INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 22, číslo 4, prosinec 2010

Informační Bulletin

České Statistické Společnosti *číslo 4, ročník 22, prosinec 2010*

Česká statistická společnost založena!

Tak zněl nadpis prvního příspěvku Informačního Bulletinu číslo 1 z května 1990. Po 51 letech od zániku první Československé statistické společnosti tak vznikla její nástupnická organizace – Česká statistická společnost (ČStS). Ustavující zasedání se konalo 29. března 1990 v Praze. Do té doby se statistici v Československu sdružovali buď v Jednotě československých matematiků a fyziků, Slovenské štatistické a demografické spoľočnosťi (SŠDS) či České demografické společnosti. Poznamenejme, že většina československých statističek a statistiků byla členem jedné z prvních dvou výše jmenovaných společností. Během jednání se slovenskými kolegy počátkem devadesátých let se jednoznačně ukázalo, že bude lepší zachovat stávající SŠDS a nově založit ČStS. Rozdělení Československa toto rozhodnutí jenom potvrdilo.

Při svém založení měla ČStS 195 členů[1]. Byly vytvořeny tři sekce: ekonomická a státní statistika, statistická teorie a metody, aplikovaná a výpočetní statistika. Přes toto formální dělení byla od samého počátku jedním z nejdůležitějších cílů společnosti snaha sdružovat všechny zabývající se „stochastikou“ všech oborů a zaměření v jedné společné organizaci. Hlavní vždy byla myšlenka, že existuje pouze „dobrá statistika“. Prosazování této myšlenky se dařilo, a s jistými výhradami se daří dodnes. S tímto cílem jsme též podporovali všechny aktivity naší společnosti.

Dalším cílem naší společnosti byla snaha o její začlenění do širšího než národního kontextu. Od samého počátku jsme úzce spolupracovali s našimi slovenskými kolegy. Od roku 2004 jsme se stali členy skupiny V6 (Visegrad 6) spolu s národními společnostmi Slovenska, Slovinska, Maďarska, Rakouska

[1] Dnes jich má kolem 250.

a Rumunska. Se zástupci těchto společností si vzájemně vyměňujeme informace, pořádáme každoročně setkání a do budoucna plánujeme i vydávání společného časopisu. V poslední době jsme navázali kontakty i s dalšími evropskými společnostmi.

Nejdůležitější funkcí každé společnosti, a tedy i ČStS, je její „obnova". To znamená zapojení a podpora mladých členů a udržení zájmu veřejnosti, především té odborné. Společnost se proto snaží o podporu mladých statistiků ze všech oborů. Každoročně jim poskytujeme finanční podporu umožňující účast na konferencích, organizujeme soutěže o nejlepší studentské práce a organizovali jsme i mezinárodní studentskou konferenci. Každé dva roky pořádáme konferenci o výuce statistiky pro nestatistiky.

Většina z příspěvků publikovaných v tomto čísle Informačního Bulletinu zazněla při slavnostním zasedání společnosti 5. září 2010 v Brně, které se uskutečnilo v rámci Brněnských statistických dnů. Ty se konaly v Akademii Sting pod záštitou rektora této akademie ve dnech 2. – 4. září 2010. Slavnostní zasedání proběhlo v důstojném prostředí novobarokní auly Vysokého učení technického pod záštitou rektora VUT Brno. Na zasedání nechyběla ani barokní hudba[2] a zahraniční hosté ze Slovenska a Rakouska. Škoda jen, že účast našich členů byla poměrně slabá. O to důstojnější a srdečnější bylo prezenční předávání několika málo pamětních diplomů předsedou společnosti na společenském večeru!

Historie ČStS je do jisté míry zachycena na webových stránkách společnosti `http://statspol.cz/`, kde lze nalézt nejenom přehled předsedů a členů výborů od roku 1990, ale také archív Informačního Bulletinu ČStS, ve kterém je zaznamenána většina událostí, kterých se společnost v uplynulých dvaceti letech účastnila. Na závěr bychom rádi naší společnosti popřáli všechno nejlepší do dalších alespoň dvaceti let úspěšné existence a resistence!

V Praze dne 31. 12. 2010

Gejza Dohnal a Jaromír Antoch

[2] Po zasedání nemohla chybět ani lidová hudba z pomezí Moravy a Slovenska.

MEZIVÁLEČNÁ ČESKOSLOVENSKÁ STATISTICKÁ SPOLEČNOST

Prokop Závodský

Adresa: VŠE, FIS, KSTP, nám. W. Churchila 4, 130 67 Praha 3

E-mail: zavodsky@vse.cz

Abstrakt

Československá statistická společnost byla založena v lednu 1929 a sdružovala několik desítek špičkových odborníků v oboru statistiky a příbuzných disciplín formou řádných a dopisujících členů. Činnost Společnosti ustala koncem roku 1938 a její obnova v r. 1948 měla jen krátké trvání.

Czechoslovak Statistical Society was founded in January 1929. Its members were several super experts in the field of statistics and related sciences. It associated both regular and corresponding members. Its activity was stopped in 1938 and its restoration in 1948 lasted only for a short time period.

Návrhy na založení statistické společnosti, jaké tehdy již existovaly v řadě vyspělých zemí, se u nás objevovaly po celá dvacátá léta minulého století. Zpočátku neměly tyto návrhy úspěch. Statistická obec byla u nás zatím nepočetná, navíc se prakticky všichni renomovaní statistici scházeli pravidelně od roku 1920 ve Statistické radě státní (SRS) – v plénu či v jejích různých výborech.

Koncem 20. let již byla situace příznivější. Především významně přibylo statistických specialistů – počet zaměstnanců Státního úřadu statistického (SÚS) vzrostl od jeho založení zhruba dvacetkrát (na 667 v r. 1929), přičemž u konceptních úředníků se vyžadovala vědecká a publikační činnost. Profesoři a docenti statistiky i příbuzných věd přibývali také na nově založených vysokých školách (Vysoká škola obchodní a Vysoká škola speciálních nauk ČVUT, nové university v Brně a Bratislavě ad.) i na stávajících. Kvalifikované statistiky v rostoucí míře zaměstnávala ministerstva a další centrální úřady čs. státu i různé hospodářské aj. instituce.

Hlavním iniciátorem založení Společnosti byl František Weyr, chystající se právě ukončit své téměř desetileté presidentství SÚS. Svým oponentům, kteří poukazovali na existenci SRS, argumentoval tím, že SRS má oficiální poslání (především projednávala a rozhodovala o programu činnosti státní statistické služby), její členové v ní působí jako zástupci institucí, které je

delegovaly atd. Naproti tomu Statistická společnost bude sdružovat nejvýznamnější statistiky, pracující v různých oborech, umožňovat jim výměnu zkušeností a podporovat vědecké bádání ve statistice.

Desetičlenný přípravný výbor se sešel 28. listopadu 1928, rozhodl o založení Společnosti a projednal návrh stanov. 30. ledna 1929 pak přípravný výbor zorganizoval ustavující schůzi Československé statistické společnosti.

Předsedou Společnosti byl zvolen Vilibald Mildschuh, 51letý profesor statistiky a národního hospodářství na Právnické fakultě UK, který měl i dlouholeté zkušenosti s úřední statistikou z pražské Zemské statistické kanceláře, kde působil v letech 1904–1917, i z práce ve SRS. Podle Weyrových Pamětí [6] se Mildschuh ucházel neúspěšně o funkci předsedy (presidenta) SÚS, když se uvolnila v r. 1919 a opět o deset let později.

Prvním místopředsedou byl zvolen Bohdan Živanský, tehdy 53letý generální tajemník ústředny obchodních komor, zabývající se dlouhodobě zejména populační, národnostní a hospodářskou statistikou. Do funkce 2. místopředsedy ustavující schůze zvolila Antonína Boháče, 47letého přednostu odboru populační statistiky SÚS, který je dnes považován za zakladatele čs. demografie.

Na schůzi byli zvoleni též následující čtyři starší pánové za členy představenstva:

- Karl Berthold – pensionovaný přednosta již zrušeného miniaturního, leč agilního Zemského statistického úřadu v Opavě;
- Cyril Horáček Starší – profesor národního hospodářství na Právnické fakultě UK;
- Rudolf Hotowetz – významný národohospodář, politik a publicista, tehdy vládní komisař Všeobecného pensijního ústavu;
- Heinrich Rauchberg – profesor statistiky na německé universitě v Praze.

Přípravný výbor na této své ustavující schůzi zvolil 30 řádných členů Společnosti (včetně 7 výše jmenovaných funkcionářů) a 22 dopisujících členů. Jednatelem Společnosti byl zvolen její nejmladší řádný člen – doc. Jaroslav Janko (* 1893), pokračovatel nedávno zemřelého prof. Josefa Beneše v přednáškách o pojistné matematice a matematické statistice na Vysoké škole speciálních nauk.

Společnost se stala podle svých stanov vědeckou akademií s omezeným počtem členů – počet řádných členů během desetileté činnosti nepřesáhl 38, dopisujících členů bylo obvykle asi o desítku méně.

Mezi řádnými členy nalezneme vysokoškolské profesory statistiky a národního hospodářství – V. Mildschuh, C. Horáček Starší i Mladší, Karel Engliš,

F. X. Hodáč, Dobroslav Krejčí, z Němců pak H. Rauchberg a Franz Xaver Weiss. Další skupinu tvořili vysokoškolští učitelé pojistné matematiky a matematické statistiky – J. Janko, Emil Schönbaum, později Ladislav Truksa, z pražské německé techniky Gustav Rosmanith. Vysokoškolské učitele z oblasti přírodních věd zastupovali mj.: matematik Karel Petr, fyzik František Nachtikal, z Němců Josef Fuhrich. Početnou skupinu řádných členů tvořili vedoucí představitelé a specialisté ze SÚS: F. Weyr, jeho nástupce ve funkci presidenta Jan Auerhan, Auerhanův náměstek a později dočasný nástupce A. Boháč, Josef Mráz, Robert Kollar ad. Nechyběli ani představitelé hospodářské politiky, podnikatelského sektoru, odborů apod. – B. Živanský, Slovák Kornel Stodola, zástupce německých podnikatelů Robert Mayer ad.

Velká většina řádných členů byla zvolena ve věku 40–53 let. Slováků najdeme mezi řádnými členy velmi málo, Německé národnosti bylo 7 členů. Nulový podíl něžného pohlaví na činnosti Společnosti nebyl projevem nějaké mizogynie mezi voliteli. V meziválečném období u nás žádná žena ve statistice nevynikla.

První valné shromáždění Čs. statistické společnosti se sešlo 26. dubna 1929 a další následovala v ročních intervalech (obvykle v červnu). Na valných shromážděních a na dalších schůzích (nejčastěji se konaly šestkrát za rok) vyslechli účastníci plánované přednášky, na něž navazovala diskuse, která často pokračovala i na dalších schůzích. Texty přednášek a diskusních příspěvků pravidelně otiskoval Čs. statistický věstník (od r. 1931 přejmenovaný na Statistický obzor) a zkrácenou verzi ve francouzštině obvykle také Revue de l'Institut international de statistique, vycházející v Haagu.

Jako první přednášel místopředseda Společnosti Boháč (Náš populační problém), převažovala témata z oblasti sociálně hospodářské statistiky a demografie, ale zazněly i výklady o matematické statistice (např. Janko – Náhodné výběry v technické praxi – 1933). Společnost navázala brzy spolupráci s podobnými společnostmi v zahraničí. Zástupci Čs. společnosti navštěvovali akce svých kolegů, zejména z blízkých evropských zemí, a Čs. statistický věstník/obzor pravidelně informoval o činnosti zahraničních statistických společností. Řada vedoucích představitelů zahraničních statistických úřadů a statistických společností byla zvolena čestnými členy Čs. statistické společnosti.

Činnost Společnosti byla z velké části dotována SÚS, členské příspěvky či pod. se neplatily. Výdaje ostatně nebyly až tak vysoké, jejich podstatnou část tvořily náklady na cesty na zahraniční konference, později byly též honorovány přednášky členů Společnosti.

Vyvrcholením činnosti Společnosti mělo být 24. zasedání Mezinárodního statistického institutu (ISI) v Praze, na jehož přípravě se velká část členů podílela. Zasedání bylo zahájeno 12. září 1938 v Rudolfinu (tehdy sídlo po-

slanecké sněmovny) za účasti 145 delegátů ze 33 zemí (což bylo více než na kterémkoli z předcházejících čtyř zasedání).

Poslední předválečné zasedání ISI (a dosud jediné na našem území) mělo nečekaný průběh: V den zahájení zasedání ISI pronesl v Norimberku na sjezdu NSDAP Hitler protičeskoslovenský válečnický projev, po něm následovaly henleinovské provokace v pohraničí. Ve strachu, že může brzy následovat válečný konflikt i bombardování Prahy, začali následujícího dne mnozí účastníci ve spěchu opouštět Prahu, takže večer (13. září) bylo rozhodnuto zasedání předčasně ukončit.

Úmrtí profesora statistiky na německé universitě Rauchberga (26. září 1938), jemuž četné spory s českými statistiky v národnostních otázkách nebránily v loajalitě k Československé republice, symbolizovalo konec jedné epochy. 24. ledna 1939 umírá předseda Společnosti Mildschuh a s následujícím zánikem Československa končí i činnost Čs. statistické společnosti.

17. listopadu 1939 jsou okupanty uzavřeny všechny české vysoké školy (přednášela zde velká část členů Společnosti), další osudy členů byly často tragické. Bývalý president SÚS Auerhan byl popraven za heydrichiády, Rudolf Tayerle zahynul téhož roku v Mauthausenu. Prof. Weiss byl z rasových důvodů vyhozen z německé university v Praze a jeho další osud mi není znám. Boháč, Kollar, Ryba a další přední statistici ze SÚS byli pensionováni, jiní ovšem pokračovali v práci (např. Leopold Šauer si zde vysloužil i přední protektorátní vyznamenání – svatováclavskou orlici, takže měl po válce co vysvětlovat). Dlouholetý německý pracovník SÚS Albin Oberschall, který pak vedl (od r. 1941) protektorátní statistický úřad, zemřel během poválečného odsunu německého obyvatelstva. V letech 1939–1945 zemřeli profesoři Horáček St., Hodáč a Nachtikal i JUDr. Hotowetz, prof. Schönbaum se zachránil před rasovým pronásledováním emigrací do Latinské Ameriky, německý profesor Leo W. Pollak, rozvíjející statistické metody v meteorologii, se vystěhoval do Irska již r. 1938.

Po osvobození byly podniknuty pokusy o obnovu činnosti Společnosti. Teprve 17. března 1948 se podařilo uskutečnit valné shromáždění České statistické společnosti s volbou funkcionářů i nových členů a přednáškou Stanislava Režného o statistice zahraničního obchodu. Ale tato labutí píseň zazněla až po „vítězném únoru“, kdy pro jakousi elitářskou statistickou akademii nebylo v organizaci socialistické vědy místo[1] .

[1] Ani demokratická myšlenka Statistické rady státní se neslučovala s totalitním vládnutím. Zatímco za protektorátu zůstala SRS v právním řádu formálně zachována, ale nepracovala, za komunistického režimu byla již r. 1948 zrušena.

Profesoři Schönbaum a Brdlík pokračovali ve své práci až za oceánem, prof. Weyr se dožil nejen svého vyakčnění z Právnické fakulty, kterou zakládal a jejímž byl prvním děkanem, ale i jejího zrušení, prof. Engliš byl (podobně jako A. Boháč za německé okupace) vystěhován z Prahy...

V. Mildschuh – převzato z [1] a B. Živanský – převzato z [9], roč. XVII (1936), s. 452

A. Boháč – převzato z [4] a J. Janko – Archiv ČVUT

Literatura

[1] Drachovský J.: Vilibald Mildchuh. Praha 1929.

[2] Kořalka J.: Historie Mezinárodního statistického institutu a jeho přínos k mezinárodní spolupráci statistiků. Diplomová práce. VŠE, Praha 1981.

[3] Podzimek J.: Biografický slovník české statistiky. Příloha časopisu Statistika, roč. 30 (1993).

[4] Šubrtová A.: Antonín Boháč – statistik a demograf. Praha 1979.

[5] Weyr F.: Paměti I. Za Rakouska (1879–1918). Brno 1999.

[6] Weyr F.: Paměti II. Za republiky (1918–1938). Brno 2001.

[7] Závodský P.: 85 let od vzniku státní statistiky na území České republiky. Statistika, roč. 42 (2005), č. 3, s. 254–262.

[8] Československý statistický věstník, roč. X (1929) – XI (1930).

[9] Statistický obzor, roč. XII (1931) – XXVIII (1948).

STATISTIKA A POČÍTAČE, STUDENTI A UČITELÉ

Jiří Anděl

Adresa: MFF UK, KPMS, Sokolovská 83, 186 00, Praha 8

E-mail: jiri.andel@mff.cuni.cz

Abstrakt

V příspěvku je pojednáno o tom, jak počítače ovlivnily statistickou praxi. Snadnost a rychlost výpočtů vyvolává řadu dalších otázek, pro jejichž řešení chybí odpovídající teorie. Je poukázáno na fakt, že u studentů přicházejících na vysokou školu klesá úroveň jejich matematických znalostí. Přitom studentské ankety naopak signalizují, že studenti nejsou spokojeni s výukou některých základních předmětů z matematické statistiky.

The paper contains some considerations how computers have influenced practical applications of statistics. Easy and fast calculations raise additional questions which can be hardly answered since the corresponding theoretical results are not known. It is demonstrated that the mathematical skills of new students decrease. However, also our students are not satisfied with some introductory statistical courses.

1. Počítače a statistická praxe

Je velice obtížné najít uspokojivou odpověď na otázku, jaké byly nejvýznamnější pokroky matematické statistiky za posledních dvacet let. Už jen samotný počet publikovaných článků a knih dramaticky rostl a roste. V každé oblasti matematické statistiky by se našly vynikající výsledky, které by však bylo těžké mezi sebou srovnávat.

Na jedné věci bychom se nepochybně shodli všichni. Tou je rozvoj výpočetní techniky a její využití ve statistické praxi. Na začátku se často zjišťovalo, zda statistické programy počítají správně. Statistici se vzájemně upozorňovali na existující nedopatření a o některých chybách se psalo i v odborném tisku. Toto období je podle mého názoru již za námi. K chybám při tvorbě programů dochází samozřejmě i nadále, ale základní procedury v renomovaných programech byly již mnohokrát odzkoušeny a patrně je na ně spolehnutí. Teď se občas dostáváme do jiné situace. Vývoj přináší nové verze programů a občas se stává, že se ztratí zpětná kompatibilita. To znamená, že prográmek, který po řadu let dobře sloužil, nyní nelze spustit. Snadnost a rychlost výpočtů vede k tomu, že si přejeme získat další informace. Tady často narážíme z toho důvodu, že není známa teoretická metoda, jak výpočet uskutečnit nebo jak výsledek vyhodnotit. Tyto obecné úvahy budu demonstrovat na příkladě, který možná bude zajímavý i sám o sobě. Výpočty byly prováděny programem `R`.

Byla sledována závislost mezi výší příjmu a spokojenosti s prací (viz Agresti 2002, tab. 2.8 na str. 57 a odst. 3.4.2 na str. 87; dále viz Presnell 2000). Spokojenost (stručně satisfac) se klasifikovala do tříd VD (very dissatisfied, tj. velmi nespokojen), LS (low satisfaction, tj. málo spokojen), MS (medium satisfaction, tj. středně spokojen), VS (very satisfied, tj. velmi spokojen). Příjem (income) se členil do tříd < 5 (do 5 000 $ ročně), $5 - 15$ (od 5 000 $ do 15 000 $ ročně), $15 - 25$ (od 15 000 $ do 25 000 $ ročně) a > 25 (více než 25 000 $ ročně). Získala se následující kontingenční tabulka, kterou budeme nazývat `jobsatis`.

```
       Satisfac
Income  VD LS MS VS
  <5     2  4 13  3
  5-15   2  6 22  4
  15-25  0  1 15  8
  >25    0  3 13  8
```

Pearsonova statistika pro test nezávislosti vyšla $\chi^2 = 11.52$, což při 9 stupních volnosti dává p-hodnotu 0.24. Zároveň však dostáváme varovné hlášení

```
Warning message:
In chisq.test(jobsatis) : Chi-squared approximation may be
                          incorrect
```

Nespokojenost počítače se dala očekávat, vždyť hodně políček kontingenční tabulky má malé četnosti. Proto aproximace rozdělení Pearsonovy statistiky rozdělením χ^2 nemusí být dostatečně přesná. Díky výpočetní technice můžeme skutečnou p-hodnotu Pearsonovy statistiky zjistit pomocí simulací. Na to je v programu `R` k disposici příkaz, takže postup je velmi jednoduchý. Na základě 10 000 simulací vyšla p-hodnota rovněž 0.24. Dokonce se také může spočítat Fisherův faktoriálový test v této tabulce. Jeho p-hodnota vyšla 0.23. Žádná z těchto variant tedy nevedla ani zdaleka k signifikantnímu výsledku.

Často se stává, že kontingenční tabulka má uspořádané kategorie, jako je tomu v případě našich dat. Někdy jsou tyto kategorie ordinální, jako je např. Likertova stupnice: vůbec nesouhlasím — nesouhlasím — neutrální — souhlasím — velmi souhlasím), v jiných případech mají pevně danou číselnou hodnotu (např. počet dětí v rodině). Tato číselná hodnota se nazývá skór. (Ve statistickém slovníku můžeme najít, že skór je záznam polohy na škále.) Nechť $u_1 \leq u_2 \leq \cdots \leq u_r$ jsou řádkové skóry a $v_1 \leq v_2 \leq \cdots \leq v_c$ jsou sloupcové skóry. Nechť r je výběrový korelační koeficient počítaný na základě četností uvedených v kontingenční tabulce. Součet všech četností v kontingenční tabulce označme n. Definujme $M^2 = (n-1)r^2$. Pak za předpokladu nezávislosti má M^2 přibližně rozdělení χ^2_1 (viz Agresti 2002, str. 87, Mantel 1963). Tento výsledek si můžeme přiblížit následujícím způsobem. Při výběru o rozsahu $n \geq 3$ z dvojrozměrného normálního rozdělení pro výběrový korelační koeficient r v případě $\rho = 0$ platí $\mathsf{E}r = 0$, $\mathsf{var}\, r = 1/(n-1)$. Z asymptotické normality koeficientu r plyne, že M^2 má asymptoticky rozdělení χ^2_1.

Pomocí korelačního koeficientu vyhodnotíme nově data uvedená v tabulce `jobsatis`. Ve výše citované literatuře bylo navrženo, aby se jako sloupcové skóry použila čísla 1, 2, 3, 4 a jako řádkové skóry čísla 2.5, 10, 20, 30. Program (či spíše prográmek) je uveden v Thompson (2007), str. 40. Program však funguje jen ve starších verzích programu `R`, např. v `R 2.4.1`. V novějších verzích, např. v `R 2.11.0`, musíme napsat vlastní program. Hodnota korelačního koeficientu `r` vyšla 0.27, hodnota M^2 vyšla 7.73 a její p-hodnota je 0.0054. Ačkoliv původní test nezávislosti vyšel zcela nevýznamně, test založený na veličině M^2 vyšel vysoce signifikantně.

Řádkové a sloupcové skóry však byly voleny poměrně libovolně. Položme si otázku, jak vypočítat takové skóry, aby jim odpovídal maximální možný korelační koeficient r. Je známo, že tyto skóry odpovídají největší kanonické

korelaci, která se počítá například v korespondenční analýze. Příslušná kanonická korelace je pak rovna hledanému maximálnímu korelačnímu koeficientu. Výpočet optimálních skórů byl proveden na datech uvedených v tabulce `jobsatis`. Hodnota maximálního korelačního koeficientu vyšla 0.32, optimální řádkové skóry jsou

```
   <5  5-15 15-25   >25
-1.05 -0.83  1.21  0.93
```

a optimální sloupcové skóry jsou

```
   VD    LS    MS    VS
-2.97 -1.17 -0.08  1.46
```

Povšimněme si, že optimální sloupcové skóry nejsou monotónní. Poslední skór 0.93 je menší než předposlední skór 1.21. V souvislosti s tímto příkladem se nabízejí následující náměty.

- Vypočítat takové optimální skóry, u nichž je dodržen požadavek monotónie. Není mi známo, že by na to existovala nějaká teoretická metoda nebo odpovídající program.
- Může se stát, že některé skóry (např. řádkové) jsou známy přesně (jako třeba počet dětí v rodině) a je třeba vypočítat optimální skóry druhé proměnné (např. sloupcové skóry). Ani zde mi není známo řešení.
- Při zběžném hledání jsem nenašel kritické hodnoty pro test nezávislosti založený na maximální hodnotě korelačního koeficientu.

Optimální skóry získané v rámci kanonické korelace se však obtížně porovnávají se skóry původními. Protože se korelační koeficient při lineární transformaci nemění, budeme hledat takové lineární transformace optimálních skórů, aby se získaly skóry co nejlépe aproximující skóry původní. Metodou nejmenších čtverců se získaly tyto výsledky. Optimální řádkové skóry nejbližší původně zadaným skórům jsou 5.56, 7.53, 25.99, 23.42, v případě sloupcových skórů pak 0.93, 2.17, 2.92, 3.98. Zejména u optimálních sloupcových skórů je překvapující jejich velká shoda s původními poměrně libovolně zvolenými skóry 1, 2, 3, 4.

2. Jací jsou noví studenti

Již řadu let koná Matematicko-fyzikální fakulta v září soustředění posluchačů, kteří nastupují do 1. ročníku. Soustředění se koná na Albeři v okrese Jindřichův Hradec. Posluchači jsou rozděleni podle oborů, které hodlají studovat. Jsou to matematika (M), fyzika (F), informatika (I) a učitelství (U).

Od roku 2004 tam posluchači píší tentýž test z matematiky. Test má 12 úloh. Za každou správně vypočítanou úlohu dostává posluchač jeden bod. Úspěšně napsal test ten posluchač, který získal alespoň 9 bodů. Má-li jen 8 bodů nebo méně, byl neúspěšný. Neúspěšným řešitelům je doporučeno, aby se účastnili dvou až třídenního kurzu středoškolské matematiky, který se koná koncem září v Praze. Účast na tomto kurzu není pro nikoho povinná a mohou se ho účastnit i ti, kteří byli v testu úspěšní. Právě proto, že je zadáván stále stejný test, jsou jeho výsledky v jednotlivých létech plně srovnatelné. Vzhledem ke konkrétním podmínkám přijímacího řízení lze téměř s jistotou usoudit, že studenti nemohou úlohy zadané v textu předávat svým kolegům nastupujícím o rok později.

Podíl úspěšných studentů v jednotlivých létech je znázorněn na obr. 1. Na obrázku není uveden graf vztahující se ke studentům učitelství, protože jejich počet byl velmi malý.

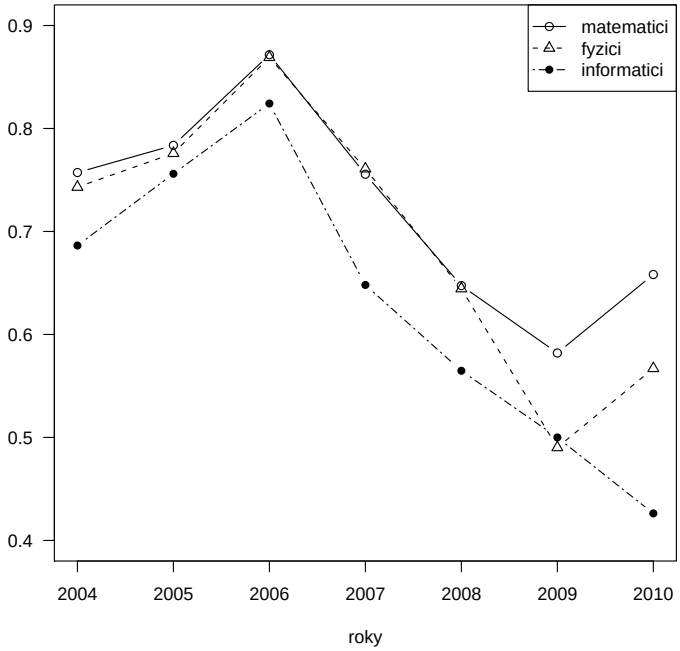
Vidíme, že nejlepší výsledky byly v roce 2006. Od té doby nastal pokles podílu úspěšných studentů. Až teprve letos, tedy v roce 2010, úroveň studentů matematiky a fyziky zaznamenala obrat. Matematická kvalita nastupujících studentů informatiky klesá lineárně dál jako po nakloněné rovině.

Jiným měřítkem kvality nastupujících studentů může být procento těch, kteří se po roce pobytu na fakultě dostanou do druhého ročníku. Graf najdeme na obr. 2. Na vodorovné ose je vynesen rok nástupu studentů na fakultu, na svislé ose je procento těch, kteří se pak po roce zapsali do druhého ročníku. Z tohoto hlediska si nejúspěšněji vedli studenti, kteří se zapisovali do prvního ročníku v roce 2006. To může být důsledkem jejich kvality, kterou jsme zaznamenali při hodnocení grafu na obr. 1. Na druhé straně se však v roce 2006 dost podstatně změnily studijní předpisy, takže nebyly tak tvrdé v prvním ročníku jako dříve. Ale vzhledem k tomu, že všechny tři grafy na obr. 1 od roku 2006 klesají, vypadá to spíše na důsledek horší kvality nastupujících posluchačů.

3. Jak studenti vidí své učitele

Zde bych se zmínil o některých výsledcích ankety, kterou připravil prof. PhDr. Jiří Mareš, CSc., pro všech 17 fakult Univerzity Karlovy. (Nakonec se toto šetření konalo pouze na 15 z nich.)

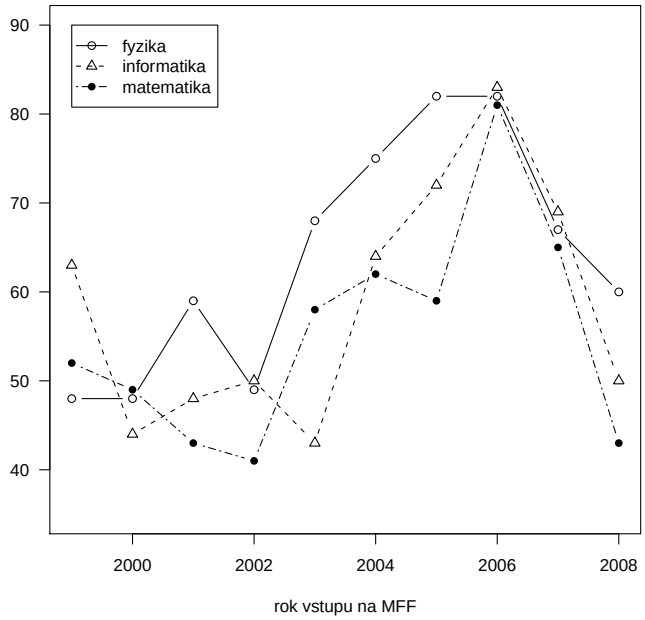
Anketa se týkala studentů prezenční formy studia. Na MFF proběhla v prosinci 2009 a v lednu 2010. Byl zadáván jednotný dotazník, opticky čitelný; studenti odpovídali začerňováním políček a časová náročnost celkové administrace dotazníku byla do 30 minut. Vedení UK rozhodlo, že na fakultách, kde dominuje strukturované studium, tj. rozdělené na část baka-



Obrázek 1: Podíl úspěšných studentů v testu z matematiky na Albeři

lářskou a navazující magisterskou, se šetření provede u bakalářského stupně. To byl i případ MFF. Šetření mělo retrospektivní charakter, tj. studenti posuzovali předměty/kurzy, které absolvovali v předchozím akademickém roce 2008/2009. Úkolem garanta na fakultě bylo vybrat pro šetření „reprezentativní“ studijní obory fakulty, které splňují následující podmínky:

a) jsou typické pro obsahové zaměření dané fakulty
b) počet studentů v jednom ročníku je „dostatečně veliký“ (tj. více než 30 osob)
c) jde o to vybrat „velké“ obory fakulty tak, aby celkový počet studentů v nich studujících tvořil dohromady alespoň polovinu všech studentů fakulty



Obrázek 2: Procento studentů, kteří se po roce zapsali do 2. ročníku

Pro posouzení kvality výuky garant vybral pro každý „reprezentativní“ obor tři předměty/kurzy, které jsou klíčové pro úspěšné absolvování daného studijního oboru, nebo jsou to předměty/kurzy blízké (či totožné) s těmi obory, které se zkoušejí u státnic. Podle těchto pokynů bylo na MFF vybráno 14 předmětů, jeden z nich však byl hodnocen ve dvou různých studijních oborech. U naprosté většiny otázek měl respondent možnost vybrat jednu z následujících odpovědí:

1. souhlasím
2. spíše souhlasím
3. těžko rozhodnout
4. spíše nesouhlasím
5. nesouhlasím
6. nelze posoudit

Pouze u otázek na počet hodin týdně domácí přípravy a na cenu za studijní materiály v tisících pro tento předmět byly nabízeny počty od nuly až po více než devět (a také možnost „nelze posoudit“). V případě celkového hodnocení, kdy měl respondent porovnat daný předmět s ostatními, byly k disposici tyto možnosti:

1. předmět byl výrazně lepší než většina ostatních
2. předmět byl o trochu lepší než ostatní
3. předmět patřil mezi průměrné
4. předmět byl o trochu horší než ostatní
5. předmět byl výrazně horší než většina ostatních

Někteří respondenti na některé otázky neodpověděli, nebo zvolili odpověď „nelze posoudit“. Pro účely našeho zpracování dat vztahujících se k MFF byly všechny takové případy klasifikovány jako chybějící pozorování. Podrobná anonymní data získaná od respondentů z MFF dodali organizátoři šetření ve formě tabulky Excelu. Další zpracování těchto údajů proběhlo na MFF v programu R. Vedení MFF o výsledky šetření projevilo velký zájem. V první řadě šlo o porovnání jednotlivých předmětů. Proto byly u odpovědí na jednotlivé otázky vypočteny aritmetické průměry.

Celkem bylo položeno 7 otázek vztahujících se ke koncepci a organizaci předmětu, 5 otázek se týkalo zkoušení a hodnocení, 10 otázek bylo zaměřeno na studijní materiály a na průběh výuky, 8 otázek mělo posoudit přínos předmětu a nakonec jedna otázka byla na celkové hodnocení.

Mezi posuzovanými předměty byl i předmět nazvaný Pravděpodobnost a statistika. Je to jednosemestrální předmět o rozsahu 4/2, který je povinný pro druhý ročník bakalářského oboru Finanční matematika. Budeme ho stručně označovat PS.

V celkovém hodnocení se nejlépe umístil předmět Teoretická mechanika. Asi je překvapující, že hned na druhém místě byl předmět Matematická analýza 2B. Náš předmět PS byl až na předposledním místě.

Z hlediska obtížnosti se studentům jevil jako nejlehčí opět předmět Teoretická mechanika. Předmět PS obsadil poslední místo.

Studenti také hodnotili, zda zkoušení bylo spravedlivé. Jako nejspravedlivější se jim jevilo zkoušení předmětu Teoretická mechanika. Předmět PS se znovu ocitl na posledním místě. Studenti tedy mají největší výhrady ke zkoušení pravděpodobnosti a statistiky.

Toto kritické hodnocení předmětu PS ale není způsobeno osobou učitele, protože ten se v jiných předmětech a v jiných anketách umisťuje pravidelně na jednom z předních míst.

Můžeme se také zajímat o to, na čem nejvíce závisí celková spokojenost respondentů. Na to byla použita regresní analýza. V literatuře se doporučuje provádět volbu optimálních regresorů interaktivně, aby se získal smysluplný model (viz Venables, Ripley 2002). Tímto způsobem vyšlo, že posluchači nejvíce oceňují (bez pořadí důležitosti)

- jasnou organizaci předmětu
- spravedlnost při zkoušení
- zkoušení, které je zaměřeno na uvažování, nikoli na faktografii

Program `R` také umožňuje automatické vyhledávání optimálních regresorů. Podle kritéria BIC vhodný počet těchto regresorů vyšel roven 3. Tyto regresory jsou (zde v pořadí důležitosti)

- jasná organizace předmětu
- předmět naučil vyjadřovat se odborně
- studenti byli vybízeni k diskusi

Statistická analýza byla samozřejmě mnohem podrobnější. Zahrnovala mj. korelační analýzu a analýzu hlavních komponent. To však v tomto příspěvku už nebudeme rozebírat.

Závěr tedy není příliš radostný. My učitelé nejsme spokojeni s klesající úrovní znalostí posluchačů a posluchačům se nelíbí naše metody výuky pravděpodobnosti a statistiky.

Literatura

[1] Agresti A. (2002): Categorical Data Analysis. 2nd Ed. Wiley, Hoboken, New Jersey.

[2] Mantel N. (1963): Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Amer. Statist. Assoc.* **58**, 690–700.

[3] Presnell B. (2000): An Introduction to Categorical Data Analysis Using R. PDF File.

[4] Thompson L. A. (2007): S-PLUS (and R) Manual to Accompany Agresti's *Categorical Data Analysis* (2002) 2nd edition.

[5] Venables W. N., Ripley B.D. (2002): Modern Applied Statistics with S. Fourth Ed., Springer, New York.

Poděkování. Práce byla podpořena výzkumným záměrem MSM 0021620839 *Metody moderní matematiky a jejich aplikace.*

QUO VADIS, VÝPOČETNÍ STATISTIKO?

Jaromír Antoch

Adresa: MFF UK, KPMS, Sokolovská 83, 186 00, Praha 8

E-mail: jaromir.antoch@mff.cuni.cz

Moto: Pokrok ve výpočetní statistice je podstatnou měrou dán pokrokem v obecném počítání, vývojem výpočetní techniky a pokrokem moderní matematické statistiky.

Abstrakt

Příspěvku je volnou úvahou na téma *Kam směřuje výpočetní statistika.* Autor se stručně soutřeďuje na vybrané aspekty daného problému, především na možností současných počítačů, aplikace, simulace, projekt R a fenomén CUDA.

The paper is an essay on selected directions of the computational statistics as author see the problem. Main attention concerns the influence of current computers, applications, simulations, project R and phenomenon of CUDA.

Nedávná výzva předsedy naší společnosti přispět do tohoto tématického čísla, spolu se vzpomínkou na nedávno absolvované symposium COMPSTAT 2010 a podobné konference či worskshopy, jakož i pravidelné čtení (nebo alespoň listování) časopisy typu Computational Statistics and Data Analysis či Journal of Computational Statistics mne přiměly se zamyslet nad tím, kam kráčí výpočetní statistika. Přes veškeré přemýšlení musím již úvodem říci, že odpověď mi příliš jasná není, bohužel.

4. Počítače kolem nás

Zdá se mi, že počítač výkonný jako CRAY I má teď pod stolem pomalu každá sekretářka. Většina uživatelů to však nebere jako výzvu k přemýšlení o tom, co vše by se s takovým nástrojem dalo spočítat, nýbrž často spíše jako prostředek hrubé síly, kterým lze problém „snáze zlomit“. Vidím to nejen na svých studentech a sobě samém[1], ale i na leckterém sofware. Studenti mi nevěří, co vše se dalo v osmdesátých letech například programem *RegAn* kolegy Karla Zváry na počítači ADT 4100[2] spočítat. Je to možná marná

[1] Proč bych namáhal hlavu, říkám si někdy, vždyť to za ni stroj do rána spočítá.

[2] Tento úspěšný československý výrobek, navržený ve VÚMS a vyráběný od roku 1972 v ZPA Trutnov, měl 32 kiloslov (šestnáctibitových) operační paměti, feritovou paměť, ovládání pomocí děrné pásky . . . Díky feritům si skutečně, na rozdíl ode mne, všechno pamatoval. Tak například ráno měl ve své feritové hlavě přesně to, s čím večer po vypnutí proudu „usínal“. O mé hlavě se to již říci vždycky nedá.

pýcha, ale omezená operační paměť a zanedbatelná rychlost procesoru člověka pořádně přiměly přemýšlet a algoritmizovat. A to se nezmiňuji o příkladech dobře známých z literatury, které popisují „výkony“ pánů Fishera či Yatese na kalkulátorech a počítačích jejich doby.

Jsem přesvědčen, že jsme nyní skoro všude obklopeni skutečně velmi výkonnými počítači. Možná až příliš výkonnými z pohledu toho, k čemu je většina z nás používá (web, mail, psaní, . . . , občas i možná trocha počítání). Ale ať jsem již měl v životě přístup k jakémukoliv počítači, vždy mne překvapovalo, a to nemile, že za „velkou vědu“ a „velké počítání“ se mnohdy vydávalo používání zbytečně složitých, často povaze problému neadekvátních postupů. Kanón na vrabce, říkával Josef Machek. A to místo toho, aby se více cenily chytré myšlenky, které dané úlohy usnadňují, ulehčují vhled do nich, výpočty podstatnou měrou urychlují atd. Situace u nás i ve světě je podobná jako před léty. Koupíte-li si nový MS Word, musíte si koupit také nový počítač, rychlejší a především s větší pamětí. Typ Vašich dopisů a zpráv se přitom moc nemění znemění. Prokážete-li a zdůvodníte velkou spotřebu strojového času, samozřejmě pro rozumný úkol, máte šanci získat podporu na nákup nového klastru počítačů. Přijdete-li s myšlenkou na podstatné urychlení těchže výpočtů, například chytrým logickým nebo matematickým postupem či obratem, „řežete pod sebou větev“. Zpravidla nedostanete nic a pouze se dozvíte, že to je přeci jasné, to by napadlo každého, případně to, že taková „drobnost“ za publikaci ani nestojí, protože . . . Nevěříte? Zkuste si to!

5. Aplikace

Aplikace bezesporu tvoří jeden z pilířů výpočetní statistiky, neboť ji každodenně zásobují novými a novými problémy – více či méně zajímavými. Poznamenejme nicméně, že většina takovýchto prací je zajímavých a užitečných spíše jako aplikace samotná, méně již pro rozvoj výpočetní statistiky, neboť se většinou jedná o více či méně rutinní použití již dříve vyvinutých postupů a algoritmů. Toto však není nutně pravidlem, neboť řada praktických úloh s sebou přináší pro výpočetní statistiku skutečné výzvy.

Osobně mezi nejdůležitější a nejzajímavější řadím především úlohy spojené s analýzou (velmi) rozsáhlých dat, která mimo jiné produkují (řazeno abecedně a ne podle důležitosti): bio-vědy, ekonomika, finance, fyzika, chemie, medicína, monitorování životního prostředí, průmysl a mnohá další odvětví. Dle mne sem například patří:

- genetika – analýza DNA, „průmyslové“ šlechtění/křížení zvířat, mapování procesů v mozku, . . . ;

- informatika – například údaje o tocích dat v sítích a intenzitě provozu;
- exponenicielně rostoucí množství informací ukládaných na web a vyhledávání informací o nich a z nich;
- meteorologie a modelování počasí;
- monitorování životního prostředí;
- velké fyzikální experimenty na urychlovačích spojené s detekcí základních částic hmoty;
- informace o chování zákazníků velkých řetězců či telekomunikačních firem, uložené v datových skladech;
- data pocházející z chemických a fyzikálních analýz komplikovaných sloučenin;
- rozsáhlé databáze o poruchách výrobků či škodních průbězích v pojišťovnách;
- časo-prostorová data a jejich apliace, ať již jsou produkována pravidelným snímkování povrchu země, epidemiologií atp.

Je třeba mít nicméně na paměti, že výpočetní statistika je zde všude pouze jedním z nástrojů. Nástrojem, který se mnohdy díky zdárnému vyřešení těchto problémů daří podstatně zkvalitnit.

6. Simulace

Simulace bezesporu tvoří další z pilířů výpočetní statistiky – pro mnohé dokonce ten hlavní. Autorovým přesvědčením však je, že skutečně efektivní (nejen statistické) simulační postupy a algoritmy jsou především výsledkem podrobné analýzy problému, dobrého nápadu a chytrého logického a/nebo matematického postupu. Výpočetní statistika jako taková se na nich především učí. Pochybuji, že je jejich motorem.

7. Projekt R

Projekt R nejenom stále žije a – ať již řízeně nebo ne(z)řízeně – kyne. Každý měsíc se objevují nové a nové balíčky, vylepšení, tu a tam i nová verze. O workshopech nemluvě. Vše přitom spojuje jediné – R. Něčeho podobného jsme byli svědky snad jenom u SASu, kde se každoročně na pravidelných výročních konferencích uživatelů scházely tisíce lidí, které vůbec nespojoval obsah toho, co dělají, nýbrž nástroj, kterým počítají. DŘÍVE SAS – NYNÍ R, mohlo by znít heslo dne. I když jak kde, v akademickém světě asi spíše než ve sféře byznysu.

Kvalita některých balíčků v R je vynikající, autoři je stále vylepšují, opravují a upravují. Jdou s dobou, dalo by se říci. O některých balíčcích je lepší na druhé straně raději nemluvit. Ale tak již tomu v životě bývá, a to nejenom ve světě nekomerčním. Celkově lze pouze doufat, že CORE team R-ka udrží tempo s vývojem operačních systémů, procesorů, překladačů, paralelizací výpočtů . . . , a neskončí na smetišti dějin, jako se stalo s řadou jeho předchůdců.

Z pohledu výuky je třeba zdůraznit, že R-ko a nadstavby nad ním se v akademickém světě spolu s Excelem prosazují jako základní programová podpora pro analýzu dat. Ve světě byznysu si tím již tak jistý nejsem. Soudím, že zde to bude Excel se SASem.

8. Fenomén CUDA

CUDA[3] je pro většinu statistiků naprosto neznámý pojem. Přesto se jedná o trend, který v poslední době začíná podstatně měnit pohled na „náročné“ počítání a v některých oblastech jej brzy zásadním způsobem podstatně změní[4]. A to všude tam, kde je třeba zpracovávat velká data nebo opakovaně provádět výpočty téhož typu, jak je tomu při simulacích, bootstrapování, analýze obrazu apod.

Není překvapením, že si lidé poměrně brzy uvědomili, že v každém počítači je vedle procesoru (CPU) také grafický procesor (GPU). Již před časem přitom GPU začaly CPU výkonnostně předhánět. A co více, GPU se přes svůj mnohdy obrovský výkon vetšinu času „fláká“, což vedlo k myšlence jej více zaměstnat. Nejprve pro specializované typy výpočtů, jakými je rychlá Fourierova transformace, tj. výpočty, na něž je mimo jiné stavěn a optimalizován, posléze i na výpočty mnohem složitější. Tak například knihovna BLAS (basic linear algebra subprograms) pro CUDA podstatně rozšířila spektrum možných aplikací. Přitom je třeba mít na paměti, že současné grafické procesory jsou v zásadě paralelní vícejádrové počítače, které jsou všudypřítomné.

V nejjednodušším případě možného použití šlo o myšlenku opustit počítání pouze na CPU a přejít na spolu-počítání (co-processing) jak na CPU tak GPU daného počítače. Dalším krokem bylo vytváření CGPU[5] jednotek, které by se k počítači připojily.

CUDA je také výkoný paralelní programovací model, který dovoluje:

- heterogení smíšené seriálně-paralelní programování;

[3] CUDA je zkratka pro *Compute Unified Device Architecture*

[4] Svědčí o tom nejenom cvrlikání brabců, ale též neformální diskuse studentů, kterým občas nediskrétně poslouchám.

[5] Cluster of Graphical Processing Units

- hierarchické počítání na více vláknech procesoru;
- používání řady knihoven, jejichž obsah a rozsah se velmi rychle rozrůstá atd.

Navíc, mnoho jazyků jako C, C++, Fortran či Python, výpočetních systémů jako Mathematica nebo Matlab a operačních systémů využití CUDA podporuje. S operačními systémy nedávno dodanými na trh, jakými jsou MS Windows 7 a Apple Snow Leopard se počítání pomocí GPU stává hlavním proudem, neboť v těchto operačních systémech GPU přestávají být pouhými grafickými procesory, ale stávají se paralelními procesory pro počítání jež jsou přístupné „jakékoliv aplikaci“. Uživatel tak zůstává při vývoji té či oné aplikace ve svém výpočetním prostředí a CUDA používá především/pouze pro specializovanou část výpočtů.

A co více? CUDA je laciná! K velkému překvapení zjistíte, že výpočetní sílu počítače CRAY II[6] můžete mít za několik tisíc dolarů. Pro Vás jako uživatele to znamená dokoupení další komponenty do počítače. Nevěříte? Zadejte do Googlu klíčové slovo CUDA a začněte jednotlivé nabídky studovat a porovnávat.

Na závěr tohoto odstavce mi nicméně dovolte jedno důrazné upozornění. K úspěchu nestačí pouze mít k dispozici příslušný hardware a software. Chce to především umět s obojím také zacházet. A k tomu se nejlépe hodí spolupráce s některým z mladších kolegů, nezatíženým klasickým programováním šedesátých let.

9. Spolupráce s jinými obory

Ukazuje se, že stále méně lidí ve statistickém světě píše své aplikace s pomocí jazyků typu C, C++, Fortran apod. Místo toho se stalo zvykem psát programy raději s použitím některého programového systému jako je Maple, Mathematica, Matlab, R či SAS. V jazyku nižší úrovně jsou, pokud vůbec, napsány pouze ty části programu, které je třeba optimalizovat, například kvůli rychlosti při simulacích, bootstrapování apod. Autoři přitom věří, že takto mají vedle správy paměti, grafiky, normalizovaného vstupu a výstupu zajištěny ty nejlepší, nejrychlejší a nejspolehlivější algoritmy. Není tomu však nutně tak, realita může být daleko od očekávání. Chcete-li příklady, stavte se někdy na diskusi a na skleničku.

Přestože je většině (výpočetních) statistiků dobře známo, že jiné obory, především numerická matematika a informatika, vyvíjejí a zdokonalují po-

[6] Se zájmem jsem si jej byl během COMPSTATu 2010 v Paříži v muzeu pořádající organizace CNAM prohlédnout.

stupy, které by se jim „mohly hodit“, vzájemné spolupráce je jako šafránu. Je to přitom škodě obou stran. Podle mne právě zde existuje velký potenciál jak pro teoretické úvahy a publikace, tak pro mnohdy zásadní vylepšení stávajících výpočetních postupů a algoritmů. Problémy a jejich řešení přicházejí nejenom se statistiky a analýzy dat samotných.

10. Výpočetní statistika a konference

Navštívíte-li libovolnou konferenci, která má slova *výpočetní statistika* v názvu nebo alespoň v podtitulku, například COMPSTAT'xx, uvidíte velké prolínání s mnoha oblastmi, které pravověrní statistici za statistiku ani nepovažují, mnohdy bohužel. Patří sem mimo jiné:

- strojové učení (machine learning);
- strojový překlad;
- matematická lingvistika;
- vyhledávání znalostí a pravidel (knowledge discovery);
- neuronové sítě;
- klasifikace a klastrování;
- grafické zobrazování rosáhlých mnohorozměrných dat atd.

O vše pokrývajícím deštníku známém pod názvem *data mining* nemluvě. Je přitom autorovým přesvědčením, že mnohé z těchto oblastí nabízejí statistice, a to nejenom výpočetní, velmi zajímavé netriviální problémy, které statistika, a matematika vůbec, svými klasickými nástroji není schopna řádně uchopit. Což je nejenom škoda, ale i velká výzva do budoucnosti (nejenom výpočetní) statistiky.

Rozvoj výpočetní statistiky si klade za svůj hlavní cíl *International Association for Statistical Computing*[7], která je jednou se sekcí jedné z nejstaších vědeckých společností na světě, jíž je *International Statistical Institute*[8]. Nejste-li doposud členy těchto společností a o statistiku, nejenom výpočetní, máte zájem, jděte prosím na níže uvedené internetové odkazy, a shledáte-li cíle těchto společností zajímavými, tak se staňte jejich aktivními členy.

Poděkování. Práce byla podpořena výzkumným záměrem MSM 0021620839 a grantem GAČR 201/09/0755. Autor by též rád poděkoval Janu Klaschkovi za veškeré poznámky a doporučení.

[7] `http://www.eco.unicas.it/strutture/iasc/`
[8] `http://isi-web.org/`

PRŮMYSLOVÁ STATISTIKA

Gejza Dohnal

Adresa: ČVUT, CQR, Karlovo nám. 13, 120 00 Praha 2

E-mail: dohnal@nipax.cz

Abstrakt

Vývoj aplikací teorie pravděpodobnosti a matematické statistiky v oblasti průmyslu a zpracování surovin je v posledních dvaceti letech více než kdy předtím poznamenán širokým nasazením počítačové techniky v těchto oblastech. Čím více této techniky řídí a monitoruje výrobní procesy, tím víde dat v tomto prostředí vzniká a je třeba je zpracovat. Na druhou stranu, výpočetní technika nám dovoluje používat v praxi metody, které ještě nedávno byly doménou teoretiků a na kterých klasické analytické metody troskotaly. Z tohoto hlediska má vývoj stochastiky v podsledních letech spíše intenzívní než extenzívní povahu. Většina používaných a rozvíjených metod pochází z dob dávno minulých a ve světle nových algoritmů dostávají další rozměr.

In the last twenty years, applications of probability theory and mathematical statistics are developed under the tumb of widely used computers and information technologies. On the other side, this trend brings more and more of data which have to be processed. The modern computer technology allows to apply a lot of stochastic methods which were a domain of theoreticians only. From this point of view, the development of stochastics has more likely an intensive than extensive character.

1. Úvod

S rozvojem hromadné výroby na přelomu 19. a 20. století se objevily první snahy využít nové matematické poznání i v této oblasti. Napomohla tomu i „pravděpodobnostní revoluce“ ve fyzice, probíhající na počátku 20.století, zejména v kontextu oblastí jako statistická fyzika, kvantová mechanika, teorie chaosu, informační fyzika, atd. Speciálně teorie pravděpodobnosti a matematicko-statistické metody zaznamenaly široký rozmach právě v této době. Počátek a první polovina 20. století přinesly práce Markova, Kolmogorova, Shewharta, Kendalla, Wienera a dalších, které znamenaly počátek aplikace pravděpodobnostních a matematicko-statistických metod v průmyslu, výrobě a obchodu. Rozvíjející se průmyslová odvětví požadovala stále více nové technologické postupy, materiály a metody řízení. Právě zde se uplatnily matematické – a stále více stochastické – modely a přístupy k vyhodnocování průmyslových experimentů a výsledků výroby. Rozvoj kybernetiky,

informatiky a počítačové techniky v 50. letech 20. století rozšířil tyto aplikace o metody Monte Carlo, které znamenaly počátek další, tentokrát „počítačové revoluce“, zasahující všechny obory vědy a lidského snažení. Stochastické simulace nás vnesly do světa virtuální reality, kdy v digitální továrně pracují virtuální dělníci a inženýři na virtuálních strojích a pozorovatel u monitoru počítače měří četnosti a důsledky poruch, časy průchodů jednotlivých součástek výrobní linkou, využití strojů, optimalizuje rozmístění a parametry strojů a provádí predikce chování výrobního systému v situacích, které by v reálném světě znamenaly neúnosné náklady či rizika.

Je však třeba konstatovat, že v oblasti stochastiky se technika, speciálně výpočetní prostředky a software, vyvíjí mnohem rychleji než teorie. Naprostá většina v současné době používaných metod má své kořeny v minulém století (řízení procesů [Gantt], vícekriteriální rozhodování [Pareto], modelování pomocí markovských procesů [Markov], regulační diagramy [Shewhart], lineární programování [Flood], dynamické programování [Bellman], matematická teorie spolehlivosti [Barlow, Proschan], teorie hromadné obsluhy [Kendall], simulační modely, analýza časových řad a další a další). Tyto metody jsou stále zdokonalovány, rozšiřovány a je kladen důraz především na jejich aplikabilitu. V posledních dvaceti letech jsme zaznamenali jejich široké uplatnění a další rozvoj díky všeobecně dostupné výpočetní technice. Poměrně bouřlivý rozvoj obecných statistických programových systémů na přelomu 80. a 90. let (BMDP, SPSS, Statgraphics, NAG, Minitab a další) pokračoval vývojem specializovaných systémů zaměřených na použití ve výrobě (především na statistické řízení procesů – SPC, analýzu časových řad, simulace). Díky tomuto vývoji lze dnes aplikovat a rozvíjet řadu metod, které byly dříve doménou pouze teoretiků a vedly k obtížným teoretickým problémům. To samo o sobě však přináší nové problémy, doposud neřešené.

Masové nasazení výrobních informačních systémů (MES – manufacturing execution systems, ERP – enterprise resource planning) v 90. letech 20. století přineslo řadu dalších teoretických i praktických problémů. Tyto systémy poskytují v reálném čase mnoho dat o výrobě, o výrobních procesech. V souvislosti s nutností zpracování velkého množství informací uložených v těchto datech se objevily pojmy jako *Data mining, Data dredging, Data fishing.* Používané metody spadají spíše do oblasti informatiky, ale i výpočetní a matematické statistiky. „Dolování dat“, jak se tyto metody označují v českém jazyce, zahrnuje řadu metod, známých už z dřívějších let dvacátého století, jako jsou neuronové sítě, genetické algoritmy, klastrování (50. léta), rozhodovací stromy (60. léta) a metody strojového učení (support vector machines) (z let osmdesátých). Nové podmínky přinesly živnou půdu pro široké nasazení těchto metod a jejich další intenzívní rozvoj.

Označení „průmyslová“ nebo též „industriální“ statistika neznamená nějakou speciální oblast stochastiky. Pod tímto pojmem dnes chápeme především aplikace metod matematické statistiky, aplikované pravděpodobnosti a stochastické simulace ve výrobě a zpracování surovin. Teoretický rozvoj těchto metod probíhá zpravidla v obecné rovině, s dopadem i do ostatních odvětví, jako je ekonomie, biologie, životní prostředí či zdravotnictví. Na druhou stranu, každá aplikace přináší nové otázky, jejichž řešení posouvá vývoj kupředu.

V případě industriální stochastiky jsou aplikace zaměřeny především do:
- průmyslové výroby,
- materiálového inženýrství,
- výroby energie,
- dopravy,
- komunikací a informatiky,
- stavebnictví.

Z hlediska metodologického, jsou to potom především:
- statistické metody pro řízení kvality
- stochastické modelování spolehlivosti, bezpečnosti a vyhodnocení rizika
- vyhodnocení experimentů, analýza stochastických procesů a časových řad, optimalizace a predikce vývoje
- simulační metody, rozvoj systémů umělé inteligence.

2. Řízení kvality

Tato oblast je poměrně široká. Ze statistického hlediska zahrnuje především měření kvalitativních znaků, jejich odhady, vyhodnocení, klasifikaci a identifikaci změn a závislostí. Sem spadá rozvoj takových metod, jako jsou neparametrické metody odhadu, regrese, analýza náhodných procesů a časových řad, klasifikační metody, rozpoznávání a zpracování obrazu a to vše jak v jednorozměrném, ale stále častěji a ve větší míře v mnohorozměrném případě.

Používání matematicko-statistických metod pro řízení kvality zaznamenalo poměrně výraznou vlnu zájmu s nástupem takzvané „Six Sigma“ metodologie. Počáteční impuls spadá už do roku 1986, kdy ve firmě Motorola zavedli „Vyhodnocování kvality na základě variability procesů a měření směrodatné odchylky“, nicméně ucelená metodologie Six Sigma byla zavedena až v roce 1995 Jackem Welschem v koncernu General Electric.

Statistické řízení procesů (SPC) zaznamenalo rozvoj nových metod pro kontrolu kvality, vícerozměrné diagramy, diagramy pro nenormální rozdělení sledovaných znaků, závislá pozorování a bayezovské a adaptivní regulační diagramy. Velká pozornost je věnována metodám optimalizace SPC.

Na druhou stranu, některé metody, jako je například *statistická přejímka*, ztrácejí téměř po sto letech svůj význam. To platí především pro velké výrobce a jejich dodavatele v automobilovém a elektronickém průmyslu, kde díky intenzivnímu sledování a udržování kvality při samotné výrobě je pravděpodobnost výstupu zmetků stlačena téměř na nulu.

Poměrně velký rozvoj lze pozorovat v oblasti *měření kvality*. Zde – opět v souvislosti s dostupnou výkonnou výpočetní technikou – jsou rozvíjeny metody analýzy obrazu a stereologie [4], [5]. To se týká především oblasti materiálového inženýrství, zkoumání kvalitativních znaků dvou a třírozměrného materiálu, hodnocení vad a kontroly kvality textilním, metalurgickém a stavebním průmyslu.

V souvislosti s měřením se stále častěji začal objevovat pojem *nejistoty měření*. Tento pojem byl navržen už v osmdesátých letech, nicméně v roze 1990 vydal Western Electricity Coordinating Council (WECC) dokument, který tento požadavek pokryl metodologicky a vyústil v normu ISO/IEC *Guide to the Expression of Uncertainty in Measurement (GUM)* [19], [20]. Z hlediska statistiky tento přístup vyvolal řadu diskuzí a protichůdných názorů, nicméně se vzrůstajícími nároky na přesnost měření je potřeba kvantifikovat chyby pomocí odhadů variability prostředí stále aktuálnější.

3. Spolehlivost, riziko

Spolehlivost – tedy v kontextu tohoto příspěvku spíše bezporuchovost – a bezpečnost provozu jsou stále více předmětem zájmu jak výrobců, tak i spotřebitelů. Zatímco celé dvacáté století bylo ve znamení zvyšování výkonu výroby a výrobků, implementace nových funkcí, redukce rozměrů, hmotnosti a spotřeby energií, ve století jednadvacátém je stále větší důraz kladen na spolehlivost, bezpečnost a snižování rizik, spojených s výrobou či používáním jejích výrobků.

S rostoucími objemy výroby a transportu nebezpečných látek se často až neúměrně zvyšují rizika, spojená s haváriemi a s používáním nebezpečných materiálů a výrobků. Zde se otevírá velký prostor pro stochastickou analýzu při modelování a optimalizaci výrobních procesů, testování dopadů na životní prostředí a zdraví lidí.

Také v této oblasti docházelo v posledních dvaceti letech spíše k intenzívnímu než k extenzívnímu rozvoji stochastických metod. Dnes již klasické modely založené na použití markovských procesů stále častěji pracují s nehomogenními procesy, semi-markovskými procesy a takzvanými *po částech deterministickými* markovskými procesy.

To souvisí s nutností modelovat bezporuchovost systémů v závislosti na v čase se měnících podmínkách (dynamic reliability) [40], modelovat systémy s vysokou spolehlivostí (ale o to s vyšším rizikem při poruše), modelování a vyhodnocování komplexních systémů, kde nevystačíme s paralelně-sériovými modely z padesátých let a u kterých rozlišujeme řadu poruchových stavů současně, modelování a vyhodnocení spolehlivosti výrobních sítí.

Současné metody umožňují i zcela nové přístupy k údržbovým strategiím (Risk-based maintenance, On–state maintenance). Konkrétně bych v této souvislosti zmínil tři oblasti, které považuji v posledních dvaceti letech za významné. První z nich je oblast spolehlivosti software. Tato oblast se začala rozvíjet už ve druhé polovině dvacátého století, ze kdy pochází i řada matematických modelů. Nicméně, v posledních dvaceti letech s masívním nasazováním výpočetní techniky ve výrobě se tyto metody intenzívně studují a rozvíjejí. Software je dnes při výrobě stejně důležitým nástrojem jako ruční nářadí či obráběcí stroje. Software řídí nejen jednotlivé stroje, ale i celé výrobní linky a provozy. Modely pro odhalování a redukci chyb v programových modulech se zásadně liší od klasických modelů vzniku poruch u strojů a zařízení. To vyžaduje jiný přístup a metody.

Druhou oblastí, která také není nová, nicméně je stále intenzívně studována a rozvíjena, je oblast spolehlivosti lidského faktoru. Tato oblast je o to složitější, že zde nelze uplatňovat řadu předpokladů, používaných při modelování výrobních systémů (například markovskou vlastnost, normalitu, nezávislost). Vedle stále spolehlivějších a bezpečnějších výrobních zařízení je lidský faktor největším zdrojem nespolehlivosti a tedy i největším zdrojem rizika při výrobě [50].

Třetí oblastí je spolehlivost zabezpečovacích systémů [48], [49]. Řada současných výrobních systémů je vysoce spolehlivá, nicméně jejich porucha může mít nedozírné následky. To se týká především výroby energií (elektrárenské bloky), transportu (železniční a letecká doprava), ale i zpracování nerostných surovin (těžba ropy a zemního plynu, ropné rafinérie). Předvídání poruchy v takovýchto provozech zabezpečují složité bezpečnostní systémy, většinou založené na elektronických a hydraulických komponentách. Nižší spolehlivost kontrolovaného zařízení může být vykompenzována vysokou spolehlivostí zabezpečovacího systému, který včas signalizuje nebezpečný stav tak, aby mohl být bezpečně odstraněn a obnoven bezpečný provoz celku. Snahy o vyhodnocení a klasifikaci zabezpečovacích systémů z hlediska spolehlivosti a bezpečnosti vyústily v roce 1998 ve vydání norem, zavádějících takzvanou *System Integrity Level* (SIL) [46], [47]. Výpočty střední pravděpodobností poruchy zabezpečovacího systému na vyžádání vyžadují aplikaci matematických modelů, zmíněných ve druhém odstavci této kapitoly.

4. Vyhodnocení experimentů a optimalizace

Tato „klasická“ oblast použití stochastických metod se v případě průmyslových aplikací v zásadě neliší od ostatních oborů, jako je přírodověda, zdravotnictví, ekonomika či demografie. Zde hrají hlavní úlohu metody bayezovské a robustní stochastiky. Studují se nová rozšíření většinou již existujících metod, posouvající jejich použití na stále širší třídu problémů na straně jedné a používající stále komlikovanější a sofistikovanější přístupy na straně druhé.

5. Simulace, systémy AI

Moderní aplikovanou statistiku si už bez metod Monte Carlo snad ani nedovedeme představit. Použití syntetických výběrů (bootstrap) pro odhady parametrů, optimalizace, simulace nejrůznějších rozdělení pravděpodobnosti, simulace náhodného chování a řada dalších simulačních metod jsou dnes běžnými nástroji tam, kde klasický analytický přístup je příliš složitý, ne-li (zatím) nemožný. Počítačová simulace se ukazuje jako velice efektivní nástroj především tam, kde klasický experimet je příliš drahý, časově náročný a často by znamenal narušení výroby a neúnosné ztráty. V řadě případů lze simulovat na počítači situace, které by v reálném životě znamenaly velké riziko a případně i ohrožení zdraví a životů. Takovéto a podobné problémy řeší koncept „digitální továrny“, vyvinutý v posledních dvaceti letech.

Na závěr uvedu opět alespň tři metody, kterým byla v posledních dvaceti letech věnována poměrně velká pozornost. Jednou z nich jsou bezesporu algoritmy využívající vlastnosti markovských řetězců pro generování pseudonáhodných čísel z různých rozdělení [8], známé pod názvem Monte Carlo Markov Chain (MCMC) jsou využívány pro takzvané pravděpodobnostní usuzování (probabilistic inference) využívaném při strojovém učení [11], [12].

V posledních dvaceti letech je stále větší pozornost věnována oblasti, která se označuje jako „strojové učení“ (machine learning) [6]. Ta zahrnuje dnes již klasické neuronové sítě [1], rozhodovací stromy a genetické algoritmy [7] spolu s novějšími metodami jako jsou bayesovské sítě a algoritmy podpůrných vektorů (support vector machines). Metody strojového učení přispívají k umělé inteligenci moderních strojů a zařízení, automatickému rozpoznávání obrazu a zabezpečovacím systémům [2] [3].

Jednou s užitečných metod pro simulaci v oblasti výrobních systémů jsou Petriho sítě [13], [14]. Metoda vyvinutá v roce 1939 tehdy třináctiletým Carl Adamem Petri byla v letech osmdesátých doplněna o stochastickou variantu a v současnosti je jedním z nejúčinějších nástrojů při modelování dynamických spolehlivostních modelů [15].

Zatímco analytické nástroje v důsleku vedou zpravidla k lineárním aproximacím nelineárních modelů, tyto a mnoho dalších metod umožňují přímé aplikace nelineárních modelů v praxi [16]. Ačkoli se většinou jedná především o numerické algoritmy, stochastika v těchto modelech hraje svoji důležitou roli.

Literatura

[1] Arbib (Ed.) (1995): *The Handbook of Brain Theory and Neural Networks.*

[2] Egmont-Petersen, de Ridder, Handels (2002): *Image processing with neural networks – a review.* Pattern Recognition

[3] Spooner, di Maggiore, Ordonez, Passino (2002): *Stable Adaptive Control and Estimation for Nonlinear Systems: Neural and Fuzzy Approximator Techniques*, John Wiley and Sons, NY

[4] Sonka, Hlaváč, Boyle (1999): *Image Processing, Analysis, and Machine Vision.* PWS Publishing.

[5] Theodoridis, Koutroumbas (2009): *Pattern Recogntition.* 4th Edition, Academic Press.

[6] Mitchell (1997): *Machine Learning.* New York: McGraw-Hill.

[7] Shu-Heng Chen et al. (2008): *Genetic Programming: An Emerging Engineering Tool*, International Journal of Knowledge–based Intelligent Engineering System.

[8] Gelfand, Smith (1990): *Sampling-Based Approaches to Calculating Marginal Densities.*

[9] Casella, George (1992): *Explaining the Gibbs sampler.*

[10] Gelman, Carlin, Stern, Rubin (1995): *Bayesian Data Analysis.* London: Chapman and Hall.

[11] Andrieu et al. (2003): *An Introduction to MCMC for Machine Learning*,

[12] Berg (2004): *Markov Chain Monte Carlo Simulations and Their Statistical Analysis.* Singapore, World Scientific.

[13] Desrochers, Al-Jaar (1995): *Applications of Petri nets in manufacturing systems*, IEEE Press.

[14] Zuravski, Zhou (1994): *Petri nets and industrial applications: A tutorial.*

[15] Bobbio, Codetta–Raiteri (2006): *Stochastic Petri Nets Supporting Dynamic Reliability Evaluation.* International Journal of Materials and Structural Reliability.

[16] Steeb (2008): *The Nonlinear Workbook: Chaos, Fractals, Neural Networks, Genetic Algorithms, Gene Expression Programming, Support Vector Machine, Wavelets, Hidden Markov Models, Fuzzy Logic with C++, Java and SymbolicC++ Programs*: 4th edition. World Scientific Publishing.

[17] Hendrickson, Lave, Matthews (2005): *Environmental Life Cycle Assessment of Goods and Services: An Input-Output Approach*. Resources for the Future Press.

[18] ISO 14040 (2006): *Environmental management: Life cycle assessment. Principles and framework*, International Organisation for Standardisation (ISO), Geneve.

[19] ISO/IEC Guide 98:1995: *Guide to the Expression of Uncertainty in Measurement (GUM)*.

[20] 2005, Grabe: *Measurement Uncertainties in Science and Technology*. Springer.

[21] 1990, Helland: *PLS regression and statistical models*. Scandivian Journal of Statistics, 17, 97-114.

[22] Kramer (1998): *Chemometric Techniques for Quantitative Analysis*.

[23] Friedman I. (1993): *A Statistical View of Some Chemometrics Regression Tools*, Technometrics, 35(2).

[24] Tenenhaus (1998): *La Regression PLS: Theorie et Pratique*.

[25] Chipman G., McCulloch (1998): *Bayesian CART model search*. JASA.

[26] Cristianini, Shawe-Taylor (2000): *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

[27] Scholkopf, Smola (2002): *Learning with Kernels*. MIT Press, Cambridge.

[28] Haenlein, Kaplan (2004): *A Beginner's Guide to Partial Least Squares Analysis*. Understanding Statistics, 3(4).

[29] Steinwart, Christmann (2008): *Support Vector Machines*. Springer-Verlag, New York.

[30] Jensen, Vedel (1998): *Local Stereology (Advanced Series on Statistical Science and Applied Probability, Vol 5). Singapore: World Scientific.*

[31] Torquato (2002): Random heterogeneous materials. *Springer-Verlag.*

[32] Wang, Ziad, Lee (2001): Data Quality. *Kluwer.*

[33] Liu (1999): Multivariate analysis by data depth: descriptive statistics, graphics and inference, The Annals of Statistics.

[34] Rafalin, Souvaine (2004): *Computational Geometry, Data Depth and Robust Statistics*, Interface 2004, Baltimore.

[35] Angeles, MacKinnon (2005): *Quality Measurement and Assessment Models including Data Provenance to grade Data Sources*. Heriot-Watt University School of Mathematical and Computer Sciences, Edinbirgh.

[36] Becker, Camarinopoulos (2000): *A semi-Markovian model allowing for inhomogenities with respect to process time*. Reliability Engineering & System Safety

[37] Droguett, Moura et all (2008): *A Semi-Markov Model with Bayesian Belief Network Based Human Reliability Modeling for Availability Assessment of Downhole Optical Systems*.

[38] Janssen, Manca (2007): *Semi-Markov Risk Models for Finance, Insurance and Reliability*. N.Y. Springer.

[39] Drougett, Moura (2009): *A faster numerical procedure for solving non-homogenous semi-Markov processes*. Reliability, Risk and Safety: Theory and Applications.

[40] Dufour, Dutuit (2002): *Dynamic Reliability: A new model*. Proceedings of ESREL2002, Lyon, France.

[41] Costa, Dufour (2003): *On the Poisson equation for piecewise-deterministic Markov processes*, SIAM J. Control Optim.

[42] Xie (1991): *Software Reliability Modeling*. World Scientific Publisher, Singapore

[43] Musa (1998): *Software Reliability Engineering*. McGraw-Hill.

[44] Ozekici, Soyer (2003): *Reliability of software with an operational profile*. European Journal of Operational Research

[45] Hu, Xie et all (2004): *Software failure prediction based on a Markov Bayesian network model*, The Journal of Systems Software.

[46] IEC 61508 (1998-2005): *Functional Safety of Electrical/Electronic/Progrsmmable Electronic Safety-Related Systems*. International Electrotechnical Commission.

[47] IAEA Safety Standard Series (2002): *Instrumentation and Control Systems Important to Safety in Nuclear Power Plants*. Safety Guide No NS-G-1.3.

[48] Rausand M. (2004): *HAZOP Hazard and Operability Study*. System Reliability Theory, Wiley (2nd ed)

[49] Hulin, Schulze (2009): *Failure analysis of software for displaying safety-relevant information*. Reliability, Risk and Safety: Theory and Applications. CRC Press.

[50] Baraldi, Conti et all (2009): *A Bayesian network model for dependence assessment in human reliability analysis*. Reliability, Risk and Safety: Theory and Applications. CRC press

HOSPODÁŘSKÁ STATISTIKA Z POHLEDU 20 LET VÝVOJE

Stanislava Hronová, Richard Hindls

Adresa: VŠE, FIS, KSTP, nám. W. Churchila 4, 130 67 Praha 3

E-mail: hronova@vse.cz, hindls@vse.cz

Abstrakt

Hospodářská statistika prošla za posledních dvacet let vývojem nesrovnatelným s ostatními částmi statistiky. Toto konstatování nevyplývá z převratných objevů, které by byly na poli hospodářské statistiky v tomto období zveřejněny, ale ze specifického postavení této disciplíny. Hospodářská statistika popisuje a analyzuje ekonomické jevy a procesy pomocí ukazatelů, jež sama definuje. Z hlediska věcné podstaty ukazatele je hospodářská statistika tedy vždy spojena s ekonomickou teorií. A když jsme v ČR po roce 1989 opustili Marxovu ekonomickou teorii, musela se zásadním způsobem změnit i hospodářská statistika. Změnily se nejen ukazatele, jejich definice a způsoby zjišťování, ale i způsoby publikování. Článek mapuje zásadní změny, ke kterým v posledních 20 let v oblasti hospodářské statistiky v ČR došlo.

Over the past twenty years, economic statistics has undergone unparalleled developments in comparison to other parts of statistics. This statement does not result from groundbreaking discoveries that would have been published in the field of economic statistics in this period, but from the special status of this discipline. Economic statistics describes and analyzes economic phenomena and processes using indicators, which itself defines. In terms of the indicators, the economic statistics is thus always associated with economic theory.

And when the Czech Republic left the Marxist economic theory after 1989, the economic statistics underwent fundamental change. Not only the indicators, their definitions and survey methods changed, but also modes of publishing. The article deals with major changes that have occurred in the field of Czech economic statistics in the last 20 years.

Hospodářská statistika prošla za posledních dvacet let vývojem nesrovnatelným s ostatními částmi statistiky. Toto konstatování nevyplývá z převratných objevů, které by byly na poli hospodářské statistiky v tomto období zveřejněny, ale ze specifického postavení této disciplíny. Hospodářská statistika popisuje a analyzuje ekonomické jevy a procesy pomocí ukazatelů, jež sama definuje. Ukazatele jsou statistickou proměnnou, která umožňuje kvantifikovat jevy a procesy, které se v ekonomice reálně odehrávají, ale které

jsou v té nejobecnější rovině popsané ekonomickou teorií. Jinak řečeno, ekonomická teorie definuje pojmy a vztahy mezi nimi a hospodářská statistika jim pak přiřazuje kvantifikovatelné veličiny – ukazatele a popisuje vztahy mezi nimi.

Z hlediska věcné podstaty ukazatele je hospodářská statistika tedy vždy spojena s ekonomickou teorií. A když jsme po roce 1989 opustili do té doby „vládnoucí“ Marxovu ekonomickou teorii, musela se zásadním způsobem změnit i hospodářská statistika. Změnily se nejen ukazatele, jejich definice a způsoby zjišťování, ale i způsoby publikování. A tato změna byla zásadní. Do té doby byly výsledky národního hospodářství publikované v socialistických zemích z hlediska své věcné definice nesrovnatelné s výsledky publikovanými ve „zbytku světa“. Nesrovnatelnost makroagregátů byla ale jen mezinárodně viditelným aspektem; odlišné byly celé systémy ukazatelů používaných v socialistických a nesocialistických zemích. Abychom pochopili hloubku těchto rozdílů, vraťme se trochu do minulosti.

Po druhé světové válce došlo z pohledu hospodářské statistiky k velmi významné změně – dosud ojedinělé či nepravidelné snahy o vytvoření systémů makroekonomických statistických ukazatelů se přeměnily v nutnost. V návaznosti na hloubku světové hospodářské krize a nutnost poválečné obnovy se v podstatě všechny země vyspělého i rozvojového světa daly na cestu budování statistických systémů, které budou schopny zajistit pravidelné, systematické a věrohodné informace o tom, co se v národním hospodářství děje, resp. dělo. Vybudovat systém makroekonomických statistických informací ale znamená nejdříve sestavit jeho základní ekonomické schéma. Ekonomická teorie totiž nabízí logiku ekonomických pohybů nezbytnou pro sestavení smysluplného systému makroekonomických statistických informací. Bez ekonomické teorie by nešlo o systémy, ale jen soubory informací bez vnitřních vazeb a nezbytné logiky popisu pohybů odehrávajících se uvnitř národního hospodářství. Vlastní podstata těchto systémů je statistická – jsou prezentovány jako systémy ukazatelů, tabulek, schémat a vztahů a definice ukazatelů vycházejí z praktických možností statistiky, tj. z pohledu jejich možného způsobu zjišťování a zpracování. Ekonomická teorie není při prezentaci systémů makroekonomických systémů statistických informací jakoby na první pohled vidět. Je ale skryta v logice stavby celého takového systému. Odlišná ekonomická teorie je tudíž základem odlišnosti makroekonomických statistických systémů ukazatelů.

V období 1945 až 1955 vznikly v jednotlivých zemích národní systémy makroekonomických statistických informací. V zemích s tržní ekonomikou se postupně rozvíjel **systém národních účtů**, založený na Keynesově ekonomické teorii, a v zemích s centrálně plánovanou ekonomikou se zaváděl

systém bilancí národního hospodářství. Oba tyto systémy měly stejný cíl – zobrazit výsledky ekonomické činnosti národního hospodářství za uplynulé období ve formě tabulek, jež mají formu účtů, resp. bilancí (ostatně srovnáme-li názvy – systém národního účetnictví a systém bilanci národního hospodářství z čistě jazykového hlediska, dojdeme k závěru, že jde o synonyma). Cíl byl sice stejný, ale způsob realizace – východiska a předpoklady, zejména pak samotná definice produktivní činnosti, charakter používaných cen atd. – činil data, pocházející z těchto systémů, nesrovnatelnými.

Systém bilancí národního hospodářství se opíral o Marxovu ekonomickou teorii, o tzv. pracovní teorii hodnoty. Podle této teorie byla za produktivní považována jen činnost vedoucí k výrobě statků (průmysl, zemědělství, stavebnictví). Služby byly považovány za produktivní jen tehdy, pokud sloužily pohybu těchto statků, popř. je doplňovaly (nákladní doprava, telekomunikační služby pro podniky, zahraniční obchod). Pojetí sféry produktivní činnosti bylo výrazně užší, než je tomu v národním účetnictví založeném na Keynesově ekonomické teorii. K ocenění vyrobených statků a služeb byly navíc používány ceny, které byly centrálně určovány a řízeny, a neměly tedy charakter cen tvořících se na trhu. Pojetí produktivní činnosti a systém cen byly dva zásadní problémy nekompatibility systému národního účetnictví a systému bilancí národního hospodářství, a tím i zásadní odlišnosti obsahu celé hospodářské statistiky. **Přechod od jednoho systému makroekonomických ukazatelů k jinému, založenému na odlišné ekonomické teorii, představuje zároveň zásadní změnu v obsahu statistických ukazatelů, tj. zásadní změnu hospodářské statistiky**[1].

Ekonomická transformace, která v tehdejším Československu započala v roce 1990, byla hned od počátku spojena i s přechodem k systému národního účetnictví. Federální vláda ČSFR na základě návrhu Federálního statistického úřadu rozhodla zavést počínaje dnem 1. 1. 1992 systém národního účetnictví a nahradit tak po čtyřiceti letech systém bilancí národního hospodářství. To představovalo první krok zásadní změny obsahu hospodářské statistiky, krok k mezinárodní standardizaci a srovnatelnosti statistických postupů a výsledků.

Dva ze základních principů ekonomické transformace – privatizace a liberalizace cen – zásadním způsobem ovlivnily změny postupů v oblasti hospo-

[1] Podíváme-li se do kterékoliv publikace hospodářské statistiky, vypadá to, že mluvíme stále jen o jedné z částí hospodářské statistiky, i když vedle systémů makroekonomických ukazatelů existuje statistika produkce, měření produktivity práce, cenová statistika a další její oblasti. Makroekonomické systémy ukazatelů sice nepokrývají zcela všechny oblasti hospodářské statistiky, ale jsou určující i pro obsah ostatních ukazatelů. Je to logické, neboť statistika nemůže poskytovat různé informace o jednom a témže jevu.

dářské statistiky. Privatizace vyvolala (již v roce 1991) vznik velkého množství malých ekonomických subjektů, rozpad a transformaci velkých podniků, a tím i problémy s aktuálností a následnými aktualizacemi registru ekonomických subjektů. Vysoká inflace znamenala tlak na spolehlivé a rychlé informace o cenovém vývoji. Přeměna malého množství velkých ekonomických subjektů na velké množství menších znamenala i zcela jiný přístup ke zjišťování, ke stylu statistické práce, k použitým metodám. Rychlost ekonomických změn vyžadovala větší podíl krátkodobých informací. Ve všech oblastech statistiky to znamenalo přechod na výběrová zjišťování, na větší podíl modelování na úkor zjišťování, na nové techniky sběru a zpracování dat.

Základem budování nového statistického systému se stal standard národního účetnictví Evropské unie z roku 1978 (ESA 1978[2]). Vzhledem k tomu, že přechod k systému národního účetnictví vyžadoval rozsáhlé změny ve všech oblastech činnosti státní statistiky, bylo rozhodnuto rozložit tento přechod do víceletého období, během něhož by souběžně existovaly oba systémy, resp. základní národohospodářské agregáty by byly počítány podle obou metodik a publikovány souběžně.

První odhady hrubého domácího produktu a složek jeho tvorby a užití za rok 1980 a 1985-1988 vycházely z agregátů bilancí národního hospodářství a byly publikovány ve Statistické ročence 1990, dopočty za rok 1989 a 1990 pak ve Statistické ročence 1991. Za léta 1990-1991 však nebyly sestaveny ani bilance národního hospodářství ani národní účty. Za toto období byly jen odhadnuty základní makroagregáty. První přímý odhad hrubého domácího produktu (tj. odhad na základě metodiky národního účetnictví) byl proveden za rok 1992. Národní účty za rok 1992 však byly publikovány až v roce 1995 a vzhledem k rozdělení ČSFR jen za Českou republiku. Národní účty za ČSFR tedy nebyly nikdy oficiálně sestaveny.

Národní účty za ČSFR nicméně existují, alespoň v podobě tzv. experimentální soustavy sektorových účtů (za 80. léta a roky 1990 a 1991), která byla výsledkem práce skupiny odborníků na Institutu ekonomie ČNB (viz NACHTIGAL, 1993). Práce, která vyjadřovala snahu výzkumných pracovníků vyřešit některé metodické problémy a sestavit alespoň provizorní účty ČSFR, byla bohužel prací na účtech země, která v okamžiku oponentury práce již vlastně neexistovala. Tato experimentální sestava byla prvním a zřejmě i posledním pokusem o sestavení úplné posloupnosti nefinančních účtů institucionálních sektorů a národního hospodářství ČSFR.

První národní účty měly být sestaveny ještě za ČSFR za rok 1992 podle standardu ESA 1978. Rozdělení Československa pak vedlo nejen ke zpomalení

[2]European System of Integrated Economic Accounts, 1979.

prací na národních účtech, ale i k jejich prostorovému omezení jen na území České republiky. Národní účty ČR 1992 byly sestaveny v běžných cenách a obsahovaly celou posloupnost tokových účtů pro všechny sektory i národní hospodářství, souhrnnou ekonomickou tabulku[3] a tabulku input-output. Při sestavování účtů za rok 1992 bylo již rozhodnuto, že národní účty za rok 1993 (publikované v roce 1996) budou konstruovány podle standardu ESA 1995[4], který přinesl celou řadu základních metodických změn. To představovalo nejen přechod na novou podobu národních účtů, ale i nesrovnatelnost některých údajů národních účtů 1992 a 1993.

Kvalita údajů v národních účtech 1993 byla negativně ovlivněna nejen překotným přechodem na nový standard, ale zejména probíhajícím procesem privatizace, o jehož průběhu neměl ČSÚ k dispozici spolehlivé informace (nedostatek informací např. o kupónové privatizaci mj. značně deformoval účet domácností). Národní účty 1993 obsahovaly účty národního hospodářství a všech rezidentských sektorů podle metodiky ESA 1995 a dále účty produkce a tvorby důchodu podle odvětví. Posloupnost účtů sektorů nebyla úplná. Nebyl sestaven účet ostatních změn aktiv a majetkový účet (rozvaha). Veškeré údaje byly v běžných cenách.

Národní účty za rok 1994 (publikované v červenci 1997) byly dalším krokem ke kvalitě národních účtů. ČSÚ publikoval vedle omezené posloupnosti účtů sektorů (stejné jako pro rok 1993) a účtů odvětví (účet produkce a účet tvorby důchodu) ještě i tzv. integrované ekonomické účty, obsahující úplnou posloupnost účtů podle ESA 1995. Doplňková část integrovaných ekonomických účtů (tj. účty ostatních změn aktiv a majetkové účty) měla ryze experimentální charakter a údaje v ní obsažené byly postupně zpřesňovány.

V následujících letech 1998 a 1999 byly publikovány účty za další období, tj. za roky 1995 a 1996. Kvalita výsledků prezentovaných ve formě národních účtů se postupně zlepšovala. Národní účty však byly dále sestavovány pouze v běžných cenách, chyběla stále tabulka input/output. Zásadním problémem využitelnosti údajů národního účetnictví v České republice zůstávaly i nadále lhůty jejich publikování. Tříletý odstup byl příliš dlouhým obdobím na to, aby odborná veřejnost plně akceptovala národní účty jako základní zdroj informací o výsledcích národního hospodářství a zejména jako jedinečný analytický nástroj.

[3]Souhrnná ekonomická tabulka (dnešní integrované ekonomické účty) je dokumentem, který ukazuje vzájemnou propojenost všech údajů národních účtů. Její sestavení (tehdy ještě nepovinné) bylo ukázkou kvality našich prvních národních účtů. Souhrnná ekonomická tabulka je výsledkem práce francouzských statistiků a před zavedením standardu SNA 1993, resp. ESA 1995 ještě nebyla jejich součástí.

[4]European System of Accounts, 1996.

Po roce 2000 se lhůty publikování ročních národních účtů postupně zkracovaly až na interval obvyklý v zemích Evropské unie, tj. devět měsíců pro tzv. předběžnou sestavu, 18 měsíců pro tzv. semidefinitivní sestavu a 30 měsíců pro definitivní sestavu. Postupně se zkvalitňovala úplnost a spolehlivost českých národních účtů. V současné době národní účty České republiky plně odpovídají standardům obvyklým ve vyspělých evropských zemích.

Poptávka po krátkodobých statistických informacích vedla postupně i k vybudování systému čtvrtletních národních účtů. Nejdříve šlo jen o odhady složek tvorby a užití HDP (tyto údaje jsou nyní k dispozici veřejnosti ve lhůtě 70 dní po konci uplynulého čtvrtletí), které jsou od roku 2008 zasazeny do obecnějšího rámci čtvrtletních národních účtů (publikovány 90 dní po skončení čtvrtletí). V roce 2008 zavedl ČSÚ ještě i zveřejňování předběžných, tzv. flash odhadů vývoje hrubého domácího produktu. Tyto předběžné odhady jsou publikovány 46. den po skončení hodnoceného čtvrtletí.

V návaznosti na změny vyvolané zavedením a rozvojem národního účetnictví došlo i k zásadním změnám i v ostatních částech hospodářské statistiky – jmenujme zejména cenovou statistiku a s tím související měření inflace, odvětvové statistiky, statistiku zaměstnanosti, mzdovou statistiku, konjunkturální průzkumy a další. Popis změn, ke kterým došlo v každé z těchto jednotlivých oblastí, by vyžadoval a zasluhoval samostatný článek. Snad na závěr upozorněme jen na některé základní aspekty těchto změn.

Liberalizace cen znamenala větší poptávku po kvalitních a rychlých informacích o vývoji cen nejen ve sféře výroby, ale zejména ve sféře spotřeby. I když vývoj cen se sledoval a cenové indexy byly publikovány i před rokem 1989[5], získala cenová statistika zcela jinou dimenzi. Nutnost **měření inflace** znamenala přechod na **index spotřebitelských cen** respektující standardy obvyklé ve vyspělých zemích (definice spotřebního koše, pravidelné revize – 1991, 1995, 2001, 2007, lhůty publikování). Vedle „národního“ indexu spotřebitelských cen (od kterého je odvozena v národním hospodářství platná míra inflace) je od roku 2004 počítán i harmonizovaný index spotřebitelských cen (od kterého je odvozena v EU srovnatelná míra inflace).

Statistika zaměstnanosti, před rokem 1990 orientovaná na evidenci zaměstnanosti, se rovněž musela přizpůsobit nastalé ekonomické situaci (velký počet ekonomických subjektů, nezaměstnanost, evropské a světové standardy apod.). První statistické ukazatele zaměstnanosti a nezaměstnanosti, plně

[5] Tzv. indexy maloobchodních cen zboží a služeb se počítaly od roku 1967 ve čtvrtletní periodicitě. Poslední index maloobchodních cen zboží a služeb vycházel ze souboru reprezentantů (celkem 1350 reprezentantů) a systému vah z roku 1984 a počítal se od roku 1987. V současný index spotřebitelských cen, platný od roku 2007 se spotřebním košem z roku 2005, má celkem 714 reprezentantů.

odpovídající mezinárodně platným definicím Mezinárodní organizace práce (ILO), byly publikovány v roce 1993. Základem pro ně bylo nově vytvořené Výběrové šetření pracovních sil prováděné v domácnostech, poprvé organizované Českým statistickým úřadem koncem roku 1992. Toto šetření se provádí kontinuálně průběhu celého roku jako panelové výběrové šetření, v současné době na vzorku 25 000 vybraných bytů, což představuje přibližně 50 000 osob ve věku nad 15 let. Dalším subjektem zjišťujícím a publikujícím údaje ze statistiky zaměstnanosti a nezaměstnanosti je Ministerstvo práce a sociálních věcí. Odlišný způsob zjišťování (výběrové šetření jako zdroj informací pro ČSÚ a evidence jako zdroj pro MPSV) vede však k odlišným (i když ne podstatně rozdílným) výsledkům. Podobné problémy provázejí i statistiku mezd.

Zcela nově se objevily v ČR tzv. průzkumy očekávání pod názvem **konjunkturální průzkumy**. Tyto průzkumy se v ČR provádějí od roku 1991, a to v průmyslu, ve stavebnictví a v obchodě, nejdříve jako čtvrtletní, od roku 1993 pak jako měsíční. Od roku 1998 se k průzkumům ve výrobní sféře připojily ještě spotřebitelské průzkumy a v roce 2002 přibyly průzkumy ve vybraných odvětvích služeb.

Cílem konjunkturálních průzkumů je doplnit tradiční krátkodobé kvantitativní informace, poznat subjektivní názory vedoucích pracovníků podniků na perspektivy jejich vývoje a v neposlední řadě pak poskytnout těmto subjektům do jisté míry prognostickou informaci. Konjunkturální průzkumy tedy nabízejí informace o tendencích vývoje, o očekáváních v hospodářské sféře apod. Formulace otázek, kvalitativní odpovědi a způsob jejich vyhodnocení eliminují vlivy, které komplikují vypovídací schopnost číselných charakteristik: kalendářní variace, náhodné výkyvy tak, aby výsledkem byla syntetická informace o očekáváních ekonomických subjektů. Cílem spotřebitelských průzkumů je zejména poznat záměry domácností ohledně nefinančních a finančních investic. Charakteristickým znakem konjunkturálních průzkumů je rychlost a malá náročnost zjišťování a zpracování. Výsledky konjunkturálních průzkumů totiž jsou vždy v předstihu před výsledky klasického statistického šetření o stejném jevu, a hrají tudíž do jisté míry roli „předstihových“ ukazatelů. Zárukou rychlosti je i kvalitativní povaha většiny otázek pokládaných v konjunkturálních průzkumech.

Hospodářská statistika prošla za uplynulých dvacet let skutečně bouřlivým vývojem, představujícím zásadní změny obsahu a definice ukazatelů, způsobů zjišťování a zpracování informací. V neposlední řadě rovněž přinesla srozumitelnost a otevřenost poskytovaných informací vůči uživatelům. Cesta to nebyla snadná a všech žádoucích změn nebylo možné dosáhnout hned. Snad právě těch dvacet let je tou dobou, kdy s jistotou můžeme říci, že se

česká hospodářská statistika skutečně dostala na úroveň standardů obvyklých ve vyspělých zemích.

Literatura

[1] European System of Accounts – ESA 1995 (Système Européen des Comptes – SEC 1995). Luxemburg, EUROSTAT, 1996.

[2] European System of Integrated Economic Accounts – ESA 1978 (Système Européen des Comptes économiques integrés). Luxemburg, EUROSTAT, 1979.

[3] HINDLS R., HRONOVÁ S.: Pět let národního účetnictví – ohlédnutí a očekávání. Politická ekonomie. 1998, roč. 46, č. 4, str. 555–565.

[4] HRONOVÁ, S.: Ke koncepci majetkových účtů institucionálních sektorů a národního hospodářství. Statistika. 1996, roč. 33, č. 7, str. 292–301.

[5] HRONOVÁ, S., HINDLS, R.: Vliv revize standardů OSN a Evropské unie na národní účty České republiky. Zpravodaj ČSÚ 1997, roč. 39, č. 3, str. 33–38.

[6] HRONOVÁ, S., FISCHER, J., HINDLS, R, SIXTA, J.: Národní účetnictví. Nástroj popisu globální ekonomiky. 1. vydání. Praha, C. H. Beck, 2009.

[7] JÍLEK J. a kol.: Nástin sociálně hospodářské statistiky. Praha, VŠE, 2005.

[8] JÍLEK J., MORAVOVÁ J., HRONOVÁ S.: Sociálně hospodářská statistika. Praha, VŠE, 1997.

[9] JÍLEK J., SOUČEK E.: Ekonomická statistika v praxi. Praha, SNTL, 1990.

[10] NACHTIGAL V: Experimentální sestavy sektorových účtů ČSFR podle metodiky ESA. Praha, Institut ekonomie ČNB, 1993.

[11] System of National Accounts 1993 – SNA 1993. New York, United Nations, IMF, OECD, EUROSTAT, World Bank 1993.

[12] www.czso.cz.

MINULOST, SOUČASNOST A BUDOUCNOST BIOSTATISTIKY

Marek Malý, Zdeněk Roth

Adresa: SZÚ, Šrobárova 42, 100 48 Praha 10

E-mail: marek.maly@szu.cz, RothZdenek@seznam.cz

Abstrakt

V článku jsou nejprve uvedeny některé klíčové události z historie rozvoje statistiky, zejména ve spojitosti s řešením problémů v biologických odvětvích vědy, které vedly ke vzniku speciálního odvětví aplikované statistiky, biostatistiky, a to jak ve světě tak i v českém prostředí. Další text je věnován oblastem biologických věd, v nichž biostatistika hraje podstatnou roli. Takovými jsou např. demografie, epidemiologie, genetika, experimentální medicína včetně farmakologie, šlechtitelské metody v zemědělství, ekologie se studiem vlivu životního prostředí a problémy při tvorbě stochastických modelů pro časové průběhy chronických onemocnění. Jsou zmíněny typické problémy, jejichž řešením se v současnosti biostatistika zabývá. Zdůrazňuje se potřeba dobrého porozumění mezi biostatistikem a biologem, resp. lékařem při plánování i analýze výsledků studií na lidech nebo jiných živých organizmech. V budoucnosti je důležitá možnost využití moderní výpočetní techniky pro tvoření a zejména ověřování statistických modelů pro mnohorozměrná data s využitím relačních databází.

At the beginning, some key points in the history of statistics are presented, especially those connected with research in biological science which led to the rise of a particular branch of applied statistics, biostatistics. First worldwide, then for the Czech country. The following text describes those parts of biological sciences where statistics plays substantial role. Those are for instance demography, epidemiology, genetics, experimental medicine and pharmacology, breeding methods in agriculture, ecology with studies of the effect of life environment and problems in forming stochastic models for time trends in chronic diseases. Problems typical for contemporary biostatistics are mentioned. The need for good understanding and communication between biostatistician and biological professional when planning and analysing the results of studies on biological subjects is stressed out. In the future, the possibility of using modern computational techniques should enable the forming and testing the complex statistical models for multinomial data with possible adaptation of relational databases.

Biostatistika je odvětví statistiky, které se zabývá zpracováním dat získaných při řešení úkolů i problémů v oblasti studia živých organizmů (živočišných

i rostlinných). Je to tedy obor uplatňující matematické a pravděpodobnostní modely při vyhodnocování pozorovaných dat. Lze říci, že je to vyšší stupeň uplatnění zdravého rozumu v takových úlohách. Biostatistika využívá velmi širokou škálu standardních statistických metod, ale povaha reálných dat často vyvolává potřebu tvorby nových postupů či zobecnění stávajících tak, aby výsledný matematicko-statistický model korektně postihoval empirická data. Důležitou součástí při řešení některých úloh jak v oblasti zemědělské tak lékařské (i zvěrolékařské) je plánování pokusů pro studium vlivu vybraných vnějších faktorů na zdravotní stav sledovaných živých organizmů.

Už v minulosti lidé usuzovali na mnohé souvislosti pozorovaných jevů, aniž ještě znali moderní statistické metody. Zajímavý je např. příklad, který uvádí Stephen M. Stigler v 56. svazku Biometrics z roku 2000 v článku The Problematic Unity of Biometrics. Tak ve Starém zákoně, na začátku knihy Danielovy, je uveden do jisté míry pohádkový příběh o tom, jak Daniel se svými třemi druhy odmítal krmi, kterou jim nabízel na příkaz krále Nabukadnezara jeho dvořan. Chtěli se živit pouze zeleninou a vodou a na obavu dvořana o jejich zdraví i vlastní bezpečnost nabídl Daniel uspořádání pokusu. Nechť mohou jíst svou vybranou zeleninovou dietu a pokud po 10 dnech bude ve srovnání s ostatními, kteří nabízenou krmi neodmítali, jejich zdravotní stav horší, nechť jsou potrestáni. Po 10 dnech však byl vzhled všech čtyř podstatně lepší než ve srovnávané kontrolní skupině. Bylo jim tedy povoleno dál se živit zeleninovou dietou. Jisté postupy dnes běžné ve srovnávacích pokusech a klinických studiích byly tak včleněny už do starodávných bájí.

Analýzu dat narození, svateb a pohřbů v Londýně publikovanou J. Grauntem v práci Observations upon the Bills of Mortality v roce 1665 lze do určité míry považovat za předobraz dnešních analýz dat státní statistiky a dat o výskytu epidemií (tehdy byl v Londýně mor). V 19. století se postup usuzování z pozorovaných dat na možné odkrytí vlivu vnějších faktorů uplatnil u vídeňského gynekologa Semmelweise, který porovnal snížení nemocnosti rodiček před a po zavedení desinfekčního mytí rukou obslužného personálu, lékařů i studentů a na základě výrazného rozdílu usoudil, že zdrojem infekce byly infikující nemyté ruce při ošetření rodiček.

Na přelomu 19. a 20. století došlo ke znovuobjevení Mendelových prací a moderní biologie potřebovala ke svému rozvoji nástroje, které dnes zahrnujeme pod biostatistiku. Práce F.Galtona, K.Pearsona a W.F.R. Weldona o dědičnosti a biologické variabilitě již obsahovaly základy statistických metod a vedly k založení časopisu Biometrika (1901). Do této doby spadají i práce W.S. Gosseta publikované pod pseudonymem „Student“ o t-rozložení a t-testech (1908) či Spearmanova práce o korelaci (1904). Nedávno tak uběhlo několik stoletých výročí zásadních okamžiků ve vývoji moderní statistiky. Co

všechno obsáhnul přední „biostatistik“ tehdejší doby, Student, nám v roce 1993 v Informačním bulletinu krásně připomněl prof. Komenda.

V roce 1925 vyšla v prvním vydání kniha R.A. Fishera Statistical Methods for Research Workers, obsahující základní koncepty matematické statistiky v podobě, kterou známe dnes. Lze říci, že rozvoj matematické statistiky byl podnícen problémy spojenými s modelováním a vyhodnocováním dat biostatistiky. To se plně projevilo po 2. světové válce, kdy byly publikovány stěžejní práce, z nichž vycházejí postupy dodnes v praxi užívané, a kdy vznikly i nové časopisy, např. Biometrics. Fisher použil v zemědělských experimentech poprvé velmi důležitý princip randomizace, který poté A.B. Hill přenesl do medicíny a navrhl první randomizovaný klinický pokus, který v roce 1948 prokázal účinnost streptomycinu při léčbě tuberkulózy. Dodnes je tento přístup v mnoha modifikacích a zobecněních základem klinických studií.

Díky plánování zemědělských pokusů a statistickému zpracování výsledků se urychlil vývoj pěstění nových odrůd, stejně tak se aplikací statistického plánování a zpracování laboratorních srovnávacích pokusů dosáhlo pokroku při rozvoji nových léků i poznatků o fyziologii života.

V českých zemích se statistické metody v biostatistice uplatňovaly zejména v poválečných díky vzrůstající potřebě statistiky ve výzkumných ústavech, v kterých se často ustavovala statistická pracoviště. Podstatnou podporu těmto pracovištím poskytovala i katedra pravděpodobnosti a matematické statistiky Karlovy univerzity. Při této příležitosti je třeba ocenit zejména již zesnulých MgMat. Marcela Josífka a Ing. Josefa Machka, CSc. Prvý jmenovaný pořádal v 50. letech 20. století pod záštitou Vědeckotechnické společnosti spolu s MgMat. Vladimírem Malým sérii přednášek věnovaných převážně aplikacím statistických metod v biologii a medicínských oborech. Ing. Machek pak po léta spolupracoval jako statistický poradce s řadou vědeckých pracovišť. V pozdějších letech byl jedním z vůdčích pracovníků v biostatistice doc. RNDr. Tomáš Havránek, DrSc. z Akademie věd. V oblasti biologických věd pracovala od 70. let řada žáků katedry matematické statistiky a pravděpodobnosti, kteří se většinou začlenili do výzkumných ústavů a jsou, pokud dosud pracují, mnozí i členy České statistické společnosti. Po roce 1989 mnoho výzkumných ústavů zaniklo (např. Výzkumný ústav pro farmacii a biochemii) či zrušilo biostatistické pozice. Proto je dnes u nás poměrně málo biostatistiků, kteří mají skutečně jako hlavní náplň práce analýzu biologických a medicínských dat a příslušnou konzultační činnost. Jejich nedostatek je suplován lékaři či inženýry, kteří mají statistické povědomí, ale přeci jen také mají určité limity. Ani některé lékařské fakulty nemají své biostatistiky.Nicméně rozvíjejí se i nová biostatistická pracoviště, jmenujme alespoň

Institut biostatistiky a analýz Masarykovy univerzity v Brně. V Praze se o podobném pracovišti zatím stále hovoří.

Současná biostatistika, tak jako i jiná odvětví statistiky, stále intenzívně využívá mnoha postupů vyvinutých v průběhu celého minulého století. K vůbec nejcitovanějším statistickým článkům patří text E.L. Kaplana a P. Meiera o odhadu křivky přežívání z roku 1958 a článek D.R. Coxe z roku 1972 popisující model proporcionálních rizik. Na tyto a podobné články pak navázalo mnoho dalších, které rozpracovávají jednotlivé detaily původní myšlenky.

Mezi aktuální témata a rychle se rozvíjející metodologie patří v posledních letech zejména postupy spojené s hodnocením observačních epidemiologických studií a klinických studií. Dotýkají se všech fází procesu realizace studie od plánování potřebného počtu subjektů a definice cílových proměnných přes metody získávání a sběru dat a způsob ošetření chybějících hodnot až po otázky volby adekvátního modelu, který je schopen zohlednit často velmi komplikovanou strukturu dat, a způsob publikace a interpretace výsledků.

V oblasti analytických postupů se jedná zejména o modely pro longitudinální data, tj. data opakovaně měřená na týchž jedincích, a proto zpravidla korelovaná, modely se smíšenými efekty a dále o Bayesovské metody, metody mnohorozměrné statistiky včetně analýzy mnohorozměrných kategoriálních dat, robustní metody, prostorovou statistiku, metody zohledňující chyby v měření potenciálních prediktorů, analýzu genetických dat, počítačově intenzivní metody opírající se o rychlý rozvoj výpočetních možností (např. bootstrap, jackknife, MCMC) a vůbec celou oblast výpočetní statistiky.

Biostatistika 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 onemocnělých osob, jejich způsobu života a životního prostředí. V posledních letech se takové studie zpracovávají lépe díky tomu, že lze pro každého jedince nebo skupinu osob ukládat i mnoho zmíněných informací do počítačových databází a ty pak používat pro podstatně sofistikovanější analýzy zaznamenaných dat. Stále se rozvíjejí metody pro analýzu cenzorovaných dat (např. se vyvíjejí různé alternativy Coxova modelu, analyzují se časově proměnlivé kovariáty, rozvíjejí se modely konkurujících si rizik). Nové typy epidemiologických studií také často vyžadují nové způsoby analýzy.

Za dobu existence České statistické společnosti, tedy za posledních 20 let, se rozvíjely mnohé výše zmíněné metody. Jde o výběr do určité míry subjektivní, každý biostatistik by asi akcentoval trochu jiný okruh. V roce 2000 jsme měli tu čest uvítat jednoho z nejpřednějších světových biostatistiků prof. N. Breslowa a vidět skutečně moderní a aktuální aplikace biostatistiky. Vy-

slechli jsme jeho přednášky na kursu Příprava a statistická analýza epidemiologických studií, který uspořádala Česká statistická společnost ve spolupráci s mezinárodní společností pro klinickou biostatistiku.

Specifickou kapitolu představují klinické studie hodnocení léků, v nichž se testuje tzv. bioekvivalence. Jde zhruba řečeno o to, zda lze nahradit starou formu léku formou novou. Je tedy potřeba prokázat obdobnost průměrných hodnot, ale požaduje se rovněž, aby si hodnoty odpovědi na testovaný a referenční lék byly dostatečně blízké u většiny jedinců. V obvyklých statistických testech je nulovou hypotézou předpoklad, že sledované proměnné žádným kontrolovaným vlivem ovlivněny nejsou a cílem experimentátorů je tento předpoklad zamítnout. Při testování bioekvivalence jsou role nulové a alternativní hypotézy prohozeny. Testovanou hypotézou je, že se oba léky liší víc, než je žádoucí, a její zamítnutí je pak potvrzením, že obě porovnávané formy téhož léku se (po předchozí aplikaci stejně velkých dávek) podstatněji neliší. I zde vznikají specifické problémy, zejména jak stanovit rozmezí, v kterém lze považovat léky za ekvivalentní. Odpověď zpravidla přináší kombinace statistických postupů s lékařskými poznatky.

Kromě čistě matematických otázek je třeba řešit i další otázky, např. přizpůsobení analýz požadavkům regulačních orgánů (zejména při vývoji nových léků), či v jaké podobě a zda vůbec aplikovat postupy mnohonásobného srovnávání. Statistici se musí zabývat i etickými otázkami realizace klinických studií, např. v souvislosti s pravidly pro včasné zastavení studie. Dlouhodobě aktuální zůstávají otázky mezioborové spolupráce. Známý statistik E.Gehan odhadl, že až 80% úsilí vynaloží na porozumění problému a komunikaci se specialisty lékaři či biology, zatímco rozhodnutí o biostatistickém přístupu k řešení problému vyžaduje pouhých 20%. Dobrý biostatistik musí kromě teoretických znalostí a schopnosti analyzovat reálná data umět dobře komunikovat s odborníky jiných oborů. Tento aspekt je přitom v rámci vzdělávání statistiků často opomíjen.

Je potřeba se stále zabývat i základními otázkami. Například význam intervalu spolehlivosti je pořád mnoha lékařům nejasný, a proto ho mnozí stále opomíjejí i přesto, že biostatistici a velké časopisy je vyžadují. Často je z celého výpočtu publikována pouze p-hodnota, což rozhodně není dostatečné. Stále se vracejí otázky vztahu mezi věcnou a klinickou významností. Epidemiologie se snaží zjistit, zda vztah mezi určitou expozicí a zdravotním jevem je kauzální, a proto často dochází k rozporům v interpretaci statistických výpočtů, které kauzalitu prokázat neumějí. Epidemiologové velmi rádi dichotomizují spojité veličiny s hlavním argumentem, že se tak zvýší přehlednost výstupů. Občas je dichotomizace přímočará a má smysl, ale častěji vede k podstatné ztrátě informace.

Statistické programové balíky jsou dnes relativně dostupné, a proto je mají a používají i mnohá lékařská a biologická pracoviště, která pak opomíjejí spolupráci se statistiky. V souvislosti s tím ale vyvstává několik problémů. Biologové a lékaři jsou často nedůslední či chybují při přípravě kontrole dat, při volbě správného statistického postupu a při interpretaci výsledků. Z podstaty věci je jasné, že nelze připravit plně vyčerpávající návod na volbu metody vhodné pro danou situaci, což v praxi vede buď k používání vysloveně chybných postupů anebo postupů nedostatečně komplexních. Biostatistik je pak často kontaktován až v případě nutnosti znovu a řádně analyzovat data při vrácení publikací autorům s námitkami recensentů časopisu. Pokud je spolupráce se statistikem byla v průběhu výzkumu opomenuta, měl by alespoň mít možnost chystané publikace vidět před jejich zasláním do časopisu, aby se mohl případně pokusit statistické nedostatky omezit. i když třeba v případě rozsahu výběru to už zpravidla není možné.

Vhodnost statistického modelu pro analyzovaná empirická data je podstatná pro veškeré aplikace statistických metod. Statistický model je propojen i s otázkami plánování pokusu nebo šetření. Souvisí to i s otázkami, kteří jedinci a v jakém počtu mají být sledováni, nebo jaká srovnání je možno při daném modelu otestovat a jak otestovat, resp. odhadnout podstatné parametry modelu a jejich přesnost. Parametry modelu mají mít reálný a interpretovatelný protějšek v datech. Důležitá je možnost dostatečně spolehlivého odhadu nahodilé variability zkoumaných dat.

Jedním z podnětů pro rozvoj biostatistiky v posledních letech byla nově objevená onemocnění, např. SARS, ale zejména AIDS. Snaha o modelování výskytu HIV/AIDS a přenosu (šíření,dynamiky) nákazy virem HIV, 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 prodlužovaná novými typy kombinované léčby. Infekčnost je proměnlivá v průběhu onemocnění. Modely analýzy přežívání jsou proto komplikované. Data jsou dvojitě cenzorovaná: až na výjimky není známa doba nákazy či sérokonverze (cenzorování zleva), někdy je známo datum posledního negativního a prvního pozitivního testu (intervalové pozorování), událost (klinický projev AIDS, úmrtí) často ještě nenastala (cenzorování zprava), dochází ke ztrátám pacientů ze sledování. Používají se zástupné indikátory progrese infekce, zejména počet CD4+ T-lymfocytů a virová nálož. Ty jsou však měřeny s velkou chybou násobenou ještě detekčními limity přístrojů. Základním zdrojem poznatků o výskytu zejména infekčních onemocnění jsou databáze registrovaných případů, které se mj. využívají k časné detekci změn v trendu. Soubor úkonů a metod zajišťujících průběžné shromažďování, upřesňování a vyhodnocování

údajů o distribuci a šíření onemocnění, se označuje termínem surveillance. Systémy surveillance slouží nejen pro analýzu shromažďovaných dat, ale pro jejich interpretaci a hlavně zpětnou vazbu, tedy využití získaných poznatků při prevenci a zavádění protiepidemických opatření. V České republice jsou všechna infekční onemocnění hlášena do celostátního systému Epidat a do několika dalších dílčích systémů, např. pro sledování výskytu HIV/AIDS. Je třeba si uvědomit, že biostatistici v našich podmínkách často nejen analyzují data, ale musí se aktivně starat i o jejich kvalitu a dohledávání údajů, což představuje mnoho hodin práce, která není navenek patrná.

Do popředí se v poslední době dostala také témata, která jsou statistiky vnímána poněkud rozporuplně. Jedním z nich je meta-analýza jakožto nástroj pro identifikaci a zhodnocení převažujících trendů na základě kombinace údajů z několika studií zkoumajících stejné či podobné hypotézy. Vznikají zde problémy statistického rázu (např. zda použít model s pevnými či náhodnými efekty), ale závažnější jsou problémy technické (jaké studie zahrnout do meta-analýzy, jaké nezahrnout, jak řešit situace, kdy např. dvě menší studie ukazují jedním směrem a jedna větší opačným). Velkým úskalím je možné zkreslení, bias. Není možné zaslepené hodnocení jako u jednotlivých klinických studií. Kvalita původních studií silně determinuje kvalitu meta-analýzy. Druhým široce diskutovaným, až módním pojmem je medicína založená na důkazech (evidence-based medicine, EBM). Označuje se jím „vědomé, zřetelné a soudné používání nejlepších současných důkazů při rozhodování o péči o jednotlivé pacienty“ (D.L.Sackett). Pro statistiky tento přístup nepředstavuje nic nového ani překvapivého, jak už v Informačním bulletinu České statistické společnosti v roce 2006 podotkl prof. Komenda. Nicméně mnozí lékaři dosud vycházejí při léčbě především z vlastních zkušeností, nikoli z výsledků epidemiologických studií. Principy EBM vyžadují, aby lékař po zformulování problému vyhledal relevantní informace („důkazy“) v databázích odborných časopisů a propojil tak klinické posouzení konkrétního pacienta s nejlepší dostupnou externí informací. Je rovněž vypracována hierarchie důkazů podle toho, z jakého typu studie pocházejí. Z tohoto pohledu lze akcentování potřebnosti důkazů považovat za užitečné a biostatistik může lékaři pomoci v orientaci.

Podstatné je, aby se statistik podílel i na přípravě epidemiologických studií, a to jak při odhadu potřebného rozsahu studie, tak i v plánu sběru dat (kromě dat přímo o nemoci je potřeba se zaměřit na údaje, jejichž nezahrnutí do analýzy může zkreslit konečné výsledky, např. věk, pohlaví, životní styl apod.). V souvislosti s metodikou plánování a analýzy epidemiologických studií je vhodné podotknout, že zejména v anglosaských zemích se epidemiologie považuje za speciální disciplinu biostatistiky. Pro úspěšné aplikace statistiky

v analýze biologických i lékařských dat je pak podstatné, aby se spolupráce statistiků a biologů nebo lékařů jak při plánování pokusů a šetření i při konečné analýze výsledných dat stala samozřejmostí.

Jak se bude v blízkém budoucnu vyvíjet biostatistika, a tedy i statistika vůbec? Je jasné, že od počátku 20. století byla teoretická část matematické statistiky těsně svázaná s počtem pravděpodobnosti a byla rozvíjena řadou autorů. Pro většinu klasických experimentálních i výběrových studií, v nichž se data považují za náhodné jevy a předmětem zkoumání jsou jejich distribuční funkce a zejména jejich parametry, byly zkoumány rozsáhlé třídy statistických modelů a vytvořeny algoritmy jak v oblasti testování hypotéz tak teorii odhadů. Empirická data jsou aproximována statistickými modely a odhady parametrů těchto modelů slouží k analýze původních empirických dat. Tak se vyvíjely aplikace statistiky na empirická data.

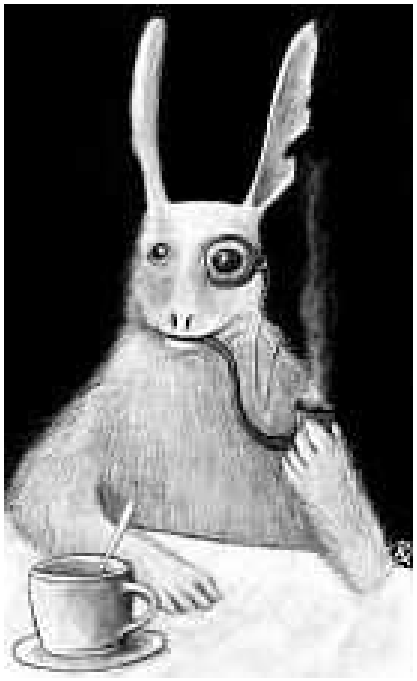
Pokrokem statistiky byly a asi i nadále budou vhodně budované statistické modely pro empirická data, která se často nedají ve své komplexnosti dosud běžně užívanými statistickými modely dostatečně dobře aproximovat. Mnohé problémy jsou v oblasti stochastických řad a procesů, nebo v oblasti analýzy vektorových proměnných (a tedy vícerozměrných množin náhodných jevů), kde je sice často možno aplikovat mnohé statistické modely, je ale obtížné ověřit, který z nich je pro daný empirický soubor dat vhodný či nejvhodnější. Velký rozmach výpočetní techniky umožňuje využívat empirické ověřování vhodnosti i velmi komplikovaných postupů spojených se snahou modelovat empirická data v co nejlepší shodě se známou empirií. Jinou oblastí je možnost vytváření velmi rozměrných a rychle přístupných relačních databází, kde bude opět akutní problém zpracování a vyhodnocení marginálních odhadů vzájemných vztahů vybraných vektorů proměnných očištěných od zkreslení možnou interakcí se zbylými datovými vektory. Zpracování takových objemných a strukturovaných dat a příprava vhodných statistických modelů bude jedním z dlouhodobých směrů rozvoje biostatistiky v budoucnosti.

Když se podíváme na dynamický vývoj statistiky v posledních 20 či dokonce 100 letech, vidíme, že je téměř nemožné předvídat, jak bude probíhat další vývoj. Určitě bude ovlivňován technickými možnostmi, ale i možnostmi lidskými. Už dnes je vidět tendence ke specializaci mezi biostatistiky. Někteří jsou spíše teoretici, jiní spíše praktici, a těch, kteří v sobě vyváženě spojují oba aspekty, ubývá. Rovněž probíhá specializace na podobory, např. genetiku, či klinické studie. Publikovaných statistických prací je tolik, že je jedinec nemůže plně obsáhnout. Nicméně v rámci podoboru vznikají často práce s obecnější platností, a proto je potřeba, aby úzké vazby mezi statistiky z různých oblastí přetrvaly. Statistika musí reagovat i na společenské proměny, které např. negativně ovlivňují úroveň respondence studií a šetření.

POZVÁNKA NA VALNÉ SHROMÁŽDĚNÍ ČESKÉ STATISTICKÉ SPOLEČNOSTI

Všichni členové České statistické společnosti jsou srdečně zváni na Valné shromáždění, které se bude konat v pondělí, 7. února 2011 v místnosti číslo 336 v Rajské budově Vysoké školy ekonomické v Praze (vchod buď přes novou budovu z nám. W. Churchilla nebo přímo z ulice Italské). Začátek bude ve 13:00.

Na programu budou zprávy o činnosti a hospodaření společnosti, volby nového výboru a odborná přednáška. Přislíbená je přednáška pana Drapala z Českého statistického úřadu o Eurostatu a sčítání lidu, domů a bytů v roce 2011.



Obsah

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Časopis je zařazen na Seznamu Rady, více viz `http://www.vyzkum.cz/`.

Předseda společnosti: doc. RNDr. Gejza Dohnal, CSc.
ÚTM FS ČVUT v Praze, Karlovo náměstí 13, Praha 2, CZ-121 35
E-mail: `gejza.dohnal@fs.cvut.cz`

Informace pro autory jsou na stránkách `www.statspol.cz`.

ISSN 1210–8022

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.