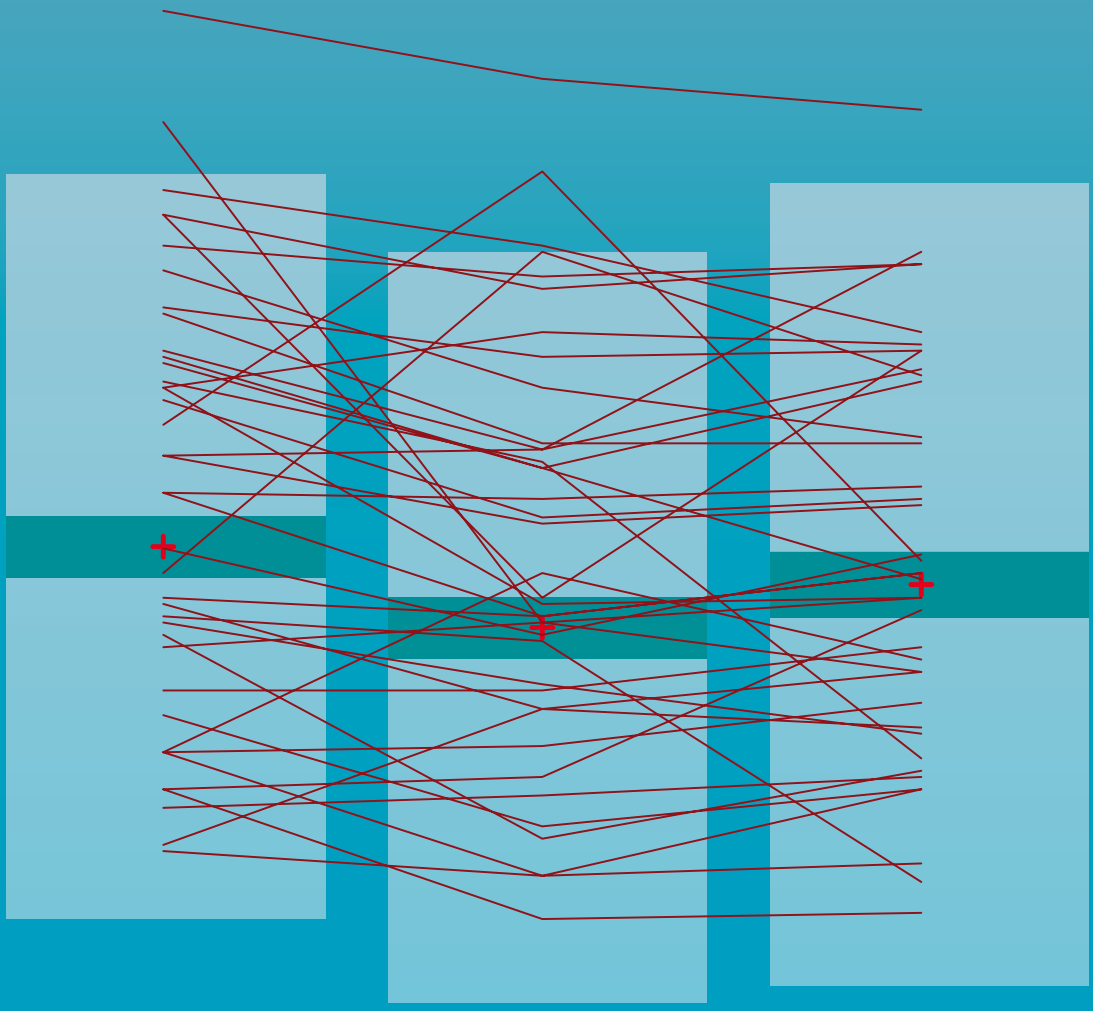
BIOSTATISTIKA PRO LÉKAŘE

PRINCIPY ZÁKLADNÍCH METOD A JEJICH INTERPRETACE S VYUŽITÍM STATISTICKÉHO SYSTÉMU R

BOHUMÍR PROCHÁZKA

KAROLINUM



Biostatistika pro lékaře
Principy základních metod a jejich interpretace s využitím statistického systému R

RNDr. Bohumír Procházka, CSc.

———

Recenzovali:
prof. RNDr. Jan Hendl, CSc.
MUDr. Zdeněk Šmerhovský, Ph.D.

Vydala Univerzita Karlova v Praze
Nakladatelství Karolinum
Obálka Jan Šerých
Sazba Bohumír Procházka
Vydání první

ISBN 978-80-246-2782-3
ISBN 978-80-246-2803-5 (online : pdf)

Univerzita Karlova v Praze
Nakladatelství Karolinum 2016

www.karolinum.cz
ebooks@karolinum.cz

Obsah

1. Úvod

V současné době se mezi lékaři skloňuje ve všech pádech pojem „medicína založená na důkazu“ a cílem je klást důraz na nejnovější znalosti a především na objektivnost hodnocení nejnovějších poznatků. Klíčovou roli tak získává vědecké uvažování často založené na principech statistické indukce.

Se statistikou se setkáváme nejen ve všech vědních oborech, ale i v běžném životě. Je často chápána zcela odlišnými způsoby – od představy, že statistika poskytuje naprosto přesné, nezvratitelné výsledky, až po názor, že statistika umožňuje dokázat jakékoliv tvrzení. Obě tyto představy jsou zcela mylné a vycházejí z neznalosti principů statistického uvažování. Snadno pak vzniká představa, že statistika je jakýsi moderní druh magie.

Největší problémy statistiky jsou spojeny s nepochopením základních principů statistického uvažování a se špatným popisem použitých pojmů nebo jejich chybným chápáním. Není důležité znát matematické vzorce, ale pochopit způsob uvažování a využít logické myšlení, které nutně musí používat každý. Cílem této knihy je přiblížit statistické uvažování a odstranit tak všechny předsudky spojené se statistikou.

Nejprve si řekněme pár slov z historie. V určitém smyslu se statistika používá od nepaměti. Již první použití čísel bylo vlastně statistikou. Člověk potřeboval popsat množství nebo velikost, vladaři potřebovali znát velikost a složení vojska, množství zásob a podobně. První zárodky disciplíny nesoucí název statistika se objevují již v 16. století. Mají především formu zeměpisného, hospodářského a politického popisu státu (Francesko Sansovin, 1562, Veit Ludvig von Seckendorf, 1662). Pojem stav (status) státu dal jméno této disciplíně a na dlouhou dobu formoval i její počáteční náplň jako slovní popis charakteristik státu (ale i zde již existovala kvantitativní vyjádření).

Později se začíná stále více prosazovat aritmetický pohled – sledují se počty obyvatel, narození a úmrtí (Johann Peter Sűßmilch, 1741, Gottfried Achenwall, 1781). To vše je zaznamenáváno především pro potřeby státu. Proto mají velký význam počty mužů schopných vojenské služby nebo hospodářské statistiky (např. počty dobytka). Touto orientací se jednoznačně formuje cíl zájmu statistiky. Cílem je popis společných vlastností skupiny objektů a jeho kvantifikativní vyjádření.

V té době začínají v mnohem těsnějším sepětí s matematikou vznikat základní myšlenky teorie pravděpodobnosti. Hlavním impulzem pro vznik této disciplíny byla odvěká snaha získat bez námahy velký majetek – nejjednodušší cestou, přístupnou každému, se zdály hazardní hry. První pokusy teorie pravděpodobnosti tedy byly snahou najít „šťastné“ číslo nebo jakoukoliv cestu, jak „spoutat“ náhodu. Teoretické základy této disciplíně dali např. Jacobi Bernoulli (1713), Christiani Hugenii (1714) nebo Abraham de Moivre (1718). Teorie pravděpodobnosti se později stala hlavním teoretickým základem statistiky.

Neměli bychom zapomínat ani na počátky statistiky ve zdravotnictví. Zde se první statistické přístupy objevují též v 17. století – anglický obchodník John Graunt pracoval s týdenními počty zemřelých na různé příčiny. V 18. století anglický lékař James Lind za pomoci sledování počtů nemocných prokázal, že konzumace citronů snižuje výskyt kurdějí u námořníků. Londýnskému lékaři Johnu Snowovi se podařilo při epidemii cholery v roce 1849 nalézt zdroj infekce v pitné

vodě pomocí zakreslování případů cholery do mapy Londýna. Florence Nightingalová, legendární postava ošetřovatelství, ale i průkopnice použití statistiky ve zdravotnictví, prokázala v době krymské války (1854–1855) významný vliv nedostatečné nemocniční hygieny na úmrtí pacientů.

Vraťme se k vlastní statistice. Původní metodou, jak získat informaci pro vytvoření statistického popisu, bylo úplné sčítání všech sledovaných charakteristik na základě úplných výkazů v celém státě. Tento přístup přežívá dodnes například v podobě pravidelného sčítání lidu. V laické společnosti je právě toto pojetí silně spojeno nejen s pojmem statistika, ale i s představou aritmetické přesnosti. Použití takovéhoto přístupu je ale spojeno s dvěma velkými problémy:

– Získání takovýchto dat je v praxi vzhledem k technické a ekonomické pracnosti často nedosažitelné.

– Aritmetická přesnost sebraných dat při úplném sčítání je stejně velmi problematická. Například když uvažujeme počet obyvatel, je údaj poplatný přesnému okamžiku (pokud vůbec) a o okamžik později je neplatný. Navíc i takto získaná čísla nemusí být přesná (obecně není možno předpokládat, že výkazy jsou bezchybné). Aritmetický součet pak může být naprosto přesným součtem nepřesných čísel.

Představa velké přesnosti je tedy pouhou fikcí a navíc ani nemá praktické použití (je zbytečné měřit hmotnost postavy s přesností na miligramy nebo velikost populace státu s přesností na jedince).

Další rozvoj statistiky se pak logicky orientoval na postupy, jak získat dostatečně přesnou představu o uvažované charakteristice sledováním pouhé části celého souboru, jak zajistit, aby tato představa platila i pro celý soubor, a jak popsat přesnost získaných výsledků. Statistika a teorie pravděpodobnosti se proto výrazně sbližují (téměř splývají). To vedlo na přelomu 19. a 20. století k přehodnocení tehdy používaných postupů a k vzniku moderního pojetí statistiky (objevují se taková jména jako například Pearson, Student, Fisher a další). V rámci této přeměny statistiky se začaly rozvíjet i metody pro studium vztahů sledovaných charakteristik. To postupně otevřelo dveře pro použití statistických metod jako nástrojů pro popis a studium hromadných jevů ve všech vědních disciplínách.

Co způsobuje vznik pověstí o statistických lžích, pověstí o tom, že statistikou je možno dokázat cokoliv?

Stejně jako v jiných disciplínách je i ve statistice možno použít její nástroje dobře i špatně. Statistické výsledky není možno chápat bez znalosti alespoň základů statistického uvažování. To ale nestačí, s publikovanými výsledky je nutno poskytnout i informace o postupech a podmínkách, za jakých byly tyto výsledky získány.

– Nezkušený čtenář se často ani nezajímá o to, v jakých podmínkách byla studie provedena, ani kterými postupy byly výsledky získány. Často použije výsledky za podmínek, které vůbec neodpovídají původní práci. Tomu se samozřejmě čtenář může bránit seznámením se základy statistického myšlení a seznámením s podmínkami, za jakých byla studie provedena.

– V mnohých pracích chybí popis podmínek studie, pak ale tuto studii není schopen použít ani znalý čtenář (stejně jako lékař nepoužije neznámý lék, byť od renomované firmy).

– Největším problémem je to, že statistické metody jsou často používány zcela neodborně. Obecně je často uznávána teze, že k provedení statistické analýzy stačí pouhá znalost aritmetiky. Proto je za statistika považován každý, kdo umí pracovat s počítačem. V posledních letech se na trhu objevilo mnoho programů, které jsou schopny provádět i složité analýzy. Široké veřejnosti se do rukou dostávají velmi silné nástroje, ale většina uživatelů neví, jak je použít, ani to, jak chápat výsledky. To je ale něco podobného, jako bychom

pacientovi nabídli všechny technické prostředky medicíny a nechali jej, ať se léčí, jak chce. V medicíně tomu brání zákon. Nekvalifikovaná léčba a případné chyby jsou jasně viditelné a v extrémním případě jsou i trestně stíhány. Chybné použití statistiky se týká pouze „populace“, následky chybného rozhodnutí na základě špatného použití statistiky se netýkají přímo jedné konkrétní osoby, ale nepřímo všech.

- K chybám dochází i vlivem špatné interpretace výsledků, například záměnou kauzality. Porovnáváme-li stravovací zvyklosti zdravých a nemocných osob, nezískáme informaci o rizikových faktorech, ale spíše zjišťujeme, zda vědomí o onemocnění způsobuje změnu chování. Pouhé technické zpracování dat nezajistí správnou interpretaci výsledků, ostatně výpočty jsou jen částí statistické práce.
- Důvodem k výroku o statistické lži nebývají chybné údaje, ale matoucí, nedostatečný popis toho, co autor publikuje, a odlišné chápání čtenáře a autora. Přispívá k tomu naše představa, že čísla dokážeme sami dobře interpretovat. Když například řekneme, že hladina glykemie v krvi je $4,5 \pm 0,8$, budeme to chápat tak, že mluvíme o průměrné hodnotě pro nějakou skupinu pacientů, nebo je to interval, v kterém očekáváme hodnoty „většiny“ těchto pacientů?
- Problémy vznikají často i zcela nevhodným použitím statistických nástrojů a přenesením výsledků do jiných podmínek, než za jakých byly pořízeny. Dochází velmi často k závažnému zkreslení (někdy i cílenému). Například sledování inflace je založeno na tzv. „spotřebním koši“ (tj. na jakési zvolené struktuře nákupů), a čím více se od tohoto koše lišíme, tím méně je pro nás informace o inflaci užitečná.
- Dalším problémem může být autocenzura, kdy se autoři rozhodli nepublikovat nevýznamné výsledky studií. Toto zkreslení skutečnosti podporují i mnohé odborné časopisy, když odmítají publikovat statisticky nevýznamné výsledky, což má za následek tzv. **publikační bias**.

Opusťme nyní úvahy o problémech špatného použití statistiky a věnujme se tomu, čím může být statistika užitečná pro medicínu.

Potřebou vědecké práce je často studovat různé hromadné jevy a jejich vztahy pomocí nástrojů a postupů, které zaručují porovnatelnost výsledků získaných i na vzdálených místech. K změření hodnot sledovaných veličin na jednotlivých objektech nestačí pouze používat porovnatelné prostředky, je nutno zajistit i srovnatelné posuzování získaných výsledků. Je nutno různé studie hodnotit a porovnávat objektivními metodami – metodami, které je možno kdykoliv a kdekoliv zopakovat jak na jiných datech, tak i v jiných podmínkách.

Na první pohled se může zdát, že biostatistika je důležitá pouze pro výzkum. To zdaleka není pravda. Obecně uznávanou skutečností je, že žádný odborník nevystačí se znalostí získanou při studiu, ale je nucen neustále sledovat vývoj vlastního oboru. Je tedy nezbytné číst odbornou literaturu a být schopen chápat publikované výsledky nejen správně, ale i kriticky. Musíme tedy znát alespoň základní principy statistického uvažování.

Dokonce i v běžném životě se často setkáváme s výsledky různých šetření. Novináři nám předkládají různá hodnocení a politici často používají jako argumenty výsledky různých statistických studií. Nevhodná interpretace pak může snadno poskytovat falešný obraz, který vzniká jak na straně posluchače, tak i neúmyslným (či úmyslným) zkreslováním skutečnosti ze strany hodnotitele. Pokud nevíme, za jakých podmínek byl publikovaný ukazatel vytvořen, nevíme vlastně nic a snadno se může stát, že máme proti sobě dvě zcela odlišná tvrzení jen kvůli tomu, že každý autor chápe zmíněný ukazatel jinak (i když používají stejný název).

Tato kniha uvádí do problematiky statistického uvažování a informuje o základních možnos-

tech a metodách matematické statistiky. Vybírá pouze některé často používané nástroje. Důraz je kladen na vysvětlení jejich principů a interpretaci výsledků, ne na přesný popis metod. Cílem tedy není popsat základní používané postupy tak, aby bylo možno podle textu napsat počítačové programy, cílem je seznámit se základními principy a metodami používanými lékaři a biology, orientovat se ve spleti statistických postupů a získané výsledky správně interpretovat.

Cílem této knihy není na rozdíl od [48] pouze popis metod a interpretace získaných výsledků, ale i seznámení s volně šiřitelným programem R, který umožňuje provádět popisované analýzy. Systém je mnohem širší a neustále se rozvijí. Tato kniha může zmínit pouze základní metody. Systém není pouze pro profesionální statistiky, existuje řada jiných odborníků, kteří jej používají.

Je připravována i stručnější varianta [49], obsahující pouze základní statistické metody, bez popisu práce s programem R a bez detailnější interpretace metod.

Pro jednodušší seznámení s programem R jsou na stránkách nakladatelství Karolinum vystavena cvičná data a skripty (viz dodatek E).

Pro prezentaci praktických příkladů a ukázek výpočtů je v této knize použit volně šiřitelný počítačový program R (The R Foundation for Statistical Computing, ISBN 3-900051-07-0), který je možno získat na adrese http://cran.r-project.org/. Tento software je jedním z předních systémů, které umožňují provádět i velmi složité analýzy, a často je používán v statistické odborné literatuře. Počítačové příkazy a výstupy jsou použity v originální formě, tj. v anglickém jazyce. Přestože kniha nemá za cíl vysvětlit principy práce s R, tak počítačový výstup obsahuje příkazy jazyka systému R.

V dodatku A je popsána jen nejzákladnější práce s R. V knize jsou veškeré texty týkající se práce s R, s výjimkou zmíněného dodatku, označeny svislou čarou vlevo s logem R na začátku tohoto bloku (stejně jako tento odstavec) a jsou psány stylem `verbatim`, tj. strojovým písmem. Vzhledem k tomu, že program je napsán pro anglické jazykové prostředí, je ve výstupech použita angličtina a reálná čísla budou tedy mít desetinná místa jsou v reálných číslech oddělena desetinnou tečkou nikoliv čárkou. V textu je použita dle české normy desetinná čárka. Pokud nastane situace, že text zobrazený programem delší než šířka stránky této publikace bude text rozdělen do dvou řádek.

Program R je programem, v kterém se příkazy zadávají na příkazovou řádku. V knize si ukážeme, že tohoto způsobu ovládání se nemusíme bát. Výhodou tohoto ovládání je veliká flexibilita a možnost konstrukce všech modelů, i těch nejsložitějších. Program ale nabízí i ovládání přes menu (balíček `Rcmdr`). Drobnou výhodou tohoto systému je cena – je volně šiřitelný.

Hlavním problémem pro statistické výpočty není ovládání počítačového programu, ale volba metody a správná interpretace získaných výsledků a to za nás nemůže udělat žádný program; je nutno znát principy a možnosti statistické analýzy.

Poděkování:

Rád bych poděkoval za konzultace, přečtení textu a připomínky Mgr. Ondřeji Vencálkovi, Ph.D., i Ing. Heleně Šebestové a dalším kolegům. Velký význam pro mne měly i reakce studentů 3. LF UK v průběhu kurzu biostatistiky. Popisované metody jsem byl schopen ukázat na praktických příkladech jen díky laskavému souhlasu řešitelů citovaných studií Státního zdravotního ústavu, Institutu postgraduálního vzdělávání ve zdravotnictví a Ústavu hematologie a krevní transfuze a dalších pracovišť. Dík patří i mé manželce a celé rodině nejen za pochopení, když jsem trávil čas psaním textu, ale i za odbornou pomoc.

Pro tvorbu vlastního textu jsem použil textový editor LaTeX. Jednotlivé výpočty a generování grafů jsem provedl za pomoci systému R.

Většina dat byla sebrána s použitím programu EpiInfo, případně pomocí programu MS Excel nebo „na míru“ vytvořených aplikací.

Připomínky či poznámky k této knize rád uvítám na e-mailové adrese bpro@post.cz.

2. Obecné úvahy

2.1. Rozdíly biologie a matematiky

Vědní disciplíny zabývající se popisem reálného světa, jako je například biologie a medicína, mají zcela jiný pohled na objekty vlastního zájmu než disciplíny matematické. Pomoc matematických disciplín je ale pro biologii nejen velmi užitečná, ale i nutná. Obzvlášť když nevystačíme s pouhým popisem dat, ale potřebujeme porozumět principům či získat nějaké závěry.

Biologie a medicína sledují velmi složité jevy, které jsou vzájemně provázané a silně závislé na prostředí. Hlavním cílem medicíny je navíc léčit, všechen její výzkum se nutně musí řídit především etikou. Tato složitost způsobila, že biomedicínské disciplíny byly v počátcích svého rozvoje zaměřeny pouze na vnější popis objektů a jevů, jejich roztřídění do různých systémů a schémat. Později se zájem zaměřil i na různé funkce a vztahy.

Složitost studovaných objektů nás nutí provést určité zjednodušení. Vždy musíme zvolit nějakou úroveň,[1] na které chápeme sledované objekty jako černé skříňky, které nelze dále dělit. Jsou sledovány jejich vlastnosti a na jejich základě je vytvářena zjednodušená představa. Tato představa vychází z použitého modelu a z toho, že zbylé rozdíly mezi jedinci jsou shrnuty do „chyby měřené veličiny“, obsahující jednak nepřesnosti měření, ale i systematické odchylky, způsobené skutečnostmi, které nedokážeme identifikovat (nebo které jednoduše nechceme do modelu zahrnout). Takto vytvářené modely sice nějak popisují realitu, ale nejsou jejím přesným obrazem. Dokonce ani ty nejsložitější modely nejsou přesným obrazem skutečnosti, pouze se jí více či méně přibližují. Cílem statistických analýz je nalézt obecné vlastnosti a vztahy sledovaných objektů. Velkým uměním je samozřejmě nalézt vhodný kompromis – úroveň zobecnění. Je nutno zvolit dostatečně podrobný model. Na druhou stranu může nadměrné zaměření na detaily způsobit, že se studie rozpadne na jednotlivé detaily a nebude možno získat obecné závěry.

Zcela opačný přístup používá matematika a teorie pravděpodobnosti. Postupně vytváří objekty svého zkoumání, od nejtriviálnějších formálních struktur k stále složitějším. To, že se matematika zabývá studiem formálních objektů, umožňuje jednak jejich přesnou znalost, jednak to dovoluje i postupně odvozovat stále složitější vztahy či zavádět složitější pojmy.

Výsledkem matematických úvah je formální odvození vztahů. Výsledky pak platí vždy, pokud jsou splněny přijaté předpoklady. Tato jistota, která je v ostatních vědních disciplínách zcela neobvyklá, je dána jednoduchostí studovaných struktur a především tzv. **deduktivním** způsobem **uvažování**.[2] Matematika pak poskytuje vědním disciplínám zabývajícím se reálnými objekty nástroje pro zjednodušený popis objektů. Samozřejmě je nutno, aby zkreslení způsobené zjednodušením bylo přijatelné.

Vědní disciplíny zabývající se skutečnou realitou sledují složité objekty a snaží se popsat jejich společné vlastnosti a vztahy mezi nimi. Tím vlastně provádí jisté zjednodušení, které umožňuje použít matematické nástroje. Na druhou stranu obvykle sledujeme jen určitou (obvykle malou) část populace. Zajímá nás ale celá populace, jinými slovy nás zajímají i další objekty, které jsme

[1] Například skupiny lidí, vzorky tělesných orgánů, buňky, molekuly, nebo dokonce atomy.

[2] V matematice se na základě určitých předpokladů odvozují postupně různé vlastnosti sledovaných objektů. Na základě jednoduchých „stavebních kamenů“ se budují stále složitější konstrukce.

nestudovali. Získané výsledky se pak snažíme zobecnit. Tento způsob **uvažování** je nazýván **induktivní**. Přírodní a lékařské vědy jsou charakteristické velkou složitostí sledovaných objektů, ty ve skutečnosti nikdy není možno popsat do detailu. Vždy je nutno na určité úrovni zahrnout do neurčitosti individuální rozdíly – přisoudit je „náhodě“. Ta může zahrnovat i vliv různých složitých vztahů, které často ani netušíme.

Největším problémem statistiky v medicíně je navázat komunikaci mezi statistikou a medicínou, tj. nalézt optimální matematický model a získané výsledky správně interpretovat.

Na tomto místě je nutno zmínit problémy spojené s použitím matematických metod pro řešení praktických úkolů. Nejen že musíme sledovanou skutečnost zjednodušit tak, aby bylo možno vytvořit adekvátní matematický model, ale také je třeba si uvědomit, že tento model nutně má různé formální předpoklady (např. že chyba měření se stejnou pravděpodobností zkresluje sledovanou hodnotu nahoru i dolů). Aby bylo možno matematický model použít, musíme tyto předpoklady přijmout. Často se jedná o triviální, lehce akceptovatelné vlastnosti. Některé je nutno důsledně zvážit, a některé jsou dokonce tak abstraktní, že vzhledem k realitě je téměř nelze posoudit. Pro řešení konkrétních problémů může existovat i více „správných“, nicméně odlišných modelů. Hlavním uměním biostatistiky je výbrat vhodný, přiměřeně složitý model (a získané výsledky správně interpretovat).

Z pohledu interpretace můžeme použít induktivní způsob popisu, jakých hodnot nabývá sledovaná charakteristika (např. výška postavy u všech dospělých osob v ČR). Řekněme, že máme jen omezenou část těchto osob. Skupina měřených osob musí samozřejmě dobře „reprezentovat“ celý soubor[3]. Pro popis celého souboru nás nezajímá pouze jeden charakteristický reprezentant, ale chceme vystihnout, jak vypadá celé spektrum hodnot v populaci, mluvíme tedy o **rozložení** hodnot sledované veličiny. Zajímá nás, jaké hodnoty můžeme očekávat. Nebo nás zajímá „skutečná hodnota“ sledované charakteristiky pro celou populaci (např. průměrná výška postavy). Tuto hodnotu nemůžeme nikdy znát zcela přesně, ale budeme chtít chybu tohoto stanovení minimalizovat. Později popsané metody umožní získat nejen její odhad, ale i představu o přesnosti tohoto odhadu, případně popsat vztah různých měřených charakteristik.

V lékařských vědách je možno sledovat různé jevy s větší nebo menší přesností. Například v oblasti farmakokinetiky můžeme stanovit v laboratorních podmínkách koncentraci sledované látky poměrně přesně – chyba zkreslení vlivem „náhody“ je poměrně malá. Na druhé straně, například v oblasti psychologie, jsou sledované charakteristiky (odpovědi na otázky) zatíženy velkou chybou.

2.2. Přístupy k řešení problémů

V praxi je možno přistupovat k hodnocení různých sledovaných jevů dvěma způsoby:

Individuálně – zajímají nás konkrétní případy jako neopakovatelné jevy. Sledujeme pacienty, o kterých pak uvažujeme jednotlivě. Výsledky nejsou přímo použitelné k žádnému zobecnění na jiný případ. Jedná se tedy o pouhý popis konkrétního případu – o kazuistiku.

Skupinově – zajímají nás obecné vlastnosti. Sledujeme skupinu objektů a jejím prostřednictvím chceme získat v určitém smyslu obecné vlastnosti. Snažíme se stanovit, jaké hodnoty se mohou v uvažované populaci vyskytovat a s jakou mírou. Právě to je prostor pro použití statistiky.

[3] Je možno odhadovat sledovanou veličinu, i pokud máme informaci o způsobu „nereprezentativnosti“ pokud je jasně popsána. Například odhadovat výšku plnoletých mužů na základě informací o odvedencích, když víme, že odvedeni byli pouze muži vyšší než 150 cm.

Statistické metody se snaží opakovaným sledováním určité skutečnosti omezit rozdílnost výsledků způsobenou vlivem „náhody“ a odhalit sledovanou zákonitost.

Statistika tedy poskytuje nástroj k popisu vlastností objektu jako člena rozsáhlejší skupiny. Přístupy k řešení takovýchto problémů je možno rozdělit do dvou skupin:

Deskriptivní – **popisná statistika.** Poskytuje pouze výčet pozorovaných jedinců a jejich vlastností, nepokouší se vyslovovat k vlastnostem jedinců, kteří nebyli sledováni. Tento přístup je používán v případě, že jsou pozorováni všichni jedinci sledované skupiny.

Induktivní – **zobecňující statistika.** Je moderní metoda matematické statistiky poskytují nástroje pro zobecňování výsledků na širší populaci. Tento přístup umožňuje zkoumat pouze část celé populace a získat pokud možno co nejpřesnější společný odhad sledovaných (obecných) charakteristik celé populace.

Někdy se setkáváme i s termínem **explorativní statistika**. Jedná se o přístup vyhledávající možné vztahy, které použitá data umožňují indikovat. Pro takovéto důsledné využívání získaných dat a generování všech možných hypotéz se někdy používá anglický termín „**data mining**“, který je možno interpretovat jako snahu o maximální „vytěžení“ informací z dat. Je to moderní, často používaný postup, má jistě svoje oprávnění při „orientaci“ v datech, ale je namístě být velmi obezřetný při prezentaci získaných výsledků. Provádíme velké množství porovnání „všeho se vším“ a následně hledáme interpretaci výsledků často za každou cenu. Snadno pak díky náhodě získáme falešné výsledky. Tyto postupy je tedy možno využít pro hledání hypotéz, ale nutně vyžadují následné ověřování na jiných datech. Jinak se můžeme později dostat do nesnází, kdy jiné studie nedokáží zopakovat naše výsledky.

Věnujme nyní pozornost jedné ze základních myšlenek, které se používají v rámci statistického uvažování.

2.3. Populace a výběr – základ statistické indukce

K vysvětlení principů **induktivní statistiky** je nutno nejprve zavést dva pojmy: **základní populace** – skupina subjektů, které nás zajímají a o kterých chceme mluvit, ale z nichž ne všechny budeme nebo jsme schopni měřit (popisovat). **Výběr** – obvykle mnohem menší skupina obsahující jedince, které máme k dispozici například pro měření či sledování.

Pokud používáme deskriptivní statistiku, týkají se naše tvrzení pouze souboru, na kterém byla prováděna měření (pozorování a podobně). V tomto případě je výběr totožný se základní populací. Získané výsledky popisují pouze zkoumaný soubor a nesnaží se o žádné zobecnění na větší nebo jinou skupinu objektů. Stačí tedy mluvit o získaných charakteristikách a ty popisují sledovaný soubor zcela přesně. Deskriptivní statistiky byly používány především v raných dobách statistiky. Jejich problémem pro velké populace je především vysoká pracnost, a v důsledku toho pak špatná validace všech dat (data mohou být systematicky zkreslena respondentem). V neposlední řadě je tu i interpretační problém – obvykle nás nezajímají objekty, které jsme měřili, ale chceme mluvit o „obdobných“ objektech. Např. průměrná porodní hmotnost nás zajímá spíš u těch novorozenců, kteří se teprve narodí, než u těch, které jsme vážili. Příkladem deskriptivní statistiky, která je používána dodnes, jsou například výsledky sčítání lidu. Jeho výsledkem je mimo jiné popis složení populace republiky podle věku, pohlaví a lokality. Právě tyto údaje se pak často používají jako základ pro hodnocení a plánování různých studií.

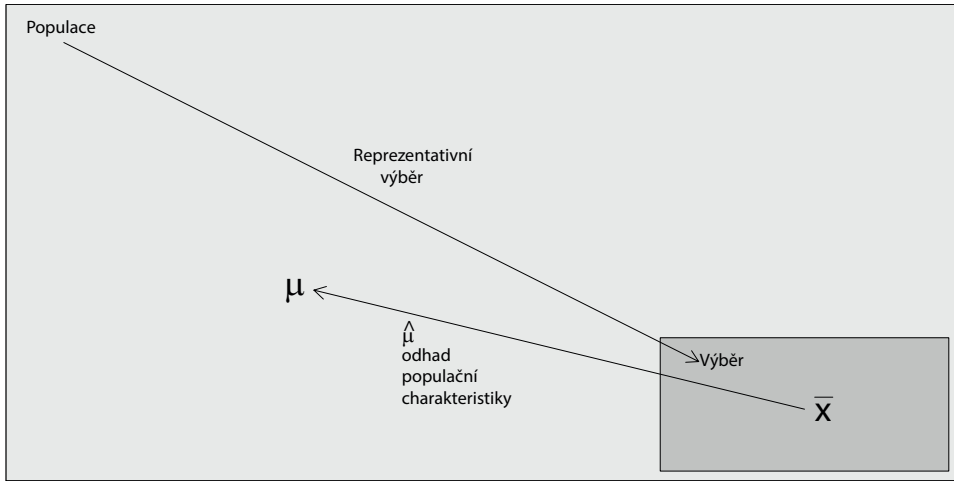
Induktivní statistika se snaží výsledky získané na výběru zobecnit (generalizovat na širší skupinu objektů) na základní populaci. Důvodů, proč je nutno pracovat pouze s výběrem, a ne s celým souborem (s celou základní populací), může být více. Jednak analýza celého souboru ne-

musí být proveditelná z ekonomických či technických důvodů, jednak můžeme chtít předpovídat vlastnosti objektů, které v době analýz ani neexistují – například můžeme na základě sledování vybraných charakteristik dětí předpovídat hodnoty pro ještě nenarozené jedince. Vlastně jsme v situaci, jako bychom reálný svět pozorovali jen malým okénkem, ale chtěli mluvit o celém „světě“. Náš pohled je pak nutně nepřesný a „kvalita“ tohoto okénka určuje, jak je naše představa „reálná“.

Podstatné je, že chceme, aby bylo možno výsledky analýzy zobecnit (přenést) na podobné jedince. K tomu je nutno výběr provést tak, aby byla zajištěna jeho reprezentativnost. Tímto pojmem se budeme podrobněji zabývat později, v kapitole 13. Jen krátce si zmíníme příklady problémů spojených s reprezentativností. Například pokud má výběr jiné věkové složení než celý soubor a pokud analyzovaná veličina závisí na věku, pak mohou být informace získané výběrem nepoužitelné pro celou populaci. Nebo pokud bychom se ptali jednotlivých osob, koho by volily, mohou zkreslit výsledky průzkumu volebních preferencí ty které se k volbám nedostaví. Ke zkreslení dojde, jestliže osoby, které se k volbám nedostaví, mají jiné preference než osoby, které skutečně budou volit.

Kvalita vztahu mezi objektem našeho zájmu (celou populací) a našimi daty (výběrem) je určena reprezentativností výběru. Tato reprezentativnost zaručuje použitelnost odhadu, který je naším hledaným cílem (obr. 2.1).

Charakteristiky týkající se základní populace obvykle označujeme pomocí malých řeckých písmen, zatímco pro jejich protějšky z výběru používáme písmen latinské abecedy. Velké písmeno latinské abecedy používáme pro sledovanou veličinu a malé s indexem pro její pozorovanou hodnotu (hodnotu pro konkrétního jedince). Pokud například označíme sledovanou veličinu písmenem X (porodní hmotnost), pak jednotlivé pozorované hodnoty (individuální porodní hmotnosti) obvykle značíme x_i. „Skutečnou“ (společnou, ideální) hodnotu populační charakteristiky X pak značíme μ (v našem případě průměrnou porodní hmotnost), její odhad pak označujeme pomocí stříšky $\hat{\mu}$, a protože máme k dispozici jen výběr, pak výběrový průměr $\bar{x}$ budeme pokládat za odhad $\hat{\mu}$ průměru sledované charakteristiky (hmotnosti). Tato úvaha je založena na tom, že výběr dobře reprezentuje celou populaci.



Obrázek 2.1.: Princip induktivní statistiky

Celá induktivní statistika je tedy založena na dvou pojmech:

Základní populace a její charakteristiky. Jedná se často o velmi rozsáhlý soubor, jehož vlastnosti nás zajímají. Můžeme je definovat dvěma způsoby:

- První je výčet prvků souboru (například soubor všech voličů, soubor evidovaných diabetiků).
- Druhou možností je popis souboru pomocí vlastností jeho členů bez omezení na konkrétní skupinu osob. Například do souboru budou patřit osoby v produktivním věku léčené na diabetes. V tomto případě neomezujeme soubor na žádnou konkrétní populaci. Jindy je soubor definován tak, že do něj patří i ti, kteří se do této skupiny třeba teprve dostanou. Tím tato skupina vlastně nemá definovanou velikost (například soubor novorozenců, soubor diabetiků, ...).

Výběr a výběrové charakteristiky. Výběr je skupina objektů, na kterých provádíme šetření. To, jak výběr odpovídá základní populaci, určuje i kvalitu výsledku – přesněji řečeno kvalitu zobecnění závěrů získaných na základní populaci. Popisné charakteristiky výběru pak slouží jako odhady charakteristik celé populace – z výběru jsou indukovány odhady těchto charakteristik na celou základní populaci.

Z pohledu induktivní statistiky nás zajímá, jaké hodnoty sledované veličiny mají jedinci z celé populace. Mluvíme pak o **rozložení** sledované veličiny. Často je používán i termín **rozdělení**. Rozložením sledované veličiny v základní populaci rozumíme souhrn všech možných hodnot této veličiny základní populace a míru, s jakou můžeme tyto hodnoty očekávat. Jde tedy o seznam všech možných hodnot této veličiny společně s četnostmi těchto hodnot v základní populaci.

Charakteristikami základní populace pak rozumíme například průměrnou hodnotu sledované veličiny v celé populaci nebo její nejčastější hodnotu či míru „rozdílnosti“ – variability hodnot sledované veličiny a podobně. Tyto charakteristiky základní populace a rozložení hodnot v základní populaci vychází z její úplné znalosti těchto údajů, a nemá tedy smysl mluvit o induktivním myšlení – jsou známy přesně. Problém je, jak již bylo zmíněno, že obvykle neznáme, a dokonce často ani nemůžeme znát, celou populaci. Proto konstruujeme výběr (ze základní populace), který nám umožní sledovanou charakteristiku pouze odhadnout.

Pokud základní populace není definována jednoznačným výčtem (např. ji definujeme jako všichni „budoucí“ novorozenci), stává se pojem rozložení celé populace více abstraktním. Rozložením pak chápeme matematickou funkci, která každé možné hodnotě (porodní hmotnosti) přiřazuje míru pravděpodobnosti, s jakou můžeme tuto hodnotu očekávat.[4] To již navozuje dva různé pohledy na pojem slova „rozložení“. Definujeme tedy:

Empirické rozložení je rozložení naměřených hodnot použitého výběru, tj. výčet všech naměřených s jejich četnostmi.

Teoretické rozložení je matematický model (například normální – Gaussovo, Poissonovo, binomické a další), který udává pro všechny možné hodnoty sledované veličiny míru pravděpodobnosti, s jakou je možno tyto hodnoty pozorovat. Matematický model se snažíme vytvořit pomocí několika málo parametrů. Těmito parametry mohou být například průměr, míra variability, pravděpodobnost onemocnění jedince atd.

Na tomto místě bychom si měli uvědomit, že odhad populačních charakteristik provádíme pomocí našeho výběru (našich dat), ale že tento odhad silně závisí jak na reprezentativnosti našeho výběru, tak i na použitém teoretickém modelu rozložení sledované veličiny.

Základním principem induktivního uvažování je tedy předpoklad, že populace, o které chceme mluvit, má stejné vlastnosti jako použitý výběr.

[4] Respektive očekávat ve zvoleném intervalu.

3. Typy sledovaných veličin

3.1. Co můžeme sledovat

Pokud popisujeme nějaké objekty, je jejich popis založen na výčtu vlastností a ty mohou být vyjádřeny různými způsoby. Ve statistice používáme pro takovéto charakteristiky nebo vlastnosti termín **jev**. Pod tímto pojmem si můžeme představit výšku postavy, její hmotnost, množství cholesterolu v krvi, vzdělání, to, zda sledovaná osoba je nemocná, rodinný stav či krevní skupinu a podobně. Abychom s těmito jevy mohli hromadně pracovat, potřebujeme je převést do nějaké formální podoby, tj. vyjádřit je číselnou hodnotou nebo nějakou skupinou kódů. Tento formální (často číselný) obraz skutečnosti nazveme **znakem**.

Formálně můžeme sledované znaky rozdělit do dvou hlavních skupin:

Kvalitativní znaky jsou charakteristiky sledovaných objektů, které popisujeme různými slovy (omezenou skupinou slov). Například pohlaví sledované osoby je muž nebo žena. Jiným příkladem je kraj, ve kterém sledovaná osoba bydlí. Vzdělání můžeme dělit na základní, středoškolské a vysokoškolské.

Podle vlastností takovéhoto seznamu je možno kvalitativní znaky dělit ještě dále:

Nominální znaky jsou takové, které není možno navzájem uspořádat. Není možno o hodnotách říci, která je větší. Příkladem může být etnický původ, rodinný stav, diagnóza nebo krevní skupina či příslušnost studenta do studijního kruhu.

Ordinální znaky jsou naopak ty, které je možno navzájem uspořádat. Je možno říci, že základní vzdělání je nižší než středoškolské a to je nižší než vysokoškolské. Jiným příkladem může být stadium nemoci. U ordinálních znaků ale nelze určit míru toho, jak jsou od sebe jednotlivé kategorie vzdáleny. Například pokud budeme hovořit o zmíněném vzdělání, nevíme nic o srovnatelnosti rozdílu mezi základním a středoškolským vzděláním na jedné straně a středoškolským a vysokoškolským vzděláním na straně druhé.

Alternativní (binární) znaky jsou ty, které mohou nabývat pouze dvou různých hodnot. Například výskyt rizikového faktoru (ano/ne), indikace nemoci (zdráv/nemocen) nebo pohlaví (muž/žena). Tyto znaky můžeme zařadit do obou předcházejících skupin. Dvě hodnoty je možno vždy uspořádat, ale pokud je hodnot více než dvě, nemusí být uspořádání možné. Tím, že u alternativní veličiny mluvíme o uspořádání, nehodnotíme, co je víc, pouze od sebe odlišujeme dva možné kódy. Někdy se používá i kódování 0/1 nebo +/–.

Kvantitativní znaky, jak vyplývá z jejich označení, jsou znaky, jejichž hodnoty jsou nejen uspořádány, ale vyjadřují dokonce i určitou míru. Příkladem může být věk, různé míry, váhy, koncentrace, počty zárodků případů a podobně. Můžeme je dále rozdělit na:

Diskrétní (celočíselné) znaky jsou ty, které nabývají pouze celočíselných hodnot. Jedná se například o počty nemocných tuberkulózou, či počty zárodků po kultivaci na plotně s živnou půdou, počet infektů horních cest dýchacích a podobně. Často vznikají jako

hromadné pozorování alternativních znaků (např. počty nemocných – z pohledu skupiny osob se jedná o diskrétní (celočíselnou) veličinu a z pohledu jedince o veličinu alternativní).

Spojité znaky jsou ty, u kterých předpokládáme, že je možno je měřit s libovolnou přesností, a které nabývají hodnot reálných čísel – například to je hmotnost, výška postavy nebo koncentrace látky. (Ve skutečnosti jsou pozorované hodnoty vždy zaokrouhlené.)

Data s neúplnou informací je termín používaný pro jevy, které nejsme schopni vždy přesně pozorovat (obvykle se jedná o kvantitativní veličiny). Můžeme jen stanovit interval, ve kterém se sledovaná hodnota nalézá. Sem je možno zařadit takové spojité jevy, které jsme schopni pozorovat, pouze nad určitým detekčním limitem; o nižších hodnotách máme pouze informaci, že tohoto limitu nedosahují. Například při měření koncentrace polétavého prachu se naměřená hodnota udává, pouze pokud je větší než detekční limit, a o hodnotách nižších se pouze říká, že jsou pod detekčním limitem. Jednou z typických veličin s neúplnou informací je doba do nějaké události, třeba doba přežití. Pokud chceme hodnotit např. délku přežití po určité léčbě, pak bychom rádi využívali nejen informace o těch pacientech, kteří již zemřeli, ale i informace o tom, jak dlouho přežívají ti živí (tj. přežili do doby, kdy jsme je viděli naposled). Někdy máme dokonce nepřesnou informaci o úmrtí (víme, že zemřel mezi dvěma daty). Touto problematikou se budeme zabývat v kapitole 15.

Rozložení na diskrétní a spojité znaky je dáno jejich jiným charakterem – diskrétní znaky představují pouze celočíselné hodnoty a spojité zase nelze vlastně nikdy změřit zcela přesně (získáme vždy jen zaokrouhlené hodnoty, vždy mají desetinná místa). Dalším důvodem je konstrukce matematických modelů, vycházející právě z typu veličiny.

Všechny výše vyjmenované typy znaků popisují různé sledované jevy. Na tomto místě bychom si měli uvědomit, že typ jevu nemusí být vždy shodný s typem znaku, který použijeme k jeho popisu, například množství protilátek můžeme popisovat pomocí alternativního znaku (větší či menší než mezní hodnota). Pro detekci leukemie používáme procento blastických lymfocytů. Počet blastických lymfocytů je z pohledu hodnocení pacienta celočíselná kvantitativní veličina, ale z pohledu hodnocení lymfocytu se jedná o veličinu kvalitativní (alternativní), která popisuje, zda lymfocyt je či není blastický. Vždy je však nutno zvažovat přípustnost z pohledu interpretace.

Jednotlivé pozorované hodnoty jsou zkresleny chybami, které mohou být jednak systematické, způsobené nehomogenitou sledované skupiny objektů, jednak nesystematické – náhodné. K určité míře zkreslení vlivem „náhody“ dochází prakticky vždy. Proto říkáme, že sledujeme **náhodné jevy** nebo že používáme **náhodné veličiny**.

3.2. Typy náhodných veličin

Vlastně se vždy zabýváme veličinami, které jsou do jisté míry ovlivněny náhodou. Mluvíme o **náhodných veličinách** a o jejich náhodném – **stochastickém chování**.

Toto náhodné zkreslení se projeví tím, že získáme jinou hodnotu, než jaká je pravdivá. U kvantitativních veličin se jedná o její „zkreslení“, tj. o to, že při teoreticky identickém opakování získáme hodnotu, která je pouze více či méně podobná (např. výška postavy místo 181 cm je třeba jen 180 cm). U kvalitativních veličin se vliv náhody projeví chybným zařazením do jiné kategorie (např. o zdravé osobě si mylně myslíme, že má diabetes).

Je zřejmé, že z formálního hlediska jsou nejjednodušší alternativní znaky, které informují o přítomnosti nějaké vlastnosti, indikují nemoc či expozici. Pro jednoduchost se nejprve omezíme právě na tyto charakteristiky typu Ano/Ne.

3.2.1. Alternativní veličiny

Jako příklad uvažujme data pediatrické studie [12]. Studie se zabývá dětmi narozenými v roce 1987 ve skupině šesti vybraných okresů. Mimo jiné zde byla dysplazie kyčelního kloubu. Uvažujme tedy alternativní (binární) veličinu, tj. veličinu, která nabývá pouze dvou hodnot (ano a ne).

Výskyt vady	Četnost n_i	Relativní četnost $p_i = \frac{n_i}{n}$
Ano	981	0,089 4
Ne	9 994	0,910 6
$\sum$	10 975	1,000 0

Tabulka 3.1.: Displazie dolních končetin na konci prvního roku věku dítěte

Klademe si otázku: „Má dítě dysplazii kyčelního kloubu?“.

Úplnou informaci o výběru (o datech, která máme k dispozici) nám poskytnou dvě čísla: počet dětí s dysplazií kyčelního kloubu v našem výběru (četnost dysplazií) a celkový počet dětí ve výběru (rozsah výběru). Výběr pak nejlépe můžeme popsat pomocí **relativní četnosti**, která je podíl:

$$\frac{\text{počet dětí s dysplazií kyčelního kloubu}}{\text{počet všech dětí ve výběru}}$$

Ekvivalentním pojmem pro celou populaci je **pravděpodobnost**. Je to teoretický podíl nemocných dětí v celé populaci, nebo jinak ji můžeme chápat jako míru očekávání, že náhodně vybrané dítě bude sledovanou diagnózu mít.

3.2.2. Nominální veličiny

Podobně jako u alternativních veličin je možno mluvit o pojmu pravděpodobnosti i u nominálních veličin. Pro ilustraci můžeme uvažovat rodinný stav matek (viz tabulka 3.2). Řekněme, že

Rodinný stav	Četnost n_i	Relativní četnost $p_i = \frac{n_i}{n_\cdot}$
Svobodná	35	0,057
Vdaná	320	0,524
Rozvedená	223	0,365
Vdova	33	0,054
$\sum$	611	1,000

Tabulka 3.2.: Rodinný stav matek dětí sledovaných DKC

tato veličina může nabývat hodnot „svobodná“, „vdaná“, „rozvedená“ a „vdova“ s pravděpodobnostmi π_1, π_2, π_3 a π_4, kde například:

$$\pi_1 = \frac{\text{počet svobodných matek v základní populaci}}{\text{počet všech matek v základní populaci}}$$

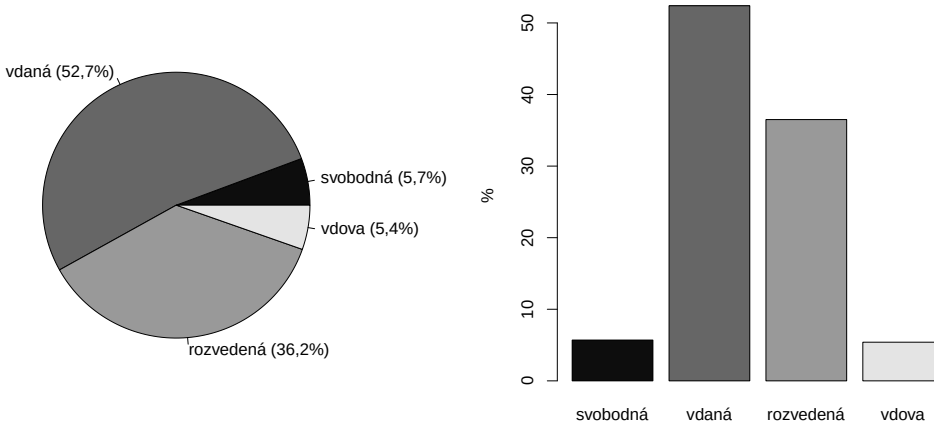
Jednotlivé kategorie označme čísly 1, 2, 3 a 4. Pak můžeme v sledovaném výběru mluvit o četnostech n_i, relativních četnostech $p_i = \frac{n_i}{n_1+n_2+n_3+n_4}$, i=1,...,4. Je zřejmé, že jejich součet je roven jedné.

$$\sum_{i=1}^{4} p_i = \sum_{i=1}^{4} \frac{n_i}{n_1 + n_2 + n_3 + n_4} = 1$$

Totéž platí i o pravděpodobnostech π_i, i=1,...,4 nebo o jejich odhadech $\hat{\pi}_i$, i=1,...,4.

Takovýto výčet pravděpodobností je úplnou informací o veličině a o jejím rozložení.

K popisu všech četností, případně relativních četností, se často používá **sloupcový graf** – tj. sloupcový graf se sloupci pro každou možnou hodnotu sledované veličiny. Výška těchto sloupců je dána jednotlivými četnostmi (případně relativními četnostmi – viz např. obrázek 3.1).



Obrázek 3.1.: Sloupcový a koláčový graf rodinného stavu matek dětí sledovaných DKC

Jindy se používá k zobrazení relativních četností takzvaný **koláčový graf**. Je to kružnice rozdělená na segmenty tak, aby velikost segmentů odpovídala podílu příslušné kategorie (viz např. obrázek 3.1). Mnozí autoři se snaží zlepšit estetický dojem použitím prostorového zobrazení těchto grafů. Je nutno si ale uvědomovat, že výpovědní hodnota pak klesne a že takovéto úpravy mohou pohled na graf zkreslovat. Pro rozumné použití koláčového grafu je užitečné, když počet segmentů není příliš velký (řekněme maximálně do deseti). Jiný interpretační problém může nastat, pokud zobrazovaný segment představuje velmi malý počet pozorování; to však není problém grafu, ale dat, lépe řečeno: jejich rozdělení do kategorií.

3.2.3. Ordinální veličiny

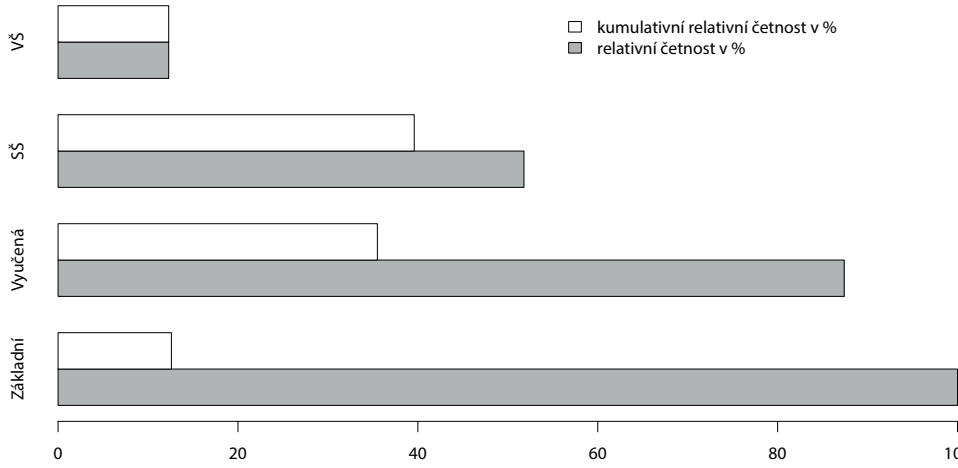
Podobně je možno uvažovat i o ordinálních veličinách. Jejich kódování je nutno provést tak, aby respektovalo přirozené uspořádání veličiny. Například pokud pracujeme s veličinou D = „dosažené vzdělání“, je přirozené uspořádat jednotlivé hodnoty takto: „základní“, „odborné“, „středoškolské“ a „vysokoškolské“. To, že sledovaná veličina je uspořadatelná, je užitečné využít. Poté,

co jednotlivé odpovědi vzestupně okódujeme (1,…,4), je možno zavést pro jednotlivé kódy i pojem **kumulativní pravděpodobnost**.[1] Kumulativní pravděpodobnost je pak pravděpodobnost toho, že libovolná osoba (ze základní populace) má nejvýše právě uvažované vzdělání.[2]

Samozřejmě nic nebrání tomu, abychom zvolili opačné pořadí kódů („vysokoškolské“, „středoškolské“, „odborné“ a „základní“).

Dosažené vzdělání	i	Četnost n_i	Kumulativní četnost $\sum_{j=i}^{4} n_j$	Relativní četnost $\frac{n_i}{n_.}$	Relativní kumulativní četnost $\sum_{j=i}^{4} \frac{n_j}{n_.}$
Vysokoškolské	4	1 360	1 360	0,124	0,124
Středoškolské	3	4 310	4 310	0,393	0,517
Vyučena	2	3 811	3 811	0,384	0,865
Základní	1	1 484	10 965	0,135	1,000

Tabulka 3.3.: Četnosti jednotlivých typů vzdělání matky



Obrázek 3.2.: Graf jednotlivých typů vzdělání matky

Pro ordinální veličinu je charakteristické, že jednotlivé sledované kategorie jsou uspořádány. To, zda uspořádání je vzestupné, či sestupné, již není tak zásadní a to se pak musí projevit v interpretaci. Relativní kumulativní četnost je výběrovou charakteristikou souboru, jejímž populačním protějškem jsou kumulativní pravděpodobnosti. Pro popis ordinálních veličin je tento kumulativní přístup vhodnější, protože reprezentuje jejich uspořádání.

Příkladem ordinální veličiny je i stadium onemocnění.

[1] V tomto případě představuje pro příslušnou kategorii (vzdělání), kolik matek má stejné nebo nižší vzdělání, než je tato kategorie.

[2] Absolvovala tedy příslušnou školu a všechny nižší stupně vzdělání.

Ze všech těchto úvah je zřejmé, že rozložení kvalitativní veličiny je možno popisovat pomocí pravděpodobností pro jednotlivé hodnoty, kterých může nabývat.

3.2.4. Kvantitativní veličiny

Tyto veličiny mohou obecně nabývat velkého množství různých hodnot. Charakterizovat rozložení výběru pomocí relativních četností pro jednotlivé pozorované hodnoty je většinou krajně nepřehledné, protože možných hodnot je mnoho. Pokud sledujeme spojitou veličinu, kterou jsme schopni pozorovat s dostatečně velkou přesností, je každá hodnota pozorována pouze jednou. Rozložení kvantitativních veličin se snažíme popsat pomocí matematického modelu tak, aby k dostatečně přesnému popisu stačilo pouze několik číselných parametrů.

Základním pojmem charakterizujícím populaci je **distribuční funkce**, obvykle označena jako $F(x)$. Je to kumulativní pravděpodobnost, že sledovaná veličina X nabývá hodnoty menší nebo rovné x

$$F(x) = P(X \leq x)$$

Například pro $x = 5000g$ je $F(x)$ je pravděpodobnost, že novorozenec má porodní hmotnost menší než $5000g$. Distribuční funkce je neklesající od nuly k jedničce. Její hodnoty nemohou být příliš „divoké“. S rostoucím x roste i $F(x)$ a „obvykle“ neobsahuje ani velké „skoky“. Byly vytvořeny různé matematické modely této funkce nazývané teoretické rozložení pravděpodobnosti. Rozložení (nebo jeho distribuční funkci) je pak možno popsat pomocí několika málo čísel – parametrů.

Distribuční funkce popisuje základní populaci. Jak jsme již řekli, většinou máme k dispozici pouze výběr. Na jeho základě můžeme sestrojit její výběrový protějšek – **empirickou distribuční funkci**. Nejprve označme n počet všech pozorování a n_x počet všech pozorování, jejichž hodnota je menší nebo rovna x (u ordinálních veličin jsme mluvili o kumulativních počtech). Empirická distribuční funkce $F_n(x)$ je schodovitá funkce, a je definována takto:

$$F_n(x) = \frac{n_x}{n}$$

Jak distribuční funkce, tak i její empirická varianta jsou charakteristiky, které podávají podrobnou informaci o rozložení kvantitativních veličin (celočíselných i spojitých).

Pro grafické zobrazení se často používá **sloupcový graf**, tedy graf, který pro každou pozorovanou hodnotu nekreslí sloupec odpovídající počtu nebo relativnímu počtu této pozorované hodnoty. Jiným, podobným typem grafu je **histogram**; ten vytvoříme tak, že číselnou osu rozdělíme na stejně dlouhé intervaly[3] a nad nimi opět zobrazíme sloupce odpovídající absolutní nebo relativní četnosti.

Rozdíl mezi histogramem a sloupcovým grafem je v tom, že histogram má sloupečky nad stejně dlouhými intervaly a sloupcový graf je vytvořen tak, že sloupce jsou konstruovány nad pozorovanými hodnotami.[4]

[3] Je důležité, abychom zvolili „rozumnou“ délku dělení. Tato délka se odvíjí od typu rozložení, ale hlavně od rozsahu použitého výběru.

[4] Tedy nepozorované hodnoty mohou z grafu vypadnout a vodorovná osa pak nepředstavuje číselnou míru, ale je pouze osou, na které jsou pozorované hodnoty.

Histogram tedy podává přehlednější informaci za cenu „zaokrouhlení“ hodnot – rozdělení do intervalů a sloupcový graf poskytuje přesný, ale méně přehledný obraz – vzdálenost sousedních naměřených hodnot bývá různá a jejich četnosti bývají malé.

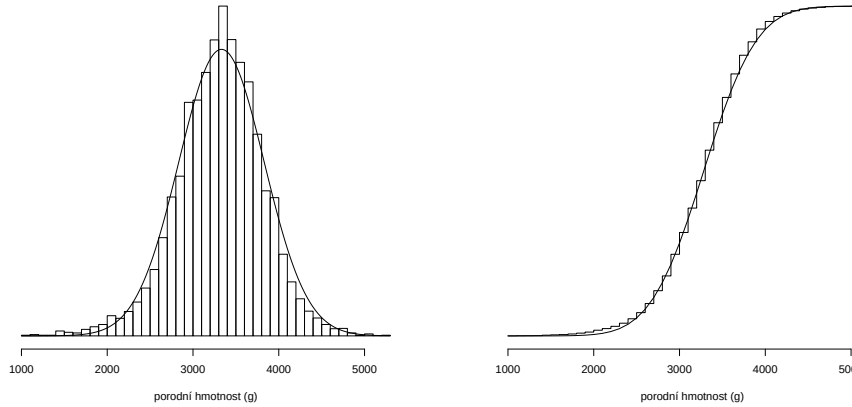
Kvantitativní veličiny – celočíselné

Proto celočíselné veličiny je možno použít jak histogram, tak i sloupcový graf. Nicméně histogram je zřejmě vhodnější a často pokud naměřené hodnoty jsou častěji shodné, je vhodné použít histogram, třeba i s délkou dělení 1, kdy do příslušného intervalu spadnou pouze identické hodnoty.

Kvantitativní veličiny – spojité

Pro spojité veličiny je sloupcový graf nepoužitelný (protože pro spojité veličiny by jednotlivé četnosti neměly být větší než jedna). Tvar histogramu silně závisí jak na zvolené délce, tak i na počátku dělení, a především na rozsahu souboru. Na rozdíl od sloupcového grafu zobrazuje i intervaly, v nichž nebylo nic pozorováno.

Histogram (v číselné nebo grafické podobě) popisuje rozložení sledované veličiny ve výběru. Slouží jako odhad rozložení veličiny v populaci, přesněji je odhadem **hustoty** ($f(x)$), popisující matematický model rozložení. Hustota je matematická funkce, která každé hodnotě přiřadí míru výskytu této hodnoty. Přesněji řečeno – plocha pod hustotou nad libovolným intervalem je rovna pravděpodobnosti výskytu hodnoty pozorované v tomto intervalu. Podíváme-li se na obrázek porodní hmotnosti 3.3 (vlevo), vidíme, že každý sloupec zobrazuje pozorovanou relativní četnost ve výběru a spojitá čára (odhad hustoty) představuje odhad pravděpodobnosti[5] výskytu hodnot v prostoru tohoto sloupce. Důsledkem toho je to, že celá plocha pod hustotou je vždy rovna jedné. Hlavním důvodem pro vytvoření matematického modelu, tedy i hustoty, je popis rozložení sledované veličiny pomocí několika málo parametrů (často se jedná o dva parametry, např. charakteristiku popisující „střed“ hodnot a míru jejich „rozházenosti“). S jemnějším dě-



Obrázek 3.3.: Histogram porodní hmotnosti chlapců narozených v obvodu Praha 4 v roce 1987 a odhad hustoty za předpokladu normálního (Gaussova) rozložení

lením a hlavně s rostoucím počtem pozorování se histogram stále více podobá hustotě, která představuje jeho „teoretickou obdobu“. To samozřejmě platí, jen pokud jsme zvolili správný matematický model.

[5]Přesněji řečeno: plocha pod hustotou nad tímto intervalem.

Nejčastěji bývá používáno takzvané **normální** (**Gaussovo**) rozložení. Bude zmíněno v následující kapitole 4. Zatím si jej můžeme představit přibližně jako model rozložení opakovaných měření nějaké veličiny (např. délky, hmotnosti, objemu), kde rozdíly mezi jednotlivými měřeními jsou dány chybou měření a biologickou variabilitou. A uvažujeme, že tyto odchylky jsou stejně pravděpodobné „nahoru“ i „dolů“.[6] Jako příklad uvažujme porodní hmotnost novorozenců (měřenou v gramech) z pediatrické studie [12]. Rozsah pozorovaných hodnot je rozdělen na 20 stejně dlouhých intervalů (včetně intervalů bez pozorovaných hodnot) a nad nimi je nakreslen sloupcový graf. Na y-ovou osu se vynáší relativní četnost, případně počet pozorování. Obrázek 3.3 (vlevo) ukazuje jednak histogram (tvořený jednotlivými sloupci), jednak odpovídající hustotu normálního (Gaussova) rozložení (hladká čára), jejíž parametry jsou odhadnuty z výběru tak, aby se tato hustota co nejvíce podobala histogramu a tvořila tak obraz hledaného matematického modelu. Graf je pořízen na základě vysokého počtu pozorování a tvar histogramu se silně podobá hustotě normálního (Gaussova) rozložení, proto můžeme předpokládat, že veličina porodní hmotnost novorozenců je normálně rozložena. Při použití normálního rozložení pro takovýto případ bychom si měli uvědomit, že se nejedná o modelování pouhého rozložení chyb měření, ale že rozložení naměřených hodnot je dáno především biologickou variabilitou sledované veličiny.

Podívejme se ještě na (empirickou) distribuční funkci. Obrázek 3.3 (vpravo) ukazuje závislost kvality jejího odhadu získaného na základě různě velkých souborů dat. Schodovitá čára je empirická distribuční funkce celého souboru, hladká, spojitá funkce vyjadřuje její odhad, realizovaný pomocí modelu normálního rozložení.

3.2.5. Celočíselné veličiny

Jedná se o veličiny vyjádřené pouze celými čísly, většinou jde o počty nějakých objektů (počty buněk, bakterií, výskytu sledované diagnózy, a pod.). Proto jsou tyto veličiny obvykle nezáporné (pokud neuvažujeme např. „změnu počtu …“). Rozložení celočíselné veličiny můžeme popsat soustavou pravděpodobností pro jednotlivé hodnoty (0, 1, 2, …). K číselné prezentaci výběrového rozložení se používají relativní nebo kumulativní relativní četnosti. Často se používá i grafické zobrazení, a to především histogram jednotlivých pozorovaných hodnot,[7] případně empirická distribuční funkce.

Stejně jako u spojitých veličin je velmi užitečné popsat studované rozložení matematickým modelem, který je dán jen několika málo parametry. Tyto parametry bude možno později snadno porovnávat (například pro různé skupiny). Nejčastěji používaným příkladem rozložení celočíselné veličiny jsou **Binomické** nebo **Poissonovo rozložení**,[8] podrobněji je popíšeme v kapitole 4. Krátce řečeno, Poissonovo rozložení je matematický model, kdy se zajímáme o velkou populaci (často neznáme její rozsah) a sledovaný jev může teoreticky nastat „mnohokrát“ a pokaždé se stejnou pravděpodobností (často velmi malou) a my sledujeme počty výskytu tohoto jevu. Binomické rozložení modeluje situaci, kdy známe rozsah populace a my, pomocí podílu nemocných v populaci, odhadujeme pravděpodobnost sledovaného jevu[9].

Příkladem celočíselné veličiny s Poissonovým rozdělením může být počet infektů horních cest dýchacích u jednoho dítěte. Pokud by tato veličina měla Poissonovo rozložení, pak předpoklá-

[6] Samozřejmě existují i jiné modely, jako například logaritmicko-normální rozložení pro koncentrace.

[7] Na x-ové ose by měly být i hodnoty s nulovým počtem pozorování, jinak bude konec grafu zkreslující. Na začátku budou například hodnoty pro 0, 1, 2 a na konci mohou být „díry“, například 25 a pak až 30. Tuto skutečnost by měl histogram respektovat. Sloupcový graf tyto „díry“ vypustí.

[8] Obě jsou popsány pouze jedním parametrem.

[9] Například uvažujeme konkrétní okres a pravděpodobnost sledované diagnózy.

dáme, že výskyt jednotlivých onemocnění je nezávislý.[10] Jiným příkladem může být počet kolonií na kultivační půdě.

Příkladem binomické veličiny může být situace, kdy sledujeme výskyt chřipkových onemocnění v daném týdnu a okrese, v němž známe počet obyvatel, a odhadujeme podíl či procento nemocných.

[10] Pravděpodobnost opakování onemocnění je tedy stejná jako pravděpodobnost prvního onemocnění a pravděpodobnost onemocnění je u všech dětí stejná.

Počty onemocnění	Četnost n_i	Kumulativní četnost $\sum_{j=1}^{i} n_j$	Relativní četnost v % $100 \cdot \frac{n_i}{\sum_{k=1}^{\infty} n_k}$	Relativní kumulativní četnost v % $100 \cdot \sum_{j=1}^{i} \frac{n_j}{\sum_{k=1}^{\infty} n_k}$
0	576	576	6,4	6,4
1	867	1 443	9,6	16,1
2	1 336	2 779	14,8	31,0
3	1 304	4 083	14,4	45,5
4	1 153	5 236	12,8	58,3
5	961	6 197	10,6	69,0
6	784	6 983	8,7	77,8
7	418	7 401	4,6	82,4
8	451	7 852	5,0	87,5
9	244	8 096	2,7	90,2
10	304	8 400	3,4	93,6
11	120	8 520	1,3	94,9
12	151	8 671	1,7	96,6
13	69	8 740	0,8	97,4
14	72	8 832	0,8	98,2
15	61	8 893	0,7	98,8
16	22	8 915	0,2	99,1
17	14	8 929	0,2	99,2
18	9	8 938	0,1	99,3
19	10	8 948	0,1	99,5
20	17	8 965	0,2	99,6
21	7	8 972	0,1	99,7
22	6	8 978	0,1	99,8
23	2	8 980	0,0	99,8
24	5	8 985	0,1	99,9
25	3	8 988	0,0	99,9
26	4	8 992	0,0	99,9
28	2	8 994	0,0	100,0
31	1	8 995	0,0	100,0
34	1	8 996	0,0	100,0
66	1	8 997	0,0	100,0

Tabulka 3.4.: Počty infektů horních cest dýchacích u dětí během prvních tří let věku

4. Základní statistické charakteristiky

V této kapitole začneme pracovat s programem R. Základní popis, jak s tímto programem pracovat, najdete v dodatku A, a použití pro konkrétní statistické výpočty v následujících kapitolách.

Dříve, než se budeme zabývat charakteristikami popisujícími námi studované subjekty, zastavme se u základního popisu souboru našich dat.

Velmi důležitý je zřejmě **rozsah našeho výběru**, jinými slovy: velikost datové matice.

Řekněme, že máme data přečtena v matici `X`. Příkazem `dim(X)` získáme její rozměr (tj. počet řádků a počet sloupců), nebo pokud máme vektor, například sloupec matice `X[,"POR_HMOT"]`), můžeme délku tohoto vektoru vypočíst příkazem `length(X[,"POR_HMOT"])`. Na tomto místě je užitečné připomenout, že program rozlišuje velká a malá písmena. Veličina `X` je tedy jiná než `x`. Problém ale nastane, pokud jsou ve sledovaném vektoru neudané hodnoty (značíme `NA`(not availble)), protože příkaz length je započte do celkové délky. Mějme například vektor `y` a vypočtěme jeho délku:

```
> y<-c(1,2,3,4,5,NA)
> y
[1] 1  2  3  4  5  NA
> length(y)
[1] 6
```

Pokud chceme, aby nezjištěná hodnota `NA` nebyla započtena do délky vektoru, můžeme použít například funkci `na.omit()`

```
> length(na.omit(y))
[1] 5
```

Naopak pokud potřebujeme zjistit, kolikrát se vyskytlo `NA`, stačí napsat příkaz:

```
> length(y)-length(na.omit(y))
[1] 1
>
```

Dalšími charakteristikami, které jsou často používány pro popis souboru dat, je **maximum** a **minimum** `max()` a `min()` . Při použití těchto funkcí nastává opět problém s hodnotami `NA`. Ten je možno řešit opět pomocí funkce `na.omit()`, tedy například příkazem `max(na.omit(x))` nebo parametrem funkce `max(x,na.rm=TRUE)`.

Z hlediska interpretace si ale musíme uvědomit, že maximum i minimum se s rostoucím rozsahem vzdalují. Není tedy možno porovnávat tyto extrémy v různě velkých studiích. Tím i ztrácí

smysl porovnávat rozsah hodnot naměřených v různě velkých studiích.

Úplnou informaci o rozložení kvantitativní veličiny v našem výběru podává buď **histogram**, nebo **empirická distribuční funkce**. Pracovat přímo s nimi je obtížné a špatně interpretovatelné, protože obě jsou popsány velkým množstvím čísel. Rádi bychom pracovali jen s několika málo jednoduchými charakteristikami. Empirickou distribuční funkci tedy nahradíme za pomoci zvoleného matematického modelu teoretickou distribuční funkcí, kterou je možno popsat pouze malým počtem číselných charakteristik. Ty pak můžeme odhadovat a porovnávat. Kvalita těchto odhadů samozřejmě závisí na tom, jak dobře model popisuje studovanou veličinu. Pokud nalezneme odhady parametrů uvažované teoretické distribuční funkce, získáme i odhad celého rozložení (pokud model platí).

Často není cílem našeho zájmu rozložení sledované veličiny, ale pouze určitá charakteristika. Tu můžeme samozřejmě také odhadovat, avšak při interpretaci výsledků si musíme uvědomit, že odhady popisují uvažované charakteristiky, ale zdaleka nemusí být dobrými charakteristikami z pohledu rozložení sledované veličiny (například může být problém s interpretací, pokud data obsahují velmi odlehlé hodnoty).

Hledáním modelu rozložení se budeme zabývat dále. Nyní se vraťme k statistickým charakteristikám. Ty jsou definovány jednak jako

Výběrové charakteristiky – ty je možno vypočíst z výběru, který máme k dispozici. Jejich hodnota je pro nás tedy jednoznačně známa. Výběrové charakteristiky nás ale zajímají pouze jako obraz populace, který závisí na volbě správného modelu a na rozsahu a reprezentativnosti použitého výběru.

Populační charakteristiky – to jsou ty, jejichž hodnotu neznáme, vlastně ani nemůžeme znát, ale které jsou cílem našeho zájmu a pokoušíme se je odhadnout prostřednictvím výběru.

Pokud chceme použít výběrové charakteristiky jako odhad populačních charakteristik, musíme si uvědomit, že kvalita této indukce (zobecnění) závisí na konstrukci výběru, vlastnostech a adekvátnosti zvoleného matematického modelu.

Statistické charakteristiky se samozřejmě liší podle typu veličiny, kterou mají popisovat. Základní rozdíl je mezi kvalitativními a kvantitativními veličinami. Nejprve se krátce soustředíme na kvalitativní veličiny. Pro ně hraje klíčovou roli charakteristika nazývaná pravděpodobnost.

4.1. Míry pro kvalitativní veličiny

Uvažujme alternativní (binární) veličinu, tj. veličinu která nabývá pouze dvou hodnot (např. ano a ne). Můžeme si třeba položit otázku: „Trpí sledovaná osoba aterosklerózou?“ Označme tuto veličinu symbolem A.

Na první pohled je zřejmé, že nelze pro populaci říci, že veličina A nabývá hodnoty „Ano“ či „Ne“. Ale můžeme říci, jak velká část populace trpí aterosklerózou. Zastavme se na chvíli u pojmu **pravděpodobnost** a **relativní četnost**. Rozložení veličiny A v základní populaci (které nás zajímá) je popsáno dvěma číselnými hodnotami – počtem osob s aterosklerózou a počtem všech osob v populaci. Získat tato čísla však není obvykle možné. Jediné, co můžeme popsat, je rozložení výběru, který máme k dispozici. Toto známé rozložení výběru nám slouží jako odhad neznámého rozložení celé populace.

4.1.1. Pravděpodobnost

K tomu, aby bylo možno s náhodou nějak pracovat, je nutno pokusit se o její kvantifikaci. Jako teoretická míra očekávání zvolené odpovědi (např. „Ano“) byl zaveden pojem **pravděpodobnost**. Sledovaná veličina nabývá hodnot „Ano“ nebo „Ne“, někdy kódovaných jako 1 a 0. Pro označení pravděpodobnosti se obvykle používá písmeno P. Je definována jako hypotetický podíl počtu pozitivních odpovědí k celkovému počtu odpovědí. Pravděpodobnost, že například veličina A (sledovaná osoba trpí aterosklerózou) nabývá hodnoty „Ano“, tedy značíme:

$$P(A = Ano)$$

Teoreticky mohou nastat dva extrémní případy $P(A = Ano) = 1$ a $P(A = Ano) = 0$. Má-li výrok pravděpodobnost rovnu 1, mluvíme o **jevu jistém**. Naopak je-li pravděpodobnost rovna nule, mluvíme o **jevu nemožném**. V reálné situaci nás zajímají veličiny s pravděpodobností mezi 0 a 1.

Samozřejmě, že pro alternativní veličinu platí $P(A = Ne) = 1 - P(A = Ano) = 0$.

Pro označení teoretické hodnoty pravděpodobnosti se často používá písmeno řecké písmeno π. Pravděpodobnost je tedy veličina, která nabývá hodnoty z uzavřeného intervalu $< 0; 1 >$.

Někdy mluvíme i o takzvané **podmíněné pravděpodobnosti**, tj. o pravděpodobnosti sledovaného jevu za podmínky, že jiná doprovodná veličina nabývá konkrétní požadované hodnoty. Pro ilustraci označme další veličinu D – „sledovaná osoba trpí diabetem“. Nyní můžeme uvažovat pravděpodobnost, že diabetik trpí aterosklerózou značíme – $P(A = Ano|D = Ano)$. Výraz před svislou čarou „$A = Ano$“ představuje jev, který nás zajímá, a výraz za svislou čarou „$D = Ano$“ nazýváme podmínkou. Podmíněná pravděpodobnost je tedy pravděpodobnost uvažovaná pouze pro objekty splňující příslušnou podmínku. Naopak o **nepodmíněné pravděpodobnosti** mluvíme, pokud nás zajímá výskyt sledovaného jevu v celé populaci (bez jakéhokoliv omezení).

Často se stává, že sledujeme současně různé jevy a ptáme se na jejich vztah. Když se budeme ptát na vztah těchto veličin, je třeba zavést pojem nezávislosti. Řekneme, že **dva jevy** A a D **jsou nezávislé**, pokud pravděpodobnost společného výskytu aterosklerózy a diabetu $P(A =$„Ano“ a současně $D =$„Ano“) je rovna součinu dílčích pravděpodobností:

$$P((A = Ano)\&(D = Ano)) = P(A = Ano) \cdot P(D = Ano)$$

Výskyt aterosklerózy u sledované osoby nezávisí na tom, zda má diabetes. Nezávislost můžeme vyjádřit i pomocí podmíněných pravděpodobností:

$$P(A = Ano|D = Ano) = P(A = Ano|D = Ne)$$

Jinými slovy: pravděpodobnost výskytu aterosklerózy u diabetiků je stejná jako pravděpodobnost aterosklerózy u osob, které netrpí diabetem. Důsledkem toho je, že tyto podmíněné pravděpodobnosti jsou rovny nepodmíněné pravděpodobnosti toho, že jakákoliv osoba má diabetes.

$$P(D = Ano|D = Ano) = P(A = Ano|D = Ne) = P(A = Ano)$$

Podobně je možno pomocí podmíněné pravděpodobnosti vyjádřit i společnou pravděpodobnost výskytu A i D:

$$P((A = Ano)\&(D = Ano)) = P(A = Ano|D = Ano) \cdot P(D = Ano)$$