ROBUST'2002, 16 – 49

AKTUÁLNÍ TÉMATA SOUČASNÉ BIOSTATISTIKY

M. BRABEC, M. MALÝ, B. PROCHÁZKA, Z. ROTH, L. TOMÁŠEK

Abstract. Biostatistics is a branch of applied statistics whose main scope lies in the development and practical use of statistical methods intended for applications in biological, medical and other healthcare-related studies. Besides general and widely used techniques, biostatistics also uses specific methodologies and approaches. These are needed for instance in design and management of clinical trials, epidemiology, demography etc. It is not only the statistical background of these techniques what matters: additional and practically important aspects arise. They involve computational, philosophical and ethical issues to name just a few. Our paper tries to review the problems which are typical of modern biostatistics. It focuses on the methods and techniques that belong to the current hot-topic list in the discipline and that are dynamically evolving at present. Several selected areas are discussed in more detail: correlated (longitudinal) data analysis, models for categorical and ordinal data, bioequivalence testing, data censoring. As an example, we will show a real life application which relates censoring and outlier detection problem.

Резюмэ. Биостатистика – это раздел прикладной статистики, направленный на применение статистических методов при решении биологической и медицинской проблематики. Наряду с общеизвестными статистическими методами для целей применяются также специфичные методы, соотвецтвующие требованиям эпидемиологии, демографии, управления клиническими экспериментами итд. Б случае специфичных методов важна не только их статистическая сущность, а также вычислительные аспекты и философские и этические вопросы, связанные на практике.

Целью нашего выступления явилось дать обзор актуальных биологических тем и уделить внимание тем общим и специфическим методам и подходам, которые в настоящее время динамично развиваются и играют ключевую роль в статистическом анализе биолого-медицинских научных исследований. Более подробно изложено моделирование коррелированных и лонгитудинальных данных, обработка порядковых и категорических данных и оценка цензурированных данных в связи с выявлением отдаленных наблюдений.

1. Úvod

1.1. **Historie.** Biostatistika je odvětví aplikované statistiky, které se zabývá použitím statistických metod při řešení biologické, lékařské a zdravotnické problematiky. V rámci aplikací tohoto typu se vedle obecně známých statistických metod používají i specifické statistické postupy potřebné v epidemiologii, demografii, při řízení klinických pokusů, atp. U mnoha specifických technik je důležitá nejen jejich statistická podstata, ale i jejich výpočetní aspekty a filozofické i etické otázky spojené

2000 *Mathematics Subject Classification.* 62P10 92B15.

Klíčová slova. Biostatistika, modely pro korelovaná data, GEE, ordinální regresní modely, klinické pokusy, bioekvivalence, epidemiologie, mnohonásobné testování.

Autoři děkují RNDr. J. Klaschkovi, Ph.D. za pečlivé prostudování práce a cenné připomínky, které přispěly ke zpřesnění textu.

s jejich použitím v reálné praxi. Podněty z praxe stojí i u zrodu nových statistických metod. Biostatistika představuje velice rozsáhlý obor, jehož složitost vyplývá z toho, že leží na pomezí několika vědních oblastí. Často se prolíná (a někdy i zaměňuje) s dalšími obory jako jsou biometrie a lékařská statistika [60] a má mnoho styčných bodů s demografií a zejména epidemiologií, která se zabývá studiem a kvantifikací výskytu nemocí ve skupinách lidí a snaží se vysvětlit příčiny nemocí a najít vazby mezi výskytem nemocí a charakteristikami lidí a jejich prostředí.

Biostatistické aspekty jsou obsaženy v širokém spektru problematiky od charakterizace základní struktury a funkce lidských organismů přes jejich interakce s prostředím, v kterém žijí a pracují, až k prevenci nemocí a terapii, k organizaci zdravotní péče a ekonomickým otázkám s tím spojeným.

Cílem našeho textu je poukázat na ty obecné i specifické postupy a metody, které se v současné době dynamicky rozvíjejí a hrají klíčovou roli při statistickém zpracování biologicko-lékařských výzkumů. O několika tématech z celkového přehledu bude pojednáno detailněji.

V rámci lékařské praxe a výzkumu vznikají obrovská množství dat se spoustou nejistot, vzájemných vazeb a variability (neboť jsou charakterizací živých bytostí a jejich životního prostředí v celé jejich komplexnosti a různorodosti), a proto je korektní analýza netriviální. To si odborníci začali uvědomovat již dávno, když rozpoznali překvapivé pravidelnosti v souhrnných charakteristikách dat, která se dosud při individuálním pohledu jevila prakticky náhodná. Jako příklad se často uvádí analýza dat narození, svateb a pohřbů v Londýně, kterou už v roce 1662 shrnul J. Graunt ve své práci *Observations upon the Bills of Mortality*. I přesto, že jeho přístup nebyl matematický, dokázal velmi citlivě a důmyslně z dat vyvodit zajímavé závěry [11]. V 19. století byli průkopníky v této oblasti W. Farr, A. Quetelet, P.Ch.A. Louis. William Farr se mj. podílel na vybudování systému pro mezinárodní klasifikaci nemocí. Práce F. Galtona, K. Pearsona a W.F.R. Weldona o dědičnosti a biologické variabilitě vyústily právě před 100 lety v založení časopisu Biometrika. V roce 1915 přednášel G.W. Snedecor kurs biometrie (orientovaný na zemědělskou problematiku) a v roce 1925 vyšla v prvním vydání kniha R.A. Fishera Statistical Methods for Research Workers. Před 50 lety navrhl A.B. Hill první randomizovaný klinický pokus, který prokázal účinnost streptomycinu při léčbě tuberkulózy. Dá se tak říci, že biostatistika stála přímo u zrodu moderní matematické statistiky.

1.2. **Aktuální témata na přelomu tisíciletí.** Jaká biostatistická témata jsou aktuální? Jak se pozná žhavé téma? Odpověď na tyto otázky bude vždy silně subjektivní; vzhledem k šíři oboru bude nepochybně odrážet osobní zkušenosti a konkrétní oblast aplikací, v níž se posuzovatel pohybuje. Nicméně lze určitě za aktuální považovat ta témata, která se často vyskytují v publikovaných článcích. V teoretické rovině jde zřídkakdy o zcela nový postup (jako v případě Coxova modelu proporcionálního rizika [29]), spíše jedná o modifikace či doplnění. Volba metod pro řešení reálných úloh do značné míry závisí na jejich softwarové dostupnosti.

Za rozvojem mnoha dnes silně využívaných technik stojí dynamický rozvoj počítačů a výpočetních možností. Je možno [64]

- uchovávat a zpracovávat stále větší objemy dat se stále komplikovanější strukturou — taková data vznikají např. při dlouhodobém sledování individuálních pacientů (onkologická či chronická onemocnění), při sledování rodinné historie v genetické epidemiologii, v souvislosti s prostorovými daty (úlohy o vztahu ukazatelů zdravotního stavu a znečištění životního prostředí),

- technicky zvládnout zpracování složitějších nejrůznějších typů modelů (např. modely pro opakovaně měřená, korelovaná data, modely s náhodnými efekty) a nových technik (bootstrap, MCMC, grafické, exaktní metody),
- navrhnout a v praxi uplatnit takové algoritmické postupy, které ještě před nedávnem nebyly realizovatelné.

Představu o tématech, která považují uznávaní odborníci za závažná, poskytují mj. práce psané u příležitosti počátku nového tisíciletí, např. [3, 24, 52], sborník připravený k výročí 50 let existence Mezinárodní biometrické společnosti [13] a určitě také kapitoly z obsáhlé šestidílné encyklopedie [12]. Vynikající americký biostatistik Norman Breslow [21, 22, 23, 24] uvádí hierarchické modely, modely se smíšenými efekty, modely BLUP (Best Linear Unbiased Prediction), GEE (Generalized Estimating Equations), dále REML (Restricted Maximum Likelihoood), problematiku chybějících dat (nejen) ve zmíněných modelech, Bayesovské metody, počítačově intenzivní metody a vůbec celou oblast výpočetní statistiky, aplikace statistických metod v genetice, toxikologii, molekulární biologii, epidemiologii životního prostředí, klinické pokusy, analýzu přežívání a další speciálnější témata jako ROC (Receiver Operating Curve), aj. Richardson [93] vyzdvihuje především zobecněné lineární modely pro korelovaná data, cenzorovaná data a analýzu přežívání, Bayesovské modelování. Podobně i Altman [8] uvádí na prvním místě modely — zobecněné aditivní modely, modely pro longitudinální data a pro hierarchická data — a dále Bayesovské metody, počítačově intenzivní metody a neuronové sítě. Článek [33] shrnuje u příležitosti stého výročí časopisu *Biometrika* oblasti, ve kterých časopis zejména přispěl k rozvoji statistické metodologie. Jde tedy vlastně o přehled aktuálních témat z pohledu předního časopisu s biostatistickým zaměřením. Mimo jiné jsou zde uvedeny Bayesovské metody, metody mnohorozměrné statistiky, robustní metody, prostorová statistika, otázky volby modelu, chybějící pozorování a v neposlední řadě zobecněné lineární modely, modely pro longitudinální data a modely pro data s nezávislými shluky korelovaných pozorování. Longitudinální, opakovaně měřená (nejen v čase), a tudíž korelovaná data poutají pozornost mnoha autorů ([41], [42], [76]).

Při všech aplikacích statistických metod (nejen v biostatistice) stále je a bude aktuální zabývat se obecnými statistickými principy, jak je shrnul např. Cox [31]. Jde zejména o otázky, kteří jedinci a v jakém počtu mají být studováni, nebo jaká srovnání a na základě jakého modelu a testu mají být provedena. U modelu je podstatné, aby dostatečně dobře odpovídal reálné situaci, aby odpovídal dosavadním znalostem a teorii i způsobu vzniku dat, aby adekvátně prokládal data, aby měl jasně interpretovatelné parametry, aby měl dostatečně realistickou strukturu chybového členu a aby umožňoval porovnání s jinými studiemi daného tématu. Je potřeba co nejvíce redukovat náhodné chyby a zabránit vzniku systematických chyb. Ukázkami různých prakticky používaných modelů se zabývá např. [18].

1.3. **Příklad: HIV/AIDS.** Velmi závažný zdravotnický problém posledních dvaceti let, onemocnění AIDS způsobené virem HIV, přináší složité úkoly i pro biostatistiky a lze na něm ilustrovat širokou škálu problémů [38]. Snaha o modelování výskytu a přenosu (šíření, dynamiky) nemoci, demografického dopadu vysoké incidence AIDS a účinků preventivních opatření je komplikována povahou nemoci. Inkubační doba je dlouhá a proměnlivá, klinicky latentní fáze může trvat roky a může být úspěšně prodlužována novými typy kombinované léčby. Infekčnost je proměnlivá v průběhu onemocnění. Modely analýzy přežívání jsou komplikované, data jsou dvojitě cenzorována: až na výjimky není známa doba nákazy či sérokonverze (cenzorování zleva), v některých případech je známo datum posledního negativního a prvního pozitivního

testu (intervalové cenzorování), událost (klinický projev AIDS, úmrtí) často ještě nenastala (cenzorování zprava); dochází ke ztrátám pacientů ze sledování. V mnoha zemích byly do nedávné doby používány monitorovací a hlásicí systémy jen pro sledování případů AIDS, a nikoli pro evidenci nakažených virem HIV. Jejich počty se odhadovaly v modelech ze známých počtů AIDS metodou tzv. zpětné kalkulace. Jenže aplikace kombinované léčby jejich platnost narušuje. Hlášení případů probíhá se zpožděním, na které je nutno při výpočtu brát zřetel. Nejsou přesně známy všechny faktory, které mají vliv na délku inkubační doby. Jedním z nich je zřejmě způsob nákazy, ale ten často není znám spolehlivě. A navíc nepochybně souvisí se sexuálním chováním, které je velmi různorodé a jakékoli dotazování na ně představuje choulostivý výzkumný problém. Používají se zástupné (surrogate) indikátory progrese infekce, zejména počet CD4+ T-lymfocytů [47] a virová nálož (počet kopií RNA) [65]. Ty jsou však vzhledem k vysoké míře biologické variability měřeny s velkou chybou, násobenou ještě detekčními limity přístrojů (počet kopií RNA lze spolehlivě stanovit zhruba v rozsahu od 500 kopií do 1 miliónu na mililitr plazmy); nemají normální rozložení; obsahují chybějící pozorování (což vede ke snahám o jejich nahrazování, např. při předpovědi doby přežívání [83]). Nové otázky přináší výzkum vakcín proti HIV, který se už dostává do fáze klinických pokusů. Vedle klinických pokusů se při výzkumu AIDS uplatňuje i mnoho dalších typů epidemiologických studií (kohortové, aj.). Používají se komplikované farmakokinetické modely pro popis sdružených účinků několika léků. Problematika HIV/AIDS se tak více či méně dotýká všech deseti témat, kterým se věnujeme v tomto textu.

2. Téma 1: Kategoriální data

2.1. **Kategoriální data v medicíně.** V některých oblastech medicíny není možnost přesného objektivního měření, a proto jsou pro ně typická především kategoriální data. Kategorizace bývá mnohdy dosti hrubá, nicméně ji přesto provází riziko chybné klasifikace do kategorií. Stále častěji se přihlíží k subjektivním informacím pacientů, které jsou většinou získávány pomocí dotazníkových šetření [36]. Epidemiologický přístup k modelování rizika se často opírá o čtyřpolní tabulky a ukazatele jako relativní riziko a odds ratio (OR; v češtině není ustálený překlad, nejčastěji se užívá křížový poměr nebo poměr šancí). Velice diskutované jsou postupy pro kontrolu vlivu rušivých, matoucích proměnných (confounding), jako jsou stratifikace a Mantelova-Haenszelova technika nebo logistická regrese [69]. Klinické pokusy přinášejí velké a řídké kontingenční tabulky. Rozvoj výpočetní techniky a vypracování příslušných výpočetních algoritmů vedly k výraznému nárůstu zájmu o tzv. *exaktní testy* pro kategoriální data [82]. Exaktní postupy ale mohou být vzhledem k diskrétní povaze úlohy dosti konzervativní [7]. Proto se prosazuje přístup označovaný jako *mid-p* [97], kdy se uvažuje standardní p-hodnota zmenšená o polovinu rozdílu mezi ní a nejbližší nižší možnou hodnotou.

Teorie kategoriálních dat je již dlouho dobře propracována (viz např. Bishop a spol. [16], Agresti [5] a další autoři jako Cochran, Goodman, Mantel). Aktuální zájem je nyní věnován snaze o větší respektování skutečné struktury dat. Pozornost se proto obrací ke kategoriálním modelům pro opakovaná pozorování [56, 116] a k modelům s uspořádanými kategoriemi [4, 6].

2.2. **Hodnocení uspořádaných kategoriálních veličin.** Ve statistické praxi se vyskytuje poměrně často hodnocení závislostí v případě, kdy jsou kategoriální odpovědi uspořádány. V oblasti biostatistiky jsou to zejména indikátory zdravotního stavu, závažnost onemocnění, míra obtíží nebo kvalitativní míra expozice. Takové

škály jsou uspořádány, ale nejsou známy kvantitativní rozdíly mezi jednotlivými kategoriemi. Někdy jsou ordinální kategorie výsledkem kategorizování spojitých veličin (např. příjem, denní počet vykouřených cigaret, konzumace alkoholu). Přitom vymezení kategorií je často značně subjektivní. Protože metody vhodné pro hodnocení uspořádaných kategoriálních veličin nejsou dostatečně známy, jsou kategoriální škály hodnoceny jako nominální, např. pomocí kontingenčních tabulek, bez využití ordinality. Podstatný rys těchto veličin je tak ignorován a důležitá informace je ztracena. Navíc při standardním hodnocení výstup (statistika χ^2) nevyjadřuje míru asociace zkoumaných veličin. Hodnocení vlivu dalších faktorů je v těchto situacích obtížné. Jinou možností je hodnocení trendu vzhledem k nějaké škále např. k přirozenému očíslování kategorií nebo jinak utvořeným numerickým hodnotám, např. jsou li známy hranice kategorií. Takový způsob kvantifikace však může značně ovlivnit výsledné závěry. Další z možností hodnocení uspořádaných kategoriálních veličin nabízí klasická logistická regrese, kdy kategorie jsou redukovány na binární případ. Přitom informace obsažená v původní několikastupňové škále je ztracena a výsledek závisí na zvolené dělicí hranici. Souhrnně metody, které hodnotí ordinální škály jako nominální, dichotomické nebo intervalové, vykazují řadu nedostatků a mohou vést k chybným statistickým závěrům. Tyto přístupy jsou v případě ordinálních škál přesto stále užívány, i když statistická teorie i software nabízejí nyní lepší možnosti.

Jednou z možností jsou tzv. *ordinální regresní modely.* V této kapitole přiblížíme dva modely anglicky nazývané „Proportional Odds“ a „Continuation Ratio“. Rozšíření modelu logistické regrese navrhli již v roce 1967 Walker a Duncan [114]. Podrobné odvození a souvislosti podobných modelů publikoval McCullagh [80].

Postupy vhodné k hodnocení uspořádaných kategorií obecně znamenají odvození skórů, které jsou přiřazovány těmto kategoriím, a jejich analýzu jako spojitých veličin. Předpokládá se existence skryté proměnné, kterou nelze přímo pozorovat a která se projevuje v ordinální škále, jejíž kategorie tvoří sousedící intervaly.

Postup budeme ilustrovat na příkladu hodnocení výsledků studie idiopatické skoliózy z nemocnice Ste-Justine v Montrealu [101]. Skupina případů idiopatické skoliózy je zde srovnávána se skupinou kontrolních osob, které byly náhodně vybrány z populace. Sledované osoby vyplňovaly dotazník, kterým se zjišťovala míra bolesti v oblasti krku a zad. Míra bolesti měla tyto kategorie: žádné obtíže (0), krátké nevyskytující se často (1), krátké opakované (2), dlouhé (3), téměř stálé (4) a stálé (5). Počty případů a kontrol v jednotlivých kategoriích jsou uvedeny v Tab. 1, která obsahuje tzv. křížové poměry (odds ratio; OR) v kategoriích 1–5 vzhledem ke kategorii 0. Tento ukazatel, užívaný běžně v epidemiologii, vyjadřuje míru asociace ve čtyřpolní tabulce

	případy	kontroly
exponovaní	a	b
neexponovaní	c	d

a je definován vztahem [69]:

$$OR = \frac{a/b}{c/d}.$$

Meze 95% intervalu spolehlivosti (CI) jsou pak

$$(OR/EF; OR \times EF)\,,$$

přičemž faktor EF se určí podle vzorce

$$EF = e^{1{,}96\cdot\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}.$$

Míra bolestí		případy	kontroly	OR^a	95% CI
žádné	0	29	41	1,00	
krátké méně časté	1	24	22	1,54	0,73 – 3,26
krátké opakované	2	18	16	1,59	0,70 – 3,63
dlouhé	3	3	5	0,85	0,19 – 3,83
téměř stálé	4	17	12	2,00	0,83 – 4,82
stálé	5	9	4	3,18	0,89 – 11,32

$^a OR$ — odds ratio

Tab. 1 *Studie idiopatické skoliózy ve vztahu k míře obtíží.*

Tzv. nominální přístup k hodnocení míry obtíží mezi případy a kontrolami, který je ekvivalentní testu χ^2 v kontingenční tabulce, vede k p-hodnotě 0,345, tedy k závěru o statisticky nevýznamných rozdílech mezi kategoriemi míry bolesti. Trend hodnocený pomocí klasické binární logistické regrese (1= případ, 0= kontrola, míra bolesti 0–5= nezávislá proměnná) vede k p-hodnotě 0,051, tedy k hodnotě hraniční.

2.3. **Model „Proportional Odds“.** Model je v podstatě rozšířením binární logistické regrese, a proto bývá též nazýván ordinální logistický model. Tento model vychází z poměrů OR ve čtyřpolních tabulkách, které postupně vzniknou dichotomizací ordinální škály s dělicími body od nejnižšího k nejvyššímu. Každý odhad OR využívá všech pozorování a závisí na poloze dělicího bodu (Tab. 2).

dichotomie	případy $x = 1$	kontroly $x = 0$	OR	95% CI
0	29	41	1,00	
1–5	71	59	1,70	0,95 – 3,06
0–1	53	63	1,00	
2–5	47	37	1,51	0,86 – 2,66
0–2	71	79	1,00	
3–5	29	21	1,54	0,80 – 2,93
0–3	74	84	1,00	
4–5	26	16	1,84	0,92 – 3,70
0–4	91	96	1,00	
5	9	4	2.37	0,71 – 7,98

Tab. 2 *Studie idiopatické skoliózy ve vztahu k míře obtíží — odds ratio (OR) v dichotomickém uspořádání.*

Označme Y ordinální odpověď nabývající hodnot $0, 1, 2, \ldots, k$ a nechť $\mathbf{x}$ označuje vysvětlující, resp. podmiňující proměnnou (obecně vektor). Dále označme π_j podmíněné pravděpodobnosti j-té odpovědi podmíněné hodnotou $\mathbf{x}$ a γ_j kumulativní podmíněné pravděpodobnosti při hodnotě $\mathbf{x}$, tj.

$$\pi_j = P(Y = j|\mathbf{x}) \qquad j = 0, \ldots, k$$

a

$$\gamma_j = P(Y \leq j|\mathbf{x}) = \pi_0 + \cdots + \pi_j.$$

V obecném modelu lineární logistické regrese se předpokládá specifická závislost kumulativních pravděpodobností γ_j na $\mathbf{x}$, tj.

$$\text{logit}(\gamma_j) = \ln\left(\frac{\gamma_j}{1-\gamma_j}\right) = \alpha_j - \boldsymbol{\beta}_j'\mathbf{x},$$

kde $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$. V modelu „Proportional Odds“ se naproti tomu předpokládají různé konstanty α_j, avšak stejné „směrnice“ $\boldsymbol{\beta}_j = \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, tj.

$$\text{logit}(\gamma_j) = \ln\left(\frac{\gamma_j}{1-\gamma_j}\right) = \alpha_j - \boldsymbol{\beta}'\mathbf{x}.$$

To znamená, že grafem kumulativních logitů vzhledem k $\mathbf{x}$ jsou paralelní nadroviny.

Jiný pohled předpokládá existenci latentní (nepozorované) proměnné Z a dělicích bodů $\alpha_1, \ldots, \alpha_k$, pro které platí

$$\alpha_{j-1} < Z \ \leq \alpha_j \Leftrightarrow Y = j,$$

přičemž

$$Z = \boldsymbol{\beta}'\mathbf{x} + \varepsilon,$$

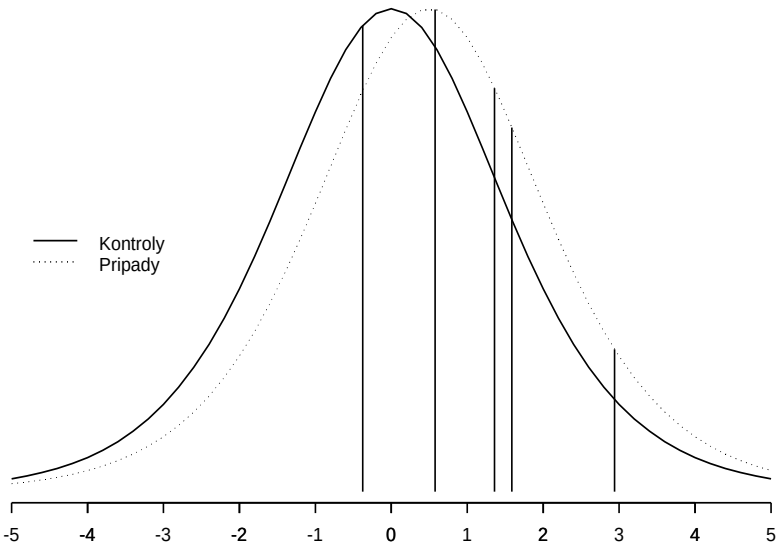
kde ε má logistické rozdělení s distribuční funkcí $F(z) = e^z/(1+e^z)$. Označíme-li

$$\text{Odds}(Y > j|\mathbf{x}) = \exp(-\alpha_j + \boldsymbol{\beta}'\mathbf{x}),$$

lze v situaci, kdy máme jedinou vysvětlující proměnnou, která slouží k rozlišení případů (=1) a kontrol (=0), vyjádřit OR jako tzv. „proportional odds ratio“ (POR) ve tvaru

$$POR = \frac{\text{Odds}(Y > j|1)}{\text{Odds}(Y > j|0)} = e^{\beta}.$$

Logistické rozdělení ve skupině případů a kontrol se tedy liší posunutím β (Obr. 1).



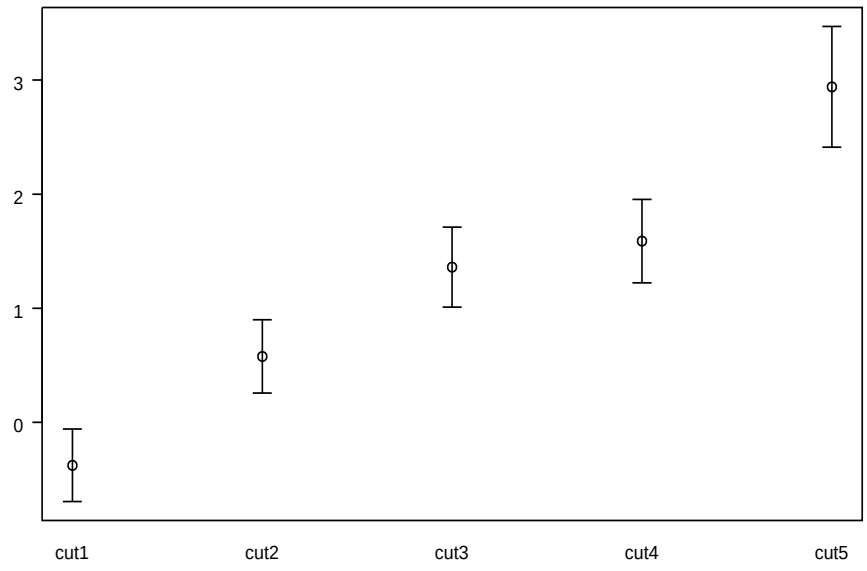
Obr. 1 *Logistická distribuce latentní proměnné Z v populaci případů idiopatické skoliózy (tečkovaná křivka), resp. kontrol (plná křivka); svislé čáry označují polohu dělicích bodů α_j a plochy pod křivkami odpovídají pravděpodobnostem π_j v každé z populací, tedy $P(Y = j|x = 1)$, resp. $P(Y = j|x = 0), j = 0, 1, \ldots, 5$.*

Obecně vektor **x** může obsahovat další důležité veličiny (např. věk). Parametry modelu se odhadují metodou maximální věrohodnosti. Tuto možnost nabízí např. software STATA (modul ologit). Výsledky pro náš příklad uvádí výpis v Tab. 3.

```
Ordered Logit Estimates                          Number of obs =    200
                                                 chi2(1)       =   3.88
                                                 Prob > chi2   = 0.0488
Log Likelihood =  -316.6824                      Pseudo R2     = 0.0061
------------------------------------------------------------------------------
        pain |      Coef.   Std.Err.      z    P>|z|     95% Conf.Interval
-------------+----------------------------------------------------------------
        beta |    0.5008   0.2550    1.964   0.050     0.00100    1.0005
-------------+----------------------------------------------------------------
       _cut1 |   -0.3768   0.1931                     (Ancillary parameters)
       _cut2 |    0.5777   0.1950
       _cut3 |    1.3600   0.2133
       _cut4 |    1.5884   0.2222
       _cut5 |    2.9410   0.3218
```

Tab. 3 *Model „Proportional Odds“ — výstup z programu* STATA *(Řádek označený beta označuje odhad β a řádky cut1 – cut5 odpovídají parametrům α_1 – α_5).*

Hodnota POR, v našem příkladu $POR = e^{0,5008} = 1,65$ (95% CI: 1,00 – 2,72), vyjadřuje společnou hodnotu OR pro jednotlivá dělení z Tab. 2. Užitečnou informaci obsahují odhady dělicích bodů a jejich chyby. Z grafického vyjádření např. vyplývá, že není velkého rozdílu mezi 3. a 4. dělicím bodem (Obr. 2).



Obr. 2 *Polohy odhadů dělicích bodů α_1 – α_5 a jejich 95% intervaly spolehlivosti.*

2.4. **Model „Continuation Ratio“.** V tomto modelu jsou hodnoceny jednotlivé kategorie ve vztahu ke kategoriím vyšším (při daném uspořádání), tj. kategorie j se srovnává s kategoriemi $j+1,\ldots,k$. Tento přístup je analogický modelu proporcionálního rizika v diskrétním čase [28]. Definujeme-li γ_j stejně jako v modelu „Proportional Odds“, bude pro $j=1,2,\ldots,k-1$

$$1-\gamma_j = P(Y>j|\mathbf{x}) = \pi_{j+1}+\ldots+\pi_k,$$

$$\delta_j = P(Y=j|Y\geq j,\mathbf{x}) = P(Y=j|\mathbf{x})/P(Y\geq j|\mathbf{x}).$$

Lineární závislost logitů podmíněných pravděpodobností δ_j vzhledem k $\mathbf{x}$ je podstatou modelu „Continuation Ratio“:

$$\mathrm{logit}(\delta_j) = \ln(\delta_j/(1-\delta_j)) = \ln(\pi_j/(1-\gamma_j)) = -\alpha_j + \boldsymbol{\beta}'\mathbf{x}.$$

V tomto modelu se předpokládá, že OR nezávisí na indexech j a jsou v případě jediné x-ové proměnné se dvěma kategoriemi vesměs rovna e^{β}. Ve srovnání s modelem „Proportional Odds“ je však význam parametrů odlišný, neboť model „Proportional Odds“ predikuje kumulativní pravděpodobnosti, zatímco model „Continuation Ratio“ podmíněné pravděpodobnosti (hazards).

Prakticky jsou v případě dichotomické proměnné x pozorování roztříděna do čtyřpolních tabulek, podobně jako v modelu „Proportional Odds“, avšak s tím rozdílem, že v každé další tabulce jsou vynechána pozorování z prvního řádku předchozí tabulky. Postup je ilustrován v Tab. 4.

dichotomie	případy	kontroly	OR	95% CI
0	29	41	1,00	
1–5	71	59	1,70	0,95 – 3,06
1	24	22	1,00	
2–5	47	37	1,16	0,57 – 2,40
2	18	16	1,00	
3–5	29	21	1,23	0,51 – 2,95
3	3	5	1,00	
4–5	26	16	2,71	0,57 – 12,90
4	17	12	1,00	
5	9	4	1,59	0,40 – 6,38

Tab. 4 *Studie idiopatické skoliózy ve vztahu k míře obtíží — odds ratio v modelu „Continuation Ratio“.*

Vzhledem k tomu, že odhady příslušné indexům $j=1,2,\ldots,k-1$ jsou asymptoticky podmíněně nezávislé [30], lze odhadovat parametry modelu technikami logistické regrese, které umožňují stratifikaci odhadů. V případě více kategorií doporučuje Greenland [55] podmíněnou logistickou regresi. Náš odhad ukazatele $CR=e^{\beta}$ („continuation ratio“) je $CR=1,47$ (95% CI: 1,02 – 2,15, p=0,041). Byl pořízen pomocí software Epicure (modul PECAN). Ve srovnání s modelem „Proportional Odds“ závisí odhady v modelu „Continuation Ratio“ na směru škály. Odhady v tomto modelu získané při modelování vzestupné škály nejsou ekvivalentní převráceným hodnotám v sestupné škále (v našem příkladě je sice $CR=0,68=1/1,47$, ale p=0,055). Při aplikacích modelu je směr škály zřejmý v situacích, které zahrnují např. čas či nevratné změny. Souhrnná míra je při těchto hodnoceních snadno interpretovatelná, neboť osoby v každé úrovni škály musely nutně projít dřívějšími úrovněmi. Tento model je tedy vhodný pro hodnocení dob selhání, resp. přežití. Oba modely nabízejí

adekvátní metodu k hodnocení uspořádaných kategoriálních veličin. Vhodnost použití prvního nebo druhého modelu závisí na konkrétních podmínkách studie. V našem případě se zdá být vhodnější první model („Proportional Odds“), neboť škála míry bolestí nemusí nutně zahrnovat čas. Použití stejného příkladu k ilustraci obou modelů bylo zvoleno pro zjednodušení výkladu. Nižší p-hodnota v modelu „Continuation Ratio“ neznamená automaticky, že tento model je vhodnější. Praktický pohled na obě metody lze nalézt v práci [14]; je jim věnována kapitola v monografii [57].

3. Téma 2: Modely pro data s nadbytkem variability nebo korelace (nebo obojího)

3.1. **Motivace.** Biomedicínská (ale i jiná) data velmi často mají netriviální (např. hierarchickou) strukturu, která vede k zajímavým problémům a komplikacím, například korelaci mezi pozorovanými proměnnými, či nadbytku jejich variability oproti situaci popisované standardními regresními modely. Povšimněme si některých příčin, které k přítomnosti korelace vedou:

- Několik zdrojů náhodné variability (strukturální variabilita a „měřicí chyba“). Zde je jasně vidět, jak mohou být problémy s přítomností korelace a s nadbytkem variability úzce propojeny.
- Jde o nezbytný rys, vyplývající ze studovaného problému (oftalmologická a jiná měření párových orgánů, data sourozenců, apod.).
- Způsob sběru dat zvolený pro „pohodlnost“ či „operativnost“ (vedoucí např. k opomenutým kovariátám, jež nebyly či ani nemohly být změřeny).

„Intraclass“ korelace ovšem ani zdaleka nemusí být jen důsledkem chyb a nedokonalostí. Může být využita zcela záměrně a efektivně při plánování experimentů. Například zkřížené pokusy (cross-over), pokusy s náhodnými bloky, „cluster randomization“, apod. V souvislosti s navrhováním experimentů/studií je dobré si uvědomit, že jednotlivá měření uvnitř téhož shluku jsou korelovaná často pozitivně. Opačná situace (s negativní korelací) je podstatně řidší. Kladná korelace pak:

- *škodí* efektivitě srovnání mezi subjekty,
- *pomáhá* efektivitě srovnání uvnitř subjektu.

Typickou situací, v níž se s korelací v datech setkáváme, je případ opakovaných měření/čtení/hodnocení téhož subjektu (tzv. repeated measures). Pro stručnost budeme dále mluvit jen o měřeních (zde pro nás není podstatné, jakým způsobem data přesně vznikla). Opakované měření se může a nemusí vztahovat k času (např. odečty několika alergických testů u téhož jedince). Když se tedy zmiňujeme o „času“, naznačujeme, že tuto roli může hrát i jiná proměnná (která nemusí nezbytně implikovat platnost relace uspořádání, apod.). Někdy se v takové obecnější situaci mluví o „clustered data“ — s poukazem na model s nezávislými „shluky“ korelovaných dat (tj. subjekty).

Jako jednoduchý příklad si vezměme, že máme opakovaná měření odpovědi, která jsou spojena s kovariátami jednoduchým regresním modelem:

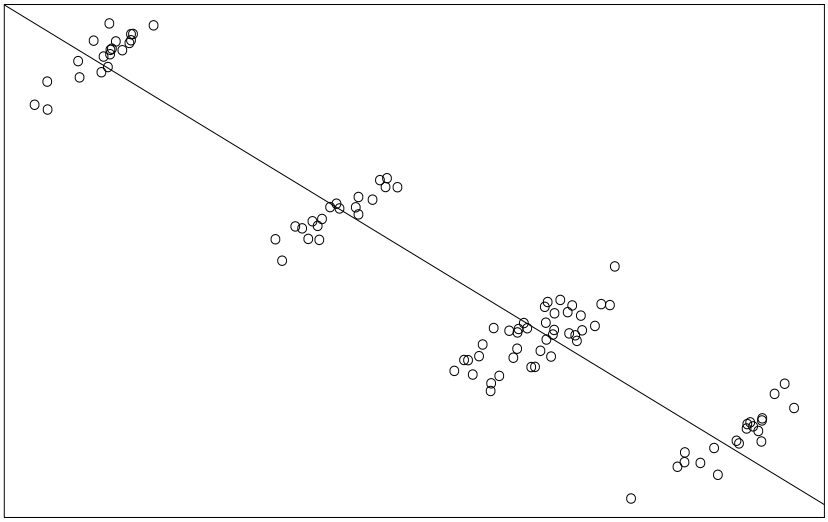
$$Y_{it} = \beta_0 + \beta x_{it} + \epsilon_{it} \quad ,$$

kde $\epsilon_{it}, \epsilon_{i't'}$ jsou pro různé subjekty ($i \neq i'$) nezávislé, zatímco uvnitř subjektů (pro $i = i'$) nikoli. Index t může (ale nemusí) být indexem času. Označíme-li počet opakování uvnitř shluku i jako n_i, můžeme mluvit o základní klasifikaci studií na dva typy:

- **Průřezové (cross-sectional)**,
  kde $n_i \equiv 1$.
  Tyto studie dovolí pouze $Y_{i1} = \beta_0 + \beta_C x_{i1} + \epsilon_{i1}$.
- **Longitudinální (longitudinal, panel data)**,
  kde $n_i > 1$ pro podstatný počet subjektů.
  Zde můžeme rozlišovat kovariáty následovně:
  - nezávislé na čase, $x_{it} = x_{i1}$
    (tak je tomu např. v typické klinické studii)
  - „časově“ proměnlivé $x_{it} \neq x_{i1}$, pro $t > 1$
    (typické pro mnoho epidemiologických šetření, zkřížené pokusy).

    Longitudinální studie s časově proměnlivou kovariátou umožňují rozklad $Y_{it} = \beta_0 + \beta_C x_{i1} + \beta_L (x_{it} - x_1) + \epsilon_{it}$.
    Interpretace β_C je stejná jako v průřezové studii, zatímco β_L popisuje střední změnu odpovědi *uvnitř* daného subjektu v „čase“ při jednotkové změně x.

Možnost výše zmíněného rozkladu je velkou výhodou a mocným argumentem pro (složitější a nákladnější) longitudinální studie. Jejich význam plyne z faktu, že nemusí nezbytně platit $\beta_C = \beta_L$ (což je problém, který úzce souvisí s jevem často zmiňovaným jako „ecological bias“). Obr. 3 tuto situaci ilustruje (bezproblémově lineární závislost má odlišné znaménko uvnitř shluku a mezi shluky).



Obr. 3 *Příklad situace s nestejnými β_C a β_L.*

Povšimněme si nyní situace, kdy máme pro nezávislé subjekty $1, \ldots, N$ napozorovaná korelovaná data: $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{it}, \ldots, Y_{i,n_i})^T$ společně s p kovariátami $\mathbf{x}_i$, kde t je čas (pro longitudinální data) nebo jiný index.

Jde nám přitom o odhad/test regresní závislosti (o studium závislosti na čase a/nebo jiných kovariátách, nikoli tedy např. o predikci využívající plně informaci z vlastní historie apod.). Korelační struktura sama o sobě není v popředí zájmu. Bere se v úvahu jen proto, že komplikuje dosažení výše zmíněných cílů. Parametry, které jí popisují jsou rušivými parametry (nuisance). O tomto přístupu se často mluví jako o tzv. marginálním modelu [35].

Na popsanou situaci lze nahlížet jako na vícerozměrný problém. Proč tedy nepoužít klasického vícerozměrného přístupu (MANOVA, apod.)? Praktickým důvodem je to, že mnohá *reálná* data (biomedicínská i jiná) jsou značně nevyvážená (s n_i nestejným pro jednotlivé subjekty). Dalším důvodem je to, že sice existuje řada různých modelů korelační struktury pro vícerozměrná normální data, ale pro rozdělení jiná (zejména diskrétní) jsou možnosti mnohem chudší. Je přitom samozřejmě žádoucí mít k dispozici unifikovaný přístup obecněji aplikovatelný (např. binomická či Poissonovská data jsou v biomedicínské oblasti velmi častá).

3.2. **GLM formulace.** Jednou z možností, jak se s výše zmíněnou situací vypořádat, je formulovat regresní model jako zobecněný lineární model, GLM (Generalized Linear Model), viz [81]. Jinými slovy, předpokládáme pro pozorování Y_{it} hustotu s parametry θ_{it}, ϕ:

$$f(y_{it}) = \exp\left(\left(y_{it}\theta_{it} - b(\theta_{it})\right)\phi + a(y_{it}, \phi)\right).$$

Povšimněme si, že konkrétní typ rozdělení je specifikován volbou $a(.), b(.)$. Z vlastností exponenciální třídy plyne, že je-li model správně specifikován, platí $E[Y_{it}] = b'(\theta_{it})$ a $Var[Y_{it}] = \frac{b''(\theta_{it})}{\phi}$. Kromě standardní normální situace tak rozptyl typicky souvisí se střední hodnotou.

Přívlastek „linear“ v názvu pochází z toho, že kovariáty vstupují ve formě lineárního prediktoru $\eta_{it} \equiv \mathbf{x}_{it}^T\boldsymbol{\beta}$. Lineární prediktor pak modeluje střední hodnotu $\mu_{it} = E[Y_{it}]$ transformovanou pomocí tzv. link funkce, $L(\mu_{it}) = \eta_{it}$. Protože link $L(.)$ je monotónní, ale obecně nelineární funkce, je výsledný model obecně nelineární (kromě speciálního případu normality).

Je zřejmé, že tato formulace zahrnuje kromě normálního i různé další prakticky důležité případy, např. logistické či Poissonovo rozdělení pro Y_{it}. Povšimněme si též, že přítomnost parametru ϕ zvyšuje poněkud flexibilitu oproti standardnímu modelu lineární exponenciální třídy. Dokáže se totiž vypořádat s nadbytkem variability (viz např. „overdispersion“ v Poissonovské regresi).

Dosavadní GLM specifikace určuje (marginální) rozdělení pro Y_{it}, a tedy i regresi (závislost střední hodnoty tohoto rozdělení na kovariátách). Složky $\mathbf{Y}_i, \mathbf{Y}_j$ pro $i \neq j$ se berou jako nezávislé. Označuje-li index i subjekt, není to nerozumný předpoklad.

Z pohledu maximálně věrohodného odhadu to však není vše. Zbývá specifikovat vztahy mezi jednotlivými složkami $\mathbf{Y}_i$ (a tedy simultánní rozdělení pro celý vektor). Standardní GLM postuluje mezi jednotlivými složkami $Y_{it}, Y_{it'}$ pro $t \neq t'$ nezávislost. V biomedicínských problémech je dosti typické, že na jedné straně sice víme, že nezávislost není splněna, ale na druhé straně nemáme k dispozici dosti informací, které by nám umožnily specifikovat plný vícerozměrný model. Jako poněkud „naivní“ se na první pohled zdá přístup, který korelaci zcela ignoruje a postupuje dále, jako kdyby data byla skutečně nezávislá.

3.3. **Pracovní model nezávislosti (IWM).** Použití pracovního modelu nezávislosti (Independence Working Model, IWM) vede k tomu, že při odhadu jakoukoli korelační strukturu nejprve prostě ignorujeme a bereme ji v úvahu až při výpočtu asymptotických středních chyb. Pro parametr $\boldsymbol{\beta}$ pak dostaneme odhad $\hat{\boldsymbol{\beta}}_I$ řešením následující soustavy rovnic odhadu:

$$\mathbf{U}_I(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T \delta_i B_i^{-1} \mathbf{S}_i = \mathbf{0},$$

kde

$$\mathbf{S}_i = \mathbf{Y}_i - \mathbf{b}'(\boldsymbol{\theta}_i),$$

$$\mathbf{b}'(\boldsymbol{\theta}_i) = (b'(\theta_{i1}), \ldots, b'(\theta_{i,n_i}))^T,$$

$$\delta_i = \operatorname{diag}\left(\frac{d\mu_{i1}}{d\eta_{i1}}, \ldots, \frac{d\mu_{in_i}}{d\eta_{in_i}}\right),$$

$$B_i = \operatorname{diag}\left(\frac{d\mu_{i1}}{d\theta_{i1}}, \ldots, \frac{d\mu_{in_i}}{d\theta_{in_i}}\right) = \operatorname{diag}\left(b''(\theta_{i1}), \ldots, b''(\theta_{in_i})\right).$$

Pokud by pracovní model nezávislosti byl skutečně správný, jednalo by se o rovnice, které dostaneme nulováním skóre (tedy o rovnice, se kterými se standardně setkáváme při hledání maximálně věrohodného odhadu). Pak by tedy získaný odhad $\hat{\boldsymbol{\beta}}_I$ byl maximálně věrohodný — se všemi přitažlivými asymptotickými vlastnostmi. Co se však pokazí v případě, že nezávislost není splněna? Průkopnické práce [75] a [121] (motivované problémy v biomedicínských aplikacích) ukázaly, že něco samozřejmě ano, ale nikoli to hlavní. Za relativně mírných podmínek regularity má totiž $\hat{\boldsymbol{\beta}}_I$ obecně, i bez nezávislosti (tedy aniž by pracovní model byl zcela správně specifikován, pokud je správně specifikována závislost střední hodnoty na kovariátách) stále ještě leckteré dobré vlastnosti:

- je konzistentní,
- je asymptoticky normální,
- $\sqrt{N}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta})$ má asymptoticky kovarianční matici

$$V_I = \lim_{N\to\infty} N \left(H_1(\boldsymbol{\beta})\right)^{-1} \left(H_2(\boldsymbol{\beta})\right) \left(H_1(\boldsymbol{\beta})\right)^{-1},$$

kde

$$H_1 \equiv H_1(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T \delta_i B_i^{-1} \delta_i X_i,$$

$$H_2 \equiv H_2(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^T \delta_i B_i^{-1} Var[\mathbf{Y}_i] B_i^{-1} \delta_i X_i,$$

- nadto lze rozptyl odhadu $\hat{\boldsymbol{\beta}}_I$ konzistentně odhadnout, a to tzv. *sendvičovým* odhadem

$$(H_1)^{-1} \left(\sum_{i=1}^{N} X_i^T \delta_i B_i^{-1} \mathbf{S}_i \mathbf{S}_i^T B_i^{-1} \delta_i X_i\right) (H_1)^{-1}.$$

Sendvič lze spočítat i bez znalosti ϕ, i když V_I na ϕ závisí.

K praktickému použití tak získáváme jakýsi „pragmatický“ přístup. I přes chybnou specifikaci modelu máme asymptoticky rozumný odhad i testy. Postup, který vypadal zpočátku naivně, je „poučený“ v tom, že sice jednoduše odhaduje parametry získané z nesprávného IWM, pak ovšem nepoužívá přímo asymptotické rozptyly, které tento model poskytuje (a které jsou typicky příliš malé), ale koriguje je sendvičováním. Ze zcela praktického pohledu tak na sendvičovou korekci lze pohlížet jako na jednoduchý nástroj proti inflaci falešně pozitivních výsledků. Kompletní ignorování korelace bez následné korekce totiž často vede (kvůli podcenění variability odhadu) k chybě prvního druhu vyšší než nominální. Tak je tomu zejména u testů efektů spojených se srovnáními *mezi* subjekty (v typické situaci kdy jsou korelace uvnitř subjektů kladné).

Pokud pracovní model není zcela správný (data nejsou nezávislá), V_I typicky není na Raově-Cramérově mezi. Cenou za nesprávnou (a „pohodlnou“) specifikaci modelu je tedy ztráta vydatnosti odhadu $\hat{\boldsymbol{\beta}}_I$. Různé simulační studie (např. už [75]) ukazují, že ztráta vydatnosti sice často není nijak katastrofální, ale zhoršuje se, pokud jsou korelace uvnitř subjektu výrazné.

3.4. **Generalized Estimating Equations (GEE).** V pozadí tohoto přístupu (viz [121]) je následující intuitivně přitažlivá myšlenka. Máme-li nějakou přibližnou představu o tom, jak korelační struktura vypadá, nabízí se možnost ji použít v podobě „realističtějšího“ pracovního modelu než je IWM s nadějí, že odměnou za „méně nesprávný“ pracovní model bude vyšší vydatnost výsledného odhadu.

Předpokládejme tedy $R_i(\boldsymbol{\alpha})$ jako korelační matici $\mathbf{Y_i}$, specifikovanou až na vektor (neznámých) parametrů $\boldsymbol{\alpha}$. Na i smí eventuálně záviset jen proto, že ne u všech subjektů musí být k dispozici všechny složky pozorovaného vektoru. Odtud máme pracovní kovarianční matici $V_i = \frac{1}{\phi} B_i^{\frac{1}{2}} R_i(\alpha) B_i^{\frac{1}{2}}$. Rovnice odhadu (GEE) pak jsou:

$$\sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} X_i^T \delta_i V_i^{-1} \mathbf{S}_i = \mathbf{0}.$$

Neznámého $\boldsymbol{\alpha}$ se lze zbavit „profilováním“:

$$\sum_{i=1}^{N} \mathbf{U}_i \left(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}} \left(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}) \right) \right) = \mathbf{0}.$$

Za mírných podmínek regularity, jako např.:

- regrese (závislost střední hodnoty na kovariátách) je správně specifikována,
- $\sqrt{N}$–konzistentní $\hat{\boldsymbol{\alpha}}, \hat{\phi}$,
- $\left| \frac{\partial \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)}{\partial \phi} \right| \leq Q(\mathbf{Y}, \boldsymbol{\beta})$ pro nějaké $Q(\mathbf{Y}, \boldsymbol{\beta}) = O_P(1)$,

se výsledný odhad (dále označovaný jako $\hat{\boldsymbol{\beta}}_R$) opět chová vcelku dobře. K tomuto chování se váží následující dobré vlastnosti:

- Odhad $\hat{\boldsymbol{\beta}}_R$ je konzistentní a asymptoticky normální.
- $\sqrt{N}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})$ má asymptoticky kovarianční matici

$$V_R = \lim_{N \to \infty} N \left(G_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right)^{-1} \left(G_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right) \left(G_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right)^{-1},$$

kde

$$G_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} X_i^T \delta_i V_i^{-1} \delta_i X_i,$$

$$G_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} X_i^T \delta_i V_i^{-1} Var[\mathbf{Y}_i] V_i^{-1} \delta_i X_i \qquad .$$

Zde si povšimněme, že pokud jsou R (a další složky modelu) specifikovány správně, $G_1^{-1} G_2 G_1^{-1}$ se redukuje na asymptotický rozptyl MLE, ale sendvičování funguje obecněji.

- Asymptotický rozptyl lze opět odhadnout sendvičem, ve kterém jsou $\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi$ nahrazeny odhady.
  Při výpočtu „vnitřku sendviče“ je $Var[\mathbf{Y}_i]$ nahrazeno svým „odhadem“ $\mathbf{S}_i \mathbf{S}_i^T$. Pro jednotlivý subjekt i je takový „odhad“ samozřejmě neúnosný, ale při celkovém výpočtu je výhodou, že efektivně dochází k „půjčování informace“ mezi subjekty.

Podotkněme, že dobré asymptotické vlastnosti platí, i když je typ marginálního rozdělení (původní GLM model) specifikován nesprávně. Jinými slovy, ani „pracovní marginální hustota“ nemusí být specifikována kompletně správně — pokud je specifikována správně závislost střední hodnoty na kovariátách (regrese). To může jít až tak daleko, že rovnice odhadu nemusí být odvoditelné ze žádné „legální“ věrohodnosti (tedy nelze je získat z věrohodnostní funkce jakéhokoli přípustného modelu). V takovém případě se mluví o tzv. quasilikelihood odhadu [117].

Výše zmíněná standardní asymptotika je založena na rostoucím N a pevném $q = \max_{i \in 1 \ldots N} n_i$. Takové asymptotické výsledky budou dobrou aproximací pro relativně velký počet malých shluků korelovaných dat (např. pro typickou longitudinální studii v klinické medicíně). Oproti tomu pro malý počet velkých shluků (uspořádání se blíží situaci s několika dlouhými časovými řadami) nebude aproximace příliš kvalitní a velké problémy mohou nastat i se zásadními vlastnostmi GEE odhadů (např. pro skupinově randomizovanou studii, kdy se typ ošetření náhodně přiřazuje nepříliš velkému počtu nemocničních center, v jejichž rámci se aplikuje velké množství zákroků na jednotlivých pacientech).

Není bez zajímavosti si povšimnout, že zatímco GEE vznikly v kontextu biomedicínských problémů, podobný přístup se objevil zřejmě nezávisle i v jiných oblastech, například ekonometrii [119], a že základy teorie vysvětlující jejich vlastnosti jsou daleko staršího data [67].

Síla GEE přístupu tkví zejména v tom, že efektivně umožňuje specifikaci modelu rozdělit do dvou kroků:

(1) specifikace regrese, tedy závislosti střední hodnoty na kovariátách (a marginálního rozdělení jednotlivého pozorování Y_{it})
(2) specifikace korelační struktury mezi složkami vektoru $\mathbf{Y}_i$.

Zatímco u vícerozměrných normálních dat je toto oddělení samozřejmostí, velkým přínosem je zejména v situacích, kdy rozdělení pozorovaných veličin není normální. Tam pak GEE přístup umožňuje specifikovat model pro vícerozměrnou situaci bez kompletní znalosti celého simultánního rozdělení. Přitom i takto nekompletní (či semi-parametrická) specifikace postačuje k odhadu (a následným testům) regrese.

3.5. **Odhady rušivých parametrů.** Ve formulaci pracovního modelu se vyskytují kromě regresních parametrů $\boldsymbol{\beta}$, o které nám jde, i rušivé parametry související s rozptylem a korelační strukturou (ϕ, $\boldsymbol{\alpha}$). Přestože nás jejich hodnoty přímo nezajímají,

jejich odhady jsou zapotřebí vzhledem k tomu, že tyto parametry vstupují do rovnic odhadu. Výše zmíněná základní asymptotika na ně (kromě $\sqrt{N}$–konzistence) neklade příliš vysoké nároky. Často se tedy používají jednoduché momentové odhady z Pearsonových reziduí $r_{it} = \frac{y_{it}-b'(\theta_{it})}{\sqrt{b''(\theta_{it})}}$. Například:

- odhad parametru měřítka jako $\frac{1}{\hat{\phi}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i} r_{it}^2}{N}$,
- odhad korelačních parametrů $\hat{\boldsymbol{\alpha}}$ jako funkce $f(R_{12}, \dots R_{q-1,q})$, kde $q = \max_{i\in 1\dots N} n_i$ a $R_{uv} = \sum_{i=1}^{N} \frac{r_{iu}r_{iv}}{N}$,
- nebo modifikace předchozího, se jmenovatelem $N-p$ (kde p je počet regresních parametrů). Ty jsou obdobou REML odhadů známých z normálních lineárních modelů.

3.6. **Volba pracovní korelace.** Praktickým problémem je samozřejmě volba konkrétní pracovní korelační matice. V biomedicínských aplikacích se v případě, že mechanismus ovládající vztahy mezi měřeními uvnitř subjektu není do detailu známý, typicky volí pracovní korelace dosti jednoduše. Kromě IWM to je například AR(1), obecněji stacionární ARMA nízkých řádů, m-závislost (stacionární korelační struktura s nenulovou korelací jen do m-tého časového posunutí, běžně označovaného jako m-tý „lag“), konstantní korelace pro všechny páry („exchangeability“). Pro situace s malými shluky korelovaných dat lze uvažovat i o „volné“ korelační struktuře, kdy vektor neznámých parametrů $\boldsymbol{\alpha}$ zahrnuje všechny prvky R nad diagonálou.

Podívejme se nyní na příklad výsledků použití různých pracovních korelací pro konkrétní data z pokusu srovnávajícího růstové křivky (měření hmotnosti ke dni 0; 7; 14; 21) krys ošetřených různou dávkou (0; 0,05; 10; 25) jistého léčiva. V normálním modelu byly předmětem zájmu vybrané lineární kontrasty odpovídající srovnáním s kontrolou (nulovou dávkou). Pro ně byly při různých volbách pracovní korelace spočteny poměry z = (odhad)/(asymptotická stř. chyba). Ty srovnává Tab. 5.

Způsob výpočtu	z pro srovnání:		
	0,05 vs. 0	10 vs. 0	25 vs. 0
naivní	2,59	1,66	3,67
IWM	1,59	1,00	2,70
AR(1) $\hat{\rho}_1 = 0,68$	1,86	1,09	1,93
konst. korelace $\hat{\rho} = 0,15$	1,71	1,19	2,70

Tab. 5 *Výsledky pro různé volby pracovní korelace.*

V tabulce vidíme v praxi dosti častou situaci. Kompletní ignorování korelační struktury — tedy výpočet odhadu i jeho střední chyby z modelu s nezávislými daty (v tabulce označeno jako naivní způsob výpočtu) vede k nadhodnocení významu testovaného efektu. Vzhledem k tomu, že z má asymptoticky normální rozdělení, srovnání dávek 0,05 a 25 s kontrolou jsou významná na 5% hladině. Srovnání 25 versus 0 zůstává v zásadě významné i při použití GEE přístupu s různými pracovními korelacemi. Oproti tomu při srovnání 0,05 versus 0 působí sendvičová korekce výrazně v konzervativním směru (vede k poklesu $|z|$).

3.7. **Výpočetní aspekty.** Odhady regresních koeficientů se standardně získávají iterativně, v zásadě obdobně jako při Fisherově skórování. V $(j+1)$-ní iteraci dostáváme $\hat{\boldsymbol{\beta}}_{j+1}$ jako:

$$\hat{\boldsymbol{\beta}}_j - \left(\sum_{i=1}^{N} X_i^T \delta_i(\hat{\boldsymbol{\beta}}_j) V_i^{-1}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\alpha}}, \hat{\phi}) \delta_i(\hat{\boldsymbol{\beta}}_j) X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^T \delta_i(\hat{\boldsymbol{\beta}}_j) V_i^{-1}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\alpha}}, \hat{\phi}) \mathbf{S}_i(\hat{\boldsymbol{\beta}}_j)\right).$$

Výhodné je, že výpočet v jedné iteraci se dá prakticky realizovat jako vážená regrese $\mathbf{Z} = \delta X \hat{\boldsymbol{\beta}}_j - \mathbf{S}$ na δX s vahou V^{-1}. Vektor $\mathbf{S}$ přitom vzniká „stohováním“ $\mathbf{S}_i$ na sebe. Podobně matice vzniká X „stohováním“ matic X_i. Matice δ, V jsou blokově diagonální s diagonálními bloky δ_i, resp. V_i. Protože však v této regresi vystupují ještě neznámé rušivé parametry $\phi, \boldsymbol{\alpha}$, je třeba je nahradit jejich odhady z předchozího kroku (v každé iteraci se po výpočtu $\hat{\boldsymbol{\beta}}_{j+1}$ počítá z reziduí odhad rušivých parametrů).

V obecnějším pojetí, odvozeném z kvadratické exponenciální třídy [123, 45, 12] se $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$ získávají simultánním řešením rovnic odhadu:

$$\sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix}^T \begin{pmatrix} V_i & C_i \\ C_i^T & L_i \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Y}_i - \boldsymbol{\mu}_i \\ \mathbf{W}_i - \boldsymbol{\nu}_i \end{pmatrix} = \mathbf{0},$$

kde $W_{itt'}$ jsou součiny Pearsonovských reziduí:

$$\mathbf{W}_i = (W_{i12}, \ldots, W_{i,(n_i-1),n_i})^T = (r_{i1} r_{i2}, \ldots, r_{i,n_i-1} r_{i,n_i})^T,$$

$$\boldsymbol{\nu}_i = E(\mathbf{W}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}).$$

Kromě V_i jako pracovní kovarianční matice $\mathbf{Y}_i$ je třeba mít specifikovanou pracovní kovarianční matici pro $\mathbf{W}_i$ (tedy L_i). A dále též pracovní kovarianci mezi oběma vektory, $C_i = \bar{\mathrm{C}}\mathrm{ov}(\mathbf{Y}_i, \mathbf{W}_i)$. Prakticky se C_i, L_i volí často velmi jednoduché (např. diagonální L_i) ve snaze vyhnout se při formulaci modelu mnoha komplikovaným vyšším momentům.

Přístup v duchu tradičního GEE (tzv. GEE1, first-order estimating equations) předpokládá ortogonalitu rovnic pro $\boldsymbol{\beta}, \boldsymbol{\alpha}$. Mimodiagonální bloky prvních dvou matic v soustavě rovnic odhadu jsou nulové ($\frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\beta}} = 0$, $C_i = 0$). Tak jsou rovnice pro odhad $\boldsymbol{\alpha}$ a $\boldsymbol{\beta}$ separovány. Předností je jednoduchost, a jak jsme se zmínili výše, pro základní asymptotiku $\hat{\boldsymbol{\beta}}$ konkrétní volba způsobu odhadu $\hat{\boldsymbol{\alpha}}$ nehraje roli (pokud je $\hat{\boldsymbol{\alpha}}$ $\sqrt{N}$–konzistentní).

Jiná je však situace z hlediska odhadu $\boldsymbol{\alpha}$ (pokud není jen rušivým parametrem a zajímá nás sám o sobě). Zmíněné jednoduché odhady nejsou příliš kvalitní. Možností, jak je za cenu komplikovanějšího modelu i výpočtů zlepšit, je tzv. GEE2 přístup, který od ortogonality rovnic odhadu $\boldsymbol{\alpha}$ a $\boldsymbol{\beta}$ upouští. Simulační studie ukazují, že zatímco vydatnost odhadu $\boldsymbol{\alpha}$ se tak zlepšuje, vliv na vydatnost $\boldsymbol{\beta}$ je malý nebo žádný. Zlepšení efektivity odhadu $\boldsymbol{\alpha}$ však není zdarma: případná špatná specifikace pracovního modelu pro $\mathbf{W}_i$ „prosakuje“ i do odhadu $\boldsymbol{\beta}$ a může ho závažně poškozovat. Vidíme zde tedy i obecněji zajímavé dilema mezi odolností vůči nesprávné specifikaci korelační struktury a efektivitou.

V literatuře, v některých programech, apod. se GEE odhad 1. řádu parametru $\boldsymbol{\beta}$ někdy označuje přívlastkem „robust“. Míněna je především robustnost vůči nesprávné specifikaci pracovní korelace. Vzhledem k tomu, že odhad je typicky založen na prvních a druhých momentech, je jasné, že robustnost např. vůči odlehlým pozorováním může být problémem. Proto se nověji objevují snahy o vylepšení, např. [66].

4. Téma 3: Neúplná nebo chybná informace

V biologickém a lékařském výzkumu je zcela typické, že se nepodaří získat všechna požadovaná data. Příčin neúplné informace je celá řada a patří k nim např.

- nedokončení longitudinální studie, neudržení spolupráce (pacient odmítne nebo není schopen pokračovat, např. pro vedlejší účinky léku),
- nenavázání kontaktu či odmítnutí účasti ve výběrovém šetření (nonresponse),
- technické důvody, odchylky od protokolu studie (u konkrétního pacienta není změřena některá proměnná anebo pacient nebyl změřen v určitém požadovaném časovém okamžiku),
- neúplná znalost určité proměnné (cenzorování, detekční limity),
- chybění dat jako důsledek plánu studie (např. když se studie provádí ve dvou krocích, kdy v prvním screeningovém kroku je provedeno jednoduché šetření u většího výběru a ve druhém kroku jsou podrobeni časově a finančně náročnému vyšetření jen někteří jedinci vytipovaní v prvním kroku).

V zásadě existují tři hlavní přístupy k analýze dat s chybějícími pozorováními [12]:

- vyloučení všech neúplných záznamů ze zpracování,
- nahrazení chybějících dat a následné zpracování doplněných dat,
- zpracování neúplných dat metodami, které takovou strukturu připouštějí.

Zejména nahrazování (imputaci) se nyní věnuje mnoho publikací, které se jím zabývají z pohledu epidemiologie, statistiky, klinických pokusů, často v kontextu určité metody. Jde o téma určitě aktuální a potřebné, ale i problematické — bezvýhradní stoupenci metody jsou schopni věřit použitelnému modelu do extrému a použít nahrazování i u proměnné, která má třeba 80 % chybějících hodnot. Kvalitativně dál je mnohonásobné nahrazování (multiple imputation), kdy je na základě prediktivní distribuce vytvořeno několik (řekněme 5) odlišných verzí původního souboru s různě doplněnými daty a výsledky z nich získané jsou kombinovány. Podrobné informace lze najít v pracích průkopníků tohoto přístupu, Rubina [99], a Schafera [104, 105]. Na internetové stránce Joe Schafera je dostupný software pro mnohonásobné nahrazování.

Rubin [98] podnítil zájem o problematiku chybějících dat, když upozornil, že způsob nahrazování by měl vycházet ze znalosti mechanismu vzniku chybějících hodnot, a charakterizoval tři hlavní mechanismy podle toho, jak pravděpodobnost chybění závisí na dalších charakteristikách. Pravděpodobnost výpadku buď nezávisí na žádném jiném údaji, který byl nebo bude zaznamenán (údaj chybí „zcela náhodně“), nebo souvisí s informacemi, které jsou již známy (jako je třeba věk, pohlaví, ale také předchozí měření pacienta), ale nemá vztah k dosud nepozorovaným datům, tj. k informacím současným ani budoucím (údaj chybí „náhodně“). V praxi je ovšem zejména v longitudinálních studiích běžná situace, kdy údaje chybějí např. v důsledku zhoršení pacientova stavu [72]. Tak nastává nejkomplikovanější možnost, kdy pravděpodobnost chybění může záviset nejen na dřívějších zjištěních, ale i na skutečnostech, které jsou právě zjišťovány nebo teprve budou zjištěny v budoucnu (údaj chybí „nenáhodně“). Zde je velký prostor pro další intenzivně probíhající výzkum adekvátních modelů.

Kromě dat chybějících představují závažný problém i data nepřesně měřená. Epidemiologové vědí již dlouho, že nepřesné stanovení expozice může ovlivnit výsledky analýzy zásadním, a navíc nepředpověditelným způsobem. Síla statistických testů se přitom snižuje. Takový problém vzniká např. v dnes velmi populárních studiích

o vztahu znečištění ovzduší a morbidity či mortality [37]. Většinou nejsou k dispozici individuální měření expozice a údaj z centrální monitorovací stanice se používá pro všechny obyvatele lokality. Jde o velmi nepřesné nahrazení, korelace ukazatelů z individuálních a centrálních měření bývá velmi nízká.

Teprve relativně nedávno (jak shrnuje Carroll a spol. [26]) došlo k výraznému pokroku ve vývoji metod, které dokáží do odhadu parametrů zahrnout opravu vzhledem k vlivu chyby v měření těch proměnných, které jsou klasickými modely považovány za přesné, tj. (v různých terminologiích) doprovodných resp. vysvětlujících proměnných, kovariát, rizikových faktorů, matoucích proměnných (confounder). Je potřeba specifikovat základní regresní model pro vztah doprovodných proměnných a odpovědi (následku), dále model chyby měření, který popisuje vztah mezi měřenými (chybou zatíženými) a skutečnými (přesnými) hodnotami ukazatelů, a konečně marginální model pro distribuci skutečných hodnot doprovodných proměnných.

5. Téma 4: Statistická genetika

Obrovské množství projektů se v poslední době zabývá genetickou problematikou, protože genetické faktory určují nejrůznější aspekty života, což je důležité nejen z pohledu medicíny nebo zemědělství, ale třeba i soudnictví. Vznikají zde extrémně velké a výpočetně náročné objemy dat, např. ve studiích lidského genomu (mapování řádově 100 000 lidských genů). Je potřeba hledat vazby a testovat mnoha genů najednou, propojovat známou genetickou informaci s konkrétní charakteristikou organismu, který daný gen nese. Statistická genetika a mikrobiologie pomáhá při řešení velice různorodých problémů. Pro představu uveďme, že editoři rozsáhlé encyklopedické příručky [15] je rozčlenili do šesti základních oblastí: bioinformatika, populační genetika, evoluční genetika, genetická epidemiologie, genetika zvířat a rostlin a aplikace (např. farmakogenetika, výpočet rizika genetického poškození). Jak poukazují různí autoři ([54], [40], [13], [118]), statistika nachází uplatnění při studiu genetických stochastických procesů, při sekvenování DNA, při měření genetické vzdálenosti či frekvence a rozmanitosti alel, atd. Závažné statistické problémy přináší potřeba extrakce informace ze silně zašuměných dat. V genetické epidemiologii slouží statistické metody jak při návrhu biologických experimentů, tak třeba při hodnocení speciálních typů studií (case-only, rodinné — typizace, atd.).

6. Téma 5: Prostorová data

Studium nemocí z hlediska jejich geografického rozšíření je jedním z trojice pilířů deskriptivní epidemiologie, která se primárně zajímá o „osobu, čas a místo“. Dnes již klasickým příkladem jsou mapy výskytu případů cholery v Londýně v roce 1853, které posléze byly jedním z důležitých vodítek k identifikaci závadného zdroje vody [110, 96]. Opět zejména technický pokrok podnítil výrazný rozvoj biostatistických metod souvisejících s mapováním nemocí [115, 73, 12] či vůbec s prostorovými daty [32]. Ožehavé téma představují tzv. shluky nemoci, tj. zjevné zvýšení výskytu nemoci v jisté oblasti oproti očekávaným počtům. Problém spočívá v malých pozorovaných i očekávaných počtech, v obtížné měřitelnosti expozice potenciálnímu rizikovému faktoru (viz téma 3) a také v tom, že čas a místo studie i formulace hypotéz se zpravidla odvozují až z dat, nejsou plánovány dopředu. Velké pozornosti se tomuto tématu dostalo na základě tvrzení o zvýšeném výskytu dětských leukémií v okolí jaderných zařízení [51], které však jiní nepotvrdili [84]. Je to i otázka zvoleného přístupu, což ilustruje na příkladu dětských leukémií Schinazi [107]. Velmi

zjednodušeně uveďme, že v provincii Columbus bylo v roce 1975 pozorováno 12 případů proti 6 očekávaným. Pravděpodobnost, že taková situace nastane právě v této oblasti, je 0,01, ale pravděpodobnost, že nastane v jednom z 200 regionů USA (v každém z nichž žije zhruba 200 000 dětí), je asi 0,80. Je zde patrná souvislost s otázkami mnohonásobného srovnávání (viz téma 10).

Tzv. small-area statistics (která se typicky zabývá situacemi, kdy se v prostoru a časovém intervalu vyskytuje méně než 20 případů nemoci) je dnes velmi žádaná. Statistické odhady v jedné oblasti jsou však nespolehlivé, což přináší nutnost kombinace a vyhlazování map, studia prostorové závislosti, atp. Zajímavé možnosti přináší aplikace Bayesovského přístupu při vyhlazování map. Příslušné hierarchicko-prostorové modely jsou realizovány pomocí MCMC simulačních algoritmů (např. software BUGS — Bayesian Inference Using Gibbs Sampling [12]).

7. Téma 6: Epidemiologické studie

7.1. **Epidemiologie a biostatistika.** Rozsáhlou oblast pro uplatnění biostatistiky představují epidemiologie a zejména epidemiologické studie, jak už vyplývá i ze zmínek u jiných zde probíraných témat. V oblasti observačních studií (studie případů a kontrol a kohortové) je zvýšený zájem věnován nestandardním výběrovým plánům (jako jsou vnořené studie případů a kontrol) a snahám o co největší kontrolu faktorů, které mohou zkreslit závěry (bias, confounding). Stále se rozšiřuje škála používaných modelů [49]. Rozpory vznikají z faktu, že statistika neumí prokázat kauzalitu, ale přitom v epidemiologii je to jeden z hlavních cílů [62].

7.2. **Detekce epidemií u diagnóz se sezónním chováním a postup využívající cenzorování dat.** Jedním z běžných pracovních nástrojů epidemiologa jsou analýzy příčin úmrtí a systémy rutinního sledování výskytu vybraných diagnóz. Častým úkolem, se kterým se tyto systémy setkávají, je studium časových trendů pro jednotlivé sledované diagnózy. Cílem je nejen nalezení případných časových trendů nebo cyklů, ale i pokus o predikci počtu výskytu onemocnění v blízké budoucnosti nebo upozornění, že pozorovaný počet nemocných je neobvykle velký (není jej možno vysvětlit trendy ani náhodným kolísáním). Bohatým zdrojem dat pro tyto analýzy je pro české epidemiology systém povinných hlášení infekčních onemocnění (EpiDat) [92]. Dalším datovým zdrojem je systém hlášení akutních respiračních onemocnění (ARO).

Nejprve si krátce ukážeme jednoduchou možnost analýzy dat ze systému EpiDat. Ten poskytuje týdenní počty hlášení velkého počtu nakažlivých onemocnění, dokonce i v jednotlivých okresech. Většina diagnóz má charakter sezónního výskytu, často se objevuje i dlouhodobý trend. Dalším faktorem je skutečnost výskytu epidemií — časově a prostorově zvýšeného výskytu počtu hlášení sledované diagnózy.

7.3. **Analýza hlášení systému EpiDat.** Pro analýzu časové řady příslušné diagnózy máme k dispozici počty hlášení Z_i v jednotlivých týdnech, kde i ($i = 1, \ldots, N$) je pořadové číslo týdne za celé sledované období. Pro jednoduchost uvažujeme, že každý rok má právě 52 týdnů. Reálná data, se kterými pracujeme, jsou již tak zatížena velkou chybou na přelomu roku. Je to způsobeno tím, že jednotlivé lékařské praxe mají v době konce roku silně omezený provoz, ale i tím, že pacienti často odkládají návštěvu lékaře až na první pracovní dny nového roku.

Protože nejjednodušším přijatelným předpokladem rozložení počtu hlášení je Poissonovo rozložení, je vhodné nepracovat přímo s pozorovanými počty hlášení Z_i, ale s jejich transformací, např. odmocninovou $X_i = \sqrt{Z_i}$.

Jedna z vhodných cest analýzy takovéto časové řady vede na sezónní modely ARMA či ARIMA. Pro epidemiology však není hlavním cílem modelování náhodné složky sledované veličiny, ale spíše modelování vlastního trendu. O to se pokouší následující přístup.

Hodnoty X_i rozložíme na čtyři základní složky:

$$X_i = T_i + V_i(S_i + \varepsilon_i) \quad ,$$

kde T_i je složka vyjadřující dlouhodobý trend, V_i její míra variability (reziduální směrodatná odchylka), S_i kolísání způsobené sezónními vlivy a ε_i náhodné kolísání. Zastavme se na chvíli u jednotlivých složek.

Dlouhodobý trend T_i může být modelován pomocí libovolné funkce nebo pomocí klouzavého průměru o délce jednoho roku — 52 týdnů (pokud chceme jako v našem případě, aby dlouhodobý trend byl oproštěn od roční periodicity). Pro odhad trendu T_i a V_i v i-tém časovém okamžiku tedy můžeme použít průměr k hodnot „okolo“ i-tého bodu pro k z intervalu $[\frac{52}{2}, N - \frac{52}{2} + 1]$. Na konci a na začátku celé řady můžeme použít např. hodnoty vypočtené pro týden $\frac{52}{2}$, resp. $\frac{N-52}{2}$.

Další problém, který musíme řešit, je nalezení periodických změn S_i sledované veličiny. Nejprve je nutno z analyzované řady odstranit nalezený trend. Označme

$$Y_i = \frac{X_i - T_i}{V_i}.$$

Pro modelování periodicity můžeme opět použít libovolné periodické matematické funkce.

Jinou možností odhadu sezónní složky o dané délce cyklu je odhadnout pro $j = 1, \ldots, m$ (kde m je počet roků, za které máme data) hodnotu S_i jako průměr Y všech stejných týdnů. Často je ale požadován odhad, který má „vyhlazený průběh“. Jako vhodná možnost se jeví jádrový odhad (například s gaussovským jádrem N(0,1)). Uvažujme pouze jádra, která jsou pro vzdálenější hodnoty nulová. Označme w_l váhy takového jádra, kde $w_l = w_{-l}$ a $w_l = 0$ pro $l > 5$. Dále zaveďme funkci $t(i) = i \bmod 52$ označující pořadí týdne v roce, které odpovídá i-tému pozorování.

Odhad S_k pro $k = 1, \ldots, 52$ vypočteme jako

$$\frac{\sum_{i=1}^{N} w_{k+t(i)} Y_i}{\sum_{i=1}^{N} w_{k+t(i)}}.$$

Takto jsme získali odhady všech komponent (T_i, V_i a S_i). Není problém získat odhad $\hat{X}_i$ a zpětnou transformací i odhad $\hat{Z}_i$ počtu hlášení v příslušném týdnu a roce.

Za předpokladu Poissonova rozložení pak snadno získáme i jednostranné toleranční intervaly, které mohou sloužit k detekci epidemií (odlehlých hodnot) tak, že hodnoty nad takto získanou toleranční mezí jsou tak velké, že je považujeme za signál výskytu epidemie.

Nástroj pro výpočet těchto odhadů jsme vytvořili v prostředí volně šiřitelných programů EpiInfo a EpiMap za pomoci dalších nástrojů připravených v jazyce C++.

7.4. **Analýza týdenních hlášení akutních respiračních onemocnění.** Jedním ze systémů rutinních hlášení je systém týdenního sledování výskytu akutních respiračních onemocnění. Tento systém není provozován celoplošně, ale každý okres podává hlášení počtu onemocnění v přibližně 20% výběru příslušného okresu. Výběr je konstruován tak, aby jej bylo možno považovat za reprezentativní. Při modelování standardizované nemocnosti jsme vycházeli z následujících úvah:

- Počty onemocnění jsou tvořeny jednak obvyklým sezónním výskytem a jednak případným epidemickým výskytem.
- K epidemickému zvýšení výskytu onemocnění dochází maximálně v 10% týdnů.
- Obvyklý sezónní výskyt má až na posunutí stejný průběh ve všech lokalitách.
- Rozložení obvyklého sezónního výskytu je Poissonova typu.

Označme $X_{i,j,k}$ relativní počet hlášení akutních respiračních onemocnění v kraji i, týdnu j a roce k na 100 000 obyvatel. Dále označme symbolem $Y_{i,j,k}$ logit

$$Y_{i,j,k} = \log\left(\frac{X_{i,j,k}}{100000 - X_{i,j,k}}\right) \quad .$$

Naším cílem bylo nalézt model pro obvyklý sezónní výskyt a vytvořit tak nástroj pro detekci epidemií (odlehlých hodnot). Základní myšlenkou je použití analýzy rozptylu dvojného třídění cenzorovaných dat (kde faktory reprezentují týden a kraj) pro logity standardizovaného počtu onemocnění na 100 000 obyvatel v příslušném kraji a týdnu. (Pokud jsme ale použili všechna $Y_{i,j,k}$ k odhadu paranetrů pomocí běžného modelu ANOVA, je získaný odhad nadhodnocen.) Cenzorování dat uvažujeme v tom smyslu, že běžný výskyt je „maskován“ zvýšením hodnot v době epidemie, víme tedy tedy v době výskytu epidemie pouze to, že pozorovaná hodnota $Y_{i,j,k}$ představuje pro běžný sezónní výskyt pouze omezení shora. Tento odhad v modelu ANOVA pro zleva (zdola) cenzorovaná data provedeme pomocí metody maximální věrohodnosti za předpokladu normality veličiny $Y_{i,j,k}$, viz například [12] nebo [63].

Jednou z možností by bylo použít od epidemiologů získanou informaci, vztahující se ke konkrétnímu týdnu a kraji, o tom, zda nastala epidemie nebo ne, a pak jednorázově provést odhad modelu. Tento přístup ale není možný, protože takovouto informaci nemáme a museli jsme proto výskyt těchto epidemií odhadovat. Nabízí se nicméně přirozená možnost považovat za epidemii období s tou kombinací indexů i, j, k, pro kterou se v modelu objevují velká rezidua, a postupně (v jednotlivých iteracích) příslušná $Y_{i,j,k}$ označovat za cenzorovaná zleva. (V období, kdy nastala epidemie, známe nemocnost, která se skládá jednak z nemocnosti běžné pro tento týden v roce a kraj, a jednak z nemocnosti vzniklé z důvodů epidemie). Tyto iterace lze provádět tak dlouho, dokud nebudou kladná rezidua „dostatečně malá“, v našem případě menší než zvolený kvantil normálního rozložení. Popišme si zmiňovaný iterační postup přesněji:

(1) Nejprve označíme všechna $Y_{i,j,k}$ za necenzorovaná.
(2) Vypočteme analýzu rozptylu dvojného třídění (s využitím metody pro zleva cenzorovaná data [63]) v modelu:

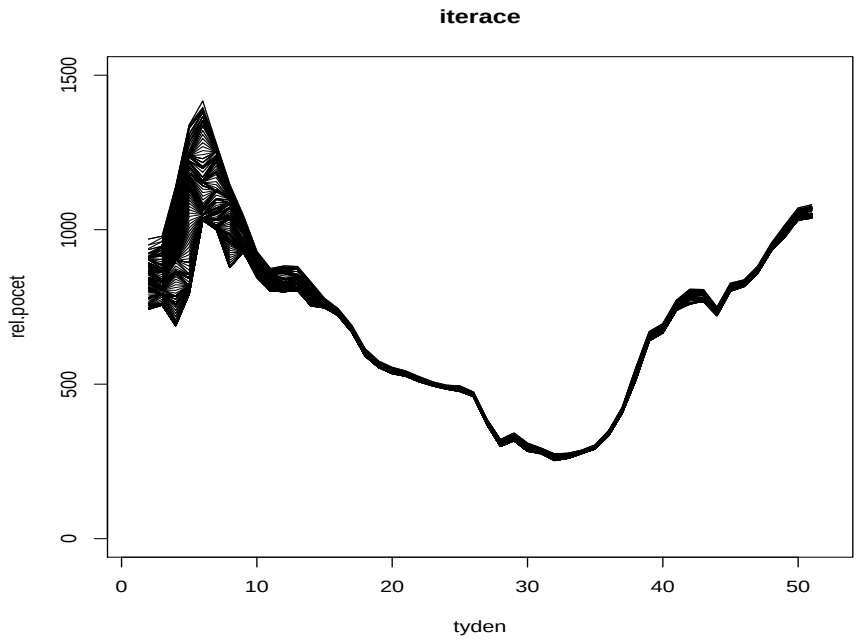
$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j,k} \quad ,$$

kde μ je absolutní člen, α_i koeficient kraje , β_j koeficient týdne v kalendářním roce a $\varepsilon_{i,j,k}$ náhodný šum nezávislý na týdnu ani na lokalitě.
(3) Nalezneme největší reziduum r_M necenzorovaných hodnot ve vypočteném modelu a označíme hodnotu odpovídající tomuto reziduu jako cenzorovanou zleva.
(4) Zjistíme, zda r_M je větší než 95% kvantil normálního rozložení $u_{0,95}$ — pokud tomu tak je, označíme hodnotu $Y_{i,j,k}$, odpovídající r_M jako cenzorovanou zleva a pokračujeme bodem 2, jinak ukončíme iterační proces.

Praktické výpočty byly provedeny programem R projektu GNU s použitím knihovny Survival.

Použití tohoto přístupu nám umožní využít předpoklad, že očekávaný počet onemocnění v konkrétním týdnu roku při neepidemickém výskytu je v daném kraji stejný v různých letech, a předpoklad, že průběh při neepidemickém výskytu je v jednotlivých krajích „paralelní“.

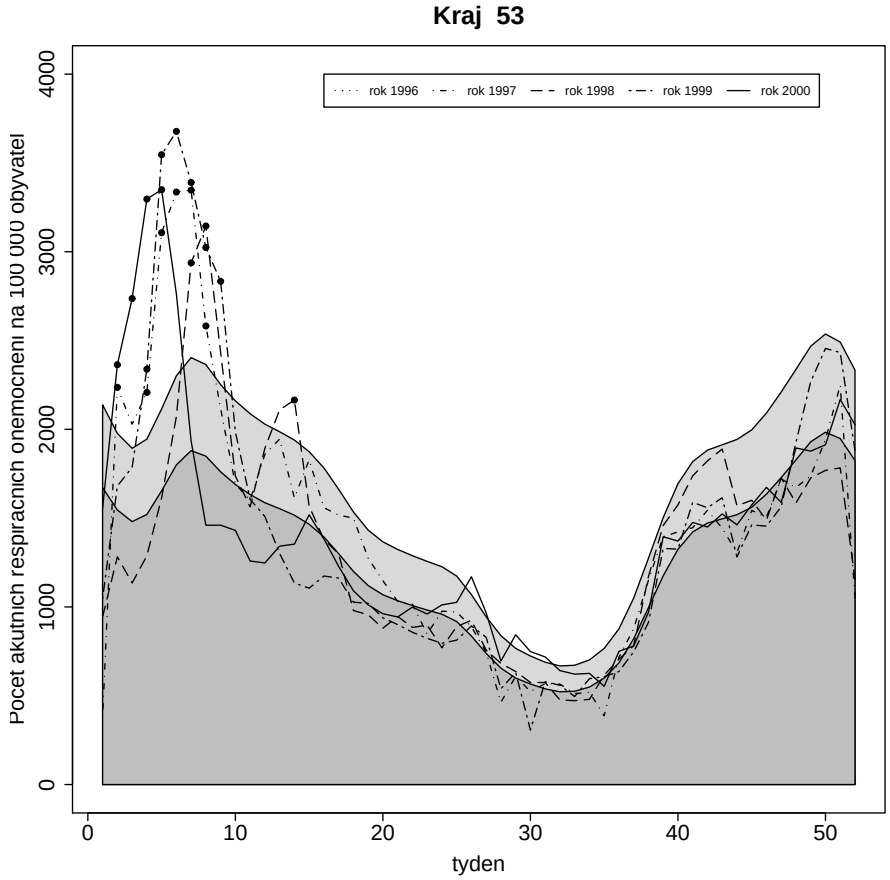
Tento výpočet jsme použili k analýze dat o počtech akutních respiračních onemocnění v letech 1996–2000. Získané odhady byly dále použity pro modelování sezónního výskytu jak pro každý kraj, tak i pro celou ČR. Pro grafické zobrazení ročních průběhů nemocnosti jsme dále průběh ještě vyhladili. K tomuto účelu jsme použili jádrový odhad s gaussovským jádrem, samozřejmě na začátku a konci roku jsou využity hodnoty z opačného konce roku.



Obr. 4 *Akutní respirační onemocnění: jednotlivé iterační kroky — odhady počtu onemocnění v jednotlivých týdnech.*

Pro detekci epidemických výskytů akutních respiračních onemocnění jsme pak použili jednostranné toleranční meze pokrývající 90% výskyt onemocnění za předpokladu, že nedošlo k epidemickému výskytu onemocnění.

Na Obr. 4 je zobrazen průběh iteračního procesu pro jednotlivé týdny roku. Každá lomená čára zobrazuje odhad $\mu + \beta_j$ v určité iteraci — je to vlastně odhad nemocnosti v jednotlivých týdnech bez korekce pro konkrétní kraj. Názorně je vidět, že výsledný odhad se od počátečního odhadu podstatně liší především v počátku roku. To odpovídá i představám epidemiologů: v letních měsících nedochází k častému výskytu akutních respiračních onemocnění, jejich výskyt se mezi jednotlivými roky nemění. Na konci roku sice dochází k častému výskytu infekcí, ale jsou způsobeny nechřipkovými viry a tato nemocnost je opět stejná v různých letech. Oproti tomu začátkem roku dochází k epidemiím, které mají každý rok jiný průběh.



Obr. 5 *Akutní respirační onemocnění: data vybraného kraje, vytvořený model a mez detekující epidemický výskyt onemocnění.*

Na Obr. 5 je zobrazen praktický výstup pro jeden z krajů (Pardubický kraj). Horní hranice tmavě šedého pásu představuje model — odhad nemocnosti v jednotlivých týdnech roku, horní hranice světle šedého pásu je jednostranná horní toleranční mez, jednotlivé čáry pozorované hodnoty a tučné body označují týdny, které byly uvažovány jako cenzorované.

8. Téma 7: Analýza přežívání a klinické pokusy

V úvodu předchozího tématu jsme zmínili observační studie. Počátky epidemiologických studií však souvisejí s vědeckým experimentem, tj. snahou vyhodnotit vliv nějaké léčby či zdravotně preventivního opatření. Takové experimentální studie mohou mít např. podobu klinických pokusů (kdy zkoumanými subjekty jsou pacienti) nebo terénních pokusů (kdy se intervence — hlavně preventivní — týká obyvatel určité komunity). Biostatistické postupy mají zásadní důležitost pro navržení takového plánu experimentální studie, který umožní správné vyhodnocení provedeného zdravotnického zásahu. Randomizovaný klinický pokus je považován za jednu z vůbec nejdůležitějších biostatistických metodologií [58, 89]. Role statistiků v intervenčních studiích výrazně roste [109]. Schémata pokusů se stále vyvíjejí a modifikují s cílem korigovat negativní vliv chybějících pozorování a různých typů možného zkreslení (bias).

Vznikají nové přístupy k hodnocení pokusů. Např. myšlenka *intention-to-treat* spočívá v tom, že všichni jedinci zařazení do určité větve randomizovaného klinického pokusu jsou skutečně dohromady hodnoceni (jako reprezentanti této větve) bez ohledu na to, zda nakonec pokus dokončili či prošli určenými procedurami a léčbou [53, 102]. V tom spočívá rozdíl oproti klasickému přístupu, kdy se za relevantní pro posouzení léčby bere jen informace o těch pacientech, kteří pokus dokončili. Jde v zásadě o rozumně konzervativní způsob použití dat motivovaný snahou zabránit vychýlení závěrů ve prospěch testované terapie. Například se ze zpracování nevylučují ani pacienti, u kterých léčba viditelně selhala. Zkoumání otázky, do jaké míry používat v analýze informaci od nedokončivších pacientů, je stále předmětem diskusí [120].

Je také potřeba reagovat i na velmi citlivé etické otázky, které se při srovnání dvou substancí, resp. substance a placeba objevují v souvislosti s požadavkem na včasné rozhodnutí o zkoumané hypotéze (ať už se jedná o její přijetí či zamítnutí). Proto by měla být věnována pozornost adaptivním schematům a sekvenčním plánům pokusů [88], které zpravidla mají různé přednosti ve srovnání s klasickými pokusy. Při aplikaci těchto metod nejsou takové charakteristiky pokusu, jako je celkový počet pozorování nebo relativní četnosti pacientů léčených porovnávanými metodami, stanoveny předem, ale rozhoduje se o nich teprve v průběhu experimentu na základě dosavadních dat. Ke statisticky korektnímu rozhodnutí o tom, který způsob léčení zasluhuje přednost, potřebují méně dat, popř. vystavují menší počet osob terapii, která se ukazuje jako pro pacienta nevýhodná.

Jednou z metodik úzce svázaných s klinickými pokusy je analýza přežívání [48]. V historii analýzy přežívání jsou dva články, které patří k nejcitovanějším statistickým pracím vůbec — Kaplanova-Meierova charakterizace funkce přežívání [70] a Coxův model proporcionálního rizika [29]. Dalším klíčovým bodem v této oblasti bylo poznání, že čítací procesy a martingaly jsou mocným nástrojem při studiu vlastností procesů přežívání [1, 46]. Od těchto základních myšlenek se odvíjejí a nepochybně i budou odvíjet žhavá témata této oblasti biostatistiky. Např. v souvislosti s Coxovým modelem se studuje problematika časově závislých doprovodných proměnných a důsledky porušení předpokladu proporcionality, či obecněji nedostatečné znalosti biologických mechanismů [3]. Zajímavou alternativou Coxova modelu je pak Aalenův neparametrický *aditivní regresní model* [2, 12]. Jsou zaváděny verze modelů známých z jiných aplikačních oblastí, zejména kontroly jakosti, např. *model se změnou měřítka (accelerated life model)*, kdy se vliv prediktorů modifikuje změnou měřítka časové osy („zrychlením či zpomalením běhu času“), nebo *model konkurujících rizik (competing risks model)*, který předpokládá, že v důsledku události jednoho typu již subjekt nemůže být v riziku události jiného typu [87]. Může jít např. o jedince ve studii o kouření a rakovině plic, který zemře na koronární onemocnění.

Analýza přežívání se ovšem používá i mimo rámec klinických pokusů, třeba v neexperimentálních studiích. Specifické rysy takových aplikací shrnuje [25].

9. Téma 8: Vývoj léků

9.1. **Biostatistická problematika při vývoji léků.** Specifickou oblastí, v níž uplatnění a používání biostatistických metod v posledních desetiletích silně vzrostlo, je vývoj léků a otázky jeho řízení [94, 102]. Na jedné straně se tady statistika účastní procesů, které se týkají i obchodní politiky a mají aspekty regulační (tato oblast biostatistiky je asi nejvíce svázána různými normami a standardy práce [86]). Na druhé

straně zde má statistika velký prostor při hodnocení pokusů spojených s vývojem nových léků, tj. s ověřováním specifické účinnosti, akutní a pak chronické toxicity na pokusných zvířatech, a konečně při klinických pokusech, čímž toto téma úzce souvisí s předchozím. Nový přípravek postupně prochází několika klinickými pokusy, které odpovídají různým stádiím testování od počátečního studia mechanismu působení a stanovení bezpečné dávky přes testy funkčnosti a bezpečnosti na nemocných osobách až po porovnání účinku s kontrolním ošetřením a posléze dlouhodobé sledování přípravku po jeho schválení k používání [12]. Každé stádium testování vyžaduje použití adekvátních biostatistických postupů.

Z problémů, kterým je věnována intenzivní pozornost, uveďme alespoň otázky uspořádání pokusu (složité křížové pokusy, multicentrické studie s kontroverzními pohledy na důležitost interakce ošetření–centrum, ekvivalenční studie, průkaz superiority přípravku) a otázky analýzy (volba výstupní proměnné, volba analyzované populace). Stále častější příklon ke klinickým pokusům řešícím souběžně několik pracovních hypotéz přináší varianty problematiky mnohonásobného testování (viz téma 10). K posunu ve volbě analyzované populace vede již zmíněný přístup *intention-to-treat*, který je nyní regulačními orgány považován za jediný bezpečný [74]. Výrazně roste zájem o hodnocení účinků léku nejen ze strany zadavatele a lékařů, ale i ze subjektivního pohledu pacientů (např. měření kvality života). K nahrazení klinicky relevantních proměnných charakterizujících výsledek pokusu, které však nelze z věcných či technických důvodů změřit, slouží tzv. zástupné (surrogate) proměnné. Aby byly použitelné, musí nejen dobře předpovídat výsledek, ale musí i dobře a v celé úplnosti postihovat efekt léčby [91]. Jak ukázaly různé příklady z praxe, může se jinak snadno stát, že zástupná proměnná neplní svou roli správně.

Specifické nároky na statistiky klade studium farmakokinetiky (kompartmentové modely) [77] a farmakoekonomiky (metody cena-zisk/cost-benefit, cena-účinnost/cost-effectiveness, cena-prospěch/cost-utility).

9.2. **Bioekvivalence.** Dříve, než je nově vyvinutý lék připuštěn do distribuce, předchází téměř pětileté období zkoušení jeho specifické účinnosti a biologické nezávadnosti, což představuje značné finanční náklady. Poměrně důležitou úlohou je inovace již zavedených léků (např. tvar a velikost pilulky, její zabarvení nebo chuť spojená často i se změnou obchodního názvu léku). Pro státní kontrolní orgány (v USA např. Food and Drug Administration) představují i tyto inovace nové léky, k jejichž zavedení by vlastně také měly být prováděny obdobné nákladné testy.

Pro inovované léky, jejichž základní chemická substance je již v praxi ověřena co do svého terapeutického účinku, se zavedl jiný typ klinického testu, v němž se nezkouší vlastní účinnost inovovaného léku na nemocných lidech, nýbrž se na zdravých dobrovolnících stará, referenční forma léku porovnává s novou, testovanou. Nesrovnávají se specifické účinky obou forem, nýbrž jejich schopnost proniknout do lidského těla, a zejména na místa jejich působení. Vzhledem k tomu, že jsou tyto léky většinou „dopravovány“ na tato místa krví, zjišťuje se časový průběh jejich koncentrace v krvi. Za měřítko jejich proniknutí je pak nejčastěji považována plocha pod časovou křivkou jejich koncentrace (Area Under the Curve — AUC, respektive její logaritmus).

Stanovení této plochy je poněkud komplikované, protože obvykle se nedá koncentrace v krvi měřit kontinuálně a jako aproximace křivky se užívá součet koncentrací naměřených v krvi v pravidelných časových intervalech. Pro farmakology je zajímavá i maximální koncentrace a čas, za jaký k ní došlo, ale tyto údaje jsou ještě hůře odhadnutelné než plocha pod křivkou.

Test na dobrovolnících se obvykle provádí jako tzv. zkřížený pokus, kdy každý dobrovolník dostane ve dvou opakováních jak testovaný, tak referenční lék, přičemž u jedné části dobrovolníků se začíná referenční a u druhé testovanou substancí. Oba léky jsou považovány za bioekvivalentní, jestliže jsou křivky koncentrací pro oba léky obdobné, resp. jsou-li obdobné hodnoty AUC (nebo jejich logaritmy).

Označme Y náhodnou proměnnou, jejíž empirické hodnoty jsou v klinickém pokusu zjišťovány. Vzhledem k obvyklému logaritmicko-normálnímu rozložení proměnné AUC je nejčastěji $Y = \ln(AUC)$. Při běžném testování shody bývá testována nulová hypotéza

$$H_0 : E(Y_T - Y_R) = \mu_T - \mu_R = \theta = 0,$$

kde indexy T a R označují testovanou a referenční lékovou formu. Tato hypotéza může být zamítnuta i tehdy, když rozdíl obou parametrů pro referenční a testovaný lék je z praktického hlediska zanedbatelný. Obdobnost (bioekvivalence) obou léků se definuje tak, že parametr θ leží v předem daném intervalu $(-\Delta_1, \Delta_2)$. Obvykle je $\Delta_1 = \Delta_2 = \Delta$. Státní řídící orgány pak určují hodnotu Δ. V USA je zadáno $\exp(\Delta) = 1.25$. Z hlediska takto definované bioekvivalence tedy není nutná naprostá shoda obou léků, ale velmi obdobný způsob jejich distribuce v organizmu po aplikaci léku.

V analýze výsledků pokusu na dobrovolnících se odhaduje i testuje parametr θ. Cílem testu je, jako obvykle ve statistice, zamítnout nulovou hypotézu, zde *neplnění bioekvivalence*. Nulová hypotéza se tedy definuje jako sjednocení dvou dílčích nulových hypotéz

$$H_0 = H_{01} \cup H_{02},$$

kde

$$H_{01} : \theta \leq -\Delta \qquad \text{a} \qquad H_{02} : \theta \geq \Delta.$$

K těmto nulovým hypotézám existují příslušné alternativní hypotézy H_{A1} a H_{A2}, takže

$$H_A = H_{A1} \cap H_{A2} : \quad -\Delta < \theta < \Delta.$$

Provede se simultánní test pro obě dílčí nulové hypotézy (obvykle dva jednostranné t-testy). Pokud je nulová hypotéza zamítnuta, považuje se bioekvivalence obou srovnávaných léků za prokázanou, a to na hladině významnosti, která je rovna větší z obou použitých hladin významnosti dílčích testů. Tento postup je ekvivalentní se stanovením $100 \cdot (1 - 2\alpha)$ procentního intervalu spolehlivosti pro odhad parametru θ. Pokud je např. 90% interval spolehlivosti uvnitř „regulačního“ intervalu $(-\Delta, \Delta)$, je hypotéza o „neekvivalenci“ zamítnuta na 5% hladině významnosti.

Výše popsaným způsobem se testuje tzv. *průměrná ekvivalence* obou léků. Pozornost je věnována pouze rozdílu středních hodnot, neuvažuje se možný vliv rozdílných rozptylů veličiny Y pro oba srovnávané léky na jejich efekt v předpokládané populaci příštích uživatelů léku.

Populační bioekvivalence se považuje za platnou, pokud jsou průměrné hodnoty obdobné, a pokud navíc je rozptyl dat u testovaného léku menší nebo stejný jako u léku referenčního. Hauck a spol. [59] zavádějí složitější nulovou, a tedy i alternativní hypotézu pro ověření bioekvivalence. Označme σ_T a σ_R standardní odchylky proměnné Y pro testovaný a referenční lék. Nulová a alternativní hypotéza se formulují takto:

$$H_0 : \quad \theta \leq -\Delta \quad \vee \quad \theta \geq \Delta \quad \vee \quad \sigma_T/\sigma_R \geq \varepsilon,$$
$$H_A : \quad -\Delta < \theta < \Delta \quad \wedge \quad \sigma_T/\sigma_R < \varepsilon.$$

Test této poměrně složité nulové hypotézy se dá provést metodou maximální věrohodnosti. Postupný test pro obdobnou hypotézu uveřejnil Vuorinen a spol. [113].

Individuální ekvivalence souvisí s takzvanou zaměnitelností, tj. možností nahradit starou formu léku formou novou u individuálního pacienta. Je tedy potřeba prokázat více než jen obdobnost průměrných hodnot; požaduje se, aby si hodnoty odpovědi na testovaný a referenční lék byly dostatečně blízké u většiny jedinců. Zkoumán individuální bioekvivalence je složitější jak z hlediska plánu studie, tak z hlediska analýzy dat. Veličiny Y_T a Y_R jsou u téhož pacienta navzájem závislé. Pro řádné posouzení individuální bioekvivalence je třeba odlišit rozptyl mezi osobami a rozptyl uvnitř osob a v analýze se soustředit pouze na rozptyl uvnitř osob. Možnou metodou je dle [106] provedení tří opakovaných pokusů na každé pokusné osobě, referenční forma léku se aplikuje dvakrát. Postupuje se tak, že se testuje, zda odchylka testované od referenční formy léku je jen zanedbatelně větší nežli pozorovaný rozdíl mezi opakovanými testy pro referenční formu léku. Autoři doporučují využít při konstrukci rozhodovacího pravidla metody bootstrapu.

Závěrem lze shrnout, že základem testování bioekvivalence je postup ne běžný při aplikaci statistických metod ve stanovení nulové, a tudíž i alternativní hypotézy [103, 108, 39]. V obvyklých statistických testech je nulovou hypotézou předpoklad, že sledované proměnné žádným kontrolovaným vlivem ovlivněny nejsou (anglický výraz „null“ znamená „nic, bez vlivu“). Cílem experimentátorů i hodnotitelů — biostatistiků je tento předpoklad zamítnout. Při testování bioekvivalence je naopak testovanou hypotézou, že oba léky se liší víc, než je zdrávo, a její zamítnutí je pak žádoucím potvrzením, že obě porovnávané formy *téhož* léku se podstatněji neliší — alespoň co do jeho přenosu do krevního řečiště po předchozí aplikaci stejně velkých dávek. To pak stačí k tomu, aby byla distribuce nové formy léku do lékáren povolena.

10. Téma 9: Biostatistik a ti druzí

V neposlední řadě je tématem, kterým se stále zabývá mnoho odborníků na celém světě, úloha a postavení biostatistiků

- v pregraduální i postgraduální výuce mediků a lékařů i biologů [9, 78, 71],
- při konzultační činnosti [79, 68, 52, 19, 100],
- jako členů nejrůznějších, často i velmi různorodých vědeckých týmů [90, 20],
- v publikačním procesu (kterého se účastní jako autoři i jako recenzenti [44, 10]).

Také encyklopedie [12] věnuje každému z těchto bodů několikastránkové heslo.

Pocock [90] rozlišuje tři hlavní kategorie odborných aktivit biostatistika: *statistické konzultace* (úkoly, při nichž je statistik limitován vymezeným časem a záběrem, i když mohou sahat od přípravy experimentu až ke zpracování a publikaci výsledků), *statistickou spolupráci* (projekty, v nichž je statistik integrálním členem vědeckého týmu a zná detailněji biologickou a lékařskou podstatu problematiky) a *metodologický výzkum* (případy originálních aplikací známých přístupů a metod v otázkách plánování či analýzy studie; případně vypracování nového postupu). Podíl času, který biostatistik věnuje jednotlivým typům těchto aktivit, silně závisí na potřebách spolupracovníků a instituce, pro kterou pracuje.

K prvořadým úkolům biostatistiků patří péče o zlepšování mezioborové spolupráce a pedagogická a osvětová činnost. Jejím cílem by mělo být, aby si nejen lékaři a biologové, ale i obecná populace, více osvojili základní principy statistického uvažování a aby lépe chápali statistické postupy. Špatná interpretace těch nejzákladnějších

statistických pojmů je bohužel velmi častá. Podobně je tomu s používáním neadekvátních metod (t-test se používá ve všech možných i nemožných situacích), které je s rostoucí dostupností statistického software stále častější.

11. Téma 10: Kontroverzní místa biostatistické praxe

Mezi dlouhodobě aktuální témata lze zařadit i několik okruhů, které mají spíše negativní publicitu, někdy tak velkou, že je van Houwelingen [64] neváhá označit za noční můry.

Biostatistici opakovaně poukazují na přehnaný důraz, který je často při hodnocení lékařských a biologických dat kladen na formální statistické testování hypotéz bez ohledu věcnou/klinickou významnost zjištěných rozdílů, na nadužívání *p-hodnoty* [8] (která je navíc mnohdy jediným výstupem) a na nutnost uvádět i intervaly spolehlivosti [50, 111]. Ty však mohou odhalit, jak obrovsky variabilní jsou data a jak nepřesné jsou proto bodové odhady. A mnohým lékařům i biologům stále činí interpretace intervalů spolehlivosti potíže. Často bývá p-hodnota chybně interpretována jako numerická míra významnosti a někteří uživatelé statistiky mají tendenci podle ní řadit a porovnávat výsledky různých testů [43]. Úplně se zapomíná na základní principy statistického testování a volbu hladiny významnosti před testem.

Nejen statistický, ale i filozofický náboj má téma *mnohonásobného testování* [112]. V lékařské literatuře lze zachytit široké spektrum názorů od nutnosti používat korekce Bonferroniho typu [17] až po jejich úplné odmítnutí [95]. Otázka mnohonásobného testování vyvstává např. u klinických pokusů s několika paralelními cíli (endpoints) [122, 88], které jsou ze statistického pohledu představovány proměnnými, jež jsou (do jisté míry) závislé. Setkáváme se s ní ale často i v běžné praxi při požadavku provést jednorozměrné porovnání mezi skupinami zvlášť pro každou z mnoha proměnných.

Vyloženě módním pojmem posledních let se stala *meta-analýza* [61, 85, 34], tedy snaha o integraci závěrů více samostatných studií. Idea kombinování informace z několika studií stejného typu, která stála u zrodu meta-analýzy (jako vyšší formy „systematického přehledu“), je jistě přínosná, protože umožní sjednocení velkého množství dosud roztříštěných informací, zvýšení přesnosti odhadů a zkrácení doby potřebné pro rozhodnutí o hypotéze. Vznikají však také velké otazníky, které úzce souvisejí s praktickou realizací metody. Správně provedená meta-analýza je technicky a časově velmi náročná, protože třeba jen vyhledávání všech relevantních studií tak, aby nedocházelo k publikačnímu zkreslení, je velmi složité. Je známo, že výsledky studií, které neprokázaly statistickou významnost, mají horší publikační možnosti, jsou publikovány v horších časopisech, později, nebo vůbec ne, a proto jsou hůře k nalezení. Problémem je i stanovení, které studie lze kombinovat, a které už ne (roli hraje typ i kvalita studie). Většina meta-analýz vychází z publikovaných sumárních charakteristik a nikoli z individuálních údajů, jejichž kombinace by byla přesnější.

Další trvalým a velmi rozšířeným problémem, objevujícím se zejména v epidemiologických aplikacích, je *dichotomizace spojitých veličin*, případně redukce většího počtu kategorií na dvě [8]. Děje se tak často pod záminkou přehlednějšího popisu a interpretace dat. Pro čistě deskriptivní účel je snad tato praxe přijatelná, ale z hlediska statistického testování jistě ne. Podle Cohena [27] je umělá ztráta informace v důsledku dichotomizace spojité proměnné ekvivalentní zahození třetiny dat. Dělicí bod bývá navíc stanovován až na základě pozorovaných dat, často se zkouší i několik různých podle hesla „co nejlépe vyjde“. Problém dichotomizace se tak propojuje s problematikou mnohonásobného testování i s tématem 1.

Velmi frekventovaným problémem je bohužel *chybná či neúplná specifikace použitého modelu* a nedostatečná kontrola jeho vlastností. Často souvisí se statistickým softwarem — používají se takové modely, které jsou v něm dostupné, respektive takové metody, které je lékař sám schopen použít bez pomoci statistika.

12. Závěr

Jaká je budoucnost biostatistiky a biostatistiků? Převládá názor, že význam biostatistických metod dále poroste a že je stále velký prostor pro zlepšování dosavadních postupů, nejen v podobě statistické metodologie, ale i v podobě organizace databází a sběru dat. Van Houwelingen [64] si ve své vizi představuje biostatistika budoucnosti jako toho, kdo modeluje individuální šance/rizika pacienta v realistickém modelu, tj. modelu, který je schopen optimálně diskriminovat mezi jedinci na základě veškerých informací o pacientech a na základě aktuálně dosaženého lékařského poznání.

V tomto příspěvku jsme se věnovali značně různorodým tématům a zazněly jak pasáže „povídavé“, tak i více matematické, jak obecné, tak i řešící konkrétní problém. I když to může působit dojmem určité roztříštěnosti, domníváme se, že je to zároveň ilustrací, jak široký je záběr pojmu biostatistika a jak široké je spektrum biostatistiků a jejich přístupů. Možná více než v jiných oborech je osobnost biostatistika formována nejen vzděláním, ale hlavně praxí a úkoly, které řeší.

Gehan [52] se domnívá, že bude docházet k ještě větší diferenciaci mezi biostatistiky věnujícími se metodologické práci (viz téma 9) a biostatistiky–techniky, jejichž hlavním úkolem bude porozumět reálnému problému a určit postup zpracování. Tak jako v jiných oborech se zřejmě biostatistici budou více specializovat pouze na určitou oblast aplikací (např. genetiku). O to více ceněný bude biostatistik se širokým záběrem, o to více bude potřebná spolupráce biostatistiků teoretičtěji zaměřených a biostatistiků orientovaných více k praxi. Nejen u nás, ale i ve světě je pociťován nedostatek biostatistiků [90] a současně i nedostatek vhodných a adekvátně oceněných pracovních míst pro ně. Zdá se, že potřeba biostatistiků všech kategorií poroste. Další rozvoj medicíny a biologie nepochybně podnítí vznik nových žhavých biostatistických témat, nicméně věříme, že většina zde zmíněných hlavních témat zůstane ještě dlouho ve středu pozornosti.

Literatura

[1] Aalen, O.O. (1975): Statistical inference for a family of counting processes. PhD thesis, University of California, Berkeley.

[2] Aalen, O.O. (1978): Non-parametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

[3] Aalen, O.O. (2000): Medical statistics — no time for complacency. *Statistical Methods in Medical Research* **9**, 31–40.

[4] Agresti, A. (1989): A survey of models for repeated ordered categorical response data. *Statistics in Medicine* **8**, 1209–1224.

[5] Agresti, A. (1990): Categorical Data Analysis. Wiley, New York.

[6] Agresti, A. (1999): Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* **18**, 2191–2207.

[7] Agresti, A. (2001): Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine* **20**, 2709–2722.

[8] Altman, D.G. (2000): Statistics in medical journals: some recent trends. *Statistics in Medicine* **19**, 3275–3289.

[9] Altman, D.G., Bland, J.M. (1991): Improving doctors' understanding of statistics. (With discussion.) *Journal of the Royal Statistical Society A* **154**, 223–267.

[10] Altman, D.G., Gore, S.M., Gardner, M.J., Pocock, S.J. (1983): Statistical guidelines for contributors of medical journals. *British Medical Journal* **286**, 1489–1493.

[11] Armitage, P. (2001): Theory and practice in medical statistics. *Statistics in Medicine* **20**, 2537–2548.
[12] Armitage, P., Colton, T., eds. (1998): Encyclopedia of Biostatistics. Wiley, Chichester.
[13] Armitage, P., David, H.A., eds. (1996). Advances in Biometry. Wiley, New York.
[14] Armstrong, B., Sloan, M. (1989): Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* **129**, 191–204.
[15] Balding, D.J., Bishop, M., Cannings, C., eds. (2001): Handbook of Statistical Genetics. Wiley, Chichester.
[16] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975): Discrete Multivariate Analysis: Theory and Practice. Massachusetts Institute of Technology, Cambridge, Mass.
[17] Bender, R., Lange, S. (2001): Adjusting for multiple testing — when and how? *Journal of Clinical Epidemiology* **54**, 343–349.
[18] Billard, L. (1994): The world of biometry. *Biometrics* **50**, 899–916.
[19] Boen, J.R., Zahn, D.A. (1997): Human Side of Statistical Consulting. Lifetime Learning Publications, Belmont.
[20] Breslow, N.E. (1978): Perspectives on the statistician's role in cooperative clinical research. *Cancer* **41**, 326-332.
[21] Breslow, N. (1990): Biostatistics and Bayes. *Statistical Science* **5** (3), 269–298.
[22] Breslow, N.E. (1996): Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14–28.
[23] Breslow, N.E. (2000): Statistics. *Epidemiologic Reviews* **22** (1), 126–130.
[24] Breslow, N.E. (2000): Statistics in the life and medical sciences. *Journal of the American Statistical Association* **95**, 281–282.
[25] Bull, K., Spiegelhalter, D.J. (1997): Survival analysis in observational studies. *Statistics in Medicine* **16**,1041–1074.
[26] Carroll, R., Ruppert, D., Stefanski, L.A. (1995): Measurement Error in Nonlinear Models. Chapman and Hall, London.
[27] Cohen, J. (1983): The cost of dichotomization. *Applied Psychological Measurement* **7**, 249–253.
[28] Cox, C. (1988): Multinomial regression models based on continuation ratio. *Statistics in Medicine* **7**, 435–41.
[29] Cox, D.R. (1972): Regression models and life tables. (With discussion.) *Journal of the Royal Statistical Society B* **34**, 187–220.
[30] Cox, D.R. (1975): Partial likelihood. *Biometrika* **62**, 629–76.
[31] Cox, D.R. (1997): The current position of statistics: A personal view. (With discussion.) *International Statistical Review* **65**, 261–290.
[32] Cressie, N.A.C. (1993): Statistics for Spatial Data. Wiley, New York.
[33] Davison, A.C. (2001): Biometrika Centenary: theory and geneal methodology. *Biometrika* **88**, 13-52.
[34] Dickersin, K. (2002): Systematic reviews in epidemiology: why are we so far behind? *International Journal of Epidemiology* **31**, 6–12.
[35] Diggle, P.J., Liang, K., Zeger, S.L. (1999): Analysis of Longitudinal Data. Oxford University Press, New York.
[36] Disman, M. (2000): Jak se vyrábí sociologická znalost. 3. vyd. Nakladatelství Karolinum, Praha.
[37] Dominici, F., Zeger, S.L., Samet, J.M. (2000): A measurement error model for time-series studies of air pollution and mortality. *Biostatistics* **1** (2), 157–175.
[38] Donnelly, C.A., Cox, D.R. (2001): Mathematical biology and medical statistics: contributions to the understanding of AIDS epidemiology. *Statistical Methods in Medical Research* **10**, 141–154.
[39] Dunnett, Ch.W., Gent, M. (1996): An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine* **15**,1729–1738.
[40] Elston, R.C., Olson, J.M., Palmer, L., ed. (2002): Biostatistical Genetics and Genetic Epidemiology. Wiley, Chichester.
[41] Everitt, B.S. (1995): The analysis of repeated measures: a practical review with examples. *Statistician* **44** (1), 113–135.
[42] Finney, D.J. (1990): Repeated measurements: what is measured and what repeats? *Statistics in Medicine* **9**, 639–644.
[43] Finney, D.J. (1994): On biometric language and its abuses. *Biometric Bulletin* **11** (4), 2–4. Český překlad: *Informační bulletin České statistické společnosti* **10** (3), 1998, 2–9.

[44] Finney, D.J. (1997): The responsible referee. *Biometrics* **53**, 715–719.
[45] Fitzmauritze, G.M., Laird, N.M., Rotnitzky, A.G. (1993): Regression models for discrete longitudinal responses. *Statistical Science*, **8**, 284–309.
[46] Fleming, T.R., Harrington, D.P. (1991): Counting Processes and Survival Analysis. Wiley, New York.
[47] Fleming, T.R. (1996): Surrogate endpoints in clinical trials. *Drug Information Journal* **30**, 545–551.
[48] Fleming, T., Lin, D.Y. (2000): Survival analysis in clinical trials: Past developments and future directions. *Biometrics* **56**, 971–983.
[49] Gail, M.H. (1991): A bibliography and comments on the use of statistical models in epidemiology in the 1980s. *Statistics in Medicine* **10**, 1819–1885.
[50] Gardner, M.J., Altman, D.G. (1986): Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal* **292**, 746-750.
[51] Gardner, M.J., Snee, M.P., Hall, A.J., Powell, C.A., Downes, S., Terrell, J.D. (1990): Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. *British Medical Journal* **300**, 423–429.
[52] Gehan, E.A. (2000): Biostatistics in the new millenium: a consulting statistician's perspective. *Statistical Methods in Medical Research* **9** (1), 3–16.
[53] Gillings, D., Koch, G. (1991): The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Information Journal* **25**, 411–424.
[54] Grant, G.R., Ewens, W.J. (2001): Statistical Methods in Bioinformatics: An Introduction. Springer, Berlin.
[55] Greenland, S. (1994): Alternative models for ordinal logistic regression. *Statistics in Medicine* **13**, 1665–77.
[56] Hagenaars, J.A. (1990): Categorical Longitudinal Data. SAGE Publications, Newbury Park.
[57] Harrell, F.E. (2001): Regression Modelling Strategies. Springer-Verlag, New York.
[58] Harrington, D.P. (2000): The randomized clinical trial. *Journal of the American Statistical Association* **95**, 312–315.
[59] Hauck, W.W., Bois, F.Y., Hyslop, T., Gee, L., Anderson, S. (1997): A parametric approach to population bioequivalence. *Statistics in Medicine* **14**, 441–454.
[60] Healy, M.J.R. (1979): Does medical statistics exist? *Bulletin in Applied Statistics* **6**, 137–182.
[61] Hendl, J. (2002): Meta-analýza v medicíně. *Časopis lékařů českých* **141** (8), 235–239.
[62] Holland, P.W. (1986): Statistics and causal inference. (With discussion and reply.) *Journal of the American Statistical Association* **81**, 945–970.
[63] Hosmer, D.W., Lemeshow, S. (1999): Applied survival analysis: Regression modeling of time to event data. Wiley, New York.
[64] van Houwelingen, H.C. (1997): The future of biostatistics: Expecting the unexpected. *Statistics in Medicine* **16**, 2773–2784.
[65] Hughes, M.D. (2000): Analysis and design issues for studies using censored biomarker measurements with an example of viral load measurements in HIV clinical trials. *Statistics in Medicine* **19**, 3171–3191.
[66] Hu, X., Lachin, J.M. (2001): Application of robust estimating equations to the analysis of quantitative longitudinal data. *Statistics in Medicine* **20**, 3411–3428.
[67] Huber, P.J. (1967): The behavior of maximum likelihood estimators under nonstandard conditions. In: LeCam, L.M., Neyman, J., eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 221–233. University of California Press, Berkeley.
[68] Hyams, L. (1971): The practical psychology of biostatistical consultation. *Biometrics* **27**, 201–211.
[69] Kahn, H.A., Sempos, Ch.T. (1989): Statistical Methods in Epidemiology. Oxford University Press, New York.
[70] Kaplan, E.L., Meier, P. (1958): Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.
[71] Khurshid, A., Sahai, H. (1993): A second bibliography on the teaching of statistics in biological, medical, and health sciences. *Statistica Applicata* **5**, 309-397.
[72] Laird, N.M. (1988). Missing data in longitudinal studies. *Statistics in Medicine* **7**, 305–316.
[73] Lawson, A. (2001): Statistical Methods in Spatial Epidemiology. Wiley, New York.
[74] Lewis, J.A. (1995): Statistical issues in the regulation of medicines. *Statistics in Medicine* **14**, 127–136.
[75] Liang, K., Zeger, S.L. (1986): Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

[76] Lindsey, J. (1993): Models for Repeated Measurements. Oxford University Press, Oxford.
[77] Lindsey, J.K., Jones, B., Jarvis, P. (2001): Some statistical issues in modelling pharmacokinetic data. *Statistics in Medicine* **20**, 2775–2783.
[78] Malý, M. (2001): Otázky spojené s výukou biostatistiky pro lékaře. *Lékař a technika* **32** (6), příloha I–IV.
[79] Malý, M., Roth, Z. (2001): Otázky komunikace statistika s lékařem. Zborník konferencie PRASTAN 2001, Kočovce, str. 98–103. Slovenská štatistická a demografická spoločnosť, Bratislava.
[80] McCullagh, P. (1980): Regression models for ordinal data. *Journal of the Royal Statistical Society B* **42**, 109–142.
[81] McCullagh, P. (1983): Generalized Linear Models. Chapman and Hall, London.
[82] Mehta, C.R., Patel, N.R. (1995): Exact logistic regression: theory and applications. *Statistics in Medicine* **14**, 2143–2160.
[83] Molenberghs, G., Williams, P.L., Lipsitz, S.R. (2002): Prediction of survival and opportunistic infections in HIV-infected patients: a comparison of imputation methods of incomplete CD4 counts. *Statistics in Medicine* **21**, 1387–1408.
[84] Muirhead, C.R. (1998): Childhood cancer and nuclear installations: a review. *Nuclear Energy* **37**, 371–379.
[85] Normand, S.-L.T. (1999): Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 321–359.
[86] North, P.M. (1998): Ensuring good statistical practice in clinical research: guidelines for standard operating procedures (an update). *Drug Information Journal* **32**, 665–682.
[87] Oakes, D. (2000): Survival analysis. *Journal of the American Statistical Association* **95**, 282–285.
[88] O'Brien, P.C., Fleming, T. (1979): A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.
[89] Piantadosi, S. (1997): Clinical Trials: A Methodologic Perspective. Wiley, New York.
[90] Pocock, S.J. (1995): Life as an academic medical statistician and how to survive it. *Statistics in Medicine* **14**, 209-222.
[91] Prentice, R.L. (1989): Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431–440.
[92] Procházka, B., Beneš, Č. (1999): Hodnocení časových trendů týdenních počtů onemocnění. *Epidemiologie, mikrobiologie a imunologie* **48** (2), 52–59.
[93] Richardson, S. (1996): Développements récents de la biostatistique. *Revue d'Epidemiologie et de Santé Publique* **44**, 482–493.
[94] Rockhold, F.W. (2000): Strategic use of statistical thinking in drug development. *Statistics in Medicine* **19**, 3211–3217.
[95] Rothman, K. (1990): No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43–46.
[96] Rothman, K., Greenland, S. (1998): Modern Epidemiology. 2nd ed. Lippincott-Raven, Philadelphia.
[97] Routledge, R.D. (1994): Practicing safe statistics with the mid-p. *Canadian Journal of Statistics* **22**, 103–110.
[98] Rubin, D.B. (1976): Inference and missing data. *Biometrika* **63**, 581–592.
[99] Rubin, D.B. (1996): Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–487.
[100] Sahai, H., Khurshid, A. (1999): A bibliography on statistical consulting and training. *Journal of Official Statistics* **15**, 587-629.
[101] Scott, S.C., Goldberg, M.S., Mayo, N.E. (1997): Statistical assessment of ordinal outcomes in comparative studies. *Journal of Clinical Epidemiology* **50**, 45–55.
[102] Senn, S. (1997): Statistical Issues in Drug Development. Wiley, New York.
[103] Senn, S. (2001): Statistical issues in bioequivalence. *Statistics in Medicine* **20**, 2785–2799.
[104] Schafer, J.L. (1997): Analysis of Incomplete Multivariate Data. Chapman and Hall, London.
[105] Schafer, J.L. (1999): Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3–15.
[106] Schall, R., Luus, H.G. (1993): On population and individual bioequivalence. *Statistics in Medicine* **12**, 1109–1124.
[107] Schinazi, R.B. (2000): The probability of a cancer cluster due to chance alone. *Statistics in Medicine* **19**, 2195–2198.
[108] Shetner, L.B. (1992): Bioequivalence revisited. *Statistics in Medicine* **11**, 1777–1788.

[109] Simon, R. (1999). The role of statisticians in intervention trials. *Statistical Methods in Medical Research* **8**, 281–286.
[110] Snow, J. (1860): On the mode of communication of cholera. 2nd ed. John Churchill, London.
[111] Sterne, J.A.C., Davey Smith G. (2001): Sifting the evidence — what's wrong with significance tests? *British Medical Journal* **322**, 226–231.
[112] Tukey, J.W. (1991): The philosophy of multiple comparisons. *Statistical Science* **6** (1), 100–116.
[113] Vuorinen, J., Turunen, J. (1997): A simple three-step procedure for parametric and nonparametric assessment of bioequivalence. *Drug Information Journal* **31**,167–180.
[114] Walker, S.H., Duncan, D.B. (1967): Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**, 167–179.
[115] Walter, S.D. (1993): Visual and statistical assessment of spatial clustering in mapping data. *Statistics in Medicine* **12**, 1275–1291.
[116] Ware, J.H., Lipsitz, S., Speizer, F.E. (1988): Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine* **7**, 95–108.
[117] Wedderburn, R.W.M. (1974): Quasi-likelihood functions, generalized linear models and the Gaussian method. *Biometrika* **61**, 439–447.
[118] Weir, B.S. (2000): Challenges facing statistical genetics. *Journal of the American Statistical Association* **95**, 319–322.
[119] White, H. (1982): Maximum likelihood estimation of misspecified models. *Econometrics* **50**, 1–25.
[120] White, I.R., Pocock, S.J. (1996): Statistical reporting of clinical trials with individual changes from allocated treatment. *Statistics in Medicine* **15**, 249–262.
[121] Zeger, S.L., Liang, K. (1986): Longitudinal data analysis using generalized linear models. *Biometrics* **42**, 121–130.
[122] Zhang, J., Quan, H., Ng, J., Stepanavage, M.E. (1997): Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials* **18**, 204–221.
[123] Zhao, L.P., Prentice, R.L. (1990): Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

Státní zdravotní ústav, Praha

Státní ústav radiační ochrany, Praha