Informační Bulletin

České Statistické Společnosti

č. 2.říjen 2001, ročník 12

Statistické dny 21.6. a 22.6. v Hradci Králové

Česká statistická společnost, Jednota českých matematiků a fyziků, pobočka v Hradci Králové a katedra matematiky PdF University v Hradci Králové uspořádaly ve dnech 21.6. a 22.6. Statistické dny. Ty byly zaměřeny k aplikacím matematicko-statistických metod v humanitních vědách. Sešly se příspěvky, které účastníky zaujaly hlavně tématem zpracovávané problematiky a volbou statistické metody. Protože o sborník referátů měla zájem hlavně pracoviště, která se teprve snaží statistických metod používat, doporučili jsme autorům příspěvků uvést stručně i základní statistické poučení. Proto doufáme, že sborníček bude užitečný i pro inspiraci užití exaktních statistických metod.

Statistických dnů se zúčastnilo asi 30 odborníků z České republiky. Od třinácti přednášejících jsme získali referáty k otištění. Tyto referáty najdete v tomto a v následujícím čísle Informačního bulletinu České statistické společnosti.

Za organizátory Statistických dnů

Prof. RNDr. PhDr. Zdeněk Půlpán, CSc.

Hradec Králové, září 2001.

METODOLOGIE DOTAZNÍKOVÝCH STUDIÍ

Marek Malý

Státní zdravotní ústav, Praha

e-mail: marek.maly@szu.cz

Abstrakt. Článek se zabývá otázkami konstrukce dotazníků pro sběr primárních dat v rámci vědeckého výzkumu a shrnuje pravidla, která by měla být v dotazníkových studiích respektována. Všímá si požadavků, které by měl splňovat dotazník jako celek i jednotlivé otázky jako jeho základní stavební prvky. Komentuje zdroje možného zkreslení, zvláště se zřetelem ke ztrátě pozorování a nízké návratnosti.

Úvod

V oblasti zdravotnictví a epidemiologického výzkumu je velmi často nutné kromě objektivně zaznamenaných údajů znát i subjektivní názory, postoje a pocity pacientů. Klíčovým nástrojem tohoto zjišťování je dotazník. I když se každý z lékařů s dotazníkem mnohokrát setkal a jeho tvorba a použití se na první pohled mohou jevit jednoduše, ve skutečnosti existuje mnoho pravidel, která je třeba při konstrukci dotazníků a metodologické přípravě dotazníkových šetření respektovat, aby výzkum skutečně mohl splnit své poslání. Design epidemiologických dotazníků a jejich začlenění do studií byly z různých pohledů opakovaně diskutovány (Dillman, 1978; Lutz, 1981; Sudman, Bradburn, 1991; Rothman, Greenland, 1998; Streiner, Norman, 1995), byly publikovány přehledy existujících dotazníků měřicích škál (McDowell, Newell, 1987). Tématika také úzce souvisí s metodikou výběrových šetření (Levy, Lemeshow, 1991; Vytlačil, 1969); i proto by měl lékař přípravu dotazníku vždy konzultovat se statistikem.

Dotazníky jsou vhodné zejména, když je potřeba shromáždit mnoho údajů logicky uspořádaných kolem hlavního problému od velkého počtu respondentů, je potřeba získat informace o preferencích jedince a jeho subjektivních postojích nebo je potřeba reagovat na předchozí fáze výzkumu či ověřit údaje získané jinou cestou. V dotazníkových šetřeních lze relativně dobře chránit soukromí respondentů, k jejich nevýhodám naopak patří jistá nepružnost, nemožnost jít do hloubky, problémy s ověřováním pravdivosti.

Dotazník jako nástroj pro sběr primárních dat

Dotazník lze popsat jako souhrn předem vybraných otázek sloužících pro shromáždění primárních dat (nových, získaných "z terénu") s vysokou efektivitou vzhledem k potřebnému času, úsilí tazatele i dotazovaného a nákladům.

Dotazník musí být přizpůsoben okolnostem, za nichž bude používán. Musí být vytvořen se zřetelem k typu statistického zjišťování, pro které slouží (úplné, nevyčerpávající/neúplné) a v případě výběrového šetření se zřetelem k typu výběru (náhodný, záměrný). Zároveň je nutno stanovit cíle a obsah celého dotazníkového šetření a formulovat pracovní hypotézy. Dále je třeba pečlivě popsat, kdo je nositelem hledané informace, jak je definován základní soubor všech nositelů informace a kolik má jedinců, jak a na základě jakého seznamu (opory výběru) má být získán vzorek (výběr) z tohoto souboru, jak má být vzorek velký (požadavek na rozsah výběru), jak má být nositel informace kontaktován, jak budou jeho údaje získány a zaznamenány. Pozornost je třeba věnovat možným problémům, tomu, jaké chyby a zkreslení (bias) mohou při sběru dat vzniknout a jak se jim bude předcházet, a kritériím pro případné vyloučení osob ze studie a pro ošetření chybějících pozorování.

K hlavním technikám pro pořizování údajů pomocí dotazníku patří jednak cílený rozhovor tazatele s dotazovaným (vedený osobně či telefonicky) a jednak dotazování, při němž záznam provádí dotazovaný sám - písemně (prezenčně, poštou), elektronicky, jako

snímek aktivit či deník (opakované záznamy). Ústní dotazování má výhodu osobního kontaktu, která dává tazateli možnost reagovat na situaci, vyložit problematická místa a zkontrolovat úplnost vyplnění. Osobní účast tazatele má zpravidla pozitivní vliv na návratnost dotazníku a často podvědomě vede probandy k větší pravdivosti odpovědí. Osoba tazatele patří ale i k rizikům postupu, protože může působit na dotazovaného negativně. Šetření s tazateli je v porovnání s písemným dotazováním (samovyplňováním) organizačně, finančně i časově náročnější. V případě, že probandi vyplňují dotazníky sami, odpovídají více s rozmyslem, protože čas vyplnění si volí sami, mají méně zábran při zodpovídání citlivých otázek a nejsou tolik ovlivněni způsobem kladení otázek. Touto formou lze zastihnout i respondenta, který není dosažitelný osobní návštěvou a lze snáze provádět opakovaná dotazování (např. průběh spotřeby potravin). Je zde však klíčová ochota dotazovaného; je snadné vynechat otázku či neodpovědět vůbec, nechat dotazník vyplnit někomu jinému, vymýšlet si odpovědi.

Příprava dotazníku

Stone (1993) popisuje kroky při přípravě dotazníku. K nim patří zejména rozhodnutí o způsobu shromáždění informací; výběr okruhu dotazů a potřebných informací z širší nabídky, a to s důsledným využitím již existujících, hlavně standardizovaných, dotazníků, případně na základě vyjádření expertů či potenciálních dotazovaných; upřesnění tvaru otázky, jejího přesného znění a nabídky odpovědí v souvislosti s požadovaným tvarem výstupů; uspořádání otázek v dotazníku z hlediska věcného sledu i grafického vzhledu; navržení způsobu kódování. Funkčnost dotazníku se ověřuje v jeho první podobě na úzkém okruhu spolupracovníků (předprůzkum) a později v pilotní studii na "průměrných" respondentech. Tak se identifikují problematické otázky a místa ve struktuře dotazníku, zjistí časová náročnost, ověří navržený způsob zpracování a vyhodnocení a jasnost instrukcí. Po úpravách vzniká konečná verze dotazníku a celou metodologii sběru dat je nutno opakovaně testovat. Je nezbytné, aby se v této fázi hodnotily i psychometrické vlastnosti dotazníků, tj. validita – výstižnost, do jaké míry měří použitý dotazník skutečně to, co je zamýšleno měřit a reliabilita - míra stability, tedy schopnosti dotazníku poskytovat stejné výsledky, kdykoli je měření opakováno za stejných podmínek. Reliabilita odráží náhodné chyby a je nutnou, ale nikoli postačující podmínkou validity, která odráží chyby systematické (Nunnally, Bernstein, 1994). Standardizací konečného dotazníku se zajistí, aby každý dotazovaný v každém výzkumu s daným dotazníkem dostal naprosto stejné otázky a aby požadovaná forma odpovědi byla stejná (Jones, 1997). Tak jsou porovnatelné výsledky z různých výzkumů navzájem a případně i s normativními daty. Standardizovat nelze dotazníky, které připouštějí jistou míru individualizace, přizpůsobení otázek respondentovi.

Typy otázek

Vlastnosti dotazníku určují především jeho základní stavební prvky – otázky. Kdosi řekl, že dobrá otázka je poloviční odpověď. Teorie dělí otázky do mnoha skupin podle různých hledisek. Otázky se klasifikují zejména podle účelu (jdoucí k jádru věci, pomocné, kontrolní), pořadí (nutno respektovat logiku věci i zaujetí respondenta, udržení jeho zájmu a různou míru přijatelnosti otázek), návaznosti (zda otázky se sebe vyplývají či mají podobnou formální strukturu), obsahu (nepřímé otázky poněkud zastírají pravý účel dotazu), závaznosti (spojeno se standardizací dotazníku, pevné formulace či adaptabilní) a formy. Členění otázek podrobně probírají Bárta (1981); Bártová, Bárta (1991); Lamser (1966); Pecáková (1995).

Zde rozlišíme otázky jen podle formy, protože toto dělení má velký význam pro praktickou realizaci. Dotazníky obsahují otázky uzavřené (kdy odpověď vybírána pouze z nabízených možností), polouzavřené (které mají předepsané varianty odpovědi, ale připouštějí i možnost formulace vlastní odpovědi) a otevřené (kdy respondent může vyslovit vše, co považuje za důležité). Tyto formy jsou doplněny ještě otázkami na kvantitativní (číselné) údaje.

Uzavřené otázky se dále dělí na alternativní neboli dichotomické (s pouze dvěma variantami odpovědi) a selektivní neboli polytomické (s více variantami). Selektivní otázky označujeme jako výběrové, pokud lze zvolit jen jednu variantu, a jako výčtové, když je současně možných více variant odpovědi (nevylučující se odpovědi). Definitivní podobu selektivní otázky určuje možnost neutrální odpovědi (lichý/sudý počet odpovědí), možnost únikové odpovědi (jiný, neznámý, atp.) a ohraničenost/neohraničenost (dle rozsahu krajní kategorie).

Uzavřené otázky jsou snadné a rychlé z hlediska výběru odpovědi, kódování i analýzy. Přehled možných odpovědí většinou dále ozřejmí položenou otázku. Na druhé straně může být obtížné sestavit úplný výčet možných odpovědí a některé závěry mohou být zavádějící z důvodu omezené nabídky odpovědí, která nutí respondenty k jednoduchým odpovědím bez možnosti upřesnění. Na tento typ otázek často vyberou nějakou odpověď i respondenti bez názoru a navíc nelze poznat, když respondent otázku chybně interpretoval. Písařské chyby snadno vzniknou, ale lze je velmi obtížně kontrolovat.

Otevřené otázky nijak neomezují počet a tvar odpovědi, nedochází ke zkreslení a respondenti mohou svou odpověď objasnit, zdůvodnit. Lze tak odhalit jejich myšlenkový postup a zachytit podstatné informace, které nebyly předvídány při plánu studie. K negativům otevřených otázek patří zejména vysoké nároky na přesnost formulace otázky a obtížná zpracovatelnost – otázky musí před vyhodnocením prohlédnout a předzpracovat odborník, což je časově náročné, únavné a s rizikem chybné či subjektivní interpretace. I respondenti musí vynaložit více úsilí a nutnost podrobnější/specifičtější odpovědi je může odrazovat. Přesnost, podrobnost, relevantnost a zaměření odpovědí se mezi respondenty silně liší a zapomětliví a špatně se vyjadřující jedinci jsou v nevýhodě.

Škálování

Subjektivní postoje, které vyjadřují respondenti svými odpověďmi na jednotlivé otázky je třeba přehledně a nekomplikovaně zaznamenat, byť za cenu jistého zjednodušení (Herzman a spol., 1995; Hindls a spol., 1999). Technika pro převod (promítnutí) subjektivně vyjádřených kvalitativních soudů, mínění, postojů a jevů zahrnujících znaky různé povahy na určitou slovně, číselně či graficky vyjádřenou stupnici pokrývající celý možný rozsah hodnot (kontinuum) se nazývá škálování. Výsledné škály mohou mít podle obecně přijaté charakterizace povahu nominální, ordinální, intervalovou či poměrovou.

Škálovací postupy mohou být založeny na vzájemném srovnávání nějakých podnětů či výpovědí (např. metoda párových srovnání), anebo na ohodnocení, oznámkování sledovaného jevu (jako u tzv. bodovacích škál). Příkladem tohoto postupu je široce využívaná Likertova metoda, při níž proband vypovídá o míře svého souhlasu či nesouhlasu s jednoduchým tvrzením, nejčastěji na pětibodové škále (např. zcela souhlasím, až na výjimky souhlasím, nelze říci jednoznačně, spíše nesouhlasím, zásadně nesouhlasím). Osgoodův sémantický diferenciál slouží k zaujetí postoje ke škálám ze souboru adjektiv opačného významu (Herzman a spol., 1995; Disman, 2000).

Funkčnost a vypovídací schopnost škál lze různými způsoby prověřovat. Zde zmíníme jen Likertův koeficient diferenciace, kterým se odhalí otázky, na něž většina odpovídá stejně, a které proto nejsou užitečné, neboť nedostatečně diferencují mezi jedinci. Koeficient se spočte jako $L = \frac{\sum x_+ - \sum x_-}{n/2}$, kde $\sum x_+$ a $\sum x_-$ představují součet 25% nejvyšších, resp. 25% nejnižších hodnot, které respondenti uvedli a n je počet respondentů. Při užití pětibodové škály nabývá koeficient hodnot z intervalu [0; 2], přičemž 0 znamená, že otázka vůbec nediferencuje. S využitím rozpětí škály R lze koeficient modifikovat ve tvaru $L^* = \frac{L}{R/2}$.

Body pro prověření každé otázky

Jelikož délka dotazníku a komplikovanost otázek zásadně ovlivňují návratnost dotazníku, je potřeba u každé jednotlivé otázky velmi pečlivě zvážit, zda je skutečně potřebná a zda v podobě, v jaké je formulovaná přináší výzkumníkovi skutečně největší užitek. Disman (2000) klade řadu „otázek o otázkách" a podobný seznam k zamyšlení představuje následující soupis:

1. Je tato otázka nezbytná, souvisí s cílem výzkumu? Není jen okrajová? Nelze odpověď získat z jiného zdroje? Není otázka v podstatě duplicitní (s výjimkou kontrolních otázek)? Jak a kde ji využijeme?
2. Ptá se otázka skutečně na to, na co chceme, aby se ptala? Je formulována adekvátně k cíli výzkumu?
3. Nepředpokládá otázka příliš velké znalosti od nějaké části respondentů? Je vůbec v silách respondentů na ni odpovědět? Nemají naopak někteří z vybraných jedinců lepší informace než jiní?
4. Nevyžaduje zodpovězení otázky příliš velké úsilí (nutnost vyhledání podkladů, výpočty, atp.)?
5. Bude otázka srozumitelná každému oslovenému a budou ji všichni chápat stejně?
6. Je otázka dostatečně konkrétní? Poskytuje dostatek informací pro rozumnou odpověď? Není příliš obecná, vágní či nejasná?
7. Je otázka jednoznačná? Neptá se na více věcí současně? Není formulována negativně, či se dvěma zápory?
8. Nepředjímá otázka určitou odpověď? Není navádějící, sugestivní, emocionálně zabarvená? Nenutí k určité odpovědi tím, že nenabízí všechny možné alternativy?
9. Pokrývají kategorie odpovědí všechny možnosti? Jsou nabízené varianty odpovědí přiměřeně podrobné a rovnoměrně vyvážené vzhledem k nabídce kladných i záporných odpovědí? Vylučují se vzájemně (u výběrových otázek)?
10. Jsou jasně rozlišeny výčtové otázky, které připouštějí více odpovědí, od otázek výběrových, které požadují jedinou volbu?
11. Je použití otevřené otázky opravdu nutné?
12. Nejedná se o otázku, na kterou lze očekávat jen velmi nespolehlivé odpovědi?
13. Je otázka správně časově zacílena? Bude se výpověď respondenta skutečně týkat požadovaného časového období? Nejedná se o otázku, k níž se postoje (a tedy i odpovědi) v čase velmi rychle mění?
14. Je otázka schopna dostatečně diskriminovat mezi jedinci? Nemá nízkou variabilitu odpovědí?
15. Není otázka citlivá, resp. pro respondenta nepříjemná či iritující? Nemůže se respondent domnívat, že přiznáním pravdivé odpovědi by mohl být ohrožen?
16. Odpovídá jazyková formulace otázky (slovník dotazníku) cílové skupině, na niž je dotazník zaměřen? Neztratil se smysl otázky přenosem z jiného jazykového či kulturního prostředí?
17. Neptá se otázka hypoteticky či na budoucí úmysly? Nevychází z falešných předpokladů?

Body pro kontrolu dotazníku

Mají-li otázky samy o sobě smysl, je ještě nutno se zabývat jejich začleněním do struktury celého dotazníku – z pohledu dramaturga, jak říká Disman (2000). Zejména je třeba se ptát:

1. Je pokryto získání všech potřebných dat pro studium zkoumaného fenoménu? Je shromažďována adekvátní informace pro nezbytnou úroveň identifikace subjektů při respektování požadavku anonymity?

2. Jsou položky (otázky) vhodně seskupeny a uspořádány v logické posloupnosti? Jsou obecnější otázky před specifickými? Jsou obtížné a citlivé otázky řazeny až ke konci?

3. Nejsou odpovědi na danou otázku (či jejich kvalita) ovlivněny sledem otázek, tj. formulací předchozích otázek a odpovědí? Nevede posloupnost příbuzných a formálně podobných otázek ke stereotypnímu zaškrtávání odpovědí (tzv. halo-efektu)?

4. Nesnaží se jediná dotazníková studie řešit příliš mnoho otázek najednou?

5. Jsou jasné instrukce pro kódování odpovědí? Je vymezený prostor pro odpověď adekvátní? Lze odlišit skutečné "nuly" od vynechané odpovědi?

6. Jsou přeskoky irelevantních otázek a organizace větvících otázek jednoduché a jasné?

7. Přispívá celková organizace dotazníku a grafická úprava k jeho zdárnému a úplnému vyplnění, k udržení pozornosti a zájmu dotazovaného? Není dotazník příliš dlouhý či časově náročný?

Zdroje zkreslení (bias)

Kromě podoby otázek a celkové struktury dotazníku jsou pro úspěch dotazníkového šetření významné ještě další faktory, které mohou zkreslit výpověď a vychýlit závěry od skutečnosti. Problémy mohou způsobit chybně zvolená populace, hypotéza, způsob či čas zjišťování; nereprezentativita či nenáhodnost výběru; rozdíly mezi tazateli; použití nestandardizovaného či neúnosně dlouhého dotazníku; nedostatečná kontrola korektnosti odpovědí, chyby v záznamech; vzájemné zaměňování odpovědí typu "nevím" s chybějícími odpověďmi. Podstatný je lidský faktor – zkoumané osoby si uvědomují, že jsou zkoumány, a nechovají se přirozeně; mají snahu předvést se v lepším světle; při dlouhodobém sledování mění podvědomě v jeho důsledku zvyklosti a postoje. Navíc je známo, že se spolehlivost dotazování liší u různých populačních skupin. Velmi negativní dopad má nízká návratnost dotazníků, tedy ztráta pozorování a s ní související a často se vyskytující rozdíly v charakteristice respondentů a nonrespondentů. Tyto problémy patří do skupiny tzv. nevýběrových chyb, které mnozí autoři považují za závažnější než chyby vyplývající z nedokonalosti výběrového plánu.

Ztráta pozorování (nonresponse) a otázky zvýšení návratnosti

Jako nonrespondent je označována osoba nereagující, odmítající vypovídat, vzdorující tlaku (tzv. renitent). Kish (1965) rozdělil příčiny ztráty pozorování do těchto kategorií:

- nepokrytí cílové populace
- absence údajů
  - proband nepřítomen doma (dočasně)
  - proband nedostupný (odmítnutí účasti při výzkumu, dlouhodobě nepřítomen - nelze získat opakovaným kontaktem)
  - proband neschopen odpovědi (z důvodů zdravotních, jazykových)
  - proband nenalezen (chyby v opoře výběru, včetně kategorií dohledání nemožné, příliš nákladné, chybí terénní pracovníci)
  - ztráta informace (dotazník neodeslán administrativní chybou, nevrácen vinou pošty, zničen, dotazník vyplněn osobou, která nepatří do výběru)

Vždy je třeba rozlišovat situace, kdy chybí celý dotazník a kdy jen některé otázky a rozlišovat, zda je proband nepřítomen nebo nedostupný, protože z charakteristik nepřítomných, kteří byli posléze dostiženi, si lze udělat představu o nedostižených. Stejně tak je třeba se snažit od zcela odmítajících získat alespoň malou část odpovědí a demografické údaje, neboť tyto dílčí údaje umožní srovnání základních charakteristik, které poskytuje nesmírně cennou výpověď o validitě výzkumu a jeho závěrů. Je třeba trvat na tom, že studie s nízkou respondencí a malou snahou o dodatečný kontakt nejsou akceptovatelné, a to i přesto, že v důsledku chybné publikační politiky mnoha časopisů se často v literatuře objevují.

V každém výzkumu je třeba hledat cesty k účinné minimalizaci vlivu chybějících pozorování (Herzman a spol., 1995), a to jednak na úrovni organizační), jednak na úrovni matematicko-statistického zpracování. Organizační prostředky ke zvýšení návratnosti sahají od psychologických až po technické: jde o vytvoření dojmu důležitosti odpovědí dotazovaného a vědomí důležitosti tématu (třeba osobním doprovodným dopisem); vytvoření atmosféry důvěry mezi dotazovaným a (detailně a pečlivě proškoleným) tazatelem (včetně např. zjevného zaručení anonymity); upozornění na výzkum v lokálních médiích; nabídnutí motivační odměny. Dotazník musí napomáhat udržení spolupráce svou strukturou a rozumnou délkou. Na nereagující probandy je třeba aktivně a různými způsoby vyvíjet tlak upomínkami (opakovaný kontakt poštou, později telefonem - Parker, Dewey, 2000), dohledáváním správného spojení. Statistické prostředky zahrnují náhradní výběr, pokud je opora výběru nedokonalá; korekci výsledků, např. vážením; využití dílčích informací o nonrespondentech. Metodou znáhodněného dotazování lze zjistit i postoje k choulostivým otázkám, na které by při přímém dotazu respondent odpověděl nepravdivě či vůbec ne (Anděl, 1995).

Zpracování dat z dotazníku

Nedílnou součástí dotazníkového šetření je zpracování dat pomocí dotazníku shromážděných. Tomu musí předcházet kontrola získaných dotazníků po formální, věcné a logické stránce. Zjištěné chyby se opravují dohledáním v materiálech, dodatečným zjištěním u tazatele či dotazovaného (při kombinovaném přístupu se např. dohledání k poštovnímu dotazníku provádí telefonem), či úsudkem z jiných odpovědí. V krajním případě jsou dotazník nebo jeho část vyřazeny ze zpracování.

Při přípravě pro zpracování jsou v dotazníku vyznačeny kódů odpovědí a příslušný odborník může převést i některé otevřené otázky. Vkládání dat do počítače má probíhat jako prostý přepis bez jakékoli interpretace či kombinování otázek – to je až věcí analýzy. První dotazníky je třeba vkládat velmi pečlivě, aby objevila případná problematické místa, resp. na neočekávané odpovědi. Je nezbytné věnovat čas kontrole vložených hodnot (dvojí zadání celého dotazníku a porovnání, inspekce extrémních odpovědí na každou otázku). Při vlastním zpracování se vychází ze základních tabelací, studia pravděpodobnostního rozložení zkoumaných jevů, ale často se aplikují i složité mnohorozměrné statistické metody a modely. Někdy pomohou speciální statistické postupy zmírnit vliv nedostatků vzniklých při realizaci dotazníkového šetření. Vždy je však možno testovat jen omezený počet hypotéz a část informace obsažené v dotazníku nutně nebude plně využita.

Závěr

Probrali jsme zde, převážně teoreticky, některé základní aspekty tvorby funkčního dotazníkového nástroje. Praxe je v této oblasti velmi rozmanitá v závislosti na vědním oboru, a proto je třeba k naplnění základní poučky „Dobrý dotazník je ten, který funguje" vykonat spousty velkých věcí i maličkostí, které se liší projekt od projektu, nelze je v úplnosti teoreticky popsat a cit pro ně se získává až na základě zkušeností.

Literatura

Anděl, M. (1995). Znáhodněné dotazování. Informační bulletin České statistické společnosti 6 (1), 14-18.

Bárta, V. (1981). Výzkum trhu. Merkur, Praha.

Bártová, H., Bárta, V. (1991). Marketingový výzkum trhu. Economia, Praha.

Dillman, D.A. (1978). Mail and telephone surveys: The total design method. Wiley, New York.

Disman, M. (2000). Jak se vyrábí sociologická znalost. 3. vyd. Nakladatelství Karolinum, Praha.

Herzman, J., Novák, I., Pecáková, I. (1995). Výzkumy veřejného mínění. Skripta VŠE, Praha.

Hindls, R., Hronová, S., Novák, I. (1999). Analýza dat v manažerském rozhodování. Grada Publishing, Praha.

Jones, P.W. (1997). Quality of life measurement: the value of standardization. Eur. Respir. Rev. 7, 46-49.

Kish, L. (1965). Survey sampling. Wiley, New York.

Lamser, V. (1966). Základy sociologického výzkumu. Svoboda, Praha.

Levy, P.S., Lemeshow, S. (1991). Sampling of populations: Methods and applications. Wiley, New York.

Lutz, W. (1981). Planning and organising a health survey. International Epidmiological Association and WHO, Geneva.

McDowell, I., Newell, C. (1987). Measuring health. A guide to rating scales and questionnaires. Oxford University Press, New York.

Nunnally, J.C., Bernstein, I.H. (1994). Psychometric theory. 3rd ed. McGraw-Hill, New York.

Parker, C.J., Dewey, M.E. (2000). Assessing research outcomes by postal questionnaire with telephone follow-up. Int. J. Epidem. 29, 1065-1069.

Pecáková, I. (1995). Statistické aspekty terénních průzkumů I. Skripta VŠE, Praha.

Rothman, K.J., Greenland, S. (1998). Modern epidemiology. 2nd ed. Lippincott-Raven, Philadelphia, 164-9.

Stone, D.H. (1993). Design a questionnaire. Brit. Med. J. 307, 1264-1266.

Streiner, D.L. Norman, G.R. (1995). Health measurement scales. A practical guide to their development and use. 2nd ed. Oxford University Press, Oxford.

Sudman, S., Bradburn, N.M. (1991). Asking questions: a practical guide to questionnaire design. Jossey-Bass, San Francisco.

Vytlačil, J. (1969). Výběrová šetření v praxi. Federální statistický úřad, SEVT, Praha.

Je statistika jen vyplňováním dotazníků či tréninkem aritmetiky?

K napsání mého příspěvku mě přivedly několikaleté zkušenosti s obhajobami diplomových a bakalářských prací studentů FaME VUT (později UTB) ve Zlíně. Že se nejedná pouze o problém této univerzity potvrzuje fakt, že na naší univerzitě studuje v magisterském studiu celá řada studentů z jiných universit.

Svého času bylo konstatováno, že každá práce by měla obsahovat nějaké statistické vyhodnocení. A tak dvě nejčetnější slova, se kterými se lze v pracích setkat, zní „marketing" a „dotazník". V souvislosti s marketingovými, ale i jinými (např. sociologickými) průzkumy, s tím lze naprosto souhlasit. Jsme ekonomická fakulta, k níž nesporně marketing patří a marketingový průzkum bez použití dotazníků a jejich vyhodnocení si jde jen stěží představit. Představy o tvorbě a vyhodnocování dotazníků jsou však, mírně řečeno, zarážející. Obecně lze konstatovat, že tvorba otázek probíhá naprosto chaoticky, s nějakou duplicitou si autoři nedělají žádné starosti, každá otázka má vždy stejnou důležitost (váhu), se škálováním si autoři hlavu nelámou. Vyhodnocení dotazníků se pak redukuje na statistický popis jednorozměrného třídění četností v podobě koláčových diagramů - i když jsou většinou barevné a hezké. S dvojným tříděním statistických dat v podobě korelačních či kontingenčních tabulek se snahou o získání představy o možných vztazích mezi sledovanými jevy jsem se dosud nesetkal. Natož pak s ověřováním závislostí na základě chí-kvadrát testu, který je založen na porovnání zjištěných sdružených četností s četnostmi očekávanými v případě nezávislosti. Jak je všeobecně známo, veličina chí-kvadrát je základem mnoha různých měr intenzity závislosti - koeficientů kontingence (Pearsonův, Čuprovův, Cramerovo V). Na naší fakultě jsou studenti v prvních dvou ročnících podrobně seznámeni s Čuprovovým koeficientem, a přesto jej u hodnocení dotazníků nevyužívají. Totéž lze říci o Spearmanově koeficientu pořadové korelace v případě analýzy vztahů mezi dvěma ordinálními proměnnými.

Nejhorší „dotazník", se kterým jsem se v poslední době setkal, nepochází bohužel z diplomových či bakalářských prací, ale slouží studentům k tomu, aby mohli hodnotit učitele. Musím podotknout, že jeho vyhodnocení nemá pro učitele nepodstatný význam. Proto považuji za vhodné jej zde zveřejnit v plné podobě (což je jednoduché, je totiž krátký).

Jméno učitele:
Předmět:
Semestr:
Datum:
Hodnocení body 0-32. Osm bodů je nejlepší hodnocení.

Otázky:					
Zná svůj předmět?	0	2	4	6	8
Dovede svůj předmět dobře učit?	0	2	4	6	8
Má náročný vztah ke studentům?	0	2	4	6	8
Má přátelský vztah ke studentům?	0	2	4	6	8

Nutno zdůraznit, že každá otázka má stejnou váhu. „Dotazník" je nepochopitelný a nelogický nejen pro mě, ale řada studentů se mi v neformálních diskusích přiznala, že jej vyplňovala jako sportku, a že na třetí otázku nevěděla, co je pro učitele lepší - zda-li je či není náročný ke studentům. Vyhodnocovaly se nejen jednotlivé otázky samostatně, ale i celkový počet bodů. A tak se lehce může stát, že učitel, který svůj předmět vůbec nezná, ale chodí se studenty na různé večírky - tudíž je k nim přátelský - je na tom průměrně, tedy dobře. Anebo naopak. Já osobně nedovedu pochopit, jak mohou studenti hodnotit, zda učitelé znají svůj předmět. To přece musí udělat někdo, kdo příslušný předmět zná lépe než hodnocený učitel. Tento můj postoj sdílí též rektor jisté soukromé brněnské univerzity.

Druhým závažným problémem, se kterým se při marketingových průzkumech v podobě dotazníků setkávám, je určení rozsahu výběru. Je jasné, že při stanovení rozsahu výběru je nutno přihlížet jak k hlediskům nestatistickým (organizačním, ekonomickým) tak i statistickým. Jako učitele statistiky mě pochopitelně zajímá kategorie druhá. Ta zdůrazňuje, že statistické přístupy ke stanovení rozsahu výběru jsou založeny na tradičních metodách statistické indukce.

Problém bych opět ilustroval na konkrétním praktickém příkladu, který se mi nedávno stal a jehož autory opět nebyli studenti, nýbrž sloužil k oficiálnímu marketingovému průzkumu. Autor průzkumu se na mě obrátil s následující otázkou:

„Kolik je třeba oslovit lidí ze souboru 11 500 (majících kabelovou televizi), abych mohl tvrdit, že výsledky mají pravděpodobnost pro celý soubor 95%?

Volební průzkumy se dělají tak, že se vyzpovídá 1000 osob a tvrdí se, že je pravděpodobnost vysoká."

Je na první pohled zřejmé, že odpověď na danou otázku nemůže být v podobě jednoho slova, v našem případě čísla. Nevím, co si autor představuje pod pojmem „výsledky". Na tom totiž záleží, jak daný rozsah stanovit. Na základě objasnění tohoto pojmu se pak statistik musí rozhodnout, zda stanoví rozsah výběru pro případ provádění intervalových odhadů či pro případ testování statistických hypotéz. To v první řadě. Následně na to pak musí stanovit o jaký konkrétní ukazatel jde, zda o střední hodnotu (což bývá nejčastěji), či o relativní četnost, rozdíl dvou středních hodnot, rozdíl dvou relativních četností, rozptyl, podíl dvou rozptylů atd..

Při běžném průzkumu je však obvykle sledován větší počet různých znaků, které mají v základním souboru odlišnou variabilitu. Každý znak tedy bude nárokovat jiný rozsah výběru. Zvolíme-li mezi možnými rozsahy ten největší, budeme mít pro všechny znaky zajištěno, že prováděné úsudky budou požadované a větší kvality. Výběr však může být neúměrně velký s ohledem na zamýšlené finanční i časové náklady průzkumu. Snížíme-li rozsah výběru, musíme počítat s tím, že některé závěry budou moci sloužit pouze jako orientační.

Vidíme tedy, že odpověď správně na otázku vyžaduje větší statistické znalosti a zkušenosti, něž aby se na první pohled mohlo laikovi zdát.

V neposlední řadě je pak zapotřebí věnovat náležitou pozornost metodám pořízení výběrového souboru. Vždy ve mně vzbudí úsměv, potkám-li na stejných místech tytéž lidi, kteří se náhodných chodců vyptávají na nejrůznější, často přihlouplé, otázky. Je zřejmé, že pokud stojí před vchodem do jistého podniku nebo na autobusovém nádraží, budou se setkávat se stále stejným (anebo aspoň přibližně stejným) vzorkem. Proto odpovědi na podobné otázky budou stejné. Správný marketingový výzkumník by se proto měl seznámit s metodami pořízení výběrového souboru a osvojit si a umět vhodně aplikovat takové pojmy, jako jsou např. náhodný výběr, výběr s vracením, výběr bez vracení, opora výběru, systematický výběr, oblastní výběr, vícestupňový náhodný výběr, samovolný výběr atd.. I tato oblast tedy vyžaduje základní znalosti statistiky.

Na závěr bych chtěl uvést jeden postřeh. Přečetl jsem několik učebnic a skript o marketingu. Vesměs bylo konstatováno, že dobrý marketingový výzkumník se musí opírat o poznatky a zkušenosti mnoha disciplín, zejména psychologie, sociologie, širokého spektra přírodních i společenských věd, měl by využívat značných možností výpočetní techniky a bohatého programového vybavení. Klíčový význam statistiky v podobě procesu sběru a analýzy informací však podle mě není dostatečně zdůrazněn.

Vladimír Rytíř
Fakulta Managementu a Ekonomiky
UTB Zlín
tel. 067-7542508
E-mail: rytir@fame.utb.cz

Lineární a logistická regrese

Zdeněk Půlpán
Univerzita Hradec Králové

Často sledujeme na týchž jedincích řadu znaků, které mohou být závislé (kouření, alkoholismus, drogová závislost). Statistická zkoumání zaměřujeme na prokázání závislosti resp. stanovení její síly (korelaci) pomocí různých ukazatelů (například korelačními koeficienty) ([10] aj.). Úkolem statistické analýzy je v případě zdůvodněného předpokladu závislosti mezi znaky i stanovení druhu závislosti. Při měření kvantitativních znaků se druh závislosti odhaduje z tvaru hypotetické křivky, plochy (nadplochy), která se hodí k napozorovaným hodnotám. „Vhodnost" se snažíme formulovat matematicky tak, aby co nejlépe vyhovovala jak vlastnostem zpracovávaných dat, tak i požadavku jednoduchosti.

Při zpracování měření dvou kvantitativních znaků je třeba stanovit, který z nich může být považován za nezávisle proměnnou (tento znak bude řídícím a označen $\mathcal{X}$), a který bude představovat závisle proměnnou (tento znak bude řízeným a označen Y). Protože předmětem statistiky je hodnocení závislostí, kdy neexistuje zcela jednoznačný vztah mezi sledovanými znaky, řídí se často tato volba zkušeností, mající svůj původ mimo statistiku. Statistika doporučuje označovat jako řídící ten znak, který je určen s větší přesností.

Bodový graf naměřených dvojic hodnot (x_i, y_i), $i = 1, 2, \cdots, n$ na n prvcích pro znaky $\mathcal{X}, Y$ napoví, jaký druh závislosti těchto znaků můžeme očekávat. Snažíme se stanovit takovou matematickou formu závislosti mezi znaky $\mathcal{X}, Y$, jejíž graf by co nejlépe aproximoval naměřené hodnoty.

Nejjednodušší typ statistické závislosti dvou znaků je vyjádřitelný ve tvaru

$$Y = \alpha + \beta \cdot \mathcal{X} + \epsilon, \tag{1}$$

kde α, β jsou vhodné konstanty a ϵ je chybová náhodná složka, vysvětlující proč všechny naměřené hodnoty (x_i, y_i) neleží v jedné přímce. Chybová složka ϵ nezahrnuje tedy jen chybu měření, ale kteroukoliv náhodnou odchylku od lineárního modelu. Předpokládáme, že kdyby nepůsobila, ležely by všechny naměřené hodnoty na přímce o rovnici $y = \alpha + \beta x$. Výsledky měření jsou však hodnoty $y_i = \alpha + \beta x_i + \epsilon_i$, lišící se od ideální o ϵ_i. O chybách předpokládáme, že jsou nezávislé, náhodné, jejich střední hodnota $E\epsilon_i = 0$ a všechna měření znaku Y jsou stejně přesná, proto chybový rozptyl $D\epsilon_i = \sigma^2$ pro $i = 1, 2, \cdots, n$.

Metodou nejmenších čtverců lze nalézt konstanty a, b tak, aby součet $Q(a,b) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ druhých mocnin rozdílů naměřených hodnot y_i od hodnot $\widehat{y}_i = y(x_i) = a + bx_i$ byl nejmenší. Označme tuto nejmenší hodnotu S_e:

$$S_e = min\ Q(a,b). \tag{2}$$

Nejmenší hodnoty S_e lze dosáhnout, když

$$b = \frac{s_{xy}}{s_x^2};\ a = \overline{y} - b\overline{x}, \tag{3}$$

kde

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;\quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2;$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Poznámka 1.: Rozdíl $y_i - \widehat{y}_i$ se nazývá *reziduum* a $Q(a,b)$ je pak *reziduální součet čtverců.* Hodnoty $\overline{x}, \overline{y}$ jsou *výběrové průměry*, s_x^2, s_y^2 jsou *výběrové rozptyly* a s_{xy} je *výběrová kovariance.* ■

Z modelu (1) vyplývá, že pro každé přesně naměřené x_i je Y_i náhodná veličina se střední hodnotou

$$E(Y_i|\mathcal{X} = x_i) = \alpha + \beta x_i \tag{4}$$

a rozptylem

$$D(Y_i|\mathcal{X} = x_i) = D\epsilon_i = \sigma^2. \tag{5}$$

Je-li navíc rozložení chybové složky ϵ normální, je také rozložení každého Y_i normální (se střední hodnotou (4) a rozptylem (5)).

Za uvedených předpokladů jsou odhady (3), získané metodou nejmenších čtverců, maximálně věrohodné a oboustranný, $100(1-\alpha)\%$ interval spolehlivosti pro parametr β modelu (1) je

$$\langle b - t_{n-2;1-\frac{\alpha}{2}} \cdot \frac{s}{s_x\sqrt{n-1}}; b + t_{n-2;1-\frac{\alpha}{2}} \cdot \frac{s}{s_x\sqrt{n-1}}\rangle, \tag{6}$$

kde $s = \sqrt{\frac{S_e}{n-2}}$; $s_x = \sqrt{s_x^2}$ a $t_{n-2;1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})\cdot 100\%$ kvantil t-rozložení při $n-2$ stupni volnosti.

Testujeme-li pak například hypotézu $H_0 : \beta = \beta_0$ proti $H : \beta \neq \beta_0$ na hladině významnosti α, zamítáme H_0 když interval (6) neobsahuje β_0; v opačném případě hypotézu H_0 nezamítáme. Speciálně pro $\beta_0 = 0$ je tento test testem lineární závislosti Y na $\mathcal{X}$; nezamítneme-li v tomto případě hypotézu H_0, můžeme předpokládat, že mezi Y a $\mathcal{X}$ neexistuje žádný lineární vztah.

Poznámka 2.: $100(1-\alpha)\%$ interval spolehlivosti pro absolutní člen α z (1) je

$$\langle a - t_{n-2;1-\frac{\alpha}{2}} \cdot \frac{s}{s_x} \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\cdot(n-1)}}; a + t_{n-2;1-\frac{\alpha}{2}} \cdot \frac{s}{s_x}\sqrt{\frac{\sum x_i^2}{n(n-1)}}\rangle. \blacksquare \tag{7}$$

Odhady parametrů α, β byly odvozeny z měření v omezeném intervalu experimentálních hodnot znaku $\mathcal{X}$ a Y. Model (1) je proto možné používat jen v rozmezí těchto intervalů.

Příklad: Mějme dva rozdílné didaktické texty A, B, kterými se má zjišťovat znalost téže oblasti učiva. O testu A víme, že je dostatečně kvalitní (je validní a reliabilní). Chceme prověřit kvalitu druhého testu. K tomu realizujeme oba testy A, B na téže reprezentatnivní skupině respondentů. Abychom vyloučili vliv pořadí řešeného testu na výsledek, bylo u každého respondenta toto pořadí určeno náhodně (s pravděpodobností 0,5 byl řešen test A).

Bylo dohodnuto, že testy budou považovány za téměř rovnocenné, když mezi výsledky $\mathcal{X}$ testu A a výsledky Y testu B bude možné předpokládat platnost modelu (1) při $\beta > 0$.

Máme k dispozici následující tabulku 1 výsledků řešení testu A a testu B u $n = 10$ respondentů

respondent	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
$\mathcal{X}$	7	6	8	10	4	4	5	7	9	9
Y	11	18	10	19	9	16	17	14	10	18

Tabulka 1

U testu A byl maximální bodový zisk 10 bodů, u testu B 20 bodů (body představují počet správně zodpovězených poožek).

Vhodnost modelu (1) posoudíme pomocí koeficientu determinace r^2, kde r je Pearsonův korelační koeficient (8)

$$r = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}. \tag{8}$$

Koeficient determinace označuje, jakou část variability Y lze modelem (1) vysvětlit. Dosazením z (3) se snadno přesvědčíme, že postupně platí

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + b^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \frac{s_{xy}^2}{s_x^2}$$

$$1 = \frac{\frac{1}{n-1}\sum(y_i - \widehat{y}_i)^2}{s_y^2} + r^2. \tag{9}$$

Výpočtem z (8) bylo zjištěno, že $r = 0,137$ a $r^2 = 0,019$. Tedy $(1 - r^2) \cdot 100\% = 98\%$ variability Y nelze vysvětlit variabilitou $\mathcal{X}$. Uvedená data nepodporují užití modelu (1).

Rovnice regresní přímky podle (3) je $\widehat{y} = 12,5 + 0,25x$. Testujme nejprve hypotézu $H_0 : \beta = 0$ proti $H : \beta \neq 0$. Interval (6) je v našem případě pro 10% hladinu významnosti

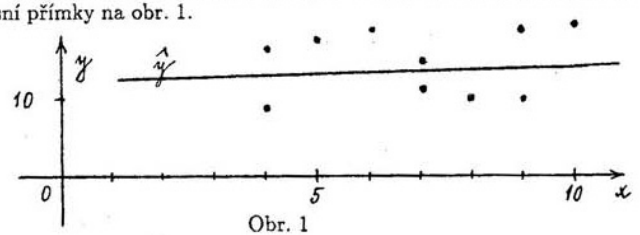
$$\langle 0,25 - 1,86 \cdot \frac{4,08}{2,31 \cdot \sqrt{9}}; 0,25 + 1,86 \cdot \frac{4,08}{2,31 \cdot \sqrt{9}} \rangle = \langle -0,85; 1,35 \rangle.$$

Obsahuje tedy nulu, a proto nezamítáme na hladině významnosti 10% nulovou hypotézu; mezi $\mathcal{X}$ a Y nemůžeme předpokládat lineární vztah.

Obsahuje-li interval spolehlivosti (6) při daném α jen kladné hodnoty, můžeme s pravděpodobností aspoň $1 - \alpha$ předpokládat, že $\beta > 0$; tj. s růstem $\mathcal{X}$ roste i Y. V našem případě je

$$b - t_{n-2;1-\frac{\alpha}{2}} \cdot \frac{s}{s_x\sqrt{n-2}} > 0, \text{ tj. } t_{8;1-\frac{\alpha}{2}} < 0,42$$

s pravděpodobností větší než 0,6. Nemůžeme tedy ani z tohoto důvodu předpokládat, že testy A, B jsou rovnocenné (v požadovém smyslu). Úvahu podporuje graf experimentálních hodnot a regresní přímky na obr. 1.



Obr. 1

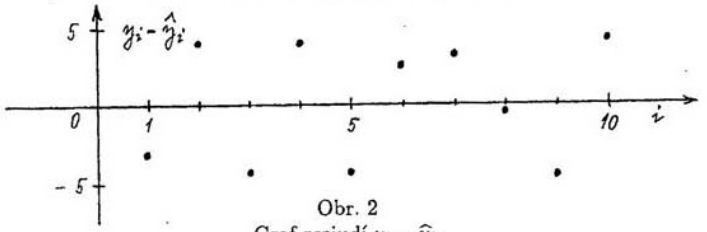
Graf experimentálních hodnot a regresní přímky.

O shodě modelu (1) s experimentálními daty se lze také přesvědčit z grafu reziduí $y_i - \widehat{y}_i$ na obr. 2, který odpovídá hodnotám z tabulky 2.

i	1	2	3	4	5	6	7	8	9	10
X_i	7	6	8	10	4	4	5	7	9	9
Y_i	11	18	10	19	9	16	17	14	10	18
$\widehat{y}_i$	14,25	14,00	14,50	15,00	13,50	13,50	13,75	14,25	14,75	14,75
$y_i - \widehat{y}_i$	- 3,25	4,00	-4,5	4,00	-4,50	2,50	3,25	-0,25	-4,75	3,25

Tabulka 2

Tabulka hodnot regresní přímky a reziduí.



Obr. 2

Graf reziduí $y_i - \widehat{y}_i$.

Z obr. 2 vidíme, že rezidua jsou sice poměrně velká, ale „rovnoměrně" rozložena okolo 0. To právě podporuje naše zjištění, že výsledky testu A neovlivňují výsledky testu B. ■

Obraťme se nyní k jediné testové položce. Hodnotíme-li její řešení respondentem jen ve dvou úrovních: správně - špatně, můžeme vzhledem k uvažované populaci označit řešení položky jako alternativní náhodnou veličinu Y s hodmotami $Y = 1$ resp. 0 (pro hodnocení například správně resp. špatně). Střední hodnota EY alternativní náhodné veličiny Y označuje podíl π správných odpovědí; její výběrovou charakteristikou je relativní počet správných odpovědí ve zkoumané populaci. Předpokládáme-li, že hodnoty náhodné veličiny Y závisí na hodnotách nějaké spojité náhodné veličiny $\mathcal{X}$ (například věku, schopnostech, znalostech, ...), pokoušíme se pro modelování podílu správných odpovědí v populaci užít opět lineární regrese. V mnoha případech však grafy naměřených hodnot $(x_i, m(x_i))$, kde x_i jsou hodnoty náhodné veličiny $\mathcal{X}$ a $m(x_i)$ jsou četnosti odpovědí, označených jako správné

při $\mathcal{X} = x_i$, připomínají S - křivku. Není proto důvod k užití lineárního modelu (1), ale ukazuje se jako výhodné užití modelu logistického:

$$Y = \pi(x) + \epsilon(x), \tag{10}$$

kde

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{11}$$

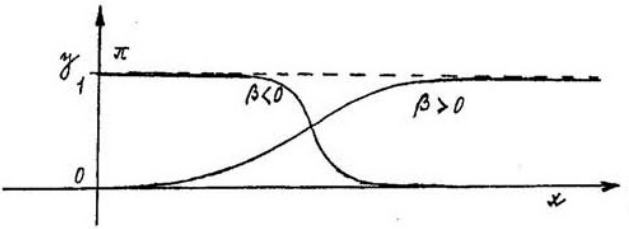
je logistická funkce. Transformace $\pi(x)$, odvozená z (11), se nazývá logitová a definuje se

$$\text{logit}(\pi(x)) = \ln \frac{\pi(x)}{1 - \pi(x)}; \tag{12}$$

z (11) a (12) vyplývá, že

$$\text{logit}(\pi(x)) = \alpha + \beta \cdot x. \tag{13}$$

Grafy závislosti π na x podle (11) jsou nakresleny na obr. 3 pro $\beta > 0$ a $\beta < 0$.



Obr. 3
Grafy $\pi(x)$ pro $\beta > 0$ a $\beta < 0$ a $x > 0$.

Tečna k logistické křivce v bodě x má směrnici

$$\beta \cdot (1 - \pi(x)) \cdot \pi(x)$$

a x-ová souřadnice inflexního bodu je $\frac{\ln 0,5 - \alpha}{\beta}$ a směrnice tečny v tomto bodě je $\frac{2}{9} \cdot \beta$. Hodnota $|\beta|$ vyjadřuje tedy „strmost" křivky (11). Pro $p = 0,5$ je logit

$$\text{logit}(0,5) = 0 = \alpha + \beta x$$

a z toho $x_0 = -\frac{\alpha}{\beta}$. Tato hodnota rozděluje všechny hodnoty x tak, že pro $\beta > 0$ a $x < -\frac{\alpha}{\beta}$ je $\pi(x) < 0,5$ a pro $x > -\frac{\alpha}{\beta}$ je $\pi(x) > 0,5$.

Model (10) se liší od modelu (1) nejen volbou funkce $\pi(x)$, ale i v tom, že chybová složka ϵ má pro pevné x v případě $Y = 1$ hodnotu $\epsilon(x) = 1 - \pi(x)$ s pravděpodobností $\pi(x)$ a hodnotou $\epsilon(x) = -\pi(x)$ s pravděpodobností $1 - \pi(x)$ když $Y = 0$. Tedy střední hodnota ϵ z (10) je sice pro každé x

$$E\epsilon(x) = (1 - \pi(x)) \cdot \pi(x) - \pi(x) \cdot (1 - \pi(x)) = 0,$$

ale rozptyl

$$D\epsilon(x) = (1 - \pi(x))^2 \cdot \pi(x) + (-\pi(x))^2 \cdot (1 - \pi(x)) = \pi(x)(1 - \pi(x))$$

závisí na x. Proto

$$0 \leq E(Y|\mathcal{X} = x) = \pi(x) \leq 1; D(Y|\mathcal{X} = x) = \pi(x)(1 - \pi(x)) \tag{13}$$

a podmíněné rozložení $Y|\mathcal{X} = x$ je binomické (ne tedy normální, jak jsme předpokládali u (1)).

Úkolem statistiky je odhadnout hodnoty parametrů α, β z n nezávislých pozorování (x_i, y_i), kde $y_i \in \{0; 1\}$ a x_i jsou hodnoty nezávisle proměnné $\mathcal{X}$. Metoda nejmenších čtverců, která se používá k odhadu koeficientů α, β u lineární regrese, se zde ukazuje jako nevhodná. Vhodnější je metoda maximální věrohodnosti, spočívající v určení takových odhadů α, β, které maximalizují věrohodnostní funkci $L(\alpha, \beta)$:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \cdot (1 - \pi(x_i))^{1-y_i}. \tag{14}$$

Stejného maxima však dosahuje i funkce (15)

$$L^*(\alpha, \beta) = \ln[L(\alpha, \beta)] = \sum_{i=1}^{n} \{y_i \ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i))\}. \tag{15}$$

Derivujeme-li (15) podle α a podle β a derivace položíme rovny nule, dostaneme dvě rovnice (16), (17):

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{16}$$

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0. \tag{17}$$

Rovnice (16) a (17) jsou pro odhady neznámých parametrů α, β nelineární a požadují speciální metody řešení. (Tyto metody jsou dnes součástí statistických výpočtů např. v programu STATISTICA resp. SPSS.) Řešení (16) a (17) označujeme $\widehat{\alpha}, \widehat{\beta}$ a pro odhad $E(Y|\mathcal{X} = x_i)$ znakem $\widehat{\pi}_i$.

K testování hypotézy $H_0 : \beta = 0$ na jisté hladině významnosti proti oboustranné alternativě $H : \beta \neq 0$ se může použít statistiky G, mající svůj původ ve věrohodnostním poměru:

$$G = -2 \ln[\frac{\text{věrohodnost při platnosti} H_0}{\text{věrohodnost při platnosti} H}] = -2 \ln[\frac{(\frac{n_1}{n})^{n_1} \cdot (\frac{n_0}{n})^{n_0}}{\prod_{i=1}^{n} \widehat{\pi}_i^{y_i} \cdot (1 - \widehat{\pi}_i)^{(1-y_i)}}] = \tag{18}$$

$$= 2 \cdot \{\sum_{i=1}^{n} [y_i \ln \widehat{\pi}_i + (1 - y_i) \ln(1 - \widehat{\pi}_i)] - [n_1 \ln n_1 + n_0 \ln n_0 - n \ln n]\}.$$

Přitom jsme značili $n_1 = \sum_{i=1}^{n} y_i, n_0 = \sum_{i=1}^{n}(1 - y_i)$.

Statistika G má asymptoticky pro dostatečně velká n χ^2 rozložení s 1 stupněm volnosti. Kritická oblast testu je tedy $\kappa_\alpha^1 = \{G; G > \chi^2_{1;1-\alpha}\}$, kde $\chi^2_{1;1-\alpha}$ je $(1-\alpha)\cdot 100\%$ kvantil.

Jinou možností je Waldův test, který využívá statistiku W, definovanou poměrem

$$W = \frac{\widehat{\beta}}{\text{standardní chyba } \widehat{\beta}}. \tag{19}$$

Statistika W má normované normální rozložení a kritickou oblastí tohoto testu je $\kappa_\alpha^2 = \{W; |W| > z_{1-\frac{\alpha}{2}}\}$, kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})\cdot 100\%$ kvantil normovaného normálního rozložení.

Oba testy využívají odhadů, získaných řešením (16) a (17), předpokládá se tedy, že je k dispozici příslušné softwarové vybavení. Užijeme-li však skórového textu, nemusíme použít statistických softwarových prostředků. Statistika S, definovaná pomocí (20):

$$S = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\cdot \sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{20}$$

má za platnsoti nulové hypotézy H_0 pro dostatečně velké n také normované normální rozložení.

Intervaly spolehlivosti pro parametry α a β se konstruují pomocí Waldovy statistiky:

$$\widehat{\beta} \pm z_{1-\frac{\alpha}{2}} \cdot \text{stand. chyba } \widehat{\beta},$$

$$\widehat{\alpha} \pm z_{1-\frac{\alpha}{2}} \cdot \text{stand. chyba } \widehat{\alpha};$$

podobně můžeme stanovit i interval spolehlivosti pro logit $(\pi(x))$

$$\widehat{\alpha} + \widehat{\beta}x \pm z_{1-\frac{\alpha}{2}} \cdot \text{stand. chyba logit } (\pi(x)).$$

Podobně jako odhady $\widehat{\alpha}, \widehat{\beta}$ i jejich standardní chyby získáme přímo ze software pro logistickou regresi.

Příklad: Hodnotíme-li výsledek řešení i-té testové položky znakem $\mathcal{X}_i = 1$ resp. 0 byla-li řešena (zodpovězena) správně resp. špatně, pak celkový testový skór je $\mathcal{X} = \sum_{i=1}^{n} \mathcal{X}_i$ (při celkem n položkách). Předpokládejme, že náhodná veličina, reprezentující celkový testový výkon $\mathcal{X}$, je spojitá (a $\sum_{i=1}^{n} \mathcal{X}_i$ je její empirické „pozadí") a hledejme pro jistou homogenní populaci odhad podmíněné pravděpodobnosti $P(\mathcal{X}_i = 1 | \mathcal{X} = x)$. Na základě předchozích úvah můžeme položit

$$P(\mathcal{X}_i = 1 | \mathcal{X} = x) = \pi(x)$$

a také psát

$$\text{logit} P(\mathcal{X}_i = 1 | \mathcal{X} = x) = \text{logit}(\pi(x)) = \alpha + \beta x.$$

Dostáváme se tak k logistické regresi, používající dvojice naměřených hodnot (x_i, y_i), kde x_i je celkový skór a $y_i = \frac{N_i^x}{N^x}$, kde N_i^x je počet jedinců, kteří vyřešili správně i-tou položku z počtu N^x těch, kteří dosáhli celkového testového výsledku x. ■

Poznámka: Volíme-li v předchozích úvahách místo funkce $\pi(x)$ funkci $\phi(\alpha+\beta x)$, kde ϕ je distribuční funkce normovaného normálního rozložení, dostáváme tzv. probitovou regresi.

■

Literatura:

[1] Komenda, S., Mazuchová, J.: Pravěpodobnostní rozdělení entropie (nit), Tvorba a testování testu, Univerzita Palackého, LF, Olomouc 1995

[2] Komenda, S., Zapletalová, J.: Analýza didaktického testu a její počítačová podpora, Lékařská fakulta, Univerzita Palackého, Olomouc 1996

[3] Půlpán, Z.: Základy sestavování a klasického vyhodnocování didaktických testů, nakl. Kotva, Hradec Králové 1991

[4] Hosmer, D., Lemeshow, S.: Applied Logistic Regression, second ed., J. Wiley & Sons, Inc. New York 2000

[5] Agresti, A., Finlay, B.: Statistical Methods for the Social Sciences, Prentice Hall, Int., Inc., New Jersey, 1997

[6] Fisher, G., H., Molenaar, J., W.: Rasch models, Foundations, recent devopments and applications, Springer - Verlag 1995

[7] Henrysson, S.: Gathering, analyzing and using data on test item, in Educational Measurement, 2nd ed., New York 1985

[8] Půlpán, Z.: K problematice vágnosti v humanitních vědách, Academia, Praha 1997

[9] Půlpán, Z.: K problematice měření v humanitních vědách, Academia, Praha 2000

[10] Půlpán, Z.: K problematice hledání podstatného v humanitních vědách, Academia, Praha 2001

[11] Spanos, A.: Probability Theory and Statistical Inference, University Press UK, Cambridge 1999

STATISTICKÉ METODY V TOXIKOLOGII

Zdeněk Roth
Státní zdravotní ústav, Šrobárova 48, 100 42, Praha 10
e-mail: roth@szu.cz

Úvod

V toxikologii je obvykle cílem zjistit, zda určitá substance, případně jiný faktor (hluk, radiace apod.), škodí normálnímu projevu života u živočichů (včetně lidí) nebo rostlin. Znalosti o toxicitě jsou založeny na výsledcích pozorování živých populací nebo výsledcích experimentu, kde jsou pozorovaní živí jedinci vystaveni definované intenzitě zkoumaného faktoru.

Kvalitativní toxikologie

V ekologii je časté porovnávání stavu zdraví u populací vystavených různé intenzitě zkoumaného faktoru s předpokládanými toxickými účinky. Výsledky se získávají tak, že se porovnávají zdravotní stavy jedinců exponovaných různou intenzitou studovanému faktoru. Tyto výsledky zjišťované na jednotlivých individuích nejsou za jinak stejných podmínek zcela totožné, vykazují náhodnou variabilitu. Předpokládáme, že tato variabilita je důsledkem nějakého statistického rozložení a že střední hodnota takových empirických dat je pro danou intenzitu působení zkoumaného faktoru dobrým odhadem jeho toxického účinku. Při tom jde zejména o závislost střední hodnoty na různé intenzitě faktoru. Pomocí statistické analýzy lze ověřit, zda případné rozdíly pozorovaných středních hodnot ukazatele zdravotního stavu nejsou jen výsledkem nahodilé variability mezi empirickými daty zjištěnými u pozorovaných jedinců.

Pro výše popsanou situaci se předpokládá, že střední hodnota kvantitativní (obvykle nějak měřitelné) proměnné charakterizující změnu zdravotního stavu se dá popsat výrazem

$$E(y_i) = \mu + \beta x_i ,$$

Kde „E" je symbolem pro střední (očekávanou) hodnotu proměnné, i je pořadové číslo empirického údaje (obvykle i = 1,2,..,n, kde n je rozsah pozorovaného vzorku), μ je průměrná hodnota pro neexponovaného (někdy tedy jen hypotetického) jedince a β je mírou vlivu velikosti toxického faktoru x_i. Statistický model pak kromě této závislosti střední hodnoty na faktoru x se ještě doplňuje tvarem statistického rozložení. Nejčastěji se předpokládá Gaussovo normální rozložení s rozptylem σ^2.

Takový jednoduchý statistický model při populačních studiích většinou nepostačuje. Proměnná y charakterizující zdravotní stav nezávisí ve většině případů jen na zkoumaném toxickém faktoru x, ale i na řadě dalších interferujících vlivů, jako jsou věk, životní styl (u lidí např. kouření, pití, dietní zvyky), a jiné faktory relevantní pro analyzovanou studii. Poměrně jednoduchý statistický model, který lépe aproximuje reálné závislosti, je lineární regresní model, který se od předchozího liší zejména v modelování střední hodnoty

$$E(y_i) = \beta_0 + \Sigma\beta_j x_{ij} ,$$

kde počet faktorů x_j ovlivňujících střední hodnotu je větší a analyzovaný toxický faktor je pouze jedním z nich. Pomocí tohoto modelu lze aspoň v hrubých rysech otestovat, zda „toxický" faktor má prokazatelně toxický vliv i za přítomnosti ostatních interferujících faktorů.

Charakter proměnné, která charakterizuje změnu zdravotního stavu, je často jen typu ano-ne. Takové binární proměnné, pro něž jednoduchým modelem je průměr dat, která označujeme 1

pro "ano" a 0 pro "ne". Takový průměr je relativní počet jedinců s nálezem "ano" a obvykle se vyjadřuje v procentech. Statistické rozložení takových dat už ovšem není Gaussovo, je to rozložení binomické s parametrem π pro střední hodnotu, který představuje teoretickou pravděpodobnost, že pozorovaný jedinec bude mít nález "ano". Takový model ovšem platí jen tehdy, je-li soubor jedinců homogenní. Pokud tuto pravděpodobnost ovlivňují další interferující faktory, lze použít statistický model obdobný lineárnímu modelu pro kvantitativní proměnné. Na rozdíl od nich však průměr binárních proměnných musí ležet v intervalu <0,1>, takže model lineární regrese pro střední hodnotu nelze použít. Přesto se pro binární veličiny dá použít obecný lineární model (GLM - General Linear Model) odvozený z maximálně věrohodnostní funkce. Používá se při tom logitové transformace, kdy se střední hodnota binomické proměnné, tedy její pravděpodobnost modeluje výrazem

$$E(y_i) = 1/(1 + \exp\{-(\beta_0 + \Sigma\beta_j x_{ij})\}\ ,$$

kde výraz v exponenciále má formu lineární funkce. Statistická metoda prokládání dat touto funkcí se nazývá *logistická regrese* kde se parametry β odhadují vzhledem k nelinearitě modelu iteračním postupem. Touto metodou se dá odhadnout parametr pro sledovaný toxický faktor i v přítomnosti interferujících faktorů a určit i interval spolehlivosti pro takový odhad, tj. i statistickou významnost jeho odchylky od nuly tedy přítomnosti toxického efektu..

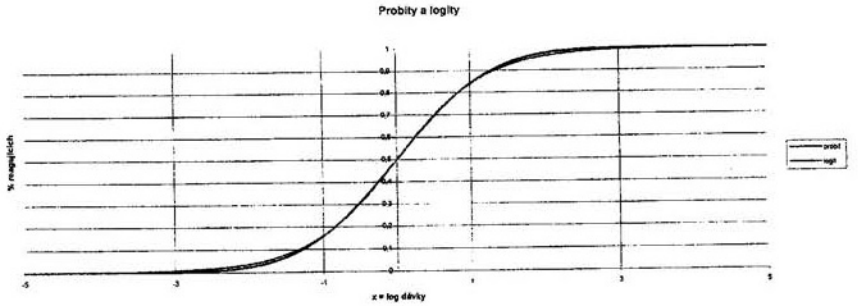
Obecné lineární modely se dají použít i pro jiná statistická rozložení. V posledních letech se pro různé biologické, tedy nejen toxikologické, vztahy konstruují velmi složité, často nelineární modely, jejichž parametr odpovídající střední hodnotě se aproximuje lineární funkcí $\beta_0 + \Sigma\beta_j x_{ij}$, jejichž vyhodnocení by dříve bylo bez moderní výpočetní techniky nerealizovatelné. Logistická regrese je jedním z takových modelů. Jejich cílem je popsat závislosti mezi výslednou proměnnou a toxickým faktorem takovým matematickým vztahem, který je biologicky, chemicky či fyzikálně zdůvodněný. Jsou to např. také modely tzv. Coxovy regrese pro vyhodnocování vlivu různých faktorů na dobu přežití, modely pro exponenciální závislosti rychlosti metabolických procesů a jejich ovlivnění léky případně jedy apod.

Poměrně složité jsou např. statistické modely, a tudíž i způsoby vyhodnocování dat při Amesově testu mutagenity.

Kvantitativní toxikologie

Při ověřování toxicity jde v shora zmíněných studiích většinou o to, zda je zkoumaný faktor prokazatelně toxický. Pokud je možno pro každého jedince určit individuální intenzitu (koncentraci, dávku) toxického faktoru, lze toxicitu faktoru definovat takovou koncentraci, která navodí toxický efekt určité velikosti. U binárních proměnných jako je uhynutí, se často za měřítko toxicity považuje koncentrace, která vyvolá reakci u poloviny exponovaných jedinců (ED50 (effective dose), LD50 (lethal dose)). U kvantitativních proměnných to může být určitá hladina nežádoucích substancí v krvi nebo moči, zrychlení srdeční nebo dýchací frekvence apod. U léčiv byly takto původně stanovovány biologické jednotky (insulin, antibiotika apod.). Je nutno poznamenat, že pro nastolení vhodných podmínek se obvykle taková stanovení provádějí v pokusném uspořádání, aby data z pokud možno malého počtu nejčastěji pokusných zvířat dala o vztahu mezi expozicí (dávkou) co nejvyšší informaci. Při vyhodnocení pokusných dat je např. důležitá ověřená znalost závislosti mezi dávkou a účinkem. Ve výše uvedeném regresním modelu se předpokládá lineární vztah mezi expozicí (dávkou) a účinkem. Je ovšem známo, že takový vztah platí (zejména u chemických substancí jako toxickém faktoru) spíše mezi logaritmem dávky a účinkem. Je tedy vhodné z dat posoudit, zda není lépe v regresním modelu vztahu mezi dávkou a účinkem použít logaritmu dávky.

Obdobně, zejména při binárních resp. binomických datech, je lépe použít vhodnou transformaci relativních počtů reagujících jedinců k získání lineárního vztahu mezi logaritmem dávky a účinkem. Jednou z možností je použít logitové transformace, která se úspěšně užívá v terénních studiích s binární veličinou pro zdravotní stav. Je třeba poznamenat, že v receptorové teorii o účinku biologicky aktivních látek se dospívá k modelu pro vztah mezi dávkou a účinkem, jehož vzorec odpovídá vztahu používanému v logitové transformaci. To se ovšem týká počtu obsazených receptorů, a tedy velikosti nějakého účinku na jednom individuu, ne však závislosti počtu reagujících jedinců na dávce. Další vhodnou transformací je transformace probitová, při které se relativní počet reagujících zvířat nahrazuje pořadnicí distribuční funkce Gaussova normálního rozložení. Předpokládá se totiž, že každé zvíře má určitou toleranci vůči testované substanci, kterou je nejnižší dávka, kterou by už zvíře nepřežilo, a že tyto individuální tolerance mají logaritmicko-normální rozložení. Čím vyšší je dávka, tím větší počet zvířat má toleranci nižší a tedy uhyne. Proložení křivky pro vztah mezi dávkou a účinkem realizované jako proložení přímky mezi logaritmem dávky a probitem odpovídajícím relativnímu počtu reagujících jedinců je vlastně odhadem logaritmicko-normálního rozložení individuálních tolerancí. Která z obou transformací je vhodnější je těžko říci, neboť obě křivky jsou velmi podobné, jak je vidět na obrázku 1.



Obr. 1: Vztah mezi dávkou a účinkem

To, že je křivka symetrická kolem středu pro logaritmy dávek, vedlo k tomu, že se navrhovaly postupy, při nichž byly dávky zvětšovány geometrickou posloupností. Z pozorovaných relativních četností reagujících pro se pak různými interpolačními metodami odhadovala dávka způsobující reakci 50% zvířat. Stejně jako u probitové či logitové metody se v těchto interpolačních postupech určoval i interval spolehlivosti pro odhadnutou dávku ED50. (Metoda Kärberova, klouzavé průměry Thompsona a Weilové).

Měření toxicity pomocí dávek LD50 a i stanovení intervalu spolehlivosti pro takovou dávku má jeden podstatný nedostatek. I v experimentálním uspořádání totiž platí, že dávka toxického faktoru není jediným faktorem, který úmrtnost zvířat ovlivňuje. Pokud jde např. o pokusné myši, je známo, že reagují jinak ráno než odpoledne, že úmrtnost ovlivňuje kromě jiného i to, zda jsou zvířata po podání jedu osamocena nebo pohromadě ve větších skupinách atd. Tyto vlivy ovlivňují výsledky pokusu a při tom ani nejsou započteny do statistické chyby, pomocí níž se počítají intervaly spolehlivosti pro dávky LD50. Z pokusu vypočtená toxicita tedy neplatí obecně jako hodnota nezávislá na okolnostech pokusu. Jedinou cestou, jak získat na pokusných podmínkách méně závislou míru toxicity, je použít v pokusu souběžně standardního přípravku o známé toxicitě a definovat toxicitu testovaného jedu relativně jako podíl hodnot LD50 obou srovnávaných substancí (relativní toxicita = LD50(standard) /

LD50(test)). Vzhledem k tomu, že interferující vlivy jsou u obou substancí totožné, a tedy obě LD50 jsou jimi stejně ovlivněné, dá se takto určená hodnota považovat za ukazatel do značné míry nezávislý na okolnostech pokusu. Problémem pouze zůstává vhodná volba standardní referenční substance.

Závěr

Toxikologie má se statistikou společný přístup ke skutečnosti. Předpokládá se, že nějaký faktor je pro živý organismus jedovatý a tato „jedovatost" či toxicita je jakási pevná míra (ve statistice parametr). Tuto pevnou míru se snaží toxikologická šetření a experimenty měřit z empirických dat. Jde tu tedy o klasický statistický postup odhadu parametru rozložení populace, kde rozložení se týká empirických měření toxického působení. Lze proto říci, že toxikologie bez statistiky je jen empirickou vědou, zatímco integrací statistických metod se teprve stává vědou obecnou.

Literatura:

Berkson,J.: A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. *J.Am.Statist.Assoc.*48 (1953) 565-569
Bliss, C. I.:The method of probits. *Science* 79 (1934) 409-410
Cox, D.R.: Regression models and life tables. *J.R.Stat.Soc.Ser.B* 34,(1972) 187-202
Finney, D.J.: Probit analysis,3rd ed. (Cambridge University Press, London 1971)
Finney, D.J.: Statistical method in biological assay; 2nd ed. (Griffin. London 1971)
Gad.S.C.: Statistivs and experimental design for toxicologists (Boca Raton, Fla.:CRC Press,1999)
Hamilton,M.: Robust estimates of the ED50. *J.Am.Statist.accoc.*74 (1979),269-278
Kaerber,G.:Beitrag zur lollektiven Behandlung pharmakologischer Reihenversuche. *Arch.Exp.Path.Pharmakol.*162 (1931) 480
Kim, B.S; Margolin, B.H.: Statistical methods for the Ames Salmonella assay:a review *Mutation Research* 436 (1999) 113-122
Salsburg ,D: Statistics for toxicologists (M.Decker.New York, 1986)
.Stead, A.G; Hasselblad,V.; Creason, J.P.; Claxton, L: Modelling the Ames test *Mutation Reasearch* 85 (1981) 13-27
Thompson,W.R.: Use of moving averages and onterpolation to estimate median-effective dose. *Bact.Rev.* 11 (1947) 115
Weil, C.S.: Tables on convenient calculation of median-effective dose (LD_{50} - ED_{50}) and instruction in their use. *Biometrics* 8 (1952) 249

ZIP regrese

Ing. Marek Brabec, PhD
Státní zdravotní ústav, Praha

Abstrakt: Článek se zabývá aplikací Poissonovské regrese na praktický problém, který vyvstává při odhadu „průchodnosti dálničních mostů pro zvěř". Cílem je vztáhnout průchodnost mostu ke kovariátám, jež určují charakter mostu (výška šířka, hloubka) a jež jsou ovlivnitelné při návrhu mostní konstrukce. Je žádoucí vliv těchto kovariát popsat jednoduchým (v praxi snadno použitelným) modelem a tento model pak testovat (např. proti konkurenčním modelům apod.). Data, jež jsou k dispozici sestávají z počtu jedinců daného druhu či skupiny druhů, jež prošli pod daným mostem. Jsou získána na základě odečtů stop ve sněhu během jistého krátkého období. Průchod zvěře pod mostem během omezeného časového intervalu (jeden až několik dnů) je „řídký jev" a volba Posissonovského modelu se tedy zdá být přirozenou. Problém však nastává s „přebytečnými nulami" – pozorovaná četnost nulových odečtů je značně vyšší než četnost předpovídaná Poisssonovským modelem. Jde tedy o komplikaci, nazývanou v literatuře ZIP (Zero Inflated Poisson). Tento výrazný rys dat je třeba vzít v úvahu jak pro odhad parametrů, tak pro výpočet následných testů. Díky přítomnosti kovariát je problém strukturován a jde tedy o analog regrese. Používáme jednoduchou modifikaci původního přístupu, jež spočívá v použití logistické regrese pro data 0/1 (neprochází nic/prochází alespoň něco) a useknuté Poissonovské regrese pro kladná data (tedy za podmínky že prochází alespoň něco). Kombinací těchto dvou regresí získáme odhady parametrů ZIP modelu, jež dovoluje jak nadbytek, tak nedostatek nul oproti Poissonovskému případu. Jde o směs degenerovaného rozdělení v nule a useknutého Poissonovského rozdělení. Model i výše zmíněný přístup k odhadu jeho parametrů ilustrujeme na výsledcích spočtených z reálných dat a srovnáváme s některými modely, publikovanými v literatuře.

1. Úvod

Jde o aplikaci, která vyvstává v souvislosti s odhadem „průchodnosti" dálničních mostů pro zvěř. Cílem je konstruovat jednoduchý model, který: i) rozumně popisuje průchodnost, ii) vztahuje ji k potenciálně důležitým kovariátám, charakterizujícím typ mostu. Vliv různých kovariát je pak testován s praktickým cílem určit takové kovariáty, jež průchodnost silně ovlivňují a dále pak i do jisté míry charakterizovat tvar této závislosti pro eventuelní praktické použití při návrhu nových mostních konstrukcí, jež by průchodu zvěře bránily co možná nejméně (v rámci daných finančních a technických omezení).

Data, jež jsou k dispozici, sestávají z počtu kusů daného druhu či skupiny druhů, prošlých pod daným mostem během jistého časového období. Takový počet je pro mnoho mostů získán z odečtu stop v okolí, zanechaných zvířaty na sněhu ve sledovaném intervalu (řádově několika dnů). Do studie jsou zařazeny mosty různých vlastností – od malých až po velké. Ke kovariátám, prakticky zajímavým pro zadavatele patří zejména výška, šířka a hloubka mostu.

Formulace modelu je na první pohled velmi jednoduchá. Relativně nízké počty zaznamenaných stop naznačují Poissonovské rozdělení jako možného kandidáta na

pravděpodobnostní část modelu. Tedy postulujeme $Y \sim Poi(\mu)$ pro napozorované počty. Vzhledem k tomu, že problém je strukturován přítomností kovariát, přímočará úvaha vede dále k Poissonovské regresi jako odpovědi na zmíněný problém. Odhady Poissonovských parametrů (intenzity průchodu) jsou formalizací „průchodnosti". Závislost takovéto průchodnosti na vybraných kovariátách lze pak snadno testovat s použitím asymptotických testů (např. poměrem věrohodností).

2. Komplikace

Komplikace zaznamenáme v okamžiku, kdy zjistíme, že data jsou sice vcelku Poissonovského charakteru pro kladné počty, ale že počet nul je příliš velký. Jinými slovy, zatímco Poi(μ) implikuje pro velké n počet nulových odečtů $\approx n\exp(-\mu)$, nulových odečtů je ve skutečnosti mnohem více. Například pro odečty kuny je počet nul 19 oproti Poissonovským 3.97 (tedy dosti dramaticky více), podobně je tomu u řady jiných druhů. Vlastník dat připouští, že nadměrný výskyt nul může být výsledkem systematického působení různých faktorů (např. kompletní neprůchodnost úzkých/nízkých mostů pro velká zvířata, nulové odečty stop v důsledku kvality povrchu, počasí apod.).

3. Model

Možným řešením je tzv. ZIP (Zero Inflated Poisson) přístup, tedy použití poněkud komplikovanějšího modelu, který nadbytek nul dovoluje. Náš model vypadá takto:

$$P(Y=k) = \begin{cases} 1-\pi & k=0 \\ \dfrac{\pi \exp(-\mu)\mu^k}{k!(1-\exp(-\mu))} & k>0 \end{cases}$$

Interpretace parametrů je přímočará:

- π popisuje „kvalitativní průchodnost"
  (pravděpodobnost že „vůbec něco" prochází)
- μ popisuje „kvantitativní průchodnost"
  (intenzitu průchodu za podmínky že „alespoň něco" prochází)

Model tedy umožňuje rozložit původně poněkud vágně definovaný pojem „průchodnosti" na dva aspekty, jež lze sledovat samostatně. Povšimněme si, že model je flexibilní v tom, že dovoluje jak nadbytek, tak i nedostatek nul oproti Poissonovské situaci. Jde o směs useknutého Poissonovského rozdělení a degenerovaného rozdělení v nule.

Protože π i μ jsou závislé na kovariátách (obecně různých), dostáváme se při kompletním popisu (strukturovaných) dat do regresního kontextu.

Přirozeným přístupem pak je použít :

- na datech 1/0 (něco prochází/neprochází nic) pro i-tý most

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \underline{x_i'\gamma}$$

tedy logistickou regresi k odhadu kvalitativní průchodnosti
- zbytek je „až na useknutí (nad nulou)" Poissonovská regrese

Odhad Poissonovské části pak lze založit na nenulových datech, maximalizací log-věrohodnosti:

$$\sum_{i:y_i>0}[-\mu_i + y_i \log(\mu_i) - \log(y_i!) - \log(1-\exp(-\mu_i))]$$

přičemž použijeme kanonický link (často používaný v Poissonovské regresi):

$$\log(\mu_i) = \underline{x}_i'\underline{\beta}$$

K vlastní numerické maximalizaci log-věrohodnosti $L(\underline{\beta})$ pak použijeme Newton-Raphsonovu metodu (kromě vyhodnocení $L(\underline{\beta})$, vyžaduje též gradient a Hessián). Ta pro nedegenerovaná data (díky regularitě modelu) vcelku rychle konverguje a poskytuje MLE odhad, spolu s odhadem tzv. pozorované Fisherovy infomace (observed information), získaným jako vedlejší produkt, z Hessiánu.

Oba kroky, tedy logistická regrese (LR) a useknutá Poissonovské regrese jsou snadno implementovatelné např. v S-plus (obecný popis tohoto prostředí, viz např. Venables, Ripley, 1994). Kombinací obou kroků dostáváme MLE odhady pro:
- všechny parametry modelu, $\underline{\beta}, \underline{\gamma}$
- intenzity a kvalitativní průchodnosti, μ_i, π_i
- snadno též pro odvozené charakteristiky, např. $E\,Y_i = \dfrac{\pi_i \mu_i}{1-\exp(-\mu_i)}$

4. Ilustrace

Ilustrací výše popsaného přístupu mohou být výsledky pro počty v kategorii drobných zvířat (ondatra, myšice, hranostaj, kolčava, kuna, tchoř, jezevec, vydra, bobr, zajíc, králík, liška, domácí šelmy). Jedním z faktorů o jejichž vliv se zadavatel intenzivně zajímal je faktor (diskretizované) šířka mostu. Definice kategorií je patrná z následující tabulky:

Kategorie	Interval šířky v metrech
1	0 až 4.9
2	5 až 14.9
3	15 a více

A) (Useknutá) Poissonovská část

1. Odhady a intervaly spolehlivosti pro kvantitativní průchodnost (intenzita průchodu za podmínky, že vůbec něco prochází) spočtené z dat za sezónu 2001.

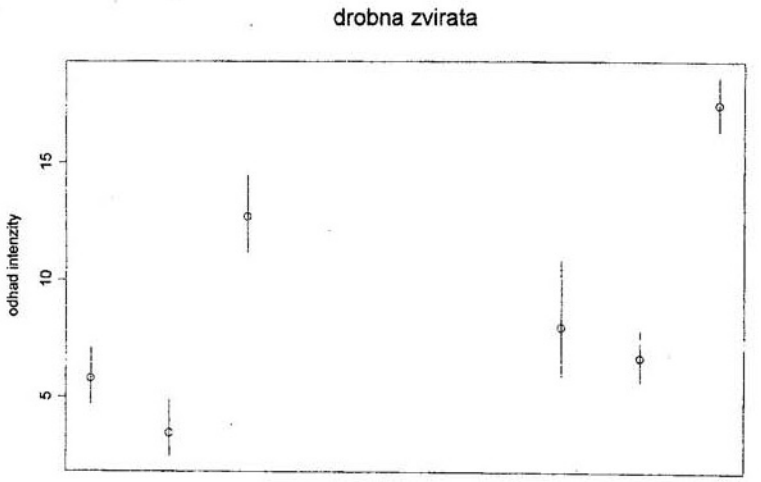
Kategorie šířky	Odhad intenzity	Střední chyba odhadu	95% konfidenční interval
1	7.997	1.172	(5.864,10.907)
2	6.706	1.088	(5.681,7.915)

3	17.469	1.035	(16.337,18.68)

Povšimněme si, že:

- model splňuje běžné podmínky regularity
- $\underline{\beta} \sim AN\left(\underline{\beta}, \frac{1}{n} I^{-1}(\underline{\beta})\right)$
- odtud lze snadno konstruovat (asymptotické) konfidenční intervaly
- $\beta_j \pm 1.96\sqrt{\left[\frac{1}{n} I^{-1}(\underline{\beta})\right]_{jj}}$
- počítat Waldovy testy či testy poměrem věrohodností

Následující obrázek srovnává 95% konfidenční intervaly pro sezóny 2000 (vlevo) a 2001 (vpravo).



2. Některé testy

Test vlivu diskretizované šířky mostu dává vysoce významný výsledek (p - hodnota<0.0001), zcela v souladu s dojmem získaným z předchozího obrázku.

Jde o (asymptotický) test poměrem věrohodností. Bereme-li pro plný model

- design matici $X = (1, X_1)$, kde X_1 je matice se sloupci danými parametrizací faktoru šířky
- a vektor parametrů $\underline{\beta}' = (\beta_0, \underline{\beta}_1')$

Je test založen na: $2\left(\max_{(\beta_0,\underline{\beta_1})} L((\beta_0,\underline{\beta_1}')) - \max_{\beta_0} L((\beta_0,\Omega'))\right) \approx \chi^2(2)$

Podobně můžeme testovat v komplikovanějších modelech.
Třeba pokud se zajímáme o vliv výšky „po očištění od vlivu šířky". Pak volíme design matici $X = (1, X_1, X_2)$, X_1 se stejným významem jako výše a:

- X_2 je matice dummy proměnných pro výšku
  Jde o analog dvoufaktorové ANOVA a pro data z roku 2001 dostáváme p-hodnota<0.0001

- $X_2 = \underline{x}_2$ je sloupec log(výška)
  Jde o analog ANCOVA a pro data z roku 2001 máme p-hodnotu 0.0002

Dále se můžeme zajímat o simultánní vliv faktoru sezóny (s úrovněmi 2000, 2001) a faktoru šířky v modelu jež je analogem dvoufaktorové ANOVA s interakcí. K tomu volíme design matici $X = (1, X_1, X_2, X_3)$, X_1 je matice dummy proměnných pro sezónu, X_2 je matice dummy proměnných pro šířku, X_3 je matice dumy proměnných pro interakci sezóna*šířka. Výsledky shrnuje následující tabulka.

Faktor	s.v.	p-hodnota
Odlišnost mezi sezónami	1	<.0001
Odlišnost mezi kategoriemi šířky	2	<.0001
Interakce sezóna*šířka	2	.2503

V souladu s grafickou prezentací výše se tedy významný jak faktor sezóny, tak faktor šířky, nikoli však jejich interakce (vliv šířky je konzistentní přes sezóny).

B) LR část (logistická regrese)

Obdobný model jako výše (analog dvoufaktorové ANOVA s interakcí) lze uvažovat pro kvalitativní průchodnost.

Faktor	p-hodnota
Odlišnost mezi sezónami	.6017
Odlišnost mezi kategoriemi šířky	.0158
Interakce sezóna*šířka	.0374

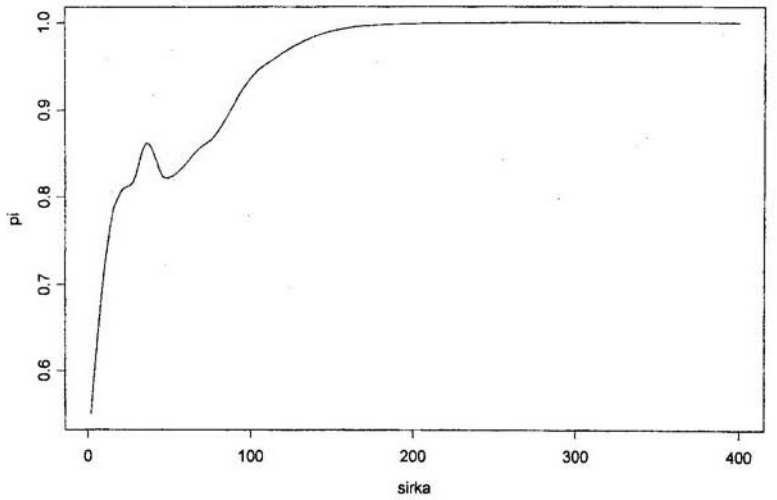
Vidíme, že zde je struktura významnosti poněkud odlišná oproti té, kterou jsme viděli u Poissonovské části (významná je zde interakce), což ilustruje dříve zmíněný fakt, že struktura kovariát může být obecně různá pro kvalitativní a kvantitativní průchodnost.

Soustředíme-li se jen na sezónu 2001, a uvažujeme-li diskretizovanou šířku, podobně jako dříve, dostáváme následující odhady kvalitativní průchodnosti:

Kategorie šířky	Odhad π	95% konfidenční interval
1	0.42	(0.18;0.69)
2	0.72	(0.54;0.86)

3	0.87	(0.76;0.94)

Alternativou je pak proložení tzv. GAM (Generalized Additive Model, viz např. Chambers, Hastie, 1992) logistického modelu, kde je šířka vzatá jako spojitá proměnná, vyhlazená loess smootherem (Cleveland, Devlin, 1988) za použití tzv. backfitting algoritmu.



Z obrázku je ihned patrné, proč jsou kvalitativní průchodnosti mezi kategoriemi výrazně odlišné. V souladu s očekáváním se kvalitativní průchodnost s rostoucí šířkou totiž zvyšuje. Překvapivě však není průběh zcela monotónní. Podobný „drnc" jako je ten na obrázku lze zaznamenat i pro jiné druhy či jejich kategorie.

5. Alternativní model

Základní model pro Poissonovská data s přebytečnými nulami lze formulovat také poněkud jinak – viz LAMBERT (1992):

$$P(Y=k) = \begin{cases} 1-\pi_L+\pi_L\exp(-\mu_L) & k=0 \\ \dfrac{\pi_L\exp(-\mu_L)\mu_L^k}{k!} & k>0 \end{cases}$$

Pro odlišení od našeho modelu, značíme parametry v modelu Lambertové s indexem L. Vidíme, že model Lambertové je směsí degenerovaného rozdělení v 0 a $Poi(\mu)$. Takováto interpretace je výhodná pro separaci generujícího nadbytečné nuly a Poissonovského procesu. Naopak restrikcí je to, že model dovoluje jen nedovoluje nedostatek nul oproti Poissonovské

situaci (jen nadbytek či s Poissonovskou situací konzistentní počet). Praktickou komplikací je i poněkud obtížnější odhad π_L parametrů.

Poznámka: pro zjednodušení píšeme μ_L a π_L bez explicitního vyjádření závislosti na kovariátách (jež mohou být obecně různé pro μ_L a pro π_L).

Je však jednoduché vidět, že (při nadbytečných nulách) jde o jednoduchou reparametrizaci modelu předchozího.

$$\mu = \mu_L$$

$$\pi = \pi_L(1-\exp(-\mu_L)) \qquad \pi_L = \frac{\pi}{(1-\exp(-\mu))}$$

Takže z pohodlně získatelných MLE odhadů (μ, π) lze snadno spočíst MLE odhady (μ_L, π_L) a naopak.

To může být užitečné zejména když proces generující nadbytečné nuly považujeme za vyloženě rušivý (např. pokud by byl ovládán jen meteorologickými podmínkami či kvalitou povrchu v době odečtu a nikoli vlastnostmi sledovaných mostů) a když chceme od tohoto rušivého procesu „čistit". Taková interpretace jistě není na místě u velkých zvířat, kde je část nul způsobená téměř fyzickou neprůchodností malých mostů, ale mohla by přicházet v úvahu u zvířat menších. Jako příklad si vezměme výsledky pro lišku v sezóně 2001.

	Kategorie šířky		
	1	2	3
$\mu = \mu_L$	1.594	2.155	4.246
π	.42	.55	.71
π_L	.52	.62	.72
$EY = \frac{\pi\mu}{1-\exp(-\mu)}$	.83	1.34	3.05

Nectností prostého ZIP modelu, tak jak byl formulován výše, a to ať v naší či v Lambertové paramerizaci je fakt, že není příliš parsimonní. Uvažujeme-li stejné kovariáty pro LR i Poi část, potřebujeme pro ZIP regresi dvakrát tolik parametrů oproti prosté Poi regresi. To může být problém při malém počtu dat a velkém modelu (s mnoha kovariátami). Proto je v práci Lambert (1992) vyvinut restringovaný model, který přikazuje jistou souvislost mezi parametry LR a Poi části.

V nejjednodušší podobě jde o to, že uvažujeme:

- kanonický, tj. log link pro Poi část, tedy $\log(\mu) = X\beta$
- kanonický, tj. logitový link pro π_L část, tedy $\log\left(\frac{\pi_L}{1-\pi_L}\right) = X\gamma$
- restringovaný model pak nařizuje

$$\log\left(\frac{\pi_L}{1-\pi_L}\right) = \tau\, X\underline{\beta}$$

tedy aby link-transformovaný π_L parametr byl proporcionální link-transformovanému μ parametru (jež je stejně jako ve standardním GLM roven lineárnímu prediktoru $X\underline{\beta}$) s konstantou proporcionality τ.

Výsledkem je model jež má jen parametry $\underline{\beta}, \tau$, místo $\underline{\beta}, \underline{\gamma}$ v nerestringovaného modelu.

Restringovaný model dává:

$$\pi_L = \frac{1}{1+\mu^{-\tau}} \qquad \pi = \frac{1-\exp(-\mu)}{1+\mu^{-\tau}}$$

Změny τ umožňují popis kvalitativně odlišných vztahů mezi π_L a μ.

- $\tau > 0$,
  π_L, π je rostoucí funkcí μ
  π_L, π, μ jsou ovládány kvalitativně podobnými mechanismy
  (působícími ve stejném směru)
- $\tau < 0$,
  π_L, π, μ jsou ovládány kvalitativně odlišnými mechanismy
  (působícími v různém směru)
- $\tau \to \infty, \qquad \pi_L \to 1$
- $\tau \to -\infty, \qquad \pi_L \to 0$

Například pro počty lišek dostáváme odhad $\tau = 0.27$, naznačující, že (v souladu s očekáváním) kvalitativní a kvalitativní průchodnost nejsou ovládány mechanismy jdoucími ve různých směrech. Tedy při zvyšování kvalitativní průchodnosti (pravděpodobnosti toho, že vůbec něco prochází) se kvantitativní průchodnost zřejmě nesnižuje.

Lambert, 1992 má i další modifikace proporcionálního modelu, které jsou opět založené na proporcionalitě mezi link transformovaným π_L a link transformovaným μ. Link pro π_L je však jiný než kanonický. Například:

- Komplementární log-log
  $\log(-\log(1-\pi_L)) = \tau\, X\underline{\beta}$
  dává
  $\pi_L = 1 - \exp(-\mu^{\tau})$

- Log-log
  $\log(-\log(\pi_L)) = -\tau\, X\underline{\beta}$
  dává
  $\pi_L = \exp(-\mu^{-\tau})$

6. Literatura

CLEVELAND, W.S – DELVIN, S.J. (1988): Locally-weighted regression: an approach to regression analysis by local fitting. JASA, **83**, 596-610.

CHAMBERS, J.M.–HASTIE, T. (eds.) (1992): Statistical models in S. Pacifik Grove, CA, Wadsworth & Brooks/Cole

LAMBERT, D. (1992): Zero-infleted Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1-14.

VENABLES, W.N.–RIPLEY, B.D. (1994): Modern applied statistics with S-plus. Springer, New York.

***

Marek Malý, Metodologie dotazníkových šetření 2
Vladimír Rytíř, Je statistika jen vyplňováním dotazníků či tréninkem aritmetiky? 9
Zdeněk Půlpán, Logistická lineární regrese 12
Zdeněk Roth, Statistické metody v toxikologii 20
Marek Brabec, ZIP regrese 24

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Předseda společnosti: Doc. RNDr. Jaromír Antoch, CSc., KPMS MFF UK Praha, Sokolovská 83, 186 75 Praha 8, e-mail: jaromir.antoch@karlin.mff.cuni.cz.
ISSN 1210 – 8022
Redakce: Doc. RNDr. Gejza Dohnal, CSc., Jeronýmova 7, 130 00 Praha 3, e-mail: dohnal@fsik.cvut.cz