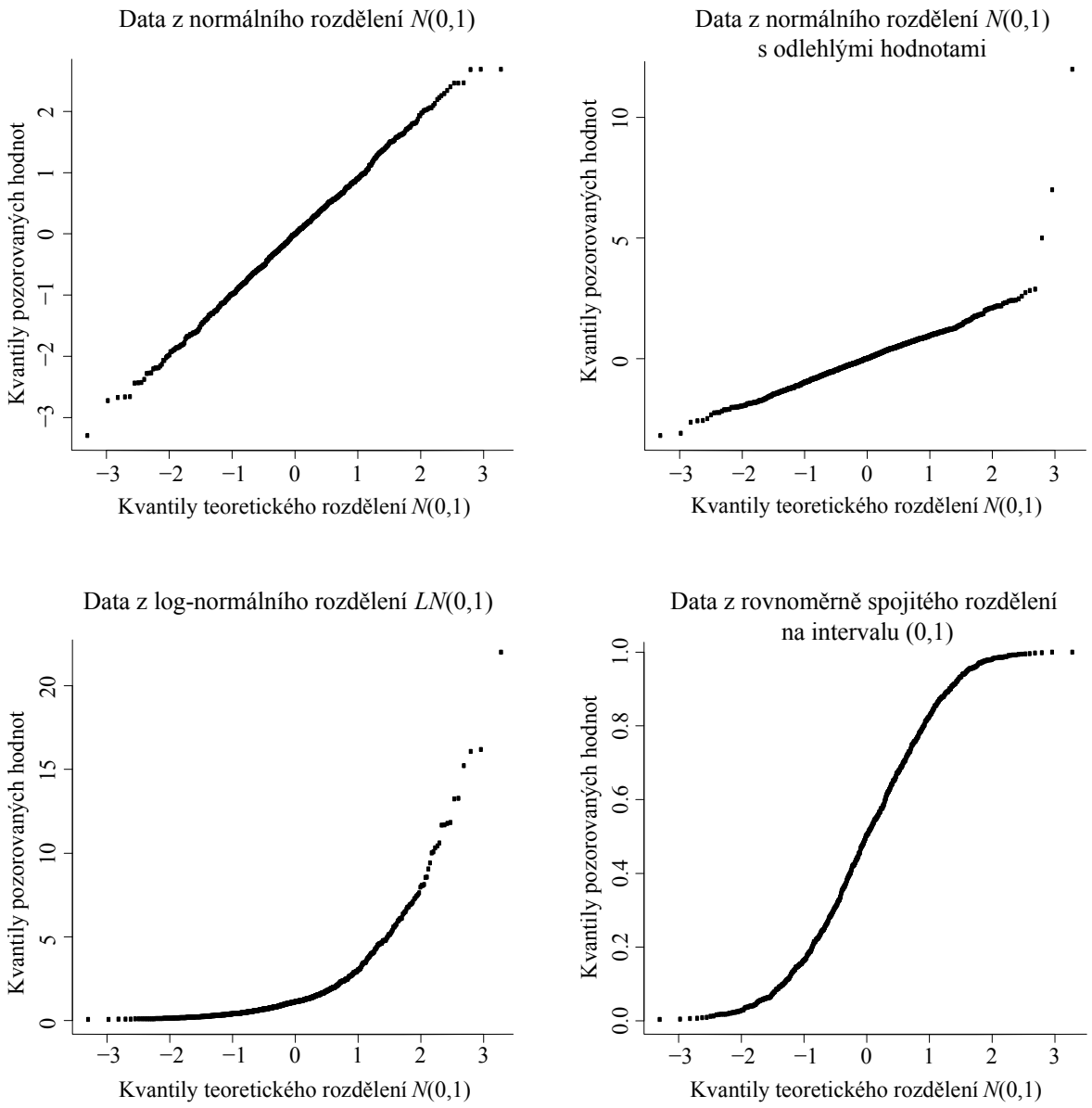
- ***Q-Q diagram***. Tento grafický nástroj [32], na rozdíl od histogramu a krabicového grafu, které jsou určeny pouze pro základní popis dat, umožňuje posoudit, zda pozorované hodnoty pochází z nějakého známého rozdělení pravděpodobnosti. Q-Q diagram proti sobě zobrazuje na ose *x* kvantily teoretického rozdělení pravděpodobnosti (v našem případě normálního rozdělení) a na ose *y* kvantily pozorovaných hodnot. V případě shody výběrového rozdělení dat s teoretickým rozdělením leží všechny body na přímce, zatímco neshodují-li se výběrové a teoretické rozdělení, budou zobrazené body vytvářet křivku odlišnou od přímky. Čtyři příklady Q-Q diagramu jsou znázorněny na obrázku 8.1, kde jsou srovnány simulované hodnoty ze čtyř různých rozdělení pravděpodobnosti s kvantily standardizovaného normálního rozdělení $N(0,1)$. Vlevo nahoře vidíme ideální shodu pozorovaných a teoretických kvantilů danou tím, že hodnoty byly simulovány taktéž z rozdělení $N(0,1)$. Vpravo nahoře jsou také zobrazeny hodnoty simulované z rozdělení $N(0,1)$, ke kterým však byly přidány tři odlehlé hodnoty. Výsledkem je graf, kde téměř všechny zobrazené body leží na přímce, výjimkou jsou právě tři odlehlé hodnoty, které lze jednoznačně identifikovat. Vlevo dole jsou v Q-Q diagramu zobrazeny simulované hodnoty z logaritmicko-normálního rozdělení s parametry 0 a 1, výsledná křivka je typická pro srovnání pozorovaných hodnot z asymetrického rozdělení pravděpodobnosti s normálním rozdělením. Vpravo dole pak vidíme Q-Q diagram pro hodnoty pocházející z rovnoměrně spojitého rozdělení na intervalu (0,1).

- ***Shapirův-Wilkův test*** [25] byl primárně odvozen pro hodnocení normality u menších výběrových souborů (n mezi 3 a 50), v roce 1982 však byl rozšířen i pro větší soubory (n do 2000). Shapirův-Wilkův test má přímou souvislost s Q-Q diagramem, neboť je založen na statistickém vyjádření toho, jak moc se křivka zobrazená Q-Q diagramem liší od ideální přímky. Jinými slovy, jedná se o proložení seřazených pozorovaných hodnot regresní přímkou vzhledem k očekávaným hodnotám normálního rozdělení. Tento test je důležitým nástrojem právě v situacích, kdy máme k dispozici pouze omezený počet pozorování (což je v případě biologických i medicínských dat časté) a na základě vizualizace pomocí Q-Q diagramu nejsme schopni jednoznačně rozhodnout o tom, zda data jsou či nejsou normálně rozdělená.

- ***Kolmogorovův-Smirnovovův test*** [36] představuje obecnější nástroj na hodnocení shody výběrového rozdělení s teoretickým rozdělením pravděpodobnosti, který je založen na srovnání výběrové distribuční funkce s teoretickou distribuční funkcí odpovídající danému (v našem případě normálnímu) rozdělení. Kolmogorovův-Smirnovovův test hodnotí maximální vzdálenost mezi těmito dvěma distribučními funkcemi. V praxi se používá modifikace dle Lillieforse, která je přímo určená pro hodnocení shody výběrového rozdělení s normálním.

V případě, že některý z předpokladů analýzy rozptylu není splněn, máme na výběr ze dvou možností, buď se pokusíme data transformovat (např. logaritmická transformace nám může pomoci s normalizací výběrového rozdělení nebo se stabilizací rozptylu u logaritmicko-normálních dat) nebo pro testování použijeme neparametrický test. Nejpoužívanější neparametrickou alternativou k analýze rozptylu je ***Kruskalův-Wallisův test***, který

nevyžaduje předpoklad normality pozorovaných hodnot. Kruskalovu-Wallisovu testu je věnována část 8.3.



Obr. 8.1 Q-Q diagramy pro srovnání výběrového rozdělení hodnot s rozdělením N(0,1).

8.3 Neparametrická alternativa analýzy rozptylu – Kruskalův-Wallisův test

Kruskalův-Wallisův test [12] je zobecněním neparametrického Mannova-Whitneyho testu pro více než dvě srovnávané skupiny. Stejně jako Mannův-Whitneyho test tak netestuje shodu konkrétních parametrů, ale shodu výběrových distribučních funkcí srovnávaných souborů s tím, že klíčovým předpokladem je zde nezávislosti pozorovaných hodnot. Je-li k počet srovnávaných výběrů, pak nulovou a alternativní hypotézu Kruskalova-Wallisova testu vyjádříme jako

$$H_0 : F_1(x) = F_2(x) = \ldots = F_k(x), \qquad H_1 : \text{nejméně jedna } F_i \text{ je odlišná od ostatních}. \tag{8.12}$$

Hlavní myšlenkou Kruskalova-Wallisova testu je, že za platnosti H_0 jsou sloučené hodnoty ze všech výběrových souborů tak dobře promíchané, že průměrná pořadí odpovídající jednotlivým souborům jsou podobná. Pro výpočet testu tedy opět seřadíme všechna pozorování podle velikosti (jako by pocházely z jednoho výběru) a přiřadíme jednotlivým hodnotám pořadí (R_{ij} bude označovat pořadí j-té hodnoty v i-té skupině). Označme k celkový počet skupin, n celkový počet pozorování a n_1, n_2, … , n_k počty pozorování v jednotlivých skupinách ($n = n_1 + n_2 + \ldots + n_k$). Dále označme T_i součet pořadí v i-té skupině:

$$T_i = \sum_{j=1}^{n_i} R_{ij}. \tag{8.13}$$

Pak testová statistika Kruskalova-Wallisova testu má tvar

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1). \tag{8.14}$$

Lze ukázat, že testová statistika Q má za platnosti nulové hypotézy chí-kvadrát rozdělení pravděpodobnosti s parametrem $k - 1$. Nulovou hypotézu H_0 tak zamítáme na hladině významnosti α, když je realizace testové statistiky Q větší než kritická hodnota (kvantil) příslušná hladině významnosti α, tedy když $Q \geq \chi^2_{k-1}(\alpha)$. Pro malé velikosti souboru je třeba srovnat statistiku Q s tabulkami pro Kruskalův-Wallisův test, které lze najít např. v [30].

8.4 Shrnutí

Analýza rozptylu je základní metodou pro testování hypotéz o středních hodnotách více než dvou skupin, která je založena na srovnání pozorované variability mezi výběry (ta je reprezentována sumou kvadrátů rozdílů výběrových průměrů od celkového průměru) a pozorované variability uvnitř výběrových souborů (ta je reprezentována sumou kvadrátů rozdílů pozorovaných hodnot od příslušného výběrového průměru). Použití analýzy rozptylu jako parametrické metody je však opět podmíněno normalitou hodnot jednotlivých výběrových souborů, navíc předpokládáme i srovnatelný rozptyl, tedy σ^2, v jednotlivých skupinách. Naštěstí existují grafické i výpočetní metody pro ověření těchto předpokladů, které by vždy mělo aplikaci analýzy rozptylu předcházet. Neparametrickou alternativou analýzy rozptylu v případě nesplnění jejích předpokladů je Kruskalův-Wallisův test, který je stejně jako Mannův-Whitneyho test založen na pořadích pozorovaných hodnot. Detailní popis všech možných situací, v nichž lze analýzu rozptylu použít přesahuje rámec těchto skript, více se o metodice analýzy rozptylu lze dovědět v [26, 36, 37].

9 Testování hypotéz o kvalitativních proměnných

Předchozí kapitoly byly věnovány hodnocení kvantitativních náhodných veličin, u nichž předpokládáme, že mohou nabývat mnoha rozdílných hodnot (v případě výšky lidské postavy teoreticky až nekonečně mnoha). V biologii a medicíně se však nezřídka setkáváme i s náhodnými veličinami kvalitativního charakteru, které mohou nabývat pouze omezeného počtu hodnot, v extrémním případě pouze dvou (binární data typu ano/ne, úspěch/neúspěch). Příklady kvalitativní náhodné veličiny jsou krevní skupina (A, B, AB, 0), pohlaví (muž, žena), druh kosatce (*Iris setosa*, *Iris versicolor*, *Iris virginica*), stadium onkologického onemocnění (I, II, III, IV) nebo dokonce i měsíc narození (leden – prosinec).

Stejně jako v případě kvantitativních veličin, tak i u kvalitativních náhodných veličin můžeme hodnotit, zda je hodnota vybrané charakteristiky náhodné veličiny rovna zvolené hodnotě, nebo zda spolu souvisí výskyt dvou náhodných veličin. Statistické hypotézy jsou tedy do značné míry obdobné jako v hodnocení kvantitativních veličin, co se však liší, jsou samozřejmě statistické testy pro jejich ověření.

9.1 Testování hypotéz o podílech

Nejjednodušší formou kvalitativní náhodné veličiny je alternativní (binární) náhodná veličina, nabývající pouze dvou hodnot, např. 0 a 1. Nezávislá opakování alternativní náhodné veličiny pak vedou k binomické náhodné veličině (viz kapitola 4), která je logicky v medicíně i biologii relativně častá, neboť popisuje situace, kdy sledujeme např. výskyt nějaké vlastnosti v dané populaci pacientů nebo výskyt živočišného druhu na daných lokalitách. Hodnocení binomické veličiny vede na tzv. ***testování hypotéz o podílech***, kdy naším cílem je hodnocení tvrzení o parametru π binomického rozdělení, který odpovídá pravděpodobnosti výskytu uvažované vlastnosti ve sledované populaci. Kromě bodového odhadu parametru π nás tedy může zajímat následující:

- Konstrukce intervalu spolehlivosti pro parametr π
- Test o parametru π proti konstantě π_0
- Test o parametru π ve dvou souborech

Při rozhodování o parametru π vycházíme z náhodné veličiny X s binomickým rozdělením pravděpodobnosti, která reprezentuje počet výskytů sledované vlastnosti (úspěchů) v posloupnosti n nezávislých experimentů (subjektů). Nás však zajímá pravděpodobnost výskytu, proto budeme uvažovat transformovanou náhodnou veličinu X / n. Její realizaci značíme malým p s tím, že se vlastně jedná o odhad parametru π, tedy

$$p = x / n = \hat{\pi} . \tag{9.1}$$

Odhad p má jako transformovaná náhodná veličina také svoje rozdělení pravděpodobnosti, kterému odpovídají následující charakteristiky

$$E(p) = E(x/n) = n\pi / n = \pi , \qquad D(p) = D(x/n) = n\pi(1-\pi)/n^2 = \pi(1-\pi)/n . \tag{9.2}$$

Obecně je rozdělení pravděpodobnosti (respektive pravděpodobnostní funkce) binomické náhodné veličiny jednoznačně dáno vzorcem (4.10), jehož výpočet je však pro větší počet nezávislých experimentů, n, nepraktický. V praxi se pro aproximaci rozdělení pravděpodobnosti binomické náhodné veličiny používá normální rozdělení, což nám umožňuje platnost centrální limitní věty (viz kapitola 5). Pouze pro připomenutí, centrální limitní věta platí pro součet n nezávislých, stejně rozdělených náhodných veličin (samozřejmě pro n jdoucí do nekonečna), což je zde splněno, neboť binomická náhodná veličina je součtem n nezávislých náhodných veličin s alternativním rozdělením. Aproximace však neplatí paušálně, podmínkou dobré aproximace normálním rozdělením je hodnota součinu $np(1 - p)$ větší než 5, nebo ještě lépe hodnota součinu $np(1 - p)$ větší než 10 [36]. Tato podmínka souvisí s množstvím informace nutné pro dosažení přibližného tvaru normálního rozdělení, tedy s množstvím informace nutné pro přesnost aproximace. Je-li podmínka dobré aproximace splněna, pak pro náhodnou veličinu Z jako transformaci X platí

$$Z = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1)\,, \tag{9.3}$$

zatímco pro Z jako transformaci p, respektive X / n platí

$$Z = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)\,. \tag{9.4}$$

9.1.1 Interval spolehlivosti pro parametr π binomického rozdělení

Při konstrukci intervalu spolehlivosti pro parametr π vycházíme (dle centrální limitní věty) z předpokladu, že p má normální rozdělení pravděpodobnosti s parametry π a $\pi(1 - \pi)/n$, tedy že platí $p \sim N(\pi, \pi(1 - \pi)/n)$. Dle vztahů (4.5) a (9.4) pak platí, že

$$P(z_{\alpha/2} \le Z \le z_{1-\alpha/2}) = 1 - \alpha\,, \tag{9.5}$$

což lze s pomocí jednoduchých úprav přepsat do tvaru

$$P(p - z_{1-\alpha/2}\sqrt{\pi(1-\pi)/n} \le \pi \le p + z_{1-\alpha/2}\sqrt{\pi(1-\pi)/n}) = 1 - \alpha\,. \tag{9.6}$$

Při konstrukci intervalu spolehlivosti samozřejmě neznáme přesnou hodnotu π, a proto je nutné ji v odhadu rozptylu náhodné veličiny, výrazu $\pi(1 - \pi)/n$, nahradit vhodným odhadem. Logicky se nabízí nahrazení bodovým odhadem, tedy hodnotou p. Při splnění podmínek pro aproximaci normálním rozdělením má $100(1 - \alpha)\%$ interval spolehlivosti pro parametr π tvar:

$$(p - z_{1-\alpha/2}\sqrt{p(1-p)/n}; p + z_{1-\alpha/2}\sqrt{p(1-p)/n})\,. \tag{9.7}$$

Příklad 9.1. Chceme pomocí 95% intervalu spolehlivosti odhadnout podíl studentů matematické biologie, kteří mají modré oči. Máme k dispozici údaje o n = 60 studentech, 17

z nich má modré oči, realizace binomické náhodné veličiny X má tedy hodnotu $x = 17$. Bodový odhad parametru π pak má hodnotu $p = 17/60 = 0{,}283$. Pro sestrojení intervalu spolehlivosti můžeme použít aproximaci normálním rozdělením, neboť $np(1 - p) = 12{,}2$, což je číslo větší než 10. Abychom mohli dosadit do výrazu pro interval spolehlivosti, je třeba spočítat standardní chybu odhadu p, tedy vypočítat

$$SE(p) = \sqrt{p(1-p)/n} = \sqrt{0{,}283(1-0{,}283)/60} = 0{,}058\,. \tag{9.8}$$

S použitím 97,5% kvantilu standardizovaného normálního rozdělení, $z_{1-\alpha/2} = 1{,}96$, pak získáme dosazením do výrazu (9.7) 95% interval spolehlivosti pro podíl studentů matematické biologie s modrýma očima ve tvaru

$$(0{,}283 - 1{,}96 * 0{,}058;\ 0{,}283 + 1{,}96 * 0{,}058) = (0{,}169;\ 0{,}397)\,. \tag{9.9}$$

Na základě našeho výběrového souboru 60 studentů tedy můžeme říci, že s pravděpodobností alespoň 95% leží podíl modrookých studentů matematické biologie v rozmezí 0,169 a 0,397.

9.1.2 Test pro podíl u jednoho výběru

Pointou testu pro podíl u jednoho výběru je stejně jako v případě jiných testů pro jeden výběr ověření rovnosti odhadu parametru π s předem danou hodnotou π_0. Vycházíme z realizace binomické náhodné veličiny X s parametry n a π, respektive z její transformace X / n, kterou značíme p. Nulová hypotéza a příslušné alternativní hypotézy (oboustranná a jednostranné) pak mají následující tvar

$$H_0 : \pi = \pi_0\,, \qquad H_1 : \pi \neq \pi_0\,, \qquad H_1 : \pi > \pi_0 \qquad H_1 : \pi < \pi_0 \tag{9.10}$$

Při splnění podmínek pro aproximaci normálním rozdělením víme, že platí vztah (9.4), což za platnosti H_0 znamená, že

$$Z = \frac{p - \pi_0}{SE(p)} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0{,}1)\,. \tag{9.11}$$

Nulovou hypotézu pak zamítáme na hladině významnosti α, když výsledná hodnota statistiky Z (v případě oboustranné alternativy absolutní hodnota statistiky Z) je větší (nebo menší) než příslušný kvantil rozdělení standardizovaného normálního rozdělení $N(0,1)$. Výraz větší nebo menší závisí na předem zvolené alternativě, příslušné možnosti jsou shrnuty v tabulce 9.1.

Tabulka 9.1 *Pravidla pro zamítnutí H_0 pro test pro podíl u jednoho výběru dle zvolené alternativy.*

Alternativa	$H_1 : \pi \neq \pi_0$	→ Zamítáme H_0, když	$\lvert Z \rvert > z_{1-\alpha/2}$
Alternativa	$H_1 : \pi > \pi_0$	→ Zamítáme H_0, když	$Z > z_{1-\alpha}$
Alternativa	$H_1 : \pi < \pi_0$	→ Zamítáme H_0, když	$Z < z_{\alpha}$

Příklad 9.2. Na hladině významnosti $\alpha = 0{,}05$ chceme testovat rovnost odhadu parametru π získaného na výběru 60 matematických biologů předem dané hodnotě $\pi_0 = 0{,}4$, jinými slovy chceme testovat, zda je podíl matematických biologů s modrýma očima roven 0,4. Splnění podmínek pro aproximaci normálním rozdělením bylo ověřeno v příkladu 9.1. Specifikace nulové a alternativní hypotézy je následující

$$H_0 : \pi = \pi_0 = 0{,}4\,, \qquad H_1 : \pi \neq \pi_0 = 0{,}4\,. \tag{9.12}$$

Pro provedení testu a rozhodnutí o platnosti H_0 vypočteme testovou statistiku Z danou vztahem (9.11):

$$Z = \frac{p - \pi_0}{SE(p)} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0{,}283 - 0{,}400}{\sqrt{0{,}4(1-0{,}4)/60}} = -1{,}85\,. \tag{9.13}$$

Vzhledem k oboustranné alternativě srovnáme absolutní hodnotu realizace testové statistiky, číslo 1,85, s 97,5% kvantilem standardizovaného normálního rozdělení, což je hodnota 1,96. V souladu s tabulkou 9.1 platí, že

$$\lvert Z \rvert = 1{,}85 < z_{1-\alpha/2} = z_{0{,}975} = 1{,}96\,, \tag{9.14}$$

což znamená, že nezamítáme H_0 na hladině významnosti $\alpha = 0{,}05$. Jinými slovy, na hladině významnosti $\alpha = 0{,}05$ nezamítáme hypotézu o tom, že podíl matematických biologů s modrýma očima je roven 0,4.

Na příkladech 9.1 a 9.2 lze demonstrovat další rozdíl v testování hypotéz o spojitých veličinách a testování hypotéz o podílech. V kapitole 6 jsme na příkladu spojité náhodné veličiny ukázali, že existuje spojení mezi testováním hypotéz a konstrukcí intervalů spolehlivosti. Toto spojení však neplatí obecně, klasickým příkladem oblasti, kde tato ekvivalence neplatí, je právě testování hypotéz o podílech. Příklady 9.1 a 9.2 nám totiž dávají protichůdné závěry. V příkladu 9.1 jsme pomocí 95% intervalu spolehlivosti odhadli, že skutečná hodnota parametru π je pokryta intervalem (0,169; 0,397) a je tedy nižší než hodnota 0,4, na druhou stranu v příkladu 9.2 jsme možnost $\pi = 0{,}4$ nevyloučili. Rozdíl v závěrech způsobil fakt, že binomické rozdělení má různý rozptyl pro různé hodnoty π. Největší rozptyl dostaneme pro $\pi = 0{,}5$, směrem k hodnotám 0 a 1 pak rozptyl binomické náhodné veličiny klesá. Pro konstrukci 95% intervalu spolehlivosti jsme ve výpočtu $SE(p)$ za odhad parametru π vzali bodový odhad π, zatímco v testu jsme ve výpočtu $SE(p)$ použili hodnotu danou H_0,

tedy hodnotu π_0, což jsou však dvě různá čísla, která ve výsledku vedou k různým závěrům. V praxi bychom se měli vždy řídit hlavním cílem naší studie nebo experimentu. Je-li tedy naším cílem zkonstruovat intervalový odhad pro sledovaný parametr, měli bychom použít vzorec pro sestrojení intervalu spolehlivosti, a naopak, je-li naším cílem testovat pozorovanou hodnotu podílu proti předpokládané hodnotě π_0, měli bychom použít test.

9.1.3 Interval spolehlivosti pro rozdíl dvou parametrů π

Máme-li nehomogenní skupinu subjektů, jinými slovy, když předpokládáme, že naše sledovaná populace je složena ze dvou populací, bude nás logicky zajímat odhad parametru π v obou jednotlivých podskupinách. Navíc nás však může zajímat i rozdíl těchto dvou odhadů opatřený intervalem spolehlivosti, který vymezuje oblast, kde se s danou pravděpodobností vyskytuje rozdíl parametrů π_1 a π_2. Můžeme tak jednoduše kvantifikovat rozdíl ve výskytu sledované vlastnosti (podílu úspěchů) v obou podskupinách. Bodové odhady parametrů π_1 a π_2 jsou následující

$$p_1 = \frac{x_1}{n_1} = \hat{\pi}_1, \qquad p_2 = \frac{x_2}{n_2} = \hat{\pi}_2, \tag{9.15}$$

kde n_1 a n_2 jsou počty nezávislých experimentů (subjektů) ve skupinách 1 a 2, x_1 a x_2 jsou příslušné počty výskytů sledované vlastnosti (počty úspěchů).

Při konstrukci intervalu spolehlivosti pro rozdíl parametrů π_1 a π_2 vycházíme opět z centrální limitní věty a využíváme aproximace na normální rozdělení, což znamená, že podmínky pro aproximaci normálním rozdělením musí být splněny v obou výběrech. Pro dobrou aproximaci tedy musí platit, že hodnota součinu $n_1p_1(1 - p_1)$ je větší než 5, stejně jako hodnota součinu $n_2p_2(1 - p_2)$.

Bodovým odhadem rozdílu parametrů π_1 a π_2 je rozdíl $p_1 - p_2$, klíčovým pro konstrukci intervalu spolehlivosti je výpočet standardní chyby tohoto rozdílu, který vzhledem k tomu, že neznáme hodnoty π_1 a π_2, má tvar

$$SE(p_1 - p_2) = \sqrt{D(p_1) + D(p_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}. \tag{9.16}$$

Při splnění podmínek pro aproximaci normálním rozdělením pak lze $100(1 - \alpha)\%$ interval spolehlivosti pro rozdíl parametrů π_1 a π_2 vyjádřit jako

$$\left(p_1 - p_2 - z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}};\ p_1 - p_2 + z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right). \tag{9.17}$$

9.1.4 Test pro rozdíl parametrů π ve dvou výběrech

Cílem testu pro podíl ve dvou výběrech je testovat hypotézu, zda jsou pravděpodobnosti výskytu uvažované vlastnosti ve dvou sledovaných populacích stejné. Vycházíme z realizace dvou binomických náhodných veličin $X_1 \sim Bi(n_1,\pi_1)$ a $X_2 \sim Bi(n_2,\pi_2)$, respektive z jejich transformací X_1 / n_1 a X_2 / n_2, které značíme p_1 *a* p_2. Nulová hypotéza tedy odráží situaci, kdy se rovnají parametry π_1 a π_2, a její ověření je založeno na testování, zda se rozdíl realizací p_1 *a* p_2 statisticky významně liší od hodnoty 0 nebo ne. H_0 a příslušné alternativy lze zapsat takto:

$$H_0 : \pi_1 = \pi_2 = \pi\,, \qquad H_1 : \pi_1 \neq \pi_2\,, \qquad H_1 : \pi_1 > \pi_2\,, \qquad H_1 : \pi_1 < \pi_2\,. \tag{9.18}$$

Pro výpočet testu potřebujeme jak odhady parametrů π_1 a π_2, které popisují obě sledované populace, tak odhad parametru π, který odpovídá situaci, kdy platí H_0. Nestranné odhady parametrů π, π_1 a π_2, jsou následující:

$$\hat{\pi} = p = \frac{x_1 + x_2}{n_1 + n_2}\,, \qquad \hat{\pi}_1 = p_1 = x_1 / n_1\,, \qquad \hat{\pi}_2 = p_2 = x_2 / n_2\,. \tag{9.19}$$

Odhad parametru π je třeba pro výpočet standardní chyby rozdílu p_1 *a* p_2, který má tvar

$$SE(p_1 - p_2) = \sqrt{\tfrac{\hat{\pi}(1-\hat{\pi})}{n_1} + \tfrac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\tfrac{p(1-p)}{n_1} + \tfrac{p(1-p)}{n_2}} = \sqrt{p(1-p)(\tfrac{1}{n_1} + \tfrac{1}{n_2})}\,. \tag{9.20}$$

Při splnění podmínek pro aproximaci normálním rozdělením (opět tyto podmínky musí být splněny v obou souborech zároveň) víme, že platí:

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} \sim N(0,1)\,, \tag{9.21}$$

Rozhodnutí o platnosti H_0 na hladině významnosti α pak závisí na výsledné hodnotě statistiky Z (v případě oboustranné alternativy absolutní hodnotě statistiky Z) a hodnotě příslušného kvantilu standardizovaného normálního rozdělení $N(0,1)$. Pravidla pro zamítnutí nulové hypotézy s ohledem na zvolenou alternativu jsou uvedena v tabulce 9.2.

Tabulka 9.2 *Pravidla pro zamítnutí H_0 pro test pro podíl ve dvou výběrech dle zvolené alternativy.*

Alternativa	$H_1 : \pi_1 \neq \pi_2$	→ Zamítáme H_0, když	$\lvert Z \rvert > z_{1-\alpha/2}$
Alternativa	$H_1 : \pi_1 > \pi_2$	→ Zamítáme H_0, když	$Z > z_{1-\alpha}$
Alternativa	$H_1 : \pi_1 < \pi_2$	→ Zamítáme H_0, když	$Z < z_{\alpha}$

Příklad 9.4. Na základě souboru 60 studentů matematické biologie chceme zjistit, zda se liší podíl modrookých studentů u aktivních a u již bývalých studentů. Na hladině významnosti α = 0,05 tak chceme testovat hypotézu o rovnosti parametrů π_1 a π_2 proti oboustranné alternativě, tedy chceme testovat následující H_0 proti H_1:

$$H_0 : \pi_1 = \pi_2 = \pi\,, \qquad H_1 : \pi_1 \neq \pi_2\,. \tag{9.22}$$

Jednotlivé počty aktivních a bývalých studentů a příslušné bodové odhady jsou uvedeny v tabulce 9.3. Je třeba poznamenat, že v tomto příkladu je použití aproximace na normální

rozdělení na samé hranici korektnosti, neboť $n_1p_1(1 - p_1) = 6$ a $n_2p_2(1 - p_2) = 6{,}1$, tedy oba součiny jsou pouze o málo větší než číslo 5.

Tabulka 9.3 *Počty studentů dle modré barvy očí a bodové odhady parametrů π_1 a π_2.*

Studenti oboru Matematická biologie	Počet studentů s modrou barvou očí	Celkový počet studentů	Bodový odhad
Současní studenti	$x_1 = 8$	$n_1 = 32$	$p_1 = \frac{x_1}{n_1} = 0{,}250$
Bývalí studenti	$x_2 = 9$	$n_2 = 28$	$p_2 = \frac{x_2}{n_2} = 0{,}321$
Celkem	$x_1 + x_2 = 17$	$n_1 + n_2 = 60$	$p = \frac{x_1 + x_2}{n_1 + n_2} = 0{,}283$

Pro výpočet testové statistiky je třeba spočítat i standardní chybu rozdílu $p_1 - p_2$, kam na rozdíl od výpočtu intervalu spolehlivosti dosazujeme odhad parametru π, který odpovídá platnosti nulové hypotézy:

$$SE(p_1 - p_2) = \sqrt{p(1-p)(\tfrac{1}{n_1} + \tfrac{1}{n_2})} = \sqrt{0{,}283(1-0{,}283)(\tfrac{1}{42} + \tfrac{1}{18})} = 0{,}117\,. \tag{9.23}$$

Nakonec vypočteme testovou statistiku Z danou vztahem (9.21)

$$Z = \frac{p_1 - p_2}{SE(p_1 - p_2)} = \frac{0{,}250 - 0{,}321}{0{,}117} = -0{,}61\,. \tag{9.24}$$

Vzhledem k oboustranné alternativě srovnáme absolutní hodnotu realizace testové statistiky Z, číslo 0,61, s 97,5% kvantilem standardizovaného normálního rozdělení, tedy s hodnotou 1,96. Nepochybně platí, že

$$|Z| = 0{,}61 < z_{1-\alpha/2} = z_{0{,}975} = 1{,}96\,, \tag{9.25}$$

proto nezamítáme nulovou hypotézu o rovnosti parametrů π_1 a π_2 a můžeme tedy říci, že na základě nám dostupných dat není statisticky významný rozdíl v podílu současných a bývalých studentů matematické biologie s modrýma očima.

9.2 Analýza kontingenčních tabulek

V předchozí části jsme se zabývali problematikou binárních znaků (přítomnost nebo nepřítomnost určité vlastnosti), což vede na hodnocení binomických náhodných veličin. Nejméně stejně tak četné jako binární znaky jsou však v přírodě i znaky s více možnými hodnotami, tedy znaky nominální a ordinální [1]. Matematicky reprezentujeme hodnoty daného znaku jako náhodnou veličinu, v případě dvou nominálních nebo ordinálních znaků pak mluvíme např. o náhodných veličinách X a Y (viz kapitola 1).

K frekvenční sumarizaci jedné nominální nebo ordinální veličiny nám slouží tabulka četností (viz kapitola 2), v případě frekvenční sumarizace kombinací dvou nominálních nebo ordinálních veličin pak mluvíme o tzv. ***kontingenční tabulce*** (*contingency table*). Kontingenční tabulky umožňují testování různých hypotéz:

1. ***Testování nezávislosti*** – pomocí testu nezávislosti můžeme rozhodnout, zda spolu souvisí výskyt dvou nominálních či ordinálních znaků, měřených na souboru n nezávislých experimentálních jednotek. Můžeme např. hodnotit nezávislost pohlaví dítěte a měsíce narození nebo již zmiňovanou souvislost modré barvy očí a období studia u studentů matematické biologie. Hlavním testem nezávislosti pro kontingenční tabulku je ***Pearsonův chí-kvadrát test***.

2. ***Testování shody struktury*** (***testování homogenity***) – o testování homogenity mluvíme v situaci, kdy nás zajímá výskyt nominálního nebo ordinálního znaku u r nezávislých výběrů z r různých populací. Příkladem je hodnocení typologie zaznamenaných nežádoucích účinků u pacientů s infarktem myokardu v několika (r) nemocnicích. Hodnocení shodnosti struktury formálně provádíme pomocí stejné testové statistiky jako testování nezávislosti, tedy také s použitím Pearsonova chí-kvadrát testu.

3. ***Testování symetrie*** – v případě, že uvažujeme opakované měření jedné náhodné veličiny na jednom výběrovém souboru n subjektů a zajímá nás hodnocení změny v jejích hodnotách, mluvíme o testování symetrie. Jedná se o obdobu párového testování u kvantitativních náhodných veličin a příkladem může být hodnocení stavu stromů (lesa) ve dvou po sobě jdoucích sezónách. Pro testování o symetrii kvalitativních náhodných veličin byl odvozen ***McNemarův test***.

9.2.1 Testování nezávislosti (Pearsonův chí-kvadrát test)

Pearsonův chí-kvadrát test je základním a nejpoužívanějším testem nezávislosti v kontingenční tabulce [37]. Nulovou hypotézou je zde tvrzení, že náhodné veličiny X a Y jsou nezávislé, což znamená, že pravděpodobnost nastání určité varianty náhodné veličiny X neovlivňuje nastání určité varianty náhodné veličiny Y. Vyjádřeno pomocí pravděpodobností tedy hypotéza nezávislosti znamená, že

$$p_{ij} = P(X = i \wedge Y = j) = P(X = i)P(Y = j) = p_i p_j, i = 1, \dots, r; j = 1, \dots, c. \qquad (9.26)$$

Test je založen na myšlence srovnání pozorovaných četností (ty jsou dány pozorováním, experimentem) a tzv. očekávaných četností (kalkulovaných za předpokladu platnosti H_0) jednotlivých kombinací náhodných veličin X a Y. Označme n_{ij} počet subjektů, u nichž nastala situace, že náhodná veličina X je rovna hodnotě i a náhodná veličina Y je rovna hodnotě j. Dále definujme tzv. ***marginální četnosti*** příslušné i-té variantě veličiny X, respektive j-té variantě veličiny Y, jako

$$n_{i.} = \sum_{j=1}^{c} n_{ij}, \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}. \qquad (9.27)$$

Za platnosti nulové hypotézy lze očekávané četnosti jednotlivých kombinací, kdy $X = i$ a zároveň $Y = j$, které budeme značit e_{ij}, vypočítat pomocí výrazu

$$e_{ij} = np_{ij} = np_i p_j = n\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}. \tag{9.28}$$

Karl Pearson již v roce 1904 [22] odvodil, že statistika

$$\mathrm{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{9.29}$$

má za platnosti nulové hypotézy o nezávislosti chí-kvadrát rozdělení pravděpodobnosti s parametrem $(r - 1)(c - 1)$, tedy že platí $\mathrm{X}^2 \sim \chi^2_{(r-1)(c-1)}$. Na rozdíl od *t*-testů, kde proti nulové hypotéze hovoří extrémně malé (většinou záporné) i extrémně velké hodnoty testové statistiky, v případě chí-kvadrát testu proti nulové hypotéze hovoří pouze extrémně velké hodnoty testové statistiky, neboť ty indikují významnou neshodu mezi pozorovanými a očekávanými četnostmi. Naopak velmi malé hodnoty testové statistiky hovoří pro nulovou hypotézu, proto nulovou hypotézu o nezávislosti X a Y zamítáme na hladině významnosti α, když hodnota testové statistiky X^2 přesáhne příslušný $100(1 - \alpha)\%$ kvantil rozdělení χ^2, tedy když

$$\mathrm{X}^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha). \tag{9.30}$$

Předpoklady Pearsonova chí-kvadrát testu, které musíme před výpočtem vždy ověřit, jsou následující:

- Jednotlivá pozorování sumarizovaná v kontingenční tabulce jsou nezávislá, tedy každý prvek výběrového souboru je zahrnut pouze v jedné buňce kontingenční tabulky.

- Alespoň 80 % buněk kontingenční tabulky má očekávanou četnost (e_{ij}) větší než 5 a všechny buňky tabulky (tedy 100 % buněk) mají očekávanou četnost (e_{ij}) větší než 2. Tento předpoklad souvisí s asymptotickými vlastnostmi statistiky X^2 a je to tedy stejně důležitý předpoklad jako např. předpoklad normality pozorovaných hodnot v případě skupiny *t*-testů.

Příklad 9.5. Při hodnocení souboru pacientů se zhoubným nádorem kůže (melanomem) chceme zjistit, zda spolu souvisí lokalizace onemocnění (část těla, na které se melanom nachází) a období, kdy bylo onemocnění pacientovi diagnsotikováno. Statisticky řečeno, chceme na hladině významnosti $\alpha = 0{,}05$ testovat nezávislost náhodné veličiny X (období diagnózy s hodnotami 1994–2000, 2001–2005 a 2006–2009) a náhodné veličiny Y (lokalizace s hodnotami horní končetina, dolní končetina, trup a hlava a krk). Tabulka 9.4 sumarizuje pozorované četnosti jednotlivých kombinací náhodných veličin X a Y, v tabulce 9.5 jsou pak uvedeny příslušné očekávané četnosti vypočtené pomocí (9.28) na základě marginálních četností z tabulky 9.4. Je vidět, že všechny očekávané četnosti jsou vyšší než 5, což znamená, že pro ověření hypotézy o nezávislosti můžeme použít Pearsonův chí-kvadrát test.

Tabulka 9.4 *Pozorované četnosti jednotlivých kombinací náhodných veličin X a Y v příkladu 9.5.*

Období = veličina X	Lokalizace = veličina Y				
	Horní končetina $Y = 1$	Dolní končetina $Y = 2$	Trup $Y = 3$	Hlava a krk $Y = 4$	Celkem
1994-2000 $X = 1$	$50 = n_{11}$	$103 = n_{12}$	$116 = n_{13}$	$7 = n_{14}$	$276 = n_{1.}$
2001-2005 $X = 2$	$106 = n_{21}$	$157 = n_{22}$	$310 = n_{23}$	$54 = n_{24}$	$627 = n_{2.}$
2006-2009 $X = 3$	$115 = n_{31}$	$142 = n_{32}$	$316 = n_{33}$	$52 = n_{34}$	$625 = n_{3.}$
Celkem	$271 = n_{.1}$	$402 = n_{.2}$	$742 = n_{.3}$	$113 = n_{.4}$	$1528 = n$

Tabulka 9.5 *Očekávané četnosti jednotlivých kombinací náhodných veličin X a Y v příkladu 9.5.*

Období = veličina X	Lokalizace = veličina Y				
	Horní končetina $Y = 1$	Dolní končetina $Y = 2$	Trup $Y = 3$	Hlava a krk $Y = 4$	Celkem
1994-2000 $X = 1$	$e_{11} = 48.95$	$e_{12} = 72.61$	$e_{13} = 134.03$	$e_{14} = 20.41$	276
2001-2005 $X = 2$	$e_{21} = 111.20$	$e_{22} = 164.96$	$e_{23} = 304.47$	$e_{24} = 46.37$	627
2006-2009 $X = 3$	$e_{31} = 110.85$	$e_{32} = 164.43$	$e_{33} = 303.50$	$e_{34} = 46.22$	625
Celkem	271	402	742	113	1528

Pro výpočet testové statistiky X^2 musíme dosadit hodnoty z tabulek 9.4 a 9.5 do vztahu (9.29), dosazení a vyhodnocení jsou následující:

$$\begin{aligned} X^2 = {} & \frac{(50-48{,}95)^2}{48{,}95} + \frac{(103-72{,}61)^2}{72{,}61} + \frac{(116-134{,}03)^2}{134{,}03} + \frac{(7-20{,}41)^2}{20{,}41} \\ & + \frac{(106-111{,}20)^2}{111{,}20} + \frac{(157-164{,}96)^2}{164{,}96} + \frac{(310-304{,}47)^2}{304{,}47} + \frac{(54-46{,}37)^2}{46{,}37} \\ & + \frac{(115-110{,}85)^2}{110{,}85} + \frac{(142-164{,}43)^2}{164{,}43} + \frac{(316-303{,}50)^2}{303{,}50} + \frac{(52-46{,}22)^2}{46{,}22} = 30{,}41 \end{aligned} \quad . \qquad (9.31)$$

Výslednou hodnotu statistiky X^2 srovnáme s kritickou hodnotou rozdělení chí-kvadrát s parametrem $(r-1)(c-1) = (3-1)(4-1) = 6$, která přísluší hladině významnosti $\alpha = 0{,}05$. Tou je kvantil $\chi^2_{(r-1)(c-1)}(1-\alpha) = \chi^2_{(6)}(0{,}95) = 12{,}59$. Vidíme, že realizace testové statistiky, číslo 30,41, překročila kritickou hodnotu, a tudíž můžeme zamítnout nulovou hypotézu o nezávislosti lokalizace onemocnění a období diagnózy. Můžeme říci, že se s obdobím částečně mění i lokalizace kožních nádorů. Tento závěr není úplně překvapivý, neboť kromě jiného může souviset i s rozvojem a oblibou solárií.

9.2.2 Test hypotézy o symetrii – McNemarův test

McNemarův test [16, 37] je test pro kontingenční tabulku v případě párového uspořádání experimentu, kdy sledujeme výskyt kvalitativní náhodné veličiny X na stejném výběrovém souboru dvakrát po sobě. Jedná se o obdobu párového t-testu. McNemarovým testem hodnotíme, zda se mezi oběma opakováními experimentu (opakovaným sledováním) liší pravděpodobnosti výskytu jednotlivých variant náhodné veličiny X. Máme-li k variant veličiny X, označme je $X_1, X_2, \ldots, X_k$, pak nulovou hypotézu McNemarova testu lze jednoduše vyjádřit jako tvrzení, že pravděpodobnost nastání varianty X_i při prvním měření a varianty X_j při druhém měření je stejná jako pravděpodobnost nastání varianty X_j při prvním měření a varianty X_i při druhém měření. Označme n_{ij} počet prvků výběrového souboru, u nichž se při prvním měření vyskytla varianta X_i a při druhém měření varianta X_j, $i,j = 1, \ldots, k$. Pak testová statistika McNemarova testu má pro obecnou kontingenční tabulku tvar

$$\mathrm{X}^2 = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}. \tag{9.32}$$

Za platnosti nulové hypotézy má statistika X^2 chí-kvadrát rozdělení s parametrem $k(k-1)/2$. Nulovou hypotézu o nezávislosti prvního a druhého měření náhodné veličiny X zamítáme na hladině významnosti α, když realizace testové statistiky X^2 překročí příslušný kvantil, tedy když $\mathrm{X}^2 \geq \chi^2_{k(k-1)/2}(1-\alpha)$.

Zvláštním případem, který je ale v biologii a medicíně relativně častý, je situace, kdy náhodná veličina X nabývá pouze dvou hodnot (např. výskyt nežádoucího účinku léčby ano/ne). V tomto případě máme kontingenční tabulku, která má pouze čtyři buňky a nazýváme ji proto ***čtyřpolní tabulkou*** (více o čtyřpolní tabulce v části 9.3). Označíme-li v souladu se vztahem (9.32) n_{12} jako b a n_{21} jako c, pak má testová statistika X^2 pro čtyřpolní tabulku tvar

$$\mathrm{X}^2 = \frac{(b-c)^2}{b+c}. \tag{9.33}$$

Za platnosti H_0 má pak testová statistika chí-kvadrát rozdělení s 1 stupněm volnosti (neboť $k(k-1)/2 = 1$). Nulovou hypotézu o nezávislosti prvního a druhého měření náhodné veličiny X tedy zamítáme na hladině významnosti α, když $\mathrm{X}^2 \geq \chi^2_1(1-\alpha)$.

Příklad 9.6. Uvažujme 20 pacientů, u nichž sledujeme ústup otoků po podání léku A a následně i ústup otoků po podání léku B. Na hladině významnosti 0,05 nás zajímá, zda je nebo není statisticky významný rozdíl v četnosti otoků po jednotlivých typech léčby. Označme b počet pacientů, u nichž došlo k ústupu otoků po léčbě A, ale nikoliv po léčbě B, a naopak, c označme počet pacientů, u nichž došlo k ústupu otoků po léčbě B, ale nikoliv po léčbě A. Realizace testové statistiky je pak následující

$$\mathrm{X}^2 = \frac{(b-c)^2}{b+c} = \frac{(2-7)^2}{2+7} = 2{,}78. \tag{9.34}$$

Vzhledem k tomu, že číslo 2,78 je menší než příslušný kvantil chí-kvadrát rozdělení ($\chi^2_{(1)}(1-\alpha) = \chi^2_{(1)}(0{,}95) = 3{,}84$) a mezi oběma srovnávanými četnostmi tak není dostatečně velký rozdíl, nezamítáme nulovou hypotézu o tom, že není rozdíl v četnosti otoků po jednotlivých typech léčby.

9.3 Analýza čtyřpolních tabulek

Definice čtyřpolní tabulky je zřejmá – je to nejjednodušší možná kontingenčí tabulka, kdy obě sledované náhodné veličiny mají pouze dvě varianty, kterých mohou nabývat. Stejně jako v případě obecné kontingenční tabulky můžeme pomocí statistických metod rozhodovat o statistické závislosti dvou sledovaných veličin, v případě čtyřpolní tabulky můžeme navíc velmi jednoduše rozhodovat i o míře této závislosti (o těsnosti statistické vazby). Příklad čtyřpolní tabulky představuje tabulka 9.6, kde jsou četnosti jednotlivých možných kombinací náhodných veličin *X* a *Y* označeny písmeny *a*, *b*, *c* a *d*.

Tabulka 9.6 *Ukázka čtyřpolní tabulky.*

Náhodná veličina *X*	Náhodná veličina *Y*		Celkem
	Y = 1	*Y* = 2	
X = 1	*a*	*b*	*a* + *b*
X = 2	*c*	*d*	*c* + *d*
Celkem	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Při rozhodování o nezávislosti ve čtyřpolní tabulce můžeme samozřejmě použít Pearsonův chí-kvadrát test, neboť tento test lze použít na jakoukoliv kontingenční tabulku, nicméně u tohoto testu je nutné hlídat jeho předpoklady: 80 % očekávaných četností, e_{ij}, větších než 5 totiž v případě čtyřpolní tabulky znamená 100 % očekávaných četností, které mají být větší než 5. Nedodržení předpokladů pro Pearsonův chí-kvadrát test může stejně jako u *t*-testu a analýzy rozptylu vést k nesmyslným závěrům. Situace s malými pozorovanými a tedy i očekávanými četnostmi jsou ale bohužel v medicíně i biologii relativně časté, a to samé platí i pro čtyřpolní tabulky. Zlatým standardem pro hodnocení čtyřpolních tabulek se proto stal jiný test, tzv. ***Fisherův exaktní test*** (*Fisher exact test*), který je založen na výpočtu přesné (exaktní) pravděpodobnosti, se kterou bychom za platnosti nulové hypotézy o nezávislosti veličin *X* a *Y* získali naší konkrétní realizaci čtyřpolní tabulky.

9.3.1 Fisherův exaktní test

Fisherův exaktní test [8, 36] byl odvozen primárně pro čtyřpolní tabulky, nicméně existuje i jeho zobecnění na libovolnou kontingenční tabulku. Jeho použití je vhodné zejména v případě, kdy máme kontingenční tabulku s malými očekávanými četnostmi, tedy pro ty, které nesplňují předpoklad Pearsonova chí-kvadrát testu. Nulovou hypotézou je v případě Fisherova testu nezávislost sledovaných veličin *X* a *Y*, což znamená, že pokud H_0 platí, měly by pozorované četnosti odpovídat očekávaným četnostem. Hlavní myšlenkou Fisherova exaktního testu je výpočet pravděpodobnosti, se kterou bychom získali čtyřpolní tabulky stejně nebo více vzdálené od nulové hypotézy při zachování pozorovaných marginálních četností. Zachování marginálních četností znamená, že se soustředíme pouze na situace, které odpovídají stejným četnostem jednotlivých variant náhodných veličin, jako jsme pozorovali v našem experimentu.

Pravděpodobnost získání konkrétního výsledku čtyřpolní tabulky s danými marginálními četnostmi lze vypočítat pomocí vzorce

$$p = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{n!a!b!c!d!}. \tag{9.35}$$

Výpočet testové statistiky potom probíhá následovně: spočítáme pravděpodobnosti p^*, příslušné všem možným tabulkám, které lze získat při zachování marginálních četností. Výsledná testová statistika, respektive p-hodnota, Fisherova exaktního testu je součtem pravděpodobností p^* menších nebo stejných jako hodnota p, která přísluší čtyřpolní tabulce sestrojené na základě pozorovaných hodnot. Sčítáme tak pravděpodobnosti možností, které jsou více nebo stejně vzdáleny od nulové hypotézy, jinými slovy tedy představují extrémnější nebo stejně extrémní variantu výsledku. Z výpočetního postupu je vidět, že Fisherův exaktní test není úplně standardním testem, neboť roli testové statistiky zde, na rozdíl od všech předchozích testů, hraje přímo p-hodnota. Tu potom pro rozhodnutí o platnosti nulové hypotézy srovnáme se zvolenou hladinou významnosti testu α, je-li p-hodnota testu menší než zvolené α, zamítáme nulovou hypotézu o nezávislosti veličin X a Y.

Příklad 9.7. Uvažujme opět skupinu 60 studentů matematické biologie s tím, že tentokrát budeme zjišťovat, zda jejich barva očí (modrá barva očí nebo jiná barva očí) souvisí s nošením brýlí (používá nebo nepoužívá brýle). Pomocí Fisherova exaktního testu chceme testovat nulovou hypotézu o nezávislosti těchto nominálních veličin. Pozorovaná data, respektive pozorovanou čtyřpolní tabulku představuje tabulka 9.7.

***Tabulka 9.7** Počty studentů matematické biologie dle modré barvy očí a nošení brýlí.*

Studenti oboru Matematická biologie	Počet studentů s modrou barvou očí	Počet studentů s jinou barvou očí	Celkový počet studentů
Studenti bez brýlí	$a = 11$	$b = 31$	$a + b = 42$
Studenti s brýlemi	$c = 6$	$d = 12$	$c + d = 18$
Celkem	$a + c = 17$	$b + d = 43$	$a + b + c + d = 60$

Pravděpodobnost příslušná pozorované čtyřpolní tabulce je dle vztahu (9.35) následující

$$p = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{n!a!b!c!d!} = \frac{42!17!18!43!}{60!11!31!6!12!} = 0{,}205. \tag{9.36}$$

Dále vypočítejme pravděpodobnosti p^*, pro jednotlivé možnosti kontingenční tabulky se zachováním marginálních četností, tedy se zachováním řádkových a sloupcových součtů. Výsledek zobrazuje tabulka 9.8.

Tabulka 9.8 *Pravděpodobnosti příslušné jednotlivým možnostem kontingenční tabulky z příkladu 9.7.*

Možnosti	a	b	c	d	p^*
1.	0	42	17	1	$4{,}6 \times 10^{-14}$
2.	1	41	16	2	$1{,}7 \times 10^{-11}$
3.	2	40	15	3	$1{,}8 \times 10^{-9}$
4.	3	39	14	4	$9{,}1 \times 10^{-8}$
5.	4	38	13	5	$2{,}5 \times 10^{-6}$
6.	5	37	12	6	$4{,}1 \times 10^{-5}$
7.	6	36	11	7	$4{,}3 \times 10^{-4}$
8.	7	35	10	8	0,003
9.	8	34	9	9	0,015
10.	9	33	8	10	0,050
11.	10	32	7	11	0,121
12.	11	31	6	12	0,205
13.	12	30	5	13	0,245
14.	13	29	4	14	0,202
15.	14	28	3	15	0,111
16.	15	27	2	16	0,039
17.	16	26	1	17	0,008
18.	17	25	0	18	$6{,}6 \times 10^{-4}$

Výsledná p-hodnota Fisherova exaktního testu je dána součtem p^* všech řádků kromě řádku 13, neboť námi pozorované hodnoty, odpovídající řádku 12, představují vzhledem k nulové hypotéze druhý nejběžnější výsledek ($p = 0{,}205$). Pro všechny řádky tabulky kromě řádku 13 tedy platí $p^* \leq p$. P-hodnotu testu tedy spočítáme jako $1 - 0{,}245 = 0{,}755$ a vzhledem k tomu, že platí $0{,}755 > 0{,}05$, nezamítáme na hladině významnosti $\alpha = 0{,}05$ nulovou hypotézu o nezávislosti barvy očí a nošení brýlí u studentů matematické biologie.

9.3.2 Senzitivita, specificita a prediktivní hodnoty

Kromě testování nezávislosti dvou náhodných veličin můžeme ve čtyřpolní tabulce hodnotit i vztah dvou náhodných veličin, u kterých jejich závislost nejen tušíme, ale dokonce ji i předpokládáme. Nejběžnější situací je statistické hodnocení správnosti diagnostických testů, kdy jsou diagnostické schopnosti testu validovány proti skutečně verifikovanému stavu testovaných osob. Srovnáváme tedy výsledky testu (pozitivní/negativní) proti skutečně prokazatelné přítomnosti/nepřítomnosti nemoci. Pro tuto situaci byla navržena sada ukazatelů správnosti, které představují číselné ohodnocení testu ve vztahu k jeho chybovosti [19, 38].

Definici těchto ukazatelů provedeme na základě značení v tabulce 9.9, kde proti sobě sumarizujeme výsledky diagnostického testu, pozitivní výsledek (označen jako A^+) a negativní výsledek (označen jako A^-), a skutečnou přítomnost nemoci, nemoc přítomna (označeno jako H^+) a nemoc nepřítomna (označeno jako H^-). Kvantifikovat skutečnou přítomnost onemocnění není vždy jednoduché, zde však tento fakt budeme považovat za bernou minci.

Tabulka 9.9 *Čtyřpolní tabulka sumarizující výsledek diagnostického testu proti skutečnosti.*

Výsledek diagnostického testu	Skutečná přítomnost nemoci		Celkem
	Ano (H^+)	Ne (H^-)	
Pozitivní (A^+)	a	b	$a + b$
Negativní (A^-)	c	d	$c + d$
Celkem	$a + c$	$b + d$	$a + b + c + d$

Prvními dvěma ukazateli správnosti testu jsou tzv. senzitivita testu a specificita testu, které definujeme pomocí podmíněné pravděpodobnosti následovně: ***senzitivita testu*** je jeho schopnost rozpoznat skutečně nemocné osoby, tedy pravděpodobnost, že test bude pozitivní, když je osoba skutečně nemocná; ***specificita testu*** je jeho schopnost rozpoznat osoby bez nemoci, tedy pravděpodobnost, že test bude negativní, když osoba není nemocná. Pomocí výše zavedeného značení definujeme senzitivitu a specificitu jako

$$\text{Senzitivita: } P(A^+ \mid H^+) = a/(a+c), \qquad \text{Specificita: } P(A^- \mid H^-) = d/(b+d). \tag{9.37}$$

Druhými dvěma ukazateli jsou tzv. prediktivní hodnoty, které také definujeme pomocí podmíněné pravděpodobnosti: ***prediktivní hodnota pozitivního testu*** je pravděpodobnost, že osoba je skutečně nemocná, když test vyjde jako pozitivní; a naopak ***prediktivní hodnota negativního testu*** je pravděpodobnost, že osoba skutečně není nemocná, když její test vyjde jako negativní. Pomocí výše zavedeného značení definujeme prediktivní hodnoty jako

$$\begin{array}{cc} \text{Prediktivní hodnota pozitivního testu:} & \text{Prediktivní hodnota negativního testu:} \\ P(H^+ \mid A^+) = a/(a+b), & P(H^- \mid A^-) = d/(c+d). \end{array} \tag{9.38}$$

Příklad 9.8. Hodnotíme přesnost vyšetření jater ultrazvukem, respektive schopnost vyšetření ultrazvukem identifikovat postižené ložisko v pacientových játrech. Přesnost je vztažena k laboratornímu ověření odebrané tkáně. Výsledky jsou dány tabulkou 9.10.

Tabulka 9.10 *Sumarizace výsledků ultrazvukového vyšetření jater vzhledem k laboratornímu ověření.*

Výsledek ultrazvuku	Histologické ověření postižení jater		Celkem
	Ložisko přítomno (H^+)	Ložisko nepřítomno (H^-)	
Pozitivní (A^+)	32	2	34
Negativní (A^-)	3	24	27
Celkem	35	26	61

Výpočet senzitivity a specificity testu je následující:

$$\text{Senzitivita testu: } P(A^+ \mid H^+) = a/(a+c) = 32/35 = 0{,}914, \tag{9.39}$$

$$\text{Specificita testu: } P(A^- \mid H^-) = d/(b+d) = 24/26 = 0{,}923. \tag{9.40}$$

Obdobně vypočítáme i obě prediktivní hodnoty testu

$$\text{Prediktivní hodnota pozitivního testu: } P(H^+ \mid A^+) = a/(a+b) = 32/34 = 0{,}941\,, \qquad (9.41)$$

$$\text{Prediktivní hodnota negativního testu: } P(H^- \mid A^-) = d/(c+d) = 24/27 = 0{,}889\,. \qquad (9.42)$$

Z hlediska interpretace je vhodné poznamenat, že senzitivita a specificita jsou spíše populační ukazatele, neboť vycházejí ze znalosti skutečné přítomnosti/nepřítomnosti onemocnění, kterou však u konkrétního testovaného pacienta stojícího v ordinaci s výsledkem testu neznáme. Více než testované osoby (potenciální pacienty) tak senzitivita a specificita zajímají lékaře, kteří mohou tyto dva ukazatele velmi dobře použít pro srovnání diagnostické správnosti dvou různých testů. Naopak prediktivní hodnoty vycházejí z konkrétního výsledku testu (pozitivní/negativní) a jsou tak zajímavé především pro pacienty. Ty totiž v případě konkrétního testu jistě zajímá, jaká je pravděpodobnost, že danou nemoc skutečně mají (respektive nemají) ve chvíli, kdy jim jejich vlastní test vyšel pozitivně (respektive negativně).

Otázkou je, jaké hodnoty senzitivity a specificity jsou dostatečné pro to, abychom označili daný test jako kvalitní nebo ještě lépe jako kvalitnější než testy, které jsou aktuálně dostupné. Odpověď není jednoduchá, neboť do značné míry závisí na stavu poznání dané oblasti a na dosažitelné správnosti dostupných testů. V určité oblasti mohou být hodnoty nad 60 % vítězstvím, v jiné se diagnostika blíží v obou ukazatelích hodnotě 100 %, což znamená, že se téměř nevyskytují falešně pozitivní a falešně negativní výsledky. V každé oblasti existují objektivní limity dané úrovní diagnostiky. Nicméně relevanci odhadu specificity a senzitivity určuje také kvalita experimentu, a to především ve dvou aspektech:

1. Dostatečná velikost experimentálního vzorku zvyšuje kvalitu a přesnost provedených odhadů. Při malém n roste pravděpodobnost, že některé specifické pacienty nezachytíme a odhady specificity a senzitivity budou zkreslené.

2. Musí být zaručena reprezentativnost vzorku vzhledem k rozdělení četností v tabulce. Je-li například podíl nemocných a zdravých jedinců v obecné populaci 1:4, měl by být takto ideálně zachován i ve výběrovém souboru, získáváme tím realistický základ pro posouzení skutečných ukazatelů testu.

Jednoduchým a přitom v odborné literatuře málo využívaným způsobem, jak vyjádřit kvalitu odhadu senzitivity a specificity, je výpočet jejich intervalu spolehlivosti. Všechny čtyři definované ukazatele totiž představují neznámé parametry, které jsou příslušné danému diagnostickému testu a které mají formu podílu. Jejich bodové odhady vypočtené na základě výběrových souborů tak můžeme jednoduše doplnit $100(1 - \alpha)\%$ intervalem spolehlivosti s pomocí postupu, který je blíže popsán v části 9.1.

Zajímavou vlastností obou prediktivních hodnot je fakt, že úzce souvisí s prevalencí sledované nemoci (nebo obecně vlastnosti) v cílové populaci. Budeme-li jednoduše uvažovat konkrétní časový okamžik (konkrétní datum), lze prevalenci vyjádřit jako procento pacientů s danou nemocí počítané ze všech osob v cílové populaci. Abychom mohli demonstrovat závislost pozitivní a negativní prediktivní hodnoty na prevalenci onemocnění (označme ji jako $P(H^+)$, pak $1 - P(H^+) = P(H^-)$), je nutno je nejdříve vyjádřit pomocí Bayesova vzorce a hodnot senzitivity a specificity následovně:

$$P(H^+ \mid A^+) = \frac{P(A^+ \mid H^+)P(H^+)}{P(A^+ \mid H^+)P(H^+) + P(A^+ \mid H^-)P(H^-)}\,, \tag{9.43}$$

$$P(H^- \mid A^-) = \frac{P(A^- \mid H^-)P(H^-)}{P(A^- \mid H^-)P(H^-) + P(A^- \mid H^+)P(H^+)}\,. \tag{9.44}$$

Odvození vztahů (9.43) a (9.44) je základním cvičením z podmíněné pravděpodobnosti, necháváme ho proto na laskavém čtenáři jako cvičení. Vliv prevalence nemoci na prediktivní hodnoty nejlépe ukážeme na příkladu.

Příklad 9.9. Vypočtěme pozitivní a negativní prediktivní hodnotu diagnostického testu na HIV pozitivitu, u kterého výrobce garantuje 98% senzitivitu a 99% specificitu. Jako první uvažujme výpočet těchto ukazatelů v zemi s vysokou prevalencí HIV pozitivity (např. jihoafrické země) a předpokládejme $P(H^+) = 0{,}2$. Prediktivní hodnoty pak jsou následující:

$$P(H^+ \mid A^+) = \frac{0{,}98 \times 0{,}20}{0{,}98 \times 0{,}20 + (1-0{,}99) \times (1-0{,}20)} = 0{,}961\,, \tag{9.45}$$

$$P(H^- \mid A^-) = \frac{0{,}99 \times (1-0{,}20)}{0{,}99 \times (1-0{,}20) + (1-0{,}98) \times 0{,}20} = 0{,}995\,. \tag{9.46}$$

Vidíme tedy, že v zemi s relativně vysokou prevalencí HIV pozitivity má kvalitní test (respektive test s vysokou senzitivitou a specificitou) velkou vypovídací hodnotu, tedy osoby s pozitivním testem (respektive negativním testem) mají vysokou pravděpodobnost, že jsou skutečně HIV pozitivní (respektive HIV negativní). Nyní uvažujme výpočet prediktivních hodnot v zemi s nízkou prevalencí HIV pozitivity (např. evropské země) a předpokládejme $P(H^+) = 0{,}002$. Hodnoty se po přepočtu změní takto:

$$P(H^+ \mid A^+) = \frac{0{,}98 \times 0{,}002}{0{,}98 \times 0{,}002 + (1-0{,}99) \times (1-0{,}002)} = 0{,}164\,, \tag{9.47}$$

$$P(H^- \mid A^-) = \frac{0{,}99 \times (1-0{,}002)}{0{,}99 \times (1-0{,}002) + (1-0{,}98) \times 0{,}002} = 0{,}999\,. \tag{9.48}$$

Máme-li zemi s nízkou prevalencí HIV pozitivity, je vidět, že kvalitní test má velmi dobrou vypovídací schopnost pro osoby, jimž vyšel negativní výsledek testu, neboť na 99,9 % jsou tyto osoby opravdu HIV negativní. Na druhou stranu osoba, jíž vyšel pozitivní výsledek testu, má i při použití kvalitního diagnostického testu pravděpodobnost pouze 16,4 %, že je skutečně HIV pozitivní.

9.4 Testy o rozdělení náhodné veličiny

Testy o rozdělení náhodné veličiny jsou potřebné nejen v situacích, kdy chceme ověřit normalitu pozorovaných hodnot kvůli následnému testování pomocí *t*-testu či analýzy rozptylu, ale také v situacích, kdy chceme výběrové rozdělení našich pozorovaných hodnot

ověřit proti dalším rozdělením pravděpodobnosti. Můžeme např. ověřovat, zda se počty bílých krvinek v 1 ml krve řídí podle logaritmicko-normálního rozdělení pravděpodobnosti nebo zda se počty pacientů, kteří přijdou do ordinace za jednotku času, řídí podle Poissonova rozdělení. Metod, jak porovnat výběrové rozdělení s teoretickým rozdělením, existuje několik, zde zmíníme pouze ty nejpoužívanější:

- ***Kolmogorovův-Smirnovovův test*** byl zmíněn již v kapitole 8 v kontextu ověření normality pro analýzu rozptylu. Jedná se o test založený na srovnání výběrové distribuční funkce s teoretickou distribuční funkcí odpovídající rozdělení, které chceme testovat. Principem ověření shody rozdělení je posouzení maximální vzdálenosti mezi těmito dvěma distribučními funkcemi. Více se lze o tomto testu dozvědět v [36].

- ***Q-Q diagram*** byl také poprvé zmíněn již v kapitole 8. V obecné podobě zobrazuje proti sobě kvantily pozorovaných hodnot a kvantily teoretického rozdělení pravděpodobnosti, v případě shody obou rozdělení zobrazené body tvoří přímku. Výhodou Q-Q diagramu je jeho jednoduchost, jeho nevýhodou je fakt, že se jedná pouze o graf a nikoliv o test. Pro korektní ověření shody výběrového a teoretického rozdělení je vhodné jej doplnit testem (v případě hodnocení normality souvisí Q-Q diagram s Shapirovým-Wilkovým testem).

- ***Chí-kvadrát test dobré shody*** je metodicky shodný s již dříve představeným Pearsonovým chí-kvadrát testem [26], neboť je také založen na myšlence srovnání pozorovaných a očekávaných četností náhodné veličiny X.

9.4.1 Chí-kvadrát test dobré shody

Stejně jako Pearsonův test je i chí-kvadrát test dobré shody primárně určen pro hodnocení diskrétních náhodných veličin, kdy předpokládáme, že náhodná veličina X nabývá r různých hodnot $A_1, A_2, \ldots, A_r$, každé s pravděpodobností $p_1, p_2, \ldots, p_r$. Zároveň platí, že $\sum_{i=1}^{r} p_i = 1$. Pokud je uvažovaný pravděpodobnostní model správný, pak by se v případě realizace náhodného výběru o rozsahu n měl počet pozorování v jednotlivých variantách, tzn. pozorované četnosti n_i, blížit hodnotě očekávaných četností $e_i = np_i$. Samozřejmě platí $\sum_{i=1}^{r} n_i = n$. V případě, že náhodná veličina X má předpokládané rozdělení pravděpodobnosti (H_0 platí), má statistika X^2 chí-kvadrát rozdělení s $r - 1$ stupni volnosti, tedy platí

$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{(r-1)} . \tag{9.49}$$

Nulovou hypotézu o shodě rozdělení veličiny X s předpokládaným teoretickým rozdělením zamítáme na hladině významnosti α, když realizace testové statistiky překročí příslušný kvantil chí-kvadrát rozdělení, tedy když $\mathrm{X}^2 \geq \chi^2_{(r-1)}(1-\alpha)$. Často jsme v situaci, kdy chceme ověřit daný typ rozdělení, ale nemáme žádnou apriorní znalost o parametrech tohoto rozdělení. Ve chvíli, kdy nulovou hypotézou specifikujeme pouze typ rozdělení, ale ne jeho parametry, pak musíme tyto parametry odhadnout z pozorovaných hodnot. Forma testové statistiky se v takovém případě nemění, nicméně za každý takto odhadnutý parametr musíme snížit počet stupňů volnosti testové statistiky o 1.

Chí-kvadrát test dobré shody lze použít i pro spojité náhodné veličiny. Ty sice nenabývají spočetně mnoha (r) hodnot, ale v případě, že rozdělíme obor možných hodnot náhodné

veličiny X do r disjunktních intervalů, lze i v tomto případě test dobré shody použít pro testování shody rozdělení. Tento postup lze nejlépe demonstrovat příkladem.

Příklad 9.10. U pacientů s nádorem kůže sledujeme jejich věk. Pro následné použití parametrických testů chceme na hladině významnosti $\alpha = 0{,}05$ ověřit, zda lze věk těchto pacientů považovat za náhodnou veličinu s normálním rozdělením pravděpodobnosti. Nemáme však žádnou apriorní informaci o parametrech normálního rozdělení, proto potenciální hodnoty μ a σ^2 odhadneme z dat. Na základě dat $n = 1536$ pacientů byl vypočten věkový průměr 56,2 let s výběrovým rozptylem 182,4. Pomocí chí-kvadrát testu dobré shody tedy ověřujeme hypotézu, že věk pacientů s nádorem kůže pochází z rozdělení $N(\mu = 56{,}2, \sigma^2 = 182{,}4)$. Pozorované a očekávané četnosti pacientů dle jednotlivých věkových kategorií jsou sumarizovány v tabulce 9.11.

Tabulka 9.11 *Pozorované a očekávané četnosti pacientů s nádorem dle věkových kategorií.*

$i$tý věkový interval	n_i	e_i	$n_i - e_i$
0,0–8,3 let	0	0,30	-0,30
8,3–16,7 let	5	2,30	2,70
16,7–25,0 let	20	13,30	6,70
25,0–33,3 let	67	53,09	13,91
33,3–41,7 let	139	146,42	-7,42
41,7–50,0 let	243	279,13	-36,13
50,0–58,3 let	336	367,95	-31,95
58,3–66,7 let	357	335,43	21,57
66,7–75,0 let	267	211,46	55,54
75,0–83,3 let	96	92,16	3,84
83,3–91,7 let	6	27,76	-21,76
91,7–100,0 let	0	6,70	-6,70

Dosadíme-li četnosti z tabulky 9.11 do vztahu (9.49), získáme realizaci testové statistiky ve tvaru

$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{(n_i - e_i)^2}{e_i} = 56{,}6\,. \tag{9.50}$$

Vzhledem k tomu, že bylo nutné odhadnout oba parametry normálního rozdělení z pozorovaných dat, počítáme stupně volnosti chí-kvadrát rozdělení testové statistiky pomocí výrazu $df = r - 1 - 2 = 9$. Srovnání realizace testové statistiky X^2 s kvantilem příslušným hladině významnosti $\alpha = 0{,}05$ je následující

$$\mathrm{X}^2 = 56{,}6 \geq \chi^2_{(r-1-2)}(1-\alpha) = \chi^2_{(9)}(0{,}95) = 16{,}92\,, \tag{9.51}$$

Hodnota X^2 překročila příslušný kvantil, proto zamítáme H_0 o normalitě rozdělení věku pacientů s nádorem kůže.

Příklad 9.11. Zaznamenáváme počty pacientů, kteří přijdou v časovém intervalu 30 minut na stomatologickou pohotovost, a chceme zjistit, zda se tyto počty řídí Poissonovým rozdělením. Celkem byly zaznamenány údaje za $n = 1200$ půlhodinových úseků (maximální počet pacientů zaznamenaných za 30 minut byl 10). Nemáme však žádnou apriorní informaci o parametru λ, proto ho odhadneme z dat jako průměrný počet pacientů počítaný přes všech

1200 půlhodinových intervalů, tedy $\hat{\lambda} = 3364/1200 = 2{,}80$. Bodový odhad parametru λ pak použijeme pro výpočet očekávaných četností, pozorované i očekávané četnosti pro jednotlivé počty pacientů jsou sumarizovány v tabulce 9.12.

Tabulka 9.12 *Pozorované a očekávané četnosti pacientů z příkladu 9.11.*

Počet pacientů	n_i	e_i	$n_i - e_i$
0	79	72,97	6,03
1	188	204,32	-16,32
2	282	286,05	-4,05
3	275	266,98	8,02
4	196	186,89	9,11
5	114	104,66	9,34
6	45	48,84	-3,84
7	10	19,54	-9,54
8 a více	11	9,75	1,25

Po dosazení četností z tabulky 9.12 do vztahu (9.49) získáme realizaci testové statistiky ve tvaru

$$\mathrm{X}^2 = \sum_{i=1}^{r} \frac{(n_i - e_i)^2}{e_i} = 8{,}5\,. \tag{9.52}$$

Opět jsme v situaci, kdy bylo nutné parametr Poissonova rozdělení odhadnout z dat, stupně volnosti budou tedy počítány z počtu uvažovaných kategorií ($r = 9$) následovně: $df = r - 1 - 1 = 7$. Následně srovnáme realizaci testové statistiky X^2 s kvantilem příslušným hladině významnosti $\alpha = 0{,}05$ takto:

$$\mathrm{X}^2 = 8{,}50 < \chi^2_{(r-1-1)}(1-\alpha) = \chi^2_{(7)}(0{,}95) = 14{,}07\,. \tag{9.53}$$

Výsledná hodnota testové statistiky X^2 tedy nepřekročila příslušný kvantil, proto nezamítáme H_0 o tom, že se počty pacientů, kteří přijdou v časovém intervalu 30 minut na stomatologickou pohotovost, řídí Poissonovým rozdělením pravděpodobnosti.

9.5 Shrnutí

Hodnocení náhodných veličin popisujících kvalitativní znaky je nejméně tak důležité jako hodnocení náhodných veličin odpovídajících kvantitativním znakům s tím, že vlastně řešíme i obdobné úlohy, tedy zjišťujeme, zda je hodnota parametru náhodné veličiny rovna zvolené hodnotě, nebo zda spolu statisticky souvisí výskyt dvou náhodných veličin. Příkladem první úlohy je testování hypotéz o podílech, kdy se zabýváme hodnocením binomické náhodné veličiny popisující četnost výskytu sledovaného znaku ve výběrovém souboru velikosti n. Příkladem druhé úlohy je testování nezávislosti v kontingenční tabulce, tedy tabulce sumarizující četnosti kombinací dvou nominálních nebo ordinálních veličin.

Speciálním případem kontingenční tabulky je čtyřpolní tabulka pro hodnocení vztahu veličin popisujících dva binární znaky. Ta nám kromě testování nezávislosti umožňuje i kvantifikaci vztahu dvou náhodných veličin, u kterých jejich závislost naopak

předpokládáme. V tomto ohledu je nejběžnější situací statistické hodnocení správnosti diagnostických testů, pro které byla navržena sada ukazatelů správnosti, které představují číselné ohodnocení testu ve vztahu k jeho chybovosti. Mluvíme pak o senzitivitě, specificitě a prediktivních hodnotách testu.

Nejpoužívanějším testem pro testování nezávislosti v kontingenční tabulce je Pearsonův chí-kvadrát test, který však lze použít i pro testování shody teoretického rozdělení pravděpodobnosti s výběrovým rozdělením pozorovaných hodnot, neboť v principu jsou oba postupy shodné. I Pearsonův test má však své předpoklady, zejména pak předpoklad o dostatečných očekávaných četnostech jednotlivých kombinací sledovaných veličin. V případě čtyřpolní tabulky máme alternativu v podobě Fisherova exaktního testu, který lze použít i při velmi malých pozorovaných a tedy i očekávaných četnostech. Tento test lze zobecnit i pro obecnou kontingenční tabulku o rozměru $r \times c$, což je ale téma přesahující rámec těchto skript. Podrobnosti o výpočtu Fisherova exaktního testu pro tabulku $r \times c$ lze nalézt v [20].

10 Základy korelační analýzy

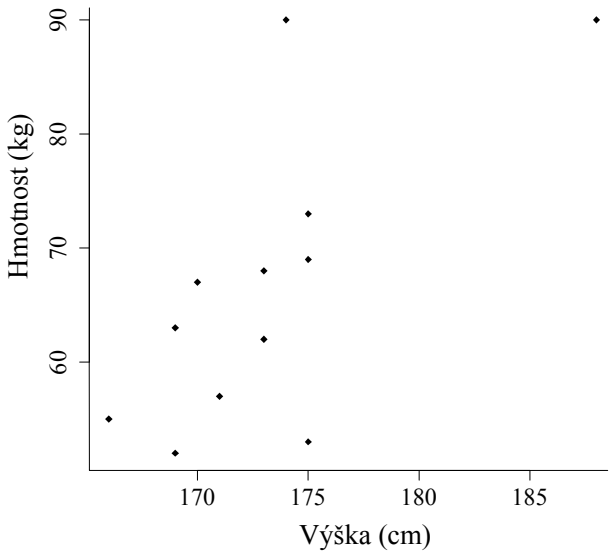
Doposud jsme se z hlediska biostatistiky zabývali hodnocením spojitých a diskrétních náhodných veličin v jedné nebo více odlišitelných experimentálních skupinách. Velmi významnou oblastí biostatistiky je však i problematika hodnocení vztahů a souvislostí mezi dvěma a více spojitými veličinami, která je základem tzv. ***korelační*** a ***regresní analýzy*** [26, 36]. Úkoly, které můžeme řešit pomocí tohoto typu metod, jsou následující:

1. Zjistit, zda mezi sledovanými spojitými veličinami existuje potenciální vztah, např. zda vyšší hodnoty jedné náhodné veličiny souvisejí s nižšími hodnotami jiné náhodné veličiny. Můžeme se např. ptát, zda výše systolického krevního tlaku souvisí s konzumací sodíku, nebo zda vyšší hladina krevní glukózy souvisí s vyšší hladinou jiné látky v krevní plazmě.

2. Predikovat hodnoty jedné náhodné veličiny na základě znalosti hodnot jiných náhodných veličin. Naším cílem může být např. predikce hodnot koncentrací nějaké těžko měřitelné látky v prostředí na základě znalosti koncentrací látek příbuzných, které však těžko měřitelné nejsou.

3. Kvantifikovat vztah mezi dvěma spojitými náhodnými veličinami, např. pro použití jedné z nich na místo té druhé jako diagnostického testu. Můžeme si např. klást otázku, jak moc spolu souvisí hladiny dvou krevních bílkovin, když bychom měření jedné z nich chtěli nahradit druhou.

Nejjednodušším způsobem, jak zjistit, zda hodnoty dvou spojitých náhodných veličin spolu nějak souvisí, je vykreslení bodového grafu (více o bodovém grafu v kapitole 2), který nám ukazuje, jak hodnoty jedné veličiny rostou nebo klesají v závislosti na druhé veličině. Příklad bodového grafu je uveden na obrázku 10.1, kde je zobrazena výška a hmotnost studentů předmětu Biostatistika pro matematickou biologii v jarním semestru 2010. Výsledek je očekávaný, s vyšší výškou má tendenci růst i hmotnost, nicméně vzhledem k tomu, že zobrazené body neleží na přímce, nelze říci, že by mezi výškou a hmotností byl přesně lineární vztah.



Obr. 10.1 Bodový graf hodnot výšky a hmotnosti studentů matematické biologie.

10.1 Pearsonův korelační koeficient

Nevýhodou bodového grafu je samozřejmě absence kvantifikace funkčního vztahu sledovaných veličin. Kvantifikace obecného funkčního vztahu je obtížná, pro kvantifikaci lineárního vztahu náhodných veličin byl zaveden tzv. ***Pearsonův korelační koeficient*** [21]. V teoretické podobě ho lze pro náhodné veličiny X a Y s nenulovým rozptylem vyjádřit následovně:

$$R(X,Y) = \frac{E((X-EX)(Y-EY))}{\sqrt{DX}\sqrt{DY}}. \tag{10.1}$$

Je důležité zdůraznit, že Pearsonův korelační koeficient charakterizuje pouze lineární vztah, jinak řečeno odráží pouze variabilitu kolem lineárního trendu. Pro kvantifikaci nelineárních závislostí je naprosto nevhodný. Základní vlastností Pearsonova korelačního koeficientu je, že nabývá pouze hodnot z intervalu $\langle -1,1 \rangle$ s tím, že hodnota $R(X,Y)$ je kladná, když vyšší hodnoty náhodné veličiny X souvisí s vyššími hodnotami náhodné veličiny Y, a naopak je záporná, když nižší hodnoty X souvisí s vyššími hodnotami Y. Hodnoty 1, respektive -1, získáme pouze v případě, kdy body zobrazené v bodovém grafu leží na přímce s kladnou, respektive zápornou směrnicí.

10.1.1 Výpočet Pearsonova korelačního koeficientu

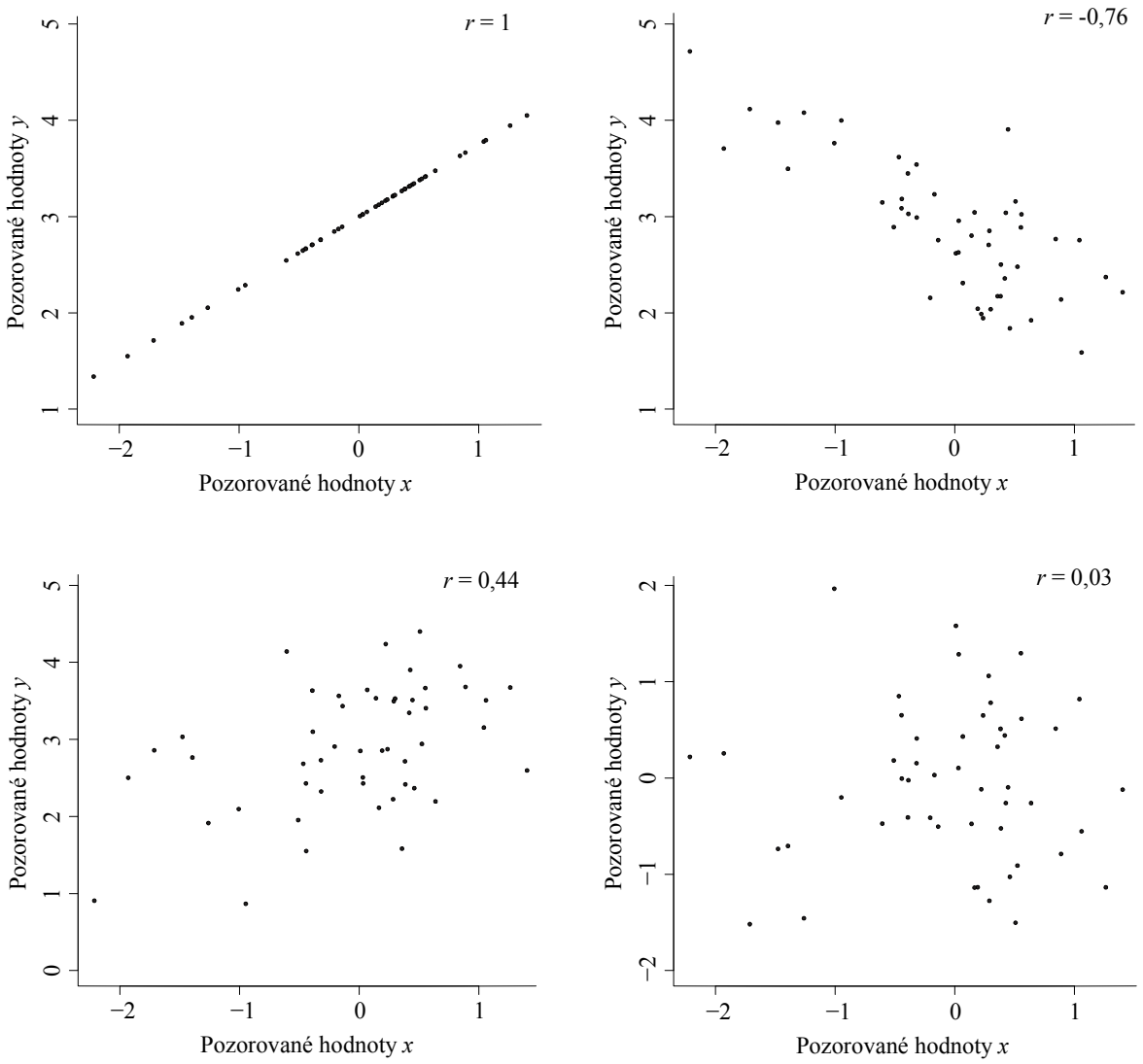
Teoretický výpočet $R(X,Y)$ je podmíněn znalostí konkrétního rozdělení pravděpodobnosti náhodného vektoru (X,Y), což se v praxi stává velmi zřídka. Lineární vztah náhodných veličin X a Y tak kvantifikujeme na základě výběrového souboru. Výběrový Pearsonův korelační koeficient standardně značíme r a při jeho výpočtu vycházíme z realizace dvourozměrného náhodného vektoru o rozsahu n, tedy dvojic pozorovaných hodnot náhodných veličin X a Y pro první až n-tou experimentální jednotku:

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}. \tag{10.2}$$

Výpočet výběrového Pearsonova korelačního koeficientu je pak následující:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}, \tag{10.3}$$

kde $\bar{x}$ a $\bar{y}$ jsou výběrové průměry, s_x a s_y jsou výběrové směrodatné odchylky. Na obrázku 10.2 jsou zobrazeny realizace náhodných veličin X a Y a k nim příslušné výběrové korelační koeficienty pro čtyři různé situace: graf vlevo nahoře odpovídá úplné lineární závislosti; graf vpravo nahoře ukazuje příklad relativně silné záporné korelace; vlevo dole pak vidíme slabě kladně korelované veličiny; vpravo dole jsou nakonec zobrazeny veličiny nekorelované.



Obr. 10.2 Ukázky realizací náhodných veličin X a Y a vypočtené výběrové korelační koeficienty.

Příklad 10.1. Vypočítejme výběrový Pearsonův korelační koeficient kvantifikující korelaci mezi výškou a hmotností studentů předmětu Biostatistika pro matematickou biologii v jarním semestru 2010. Pozorované hodnoty (realizace náhodného vektoru o rozsahu $n = 13$) jsou uvedeny v tabulce 10.1, navíc jsou předmětem obrázku 10.1.

Tabulka 10.1 *Pozorované hodnoty výšky a hmotnosti 13 studentů.*

175	166	170	169	188	175	176	171	173	175	173	174	169
69	55	67	52	90	53	57	57	68	73	62	90	63

Výpočet výběrových statistik pro jednoduchost vynecháme (laskavý čtenář si je může jednoduše dopočítat na základě dat v tabulce 10.1), dosazením do vztahu (10.3) získáme následující hodnotu výběrového Pearsonova korelačního koeficientu:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{148\,929 - 148\,417{,}2}{(13-1)*5{,}3*12{,}5} = 0{,}64\,. \tag{10.4}$$

Hodnota r =0,64 ukazuje na silnou korelaci, kdy s vyšší výškou roste i hmotnost, což odpovídá očekávání, nicméně je třeba si uvědomit malou velikost výběrového souboru a dvě odlehlé hodnoty na obrázku 10.1 odpovídající hmotnosti 90 kg, které úplně nekorespondují se zbytkem souboru. Obě tyto skutečnosti ovlivňují výslednou hodnotu r.

10.1.2 Interval spolehlivosti pro Pearsonův korelační koeficient

Jako každou výběrovou statistiku je i výběrový Pearsonův korelační koeficient r vhodné doplnit $100(1 - \alpha)\%$ intervalem spolehlivosti, který nám dá informaci o variabilitě tohoto odhadu. Na rozdíl od výpočtu bodového odhadu, který lze vypočítat na datech z různých rozdělení, je však v případě, že chceme rozhodovat o vlastnostech Pearsonova korelačního koeficientu (např. konstruovat interval spolehlivosti pro r nebo testovat hypotézy o r), nutné učinit předpoklad o normalitě náhodných veličin X a Y. Jinými slovy, při výpočtu r předpokládáme realizaci dvourozměrného náhodného vektoru z dvourozměrného normálního rozdělení o rozsahu n. Dalším problémem při konstrukci intervalu spolehlivosti pro r je fakt, že výběrové rozdělení výběrového korelačního koeficientu není normální. Abychom byli schopni interval spolehlivosti zkonstruovat, je třeba použít transformaci na náhodnou veličinu W, přičemž transformace je následující:

$$W = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right). \tag{10.5}$$

Lze ukázat, že náhodná veličina W má normální rozdělení s rozptylem přibližně $D(W) = 1/(n-3)$, kde n je velikost výběrového souboru. Vzhledem k normalitě veličiny W má $100(1 - \alpha)\%$ interval spolehlivosti pro její střední hodnotu tvar

$$(d^*, h^*) = w \pm z_{1-\alpha/2}\frac{1}{\sqrt{n-3}}, \tag{10.6}$$

kde $z_{1-\alpha/2}$ je příslušný kvantil standardizovaného normálního rozdělení. Výsledný $100(1 - \alpha)\%$ interval spolehlivosti pro r pak dostaneme zpětnou transformací ve tvaru

$$(d, h) = \left(\frac{\exp(2d^*)-1}{\exp(2d^*)+1}; \frac{\exp(2h^*)-1}{\exp(2h^*)+1}\right), \tag{10.7}$$

Příklad 10.2. Navážeme na příklad 10.1, kde byl vypočítán výběrový korelační koeficient pro vztah výšky a hmotnosti studentů biostatistiky. Nyní pro r = 0,64 zkonstruujeme 95% interval spolehlivosti. Realizace transformované náhodné veličiny je následující:

$$w = \frac{1}{2}\ln\frac{1+0{,}64}{1-0{,}64} = 0{,}758\,, \tag{10.8}$$

Interval spolehlivosti pro střední hodnotu náhodné veličiny W s $\alpha = 0{,}05$ má tvar

$$(d^*, h^*) = 0{,}758 \pm 1{,}96/\sqrt{13-3} = (0{,}138; 1{,}377)\,, \tag{10.9}$$

z čehož plyne výsledný 95% interval spolehlivosti pro výběrový korelační koeficient vztahu výšky a hmotnosti studentů biostatistiky

$$(d, h) = \left(\frac{\exp(2d^*)-1}{\exp(2d^*)+1}; \frac{\exp(2h^*)-1}{\exp(2h^*)+1}\right) = (0{,}14; 0{,}88)\,. \tag{10.10}$$

Z výsledku vidíme, že 95% interval spolehlivosti je velmi široký, neboť připouští jak hodnoty odpovídající silné korelaci ($r = 0{,}88$), tak hodnoty odpovídající velmi slabé, nebo spíše žádné korelaci ($r = 0{,}14$). Zde je na vině zejména malý rozsah výběrového souboru, neboť je zřejmé, že na základě $n = 13$ pozorování je velmi obtížné dělat zásadní závěry ohledně vztahu dvou náhodných veličin.

10.1.3 Test hypotézy o nulové korelaci dvou náhodných veličin

I v případě malého výběrového souboru, jaký byl použit např. v příkladech 10.1 a 10.2, je logické klást si otázku, zda je či není korelace dvou sledovaných veličin nulová. Tato situace vede na testování následujících hypotéz:

$$H_0 : r = 0\,, \qquad H_1 : r \neq 0\,. \tag{10.11}$$

Pro testování je nezbytný předpoklad realizace dvourozměrného náhodného vektoru o rozsahu n z normálního rozdělení, což znamená, že máme k dispozici náhodný vektor

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}, \qquad \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}\right). \tag{10.12}$$

Za platnosti nulové hypotézy pak má statistika

$$T = r\sqrt{\frac{n-2}{1-r^2}} \tag{10.13}$$

Studentovo t rozdělení pravděpodobnosti s $n - 2$ stupni volnosti. Pro oboustrannou alternativu zamítáme nulovou hypotézu na hladině významnosti $\alpha = 0{,}05$, když hodnota testové statistiky přesáhne v absolutní hodnotě kvantil $t_{1-\alpha/2}^{(n-2)}$. Je třeba poznamenat, že testovou statistiku T

nelze použít pro testování obecné hypotézy $H_0 : r = r_0 \neq 0$, neboť pro r různé od nuly nemá testová statistika Studentovo t rozdělení. Postup pro testování hypotézy $H_0 : r = r_0 \neq 0$ lze najít např. v [26, 36].

Příklad 10.3. Provedení testu o nulové korelaci dvou náhodných veličin opět demonstrujeme na datech výšky a hmotnosti studentů biostatistiky. Realizace testové statistiky dané vztahem (10.13) je následující

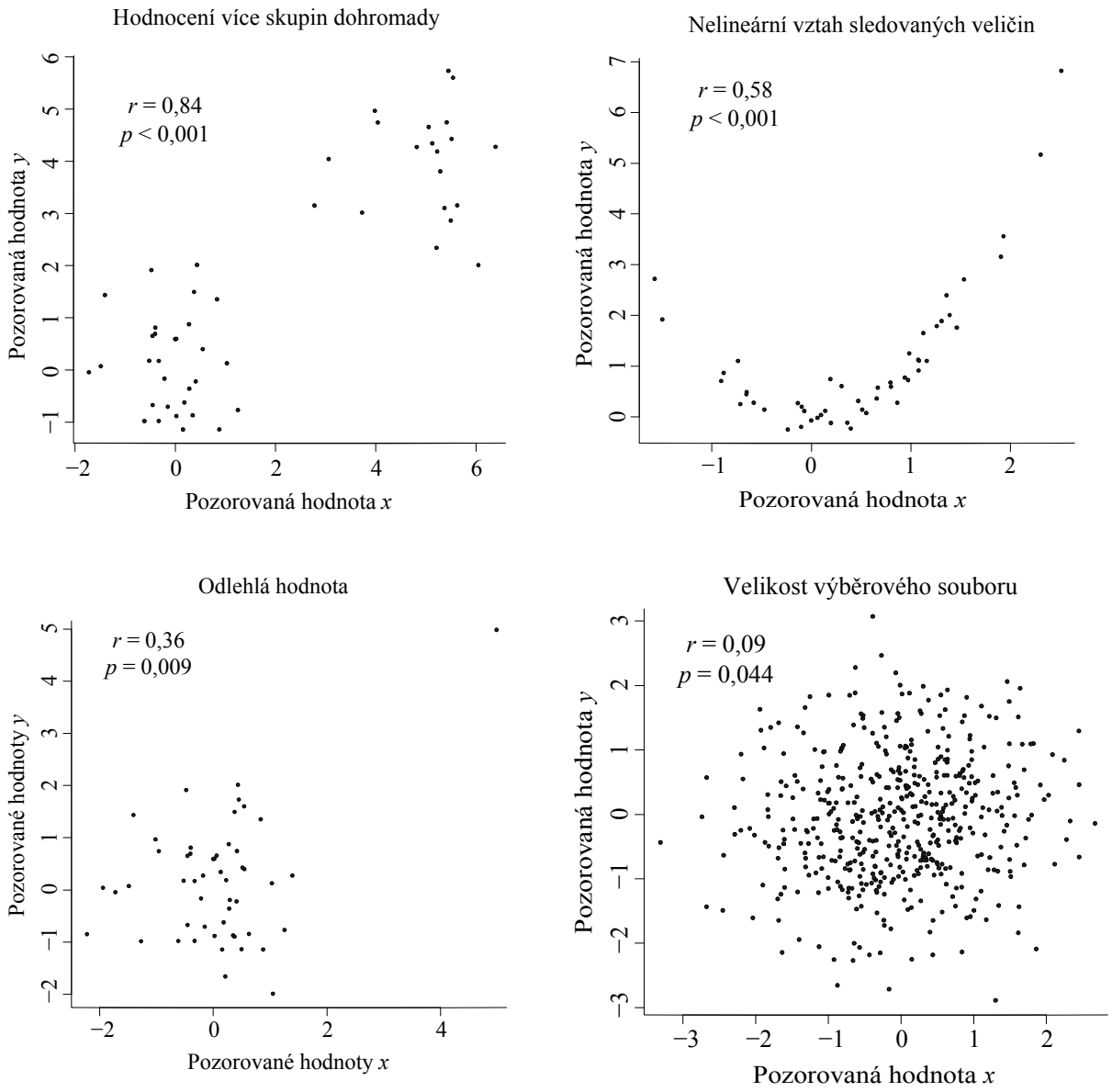
$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}64\sqrt{\frac{13-2}{1-0{,}64^2}} = 2{,}76. \tag{10.14}$$

Srovnáme-li výslednou hodnotu testové statistiky t s kvantilem Studentova t rozdělení příslušným hladině významnosti $\alpha = 0{,}05$, tedy provedeme-li srovnání

$$t = 2{,}76 > 2{,}20 = t_{0{,}975}^{(11)} = t_{1-\alpha/2}^{(n-2)}, \tag{10.15}$$

zamítáme H_0 o tom, že mezi výškou a hmotností studentů biostatistiky je nulová korelace.

Jak bylo uvedeno výše, Pearsonův korelační koeficient kvantifikuje míru lineárního vztahu mezi náhodnými veličinami X a Y. Jeho výpočet je tedy naprosto nevhodný v situacích, kdy se o lineární vztah mezi X a Y nejedná. Obrázek 10.3 ukazuje čtyři situace, kdy výpočet výběrového Pearsonova korelačního koeficientu nemá smysl, respektive kdy může být jeho výpočet z hlediska interpretace zavádějící. Graf vlevo nahoře znázorňuje situaci, kdy výběrový soubor obsahuje dvě skupiny subjektů s odlišnými hodnotami náhodných veličin X i Y. Ve chvíli, kdy si tohoto nejsme vědomi, výpočet výběrového Pearsonova korelačního koeficientu indikuje silnou korelaci X a Y ($r = 0{,}84$), která je dokonce na daném souboru vysoce statisticky významná ($p < 0{,}001$). Tento výsledek je však statistický artefakt a ve skutečnosti není relevantní. Ideální by v tomto případě bylo soubor rozdělit a kvantifikovat korelaci v obou podsouborech zvlášť (podle obrázku je korelace X a Y v podsouborech naopak velmi malá). Graf vpravo nahoře ukazuje situaci, kdy je mezi veličinami X a Y nelineární vztah. Také zde je výsledný korelační koeficient ($r = 0{,}58$) relativně vysoký, statisticky významný a zároveň neodpovídá skutečnosti. Vlevo dole pak vidíme, jaký vliv má odlehlá hodnota v případě dvou nezávislých (a tedy i nekorelovaných) veličin X a Y. Vzhledem k nezávislosti bychom čekali realizaci r kolem 0, nicméně zde vidíme výsledné r rovno 0,36, opět statisticky významné ($p = 0{,}009$). Konečně, graf vpravo dole ukazuje vliv velikosti výběrového souboru na statistickou významnost korelačního koeficientu. V tomto případě je korelace mezi veličinami X a Y velmi slabá až žádná ($r = 0{,}09$), nicméně velikost výběrového souboru je tak velká ($n = 500$), že statistický test indikuje statisticky významný rozdíl r od hodnoty 0. Toto je klasický příklad rozporu mezi statistickou a praktickou významností výsledku, kdy je nezbytné kromě statistiky do výsledné interpretace zapojit i znalost dané problematiky. Všechny čtyři problematické případy lze velmi dobře odhalit s použitím bodového grafu, který by měl být jedním z prvních kroků při hodnocení vzájemného vztahu dvou spojitých náhodných veličin.



Obr. 10.3 Problematické situace pro výpočet Pearsonova korelačního koeficientu.

10.2 Spearmanův korelační koeficient

Zatímco první situaci na obrázku 10.3 lze řešit rozdělením souboru na dva a následným výpočtem korelačního koeficientu v obou podsouborech, v situaci odpovídající grafu vpravo nahoře nemá smysl Pearsonův korelační koeficient počítat vůbec, neboť ten odráží pouze lineární závislost. Rozšíření směrem k hodnocení určitých forem nelineární závislosti představuje tzv. ***Spearmanův korelační koeficient***. Jedná se o neparametrický korelační koeficient, který je robustní vůči odlehlým hodnotám a obecně odchylkám od normality, neboť stejně jako řada dalších neparametrických metod pracuje pouze s pořadími pozorovaných hodnot [27]. Na rozdíl od Pearsonova koeficientu korelace, který popisuje lineární vztah veličin X a Y, Spearmanův koeficient korelace popisuje, jak dobře vztah veličin X a Y odpovídá monotónní funkci, která může být samozřejmě nelineární.

Při výpočtu opět vycházíme z realizace dvourozměrného náhodného vektoru o rozsahu n, tedy dvojic pozorovaných hodnot náhodných veličin X a Y pro n subjektů. Dále definujme

číslo x_{ri} jako pořadí hodnoty x_i v rámci vzestupně uspořádaných hodnot x_1, ..., x_n, číslo y_{ri} jako pořadí hodnoty y_i v rámci vzestupně uspořádaných hodnot y_1, ..., y_n, čísla $\bar{x}_r$ a $\bar{y}_r$ jako průměry hodnot x_{ri}, respektive y_{ri} (tedy jako průměrná pořadí), a čísla s_{x_r} a s_{y_r} jako odpovídající směrodatné odchylky. Spearmanův korelační koeficient, označme ho r_s, pak vypočítáme pomocí vzorce

$$r_s = \frac{\sum_{i=1}^{n} x_{ri} y_{ri} - n\bar{x}_r \bar{y}_r}{(n-1)s_{x_r} s_{y_r}}, \tag{10.16}$$

což není nic jiného než vzorec pro výběrový Pearsonův korelační koeficient počítaný na pořadích pozorovaných hodnot. Hodnoty r_s se pohybují stejně jako v případě koeficientu r v rozmezí od -1 do 1. Hodnot kolem nuly nabývá Spearmanův korelační koeficient v případě, že pořadí hodnot x_i a y_i jsou náhodně zpřeházená a mezi sledovanými veličinami není žádný vztah. Naopak hodnot -1 a 1 nabývá Spearmanův korelační koeficient v případě, že jedna z veličin je monotónní funkcí druhé veličiny.

Výpočetní alternativou ke vzorci (10.16) je výpočet založený na diferencích pořadí pozorovaných hodnot, které definujeme následovně:

$$d_i = x_{ri} - y_{ri}. \tag{10.17}$$

Hodnotu Spearmanova korelační koeficient pak odhadneme pomocí vztahu

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{10.18}$$

Tento výpočet r_s platí přesně pouze pro neopakovaná pozorování, což znamená, že je citlivý na opakující se hodnoty, které vedou k průměrování pořadí. Vyskytuje-li se mezi hodnotami x_1, ..., x_n, respektive y_1, ..., y_n, množství shodných hodnot, je vhodnější použít k výpočtu Spearmanova korelačního koeficientu definiční vztah (10.16).

Příklad 10.4. Pro srovnání s hodnotou r = 0,64 vypočtenou v příkladu 10.1 odhadneme korelaci výšky a hmotnosti studentů biostatistiky také pomocí Spearmanova koeficientu korelace. Hodnoty potřebné k výpočtu jsou uvedeny v tabulce 10.2. Vzhledem k přítomnosti opakovaných hodnot u výšky i hmotnosti vypočteme nejprve Spearmanův korelační koeficient s použitím vzorce (10.16):

$$r_s = \frac{\sum_{i=1}^{n} x_{ri} y_{ri} - n\bar{x}_r \bar{y}_r}{(n-1)s_{x_r} s_{y_r}} = \frac{721{,}5 - 637}{(13-1) * 3{,}86 * 3{,}88} = 0{,}47. \tag{10.19}$$

Dále vypočteme hodnotu r_s i pomocí vztahu (10.18). V tomto případě dosadíme hodnoty z tabulky 10.2 následovně:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6*191}{13(13^2-1)} = 0{,}48\,, \tag{10.18}$$

Je vidět, že v tomto případě dávají oba výpočty koeficientu r_s velmi podobné výsledky, které odpovídají střední korelaci mezi výškou a hmotností. Oba výsledky se však liší od původně vypočtené hodnoty r = 0,64. Důvodem jsou dvě pozorování odpovídající hmotnosti 90 kg, které úplně nekorespondují se zbytkem souboru (viz obrázek 10.1). V tomto případě, kdy máme velmi limitovanou velikost výběrového souboru, je tedy lepší dát přednost neparametrické variantě, tedy hodnotě Spearmanova koeficientu korelace.

Tabulka 10.2 *Hodnoty pro výpočet Spearmanova koeficientu korelace výšky a hmotnosti studentů.*

Student	Výška: x_i	Pořadí výšky	Hmotnost: y_i	Pořadí hmotnosti	Rozdíl d_i	d_i^2
1	175	10	69	10	0	0
2	166	1	55	3	-2	4
3	170	4	67	8	-4	16
4	169	2,5	52	1	1,5	2,25
5	188	13	90	12,5	0,5	0,25
6	175	10	53	2	8	64
7	176	12	57	4,5	7,5	56,25
8	171	5	57	4,5	0,5	0,25
9	173	6,5	68	9	-2,5	6,25
10	175	10	73	11	-1	1
11	173	6,5	62	6	0,5	0,25
12	174	8	90	12,5	-4,5	20,25
13	169	2,5	63	7	-4,5	20,25

Konstrukce 100(1 – α)% intervalu spolehlivosti i test nulové hypotézy H_0: $r_s = 0$ probíhá pro Spearmanův korelační koeficient stejně jako pro koeficient Pearsonův. Co se týče konstrukce intervalu spolehlivosti, výběrové rozdělení r_s je pro výběry o velikosti alespoň 10 stejné jako výběrové rozdělení r. Pro větší vzorky, kdy je velikost souboru alespoň 30, je pak možné použít pro ověření nulové hypotézy $r_s = 0$ stejnou testovou statistiku jako v případě r danou vztahem (10.13). Pro zamítnutí H_0: $r_s = 0$ pak platí také stejná pravidla jako pro koeficient r.

10.3 Shrnutí

Korelační koeficienty jsou základním nástrojem, jak kvantifikovat vztah dvou spojitých náhodných veličin, i když abychom byli úplně přesní, Spearmanův korelační koeficient pracující s pořadími hodnot lze použít i v případě diskrétních náhodných veličin s ordinální škálou hodnot. Pro oba zmíněné koeficienty, Pearsonův i Spearmanův, byly uvedeny postupy pro konstrukci intervalu spolehlivosti a test hypotézy o tom, že příslušný korelační koeficient je roven nule. Opět je třeba zdůraznit omezení těchto výpočtů. Hodnotu výběrového Pearsonova korelačního koeficientu sice můžeme jako číslo vypočítat na datech z různých rozdělení, ale chceme-li se pustit do sestrojení intervalu spolehlivosti nebo testu hypotézy $r = 0$, je nutné učinit předpoklad o normalitě sledovaných náhodných veličin. Pro konstrukci intervalu spolehlivosti a testu hypotézy $r_s = 0$ pro Spearmanův korelační koeficient je zase

třeba zajistit dostatečnou velikost výběrového souboru, aby se výběrové rozdělení r_s dostatečně podobalo výběrovému rozdělení koeficientu r.

Nakonec poznamenejme, že korelace dvou náhodných veličin se často interpretuje také pomocí druhé mocniny Pearsonova korelačního koeficientu, tedy pomocí r^2. Hodnota r^2 vyjadřuje, kolik procent své variability sdílí jedna náhodná veličina s druhou, což lze ještě vyjádřit jako procento variability jedné náhodné veličiny, které může být predikováno pomocí té druhé. Z tohoto hlediska má hodnota r^2 význam a interpretaci zejména ve stochastickém modelování.

Literatura

[1] Agresti, A.: Categorical Data Analysis. John Wiley & Sons, New Jersey (2002)

[2] Altman, D. G.: Practical Statistics for Medical Research. Chapman and Hall, London (1999)

[3] Anděl, J.: Matematická statistika. SNTL/Alfa, Praha (1978)

[4] Cox, D.R., Oakes, D.: Analysis of Survival Data. Chapman & Hall/CRC, New York (1998)

[5] Dwass, M.: Some k-sample rank-order tests. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG & Mann HB (eds.): Contributions to probability and statistics. Stanford University Press, Stanford, 198–202 (1960)

[6] Efron, B., Tibshirani R. J.: An Introduction to the Bootstrap. Chapman & Hall/CRC, New York (1994)

[7] Estève, J., Benhamou, E., Raymond, L.: Statistical methods in cancer research. Volume IV: Descriptive epidemiology. International Agency for Research on Cancer, Lyon (1994)

[8] Fisher, R. A.: On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of Royal Statistical Society, 85(1): 87-94 (1922)

[9] Fisher, R. A.: Statistical Methods for Research Workers. Oliver & Boyd, London (1925)

[10] Ge, Y., Dudoit, S., Speed, T. P.: Resampling-based multiple testing for microarray data analysis. Technical report #633, University of California at Berkeley (2003)

[11] Hemkens, L. G., Grouven, U., Bender, R., et al.: Risk of malignancies in patients with diabetes treated with human insulin or insulin analogues: a cohort study. Diabetologia, 52: 1732–1744 (2009)

[12] Kruskal, W., Wallis, W. A.: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47 (260): 583–621 (1952)

[13] Lehmann, E. L., Casella, G.: Theory of Point Estimation. Springer-Verlag, New York (1998)

[14] Mann, H. B., Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18 (1): 50–60 (1947)

[15] Marubini, E., Valsecchi, M. G.: Analysing Survival Data from Clinical Trials and Observational Studies. John Wiley & Sons, New York (2004)

[16] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2): 153–157 (1947)

[17] Motulsky, H.: Intuitive Biostatistics, A Nonmathematical Guide to Statistical Thinking. 2nd edition, Oxford University Press, New York (2010)

[18] N. E. J. M. Editors: Looking Back on the Millennium in Medicine. N Engl J Med, 342, 42-49 (2000)

[19] Pagano, M., Gauvreau, K.: Principles of biostatistics. 2nd edition, Brooks/Cole, Cengage Learning, Belmont (2000)

[20] Pagano, M., Halvorsen, K. T.: An Algorithm for Finding the Exact Significance Levels of r × c Contingency Tables. Journal of the American Statistical Association, 76: 931-934 (1981)

[21] Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50 (302): 157–175 (1900)

[22] Pearson, K.: On the theory of contingency and its relation to association and normal correlation. In: Mathematical contributions to the theory of evolution. Dulau and Co., London (1904)

[23] Pocock, S. J., Geller, N. L., Tsiatis, A. A.: The Analysis of Multiple Endpoints in Clinical Trials. Biometrics, 43(3): 487-498 (1987)

[24] Schucany, W. R.: Jackknife Method. In: Armitage, P., Colton, T. (eds.): Encyclopedia of Biostatistics, 2nd Edition, Vol 4: 2651-2652. John Wiley & Sons, New Jersey (2005)

[25] Shapiro, S. S., Wilk, M. B.: An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52: 591-611 (1965)

[26] Sokal, R. R., Rohlf, F. J.: Biometry, The principles and practice of statistics in biological research. 3rd edition, W. H. Freeman and company, New York (1995)

[27] Spearman, C.: The proof and measurement of association between two things. Amer J Psychol, 15: 72–101 (1904)

[28] Steel, R. G. D.: A rank sum test for comparing all pairs of treatments. Technometrics, 2: 197–207 (1960)

[29] Student: The Probable Error of a Mean. Biometrika, 6(1), 1-25 (1908)

[30] Welch, B. L.: The Significance of the Difference Between Two Means when the Population Variances are Unequal. Biometrika, 29 (3–4): 350–362 (1938)

[31] Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1: 80-83 (1945)

[32] Wilk, M. B., Gnanadesikan, R.: Probability plotting methods for the analysis of data. Biometrika, 55(1): 1–17 (1968)

[33] Wooding, W. M.: Planning Pharmaceutical Clinical Trials: Basic Statistical Principles. John Wiley & Sons, New York (1994)

[34] Woodworth, G. G.: Biostatistics, A Bayesian Introduction. John Wiley & Sons, New Jersey (2004)

[35] Xu, F., Garcia, V.: Intuitive statistics by 8-month-old infants. PNAS, 105, 5012–5015 (2008)

[36] Zar, J. H.: Biostatistical Analysis. 5th edition, Pearson Prentice-Hall, New Jersey (2010)

[37] Zvára, K.: Biostatistika. Nakladatelství Karolinum, Praha (2006)

[38] Zvárová, J.: Základy statistiky pro biomedicínské obory. Nakladatelství Karolinum, Praha (2004)

Obsah

Summary

The publication *Biostatistics* was funded as a part of the ESF project no. CZ.1.07/2.2.00/07.0318 entitled „MULTIDISCIPLINARY INNOVATION OF STUDY IN COMPUTATIONAL BIOLOGY“, which was investigated at the Faculty of Science, Masaryk University. This project aimed to improve study courses that form a core of the Computational Biology study programme at the Masaryk University.

Therefore, our target readers are the students of Computational Biology, to whom we want to provide a comprehensive overview of the essentials of biostatistics in the context of real-life biological and clinical data. However, given the scope of the publication, only the key topics and methods of biostatistics could be included, which caused, on the other hand, that a number of frequently used methods are not covered in this textbook. First chapter of this publication serves as an introduction to the field of biostatistics as well as to key aspects that are necessary to know when evaluating data. Second chapter is devoted to data types and data summarization. Chapters 3 and 4 are focused on random variables as well as the most common probability distributions. Fifth chapter introduces elementary theory of the point and interval estimates and shows how to find interval estimates for parameters of the normal probability distribution. Chapter 6 is devoted to key aspects of hypotheses testing, namely to difference between the null and the alternative hypothesis, *p*-value, and the relationship between statistical and practical significance. Chapters 7 and 8 focus on testing hypotheses on quantitative data whereas chapter 9 presents methods for testing hypotheses on qualitative data. An introduction to correlation analysis is given in chapter 10.

This publication has no ambition to replace any existing textbook on statistics or biostatistics. It is always better for students as well as other readers to learn from multiple sources and to acquire a comprehensive picture of the considered issue. This is especially true in biostatistics, where each author represents a different perspective of the field and different experience from real practice. Moreover, the textbook does not substitute lectures, but serves only as their complement.

Our goal was to explain the methods of biostatistics correctly from both the theoretical and practical point of view. Therefore, the textbook focuses also on the practical aspects of data assessment in addition to the methodical background; it particularly focuses on the pitfalls that can anyone meet during the process of data preparation and calculation and interpretation of results. We hope that *Biostatistics* will serve to the students not only as a material for passing the exam, but also as a reference text for their own data assessment in bachelor's and master's theses.

Biostatistika
RNDr. Tomáš Pavlík, Ph.D.; doc. RNDr. Ladislav Dušek, Dr.

Recenzenti: doc. Vladimír Rogalewicz, CSc.; Mgr. Ondřej Pokora, Ph.D.
Jazyková korekce: ing. Marie Juranová
Obálka: Radim Šustr, DiS
Vydalo: AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o. Brno,
Purkyňova 95a, 612 00 Brno
www.cerm.cz
Tisk: FINAL TISK s.r.o. Olomučany
Náklad: 200 ks
Vydání: první
Vyšlo v roce 2012

ISBN 978-80-7204-782-6