Informační Bulletin

České Statistické Společnosti č. 2. září 2002. ročník 13

STATISTIKA –

časopis českých statistiků?

Určitě všichni čtenáři tohoto bulletinu vědí o existenci časopisu STATISTIKA, vydávaného každý měsíc Českým statistickým úřadem. Časopis je pokračovatelem v řadě odborných ekonomicko-statistických časopisů vydávaných státní statistickou službou už od roku 1920 (*Československý statistický věstník* v letech 1920–1930, *Statistický obzor* 1931–1961, *Statistika a kontrola* 1962–1963 a *Statistika* od roku 1964). Od příštího roku jej čeká další změna, tentokrát ne v názvu, ale v obsahu. Z *ekonomicko-statistického časopisu* by se měl stát *časopis českých statistiků*. Co to znamená? Záměrem je vytvořit časopis, který by reprezentoval *celou* naši statistickou obec a ne *pouze* její státní a ekonomickou část (i když ta je velmi podstatná). Časopis bude obsahovat rubriky o obecných problémech a otázkách z různých oblastí statistiky, zajímavé aplikace a metodické přínosy, novinky ze státní statistiky, anotace nově vyšlých knih a mnohé další. Ústřední částí každého čísla však bude monotematicky zaměřená rubrika s vyžádanými originálními příspěvky na předem dané téma.

Jedním z důvodů připravované reformy je snaha o zvýšení atraktivity časopisu nejen v rámci české statistické veřejnosti, ale i v mezinárodním měřítku. Proto budou o práci v redakční radě časopisu požádáni i přední odborníci ze zahraničí. Ediční rada věří, že větší než dosud autorskou podporu novému časopisu poskytnou i další odborná pracoviště, jako jsou statistické katedry na českých vysokých školách, výzkumná pracoviště, instituce a firmy.

Ideálně by časopis *Statistika* měl být svou úrovní a prestiží nejen srovnatelný se zahraničními časopisy podobného typu, ale měl by být i užitečným pomocníkem statistiků při plnění běžných pracovních i výzkumných úkolů z různých oblastí. Při vytváření nové koncepce je optimismus asi na místě, takže lze trochu odvážně říci, že navrhované změny jsou ovlivňovány snahou:

- Zvýšit autorský a čtenářský zájem o časopis u českých i zahraničních statistiků.
- Rozšířit a zlepšit odborné zaměření časopisu.
- Nabídnout statistikům prestižní a mezinárodně uznávaný statistický časopis.
- Motivovat zahraniční kolegy k publikování v České republice.
- Vytvořit prostor pro kvalifikované diskuse.
- Reagovat na zásadní změny jimiž dnes statistika ve světě prochází.
- Informovat o metodických, softwarových i pedagogických novinkách.
- Zabezpečit jednotu statistiky a zmenšit rozdíly mezi různými vědními obory.

Závažných statistických problémů a otázek je určitě celá řada. Očekávané zvýšení prestiže by mělo přinést i zvýšení impakt-faktoru časopisu.

Změní se i periodicita – časopis bude vycházet čtyřikrát do roka v předpokládaném rozsahu 60 stran. Příspěvky budou nejen v češtině, ale i v angličtině a slovenštině.

Časopis bude i nadále vydáván Českým statistickým úřadem, nicméně Česká statistická společnost dostala možnost se podílet na jeho náplni prostřednictvím svých členů v redakční radě a vlastních odborných příspěvků. Vedle Informačního Bulletinu tak bude mít naše společnost další možnost prezentovat výsledky práce svých členů. Tím jsme se na závěr dostali k onomu otazníku v názvu tohoto příspěvku. Bude *Statistika* opravdu *časopisem českých statistiků*? To nyní záleží především na Vás!

JA&GD&PH

O heterogenitě v modelech pro intenzity výskytu událostí

Petr Volf

ÚTIA AV ČR a TU Liberec

Abstract: In the present paper we first recall the notion and examples of random point processes, including the counting process, the cumulative process and marked process. The main body of the paper deals with the analysis of heterogeneity in the framework of counting process intensity model, and with the methods of solution – evaluation of individual heterogeneities and corresponding probability distribution.

Připomeneme nejdříve některé typy bodových náhodných procesů, čítací procesy a model jeho intenzity. Dále zopakujeme i modely pro procesy kumulativní a marked procesy. Hlavně se však budeme věnovat způsobu modelování heterogenity, tj. odlišné individuální intenzity čítacího procesu, a metodám evaluace takových modelů. Budeme je ilustrovat na jednoduchém příkladě.

1 Náhodné bodové a čítací procesy

Náhodný bodový proces je, obecně vzato, posloupnost náhodných "hodnot" $0 < T_1 < T_2 < \ldots$. Jak tyto hodnoty vznikají, to je dáno typem procesu. Čítací proces je pak proces s bodovým jednoznačně svázaný (lze je vlastně ztotožnit, bodový proces popsat jako čítací a naopak): $N(t) = \sum_{i=1}^{\infty} 1[T_i \leq t]$, tj. tento proces si připočte $+1$ v každém bodě odpovídajícího bodového procesu, při t rostoucím od 0 (nejčastěji jde skutečně o čas, ale v jiných aplikacích může jít o vzdálenost, rostoucí sílu působící na objekt apod.).

Z bodových procesů je nejznámější homogenní Poissonův proces, kdy $T_i = T_{i-1} + X_i$, X_i, $i = 1, 2, \ldots$, jsou i.i.d. náhodné veličiny s exponenciálním rozdělením (a $T_0 = 0$). Parametr λ exponenciálního rozdělení ($\lambda = 1/EX_i$) je intenzita procesu (zároveň je to intenzita rozdělení n.v. X_i dle definice $\lambda(x) = f(x)/(1 - F(x))$, když $f(x)$ je hustota a $F(x)$ je distribuční funkce). Obecnější procesy jsou nehomogenní Poissonův (intenzita již není konstantní, ale je to stále deterministická funkce) a pak případy, kdy intenzita je náhodná: Tak smíšený (mixed) Poissonův proces je Poissonův proces s intenzitou - náhodnou veličinou, viz např. definice 8.5.1, Rolski et al (1999). Rozdělením této náhodné veličiny (řekněme, s distribuční funkcí $G(\lambda)$) vlastně mícháme různé homogenní Poissonovy procesy. Pro každé $t > 0$ a celé $k \geq 0$ je

$$P(N(t) = k) = \int_0^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} dG(\lambda).$$

Všimněme si, že to odpovídá i Bayesovskému pohledu na situaci s neznámým λ a jeho priorem $G(\lambda)$.

Dvojitě stochastický (double stochastic) Poissonův proces, neboli Coxův proces, je pak nehomogenní variantou mixed procesu: Existuje náhodný proces $\lambda(t)$ tak, že pro $t > 0$ a celé $k \geq 0$ je

$$P(N(t) = k) = E\{\frac{(\int_0^t \lambda(s)ds)^k}{k!} exp(-\int_0^t \lambda(s)ds)\}.$$

Ta "dvojitá stochastičnost" znamená, že tu působí dva náhodné mechanizmy $\lambda(t)$ a $N(t)$, "nezávislá" realizace $\lambda(t)$ je intenzitou pro $N(t)$. Definici a rozbor situace viz např. opět Rolski et al, kapitola 12.2.2.

Pokud jde o čítací proces, budu zde tento pojem používat ve stejném smyslu jako například Andersen et al (1993). Tedy jako proces, jehož body přírůstků +1 jsou řízeny náhodnou intenzitou, která je predikovatelná a adaptovaná na neklesající posloupnost σ-algeber, filtraci $\mathcal{F}_{t-}$, tj. chápanou jako spojitou zleva. Můžeme tento proces považovat i za zobecnění dvojitě stochastického Poissonova procesu. Predikovatelnost zde znamená, že proces je pozorovatelný a hodnota v t je odvoditelná z hodnot těsně před t (takže například spojitost trajektorií zleva dostačuje). Základní vztah mezi pravděpodobností výskytu bodu v malém časovém intervalu Δt a momentální intenzitou $\lambda(t)$ je nyní podmíněný:

$$P\left(N(t+\Delta t) - N(t^-) = 1 | \mathcal{F}_{t-}\right) = \lambda(t)\,\Delta t + o(\Delta t).$$

Interpretace je taková, že $\mathcal{F}_{t-}$ obsahuje informaci o událostech, které se staly před t, tj. historii procesu, která má vliv na pravděpodobnost skoku $N(t)$ v t a nejbližší budoucnosti, tj. na "inovaci" procesu. Význam tohoto matematického aparátu je, že za těchto okolností je kumulovaná intenzita $L(t) = \int_0^t \lambda(s)\,ds$ kompenzátorem čítacího procesu, tj. proces $M(t) = N(t) - L(t)$ je martingal (adaptovaný na σ algebry $\mathcal{F}_{t+}$ obsahující historii i inovaci). Navíc, dík tomu, že v každém $[t, t+\Delta t)$ je $\Delta N(t)$ "nulajedničková" náhodná veličina (přesněji, s $P \to 1$ při $\Delta t \to 0$), tak varianční proces je $var(N(t)|\mathcal{F}_{t-}) = \langle N \rangle(t) = \langle M \rangle(t) = L(t)$. Podrobněji je toto vše probráno v mnoha článcích i knihách (Andersen et al, 1993, Fleming a Harrington, 1991).

1.1 Kumulativní proces, marked proces

Připomeňme si další typ procesu, a to

$$C(t) = \int_0^t Y(s)\,dN(s) = \sum_{i=1}^{n} Y(T_i) \cdot 1[T_i \le t]$$

kde $N(t)$ je čítací proces a $Y(t)$ jsou náhodné veličiny. Tento proces tedy sčítá přírůstky dané hodnotami $Y(T_i)$ v náhodných bodech T_i. Budeme jej nazývat kumulativní proces (Embrechts et al, 1997, Rolski et al, 1999, Volf, 2000). Užívá se pro popis některých finančních procesů (transakce, pojistné události), ale například i v modelování nárůstu poškození technických zařízení (Kahle and Wendt, 2000) a pro podobné účely v biologických či environmentálních studiích. Jeho nejjednodušší verze je složený (compound) Poissonův proces, kde body T_i jsou dány homogenním Poissonovým procesem a $Y(T_i)$ jsou i.i.d. náhodné veličiny, nezávislé i na T_i. V obecné verzi je $N(t)$ čítací proces odpovídající popisu v předchozím odstavci, a pokud jde o náhodné veličiny $Y(t)$, můžeme i jejich rozdělení popsat podmíněně, při daném $\mathcal{F}_{t-}$, nějakou hustotou $f(y; t, Z(t))$. Většinou potřebujeme i podmíněné momenty

$$E(Y(t)|\mathcal{F}_{t-}) = \mu(t, Z(t)), \quad \text{var}(Y(t)|\mathcal{F}_{t-}) = \sigma^2(t, Z(t)).$$

V tomto dosti obecném modelu se tedy předpokládá, že přírůstky $Y(t)$ a intenzita pro $N(t)$ mohou záviset na své vlastní (společné) historii, prostřednictvím kovariát $Z(t)$. Při vhodném popisu této závislosti na historii se i pro kumulativní proces odvodí rozklad na jeho kumulovanou intenzitu – kompenzátor – a na martingal (Volf, 2000).

Marked proces. Protože neznám vhodný výraz (snad "značkovaný" bodový proces, či proces označených bodů ?), užívám zde pojem marked proces. Je definován jako posloupnost bodů se značkami (markry) T_i, X_i, $i = 1, 2, \ldots$, kde T_i jsou body náhodného bodového procesu, $0 < T_1 < T_2 < \ldots$, a markry X_i jsou náhodné veličiny, které zpravidla říkají, co se v bodě T_i stalo (Arjas, 1989, Snyder, 1975, zmínka je i v Rolski et al, 1999).

Pojem marked procesu nepřináší vlastně nic nového z hlediska bodových procesů, ale nabízí nástroj pro klasifikaci bodového procesu na "podprocesy" rozlišené hodnotami markru (nejčastěji markry tvoří náhodný proces související se sledovaným bodovým procesem).

Zjevně nastává teoretický problém, jak takovýto proces adaptovat na nějakou společnou historii (popsanou zase pomocí filtrace – neklesající posloupnosti σ-algeber). Jde například o to, aby pro každé t a každou (borelovskou) množinu A bylo

$$N_A(t) = \sum_{i=1}^{\infty} 1[T_i \leq t,\ X_i \in A]$$

rozlišitelné na predikovatelnou část, kompenzátor, a na "nepredikovatelný" přírůstek, který má ovšem predikovatelné charakteristiky – jako například martingal v teorii čítacích procesů. Arjas (1989) možná i proto omezil prostor hodnot X_i na spočetnou množinu. Jiná cesta je taková, kterou jsme zvolili v popisu kumulativních procesů. I kumulativní proces $\{N(t), Y(t)\}$ lze chápat jako marked proces, kde přírůstky $Y(T_j)$ v bodech skoků T_j čítacího procesu $N(t)$ jsou markry a $\{T_j, Y(T_j),\ j = 1, 2, \ldots\}$ jsou body – realizace tohoto marked procesu (srovnej i Scheike, 1994).

Modely marked procesů začnou být zajímavé v případech, kdy typ události (tj. markr) v jednom procesu ovlivňuje intenzitu jiného procesu, což je typický případ v modelování spolehlivosti technických systémů.

2 Heterogenita v modelech pro intenzitu událostí

Představme si, že sledujeme intenzitu výskytu určitých (opakovatelných) událostí, pro skupinu objektů. Jelikož realita je rozmanitější než veškeré modely, v mnoha případech narazíme na to, že objekty se nějakým společným modelem dají popsat jen částečně. I když mnohé faktory ovlivňující individuální intenzitu procesu sledovaných událostí známe, jsme schopni je měřit a zahrnout jako kovariáty do regresního modelu pro intenzitu, stále mohou zbývat další vlivné faktory. Buď jsme je "zatím" vynechali, nebo je nedokážeme popsat a ani přesně rozpoznat – to je častá situace u živých organismů, kdy každý může mít svou individualitu, obtížně měřitelnou, která není popsatelná jen rozptylem v rámci daného modelu. Může samozřejmě jít i o charakteristiku skupinovou, tj. faktor společný celé třídě objektů.

Tuto **heterogenitu** tedy budeme zkoumat jako charakteristiku rozmanitosti objektů poté, co popíšeme to co mají společné. Pojem heterogenita se vyskytuje (v tomtéž smyslu jako zde) zejména v sociologických a demografických studiích (Gourieroux, 1988), synonymem je "frailty" model v biostatistické literatuře (Andresen et al, 1993, Ch. IX, Wang and Chang, 1999), či náhodný faktor (random effect) – dost často nyní používaný v lineárních regresních modelech, například i ve statistické analýze spolehlivosti. Tato veličina se zavádí do modelu i s nadějí, že se podaří ji nějak vysvětlit po provedené analýze. Podstatné pro úspěch analýzy je, aby sledované události byly opakovatelné a u každého objektu (či v homogenní skupině

objektů) jich bylo pozorováno více, např. poruchy a opravy, opakované události v lidském životě (změny práce či pobytu, nemoci apod.).

Můžeme začít tím, že každému sledovanému objektu j přiřadíme jeho (zpočátku neznámý) parametr v_j (tj. nebudeme mluvit o *náhodném* faktoru). Předpokládejme tedy, že pozorujeme následující data: Pro každý z $j = 1, \ldots, M$ objektů jsme sledovali $i = 1, \ldots, N_j$ událostí, v časech T_{ij}, přitom j-tý objekt v čase t je ovlivňován procesem kovariát $Z_j(t)$ (může být vícerozměrný a záviset na čase). Čas může být individuální (věk) nebo kalendářní, to vždy vyplyne z kontextu, osudy objektů považujeme za vzájemně podmíněně nezávislé, při daných $Z_j(t)$ (tj. v případě, že by stav jednoho objektu ovlivňoval budoucnost druhého, vyjádřili bychom ten vliv pomocí některé z kovariát). Dále uvažujme, jak je zvykem, i procesy $I_j(t) = 1$, pokud je j-tý objekt pozorován, $I_j(t) = 0$ jinak. Předpokládejme dále Coxův tvar regresního modelu pro rizikovou funkci, $h(t,z) = h_0(t) \cdot \exp(b(z))$, a přidejme neznámé parametry heterogenity, v_j, $j = 1, \ldots, M$. (Mimochodem, už teď vidíme, že model lze vždy dále zobecňovat, zde například tím, že heterogenita se může – individuelně různě – měnit v čase. Tuto možnost zatím vynecháme.) Model tedy předpokládá, že intenzita výskytu události v čase t pro objekt j je $v_j \cdot \lambda_j(t)$, kde $\lambda_j(t) = h_0(t) \cdot \exp(b(Z_j(t))) \cdot I_j(t)$. V souladu s teorií pro modely čítacích procesů (Andersen et al, 1993, Fleming and Harrington, 1991, aj.), $Z_j(t)$ a $I_j(t)$ jsou brány jako náhodné procesy pozorovatelné a predikovatelné. Např. postačí spojitost jejich trajektorií zleva, pozorovatelnost $I_j(t)$, a pozorovatelnost $Z_j(t)$ pokud $I_j(t) = 1$. Dále označíme $H_0(t) = \int_0^t h_0(s)\,\mathrm{d}s$, kumulativní společnou ("baseline") rizikovou funkci, $L_j(t) = \int_0^t \lambda_j(s)\,\mathrm{d}s$ kumulativní intenzity (náhodné procesy, se spojitými trajektoriemi, neklesajícími). Je vidět, že v_j a $h_0(t)$ by nebyly jednoznačné, proto je nutné provést nějakou "normalizaci", například položit $v_1 = 1$, nebo držet $\sum_{j=1}^M v_j/M = 1$, apod.

Pozorovaná data jsou tedy $Z_j(t)$, $I_j(t)$, $N_j(t)$, skoky čítacích procesů $\Delta N_j(t) = 1$ jsou v bodech – časech pozorovaných událostí T_{ij}, $i = 1, \ldots, N_j = N_j(T)$, $j = 1, \ldots, M$, T je čas ukončení pozorování. Věrohodnostní proces – aspoň jeho relevantní část, tj. část obsahující neznámé parametry v_j a funkce $h_0(t)$, $b(z)$, je

$$\mathcal{V} = \prod_{j=1}^{M} \left\{ \prod_{t \le T} (v_j \lambda_j(t))^{\mathrm{d}N_j(t)} \cdot exp\left(- \int_0^T v_j \lambda_j(t)\,\mathrm{d}t \right) \right\},$$

logaritmus pak je

$$\mathcal{L} = \ln \mathcal{V} = \sum_{j=1}^{M} \left\{ \int_0^T (\ln v_j + \ln \lambda_j(t))\,\mathrm{d}N_j(t) - \int_0^T v_j \lambda_j(t)\,\mathrm{d}t \right\}. \qquad (1)$$

Necháme zatím stranou odhad regresní funkce $b(z)$, která se odhaduje z tzv. parciální (částečné) věrohodnostní funkce. Pokud bychom $b(z)$ považovali za známou, tak technikou MVO dojdeme ke skokovitému odhadu $H_0(t)$ (Nelson - Aalen či Breslow - Crowley odhad)

$$\hat{H}_0(t) = \int_0^T \sum_{j=1}^{M} \frac{\mathrm{d}N_j(t)}{\sum_{k=1}^{M} I_k(t)\, e^{b(z_k(t))} \cdot v_k} \qquad (2)$$

tj. se skoky v bodech T_{ij}. Zároveň, pokud vhodně přepíšeme $\mathcal{V}$:

$$\mathcal{V} = \prod_{j=1}^{M} \left\{ v_j^{N_j(T)} \exp(-v_j L_j(T)) \cdot \prod_t \lambda_j(t)^{\mathrm{d}N_j(t)} \right\}, \tag{3}$$

vidíme, že při daných hodnotách $L_j(T)$ lze snadno řešit úlohu MVO pro každé v_j zvlášť. Hledáme v_j maximalizující $N_j(T) \cdot \ln v_j - v_j L_j(t)$, což vede na logický výsledek $\hat{v}_j = \frac{N_j(T)}{L_j(T)}$, neboli individuální náchylnost k události je odhadnuta jako poměr počtu událostí, ke kterým došlo v $[0,T]$, k celkové kumulované intenzitě dané modelem, což by měla být predikce očekávaného počtu událostí. Tj. $\hat{v}_j$ měří neshodu mezi modelem ($L_j(t)$) a pozorovanou skutečností ($N_j(t)$), a to poměrem hodnot dosažených na konci intervalu $[0,T]$ (zatímco tzv. zobecněné reziduály, běžně využívané v testech shody dat a modelu – viz např. Volf (1996) – jsou procesy definované rozdílem $N_j(t) - L_j(t)$, pro všechna $t \in [0,T]$).

Přirozený postup odhadování je iterativní, při daných odhadech v_j odhadneme $\hat{H}_0(t)$, tím odhadneme i $L_j(t)$ a můžeme znovu odhadnout hodnoty v_j. Toto schema lze považovat i za variantu tzv. EM algoritmu. Tam se většinou vychází ze situace, kdy chybí část "dat" a známe podmíněné rozdělení těchto chybějících dat při daných parametrech modelu a pozorovaných datech. Zde za onu chybějící část dat můžeme považovat v_j, $j = 1, \ldots, M$, kdežto $H_0(t)$ a $b(z)$ jsou "parametry" modelu. Tento pohled už zachází s v_j jako s realizacemi náhodných veličin V_j, které v případě homogenity objektů by měly být i.i.d. Zároveň, jejich složení přes celou populaci objektů charakterizuje rozptyl okolo modelu. Otázkou, ke které se ještě vrátíme, je, kdy má smysl heterogenitu do modelu zavádět, tj. kdy je statisticky významná.

Řešení pomocí EM algoritmu – sestává z opakování dvou kroků: Při daných odhadech v_j provedeme MVO parametrů resp. funkcí modelu, tj. $b(z)$ maximalizací logaritmu parciální věrohodnostní funkce

$$\mathcal{L}^P = \sum_{j=1}^{N} \int_0^T \ln \left\{ \frac{e^{b(z_j(t))}}{\sum_{k=1}^{N} v_k e^{b(z_k(t))} I_k(t)} \right\} \mathrm{d}N_j(t) \tag{4}$$

a odhad $H_0(t)$ z (2), při dosazení odhadu $b(z)$ z (4). Toto je tzv. M-krok EM algoritmu. Následuje E-krok, kde se při aktuálních hodnotách odhadů udělá podmíněná střední hodnota chybějících dat. Všimneme si, že (3) má vlastně pro každé v_j tvar hustoty gamma rozdělení (až na konstantu), s parametry $N_j(T) + 1$ a $L_j(T)$. Takže zatímco příslušný modus (bod maxima) byl $N_j(T)/L_j(T)$, střední hodnota $E(v_j|\ldots) = (N_j(T) + 1)/L_j(T)$.

Odhady takto získané většinou konvergují. Nyní by bylo vhodné odbočit, říci něco z teorie EM algoritmu a jeho konvergence k (některému) maximu věrohodnostní funkce (viz. např. Dempster et al, 1977), a pokusit se tuto teorii aplikovat na náš případ. Samozřejmě, takové pokusy byly činěny (viz reference v Andersen et al, či speciální literatura o frailty modelech). Jedna věc je při daných datech konvergence EM algoritmu. Ta vyplývá z toho, že každý krok zvětší hodnotu věrohodnostní funkce, která je přitom ohraničená. Něco jiného je konzistence takto získaných odhadů, při rostoucím počtu dat. Růst počtu dat ("hustoty" dat celkově) je potřeba ke konzistenci odhadu $H_0(t)$ i $b(z)$. Ale pro zpřesňování odhadů v_j by měl růst počet dat i pro každý objekt, což v praktických příkladech na ohraničeném časovém intervalu není reálné. Dále, pro odhad rozdělení heterogenity bychom potřebovali velký

počet pozorovaných objektů. Při teoretických úvahách se tedy předpokládá (stejně jako v mnoha obdobných případech), že roste i počet objektů (tj. počet parametrů), ale pomaleji než počet dat. Pak je možné dojít i k nějakému teoretickému výsledku. Jednodušší je případ, kdy v_j charakterizují odlišnost různých skupin, kterých je stálý počet.

Bayesovské řešení. Ještě se na problém veličin v_j podívejme z dalšího hlediska. Nabízí se pochopitelně hledisko Bayesovské, které obchází problém konzistence a řeší úlohu optimálního využití informace z dat jiným způsobem. Uvažujme náhodnou veličinu V, s apriorním rozdělením gamma (tak dostaneme i aposteriorní rozdělení typu gamma) a s $EV = 1$ – to abychom odstranili onu již zmíněnou nejednoznačnost V a $h_0(t)$. Respektive, uvažujme M i.i.d. veličin V_j s tímtéž apriorním rozdělením, které je dáno hustotou $f(v) = b^a v^{a-1} e^{-bv}/\Gamma(a)$, s parametrem $a = b > 0$. Ten můžeme buď zvolit (pak $EV = a/b = 1, a/b^2 = 1/a$ jsou střední hodnota a rozptyl apriorního rozdělení), nebo s ním také zacházet jako s neznámým parametrem. Tuto druhou cestu volí ony již zmíněné frailty modely. Věrohodnostní funkce pro $h_0(t)$ a $b(z)$, při daných v_i, je tatáž jako předtím. Aposteriorní rozdělení pro veličiny V_i, při daných datech a $H_0(t)$, $b(z)$, je pak

$$g(\mathbf{v}|\ldots) = C \cdot \prod_{j=1}^{M} \left\{ v_j^{a-1} \prod_t v_j \lambda_j(t)^{dN_j(t)} \cdot \exp(-av_j - v_j L_j(T)) \right\},$$

takže aposteriorní rozdělení pro jednotlivé V_j jsou podmíněně nezávislá a jsou typu gamma s parametry $N_j(T)+a$, $L_j(T)+a$. Příslušný Bayesův odhad (modus aposteriorního rozdělení) je tedy $\tilde{v}_j = (N_j(T)+a-1)/(L_j(T)+a)$. Vidíme, že výsledek je prakticky tentýž jako v předchozím způsobu řešení. Pokud parametr a volíme malý, blízko k 0, apriorní rozdělení se přiblíží "neinformativnímu".

Ve frailty modelech se situace usložití procedurou průběžného odhadu a, která ale zřejmě nic podstatného pro analýzu nepřinese a naopak v praxi spíš "přispěje" k nestabilitě výpočtů. Proto tento přístup nebudu dále rozebírat.

Odhadem individuálních heterogenit v_j analýza nekončí, měla by následovat i jakási interpretační část. Získané hodnoty $\hat{v}_j$ ($\tilde{v}_j$) můžeme považovat za reprezentaci nějaké náhodné veličiny V – heterogenity celé populace analyzovaných objektů. Takže nyní je ta pravá chvíle z "dat" $\hat{v}_j$ odhadnout rozdělení veličiny V. Můžeme udělat neparametrický odhad (empirickou distribuční funkci, jádrový odhad pro hustotu), i zkusit gamma rozdělení. Takto získané rozdělení je pak dobré použít například při predikování událostí pro nový objekt. Intenzita poruch sama o sobě charakterizuje průměrnou trajektorii čítacího procesu. Heterogenita k tomu vlastně dodává charakteristiku rozptýlenosti trajektorií, tedy jakoby další parametr. Srovnej, že pro homogenní Poissonův proces chápaný jako čítací proces je střední trajektorie $\lambda \cdot t$, a rozptyl v t je také $\lambda \cdot t$. Pomocí heterogenity dostaneme rozptyl $\mathrm{var}V \cdot (\lambda t)^2 + EV \cdot \lambda t$. Tento vztah by bylo možné použít k analýze, zda je příspěvek heterogenity významný, tj. zda se frekvence událostí vymyká modelu. Takovéto testy jsou součástí zde uvedeného příkladu.

Dále, hodnoty $\hat{v}_j$ můžeme použít k odhalení outlierů, tj. objektů, jejichž proces událostí se výrazně liší (co do průměrné intenzity v $[0,T]$) od ostatních. Úlohu detekce netypických trajektorií procesu jsme převedli na úlohu detekce netypických bodů v souboru v_j, které by měly reprezentovat jednu náhodnou veličinu. To je už standardní úloha řešená běžně pomocí kvantilů rozdělení maxima apod. (v Shewhar-

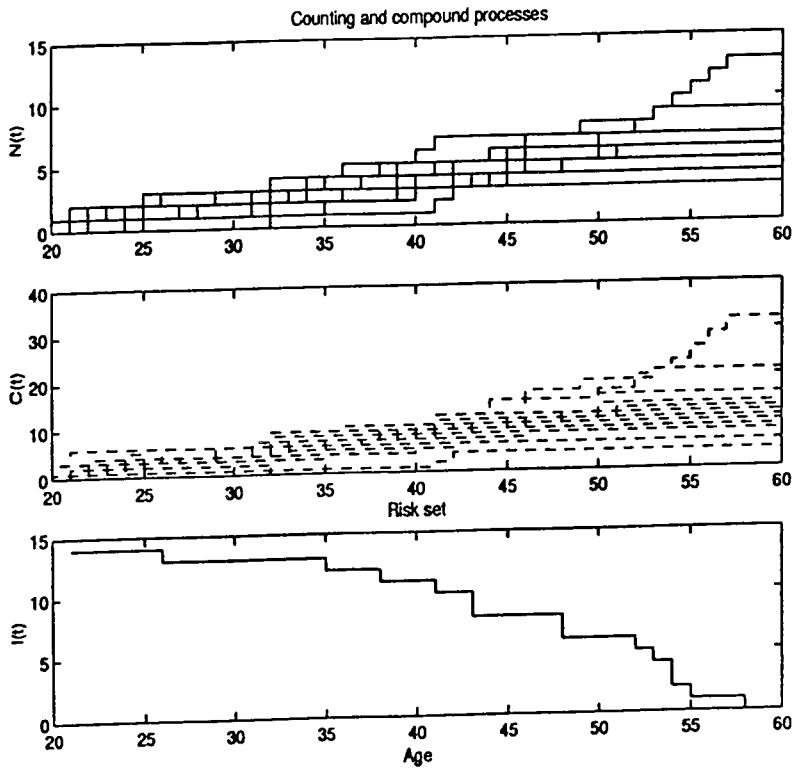


Figure 1: Pozorované procesy

tových diagramech je datum "podezřelé" z odchýlenosti už když překročí zvolený (97,5 % třeba) kvantil základního rozdělení veličiny V).

3 Příklad

V rámci přípravy dat na symposium "Stav účastníků Robustů po dvaceti letech", které proběhlo paralelně s konferencí Robust 2000, někteří kolegové dobrovolně (anonymně) vypovídali o sledu důležitých událostí ve svém životě (nebylo specifikováno jakých, zřejmě vesměs nepříznivých). Čítací proces v příkladě je proces těchto událostí, registrovaných od věku 10 let, čas t je věk. Obr. 1 ukazuje data - čítací i kumulativní procesy (sčítají se ohodnocení závažnosti událostí) pro jednotlivé osoby (bylo jich 14) a proces - součet indikátorů, tedy počet osob daného věku. Neprováděli jsme regresní analýzu, použili jsme metodu EM algoritmu pro analýzu heterogenity. Výsledky jsou na obr. 2, konkrétně odhad kumulativní (baseline) rizikové funkce a odhady hodnot heterogenity v_j pro jednotlivé osoby. K těmto hodnotám jsme pak našli metodou MVO gamma rozdělení – je však fitováno na hodnoty v_j bez jejich maxima. Cílem je zvětšení citlivosti detekce maxima jako případného out-

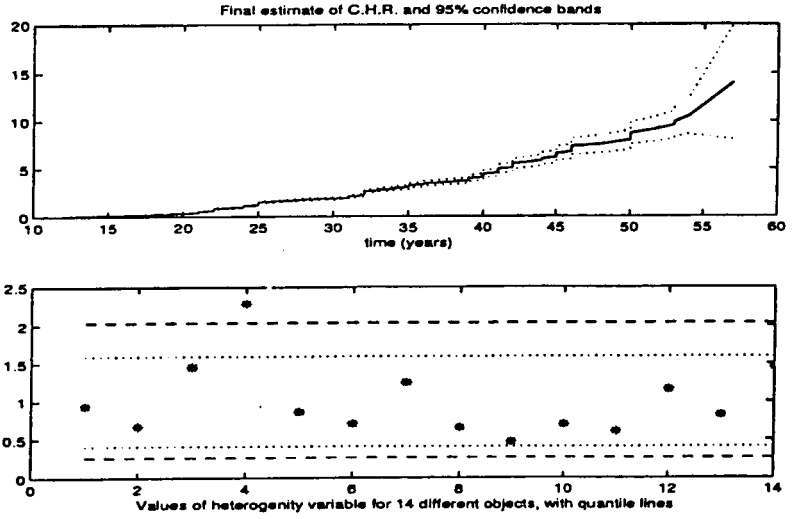


Figure 2: Odhad kumulativní rizikové funkce a hodnot heterogenity

lieru. Obrázek obsahuje také horní a dolní 0.025 kvantily tohoto gamma rozdělení (tečkovaně) a kvantily pro rozdělení maxima ze 14 položek (čárkovaně). Obr. 3 ukazuje ono gamma rozdělení, odhadnuté parametry byly $a = 8.703, b = 9.615$, tj. odhad $EV = a/b = 0.905$ a $\mathrm{var}V = a/b^2 = 0.0941$. Na obr. 4 jsou 3 vybrané trajektorie, konkrétně pro osoby 4, 5 a 9, spolu s odpovídajícím průměrným modelem za předpokladu homogenity, tj. s průběhem kumulativní intenzity (stejné jako na obr. 2a) spočtené za předpokladu stejného rizika pro všechny osoby. Porovnání ukazuje odchylky od takového modelu, tedy "průběžnou" heterogenitu. Tečkovaně jsou doplněny, v každém t (a spojeny), horní i dolní 0.025 kvantily Poissonova rozdělení s parametrem $\Lambda(t) = \int_0^t I_j(t) dH_0(t)$. Tento obrázek je grafickou variantou testu dobré shody dat s modelem v rámci čítacích procesů, viz Volf (1996). Osoba č. 4 má registrovány jen 3 události, ale během období, kdy kumulovaná intenzita, která reprezentuje vlastně očekávaný průběh čítacího procesu, je nízká. (Jako jediný(á) uvedl(a) důležitou událost již ve věku 12 let, zřejmě prudký náraz puberty, na který se ostatní asi již nerozpomněli(y)). Skutečný čítací proces této osoby je tedy zpočátku vychýlený vzhůru. Osoba 9 má také registrováno málo událostí, ale za poměrně dlouhou dobu. Srovnání s kumulativní intenzitou ukazuje, že příslušný čítací proces je odchýlený naopak dolů. U osoby č. 5 je sice registrován daleko největší počet významných životních událostí (13), ale jak ukazuje graf, zcela to odpovídá očekávání za poměrně dlouhou dobu, po kterou nás tento již zocelený kolega provází. Jeho hodnota heterogenity je blízko k jedné (viz. obr. 2).

Výzkum této tematiky je podporován granty MŠMT ČR č. MSM 245 100 303 a GA ČR č. 402/01/0539.

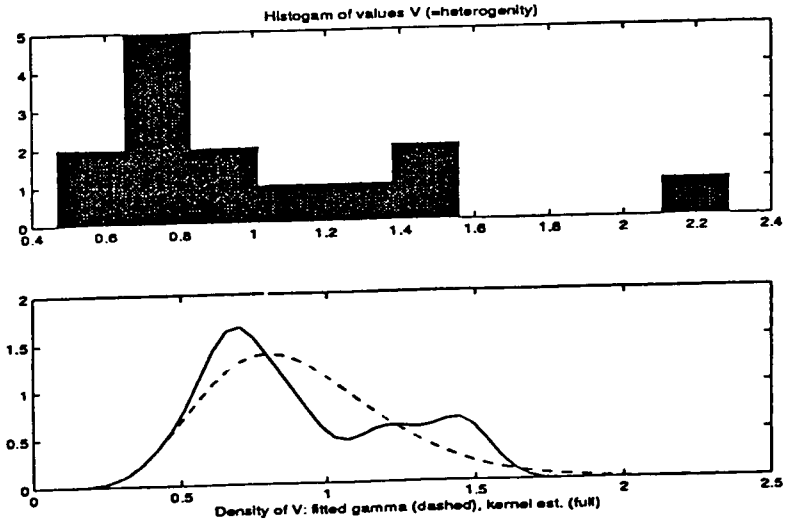


Figure 3: Odhad rozdělení heterogenity

Literatura

Andersen P.K., Borgan O., Gill R.D. and Keiding N. (1993). Statistical Models Based on Counting Processes. Springer, New York.

Arjas E. (1989). Survival models and martingale dynamics. Scand. J. Statist. 16, 177–225.

Dempster A.P., Laird N.M. and Rubin D.B. (1997): Maximum likelihood estimation from incomplete data via the EM algorithm. J.R.S.S., Ser. B 39, 1–22.

Embrechts P., Klüppelberg C. and Mikosch T. (1997). Modelling Extremal Events. Springer.

Fleming T.R. and Harrington D.P. (1991): Counting Processes and Survival Analysis. Wiley, New York.

Gourieroux Ch. (1988). In: Analyse Statistique des Durees de Vie, Univ. d'Aix – Marseille II.

Kahle W. and Wendt H. (2000): Statistical analysis of damage processes. In Recent Advances in Reliability Theory (N. Limnios and M. Nikulin, editors), Birkhauser, 199–212.

Rolski T., Schmidli H., Schmidt V. and Teugels J. (1999). Stochastic Processes for Insurance and Finance. Wiley.

Scheike T.H. (1994). Parametric regression for longitudinal data with counting process measurement times. Scand. J. Statist. 21, 245–263.

Snyder D. L. (1975). Random Point Processes. Wiley, New York.

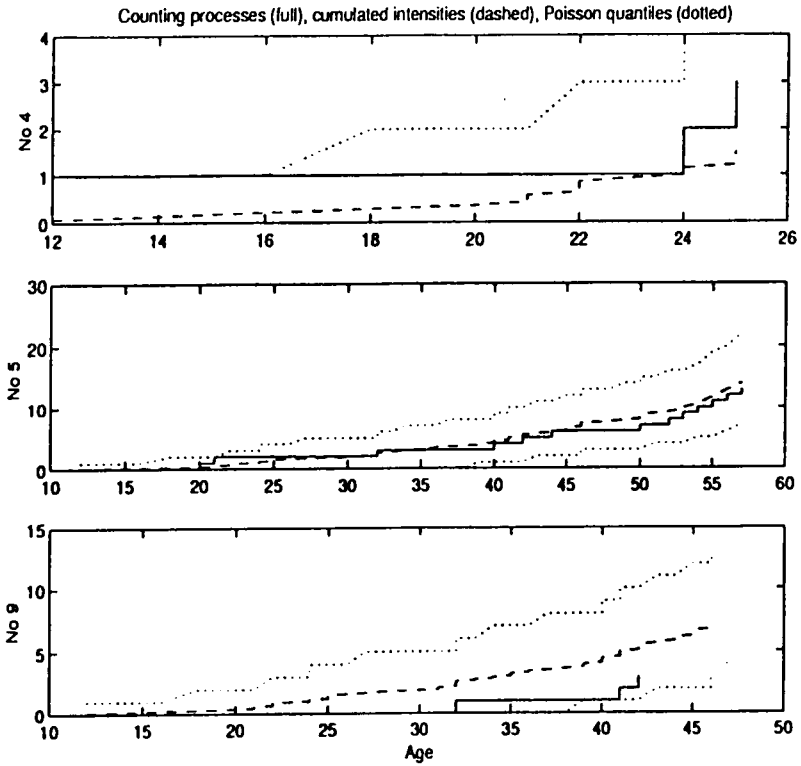


Figure 4: Grafické testy shody pozorovaných $N_j(t)$ (plně) s modelovými $L_j(t)$ (čárkovaně), pro $j = 4, 5, 9$

Volf P. (1996). Analysis of generalized residuals in hazard regression models. Kybernetika 32, 501–510.

Volf P. (2000). On cumulative process model and its statistical analysis. Kybernetika 36, 165–176.

Wang M. Ch. and Chang Sh. H. (1999). Nonparametric estimation of recurrent survival function. J.A.S.A. 94, 146–153.

MĚŘENÍ RELIABILITY ANEB BACHA NA CRONBACHA

KAREL ZVÁRA

1. Reliabilita

Při popisu nespolehlivosti měření spojitého znaku se zpravidla vychází z představy, že měření $Y = A + e$ je součtem měřené náhodné hodnoty znaku (vlastnosti) A a náhodné chyby měření e. Předpokládá se, že A, e jsou nezávislé náhodné veličiny s kladným rozptylem. Označme

$$\mathsf{E}\,A = \mu, \qquad \operatorname{var} A = \sigma_A^2,$$
$$\mathsf{E}\,e = 0, \qquad \operatorname{var} e = \sigma^2.$$

Spolehlivost veličiny Y jako měření A se vyjadřuje pomocí **reliability** (např. Hniličková a kol. (1972))

$$\operatorname{rel}(Y) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}. \tag{1}$$

Reliabilita tedy porovnává variabilitu chybového členu s variabilitou měřené vlastnosti. Nevyjadřuje kvalitu měření absolutně, ale relativně, na což se mnohdy v aplikacích zapomíná.

Pokud bychom měli dvě nezávislá měření stejné měřené hodnoty A, pak snadno zjistíme, že pro jejich korelační koeficient platí

$$\operatorname{cor}(A + e_1, A + e_2) = \operatorname{rel}(Y),$$

což dá přirozenou interpretaci pojmu reliabilita.

Nespolehlivost měření způsobená jeho náhodnou složkou má nepříjemné důsledky. Odhadujme například korelační koeficient mezi dvěma měřenými vlastnostmi A, B pomocí X resp. Y s příslušnými reliabilitami $\operatorname{rel}(A), \operatorname{rel}(B)$.

Autor děkuje Markovi Malému, bez jehož pomoci při pořizování klíčových historických publikací by publikace nemohla vzniknout a Zdeňku Rothovi za velice užitečnou diskusi.

Práce vznikla za podpory výzkumného záměru CEZ:J13/98:113200008.

Předpokládáme samozřejmě, že chybové složky měření jsou mezi sebou i s měřenými vlastnosti nezávislé. Snadno zjistíme, že platí

$$\begin{aligned}\operatorname{cor}(X,Y) &= \frac{\operatorname{cov}(A+f,B+e)}{\sqrt{\operatorname{var}(A+f)\operatorname{var}(B+e)}} \\ &= \operatorname{cor}(A,B)\sqrt{\operatorname{rel}(A)\operatorname{rel}(B)},\end{aligned}$$

takže se sledovaná závislost viděná prostřednictvím závislosti X, Y poněkud rozředí. Absolutní hodnota korelačního koeficientu odhadu je tím menší, čím jsou méně spolehlivá měření.

Podobně uvažujme o plánování rozsahu výběru, který má sloužit k výpočtu intervalu spolehlivosti pro střední hodnotu μ normálního rozdělení (při známém rozptylu σ_A^2) s předem danou poloviční délkou δ. Porovnejme potřebné rozsahy, kdybychom měřili přímo A a když měříme pomocí $Y = A+e$. V tomto druhém případě potřebujeme aspoň

$$n \geq (\sigma_A^2 + \sigma^2)\left(\frac{u(\alpha/2)}{\delta}\right)^2 = \frac{1}{\operatorname{rel}(Y)}\sigma_A^2\left(\frac{u(\alpha/2)}{\delta}\right)^2,$$

takže s klesající spolehlivostí $\operatorname{rel} Y$ pochopitelně požadované n roste.

2. Opakovaná pozorování

Vzorec (1) nenabízí žádnou přímou metodu odhadu reliability, protože veličinu A nemůžeme pozorovat přímo, takže nemůžeme přímo odhadnout ani její rozptyl.

Uvažujme nyní situaci, kdy jsou možná opakovaná pozorování stejné hodnoty měřené vlastnosti. Lze předpokládat nezávislost příslušných chybových členů (při pevné hodnotě měřené veličiny). Bohužel, v psychologických či pedagogických vědách se s takovou situací nesetkáváme, kdežto vzdálenost dvou bodů na neostrém rentgenovém snímku může lékař změřit několikrát.

Uvažujme nejprve právě dvě opakování každého měření $Y_1 = A + e_1, Y_2 = A+e_2$, přičemž jsou náhodné veličiny A, e_1, e_2 nezávislé. Zaveďme $P = Y_1 + Y_2$ a $Q = Y_1 - Y_2$. Zřejmě platí $\operatorname{var} P = \operatorname{var}(2A + e_1 + e_2) = 4\sigma_A^2 + 2\sigma^2$, takže je

$$\operatorname{rel}(P) = \frac{\operatorname{var}(2A)}{\operatorname{var}(2A + e_1 + e_2)} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/2}. \tag{2}$$

Reliabilita součtu je tedy větší než reliabilita jednotlivých sčítanců. Je zřejmé, že při sčítání m takových nezávislých měření stejné hodnoty A bychom dostali

$$\operatorname{rel}\left(\sum_{i=1}^{m} Y_i\right) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/m}.$$

Zřejmě platí $\operatorname{var} Q = \operatorname{var}(e_1 - e_2) = 2\sigma^2$. Reliabilitu jednoho měření můžeme tedy vyjádřit jako

$$\operatorname{rel}(Y) = \frac{\operatorname{var} P - \operatorname{var} Q}{\operatorname{var} P + \operatorname{var} Q}, \tag{3}$$

reliabilitu součtu obou měření (a také jejich průměru) jako

$$\operatorname{rel}(P) = \frac{\operatorname{var} P - \operatorname{var} Q}{\operatorname{var} P} = 1 - \frac{\operatorname{var} Q}{\operatorname{var} P}. \tag{4}$$

Rozptyl P i Q lze odhadnout, umíme tedy odhadnout také obě reliability. Podle Cronbach (1951) pochází vztah (4) z článku Rulon (1939).

3. Několik položek

V psychologickém výzkumu (podobně ve výzkumu pedagogickém) je zvykem hodnotit několik položek. Následuje model pro tuto situaci, který je velmi podrobně vyšetřen ve výzkumné zprávě Novicka a Lewise (1966).

Předpokládejme, že u každého měřeného objektu (jedince) hodnotíme m vlastností $B_1, \ldots, B_m$ pomocí chybou ovlivněných měření $Y_j = B_j + e_j$, přičemž nás zajímá součet $B_\bullet = \sum_j B_j$ odhadovaný pomocí součtu $Y_\bullet = \sum_j Y_j$. Předpokládáme opět, že všechny chybové členy jsou nezávislé, že nezávisí ani na měřených hodnotách. Mezi měřenými vlastnostmi obecně předpokládáme závislost. Reliabilita součtu je dána vztahem

$$\operatorname{rel}(Y_\bullet) = \frac{\operatorname{var} B_\bullet}{\operatorname{var} Y_\bullet} = \frac{\operatorname{var} B_\bullet}{\operatorname{var} B_\bullet + m\sigma^2}. \tag{5}$$

Důležitým výsledkem je tvrzení věty 2.1 zmíněné zprávy, podle kterého platí nerovnost

$$\operatorname{rel}(Y_\bullet) \geq \frac{m}{m-1}\left(1 - \frac{\sum_j \operatorname{var} Y_j}{\operatorname{var} Y_\bullet}\right) = \alpha, \tag{6}$$

kde výraz uvedený na pravé straně je funkcí přímo pozorovaných veličin (lze jej tedy odhadnout) a nazývá se **Cronbachovo alfa** podle článku Cronbach (1951), kde bylo toto označení nejspíš prvně použito.

Ještě zajímavějším výsledkem je tvrzení věty 3.1 stejné zprávy, podle kterého ve vztahu (6) nastává právě tehdy, když současně platí

$$\operatorname{var} B_1 = \ldots = \operatorname{var} B_m, \tag{7}$$

$$\operatorname{cor}(B_i, B_j) = 1, \qquad 1 \leq i, j \leq m. \tag{8}$$

Prakticky to znamená (důkaz naznačen v posledním odstavci), že existují konstanty $\beta_1, \ldots, \beta_m$ (můžeme požadovat splnění požadavku $\sum_j \beta_j = 0$) a

náhodná veličina A taková, že lze pro každé j lze s jednotkovou pravděpodobností psát

$$B_j = A + \beta_j. \tag{9}$$

4. Dvojné třídění

Pro skutečná měření několika položek nyní zavedeme smíšený model analýzy rozptylu dvojného třídění.

Předpokládejme, že jednotlivé položky jsou vázány vztahem (9), přičemž pro t-tou osobu se A realizuje jako A_t, konstanty β_j jsou pro všechny osoby stejné. Potom j-té měření t-té osoby lze vyjádřit jako

$$Y_{tj} = A_t + \beta_j + e_{tj}, \qquad 1 \le t \le n, 1 \le j \le m, \tag{10}$$

kde $e_{tj} \sim \mathrm{N}(0, \sigma^2)$ a $A_t \sim \mathrm{N}(\mu, \sigma_A^2)$ jsou nezávislé náhodné veličiny.

Pro součty čtverců ($\bar{Y}_{t\bullet}, \bar{Y}_{\bullet j}, \bar{Y}_{\bullet\bullet}$ značí průměry přes indexy nahrazené tečkami) platí

$$SA = \sum_{t=1}^{n}\sum_{j=1}^{m} \left(\bar{Y}_{t\bullet} - \bar{Y}_{\bullet\bullet}\right)^2 \sim (m\sigma_A^2 + \sigma^2)\chi^2(n-1),$$

$$SE = \sum_{t=1}^{n}\sum_{j=1}^{m} \left(Y_{tj} - \bar{Y}_{\bullet j} - \bar{Y}_{t\bullet} + \bar{Y}_{\bullet\bullet}\right)^2 \sim \sigma^2\chi^2((n-1)(m-1)).$$

Průměrné čtverce MSA, MSE mají střední hodnoty

$$\mathrm{E}\, MSA = \mathrm{E}\, SA/(n-1) = m\sigma_A^2 + \sigma^2, \tag{11}$$

$$\mathrm{E}\, MSE = \mathrm{E}\, SE/((n-1)(m-1)) = \sigma^2, \tag{12}$$

takže Cronbachovo alfa lze pomocí těchto středních hodnot vyjádřit jako

$$\alpha = \frac{\mathrm{E}\, MSA - \mathrm{E}\, MSE}{\mathrm{E}\, MSA}.$$

Když nyní nahradíme střední hodnoty jednotlivými statistikami, dostaneme odhad Cronbachova alfa

$$\hat{\alpha} = \frac{MSA - MSE}{MSA} = 1 - \frac{MSE}{MSA} = 1 - \frac{1}{F_A}, \tag{13}$$

kde F_A je statistika běžně používaná k rozhodování o hypotéze $\sigma_A^2 = 0$.

5. Klasický odhad

Označme symbolem $\mathbf{Y}_{t\bullet} = (Y_{t1}, \ldots, Y_{tm})'$ pozorování u t-té osoby. Z našich předpokladů plyne

$$\mathbf{Y}_{t\bullet} \sim \mathrm{N}_m(\beta + \mu\mathbf{1}, \Sigma),$$

kde varianční matice má tvar $\Sigma = \sigma^2\mathrm{I} + \sigma_A^2\mathbf{1}\mathbf{1}'$. Označíme-li symbolem S výběrový odhad matice Σ s prvky

$$s_{ij} = \frac{1}{n-1}\sum_{t=1}^{n}\left(Y_{ti} - \bar{Y}_{\bullet i}\right)\left(Y_{tj} - \bar{Y}_{\bullet j}\right),$$

lze potom kupodivu zapsat oba průměrné čtverce pomocí složek matice S:

$$MSA = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij},$$

$$MSE = \frac{1}{m-1}\sum_{j=1}^{m} s_{jj} - \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}.$$

Dosadíme-li tyto výrazy do $\hat{\alpha}$, dostaneme po úpravě

$$\hat{\alpha} = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} s_{jj}}{\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}}\right), \tag{14}$$

což je asi nejpoužívanější zápis odhadu Cronbachova alfa.

Vyjádření $\hat{\alpha}$ pomocí statistik analýzy rozptylu umožní snadno zjistit, pro jaká data vyjde $\hat{\alpha} = 1$. Je to právě tehdy, když je reziduální součet čtverců SE nulový. To je zřejmě jen tehdy, když je hodnota Y_{tj} součtem „osobního" čísla a_t a „položkového" čísla b_j. Pak by stačilo měřit jedinou položku, ostatní jsou nadbytečné!

Na druhé straně nás nesmí překvapit záporný odhad Cronbachova alfa, který dostaneme tehdy, když se hodnocené osoby liší nedostatečně, když vyjde $F_A < 1$. Proto je vhodné upravit definici tohoto odhadu tak, že na pravé straně (13) a (14) uvedeme maximum z tam uvedeného výrazu a nuly. To je ostatně konzistentní s používaným odhadem komponent rozptylu. Snad mohu v této souvislosti připomenout výhradu k definici reliability, že porovnává variabilitu chybové složky měřeného s variabilitou měřeného.

6. Příklady

Konzultační činnost přináší nejrůznější data. Před časem jsem získal výsledky 231 testů zkoumajících znalosti z biologie. Test sestával z 29 položek, z nichž 11 položek připouštělo pouze odpovědi ano/ne. Na druhé straně tam

byla jedna položka šestibodová, jedna pětibodová, několik čtyřbodových atd. Výsledná hodnota odhadu Cronbachova alfa je 0,813, což je jistě slušný výsledek.

Z nestejných bodových ohodnocení jednotlivých položek je zřejmé, že zřejmě nebude splněn předpoklad stejných rozptylů uplatněný v modelu dvojného třídění (10). Proto je zajímavé konstatovat, že odhad alfa spočítaný z normovaných veličin (varianční matici v (14) prostě nahradíme maticí korelační) vyjde číselně velmi podobný, totiž 0,805. Kdybychom každou odpověď hodnotili procentem maximálního možného počtu bodů, dostali bychom odhad alfa pouze 0,765.

Nyní se věnujme některým možným problémům. Použijeme zmíněných jedenáct položek s nulajedničkovými odpověďmi. Na těchto datech lze velmi dobře ilustrovat závislost Cronbachova alfa na variabilitě měřeného znaku. Prostý výpočet Cronbachova alfa vede k hodnotě 0,403. Kdybychom se omezili jen na ty studenty, kteří měli jen 5 nebo 6 správných odpovědí, vzorec (14) by u 80 studentů s touto vlastností dal $-8{,}581$! Takové studenty bychom tedy nedokázali rozlišit. Mají příliš podobné hodnoty statistiky, podle které je posuzujeme. Na druhé straně, u celkem 28 studentů s nejvýš dvěma nebo aspoň devíti body (tedy směs velmi slabých a velmi dobrých) vyšlo 0,898.

Uvedenou situaci dokumentuje i další příklad. Studentka učitelství na MFF UK se ve svojí diplomové práci zabývala nejrůznějšími pohledy na výsledky přijímacích zkoušek ke studiu fyziky na této fakultě (Brychtová (2001)). V daném ročníku psali u přijímacích zkoušek písemku z fyziky o čtyřech úlohách celkem 103 uchazeči, z nich pouze 66 bylo přijato a skutečně nastoupilo. V prvním případě, kdy je variabilita měřené vlastnosti nepochybně větší, vyšlo 0,735. Pro skutečně nastoupivší dostaneme pouze 0,506. Jaké Cronbachovo alfa bychom dostali ze všech písemek, kdyby je museli psát všichni uchazeči, tedy i ti, které fakulta (vzhledem k vynikajícímu prospěchu na střední škole nebo úspěchům v matematické olympiádě) přijala bez zkoušení?

Z obou příkladů je zřejmé, že Cronbachovo alfa je skutečně relativní mírou přesnosti našich měření. Navíc hodnota jeho odhadu může záviset na měřítku, jaké zvolíme pro jednotlivé položky.

7. Závěr?

Má vůbec smysl měřit reliabilitu pomocí Cronbachova alfa, když odhaduje pouze dolní mez pro hledanou charakteristiku? Jaké máme právo použít postup založený na představě o spojitých veličinách, který kromě jiného předpokládá, že rozptyl je nezávislý na úrovni měřené veličiny, když naše měření

jsou nulajedničkové veličiny? Neměli bychom raději v modelu logistické regrese s dvojným tříděním podle testovaných jedinců a podle položek testovat podmodel, v němž na testovaných jedincích nezáleží? K tomu se používá jako testová statistika rozdíl deviancí v podmodelu a modelu $D(B) - D(A+B)$, který má za platnosti nulové hypotézy rozdělení $\chi^2(n-1)$. Když se necháme inspirovat vztahem (13), dostaneme pro test s nulajedničkovými odpověďmi upravené (logistické?) Cronbachovo alfa

$$\hat{\alpha}_{\text{logist}} = 1 - \frac{230}{396{,}8868} = 0{,}420,$$

neboť test psalo 231 studentů a rozdíl deviancí podmodelu a modelu je roven 396,8868. Připomeňme, že klasický výpočet dal poněkud menší hodnotu 0,403.

Nebo jiný námět – neměli bychom místo prostého součtu bodů z jednotlivých položek hledat takovou jejich lineární kombinaci, která dané jedince rozliší co nejlépe? Šlo by i potom o odhad Cronbachova alfa?

Jak mi právem naznačil recenzent tohoto textu, je na místě pokusit se o odpověď na položenou otázku. Jako u každého statistického postupu je třeba i v případě hodnocení reliability pídit se po přiměřenosti zvoleného postupu. Například, když odhadnu parametry regresní přímky, podívám se také na diagramy reziduí. Navíc, získané odhady si troufnu použít k predikci nanejvýš v rozsahu hodnot nezávisle proměnné, který jsem měl k disposici při výpočtu odhadu. Ani Cronbachovo alfa nelze používat mechanicky, v každé situaci, v níž by čistě technicky počítač uměl spočítat jeho hodnotu. Vypovídá jen o reliabilitě měření prováděných za stejných podmínek, tedy u výběru ze stejné populace. Pokusil jsem se ukázat, k čemu může vést nevhodné použití.

8. Náznak důkazu vztahu (6)

Abychom dokázali nerovnost (6), dokážeme nejprve, že pro prvky σ_{ij} varianční matice Σ náhodného vektoru $\mathbf{B} = (B_1, \ldots, B_m)'$ platí nerovnost

$$\sum_j \sum_i \sigma_{ji} \geq \frac{m}{m-1} \sum \sum_{i \neq j} \sigma_{ij}. \tag{15}$$

Vyjdeme z nerovnosti

$$0 \leq (\sqrt{\sigma_{ii}} - \sqrt{\sigma_{jj}})^2 = \sigma_{ii} + \sigma_{jj} - 2\sqrt{\sigma_{ii}\sigma_{jj}},$$

v níž nastává rovnost právě tehdy, když jsou oba rozptyly stejné. Protože je korelační koeficient shora ohraničen jedničkou, můžeme poslední nerovnost upravit na

$$\sigma_{ii} + \sigma_{jj} \geq 2\sqrt{\sigma_{ii}\sigma_{jj}} \geq 2\sigma_{ij}, \tag{16}$$

přičemž k rovnosti musí být nejen stejné rozptyly, ale navíc obě náhodné veličiny musí být s jednotkovou pravděpodobností vázány lineárním vztahem

$$B_j = B_i + \gamma_{ji} \tag{17}$$

pro nějaké konstanty γ_{ji}. Abychom došli k (9), stačí zvolit za A průměr z náhodných veličin B_i a sečíst (17) přes i. Dostaneme tak

$$B_j = A + \sum_{i=1}^{m} \gamma_{ji}/m = A + \beta_j.$$

Dalším sčítáním přes j zjistíme, že takto zavedené konstanty β_j mají nutně nulový součet.

Když krajní výrazy v nerovnosti (16) sečteme přes všechna $i \neq j$, dostaneme po úpravě nerovnost

$$\sum_j \sigma_{jj} \geq \frac{1}{m-1} \sum_j \sum_{i \neq j} \sigma_{ji}.$$

Odtud se již snadno dostane nerovnost (15). Reliabilitu součtu $B_\bullet$ pak můžeme odhadnout zdola:

$$\begin{aligned} \mathrm{rel}\,(B_\bullet) = \frac{\sum_j \sum_i \sigma_{ij}}{\mathrm{var}\,(\sum_j Y_j)} &\geq \frac{m}{m-1} \frac{\sum \sum_{j \neq i} \sigma_{ij}}{\mathrm{var}\,(\sum_j Y_j)} \\ &= \frac{m}{m-1} \frac{\mathrm{var}\,(\sum Y_j) - \sum_j \mathrm{var}\, Y_j}{\mathrm{var}\,(\sum_j Y_j)} \\ &= \frac{m}{m-1} \left(1 - \frac{\sum_j \mathrm{var}\, Y_j}{\mathrm{var}\,(\sum_j Y_j)} \right). \end{aligned}$$

9. Literatura

[1] K. Brychtová. Reliabilita přijímací zkoušky z fyziky na MFF UK. Diplomová práce, MFF UK, Praha.

[2] L. J. Cronbach (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

[3] J. Hniličková, M. Josífko, A. Tuček (1972). *Didaktické testy a jejich statistické zpracování*. SPN, Praha.

[4] M. R. Novick, Ch. Lewis (1966). Coefficient alpha and the reliability of composite measurements. Technical report, Education Testing Service, Princeton, New Jersey.

[5] P. J. Rulon (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educ. Rev.*, 9, 99–103.

ROBUST'2002 z pohledu mladého nadšence – aneb...

Pavel Stříž

Podle očekávání všechno od začátku probíhalo špatně. Občas vysvitl i *paprsek* naděje, že úplně všechno úplně špatně není. Ale o tom až za okamžik.

ROBUST'2002 (konaný v Mezinárodním centru duchovní obnovy v Hejnicích, ve dnech 21.–25. ledna) nebyl jen o ROBUSTu, ale v našich případech (mluvím za nás Zlíňáky – za sebe a za pana akademika Rytíře) to bylo o věcech před ním i po něm. Začalo to reagovat, jako řada jiných věcí, zhola nevině vtipem a malou narážkou pana akademika Rytíře, že bych se měl podívat do světa na zkušenou. To jsem ovšem hned zamítl, ale později se to ukázalo jako ne úplně špatná myšlenka. Tehdy jsem dodělal první dvě mapy hlavního hráče k bridži k jejich dalšímu použití[1] a tak jsem si říkal, že bych to mohl prezentovat odborníkům na slovo vzatým. S maximálním rizikem, že to bude ostuda jako hrom.

Po odeslání několikrát rozšířeného abstraktu a podepsání se pod svou práci se to začalo opravdu dít. Na fakultě projevilo pár skvělých lidí zájem a vskutku nám i část nákladů proplatili. Avšak běhání kolem toho bylo docela dost, i když to vypadalo na docela směšné záležitosti. Abych nelhal, nakonec to dopadlo, no ani Vám nebudu říkat jak. Kompletní odjezdové materiály jsme posbírali až v pátek odpoledne před odjezdem. V neděli jsme vyrazili vlakem směrem na Prahu – zvláště pak se zátěží běžek a příznivým zimním počasím nám vlak před nosem skoro ujel.

Během cesty, značně posilněni, jsme se jako vzorní Moraváci, návštěvníci z vesmíru na Zemi, v pražském systému meter ne úplně neztratili. Během setkání a představení s panem akademikem Klaschkou na autobusovém nádraží v Praze jsem se uvedl jako vzorný buran z mimoměsta. A to jsem si říkal, že právě na slušné maminčino vychování si mám dát zvláštní pozor (priorita číslo jedna). To bylo velmi dobře známé první varování. A pak se spustil kolotoč zajímavých událostí. Během různých úvah svědomí, jak nevhodné bylo podat ruku první a to ještě navíc muži v přítomnosti dámy, jsem zapomněl zavazadla v Praze na Florenci. Kvůli tomu jsem čekal v Liberci (blahopřeji postupu do hokejové extraligy a velmi významnému úspěchu i ve fotbale, vlastně kopané) na další autobus *z Prahy* (úspěšně podle instrukcí nalezeno a vráceno, uctivá poklona pracovníkům ČSAD), nedal jsem ovšem zprávu ostatním a tak mi přistavený mikrobus byl nechat tiše ujet...

[1] Jednotlivé ukázkové diagramy z map zobecnit a vytvořit generátor rozdání s účastní hrací metody a otestovat použitelnost některých algoritmů umělé inteligence.

Jeden z těch publikovatelných příběhů[2] je setkání, první kontakt, s panem akademikem Tvrdíkem[3].

Až jsem si dostatečněkrát prohlédl krásy Liberce a dorazil mi expres autobus z Prahy se zavazadly, tak jsem si musel zavelet a vybrat: s ostudou to zabalit hned a hurá domů (↩) a nebo vyrazit do Hejnic (↬). Nakonec náhodným výběrem jsem zvolil druhou variantu (↬). Vlak mě zastihl a při průměrném zjišťování, zda-li nejsem již v Polsku, jsem dokonce do Hejnic dorazil. Po přeptání se několika domorodců, kde se vyskytuje místní grandhotel, jsem se dostal až do bodu P (viz obrázek 1 a dále se žíhám již jen zde, osa x představuje okolí a osa y představuje náročnost terénu). Tvrz jsem podle slov domorodců poznal, do bodu P jsem dorazil již za tmy a tak se mi tam toto skvostné sídlo ukazovalo v celé své osvětlené kráse. Po prvních dojmech jsem si jen s povzdechem pomyslel: „*Nó to zas bude vostudý*".

Při přechodu (směr tučná šipka) jsem narazil na pana akademika Rytíře, který kráčel z kontrolní prohlídky po okolí (tato budova na obrázku vskutku není, ale pivo tam točili). V prostoru kolem bodu 0, asi jste to také, vážení účastníci, nepřehlédli, se dalo vskutku velmi pěkně zapadnout do sněhového korýtka – raději jsem zapadnutí otestoval nadvakráte.

Spolu se svým průvodcem (již s okolím blíže obeznámen) jsme se dožihali k vstupní bráně tvrze (bod 1). Nastalo srdečné uvítání od pořadatelů a k zadání prvního bojového úkolu od pana akademika Antocha: „Dobrý den pane kolego, támhle na druhé straně uvázl pan Tvrdík, prosím Vás, zajděte tam a ukažte mu cestu.", na kterém se vskutku nedalo nic zamítnout.

Zrovna v té chvíli jsem neměl nic lepšího na práci a tak jsem vyrazil na obchůzku kolem tvrze. Vydal jsem se tedy z klíčového bodu 1 kolem bodu 2 do bodu 3. Prostor 0 jsem ověřil preventivně ještě jednou - stále fungoval a asi se nechtěl ani přesunout. Jakmile jsem se vyhrabal, tak jsem si říkal že to budou asi zadní úniková vrátka a že některý z účastníků vyrazil z úplně jiného směru než původně já. Proto se nedostal k bodu 1 okamžitě.

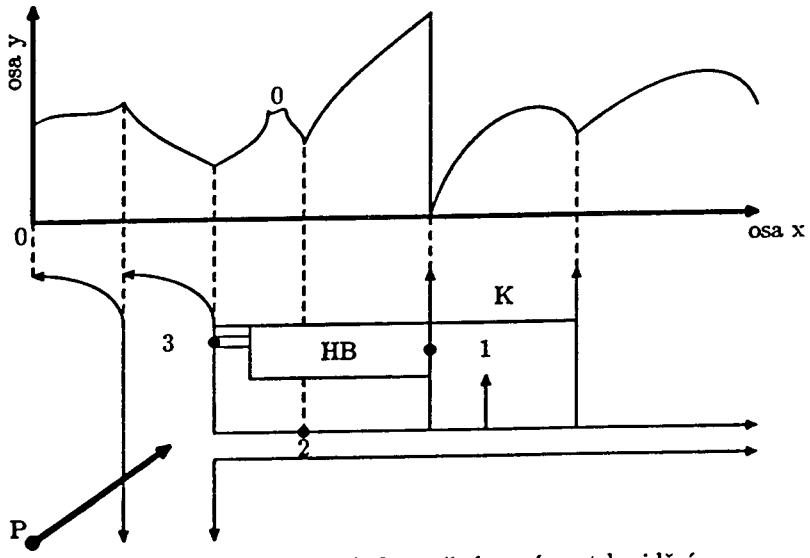
„Dobrý den pane, taky účastník ROBUSTu?" jsem se dotázal oběti hradeb tvrze. „Ano, ano." odvětila oběť. Dříve poučen jsem raději dále pokračoval se vší zdvořilostí: „Pan Antoch Vás pozdravuje a že to máte celé obejít."

---

[2] Jména, události i období našeho letopočtu jsou smyšlena a jinak dále upravena. Tím se tedy omlouvám za své nepřesné citace, ale snad jsem význam nezměnil.

[3] Tímto srdečně pana akademika zdravím. Článek pana Tvrdíka o průměrném věku účastníka ROBUSTu dokumentuje, že jsem snad stále ještě mladý (výběrově). Pod slovem nadšenec v názvu příspěvku se skrývá nadšení z budoucí akce, i když by tam mělo být spíše nadchnutý z budoucí akce.

Nastala minuta pravdy: „Aha, já jsem zkoušel tu bránu[4] a ono nic.“ To už jsem pochopil a jen jsem si myslel, že to určitě bude nějaký ten matematik či nejlépe odborník na umělé neuronové sítě a že se prostě z bodu 3 do bodu 1 (skrz bod 2) jednoduše neprožíhal. Jen mi v té chvíli blikl hlavou parametr (jak by řekli naši kolegové ze vzdálených koutů ciziny) momentum[5]. A zrovna jak naschvál špatně (říká se neoptimálně) nastaven... Neprožíhal se po uvedené křivce přes některé vrcholy a uvázl v lokálním minimu (bod 3) a tak musel nastoupit učitel, tvrdší žíhatel... Když už jsem u zmíněné křivky, je vidět, že z pravé strany mapy je vstup do tvrze (bod 1) mnohem snazší.



Obrázek 1. I takto to bylo možné servírovat k vidění.

Tím úvodní prohlídka okolí Hejnic skončila i pro dalšího účastníka ROBUSTu. Na velmi podrobné (skoro vojensky orientované) mapě je pak *K* rozhledna, kostel či chrám a *B* hlavní generální štáb, hlavní budova či ubytovna. A nebo naopak...

[4] Bod 2 na obrázku 1 pro opravdu zaujaté čtenáře.

[5] Moment zohledňuje při výpočtu změny váhy ve směru gradientu chybové funkce navíc i předchozí změnu vah. Parametr momentu určuje míru vlivu předchozí změny. Pomocí momentu gradientní metoda lépe opisuje tvar chybové funkce, protože bere do úvahy předchozí gradient.

A jak to probíhalo dále? To jste, milí čtenáři, možná také shlédli. Přednášky na velmi vysoké úrovni, výborná společnost, velmi vstřícný personál a skvělé jídlo (to bych speciálně vyzdvihl). Příjemným zjištěním byla skutečnost kolik doktorandů (aspirantů) podpořilo své učitele a spolupracovníky v jejich práci.

Některým našim studentům bych věnoval návštěvu ROBUSTu za trest a některým vskutku za odměnu – každému podle zásluh. Protože kolikrát v životě lze přednášet před desetinásobnou státnicovou komisí; kde v první řadě sedí pan akademik Saxl, jako fiktivní nejvyšší dozorčí, předseda této komise a posluchač nejpozornější, a za ním česká elita v oborech studentem běžně studovaných?

ROBUST'2002 oficiálně skončil, pro nás Zlíňáky, naším stejně rychlým odjezdem (jako příjezdem) v pátek odpoledne, vlakem s panem akademikem Tvrdíkem. Pan akademik nad očekávání nečekaně nepřijel do Hejnic z jiného směru než Zlíňáci.

Vřele děkuji za příležitost vystoupit, kterou jsem rád a snad i plně využil a za možnost býti vyslyšen. Rád bych také poděkoval všem účastníkům ROBUSTu za neobyčejně skvělou atmosféru jak u příspěvku mého, tak u příspěvků ostatních.

U svého drobného příspěvku jsem (z hlediska místa a času) očekával maximálně (i s mluvčím) pět lidí (koeficient nervozity maximálně 5/100), což se vskutku skoro naplnilo s malou chybou, a to že posluchárna byla, vhodně řečeno, plná. Výpočet koeficientu nervozity během mého příspěvku nechám na vážených čtenářích. Před „startem" jsem se nedožíhal ani k tomu, abych zapnul notebook (na znázornění by nestačil jen 2D obrázek). Radost to byla převeliká.

Jako řečník jsem příliš neuspěl. Nejen že jsem přednášku přetáhl (a krátil tak čas profesoru Hájkovi a jeho robustnímu varhannímu koncertu), ale měl jsem v přednášce i hluchá místa. Je to takové místo, které vyžaduje hlubší analýzu a komentář. Měl jsem již dříve očekávat, že posluchač nemusí být s tímto problémem obeznámen a přeptá se. Tím se z profilu dostávám na konkrétní šrouby, které jsou samy o sobě zajímavé, avšak záleží již jen na zkušenostech mluvčího, aby o šroubu přestal mluvit a vrátil se k původnímu profilu (koeficient zkušenosti byl významně podobný očekávanému koeficientu nervozity).

Můj cíl, aby to pro posluchače nebylo příliš náročné (což při zkušenostech publika by bylo téma již alespoň částečně řešené), ani příliš jednoduché (to by bylo nevhodné a drzé), jsem snad splnil – nechť vážený posluchač posoudí sám.

I tak nemohu nepoděkovat za to, že jsem alespoň na chvíli mohl

být součástí této robustní akce.

„Proč není ROBUST každý akademický rok?" se na závěr sám sebe táži. Protože akce takového ražení a kalibru se nedá za jeden normální člověkorok normálním smrtelníkem rozchodit.

Jaký je tedy tzv. karetní post mortem? Až pojedete na další ROBUST, pozor ať se nedožíháte až na Slovensko (do Polska), do Olomouce (Prahy) či jen do Zlína (Liberce). Proto raději nechtějte ...

... v lokálním minimu uváznout!

Na moravské nástrahy se těšte možná již na příštím ROBUSTu!

Univerzita Tomáše Bati, FaME Zlín; p_striz@fame.utb.cz, p-aj-a@email.cz

Omluva redakce

Na tomto místě původně mělo být oznámení o chystaných akcích na podzim tohoto roku. Vzhledem k tomu, že tyto akce byly připravovány ve spolupráci s Českým statistickým úřadem a jeho pracovníci měli být hlavními přednášejícími, z důvodů všem dobře známých se tyto akce odkládají na neurčito. (Blíže na poslední stránce)

V minulém čísle Informačního Bulletinu jsme čtenářům připravili (nechtěně) malý kvíz. Uveřejnili jsme jednu z prvních verzí příspěvku ‚O objektivních příčinách nespokojenosti v divadelních souborech' *kolegy Tvrdíka, obsahujícím některé – řekněme n e p ř e s n o s t i (za což se mu tímto omlouváme). Ti z Vás, kteří na tyto chyby přišli, si mohou připsat deset kladných bodů a budou zařazeni do dalšího kola ke slosování hodnotných cen. Ti, kteří si žádných chyb v minulém čísle dosud nevšimli, mají nyní poslední možnost je najít (pokud nečtou bulletiny a časopisy odzadu a nepřečetli si následující příspěvek dříve než toto oznámení :). Pro ověření Vaší bystrosti (a uvedení věcí na pravou míru) uveřejňujeme nyní tu správnou verzi příspěvku a omluvu autorovi. -gd-*

Objektivní příčiny nespokojenosti v divadelních souborech

Josef Tvrdík, OU

Na oslavu padesátin našeho kamaráda-učitele napsal jiný kamarád, právník, veršovanou jednoaktovku nazvanou Komenského hledání. Ve hře vystupuje sedm postav a jejich obsazení vzniklo vcelku spontánně. Roli nejhlavnější samozřejmě uchvátil autor a současně režisér pro sebe. Jelikož jsem dostal jednu z pěti vedlejších (ale významných) rolí, považoval jsem obsazení za správné. Narozeninová premiéra měla velmi příznivý ohlas. Celkem pochopitelně mezi neúčinkujícími pak zazněly poznámky o jejich nespravedlivém opomenutí. Proč zrovna oněch sedm bylo obsazeno, když kmenový stav potenciálních herců byl 22? Jistě bylo možné i obsazení jiné, které by oni považovali za lepší. Těch jiných obsazení je opravdu mnoho. Hra nevyžaduje, aby dámské role hrály dámy a pánské role pánové, dokonce ani při premiéře to tak nebylo, takže sedmička herců mohla být vybrána $\binom{22}{7}$=170544 způsoby. Z těchto sedmic lze pak pak sestavit $\binom{22}{7}7!$ = 859541760 různých obsazení.

V opravdových divadelních souborech je situace pro spokojenost členů souboru většinou trochu příznivější. Herci i role jsou rozděleni do kategorií podle věku, pohlaví atd. Uvažujme, že je k takových kategorií, v kategorii je m_i rolí a soubor disponuje n_i herci v kategorii, $0 \leq m_i \leq n_i$, i=1, 2, ..., k. Počet možných obsazení N je pak

$$N = \prod_{i=1}^{k} \binom{n_i}{m_i} m_i!$$

Vezmeme-li dva docela realistické příklady s číselnými hodnotami z následující tabulky, dostaneme N_1 = 2 592 , resp. N_2 = 54 000 možných obsazení.

i	1	2	3	4	Příklad
m_i	2	3	1	1	1,2
n_i	4	4	3	3	1
n_i	6	5	6	5	2

Naději na spokojenost snad má divadlo jednoho herce, soubory s jedním hercem v každé kategorii a tomu přizpůsobeným repertoárem nebo možná soubory nehrající vůbec nic. V ostatních případech je často mnoho možných obsazení a mnoho nespokojených herců.

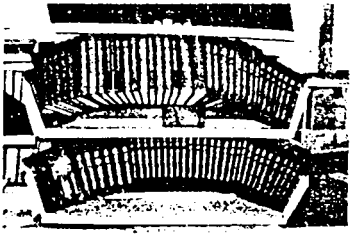
Statistika ~~na~~ ve vodě

O letošních záplavách se toho napsalo a namluvilo už dost, nicméně jsou tu dvě věci, o kterých se příliš nemluvilo a které se přímo dotýkají české statistiky. První z nich je Český statistický úřad, který se ocitl z poloviny pod vodou, „z poloviny" ve smyslu vertikálním.

Kromě obrovských problémů při jeho fungování v provizorních podmínkách, postihlo to i naši společnost, které ČSÚ laskavě vyráběl ve svém reprografickém středisku Informační bulletin, a toto středisko se nacházelo právě v té „spodní" zatopené polovině. Číslo, které právě dostáváte, bylo vyrobeno „na koleně" v provizorních podmínkách, a proto jsme se rozhodli už definitivně přestat posílat IB těm členům, kteří se dosud neuhradili členský příspěvek za minulá léta.

Druhá pohroma se snesla na knihovnu matematicko-fyzikální fakulty v pražském Karlíně. Ta byla z větší části zničena a zbytek bude jen těžko zachraňován. Účastníci semináře STAKAN jistě pamatují na příjemné prostředí profesorské knihovny, v jejíž regálech se skrývalo mnoho vzácných knih. Stačilo pár dní a knihovna i knihy jsou pryč. (Na stránce www.karlin.mff.cuni.cz/fakulta/lib je uveřejněno několik fotografií) -gd-

Voda opadla a nastoupila plíseň



Jak je dlouhý metr knih? Tenhle 108cm.

Profesorská knihovna ...

.....

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