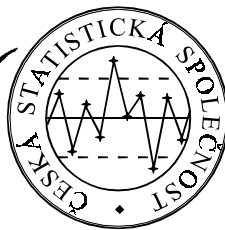


Informační Bulletin



České Statistické Společnosti

č. 3. červen 1999, ročník 10

KAŽDÝ PĚTADVACÁTÝ ČECH JE SLOVÁK

Josef Tvrđík

Včera jsem v závěru pořadu Oko uviděl na televizní obrazovce údaj, že v České republice nejpočetnější menšinou jsou Slováci, jichž je asi 400 000. Okamžitě mi bleskla hlavou věta “Každý pětadvacátý Čech je Slovák”. A bezprostředně na to jsem se zděsil, jak nepřesného a nesprávného, prostě nekorektního, vyjadřování se dopouštím, byť jen v myšlenkách.

Výrok je v rozporu se základními představami popisné statistiky. Je vůbec součet relativních četností roven jedné? Vždyť na Čechy v Česku zbývá jen nanejvýše 24/25, a kde je ten zbytek do jedné? Nebo jsou snad všichni Slováci žijící v Česku Češi? Tak tudy raději ne, ať nenarazím.

Musím v tom udělat pořádek a jasně oddělit dvě různé veličiny, totiž státní příslušnost (anglicky nationality) a národnost (anglicky nationality). Pak by ona věta mohla znít “Každý pětadvacátý občan České republiky je Slovák”. Ale je to opravdu už v pořádku? Co já vím, zda všichni Slováci žijící v Česku jsou i českými občany? Od jakého věku se vlastně člověk začne počítat mezi občany podle byrokratických zvyklostí? Tak snad raději “obyvatel” místo “občan”. Ale kdo je obyvatel? Musí mít trvalé bydliště v Česku? A co problematická formulace “každý pětadvacátý”? Copak jsou všichni uspořádáni do řady s pravidelnými rozestupy podle národnosti? Už to mám, snad v nenapadnutelné podobě: “Přibližně 4 procenta lidí žijících v České republice jsou slovenské národnosti”.

Ale snahou o korektnost vyjadřování se něco ztratilo. Především stručnost, ale i něco dalšího. V původní formulaci je skrytě přítomný i postoj, že Slováci k nám patří a beru je jako našince. Snad podobný postoj mám nebo se naučím mít i k jiným lidem. Vracím se k nekorektní formulaci, že každý pětadvacátý Čech je Slovák.

O JAZYKU BIOMETRIE A JEHO ZNEUŽÍVÁNÍ

David J. Finney

Souhrn

Každý vědní obor vyžaduje pro své pojmy, procedury a metody zvláštní názvy. Ty se často přebírají z hovorového jazyka, takže výsledkem neschopnosti rozlišit mezi specializovaným a běžným významem může být zmatek v uvažování. Lavoisier vyvinul systematické a rozšiřitelné názvosloví pro chemii. Toto názvosloví našlo širší užití v podobě metafor, takže dokonce ani chemikovi nemusí připadat nepřiměřené, když čte, že ... *hrdina novely, muž se rtuťovitým temperamentem, upřel olověný pohled na ocelové svaly svého soka.*

Matematický slovník se vyvinul vypůjčováním jednoduchých běžných slov, a statistika a biometrie pokračovaly v této praxi. Velmi rozšířený odpor a nedůvěra k statistickým metodám, běžné mezi biology, mohou být dále zhoršovány jazykovými dvojsmysly a obavami, že správné používání technické terminologie je příliš obtížné.

Zakládací listina Mezinárodní biometrické společnosti stanoví naše cíle jako „... *povznesení biologické vědy rozvíjením kvantitativních teorií, aplikací, vývoj a šíření efektivních matematických a statistických metod*“. Měli bychom proto podporovat správné užívání korektního technického jazyka, jak v mluvené, tak v psané podobě a potlačovat terminologické nejasnosti a dvojsmysly. Při výuce, recenzování i při spolupráci s jinými vědci potřebujeme bez přílišného puntičkářství spojit osobní péči s konstruktivně kritickým přístupem; to znamená zavržení rozvlácnosti, která často činí části vědeckého textu nepřesnými a potenciálně zavádějícími. Moje stanovisko je zřetelně vyjádřeno v příkladech. Budu přítom citovat z publikovaných zdrojů, ale neuvedu přesné odkazy, aby se nezdálo, že jsem příliš kritický k některým autorům.

2. Jazyk vědy

Z důvodu celistvosti a jasnosti by autorův psaný vědecký projev měl hladce plynout, ať je či není psán jeho mateřským jazykem. Autor se musí vyjadřovat přesně a srozumitelně a proto by měl vybírat taková

©International Biometric Society, 1994.

Přeložili Marek Malý, Zdeněk Roth a Ladislav Tomášek.

slova a slovní spojení, která vyjadřují myšlenky přesně, zejména se zřetelem k odborným pojmům jeho vlastního oboru nebo jakékoli jiné jím používané vědecké disciplíny.

Každý vědec se pravděpodobně setká s nesprávným použitím svého technického slovníku novináři, občas přitom může být zneužití tak velké, že vyžaduje publikovanou opravu. Žádný vědec však nemůže být spokojen, když je jeho specializovaný slovník chybně užíván pracovníkem jiného vědního oboru: takto vzniklé zmatky v terminologii a symbolice se objevují v mnoha publikacích, nejen v populárních tiskovinách, ale pohříchu i ve vědeckých článcích, které tématicky pokrývají více než jednu disciplínu. Dobře si pamatuji jednoho farmáře, člena komise pro výzkum cukrové řepy, který způsobil veselí, když popsal půdu jako chudou na obsah magnesia, zatímco měl říci manganu. Kéž bych si mohl být jistý, že chemici v této komisi jsou zvyklí pečlivě rozlišovat mezi *směrodatnou odchylkou* a *střední chybou*!

3. Biometrie

Moderní biolog nemusí být vzdělán v základech a interpretaci statistických technik vhodných pro jeho výzkum. Na svém počítači může používat softwarový systém, který mu byl doporučen jako prostředek zahrnující tyto techniky, a přesto nemusí být obeznámen se základy a činnostmi tohoto systému nebo systému použitého pro jeho data statistikem, se kterým konzultuje. V důsledku toho mohou jeho publikace obsahovat fráze a symboly, o nichž si myslí, že jsou to standardní statistické formule, ale které jsou ve skutečnosti zavedeny jen v jeho softwaru či dokonce v nějaké nestandardní učebnici. *Toto může čtenáře mystifikovat!*

Nedávno jsem poprvé použil MINITAB a byl jsem udiven závěrem uvedeným ve výstupu: *Test je signifikantní pro $p = 0.000\ 0$!* Tento pozoruhodný výstup je možná původcem tvrzení v časopise CANCER z roku 1991, že $p \leq 0.000\ 0$. Ostatně rukopis, který jsem recenzoval pro biologický časopis, obsahoval častá tvrzení, že $p < 0.000\ 0$. To může být výsledkem zaokrouhlování na čtyři desetinná místa. Nevidím však žádnou omluvu pro autory software, kteří si dovolují vytvářet absurdní výstupy svádějící uživatele k domněnce, že se jedná o autentickou statistickou terminologii. Některé výstupy dokonce pomíjejí rozlišení mezi pojmy *test* a *testová statistika*.

4. Redigování časopisů, vyučování a negramotnost

Víra, že hlavní náplní statistické analýzy je test statistické významnosti, je stále příliš rozšířená. Je tato víra odpovědná za zvyšující se výskyt formulací *významnost je 0.0211* nebo *významnost byla jen 0.33*? Mám podezření, že takové statistické vulgarismy jsou transkripce výstupů z programů, jejichž autoři jsou schopni naprogramovat statistickou aritmetiku, ale nedostává se jim porozumění jejího významu. V jedné nedávné vynikající přednášce ukázal význačný vědec v oblasti medicíny diapozitiv s výčtem možných úmrtností a komentáři; ve sloupcích s úmrtnostmi byly údaje $1 \cdot 10^2$ a $1 \cdot 10^3$, které četl jako *1 ze sta a jeden z tisíce*. Editor biologického časopisu a jeho recenzenti, vybraní obvykle pro své vynikající postavení v biologii, mohou přehlédnout opomenutí a neobvyklosti ve značení a terminologii; nicméně by měli vyžadovat vyjasnění textu před přijetím k publikaci (Finney, 1993 či Finney a Harper, 1993).

Politováníhodná praxe, při níž jsou z článků vypouštěny prakticky všechny numerické tabulky a místo nich jsou průměry prezentovány pouze v blokových diagramech připomínajících mrakodrapy na Manhattanu, vede často k tomu, že jsou tyto diagramy označovány jako *histogramy*! Nevím, zda je či není toto mylné označení užíváno nějakým konkrétním grafickým programovým systémem. Nedávno publikovaná kniha Billard et al., 1994, sponzorovaná Mezinárodní biometrickou (*sic*) společností a zamýšlená jako ilustrace dobré biometrické praxe, nejenom dokumentuje (na straně 196) neznalost pojmu histogram, ale také obsahuje žalostné pravopisné nedostatky i jiné presentační chyby.

Biologové jsou běžně seznamováni s koncepcí regrese odkazem na čáru představující závislost hodnot y -ových na x -ových. Použití slova *směrnice* jako synonyma pro *regresní koeficient* je nežádoucí, neboť tento termín není snadno zobecnitelný na nelineární regresi. Pokud programový balík používá tento termín ve svém tištěném výstupu, může to biologa podpořit ve víře, že je obecně akceptovatelný. Občas lze vidět samotné slovo *regrese* použité na místě výrazu *regresní koeficient*, což může rovněž podpořit ležerní způsob uvažování a vést dokonce k zanedbání důležitého faktu, že u koeficientu je potřeba jasně specifikovat jednotky měření (např. cm/mg).

5. Významnost

Biometrie je vázána ke speciálnímu užití slova *významný*. Mnoho biometrů a biologů automaticky předpokládá, že jakékoli publikované tvrzení popisující rozdíl jako *významný*, znamená *statisticky významný* podle běžně užívané či uvedené testové procedury a možná se dokonce domnívá, že *s pravděpodobností 0.05*. Nedávno jsem viděl vytištěno: *Rozdíl je lehce statisticky významný ($p \leq 0.0454$)*. Na jiném místě jsem četl $P = NS$.

Harris, 1976, uvádí následující hlubokou myšlenku:

Při procentním vyjádření není žádný údaj menší než 40 procent považován za statisticky významný. To vysvětluje, proč 40 procent je hranice pro složení universitní zkoušky.

V jednom článku, nedávno publikovaném ve významném časopise, zůstává nejasné, zda autoři pojmem *signifikantní* míní: *dostatečně velký, aby byl prakticky zajímavý*, jako je to běžné v hovorové mluvě, anebo *svědčící proti nevyslovené, leč implicitně zjevné nulové hypotéze*. Autoři napsali:

Pro účely těchto experimentů jsme definovali signifikantní stálou změnu odpovědi na testovací podnět následovně. Jestliže ($B \pm b$) představuje průměr a směrodatnou odchylku 12 odpovědí v průběhu první periody a ($A \pm a$) jsou průměr a směrodatná odchylka 12 odpovědí v průběhu druhé periody, pak signifikantní stálé změny jsou definovány jako ty, při kterých platí $(A - B) > (a + b)$.

Bylo by rozumné, aby editoři, autoři statistických programových systémů a všichni, kdo učí biologie nebo jim radí, vyžadovali vkládání slova *statisticky* kamkoli, kde mohou vzniknout pochybnosti o významu textu.

Jen málo biologů shledává logiku testů významnosti snadnou. Četl jsem článek, který obsahoval testy významnosti pro množství klasifikací. Autor komentoval rozdíl, který byl *nejméně významný*, snad s odkazem na ten test, který měl největší p -hodnotu (p -value) mezi všemi, pro které $p < 0.05$. Jiný autor, snad inspirován špatně napsaným softwarem, pohlížel na p jako na numerickou míru významnosti: uvedl dlouhou souhrnnou tabulku se sloupcem nadepsaným *Významnost*, v němž vypsál p -hodnoty pro skupinu standardních testů. Recenzent, který netrvá na opravě a zlepšení takovýchto věcí, nepomáhá ani čtenáři ani autorovi.

Přehnaný důraz na statistické testování biologických dat může vést ke komentářům, že se testu ... *nepodařilo dosáhnout významnosti*. Chápe

snad autor testy jako atlety, kteří se snaží ze všech sil uspět, ale občas se jim to nepovede? Opačný extrém nelogičnosti se objevil v nedávné reklamní kampani firmy SPSS. Úhledně vyvedený histogram byl doplněn komentářem: *Prošel testem chí-kvadrát na normalitu*, představujícím zjevně tendenci považovat nevýznamný χ^2 za silný důkaz toho, že výchozí rozdělení bylo skutečně normální!

6. Náhodnost a pravděpodobnost

Velice často se objevují nepřesnosti při psaní o náhodnosti (randomness). Pevně věřím v důležitost randomizace a nechci nyní přinášet argumenty ve prospěch randomizace, ale apelovat na nezkreslené uvádění faktů. Snažme se zajistit, aby se každý údaj o randomizaci nebo náhodném výběru vztahoval k explicitnímu užití zcela náhodného postupu, tak náhodného, jak jen je možné. Čtenář by měl vždy být na základě příslušného údaje informován, že proběhlo cosi velmi odlišné od čiré náhody či nerozlišujícího vybírání.

Jiným běžným zdrojem nedorozumění je vyjadřování četnosti úspěšných výsledků prostřednictvím termínu *odds* (*šance*) místo pojmu *pravděpodobnost*. Autor může být veden vírou, že *odds* je srozumitelnější širšímu okruhu lidí než *pravděpodobnost*. To skutečně může platit v anglofonním prostředí pravidelných návštěvníků dostihů a karetních hráčů, ale má to svá úskalí při použití ve vztahu k přesnému kvantitativnímu uvažování. Slova podstatná *pro* a *proti* jsou příliš často vypouštěna či ponechána čtenářově představivosti. Jestliže čtu formulaci: *Šance, že nastane jev E, je m ku n*, chápu ji tak, že *ve velmi dlouhé řadě nezávislých pokusů se E častěji objeví než neobjeví v poměru m : n*. V duchu si tu formulaci překládám jako: *Pravděpodobnost výskytu jevu E je $m/(m+n)$* .

V běžné mluvě bývají slova *naděje* (*chance*, *šance* (*odds*)), *pravděpodobnost* (*probability*, *likelihood*) a dokonce i *riziko* (*risk*) používána jako synonyma. Další zmatky mohou vzniknout v souvislosti s epidemiologickým termínem *odds-ratio*.¹ Nemohu si odpustit zmínku o jednom příkladu ilustrujícím, jak může být jazyk statistiky zneužit novináři a tvůrci reklam. Britský systém podněcující plátce daně, aby státu poskytli bezúročné půjčky, jinak známý jako Premium Bonds (výherní poukázky),

¹Tento termín bývá do češtiny překládán jako *poměr šancí*, *křížový poměr*, *sázkové riziko*, *poměr pravděpodobností expozice*, *interakce* apod. Pozn. překl.

nedávno změnil strukturu cen, takže v lednu 1994 nový oficiální leták obsahoval atraktivní tvrzení, že:

Každá obligace v hodnotě jedné L má svou vlastní a pro všechny obligace stejnou šanci, že bude tažena jako prémie v měsíčním slosování. Šance, že každá jednotlivá obligace v hodnotě jedné L vyhraje cenu, je 15 000 ku 1 v každém měsíčním tahu.

Nekontroloval jsem aritmetiku, ale řekl bych, že pravdou může být, že šance proti výhře je „jen“ 15 000 ku 1.

7. Různé

Zneužitý mohou být dokonce i základní statistické pojmy. V časopise zabývajícím se lidskou genetikou jsem četl:

Vážený pane,

podle definice je průměrná hodnota libovolného znaku optimální pro daný druh a dané prostředí. Jaká je pak biologická nevýhoda vysokého IQ?

Sherwin, 1994, pokáral statistiky za matení techniků slovy: „... určitý druh pedantství, které trvá na tom, aby byl běžný průměr nazýván aritmetickým průměrem“. Není větší pedantství v tom, když chemik přejmenuje louh sodný na hydroxid sodný? Musíme se nicméně mít na pozoru před rizikem, že technický termín může být zavádějící, pokud je téměř shodný s výrazem běžně užívaným v obecném hovoru. Jen málo lze udělat, abychom zabránili nestatistikům v chybném užívání naší terminologie. Altman a Bland, 1991, například zjistili, že ... *analýza rozptylu byla použita jako jméno pro určitý postup při hospodaření s nemocničními zdroji!*

Čas od času sice můžeme my sami býti inovátory, ale třebaže mezery ve vzdělání limitují naše možnosti odvozovat termíny z klasických jazyků, měli bychom vyvinout dostatečné úsilí a zabránit nejasnostem volbou takových slov, která pravděpodobně nebudou svádět novináře, aby je používali jako nové metafory či prázdná hesla. Povšiml jsem si, že dokonce jeden z papežů moderní aplikované statistiky (Taguchi, 1987) používá slova *variance (variation)* jako kdyby to byl přesně definovaný pojem, snad synonymum pro rozptyl či směrodatnou odchylku.

Vyjadřování změn pomocí procent může být matoucí, pokud se poruší konvence, že procentuální změna musí vždy být určena relativně vzhledem k základní hladině. Uvažujme vlastnost (nemoc či jinou charakteristiku), která se vyskytovala u 65% lidí v populaci. O rok později se

tento podíl změnil na 40%. Má se uvedená změna popisovat jako 25%-ní nebo jako 38%-ní ($=25 \cdot 100/65$) snížení? Pečlivý vědec by to měl popsat tak, aby nezůstaly žádné pochybnosti o jeho interpretaci. Podobný slovní obrat, oblíbený mezi novináři, ale neznámý ve vědeckých časopisech, je konstatování, že určitý typ události byl v roce 1994 třikrát méně častý než v roce 1993; domnívám se, že to má znamenat, že četnost v roce 1994 byla jednou třetinou četnosti v roce 1993, ale proč to tedy tak neříci? O nic lepší se mi nezdá ani výrok . . . *V průběhu tohoto období byla úmrtnost na XXXX snížena více než na polovinu.*

Ani matematické symboly nejsou odolné vůči nebezpečí nesprávného užití. Symbol $<$ je matematickou zkratkou pro slova *je menší než*, a v tomto významu by měl být umístěn mezi dva numerické či symbolické výrazy, ale ne někam jinač. V časopise pro biologický výzkum jsem nedávno viděl nevhodný zápis:

Všechny chyby byly $< 5\%$.

Přemýšlel jsem, zda se autor později nebude odkazovat na *jistý* \int *krok ve svých argumentech* či zda by nějaký rostlinný fyziolog nemohl příležitostně napsat o výzkumu obilninových $\sqrt{u^2}$.

Nepořádný pravopis, i když snad zřídka bývá příčinou nedorozumění, může vést k chybnému užití jazyka, ať již ve vědeckém kontextu či mimo něj. Při citování jiných vědců a při uvádění eponymicky pojmenovaných metod považují chybný pravopis za nezdvořilost. U terminologie, která pochází z jiných jazyků než z angličtiny, musíme mít na paměti, že vypuštění čárek či jiných diakritických znamének může zmást některé čtenáře, ačkoliv se musíme smířit s tím, že u mnoha jazyků s málo běžným písmem je přijetí verze psané latinkou nevyhnutelné.

Literatura

- Altman D. G., Bland J. M. (1991): *Improving doctors understanding of statistics*. Journal of the Royal Statistical Society, A154, 223–267.
- Anonym (1994): *Premium Bonds (Go on, have a flutter)*. Department of National Savings DNS 765/94/01.
- Billard L. et al. (eds.) (1994): *Case Studies in Biometry*. J. Wiley, New York.
- Finney D. J. (1993): *On presenting the results of statistical analyses in scientific journals*. Biometrical Journal, 35, 499–509.

²Angl. root=kořen. Pozn. překl.

- Finney D. J., Harper J. L. (1993): *Editorial code for presentation of statistical analyses*. Proceedings of the Royal Society, London, B254, 287–288.
- Harris M. (1976): *Fluoridation: Statistical evaluation of data*. The Probe, 18, 85–87.
- Sherwin D. J. (1994): *Engineers ignorance of statistics*. RSS News, 21, 6.
- Taguchi G. (1987): *System of Experimental Design*. Kraus Int. Publ., White Plains NY.

Autobiografická poznámka

Prof. David John Finney, ScD (* 1917) patří k předním britským biostatistikům. Pracoval v Rothamstedu, Oxfordu, Aberdeenu a v letech 1966–84 byl profesorem statistiky a vedoucím katedry statistiky na universitě v Edinburghu. Je členem mnoha statistických institucí, nositelem Řádu britského impéria (CBE), řady dalších vyznamenání a čestných doktorátů. Byl prezidentem Mezinárodní biometrické společnosti (1964–65) a Královské statistické společnosti (1973–74). Je autorem mnoha článků ve statisticky a biologicky orientovaných časopisech a několika monografií, zejména: Probit Analysis (1947, 1971), Statistical Method in Biological Assay (1952, 1978), Statistics for Biologists (1980). Kromě vlastní statistické práce dlouhodobě usiluje o čistotu statistického uvažování, vyjadřování a praxe, kromě tohoto textu např. v člancích:

Jedná se o překlad článku *On Biometric Language and its Abuses*, jež byl publikován v *Biometric Bulletin* 11 (4), 1994, pp. 2–4. Biometric Bulletin je informačním bulletinem Mezinárodní biometrické společnosti, která nám laskavě poskytla souhlas s publikací tohoto překladu, souhlas poskytl i profesor Finney osobně.

- Finney D. J. (1968): *Teaching biometry in the university*. Biometrics 24, 1–12.
- Finney D. J. (1991): *Ethical aspects of statistical practice*. Biometrics 47, 331–339.
- Finney D. J. (1995): *Statistical science and effective scientific communication*. Journal of Applied Statistics 22, 293–308.
- Finney D. J. (1997): *The responsible referee*. Biometrics 53, 715–719.

ASYMPTOTICKÉ TESTY

JIŘÍ ANDEĚL

Při významných výročí známých osobností bývá zvykem, že žáci, přátelé a spolupracovníci přispějí publikací odborného či vědeckého článku k počtě takové události. Není pochyb o tom, že Ing. Josef Machek, CSc., takovou známou osobností české statistiky opravdu je. Není pochyb o tom, že sedmdesátka je významným výročím. Ing. Machek je znám jako vynikající pedagog, který umí přiblížit svým posluchačům klasické i moderní partie matematické statistiky s velkým citem pro jejich aplikovatelnost.

V posledních desetiletích se setkáváme ve statistické literatuře s některými novými odvozeními známých testových procedur i s publikací originálních a užitečných testů. Ačkoliv metody, na nichž to je založeno, se pokládají za všeobecně známé, není snadné najít příslušné literární odkazy. Následující text se pokouší přiblížit celou tuto problematiku. Děkuji Ing. Machkovi za zájem, s jakým sledoval vytváření předchozích verzí tohoto příspěvku, i za poznámky, které vedly k jeho vylepšení. Domnívám se, že je to typ statistických metod, které mu vždy byly blízké.

1. LR TEST

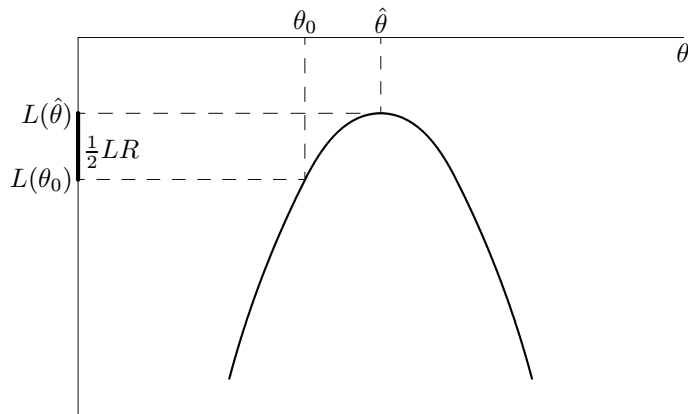
Nechť $\theta \in \Omega$, kde Ω je jednorozměrný parametrický prostor. Nechť $l(\theta)$ je věrohodnostní funkce a $L(\theta) = \ln l(\theta)$ logaritmická věrohodnostní funkce. Pro testování jednoduché hypotézy $H_0 : \theta = \theta_0$ proti alternativě $H_1 : \theta \neq \theta_0$ se používá test založený na věrohodnostním poměru (tzv. LR test, čili likelihood ratio test)

$$\lambda = \frac{l(\theta_0)}{l(\hat{\theta})}.$$

Dá se dokázat, že za platnosti H_0 má za velmi obecných podmínek náhodná veličina

$$(1.1) \quad LR = -2 \ln \lambda = 2[L(\hat{\theta}) - L(\theta_0)]$$

asymptoticky χ_1^2 rozdělení. Schématicky jsou některé veličiny znázorněny na obr. 1.



Obr. 1

Pokud θ je k -rozměrný parametr, θ_0 jeho skutečná hodnota a $\hat{\theta}$ jeho maximálně věrohodný odhad, pak za platnosti H_0

$$(1.2) \quad LR = 2[L(\hat{\theta}) - L(\theta_0)]$$

má asymptoticky χ_k^2 rozdělení.

2. WALDŮV TEST

Z Taylorova vzorce dostaneme aproximaci

$$L(\theta_0) \doteq L(\hat{\theta}) + (\theta_0 - \hat{\theta})L'(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2 L''(\hat{\theta}).$$

Protože $L'(\hat{\theta}) = 0$, máme odtud

$$2[L(\hat{\theta}) - L(\theta_0)] \doteq -(\hat{\theta} - \theta_0)^2 L''(\hat{\theta}).$$

Veličina na levé straně je rovna LR . Vzhledem k tomu, že veličina $-L''(\hat{\theta})$ je odhadem Fisherovy informace $I(\hat{\theta})$, je Waldův test založen na statistice

$$(2.1) \quad W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}),$$

která má za platnosti H_0 asymptoticky χ_1^2 rozdělení.

V případě k -rozměrného vektoru parametrů θ je Waldova statistika definována vzorcem

$$(2.2) \quad W = (\hat{\theta} - \theta_0)' \mathbf{I}(\hat{\theta})(\hat{\theta} - \theta_0),$$

kde $\mathbf{I}(\theta)$ je Fisherova informační matice. Za platnosti H_0 má W asymptoticky χ_k^2 rozdělení. Výhodou vzorce (2.2) ve srovnání s dále uvedeným vzorcem (3.3) je to, že matici $\mathbf{I}(\theta)$ není nutné invertovat. Nevýhodou je to, že je třeba počítat hodnotu této matice v bodě $\hat{\theta}$. Jak výpočet tohoto bodu, tak i výpočet matice $\mathbf{I}(\hat{\theta})$ může být komplikovaný.

3. SKÓROVÝ TEST

Tomuto testu se často říká test založený na Lagrangeových multiplika-torech. Zdůvodnění tohoto názvu se najde v literatuře (např. Judge a kol. 1988). Vyjde se z aproximace založené na Taylorově vzorci

$$L'(\hat{\theta}) \doteq L'(\theta_0) + (\hat{\theta} - \theta_0)L''(\theta_0).$$

Jelikož $L'(\hat{\theta}) = 0$, máme odtud

$$\hat{\theta} - \theta_0 \doteq -\frac{L'(\theta_0)}{L''(\theta_0)}.$$

Proto

$$(3.1) \quad W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}) \doteq \frac{[L'(\theta_0)]^2}{[L''(\theta_0)]^2} I(\hat{\theta}).$$

Víme, že $-L''(\theta_0)$ je odhadem Fisherovy informace $I(\theta_0)$. Platí-li H_0 , pak $\hat{\theta}$ je konsistentním odhadem θ_0 a $I(\hat{\theta})$ je odhadem $I(\theta_0)$. Proto se pravá strana vzorce (3.1) dá aproximovat výrazem

$$(3.2) \quad LM = \frac{[L'(\theta_0)]^2}{I(\theta_0)}.$$

Veličina LM (jejíž označení vzniklo z Lagrange multiplier) má za platnosti H_0 asymptoticky χ_1^2 rozdělení.

Poznamenejme, že se u regulárních systémů hustot předpokládá, že $EL'(\theta) = 0$ a definuje se $I(\theta) = E[L'(\theta)]^2$. Tyto střední hodnoty se berou při skutečné hodnotě parametru θ . Platí-li H_0 , vzhledem k asymptotické normalitě má pak $L'(\theta_0)/\sqrt{I(\theta_0)}$ asymptoticky $N(0, 1)$, takže čtverec této veličiny, který je roven LM , má vskutku asymptoticky χ_1^2 .

Je-li θ vektor o k složkách, pak

$$(3.3) \quad LM = [L'(\theta_0)]' [I(\theta_0)]^{-1} L'(\theta_0)$$

má za platnosti H_0 asymptoticky χ_k^2 rozdělení. Výhodou vzorce (3.3) je to, že se v něm vůbec nevyskytuje hodnota $\hat{\theta}$, takže ji není nutné počítat. Na druhé straně je vyžadována inverze Fisherovy informační matice. To ale při nevelkém počtu parametrů nebývá problém.

4. PŘÍKLAD: VÝBĚR Z $N(\theta, 1)$

Nechť X_1, \dots, X_n je výběr z $N(\theta, 1)$. Věrohodnostní funkce je

$$l(\theta) = \frac{1}{(2\pi)^{n/2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2 \right\}$$

a logaritmická věrohodnostní funkce je

$$L(\theta) = \ln l(\theta) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2.$$

Zabývejme se testem hypotézy $H_0 : \theta = \theta_0$ proti alternativě $H_1 : \theta \neq \theta_0$. Nejprve vypočteme

$$L'(\theta) = \sum (X_i - \theta), \quad L''(\theta) = -n, \quad I(\theta) = E[-L''(\theta)] = n.$$

Z toho plyne, že $\hat{\theta} = \bar{X}$. Dále dostaneme

$$LR = 2[L(\hat{\theta}) - L(\theta_0)] = \sum (X_i - \theta_0)^2 - \sum (X_i - \bar{X})^2 = n(\bar{X} - \theta_0)^2,$$

$$W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}) = (\bar{X} - \theta_0)^2 n,$$

$$LM = \frac{[L'(\theta_0)]^2}{I(\theta_0)} = \frac{1}{n} \left[\sum (X_i - \theta_0) \right]^2 = n(\bar{X} - \theta_0)^2.$$

5. PŘÍKLAD: VÝBĚR Z $Po(\lambda)$

Nechť X_1, \dots, X_n je výběr z Poissonova rozdělení s parametrem λ . Věrohodnostní funkce je

$$l(\lambda) = \frac{\lambda^{\sum X_i}}{X_1! \dots X_n!} e^{-n\lambda}.$$

Proto logaritmická věrohodnostní funkce a její derivace podle λ jsou

$$\begin{aligned} L(\lambda) &= -\ln(X_1! \dots X_n!) + n\bar{X} \ln \lambda - n\lambda, \\ L'(\lambda) &= \frac{n\bar{X}}{\lambda} - n\lambda, \quad L''(\lambda) = -\frac{n\bar{X}}{\lambda^2}. \end{aligned}$$

Fisherova informace je

$$I(\lambda) = -EL''(\lambda) = \frac{n}{\lambda^2} E\bar{X} = \frac{n}{\lambda}.$$

Ze vzorců (1.1), (2.1) a (3.2) postupně dostaneme testové statistiky pro test $H_0 : \lambda = \lambda_0$ proti $H_1 : \lambda \neq \lambda_0$ ve tvaru

$$\begin{aligned} LR &= 2n[\bar{X}(\ln \bar{X} - \ln \lambda_0) - (\bar{X} - \lambda_0)], \\ W &= \frac{n(\bar{X} - \lambda_0)^2}{\bar{X}}, \\ LM &= \frac{n(\bar{X} - \lambda_0)^2}{\lambda_0}. \end{aligned}$$

V tomto případě se všechny tři statistiky od sebe liší. Přitom každá z nich má za platnosti H_0 asymptoticky χ_1^2 rozdělení.

6. ZNAMÉNKOVÝ TEST DOPLNĚNÝ TESTEM NA POŘADÍ

Pokusné osoby jsou dotazovány, kterému ze dvou ošetření A a B dávají přednost. Přitom daný počet osob n_1 nejprve vyzkouší A (a pak B) a daný počet osob n_2 nejprve vyzkouší B (a pak A). Celkový počet dotazovaných osob je $n = n_1 + n_2$. Odpovědi jsou sumarizovány v tab. 1.

Tab. 1

| | Počet osob preferujících A | Počet osob preferujících B | Celkem |
|-----------|-------------------------------|-------------------------------|--------|
| Nejdřív A | N_{11} | N_{12} | n_1 |
| Nejdřív B | N_{21} | N_{22} | n_2 |
| Celkem | $N_{.1}$ | $N_{.2}$ | n |

V praxi bývá obvyklé volit $n_1 = n_2$. Pokud preference nezávisí na pořadí, k hodnocení dat se používá znaménkový test. Ten je založen na tom, že v případě stejných preferencí A i B platí $N_{.1} \sim Bi\left(\frac{1}{2}, n\right)$. Viz Anděl (1998), str. 80. Test, který bere v úvahu možný vliv pořadí, publikovali Rayner a Best (1997).

Obecně však u osob, které dostanou nejdřív A, je pravděpodobnost preference A rovna p_{11} a pravděpodobnost preference B je rovna $p_{12} = 1 - p_{11}$. Podobně u osob, které dostanou nejdřív B, je pravděpodobnost preference A rovna p_{21} a pravděpodobnost preference B je $p_{22} = 1 - p_{21}$. Protože jde o dvě nezávislá binomická rozdělení, je v tomto případě věrohodnostní funkce

$$l = \frac{n_1!}{N_{11}!N_{12}!} p_{11}^{N_{11}} p_{12}^{N_{12}} \frac{n_2!}{N_{21}!N_{22}!} p_{21}^{N_{21}} p_{22}^{N_{22}}.$$

Místo pravděpodobností p_{11} a p_{12} zavedme dva nové parametry θ a ϕ . Parametr ϕ bude vyjadřovat vliv pořadí, v jakém byla ošetření aplikována, na preference. Parametr θ pak bude vyjadřovat vliv rozdílu mezi přípravky. Položme

$$\begin{aligned} p_{11} &= \frac{1}{2} + \theta + \phi, & p_{12} &= \frac{1}{2} - \theta - \phi, \\ p_{21} &= \frac{1}{2} + \theta - \phi, & p_{22} &= \frac{1}{2} - \theta + \phi. \end{aligned}$$

Pak logaritmická věrohodnostní funkce bude

$$\begin{aligned} L(\theta, \phi) = \ln l &= \ln \left(\frac{n_1!}{N_{11}!N_{12}!} \frac{n_2!}{N_{21}!N_{22}!} \right) \\ &+ N_{11} \ln \left(\frac{1}{2} + \theta + \phi \right) + N_{12} \ln \left(\frac{1}{2} - \theta - \phi \right) \\ &+ N_{21} \ln \left(\frac{1}{2} + \theta - \phi \right) + N_{22} \ln \left(\frac{1}{2} - \theta + \phi \right). \end{aligned}$$

Odtud se dostanou věrohodnostní rovnice

$$\begin{aligned}\frac{\partial L(\theta, \phi)}{\partial \theta} &= \frac{N_{11}}{\frac{1}{2} + \theta + \phi} - \frac{N_{12}}{\frac{1}{2} - \theta - \phi} + \frac{N_{21}}{\frac{1}{2} + \theta - \phi} - \frac{N_{22}}{\frac{1}{2} - \theta + \phi}, \\ \frac{\partial L(\theta, \phi)}{\partial \phi} &= \frac{N_{11}}{\frac{1}{2} + \theta + \phi} - \frac{N_{12}}{\frac{1}{2} - \theta - \phi} - \frac{N_{21}}{\frac{1}{2} + \theta - \phi} + \frac{N_{22}}{\frac{1}{2} - \theta + \phi}.\end{aligned}$$

6a. Test na efekt ošetření.

Chceme testovat hypotézu $H_0 : \theta = 0$ proti alternativě $H_1 : \theta \neq 0$. Víme, že LM test je založen na

$$U = \left. \frac{\partial L}{\partial \theta} \right|_{\theta=0} = \frac{N_{11}}{\frac{1}{2} + \phi} - \frac{N_{12}}{\frac{1}{2} - \phi} + \frac{N_{21}}{\frac{1}{2} - \phi} - \frac{N_{22}}{\frac{1}{2} + \phi}.$$

Za platnosti H_0 je

$$\begin{aligned}N_{11} &\sim Bi\left(\frac{1}{2} + \phi, n_1\right), & N_{12} &= n_1 - N_{11}, \\ N_{21} &\sim Bi\left(\frac{1}{2} - \phi, n_2\right), & N_{22} &= n_2 - N_{21},\end{aligned}$$

takže

$$\begin{aligned}U &= \left(\frac{1}{\frac{1}{2} + \phi} + \frac{1}{\frac{1}{2} - \phi}\right) N_{11} + \left(\frac{1}{\frac{1}{2} - \phi} + \frac{1}{\frac{1}{2} + \phi}\right) N_{12} - \frac{n_1}{\frac{1}{2} - \phi} \\ &\quad - \frac{n_2}{\frac{1}{2} + \phi}, \\ \text{var } U &= \left(\frac{1}{\frac{1}{2} + \phi} + \frac{1}{\frac{1}{2} - \phi}\right)^2 n_1 \left(\frac{1}{2} + \phi\right) \left(\frac{1}{2} - \phi\right) \\ &\quad + \left(\frac{1}{\frac{1}{2} - \phi} + \frac{1}{\frac{1}{2} + \phi}\right)^2 n_2 \left(\frac{1}{2} + \phi\right) \left(\frac{1}{2} - \phi\right) \\ &= \frac{n}{\left(\frac{1}{2} + \phi\right) \left(\frac{1}{2} - \phi\right)}.\end{aligned}$$

Veličina U i její rozptyl obsahují neznámý parametr ϕ . Ten za platnosti H_0 odhadneme z věrohodnostní rovnice $\frac{\partial L}{\partial \phi} = 0$, tj. z

$$\frac{N_{11}}{\frac{1}{2} + \phi} - \frac{N_{12}}{\frac{1}{2} - \phi} + \frac{N_{21}}{\frac{1}{2} - \phi} - \frac{N_{22}}{\frac{1}{2} + \phi} = 0.$$

Odtud dostaneme pro ϕ maximálně věrohodný odhad

$$\hat{\phi} = \frac{N_{11} + N_{22} - N_{12} - N_{21}}{2n}.$$

Dále máme

$$\frac{1}{2} + \hat{\phi} = \frac{N_{11} + N_{22}}{n}, \quad \frac{1}{2} - \hat{\phi} = \frac{N_{12} + N_{21}}{n}.$$

Testová statistika pro H_0 je

$$U^* = \frac{\frac{N_{11} - N_{22}}{N_{11} + N_{22}} - \frac{N_{12} - N_{21}}{N_{12} + N_{21}}}{\sqrt{\frac{\frac{N_{11} + N_{22}}{n} \frac{N_{12} + N_{21}}{n}}{n}}} = \frac{2(N_{11}N_{21} - N_{12}N_{22})}{\sqrt{n(N_{11} + N_{22})(N_{12} + N_{21})}}.$$

Pokud $|U^*| \geq u\left(\frac{\alpha}{2}\right)$, zamítáme H_0 na hladině, která je asymptoticky rovna α . Analogicky se dají konstruovat i jednostranné testy.

6b. Test na efekt pořadí.

Jde o test hypotézy $H'_0 : \phi = 0$ proti alternativě $H'_1 : \phi \neq 0$. Test se založí na veličině

$$V = \left. \frac{\partial L}{\partial \phi} \right|_{\phi=0} = \frac{N_{11}}{\frac{1}{2} + \theta} - \frac{N_{12}}{\frac{1}{2} - \theta} - \frac{N_{21}}{\frac{1}{2} + \theta} + \frac{N_{22}}{\frac{1}{2} - \theta}.$$

Platí-li H'_0 , je

$$\begin{aligned} N_{11} &\sim Bi\left(\frac{1}{2} + \theta, n_1\right), & N_{12} &= n_1 - N_{11}, \\ N_{21} &\sim Bi\left(\frac{1}{2} + \theta, n_2\right), & N_{22} &= n_2 - N_{21}. \end{aligned}$$

Proto

$$\begin{aligned}
 V &= \left(\frac{1}{\frac{1}{2} + \theta} + \frac{1}{\frac{1}{2} - \theta} \right) N_{11} - \left(\frac{1}{\frac{1}{2} + \theta} + \frac{1}{\frac{1}{2} - \theta} \right) N_{21} \\
 &\quad - \frac{n_1}{\frac{1}{2} - \theta} + \frac{n_2}{\frac{1}{2} - \theta}, \\
 \text{var } V &= \left(\frac{1}{\frac{1}{2} + \theta} + \frac{1}{\frac{1}{2} - \theta} \right)^2 n_1 \left(\frac{1}{2} + \theta \right) \left(\frac{1}{2} - \theta \right) \\
 &\quad + \left(\frac{1}{\frac{1}{2} + \theta} + \frac{1}{\frac{1}{2} - \theta} \right)^2 n_2 \left(\frac{1}{2} + \theta \right) \left(\frac{1}{2} - \theta \right) \\
 &= \frac{n}{\left(\frac{1}{2} + \theta \right) \left(\frac{1}{2} - \theta \right)}.
 \end{aligned}$$

Parametr θ odhadneme z věrohodnostní rovnice $\frac{\partial L}{\partial \theta} = 0$, což při $\phi = 0$ dává

$$\frac{N_{11}}{\frac{1}{2} + \theta} - \frac{N_{12}}{\frac{1}{2} - \theta} + \frac{N_{21}}{\frac{1}{2} + \theta} - \frac{N_{22}}{\frac{1}{2} - \theta} = 0.$$

Po vyřešení se pro θ dostane maximálně věrohodný odhad

$$\hat{\theta} = \frac{N_{11} + N_{21} - N_{12} - N_{22}}{2n}.$$

Proto

$$\frac{1}{2} + \hat{\theta} = \frac{N_{11} + N_{21}}{n}, \quad \frac{1}{2} - \hat{\theta} = \frac{N_{12} + N_{22}}{n}.$$

Testová statistika pro H'_0 je

$$V^* = \frac{\frac{N_{11} - N_{21}}{N_{11} + N_{21}} - \frac{N_{12} - N_{22}}{N_{12} + N_{22}}}{\sqrt{\frac{\frac{n}{N_{11} + N_{21}} \frac{n}{N_{12} + N_{22}}}{n}}} = \frac{2(N_{11}N_{22} - N_{12}N_{21})}{\sqrt{n(N_{11} + N_{21})(N_{12} + N_{22})}}.$$

V případě $|V^*| \geq u\left(\frac{\alpha}{2}\right)$ zamítáme H'_0 na hladině, která je asymptoticky rovna α . Analogicky lze konstruovat i jednostranné testy.

6c. Numerická ilustrace.

Po běžeckých závodech dostalo 60 atletů nápoje A a B. Dá se očekávat, že po závodě budou atleti žízniví a že víc ocení první podaný nápoj tišící jejich žízeň. Nelze tedy vyloučit efekt pořadí. Výsledky průzkumu jsou v tab. 2.

Tab. 2

| | Počet osob preferujících A | Počet osob preferujících B | Celkem |
|-----------|----------------------------|----------------------------|--------|
| Nejdřív A | 27 | 3 | 30 |
| Nejdřív B | 10 | 20 | 30 |
| Celkem | 37 | 23 | 60 |

Po dosažení vyjde

$$U^* = \frac{2(27 \times 10 - 3 \times 20)}{\sqrt{60(27 + 20)(3 + 10)}} = 2.19,$$
$$V^* = \frac{2(27 \times 20 - 3 \times 10)}{\sqrt{60 \times 37 \times 23}} = 4.51.$$

V obou případech je překročena kritická hodnota $u(0.025) = 1.96$. Je tedy prokázán jak rozdíl v preferenci nápojů, tak efekt pořadí. Kdyby se aplikoval na data jen znaménkový test, hodnotilo by se jen to, že 37 atletů preferuje A a 23 atletů preferuje B. Podle vzorce (7.10) na str. 81 knihy Anděl (1998) by se vypočetla hodnota

$$\frac{2 \times 37 - 60}{\sqrt{60}} = 1.81.$$

Absolutní hodnota z tohoto čísla nepřekračuje kritickou hodnotu 1.96, takže samotný znaménkový test by efekt ošetření neprokázal.

7. McNEMARŮV TEST

Nechť četnosti n_{ij} ve čtyřpolní kontingenční tabulce (viz tab. 3) mají multinomické rozdělení s pravděpodobnostmi p_{ij} uvedenými v tab. 4.

Tab. 3

| | | |
|----------|----------|----------|
| n_{11} | n_{12} | $n_{1.}$ |
| n_{21} | n_{22} | $n_{2.}$ |
| $n_{.1}$ | $n_{.2}$ | n |

Tab. 4

| | | |
|----------|----------|----------|
| p_{11} | p_{12} | $p_{1.}$ |
| p_{21} | p_{22} | $p_{2.}$ |
| $p_{.1}$ | $p_{.2}$ | 1 |

Hypotéza $H'_0 : p_{1.} = p_{.1}$ je ekvivalentní s hypotézou $H''_0 : p_{12} = p_{21}$. Provedeme reparametrizaci modelu. Položíme $p_{12} = p + \delta$, $p_{21} = p - \delta$ a třídy (1,1) a (2,2) sdružíme v jedinou. Náhodný vektor $(n_{12}, n_{21}, n_{11} + n_{22})'$ má pak trinomické rozdělení s parametry $(n; p + \delta, p - \delta, 1 - 2p)$. Hypotézy H'_0 i H''_0 jsou ekvivalentní s $H_0 : \delta = 0$. Věrohodnostní funkce je

$$l(p, \delta) = \frac{n!}{n_{12}! n_{21}! (n_{11} + n_{22})!} (p + \delta)^{n_{12}} (p - \delta)^{n_{21}} (1 - 2p)^{n_{11} + n_{22}}$$

a logaritmická věrohodnostní funkce je

$$L(p, \delta) = \ln l(p, \delta) = \text{const.} + n_{12} \ln(p + \delta) + n_{21} \ln(p - \delta) + (n_{11} + n_{22}) \ln(1 - 2p).$$

Odtud dostaneme

$$\begin{aligned} \frac{\partial L(p, \delta)}{\partial \delta} &= \frac{n_{12}}{p + \delta} - \frac{n_{21}}{p - \delta}, \\ \frac{\partial L(p, \delta)}{\partial p} &= \frac{n_{12}}{p + \delta} + \frac{n_{21}}{p - \delta} - \frac{2(n_{11} + n_{22})}{1 - 2p}. \end{aligned}$$

K testu H_0 se využije veličina

$$S = \left. \frac{\partial L(p, \delta)}{\partial \delta} \right|_{\delta=0} = \frac{n_{12} - n_{21}}{p}.$$

Její rozptyl je za platnosti H_0 roven

$$\begin{aligned} \text{var } S &= \frac{1}{p^2} [\text{var } n_{12} - 2\text{cov}(n_{12}, n_{21}) + \text{var } n_{21}] \\ &= \frac{1}{p^2} [np(1-p) + 2np^2 + np(1-p)] = \frac{2n}{p}. \end{aligned}$$

Alternativně bychom při výpočtu tohoto rozptylu mohli postupovat také tak, že bychom nejprve vypočetli

$$\frac{\partial^2 L(p, \delta)}{\partial \delta^2} = -\frac{n_{12}}{(p + \delta)^2} - \frac{n_{21}}{(p - \delta)^2}.$$

Odtud bychom dostali Fisherovu míru informace o parametru δ

$$-E \frac{\partial^2 L(p, \delta)}{\partial \delta^2} = \frac{n(p + \delta)}{(p + \delta)^2} + \frac{n(p - \delta)}{(p - \delta)^2},$$

což při $\delta = 0$ dává rovněž hodnotu $2n/p$.

Parametr p odhadneme z věrohodnostní rovnice $\partial L(p, \delta)/\partial p = 0$, která za platnosti H_0 zní

$$\frac{n_{12}}{p} + \frac{n_{21}}{p} - \frac{2(n_{11} + n_{22})}{1 - 2p} = 0.$$

Její řešení dostaneme maximálně věrohodný odhad \hat{p} parametru p ve tvaru

$$\hat{p} = \frac{n_{12} + n_{21}}{2n}.$$

Jako testovou statistiku nemůžeme použít rovnou $S/\sqrt{\text{var } S}$, protože obsahuje neznámý parametr p . Nahradíme-li ho maximálně věrohodným odhadem, dostaneme statistiku

$$MN = \frac{n_{12} - n_{21}}{\hat{p}} \bigg/ \sqrt{\frac{2n}{\hat{p}}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}},$$

což je McNemarova statistika mající za platnosti H_0 asymptoticky rozdělení $N(0, 1)$.

Poděkování. Práce byla vytvořena v rámci grantu GAČR 201/97/1176 a výzkumného záměru CEZ:J13/98:113200008.

LITERATURA

- Anděl J. (1978): *Matematická statistika*. SNTL/ALFA, Praha.
 Anděl J. (1998): *Statistické metody*. Matfyzpress, Praha. (2. vyd.)
 Judge G. G., Hill R. C., Griffiths W. E., Lütkepohl H., Lee T.-Ch. (1988): *Introduction to the Theory and Practice of Econometrics*. Wiley, New York. (2. vyd.)
 Rayner J. C. W., Best D. J. (1997): How order affects the sign test. *Biometrics* **53**, 1416–1421.

UŽITÍ COXOVA MODELU PŘI ANALÝZE PROGNOSTICKÝCH FAKTORŮ DIABETICKÉ RETINOPATIE

EVA JAROŠOVÁ, MAREK MALÝ

Abstrakt. Na základě dat z pražského diabetologického registru je u diabetiků II. typu pomocí Coxova regresního modelu modelována doba do vzniku retinopatie a jsou analyzovány prognostické faktory, které mohou s výskytem této pozdní orgánové komplikace diabetu souviset. Z výsledků prezentovaných modelů vyplývá, že k statisticky významnému zvýšení rizika retinopatie dochází zejména u pacientů, kteří museli být hned po záchytu léčeni inzulinem nebo perorálními antidiabetiky, ale lze je též pozorovat u žen a u pacientů zachycených ve vyšším věku. Vliv hypertenze a body-mass indexu nebyl prokázán. Dále jsou diskutovány otázky související s porušením předpokladů užitého modelu. Práce je rozšířeným zněním přednášky uvedené na Českobudějovických statistických dnech v září 1998.

1. ÚVOD

Diabetes mellitus (DM) je jedním ze závažných onemocnění se zdravotními, sociálními i ekonomickými důsledky. Počet diabetiků evidovaných v České republice se neustále zvyšuje. Zatímco v roce 1975 bylo evidováno 234 071 nemocných, v roce 1986 to bylo již 406 919 osob a k 31.12.1996 bylo registrováno 582 426 diabetiků ([19]). Výskyt cukrovky je poněkud vyšší u žen, které tvoří více než 55 % všech nemocných. Přes velké pokroky v léčbě diabetu se stále jedná o onemocnění, k jehož zvládnutí je nutná trvalá léčba a jehož závažnost je znásobena průvodními zdravotními komplikacemi, zejména očními, ledvinovými a nervovými a aterosklerotickým onemocněním malých i velkých tepen.

Aby bylo možno lépe předcházet vzniku komplikací, je nutné podrobně analyzovat souvislosti jejich vzniku s mnoha charakteristikami a faktory. Jako jeden z primárních zdrojů dat mohou posloužit údaje z diabetologických porad. Data z pražských ordinací jsou dostupná díky činnosti regionálního diabetologického registru, založeného v roce 1989 [14]. Předkládaná práce je založena na analýze souboru diabetiků, kteří byli nebo jsou pacienti jedné konkrétní běžné diabetologické poradny. Hlavním cílem práce bylo zjistit, zda

Date: VŠE Praha, SZÚ Praha.

REGIONÁLNÍ DIABETOLOGICKÝ REGISTR

Příjmení: _____ Jméno: _____ Ordinační Rod. číslo: _____
 Ulice: _____ Město: _____ PSC: _____
 Datum záchytu: _____ Výška: _____ cm Váha: _____ kg
 Typ DM: 1. I 2. II 3. PGT 4. GDM 5. sekund DM
 Způsob manifestace: 1. náhod. výskyt 2. klin. symptomy 3. koma 9. neznámo
 Rodinná anamnéza: 1. prarodiče 2. rodiče 3. sourozenci 4. děti 9. nezn.

Terapie: základní změny (5 posledních terapií) PAD:
 rok změny: []
 PAD: 1. Dirastan
 2. Glibenklamid
 3. Diamicon
 4. Minidab
 5. Gluconom
 6. Amaryl
 BISULANIDY: 1. Meformin 2. Nejjou 3. Bufornin
 INZULIN: 1. 1x denně 2. 2x denně 3. IT 4. Pumpa

Retinopatie: od roku []
 neovaskulární proliferativní slepota laser vitrektomie katarakta
Nefropatie: od roku []
 proteinurie renální insuf. (do 200) renální insuf. (nad 200) nefro. dispenzace 1. Strahov 2. IKEM 3. obvod
Selfmonitoring: od roku []
 moče
Neuropatie: od roku []
 Hypertenze ISCH I.M. CMP ICHDK
 amputace DK L [] [] P [] []
Insulinové pero: []
 1. Aro 2. NOVO 3. BD 4. Ostipen 5. Jine

Vyšetření:
 MA rok []
 náleží []
 GLP rok []
 náleží []
 HbA1c rok []
 náleží []
 1. náleží v normě 2. náleží středně zvýšen 3. náleží značně zvýšen

OBRÁZEK 1. Dotazník pražského diabetologického registru

výsledky analýzy dat získaných z diabetologického registru odpovídají v literatuře uváděným tvrzením ohledně prognostických faktorů a střední doby do vzniku komplikací. U dostupných faktorů tedy bylo třeba posoudit, které z nich mají vliv na dobu do vzniku prvních potíží a zjistit, jak které ukazatele ovlivňují riziko, že se u pacienta projeví potíže způsobené diabetem.

Představu o tom, jaké údaje jsou o diabetických shromažďovány, poskytuje dotazník regionálního diabetologického registru (obr. 1). Dotazník lze rozdělit do několika částí. V první z nich jsou uvedeny základní charakteristiky pacienta (jednoznačně identifikovaného rodným číslem) a jeho onemocnění. Kromě typu diabetu, způsobu manifestace a rodinné anamnézy lze odvodit věk při záchytu diabetu. Z hlediska typu diabetu se rozlišuje několik skupin, především inzulin-dependentní diabetes I. typu, inzulin-nondependentní diabetes II. typu a sekundární DM. V registru jsou též sledováni nemocní s porušenou glukózovou tolerancí (PGT). Diabetes I. typu se zpravidla projevuje v dětském či mladistvém věku a je nutné jej léčit inzulinem, zatímco diabetes II. typu je poruchou charakteristickou pro starší věkové třídy. Mechanismus vzniku obou typů onemocnění je podstatně odlišný: u typu I se jedná hlavně o imunitně zprostředkovanou selektivní destrukci buněk produkujících inzulin, typ II je heterogenní skupinou poruch (často spojenou s obezitou), které

jsou výsledkem jak snížené sekrece inzulínu po podání glukózy, tak snížení účinnosti inzulínu ([2]).

V druhé části dotazníku je zaznamenána terapie; kromě původní terapie nasazené po záchytu onemocnění lze uvést maximálně 5 posledních skladeb terapie. Lehčí stádia onemocnění mohou být léčena pouze dietou, těžší pak různými dávkami tablet, tzv. perorálních antidiabetik (PAD), inzulínu či jejich kombinace. Existují dvě hlavní skupiny PAD: deriváty sulfonylmočoviny a biguanidy.

K hlavním problémům způsobeným diabetem patří (viz třetí část dotazníku) komplikace oční (retinopatie) a ledvinové (nefropatie), neuropatie a diabetické koma. S vývojem nemoci může souviset také hypertenze, ischemická choroba srdce (ISCH), ischemická choroba dolních končetin (ICHDK) a další onemocnění, která se mohou projevit před záchytem diabetu i po něm. Retinopatie je klasifikována do tří uspořádaných kategorií jako prostá (neproliferativní) a proliferativní retinopatie a terminální stadium bezprostředně ohrožující pacienta slepotou. Obdobně je nefropatie klasifikována jako proteinurie a dva stupně renální insuficience (rozlišené podle hladiny kreatininu v krvi).

Poslední část dotazníku je určená pro zaznamenání nálezů při jednotlivých vyšetřeních pacienta. Tyto údaje nejsou bohužel v databázi prakticky k dispozici (nejsou vkládány do počítače zpravidla z důvodu nedostatečné pracovní kapacity). Do statistických modelů proto nelze začlenit údaje o stupni kompenzace diabetu, který je významným rizikovým faktorem s velkým praktickým dosahem. Např. zahrnutí periodických měření hladiny glykovaného hemoglobinu, z níž lze odhadovat kontrolu glykémie v posledních týdnech, by nepochybně výrazně zpřesnilo výsledky analýzy.

Studovaný soubor obsahuje celkem 3107 pacientů, z nichž u naprosté většiny (2876, tj. 92.6 %) byl při záchytu zjištěn DM II. typu, u 38 (1.2 %) DM I. typu, u 14 (0.5 %) sekundární DM a u 179 (5.8 %) PGT. V celostátním měřítku tvoří I. typ 6.6 % evidovaných případů DM, II. typ 92.1 % a sekundární diabetes 1.3 % ([19]).

Jelikož se v daném případě nejedná o specializovanou ordinaci, jsou v ní některé specifické populační skupiny (jako mladiství či silně riziková jedinci) zastoupeny méně než by odpovídalo celostátním poměrům. V důsledku toho je také dotčen počet léčených diabetiků I. typu. U diabetu II. typu lze zřejmě daný soubor do značné míry považovat z hlediska zkoumaných znaků za náhodný výběr z širší populace diabetiků.

Rozbor komplikací u diabetu I. typu byl pro celý pražský registr proveden v [16]. Dále se omezíme na nejrozsáhlejší skupinu - diabetes II. typu (2876 osob, z toho 1344 mužů, tj. 46.7 %). Soustředíme se především na studium nejčastější komplikace, diabetické retinopatie, prostřednictvím doby od záchytu

| | abs. | rel. |
|------------------|------|--------|
| Retinopatie | 264 | 9.2 % |
| Nefropatie | 124 | 4.3 % |
| Neuropatie | 180 | 6.3 % |
| dieta | 1915 | 66.6 % |
| PAD | 838 | 29.1 % |
| Inzulín | 121 | 4.2 % |
| PAD + inzulín | 2 | 0.1 % |
| Rodinná anamnéza | 930 | 31.9 % |

TABULKA 1. Počty komplikací a typ léčby u diabetiků II. typu

DM do vzniku prvních retinopatických obtíží (klasifikovaných do kterékoli ze tří kategorií). Na základě dotazníku lze tuto dobu stanovit pouze s přesností na roky. Proměnné charakterizující léčbu budou ve shodě s literaturou ([12]) uvažovány pro stav po záchytu nemoci (první nasazená léčba) s tím, že později je možno do zpracování zahrnout informace z dalších časových okamžiků. Tab. 1 udává, u kolika pacientů se objevily potíže, u kolika byla nutná jiná léčba než dieta (tablety a inzulín samostatně, kombinace) a u kolika byla známa pozitivní rodinná anamnéza. Současný výskyt retinopatie a nefropatie byl zaznamenán u 60 pacientů, přičemž u 41 z nich byla jako první zjištěna retinopatie. Celostátně je pro všechny typy diabetu udáván výskyt retinopatie u 11.0 % a nefropatie u 6.4 % pacientů ([19]). Léčba aktuálního stavu pouze dietou postačuje ve 40.2 % případů, což je samozřejmě v důsledku vývoje nemoci podstatně méně než v tabulce uvedený stav při záchytu. Data byla zpracována statistickým programem SPSS [13].

Pro základní analýzu byly reálně k dispozici následující proměnné:

| | |
|-----------------|---|
| POHL: | pohlaví (1=muž, 2=žena) |
| VEK: | věk při záchytu DM |
| BMI: | body-mass index |
| RA: | rodinná anamnéza (0=neg., 1=poz.), |
| TRET,TNEF,TNEU: | doba do události (retinopatie, nefropatie, neuropatie) |
| IRET,INEF,INEU: | odpovídající indikátory události (0=cenzorováno, 1=pozorována událost) |
| IRETNEF: | indikátor, zda retinopatie předchází nefropatii (0=ne, 1=ano) [u osob s oběma typy komplikací] |
| TERP,TERI: | indikátory terapie: PAD, inzulín (0=ne, 1=ano) |
| TLAK: | hypertenze při záchytu DM (0=ne, 1=ano). |

2. STATISTICKÉ METODY

Postupy pro analýzu přežívání se zabývá mnoho monografií; k dnes již klasickým patří [10], [6] a [11]. Typickým rysem dat popisujících doby do událostí, které charakterizují průběh choroby, (tzv. doby do selhání) je, že jsou (často silně) cenzorována. Je tomu tak i s daty diabetologického registru, protože potíže nastávají jen u některých diabetiků. Ze statistického pohledu je vznik obtíží chápán jako událost a případy bez manifestace komplikací jsou cenzorovány buď úmrtím pacienta (zemřelí tvoří asi 29 % diabetiků II. typu v souboru) nebo ukončením sledování (září 1998). U každého pacienta tedy známe buď dobu do události nebo dobu do ukončení sledování (časový cenzor) a navíc máme k dispozici informaci, který časový údaj byl pozorován (tj. zda došlo k události). Doba do události je vlastně časovou vzdáleností mezi dobou diagnózy diabetu a vznikem komplikací. Dalším charakteristickým rysem daných dat je, že pacienti sledování po různě dlouhou dobu.

Vhodnou základní statistickou popisnou technikou je proto metoda životních a úmrtnostních tabulek (life tables, [20]) vycházející z požadavku, aby průběh procesu přežívání byl v čase neměnný a aby se cenzorovaná pozorování nelišila charakterem od necenzorovaných. Metoda vychází z myšlenky rozkladu doby sledování po vstupu do registru na menší dílčí intervaly, v našem případě roky. V každém časovém intervalu přispívají všichni jedinci, kteří byli pozorováni alespoň tak dlouhou dobu, k výpočtu pravděpodobnosti výskytu události, tj. komplikací diabetu. Pravděpodobnosti vypočtené pro jednotlivé intervaly pak slouží k odhadu celkové pravděpodobnosti výskytu události v různých časových okamžicích. Intenzita (hazard rate) je podmíněná pravděpodobnost, že jedinec, který se počátku určitého časového intervalu dožil bez komplikací, bude jimi postižen právě v tomto intervalu. Počet osob v riziku je počítán jako počet osob vstupujících do období mínus polovina počtu těch, jejichž sledování v tomto období skončilo (tj. byli cenzorováni). Podíl osob přežívajících v daném intervalu bez komplikací pak slouží jako jeden ze způsobů odhadu funkce přežití $S(t)$, udávající pravděpodobnost dožití do času t bez komplikací.

Pro zkoumání vlivu rizikových faktorů na dobu do vzniku komplikací jsme zvolili jeden z nejčastěji užívaných přístupů, Coxův regresní model [4]

$$h_i(t) = h_0(t) \exp(\beta' x_i),$$

kde $h_i(t)$ je intenzita výskytu události u i -tého pacienta (riziková funkce),
 $h_0(t)$ základní intenzita,
 β vektor parametrů regresního modelu,
 \mathbf{x}_i vektor hodnot vysvětlujících proměnných u i -tého pacienta.

Tento model (poprvé navržený D.R.Coxem [5]) je založen na předpokladu proporcionálního rizika, ale neklade požadavky na tvar distribuční funkce dob přežívání, a proto bývá označován jako semiparametrický. V české lékařské literatuře na něj upozornil již Hausner a spol. [8].

Jeho parametry se odhadují metodou maximální věrohodnosti, jako věrohodnostní funkce se v případě, kdy pro necenzorované doby do vzniku komplikací platí $t_{(1)} < t_{(2)} < \dots < t_{(r)}$, použije funkce

$$L(\beta) = \prod_{i=1}^n \left[\frac{\exp(\beta' \mathbf{x}_i)}{\sum_{l \in R(t_i)} \exp(\beta' \mathbf{x}_l)} \right]^{\delta_i},$$

kde $R(t_i)$ je množina pacientů vystavených riziku v okamžiku t_i , tj. pacientů, u nichž do tohoto okamžiku komplikace nenastaly,

\mathbf{x}_i je vektor hodnot vysvětlujících proměnných pro pacienta, u něhož v okamžiku t_i vznikly komplikace,

δ_i je indikátorová proměnná nabývající hodnoty 0, je-li údaj o době do vzniku komplikací u i -tého pacienta cenzorován, a hodnoty 1 jinak.

Existují-li shody, použije se aproximace, např. podle Breslowa [3]. Maximalizace logaritmu věrohodnostní funkce se provádí obvykle Newtonovou-Raphsonovou metodou.

Při rozhodování o tom, které proměnné v modelu ponechat, se vedle Waldových testů o jednotlivých parametrech se statistikou

$$\frac{b_j}{s(b_j)},$$

jejichž závěry jsou pro vzájemnou závislost odhadů regresních parametrů b_j obtížně interpretovatelné, používá zejména test věrohodnostním poměrem se statistikou

$$-2 \ln \frac{\hat{L}(\mathbf{1})}{\hat{L}(\mathbf{2})},$$

přičemž $\mathbf{1}$ označuje širší model s proměnnými $x_1, x_2, \dots, x_p, x_{p+1}, \dots, x_{p+q}$ a $\mathbf{2}$ podmodel s proměnnými x_1, x_2, \dots, x_p . Testová statistika má pak při platnosti nulové hypotézy $H_0 : \beta_{p+1} = \dots = \beta_{p+q} = 0$ chí-kvadrát rozdělení s q stupni volnosti.

Kromě toho lze použít automatických procedur pro výběr proměnných, jako je metoda forward, backward, stepwise aj. Při zařazování interakcí se doporučuje dodržovat princip hierarchie, tj. obsahuje-li model interakci, musí obsahovat příslušné hlavní efekty.

Pro posouzení vhodnosti modelu se provádí diagnostika reziduí; zde se omezíme na jejich dva typy. Coxova-Snellova rezidua jsou pro i -tého jedince dána vztahem

$$r_i^C = \hat{H}_i(t_i) = \hat{H}_0(t_i) \exp(\mathbf{b}'\mathbf{x}_i) = -\ln \hat{S}_i(t_i),$$

kde $\hat{H}_0(t_i)$ a $\hat{S}_0(t_i)$ jsou odhady kumulativní základní intenzity a funkce přežití v okamžiku t_i .

Lze odvodit, že tato rezidua mají v případě vhodného modelu exponenciální rozdělení s jednotkovou střední hodnotou. Ověření se provádí pomocí grafu funkce přežití aplikované na tato rezidua. Stupnice na obou osách jsou konstruovány tak, aby byl při vhodně zvoleném modelu pozorován lineární průběh.

Jiný typ reziduí, parciální neboli skórová rezidua ([17]), se určuje pro každou proměnnou zvlášť a je dán vztahem

$$r_{ki}^S = \delta_i(x_{ki} - a_{ki}),$$

kde

$$a_{ki} = \frac{\sum_{l \in R(t_i)} x_{kl} \exp(\mathbf{b}'\mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\mathbf{b}'\mathbf{x}_l)}.$$

Tato rezidua se vynášejí v závislosti na době do události a v případě oprávněnosti příslušné proměnné v modelu by seskupení bodů v grafu mělo být zcela náhodné a symetrické kolem nuly, jedná-li se však o dichotomickou vysvětlující proměnnou, leží body ve dvou vodorovných pásech symetricky vůči vodorovné ose.

Základním předpokladem použití Coxova modelu je proporcionalita rizika, což znamená, že lineární složka (a tedy všechny vysvětlující proměnné) musí být v čase konstantní. K prvotnímu ověření tohoto předpokladu jsou vhodné grafické metody [9]. U kategoriálních vysvětlujících proměnných lze např. využít grafu závislosti transformované odhadnuté funkce přežití ($\ln(-\ln \hat{S}(t))$) na době do události pro každou kategorii vysvětlující proměnné. Při splnění předpokladu budou průběhy přibližně rovnoběžné. Jinak lze předpoklad proporcionality ověřit prostřednictvím zařazení tzv. časově závislé proměnné; tato

| Věk | BMI | | | Celkem |
|--------|-----|-------|-----|--------|
| | -25 | 25-30 | 30+ | |
| -45 | 11 | 52 | 39 | 102 |
| 45-60 | 59 | 264 | 233 | 556 |
| 60-75 | 71 | 303 | 176 | 550 |
| 75+ | 38 | 73 | 25 | 136 |
| Celkem | 179 | 692 | 473 | 1344 |

(a) muži

| Věk | BMI | | | Celkem |
|--------|-----|-------|-----|--------|
| | -25 | 25-30 | 30+ | |
| - 45 | 11 | 17 | 50 | 78 |
| 45-60 | 36 | 132 | 248 | 416 |
| 60-75 | 78 | 283 | 425 | 786 |
| 75+ | 40 | 112 | 100 | 252 |
| Celkem | 165 | 544 | 823 | 1532 |

(b) ženy

TABULKA 2. Rozložení diabetiků II. typu podle pohlaví, věku a BMI

proměnná se modeluje pomocí interakce ověřované proměnné s dobou do události a testuje se regresní parametr u této interakce. Statisticky významný výsledek testu naznačuje porušení předpokladu proporcionálního rizika.

3. VÝSLEDKY

Průměrný věk (a směrodatná odchylka) mužů a žen při manifestaci diabetu II. typu je ve studované části souboru 60.4 (11.1), resp. 64.4 (11.1) let. Vztah věkových a výško-váhových charakteristik je uveden v tab. 2 (u několika pacientů chybí příslušné údaje). Body-mass index (BMI) je poměrem hmotnosti vyjádřené v kilogramech ku druhé mocnině výšky (v metrech). Hodnoty 25, resp. 30 (u mužů) a 24, resp. 29 (u žen) bývají chápány jako hranice nadváhy, resp. obezity. V soulase s [15] nacházíme mezi diabetiky II. typu značné procento obézních (asi 35 % mužů a 54 % žen) včetně několika monstrózně obézních (BMI přes 39). Průměrná délka trvání diabetu v době analýzy je u daných pacientů 9.6 let a prakticky se neliší v závislosti na pohlaví. U 32.2 % nemocných je délka trvání do 5 let, zatímco u 8.2 % činí 20 let či více.

Základní informaci o délce přežívání bez komplikací u diabetiků II. typu ve sledovaném souboru přináší tab. 3. Odhad funkce přežití v jednotlivých letech po záchytu byl proveden metodou úmrtnostních tabulek. Nejdůležitější ukazatele, intenzita a kumulativní podíl přežívajících, jsou doplněny příslušnými středními chybami (SE).

Medián doby přežití je čas, ve kterém je hodnota kumulativní funkce přežití rovna 0.5. Je to tedy čas, v němž je očekáváno, že u poloviny sledovaných jedinců dojde k události. Pro tento soubor diabetiků je jeho hodnota 31.66 let.

| Zač. čas. intervalu (roky) | Počet při vstupu do int. | Počet cenzovaných v int. | Počet osob v riziku | Počet událostí | Podíl událostí | Podíl přeživších | Kumul. podíl přeživších | SE pro kum. podíl | Intenzita | SE pro intenzitu |
|----------------------------|--------------------------|--------------------------|---------------------|----------------|----------------|------------------|-------------------------|-------------------|-----------|------------------|
| 0 | 2876 | 126 | 2813.0 | 22 | 0.0078 | 0.9922 | 0.9922 | 0.0017 | 0.0079 | 0.0017 |
| 1 | 2728 | 134 | 2661.0 | 7 | 0.0026 | 0.9974 | 0.9896 | 0.0019 | 0.0026 | 0.0010 |
| 2 | 2587 | 168 | 2503.0 | 16 | 0.0064 | 0.9936 | 0.9832 | 0.0025 | 0.0064 | 0.0016 |
| 3 | 2403 | 167 | 2319.5 | 14 | 0.0060 | 0.9940 | 0.9773 | 0.0029 | 0.0061 | 0.0016 |
| 4 | 2222 | 182 | 2131.0 | 12 | 0.0056 | 0.9944 | 0.9718 | 0.0033 | 0.0056 | 0.0016 |
| 5 | 2028 | 142 | 1957.0 | 11 | 0.0056 | 0.9944 | 0.9663 | 0.0037 | 0.0056 | 0.0017 |
| 6 | 1875 | 166 | 1792.0 | 8 | 0.0045 | 0.9955 | 0.9620 | 0.0040 | 0.0045 | 0.0016 |
| 7 | 1701 | 139 | 1631.5 | 9 | 0.0055 | 0.9945 | 0.9567 | 0.0043 | 0.0055 | 0.0018 |
| 8 | 1553 | 152 | 1477.0 | 10 | 0.0068 | 0.9932 | 0.9502 | 0.0048 | 0.0068 | 0.0021 |
| 9 | 1391 | 142 | 1320.0 | 14 | 0.0106 | 0.9894 | 0.9402 | 0.0054 | 0.0107 | 0.0028 |
| 10 | 1235 | 197 | 1136.5 | 17 | 0.0150 | 0.9850 | 0.9261 | 0.0063 | 0.0151 | 0.0037 |
| 11 | 1021 | 130 | 956.0 | 15 | 0.0157 | 0.9843 | 0.9116 | 0.0072 | 0.0158 | 0.0041 |
| 12 | 876 | 115 | 818.5 | 17 | 0.0208 | 0.9792 | 0.8926 | 0.0084 | 0.0210 | 0.0051 |
| 13 | 744 | 93 | 697.5 | 13 | 0.0186 | 0.9814 | 0.8760 | 0.0094 | 0.0188 | 0.0052 |
| 14 | 638 | 78 | 599.0 | 16 | 0.0267 | 0.9733 | 0.8526 | 0.0109 | 0.0271 | 0.0068 |
| 15 | 544 | 95 | 496.5 | 13 | 0.0262 | 0.9738 | 0.8303 | 0.0122 | 0.0265 | 0.0074 |
| 16 | 436 | 56 | 408.0 | 7 | 0.0172 | 0.9828 | 0.8160 | 0.0131 | 0.0173 | 0.0065 |
| 17 | 373 | 64 | 341.0 | 15 | 0.0440 | 0.9560 | 0.7801 | 0.0155 | 0.0450 | 0.0116 |
| 18 | 294 | 43 | 272.5 | 6 | 0.0220 | 0.9780 | 0.7630 | 0.0167 | 0.0223 | 0.0091 |
| 19 | 245 | 40 | 225.0 | 7 | 0.0311 | 0.9689 | 0.7392 | 0.0184 | 0.0316 | 0.0119 |
| 20 | 198 | 26 | 185.0 | 2 | 0.0108 | 0.9892 | 0.7312 | 0.0190 | 0.0109 | 0.0077 |
| 21 | 170 | 21 | 159.5 | 5 | 0.0313 | 0.9687 | 0.7083 | 0.0210 | 0.0318 | 0.0142 |
| 22 | 144 | 32 | 128.0 | 2 | 0.0156 | 0.9844 | 0.6972 | 0.0221 | 0.0157 | 0.0111 |
| 23 | 110 | 18 | 101.0 | 0 | 0.0000 | 1.0000 | 0.6972 | 0.0221 | 0.0000 | 0.0000 |
| 24 | 92 | 23 | 80.5 | 1 | 0.0124 | 0.9876 | 0.6886 | 0.0235 | 0.0125 | 0.0125 |
| 25 | 68 | 19 | 58.5 | 0 | 0.0000 | 1.0000 | 0.6886 | 0.0235 | 0.0000 | 0.0000 |
| 26 | 49 | 13 | 42.5 | 2 | 0.0471 | 0.9529 | 0.6562 | 0.0316 | 0.0482 | 0.0341 |
| 27 | 34 | 12 | 28.0 | 1 | 0.0357 | 0.9643 | 0.6327 | 0.0382 | 0.0364 | 0.0364 |
| 28 | 21 | 5 | 18.5 | 1 | 0.0541 | 0.9459 | 0.5985 | 0.0491 | 0.0556 | 0.0555 |
| 29 | 15 | 7 | 11.5 | 0 | 0.0000 | 1.0000 | 0.5985 | 0.0491 | 0.0000 | 0.0000 |
| 30 | 8 | 4 | 6.0 | 0 | 0.0000 | 1.0000 | 0.5985 | 0.0491 | 0.0000 | 0.0000 |
| 31 | 4 | 0 | 4.0 | 1 | 0.2500 | 0.7500 | 0.4489 | 0.1347 | 0.2857 | 0.2828 |
| 32 | 3 | 1 | 2.5 | 0 | 0.0000 | 1.0000 | 0.4489 | 0.1347 | 0.0000 | 0.0000 |
| 33 | 2 | 1 | 1.5 | 0 | 0.0000 | 1.0000 | 0.4489 | 0.1347 | 0.0000 | 0.0000 |
| 34 | 1 | 0 | 1.0 | 0 | 0.0000 | 1.0000 | 0.4489 | 0.1347 | 0.0000 | 0.0000 |
| 35 | 1 | 0 | 1.0 | 0 | 0.0000 | 1.0000 | 0.4489 | 0.1347 | 0.0000 | 0.0000 |
| 36 | 1 | 1 | 0.5 | 0 | 0.0000 | 1.0000 | 0.4489 | 0.1347 | 0.0000 | 0.0000 |

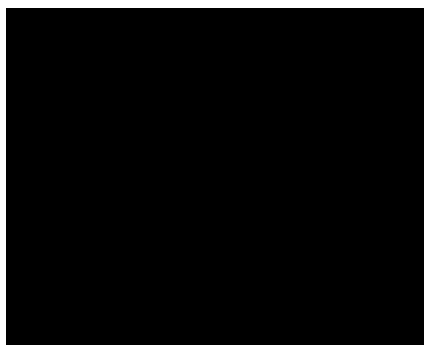
TABULKA 3. Životní tabulka pro dobu do retinopatické komplikace (DM II. typu)

| Proměnná | b | SE | Wald | st | | exp(b) | 95%CI pro exp(b) | |
|----------|---------|--------|---------|-----|--------|--------|------------------|--------|
| | | | | vol | p-hodn | | dolní | horní |
| BMI | 0.0060 | 0.0169 | 0.1256 | 1 | 0.7230 | 1.0060 | 0.9733 | 1.0399 |
| POHL | 0.4372 | 0.1386 | 9.9439 | 1 | 0.0016 | 1.5483 | 1.1799 | 2.0317 |
| RA | 0.0057 | 0.1371 | 0.0017 | 1 | 0.9668 | 1.0057 | 0.7687 | 1.3158 |
| TERI | 1.7568 | 0.2177 | 65.1221 | 1 | 0.0000 | 5.7940 | 3.7816 | 8.8774 |
| TERP | 1.0774 | 0.1426 | 57.1191 | 1 | 0.0000 | 2.9371 | 2.2211 | 3.8838 |
| TLAK | 0.0319 | 0.1463 | 0.0475 | 1 | 0.8274 | 1.0324 | 0.7750 | 1.3753 |
| VEK | -0.0185 | 0.0066 | 7.9217 | 1 | 0.0049 | 0.9817 | 0.9691 | 0.9944 |

TABULKA 4. Coxova regrese - výchozí model

| Proměnná | b | SE | Wald | st | | exp(b) | 95%CI pro exp(b) | |
|----------|---------|--------|---------|-----|--------|--------|------------------|--------|
| | | | | vol | p-hodn | | dolní | horní |
| POHL | 0.4463 | 0.1364 | 10.7048 | 1 | 0.0011 | 1.5626 | 1.1960 | 2.0414 |
| TERI | 1.7497 | 0.2163 | 65.4600 | 1 | 0.0000 | 5.7527 | 3.7650 | 8.7903 |
| TERP | 1.0800 | 0.1423 | 57.6342 | 1 | 0.0000 | 2.9447 | 2.2280 | 3.8919 |
| VEK | -0.0185 | 0.0063 | 8.6149 | 1 | 0.0033 | 0.9817 | 0.9696 | 0.9939 |

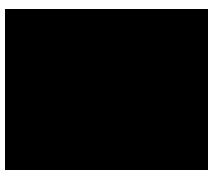
TABULKA 5. Coxova regrese - významné prognostické proměnné



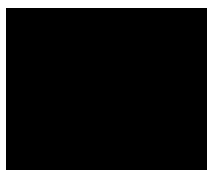
OBRÁZEK 2. Funkce přežití podle pohlaví

| Proměnná | b | SE | Wald | p-hodn | 95%CI pro exp(b) | | |
|-----------|---------|--------|---------|--------|------------------|--------|---------|
| | | | | | exp(b) | dolní | horní |
| POHL | 0.3510 | 0.1445 | 5.9018 | 0.0151 | 1.4205 | 1.0702 | 1.8856 |
| RA | 0.1338 | 0.1442 | 0.8608 | 0.3535 | 1.1431 | 0.8617 | 1.5165 |
| TERI | 1.6739 | 0.3754 | 19.8857 | 0.0000 | 5.3329 | 2.5553 | 11.1294 |
| TERP | 1.0849 | 0.1430 | 57.5577 | 0.0000 | 2.9591 | 2.2358 | 3.9164 |
| VEK | -0.0207 | 0.0066 | 9.9156 | 0.0016 | 0.9796 | 0.9671 | 0.9922 |
| POHL*TERI | 0.9808 | 0.4329 | 5.1327 | 0.0235 | 2.6667 | 1.1414 | 6.2300 |
| RA*TERI | -1.2411 | 0.4310 | 8.2906 | 0.0040 | 0.2891 | 0.1242 | 0.6728 |

TABULKA 6. Coxova regrese - model s interakcemi



OBRÁZEK
3. Graf intenzity (Coxova-Snellova rezidua)



OBRÁZEK
4. Ukázka partiálních reziduí pro spojitou proměnnou

Detailnější analýza doby přežívání bez komplikací, hlavně z hlediska vlivu prognostických faktorů, byla provedena pomocí Coxova modelu. Nejprve jsme do modelu zařadili všechny dříve zmíněné vysvětlující proměnné. Výsledky jsou uvedeny v tab. 4. Zahrnují odhady parametrů modelu včetně středních chyb a intervalů spolehlivosti pro ukazatel e^b , který má snadnější interpretaci než vlastní parametr b .

Podle dosažené hladiny významnosti (p-hodnoty) u Waldova testu jednotlivých parametrů lze soudit, že proměnné POHL, TERI, TERP a VEK jsou důležité prognostické proměnné. Tab. 5 odpovídá modelu pouze s těmito proměnnými. Hodnota $e^{0.4463} = 1.5626$ u proměnné indikuje, že u žen je riziko vzniku komplikací přibližně 1.56 krát větší než u mužů (za předpokladu, že hodnoty ostatních proměnných jsou stejné). U pacientů léčených od počátku



OBRÁZEK
5. Ukázka parciálních reziduí pro diskrétní proměnnou



OBRÁZEK
6. LML funkce podle pohlaví

| Proměnná | b | SE | Wald | p-hodn | 95%CI pro exp(b) | | |
|----------------|---------|--------|---------|--------|------------------|--------|---------|
| | | | | | exp(b) | dolní | horní |
| POHL | 0.1057 | 0.2793 | 0.1432 | 0.7051 | 1.1115 | 0.6430 | 1.9213 |
| RA | 0.1094 | 0.2863 | 0.1460 | 0.7024 | 1.1156 | 0.6365 | 1.9552 |
| TERI | 2.4257 | 0.6992 | 12.0360 | 0.0005 | 11.3106 | 2.8729 | 44.5296 |
| TERP | 1.7093 | 0.2910 | 34.5069 | 0.0000 | 5.5253 | 3.1237 | 9.7734 |
| VEK | 0.0112 | 0.0123 | 0.8271 | 0.3631 | 1.0112 | 0.9872 | 1.0359 |
| POHL*TERI | 1.0873 | 0.7613 | 2.0401 | 0.1532 | 2.9663 | 0.6672 | 13.1891 |
| RA*TERI | -1.0988 | 0.7544 | 2.1212 | 0.1453 | 0.3333 | 0.0760 | 1.4621 |
| TCOV*POHL | 0.0219 | 0.0224 | 0.9512 | 0.3294 | 1.0221 | 0.9781 | 1.0681 |
| TCOV*RA | 0.0029 | 0.0228 | 0.0165 | 0.8976 | 1.0029 | 0.9591 | 1.0488 |
| TCOV*TERP | -0.0587 | 0.0232 | 6.4127 | 0.0113 | 0.9430 | 0.9011 | 0.9868 |
| TCOV*TERI | -0.0797 | 0.0692 | 1.3254 | 0.2496 | 0.9234 | 0.8063 | 1.0576 |
| TCOV*VEK | -0.0032 | 0.0010 | 9.6112 | 0.0019 | 0.9968 | 0.9948 | 0.9988 |
| TCOV*POHL*TERI | -0.0138 | 0.0782 | 0.0312 | 0.8599 | 0.9863 | 0.8462 | 1.1496 |
| TCOV*RA*TERI | -0.0121 | 0.0784 | 0.0238 | 0.8774 | 0.9880 | 0.8473 | 1.1521 |

TABULKA 7. Coxova regrese - interakce s časově závislou proměnnou

inzulínem je riziko 5.75 krát větší než u pacientů bez inzulínu. Podíl rizik u pacientů s léčbou PAD a bez ní je 2.94.

Vliv různých faktorů na dobu do vzniku komplikací lze dokumentovat graficky. Jako ukázkou uvádíme obr. 2 s funkcí přežití při stratifikaci podle pohlaví. Riziko vzniku komplikací je pro ženy systematicky větší.

V další fázi jsme zkoumali vliv dvoufaktorových interakcí. Jako významné se ukázaly interakce POHL*TERI a RA*TERI. Odhady parametrů nového

modelu s testy a konfidenčními intervaly jsou v tab. 6. Protože se při uvažování interakce v modelu doporučuje zařadit i odpovídající hlavní efekty faktorů ([4]), obsahuje nový model znovu proměnnou RA. Hodnoty parametrů u interakci se potom interpretují v souvislosti s parametry u hlavních efektů takto: Podíl rizik u žen vůči mužům bez léčby inzulinem je $e^{0.3510} = 1.4205$, při léčbě inzulinem $e^{(0.3510+0.9808)} = 3.7879$, podíl rizik u pacientů s inzulinem vůči pacientům bez něj je u mužů $e^{1.6739} = 5.3329$, u žen $e^{(1.6739+0.9808)} = 14.2207$. Podíl rizik u pacientů s výskytem diabetu v rodině vůči pacientům, u nichž se v rodině diabetes nevyskytl, je u pacientů bez inzulinu $e^{0.1338} = 1.1431$, u pacientů s inzulinem $e^{(0.1338-1.2411)} = 0.3304$.

Ukázky grafů pro diagnostiku reziduí jsou na obr. 3 až 5. Takové grafy byly sestrojeny pro všechny proměnné v modelu a na jejich základě lze říci, že žádná ze zařazených proměnných nevykazuje zásadní negativní vliv na chování reziduí.

Z grafu 6 je možno usoudit, že proměnná POHL nevede k porušení předpokladu proporcionálního rizika. Podobný závěr lze učinit i u ostatních kategoriálních proměnných. Pro další ověření předpokladu proporcionality byly k proměnným a interakcím v modelu z tab. 6 přidány jejich interakce s proměnnou TCOV představující měnící se čas. Výsledky Waldova testu (tab. 7) naznačují, že k porušení předpokladu zřejmě dochází u proměnných VEK a TERP.

V tab. 8 jsou všechny popisované modely porovnány na základě věrohodnostní funkce pomocí hodnot statistiky $-2 \ln \hat{L}$. Z údajů pro modely v tab. 4 a tab. 5 plyne, že vypuštění proměnných BMI, TLAK a RA ze základního modelu v tab. 4 nezpůsobí významnou změnu. Naopak přidání dvou interakcí do modelu z tab. 5 statisticky významně snížilo hodnotu sledovaného ukazatele, a proto složitější model v tab. 6 lépe postihuje existující vazby.

| Model | Proměnné v modelu | $-2 \ln \hat{L}$ |
|--------|--|------------------|
| | žádná | 3124.446 |
| Tab. 4 | POHL,TERI,TERP,VEK,BMI,RA,TLAK | 3019.407 |
| Tab. 5 | POHL,TERI,TERP,VEK | 3019.602 |
| Tab. 6 | POHL,TERI,TERP,VEK,RA, POHL*TERI,RA*TERI | 3007.258 |
| Tab. 7 | POHL,TERI,TERP,VEK,RA, POHL*TERI,RA*TERI,interakce s TCOV | 2988.299 |

TABULKA 8. Porovnání modelů pomocí logaritmu věrohodnostní funkce

4. ZÁVĚR

Tato práce představuje prvotní zhodnocení daného souboru dat prostředky analýzy přežití. Je třeba mít na paměti, že se jedná o data z pouze jedné ze zhruba 30 pražských diabetologických ordinací a že analýza celého registru by nepochybně zpřesnila výsledky. Speciální pozornost by bylo možno zaměřit na již zemřelé pacienty, u nichž lze definitivně rozhodnout, zda určitými komplikacemi trpěli či nikoliv.

Výsledky naší analýzy se v některých bodech zcela neshodují s názory publikovanými v lékařské literatuře. Zřetelné jsou např. rozdíly v mediánu doby přežití bez komplikací. Často prezentovaná hodnota 10-15 let však vychází pouze z údajů pacientů, u nichž komplikace opravdu nastaly, zatímco zde uvedený výsledek bere v úvahu cenzorování. Je ovšem otázkou, která z hodnot má větší praktický význam. Vliv pohlaví na výskyt retinopatie a na riziko jejího vzniku se obecně nepovažuje za prokázaný. Nicméně analyzovaná data indikují zvýšení rizika u žen v porovnání s muži. Statisticky významný vliv proměnných TERI a TERP na dobu do komplikací prokazující větší riziko u pacientů léčených inzulinem či PAD byl pozorován i v jiných studiích a lze jej nejspíše vysvětlit horším stavem pacientů v době záchytu diabetu; způsob terapie je pouze zástupným ukazatelem stavu pacienta. V literatuře je opakovaně poukazováno na zásadní význam, který má důsledná regulace a kontrola glykémie na snížení rizika retinopatie (viz např. [7]). Data dostupná v registru však neumožnila zkoumání této souvislosti.

Podle modelu v tab. 7 s časově závislou proměnnou TCOV je riziko vzniku retinopatie větší u pacientů, u nichž se DM manifestoval později. To je v soulase s literaturou (např. [18]), neboť se sítnicový cévní systém během procesu stárnutí stává zranitelnějším. Získané výsledky však nasvědčují tomu, že se uvedené riziko v průběhu trvání nemoci mění.

Hlavním nástrojem předkládané analýzy je Coxova regrese. Data o komplikacích diabetu bývají studována i pomocí jiných přístupů, např. v [12] je ilustrováno použití logistického modelu (viz též [1]). Další zpracování dat z diabetologického registru by však především mělo zohlednit longitudinální charakter dat a odstranit tak omezení zde použitých modelů, do nichž byly kromě časových údajů zařazeny pouze proměnné s údaji zjištěnými u pacienta při záchytu diabetu či krátce po něm, tj. na počátku sledování. Na základě předchozích výsledků naznačujících závislost rizika na čase by tedy bylo vhodné využít dalších údajů z registru, týkajících se zejména změn terapie (v ojedinělých případech byl překlasifikován i typ diabetu), a použít složitější model, v němž se budou vyskytovat proměnné, jejichž hodnoty se u jednotlivých pacientů s časem mění.

Postup analogický hodnocení retinopatických obtíží lze aplikovat i v případě neuropatie, resp. nefropatie. Podrobná analýza těchto komplikací bude předmětem další práce, nicméně předběžně lze říci, že indikátory terapie (TERI i TERP) a věk při záchytu diabetu budou i zde statisticky významnými prognostickými faktory. K nim se navíc při modelování doby do nefropatie přiřazuje TLAKE a také proměnná IRETNEF, tj. indikátor, zda u pacientů s oběma typy komplikací předcházela retinopatie nefropatii.

Vzhledem ke způsobu zaznamenávání doby vzniku potíží, kdy je uveden pouze rok, lze získaná data považovat za intervalově cenzorovaná. Tento fakt byl v dosavadním zpracování ignorován. Podle [4] přicházejí v úvahu další možnosti, např. modelování ln-ln transformace doplňku pravděpodobnosti vzniku komplikací ve sledovaném období (od záchytu do určitého okamžiku) nebo modelování ln-ln transformace doplňku podmíněné pravděpodobnosti vzniku komplikací v určitém intervalu (za předpokladu, že před tímto intervalem komplikace nenastaly) pomocí zobecněného lineárního modelu. Protože SPSS neumožňuje snadné zařazení výše zmíněných časově závislých proměnných ani přímé použití ln-ln transformace, chtěli bychom při další analýze složitějších modelů využít statistického balíčku SAS.

REFERENCE

- [1] Agresti A. (1990): *Categorical Data Analysis*. J. Wiley, New York.
- [2] Berkow R. (ed.) (1996): *MERCK Manual*. Kompendium klinické medicíny. 1. české vydání. X-Egem, Praha.
- [3] Breslow N. E. (1974): Covariance analysis of censored survival data. *Biometrics* 30, 89–100.
- [4] Collett D. (1994): *Modelling Survival Data in Medical Research*. Chapman & Hall, London.
- [5] Cox D. R. (1972): Regression Models and Life-Tables. *J. of the Royal Stat. Soc.*, 34, 187–220.
- [6] Cox D. R., Oakes D. (1984): *Analysis of Survival Data*. Chapman & Hall, London.
- [7] Guillausseau P. J. et al. (1998): Glycaemic control and development of retinopathy in type 2 diabetes mellitus: a longitudinal study. *Diabetic Medicine* 15, 151–155.
- [8] Hausner P., Klein P. a kol. (1982): Využití Coxovy regresní analýzy v medicíně. *Čas. lék. čes.*, 121, č.28, 870–874.
- [9] Hess K. R. (1995): Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine* 14, 1707–1723.
- [10] Kalbfleisch J. D., Prentice R. L. (1980): *The Statistical Analysis of Failure Time Data*. J. Wiley, New York.
- [11] Lawless J. F. (1982): *Statistical Models and Methods for Lifetime Data*, J. Wiley, New York.
- [12] Lee E. T. (1992): *Statistical Methods for Survival Data Analysis*. J. Wiley, New York.
- [13] Norušis, M.J. (1993): *SPSS for Windows*. Release 6.0. SPSS Inc., Chicago.

- [14] Perušičová J., Neuwirt K. (1992): Pražský registr diabetiků - Využití počítače: základní údaje o souboru diabetiků a nemocných s porušenou glukózovou tolerancí. Čas. lék. čes. 131, 662–667.
- [15] Perušičová J., Neuwirt K. (1993): Pražský registr diabetiků - 2. část. Body mass index, rodinná anamnéza, způsob manifestace diabetes mellitus. Čas. lék. čes. 132, 74–77.
- [16] Perušičová J., Neuwirt K. (1993): Pražský registr diabetiků - 3. část. Retinopatie, nefropatie a neuropatie u diabetiků I. typu. Čas. lék. čes. 132, 489–493.
- [17] Schoenfeld, D.A. (1982): Partial residuals for the proportional hazard regression model. Biometrika 69, 239–241.
- [18] Řehák, J. (1998): Se stoupajícím věkem ohrožuje diabetiky ztráta zraku. Příloha Zdravotnických novin č.31, 7. 8. 1998.
- [19] ÚZIS (1997): Péče o nemocné cukrovkou 1996. Ústav zdravotnických informací a statistiky ČR, Praha.
- [20] Žáček A. (1984): Metody studia zdraví a nemocí v populaci. Avicenum, Praha.

| | |
|--|----|
| <i>Josef Tvrđík, Každý pátadvacátý ěech je Slovák</i> | 1 |
| <i>David J. Finney, O jazyku biometrie a jeho zneužívání</i> | 2 |
| <i>Jiří Anděl, Asymptotické testy</i> | 10 |
| <i>Eva Jarošová, Marek Malý, Užítí Coxova modelu při analýze prognostických faktorů diabetické retinopatie</i> | 22 |

Informační Bulletin ěeské statistické společnosti vychází ětyřikrát do roka v ěeském vydání.
Předseda společnosti: Ing. Zdeněk Roth, CSc., SZÚ Praha, MSP, Šrobárova 48, 100 42 Praha 10,
e-mail: marek.maly@szu.cz. ISSN 1210 – 8022
Redakce: RNDr. Gejza Dohnal, CSc., Jeronýmova 7, 130 00 Praha 3, e-mail: dohna1@fsik.cvut.cz.