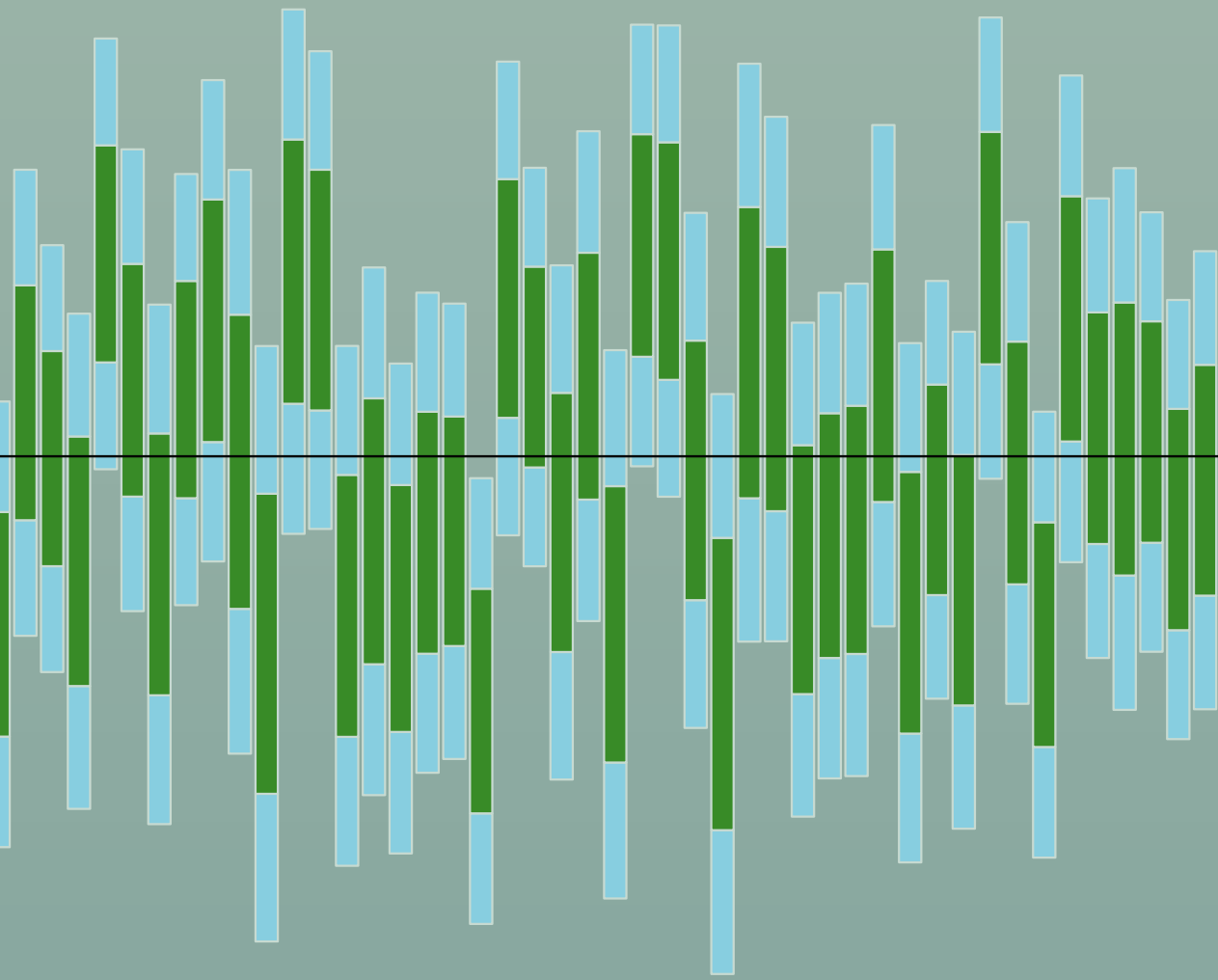
STRUČNÁ BIOSTATISTIKA PRO LÉKAŘE

BOHUMÍR PROCHÁZKA

KAROLINUM



Stručná biostatistika pro lékaře

RNDr. Bohumír Procházka, CSc.

———

Recenzovali:
MUDr. Jiří Keller, Ph.D.
MUDr. Zdeněk Šmerhovský, Ph.D.

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

ISBN 978-80-246-2783-0
ISBN 978-80-246-2804-2 (online : pdf)

Univerzita Karlova v Praze
Nakladatelství Karolinum 2015

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 cílemje klást důraz na nejnovější znalosti a především na objektivnost hodnocení nejnovějších poznatků. Podrobnější popis medicíny založené na důkazech najdete v [56]. 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. V tomto stručném souhrnu se pokusím seznámit se základními principy. Podrobněji jsou popsány např. v [42].

Vraťme se k vlastní statistice. Původní metodou jak získat informace 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 iluzí a navíc ani nemá praktické použití (je zbytečné měřit hmotnost postavy s přesností na miligramy nebo velikost populace státu s přesností na jedince).

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. To ale není pravda, je nutno vědět, jaký nástroj a kdy je vhodné použít.

- 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.
- 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.

Tato kniha uvádí velmi stručně do problematiky statistického uvažování a informuje o základních možnostech a metodách matematické statistiky. Vznikla jako stručný přehled statistických metod používaných v medicíně. Podrobnější výklad s poukázáním na problémy interpretace výsledků a s příklady je v knize [42], případně v předchozím vydání [41] a v další literatuře.

Provádět statistické výpočty je možno pomocí mnoha programů, ale i pomocí programu R, který je popsán v zmíněné knize [42] a je OpenSource, tedy zdarma. Tato kniha ukazuje, že práce s tímto programem je jednoduchá i pro nestatistiky. Problém je ale u všech (tedy i komerčních) programů s volbou správné metody a s tím související interpretací výsledků.

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 LaTeXa 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

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.

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, ale dovoluje i postupně odvozovat stále složitější vztahy či zavádět složitější pojmy.

Vědní disciplíny zabývající se skutečnou realitou sledují složité objekty a snaží se popsat jejich společné vlastnosti. Tím vlastně provádíme jisté zjednodušení, které umožňuje použít matematické nástroje. Obvykle sledujeme jen určitou (obvykle malou) část populace. Zajímá nás ale celá populace, jinými slovy nás zajímají i další objekty, které jsme nestudovali. Získané výsledky se pak snažíme zobecnit. Použijeme k tomu takzvané **induktivní uvažování**. Přírodní a lékařské vědy jsou charakteristické velkou složitostí sledovaných objektů, ty ve skutečnosti nikdy není možno popsat do detailu. Vždy je nutno je na určité úrovni zahrnout do neurčitosti individuální rozdíly – přisoudit „náhodě“. Ta může pokrývat 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 modelu 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 vbrat 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. 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 ve vzorku poměrně přesně – chyba zkreslení vlivem „náhodou“ je poměrně malá, ale různé vzorky, byť jedné osoby se mohou výrazně lišit. Na druhé straně, například v oblasti psychologie, jsou sledované charakteristiky (odpovědi na otázky) zatíženy velkou chybou.

2.1 Přístupy k řešení problémů

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

Individuálně – zajímají nás konkrétní případy jako neopakovatelné jevy. Jedná se tedy o pouhý popis konkrétního případu – o kazuistiku.

Skupinově – zajímají nás obecné vlastnosti. Hledáme obecné vlastnosti. Zde je prostor pro použití statistiky.

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

Statistika tedy může být:

Deskriptivní – **popisná statistika**, se nepokouší vyslovovat k vlastnostem jedinců, kteří nebyli sledováni.

Induktivní – je moderní přístup matematické statistiky poskytující nástroje pro zobecňování výsledků na širší populaci.

Věnujme nyní pozornost jedné ze základních myšlenek, která se používá v rámci statistického uvažování.

2.2 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.

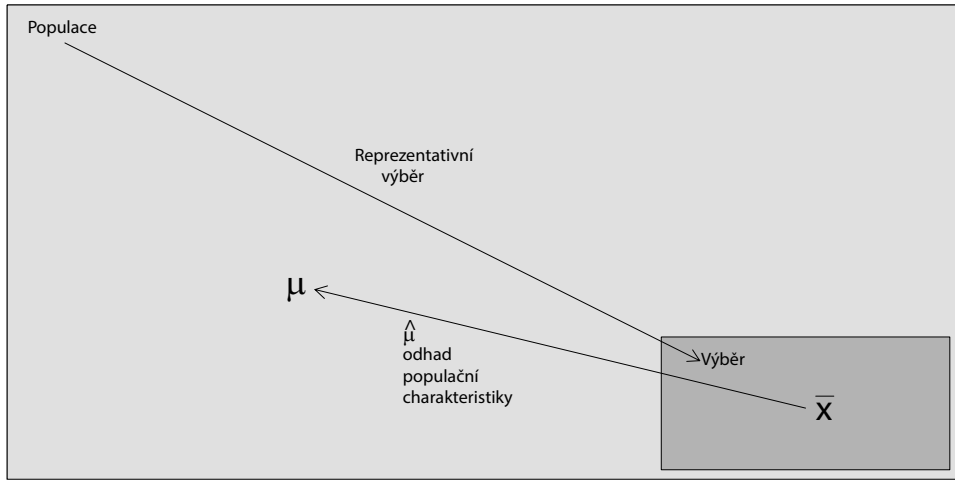
Pokud používáme deskriptivní statistiku, týkají se naše tvrzení pouze souboru, na kterém byla prováděna měření (pozorování a podobně). V tomto případě je výběr totožný se základní populací. Získané výsledky popisují pouze zkoumaný soubor a nesnaží se o žádné zobecnění na větší nebo jinou skupinu objektů. Stačí tedy mluvit o získaných charakteristikách a ty popisují sledovaný soubor zcela přesně.

Induktivní statistika se snaží výsledky získané na výběru zobecnit (generalizovat na širší skupinu objektů) na základní populaci. 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.

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" 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.

Výběr a výběrové charakteristiky. Výběr je skupina objektů, na kterých provádíme šetření.

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.

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 matematický model. „rozložení“. Definujeme tedy:

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

Teoretické rozložení je matematický model (například normální – Gaussovo, Poissonovo, binomické a další), který udává pro všechny možné hodnoty sledované veličiny pravděpodobnost, 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 tedy je předpoklad, že populace, o které chceme mluvit, má stejné vlastnosti jako použitý výběr.

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, skutečnost, 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 ještě dále dělit:

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 méně než středoškolské a to je méně 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). Někdy se používá i kódování 0/1 nebo +/–.

Kvantitativní znaky, jak již je řečeno v jejich názvu, 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 vždy schopni 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.

Rozdělení 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.

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 [9]. Výběr obsahuje děti narozené v roce 1987 ve skupině šesti vybraných okresů. Mimo jiné zde byla sledována dysplazie kyčelního kloubu. Uvažujme tedy alternativní (binární) veličinu.

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

Úplnou informaci o výběru (tj. 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:

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

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

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

Jindy se používá k zobrazení relativních četností takzvaný **koláčový graf**. Je to kružnice rozdělená na segmenty tak, aby velikost segmentů odpovídala podílu příslušné kategorie.

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 „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“).

Pro ordinální veličinu je charakteristické, že jednotlivé sledované kategorie jsou uspořádány. To, zda uspořádání je vzestupné, či sestupné, již není tak zásadní, to se pak musí projevit v interpretaci.

Příkladem ordinální veličiny je i stádium onemocnění.

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

[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.

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 tedy 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 označme jako$F(x)$. Je to kumulativní pravděpodobnost, že sledovaná veličina X nabývá hodnoty menší nebo rovné x.

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

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. Empirická distribuční funkce $F_n(x)$ je schodovitá, neklesající 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 nakreslí 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]

Kvantitativní veličiny – celočíselné

Pro celočíselné veličiny je možno použít jak histogram, tak i sloupcový graf.

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

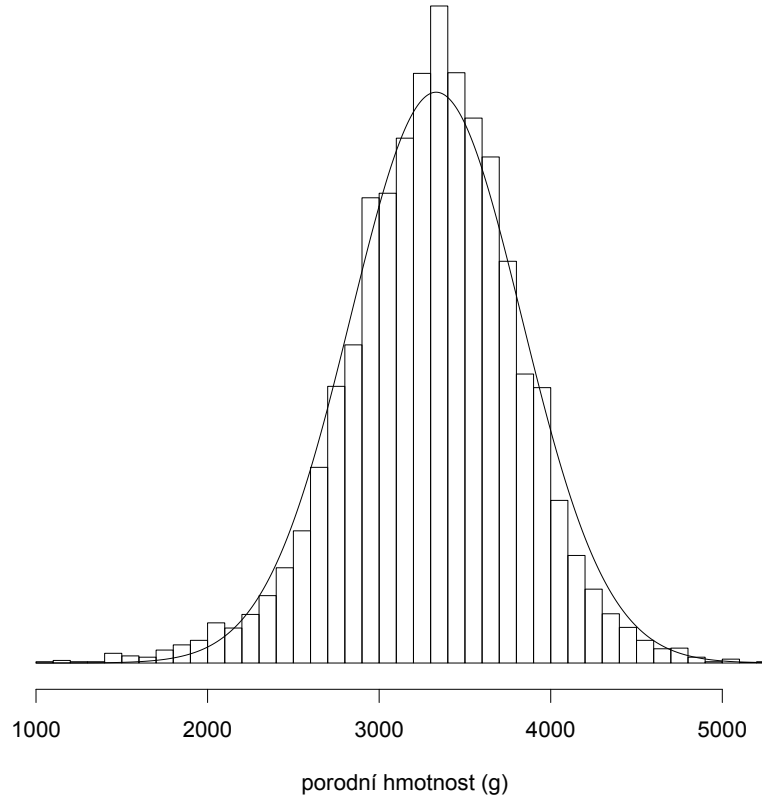
[3] Je důležité, abychom zvolili „rozumnou“ délku dělení. Tato délka se odvíjí od typu rozložení, ale hlavně od rozsahu použitého výběru.

[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 četnos bývají malé.

i na počátku dělení, a především na rozsahu souboru. Na rozdíl od sloupcového grafu zobrazuje i intervaly, kde 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. Ukázkou histogramu (sloupečky) a hustoty (spojitá čára) pro porodní hmotnost je na obrázku 3.1.



Obrázek 3.1: Histogram porodní hmotnosti chlapců narozených v obvodu Praha 4 v roce 1987 (sloupečky) a odhad hustoty (spojitá čára) za předpokladu normálního (Gaussova) rozložení

S jemnějším dělením 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.

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“. Jako příklad uvažujme porodní hmotnost novorozenců (měřenou v gramech). Rozsah pozorovaných hodnot je rozdělen na 20 stejně dlouhých intervalů (včetně intervalů bez pozorovaných hodnot) a nad nimi je nakreslen sloupcový graf viz 3.1.

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 možných hodnot, případně empirická distribuční funkce.

Stejně jako u spojitých veličin je velmi užitečné popsat studované rozložení matematickým modelem, který je dán jen několika málo parametry. **Poissonovo rozložení**.[5], které si podrobněji popíšeme v následující kapitole 4 je matematický model, kdy se zajímáme o velkou populaci (často neznáme její rozsah) a pravděpodobnost sledovaného jevu je velmi malá a my sledujeme počty výskytu tohoto jevu. Naopak **binomické rozložení** modeluje situaci, kdy známe velikost populace a všichni mají stejnou pravděpodobnost například onemocnění a my odhadujeme tuto pravděpodobnost pomocí podílu nemocných v této populaci.

Příkladem celočíselné veličiny s Poissonovým rozdělením může být počet infektů horních cest dýchacích u jednoho dítěte za nějaké časové období. 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.

[5] Je popsáno pouze jedním parametrem.