

Moderní analýza medicínských dat

Silvie Bělášková, Lenka Blažková

Statistická analýza dat umožňuje odpovídat na otázky týkající se pochopení světa kolem nás. Statistické metody jsou součástí každého výzkumu. Objevují se všude tam, kde je potřeba něco zhodnotit, posoudit nebo vyvodit závěry s požadovanou přesností, a to nejenom v lékařství, ale napříč mnoha různými obory. K dispozici je proto i široká nabídka softwaru, který nabízí jednoduchý způsob, jak data zpracovat. Je nutné využívat software korektně a zároveň nepřeceňovat jeho možnosti.

Jak slouží statistika lékařům?

Těžko určíme, kdy se statistika poprvé výrazně podílela na výzkumu v oblasti medicíny nebo kdy poprvé prokázala medicíně službu. Jedním z prvních úspěchů statistiky je mapování případů cholery při epidemii v Londýně v roce 1853. Díky statistickým metodám došlo ke správné alokaci zdroje nákazy. Mezi další nevyvratitelné úspěchy patří první randomizovaný klinický pokus navržený A. B. Hillem, který prokázal účinnost streptomycinu při léčbě tuberkulózy. V současnosti statistické metody využívají lékaři jak v oblasti výzkumu, kde většina studií používá nějaký typ statistické analýzy, tak při běžném zpracování zdravotních údajů na úrovni jedné nemocnice nebo regionu, kde postačí softwarové nástroje na dotazování do databází a jednoduché údaje popisné statistiky. Kritické posuzování informací o efektivitě terapie nebo kvalitě diagnostického pokusu vyžaduje nejen kvalitní software, ale zejména znalost metodologie a statistiky. Kvalitnímu výzkumu předchází pečlivá příprava, předpokladem úspěchu je i jeho správné načasování a realizace. Čím větší ambice má plánovaná analýza, tím potřebnější je poradit se s odborníkem – nejlépe před započítím samotného šetření. Lékaři jsou často proti spolupráci s externistou, výsledky statistiků totiž nemusí vždy splňovat původní představu lékaře a ten považuje zásah statistika za znehodnocení svého výzkumu. Kromě statisticky přesných a korektních výsledků je důležitá jejich správná interpretace.

Fáze statistiky

Statistické šetření probíhá ve třech fázích: sběr dat, analýza dat a statistické usuzování (obr. 1).

Experimentální design	Určení velikosti vzorku nezbytného pro statisticky relevantní výsledek.
Vzorkování	Vzorkovací plán zabezpečující náhodnost a reprezentativnost vzorku.
Uložení a správa dat	Uložení dat ve vhodném formátu a jejich očištění.
Vizualizace	Grafický přehled o rozdělení dat, vazbách proměnných apod.
Popisná analýza	Porovnání popisných statistik s údaji z existující literatury.
Testování hypotéz	Testování vazeb s cílem vzájemně vysvětlit variabilitu proměnných.
Modelování	Využití získaných znalostí k vytvoření prediktivních modelů.

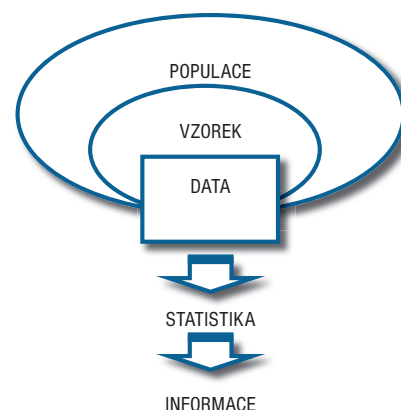
Obr. 1

Sběr dat

Statistika nabízí možnost nahlížení na realitu prostřednictvím vzorku. Výsledky analýzy jsou jen tak dobré, jak dobrý je samotný vzorek (obr. 2), proto je tato fáze považována za nejdůležitější. Reprezentativnost a náhodnost vzorku spolu s jeho velikostí patří mezi nejdůležitější faktory ovlivňující věrohodnost našeho závěru. Statistika není schopna činit závěry o jevech neobsažených ve vzorku. Nedostatečnost vzorku je nejčastější

chybou při statistickém šetření. Před vlastním sběrem dat je nutné zformulovat otázky, na které hledáme odpověď, a určit cílovou skupinu výzkumu.

Softwarové nástroje pomáhají vytvářet výběrové schéma, pomocí generátorů pseudonáhodných čísel umí vytvořit náhodný vzorek z celkového datového souboru, mohou být použity k rozdělení pacientů do kontrolní a testované skupiny, anebo umožní z malého vzorku nasimulovat data s potřebným počtem pozorování, která zachovají původní vazby. Využijeme je také pro snadnou a rychlou identifikaci chybných měření a záznamů či odhalení odlehklých hodnot.

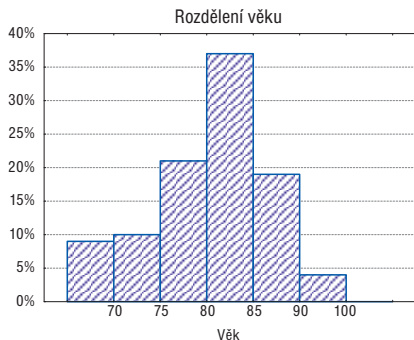


Obr. 2

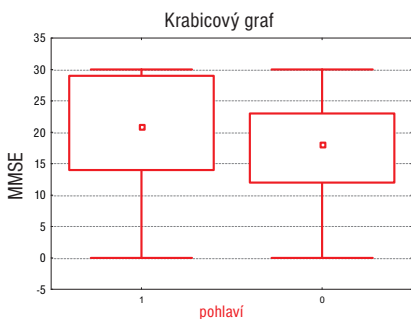
Analýza dat

Každý statistický software poskytuje rychlý přístup k hodnotám nejčastějších deskriptivních statistik, jako jsou průměr a medián, modus (nejčtenější hodnota) či směrodatná odchylka. Účelem je nalézt číselné charakteristiky, které shrnují vlastnosti dat. Průměr je ideální číselná charakteristika, ale má smysl pouze u dat kvantitativních. Jeho nevýhodou je citlivost na extrémní hodnoty. Průměr představuje jakési těžiště dat, ale jeho použití je třeba předem zvážit. Pro lepší orientaci v datech slouží grafické znázornění. Statistické softwary bývají vybaveny širokou škálou různých grafů, je třeba pouze správně vybrat typ grafu. Nejčastěji používaný je histogram (obr. 3), který je

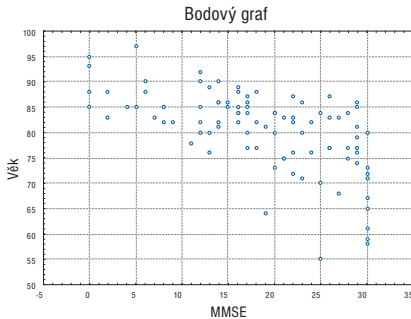
vhodný především pro data intervalového typu. Lze sledovat empirické rozdělení dat a další charakteristiky. Pro analýzu závislosti dvou znaků lze použít krabicový nebo bodový graf (obr. 4 a 5).



Obr. 3: Histogram



Obr. 4: Krabicový graf



Obr. 5: Bodový graf

Obrázek 4 je příkladem porovnání mediánů MMSE (mentální charakteristiky pozorované veličiny) v závislosti na pohlaví. Liší se mediány pohlaví, proto je třeba otestovat, zda se jedná o statisticky významnou odlišnost. Na obrázku 5 pozorujeme v datech klesající tendenci a určitou lineární závislost, kterou by bylo vhodné rovněž statisticky otestovat a vyjádřit přesný vztah pomocí regresní analýzy. Pro vícerozměrná data poskytuje software moderní explorační techniky, hodnoty různých statistik i grafy závislosti. Pokročilé dataminingové metody umožňují dokonce automatický výběr pouze několika nezávisle proměnných, které nejlépe popisují variabilitu zkoumané závislé veličiny. Jakkoli pokročilý používáme software, rozhodování o zahrnutí či nezahrnutí



proměnných do modelu i výběr statistických metod musíme zvážit sami. Statistický software rozhodování usnadňuje, a pokud váháme, pro který z několika možných modelů se rozhodnout, pomůže nám jednotlivé modely vyhodnotit a porovnat.

Statistické usuzování

Cílem statistického usuzování je vyvodit na základě vzorku dat závěry, které jsou platné pro celou populaci, ze které vzorek pochází. Krom toho je výsledkem statistického usuzování i zhodnocení, jak jsou dané závěry spolehlivé. Základním principem statistiky je pravděpodobnost výskytu nějaké události. Prostřednictvím vzorku se snažíme odhadnout skutečnou pravděpodobnost události. Klíčovou roli pro správnost vzorku hraje náhoda. Předpokladem analýz je, že vzorek pochází z tzv. náhodného výběru. Při opětovně nezávislé realizaci téhož náhodného experimentu se podíl jevu mezi všemi doposud provedenými realizacemi zpravidla ustuluje kolem konstanty. V lékařském prostředí a výzkumu vzniká obrovské množství dat se spoustou nejistot, vzájemných vazeb a variabilitou, korektní analýza je tudíž netriviální. Dnes je každý uživatel statistického softwaru schopen získat výsledky libovolně náročných analýz bez nahlížení do statistických tabulek a sestavování složitých vzorců. Důležité je ovšem, aby uživatel rozuměl principům metod, znal a ověřoval jejich předpoklady. Úskalím nejen v lékařství je, že software uživatele neupozorní, když provádí analýzu, která nemá pro konkrétní vzorek dat smysl.

Statistické testy

Na počátku klinického výzkumu stojí hypotéza – tvrzení týkající se určité cílové populace. Výzkum má v experimentu tuto hypotézu potvrdit, nebo vyvrátit. Používají

se k tomu statistické testy – rozhodovací pravidla, která na základě vzorku dat zamítnou, nebo nezamítnou testovou hypotézu, kterou vycházejí z primárního cíle studie přesně definuje statistik. Software nabízí pro běžně testované hypotézy, jako je shoda průměrných hodnot zvoleného ukazatele v různých podskupinách výběru, více možných postupů. Etické zásady zakazují přístup, kdy se na data aplikují všechny dostupné testy a následně se zvolí ten, který nejvíce odpovídá původní představě lékaře. Metodika je navrhována na počátku experimentu a je třeba se jí důsledně držet. Všechny nesprávné výsledky a interpretace totiž ohrožují pověst vědy a nesou v sobě potenciální nebezpečí pro přírodu a společnost při jejich aplikaci.

Další využití IT v lékařství

Závěrem zmíníme, že význam pro medicínu nemá pouze využití možností softwaru pro statistickou analýzu záznamů o pacientech či dat z experimentů. Lékaři při náročných operacích využívají množství rozličných mechanických implantátů, které musí splňovat ta nejpřísnější kritéria, neboť jejich správné fungování a přesné parametry mohou být životně důležité. Například při výrobě kardiostimulátorů je proto využíván pouze takový software, jenž splňuje vysoké nároky na správnost výpočtů. ■



Autorky působí ve společnosti StatSoft CR. Silvie Bělášková jako externí lektorka, Lenka Blažková jako konzultantka a analytička. Grafy použité v textu byly vytvořeny pomocí softwaru Statistica 9.1.