

STAKAN 2007 – Sborník příspěvků

STAKAN 2007 – Sborník příspěvků

Sborník příspěvků konference konané na Rusavě,
Česká republika, ve dnech 25. – 27. května 2007.

Organizátory byla
Česká statistická společnost
(<http://www.statspol.cz/>)

a

Slovenská statistická a demografická spoločnosť
(<http://www.ssds.sk/>).

Editoři sborníku jsou
Pavel STRÍŽ, Gejza DOHNAL a Jaromír ANTOCH.

Vydáno v XII/2007, nakladatelstvím
Martin STRÍŽ, Bučovice. 1. vydání.

Řada článků byla o něco dříve publikována v časopisech *Forum Statisticum
Slovacum* Slovenské statistické a demografické spoločnosti a *Informačním
Bulletinu* České statistické společnosti. Editoři tímto děkují za souhlasy
k přetištění autorům i oběma společnostem.

ISBN 978-80-87106-07-5 (DVD-ROM)

PŘEDMLUVA

Dobrý den čtenáři!

Do rukou se Vám dostávají příspěvky ze společné česko-slovenské statistické akce STAKAN 2007 konané na Rusavě. Zvanými přednáškami byly profesor Petr Hebák *Výuka statistiky 2007* a profesor Gejza Wimmer *Matematické modelovanie v jazykovede*.

Než se, milý čtenáři, vrhneš do četby článků, tak něco málo poznámek.

Editori by rádi poděkovali rodině Strážové za pomoc před, během i s konferenčními pracemi.

Také děkuji panu Petru Nevřivovi, t. č. student Fakulty managementu a ekonomiky Univerzity Tomáše Bati ve Zlíně, s překlápěním z Microsoft Wordu do $\text{\TeX}$ u. Překlápění dva lidi stálo více než 230 člověkohodin. Editori si nejsou jisti, jestli bude možné tuto pomoc v takovém rozsahu nabídnout i v budoucnu. Proto editori apelují na matematiky a statistiky k dostudování základů sazby v $\text{\LaTeX}$ u.

Přikládáme proto na DVD anglickou a slovenskou příručku. Český překlad lze nalézt na adrese <http://www.penguin.cz/~kocer/texty/lshort2e/>.

Pokud si chcete vyzkoušet $\text{\TeX}$ záležitosti bez instalace programu, tak prosím navštivte webové rozhraní <http://live.contextgarden.net/> nebo <http://tex.mendelu.cz/>.

Ke studijním účelům přikládáme i $\text{\LaTeX}$ zdrojové kódy této publikace (včetně všech chyb a kliček), dle vzoru profesora Knutha, tvůrce $\text{\TeX}$ u.

A co bude dál? Mnoho toho bude! Minimálně bude Robust v roce 2008 (viz Upoutávka na další straně).

Srdečně Vás zdraví editoři!

V Bučovicích 1. prosince 2007

Pavel STRÁŽ, Gejza DOHNAL a Jaromír ANTOCH

REZERVUJTE SI ČAS JIŽ NYNÍ

ROBUST 2008 :: ZÁŘÍ 8. – 12., 2008

Vážené kolegyně a vážení kolegové,

dovolujeme si Vám oznámit, že ROBUST 2008 se uskuteční ve dnech 8. – 12. 9. 2008 v nádherném prostředí Ráckovy doliny v Roháčích za příchodu babího léta a především pak nových stochastických myšlenek jež, jak doufáme, dorazí s Vámi.

Tak jako i v minulých letech nabízíme příjemné prostředí, bohatý program odborný i společenský, nízké ceny a především neopakovatelnou atmosféru setkání stochastiků všeho druhu z Česko-Slovenska.

Již dnes Vám můžeme slíbit přednášku o historii matfyzáckých táborů u lovecké chaty poblíž druhé salaše, která se uskuteční v prostorách bývalého tábora. Prosíme Vás tímto, abyste ve svých „skrýších“ z mládí hledali dobové fotografie a jiné tajné relikvie z oněch dob a zapůjčili nám je na výstavku.

V nejbližších dnech dostanete podrobnější informace. Sledujte take náš web <<http://www.karlin.mff.cuni.cz/~antoch/>>

Na setkání se těší

J. Antoch, G. Dohnal a V. Witkovský

P.S. Hotel je připraven se postarat jak o ty, kteří přijedou dříve, tak o ty, kteří by se rádi po konferenci zdrželi déle.

REZERVUJTE SI PROTO SVŮJ ČAS JIŽ NYNÍ

STAKAN 2007

V DROBNOHLEDU DVOU ČLÁNKŮ

STAKAN 2007 – DOJMY A REFLEXE

Jaromír Běláček

Na seminář *STAKAN 2007* konaný ve dnech 25.–27. 5. 2007 mě pozval můj dlouholetý kamarád, přítel (a t. č. vedoucí) Honza K. Existenci semináře s výše uvedeným názvem jsem zaregistroval již dříve na seznamu akcí pořádaných Slovenskou statistickou a demografickou společností (zkr. SŠDS – viz www.ssds.sk), kterou již po delší dobu považuji – podobně jako Českou statistickou společnost (zkr. ČStS – viz také www.statapol.cz) dnes – za reprezentativní a důstojnou odbornou platformu demografie a statistiky (nejen pro Slovensko). Že název semináře znamená něco jako *STATistika & KANtoři*, jsem mohl Honzovi do té doby pouze věřit. Právě tak by totiž samotný název semináře mohl někomu podsouvat *SKLENICI, ne nutně neprázdnou...* Obzvláště když se seminář měl konat někde v lesích na moravsko-slovenském pomezí (přesněji v rekreačním zařízení Jestřábí-Rusava nedaleko Bystřice pod Hostýnem).

Odborný seminář byl *naplněn* především překvapivě širokým spektrem profesionálů z oblasti teorie, výuky a aplikací statistiky, povětšinou z českých a slovenských vysokých škol. Pro ilustraci uvádím na semináři aktivně působící reprezentace za ČStS (prof. RNDr. Jaromír Antoch, CSc., z MFF UK Praha, bývalý předseda ČStS; doc. RNDr. Gejza Dohnal, CSc., z FS ČVUT, stávající předseda ČStS), za SŠDS (Doc. Ing. Jozef Chajdiak, CSc., – vědecký tajemník – a další přítomní kolegové ze SR, vesměs z Univerzity Komenského v Bratislavě), z pozvaných přednášejících za VŠE Praha (prof. Ing. Petr Hebák, CSc.; Doc. Ing. Hana Řezanková, CSc., a další) a z regionu organizátorů semináře (prof. Ing. Zdeněk Molnár, CSc., a Ing. Pavel Stříž, Ph.D., z Univerzity Tomáše Bati ve Zlíně). Výše uvedenému odpovídá rovněž spektrum problematik referovaných účastníky v rámci cca 25 přednesených příspěvků. Protože ze semináře byl pořizován videozáznam a většina příspěvků by měla být publikována ve slovenském časopise *Forum Statisticum Slovaca*, omezím se dále jen na to, co mě osobně nejvíc oslovilo a zaujalo.

Výuka statistiky na VŠ je v současné době *sycená* již zcela samozřejmou dostupností statistického software s širokou podporou internetových aplikací. Mělo by tedy být rovněž samozřejmostí, že se studenti VŠ naučí používat všechny základní statistické metody potřebné pro jejich budoucí praxi. Jak podtrhl v rámci syntetické úvodní přednášky „Výuka statistiky dříve, dnes a zítra“ prof. P. Hebák, je toto samozřejmě záležitostí nejen zájmu samotné profesionální obce statistiků, ale rovněž zodpovědností vysoce postavených manažerů vědy, výzkumu a vzdělávání, i kompetentních vládních představitelů, aby se s mnohem větší měrou *opájeli* objektivními statistickými daty,

neboť se nacházíme v situaci, kdy samotná matematika je (cit.) „... potlačována anebo dokonce vytlačována z výuky na středních školách ...“. Rozvoj statistiky samotné by měl vycházet ze stejných teoretických základů, na kterých byla postavena již v 19. a 20. století (perspektivy v syntéze bayesovských a „klasických parametrických“ postupů?).

Certifikovanou možností spolupráce mezi statistickými orgány a vysokými školami by bylo možno *zapít* např. ve vztahu k nedávno podepsané Dohodě mezi ČSÚ a UTB ve Zlíně (Hrbáček, Stříž). Metody statistické analýzy, Aplikovaná statistika, Počítačové zpracování dat nebo Ekonometrie jsou předměty vyučované podle „starých i nových“ osnov (Klímek, Stříž). V těchto souvislostech byli účastníci semináře seznámeni rovněž s rozšířením možností, které skýtá v oblasti statistiky nejnovější nástavba programové aplikace Microsoft Excel (přítomnými expertně oceněno jako přínosné). Od března 2007 bylo na FaME UTB vytvořeno centrum, které si klade za cíl „propojení aktivit na poli aplikovaného ekonomického výzkumu“ (Molnár). Těmto závěrům konvenovalo rovněž vystoupení J. Chajdiaka, kde byla podtržena úloha statistiky: „Ukázat svět, jaký je ... s cílem: vytvářet podporu, resp. doporučení procesu objektivizace manažerského rozhodování.“

Řada příspěvků prezentovaných na semináři měla poměrně sofistikovaný a specializovaný *statistický obsah*. Toto se týkalo například přednášky prof. Wimmera (Matematický ústav SAV) „Matematické modelování v jazykovede“, kde bylo představeno rozsáhlé spektrum aplikací speciálních pravděpodobnostních rozložení, profi-přednášky M. Urbaníkové „Využitie štatistiky v poistnej matematike“, referencí H. Řezankové (VŠE Praha) týkající se „Výuky jednorozměrné a dvourozměrné analýzy kategoriálních dat“, oživení „Metody latinských čtverců“ v podání prof. D. Jaruškové (ČVUT) anebo přednášky doc. Tvrdíka (Ostravská univerzita) „Grafy a tabulky ve statistice (aneb na co ve výuce většinou není čas)“ všímající si aspektů ne zcela postačujících a adekvátních způsobů prezentace statistických dat a výsledků analýz.

Až překvapivě *osamoceně* (z hlediska informatiky) na mne zapůsobilo na této platformě vystoupení doc. Žváčka (VŠE Praha) „Statistické výpočetní prostředí 2007“, ale i M. Dolejšové (UTB Zlín – její vystoupení pod názvem „Zařazení geografických informačních systémů do výuky předmětu Informatika ve veřejné správě“), které sice odlišným, ale zcela transparentním způsobem reflektovaly internetovské směry a potřeby vizuálního rozvoje statistiky (ve světě, a tedy i v ČR?). .. Že by jen *do prázdna* měl vyznít návrh doc. Žváčka, že by i profesionálové z ČR mohli být více aktivní při rozvoji a tvorbě některých speciálních stránek typu www.wikipedia.org v oblasti statistiky? (Pravděpodobně jsme to doposud považovali za záležitost více-

méně individuálního přístupu k uvedené problematice.) V tomto smyslu zapůsobilo nepochybně hodnověrně vystoupení M. Vojtkové a I. Stankovičové (UK Bratislava) „Viackriteriálne hodnotenie krajín EU na základe vybraných ukazateľov Lisabonskej stratégie“ vedoucí k typologii zemí EU podle objektivních rozvojových a sociálně-ekonomických ukazatelů.

Aplikace v medicíně byly na *STAKANU 2007* pokryty vystoupeními M. Malého (Státní zdravotní ústav) „K otázkám výuky a statistických konzultací“ zhodnocující jak praktické, tak i teoretické zkušenosti z biomedicíny a J. Běláčka (1. LF UK Praha + VFN) usilujícího o demonstraci obdobné problematiky v rámci příspěvku s názvem „Analýza laterality u dětí ve věku 9-11 let (aplikace statistických metod)“. Na *zdraví* se ve své aplikační části orientovalo rovněž vystoupení M. Žambochové (FSE ÚJEP v Ústí nad Labem) „Jak na rozhodovací stromy“ (zde byla shrnuta dlouhodobě nedostatečně průhledná metodika analýzy dat prostřednictvím regresních stromů) a přednáška I. Saxla (MFF UK Praha) „Alfred Binet a počátky testování inteligence“, která překročila rámec statistiky až do oblasti *historie psychiatrie*.

Kdybych byl býval věděl v jaké atmosféře na nás letošní *STAKAN čeká*, tak bych určitě na tuto akci přijel o den dříve. . . Z hektického běhu každodenních pracovních povinností byli všichni účastníci doslova od prvního vstupu *napojeni* na poklidnou a konstruktivně rehabilitující diskusní atmosféru, což bylo samozřejmě zásluhou všech oficiálních i neoficiálních pořadatelů akce. Týkalo se to nejen dopravy na místo, ubytování, stravování a vydařeného počasí, ale i organizace kulturního programu: vystoupení Taneční skupiny Kirké předvádějící břišní tance, pátečního táboráku, sobotního odpoledního výletu do okolí Rusavy. Do této kategorie lze přiřadit i vizuálně velmi atraktivní přednášku „Velkorozměrné struktury vesmíru – zpracování dat o polohách galaxií“ (J. Vašátko – Zlínská astronomická společnost), která byla pro zájemce rozšířena o praktické pozorování objektů noční oblohy. Hvězdářským dalekohledem zvětšená planeta Jupiter (*k vidění dvakrát* se svými čtyřmi největšími měsíci se změnou jejich vzájemného postavení), na mě rozhodně zapůsobila pozitivněji, než by dokázala sklenice slivovice (*viděná nadvakrát* v obdobném postavení). . .

Na závěr se omlouvám všem přednášejícím (a zejména těm), jejichž vystoupení na semináři *STAKAN 2007* bylo dostatečně smysluplné a originální a v mém rádobychaotickém přehledu o nich nebyla zmínka. Účelem tohoto shrnutí nebyla *úplnost* nebo *úsilí o plnou* objektivitu (a už vůbec ne *degustace či hodnocení!*). Výuka statistiky v ČR má nepochybně silný a odborně erudovaný potenciál. V tomto smyslu byl letošní *STAKAN vyzvednut* určitě hodně vysoko. Těšme se na další. . .

Adresa: Ústav biofyziky a informatiky, I. lékařská fakulta, Univerzita Karlova

STAKAN 2007 (RUSAVA, JESTŘÁBÍ, ČESKÁ REPUBLIKA, 25. – 27. 5. 2007)

Kráľ a kolektív doktorandov z Banskej Bystrice

Niekoľko (vôbec nie subjektívnych) pohľadov na konferenciu STAKAN 2007 (alebo Ako nám bolo dobre...)

Na úvod by sme sa hádam mali trošku predstaviť. Sme dosť konzistentná skupinka šiestich „večných študentov“ zo samotného srdca Slovenska, pozostávajúca z piatich doktorandov a jedného „doktoranda už ukončeného“ (pozícia jediného vysokoškolského učiteľa ho automaticky nominovala do funkcie vedúceho výpravy). Mimochodom, pomer žien a mužov, na prírodovedcov dosť netypický, je 4:2.

Účasť na konferencii STAKAN 2007 sme si naplánovali z viacerých (ako inak, dobrých) dôvodov. Vymenujeme aspoň niekoľko z nich:

1. téma konferencie (uznávame, že to asi nebol úplne ten najhlavnejší dôvod, ale uviesť ho ako prvý je minimálne slušnosť),
2. zloženie účastníkov (evokovalo príležitosť osobne sa „dotknúť“ žijúcich osobností českého a slovenského štatistického neba),
3. miesto konania konferencie (niekde na chate blízko Rusavy, pre nás miesto, ktoré by sa dalo označiť rýdzo slovensky ako Pánu Bohu za chrbtom, čo dráždilo našu zvedavosť),
4. skupinovú účasť (a dokonca bez prítomnosti školiteľov).

Aj napriek zodpovednému plánovaniu našej cesty (ktoré vedúci skupiny štandardne presunul na výkonných členov ako inak ženského pohlavia, ktoré túto náročnú úlohu primerane zodpovedne zvládli) sa pri cestovaní do Rusavy vyskytli neočakávané problémy. Napodiv nezačali už pri nastupovaní do vlaku v Banskej Bystrici, ale až na území nášho bratského národa. Výsledkom bolo, že sme na miesto určenia dorazili asi s dvojhodinovým meškaním, keďže naša trasa nadobudla charakter špirály. Našťastie organizátori sa nás hneď ujali a po vysvetlení nejasností týkajúcich sa vekovej kategorizácie (pohlavie bolo určené správne) nám prideliť ubytovanie. S nadšením sme „rozzipovali“ batožinu a pridali sa k poslucháčom už prebiehajúcich prednášok. Veľmi príjemne nás prekvapila srdečná a neformálna atmosféra prednášok. Diskusia bola otvorená všetkým vekovým skupinám a názorovým orientáciám. Blížil sa čas večere, tak sme čakali, čo bude. Ako už tušíte, chystáme sa pokračovať v udeľovaní superlatívov – jedlo bolo výborné! Večer nás svojou prednáškou uchvátil pán profesor Žváček a väčšinu účastníkov konferencie potom uchvátilo vystúpenie tanečnej skupiny Kirké. Samozrejme, debata (jednoznačne

vysoko odborná) našej šesťčlennej výpravy pokračovala až do skorých ranných hodín pred chatkou.

V sobotu (po pozornej účasti na skvelých doobedňajších prednáškach) sme sa postarali o všeobecné pobavenie účastníkov konferencie na výlete. Podarilo sa nám totiž na poslednej križovatke turistických chodníkov zvoliť úplne odlišný smer ako zvolili všetci ostatní, ale uistení správnou farebnou kombináciou značiek (aj keď v inom poradí) sme pokračovali v ceste. Keďže sme však nasledujúcich 30 minút namiesto stúpania ustavične šli smerom nadol, pochybnosti začali nahlodávať našu istotu. Nakoniec sme sa rozhodli ujasniť si naše ďalšie smerovanie u prvého človeka, ktorého stretneme. Po otázke „Kam sa dostaneme touto cestou?“ položenej miestnym chatárom nás odpoveď uistila, že ideme nesprávne. Uvažovali sme, čo spraviť... Nesmelé návrhy zmeniť cieľ výletu na najbližšiu obec boli v zárodku potlačené. Rozvážne sme sa rozhodli zavolať organizátorom (ktorí s celou skupinou boli už dávno v celi) a poslúchnuť ich pokyny. Tie boli jasné – vrátiť sa a pokračovať v ceste nahor! Tak sme sa vzchopili, a cestu k spomínanej križovatke sme zvládli za 20 minút! K vytúženému cieľu to bol potom už len kúsok. Keď sme zbadali členov našej výpravy (a oni nás), odľahlo nám (a na poznámku „Kam sa nám ztratila celá Banská Bystrica?“ sme reagovali už len so smiechom).

Po večernej sérii prednášok nás očarila prednáška trochu z iného súdka – o galaxiách, a keďže sa aj počasie umúdrilo, mohli sme sledovať nočnú oblohu hviezdárskym ďalekohľadom.

V nedeľu sme už (aj keď nie veľmi radi) museli pomýšľať na odchod. Konferencia bola ukončená vynikajúcim obedom, po ktorom sme sa skvele organizovanou dopravou dostali do Rusavy. Cesta domov už bola z hľadiska dopravy menej komplikovaná (ale rovnako veselá) ako cesta do Rusavy, pričom sa nám na trase Rusava – Banská Bystrica podarilo dosiahnuť úctyhodnú priemernú rýchlosť 30 km/h.

Po návrate sme sa všetci zhodli na jednej veci. Nasledujúci STAKAN si nesmieme nechať ujsť. Táto konferencia pre nás totiž predstavuje vzor toho, ako má konferencia vyzerieť. Skvelá organizácia. Priateľská a tvorivá atmosféra. Zaujímavé, originálne a podnetné prednášky. A nie nezanedbateľná vec, pomerne nízke náklady.

Tento náš článok prosím chápťe najmä ako VELKÉ POĎAKOVANIE organizátorom konferencie za skvele zvládnutý priebeh konferencie a srdečnú atmosféru, ktorá v nás zanechala nádherné spomienky.

Kolektív autorov z Banskej Bystrice

P.S. Prípadné faktografické nepresnosti v článku sú dôsledkom extrémneho množstva pozitívnych spomienok, ktoré sa s odstupom času navzájom trochu môžu miešať.

STAKAN 2007
PŘÍSPĚVKY

STATISTIKA NA FAKULTĚ MANAGEMENTU VŠE

STATISTICS AT THE UNIVERSITY OF ECONOMICS (CZECH REPUBLIC), THE FACULTY OF MANAGEMENT

Jitka Bartošová

Abstrakt: Statistika je jedním ze základních stavebních kamenů ekonomických teorií v oblasti mikroekonomie i makroekonomie. Z tohoto faktu vyplývá nezbytnost realizace výuky tohoto předmětu v návaznosti na výuku ekonomických předmětů. Při přípravě výuky statistických předmětů i při samotném vyučovacím procesu jsou také využívány statistické a matematické programy. Tento příspěvek se zabývá náplní výuky statistických a dalších navazujících předmětů na Fakultě managementu v Jindřichově Hradci a aplikacemi vhodných a snadno dostupných programů při výuce.

Klíčová slova: interdisciplinární předmět, seminární práce, statistika, výuka.

Abstract: Statistics forms one of the basic grounds of economic theories in micro- and macroeconomic area. This fact makes it necessary to link teaching of this subject matter to economic courses. For professor's preparation for statistic courses and for teaching itself, also statistical and mathematical programs are used. This contribution concerns the content of courses on statistics and related subjects in the Faculty of Management in Jindřichův Hradec (Czech Republic) and application of suitable and easily available programs for the purpose of teaching.

Key words: education, multispecialty subject, seminary work, statistics.

1. Úvod

Před dvěma roky (tj. v akad. roce 2005/06) proběhla na Fakultě managementu VŠE v Jindřichově Hradci transformace výuky podle požadavků ECTS. Důsledkem tohoto přechodu byla změna jak ve skladbě a náplni předmětů, tak i ve způsobu výuky a testování nabytých znalostí. Transformační proces vnesl významné změny do výuky všech předmětů, a to jak na bakalářském, tak i na magisterském stupni studia, tedy i do matematických, statistických

a informatických předmětů, jejichž výuku na fakultě zajišťují pracovníci katedry managementu informací (KMI).

Cílem výuky výše uvedených předmětů na ekonomických fakultách je vybavit studenty vhodným arzenálem prostředků potřebných pro kvantitativní vyjádření ekonomických zákonitostí a ověření jejich platnosti na reálných datech. Kvantitativní popis ekonomických vztahů by měl tvořit doplněk k popisu verbálnímu a grafickému, s nímž se studenti setkávají v ekonomických předmětech. Tento popis umožňuje využívat kvantitativní metody rozhodování, metody statického a dynamického modelování a prognózování. Důležitým předpokladem praktického využití kvantitativních metod je rovněž schopnost studentů efektivně využívat nejmodernější prostředky IT při získávání a zpracovávání dostupných kvantitativních i kvalitativních informací.

2. Výuka matematických, statistických a informatických předmětů na FM VŠE

Tento náročný úkol – vybudování kvantitativního ekonomického myšlení – plní na FM předměty vyučované pracovníky katedry managementu informací. Jedná se především o předměty povinné, které zajišťují jednotný znalostní standard, ale svou nezastupitelnou roli zde hrají rovněž předměty povinné či volně volitelné, které jsou vnímány do jisté míry jako předměty „nadstandardní“ a které studentovi umožňují individuální volbu a profilování (podle svých potřeb, schopností a zájmů). V současné době je na fakultě k dispozici následující nabídka předmětů garantovaných a vyučovaných pracovníky KMI:

1. Matematické předměty

(a) povinné

- Matematika pro ekonomy
- Aplikovaná matematika

(b) volitelné

- Základy matematického myšlení

2. Statistické předměty

(a) povinné

- Analýza dat
- Statistika pro manažery
- Manažerské rozhodování

(b) volitelné

- Pravděpodobnostní modely
- Stochastické modely
- Základy ekonometrické analýzy
- Kvantitativní metody z operačního managementu
- Modelování z reálných dat

3. Informatické předměty

(a) povinné

- Informatika 1
- Informatika 2
- Manažerská informatika
- Management znalostí

(b) volitelné

- Informatika 3
- Geografické informační technologie
- Informační systémy veřejné sféry
- Technologie WWW
- Technologie informačních systémů
- Základy umělé inteligence
- Počítačový design a reklama
- Technologie WWW 2

4. Interdisciplinární předměty

(a) povinné

- Modelování v ekonomii

(b) volitelné

- Techniky ekonomického modelování na PC
- Modelování z reálných dat

Jak je patrné z uvedeného přehledu, na FM je nejbohatší nabídka předmětů informatických a statistických. Ze statistických předmětů studenti absolvují povinně na bakalářském stupni studia (ve 3. a 4. semestru) dva předměty – Analýzu dat a Statistiku pro manažery. Předmět **Analýza dat** je zaměřen na postupy potřebné pro prezentaci dat a jejich základní analýzy. Součástí

předmětu je též úvod do pravděpodobnostního a induktivního uvažování a seznámení se s vybraným statistickým softwarem. Na Fakultě managementu byl k tomuto účelu vybrán program $\mathbb{R}$, který patří mezi open source a student ho tedy bude mít k dispozici i po absolvování fakulty. Další předmět – **Statistika pro manažery** – je zaměřen na statistické postupy používané v analýze závislostí. Součástí předmětu je též aplikace těchto metod na reálná data. Podmínkou úspěšného zakončení je v obou uvedených kurzech získání dostatečného počtu bodů ze závěrečné písemné práce a vypracování a včasné odevzdání tří domácích úkolů vypracovávaných ve skupinách o 1 – 3 studentech. (Podrobná náplň je obsažena např. v Bartošová, 2006, Komárková, Komárek, Bína, 2006, Komárek, Komárková, 2006.)

Třetí povinný statistický předmět – **Manažerské rozhodování** – je zařazen do magisterského stupně studia. Studenti se v něm seznámí se základními pojmy a poznatky manažerského rozhodování, rozhodovacími procesy a jejich strukturou, racionálními postupy řešení rozhodovacích problémů, základními metodami rozhodování za jistoty, rizika a nejistoty, managementem rizika a volbou úspěšného stylu rozhodování. Získávají také informace o významu znalostních systémů pro manažerské rozhodování, základech učících se přístupů k rozhodování a metodách výběru nejdůležitějších informací pro rozhodování na základě učení (pomocí dobývání znalostí z dat).

3. Interdisciplinární předměty

Na FM VŠE jsou do výuky zařazeny rovněž předměty, které využívají znalosti získané v matematických, statistických a informatických předmětech – interdisciplinární předměty. Jedním z těchto předmětů je povinně volitelný předmět **Techniky ekonomického modelování na PC**, který je zařazen do 4. semestru. Tento předmět v sobě propojuje teoretické znalosti získané studiem mikroekonomie a makroekonomie s nástroji pro jejich kvantitativní rozbor (viz Bartošová, 2007, Stankovičová, 2006, Vlčková, 2006). Důraz je kladen na praktické využití nabytých znalostí při řešení konkrétních ekonomických problémů na PC.

Součástí úspěšného zakončení kurzu je vypracování kvalitní **seminární práce** a její prezentace na veřejnosti. Úkolem studenta je prokázat schopnost vyhledat zajímavý aktuální problém ze zadané teoretické oblasti, získat potřebné informace z dostupných datových zdrojů a za pomoci nabytých znalostí a vhodných softwarových prostředků tento problém vyřešit. V práci je nezbytné propojit teoretické a praktické znalosti z ekonomie, matematiky, statistiky a informatiky. Tento předmět si studenti FM mohli vybrat v letošním akad. roce poprvé. Je zajímavé, že přestože se jedná o předmět značně

náročný, studenti se s požadavky na jeho zakončení (získání dostatečného počtu bodů ze znalostního testu a vypracování a prezentace seminární práce) dokázali vypořádat velmi dobře. Při vypracování semestrální práce nebylo největším problémem vyhledání zajímavého tématu ani jeho kvantitativní zpracování na počítači, ale nalezení potřebných datových souborů. Získat veřejně dostupná kvalitní data, která by poskytovala dostatek informací, postihovala delší časový úsek a byla srovnatelná, není snadné. Požadavky, které byly kladeny na studenty v tomto interdisciplinárním předmětu, si můžeme demonstrovat na konkrétní ukázce seminární práce, která byla vypracována v rámci předmětu Techniky ekonomického modelování na PC.

3.1. Ukázka seminární práce z předmětu Techniky ekonomického modelování na PC

Struktura seminární práce i požadavky na její formální provedení jsou předem dané. Každá seminární práce se skládá ze dvou částí – z teoretické a praktické, dále pak ze seznamu použitých zdrojů (knižních a internetových) a příloh, které obsahují datové soubory a úplné znění motivačního článku. V **teoretické části** je vybraný problém kvantitativně i kvalitativně charakterizován z hlediska ekonomické teorie a jsou zde popsány kvantitativní metody a softwarové produkty, které budou při jeho řešení použity. **Praktická část** obsahuje motivaci, která vedla k výběru dané problematiky, postup řešení problému na PC, výsledky a závěry (tj. interpretaci získaných výsledků).

Témata, která byla studenty zpracovávána, lze rozdělit do tří skupin – na témata zaměřená na modelování situace ve výrobě, ve spotřebě a na trhu. Na ukázkou zde byla vybrána seminární práce, která se zaměřila na problematiku trhu práce. Práce byla motivována článkem „Nobelova cena 2006 – Nobelův výbor s cenou za ekonomii zaspal dobu“, který byl publikován v Hospodářských novinách dne 10. 10. 2006. Autor se v článku vyjadřuje k udělení Nobelovy ceny za ekonomii, kterou získal ekonom Edmund S. Phelps za „analýzu intertemporálních substitučních vztahů v makroekonomické politice“. Ve své práci Phelps zkoumal, zda i v makroekonomii platí známé rčení „něco za něco“. Konkrétně zjišťoval, zda je možné, aby tvůrci hospodářské politiky dokázali „vyměnit“ úroveň některých veličin, pokud by shledali, že je to výhodné – tj. zda je např. možné vyměnit nižší nezaměstnanost za vyšší inflaci a naopak. Tím se autor článku dostává k Phillipsově křivce a popisuje její vývoj s tím, že Nobelova cena měla být udělena dříve, protože dnes už je ekonomická teorie trochu dál a je zřejmé, že Phillipsova křivka současný vývoj ekonomiky vždy dobře nevystihuje.

K ověření platnosti tohoto intertemporálního substitučního vztahu v podmínkách současné české ekonomiky bylo v práci použito několik modelů Phillipsovy křivky:

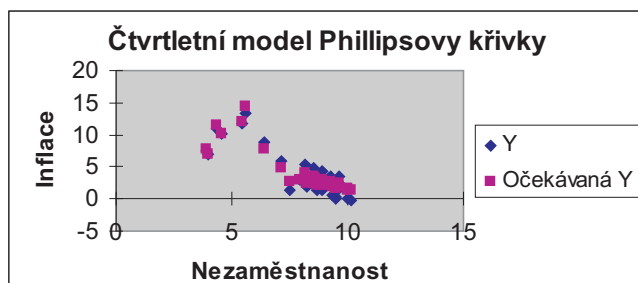
$$\hat{\pi}_t = 5,910 - 0,442U_t + 0,903(\pi_t^e)^2,$$

kde $\hat{\pi}_t$ je odhad skutečné inflace v %
 U_t je úroveň nezaměstnanosti v %
 π_t^e je očekávané inflace v %

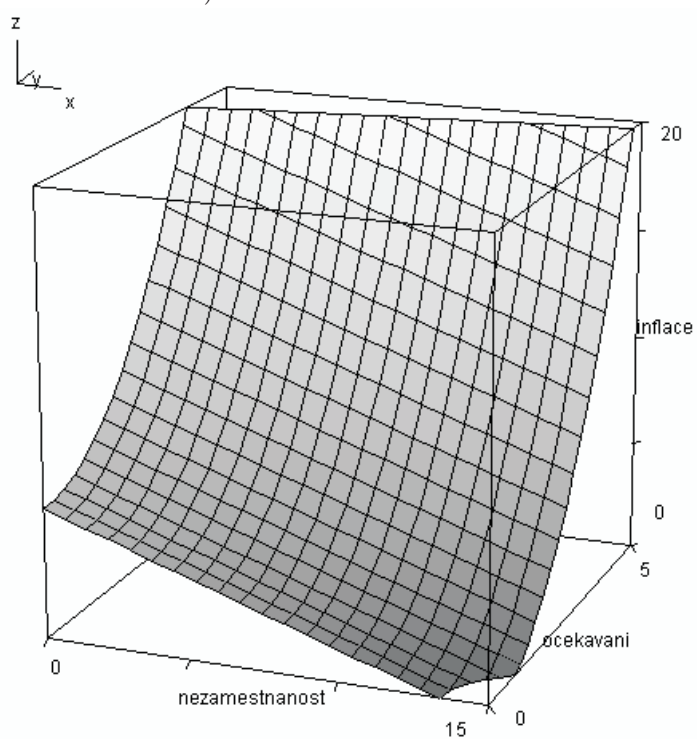
Tento model (zobrazený v grafech č. 1 a 2) vystihuje 93 % celkové variability a je statisticky významný, stejně jako všechny jeho parametry.

Ačkoliv příslušný neoklasický model je rovněž statisticky významný a vysvětluje závislost dokonce z 96 %, ukázalo se, že u této podoby Phillipsovy křivky je vývoj inflace závislý pouze na očekávané inflaci, neboť ostatní dva odhadnuté regresní parametry nejsou statisticky významné (viz obrázek č. 1). Naproti tomu v upraveném čtvrtletním modelu je vývoj inflace významně ovlivněn jak očekávanou inflací, tak i nezaměstnaností. Ovšem i v tomto modelu mají očekávání na vývoj skutečné inflace v ČR v období 1997–2006 mnohem výraznější vliv, neboť ji ovlivňují kvadraticky a oproti nezaměstnanosti mají prakticky dvojnásobnou hodnotu regresního koeficientu.

Obrázek 1: Závislost skutečné inflace na nezaměstnanosti ve čtvrtletním modelu krátkodobé Phillipsovy křivky speciálně upraveném pro českou ekonomiku (Výstup z programu MS Excel)



Obrázek 2: Graf odhadnutého modelu čtvrtletní krátkodobé Phillipsovy křivky speciálně upraveném pro českou ekonomiku (Výstup z programu Derive5)



Regresní statistika			
Násobné R	0,97983512		
Hodnota spolehlivosti R	0,96007686		
Nastavená hodnota spolehlivosti R	0,95939441		
Chyba stř. hodnoty	0,66939053		
Pozorování	120		
	<i>Koeficienty</i>	<i>t stat</i>	<i>Hodnota P</i>
Hranice	0,947938896	1,510686471	0,133564470
Soubor X 1	-0,084581161	-1,332168719	0,185393806
Soubor X 2	0,947050148	29,677060190	2,75540E-56
ANOVA			
	<i>Rozdíl</i>	<i>SS</i>	<i>MS</i>
Regrese	2	1260,74213	630,37106300
Rezidua	117	52,42579	0,44808368
Celkem	119	313,16792	
			<i>Významnost F</i>
			1,48473E-82

Tabulka 1: Výsledky odhadu parametrů neoklasického modelu Phillipsovy křivky
(Výstup z programu MS Excel)

V ukázkové seminární práci byly použity dva knižní a tři internetové informační zdroje. Volba nejvhodnějšího modelu a odhad jeho parametrů byly provedeny v programu MS Excel (obrázek č. 1), pro grafická zobrazení odhadnutých modelů byl použit kromě Excelu (graf č. 1) také jednoduchý matematický program Derive5, který umožňuje tvorbu 3D grafů (graf č. 2).

4. Závěr

Můžeme konstatovat, že nabídka statistických předmětů, vyučovaných na Fakultě managementu v Jindřichově Hradci pracovníky katedry managementu informací, je široká a zahrnuje v sobě všechny základní matematické, statistické a informatické metody potřebné pro kvantitativní zpracování ekonomických dat i ověření platnosti ekonomických teorií.

Ze softwarových produktů, které jsou při výuce těchto předmětů na FM využívány, je dávana přednost široce rozšířeným a snadno dostupným produktům, jako jsou produkty firmy Microsoft, dále pak jednoduchý a levný matematický program Derive a volně šiřitelný nástroj $\mathbb{R}$, který je vhodný pro provádění statistických výpočtů a tvorbu grafických výstupů. Volba výukového softwaru vychází z požadavku, aby student mohl s tímto produktem pracovat i po absolvování fakulty. Tím se přístup na naší fakultě odlišuje od přístupu na Fakultě informatiky a statistiky VŠE, kde kromě programu MS Excel (viz Marek, 2006) využívají při výuce rovněž drahé softwarové produkty na profesionální úrovni, jako je SAS (viz Jarošová, Marek, Pecáková, Pourová, Vrabec, 2005) nebo SPSS (viz Řezanková, 2005).

Na závěr příspěvek poukazuje na přednosti zařazování interdisciplinárních předmětů do učebních plánů fakult s ekonomickým zaměřením, a to jak na bakalářském, tak i na magisterském stupni studia. Jejich nespornou výhodou je to, že umožňují těsné propojení matematických, statistických a informatických znalostí s ekonomickými teoriemi a praxí.

Reference

- [1] Bartošová, J. 2007. *6MI420 Modelování v ekonomii (Podpůrný učební text k on-line kurzu)*. Oeconomica, Praha, 255 s. ISBN 978-80-245-1162-7.
- [2] Bartošová, J. 2006. *Základy statistiky pro manažery*. Oeconomica, Praha, 198 s. ISBN 80-245-1019-7.
- [3] Hušek R., Pelikán J. 2003. *Aplikovaná ekonometrie*. Professional Publishing, ISBN 80-86419-29-0.

- [4] Jarošová, E., Marek, L., Pecáková, I., Pourová, Z., Vrabec, M. 2005. *Statistika pro ekonomy – Aplikace*. 1. vyd. Professional Publishing, Praha, 423 s. ISBN 80-86419-68-1.
- [5] Komárková L., Komárek A., Bína V. 2006. *Základy analýzy dat a statistického úsudku, s příklady v R*. Elektronický výukový materiál.
- [6] Komárek A., Komárková L. 2006. *Statistická analýza závislostí, s příklady v R*. Elektronický výukový materiál.
- [7] Marek, L. 2006. Pravděpodobnostní rozdělení v MS Excel. *Statistika, 86(6)*, s. 497-522. ISSN 0322-788X.
- [8] Řezanková, H. 2005. Testy pro alternativní proměnné ve statistických programových systémech. *Forum Statisticum Slovaca* 1(2), s. 114-118. ISSN 1336-7420.
- [9] Stankovičová, I. 2006. Základné princípy stratégie modelovania. *Forum Statisticum Slovaca* 2(1), s. 78-81. ISSN 1336-7420.
- [10] Vlčková, V. 2006. Analýza dat. Strategický marketingový management. *CIMA, Praha, 12.1-12.20*, Výukové materiály v rámci projektu CZ.04.3.07/3.2.01.2/2145. ISBN 80-239-8387-3.

Adresa: RNDr. Jitka Bartošová, Ph.D.
 VŠE Praha, Fakulta managementu
 Jarošovská 1117/II
 377 01 Jindřichův Hradec
 E-mail: bartosov@fm.vse.cz

DYNAMICKÁ VERSUS KLASICKÁ SIMULAČNÁ METÓDA MONTE CARLO ¹

Mária Bohdalová

Abstrakt: V príspevku popisujeme známu simulačnú metódu Monte Carlo a jej zovšeobecnenie, ktoré využíva dynamické modelovanie vstupných údajov a kopula funkcie pre popis vzájomných vzťahov medzi vstupnými údajmi. Metódy budeme prezentovať na modelovaní rizika finančného portfólia.

1. Úvod

Stochastická simulačná metóda (alebo Monte Carlo simulácia)^{2,3}, nachádza svoje uplatnenie aj v meraní finančných rizík pomocou hodnoty VaR⁴. Metódou Monte Carlo je možné simulovať scenáre pre zmeny rizikových faktorov⁵ z vhodného rozdelenia viazané ku konkrétnemu dátumu. Pre odhad hodnoty VaR používa veľký počet simulácií vývoja hodnoty portfólia⁶. Tie sú určené veľkým počtom náhodne generovaných rizikových faktorov, ktorých rozdelenia sú známe. Scenáre je možné generovať buď náhodným spôsobom (známa Monte Carlo metóda) alebo iným, systematickejšim spôsobom. Simulácie môžu generovať vysoko pravdepodobné odhady hodnoty VaR. Ako vstupné údaje pre simulácie môžu slúžiť najnovšie informácie a historické údaje o rizikových faktorech. Metódou Monte Carlo môžeme testovať napríklad jednodenné zmeny hodnoty portfólia na základe veľkého počtu náhodne zvolených kombinácií rôznych situácií rizikových faktorov. Následne stanovíme jednodňovú stratu s pravdepodobnosťou napr. 1%, ktorá sa rovná jed-

¹Tento článok je podporený grantami č. VEGA-1/3014/06, VEGA-1/4024/07 a APVV-0375-06.

²JÍLEK, J.: *Finanční rizika*. Praha, Grada Publishing, 2000, s. 424-425.

³CROUHY, M. – GALAI, D. – MARK, R.: *Risk Management*. New York, McGraw-Hill Companies, 2001, s. 198.

⁴Hodnota VaR (Value at Risk) – hodnota v riziku. Je to potenciálna strata, ktorá môže nastať s určitou pravdepodobnosťou v priebehu nasledujúceho obdobia držania, stanovená na základe určitého historického obdobia, pre dané portfólio pri nepriaznivých trhových zmenách. (JÍLEK, J.: *Finanční rizika*. Praha, GRADA Publishing, 2000, s. 603.)

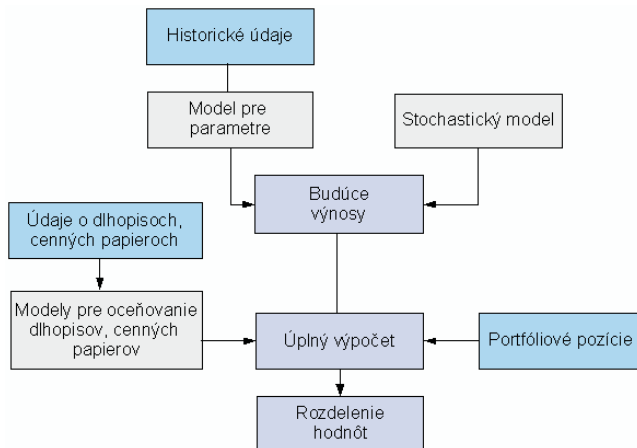
⁵Rizikový faktor predstavuje premennú, ktorej hodnota ovplyvňuje hodnotu jednotlivých nástrojov portfólia. Trhovými faktormi môžu byť úrokové miery, menové kurzy, ceny akcií a komodít.

⁶Portfóliom rozumieme istý súbor finančných nástrojov (aktív) v držbe individuálneho investora – buď fyzickej alebo právnickej osoby, určený nielen zložkami, ktoré zahŕňa ale aj ich objemom. (SKŘIVÁNKOVÁ, V. – SKŘIVÁNEK, J.: *Kvantitativne metody finančních operací*. Bratislava, IURA Edition, 2006, s. 76.)

nodňovej hodnote v riziku. Metóda Monte Carlo je flexibilná a preto je ju možné použiť pre modelovanie rôznych druhov rizík, na výpočet expozitúry, na výpočet trhového rizika vrátane určenia nelineárneho cenového rizika, rizika volatility a pod. Môžeme ich použiť aj pri modelovaní VaR pre dlhšie časové horizonty, čo je dôležité pri meraní kreditného rizika a tiež ich môžeme použiť pri meraní operačného rizika⁷.

Metódou Monte Carlo modelujeme stochastické procesy – procesy zahrňajúce ľudský výber, či procesy s neúplnými informáciami. Metóda Monte Carlo je teoreticky najlepšou metódou pre výpočet hodnoty VaR. Jej nevýhodou je, že je časovo, respektíve výpočtovo náročnejšia v porovnaní s inými metódami na výpočet hodnoty VaR⁸. Naviac, v porovnaní s inými metódami, vyžaduje najviac skúseností a profesionality od jej tvorcov⁹. Kroky simulačnej metódy Monte Carlo vidíme na obrázku 1¹⁰:

Obrázek 1: Kroky simulačnej metódy Monte Carlo



⁷JORION, P.: *Value at Risk: The Benchmark for Controlling Market Risk*. Blacklick, OH, USA: McGraw-Hill Professional Book Group, 2000. s. 291.

⁸Napríklad s kovariančno-variančnou metódou, metódou historickej simulácie a pod.

⁹DELIANEDIS, G. – LAGNADO, R. – TIKHONOV, S.: *Monte Carlo Simulation of Non-normal Processes*, working paper, London: Midas-Kapiti Intl., 2000, s. 3.

¹⁰JORION, P.: *Value at Risk: The Benchmark for Controlling Market Risk*. Blacklick, OH, USA. McGraw-Hill Professional Book Group, 2000, s.225.

2. Základné kroky simulačnej metódy Monte Carlo pre určenie VaR

Označme hodnotu portfólia v aktuálnom čase t symbolom V_t . Predpokladajme, že hodnota V_t závisí od n rizikových faktorov, z ktorých uvedieme napríklad úrokovú sadzbu (interest rate), výmenný kurz (foreign exchange – FX), ceny akcií (share prices) a podobne.

Výpočet VaR metódou Monte Carlo pozostáva z nasledujúcich krokov:

Krok 1: Zvolíme hladinu významnosti α , na ktorú sa bude hodnota VaR odvolávať.

Krok 2: Simulujeme vývoj rizikových faktorov od aktuálneho času t po čas $t + 1$ tým, že vygenerujeme n -ticu pseudonáhodných¹¹ čísel pre odpovedajúce marginálne rozdelenie a odpovedajúce združené rozdelenie, ktoré zodpovedá správaniu sa rizikových faktorov. Počet m týchto n -tíc je obvyčajne veľké číslo, rádovo tisíc ($m = O(1000)$).

Krok 3: Vypočítame m rôznych hodnôt portfólia v čase $t + 1$ použitím simulovaných n -tíc hodnôt rizikových faktorov. Tieto hodnoty označíme $V_{t+1,1}, V_{t+1,2}, \dots, V_{t+1,m}$.

Krok 4: Vypočítame simulované výnosy/straty, tzn. rozdiely medzi simulovanými budúcimi hodnotami portfólia a aktuálnymi hodnotami portfólia, $\Delta V_{t,i} = V_{t+1,i} - V_{t,i}$, pre $i = 1, \dots, m$.

Krok 5: Ignorujeme časť α najhorších zmien $\Delta V_{t,i}$. Minimum zo zvyšných $\Delta V_{t,i}$ je hodnota VaR portfólia v čase t . Označíme ho $\text{VaR}(\alpha, t, t + 1)$.

So zmenou času z t na $t + 1$, reálna hodnota (nezmenená) portfólia sa zmení z V_t na V_{t+1} . S týmito údajmi na druhej strane môžeme spätne testovať hodnotu $\text{VaR}(\alpha, t, t + 1)$ porovnaním s $\Delta V_{t,i}$.

Je zrejmé, že ťažisko uvedeného postupu spočíva v kroku dva, čiže v generovaní pseudonáhodných čísel v súlade s odpovedajúcimi rozdeleniami. Väčšina metód využívajúca princíp Monte Carlo pre výpočet VaR používa v tomto kroku postup, ktorý detailne popisujeme v nasledujúcej časti, a na ktorý sa budeme v ďalšom texte odvolávať ako na „tradičný“ postup.

¹¹O pseudonáhodných číslach hovoríme pretože ich generujeme pomocou algoritmu s vopred určenými pravidlami, pričom ak je počiatočná hodnota pre spustenie generovania rovnaká pre viac opakovaní, tak dostaneme rovnaké čísla.

3. Tradičná simulačná metóda Monte Carlo pre výpočet VaR

Úlohou druhého kroku výpočtu VaR metódou Monte Carlo je vygenerovať n -ticu pseudonáhodných čísel pre vhodné marginálne a vhodné združené rozdelenie, ktoré opisujú správanie sa rizikových faktorov (n je počet rizikových faktorov).

Ak vezmeme do úvahy len hodnoty výmenného kurzu a úrokovú mieru ako rizikové faktory, **tradičný postup** je nasledujúci [RAN02]:

Krok 1: Zhromaždíme historické údaje pre n rizikových faktorov, tzn. dostaneme n časových radov s rozsahom $N + 1$ obchodných dní. Tieto údaje označíme $x_{i,0}, x_{i,1}, \dots, x_{i,N}$, pre $i = 1, 2, \dots, n$, pričom dnešná hodnota je $x_{i,N}$. Obyčajne volíme $N + 1 = 250$, alebo viac.

Krok 2: Za predpokladu, že $x_{i,j} \neq 0$ vypočítame relatívne zmeny rizikových faktorov (výnosnosť rizikových faktorov)¹²:

$$r_{i,j} = \frac{x_{i,j} - x_{i,j-1}}{x_{i,j-1}}, \text{ prípadne } r_{i,j} = \ln \frac{x_{i,j}}{x_{i,j-1}}, \quad (1)$$

pre $i = 1, 2, \dots, n$ a $j = 2, \dots, N$

príčom hodnoty $r_{i,1}, r_{i,2}, \dots, r_{i,N}$ pre každé $i = 1, 2, \dots, n$ prislúchajú náhodnej premennej r_i .

Krok 3: V tomto kroku zavedieme predpoklad o marginálnych rozdeleniach $f_1, f_2, \dots, f_n$ pre náhodné premenné $r_{i,1}, r_{i,2}, \dots, r_{i,N}$. Obyčajne vo finančných aplikáciách predpokladáme, že jednotlivé marginálne rozdelenia f_i pochádzajú z normálneho rozdelenia $N(\mu_i, \sigma_i^2)$:

$$f_i(r_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(r_i - \mu_i)^2}{2\sigma_i^2}\right), \quad (2)$$

kde $i = 1, 2, \dots, n$, μ_i je priemer a σ_i^2 je rozptyl náhodnej premennej r_i , pričom

$$\mu_i = E(r_i) \text{ a } \sigma_i^2 = E((r_i - \mu_i)^2). \quad (3)$$

¹²CIPRA, T.: *Matematika cenných papírů*. Praha, Nakladatelství HZ Praha, 2000, s. 120.

Krok 4: Vo všeobecnosti sú parametre marginálnych rozdelení neznáme. Našou úlohou je určiť odhad týchto parametrov z historických údajov. V prípade normálneho rozdelenia je táto úloha jednoduchá¹³, pretože parametre normálneho rozdelenia sú očakávaná hodnota a rozptyl rozdelenia historických údajov. Ich odhady spočítame nasledovne:

$$\hat{\mu}_i = \frac{1}{N} \sum_{j=1}^N r_{i,j}, \quad (4)$$

$$\hat{\sigma}_i^2 = \frac{1}{N-1} \sum_{j=1}^N (r_{i,j} - \hat{\mu}_i)^2, \quad (5)$$

kde $i = 1, 2, \dots, n$.

Poznamenajme, že priemer a rozptyl ľubovoľného rozdelenia môže byť definovaný rovnakými vzťahmi ((4), (5)). Vyplýva to zo zákona veľkých čísel¹⁴. Vo všetkých prípadoch sú tieto odhady nevychýlené a konzistentné¹⁵.

Krok 5: V tomto kroku potrebujeme generovať n -tice pseudonáhodných čísel združeného rozdelenia. Preto zavedieme ďalšie predpoklady o tom, čo určuje závislostnú štruktúru medzi náhodnými premennými. Pretože marginálne rozdelenia už máme zvolené, stačí zvoliť (nie úplne ľubovoľne) združené rozdelenie. Nech $f(\vec{r})$ označuje príslušné združené rozdelenie. Je zrejmé, že nasledujúca podmienka je splnená:

$$f_i(r_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\vec{r}) dr_1 \dots dr_{i-1} dr_{i+1} \dots dr_n \quad (6)$$

pre každé $i = 1, 2, \dots, n$.

V tradičných Monte Carlo simuláciách predpokladáme, že združené rozdelenie je multinormálne rozdelenie, tzn. má tvar:

¹³ „Jednoduchá“ v tom zmysle, že odhady počítame priamo z historických údajov bez použitia akéhokoľvek numerického algoritmu (ide o tzv. bodové odhady).

¹⁴ LAMOŠ, F. – POTOCKÝ, R.: *Pravdepodobnosť a matematická štatistika.*, Bratislava, ALFA, 1989, s. 45.

¹⁵ F. – POTOCKÝ, R.: *Pravdepodobnosť a matematická štatistika.*, Bratislava, ALFA, 1989, s. 80-81.

$$f(\vec{r}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(\frac{-1}{2}(\vec{r} - \vec{\mu})^T \Sigma^{-1}(\vec{r} - \vec{\mu})\right), \quad (7)$$

kde $\vec{r} = (r_1 \dots r_n)^T$,

$$\vec{\mu} = (\mu_1 \dots \mu_n)^T \quad (8)$$

a Σ je variačno-kovariančná matica:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & c_{1,2} & c_{1,3} & \dots & c_{1,n} \\ c_{1,2} & \sigma_2^2 & c_{2,3} & \dots & c_{2,n} \\ c_{1,3} & c_{2,3} & \sigma_3^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & c_{n-1,n} \\ c_{1,n} & c_{2,n} & \dots & c_{n-1,n} & \sigma_n^2 \end{pmatrix} \quad (9)$$

kde σ_i^2 je rozptyl výnosností r_i ,
 $c_{i,j}$ sú kovariancie medzi výnosnosťami r_i a r_j :

$$c_{i,j} = E((r_i - \mu_i)(r_j - \mu_j)). \quad (10)$$

Pre úplnosť zavedieme korelačný koeficient $\rho_{i,j}$:

$$\rho_{i,j} = \frac{c_{i,j}}{\sigma_i \sigma_j}, \quad (11)$$

pre $\sigma_i, \sigma_j \neq 0, i = 1, 2, \dots, n$ a $j = 1, 2, \dots, N$.

Lahko sa dá ukázať, že funkcia f daná vzťahom (7) spĺňa podmienku (6).

Krok 6: V tomto kroku určíme kovariancie (11). Kovariancie môžeme odhadnúť podobne ako priemer a rozptyl z historických údajov:

$$\hat{c}_{i,j} = \frac{1}{N-1} \sum_{k=1}^N (r_{i,k} - \hat{\mu}_i)(r_{j,k} - \hat{\mu}_j) \quad (12)$$

pre $i = 1, 2, \dots, n$ a $j = 1, 2, \dots, N$.

Krok 7: Teraz máme určené marginálne rozdelenia (krok 3), združené rozdelenie (krok 5) a tiež máme odhadnuté aj ich parametre (kroky 4 a 6). n -tice pseudonáhodných čísel generujeme štandardnou procedúrou, ktorá využíva Choleského dekompozíciu matíc a nájdeme ju napríklad v prácach [MIK85, str. 80], [FIN96]. Predpokladajme teraz, že máme vygenerované potrebné n -tice hodnôt r_i . Označíme ich nasledovne:

$$\vec{r}^k = (r_1^k \dots r_n^k)^T, \quad (13)$$

kde $k = 1, \dots, m$ a m je počet Monte Carlo iterácií.

Krok 8: V predchádzajúcom kroku sme získali m nezávislých n -tíc náhodných čísel. Tieto čísla sa vzťahujú na relatívne zmeny v údajoch (výnosnosť) (pozri krok 2) a preto je na záver potrebné simulované rizikové faktory pretransformovať. Úpravami (1) a (13) získame m simulovaných hodnôt pre n rizikových faktorov v čase $N + 1$ pre aritmetickú alebo geometrickú (spojitú) výnosnosť:

$$x_{i,N+1}^k = x_{i,N}(1 + r_i^k), \text{ alebo } x_{i,N+1}^k = \exp(r_{i,j}^k) \cdot x_{i,j-1} \quad (14)$$

kde $i = 1, 2, \dots, n; j = 2, 3, \dots, n$ a $k = 1, \dots, m$.

Vyššie uvedený postup simuluje rozdelenie výnosností jednotlivých rizikových faktorov. O jednotlivých marginálnych rozdeleniach predpokladá, že pochádzajú z normálneho rozdelenia (pozri krok 3, vzťah (2)). O združenom rozdelení výnosností rizikových faktorov zasa predpokladá, že je multinormálne (pozri krok 5, vzťah (6)). Je nutné poznamenať, že ak sú marginálne rozdelenia normálne, z toho ešte nevyplýva, že združené rozdelenie musí byť multinormálne (dá sa to ukázať priamo zo Sklarovej vety¹⁶). Avšak platí, že ak je združené rozdelenie multinormálne, tak marginálne rozdelenia sú normálne. Toto je podstatné pre pochopenie prístupu pomocou kopúl. Naviac, Embrechts, McNeil a Straumann vo svojich prácach ([EMB99], [EMB02]) ukázali, že závislostná štruktúra určená koreláciami spôsobuje problémy a podobne aj multinormálne rozdelenie spôsobuje problémy. Ďalšie problémy spôsobuje to, že rizikové faktory majú tzv. „ťažké“ chvosty¹⁷ a generovanie

¹⁶NELSEN, R. B.: *An introduction to Copulas*. New York, Springer, 1999.

¹⁷„Ťažké“ chvosty rozdelenia znamenajú, že v praxi sa vyskytujú častejšie extrémne udalosti ako sa očakáva podľa normálneho rozdelenia.

údajov pomocou pseudonáhodných generátorov nie je vždy „najvhodnejšie“. Preto sa v modernej teórii manažmentu rizika hľadajú nové prístupy, ktoré tieto nedostatky odstraňujú. Jeden z nich uvádzame v nasledujúcej časti príspevku.

4. Zovšeobecnená metóda Monte Carlo

Zovšeobecnená metóda Monte Carlo dynamicky modeluje výnosnosti rizikových faktorov, pričom

- zachováva nenormalitu rozdelení výnosností rizikových faktorov,
- poskytuje prepojenie medzi výnosnosťami rizikových faktorov, ktoré môžu mať normálne ale i nenormálne rozdelenie,
- dynamicky kontroluje chyby každého modelu pre výnosnosť rizikového faktora¹⁸.

Dynamické modelovanie rizikových faktorov zrozumiteľne vysvetľuje manažerom aké veľké je riziko v portfóliu. Rizikové faktory umožňuje modelovať buď samostatne, alebo súčasne. Modelované rizikové faktory môžu, ale nemusia mať normálne rozdelenie, a ich závislostnú štruktúru popisuje kopula funkciami. Tento prístup je možné použiť pre odhad nielen trhového ale i kreditného rizika obchodných portfólií a tiež pre rôzne finančné nástroje¹⁹.

Zovšeobecnené riešenie:

Vyššie uvedené kroky tradičného algoritmu nahradíme nasledovnými²⁰:

Krok 3: Určíme vhodný stochastický model pre výnosnosť každého rizikového faktora (hranice združeného rozdelenia)²¹

$$v_i = f_i(\mathbf{x}, \mathbf{y}, \theta_i) + \varepsilon_i, \quad (15)$$

¹⁸SAS® RISK Dimensions®: *Dynamic Risk factor Modeling Methodology*. White Paper, dostupné na <http://www.riskadvisory.com/pdfs/sasriskdimensionsriskfactor.pdf>, navštívené dňa 20. 9. 2006.

¹⁹Finančný nástroj predstavuje samostatný komponent portfólia. Najpoužívanejšími nástrojmi sú aktíva: akcie, cenné papiere, dlhopisy (obligácie), menové pozície. Medzi nástroje patria aj deriváty (kontrakty, ktorých hodnota závisí na cene podkladového nástroja) ako *forwardy*, *futurity*, *opcie*, *swapy* a podobne. (JÍLEK, J.: *Finanční rizika*. Praha, GRADA Publishing, 2000, s. 602, s. 599.)

²⁰Pozri SAS online dokumentáciu <http://www.sas.com/>.

²¹V ekonometrickej literatúre poznáme dva druhy pozorovaných premenných: exogénne (nezávislé) premenné (obyčajne ich označujeme X) a endogénne (závislé) premenné (obyčajne ich označujeme Y).

kde $i = 1, 2, \dots, m_{en}$; m_{en} je počet endogénnych premenných,
 $\mathbf{x}$ sú exogénne premenné (rizikové faktory),
 $\mathbf{y}$ sú endogénne premenné,
 θ vektor odhadnutých parametrov modelu
 $\mathbf{a}$ je vektor chýb stochastického modelu pričom

$$\varepsilon_i \sim \mathbf{F}_i(\chi_i), \quad (16)$$

kde $F_i(\chi_i)$ je vopred špecifikované rozdelenie, určené vektorom parametrov χ_i .

Krok 4: Pre odhad vektora parametrov stochastického modelu θ použijeme známe štatistické metódy^{22,23} (napr. metódu najmenších štvorcov, zovšeobecnenú metódu momentov alebo metódu maximálnej vierohodnosti).

Krok 5: Odhadneme združené rozdelenie $F(\cdot)$ vektora chýb určeného kopulou $C : F(\varepsilon) = C(F_1^{-1}(\varepsilon_1), \dots, F_m^{-1}(\varepsilon_m))$. Jednotlivé marginálne rozdelenia rizikových faktorov sú určené endogénnou premennou y , ktorej parametre sme odhadli v predchádzajúcom kroku.

Krok 6: Ak má vektor chýb ε napríklad normálne rozdelenie, tak neznáme parametre rozdelenia sú μ_i a σ_i^2 (priemer a rozptyl). Odhady týchto parametrov $\hat{\mu}_i$ a $\hat{\sigma}_i^2$ môžeme spočítať pomocou vzťahov (4) a (5), ako sme sa zmienili vo štvrtom kroku pôvodného algoritmu.

Krok 7: Odhadneme korelačnú maticu $\sum$ vektorov chýb (pozri vzťah (12)). Vygenerujeme nezávislé pseudonáhodné (kvázináhodné) čísla u, w z normálneho rozdelenia $N(0, 1)$. Pomocou korelačnej matice $\sum$ ich pretransformujeme do korelovaných premenných a z inverznej Gaussovej kopuly vypočítame číslo $v (v = C_u^{-1}(w))$. Nakoniec z inverzných distribučných funkcií rizikových faktorov určíme dvojicu chýb $(e_1, e_2) : (e_1 = F_1^{-1}, e_2 = F_2^{-1}(v))$. Dvojica chýb (e_1, e_2) je vygenerovaná pomocou pseudonáhodných (kvázináhodných) čísel a má želanú závislostnú štruktúru určenú kopulami.

²²WONNACOTT, T. H. – WONNACOTT, R. J.: *Statistika pro obchod a hospodářství*. New York, Victoria Publishing, 1992, s. 385-505, 611-630.

²³LAMOŠ, F. – POTOCKÝ, R.: *Pravdepodobnosť a matematická štatistika.*, Bratislava, ALFA, 1989, s.79-107, s. 177-200.

Teraz poznáme chyby, poznáme model a spätne určíme odpovedajúce výnosnosti rizikových faktorov, tzn. pokračujeme krokom 8 tradičnej metódy Monte Carlo.

Na to, aby sme získali predstavu o vygenerovanej združenej funkcii a o tom ako vplýva na hodnotu portfólia zopakujeme dostatočne veľa krát siedmy krok Monte Carlo simulácie (napr. použijeme od 1000 do 10000 opakovaní). Výsledné hodnoty portfólia usporiadame podľa veľkosti a určíme príslušný kvantil (hodnotu VaR).

Efektívnu Monte Carlo simuláciu získame, keď dimenzia modelovaného priestoru (určená počtom rizikových faktorov) bude najmenšia možná. Nakoľko dimenzia modelovaného priestoru priamo vplýva na počet opakovaní krokov 3-6, je vhodné za účelom zníženia dimenzie modelovaného priestoru najskôr použiť metódu hlavných komponentov (PCA metódu)^{24,25}, a až potom pristúpiť ku generovaniu možných scenárov.

Tiež je dôležité, aby zvolené stochastické modely modelovali náhodné procesy určené historickými údajmi jednotlivých rizikových faktorov čo najpresnejšie. Presnosť modelu určuje, ako dobre opisuje viacrozmerná pravdepodobnostná funkcia hustoty aktuálne pravdepodobnosti budúcich udalostí (a tým aj správnosť odhadu hodnoty VaR). Porovnanie oboch tu uvádzaných metód nájdeme v Tabuľka 1.

5. Záver

Pre porovnanie oboch simulačných prístupov sme vytvorili portfólio obsahujúce 10 britských štátnych pokladničných poukážok s nulovým kupónom pri úrokovej sadzbe LIBORu 5,375 %, pričom každá poukážka má nominálnu hodnotu 100 000 GBP a splatnosť 1 mesiac (tzn. sú splatné k 1. 5. 2006). Portfólio chceme držať v slovenských korunách. Rizikové faktory vplývajúce na toto portfólio sú výmenný kurz britskej libry voči slovenskej korune²⁶ a úroková sadzba LIBOR²⁷. Časové rady analyzovaných mien a úrokových sadzieb zahŕňajú obdobie od 2. 1. 2004 do 31. 3. 2006, čo predstavuje 580 obchodných dní pre každý časový rad. Porovnanie odhadov 1-dňových 99%

²⁴LAMOŠ, F. – POTOCKÝ, R.: *Pravdepodobnosť a matematická štatistika.*, Bratislava, ALFA, 1989, s. 260-268.

²⁵BOHDALOVÁ, M. – STANKOVIČOVÁ, I.: *Using the PCA in the Analyse of the risk Factors of the investment Portfolio.* In: Forum Statisticum Slovakum, 3/2006, s. 41-52, ISSN 1336-7420.

²⁶Údaje o dennom výmennom kurze sú čerpané z <http://www.nbs.sk/>, navštívené dňa 30. 4. 2006

²⁷Údaje o dennej úrokovej sadzbe sú čerpané z <http://www.bba.org.uk/bba/jsp>, navštívené dňa 30. 4. 2006.

Tradičná Monte Carlo simulačná metóda	Zovšeobecnená Monte Carlo simulačná metóda
Každý rizikový faktor má normálne (lognormálne) rozdelenie.	Pre každý rizikový faktor vieme určiť vhodný, známy model.
Závislostná štruktúra medzi rizikovými faktormi je určená var.-kovar. maticou za predpokladu multinormálneho rozdelenia.	Závislostná štruktúra medzi chybami modelov rizikových faktorov je určená Gaussovou kopulou.
Vyžaduje numerický odhad korelačných koeficientov a parametrov normálnych rozdelení odpovedajúcich jednotlivým výnosnostiam rizikových faktorov.	Vyžaduje numerický odhad parametrov modelov výnosností rizikových faktorov, korelačných koeficientov a parametrov Gaussovej kopule pre rozdelenie vektorov chýb odpovedajúcim modelom.
Scenáre generuje pomocou pseudonáhodných čísel a korelačných koeficientov.	Scenáre generuje pomocou pseudonáhodných alebo kvázináhodných čísel a kopúl.

Tabuľka 1: Porovnanie tradičnej a zovšeobecnenej Monte Carlo simulačnej metódy

hodnôt VaR získaných tradičnou metódou Monte Carlo a zovšeobecnou metódou Monte Carlo uvádzame v Tabuľka 2. Pre obe metódy sme použili generovanie scenárov buď pseudonáhodnými alebo kvázináhodnými²⁸ (Faurovými, Sobolovými) postupnosťami čísel.

Hodnoty VaR získané tradičnou metódou Monte Carlo sú v porovnaní so zovšeobecnou metódou Monte Carlo oveľa vyššie, čo v konečnom dôsledku núti investorov vyhradiť oveľa vyššie rezervy pre zabezpečenie portfólia. Tieto peniaze neprinášajú zisky, čo môže viesť investorov ku krachu. Dôvod prečo vznikli tieto rozdiely spočíva v tom, že tradičná metóda Monte Carlo predpokladá, že rizikové faktory majú normálne rozdelenie, čo pre dané rizikové faktory nebolo splnené. Pre zovšeobecnú metódu Monte Carlo sme použili konštantný (mean) model pre popis výnosov rizikových faktorov.

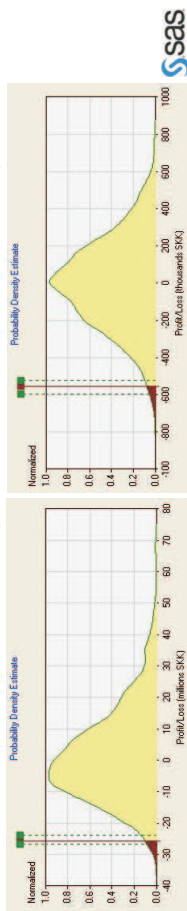
Rozdelenie výnosov a strát získaných oboma metódami vidíme na grafoch Graf 1-Graf 3. Všetky výstupy sú získané pomocou softvérového riešenia systému SAS® RISK Dimensions®.

²⁸Generátor kvázináhodných čísel generuje čísla, ktoré nemajú náhodnú zložku. Používanie generátora kvázináhodných čísel môže viesť k zvýšeniu výkonnosti Monte Carlo simulácií v tom zmysle, že sa zníži čas a/alebo sa zvýši presnosť výpočtu.

Methods	Mark to Market Value (SKK)	At-Risk Value (SKK)	At-Risk Value as percent of Base Value	Lower Tolerance Limit of At-Risk Value (SKK)	Upper Tolerance Limit of At-Risk Value (SKK)	Lower Tol Limit of VaR as percent of Base	Upper Tol Limit of VaR as percent of Base
MMC-Vš-SkN	53993907,96	553048,05	1,02	481172,83	666391,24	0,89	1,23
MMC-Vš-FkN	53993907,96	583371,13	1,08	552263,85	612294,63	1,02	1,13
MMC-VK-FkN	53993907,96	24755464,62	45,85	22975382,26	27097004,14	42,55	50,19
MMC-Vš-PsN	53993907,96	583261,93	1,08	519808,82	639833,05	0,96	1,19
MMC-Vš-SkN	53993907,96	553048,05	1,02	481172,83	666391,24	0,89	1,23
MMC-Vš-FkN	53993907,96	583371,13	1,08	552263,85	612294,63	1,02	1,13

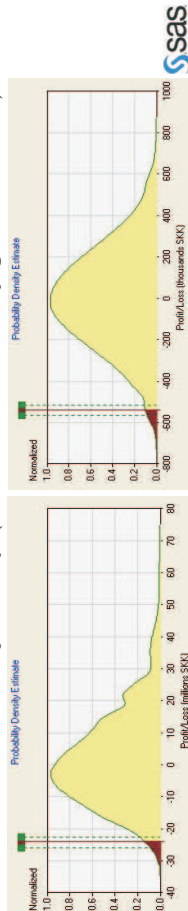
Tabulka 2: Prehľad odhadnutých 1-dňových 99 % hodnôt VaR jednotlivými metódami pre portfólio držané v SKK

Obrázek 2: VaR vypočítané pomocí Monte Carlo tradiční a zrychlené metody (pseudonáhodný generátor čísel)



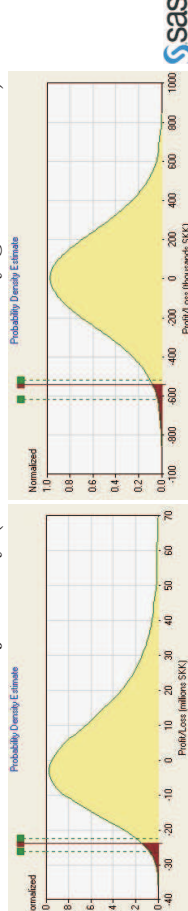
sas

Obrázek 3: VaR vypočítané pomocí Monte Carlo tradiční a zrychlené metody (Faurov kvázináhodný generátor čísel)



sas

Obrázek 4: VaR vypočítané pomocí Monte Carlo tradiční a zrychlené metody (Sobolov kvázináhodný generátor čísel)



sas

V príspevku sme uviedli zovšeobecnenie známej metódy Monte Carlo, ktorá je vhodná na popísanie rizikových faktorov s tzv. „ťažkými“ chvostami, pričom sa zachová závislostná štruktúra medzi rizikovými faktormi.

Reference

- [BOHN06] BOHDALOVÁ, M. – NÁNÁSIOVÁ, O.: *A Note to Copula Functions*. In: E-leader Bratislava, 11. – 15. jun 2006, <http://www.g-casa.com/BratislavaSlovakia.php>.
- [BOHS06] BOHDALOVÁ, M. – STANKOVIČOVÁ, I.: *Using the PCA in the Analyse of the risk Factors of the investment Portfolio*. In: Forum Statisticum Slovakum, 3/2006, s. 41-52, ISSN 1336-7420.
- [CIP00] CIPRA, T.: *Matematika cenných papírů*. Praha, Nakladatelství HZ Praha, 2000, 241 s., ISBN 80-86009-35-1.
- [DEL00] DELIANEDIS, G. – LAGNADO, R. – TIKHONOV, S.: *Monte Carlo Simulation of Non-normal Processes*, working paper, London: Midas-Kapiti Intl., 2000.
- [EMB99] EMBRECHTS, P. – DE HAAN, L. – HUANG, X.: *Modeling Multivariate Extremes*. Working paper. <http://www.math.ethz.ch/~baltes/ftp/papers.html> 1999, navštívené 27. 5. 2006.
- [EMB01] EMBRECHTS, P. – LINDSKOG, F. – McNEIL, A. J.: *Modeling Dependence with Copulas and Applications to Risk Management*. Zürich, 2001, <http://www.math.ethz.ch/finance>, navštívené 27. 5. 2006.
- [EMB02] EMBRECHTS, P. – McNEIL, A. J. – STRAUMANN, D.: *Correlation and Dependence in Risk Management: Properties and Pitfalls*. In: Dempster, M. A.H.: *Risk management: Value at Risk and Beyond*. West Nyack, NY, USA: Cambridge University Press, 2002, s. 176-223.
- [FIN96] FINGER, CH. C.: *Monte Carlo Simulation*. Morgan Guaranty Trust Company. Risk management Research, 1996.
- [JIL00] JÍLEK, J.: *Finanční rizika*. Praha, GRADA Publishing, 2000, 635 s., ISBN 80-7169-579-3.

- [JOR00] JORION, P.: *Value at Risk: The Benchmark for Controlling Market Risk*. Blacklick, OH, USA: McGraw-Hill Professional Book Group, 2000, 535 s., ISBN 0-07-137921-5.
- [KOM98] KOMORNÍK, J. – KOMORNÍKOVÁ, M. – MIKULA, K.: *Modelovanie ekonomických a finančných procesov*. Bratislava, skriptá Univerzita Komenského Bratislava, 196 s., ISBN 80-223-1259-2.
- [LAM89] LAMOŠ, F. – POTOCKÝ, R.: *Pravdepodobnosť a matematická štatistika*. Bratislava, ALFA, 1989; 342 s., ISBN 80-05-00115-0.
- [MEL05] MELICHERČÍK, I. – OLŠAROVÁ, L. – ÚRADNÍČEK, V.: *Kapitoly z finančnej matematiky*. Bratislava, EPOS, 2005, 242 s., ISBN 80-8057-651-3.
- [MÍK85] MÍKA, S.: *Numerické metody algebry*. Praha, SNTL, 1985, 169 s.
- [NEL99] NELSEN, R.B.: *An introduction to Copulas*. New York, Springer, 1999.
- [NAN03] NÁNÁSIOVÁ, O.: *Map for simultaneously measurements for a Quantum logic*, Int. Jour. of Theor. Phys., vol. 42, 2003, s. 1889-1903.
- [RAN02] RANK, J.: *Improving VaR Calculations by using Copulas and Non-Gaussian Margins*. MathFinance Workshop, 3. apríla, 2002, http://workshop.mathfinance.de/2002/papers/-Joern_Rank_Copula_20020403.pdf, navštívené dňa 25. 6. 2006.
- [URB06] URBANÍKOVÁ, M.: *Financial derivatives and their usage in corporate practice*. In: CO-MAT-TECH 2006. 14. medzinárodná vedecká konferencia (Trnava, 19.–20. 10. 2006). Bratislava: STU v Bratislave, 2006, s. 1449-1456. ISBN 80-227-2472-6.
- [WON95] WONNACOTT, T. H. – WONNACOTT, R. J.: *Statistika pro obchod a hospodářství*. Praha, Victoria Publishing, 1995, 891 s., ISBN 80-85605-09-0.
- [1] SAS® RISK Dimensions®: Dynamic Risk factor Modeling Methodology. White Paper, dostupné na www.riskadvisory.com/pdfs/sasriskdimensionsriskfactor.pdf, navštívené dňa 20. 9. 2006.
- [2] <http://www.bba.org.uk/bba/jsp>

[3] <http://www.finance-research.net/>

[4] <http://www.nbs.sk/>

[5] <http://www.sas.com/>

Adresa: RNDr. Mária Bohdalová, Ph.D.
KIS FM UK, Odbojárov 10
820 05 Bratislava
E-mail: `maria.bohdalova@fm.uniba.sk`

ZAŘAZENÍ GEOGRAFICKÝCH INFORMAČNÍCH SYSTÉMŮ DO VÝUKY PŘEDMĚTU INFORMATIKA VE VEŘEJNÉ SPRÁVĚ

INCLUSION OF THE GEOGRAPHICAL INFORMATION SYSTEMS IN INFORMATICS IN THE PUBLIC ADMINISTRATION

Miroslava Dolejšová

Abstract: This paper deals with a new conception of the subject "Informatics in the public administration" taught in the second year of the field of study "Public administration and the regional development" and guaranteed by the Institute of the Informatics and Statistics at the Tomas Bata University in Zlín. The main purpose is to point out the significance of the geographical information systems in the public administration, to describe and clearly demonstrate the ways of their application both from the eye of citizens and the public authorities and to indicate the further development in this area.

Отвлеченное понятие

Включение географических информационных систем в обучении предмета Информатика в публичной администрации

Статья занимается новым предмета Информатика в публичной администрации во втором курсе учебной специальности Публичная администрация и региональное развитие. Предмет обеспечивает Институт информатики и статистики Университета Фомы Бата в Злине. Основная цель настоящей статьи состоит в ассигновании значения географических информационных систем в области публичной администрации, характеристике и демонстрации метод их использования с точки зрения гражданина и организации публичной администрации и также в подсказе следующего возможного развития в этой области.

V současné době informační a znalostní společnosti se bez kvalitních informací prakticky neobejdeme. Na jedné straně existuje dostatek informací, ale stačí nám to k rozhodování? Víme, které informace máme použít, abychom

učinili správné rozhodnutí? Máme k dispozici nástroje, které by nám byly schopny pomoci rozhodování usnadnit?

Odpověď je poměrně jednoduchá. Tyto nástroje k dispozici jsou a nazývají se geografické informační systémy.

Geografické informační systémy mají uplatnění v řadě oblastí. Jednou z nejvyužívanějších oblastí je oblast veřejné správy. Cílem tohoto příspěvku je popsat možnosti zavedení geografických informačních systémů do předmětu Informatika ve veřejné správě, který je vyučován na Fakultě ekonomiky a managementu Univerzity Tomáše Bati ve Zlíně.

Stručné představení předmětu IVS

Ústav informatiky a statistiky zajišťuje výuku povinného předmětu Informatika ve veřejné správě ve 2. ročníku zimního semestru studijního programu Hospodářská politika a správa ve studijním oboru Veřejná správa a regionální rozvoj, a to ve všech formách studia. Výuka v prezenční formě studia probíhá v rozsahu dvou cvičení po dobu 14 týdnů na počítačových učebnách, pro kombinovanou a celoživotní formu studia je vymezeno 9 konzultací.

Předmět je v první řadě zaměřen na práci s informačními zdroji, které jsou potřebné v oblasti veřejné správy. Neméně důležitým cílem tohoto předmětu je rozšíření počítačové gramotnosti studentů o kreslení diagramů v programu Microsoft Word, rozšíření znalostí programu Excel (grafy, jednoduché databázové funkce), případně další problémy, které budou potřeba pro vypracování bakalářské nebo diplomové práce.

Předmět je ukončen klasifikovaným zápočtem. Podmínkou pro jeho získání je vypracování seminární práce a úspěšné absolvování znalostního testu. Charakter seminární práce je zcela tvůrčí. Jejím cílem je provést analýzu toku informací ve studentem vybrané organizaci veřejné správy. V případě, že se studentům vypracování seminární práce zalíbí, mohou pokračovat v jejím podstatném rozšíření i v podobě bakalářské nebo diplomové práce.

Proč právě GIS v předmětu IVS?

Zásadním nedostatkem oboru Veřejná správa a regionální rozvoj jsou chybějící předměty, které jsou vyučovány Ústavem informatiky a statistiky ve studijním oboru Management a ekonomika. V oboru Management a ekonomika jsou navíc vyučovány předměty Databáze a programování (letní semestr 1. ročníku) a Aplikovaná informatika (zimní semestr 2. ročníku), který je věnován řešení konkrétních manažerských úloh v programu Excel. Studenti oboru Veřejná správa a regionální rozvoj mají pouze předmět Informatika

pro ekonomy (pro oba obory je předmět společný) a ve druhém ročníku pak předmět Informatika ve veřejné správě.

Pouhé vyhledávání informačních zdrojů, rozšíření počítačové gramotnosti o možnosti Wordu a Excelu podle názoru autorky nestačí. Návaznost na další předměty oboru Veřejná správa a regionální rozvoj (především předmět Regionální analýza) vyvolala potřebu úpravy obsahu tohoto předmětu, který by propojil obsah obou předmětů z různých úhlů pohledu. Předmět Informatika ve veřejné správě by se mohl stát doplňkem předmětu Regionální analýza, případně dalších předmětů, které jsou vyučovány Ústavem veřejné správy a regionálního rozvoje. V každém případě by se jednalo o obohacení obsahu obou uvedených předmětů, rozšíření znalostí studentů a navázání úzké pedagogické i vědeckovýzkumné spolupráce mezi oběma ústavu.

Proč využívat GIS?

K nejčastějším typům informací, které potřebuje každý člověk i organizace, jsou geografické informace. Pravidelně hledáme cestu do konkrétního místa, rozhodujeme se, kde umístíme další provozovnu, chceme vědět, jak dlouho nám přeprava do zvoleného místa bude trvat a také chceme znát vzdálenost mezi oběma místy. Geografické informační systémy nám tyto (a nejenom tyto) typy informací poskytují.

Možná se na první pohled bude zdát, že geografické informační systémy jsou jen software. GIS však možnosti klasického programu převyšují. Nejen, že nám usnadňují rozhodování, ale především se jedná o kvalitní grafický nástroj pro analýzu a modelování dat, která jsou ve většině případů volně k dispozici na Internetu ke stažení.

Geografické informační systémy jsou obvykle spojovány s mapami a navigačním systémem. Mapy jsou však výstupem geografických informačních systémů a navigační systém (myšleno systém globální navigace GPS) nebývají zahrnuty do těchto systémů.

Základem jsou však data

V minulosti i dnes lidé získávali znalosti a sdíleli je v různých podobách. Dorozumívali se prostřednictvím slov, textu, symbolů v podobě hieroglyfů, hudby, obrazů a kreseb. V dnešní době digitalizace sdílíme znalosti prostřednictvím různých sítí (především World Wide Web). Postupně se možnosti rozšiřují o počítačové modelování a simulace, digitální zpracování obrazu i textu, správu obsahu a zejména o využívání metod statistické analýzy.

Tři pohledy na GIS

Na geografické informační systémy lze pohlížet ze tří hledisek: mapového, databázového a modelového.

Mapový pohled je nejjednodušší. Pomocí geografických informačních systémů můžeme mapy vytvářet a upravovat. Výsledná mapa je pak souborem námi vybraných vrstev, které chceme v mapě zobrazit.

Mapa jako obraz nám nebude stačit. Potřebujeme ji doplnit o konkrétní údaje. Tato data jsou uložena v různých databázích, které se často označují jako geodatabáze.

Pokud máme k dispozici obraz i data, můžeme prostřednictvím různých analytických nástrojů získávat další informace potřebné pro rozhodování. Výběrem dat, různými dotazy, aplikací statistických a analytických funkcí a dalších vhodných nástrojů získáme nová data, která nám umožní získat komplexnější pohled na studovaný problém.

Geografické informační systémy lze proto definovat jako prostředek zobrazení, manipulace a analýzy prostorových dat.

Typy dat, se kterými GIS pracují

Geografické informační systémy pracují s různými typy dat. K nejpoužívanějším z nich patří vektorová a rastrová data. Vektorová data mohou být vyjádřena jako polygony (typické pro lesy, území krajů a velkých měst, vodní plochy), čáry pro zobrazení silnic, železnic, řek, ulic nebo jako body (stromy, obchody, malá města a obce). Rastrová data jsou vyjádřena jako matice bodů. K rastrovým datům patří družicové a letecké snímky, zobrazení terénů a sítí.

Co můžeme zjišťovat prostřednictvím GIS?

- Zjišťovat počty (KOLIK?): počet nemocnic, počet úřadů.
- Zjišťovat hustotu (JAK MNOHO?): počet lékařů připadajících na 1 000 pacientů.
- Zjišťovat, co je uvnitř určité oblasti: počet obchodů v určitém regionu.
- Zjišťovat, co je poblíž: nejbližší nemocnice, nejbližší rekreační středisko.
- Modelovat změny (obvykle předpovědi počasí).

Praktické ukázky

Další část příspěvku bude popisovat možné aplikace geografických informačních systémů (nebo lépe řečeno ne přímo jejich podoby) v předmětu Informatika ve veřejné správě. Jelikož se studenti v tomto předmětu seznamují

také s nejpoužívanějším informačním zdrojem ve veřejné správě, Portálem veřejné správy České republiky, bude potřeba jejich znalosti rozšířit i o aplikaci mapových služeb. Příspěvek bude charakterizovat ilustrativní příklady, které byly vytvořeny Portálem veřejné správy [1] a prostřednictvím internetové stránky <http://www.mapy.cz/> [2].

Protože se již druhým rokem konference STAKAN koná v krásném prostředí Rusavy, rozhodla jsem se ukázat možnosti, které geografické informační systémy nabízejí, právě v okolí Rusavy.

Při pokusu najít přesnou polohu rekreačního zařízení Rusava (chata Jestřabí) budeme úspěšní přes internetové stránky <http://www.mapy.cz/>. Pokud se o totéž pokusíme přes Portál veřejné správy, zobrazí se jen turistická mapa. Práce s mapovými službami není vůbec složitá. Mapové služby spustíme kliknutím na odkaz Mapy v pravé horní části Portálu veřejné správy. Největší část této aplikace tvoří prostor pro zobrazení mapy. Postupným kreslením obdélníků pomocí myši specifikuje region, který chceme prozkoumat. Jakmile jsme s výběrem zcela spokojeni, vybereme z části Funkce aplikace odkaz na obrázek tří map, který představuje seznam tematických úloh. K dispozici je jich celkem padesát.

Velmi důležitou součástí aplikace jsou karty Vrstvy a Legenda. Na kartě Vrstvy lze vybrat, co konkrétního si přejeme zobrazit, na kartě Legenda vidíme vysvětlení popisků mapy a vybrané statistické údaje. Zrušením zaškrtnutí před příslušnou vrstvou odstraníme z mapy všechno, co v ní vidět nechceme. Změna se však provede pouze překreslením mapy (zakulacená šipka nacházející se vlevo od karty Vrstvy). Podle barevného rozlišení jsme schopni zjistit potřebné informace. Rozsah využívaných dat je však omezen pouze na údaje Českého statistického úřadu a na informace týkající se životního prostředí. Přesto můžeme zjišťovat velmi zajímavé informace.

Budeme-li se zajímat o hustotu zalidnění v okolí Rusavy, vybereme tematickou úlohu Hustota zalidnění. Z karty Vrstvy necháme zobrazeny hranice územních jednotek, obce a komunikace a hustotu zalidnění (obr. 1). Z obrázku 1 je patrné, že v okolí Rusavy je hustota obyvatel asi 2 až 14 obyvatel na km².

Autorku příspěvku také zajímalo, zda má obec Rusava veřejnou knihovnu. Přímě v Rusavě i v jejím okolí se nachází jedna knihovna (obr. 2). Další zajímavou otázkou je vybavení obce Rusava kanalizací. K velkému překvapení bylo zjištěno, že v okolí Rusavy není kanalizace vybudována vůbec (obr. 3).

K dalším součástem mapových služeb patří i vektorová mapa pozemních komunikací. Tato mapa umožňuje zkoumat, jaké objekty se v dané lokalitě vyskytují. Z obrázku 4 je patrné, že v Rusavě je pouze 7 mostů (malá kolečka na silnici).

Jinou, velmi zajímavou možností, je měření vzdálenosti mezi dvěma náhodně zvolenými místy. Protože máme zájem vidět dvě konkrétní místa v okolí Rusavy, musíme si vybrat jiný, způsob zobrazení mapy. Zobrazení podrobnějších informací nejlépe nabízí topografická mapa Armády České republiky. Hledáme možnosti, jak se dostat z místa A (zelený text RUSAVA pod kótou 536 vlevo nad obcí Rusava) do místa B (část Ráztoka, kóta 422). První možností je chůze po silnici (obr. 5). Délka trasy je delší než 4 km (přesněji 4 406 m). Druhou možností je vydat se přímo bez ohledu na překážky v terénu, což bude činit přes 2,5 km (přesněji 2 566 m). To znázorňuje obrázek 6.

To, co mapové služby Portálu veřejné správy nedovedou, je plánování tras. Zde byly použity údaje internetové stránky <http://www.mapy.cz/>. Můžeme si naplánovat trasu ze Zlína do Rusavy a zjišťovat jednak nejrychlejší trasu (obr. 7), která trvá 40 minut a její délka je 30,88 km a jednak nejkratší trasu (obr. 8), která trvá 45 minut a její délka je téměř 30 km (přesněji 29,48 km). Můžeme si detailně prohlédnout celou trasu.

Internetová stránka <http://www.mapy.cz/> umožňuje také vyhledávat nejbližší objekty v okolí. Pokud se budeme zajímat o nejbližší ubytování v Rusavě, najdeme celkem tři možnosti, které jsou označeny číselně. V pravé části obrazovky si můžeme prohlížet i detailní informace o daném místě. Chata Jestřabí však nebyla v seznamu ubytování na Rusavě nalezena.

Po zadání klíčového slova „rekreační zařízení“ se chata Jestřabí již objeví (obr. 9). Nyní si z tohoto místa naplánujeme cestu do Zlína. Výsledkem bude popis trasy s uvedením počtu kilometrů na jednotlivých úsecích trasy a odhadovaný čas přepravy. Celková délka trasy bude cca 31 km (přesněji 30,88 km) a cesta bude trvat 40 minut, což je vlastně nejrychlejší cesta (obr. 10).

Jak již bylo řečeno, je rozsah zjišťovaných informací, které jsou k dispozici na Portálu veřejné správy, dost omezený. Můžeme čerpat podklady o životním prostředí, údaje z integrovaného registru znečišťování, hranice územních jednotek, volebních obvodů a příslušných úřadů, lze využívat vojenské i staré mapy, údaje Českého statistického úřadu týkající se obyvatelstva a vybavenosti obcí, údaje České pošty, zjišťovat informace o kvalitě koupacích vod a omezené informace o dopravě. I když není tématických úloh tak velký počet, umožňují i tyto informace pomoci nalézt odpovědi na otázky, které nás zajímají. Uživatelé, kteří nedovedou ještě pracovat s geografickými informačními systémy, získají alespoň základní přehled možností mapových služeb Portálu veřejné správy i informační stránky <http://www.mapy.cz/>. Dalším krokem pro ně již bude naučit se pracovat s konkrétním geografickým informačním systémem.

Další možnosti výuky předmětu Informatika ve veřejné správě

Z výše uvedených popisů je patrné, že geografické informační systémy si pozornost určitě zasluhují. Ukázky uvedené na přednášce STAKAN 2007, které doplňuje krátká videoukázka, jsou první vlaštovkou pro inovaci předmětu Informatika ve veřejné správě.

Přesto je vhodné ukázat a naučit studenty oboru Veřejná správa a regionální rozvoj pracovat prakticky s konkrétním produktem geografického informačního systému minimálně v rozsahu dvou až tří cvičení. Tento krok bude podle názoru autorky velkým oživením výuky.

V souvislosti s vypracováním seminární práce na téma analýzy informačních toků se nabízí otázka rozšíření obsahu tohoto předmětu o analýzu rizik informačních toků a informačních systémů v oblasti veřejné správy. Předmět lze rozšířit i o aplikaci metod kontroly a auditu ve veřejné správě, jednoduchých metod pro hodnocení veřejných projektů a veřejných zakázek, pomocí nichž by bylo možné řešit studenty vybraný problém týkající se veřejné správy. Zvažuje se i výuka krizového řízení ve veřejné správě v podobě konkrétního softwarového produktu. Nabízí se celá řada možností, které obohacují obsah tohoto předmětu ve všech jeho stránkách.

V neposlední řadě bude tento předmět inspirací pro vypracování bakalářských i diplomových prací a současně jednou z oblastí vědeckovýzkumné činnosti Ústavu informatiky a statistiky.

Autorka příspěvku přivítá jakékoliv podněty i kritické připomínky týkající se geografických informačních systémů a problematiky veřejné správy.

Internetové zdroje: Odkazy byly funkční k 15. červnu 2007.

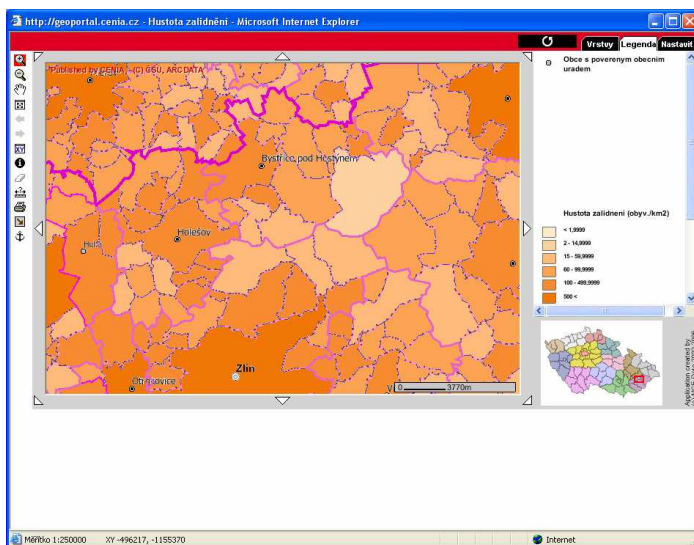
Reference

- [1] Ministerstvo vnitra. *Portál veřejné správy České republiky* [online]. 2003-2007 [cit. 2007-06-15]. Dostupný z WWW: <<http://geoportal.cenia.cz/mapmaker/cenia/portal/>>.
- [2] Seznam.cz. *Mapy.cz – mapa Evropy, České republiky, plány měst a obcí v ČR* [online]. 1996-2007 [cit. 2007-06-15]. Dostupný z WWW: <<http://www.mapy.cz/#>>.

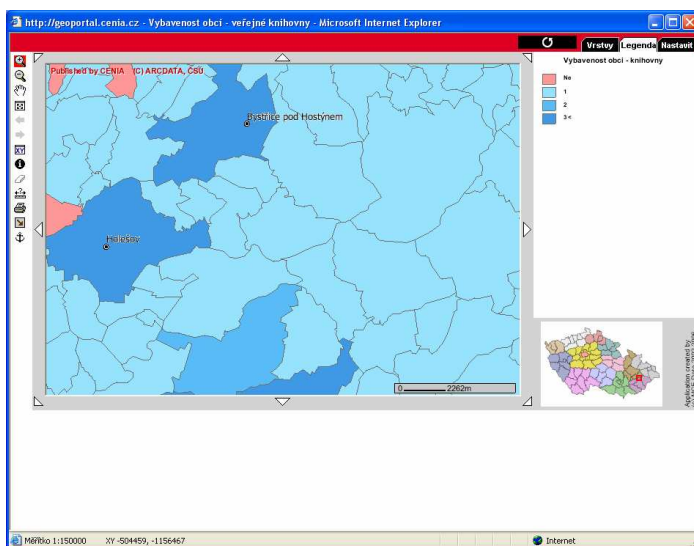
Adresa: Ing. Miroslava Dolejšová, Ph.D., ÚIS FaME, Univerzita Tomáše Bati ve Zlíně, Náměstí T. G. Masaryka 1279, 760 01 Zlín

Telefon: +420 576 037 430

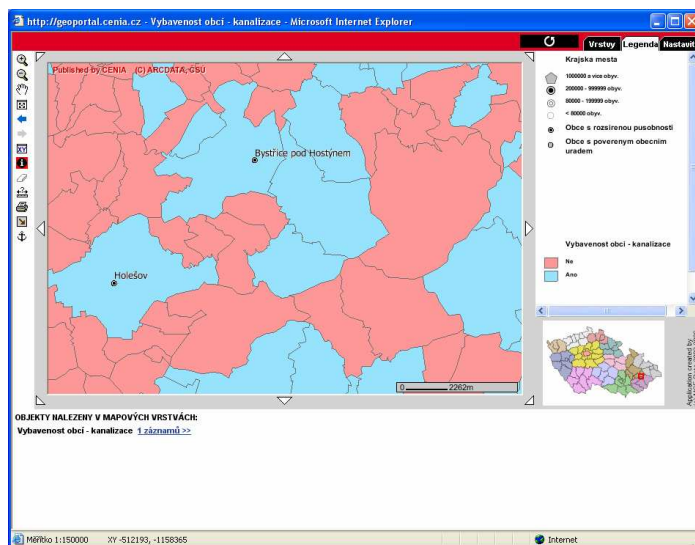
E-mail: dolejsova@fame.utb.cz



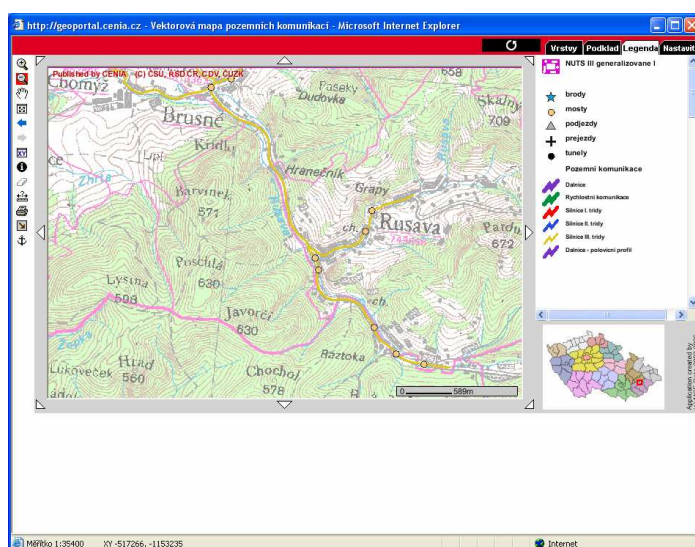
Obrázek 1: Hustota zalidnění v obci Rusava [1]



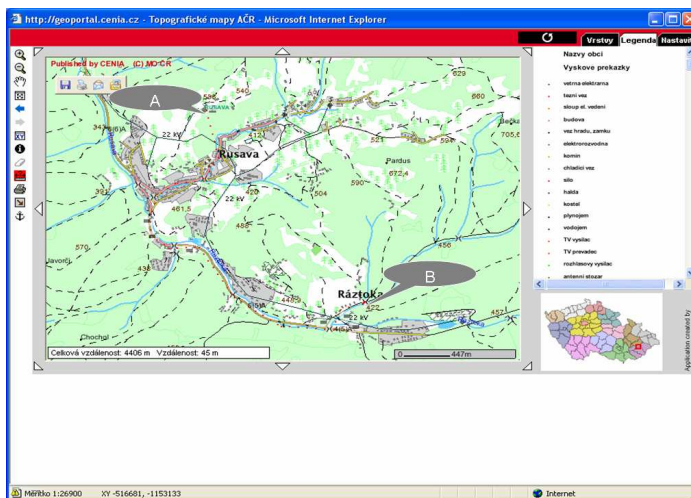
Obrázek 2: Vybavenost obce Rusava veřejnou knihovnou [1]



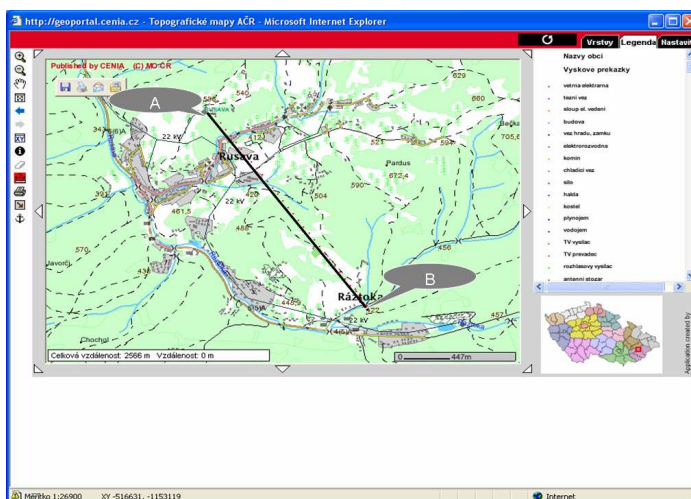
Obrázek 3: Vybavenost obce Rusava kanalizací [1]



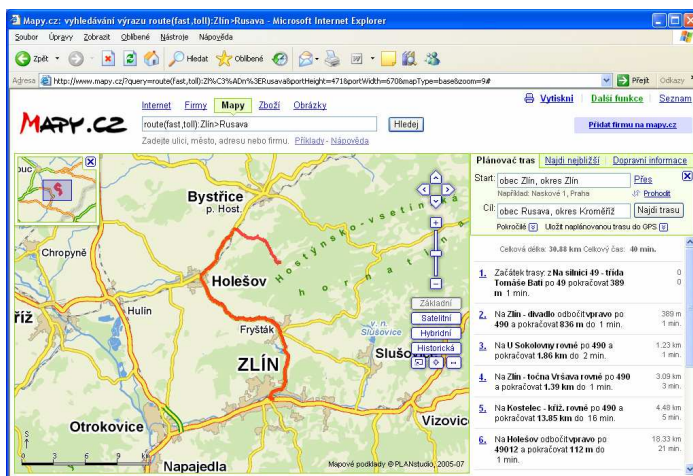
Obrázek 4: Vektorová mapa pozemních komunikací [1]



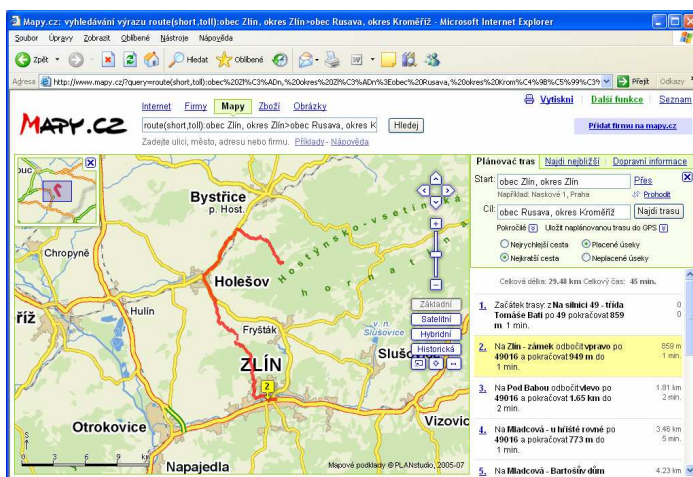
Obrázek 5: Měření vzdálenosti mezi dvěma zvolenými místy: delší trasa [1]



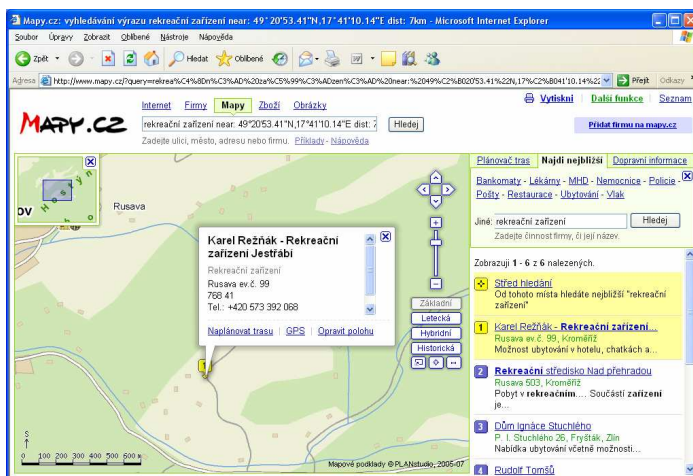
Obrázek 6: Měření vzdálenosti mezi dvěma zvolenými místy: přímá trasa [1]



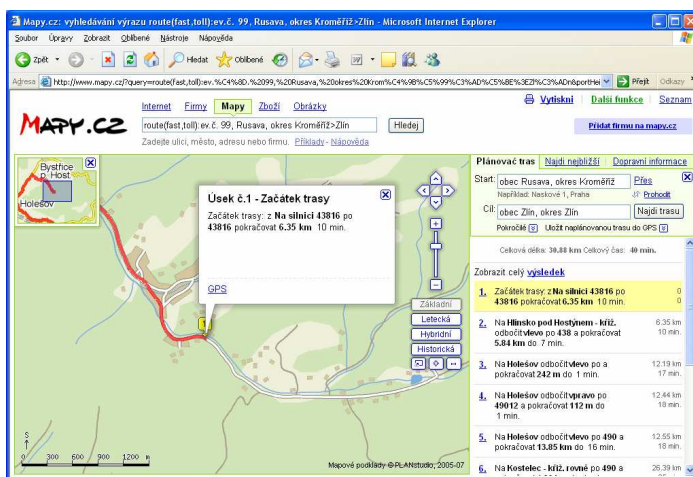
Obrázek 7: Nejrychlejší cesta ze Zlína do Rusavy [2]



Obrázek 8: Nejkratší cesta ze Zlína do Rusavy [2]



Obrázek 9: Zobrazení rekreační chaty Jestřabí [2]



Obrázek 10: Zobrazení trasy z Rusavy do Zlína [2]

VÝUKA STATISTIKY 2007

Petr Hebák

Vážení kolegové,

opakovane a možná mylně či neoprávněně předpokládám, že zdejší jednání navazuje na naše dvě předchozí setkání. Na Stakanu v roce 1999 se z různých hledisek poměrně dlouze a (podle mého názoru) důkladně diskutovalo o různých otázkách a problémech souvisejících s výukou statistiky na různých školách či úrovních studia, jakož i o užitečnosti využití statistiky nestejně obsahově i datově orientovanými (současnými či očekávanými budoucími) uživateli statistických postupů a metod. Rovněž dnes považuji za užitečné si nejdříve připomenout základní myšlenky tehdejšího příspěvku, který pak pod názvem *Učíme statistiku* vyšel ve statistickém bulletinu a je i zde pro případné zájemce v počítači k dispozici. Hlavně se však budu v tomto úvodu snažit posoudit (dnes, stejně jako o čtyři roky později v roce 2003) stav, ke kterému jsme došli za osm let od mého prvního vystoupení na toto téma. Těmito a podobnými sliby jsem zahájil i přednášku na Stakanu v roce 2003 pod názvem *Výuka nestatistiků*. Již tehdy jsem se společně s účastníky setkání zamýšlel nad změnami a zkušenostmi v oblasti významu a výuky statistiky, jakož i nad oprávněností obav o budoucí postavení a přežití *statistiky* v jednadvacátém století, které jsou i dnes (z mnoha článků renomovaných statistiků ve význačných odborných časopisech) stále zřetelnější. Jak však uvidíme, není to všude stejné a zvláště na otázku o postavení statistiky odpovídají někteří jinak, než bychom asi odpověděli dnes my. V roce 1999 jsem řekl, že některé pesimistické úvahy úzce souvisí nejen se zaměřením výuky, ale rovněž s užitečností a všeobecnou úctou k výsledkům naší činnosti, jež jsou v různých podobách předkládány veřejnosti. Důsledky našeho pedagogického statistického působení částečně vycházejí z postoje, který máme sami k sobě a k našemu oboru, ale i z postoje, který oprávněně či neoprávněně mají jiní k nám a naší disciplíně. Stále jsem totiž přesvědčen, že otázka *postavení statistiky v budoucnosti* je neoddelitelná od obsahu i způsobu výuky *nestatistiků*. Nevím, zda na VŠE v Praze je situace netypická (ve skrytu duše doufám, že tomu tak je), protože u nás (podle mého hodnocení) situace v oblasti výuky statistiky pro *nestatistiky* a možná brzy i pro *statistiky* se ještě výrazně zhoršila. Co říci, když místo dvousemestrového předmětu 3/1 je zařazen jednosemestrový předmět 2/2, když sjednocování způsobu zkoušení různých předmětů ve svých důsledcích vede k situaci, kdy od studentů nelze příliš mnoho chtít a navíc

některé fakulty či katedry se snaží minimalizovat výuku matematiky, statistiky či jiných i jen velmi málo exaktně orientovaných předmětů. Ve smyslu bývalého hesla „*Za masovost – za rekordy!*“ u nás začíná převládat snaha o masové (velice všeobecné) tříleté bakalářské studium s nejrůznějšími předměty údajně manažerského typu před klasickým univerzitním vzděláním. Do pozadí (myslím si, že chybně) ustupuje magisterské, jakož i ucelené pětileté studium s hlubší orientací na zvolený obor studia. Ukazatel počet studentů na jednoho učitele ani nechci komentovat, i když jeho důsledky všichni velmi dobře známe. Je určitě velice dobré, že přicházející absolventi středních škol mají výrazně lepší jazykové znalosti (zvláště angličtiny) než ti dřívější či někdy i důkladnější než jejich učitelé a je užitečné, že vědí poměrně hodně o práci a možnostech počítačů. Méně je už povzbudivé, že tradičně k nám přicházejí studenti z různých typů středních škol bez aspoň minimalistického matematického základu. Myslím, že předmět *logika* či nějaký podobný dnes na gymnáziu ani není vyučován. Odhaduji, že pokud jde o význam pravděpodobnostního nebo statistického myšlení více než 90 % zájemců o studium (bohužel i těch, kteří se hlásí na obor statistika) ani nemá tušení, že něco takového existuje. Nejen starší politici, vystudovaní novináři či komentátoři v rozhlase a televizi (jak až na vzácné výjimky neustále vidíme), ale i žadatelé o studium, kteří se narodili buď těsně před, nebo už po roce 1989 nemají ani elementární schopnost posoudit význam *čísel* různého typu a mají k nim spíše až podvědomý odpor. Po přijetí ke studiu je tím už dopředu dána velmi neurčitá (spíše velmi malá) naděje, že v průběhu krátkého (často zcela jinak orientovaného) studia se situace výrazně změní. Zkušenosti říkají, že pokud sami nebo díky rodičům, pedagogům či kamarádům někteří z nich zásadně nezmění svůj *osobní postoj* k exaktnímu a kvantitativnímu uvažování, lze jen těžko předpokládat, že po ukončení studií ve své budoucí běžné činnosti nebo odpovědné funkci budou sledovat nějaké numerické analýzy či dokonce je doporučovat a s úspěchem využívat.

Před čtyřmi i osmi roky obdrželi účastníci Stakanu jako jeden z podkladů pro diskusi rozsáhlé příspěvky k výuce statistiky, takže jsem se při slovním doprovodu textu zaměřil jen na některé sporné nebo zajímavější oblasti. Týkalo se to nejen hlavního tématu, kterým byla *výuka statistiky pro nestatistiky*, ale i některých otázek studijních programů na oborech s převážně či výhradně statistickým zaměřením. Obsahově byla celá problematika rozdělená do následujících (myslím si, že pořád aktuálních) bodů. Některé z nich přesahovaly obsahové pojetí tehdy mnohem úžeji zaměřené debaty, takže jim byla věnována menší nebo téměř žádná pozornost. Pro připomenutí to byly tyto body:

1. Jaké je postavení statistiky v zahraničí a u nás?
2. V čem se statistika za posledních dvacet let změnila a co lze očekávat?
3. Co se od nás statistiků při výuce statistiky požaduje?
4. Jsme pedagogové, vědci nebo obojí?
5. Jakou formu výuky zvolit?
6. Jak jsme na tom se statistickou literaturou?
7. Jaký obsahový a časový rozsah výuky statistiky bychom doporučili?
8. Kdo a jak statistiku učí kromě nás?
9. Jaký by měl být obsah základního kurzu statistiky pro bakaláře?
10. Jaký by měl být obsah kurzu statistiky pro magistry na různých oborech?
11. Jaké máme naděje na obchodních, podnikatelských či MBA kurzech?
12. Jak organizovat studium na oboru *Statistika*?
13. Jak organizovat doktorské či jiné postgraduální studium statistiků?

Už v minulém vystoupení bylo jedním ze záměrů ukázat nejen na současný stav, ale zamyslet se nad obecně možným vývojem různých vědních oborů, nad interdisciplinárním přístupem k některým obecnější otázkám zpracování dat a v této souvislosti i nad očekávaným postavením statistiky v budoucnosti. Vzhledem k mému zájmu, jsem se i minule snažil vyvolat debatu o možnosti využití *bayesovského* pojetí pravděpodobnosti v základním kurzu pro nestatistiky a zmínit nezbytnost propojení některých metod s dalšími informacemi teoretického i empirického charakteru (tedy s informacemi pocházejícími odjinud než z analyzovaných dat) při větším zapojení lidského úsudku, a to zvláště při řešení úloh postrádajících bezprostřední analogii. Souvisí to nejen s filozoficky orientovanými debatami o *subjektivní pravděpodobnosti* a s odlišnými názory na různé možnosti pojetí výuky statistiky, na které jsem se v přednášce odvolával současně s prezentací jedné takto orientované učebnice. Má se to obecněji dotýkat otázky využití statistiky v oblasti společenských věd a inspirovat k zamyšlení nad všeobecně *nedostatečným využíváním* dat a nad běžnou absencí datových analýz při *rozhodování*.

Na závěr tohoto úvodu připomínám pár shrnujících poznámek k příspěvku z roku 1999, o kterých po letech s uspokojením konstatuji, že bych na nich dnes změnil pouze drobnosti, zatímco v zásadních bodech bych s nimi stále souhlasil.

Co chceme, aby si studenti z výuky statistiky odnesli?

A) Budoucí statistici absolventi oboru statistiky (na VŠE či jinde):

- Solidní matematické základy.
- Různé aspekty pravděpodobnostního a statistického myšlení.
- Široké znalosti statistických metod a technik.
- Způsob přípravy experimentu a organizace různých šetření a zjišťování.
- Vše co souvisí s přípravou a pořizováním dat.
- Ovládání statistického počítačového zázemí.
- Práce s velkými databázemi a obecně rozsáhlými soubory.
- Znalost problematiky oboru či oblastí používání.

B) Nestatisticy (jedno semestrový až tři semestrový kurz statistiky typu 2/0 až 2/2):

- Základy pravděpodobnostního a statistického myšlení.
- Pochopení významu čísel, charakteristik či ukazatelů.
- Argumenty proti přílišné demokratizaci statistiky.
- Schopnost předcházet bagatelizaci a vulgarizaci při používání statistiky.
- Odlišení různých typů dat a způsobů jejich pořízení.
- Důsledky agregace a kategorizace dat.
- Odlišení možností různých statistických metod.
- Plné pochopení jednoduchých ilustrativních úloh s výsledky z počítače.
- Schopnost debat nad výstupy z počítačů (vstup a výstup metod).

Co z toho plyne pro výuku statistiky pro nestatistiky?

- Ústup od převážně formálně matematického přístupu k výkladu.
- Výklad ve smyslu „les místo stromů a nadhled před podrobnostmi.“
- Minimum vzorců a podrobností o výpočetních algoritmech.
- Důkazy jen zcela výjimečně pro objasnění myšlenky.
- Významná úloha výběru a statistické literatury a kvality výkladu.
- Hluboké znalosti přednášejících a malá demokratizace statistiky.
- Výklad nemá být zaměřen tak, aby se absolventi kurzu stali statistiky.
- Užitečnost výuky statistiky u počítače je při nejmenším sporná.
- Čas nelze ztrácet psaním vzorců či výkladem počítačového postupu.
- Jsem přesvědčen, že teorii opakovaných výběrů student téměř nikdy nepochopí.

Undergraduate Statistical Education (podle S. S. Wilkse před 50 lety)

Na rozdíl od mých pesimistických obav o budoucnost uplatnění statistiky u nás je mnohem radostnější si znovu přečíst příspěvek *Samuela Stanleyho Wilkse* na pravidelném ročním setkání *American Statistical Association*, i když byl přednesen už v roce 1950! Proslov Wilkse byl velice zajímavý a hlavně je tak stále aktuální, že jej znovu uvedl prestižní časopis *The American Statistician* v prvním čísle minulého roku pod názvem *Undergraduate Statistical Education*. Wilks tehdy vyšel z jakoby už v té době nesporné skutečnosti, že vývoj přinesl všeobecné široké uplatnění statistiky a statistických metod ve všech oblastech veřejného života a vědeckého výzkumu. Wilks konstatoval (jakoby už na přelomu minulého století samozřejmou) skutečnost, že statistika a její uplatnění se už netýká jen profesionálních statistiků, ale tisíců lidí, kteří poznali nutnost stát se *inteligentními konzumenty* statistiky a statistických metod. Za vážný a dlouhodobě vleklý problém však už tehdy označil situaci v oblasti statistického vysokoškolského vzdělávání. Název i obsah příspěvku ukazuje, že se Wilks zabýval přesně stejnými otázkami, o kterých my už (říkejme o téměř 60 let později) stále debatujeme. Nevím, zda jeho myšlenky tehdy vstoupily do amerického vzdělávacího systému, ale každopádně jsem považoval za užitečné si jeho slova aspoň částečně připomenout.

Než přejdu k systému, který Wilks na setkání členů americké statistické asociace předložil, bych rád poznamenal, že kromě výše zmíněného všeobecného rozšíření oblastí využití statistiky byly pro mne rovněž velice překvapující další odstavce příspěvku, ve kterých Wilks řekl, že nemá v úmyslu zabývat se postgraduálním statistickým vzděláním, protože *podle jeho slov je tato oblast široce a úspěšně vyřešena*. Důvodem byla (a snad stále jsou) *silná centra pro pokročilou výuku aplikované i teoretické statistiky s rostoucím počtem úspěšných absolventů a dalších schopných zájemců o studium*. V této souvislosti Wilks zmínil i *užitečný důsledek dřívějších rozsáhlých debat o způsobu výuky pokročilé statistiky a obrovské zásluhy řady významných (v článku jmenovaných) statistických institucí i skutečnost, že na trhu je dostatek velice kvalitních knih a studenti mají k dispozici spoustu užitečných studijních materiálů*. Mohu jen konstatovat, že je mi velice líto, že pro argumentaci týkající se našeho dnešního postgraduálního studia na oboru statistika jsem nikdy neměl aspoň k nahlédnutí výsledky zmiňovaných debat o výuce pokročilé statistiky ani důsledky jím navrženého systému vysokoškolského statistického vzdělávání.

Méně už pro mne bylo překvapivé konstatování Wilkse, že zmíněný i další (v článku podrobně popsany) růst zájmu o uplatnění statistiky odhalil, že

v USA (a určitě nejen tam) je spousta (říká tisíce) lidí nedostatečně statisticky vzdělaných a velmi málo vybavených elementárními základy statistiky. Většina z nich totiž nezískala téměř žádné nebo zcela žádné statistické vzdělání na střední ani vysoké škole. Wilks se pak nediví velkému zájmu o večerní studium o krátké kurzy nebo jednoduché knížky či jiné studijní materiály, který dodatečně projevují mnozí lidé s nejrůznějším zaměstnáním, zaměřením či odlišným odborným vzděláním. Někteří univerzitní učitelé statistiky se pak na této *ad hoc* zaměřené výuce aktivně podílejí, což je sice jistě chvályhodné, ale vůbec neřeší dlouhodobé potřeby v této oblasti. Podle Wilkse takové statistické vzdělávání připomíná (cituji volným překladem) *skupinu dočasných kasárenských baráků, které byly postaveny ve spěchu kousek po kousku, nepěkně vyhlížejících a stojících na velmi chatrných základech*. Wilks shrnuje, že pokud to je (říkejme *tehdy bylo*) zapotřebí nebo není k dispozici lepší řešení velkého zájmu o získání elementárních informací o statistice, nelze k tomu nic říci či namítat, ale dříve či později to stejně bude vyžadovat nějaký propracovanější systém.

Nemusím asi příliš vysvětlovat, že mi jeho slova zvláště v této části připadají velice současná. Dokonce natolik současná, že jsem se při prvním čtení mylně domníval, že jeho příspěvek je z poslední doby a nikoli přes půl století starý.

Podle Wilkse ve výuce základů statistiky je zapotřebí (cituji bez překladu):

„... *We need in statistics elementary courses at elementary levels in which the student can concentrate on fundamental concepts and basic skills in a graduated manner, doing just enough problems and laboratory exercises to fix these ideas without losing himself in the meaningless manipulation of formulas. If these elements are presented clearly and systematically to a student early in his college career he will be in position to use them with facility and understanding in later courses, in thesis work, and in life-sized problems. If properly organized this basic material can be presented eventually in a sequence of two full-year courses, just as the basic mathematics for students in the physical sciences and engineering ...*“

Jaké s tím souvisí organizační aspekty výuky elementární statistiky?

a) Kdo a jak má učit základní kurzy statistiky?

Základy statistické analýzy a principy statistických úsudků jsou v zásadě stejné, když se odpovídajícím způsobem vyučují v biologii, ekonomii, matematice, psychologii či sociologii. Jde o elementární vysokoškolské kurzy a na různých školách mohou být odlišné přístupy k jejich zabezpečení. Někde existuje katedra statistiky, jinde může pro tento účel takto zaměřená katedra vzniknout a vzácností není, že taková výuka je svěřená katedře matematiky. Může se dokonce i stát, že (v rámci tzv. demokratizace statistiky) se budou snažit výuku statistiky nabízet i zástupci jiných (více či méně příbuzných) kateder a už teď je možné říci, že nejsou dobré zkušenosti s tím, když statistiku učí někdo jiný než statistik. Například katedry matematiky jen zřídka mají členy zaměřené výhradně nebo převážně na počet pravděpodobnosti a jsou tak soustředěny na klasickou matematickou výuku. Je pro ně velice obtížné si představit, že studenti ať už třeba v biologii či ve společenských vědách potřebují trénink kvantitativních metod, a pokud si takovou potřebu uvědomují, pak většinou nevidí žádné důvody k tomu, aby se výuka studentů ekonomie, sociologie či biologie nějak zásadněji odlišovala od klasického matematického vzdělávání fyzikálně či technicky orientovaných studentů. Poznamenejme, že pod dojmem existujících statistických paketů a *kuchařkovitých* návodů k jejich použití, se vedoucí pracovníci některých fakult či kateder (více než kdy dříve) domnívají, že si elementární statistické vzdělávání zabezpečí i sami vlastními silami bez statistiků. Osobně si myslím, že my statistici nutně potřebujeme najít takový způsob výuky základů statistiky, který bude na jedné straně plnohodnotným výkladem principů pravděpodobnostního i statistického myšlení a na druhé straně bude atraktivním seznámením s možnostmi statistických metod. Jinak riskujeme, že nám ujede vlak a může se stát i těm mladším, že nová šance už nepřijde.

Pozoruji už delší dobu, že náš přístup k výuce statistiky je poznamenán chybami, které se ve větší či menší míře projevují (podle mého názoru asi) téměř ve všech běžných způsobech výuky statistiky pro nestatistiky:

b) Tradiční výuka základního kurzu (říkejme aspoň ročního v rozsahu 2/2)

Problém je už na samém začátku, že výuka základů statistiky vychází z často neoprávněného předpokladu, že většina studentů je schopna logicky vnímat

výklad, trochu rozumí *číslům* a má aspoň minimalistické znalosti z matematiky (třeba v rozsahu nepovinného předmětu na nepřírodovědném gymnáziu). Při začátku kurzu se naráží na další rovněž mnohdy neoprávněný předpoklad, že je k dispozici taková učebnice *základů statistického myšlení*, kde je nejenom kvalitní výklad všech používaných pojmů, přehled potřebných vzorců, řada ilustrativních řešených příkladů i neřešených cvičení s výsledky, Ale i ukázkové úlohy, kde v rámci studovaného oboru a zaměření školy je možné se přesvědčit o užitečnosti kvantitativního přístupu. Pedagog, který si tyto skutečnosti uvědomuje, se pak (asi rovněž většinou neoprávněně) snaží suplovat výchozí neznalosti studentů, neexistenci odpovídající učebnice (*ušité na tělo probíhajícímu kurzu*) a vyložit na přednáškách vše tak, aby ti co si všechno z přednášek zaznamenají byli dostatečně vybaveni *povinnou literaturou*. Potíž je v tom, že toho je pak mnoho a v daném čase to přednášející ani technicky nemůže stihnout, natož aby studenti stihli si dělat jakýsi nadhled nad celým předmětem a vnímat pocit užitečnosti svých rozsáhlých poznámek z přednášek. Ti, co občas nebo dokonce vždy nepřijdou ani nemají možnost tento pocit získat, a ti co chodí pravidelně si spíše jen píšou než by o tom hlouběji přemýšleli. Na cvičeních to potom vypadá tak, že studenti nikdy nic nikde o statistice neslyšeli ani nepřečetli, takže cvičící si opět chybně vytvoří názor, že to teď bude muset studentům pořádně vysvětlit a navíc jim i ukázat, že leccos z toho počítač umí udělat po příslušných příkazech za nás. Domácí cvičení se už tradičně v českých podmínkách stávají spíše výzvou pro ty nejlepší, aby těm druhým pomohli odevzdat aspoň minimum. Toto vše každý z nás důvěrně zná a není potom vůbec divné, že u zkoušek to potom (opět většinou, ale naštěstí nikoli výhradně) působí dojmem, že žádná výuka nebyla, studijní materiály nejsou a počítačové zázemí příliš nepomáhá. Přes tento stav většina studentů zkoušku dříve či později udělá, aby členové jiných kateder i vedení fakult či školy nedošli k názoru, že ta statistika studentům příliš nepomáhá, pravděpodobnostní ani statistické myšlení nenabízí a konkrétní úlohy daného oboru neřeší. Po absolvování školy tito studenti s „úspěšně“ absolvovanou zkouškou ze statistiky konstatují, že jim to nic nedalo a svůj vnitřní odpor k *číslům* se promítá do jejich činnosti a stanou-li se později vedoucími pracovníky firmy, institucí, odborů atd. i do činnosti týmu, který řídí. Namítnete, že to není tak zlé, jak to zde popisuji, že je celá řada talentovaných, vědění chtivých studentů, kteří dobře vnímají potřebu kvantitativních znalostí. Máte jistě pravdu, ale to neřeší skutečnost, že způsob, kterým se statistika učí je částečně velice zastaralý a neodpovídající současným možnostem a částečně naopak příliš zahleděný do nutnosti (a neexistujícího tlaku) naučit statistiku nestatistiky dělat přímo s počítačem a tak je vybavit pro (jak se s oblibou říká) *praktický život*.

Wilks tedy žádal už v roce 1950 aspoň minimální znalosti z pravděpodobnosti a logiky ze střední školy. Dále pak na vysoké škole dvouleté plnohodnotné kurzy statistiky (nejlépe hned v prvním a druhém ročníku studia), které budou zabezpečené dobrými učebnicemi a kvalitními pedagogy (jak říká *netoužícími po výzkumu ani význačných institucích*), ale připraveni právě na závažný úkol popularizace statistiky ve vědomí veřejnosti s tím, že zároveň nabídnou nejlepším studentům možnost získat základy pro hlubší pochopení metod výzkumu v dalším studiu i rozhodovací znalosti v běžném životě. Vzhledem ke stále aktuálnosti tohoto již 56 let starého příspěvku pro *110th Annual Meeting of the American Statistical Association, Chicago, 28. 12. 1950* si dovoluji shrnout hlavní myšlenky Wilkse do krátkého závěru této části mého vystoupení.

1. Už v roce 1950 považuje Wilks otázku statistického vzdělávání na vysokých školách za největší problém a hlavní úkol pro statistiky ve druhé polovině 20. století. Úkol může být splněn, když se najde schopnost reagovat na úlohu statistiky ve výrobních a obchodních úlohách, jakož i ve všech oblastech výzkumu. Na vzdělávací škále to musí jít dostatečně *dolu a hluboko, protože to je učeno pozdě a povrchně!* Mnozí opouštějí vysokou školu nedotčení nejen speciálně statistikou, ale kvantitativním myšlením vůbec.
2. Podstatou řešení jsou two full years courses, obsahující základy pravděpodobnosti, statistiky, logiky a experimentální filozofie, přičemž je nutné mít aspoň nějaké znalosti z matematiky.
Poznámka PH: Dnes k tomu přistupuje (cituji z materiálů QMSS při ESF) *katastrofický nedostatek kvantitativního vzdělání pracovníků ve společenských vědách*, jakož i zaběhnutá snaha statistiků naučit studenty mačkat *správné* počítačové klávesy a provádět *nejlepší* volby z nabídek statistických paketů. Již mnohokrát jsme diskutovali problémy výuky, ale jakékoli zlepšení nevidím.
3. Elementární kurzy statistiky by měli učit statistici. Vzhledem k tomu, že ti dobří odcházejí jinam, by to mohli být přednostně B. A. či M. A. studenti anebo skupina učitelů, která se právě pro tuto výuku nejlépe hodí.
4. Dvouletý kurz by měl být nejlépe v prvním a druhém ročníku VŠ studia, takže přicházející studenti musí mít už ze střední školy jisté základy, aby to mohli dobře chápat.

5. Pro střední školy to znamená úkol vytvořit časový prostor pro výuku, přičemž si Wilks myslí, že zde by měli matematici upustit od výkladu třeba trigonometrických rovnic, prostorové geometrie, vypustit neefektivní důkazy ve prospěch myšlenek elementární pravděpodobnosti, statistiky a hlavně logiky.

Wilks končí slovy . . . *the challenge is great and it must be met.*

Poznámka PH: Pochopitelně Wilkse tehdy nemohlo napadnout zmínit nebo dokonce doporučit vhodné řešení pro situaci, kdy matematika je zcela vytlačována ze středních i ekonomických VŠ, na gymnázium často existuje jen jako nepovinný předmět, ze kterého tedy většina studentů ani nematuruje a výklad logiky či jiných příbuzných oblastí téměř neexistuje. Netušil, že po 57 letech budeme stát před situací, kdy postoj univerzit či jiných vysokých škol ke kvantitativnímu vzdělání obecně je více než sporný a získat dvouletý plnohodnotný prostor pro výuku nestatistiků téměř nepřichází v úvahu.

Klasický či bayesovský přístup k výkladu elementární statistiky?

Asi před čtrnácti dny po ukončení nepovinného kurzu Bayesovské statistiky jsem dostal pro mne velice příjemné hodnocení a vyjádření jednoho z účastníků tohoto kurzu, který je ve čtvrtém ročníku našeho oboru. Z jeho dvoustránkového dopisu vybírám jednu část, kde říká.

„. . . úplně jsem nevěděl, co bych měl od tohoto předmětu očekávat a spíše jsem si ho zapsal ze zkušenosti s předměty absolvovanými s Vámi. Musím ale říct, že jsem byl velmi příjemně překvapen. Po absolvování předmětu se pro mě sice pan Bayes a jeho přístup ke statistice nenastali ničím, čemu bych bezmezně důvěřoval a zavrhl veškeré poznatky o klasických přístupech, ale znamená to pro mě poznání, že bayesiánství je více než jen Bayesův vzorec a je mnohdy velmi užitečné se na statistiku dívat i přes bayesovské brýle. Navíc to byla opravdu příjemná změna, kdy se člověk může dívat na statistiku v trochu lidštější formě. V rámci výuky na VŠE na to bohužel ve většině případů nezbývá čas. Zajímavý byl pak článek o výuce statistiky už od počátku na bayesovských základech. Z vlastních zkušeností vím, že orientovat se ve statistických základech je pro začátečníka velmi obtížná záležitost. Neříkám, že by se měla změnit koncepce výuky statistiky na bayesovský přístup, ale pro obor Statistika by tento předmět měl být povinný – podat studentům statistický základ tak, jak to bylo v článku pojato, tedy že na většinu věcí

mohou přijít sami. To v klasické statistice z pohledu začátečníka jde opravdu ve velmi málo případech ...“

Za téměř třicet let (postupně menšího až dnes většího) zájmu o bayesovský přístup k pojetí pravděpodobnosti a induktivních, deduktivních a reproduktivních úsudků jsem postupně přecházel od osoby Thomase Bayese a začátků neobayesiánství či pozdějšího bayesiánství, přes bayesovskou teorii, bayesovské výpočty až k bayesovským aplikacím a dnešním snahám vyučovat základy počtu pravděpodobnosti a základů statistiky z bayesovského hlediska. Článků na toto téma mám desítky, ale zde zmiňuji jen jeden, ale pokud jde o bayesovsky orientované knihy, jež mne zaujaly, uvádím je v pořadí jak vycházely:

Edward E. Leamer.: Ad Hoc Inference with Nonexperimental Data. Wiley 1978.

Chamont Wang.: Sense and Nonsense of Statistical inference. Dekker 1993.

José M. Bernardo – Adrian F. M. Smith.: Bayesian Theory. Wiley 1994.

Christian P. Robert.: The Bayesian Choice. Springer 1994.

Donald A. Berry.: Statistics – A Bayesian Perspective. Duxbury Press 1996.

Mike West – Heft Harrison.: Bayesian Forecasting and Dynamic Models 1997.

Bradley Efron.: R. A. Fischer in the 21st Century. Statistical Science 1998.

S. James Press – Judith M. Tanur.: The Subjectivity of Scientists and the Bayesian Approach. Wiley 2001.

William M. Bolstad.: Introduction to Bayesian Statistics. Wiley 2004.

Odmysleme si souboj mezi *subjektivisty* a *objektivisty*, který trval přibližně dvě století a respektujeme současný stav, kdy kritika bayesovského způsobu myšlení je už spíše *umíněností*, částečnou nebo úplnou *neznalostí argumentů* anebo jen *neochotou* některých klasiků respektovat vývoj, ke kterému v této oblasti nesporně došlo. Bayesovský přístup přestal být veřejně kritizován, i když se ještě dnes v soukromých debatách či polemikách setkám s výroky typu ... *je to možná zajímavé, ale já se zabývám něčím jiným; není to můj šálek kávy; příliš tomu nevěřím* ... či s podobnými dalšími. Za zlomový považuji rok 1997, kdy M. Kendall do slavné série knih (pod společným zastřešujícím názvem *Advance Theory of Statistics*) zařadil díl 2e pod názvem *Bayesian Inference*, jehož autorem je *D. W. Lindley*. Je možné bez přehánění říci, že dnes neexistuje významný statistický časopis, který by pravidelně nezařazoval nejrozněji zaměřené články k rozvoji bayesovské statistiky. Výše uvedená kniha o subjektivitě vědců je nádherným důkazem, že dvanáct pro knihu vybraných osobností (Aristoteles, Galileo Galilei, Viliam Harvey, Isaac Newton, Antoine Lavoisier, Alexander von Humboldt, Michael Faraday, Charles Darwin,

Louis Pasteur, Sigmund Freud, Marie Curie a Albert Einstein), všeobecně považovaných za ikony vědy, jednoznačně svým přístupem demonstrují potřebu i schopnost vědeckého využití osobního přesvědčení, intuice, předchozích znalostí. Tito odborníci ve své profesi prokázali význam subjektivity pro získání nových poznatků a schopnost její kombinace s empirickými výsledky bádání.

Přibližně rok 1996 znamenal i začátek období rozsáhlých debat o přednostech různých způsobů výuky statistiky pro nestatisticky, a tedy i o možnosti vyučovat statistiku pro nestatisticky z bayesovského hlediska. Od tohoto roku také začaly častěji vycházet takto zaměřené učebnice kombinované s argumenty ve prospěch bayesovského přístupu, i když takové publikace existovaly už v šedesátých letech (např. učebnice *D. Blackwella: Basic Statistics*. McGraw-Hill 1969). Tyto knihy pochopitelně využívají apriorní znalosti o posuzované skutečnosti a používají Bayesův vzorec jako nástroj kombinace dosavadních znalostí s výsledky provedených pokusů či získaných nových pozorování a zjišťování. Cílem příznivců bayesovského způsobu myšlení v těchto debatách o vhodném způsobu výuky statistiky výuce bylo a je prokázat, že klasická teorie opakovaných pokusů nevyužívá pro výpočet pravděpodobností nic z daného konkrétního výběru, ale opírá se výhradně jen o hůře představitelnou situaci *všech možných výběrů* z dané populace. Bayesovci tvrdí, že výsledky *opakovaných výběrů* nebo pojmy typu výběrové rozdělení jsou pro začátečníka mnohem méně pochopitelné a představitelné, a navíc jejich využití vede k méně přesným výsledkům než jednodušší bayesovský přístup. Podle bayesovců by tedy pro rozhodnutí, jak učit statistiku pro nestatisticky, mělo být podstatné, zda jednodušší, pochopitelnější a přesnější jsou *klasické* úsudky (založené na teorii opakovaných výběrů) anebo bayesovské úsudky opírající se o kombinaci apriorní informace a nově získaných dat při využití Bayesova vzorce pro získání posteriorního rozdělení a jeho charakteristik.

Pro aspoň orientační představu o jedné části bayesovské argumentace si ukažme aspoň jeden z řady příkladů výše jmenované knihy Williama Bolstada (*Introduction to Bayesian Statistics*. Wiley 2004), který se týká odhadu populačního podílu:

Tři studenti byli požádáni, aby vyjádřili svůj postoj k π , což je podíl trvale bydlících osob v Hamiltonu podporujících výstavbu kasina v jejich městě. *Anna* si myslí, že její apriorní průměr (podíl) je 0,2 a její směrodatná odchylka je 0,08. Použití beta rozdělení jako modelu jejího postoje vede k parametrům $a = 4,8$ a $b = 19,2$. *Bart* žije v Hamiltonu teprve krátce, nezná kritické debaty k myšlence výstavby kasina a na dotazovaný podíl nemá žádný názor a nic o něm neví. Zná teorii doporučující jako model beta rozdělení a pro něj parametry jsou $a = b = 1$. *Chris* neumí použít beta roz-

dělení a svůj postoj vyjádřil pomocí vah, které upravil tak, aby získal spojitě apriorní rozdělení. Po úpravách dostáváme apriorní rozdělení ve tvaru

$$g(\pi) = \begin{cases} 20\pi & \text{pro } \pi \text{ od } 0,0 \text{ do } 0,2 \\ 0,2 & \text{pro } \pi \text{ od } 0,2 \text{ do } 0,3 \\ 5 - 10\pi & \text{pro } \pi \text{ od } 0,3 \text{ do } 0,5 \end{cases}$$

Úmyslně je použito apriorní rozdělení ve třech různých podobách a z prvního obrázku (označený jako Figure 8.2 na následující straně) je vidět, že se tato apriorní rozdělení zřetelně liší. Tito tři studenti dostali náhodný výběr $n = 100$ trvale bydlících osob Hamiltonu, ze kterých jich 26 vyjádřilo podporu výstavbě kasina, zatímco zbývajících 74 se vyjádřilo proti. *Anna* má posteriorní rozdělení beta s parametry 30,8 a 93,2, *Bart* má posteriorní rozdělení rovněž beta, ale s parametry 27 a 75, zatímco výpočet posteriorního rozdělení *Chrise* vyžaduje numerickou integraci součinu apriorního rozdělení a věrohodnostní funkce (Bolstad nabízí na internetu dostupné makro Minitabu). Druhý obrázek (označený jako Figure 8.3 na následující straně) ukazuje, že opticky se od sebe posteriorní rozdělení všech tří studentů málo liší a potvrzuje to i tabulka charakteristik i následující tabulka charakteristik posteriorního rozdělení.

Pro větší přehlednost shrnutí hlavních bodů *Bolstada* (týkajícího se odhadu podílu)

(podrobnější velmi jednoduchý a srozumitelný výklad je v příslušné kapitole citované knihy)

- Vztah: *posteriorní rozdělení je úměrné součinu apriorního rozdělení a věrohodnostní funkce* je podstatný pro určení tvaru posteriorního rozdělení. Potřebné konstanty je třeba vypočítat tak, aby se integrál z hustoty pravděpodobnost rovnal jedné.
- Je-li apriorní rozdělení beta s parametry a, b je posteriorní rozdělení rovněž beta s parametry $a+y, b+y$, kde y je počet výskytů sledovaného jevu z n náhodných pokusů (n náhodně vybraných jednotek souboru z populace).
- Nevíme-li nic o π , můžeme použít beta rozdělení s parametry $a = b = 1$.
- Máme-li nějakou apriorní znalost o neznámém podílu, můžeme ji vyjádřit pomocí vah, které lze lineární interpolací převést na spojitě apriorní rozdělení.
- Posteriorní střední hodnota je odhad, který má nejmenší posteriorní čtvercovou chybu. Je to optimální *post-data* odhad.

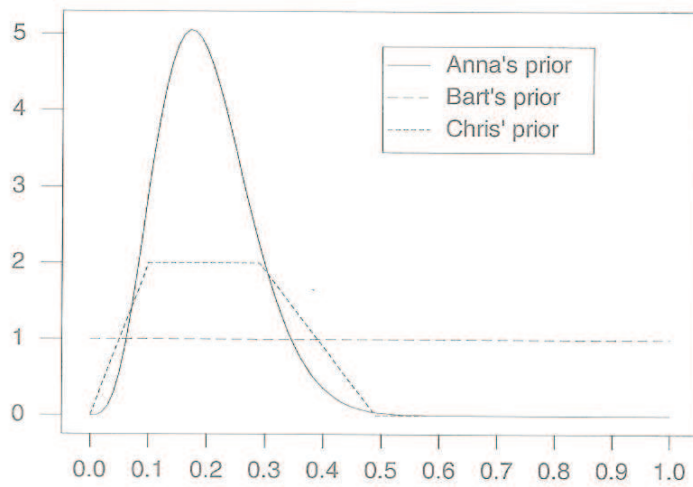


Figure 8.2 Anna's, Bart's, and Chris' prior distributions.

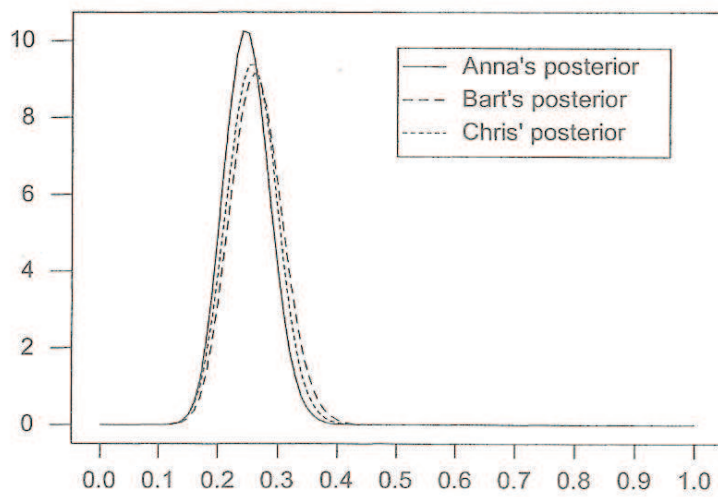


Figure 8.3 Anna's, Bart's, and Chris' posterior distributions.

Osoba	Posteriorní rozdělení.	Střední hodnota	Medián	Směrodatná odchylka	Kvadrilové rozpětí
Anna	$B(30,8;93,2)$	0,248	0,247	0,039	0,053
Bart	$B(27,0;75,0)$	0,270	0,263	0,044	0,059
Chris	Numerické	0,261	0,255	0,041	0,057

Anna, Bart i Chris vypočítali svůj dvoustranný 95% *credible interval* (pojem, který pro odlišení od klasického *intervalu spolehlivosti* Bolstad úmyslně používá). Výpočet je provedený přesně (na základě uvedených hustot) i s využitím normální aproximace, aby bylo možné srovnání rozdílů obou výpočtů. Výsledky jsou v následující tabulce.

Osoba	Posteriorní rozdělení.	<i>Credible Interval</i> – přesně	<i>Credible interval</i> – normální aproximace
Anna	$B(30,8;93,2)$	0,177 až 0,328	0,172 až 0,324
Bart	$B(27,0;75,0)$	0,184 až 0,354	0,183 až 0,355
Chris	Numerické	0,181 až 0,340	0,181 až 0,341

- $(1 - \alpha)100\%$ bayesovský *credible interval* je interval, který má posteriorní pravděpodobnost $1 - \alpha$, že obsahuje hodnotu odhadovaného parametru.

Klasický interval spolehlivosti

Klasické intervaly spolehlivosti všichni dobře známe, takže jakýkoli výklad je snad zbytečný. Vyjděme z toho, že běžném klasickém vnímání je parametr (v tomto případě π alternativního rozdělení) nějaká neznámá *konstanta*. Proti tomu krajní body intervalu spolehlivosti (D, H) jsou před provedením výběru náhodné veličiny, zatímco po provedení výběru jsou to vypočtené hodnoty těchto veličin. Jakmile tedy konkrétní výběr byl proveden už nic náhodného není. Pak vypočítaný interval buď obsahuje neznámou hodnotu parametru či nikoli, ale my nevíme, který z těchto dvou případů nastal. Na tento interval se už tedy dále nemůžeme dívat jako na náhodný. Podle klasického (četnostního) paradigma je správná interpretace, že $(1 - \alpha)100\%$ náhodných intervalů (ze všech možných), vypočítaných tímto způsobem, bude obsahovat skutečnou hodnotu neznámého parametru π . V tomto smyslu máme tedy $(1 - \alpha)100\%$ *důvěru* (úmyslně využívám možnosti neoznačit slovo *confidence* za spolehlivost, ale častějším slovníkovým překladem – poznámka PH), že právě náš interval hodnotu π obsahuje. Bayesovci říkají, že činit pravděpodobnostní úsudky na základě takto pojímaných intervalů spolehlivosti představuje chybnou a zavádějící interpretaci. K tomu se ještě vrátíme v závěrečné diskusi o výhodách a nevýhodách jednotlivých přístupů.

Po použití normální aporoximace interval spolehlivosti pro π je známý běžně používaný interval

$$p \pm u_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}},$$

kde $p = y/n$ je výběrová relativní četnost a $u_{1-\alpha/2}$ je $(1 - \alpha/2)100\%$ kvantil normovaného normálního rozdělení.

Srovnání klasického intervalu spolehlivosti s bayesovským *credible* intervalem
Pravděpodobnostní výpočty pro interval spolehlivosti jsou založené na *výběrovém rozdělení* použité statistiky. Jinými slovy, jak se hodnoty této statistiky mění (liší) ve všech možných výběrech. Z toho vyplývá, že pravděpodobnosti s tím souvisící jsou *pre-data*, protože nezávisí na konkrétním posuzovaném výběru. To je zásadní rozdíl ve srovnání s bayesovským *credible* intervalem, který se určuje na základě posteriorního rozdělení, takže má přímou pravděpodobnostní interpretaci (ve smyslu bayesovského *degree of belief*), která je podmíněná napozorovanými daty. To je pro výzkumníka velmi užitečné.

Může se (ale nemusí) zajímat se i o skutečnosti, které nenastaly, ale mohly nastat. Bayesovský přístup je *post-data*, protože sumarizuje informaci, která je získaná z dat, jež bylo úkolem posoudit.

V našem příkladu klasický 95% interval spolehlivosti je

$$0,26 \pm 1,96 \sqrt{\frac{0,26 \cdot 0,74}{100}},$$

což je tedy interval od 0,174 do 0,346. Rozdíly nejsou velké pro $n = 100$, ale přesto je vidět, že z hlediska délky intervalu (přesnosti odhadu), je klasický interval srovnatelný jen s intervalem *Barta* (který žádnou apriorní představu neměl), ale je zřetelně horší než *credible* intervaly obou zbývajících studentů (kteří jistou výchozí představu využívali).

Na závěr si ještě jen velice stručně připomeňme hlavní a nejčastější argumenty, které byly probírány v referátech i diskusi po přednáškách Donalda A. Berryho, Davida S. Moora a Jima Alberta při příležitosti setkání statistiků v Chicagu v srpnu 1996 a uvedl je *The American Statistician* v čísle 3 v roce 1997 na str. 241 – 268.

Výchozí referát přednesl Donald A. Berry pod názvem *Teaching Elementary Bayesian statistics with Real Applications in Science*. Berry je i autorem výše uvedené učebnice základů statistiky z bayesovského hlediska a před přibližně deseti roky už byl jedním z nejvýraznějších propagátorů výuky kurzu statistiky z bayesovského hlediska. Berry polemizuje s názorem, že bayesovská statistika není vhodná pro výuku elementárního kurzu. Snaží se ukázat výhody, které může právě tento přístup studentům přinést. Podle jeho názoru je bayesovský přístup ve vědě vhodnější než klasický. Platí to především ve výuce, kdy lze jen velmi těžko přenést na studenty svoje osobní zkušenosti a cit pro volbu vhodné metody, případně její modifikaci při řešení konkrétní úlohy klasickými postupy. Postupně se zabývá nejčastějšími důvody skutečnosti, že existuje jen málo základních (všichni víme, že i pokročilých – PH) bayesovských kurzů. Říká, že (srovnejme s ČR – PH) na většině univerzit existují bayesovské kurzy, ale jen velmi málo z nich nabízí tyto kurzy i začínajícím. Dokonce i skalní zastánci bayesovského přístupu vyučují základní kurzy většinou z klasického pohledu. Jmenuje tyto nejčastěji uváděné důvody této skutečnosti.

1. Bayesovská statistika je příliš komplikovaná na to, aby byla přednášena v základních kurzech.

Právě naopak. Bayesovský přístup je založen pouze na některých základních myšlenkách, ze kterých se odvíjí vše ostatní. Studenti musí být schopni sledovat logický vývoj problému a musí být ochotni myslet, ale to je vše. Na rozdíl od logického vývoje a intuitivní interpretace výsledků bayesovského přístupu, jsou klasické metody téměř nepochopitelné i pro ty nejlepší studenty. Například intervaly spolehlivosti. Mnoho vyučujících (a dokonce i některé knihy) se dopouští nepřesností při interpretaci intervalů spolehlivosti. Vypočítat interval spolehlivosti je jednoduché, ale téměř každý (kromě odborníků) se domnívá, že 95% interval spolehlivosti např. 2,6 až 7,9 říká, že s pravděpodobností 95% zjišťovaný parametr leží v intervalu od 2,6 do 7,9. p-hodnoty jsou stejně podivné a všech (snad kromě statistiků) dávají klasickým výsledkům inverzní bayesovskou interpretaci. Někteří statistici se domnívají, že bayesovská statistika je obtížná, protože ji sami nerozumějí, nebo zastávají názor, že bayesovský přístup má být vyučován v pokročilých kurzech, podobně jako regresní analýza nebo neparametrické metody. Ve skutečnosti se však bayesovský pohled uplatňuje na celou statistiku a také na celou vědu.

2. Bayesovský přístup je subjektivní.

Ano je aspoň podle mého názoru (říká Berry). Základem všech bayesovských úsudků a rozhodnutí je současné rozdělení různých neznámých (myslí se apriorní, pokud se vztahuje k začátku experimentu a posteriorní, pokud závisí na výsledcích pokusu). Každý jednotlivec má své vlastní rozdělení pravděpodobností neznámých. Posteriorní rozdělení dvou lidí účastnících se jednoho pokusu jsou obvykle bližší než rozdělení výchozí, ale nikdo nemůže zaručit, že se názory lidí budou shodovat, a to dokonce ani pod tíhou přesvědčivých důkazů. V tomto smyslu bayesovský princip odpovídá vědeckému přístupu. Mezi lidmi převládá představa, že věda je objektivní. Proto by i statistici měli být objektivní jako u klasického přístupu. Tato představa je však mylná ve většině vědeckých přístupů se používají výrazy jako ... bylo všeobecně přijato, věřili jsme nebo pokud věříme. To, co je ve vědě známo, je obvykle to, čemu většina vědců věří, i když možná ne všichni. Věda se vyvíjí upravováním a opravováním názorů po získání nových informací. Vědci málokdy používají Bayesův vzorec, ale kdyby ho použili byla by jejich komunikace mnohem efektivnější.

3. Bayesovci se nemohou shodnout na výchozím rozdělení.

Ano nemohou. Neexistuje žádné jediné apriorní rozdělení, které by bylo vhodné pro každou situaci. Pokud by existovalo, ztratily by se tím mnohé výhody bayesovského přístupu. Bayesovci používají informace, získané mimo daný pokus. Tyto informace mohou být těžko shodné ve dvou situacích. Touha po jediném výchozím rozdělení je přenesena z klasického přístupu a svědčí o nepochopení velmi důležitého rozdílu mezi těmito dvěma pohledy: Bayesovský přístup používá i k vytváření úsudku všechny dostupné informace, zatímco klasický přístup využívá pouze data získaná experimentem nebo zjišťováním. Neexistence předepsaného výchozího rozdělení je velkou silou bayesovského pohledu, protože zabraňuje mechanické analýze sebraných dat. Statistici jsou nuceni zjistit vše, co znají vědci i jak k tomu došli a to zvyšuje míru spolupráce mezi příslušníky jiných oborů a statistiky. Také to lépe umožňuje statistikům navrhnout další směr vývoje pokusů, které by jinak nemuseli provádět.

4. Návrh nového kurzu vyžaduje vyšší úsilí.

To je zajisté pravda. A možná to je jeden ze základních důvodů, proč se základní bayesovské kurzy nepřednášejí. Ale to se možná brzy změní.

(Zatím se tak nestalo i když sylaby a učebnice přibývají – poznámka PH.)

5. Studenti potřebují znát klasické metody a přístupy.

Dnešní studenti (říká Berry v roce 1996) dostudují v době, kdy klasické metody v Prazi převažují. Je student, který absolvoval kurzy z bayesovské statistiky znevýhodněn? Například intervaly spolehlivosti. Studenti bayesovských kurzů se naučí určovat posteriorní rozdělení a pravděpodobnostní intervaly těchto rozdělení. Je snadné vysvětlit, že interval spolehlivosti je vlastně pravděpodobnostní interval při výchozím rovnoměrném rozdělení. Podobný vztah existuje i pro testování hypotéz. Ocitnou se vlastně ve výhodě. Vědí, že obvyklá (a nesprávná) interpretace intervalu spolehlivosti (že zjištěný interval obsahuje neznámý parametr s jistou pravděpodobností) platí pouze při zvláštních výchozích informacích. Někteří lidé tvrdí, že bayesovský přístup je ve vědě málo využíván, a proto by neměl být vyučován. Tento argument je nejen nepodstatný, ale navíc ani není pravdivý. Vědci sice Bayese většinou odmítají, ale obvykle sami uvažují jako Bayes, ať již jeho vzorec znají či nikoli (viz uvedená kniha o subjektivitě vědců PH). Kupříkladu si upravují názor podle výsledků pokusu.

6. Neexistují žádné vhodné výukové materiály.

*To může být problém. Učebních textů opravdu není mnoho, ale tento problém se možná podaří časem odstranit. Tím se však dostáváme do určitého kruhu. Nedo-
statek textů vede k vyučování klasického přístupu, a tím pádem žádné bayesovské texty nejsou zapotřebí.*

Pokračování Berryho příspěvku i dalších dvou, které přednesli (zastánce klasického způsobu výuky) David S. Moore pod názvem Bayes for Beginners? Some Reasons to Hesitate a (rovněž příznivec bayesovského pohledu) Jim Albert pod názvem Teaching Bayes' Rule: A Data-Oriented Approach, jakož i reakce v následné diskusi si nechám v případě zájmu a času až pro samotnou přednášku. Nemyslím však, že je to nutné, protože zájemci mají možnost si vše sami důkladně přečíst a zaujmout názor podle originálu, což je určitě vhodnější.

Adresa: Petr Hebák, Katedra statistiky a pravděpodobnosti, Fakulty informatiky a statistiky, Vysoká škola ekonomická v Praze

E-mail: hebak@centrum.cz

Telefon: +420 606 657 456

SPOLUPRÁCE MEZI ČSÚ A UNIVERZITOU TOMÁŠE BATI VE ZLÍNĚ

Pavel Hrbáček a Pavel Stříž

Abstract: The article informs about three-year cooperation between Czech Statistical Office in Zlín and Tomas Bata University in Zlín. At the beginning of this cooperation, both institutions agreed with and signed the official document listing specific and also general items of the agreement.

1. O spolupráci

V polovině září 2004 navštívil pracoviště informačních služeb tehdejší Krajské reprezentace Českého statistického úřadu ve Zlíně doktorand Fakulty managementu a ekonomiky Univerzity Tomáše Bati ve Zlíně ing. Pavel Stříž, za účelem získání statistických podkladů pro zpracování své disertační práce. Z následné diskuse, která pak proběhla u ředitele krajské reprezentace ČSÚ (dále jen KR) ing. Pavla Hrbáčka bylo nejen dohodnuto předání potřebných dat ke zpracování zadaného úkolu, ale nastínily se i možnosti spolupráce mezi oběma institucemi, jako např. poskytování dat studentům pro vypracování seminárních nebo diplomových prací, bezplatné předávání potřebných informačních materiálů pro studijní potřeby fakulty, hostování zaměstnance ČSÚ v předmětu Metody statistické analýzy, či přidání odkazu internetových stránek KR Zlín na webové adrese fakulty.

2. O dohodě

Ještě koncem měsíce proběhlo na půdě zlínské univerzity první oficiální jednání představitelů fakulty a KR, na kterém zástupci obou subjektů potvrdili velký zájem o spolupráci s tím, že kromě již sjednaných okruhů by obě strany dle potřeb a svých možností vzájemně spolupracovaly i v dalších oblastech (hledání vhodných témat pro zpracování analytických publikací a studentských prací, konzultování možných přístupů při vyhodnocování statistických dat, využívání vhodných matematicko-statistických metod v regionálních analýzách, účastí na vybraných jednorázových akcích druhé strany aj.).

Vzhledem k tomu, že se jednalo o poměrně široký okruh zájmů, bylo dohodnuto, že se sjedná písemná smlouva o spolupráci. Návrh dohody se při konzultacích musel několikrát přepracovávat, to však neovlivnilo postupné naplňování předjednané formy spolupráce. V první polovině roku 2005 se

KR prezentovala v první hodině letního semestru studentům 1. a 2. ročníku studia v předmětech Metody statistické analýzy a Aplikovaná statistika. Fakulta vypsalala bakalářské a diplomové práce z oblasti demografické statistiky a pomáhala při stanovení vah pro párové srovnávání u demografických ukazatelů do publikace KR Demografický, sociální a ekonomický vývoj Zlínského kraje v letech 2000 až 2004.

Předseda ČSÚ, ing. Jan Fischer, CSc., souhlasil s písemnou dohodou s tím, že bude uzavřena na nejvyšší úrovni, to je mezi ČSÚ a univerzitou. Rektor univerzity, prof. Ing. Petr Sába, CSc., s tímto návrhem rovněž souhlasil, takže po dalších dílčích textových úpravách a zapracování připomínek právních oddělení byl dokument připraven k podpisu. Bylo dohodnuto, že k oficiálnímu setkání představitelů obou institucí a k slavnostnímu podepsání dokumentu dojde na půdě univerzity.

Vzhledem k vysokému pracovnímu vytížení předsedy a rektora se nakonec podařilo zorganizovat setkání až na třetí termín 23. května 2005. Za ČSÚ se s panem předsedou zúčastnili tohoto slavnostního aktu ještě vrchní ředitel sekce regionálních orgánů ing. Jiří Rolenc, ředitel KR ing. Pavel Hrbáček a vedoucí oddělení informačních služeb KR ing. Soňa Vařeková. Delegaci univerzity vedenou rektorem tvořili dále proděkan Fakulty managementu a ekonomiky pro vědecko-výzkumnou činnost a propagaci doc. Ing. Roman Bobák, Ph.D. (za nemocného děkana doc. PhDr. Vnislava Nováčka, CSc.), dále ředitel ústavu informatiky a statistiky doc. Ing. Rudolf Pomazal, CSc., a lektor statistických předmětů ing. Pavel Stříž. Po úvodních slovech nejvyšších představitelů těchto institucí a podepsání dohody, následovala diskuze, při které obě strany mimo jiné vysoce vyzdvihly možnosti spolupráce, a na závěr proběhla tisková konference s novináři. O uzavření dohody byly zveřejněny tiskové zprávy a také informace na intranetu ČSÚ a v místním univerzitním časopise Universalia.

3. Forma a způsoby spolupráce

Postupně se naplňovaly a dále rozvíjely dohodnuté formy a způsoby spolupráce:

- na konferenci Ústavu veřejné správy a regionálního rozvoje pracovníků UTB vyšel článek ing. Stříže a ing. Kasala z FaME o spolupráci mezi UTB a ČSÚ a o vydaném Lexikonu obcí ČR,
- děkanovi FaME byl ředitelem KR předán dárkový výtisk publikace Lexikon obcí ČR a Pramenné dílo ze sčítání lidu domů a bytů v ČR 2001,
- ČSÚ Zlín průběžně předává do knihovny univerzity vlastní publikace v tištěné i elektronické podobě,

- fakulta vytiskla skriptu pro studenty „Metody statistické analýzy“ s desetistránkovou přílohou prezentace o ČSÚ,
- fakulta nabídla možnost výuky předmětu Metody statistické analýzy zaměstnanci ČSÚ Zlín,
- pro studenty denního a kombinovaného studia připravil ČSÚ Zlín na jaře 2006 i v letošním roce jednohodinové prezentace o činnosti ČSÚ se zaměřením na krajské pracoviště (první ročníky) a o používaných databázích na krajských pracovištích (druhé ročníky),
- v loňském zimním semestru se poprvé uskutečnila prezentace ČSÚ pro studenty 4. ročníku na téma Zahraniční databáze se zaměřením na evropská data v rámci předmětu Ekonometrie v podání ing. Martina Černého z odboru veřejných databází ČSÚ. Přitom bylo dohodnuto její pravidelné každoroční opakování,
- na webových stránkách ČSÚ Zlín mají studenti FaME odkaz na soubory prezentací, které jim byly odpřednášeny a které si mohou stáhnout,
- po prezentacích vždy vzrostl zájem studentů o statistické informace, který se mimo jiné projevil v posledních dvou letech ve zvýšené návštěvnosti webových stránek ČSÚ Zlín v měsících únor až duben,
- fakulta na vyžádání ČSÚ Zlín vypracovala další posudek ke stanovení vah při párovém srovnávání požadovaných ukazatelů tentokrát pro další statistickou publikaci Vývoj lidských zdrojů ve Zlínském kraji v letech 2000 až 2005,
- zástupce fakulty je zván na veřejné prezentace ČSÚ ve Zlíně k vydávaným publikacím, naopak zástupce ČSÚ se zúčastňuje konferencí, které pořádá fakulta,
- v dubnu 2007 se uskutečnilo setkání děkana fakulty a regionálního zmocněnce ČSÚ,
- poslední akcí, která byla společná, byl příspěvek na semináři o výuce a aplikacích statistiky STAKAN 2007, který pořádala Česká statistická společnost ve dnech 25. – 27. května na Rusavě (okres Kroměříž). V prezentaci pro kantory statistiky byly zmíněny zkušenosti, úspěchy i neúspěchy za poslední tři roky, kdy byla připravována a realizována dohoda o spolupráci mezi oběma subjekty.

4. Shrňme-li

Bilancujeme-li poslední dva roky od oficiálního podepsání dohody, musíme ale i konstatovat, že ne všechny přijaté závazky se plní. Zatím se nám dobře

nedaří získat studenty ke zpracování bakalářské nebo diplomové práce ze statistické oblasti a naplňovat dohodu v oblasti analyticko-publikační činnosti, to je zpracování společných materiálů, využívání vhodných matematicko-statistických metod v analýzách ČSÚ Zlín, či ve společném hledání nových témat pro zpracování analytických materiálů nebo studentských prací s regionální nebo euroregionální tematikou.

Dohodu ale nepojímáme jako neměnnou, naopak může být dále rozvíjena, doplňována o nové iniciativy, resp. mohou být některá ustanovení dohody vypuštěna, pokud to obě strany uznají za vhodné.

E-mail: pavel.hrbacek@czso.cz, striz@fame.utb.cz

Příloha: Dohoda o spolupráci

Mezi subjektem 1

se sídlem: / jehož jménem jedná: / IČO: / (dále jen) / na straně jedné
a subjektem 2

se sídlem: / jehož jménem jedná: / IČO: / (dále jen) / na straně druhé.

Realizace spolupráce: krajská reprezentace (dále jen) a fakulta (dále jen).

Odpovědní pracovníci: osoba 1 a osoba 2.

I. Předmět dohody

Účelem této dohody je vytvoření užších všestranných vazeb mezi oběma stranami v oblasti využití, zpracování a prezentace statistických informací charakterizujících region a jeho postavení v České republice.

V rámci dohody je realizována vzájemná výměna potřebných informačních a studijních materiálů.

II. Hlavní formy spolupráce

Subjekt 1 se zavazuje:

- Poskytovat statistické informace a anonymní údaje v tištěné i elektronické podobě, a to v souladu se zákonem č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů a zákonem č. 101/2000 Sb., o ochraně osobních údajů a o změně některých zákonů, ve znění pozdějších předpisů. Předmětem této dohody není poskytování důvěrných statistických údajů.

- Informovat studenty o tom, za které oblasti a v jakém územním členění má data subjekt 1 k dispozici, k tomuto připravit jednou v roce prezentaci pro studenty 1. ročníku subjektu 2.
- Spolupracovat se subjektem 2 při navrhování témat pro bakalářské a diplomové práce studentů s ohledem na dostupnost dat v požadovaném členění.
- Metodicky pomáhat studentům a lektorům při získávání statistických dat pro bakalářské, diplomové a disertační práce.
- Umožnit lektorům publikování odborných článků se statistickou tematikou v časopisu Statistika.

Subjekt 2 se zavazuje:

- Zapojoovat se do oponentních řízení publikací regionálního kraje.
- Odborně a metodicky pomáhat při využívání matematicko-statistických metod v analytických materiálech subjektu 1.
- Konzultovat se subjektem 1 možné přístupy při vyhodnocování statistických dat v připravovaných rozbořech a analýzách subjektu 1.

Obě strany se zavazují:

- Spolupracovat na zpracování předem dohodnutých společných publikací v souladu s právními předpisy.
- Společně hledat nová vhodná témata pro zpracování analytických publikací a studentských prací s regionální i euroregionální problematikou.
- Vzájemně se informovat a účastnit se vybraných jednorázových akcí druhé strany (např. odborných konferencí, seminářů, prezentací, setkání v regionu – možné využití poslucháren subjektu 2).
- Při zveřejňování informací ze zdrojů druhé strany uvádět stranu jako zdroj informací.

III. Závěrečné ujednání

Tato dohoda se uzavírá na dobu neurčitou a nabývá účinnosti dnem podpisu oběma stranami. Vyhotovuje se v šesti vyhotoveních, tři pro každou stranu.

Změny a doplňky této dohody lze provádět pouze formou písemných dodatků, schválených a podepsaných oběma stranami.

Dohodu lze ukončit se souhlasem obou stran na návrh kterékoliv z nich s tím, že rozpracované akce (úkoly) se reálně dokončí.

Dohoda nabývá platnosti a účinnosti dnem podpisu obou stran.

Kde, kdy, osoba 1, osoba 2 a jejich podpisy, případně razítka.

NIEKOLKO POZNÁMOK K VÝUČBE ZÁKLADNÉHO KURZU STATISTIKY

Jozef Chajdiak

Úlohou štatistiky je ukázať svet taký aký je! Štatistika je súčasťou procesu rozhodovania – nerozhoduje, dáva len podklady pre rozhodovanie, podporuje rozhodovanie:

- dáva číselné podklady,
- poukazuje na atypické merania a javy,
- poukazuje na závislosť resp. nezávislosť a jej mieru,
- prezentuje hromadne napozorované údaje,
- prezentuje trend a sezónnosť vývoja,
- modeluje stav alebo vývoj skúmaných objektov,
- ...

Štatistika nachádza svoje vyjadrenie v štatistickom skúmaní. To má:

- cieľ: podpora procesu rozhodovania,
etapu zisťovania,
etapu spracovania,
etapu rozboru,
- odporúčania pre rozhodovací proces.

V procese výučby sa čiastočne vytráca zdôraznenie faktu, že štatistika je súčasťou procesu rozhodovania, jej cieľom je podpora procesu rozhodovania a výsledkom sú odporúčania pre rozhodovací proces.

V rozhodovacom procese principiálnym je potreba sa rozhodnúť. Existuje viacero schém prijatia rozhodnutia. Jedna z nich využíva štatistické podklady. Úlohou je ukázať svet taký aký je a odhadnúť, aký bude svet po realizácii prijatého rozhodnutia.

V etape zisťovania treba špecifikovať:

- množinu zisťovaných znakov (model „sveta“),
- spôsob zisťovania (úplné a neúplné zisťovanie),
 - výberové zisťovanie (poznáme pravdepodobnosť zaradenia každej jednotky do vyberanej vzorky),
 - ostatné neúplné zisťovania.

Problematika špecifikácie množiny analyzovaných znakov sa čiastočne prenáša na iné disciplíny (na tie, ktorých sa analýza obsahovo dotýka). Na druhej strane, bez množiny zisťovaných znakov nemáme čo zisťovať, potom analyzovať a nakoniec vypracovávať podporné stanoviská pre rozhodovanie. Tak špecifikovanie množiny zisťovaných znakov je podstatnou súčasťou štatistiky.

Štatistika je súčasťou rozhodovacieho procesu.

Z tohto pohľadu je dôležité, že:

- Výsledky analýz (a teda aj odporúčania pre rozhodovanie) sú tak kvalitné, ako sú kvalitné zistené a spracované údaje, alebo horšie!
- Nezistenie, chybné zistenie, vynechanie, zmena zisteného znamená analýzu iného sveta než v skutočnosti je!

Môžeme podrobnejšie pozerať aj na obsah (kvalitu) analyzovaných údajov, ktoré obsahujú:

- **objektívny odraz** obsahu analyzovanej skutočnosti a **chyby**:
- náhodná chyba ($E[e] = 0$),
- systematická chyba (záujmy),
- nepoznanie/skreslenie časti skutočnosti (nedostatočné poznanie).

V etape spracovania sa realizuje:

- zaznamenanie zistených hodnôt,
- kontrola zistených hodnôt,
- korekcie chybných resp. nejasných hodnôt,
- vytvorenie počítačového súboru údajov.

Z tohto pohľadu sa zdá potrebné presunúť časť výučby z oblasti IT (vytváranie, napĺňanie a vyberanie údajov z databáz resp. databáňk) do štatistiky.

Štatistická podstata údajov:

- **každý** údaj patrí do nejakého súboru údajov,
- hodnoty v súbore údajov majú nejaké **rozdelenie**,
- o rozdelenia hodnôt údajov sa modelujú zákonom rozdelenia pravdepodobností výskytu hodnôt (hodnota x sa vyskytuje s pravdepodobnosťou $p(x)$),
- zákon rozdelenia pravdepodobností výskytu hodnôt (hodnota x sa vyskytuje s pravdepodobnosťou $p(x)$) má silu **prírodného** zákona:

- nevieme aká hodnota sa vyskytne v individuálnom meraní, ale v stovke meraní podiel výskytu hodnoty x bude okolo $p(x)$,
- štatistická podstata znamená „**okolo**“ $p(x)$.

Rastom (n) počtu napozorovaných hodnôt empirické zistenia (u_n) **konvergujú** k objektívnym hodnotám (Q).

$$\lim_{n \rightarrow \infty} \Pr\{|u_n - Q| < \varepsilon\} = 1$$

Vzťah matematiky a štatistiky:

- sú samostatné disciplíny,
- pravdepodobnosť je čisto matematická disciplína,
- matematická štatistika predstavuje aplikáciu matematiky v štatistike resp. riešenie štatistických úloh matematikou,
- zisťovanie a spracovanie údajov sú čisto štatistické disciplíny.

Niekedy máme sklony zjednodušujúco za štatistiku považovať len matematickú štatistiku.

Uvedené poznámky nie sú výkladom Ústavného súdu k problematike čo a ako učiť v základnom kurze štatistiky. Sú len autorovým odporúčaním pri rozhodovaní sa k niektorých aspektom výučby štatistiky. „Čo a ako“ sa musí rozhodnúť učiteľ.

Adresa: Doc. Ing. Jozef Chajdiak, CSc., Statis Bratislava

E-mail: chajdiak@statis.biz

DATA MINING AND SOFTWARE TOOLS

Ing. Petr Klímek, Ph.D.

Adresa: UTB, Fame, ÚIS

E-mail: klimek@fame.utb.cz, pklimek@seznam.cz

1. Introduction

1.1. Data Mining: In Brief

The potential returns of data mining are enormous. Innovative organisations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimise losses due to error or fraud.

Data mining is a process that uses a variety of data analysing tools to discover patterns and relationships in data that may be used to make valid predictions. The first and simplest analytical step in data mining is to **describe** the data – summarise its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together). As emphasised in the section on the data mining process, collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. You must **build a predictive model** based on patterns determined from known results, then test that model on results outside the original sample. A good model should never be confused with reality (you know a road map isn't a perfect representation of the actual road), but it can be a useful guide to understanding your business. The final step is to empirically **verify** the model. [2], [6]

1.2. Data Mining: What It Can't Do

Data mining is a tool, not a magic wand. It won't sit in your database watching what happens and send you an e-mail to get your attention when it sees an interesting pattern. It doesn't eliminate the need to know your business, to understand your data, or to understand analytical methods. Data mining assists business analysts with finding patterns and relationships in the data – it does not tell you the value of the patterns for the organisation. Furthermore, the patterns uncovered by data mining must be verified in the real world. [3]

Remember that the predictive relationships found via data mining are not necessarily causes of an action or behaviour. For example, data mining

might determine that males with incomes between \$50,000 and \$65,000 who subscribe to certain magazines are likely purchasers of a product you want to sell. While you can take advantage of this pattern, say by aiming your marketing at people who fit the pattern, you should not assume that any of these factors cause them to buy your product.

1.3. Data Mining, Machine Learning and Statistics

Data mining takes advantage of the advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification. Both communities have made great contributions to the understanding and application of neural networks and decision trees. [7]

Data mining does not replace traditional statistical techniques. Rather, it is an extension of the statistical methods that are in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on an elegant theory and analytical methods that worked quite well on modest amounts of data being analysed. The increased power of computers and their lower cost, coupled with the need to analyse enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions. [4]

New techniques include relatively recent algorithms like neural networks and decision trees [1], and new approaches to older algorithms such as discriminant analysis. By virtue of bringing to bear the increased computer power on the huge volumes of available data, these techniques can approximate almost any functional form or interaction on their own. Traditional statistical techniques rely on the modeler to specify the functional form and interactions.

2. Data Mining and Hardware/software Trends

The drop in the cost of computer processing has been equally dramatic. Each generation of chips greatly increases the power of the CPU, while allowing further drops on the cost curve. This is also reflected in the price of RAM (random access memory), where the cost of a megabyte has dropped from hundreds of dollars to around a dollar in just a few years. While the power of the individual CPU has greatly increased, the real advances in scalability stem from parallel computer architectures. Virtually all servers today support multiple CPUs using symmetric multi-processing, and clusters of these

SMP servers can be created that allow hundreds of CPUs to work on finding patterns in the data.

2.1. Data Mining Applications

Data mining offers value across a broad spectrum of industries. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances which may be candidates for development as agents for the treatment of disease. [8]

2.2. Successful Data Mining

There are two keys to success in data mining. First, is coming up with a **precise formulation** of the problem you are trying to solve. A focused statement usually results in the best payoff. The second key is using the **right data**. After choosing from the data available to you, or perhaps buying external data, you may need to transform and combine it in significant ways.

The more the model builder can "play" with the data, build models, evaluate results, and work with the data some more (in a given unit of time), the better the resulting model will be. Consequently, the degree to which a data mining tool supports this interactive data exploration is more important than the algorithms it uses. [5]

2.3. Data Mining Methodology

Recognising that a systematic approach is essential to successful data mining, many vendor and consulting organisations have specified a process model designed to guide the user (especially someone new to building predictive models) through a sequence of steps that will lead to good results.

SPSS uses the 5A's - Assess, Access, Analyze, Act and Automate - and SAS uses SEMMA - Sample, Explore, Modify, Model, Assess.

Recently, a consortium of vendors and users consisting of NCR Systems Engineering Copenhagen (Denmark), Daimler-Benz AG (Germany), SPSS/Integral Solutions Ltd. (England) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands) has been developing a specification called CRISP-DM - Cross-Industry Standard Process for Data Mining. CRISP-DM is similar to process models from other companies including the one from Two Crows Corporation. As of September 1999, CRISP-DM is a work in progress. It is a good start in helping people to understand the necessary steps in successful data mining. [6]

The basic steps of data mining for knowledge discovery are:

1. Define business problem
2. Build data mining database
3. Explore data
4. Prepare data for modelling
5. Build model
6. Evaluate model
7. Deploy model and results

2.4. Strengths and Weaknesses of Leading Data Mining Tools

We have compared in this paper, 8 leading data mining tools. Table 1 shows the list of them, describes their developers, URL and their version. In Table 2 we can see the strengths and weaknesses of them.

Product	Company	URL	Version
Clementine	SPSS	http://www.spss.com/clementine	4
Enterprise Miner	SAS institute	http://www.sas.com/software/components/miner.html	Beta
Intelligent Miner	IBM	http://www.software.ibm.com/data/iminer/	2
Mine Set	Silicon Graphics	http://www.sgi.com/Products/software/MineSet	2,5
Model1	Group1/Unica Technologies	http://www.unica-usa.com/model1.htm	3,1
PRW	Unica Technologies	http://www.unicausa.com/prodinfo.htm	2,1

Product	Company	URL	Version
NeuroShell	Ward Systems Group, Inc.	http://www.wardsystems.com/neuroshel.htm	3
S-Plus	Mathsoft	http://www.mathsoft.com/splus/	4

	Strengths and Weaknesses	
Product	Strengths	Weaknesses
Clementine	visual interface; algorithm breadth	scalability
Enterprise Miner	depth of algorithms	harder to use
Intelligent Miner	graphical tree/cluster output	few algorithm options
Mine Set	data visualisation	no model export
Model1	easy to use	really a vertical tool
PRW	extensive algorithms	limited visualisation
NeuroShell	multiple neural architectures	only neural networks
S-Plus	programable, extendable	limited inductive models

3. Summary

Data mining offers great promise in helping organisations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs.

Building models is only one step in knowledge discovery. It's vital to accurately collect and prepare the data, and to check the models against the real world. The "best" model is often found after building models of several different types, or by trying different technologies or algorithms.

Choosing the right data mining products means finding a tool with good basic capabilities, an interface that matches the skill level of the people who'll be using it, and features relevant to your specific business problems. After you've narrowed down the list of potential solutions, get a hands-on trial of the likeliest ones.

References

- [1] COLES, S., ROWLEY J. Revisiting decision Trees. Management Decision, 1995, Vol. 33, No. 8, p. 22-26. ISSN 0025-1747
- [2] FAYYAD, U. M., PIATETSKI-SHAPIO, G. Advances in Knowledge Discovery and Data Mining. MIT Press, 1996. ISBN 0262560976
- [3] GARGANO, M.L., RAGGAD, B.G. Data mining-a powerful information creating tool. OCLC Systems and Services, 1999, Vol. 15, No. 2, pp. 91-90. ISSN 1065-075X
- [4] HAN, J., KAMBER, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2000. ISBN 1558604898
- [5] LEE, S. J., SAU, K. A review of data mining techniques. Industrial Management & Data systems, 2001, Vol. 2001, No. 1, pp. 44-46. ISSN 0263-5577
- [6] PARR RUD, O. Data mining Cookbook: Modelling Data for marketing, Risk and Customer Relationship. Management. John Wiley & Sons, 2000. ISBN 0471385646
- [7] VENUGOPAL, V., BAETS, W. Neural networks and statistical Techniques in Marketing Research: A Conceptual Comparison. Marketing Intelligence and Planning, 1994, Vol. 12, No. 7, pp. 30-38. ISSN 0263-4503
- [8] WADDELL, D., SOHAL, A.S. Forecasting: The Key to Managerial Decision Making. Management Decision, 1994, Vol 32, No. 1, pp. 41-49. ISSN 0025-1747
- [9] Relevant URLs are mentioned in Table 1.

HISTORIE A SOUČASNOST VÝUKY STATISTIKY NA FAME, UTB VE ZLÍNĚ

Petr Klímek

Abstrakt: Tento příspěvek si klade za cíl seznámit čtenáře s výukou statistiky a dalších souvisejících předmětů na Fakultě managementu a ekonomiky Univerzity Tomáše Bati ve Zlíně, kterou zajišťují pracovníci Ústavu informatiky a statistiky. V úvodu je stručně popsána historie vývoje UTB a FaME. V dalších částech příspěvku jsou uvedeny studijní opory, programy a další detaily o statistických předmětech. Taktéž jsou v závěru vyjmenovány potíže, se kterými se vyučující setkávají ve své pedagogické praxi.

Klíčová slova: UTB, FaME, ÚIS, metody statistické analýzy, aplikovaná statistika, počítačové zpracování dat.

1. O Univerzitě Tomáše Bati ve Zlíně

Univerzita Tomáše Bati ve Zlíně (UTB) byla zřízena ke dni 1. 1. 2001. UTB ve Zlíně je v současné době tvořena pěti fakultami: Fakultou technologickou, Fakultou managementu a ekonomiky, Fakultou multimediálních komunikací, Fakultou aplikované informatiky a Fakultou humanitních studií. Zájemcům o studium nabízí univerzita celkem 61 studijních oborů v 32 akreditovaných studijních programech. Všechny studijní programy byly nově akreditovány nebo reakreditovány ve smyslu Boloňské deklarace, jako navazující dvou-, resp. třístupňové studium bakalářské, magisterské a doktorské (3+2+3). Na UTB působí celkem 343 akademických a vědeckých pracovníků. Personální strukturu tvoří 40 profesorů, 63 docentů, 133 odborných asistentů, 48 asistentů, 34 lektorů a 25 vědeckých pracovníků. Celkový počet zaměstnanců je 649 (k 25. 8. 2006).

V akademickém roce 2006/07 na UTB ve Zlíně studuje 10 158 studentů. UTB má kvalitní informační zázemí. K univerzitní počítačové síti je připojeno 1720 počítačů, z toho více než 600 v počítačových učebnách a internetových studovnách. Vysokoškolské koleje mají kapacitu 758 přípojných míst. Studenti mají možnost připojení vlastních počítačů a notebooků.

Ústřední knihovna UTB registruje okolo 5000 čtenářů. Ročně vyřídí asi 30 000 výpůjček, z toho více než 600 meziknihovních. Knihovna disponuje více než 33 000 knihami a periodiky a téměř 4 000 ostatními dokumenty. Roční přírůstek knihovních jednotek činí asi 3 000. Knihovna ročně zorganizuje na 30 vzdělávacích a výchovných akcí. V knihovně, ve studovnách a čítárně je

110 míst k sezení. Součástí celoškolské knihovny je i 12 menších, ústavních knihoven s odbornou literaturou. Studentům slouží také 2 areálové studovny v budovách na Mostní ulici a na Jižních Svazích.

Kongresové centrum s názvem Academia Centrum UTB organizuje na 200 akcí ročně pro externí objednavatele (přednášky, semináře, školení, prezentace firem, kongresy, konference apod.) Academia Centrum dále zajišťuje asi 100 neziskových akcí podle požadavků jednotlivých fakult a pracovišť univerzity. Významným přínosem pro studenty UTB všech forem studia je ediční činnost Academia Centra: vydávání skript (na 50 ročně),

Projekt vzniku budoucí univerzity byl připravován v 90. letech. Tehdy byla založena Fakulta managementu a ekonomiky a Institut reklamní tvorby a marketingových komunikací (budoucí FMK). K úspěšnému završení snah o zřízení univerzity došlo dne 14. listopadu 2000, kdy tehdejší prezident Václav Havel podepsal zákon o zřízení Univerzity Tomáše Bati ve Zlíně (UTB) ke dni 1. 1. 2001. Zlín se tak stal univerzitním městem, poskytujícím zájemcům vysokoškolské vzdělání v širokém spektru oborů. Slavnostní inaugurace Univerzity Tomáše Bati a jejího prvního rektora, prof. Petra Sáhy, proběhla 16. května 2001 a zúčastnilo se jí přes 400 hostů z řad reprezentantů všech českých univerzit a vysokých škol, členů parlamentu, zástupců místní samosprávy, podnikatelů, ale i mnoho hostů ze zahraničí.

V lednu roku 2002 se univerzita rozrostla o třetí fakultu, Fakultu multimediálních komunikací, jež je unikátním pracovištěm svého druhu v ČR nabízejícím studijní programy Mediální a komunikační studia a Výtvarná umění. Od 1. 1. 2006 je v provozu čtvrtá fakulta s názvem Fakulta aplikované informatiky a od 1. 1. 2007 pátá fakulta s názvem Fakulty humanitních studií. Na UTB nyní studuje více než 10 000 studentů.

2. O Fakultě managementu a ekonomiky

Fakulta managementu a ekonomiky (FaME) byla založena v roce 1995. Fakulta je složena z osmi ústavů: Ústav managementu, Ústav ekonomie, Ústav podnikové ekonomiky, Ústav managementu výroby – průmyslového inženýrství, Ústav financí a účetnictví, Ústav veřejné správy a regionálního rozvoje, Ústav informatiky a statistiky, Ústav tělesné výchovy.

V akademickém roce 2005/6 na FaME ve Zlíně studuje:

2056	studentů v bakalářských studijních programech
953	studentů v navazujících magisterských studijních programech
824	v bakalářských studijních programech realizovaných na VOŠE Zlín
97	studentů v doktorských studijních programech

Na FaME působí celkem 88 akademických pracovníků. Personální strukturu tvoří 12 profesorů, 20 docentů, 36 odborných asistentů, 17 asistentů, 3 lektori. Neakademických pracovníků (ostatní zaměstnanci) má FaME 23. Celkový počet zaměstnanců je 111.

Fakulta managementu a ekonomiky prošla od svého založení v roce 1995 dynamickým vývojem z pohledu počtu studentů, počtu pracovníků i počtu zabezpečovaných studijních programů.

V současné době nabízí ve vzdělávací činnosti v prezenční a kombinované formě studia dva bakalářské, dva navazující magisterské a jeden doktorský studijní program. Rozvíjí formy celoživotního vzdělávání identické bakalářským a navazujícím magisterským studijním programům. Garantuje v rámci Univerzity Tomáše Bati ve Zlíně dva bakalářské studijní programy realizované na Vyšší odborné škole ekonomické ve Zlíně v prezenční i kombinované formě. Zejména při výuce v kombinované formě studia a celoživotním vzdělávání jsou postupně aplikovány formy distančního e-learningového studia přes portál <http://education.utb.cz/>. Fakulta má od roku 2003 oprávnění uskutečňovat habilitační řízení v oboru management a ekonomika podniku.

Vědecko-výzkumná činnost fakulty se postupně rozvíjela v zaměření na transformační marketing a inovační management řešením dvou projektů agentur GAČR v období 1995-1998 a projektů Fondu rozvoje VUT v Brně. Současné dlouhodobé cíle výzkumu vycházejí z výsledků řešení výzkumného záměru MSM 265300021 Výzkum konkurenční schopnosti českých průmyslových výrobců a jsou orientovány na formulaci teoretických zásad a východisek pro koncipování průmyslové politiky euroregionu ve vztahu ke globalizaci ekonomiky a na vypracování metodických materiálů pro české firmy k zavedení systémů modernizace podnikových procesů, směřujících ke zvýšení jejich konkurenceschopnosti. Pozornost je věnována problémům teorie konkurenceschopnosti, specifikům konkurenceschopnosti malých a středních firem, úloze řízení lidských zdrojů, marketingu a marketingového řízení, ekonomických (hodnotových) procesů, logistické a informační podpory konkurenceschopných procesů u českých průmyslových výrobců. Vytvoření samostatné Fakulty managementu a ekonomiky VUT BRNO se sídlem ve Zlíně vyrostlo z naléhavé potřeby rozvoje moravských regionů Zlín, Uherské Hradiště, Kroměříž, Vsetín a Hodonín, zvláště pak z nutnosti dynamického rozvoje průmyslových, obchodních a veřejně správních institucí ve východomoravských regionech.

3. Ústav informatiky a statistiky

ÚIS zajišťuje výuku v oblastech informatiky, matematiky, statistiky, kvantitativních metod a ekonometrie na všech oborech studia FaME (PS, KS, CŽV, e-learning (EDEN, RIUS), Sokrates-Erasmus, PGS). Na ÚIS působí celkem 14 zaměstnanců (2 profesori, 2 docenti, zbytek OA + 3 doktorandi). Ředitelem je doc. Ing. Rudolf Pomazal, CSc.

Ústav informatiky a statistiky byl založen v roce 1998 s cílem rozvinout a výrazně zkvalitnit výuku informatiky a statistiky na Fakultě managementu a ekonomiky na Univerzitě Tomáše Bati ve Zlíně.

Ústav byl od začátku ústavem mladých dynamicky se rozvíjejících pracovníků, zejména doktorandů, kteří postupně přecházeli na pozice asistentů. Do garance ústavu byly zařazeny nejen dosavadní předměty z informatiky, statistiky a operačního výzkumu vyučované na FaME, ale zejména byly postupně budovány předměty nové, které výrazně modernizovaly dosavadní výuku v oblasti informatiky a to zejména směrem k využívání databázových a internetových technologií, podnikových informačních systémů a jejich ochrany a bezpečnosti. Učitelé ústavu také garantují výuku na studiu doktorandském, a to předměty Informační systémy pro vědu a výzkum, Metodika vědecké práce a Matematické metody v ekonomickém výzkumu.

Pracovníci ústavu se věnují výzkumu v oblasti efektivnosti informačních systémů, metodám dolování znalostí z dat a ochraně podnikových znalostí a dat a zejména zavádění metod e-learningu. Právě v této poslední oblasti zaznamenali pracovníci ústavu velký úspěch řešením grantů Fondu rozvoje vysokých škol a grantu LEONARDO Batcos.

Pracovníci ústavu pravidelně garantují a organizují konference s názvem Internet a konkurenceschopnost podniku, které se již od roku 1999 konají ve Zlíně v rámci celostátní akce Březen měsíc Internetu.

Další rozvoj ústavu by se měl ubírat cestou vyšší vnitřní odborné integrace a synergie informačních technologií a metod statistického a operačního výzkumu. Postupně by také výuka všech ústavem zabezpečovaných předmětů měla být převedena do virtuálního vzdělávacího prostředí v Learning Management Systemu EDEN.

Ústav dále spolupracuje s řadou firem a s Českým statistickým úřadem, krajskou reprezentací Zlín.

4. Výuka statistiky na ÚIS

Výuka statistiky je zajišťována 4 pedagogy. Ve studijních programech nalezneme dva hlavní předměty:

- **Metody statistické analýzy** (Statistical Analysis Methods, garant doc. Pomazal, rozsah 2-2-0, z, zk);
- **Aplikovaná statistika** (Applied Statistics, garant dr. Klímek, rozsah 2-2-0, z, zk).

Jsou vyučovány v prezenční a kombinované formě BS (MSA pro oba obory, APS pro obor ekonomika a management) a také v CŽV.

A. Metody statistické analýzy (Statistical Analysis Methods)

Jedná se o první statistiku. Je povinná pro oba obory: tzn. pro cca 300 posluchačů PS a 150 KS.

Pro tento kurz byly vydány studijní opory (skripta):

1. STŘÍŽ P., RYTÍŘ, V., KLÍMEK, P., KASAL, R.: *Přednášky z Metod statistické analýzy*. 2. upravené vydání. Zlín: FaME, 2006.
2. STŘÍŽ P. a kolektiv: *Cvičebnice do Metod statistické analýzy*. Zlín: FaME, 2007.

Další opory jsou připraveny na portálu ÚIS <http://uis.fame.utb.cz/>.

Program MSA obsahuje tyto hlavní oblasti:

1. Jednoduché prostředky popisu statistických souborů.
2. Úvod do teorie pravděpodobnosti.
3. Náhodné veličiny.
4. Podstata a základní úlohy matematické statistiky (odhady a testování hypotéz).

Cvičení pro MSA je rozděleno do 8 skupin (2 hodiny týdně). Vede je ing. Beranová spolu s ing. Kasalem. Výuka je založena na ručních výpočtech a vlastních seminárních úkolech – probíhají prezentace. Dále je základní statistika procvičována v Excelu (funkce, analýza dat) a v demoverzích statistických programů.

B. Aplikovaná statistika (Applied Statistics)

Aplikovaná statistika je povinná pro jeden obor: 200 PS, 100 KS+10 CŽV

Pro tento kurz byly vydány studijní opory (skripta):

1. KLÍMEK, P. *Aplikovaná statistika pro ekonomy*. Skripta pro 2. ročník denního studia. Zlín: UTB, FaME, 2003.

2. KLÍMEK, P. *Aplikovaná statistika – cvičení*. 2. upravené vydání. Zlín: FaME, 2004.
3. KLÍMEK, P. *Aplikovaná statistika – studijní pomůcka pro distanční studium*. Zlín: FaME, 2005.

Další studijní opory lze nalézt na následujících adresách:
na portálu ÚIS (PS + KS) <http://uis.fame.utb.cz/>,
na portálu e-learningu (KS) <http://education.utb.cz/> a
na ftp serveru (PS + KS) <ftp://study.uis.fame.utb.cz/>.

Program Aplikované statistiky obsahuje tyto části:

1. Analýza závislostí.
 - (a) χ^2 test v kombinační tabulce.
 - (b) ANOVA (1 faktor, 2 faktory).
 - (c) Regresní a korelační analýza.
2. Neparametrické testy.
3. Časové řady.
4. Indexy a difference.

Cvičení pro APS je rozděleno do 7 skupin (2 hodiny týdně). Vede je dr. Klímek a ing. Kasal. Ve výuce je probírána sbírka příkladů, dále si studenti připravují vlastní seminární práce (presentace + diskuse). Během cvičení je používán MS Excel (funkce, analýza dat – česká terminologie!), dále nadstavba XLStats (pracuje pod Excelem, dobré zkušenosti), z ryze statistických programů Statistica 7Cz, a na doplnění demo Statgraphics Centurion XV (StatReporter).

5. Výuka statistice příbuzných disciplín

Kromě výše uvedených kurzů ÚIS zajišťuje výuku i následujících příbuzných disciplín:

- Počítačové zpracování dat (PZD – 2. ročník, LS, paralelně s APS, prof. Molnár + dr. Stríž, rozsah 2-1-0, klz).
- Kvantitativní metody rozhodování (KMR – 4. ročník, ZS, všechny obory, garant dr. Zimola + dr. Kolčavová, rozsah 2-2-0, z, zk).
- Manažerské rozhodování při riziku a nejistotě (MRRN – 4. ročník, ZS, vybrané obory doc. Pomazal, dr. Brychta + ing. Beranová, rozsah 2-1-0, klz).

- Ekonometrie (EKN – 5. ročník, ZS, vybrané obory, dr. Klímek, rozsah 2-1-0, klz).

6. Problémy při výuce statistiky

Při výuce statistiky se můžeme setkat na FaME s těmito problémy:

- vysoká fluktuace garantů MSA,
- doktorandi – časté odchody do komerční oblasti,
- studenti – navyšování počtů, pokles kvality, malý podíl gymnázií (nedostatek v základní matematice a teorii pravděpodobnosti),
- česká terminologie v Excelu.

Abychom nejmenovali pouze problémy, můžeme hledět s nadějí do budoucnosti. Sledujeme zvyšující se zájem o statistiku u výborných studentů. Máme řadu SVČ, diplomových prací (BDP, MDP), nové studenty na PGS. Také probíhá neustálé rozšiřování výuky formou e-learningu: EDEN (UTB) a RIUS (Rozběh interuniverzitního studia v síti vybraných univerzit ČR: ZU v Plzni, UHK a UTB). Tento zájem zaznamenáváme nejen u českých, ale i zahraničních studentů – souběžně totiž probíhá výuka obou statistik (Statistical Analysis Methods, Applied Statistics) v rámci projektů Sokrates-Erasmus (tvorba nových materiálů v angličtině).

7. Literatura

Studijní opory pro výuku statistiky na FaME jsou uvedeny přímo v textu, rovněž i relevantní URL adresy jsou psány kurzívou tamtéž.

Adresa: Ing. Petr Klímek, Ph.D.

Ústav informatiky a statistiky, Fakulta managementu a ekonomiky, Univerzita Tomáše Bati ve Zlíně, Mostní 5139, 760 01 Zlín

E-mail: klimek@fame.utb.cz

XLSTATISTICS FOR TEACHING STATISTICS IN FAME, TBU IN ZLÍN

Petr Klímek

Address: Ing. Petr Klímek, Ph.D., Department of Informatics and Statistics, Faculty of Management and Economics, Tomas Bata University in Zlín, Mostní 5139, Zlín, 76001, The Czech Republic

E-mail: klimek@fame.utb.cz, pklimek@seznam.cz

Foreword:

Paper [3] was discussed at the conference, STAKAN 2007 in Rusava. For the conference the following special paper was prepared in English for the News Bulletin. XLStatistics is used for education of statistical courses in FaME, TBU in Zlín. It is a very simple programme which is for education and evaluation purposes free.

1. Introduction

XLStatistics is a set of workbooks for Microsoft Excel (Versions 5 and above) that can be used to carry out most of the commonly encountered types of statistical analyses of data.

Its author is Dr. Rodney Carr from Deakin University Warrnambool (Australia) (e-mail: rodneyc@deakin.edu.au). XLStatistics is for education and evaluation purposes free, the licence cost is 30 AUD or 20 USD. The more licences bought the lower the payment. Registered users who have purchased a copy of XLStatistics are entitled to free upgrades to successive versions of the XLStatistics workbooks as they are produced. This price is very low compared to commercial statistics software products.

The set comprises of 11 basic analysis workbooks (Data Analysis Workbooks) all designed to handle data with a particular combination. The two basic types of variables are Numerical (Num) and Categorical (Cat). Other workbooks are linked into the basic ones. There are five other workbooks that perform special tasks (e.g. probability functions, sample selection, etc.). You can find a full list of them in Figure 1.

2. Hardware Requirements and Installation

XLStatistics is compatible with MS Excel and MS Word 97 and above. Hardware requirements are therefore similar to these products. After downloading

Data Analysis Workbooks				
Number of Variables	1	1 Numerical 1Num	1 Categorical 1Cat	
	2	1 Numerical 1Num1Cat 1 Categorical	2 Categorical 2Cat	2 Numerical 2Num
	3	1 Numerical 1Num2Cat 2 Categorical	2 Numerical 2Num1Cat 1 Categorical	
	n	1 Numerical 1NumnCat n Categorical	n Categorical nCat	n Numerical nNum
		n Numerical nNum1Cat 1 Categorical		
Other Workbooks				
Probability Functions		PDF	Transform	Options Help
Sample Selection		SampSel	Populate	☐ Hide Launchpad
Quality Control		Control		☐ Zoom sheets to fit window

Figure 1: XLStatistics main menu (workbook)

from the web address <http://www.deakin.edu.au/~rodneyc/XLStats.htm> we get archive XLS5.zip (cca 6 MB). We extract this archive to our created folder. There about 70 files in this folder. Now, we can start with XLStatistics in the following steps:

1. Open file XLStats.xls and create an icon on the computer desktop for easier opening.
2. Hide Launch-pad (optional). The Launch-pad is a good way of seeing the overall structure of XLStatistics. It could be more comfortable for some users (see Figure 2).
3. Organise your data in a separate workbook.
4. Highlight the data on your data worksheet in Step 3. Highlight one block of data only making sure to cover the variables you wish to analyse.
5. Select the appropriate analysis workbook from the XLStatistics menu (see Figure 1 or 2).

Note:

Users can alter anything in cells with blue text in the workbooks, as needed, to modify the analyses, and they can apply their own formatting to all cells and charts. In the workbooks there is also on-line help available by double-clicking on any cell with a red label.

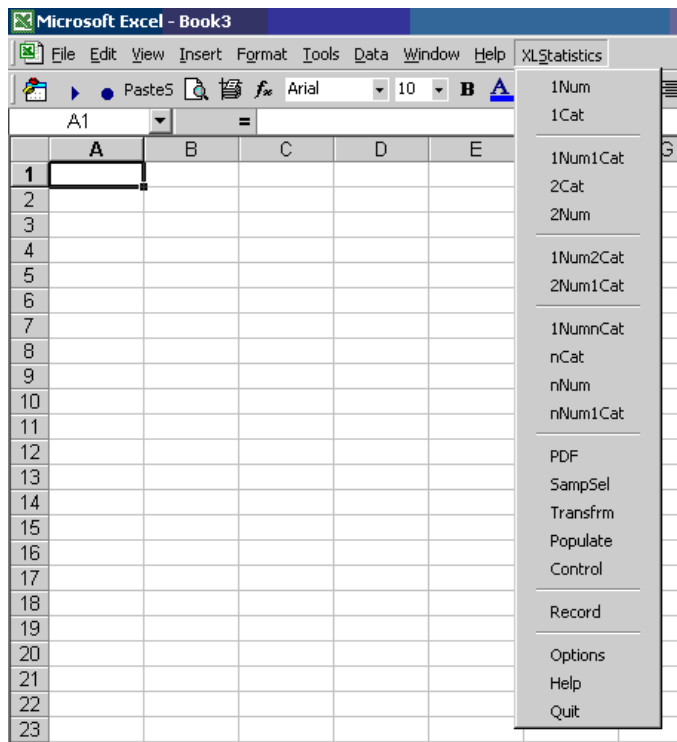


Figure 2: XLStatistics menu on Excel's main menu bar

6. View results and modify as needed.
7. Let's review the results and make any necessary modifications.
8. Save your results. Do not save the XLStatistics workbook (because they are linked together) to save your analyses – instead copy the required results out and save them individually. You can use regular copy-and-paste operations to do this (though results should be pasted as pictures, if possible).

If we want to uninstall this product, we simply delete all the files in the folder where XLStatistics was installed.

We can see a complete list of workbooks and statistical procedures which XLStatistics offers in Table 1.

Table 1: Complete list of sheets and procedures of XLStatistics

Sheets	Procedures
One numerical variable (Single-sample data) 1Num	Numerical and graphical summaries – Sample means, standard deviations, etc – Frequency charts (many variations) – Box-and-whisker plots Tests, etc – t-test for μ – Confidence, Tolerance and Prediction intervals – Sign test and confidence interval for the media – Chi-square test for variance – Residuals plots – Power analysis/sample size selection – Test for normality Other graphs – Bar chart with error bars – Cumulative frequency plots of the data or log of the data on a probability scale.
One categorical variable 1Cat	Numerical and graphical summaries – Sample proportions tables, bar charts, etc Tests – Large- and small- sample tests for proportions – Confidence intervals for proportions – Power analysis/sample size selection – Goodness-of-fit test – Runs test Other graphs – Pie chart – Frequency chart with error bars
One numerical variable, one	Numerical and graphical

Table 1: Tabular continuing

Sheets	Procedures
categorical variable 1Num1Cat	summaries – Sample means, etc – Frequency charts – Box-and-whisker plots – Plots of means with error bars Tests – Oneway Analysis of Variance and Fixed)(Random effects – Test for intercept term – Kruskal-Wallis test – Residuals plots – Hartley's test – Two-sample t-tests and confidence intervals for the difference between two means – Mann-Whitney test – F-test – Power analysis/sample size selection (2 level only) – Randomised 2-group test Other graphs – Means plots with error bars or bands of various types – Frequency charts
Two categorical variables 2Cat	Summary tables – Counts, proportions, Row- or Column % tables, etc Tests – Chisquare test – Two-sample t-tests and confidence intervals for the difference between two proportions – Small-sample tests for the difference between two proportions – Analysis of 2x2 tables (Odds ratio, Fisher's exact test)

Table 1: Tabular continuing

Sheets	Procedures
	Other graphs – Plots of proportions with error bars of various types
Two numerical variables 2Num	One-and two-variable numerical summaries Scatterplots with axes and/or their direction swapped Linear regression, with or without constant term – Estimates, confidence intervals, tests for the parameters – Scatterplot with regression line, etc – Residuals plots – Prediction and inverse prediction. Prediction bands Fitting a user-defined function to data using various methods: – Linearizable functions with linear least squares regression – Polynomials – Non-linear least squares regression – Least squares regression on straight lines allowing for a possible break-point Fitting a smooth curve to data using various methods: – Moving Average (mean or median) – Means of Grouped data with error bars/bands – Locally weighted regression - LOWESS Analysis for ordinal variables: – Spearman's rho – Kendall's tau-b Adding labels to points in a scatterplot
Two numerical variables,	(Multiseries) scatterplots with

Table 1: Tabular continuing

Sheets	Procedures
one categorical variable 2Num1Cat	axes and/or their direction swapped Multiline/multiaxes plots Linear regression (Analysis of Covariance). – Various choices (with or without constant term, etc) for the analysis – Scatterplot with regression lines, etc – Plots of the slopes and intercepts with error bars – Hypothesis tests Fitting a user-defined function to data (group-by-group) using various methods: – Linearizable functions with linear least squares regression – Polynomials – Non-linear least squares regression – Least squares regression on straight lines allowing for a possible break-point Fitting a smooth curve to data (group-by-group) using various methods: – Moving Average (mean or median) – Means of Grouped data with error bars/bands – Locally weighted regression - LOWESS
One numerical variable, two categorical variables 1Num2Cat	Summary tables – Means, standard deviations, etc. Basic tests – Two-way Analysis of Variance (Balanced data only). Choices of Fixed and/or Random effects, Interaction

Table 1: Tabular continuing

Sheets	Procedures
	– Residuals plots Graphs – Means plots with error bars or bands, etc.
n Numerical variables (Multiple regression package) nNum	Single-variable summaries – Frequency charts and combined frequency – Side-by-side boxplots – Means plots – Numerical summaries Two-variable summaries – Scatterplots, scatterplot matrix – Two-variable regression – Correlation, correlation matrix – Multiline/axis plots Multiple regression – Regression equation, usual summary statistics and tests – Prediction – Residuals analyses – Plots of residuals – Jarque-Bera test for normality – Ramsey test for model specification error – Durbin-Watson test for serial correlation. Corrections for serial correlation (Cochran-Orcutt) – Glejser test for heteroscedasticity – Analysis of variance and Model comparison Analysis of repeated measures data – Case-by-case plots – Plots of means – Side-by-side boxplots

Table 1: Tabular continuing

Sheets	Procedures
	– Analysis of variance
n Categorical variables nCat	Pivot Tables showing Counts or Percentages (uses Excel's built-in Pivot Wizard)
1 Numerical and n Categorical variables 1NumnCat	Pivot Tables showing appropriate numerical variable (summaries of the numerical uses Excel's built-in Pivot Wizard)
n Numerical variables and 1 Categorical variable 1NumnCat	– Numerical summaries of all the numerical variables category-by-category – Multiple plots of means
Propability PDF	Probabilities and inverse probabilities, graphs, etc, for the most common distributions: – Beta – Binomial – Chisquare – Exponential – F-distribution – Gamma – Hypergeometric – Lognormal – Normal – Poisson – t-distribution – Weibull – User defined
Sample selection SampSel	Selecting random samples from specified distributions Selecting random samples from a population (with or without replacement) Free-hand forming or altering of data

Table 1: Tabular continuing

Sheets	Procedures
Transforming data Transform	Various transformations of data, with associated summary information
Population (census data) Populate	Summarize data for a whole population (1 numerical variable or 1 categorical variable)
Quality Control Charts Control	For drawing the common control charts: $\bar{x}$, S, R, I, %MR, p&np, c and u.

[following [2] software documentation]

3. A Statistical Education Use of XLStatistics in FaME, TBU in Zlín

You can find in the bachelor study programmes at FaME, TBU in Zlín two main courses from statistics:

- **Statistical Analysis Methods**, lecturer As. prof. Rudolf Pomazal, CSc., hours per week: 2-2-0, exam (written, oral) – 1st grade);
- **Applied Statistics**, lecturer Ing. Petr Klímek, Ph.D., hours per week: 2-2-0, exam (written) – 2nd grade).

They are obligatory in both the present and combined form of bachelor programme studies.

Several years ago we faced the question of which software product to use for the statistical courses in education. The supply of these products was and still is very large. We tried to use demo and trial versions of commercial products such as Statistica, Stagraphics, SPSS, etc. In the end we choose XLStats because it is free, user friendly, students found it easy to use. Students are also used to working with MS Excel from other courses of informatics. The range of statistical procedures of XLStatistics covers all courses of statistics at FaME, TBU in Zlín.

4. Other products – XLMath

From Mr. Carr we can also use XLMaths in courses of mathematics at universities. It is a very useful and user friendly tool for mathematical calculations especially in economics faculties. It has the same advantages as previous XLStatistics. It is free for education and evaluation purposes. The license cost is cheap – just 7 USD or 10 AUD. It works as a set of linked workbooks in MS Excel. Installation is similar to XLStatistics. Users can alter anything in cells with **blue** text in the workbooks as needed to modify the analyses and they can apply their own formatting to all cells and charts. In the workbooks there is on-line help available by double-clicking on any cell with a **red** label. Figure 3 displays the main menu workbook of this programme. All features of mathematics courses can be found in economics faculties in the Czech Republic.

XLMathematics Ver 2 - Excel workbooks for mathematical computations

Express	Expression evaluation	Limit	Limits of functions
TabFun	Tabulating functions	Seq&Ser	Limits of sequences and series
SysLin	Solving systems of linear equations	Deriv	Differential calculus
Graph	Graphing	Newton	Solving equations using Newton's Method
FunOps	Operations on Functions	Integ	Integral calculus
GrafPapr	Producing graph paper		

Documentation
XLMath.doc

Options
Help

© Rodney Carr 1996-99

Figure 3: Main workbook of XLMaths

There is a list of XLMaths sheets and procedures in Table 2.

Table 2: Complete list of sheets and procedures of XLMath

Sheets	Procedures
Graphing Graph.xls	Many different varieties of graphs: – Single function

Table 2: Table continuing

Sheets	Procedures
	– Functions with data – Many functions on the same axes – Functions defined only at a discrete set of points – Graphs in polar coordinates – Conic sections – Parametric curves – Graphs of functions of two variables
Expression evaluation Express.xls	For using Excel as a simple calculator
Tabulation of functions TabFun.xls	Table of functions using Excel
Graph Paper GrafPapr.xls	Use for producing and printing graph paper of various types (linear/log)
Operations on functions FunOps.xls	Many different varieties of graphs: – Plots of functions with their inverses – Transforms of functions – Composites of functions
Solving systems of linear equations SysLin.xls	Solve systems of linear equations using Gauss-Jordan elimination.
Limits of functions Limit.xls	Determine limits of functions numerically. Handles limits at infinity, etc. Includes graphs with asymptotes, etc.
Differential calculus Deriv.xls	Functions and their derivatives with – Tangent lines and secant line approximations – Numerical calculation of derivatives – Graphs of the derivative
Integral calculus Integ.xls	Numerical integration – Riemann sums – Trapezoidal rule – Simpson's rule – Integration of functions defined only at discrete points

Table 2: Table continuing

Sheets	Procedures
Newton's Method Newton.xls	Numerical solution of equations using Newton's method (with graphs, etc.)
Sequences and Series SeqSer.xls	Determine limits of sequences and series numerically

[following [1] software documentation]

References

- [1] CARR, R. *XLMathematics*. Excel workbooks for Mathematical Analysis. Version 2. Official documentation.
- [2] CARR, R. *XLStatistics*. Excel workbooks for Data Analysis. Version 5. Official documentation.
- [3] KLÍMEK, P. *History and Present Teaching of Statistics at FaME, TBU in Zlín*. Lecture and article from the STAKAN 2007 conference held in Rusava, The Czech Republic. (In Czech) Historie a současnost výuky statistiky na FaME, UTB ve Zlíně. Příspěvek a prezentace na konferenci STAKAN 2007 na Rusavě.

O VYUČOVANÍ VIACROZMERNÝCH ŠTATISTICKÝCH METÓD NA ŠKOLÁCH EKONOMICKÉHO ZAMERANIA

Pavol Kráľ, Gabriela Nedelová

Abstrakt: V našom príspevku sa pokúsime poukázať na možnosti výučby viacrozmerných štatistických metód v podmienkach ekonomických fakúlt, ktoré nemajú akreditované kvantitatívne zamerané odbory. Budeme diskutovať o spôsoboch, ktoré nám môžu pomôcť eliminovať problém s často nedostatočnými matematickými i štatistickými základmi študentov ekonómie, najmä o použití systému Moodle v kombinácii so štatistickým softvérom (R, SPSS, ...) pri vyučovaní.

Kľúčové slová: viacrozmerné štatistické metódy, R, SPSS, Moodle.

1. Úvod

Ako napovedá názov príspevku, budeme sa v našom článku venovať spôsobu a možnostiam vyučovania viacrozmerných štatistických metód na vysokých školách ekonomického zamerania. Hoci by to mohol názov článku naznačovať, nejde o žiadnu vyčerpávajúcu komparatívnu štúdiu venujúcu sa vyučovaniu viacrozmerných štatistických metód na fakultách s ekonomickým zameraním rôznych vysokých škôl a univerzít, ale o niekoľko úvah učiteľov štatistiky na tému viacrozmerné štatistické metódy v príprave budúcich ekonómov.

Naše myšlienky a názory budeme ilustrovať na príklade materskej fakulty autorov, Ekonomickej fakulty Univerzity Mateja Bela v Banskej Bystrici (v ďalšom texte EF UMB), ktorú si dovoľíme považovať za pomerne dobrú reprezentáciu súčasného stavu problematiky nášho článku. Aby sme sa vyhli prípadnému nedorozumeniu, pokúsime sa charakterizovať typ fakúlt, ktorým sa budeme v našom príspevku venovať.

Inými slovami, pokúsime sa vysvetliť, čo presne myslíme pod školou ekonomického zamerania, ktorá nemá akreditované kvantitatívne orientované odbory. Pre jednoduchosť budeme za kvantitatívne orientovaný odbor považovať odbor, ktorého študenti absolvujú základný kurz matematiky a štatistiky v úvodných dvoch ročníkoch svojho štúdia a počas svojho ďalšieho štúdia sa s ďalšími predmetmi obsahujúcimi dopĺňujúce poznatky z matematiky a štatistiky stretávajú najmä v podobe povinne voliteľných a výberových predmetov. Za postačujúce identifikačné znaky považujeme to, že odbor nemá

v názve kvantitatívne metódy a pracovisko (katedra) zabezpečujúca vyučovanie kvantitatívnych metód nie je odborovou katedrou.

2. Kvantitatívne metódy na EF UMB

Domnievame sa, že takúto definíciu spĺňajú všetky odbory, ktoré je možné v súčasnosti študovať na EF UMB. Existujú tam nasledujúce odbory:

- Verejná ekonomika a služby (VES),
- Verejná správa a regionálny rozvoj (RRVS),
- Cestovný ruch (CR),
- Ekonomika a manažment podniku (EMP),
- Financie, bankovníctvo a investovanie (FBI).

Pre všetky tieto odbory zabezpečuje vyučovanie kvantitatívnych metód Katedra kvantitatívnych metód a informatiky. S kvantitatívnymi metódami sa študenti jednotlivých odborov stretnú prvýkrát, keď absolvujú základný kurz matematiky a štatistiky v nasledujúcich predmetoch:

- VES (Matematika 1, Matematika 2, Štatistika 1),
- RRVS (Matematika 1, Matematika 2, Štatistika 1),
- CR (Matematika 1, Matematika 2, Štatistika 1, Štatistika 2),
- EMP (Matematika 1, Matematika 2, Štatistika 1, Štatistika 2),
- BEM (anglická mutácia programu EMP, Mathematics 1, Mathematics 2, Statistics 1, Statistics 2),
- FBI (Matematika 1, Matematika 2, Štatistika 1, Štatistika 2).

Obsah základného kurzu matematiky a štatistiky je typický pre školy tohto zamerania. V matematike sú to základy limitného počtu, diferenciálneho a integrálneho počtu jednej premennej a viacerých premenných, základné informácie o postupnostiach a nekonečných radoch, základy lineárnej algebry. Všetky preberané témy obsahujú časti ilustrujúce možné aplikácie preberaných častí matematiky v ekonómii.

Obsahom základného kurzu štatistiky je popisná štatistika, základy pravdepodobnosti, indexy, časové rady, základy indukčnej štatistiky (bodové a intervalové odhady, testovanie hypotéz), regresná a korelačná analýza, analýza kategoriálnych znakov. Drtivú väčšinu príkladov, ktoré sú používané pri výučbe tohto kurzu, tvoria príklady s ekonomickou problematikou. Môžeme povedať, že sa v podstate jedná o kurz jednorozmernej štatistiky (okrem častí viacrozmerná lineárna regresia, korelačná analýza, jednofaktorová ANOVA,

analýza kategoriálnych znakov), ktorý je porovnateľný s obdobnými úvodnými kurzami štatistiky na fakultách rovnakého typu.

Na rozdiel od matematiky ale nie je základná štatistická príprava všetkých odborov rovnaká, pretože odbory VES a RRVŠ majú štatistiku len v jednom semestri. Z toho vyplýva, že tieto odbory majú niektoré časti štatistiky z časových dôvodov „oklieštené“ alebo dokonca úplne vynechané.

Na druhej strane prekvapujúco rozsah a obsah výučby štatistiky odboru CR je úplne rovnaký ako pri odboroch EMP a FBI. Vďaka tomu študenti rôznych odborov vstupujú do ďalšieho štúdia s rôznym rozsahom predpokladaných štatistických znalostí. Bez ohľadu na rozsah výučby majú ale študenti všetkých odborov v priemere pomerne rezervovaný vzťah k matematicko-štatistickým metódam. To je z veľkej časti spôsobené tým, že absolvovať úspešne vyššie uvedené predmety predstavuje pre časť študentov pomerne veľký problém.

To má za následok, že štatistika je často považovaná za nepríjemnosť pri ceste za titulom, pričom sa mnohí študenti domnievajú, že pri svojom budúcom uplatnení ju nebudú potrebovať. Tento názor je žiaľ často podporovaný aj niektorými kolegami z ďalších katedier, ktorí majú voči kvantitatívnym metódam a obzvlášť štatistike pomerne chladný postoj.

V dôsledku toho sa aplikácie štatistických metód často neobjavujú ani v predmetoch, kde by to bolo možné, prirodzené, ba dokonca žiaduce. Predpokladáme, že uvedené skutočnosti nie sú špecifické len pre EF UMB, ale že podobný problém je typický a rôznym spôsobom sa ho pokúšajú riešiť aj ďalšie fakulty s ekonomickým zameraním. Predpokladáme, že práve v dôsledku uvedených skutočností sa štatistické znalosti študentov prudko v priebehu štúdia redukujú až na tak nízku úroveň, že v podstate neumožňuje bezproblémovú štatistickú konzultáciu diplomovej práce, pretože študent má často už vo štvrtom a piatom ročníku problém so základnými štatistickými pojmami.

Okrem základného kurzu sa matematicko-štatistické metódy vyskytujú v množstve povinných, povinne voliteľných a výberových predmetov, napríklad: Ekonometria, Optimalizačné metódy v ekonómii, Analýza časových radov, Matematicko-štatistické metódy, Poistná matematika, Poistná matematika a štatistika, Finančná matematika 1, Finančná matematika 2, Kvantitatívny manažment. . .

3. Ako na viacrozmerné štatistické metódy?

Viacrozmerné štatistické metódy na EF UMB sú obsahom práve predmetu Kvantitatívny manažment, ktorý je určený ako povinne voliteľný predmet

pre štvrtý alebo piaty rok štúdia. Tento predmet bol zahrnutý do študijných plánov, pretože považujeme za dôležité, aby budúci ekonómovia získali určitý prehľad aj o viacrozmerných štatistických metódach. Najmä z hľadiska ich možného využitia ako podporného nástroja pri prijímaní rozhodnutí.

Z toho vyplývajú aj hlavné ciele tohto predmetu, ktorým je poskytnúť študentom pomerne ucelený prehľad najčastejšie používaných viacrozmerných štatistických metód a do určitej miery eliminovať ich negatívne skúsenosti so štatistikou v nižších ročníkoch. Okrem viacrozmerných štatistických metód sú obsahom tohto predmetu aj základy teórie rozhodovania. Študent by teda mal vďaka tomuto predmetu získať predstavu o spôsobe využitia matematicko-štatistických metód v prípade, že má k dispozícii rozsahom malé i primerané dátové súbory.

Pri zavedení predmetu bolo potrebné vyriešiť niekoľko kľúčových otázok. Predpokladáme, že rovnaký typ problémov je spojený so zavedením predmetu obsahujúcom viacrozmerné štatistické metódy aj na iných školách ekonomického zamerania. Základnou otázkou bolo určiť, ktoré viacrozmerné štatistické metódy vzhľadom na obmedzenú hodinovú dotáciu predmetu (80 minút týždenne) vybrať. Inšpirovali sme sa výbornou trilógiou prof. Hebáka a kol. Viacrozmerné štatistické metódy 1, 2, 3, ktorá je aj základnou študijnou literatúrou predmetu Kvantitatívny manažment. Momentálne tvoria obsah predmetu nasledujúce metódy: diskriminačná analýza, logistická regresia, metóda hlavných komponentov, faktorová analýza, korešpondenčná analýza, zhluková analýza, ANOVA.

Základnou motiváciou pre ich zaradenie do kurzu bola skutočnosť, že sa často vyskytujú problémy ekonomickej praxe, ktoré môžu byť riešené práve s použitím uvedených metód. Ďalším problémom bolo určiť spôsob, akým uvedené metódy vyučovať. Museli sme prihliadať na to, že matematický aparát našich študentov nie je dostatočný na podrobné matematické odvodenie spomínaných metód. Rozhodli sme sa preto, že pri každej metóde uvedieme len základné princípy, na ktorých je založená a podmienky, ktoré musia byť splnené, aby sme ju mohli použiť. Postup riešenia sme sa rozhodli neodvodzovať, len ukázať možnosti riešenia vybraných metód pomocou vhodného štatistického softvéru.

Zvažovali sme voľne dostupný štatistický softvér R a komerčný program SPSS 13.0, ktorý bol fakultou zakúpený z projektu ESF 11230100060. Najmä vzhľadom na kvalitu grafického rozhrania sme sa zatiaľ rozhodli pre použitie programu SPSS 13.0. Použitiu tohto softvéru sme prispôbili aj spôsob vysvetlenia jednotlivých metód, keď sme rešpektovali obmedzenia vyplývajúce z implementácie jednotlivých metód v tomto programe. Vysvetlenie každej metódy sme sa rozhodli založiť na reálnom (pokiaľ je to možné nielen na

potenciálne existujúcim, ale skutočne sa vyskytujúcim) ekonomickom probléme. Študent by tak mal získať určitý receptár na použitie viacrozmerných štatistických metód s využitím programu SPSS 13.0.

Presnejšie povedané, na aké problémy je daná metóda vhodná, ako je možné daný problém riešiť s použitím programu SPSS 13.0 a najmä ako interpretovať získané výsledky zo štatistického a ekonomického hľadiska. Keďže sa jedná o povinne voliteľný predmet, je dôležité, aby sme zabezpečili aj dostatočný počet študentov, ktorí budú chcieť tento predmet absolvovať, čím by sa umožnil jeho kontinuálny rozvoj. Stratégia ako to dosiahnuť je v podstate zrejmá. Prvým krokom je vhodná voľba názvu predmetu. Pokiaľ je to možné, v názve by sa vôbec nemali priamo objavovať štatistické metódy, ale pokiaľ je to čo i len trochu možné najmä známe ekonomické pojmy. Týmto spôsobom sa čiastočne eliminuje dopad negatívnej skúsenosti študentov so štatistikou v nižších ročníkoch.

Študenti samozrejme v prípade záujmu o daný predmet pravdepodobne podrobnejšie preskúmajú syllabus predmetu a veľmi skoro zistia, čo sa skutočne skrýva za nami zvoleným názvom. Dosiahneme tak ale aspoň to, aby o zapísaní daného predmetu uvažovali dlhší čas. Za ukážku vhodne zvoleného názvu považujeme napríklad názov predmetu Kvantitatívny manažment.

Ďalším významným determinantom záujmu o predmet je nepochybne aj osobnosť vyučujúceho. Pokiaľ je to možné, mal by to byť vyučujúci, ktorého študenti poznajú zo základného kurzu štatistiky, považujú ho za erudovaného v oblasti štatistiky a majú s ním v podstate len pozitívne osobné skúsenosti najmä s ohľadom vo vzťahu k študentom. Túto časť stratégie je vo väčšine prípadov ťažko naplniť.

V neposlednom rade a otvorene môžeme priznať, že pre študentov je jedným z najdôležitejších kritérií pri voľbe voliteľného alebo povinne voliteľného predmetu charakter jeho záverečného hodnotenia. Na základe vlastných skúseností so skúšaním v tomto predmete považujeme za optimálne riešenie seminárnu prácu so stanoveným ohraničením počtu strán, napr. na maximálne päť, pričom všetko presahujúce tento rozsah môže byť zaradené ako elektronická príloha. Praktickú tému seminárnej práce si po konzultácii s vyučujúcim volia študenti samostatne, pričom sa prihliada na kvalitu a samostatnosť riešenia.

Na zjednodušenie komunikácie medzi študentom a vyučujúcim je vhodné použiť niektorý LMS systém. Vhodnou voľbou je napríklad elearningový systém MOODLE, v ktorom bol implementovaný aj predmet Kvantitatívny manažment. Je zrejmé, že študenti obľubujú elektronickú komunikáciu, ktorá im často šetrí čas, ktorý takto môžu teoreticky venovať štúdiu. Veľkou nevýhodou ale je, že väčšina študentov ešte nemá vyvinutý zmysel pre kontrolu

svojej elektronickej komunikácie, t.j. nepoužívajú voľne dostupné nástroje na kontrolu, či nimi zadanú otázku alebo uploadovanú prácu skutočne dostal predpokladaný adresát (vyučujúci daného predmetu).

Napriek tomu Moodle vďaka svojej jednoduchosti a užívateľskej prívetivosti umožňuje pomerne ľahko získať silný podporný nástroj pri vyučovaní študentov denného štúdia a zároveň môžeme na ňom založiť aj vzdelávanie externých študentov. Môže slúžiť ako úložisko materiálov pre študentov daného predmetu, miesto ich sebahodnotenia i odborných diskusií.

Posledným problémom, ktorý je vždy potrebné riešiť je voľba vhodnej literatúry. Za najlepšiu alternatívu pre danú situáciu asi môžeme považovať už spomínanú trilógiu prof. Hebáka, ale stále narastá tlak na zabezpečenie predmetov vlastnou študijnou literatúrou. Tieto požiadavky viedli k podaniu projektu Kega 3/5214/07: Platformovo nezávislá voľne prístupná elektronická učebnica: *Viacrozmerné štatistické metódy so zameraním na riešenie projektov ekonomickej praxe*.

Výsledkom riešenia tohto projektu by mala byť voľne prístupná učebnica viacrozmerných štatistických metód vo formáte PDF, implementovaná zároveň v LMS Moodle. Vzhľadom na požadovanú platformovú nezávislosť a voľnú dostupnosť sme sa pre učebnicu rozhodli pripraviť riešenia problémov nielen v programe SPSS 13.0, ale aj vo voľne dostupnom softvéri R.

Základným predpokladom pre vznik kvalitnej učebnice je podľa nášho názoru hlavne dostatok vhodných praktických problémov. Na ich získanie je nevyhnutné rozvinúť hlbšiu spoluprácu s kolegami z ekonomických katedier.

Pomerne vhodným spôsobom sa nám javí vytváranie kombinovaných autorských kolektívov pri riešení projektových úloh, pri príprave článkov na publikovanie i vedení diplomových prác.

Domnievame sa, že napriek pozitívnym náznakom v tejto oblasti ešte stále množstvo vecí ostáva v rámci ekonomických fakúlt nedokončených.

4. Záver

V tomto článku sme sa snažili naznačiť spôsob ako implementovať vyučovanie viacrozmerných štatistických metód do odbornej prípravy budúcich ekonómov. Vhodným sa nám javí najmä zaradenie povinne voliteľného predmetu zameraného na tieto metódy v kombinácii so základmi teórie rozhodovania do štvrtého ročníka, kde finišuje špecializácia budúcich absolventov. Tento spôsob je podľa nášho názoru vhodný najmä preto, že sa viacrozmerné štatistické metódy môžu dostať do povedomia študentov ako účinná podpora rozhodovania.

Vďaka použitiu štatistického softvéru SPSS 13.0 je asi možné sa v budúcnosti priblížiť aj ideálnej situácii, keď sa absolventi v prípade možnosti už nebudú vyhýbať použitiu viacrozmerných štatistických metód na riešenie vhodných štatistických problémov, s ktorými sa vo svojej budúcej praxi stretnú.

Samozrejme je dôležité, aby zároveň správne odhadli aj limity svojich vedomostí zo štatistiky a v prípade potreby vyhľadali odbornú pomoc štatistika, napríklad na fakulte, na ktorej získali svoje vzdelanie. Z dlhodobého hľadiska by tak mohlo dôjsť k vytvoreniu hromadného aktívne fungujúceho prepojenia ekonomickej praxe reprezentovanej absolventami so štatistikou a štatistikmi pôsobiacimi na univerzitách a vysokých školách.

Tento, dúfajme, že nie utopický stav, by pravdepodobne priniesol nesporné výhody každej zo zúčastnených strán. Na jednej strane by firmy mohli ušetriť alebo získať finančné prostriedky, na druhej strane by to minimálne prospelo odbornému rastu vyučujúcich štatistiky a zlepšilo štatistickú prípravu budúcich ekonómov.

Adresa: Pavol Král a Gabriela Nedelová
Ekonomická fakulta, Tajovského 10
975 90 Banská Bystrica, Slovenská Republika

E-mail: pavol.kral@umb.sk, gabriela.nedelova@umb.sk

Podakovanie: Tento článok vznikol s podporou grantu Kega 3/5214/07.

K OTÁZKÁM VÝUKY STATISTICKÝCH KONZULTACÍ

Marek Malý

Abstrakt: Běžné univerzitní vzdělávací programy pro statistiky se zpravidla nevěnují výuce umění statistických konzultací a nepřipravují své absolventy na skutečnost, že budou muset efektivně komunikovat a spolupracovat se specialisty z jiného oboru. Argumentem bývá, že absolvent znalý teoretické statistiky se praktické záležitosti snadno doučí. Zkušenosti literární i z naší praxe ukazují, že tomu tak často není a že do určité míry lze vhodnou výukou na dráhu statistického konzultanta připravit. Příspěvek se zabývá otázkami spojenými s takovou výukou a její možnou náplní i problematikou statistických konzultací obecněji.

Úvod

Statistické konzultace provázejí vývoj statistiky a pravděpodobnosti prakticky od počátku a stály u zrodu nejednoho originálního postupu. I mnohé práce K. Pearsona a R. A. Fishera byly motivovány praktickými požadavky na analýzu dat vycházejícími z konzultací. Přitom dodnes stojí výuka statistických konzultací a její výklad ve statistické literatuře na okraji zájmu, i když jsou již k dispozici podnětné monografie [2, 5, 6] či přehledové články [20]. Často se setkáváme s tvrzením, že absolvent matematické statistiky s dobrou teoretickou výbavou se praxi statistických konzultací a aplikací rychle naučí. Ve skutečnosti jde o dlouhodobý, mnohaletý proces, který zřejmě nelze plně naučit ve škole, nicméně je dobré, když absolvent nemusí na vše přicházet intuitivně; škola jej rozhodně může alespoň částečně připravit a nasměrovat. Širokým tématem statistických konzultací se zabývá velmi bohatá škála autorů, jak shrnují např. přehledy [12, 19, 24].

V tomto textu chceme jednak shrnout základní požadavky kladené v praxi na statistického konzultanta (se zaměřením na medicínu a biologii) a jednak poukázat na některé aspekty spojené s výukou konzultací.

Statistik se musí v procesu statistické konzultace seznámit s problémem v podobě prezentované klientem. V interakci s klientem se pak pokouší odvodit reálnou podobu problému a převést ji do statistické formulace. Po vlastním matematicko-statistickém řešení následuje sepsání zprávy o provedených analýzách a jejich výsledcích a interpretace, která by měla používat pojmů srozumitelných klientovi při zachování správnosti ze statistického pohledu.

Požadavky kladené na biostatistického konzultanta

V literatuře je popsána spousta vlastností a dovedností, které by měl konzultant mít, od schopností matematických přes komunikační až po osobnostní. Základní je asi zájem a schopnost řešit reálné problémy a pomáhat klientům s jejich řešením. Podle [11, 18] lze shrnout, že dobrý statistický konzultant by měl především:

- umět pracovat s reálnými problémy, vyhledávat jejich podstatu a převádět je do statistické mluvy, tedy umět nazírat problémy jak globálně, tak v detailu [4],
- umět data vhodně zpracovávat, a to i nestandardními či originálními postupy,
- mít cit pro interpretaci a schopnost formulovat závěry v mluvené i psané podobě,
- umět komunikovat, a to i s klienty, které třeba statistika odrazuje, ale jsou nuceni se jí zabývat.

Podrobněji můžeme zmínit následující ne vždy reálně splnitelné požadavky:

- schopnost konzultovat včetně ovládnutí psychologických aspektů konzultační praxe, tj. klást správné a cílené otázky (lékaři a biologové často nejsou schopni přesně sdělit své požadavky, dokonce ani cíle výzkumu, viz též [7]), vyhledávat problémová místa, naslouchat – a to zejména nestatistikům, pozorovat, diskutovat a vést diskusi,
- schopnost z volného rozhovoru s klientem extrahovat podstatné informace a zformulovat je tak, aby to odpovídalo klientovým představám, případně s taktem usměrnit jeho nereálné požadavky,
- schopnost navrhnout optimálním způsobem metodiku realizace studie včetně otázek stanovení rozsahu výběru a způsobu sběru dat,
- schopnost nazírat na problémy jako na obecně vědeckou problematiku, ne čistě statistickou,
- schopnost technické i statistické práce s reálnými datovými soubory se všemi jejich typickými rysy a nečistotami (nečistoty a chyby v datech, chybějící pozorování) – tj. znalosti postupů čištění, kontroly a přípravy dat pro konkrétní software,
- umění vytipovat problematická místa studie i dat samotných a zjistit si podrobnosti, ochota prověřovat postupy přípravy i zpracování dat,
- dobré teoretické znalosti statistiky a široké škály metod včetně schopnosti vyhledat a správně aplikovat ne zcela běžné statistické techniky,

případně rozvíjet statistickou metodologii, to vše s ohledem na předpoklady použitých metod a důsledky jejich porušení,

- znalost základních pojmů a reálií z oboru, v němž je statistika aplikována,
- schopnost efektivně řešit problémy a pracovat v reálných podmínkách se všemi z toho vyplývajících omezeními, zejména časovými,
- schopnost samostatné práce, ochota se průběžně vzdělávat,
- schopnost aktivní tvořivé a nekonfliktní vědecké komunikace a spolupráce v interdisciplinárním týmu při potlačení primární orientace na své vlastní odborné cíle,
- organizační schopnosti potřebné při zpracování rozsáhlejších analýz i při usměrňování chodu celé studie,
- schopnost formulovat závěry a prezentovat je v mluvené, psané i grafické podobě,
- schopnost pedagogicky působit (formálně i neformálně v průběhu konzultací),
- aplikace postupů v souladu s etickými standardy.

Statistická konzultace, tak jako obecně každá konzultace, je dvousměrný proces, v jehož rámci musí jak klient tak konzultant být schopni vzájemné komunikace. Podoba konkrétní statistické konzultace se vždy odvíjí od vlastností, schopností a možností konzultanta i jeho klienta a záleží také na povaze jejich spolupráce. Podle [8, 14, 17, 22] může být biostatistik v různém postavení vůči klientům – výzkumníkům jiných oborů. Často je „pouze“ pomocníkem, statistické konzultace s klientem jsou limitovány vymezeným časem a záběrem, mnohdy jde o jednorázovou konzultaci, při níž se statistik setkává s problémem poprvé a musí ho vyřešit hned či v krátké době. Statistik se bez podrobné znalosti problematiky vyjadřuje k nějakému detailu práce (ale zákonitě nemůže prověřit všechny souvislosti). Existují i klienti, kteří statistika nutí do vedoucí role, předají mu komplikovaná data a chtějí, aby z nich vše vyčetl, přesouvají na něj zodpovědnost. Přitom prakticky vždy klient ví o datech a jejich souvislostech více než statistik.

Kimball [13] zavedl pojem chyba třetího druhu pro označení chyby vyplývající ze správné odpovědi na nesprávnou otázku. Tento problém typicky vzniká, pokud výzkumník (např. z časových důvodů) nevysvětlí statistikovi podstatu věci dostatečně podrobně a statistik se na podrobnosti nevyptá. Kimball uvádí příklad statistika, který v časové tísní poradil telefonicky klientovi, jak testovat shodu dvou korelačních koeficientů a až posléze při prezentaci zjistil, že se nejedná o nezávislé koeficienty, jak implicitně předpokládal.

Zde se jednoznačně jedná o problém v komunikaci, který právě u jednorázových konzultací může snadno nastat. Proto také jeden z našich nejzkušenějších statistických konzultantů MgMat. M. Josífků vždy zdůrazňoval, že statistik musí s klientem obsírně hovořit, aby odhalil problematická místa a zjistil co nejvíce o povaze celého problému i jednotlivých analyzovaných veličin. Tato rada dnes ještě nabývá na důležitosti, protože data se zasílají elektronicky a statistik s klientem se třeba ani nesejdou.

Kvalitativně výše nad popsanými nevyváženými vztahy stojí statistická spolupráce s klientem, k níž zpravidla dochází při závažnějších projektech a po dlouhodobější spolupráci. Statistik je do značné míry obeznámen s problematikou a často je integrálním členem vědeckého týmu, klienty je chápán jako kolega. Ne vždy lze ale tohoto ideálu dosáhnout, i jednorázové konzultace budou vždy existovat a konzultant si s nimi musí umět poradit co nejlépe.

L. Hyams [9, 10] už před více než 35 lety vymezil se svěžím humorem hlavní typy statistiků a klientů potkávajících se v průběhu statistických konzultací. Jeho klasifikace jsou trvale platné, a proto je zde ve stručné podobě přebíráme. Navzdory svému ladění přináší článek mnoho závažných otázek a jednotlivé charakteristiky, i když úsměvné, by nás měly vést k zamyšlení, k jakým situacím může v průběhu konzultací docházet, na co by měl být statistický konzultant připraven a čeho by se měl vyvarovat.

Stereotypy klientů:

1. **Pravděpodobnostník** (The Probabilist). Neví sice moc o statistice, ale i jeho tvrdý obličej se rozzáří pohledem na $p < 0.001$.
2. **Sběrač čísel** (The Numbers Collector). Přichází s plnou náručí datových formulářů, na nichž pracoval 3 roky a provedl N^N pokusů. Těžko vysvětlí, co vlastně dělal, je to moc složité. V haldě jaloviny jsou jistě ukryty drahokamy – statistiku najdi je. Nutno provést do 3 dnů, aby se nepromeškala lhůta!
3. **Občasná pijavice** (The Sporadic Leech). Nepřichází do pracovny, potká jako mimochodem statistika u oběda, probere počasí a poslední události a pak se zeptá „Co myslíte, že bych měl udělat s mými daty?“. Nedbale vyslechne radu, ale pak vše udělá sám a jen požádá o přehled literatury. Naštěstí nikdy svého poradce nezařadí mezi autory, ani mu nepoděkuje.
4. **Statistik amatér** (The Amateur Statistician). Dle něj každý může být politikem, psychologem nebo statistikem i bez odborné průpravy. Přichází proto, že nemá čas věnovat 7-8 hodin na prostudování statistické literatury. Žádá pouze o technickou pomoc s výpočty, jinak všechno ví sám. Neříkejte mu, že nemá pravdu, to by se ho velmi dotklo.

5. **Vytrvalec** (The Long Distance Runner). Přijímá statistika ve své přepychové pracovně, pohovoří s hřejivým úsměvem a potřesením ruky a pět minut vypráví o vzrušujících perspektivách své práce. O práci samé nic a rychle končí rozhovor pro neodkladný telefon. Statistika dovede k některému ze svých asistentů, aby mu řekl o špinavých detailech. Nemá pro statistika čas, ale je milý a dobře platí.

Stereotypy statistiků:

1. **Stavitel modelů** (The Model Builder). Na jakákoli data použije statistický model, kterým se momentálně zabývá nebo o kterém něco ví. Vůbec nahraje roli, zda to souvisí s otázkami, které klient klade nebo které jsou biologicky důležité. Připomíná opilce, který hledá v tmavé uličce ztracené klíče pod lucernou, protože je tam světlo.
2. **Lovec** (The Hunter). Je to statistický protipól sběrače dat. Jakákoli data podrobí rozsáhlé a vyčerpávající počítačové analýze. I když jsou data prostá, zavalí klienta půlmetrovým štosem počítačových výstupů se 17 signifikantními výsledky, které však nejsou v žádném vztahu s klientovými otázkami.
3. **Zvonař** (The Gong). Začíná každou poradou kreslením zvonovitých křivek.
4. **Tradicionalista** (The Traditionalist). Je přesvědčen, že se od časů R. A. Fishera ve statistice nic důležitého nestalo. Má proto omezený odborný slovník a počítače považuje za ďáblův vynález.
5. **Randomofil** (The Randomophilic). Pevně věří, že nezáleží na tom, co děláte, pokud jste dobře „randomizovali“. Připomíná matku, která zastihne 14-tiletou dceru v sexuálně choulostivé situaci a pochválí ji „Hlavně že nekouříš, miláčku“.
6. **Quantofrenik** (The Quantophreniac). Není podstatné, zda měříš to, co chceš, hlavně že jsou to „tvrdá“ data.
7. **Řvoun žádající více dat** (The More Data Yeller). Název říká vše.
8. **Hnidopich** (The Nit Picker). Soustřeďuje se vždy na to, jak najít v datech nedostatky. U podružných věcí dělá z komára velblouda, zatímco rozumným výsledkům nevěnuje pozornost.

Kategorizaci klientů a statistiků lze najít i v dílech dalších autorů. Van Belle [23] uvádí např. klienta, který oč méně ví i o statistice, o to více požaduje – nerozumí sice ani t-testu, ale požaduje faktorovou analýzu 50 proměnných na 20 případech (Innocent Abroad). Pro takového klienta je adekvátní statistik – kouzelník. Jiný typ zákazníka (Tinkerer) chce „jen“ nějakou drobnou úpravu

v datech a poté provést analýzu znovu. Ta ovšem zabrala 5 týdnů. Toho je schopen jen statistik – Sisýfos.

Otázky výuky statistických konzultací

K hlavním problémům absolventů statistického vzdělávání, kteří začínají s praxí statistických konzultací patří zejména to, že neumějí zacházet se skutečně reálnými daty majícími typicky mnoho nedostatků a že nejsou připraveni komunikovat s nestatistiky nematematickým jazykem o tématech jiných oborů. I když znají mnohé metody teoreticky (včetně odvození), neumějí je dobře aplikovat. Spektrum metod potřebných v praxi je navíc podstatně širší než to, s čímž se setkali při studiu. Např. statistické ukazatele používané v epidemiologii, jako relativní riziko či poměr šancí, podobně jako analýza longitudinálních a korelovaných dat nebývají v centru statistických výukových plánů. Zákazníci přicházejí mnohdy s daty, na která nelze jednoduše aplikovat některou z běžných metod či která nesplňují předpoklady běžných metod.

Jak shrnul Kulich [15], studenti statistiky by se měli naučit aplikovat všechny metody, které probírají na přednáškách. Zejména by se měli učit:

- aplikovat statistické metody na úlohy s neurčitou formulací,
- analyzovat reálná nesimulovaná data obsahující nečistoty, chyby, chybějící hodnoty, veličiny nejasného původu,
- zamýšlet se nad platností předpokladů a důsledky jejich porušení,
- porovnat alternativní přístupy řešení a umět zvolit a obhájit ten optimální.

Možné přístupy k výuce konzultací jsou popsány např. v instruktivní monografii Janice Derr [5], která je doplněna i ukázkovými videosekvencemi modelových situací. Obecně jde především o přechod od pasivní k aktivní výuce, k tomu, aby studenti byli nuceni řešit opravdu reálné problémy. Jedině tím, že se aktivně utkají se skutečně reálnými daty se všemi jejich záludnostmi se mohou dostat na kvalitativně vyšší úroveň v chápání procesu statistické analýzy dat. Studenti by se měli seznámit se zkušenostmi statistiků, kteří běžně vedou konzultace. Měli by mít možnost se reálných konzultací účastnit jako pozorovatelé. Případně se mohou přímo na takových konzultacích a jejich vyhodnocení podílet a zkusit pod dohledem učitele vyřešit reálný problém, s nímž klient přichází. Zkušenost učitele z konzultační praxe nepochybně velmi přispívá ke kvalitě jeho výuky – jednak získá cit pro potřeby lékařů (např. dovede vnímat rozdíly mezi statistickou a klinickou významností) a jednak má dobrý zdroj reálných příkladů.

Další možnost výuky skýtá uspořádání výukových konzultací, při nichž jako klient vystupuje např. pedagog jiného oboru. V jiném modelu zastává roli klienta pedagog – statistik a studenti mohou přispívat k průběhu konzultace kolektivně. Studenti mohou pod vedením pedagoga sehrát krátké scénky statistik – klient. Mohou sledovat a probírat videozáznamy reálných konzultací. Každopádně musí pochopit nezbytnost a komplikovanost vzájemné interakce mezi statistikem a biologem.

Studenti by se měli intenzivně cvičit v psaní zpráv o provedených analýzách [1]. Nemělo by ovšem jít o pouhý výčet, kde co je ve výstupu ze statistického programu, ale o skutečnou interpretaci – v reálné situaci je k úspěšné interpretaci zpravidla nutná spolupráce se zadavatelem. Studenti by poté měli své zpracování prezentovat a obhajovat v diskusi s vyučujícím a spolužáky. I výuka slovního projevu a sdělování výsledků analýz menším či větším skupinám posluchačů – nestatistiků za pomoci presentačního software je velmi důležitá a přitom opomíjená [4]. K tomu patří i umění přehledného a věcně správného grafického znázornění výsledků. Pro slovní vyjádření může studentům napomoci studium různých interpretací výstupů a rozbor publikovaných článků prezentujících výsledky statistického zpracování v časopisech daného oboru aplikace. Obecně by se studenti měli seznámit s literaturou týkající se statistické praxe a učebnice o konzultacích [21]. Někteří autoři navrhnou kurzy na téma jako *Biologie pro statistiky*, atp. [3], čímž je poukazováno na jistou asymetrii, neboť studenti biologie či medicíny zpravidla absolvují základní kurs statistiky, i když mnohdy v nedostatečné míře či v nevhodné fázi studia [16]. Ovšem i biologové a lékaři by se měli učit nejen o statistice, ale i o konzultacích. V neposlední řadě by se studenti měli zabývat etickými otázkami statistických analýz a měli by být důsledně vedeni k pečlivé, přesné a korektní práci s daty.

Požadavky zde vyřčené jdou zřejmě nad rámec reálných možností výuky statistiků, nicméně by bylo dobré, kdyby vedly k zamyšlení, zda alespoň částečně nelze výuku modifikovat ve prospěch usnadnění vstupu absolventů do praxe.

Reference

- [1] Baskerville, J. C. (1981). A systematic study on the consulting literature as an integral part of applied training in statistics. *Am. Statist.* 35, 121-123.
- [2] Boen, J., Zahn, D. A. (1982). The human side of statistical consulting. Lifetime Learning Publications, Belmont.
- [3] Cox, C. P. (1968). Some observations on the teaching of statistical consulting. *Biometrics* 24, 789-801.
- [4] DeMets, D. L., Anbar, D., Fairweather, W., Louis, T. A., O'Neill, R. T. (1994). Training the next generation of biostatisticians. *Amer. Statist.* 48, 280-284.
- [5] Derr, J. (2000). Statistical Consulting: A Guide to Effective Communication. Duxbury, Pacific Grove.
- [6] Hand, D. J., Everitt, B. S., eds. (1987). The statistical consultant in action. Cambridge University Press, Cambridge.
- [7] Hand, D. J. (1994). Deconstructing statistical questions. With discussion. *J. R. Statist. Soc. A* 157, 317-356.
- [8] Hunter, W. G. (1981). The practice of statistics: The real world is an idea whose time has come. *Am. Statist.* 35, 72-76.
- [9] Hyams, L. (1969). Letter to the Editor. *Biometrics* 25, 431-434.
- [10] Hyams, L. (1971). The practical psychology of biostatistical consultation. *Biometrics* 27, 201-211.
- [11] Kenett, R., Thyregod, P. (2006). Aspects of statistical consulting not taught by academia. *Statistica Neerlandica* 60, 396-411.
- [12] Khurshid, A., Sahai, H. (1993). A second bibliography on the teaching of statistics in biological, medical, and health sciences. *Statistica Applicata* 5, 309-397.
- [13] Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *J. Amer. Statist. Assoc.* 52, 133-142.
- [14] Kirk, R. E. (1991). Statistical consulting in a university: dealing with people and other challenges. *Am. Statist.* 45, 28-34.

- [15] Kulich, M. (2000). Úvahy nad výukou matematické statistiky na KPMS MFF UK. In: Sborník semináře TPA'2000 Diskuse k výuce statistiky. Preprint.
- [16] Malý, M., Roth, Z. (2001). Otázky komunikace statistika s lékařem. Sborník Prastan.
- [17] Pocock, S. J. (1995) Life as an academic medical statistician and how to survive it. Stat. Med. 14, 209-222.
- [18] Russell, K. G. (2001). The teaching of statistical consulting. J. Appl. Probab. 38A, 20-26.
- [19] Sahai, H., Khurshid, A. (1999). A bibliography on statistical consulting and training. Journal of Official Statistics 15 (4), 587-629.
- [20] Sprent, P. (1970). Some problems of statistical consultancy (with discussion). J. Roy. Stat. Soc. A 133, 139-165.
- [21] Taplin, R. H. (2003). Teaching statistical consulting before statistical methodology. Aust. N. Z. J. Stat. 45, 141-152.
- [22] Tobi, H., Kuik, D. J., Bezemer, P. D., Ket, P. (2001). Towards a curriculum for the consultant biostatistician: identification of central disciplines. Statist. Med. 20, 3921-3929.
- [23] van Belle, G. (1982). Some aspects of teaching biostatistical consulting. In: Rustagi, J. G., Wolfe, D. A. (eds.). Teaching Statistics and Statistical Consulting. Academic Press, New York, 343-365.
- [24] Woodward, W. A., Schucany, W. R. (1977). Bibliography of statistical consulting. Biometrics 33, 564-565.

Adresa: Marek Malý

Státní zdravotní ústav Praha a Ústav informatiky AV ČR Praha

Poděkování: Práce byla částečně podpořena výzkumným záměrem Ústavu informatiky AV ČR AV0Z10300504.

VYUČOVANIE ŠTATISTIKY V NEMATEMATICKÝCH ODBOROCH

Michal Munk, Marta Vrábelová

Abstrakt: Príspevok pozostáva zo štyroch častí. V prvej časti je zdôvodnená potreba vytvorenia špecifického učebného materiálu na podporu vyučovania výpočtovej štatistiky aj pre nematematickov. V druhej časti je predstavená elektronická kniha *Analýza dát*, ktorá okrem všeobecného základu zo štatistiky (*Získavanie dát, Exploračná analýza, Inferenčná analýza*) obsahuje popis konkrétnych modulov (*Základná štatistika (popisná štatistika, parametrické a neparametrické metódy), Analýza spoľahlivosti/prvkov (položiek), Regresná analýza, Analýza rozptylu a kovariancie*), ktorých použitie je ilustrované na riešených príkladoch prostredníctvom štatistického programového systému. V tretej časti je popísaná elektronická pomôcka *Stromový graf analytických metód*, ktorá slúži k výberu správnej metódy a je prepojená s obsahom knihy. V závere sú naznačené možnosti rozšírenia jej obsahu o aplikovanie metód strojového učenia za účelom analýzy dát z databáz (objavovanie znalostí z databáz, hĺbková analýza) a viacrozmerných prieskumných techník.

1. Učebný materiál na podporu vyučovania štatistiky alebo HelpDesk k analýze dát

Učebný materiál *Analýza dát* na podporu vyučovania výpočtovej štatistiky má podobu elektronickej knihy, ktorá môže byť publikovaná ako na CD nosiči, tak aj priamo na webe:

<http://www.stat.studnet.sk/>

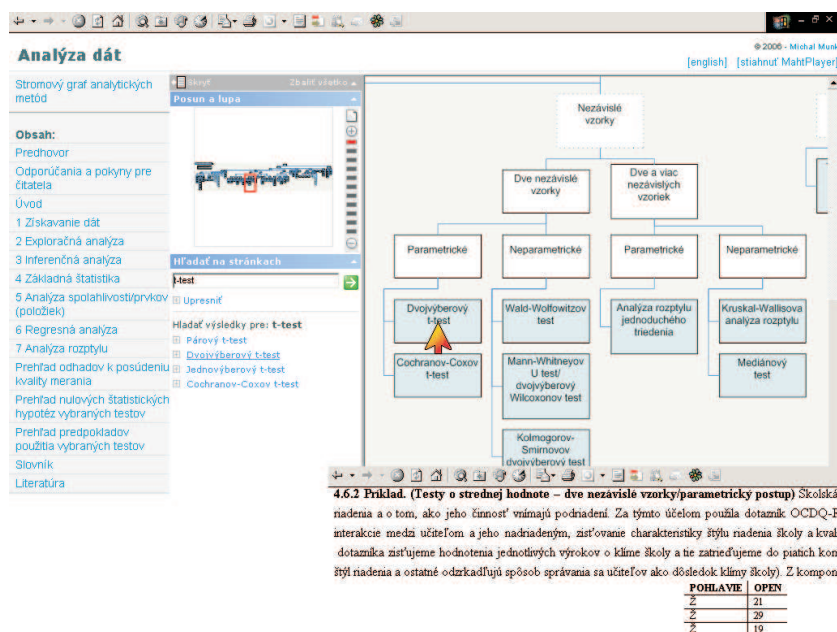
Táto elektronická kniha je primárne určená pre univerzitné kurzy zo štatistiky aj pre nematematickov. Snažili sme sa ňou, čo najlepšie pokryť hlavne najpoužívanjšie štatistické metódy. V podobe webu sekundárne slúži ako HelpDesk k analýze dát, napr. pri realizácii výskumu k záverečným prácam tým, že je do nej implementovaný *Stromový graf analytických metód*, ktorý slúži k výberu správnej metódy s prepojením na jej obsah. Z obrázka 1 je vidieť výber metódy a následné prepojenie s obsahom – v tomto prípade s riešeným príkladom, ktorý ilustruje použitie dvojvýberového t-testu.

Využívať „silné“ nástroje, ktoré nám dnes ponúka štatistický softvér sa dá iba za predpokladu, že budeme mať komplexný prehľad o analytických metódach, budeme si vedieť vybrať správnu metódu a správny postup pri

analýze dát. Vychádzali sme z potreby vytvoriť materiál, ktorý by bol aj pomôckou pri analýze dát a nie len klasickým učebným materiálom, pričom sme veľký dôraz kládli na vizualizáciu problematiky.

Technické požiadavky na spustenie:

- Pre prístup k elektronickej knihe je možný akýkoľvek počítač s prehliadačom Internet Explorer 6.0 a novším.
- MathPlayer, ktorý zabezpečí zobrazenie MathML v prehliadači.
- Doporučené grafické rozlíšenie 1024x768.



Obr. 1: HelpDesk k analýze dát

2. Obsah, štruktúra a odporúčania k vytvorenej elektronickej knihe

Naším cieľom bolo napísať prehľadovú publikáciu zo štatistiky, ktorá by prostredníctvom riešených príkladov oboznámila čitateľa s analýzou dát a s prácou so štatistickým softvérom.

Text je členený do dvoch častí.

Prvá viac-menej teoretická časť sa delí na získavanie dát, exploračnú analýzu a inferenčnú analýzu. V tejto časti čitateľ získa celkový prehľad zo štatistiky a v nasledujúcej praktickej časti si potom môže prehľbovať poznatky o konkrétnych štatistických metódach na riešených príkladoch.

V kapitole *Získavanie dát* sa čitateľ oboznámi s meracími procedúrami a posudzovaním ich kvality, so základnými výskumnými plánmi a postupmi pri ich realizácii.

V ďalšej kapitole *Exploračná analýza* sa oboznámi s jej základnými zložkami: popisnou štatistikou, vizualizáciou dát, analýzou reziduálnych hodnôt, transformáciou dát a viacrozmernými prieskumnými technikami.

V poslednej kapitole tejto časti *Inferenčná analýza* sa oboznámi s pravdepodobnosťou ako teoretickým základom inferenčnej analýzy, s odhadmi parametrov a testovaním hypotéz.

Druhá obsiahlejšia časť, sa venuje konkrétnym štatistickým metódam. Čitateľ získa teoretické poznatky o konkrétnych metódach a na riešených príkladoch sa oboznámi s postupom pri riešení konkrétnych problémov, s overovaním validity použitých metód, interpretáciou výsledkov a to prostredníctvom štatistického softvéru.

V kapitole *Základná štatistika* sa venujeme popisnej štatistike a základným parametrickým a neparametrickým metódam. V tejto časti je uvedená séria jedenástich príkladov z danej problematiky. V príkladoch okrem vysvetlenia a interpretovania použitých štatistík, mier a grafov nájdeme aj návody ako overiť predpoklady použitia jednotlivých metód a ako riešiť ich prípadné porušenia.

V kapitole *Analýza spoľahlivosti/prvkov (položiek)* sú uvedené dva príklady. V prvom prípade je vysvetlené posúdenie kvality škály dotazníka, v druhom kvality testu.

V predposlednej kapitole *Regresná analýza* je uvedená séria ôsmich príkladov. V prvom prípade sú vysvetlené a interpretované všetky možné štatistiky, miery a grafy, ktoré ponúka modul viacnásobná lineárna regresia (Multiple Regression) systému STATISTICA. V druhom nájdeme ukážku ako overiť predpoklady pre regresnú analýzu. Ďalšie príklady sa zaoberajú špeciálnymi prípadmi a metódami a to: regresnou priamkou prechádzajúcou počiatkom, kvadratickou regresiou, overovaním stability modelu, transformáciou premenných, metódou umelých premenných, korelačnou maticou ako vstupným súborom, krokovou regresnou analýzou a predpovedaním závislej premennej.

V poslednej kapitole *Analýza rozptylu* je uvedená séria siedmich príkladov. V prvom prípade sú vysvetlené a interpretované všetky možné štatistiky, miery a grafy, ktoré ponúka modul analýza rozptylu (ANOVA/MANOVA)

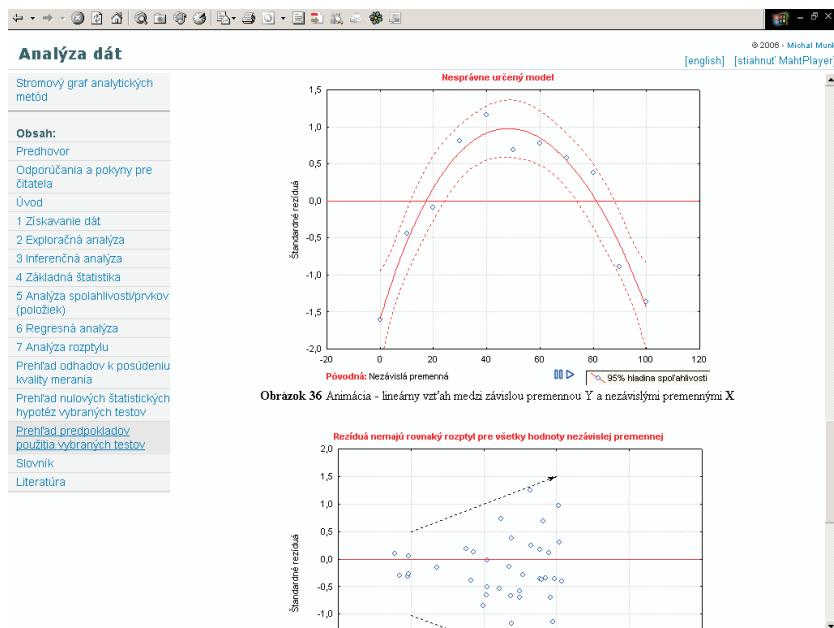
systému STATISTICA. V nasledujúcich troch nájdeme návod ako overiť predpoklady použitia (validity) jednotlivých analýz (ANOVA, MANOVA, opakované merania ANOVA, ANCOVA/MANCOVA). Ďalšie dva príklady sa zaoberajú špeciálnymi dizajnmi analýzy rozptylu (dizajn s náhodným efektom a hniezdny dizajn). Posledný príklad je ukážkou použitia kontrastnej analýzy – plánované porovnania.

Na záver sú ponúknuté ešte tri súhrny – *Prehľad odhadov k posúdeniu kvality merania*, *Prehľad nulových štatistických hypotéz vybraných testov* a *Prehľad predpokladov použitia vybraných testov*, ktorý je doplnený animáciami (obr. 2) pre zdôraznenie potreby overovať predpoklady použitia a potreby vizualizácie dát.

Pri tvorbe publikácie sme predovšetkým vychádzali z vlastných skúseností s analýzou dát a s prácou so štatistickým softvérom STATISTICA. Tomu odpovedá i dôraz na oblasť výpočtovej štatistiky. Pri spracovaní tém sme čerpali z reprezentatívnej slovenskej, českej a anglickej literatúry, ale i z množstva odborných článkov. Za všetky zdroje by sme chceli na tomto mieste uviesť knihy *Štatistické metódy* od Jiřího Anděla, *Pravdepodobnosť a štatistika* od Marty Vrábelovej a Dagmar Markechovej, *Přehled statistických metod zpracování dat* od Jana Hendla, *Štatistické metódy v pedagogike* od Gejzu Wimmera, *Basic practice of statistics* od Davida S. Moora, *Exploratory data analysis* od Johna W. Tukeya a *Electronic Statistics Textbook* od spoločnosti StatSoft.

Jeden z problémov, na ktorý sme pri písaní narazili, bola malá alebo skôr žiadna dostupnosť štatistického softvéru lokalizovaného do slovenčiny. Sčasti výnimku tvorí softvér STATISTICA od spoločnosti StatSoft, ktorý je lokalizovaný do češtiny, rovnako ako doplnok MS Excelu Analýza dát. Problémom lokalizovaných programov je, že nie ku všetkým anglickým termínom existujú jednoznačné slovenské, respektíve české ekvivalenty. Okrem toho sa často vyskytujú terminologické nepresnosti, ako príklad môžeme uviesť českú lokalizáciu Analýzy dát v Exceli, kde „df – degrees of freedom“ (stupne voľnosti) sú preložené ako „rozdíl“. Z týchto dôvodov sme pri riešení príkladov používali štatistický softvér lokalizovaný do angličtiny, čím sa tento materiál stáva použiteľný pre širšiu skupinu užívateľov, vzhľadom na fakt, že v tomto jazyku je k dispozícii väčšina štatistických programov. Anglické termíny sme sa snažili, čo najsprávnejšie preložiť a vysvetliť. Súčasťou elektronickej knihy *Analýza dát* je aj *Štatistický slovník*, ktorý obsahuje cez 300 odborných anglických termínov preložených do slovenčiny. Čitateľ má tak možnosť rozšíriť si odbornú terminológiu zo štatistiky v angličtine.

Práca nie je určená iba užívateľom programov radu STATISTICA, ale všetkým tým, ktorí k analýze dát používajú nejaký štatistický softvér, vzhľa-



Obrázok 36 Animácia - lineárny vzťah medzi závislou premennou Y a nezávislými premennými X

Obr. 2: Animácie doplnujúce Prehľad predpokladov použitia vybraných testov

dom na ich veľkú podobnosť. STATISTICU sme si vybrali z dôvodu, že obsahuje širokú paletu metód a nástrojov a tým predpokladáme, že aj užívatelia iných programov tu nájdu interpretáciu potrebných mier, štatistík a grafov.

Čitateľovi, ktorý nemá základné znalosti zo štatistiky odporúčame pred zoznamovaním sa s konkrétnou metódou v kapitolách 4, 5, 6 a 7, preštudovať si najskôr kapitoly 1, 2 a 3.

Pri realizácii výskumného plánu môže použiť metodiky uvedené v kapitole 1.

Tabuľky a obrázky (grafy, schémy a pod.) sú číslované a pomenované v celej práci okrem riešených príkladov, vzhľadom na to, že názvy tabuliek a grafov sú zhodné s príslušnou vysvetľovanou voľbou z ponuky príslušnej metódy.

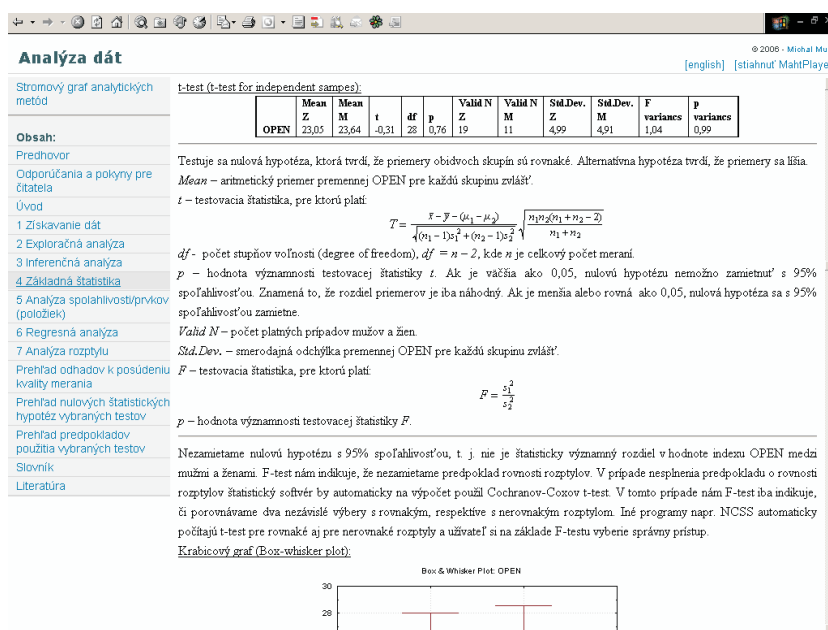
Štatistický softvér nám ponúka v rámci každej metódy/analýzy ponuku s množstvom volieb. Použité voľby v riešených príkladoch sú pod oddeľovacou čiarou vždy preložené, stručne popísané a vysvetlené, ak už tak nebolo

učinené v teoretickej časti. Pod ďalšou čiarou čitateľ nájde interpretáciu výsledkov vybranej voľby (obr. 3).

Napríklad:

Slovenský názov voľby (anglický ekvivalent):

Výstup v podobe tabuľky alebo grafu.
Popis a vysvetlenie.
Interpretácia.



Obr. 3: Výstup, popis, vysvetlenie a interpretácia voľby pre výpočet t-testu

3. Výber metódy

V prípade, že čitateľ má záujem o použitie konkrétnej analytickej metódy (štatistické metódy a metódy strojového učenia), respektíve nevie akú metódu má použiť na riešenie problému, odporúčame použiť *Stromový graf ana-*

lytických metód, ktorého úlohou je navigovať používateľa pri výbere správnej metódy.

Zobrazuje všetky dnes používané metódy, ktoré sú zatriedené do skupín. Jednotlivé úrovne v grafe majú rovnaké orámovanie. Konkrétne metódy (analýzy, testy, miery a grafy) sú zobrazené v jednej farbe. Korene, ktoré zatriedujú metódy a pomáhajú pri výbere tej správnej, nie sú farebne zvýraznené vzhľadom na ich pomocnú funkciu. Graf je doplnený o nástroje na vyhľadávanie metód („hľadať na stránkach“, „posun a lupa“), ktoré v ňom zjednodušujú orientáciu.

Graf môžeme prehľadávať dvojakým spôsobom:

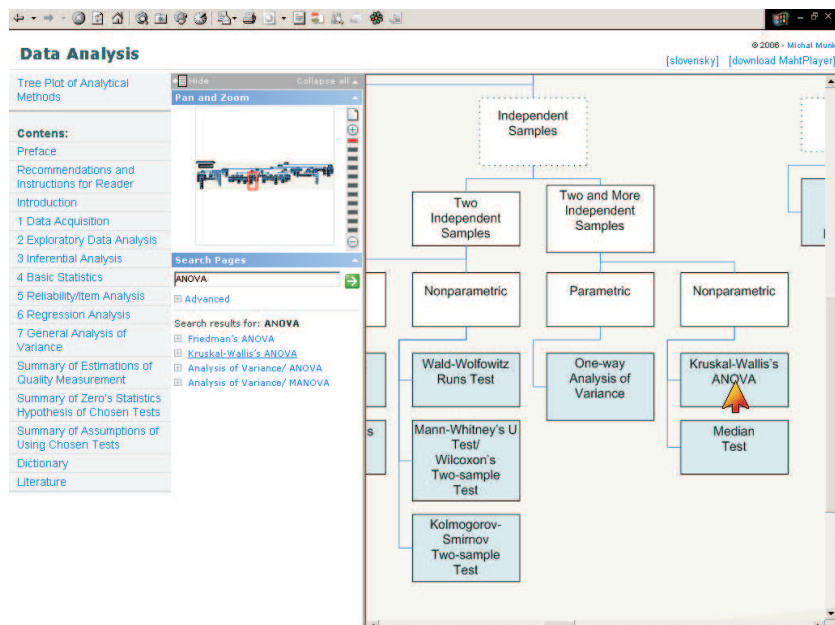
- Od všeobecného ku konkrétnemu (zhora dole), napríklad: *Štatistické metódy – Základná štatistika – Testy rozdielov medzi premennými – Testy o strednej hodnote a ich neparametrické alternatívy – Nezávislé vzorky – Dve a viac nezávislých vzoriek – Neparametrické – Kruskalova-Wallisova analýza rozptylu*.
- Od konkrétneho ku všeobecnému (zdola hore), napríklad: zadáme do vyhľadávača metód – *Znamienkový test* a zistíme na testovanie čoho a za akých podmienok (parametrické verzus neparametrické štatistiky, závislé verzus nezávislé vzorky) sa používa.

Po výbere konkrétnej metódy z grafu je čitateľovi ponúknutý materiál z danej problematiky za predpokladu, že táto metóda bola nami spracovaná. Vzhľadom na to, že graf obsahuje takmer všetky metódy, ktoré sa dnes používajú k analýze dát nebolo možné všetky spracovať. Z tohto dôvodu sme sa snažili spracovať iba tie najpoužívanejšie, a to v rozsahu 300 strán. Tým, že sú materiály v elektronickej podobe publikované na webe sú neustále prístupné študentom a navzájom prepojené. V inej podobe ako elektronickej nie je možné *Stromový graf analytických metód* publikovať, vzhľadom na jeho rozmery 1400 mm x 290 mm, pri veľkosti písma 10 b.

Elektronická kniha *Analýza dát* je z časti lokalizovaná do anglického jazyka, konkrétne bola lokalizovaná úvodná strana a *Stromový graf analytických metód* (obr. 4), na ktorom je celý materiál postavený.

4. Rozšírenie elektronickej knihy

V budúcnosti by sme radi rozšírili elektronicкую knihu *Analýza dát* o možnosti aplikovania metód strojového učenia za účelom analýzy dát z databáz. To si vyžaduje rozšíriť kapitolu *Získavanie dát* o poznatky z databáz, objavovania znalostí z databáz (Knowledge Discovery in Databases, KDD) a doplniť ďalšiu kapitolu *Hľbková analýza (Data Mining)*.



Obr. 4: *Tree Plot of Analytical Methods*

V literatúre sa často odlišujú tieto dva prístupy k analýze dát a tieto prístupy nebývajú predmetom jednej publikácie. Nepopierame, že sú tu podstatné rozdiely, napr. kým pri výskumných plánoch sú dáta získané cielene, tak aby odpovedali na dané ciele (hypotézy), v KDD sú získané z databáz, ktoré nevznikli s cieľom ich následnej analýzy, ale sú zhromaždené primárne z iných dôvodov a nemusia obsahovať požadované informácie, v KDD sa spracovávajú a analyzujú „veľké dáta“ (GB, TB) a použitie analytických metód v KDD je formalizované, kým pri výskumných plánoch použité metódy závisia na hypotézach a ich návrh je súčasťou plánu.

My sa však snažíme hľadať spoločné body v týchto na prvý pohľad odlišných prístupoch k analýze dát. V podstate môžeme KDD prirovnať k výskumným plánom. Rovnako ako v prípade výskumných plánov musíme získať dáta, napr. z nejakej výskumnej vzorky, tak aj v prípade KDD musíme zhromaždiť validné dáta z databáz. Výskumné plány rovnako ako aj KDD predstavujú určitý proces a vznikajú k nim metodiky, ktoré nám umožňujú prenášať skúsenosti z úspešných projektov. V obidvoch prístupoch na získané a spracované

dáta aplikujeme štatistické metódy, ako metódy exploračnej analýzy, tak aj metódy inferenčnej analýzy.

V KDD však okrem tradičných štatistických metód sa na analýzu dát z databáz začali používať aj metódy strojového učenia. Tieto metódy vychádzajú z empirického učenia, ktoré sa používajú pri učení sa konceptom na základe príkladov, pozorovania a objavovania. Tieto metódy sú použiteľné vďaka tomu, že v prípade analýz dát z databáz vychádzame z veľkého objemu dát. Preberaním takýchto metód, ktoré sú založené na heuristickom prieskume dát, rozvíjame otvorenú matematiku a dovoľíme si tvrdiť, že aj popularizujeme analýzu dát ako takú. Aplikáciou metód hĺbkovej analýzy študenti dokážu objaviť skryté znalosti vo veľkých dátach. A hlavne v prípade symbolických metód sa výsledky veľmi ľahko interpretujú – rozhodovacie a asociačné pravidlá sa dajú vyjadriť v prirodzenej reči a teda ľuďom ľahko rozumejú, čo je veľmi dôležité na to, aby sa získané znalosti mohli uplatniť v praxi. Napríklad aj OLAP analýzy sú v praxi veľmi často preferované len preto, že výstupy sa veľmi ľahko interpretujú a hlavne nevyžadujú hlboké vedomosti zo štatistiky a analýzy dát. Ale na druhej strane ich použitím získa užívateľ iba sumarizáciu napr. objemu predaja podľa rôznych dimenzií.

Vzhľadom na to, že chceme, aby kniha slúžila čo najširšiemu okruhu študentov i doktorandov našej univerzity, potrebujeme ju rozšíriť o niektoré mnohorozmerné štatistické metódy a to metódy klasifikačné (hlavne pre biológov a geografov) a metódy ordinačné (pre ekológov, študenti sociológie, sociálnej práce, doktorandi rôznych odborov teórie vyučovania tiež potrebujú faktorovú analýzu). Keď bude treba, budeme používať špeciálny softvér. Faktorovú analýzu ([4]) i ďalšie metódy máme čiastočne pripravené. Touto problematikou sa zaoberajú Marhold – Suda v knihe *Štatistické spracovanie mnohorozmerných dát v taxonomii. Fenetické metódy* prístupnej napr. na stránke <http://botany.upol.cz/prezentace/duch/analyza.pdf>, Lepš – Šmilauer v učebnom texte *Mnohorozmerná analýza ekologických dát* prístupnej na stránke <http://botany.upol.cz/prezentace/duch/leps.pdf>, Meloun – Milický – Hill v knihe *Počítačová analýza vícerozmerných dát v príkladoch* (Academia 2005). Túto problematiku by sme však chceli preštudovať hlbšie, prečítať aj články od Tera Braaka, napr. [2].

5. Záver

Teoretickú časť k databázam, KDD a hĺbkovej analýze máme v súčasnosti už spracovanú. V budúcnosti by sme radi spracovali konkrétne metódy zo symbolických metód strojového učenia, aby sme prezentovali použitie aj ne-

štatistických metód na analýzu dát, rozvíjali otvorenú matematiku a popularizovali analýzu dát vzhľadom na jednoduchú interpretáciu výstupov a tým, že študenti dokážu objaviť skryté znalosti vo veľkých dátach. Zaoberať by sme sa tiež chceli viacrozmernými prieskumnými technikami využiteľnými pri spracovaní výsledkov výskumu v biologických, ekologických a pedagogických vedách.

Literatúra

- [1] MUNK, M. – KAPUSTA, J.: Virtuálna škola „Štatistika“. *Forum Statisticum Slovacum*, 2005, č. 3, s. 44-49, ISSN 1336-7420.
- [2] TER BRAAK, C. J. F.: Canonical Correspondence Analysis: A new Eigenvector technique for multivariate direct gradient analysis. *Ecology* 67 (5), 1986, s. 1167-1179.
- [3] URBANÍKOVÁ, M.: O testování štatistických hypotéz trochu jinak. In *Zborník: 25. konferencia VŠTEP*. Praha: JČMF, 1998, s. 323-328.
- [4] VRÁBELOVÁ, M.: Faktorová analýza a jej výpočet v počítačových systémoch. *IKT vo vyučovaní matematiky*, UKF Nitra 2005, s. 131-142, ISBN 80-8050-925-5.

Adresa: RNDr. Michal Munk
ÚTV PF UKF v Nitre, Drážovská 4, 949 74 Nitra
E-mail: mmunk@ukf.sk

Adresa: doc. RNDr. Marta Vrábelová, CSc.
KM FPV UKF v Nitre, Tr. A. Hlinku 1, 949 74 Nitra
E-mail: mvrabelova@ukf.sk

A MULTIVARIATE TWO-SAMPLE TEST IN NONPARAMETRIC METHODS

Hidetoshi Murakami

E-mail: murakami@gug.math.chuo-u.ac.jp

Address: Department of Mathematics, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan.

Abstract: In this paper, a novel nonparametric multivariate rank test based on a Baumgartner type statistic is proposed. Simulations are used to investigate the power of suggested statistics for various population distributions.

1. Introduction

The purpose of this paper is to consider a multivariate two-sample problem, which is one of the most important statistical problems. Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$ be two random samples of size n and m independent observations, each of which has a continuous distribution $F(x)$ and $G(y)$, respectively. In nonparametric methods, the Wilcoxon test (Hollander and Wolfe; 1999) is a standard test for the location parameters such as $F(x) = G(y - \theta)$. Baumgartner *et al.* (1998) introduced a nonparametric two-sample rank test, and the power of the Baumgartner statistic is almost equivalent to the Wilcoxon test. The aforementioned authors asserted the Baumgartner statistic could be applied for a scale parameter such as $F(x) = G(y/\sigma)$ and was more powerful than the Kolmogorov-Smirnov (Gibbons; 2003) and the Cramér-von Mises (Hájek *et al.*; 1999) tests.

Let $R_1 < \dots < R_n$ and $H_1 < \dots < H_m$ denote the combined-samples ranks of the X -value and Y -value in an increasing order of magnitude, respectively. The test statistic proposed by Baumgartner *et al.* is

$$B = \frac{1}{2}(B_X + B_Y),$$

where

$$B_X = \frac{1}{n} \sum_{i=1}^n \frac{\left(R_i - \frac{n+m}{n}i\right)^2}{\frac{i}{n+1} \left(1 - \frac{i}{n+1}\right) \frac{m(n+m)}{n}}$$

and

$$B_Y = \frac{1}{m} \sum_{j=1}^m \frac{\left(H_j - \frac{m+n}{m}j\right)^2}{\frac{j}{m+1} \left(1 - \frac{j}{m+1}\right) \frac{n(m+n)}{m}}.$$

Recently, Murakami (2006) defined a k -sample Baumgartner statistic. In addition, Neuhäuser (2003) suggested the Baumgartner statistic in the presence of ties. Additionally, Neuhäuser (2001) investigated the behavior of a modified Baumgartner statistic in a one-sided test. In many cases, the location and scale parameters are tested at the same time. Then Neuhäuser (2000) introduced a modified Lepage statistic, namely L_B , which was combined with the Baumgartner and Ansari-Bradley (1960) statistics. In addition, Murakami [8] suggested a modification of L_B statistic which was combined with another modified Baumgartner statistic and the Mood (1954) statistic. A modified Baumgartner statistic proposed by Murakami (2006) was defined as

$$B^* = \frac{1}{2}(B_X^* + B_Y^*),$$

where

$$B_X^* = \frac{1}{n} \sum_{i=1}^n \frac{\left(R_i - \frac{n+m+1}{n+1}i\right)^2}{\frac{i}{n+1} \left(1 - \frac{i}{n+1}\right) \frac{m(n+m+1)}{n+2}}$$

and

$$B_Y^* = \frac{1}{m} \sum_{j=1}^m \frac{\left(H_j - \frac{m+n+1}{m+1}j\right)^2}{\frac{j}{m+1} \left(1 - \frac{j}{m+1}\right) \frac{n(m+n+1)}{m+2}}.$$

The B^* statistic is used with the exact mean and variance of R_i and H_j . The B^* statistic is more powerful than the B statistic for a location parameter when sample sizes are unequal. In addition, it is also important with a statistical problem to consider a multivariate case. For a bivariate case, Murakami [9] proposed the bivariate Baumgartner statistic and derived the limiting distribution. In this paper, we propose a multivariate nonparametric rank test in Section 2. To investigate the power of the multivariate Baumgartner statistic, we carry out simulation studies of various population distributions in Section 3. All the simulations are 10,000 repeated times and there are 10,000 permutations in this paper.

2. A multivariate statistic

In this section, we propose a multivariate Baumgartner statistic, namely B_p . Let $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(p)})'$ and $\mathbf{Y} = (\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(p)})'$, where $\mathbf{x}^{(d)} = (x_{d1}, \dots, x_{dn})'$, $\mathbf{y}^{(d)} = (y_{d1}, \dots, y_{dm})'$, $d = 1, 2, \dots, p$ are two random samples of size n and m independent observations from different populations and with p -dimensional continuous distribution $F(\mathbf{x})$ and $G(\mathbf{y})$, respectively. Suppose that $R_1^{(d)} < \dots < R_n^{(d)}$ and $H_1^{(d)} < \dots < H_m^{(d)}$ are the combined-sample ranks of the $\mathbf{X}$ -value and $\mathbf{Y}$ -value in increasing order of magnitude, respectively. This means that it is possible to obtain a separate ranking for each variable (Puri and Sen; 1971). Now we define a multivariate Baumgartner statistic as follows:

$$B_p = \sum_{d=1}^p \frac{1}{2} \left(B_X^{(d)} + B_Y^{(d)} \right),$$

where

$$B_X^{(d)} = \frac{1}{n} \sum_{i=1}^n \frac{\left(R_i^{(d)} - \frac{n+m+1}{n+1} i \right)^2}{\frac{i}{n+1} \left(1 - \frac{i}{n+1} \right) \frac{m(n+m+1)}{n+2}}$$

and

$$B_Y^{(d)} = \frac{1}{m} \sum_{j=1}^m \frac{\left(H_j^{(d)} - \frac{m+n+1}{m+1} j \right)^2}{\frac{j}{m+1} \left(1 - \frac{j}{m+1} \right) \frac{n(m+n+1)}{m+2}}.$$

We use the permutation test to estimate the p -value because it is difficult to calculate the exact critical values of B_p statistic.

3. Simulation study

Next, we investigate the behaviour of the B_p statistic. For power comparison of the statistics, we conduct a simulation study for some distributions as in different populations. In particular, we tested the hypothesis $H_0 : F(\mathbf{x}) = G(\mathbf{y})$ against $H_1 : \text{not } H_0$. We assumed that $F(\mathbf{x})$ and $G(\mathbf{y})$ described the following distributions.

1. $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$: the Normal distributions
2. $\eta(\boldsymbol{\lambda}_1)$ and $\eta(\boldsymbol{\lambda}_2)$: the exponential distributions

Suppose $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p)$. Generally, the location and scale parameters of the $\mathbf{X}$ and $\mathbf{Y}$ samples are unequal. We examined the power at which the location and scale parameters differed. The following Tables show the results of power of the multivariate Wilcoxon, namely W_p ,

$$W_p = \sum_{d=1}^p \left| \sum_{i=1}^n R_i^{(d)} - \frac{n(n+m+1)}{2} \right|$$

and the B_p statistics where $n = m = 10$ and $n = 10, m = 5$. For all cases, 10,000 times permutations in each simulation were performed, and we simulated 10,000 times to obtain the actual significance level. We treat the case of $p = 3, 4$, and 5 in this paper.

Table 1-a. Case of $n = m = 10$ for $N_3(\mathbf{0}, \mathbf{I}_3)$ and $N_3(\mu_2, \sigma \mathbf{I}_3)$

μ_2			σ			
			1.0	5.0	10.0	15.0
0.0	W_p	0.05	0.07	0.08	0.09	0.09
	B_p	0.05	0.20	0.43	0.58	0.69
0.5	W_p	0.28	0.13	0.11	0.11	0.11
	B_p	0.27	0.31	0.49	0.62	0.72
1.0	W_p	0.88	0.39	0.24	0.20	0.17
	B_p	0.85	0.59	0.63	0.71	0.77
1.5	W_p	1.00	0.74	0.48	0.35	0.29
	B_p	1.00	0.87	0.83	0.84	0.86

Table 1-b. Case of $n = 10, m = 5$ for $N_3(\mathbf{0}, \mathbf{I}_3)$ and $N_3(\mu_2, \sigma \mathbf{I}_3)$

μ_2		1.0	σ			20.0
			5.0	10.0	15.0	
0.0	W_p	0.05	0.11	0.14	0.16	0.17
	B_p	0.05	0.22	0.37	0.47	0.53
0.5	W_p	0.18	0.15	0.16	0.17	0.18
	B_p	0.19	0.29	0.41	0.49	0.55
1.0	W_p	0.67	0.31	0.25	0.23	0.23
	B_p	0.64	0.47	0.51	0.55	0.59
1.5	W_p	0.97	0.55	0.38	0.32	0.29
	B_p	0.95	0.69	0.64	0.65	0.66

Table 2-a. Case of $n = m = 10$ for $N_4(\mathbf{0}, \mathbf{I}_4)$ and $N_4(\boldsymbol{\mu}_2, \sigma \mathbf{I}_4)$

$\boldsymbol{\mu}_2$		1.0	σ			20.0
			5.0	10.0	15.0	
0.0	W_p	0.05	0.06	0.08	0.08	0.09
	B_p	0.05	0.24	0.53	0.72	0.82
0.5	W_p	0.33	0.15	0.13	0.12	0.12
	B_p	0.32	0.38	0.61	0.75	0.84
1.0	W_p	0.95	0.46	0.28	0.22	0.19
	B_p	0.92	0.69	0.76	0.83	0.89
1.5	W_p	1.00	0.84	0.56	0.42	0.34
	B_p	1.00	0.94	0.91	0.92	0.94

Table 2-b. Case of $n = 10, m = 5$ for $N_4(\mathbf{0}, \mathbf{I}_4)$ and $N_4(\boldsymbol{\mu}_2, \sigma \mathbf{I}_4)$

$\boldsymbol{\mu}_2$		1.0	σ			20.0
			5.0	10.0	15.0	
0.0	W_p	0.05	0.13	0.16	0.18	0.19
	B_p	0.05	0.27	0.47	0.58	0.66
0.5	W_p	0.21	0.18	0.19	0.20	0.21
	B_p	0.21	0.35	0.50	0.60	0.66
1.0	W_p	0.77	0.38	0.30	0.27	0.26
	B_p	0.73	0.57	0.63	0.68	0.72
1.5	W_p	0.99	0.65	0.46	0.39	0.35
	B_p	0.99	0.80	0.75	0.76	0.79

Table 3-a. Case of $n = m = 10$ for $N_5(\mathbf{0}, \mathbf{I}_5)$ and $N_5(\boldsymbol{\mu}_2, \sigma \mathbf{I}_5)$

$\boldsymbol{\mu}_2$		1.0	σ			20.0
			5.0	10.0	15.0	
0.0	W_p	0.05	0.06	0.08	0.09	0.09
	B_p	0.05	0.28	0.64	0.82	0.91
0.5	W_p	0.39	0.16	0.13	0.13	0.13
	B_p	0.36	0.44	0.70	0.84	0.91
1.0	W_p	0.97	0.53	0.33	0.25	0.22
	B_p	0.96	0.78	0.85	0.91	0.94
1.5	W_p	1.00	0.90	0.63	0.48	0.38
	B_p	1.00	0.97	0.95	0.96	0.97

Table 3-b. Case of $n = 10, m = 5$ for $N_5(\mathbf{0}, \mathbf{I}_5)$ and $N_5(\boldsymbol{\mu}_2, \sigma \mathbf{I}_5)$

μ_2		1.0	σ			20.0
			5.0	10.0	15.0	
0.0	W_p	0.05	0.13	0.17	0.20	0.21
	B_p	0.05	0.31	0.53	0.66	0.74
0.5	W_p	0.24	0.20	0.21	0.22	0.23
	B_p	0.24	0.42	0.59	0.70	0.76
1.0	W_p	0.84	0.42	0.33	0.30	0.29
	B_p	0.81	0.64	0.70	0.76	0.81
1.5	W_p	1.00	0.72	0.52	0.43	0.39
	B_p	0.99	0.86	0.82	0.84	0.86

In this case, when the location (but not the scale) was shifted, the power of the W_p statistic is greater than the B_p statistic but the difference between these two statistics was small. Furthermore, the B_p statistic was more powerful for scale and location-scale parameter shifts. Therefore, the B_p statistic is more suitable than the W_p statistic for treating the parameters associated with Normal distribution.

Table 4-a. Case of $n = m = 10$ for $\eta_3(\mathbf{1})$ and $\eta_3(\sigma \mathbf{1})$

		σ			
	1.0	2.0	3.0	4.0	5.0
W_p	0.05	0.42	0.82	0.96	0.99
B_p	0.05	0.42	0.82	0.96	0.99

Table 4-b. Case of $n = 10, m = 5$ for $\eta_3(\mathbf{1})$ and $\eta_3(\sigma \mathbf{1})$

		σ			
	1.0	2.0	3.0	4.0	5.0
W_p	0.05	0.28	0.63	0.82	0.92
B_p	0.05	0.30	0.65	0.84	0.93

Table 5-a. Case of $n = m = 10$ for $\eta_4(\mathbf{1})$ and $\eta_4(\sigma \mathbf{1})$

		σ			
	1.0	2.0	3.0	4.0	5.0
W_p	0.5	0.50	0.91	0.99	1.00
B_p	0.5	0.49	0.91	0.99	1.00

Table 5-b. Case of $n = 10, m = 5$ for $\eta_4(\mathbf{1})$ and $\eta_4(\sigma\mathbf{1})$

		σ			
	1.0	2.0	3.0	4.0	5.0
W_p	0.5	0.33	0.73	0.91	0.96
B_p	0.5	0.35	0.74	0.92	0.97

Table 6-a. Case of $n = m = 10$ for $\eta_5(\mathbf{1})$ and $\eta_5(\sigma\mathbf{1})$

		σ			
	1.0	2.0	3.0	4.0	5.0
W_p	0.05	0.56	0.96	1.00	1.00
B_p	0.05	0.56	0.95	1.00	1.00

Table 6-b. Case of $n = 10, m = 5$ for $\eta_5(\mathbf{1})$ and $\eta_5(\sigma\mathbf{1})$

		σ			
	1.0	2.0	3.0	4.0	5.0
W_p	0.05	0.37	0.79	0.94	0.98
B_p	0.05	0.39	0.80	0.95	0.99

From the results of the simulation study, the power of B_p statistic is equivalent to W_p statistic when sample sizes are equal. However, when $n \neq m$, the B_p statistic is more efficient than the W_p statistic. Therefore, the B_p statistic is more suitable than the W_p statistic for evaluating the parameters associated with the exponential distribution.

Conclusion and discussion

In this paper, we proposed a multivariate nonparametric test based on the Baumgartner type statistic. The results of our simulations of using the permutation for Normal and exponential distributions indicated that the multivariate Baumgartner statistic was more suitable than the multivariate Wilcoxon test. In the future, it will be important to derive the limiting distribution of the multivariate Baumgartner statistic.

Reference

- [1] Ansari, A. R. and Bradley, R. A. (1960), Rank sum tests for dispersion, *Annals of Mathematical Statistics*, **31**, 1174–1189.
- [2] Baumgartner, W., Weiß, P., Schindler, H. (1998), A nonparametric test for the general two-sample problem. *Biometrics*, **54**, 1129–1135.
- [3] Gibbons, J. D. and Chakraborti, S. (2003), *Nonparametric Statistical Inference*, 4th edition. Dekker, New York.
- [4] Hájek, J., Šidák, Z. and Sen, P. K. (1999), *Theory of rank tests*, 2nd edition. Academic Press, San Diego.
- [5] Hollander, M., Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, 2nd edition. John Wiley & Sons, New York.
- [6] Mood, A. M. (1954), On the asymptotic efficiency of certain nonparametric two-sample tests, *Annals of Mathematical Statistics*, **25**, 514–522.
- [7] Murakami, H. (2006), A k -sample rank test based on modified Baumgartner statistic and its power comparison. *Journal of the Japanese Society of Computational Statistics*, **19**.
- [8] Murakami, H. (Printing), Lepage type statistic based on the modified Baumgartner statistic. *Computational Statistics and Data Analysis*.
- [9] Murakami, H. (Submitting), A bivariate two-sample Baumgartner statistic. *Journal of Nonparametric Statistics*.
- [10] Neuhäuser, M. (2000), An exact two-sample test based on the Baumgartner-Weiss-Schindler statistic and a modification of Lepage's test. *Communications in Statistics. – Theory and Methods*, **29**, 67–78.
- [11] Neuhäuser, M. (2001), One-sided two-sample and trend tests based on a modified Baumgartner-Weiss-Schindler statistic. *Journal of Nonparametric Statistics*, **13**, 729–739.
- [12] Neuhäuser, M. (2003), A note on the exact test based on the Baumgartner-Weiß-Schindler statistic in a presence of ties. *Computational Statistics and Data Analysis*, **42**, 561–568.
- [13] Puri, M. L. and Sen, P. K. (1971), *Nonparametric Methods in Multivariate Analysis*, John Wiley & Sons, New York.

PROBABILITY AND QUANTUM LOGIC

Olga Nánásiová, Mária Minárová, Ahmad Mohammed

Abstrakt: This paper is a short report of some results in the theory of quantum logic. The theory of quantum logic is a generalisation of the classical probability theory.

Introduction

It is well-known fact that the classical probability space $(\Omega, \mathcal{S}, P)$ is constructed for compatible random events, it means for every couple of random events $A, B \in \mathcal{S}$ $A = (A \cap B^c) \cup (A \cap B)$ and so

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Indeed we can rewrite this as a function of two variables $Q : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$ such that $Q(A, B) = P(A \cup B)$. The same situation we obtain for intersection and symmetric difference.

Theory of quantum logic includes various algebraic structures which describe the alternative approach to random events. We consider as a basic structure an orthomodular lattice with a state, briefly an OML. In the language of quantum logic a set σ -algebra is a Boolean algebra. Each Boolean algebra is an OML but not the reverse. If we consider the function of union, intersection and symmetric difference as a function of two variables, than we can introduce such a measure also on an OML.

If we compare approaches to random events in the probability theory and in the theory of quantum logic we can find the same source. The theory of quantum logic works with the similar notions as the probability theory, but the algebraic structure is more general. Let $(\Omega, \mathcal{S}, P)$ be a probability space and $(L, I, O, \vee, \wedge, \perp)$ be an orthomodular lattice with a state (briefly quantum logic). Some relevant notions follow:

1. Ω is relevant to I , $\emptyset$ is relevant to O and $\mathcal{S}$ is relevant to L ;
2. a probability measure P is relevant to a state m ;
3. $\cup$ is relevant to $\vee$ and $\cap$ is relevant to $\wedge$;
4. A^c , where $A \in \mathcal{S}$ is relevant to $a^\perp$, where $a \in L$;
5. a probability of intersection two random events is relevant to s-map p ;
6. a probability of union two random events is relevant to j-map q ;
7. a probability of symmetric differential two random events is relevant to d-map d .

We can find some results in [2]-[9].

1. Basic notions and properties

Definition 1. 1. Let L be a nonempty set endowed with a partial ordering $\leq$. Let the greatest element (I) and the smallest element (O) exist and let the operations supremum ($\vee$), infimum ($\wedge$) (the lattice operations) be defined. Let $\perp: L \rightarrow L$ be a map with the following properties:

- (i) $\forall a, b \in L \ a \vee b, a \wedge b \in L$.
- (ii) $\forall a \in L \ \exists! a^\perp \in L$ such that $(a^\perp)^\perp = a$ and $a \vee a^\perp = I$.
- (iii) If $a, b \in L$ and $a \leq b$, then $b^\perp \leq a^\perp$.
- (iv) If $a, b \in L$ and if $a \leq b$, then $b = a \vee (a^\perp \wedge b)$ (orthomodular law).

Then $(L, O, I, \vee, \wedge, \perp)$ is called an orthomodular lattice (briefly an OML).

Let L be an OML. Then elements $a, b \in L$ will be called:

- *orthogonal* ($a \perp b$) if $a \leq b^\perp$;
- *compatible* ($a \leftrightarrow b$) if $a = (a \wedge b) \vee (a \wedge b^\perp)$

If $a, b_1, b_2 \in L$, such that $a \leftrightarrow b_i$ for $i = 1, 2$, then $a \leftrightarrow b_1 \vee b_2$ and

$$a \wedge (b_1 \vee b_2) = (a \wedge b_1) \vee (a \wedge b_2) \quad ([3]).$$

Definition 1. 2. A map $m: L \rightarrow [0, 1]$ such that

- (i) $m(O) = 0$ and $m(I) = 1$;
- (ii) If $a \perp b$ then $m(a \vee b) = m(a) + m(b)$;

is called a state on L .

Definition 1. 3. Let L be a OML. A map $d: L \times L \rightarrow [0, 1]$ will be called a difference map (d -map) if the following conditions hold:

- (d1) $d(O, I) = d(I, O) = 1 \ \forall a \in L \ d(a, a) = 0$;
- (d2) $d(a, b) = d(a, O) + d(O, b)$ for $a \perp b$;
- (d3) If $a \perp b, c \in L$,

$$d(a \vee b, c) = d(a, c) + d(b, c) - d(O, c)$$

$$d(c, a \vee b) = d(c, a) + d(c, b) - d(c, O).$$

It is possible to show that

- (1.) $\forall a \in L \ d(O, a) = d(a, O)$ is a state on L ;
- (2.) if $a \leftrightarrow b$, then $d(a, b) = d(a \wedge b^\perp, O) + d(O, a^\perp \wedge b)$;
- (3.) $\forall a \in L \ d(a, a^\perp) = 1$;
- (4.) If $a \perp b, c \in L$,

$$d(a \vee b, c) \leq d(a, c) + d(b, c)$$

$$d(c, a \vee b) \leq d(c, a) + d(c, b);$$
- (5.) $\forall a \in L \ d(a, I) = d(a^\perp, O)$;
- (6.) if $a \leftrightarrow b$, then $d(a, b) = d(b, a)$.

We can see, that for compatible elements, a d-map is a measure of a symmetric difference. One of the important notion on an OML is an s-map (the relevant notion to a measure of intersection on a Boolean algebra and a j-map (the relevant notion to a measure of union on a Boolean algebra). In the following, we introduce an s-map and a j-map and we show that it is possible to define these functions with a d-map.

Definition 1. 4 (7). *Let L be a OML. The map $p : L^2 \rightarrow [0, 1]$ will be called s-map if the following conditions hold:*

- (s1) $p(I, I) = 1$;
- (s2) if $a \perp b$, then $p(a, b) = 0$;
- (s3) if $a \perp b$ and $\forall c \in L$

$$p(a \vee b, c) = p(a, c) + p(b, c)$$

$$p(c, a \vee b) = p(c, a) + p(c, b).$$

In the paper [5] has been proved, that

- 1. For $\forall a \in L$ a map $\nu(a) = p(a, a)$ is a state on L ;
- 2. If $a \leftrightarrow b$, then $p(a, b) = \nu(a \wedge b) = p(b, a)$;
- 3. For $\forall a, b \in L \ p(a, b) \leq p(b, b)$.

There is an example that shows, that $p(a, b)$ is not equal to $p(b, a)$ in general. We say that p is commuting if $\forall a, b \in L \ p(a, b) = p(b, a)$.

In the following part we put some properties for the map q , which is defined as follow: $q(a, b) = p(a, a) + p(b, b) - p(a, b)$.

In [8] has been shown that

1. For $\forall a \in L \ q(a, O) = q(O, a) = q(a, a)$;
2. For $\forall a \in L \ q(a, I) = q(I, a) = 1$.
3. if $a \leq b$, then $q(a, b) = q(b, b)$;
4. for $\forall a \in L, \ q(a, a) = p(a, a) = \nu(a)$;
5. if $a \leftrightarrow b$, then $q(a, b) = q(a \vee b, a \vee b)$.

Definition 1. 5 (8). *Let L be an OML. A map $q : L \times L \rightarrow [0, 1]$ will be called a join map (j -map) if the following conditions hold:*

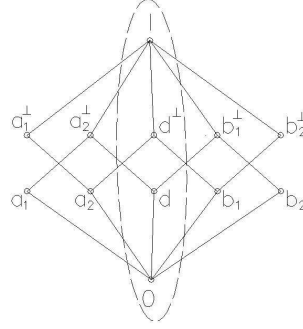
- (q1) $q(O, O) = 0$ and $q(I, I) = 1$;
- (q2) If $a, b \in L$ and $a \perp b$, then $q(a, b) = q(a, a) + q(b, b)$;
- (q3) If $a, b \in L$, then for each $c \in L$

$$q(a \vee b, c) = q(a, c) + q(b, c) - q(c, c)$$

$$q(c, a \vee b) = q(c, a) + q(c, b) - q(c, c).$$

Proposition 1. 1. *Let L be an OML and let p be an s -map. Let $q(a, b) = p(a, a) + p(b, b) - p(a, b)$. Then*

- (i.) if $\forall a, b \in L \ d(a, b) = p(a, b^\perp) + p(a^\perp b)$, then d is a d -map;
- (ii.) if $\forall a, b \ d(a, b) = q(a, b) - p(a, b)$, then d is a d -map;
- (iii.) $\forall a \in L \ p(a, a) = q(a, a) = d(a, O)$.



2. Examples

In this part we show two examples of quantum logics.

Example 2. 1. Let $\Omega = \{1, 2, 3, 4, 5\} = I$ and $\emptyset = O$. Let $a_1 = \{1, 2\}$, $a_2 = \{3, 4\}$, $d = \{5\}$, $b_1 = \{1, 3\}$ and $b_2 = \{2, 4\}$. The orthomodular lattice L , which is generated by these elements is in the following picture.

It is clear, that in this case $c \in L$ implies $c \in 2^\Omega$. Let P be a probability measure on 2^Ω . Let us denote

$$p(u, v) = P(u \cap v),$$

for example $p(a_1, b_1) = P(\{1, 2\} \cap \{1, 3\}) = P(\{1\})$. We can see, that $a_1 \wedge b_1 = O$, but $a_1 \cap b_1 = \{1\}$. It is not difficult to show, that p is s -map.

Example 2. 2. Let L be the set of all symmetrical matrices over R^n such that $AA = A$. It can be shown, that a map $p(A, B) = \frac{\text{Tr}(AB)}{n}$ is the s -map. Then

$$q(A, B) = \frac{\text{Tr}(A)}{n} + \frac{\text{Tr}(B)}{n} - \frac{\text{Tr}(AB)}{n}$$

is the j -map. Especially, if $n = 2$ and $A \neq O, I$, where I is the identity matrix and O is the zero matrix, then

$$A = \begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix},$$

where $a^2 + b^2 = 1$ and I, O .

If we put $a = \cos(\alpha)$, then

$$A = \begin{pmatrix} \cos^2 \alpha & \cos \alpha \sin \alpha \\ \cos \alpha \sin \alpha & \sin^2 \alpha \end{pmatrix},$$

where $\alpha \in R$ ($A := a(\alpha)$). $\mathcal{B}_a = \{O, I, A, A^\perp = I - A\}$ is the Boolean sub-algebra of L . In this case

$$p(a(\alpha), a(\beta)) = \frac{\cos(\alpha - \beta)^2}{2}$$

$$q(a(\alpha), a(\beta)) = \frac{1 + \sin(\alpha - \beta)^2}{2}$$

$$d((\alpha), a(\beta)) = q(a(\alpha), a(\beta)) - p(a(\alpha), a(\beta)).$$

Acknowledgement

This work was supported by Science and Technology Assistance Agency under the contract No. APVV-0375-06, VEGA 1/4024/07, VEGA-1/3321/06.

References

- [1] Adenier G., Khrennikov A., Yu.: Anomalies in EPR-Bell Experiments. AIP Conference Proceedings, QUANTUM THEORY: Reconsideration of Foundations 3, Jan.4, 810 (2006), pp. 283-293.
- [2] Dvurečenskij A., Pulmannová S.: New Trends in Quantum Structures. Kluwer Acad. Publ., 516 (2000).
- [3] Pták, P., Pulmannová S.: Quantum Logics. Kluwer Acad. Press, Bratislava (1991).
- [4] Riečan, B., Neubrun, T. Integral, Measure and Ordering, Kluwer Acad. Press, Bratislava (1997).
- [5] Khrennikov, A., Nánásiová, O.: Representation theorem of observables on a quantum system. Int. Journ. of Theor. Phys. (2006).
- [6] Nánásiová, O.: Principle conditioning. Int. Jour. of Theor. Phys., 43 (2004), pp. 1383-1395.
- [7] Nánásiová, O.: Map for Simultaneous Measurements for a Quantum Logic. Int. Journ. of Theor. Phys., 42 (2003), pp. 1889-1903.

- [8] Bohdalová M., Minárová M., Nánásiová O.: A note to algebraic approach to uncertainty. Forum Stat. Slovaca, (2006), 3, pp. 31-39.
- [9] Nánásiová O., Minárová M., Mohammed A.: Measure of "symmetric difference", Proc. Magia, (2006), ISBN 978-80-227-2583-5, ISBN 80-227-2583-8, pp. 55-60.

Address: Olga Nánásiová, Mária Minárová, Ahmad Mohammed
Department of Mathematics, FCE Slovak University of Technology,
Radlinského 11, 813 68 Bratislava, Slovak Republic
E-mail: olga@math.sk, minarova@math.sk, mohammed@math.sk

VÝUKA JEDNOROZMĚRNÉ A DVOUROZMĚRNÉ ANALÝZY KATEGORIÁLNÍCH DAT

Hana Řezanková

Abstract: The paper is focused on teaching one-way and two-way analysis of categorical data. It is based on experiences with teaching at the University of Economics, Prague. The preparing of teaching, the choice of the software and data files and the comparison of some software packages are dealt with. The possibilities of systems SAS Enterprise Guide, SPSS, S-PLUS, STATGRAPHICS and STATISTICA are compared in the area of categorical data analysis.

1. Příprava předmětu

Tento příspěvek se věnuje jednak obecně předmětům zaměřeným na analýzu dat, jednak konkrétně výuce základů analýzy kategoriálních dat v bakalářském, event. magisterském studiu. Předpokládá se tedy, že buď celá nebo část výuky se koná na počítačové učebně vybavené vhodným programovým vybavením. Mnohé úvahy vycházejí ze specifické organizace studia na Vysoké škole ekonomické v Praze. Avšak vzhledem k tomu, že kreditní způsob studia, včetně ECTS kreditů, má jistý obecný základ, neměly by být mezi vysokými školami zásadní odlišnosti v oblastech, o kterých bude dále pojednáno.

Připravuje-li pedagog na vysoké škole výuku předmětu, je často limitován různými faktory. Málokdy má plnou volnost, aby předmět akreditoval podle svých představ. Někdy je již předmět akreditován a pedagog má k dispozici obsah, který v rámci stanoveného počtu hodin naplňuje výkladem teorie a řešením příkladů.

Pokud je pedagog garantem a připravuje podklady pro akreditaci předmětu, pak je třeba zohlednit zejména následující:

- zda je pro předmět určeno, *kolik hodin týdně* má být vyučován, případně je-li dána dolní či horní hranice této hodinové dotace (pokud je v této oblasti určitá volnost, pak záleží na rozsahu a hloubce látky, kterou by v rámci daného tématu měli zvládnout studenti určitého oboru, stupně studia, případně semestru či ročníku),
- zda je u předmětu vymezeno *členění na přednášky a cvičení* a pokud ano, zda se bude na počítačové učebně konat také přednáška, či nikoli

(nemusí být součástí akreditace, lze měnit operativně například podle počtu přihlášených studentů v semestru, viz níže),

- jaký *software* je pro analýzu dat k dispozici (v době akreditace nemusí být žádný, může být jeden, či více specializovaných programových produktů, používaný software lze měnit operativně podle znalostí či zájmů studentů účastnících se kurzů),
- zda existuje vhodný *studijní materiál* dostupný v potřebném počtu studentům.

K prvnímu bodu zřejmě není potřeba žádný komentář. Přejdeme tedy k bodu druhému. Pokud má smysl rozlišovat přednášky a cvičení, pak můžeme uvažovat dvě základní situace. Je-li studentů více, než předpokládaná kapacita učebny na cvičení, a je menší kapacita pedagogů, pak je výhodné zorganizovat přednášku pro všechny studenty daného předmětu v semestru (varianta A). Na druhou stranu, pokud kapacita počítačových učeben s vhodným programovým vybavením stačí pojmout všechny zájemce o předmět a je dostatečná kapacita pedagogů, je výhodné, aby se i přednášky konaly na počítačové učebně, případně se přednášky a cvičení nerozlišovaly, tj. forma výuky může být buď označena jako přednáška nebo jako cvičení (varianta B). Pak lze kombinovat teoretický výklad s bezprostředním praktickým využitím určité metody. Nevýhodou může být někde menší tabule na počítačových učebnách, což lze kompenzovat například promítáním vzorců z dokumentu na počítači pomocí datového projektoru.

Není-li zatím k dispozici *specializovaný software*, je vhodné podniknout kroky k jeho získání (zakoupit z prostředků školy, resp. fakulty, požádat o grant FRVŠ apod.). Otázkou je, který produkt pořídit. Problematika výběru vhodného softwaru bude probrána v dalším odstavci a následně později podrobně v souvislosti s výukou analýzy kategoriálních dat. Na většině škol¹ je hlavním hlediskem cena. Ovšem když vezmeme v úvahu, že lze o pořízení programového systému požádat prostřednictvím grantu, nemusí být nutně cena limitujícím faktorem. Je důležité zohlednit, zda by mohl být software využit i v jiných předmětech. Pokud by se nepodařilo software získat do začátku výuky, pak lze provádět některé výpočty v systému MS Excel (zejména transformace dat, tabulky rozdělení četností, dosazování do vzorců), využívat prostředků na Internetu (viz [7]) aj.

Je-li softwarových produktů k dispozici více, je třeba zvážit, který je pro studenty daného oboru a stupně studia nejvhodnější. Buď ten, který již znají,

¹Termínem „škola“ bude nadále označován vysoká škola či fakulta, přesněji řešeno subjekt, který spravuje počítačové učebny, zabezpečuje nákup a instalaci programových systémů aj.

nebo ten, ve kterém jsou vyučované metody zastoupeny co nejvíce, a to jak pokud jde o počet metod, tak co se týká použitých postupů, dílčích možností, grafických výstupů apod. Nejsou-li studenti se softwarem obeznámeni, dalším zvažovaným faktorem by měla být snadnost ovládání. Studenti by se měli soustředit na analýzu dat a interpretaci výsledků a neměli by se příliš „rozptylovat“ výukou programového systému (s výjimkou situace, kdyby byl software využíván v dalších navazujících předmětech a součástí stávajícího předmětu by měla být výuka samotného programového systému).

Kromě samotných metod je třeba vzít v úvahu možnosti programového systému v oblasti přípravy dat, jejich popisu, transformací a práce s chybějícími údaji. Důležitým faktorem je také to, jaké *existující datové soubory* by měly být při výuce využívány. Převést samotná data z jednoho systému do druhého problém obvykle není. Problémem je to, že se třeba nepřevdou popisy proměnných, popisy použitých kódů (číselníky), či identifikace chybějících údajů, což může znesnadnit výuku.

Se softwarem úzce souvisí dostupný studijní materiál. Pokud jsou k dispozici skripta nebo existuje cenově dostupná kniha zaměřená na vyučované téma a obsahuje příklady s využitím určitého programového systému, pak při možnosti výběru produktu by tato skutečnost měla být zohledněna. Jinak je vhodné nějaký materiál připravit, alespoň formou dílčích dokumentů poskytovaných pouze účastníkům příslušných kurzů. I když si v současné době ještě někteří studenti zapisují při výuce poznámky, stále více se dožadují skript nebo knih obsahově přesně korespondujících s vyučovaným předmětem.

Dosud nerozsáhlejším studijním materiálem k předmětům zaměřeným na analýzu kategoriálních dat, který je dostupný v češtině, je kniha [4]. Na VŠE jsou používána například skripta [5]. Částečný výklad některých pasáží je též k dispozici elektronicky na Internetu v rámci interaktivní učebnice IASTAT, viz [7]. Text však již nebyl delší dobu aktualizován a některé úpravy by bylo vhodné provést. Při přípravě studijních materiálů se lze kromě anglické a české literatury (seznam základní viz [5]) inspirovat i na Slovensku, viz např. [2], [3] a [8].

2. Dopad volné tvorby studijního plánu na obsah a způsob výuky

Vezměme konkrétnější situaci týkající se VŠE v Praze. Studijní plán studenta je individuální, jediným omezujícím faktorem je splnění studijních povinností ve formě počtu získaných kreditů a počtu složených zkoušek. I když s novým způsobem studia založeném na ECTS kreditech je v prvním semestru tzv. pevný rozvrh, již ve druhém a dalších semestrech se může skladba předmětů

jednotlivých studentů lišit vzhledem k tomu, že někteří neuspěli u zkoušek, jiným z důvodu nemoci byly některé předměty omluveny, další přerušili studium.

Na přednáškách a cvičeních se pak setkávají studenti různých semestrů a ročníků jednoho oboru. Pokud je předmět určen jako oborově povinný, nebo je součástí skupiny předmětů, z níž si student musí něco vybrat (oborově volitelný), jeho volná kapacita (volná místa v učebně po zapsání studentů určitého oboru) je dána k dispozici studentům ostatních oborů v podobě celoškolně volitelného předmětu. Vzhledem k malým počtům studentů v jednotlivých oborech se také často stává, že je předmět akreditován jako povinný či volně volitelný pro více studijních oborů, resp. specializací². Na přednáškách a cvičeních se pak setkávají také studenti různých oborů.

Nejde samozřejmě o to, že se tam studenti setkají, jde o to, že je potřeba pro ně připravit výuku tak, aby byla pro všechny srozumitelná a aby se pokročilejší nenudili. Je potřeba vzít v úvahu, že někteří studenti již absolvovali různé předměty a jiní ne. To souvisí i se znalostí určitého softwarového produktu. V případě, že je na škole k dispozici více programových systémů, které je možno využít, každý student může znát jiný a někdo třeba žádný. Ideální řešení pro tuto situaci asi nalézt nelze a pedagog je ve velmi obtížné pozici.

Na VŠE se navíc situace zkomplikovala tím, že byl zaveden nový způsob studia s jinými pravidly a vzhledem k malým počtům studentů v jednotlivých oborech je potřeba slučovat výuku pro různé obory. Například v původním způsobu studia byl akreditován předmět se dvěma hodinami přednášek a dvěma hodinami cvičení za 14 dní, což bylo realizováno střídavou výukou přednášek a cvičení podle lichých a sudých týdnů. V novém způsobu studia se již liché a sudé týdny nerozlišují, takže vzhledem k tomu, že jsou při výuce využívány počítače, je nový předmět akreditován se dvěma hodinami cvičení týdně. Jednoho konkrétního kurzu by se tedy měli účastnit studenti různých oborů s tím, že u některého by rozdělení na přednášky a cvičení bylo již pouze formální (výuka ve variantě B). Bohužel výše uvedené nelze zobecnit a v některých případech jsou buď nové předměty akreditovány s jiným počtem hodin výuky, nebo jsou akreditovány předměty zcela odlišné.

3. Výběr softwaru a datových souborů

Problematiku výběru softwaru lze ilustrovat na předmětu, jehož cílem je seznámit studenty především bakalářského stupně studia se základy analýzy dat uspořádaných do kontingenčních tabulek a se související problematikou.

²Na VŠE existuje tzv. vedlejší specializace, kterou si volí studenti magisterského stupně studia. Z této vedlejší specializace skládají státní zkoušku.

Na VŠE je již léta takový předmět vyučován s využitím programového systému SPSS. Dále jsou k dispozici starší systém STATGRAPHICS a systém SAS. Před několika lety byl výpočetním centrem zaveden nový způsob evidence softwaru, který vystřídal způsob sledování počtu současně spuštěných licencí s možností používání produktů s omezeným počtem licencí pouze na určitých učebnách. To je potřeba zohlednit při sestavování rozvrhu. V minulém semestru jsem jednak zapomněla dát takový požadavek na učebnu se systémem SPSS, jednak bylo z důvodu velkého počtu zájemců o předmět přidáno cvičení. Přesunutí výuky na potřebnou učebnu bylo z různých důvodů nerealizovatelné (hlavní roli hrál fakt, že počítačové učebny mají rozdílné kapacity). Při výuce byl tedy využíván především systém SAS Enterprise Guide (v dalším textu bude zkratka SAS používána výhradně pro tento produkt) a částečně též STATGRAPHICS.

Pro analýzy jsou používány různé *datové soubory*. Především to jsou datové soubory dodávané se systémem SPSS, jejichž součástí jsou popisy proměnných, popisy použitých kódů (číselníky) a specifikace kódů pro chybějící údaje. Dále je to soubor odpovědí řešitelů chemického korespondenčního semináře na několik jednoduchých otázek. Dotazník pro anketu připravili synové, kteří mi poskytli získané údaje. Soubor ve formátu SPSS je řádně popsán, ke kategoriálním proměnným existují číselníky. Na základě této tematiky jsem pro studenty připravila vzorové příklady k probírané látce, které s upravenými texty ze skript poskytují studentům jako PDF soubory.

Zmíněné datové soubory lze v jiných souborech používat pouze omezeně, protože se převedou pouze data, nikoli popisy. Postupně jsem v systému SAS přidala popisy alespoň k proměnným souboru týkajícího se korespondenčního semináře. Ve stejném semestru mi jeden student kurzu nabídl data, která získal pro svou diplomovou práci od jedné cestovní kanceláře. Data byla pořízena na základě dotazníku, který nebyl profesionální a volba otázek a nabízených odpovědí nebyla v některých případech příliš šťastná. Také vložení do systému MS Excel nebylo provedeno způsobem odpovídajícím datovým souborům vytvořeným na základě profesionálních sociologických šetření. Přesto jsem začala tento soubor s oblibou používat, dokonce i v následujícím semestru, abych jednak poukázala na rozdíly mezi programovými systémy, jednak odůvodnila, proč je v případě kategoriálních dat je vhodnější vložit do tabulky pouze kódy a používat číselník.

Omezená možnost používání zavedených datových souborů vedla k tomu, že jsem ve větší míře začala používat zadání ve formě kontingenční tabulky. Tím se mohou studenti seznámit s rozdíly mezi systémem STATGRAPHICS, do jehož datového editoru lze tuto tabulku zadat přímo, a ostatními systémy (SAS, příp. SPSS), kde je třeba data vložit jako kombinaci kódů a jejich

četností. Zadání v podobě kontingenční tabulky používám například u souborů, které nemohu studentům poskytnout. Je to zejména datový soubor, který jsem na základě smlouvy získala z archivu Sociologického ústavu AV ČR. K tomu později přibyl datový soubor z průzkumu o uplatnění absolventů vysokých škol. Dále lze tímto způsobem vkládat různá data publikovaná v literatuře, například známé příklady na paradoxy poměru šancí, viz [1].

Doplňkem k výše uvedeným pomůckám je již dříve zmíněná interaktivní učebnice IASTAT, pomocí níž jsou ilustrovány zejména míry variability pro nominální a ordinální proměnné, které jsou základem při zkoumání asymetrické závislosti. Možnosti jsou ovšem omezené, výpočty lze provádět pouze pro maximálně pět kategorií.

4. Porovnání programových systémů a datových souborů

V předchozí části byly zmíněny programové systémy SPSS, SAS a STATGRAPHICS (dále SG), které lze využít pro analýzu kategoriálních dat. Do dalšího porovnání bude navíc zahrnut systém STATISTICA (dále ST) a částečně také systém S-PLUS (možnosti nabídkového režimu), tj. systémy s jednoduchým ovládáním vhodným pro výuku.

Hledisek pro porovnání programových systémů lze vymyslet velké množství. V tomto příspěvku budou vybrány pouze některé. Výše byl již například zmíněn *způsob vstupu dat*. Programové systémy obvykle umožňují vstup zdrojových dat, tj. vstup datové matice, v níž řádky odpovídají statistickým jednotkám a sloupce statistickým znakům (proměnným). Není to však pravidlem.

Pokud jde o *jednorozměrnou analýzu*, pak k provedení *binomického testu* některé systémy požadují zadat již zjištěné četnosti. Systém SG vyžaduje relativní četnost sledované kategorie (předtím tedy musí být vytvořena tabulka četností jiným způsobem), pro S-PLUS jsou vstupními parametry absolutní četnost sledované kategorie a celkový rozsah výběru. Systém ST tímto testem zřejmě nedisponuje, pro soubory většího rozsahu může být pro použití *chí-kvadrát test dobré shody*. Pokud jde o *chí-kvadrát test*, pak v SG existuje pouze jeho aplikace na shodu s konkrétním pravděpodobnostním rozdělením, systém S-PLUS touto nabídkou také přímo nedisponuje. V systému ST je potřeba zadat do jedné „proměnné“ (sloupce) zjištěné četnosti a do druhé četnosti očekávané. To je však z hlediska vstupu dat v systému ST výjimka.

Obecně ST, stejně jako systémy SAS a SPSS, umožňují jak vstup zdrojových dat, tak *vstup četností*. Ve druhém případě se do jedné proměnné zadají varianty hodnot a do druhé četnosti odpovídající jednotlivým varian-

tám. V SPSS se pak najednou před všemi analýzami specifikuje, že proměnná s četnostmi obsahuje *váhy*. V systému SAS se obdobná specifikace musí provést před každou analýzou tím, že se proměnná obsahující četnosti definuje jako *frequency variable*. V systému ST je tato možnost obsažena v rámci každé analýzy s tím, že lze specifikovat, zda mají být váhy použity pouze pro danou analýzu, nebo i pro všechny další (bohužel tuto možnost nelze využít pro chí-kvadrát test).

V případě *dvourozměrné analýzy* lze v systémech SAS, SPSS a ST postupovat zcela stejným způsobem. Zadáváme tedy buď zdrojová data, nebo kombinace variant hodnot do dvou proměnných a do třetí četnosti odpovídající jednotlivým kombinacím. Systém ST navíc v případě čtyřpolních tabulek umožňuje zadávat přímo sdružené četnosti, a to do čtyř speciálních políček. V systémech SG a S-PLUS může datový editor obsahovat buď zdrojová data, nebo sdružené četnosti uspořádané do kontingenční tabulky. V systému SG je třeba zvolit vhodnou proceduru, v S-PLUS lze při analýze specifikovat, že datový editor obsahuje kontingenční tabulku.

Zdánlivě podružnější se může jevit *možnost popisu jednotlivých variant hodnot*, která je nejvíce propracována v SPSS. Tento systém je primárně určen ke zpracování dat z dotazníků, kde převažují kategoriální proměnné. Vytvoření datového souboru pomocí číselných kódů odpovědí a číselníků těchto kódů poskytuje řadu výhod. Jedna výhoda se týká vkládání a uchování dat, kdy jsou data snadněji kontrolovatelná a menší co do objemu. Druhá výhoda je při analýzách, zejména v případě ordinálních proměnných. Tabulky a grafy četností se v případě textových hodnot vytvářejí tak, že se kategorie uspořádají podle abecedy, což nemusí odpovídat jejich pořadí na ordinální škále. Z toho vyplývá potřeba použití číselných kódů. Bez odpovídajících číselníků však jsou výsledné tabulky a grafy obtížně interpretovatelné, analytik je musí v konečné fázi jednotlivě popsat. Číselníky v systému SPSS umožňují výsledné tabulky a grafy popisovat automaticky.

Dalším specifikem datových souborů vytvořených na základě dotazníků je *velký podíl chybějících údajů*. Pro proměnné obsahující číselné kódy (a čísla) umožňují některé systémy specifikovat kódy pro chybějící údaje. Výhodou systému SPSS je, že umožňuje specifikovat až tři takové kódy, případně celý interval, jehož hodnoty lze označit jako chybějící údaje. Je-li datový soubor tvořen textovými hodnotami a některý údaj chybí, pak v políčku není vložena žádná hodnota. V *tabulce četností* je pak v SPSS tato varianta uvedena jako první a její četnost se zahrnuje do výpočtu relativních a kumulativních četností. Číselné kódy jsou pro vytvoření tabulky četností přímo nutné. Při jejich použití a definování kódů pro chybějící údaje se v jednorozměrné ta-

bulce počítají dvě varianty relativních četností, jedna pro všechny hodnoty včetně chybějících údajů a druhá pouze pro tzv. platné hodnoty.

V systému SAS se v případě chybějících údajů nekládá nic. Pro jednorozměrnou tabulku četností se řádky s takovými políčky nezahrnují do analýzy, pouze se pod tabulkou vypíše jejich počet. Jsou-li v proměnné vyjadřující určité aktivity pouze jedničky a prázdná políčka, pak výsledná tabulka četností obsahuje pouze jeden řádek a nelze přímo vyčíst, kolik procent respondentů na danou otázku odpovědělo kladně (řešením je nahradit prázdná políčka nulovou hodnotou). Obdobný výsledek pro data obsahující prázdná políčka získáme v systému SG s tím, že se nevypisuje počet chybějících údajů. Ten si tedy musíme zjistit odečtením zjištěného počtu platných hodnot od celkového rozsahu výběru (počtu řádků v tabulce).

Systém ST chybějící údaje bere do úvahy. Tabulku četností můžeme získat dvěma různými způsoby. Při způsobu, kdy systém sám navrhuje intervaly (bez ohledu na počet variant hodnot), se zobrazují dvě varianty četností stejně jako v SPSS, tj. pouze pro platné hodnoty a včetně chybějících údajů. Při zobrazení četností pro jednotlivé kategorie se při výpočtu relativních četností vychází z celého rozsahu výběru, tj. včetně chybějících údajů.

Pokud jde o porovnání systémů z hlediska samotné analýzy kategoriálních dat, pak se v tomto příspěvku zaměříme pouze na její základy. V oblasti **jednorozměrné analýzy** to jsou kromě tabulek a grafů četností již výše binomický test a chí-kvadrát test dobré shody. V oblasti dvourozměrné analýzy pak testy a míry závislosti pro kontingenční tabulky, případně některé další neparametrické testy.

Binomický test se v programových systémech vyskytuje ve třech implementacích, a to jako exaktní s využitím binomického rozdělení, jako asymptotický s využitím aproximace normovaným normálním rozdělením bez korekce a jako asymptotický s korekcí (při výpočtu testové statistiky se v čitateli přičítá hodnota 0,5). Podrobnější popis této problematiky lze nalézt v [6]. SAS zahrnuje první dvě možnosti s tím, že jejich výběr zcela závisí na uživateli. SPSS dle popisu algoritmů disponuje možností první a třetí, přičemž pro rozsah výběru do 25 včetně je na výstupu uvedeno, že je použito binomické rozdělení, a pro větší výběry se uvádí, že byla použita aproximace. Výsledky pro tyto větší výběry však odpovídají hodnotám distribuční funkce binomického rozdělení. Systémy SG a S-PLUS aplikují pouze exaktní test. Tyto systémy také umožňují, aby si uživatel zvolil jednu ze tří variant alternativní hypotézy. SAS zobrazuje výsledky pro relevantní jednostrannou a pro oboustrannou hypotézu, SPSS zobrazuje výsledky pouze jedné varianty dle kontextu, viz [6].

Chí-kvadrát test dobré shody může (stejně jako binomický test) v různých systémech vyžadovat různý vstup dat, viz výše. SAS umožňuje testovat pouze shodu s diskrétním rovnoměrným rozdělením (tj. shodu relativních četností pro všechny kategorie). SPSS má tuto možnost sice prioritně nastavenou, ale umožňuje též zadat seznam očekávaných četností. Systém ST vychází z četností zadaných do datového editoru, viz výše. SAS na rozdíl od ostatních systémů nabízí navíc exaktní variantu, jejíž výsledek je v případě dvou kategorií shodný s výsledkem exaktního binomického testu.

Zajímavé je také zařazení výše uvedených testů do nabídek systému. V SPSS jsou oba testy zařazeny do skupiny neparametrických testů. V systému SAS jsou testy nabízeny v rámci možností jednorozměrné tabulky četností. Binomický test je v systému SG zařazen k testování hypotéz pro jednu proměnnou (tedy do stejné skupiny jako parametrické testy) a v S-PLUS ve stejné skupině jako kontingenční tabulky, viz níže. Chí-kvadrát test v systému ST najdeme v nabídce *Neparametrická statistika*.

Kontingenční tabulky zahrnují vždy dvě základní oblasti, a to charakteristiky políček tabulky a charakteristiky závislostí dvou sledovaných kategoriálních proměnných. V prvním případě jde kromě sdružených a marginálních zjištěných absolutních četností též o různé varianty relativních četností (řádkové, sloupcové a na základě celé tabulky), o četnosti očekávané, rezidua a dílčí výpočty pro chí-kvadrát test o nezávislosti. I když možnosti jednotlivých uvažovaných systémů se i v této oblasti liší, nebudou zde podrobně rozvedeny, protože je zde pouze o pomocné nástroje pro sledování závislostí.

Ve druhém případě jde o různé testy a o výběrové míry závislosti. Z testů jde především o testy nezávislosti v kontingenční tabulce, případně o McNemarův test shody četností v políčkách na (vedlejší) diagonále ve čtyřpolní tabulce. Další testy se týkají testování nulovosti některých koeficientů závislosti (resp. logaritmu odhadu míry, jako v případě poměru šancí). Pokud jsou však tyto testy implementovány, jejich výsledky jsou zobrazovány spolu s příslušnými výběrovými koeficienty.

První dva typy testů systémy obvykle nabízejí odděleně od měř závislosti (výjimkou je systém ST). Základem u testů *nezávislosti* je Pearsonova statistika chí-kvadrát. Pro čtyřpolní tabulku bývá navíc součástí výstupu statistika s Yatesovou korekcí a výsledek Fisherova exaktního testu (pro relevantní jednostrannou a pro oboustrannou alternativní hypotézu). Tyto možnosti zahrnují všechny zde uvažované programové systémy. Odlišnosti existují, ale nejsou rozsáhlé. S výjimkou S-PLUS je v systémech zahrnut věrohodnostní poměr, v systémech SAS a SPSS navíc Mantelova-Haenszelova statistika chí-kvadrát. SAS zobrazuje u Fisherova exaktního testu výsledky pro obě jednostranné alternativní hypotézy.

Test McNemarův není obsažen v systému SG. Protože tento test je v podstatě speciálním případem binomického testu pro shodu četností, rozdíly odpovídají rozdílům binomického testu. Můžeme tedy rozlišit exaktní test s využitím binomického rozdělení, asymptotický s využitím chí-kvadrát rozdělení bez korekce a asymptotický s korekcí (při výpočtu testové statistiky se v čitateli před umocnění odečítá hodnota 1). Podrobnější popis této problematiky lze nalézt v [6]. SAS zahrnuje první dvě možnosti s tím, že jejich výběr zcela závisí na uživateli. SPSS disponuje možnostmi první a třetí, přičemž pro součet četností do 25 včetně je použito binomické rozdělení, a pro větší výběry aproximace. Tak je tomu ovšem pouze v případě implementace zařazené k neparametrickým testům. V rámci kontingenčních tabulek je implementován McNemarův-Bowkerův test, který umožňuje porovnávat četnosti v políčkách označených vzájemně opačným pořadím indexů. McNemarův test je tedy speciálním případem pro čtyřpolní tabulku a vždy je prováděn jako exaktní. V S-PLUS je nabízena druhá a třetí varianta (standardně nastavená je možnost s korekcí, lze vypnout). V systému ST je použita třetí varianta, tj. asymptotický chí-kvadrát s korekcí. Zvláštností je, že se kromě shody četností v políčkách na vedlejší diagonále testuje také shoda četností v políčkách na hlavní diagonále. Jak je vidět, každý systém zaujímá jiný přístup.

Větší odlišnosti se vyskytují u *měr závislosti*. Nabídkový režim systému S-PLUS nenabízí žádné. U ostatních jsou samozřejmostí *míry založené na Pearsonově chí-kvadrát statistice*, jako Pearsonův kontingenční koeficient a Cramérovo V, příp. koeficient f_i . V systému SAS jsou tyto koeficienty součástí výstupu týkajícího se výsledku chí-kvadrát testu, v systému SG jsou zařazeny k symetrickým mírám. V systémech SPSS a ST je třeba vybrat je z nabídky.

Dále můžeme míry rozlišit jednak podle typu proměnných, jednak podle toho, zda jde o závislost vzájemnou (symetrické míry), nebo jednostrannou (asymetrické míry). SAS a ST neuplatňují žádnou z těchto klasifikací. Lze pouze odlišit asymetrické míry, u nichž jsou v systému SAS uváděny symboly $C|R$, resp. $R|C$ (závislost sloupcové proměnné na řádkové nebo řádkové proměnné na sloupcové) a v systému ST symboly $X|Y$, resp. $Y|X$. Systém SG organizuje výstup do dvou částí, přičemž první zahrnuje asymetrické míry (včetně jejich symetrických variant – pokud existují) a druhá míry symetrické (včetně měr založených na Pearsonově chí-kvadrát statistice). Nabídka SPSS je členěna podle typů proměnných, výstup pak primárně podle symetrických a asymetrických měr a v rámci těchto skupin pak podle typů proměnných. Jsou rozlišeny míry pro dvě nominální, dvě ordinální a dvě kvantitativní (intervalové) proměnné a míra pro závislost kvantitativní (intervalové)

proměnné na nominální (odmocnina z poměru determinace počítaného při analýze rozptylu).

Z *asymetrických měr pro nominální proměnné* obsahují všechny čtyři systémy koeficient nejistoty. V systému ST je to jediná míra pro tento typ proměnných. Ostatní tři systémy obsahují ještě koeficient lambda a SPSS navíc koeficient tau (podrobněji viz [5]). Největší zastoupení je u *měr pro ordinální proměnné*. Všechny čtyři systémy zahrnují symetrické koeficienty gama, Kendallovo tau-b a tau-c a asymetrické Somersovo d. Kromě systému SG dále obsahují Spearmanův korelační koeficient. *Míra pro kvantitativní proměnné* je zastoupena jedna, a to Pearsonův korelační koeficient (v SPSS jsou korelační koeficienty nabízeny odděleně). Tento koeficient chybí v systému ST. Dále můžeme v systémech nalézt koeficient éta určující *míru jednostranné závislosti kvantitativní vysvětlované proměnné na proměnné nominální*. Je obsažen pouze v systémech SPSS a SG.

Systém SPSS a zvláště systém SAS poskytují ještě některé další analýzy. Jsou to především charakteristiky souhlasu ve čtvercových tabulkách. Kromě již výše uvedeného McNemarova testu sem patří koeficient souhlasu kappa a analýza poměru šancí ve čtyřpolní tabulce. Ta je velmi detailně implementována právě v systému SAS.

Jak již bylo uvedeno dříve, některé koeficienty lze *testovat na nulovost* (netýká se koeficientů určených pro kvantitativní proměnné, u nichž vzhledem k diskrétnímu charakteru není splněn předpoklad normality). SPSS uvádí výsledky testů pro všechny koeficienty určené pro nominální a ordinální proměnné automaticky, v systému SAS lze tuto možnost zvolit. Výsledky jsou uvedeny pouze pro koeficienty týkající se ordinálních proměnných a koeficientu kappa. Systém SG uvádí výsledky u koeficientů korelace, tj. Pearsonova a Kendallova tau-b, systém ST u korelačního koeficientu Spearmanova. SAS na rozdíl od ostatních systémů uvádí navíc dolní a horní meze intervalového odhadu.

Závěrem této problematiky si naznačme, pod jakými nabídkami se analýza kontingenčních tabulek skrývá. V systému SAS je to jedna z hlavních analýz pod názvem *Table Analysis*. V systému SPSS jde o dílčí analýzu v rámci popisných metod, tj. v části analýz se vybírá *Descriptive Statistics* a *Crosstabs*. Obdobné je to v systému SG, kde jde o nabídky *Describe*, *Categorical Data* a *Crosstabulation* (resp. *Contingency Tables* při zadávání již zjištěných sdružených četností). V systému ST v české verzi vybereme *Základní statistiky/tabulky* a v rámci nich *Kontingenci tabulky* (nabídka testů a měr závislosti je k dispozici až po specifikaci proměnných a přechodu do druhé fáze analýzy). Poněkud jinak je tomu v S-PLUS, kde v nabídce statistických metod je třeba zvolit *Compare Samples, Counts and Proportions*

a *Chi-square Test*, případně jiný test. K dispozici jsou Fisherův exaktní a McNemarův test.

Další *neparametrické testy* použitelné pro kategoriální data a neuvedené výše jsou implementovány v systémech SPSS a ST. V oblasti dvourozměrné analýzy jde o testy pro dva či více nezávislých výběrů (sleduje se závislost ordinální proměnné na nominální) a testy pro dva závislé výběry (pro dvě ordinální proměnné, v SPSS zahrnut i McNemarův test).

Při obecném porovnání programových systémů by bylo dalším hlediskem organizace a formát výstupů, možnost jejich úpravy a převodu do systémů pro přípravu textových dokumentů a prezentací. Z hlediska výuky jde však o záležitost podružnou. Hraje sice důležitou úlohu při zpracování seminárních prací, ale ve vztahu k výše uvedeným faktorům je její význam menší.

Při výuce je kladen důraz především na získání výsledků a jejich interpretaci. Z tohoto hlediska lze drobné nedostatky vytknout systému ST v oblasti kontingenčních tabulek, kde například u koeficientu nejistoty jsou 3 varianty označeny jako X, Y a X|Y, kde poslední symbol je určen pro vzájemnou závislost, kdežto u Somersova d symboly X|Y a Y|X označují závislost jednostrannou. U výsledků testů je jeden ze sloupečků nadepsán „sv“ (stupně volnosti), ale obsahy políček zůstaly nepřeloženy, takže se vyskytuje např. „df = 1“. Dále u procedury určené pouze pro čtyřpolní tabulky je výsledek McNemarova testu označen názvem „Chí-kvadrát“ (tento název se tedy ve výstupu objevuje dvakrát, v prvním případě označuje Pearsonovu statistiku). Pod názvem „McNemarův chí-kvadrát“ se skrývá neobvyklá varianta testování shody četností v políčkách na hlavní diagonále.

5. Závěr

Pokud pedagog vyučuje již akreditovaný předmět, je omezen počtem hodin, případně rozdělením výuky na přednášky a cvičení. Obvykle je omezen také softwarovým vybavením. To však nemusí být trvalý stav, škola může získat finanční prostředky a vybavit počítačovou učebnu jiným systémem, či novější verzí stávajícího. I když pro základní výuku statistiky disponují široce zaměřené statistické systémy potřebnými metodami, v určitých oblastech se mohou implementace lišit, jak bylo uvedeno výše.

Z hlediska výuky je třeba se zaměřit, zda testy jsou prováděny jako exaktní, či jako aproximace, event. zda je při této aproximaci použita korekce, či nikoli. To se týká například binomického a McNemarova testu. Pokud jsou přednášky organizovány odděleně od cvičení, je vhodné výklad látky zaměřit primárně na možnosti používaného softwaru (a v rámci časových možností zmínit také možnosti jiné).

Pokud jde o výuku analýzy kategoriálních dat, tak z porovnávaných systémů je co do rozsahu metod je nejvíce vybaven systém SAS, v některých směrech je ovšem omezen. Chí-kvadrát test dobré shody umožňuje testovat pouze shodu četností, nejsou používány korekce při aproximaci binomického rozdělení normálním a chí-kvadrát. Dále SAS neobsahuje všechny koeficienty závislosti, které jsou zahrnuty v SPSS, a neuvádí výsledky testů na nulovost těchto koeficientů v případě měr pro nominální proměnné. Na druhou stranu jsou součástí výstupu intervaly spolehlivosti.

Systém SPSS je zase uživatelsky příjemnější v oblasti práce s popisy kódů a s chybějícími údaji. Každý systém má své určité přednosti, ať už co se týká možnosti vstupu dat, či práce s výstupy. K výuce lze tedy použít různé systémy. Ideální jsou alespoň dva, aby si studenti uvědomili, v čem se mohou programové systémy lišit a co je potřeba zohlednit při interpretaci výsledků.

Reference

- [1] Anděl, J.: Statistické modely. *Statistika*, 2003, č. 2, s. 1-17.
- [2] Luha J.: Metódy štatistickej analýzy kvalitatívnych znakov. *EKOM-STAT '93*. SŠDS, Trenčianske Teplice 1993.
- [3] Luha J.: Analýza nominálnych a ordinálnych znakov. *EKOMSTAT 2000*. SŠDS, Trenčianske Teplice 2000
- [4] Řehák, J., Řeháková, B.: *Analýza kategorizovaných dat v sociologii*. Academia, Praha 1986.
- [5] Řezanková, H.: *Analýza kategoriálních dat*. Oeconomica, Praha 2005.
- [6] Řezanková, H.: Testy pro alternativní proměnné ve statistických programových systémech. *Forum Statisticum Slovaca*, 2005, č. 2, s. 114-118.
- [7] Řezanková, H., Marek, L., Vrabec, M.: *IASSTAT – Interaktivní učebnice statistiky*. <http://iastat.vse.cz/>.
- [8] Stankovičová, I.: Ako robiť štatistiku v systéme SAS. *Výpočtová štatistika 2000*, SŠDS, Bratislava 2000, s. 74-78.

Adresa: doc. Ing. Hana Řezanková, CSc.

Katedra statistiky a pravděpodobnosti, Vysoká škola ekonomická v Praze, Nám. W. Churchilla 4, 130 67 Praha 3

E-mail: rezanka@vse.cz

VIACKRITERIÁLNE HODNOTENIE ZAMESTNANOSTI ČLENSKÝCH KRAJÍN EÚ NA ZÁKLADE VYBRANÝCH UKAZO- VATEĽOV LISABONSKEJ STRATÉGIE

Iveta Stankovičová

Abstract: The European Council held in Brussels in March 2004 invited the Commission to establish a High Level Group headed by Mr. Wim Kok to carry out an independent review to contribute to the mid-term review. Its report should identify measures, which together form a consistent strategy for the European economies to achieve the Lisbon objectives and targets. More than a hundred indicators have been associated with the Lisbon process. This makes this instrument ineffective. Simplification is vital. The establishment by the European Council of a more limited framework of 14 targets and indicators offers the opportunity to improve the working of the instrument of peer pressure. The European Commission should present to the Heads of State or Government and the wider public annual updates on these key 14 Lisbon indicators in the format of league tables with rankings (1 to 25, now 1 to 27), praising good performance and castigating bad performance. These 14 indicators offer the opportunity for Member States to further emphasise the growth and employment dimension of Lisbon if they choose. Not all Member States start from the same position, especially those who have recently joined.

The goal of this paper is multidimensional evaluation of employment in EÚ 27. We used the principal component analysis and the method of distance from fictive object.

Key words: the Lisbon strategy for growth and employment, the 14 indicators, analysis of employment in EU 27, multidimensional methods, principal component analysis, method of distance from fictive object.

1. Úvod

Európska rada na svojom zasadnutí v Bruseli v marci 2004 požiadala Európsku komisiu, aby zriadila skupinu špičkových expertov, ktorá vypracuje nezávislý príspevok ku hodnoteniu polčasu Lisabonskej stratégie. Team dostal za úlohu vypracovať správu a v nej načrtnúť stratégiu, ktorá by umožnila

európskym ekonomikám, aby lisabonské ciele splnili. Skupina mala byť obsadená expertmi, ktorí reprezentujú záujmy všetkých zainteresovaných. Vedením skupiny bol poverený Wim Kok, bývalý ministerský predseda Holandska a mala spolu 13 členov. Správa komisie mala byť vypracovaná do 1. novembra 2004 a zverejnená. Team pracoval od marca do októbra 2004, zišiel sa šesťkrát a výsledky predložili Európskej komisii 3. októbra 2004.

V úvode záverečnej správy skupiny expertov sa píše, že v marci 2000 vtedajší európsky lídri poverili EÚ stať sa do roku 2010 „najdynamickejším a najviac konkurencieschopným, na najnovších poznatkoch založeným hospodárstvom na svete, schopným udržateľného rozvoja, s viacerými a lepšími pracovnými príležitosťami a väčšou sociálnou kohéziou a rešpektom k životnému prostrediu“ [1]¹. Tento plán sa všeobecne volá Lisabonská stratégia a je v podstate súborom navzájom sa ovplyvňujúcich reforiem.

S Lisabonskou stratégiou je spojených viac ako 100 ukazovateľov, na základe ktorých sa má proces reforiem hodnotiť. Je teda pravdepodobné, že každá krajina môže byť najlepšia v tom či onom ukazovateli. V dôsledku toho sa tento nástroj stáva neúčinným. Členské štáty nemajú motiváciu na zlepšenie svojich výsledkov. Zjednodušenie systému ukazovateľov je veľmi dôležité, a preto skupina expertov navrhla zaviesť rámec len 14-tich cieľových ukazovateľov. Skupina špičkových odborníkov považuje tento zoznam hlavných ukazovateľov za najlepší kompromis ako zachovať stratégiu Lisabonu a zároveň postihnúť jej ambicióznosť a komplexnosť. Európska komisia by mala hlavám štátov i širšej verejnosti každoročne predkladať aktualizácie týchto kľúčových 14-tich ukazovateľov Lisabonu vo forme „ligových“ tabuliek s hodnotením (od 1 do 25, resp. dnes už od 1 do 27), ktoré bude oceňovať dobré a kritizovať zlé výsledky. Týchto 14 ukazovateľov poskytuje členským štátom príležitosť zdôrazniť dimenziu Lisabonu zameranú na celkový rast ekonomiky EÚ a rast zamestnanosti. Nie všetky členské štáty však začínajú na rovnakej priečke. To platí predovšetkým pre novoprijaté krajiny.

2. Obsah súboru dát a ciele viackriteriálnej analýzy

V súčasnosti má Európska únia už 27 členských štátov² (štáty EÚ 27). Na internetovej stránke Eurostatu (viď zdroje dát [1] a [2]) sa dajú vyhľadať údaje pre všetkých 15 ukazovateľov Lisabonskej stratégie za roky 2003 až 2006 za všetky členské štáty (EÚ 27), ale aj pre ostatné európske štáty a pre porovnanie aj pre USA, Kanadu a Japonsko. V údajoch sa však vyskytuje

¹http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf, str. 6

²Na začiatku roku 2007 boli prijaté do EÚ štáty Bulharsko a Rumunsko, ktoré zmenili EÚ 25 na EÚ 27.

aj niekoľko chýbajúcich hodnôt za niektoré štáty (napr. Malta) a za niektoré ukazovatele (napr. regionálny rozptyl miery nezamestnanosti).

V správe skupiny expertov sa síce uvádza počet vybraných 14-tich ukazovateľov, ale pôvodne bolo týchto ukazovateľov 15. Posledný ukazovateľ v zozname, regionálny rozptyl miery nezamestnanosti, je však ťažko zisťiteľný. Ukazovateľ nebol vykázaný pre 9 z 27 krajín EÚ za sledované roky 2003 a 2006 a preto musí byť z viacrozmerných analýz vylúčený.

Ako sme už spomínali v úvode, cieľom Lisabonskej stratégie je rast ekonomiky a zamestnanosti v celom spoločenstve krajín EÚ. Tieto dva aspekty sa pokúsime analyzovať na základe zjednodušeného systému vybraných 14-tich ukazovateľov podľa expertnej skupiny Wima Koka a to pomocou vybraných metód viackriteriálneho hodnotenia v krajinách EÚ 25, resp. EÚ 27. Cieľom tohto príspevku je analýza zamestnanosti vzhľadom k vytýčeným cieľom pre rok 2010 v tejto oblasti. Obsahom príspevku M. Vojtkovej: „Hľadanie podobnosti krajín EÚ podľa vybraných ukazovateľov Lisabonskej stratégie“, ktorý sa nachádza tiež v tomto zborníku STAKAN 2007, je zhľuková analýza krajín EÚ 25 na základe vybraných 14-tich ukazovateľov Lisabonskej stratégie za rok 2003.

3. Zoznam vybraných ukazovateľov lisabonskej stratégie

Súbor vybraných 15-tich ukazovateľov skupinou expertov EÚ pod vedením Wima Koka je nasledovný:

1. **HDP na obyvateľa v parite kúpnej sily (HDP_obyv)** – Hodnoty hrubého domáceho produktu sú v parite kúpnej sily na obyvateľa. Dáta sú poskytnuté členskými štátmi podľa normy ESA 95 (ročné dáta, národné účty a hlavné agregáty). Časový rad je vykazovaný od roku 1995, pre niektoré krajiny od roku 1991.
2. **Produktivita práce na zamestnaného v parite kúpnej sily (Prod_prac)** – Porovnanie rastu jednotkových nákladov práce k odmeňaniam zamestnancov a produktivity práce tak, aby zobrazili aký stupeň odmien zamestnancov je v relácii k ich produktivite práce. Teda ide o zobrazenie relácie koľko je zamestnanec platený ku produktivite jeho práce. Dáta sú poskytnuté členskými krajinami podľa normy ESA 95 (ročné dáta, národné účty a hlavné agregáty). Časový rad je vykazovaný od roku 1995, pre niektoré krajiny od roku 1991.
3. **Miera zamestnanosti celková (v %) (Zam)** – Zamestnané osoby vo veku od 15-64 rokov ako podiel na celkovej populácii v rovnakej

vekovej skupine – harmonizované s dátami obyvateľstva a národnými účtami. Dáta sú vykazované 12 týždňov po skončení príslušného roka ako ročný priemer.

4. **Zamestnanosti žien (v %) (Zam_zeny)** – Zamestnané ženy vo veku od 15-64 rokov ako podiel na celkovej populácii v rovnakej vekovej skupine – harmonizované s dátami obyvateľstva a národnými účtami. Dáta sú vykazované 12 týždňov po skončení príslušného roka ako ročný priemer.
5. **Miera zamestnanosti starších pracovníkov (v %) (Zam_starsi)** – Zamestnané osoby vo veku od 55-64 rokov ako podiel na celkovej populácii v rovnakej vekovej skupine – harmonizované s dátami obyvateľstva a národnými účtami. Dáta sú vykazované 12 týždňov po skončení príslušného roka ako ročný priemer.
6. **Dosiahnutý (stredoškolský) stupeň vzdelanosti mladých ľudí vo veku 20-24 rokov (Stup_vzdel)** – Osoby vo veku 20 až 24 rokov ktoré dosiahli stredoškolské vzdelanie ako percento populácie v rovnakej vekovej skupine. Časový rad je publikovaný od roku 1992 pre členské krajiny, a od 1995 postupne aj pre krajiny prístupujúce.
7. **Výdaje na výskum a rozvoj ako percentuálny podiel na HDP (Vydaje_vyskum)** – Skladajú sa zo súkromných výdajov na výskum a rozvoj, vysokoškolského vzdelávania vo výskume a rozvoji, vládnych výdajov a výdajov súkromných neziskových spoločností.
8. **Investície (v súkromnom sektore) ako percentuálny podiel na HDP (Investicie)** – Ukazovateľom sa vykazujú investície (formovanie hrubého fixného kapitálu) súkromného sektora ako percentuálny podiel na HDP.
9. **Porovnanie cenovej úrovne s úrovňou v EÚ-15 (Porovnanie_cien)** – Parity kúpnej sily sú vykazované Eurostatom. Eurostat je zodpovedný za koordináciu a kalkuláciu finálnych parít na základe dát a cenových prieskumov Národných štatistických inštitútov (NSIs).
10. **Miera riziku chudoby (Risk_chudoby)** – Percento obyvateľstva s disponibilným príjmom pod prahom chudoby, ktorý je stanovený ako 60% priemerného disponibilného príjmu po sociálnych transferoch. Tento podiel je počítaný pred sociálnymi transfermi (príjem aj s dôchodkami ale bez iných sociálnych transferov) a so sociálnymi transfermi (celkový príjem).
11. **Dlhodobá miera nezamestnanosti (Dlhodob_nezam)** – Podiel dlhodobo nezamestnaných (12 mesiacov a viac) na ekonomicky aktív-

nom obyvateľstve – harmonizované s národnými mesačnými odhadmi nezamestnanosti.

12. **Emisie skleníkových plynov (Sklen_plyny)** – Ukazovateľ je vykazovaný na základe Kyótskeho koša (Kyoto basket) šiestich skleníkových plynov (CO_2 , N_2O , CH_4 , HFCs, PFCs, SF_6) vážených na základe ich príspevku ku tvorbe globálneho otepľovania, v 1000 ton CO_2 ekvivalentu, delené emisiou v bázičkom roku.
13. **Celková vnútroštátna spotreba energie ako podiel na HDP (Spotr_energie)** – Podiel vnútroštátnej spotreby energie na HDP. Hrubá vnútorná spotreba energie je meraná v ekvivalentoch kilogramu oleja (kgoe) a HDP v 1000 EUR, výsledný pomer je kgeo na 1000 EUR. HDP je vyjadrené v stálych cenách.
14. **Množstvo nákladnej prepravy pomerne ku HDP (Nakl_preprava)** – Index množstva vnútornej nákladnej prepravy (železničná, lodná a pozemná preprava) ako podiel ku HDP, vyjadrený v tonokilometroch/HDP (v stálych cenách z roku 1995, v Eurách, 1995 = 100).
15. **Regionálny rozptyl miery zamestnanosti** – Celkový variačný koeficient ukazovateľa miera zamestnanosti (vo vekovej skupine 15-64 rokov) v regióne v rámci krajín. Ukazovateľ nebol vykázaný u 9 z 27 krajín, preto nie je možné ho využiť vo viacrozmerných analýzach.

4. Analýza zamestnanosti v členských krajinách EÚ

Na hodnotenie úrovne zamestnanosti v krajinách EÚ skupina expertov vybrala 3 štrukturálne ukazovatele zamestnanosti:

- miera zamestnanosti celková (v %) (Zam),
- miera zamestnanosti žien (v %) (Zam_zeny),
- miera zamestnanosti starších pracovníkov (v %) (Zam_starsi).

Na základe popisných štatistík pre jednotlivé ukazovatele zamestnanosti v krajinách EÚ 27 (Tabuľka 1) môžeme konštatovať, že priemerná celková zamestnanosť v roku 2006 oproti roku 2003 vzrástla (zo 63% na 65%) a znížila sa variabilita tohto ukazovateľa. Podobný vývoj bol aj v čiastkových ukazovateľoch. Rozdelenie početnosti ukazovateľov zamestnanosti je pomerne symetrické (porovnanie mediánu (Median) a aritmetického priemeru (Mean)). Extrémne nízke hodnoty sa vyskytujú len v miere zamestnanosti žien na Malte (Tabuľka 9).

Vybrané ukazovatele zamestnanosti sú výpočtovo previazané, a preto sú korelovné. Hodnoty párových Pearsonových koeficientov korelácie v sledovaných rokoch 2003 a 2006 sú vyššie ako 0,68 a všetky sú významné (Tabuľka 2 a Tabuľka 3). Vzhľadom k tejto skutočnosti môžeme skonštruovať „integrálny“ ukazovateľ zamestnanosti pomocou metódy hlavných komponentov.

Metódu hlavných komponentov sme aplikovali na korelačné matice ukazovateľov v sledovaných rokoch. Na základe vlastných čísiel môžeme konštatovať, že prvý hlavný komponent PRIN1 vysvetľuje až 84,6% (Tabuľka 4) celkovej variability dát o zamestnanosti v roku 2003 a 87,9% (Tabuľka 5) v roku 2006. Môžeme ho teda označiť za „integrálny“ ukazovateľ zamestnanosti. Všetky tri pôvodné ukazovatele sú s PRIN1 silno a pomerne rovnomerne korelované (Tabuľka 6 a Tabuľka 7).

Hodnoty komponentných skóre prvého hlavného komponentu PRIN1 pre členské krajiny EÚ 27 nám umožňujú zostrojiť grafy, na ktorých je názorne vidieť poradie krajín podľa veľkosti „integrálneho“ ukazovateľa zamestnanosti v roku 2003 (Obrázok 1) a v roku 2006 (Obrázok 2). Môžeme konštatovať, že najvyššie hodnoty vybraných štrukturálnych mier zamestnanosti dosahovali v roku 2003 štáty Švédsko, Dánsko a Veľká Británia. Na konci rebríčka sa nachádzali štáty Malta, Poľsko, Taliansko, Bulharsko a tiež Slovensko. Česká republika sa nachádzala na 13. mieste. V roku 2006 sa rebríček krajín EÚ 27 nepatrne zmenil. Švédsko, Dánsko a Veľká Británia zostali na čele rebríčka s najvyššími hodnotami integrálneho ukazovateľa zamestnanosti. Situácia sa zhoršila v Portugalsku, ale aj v Čechách a v Maďarsku. Bulharsko sa zlepšilo o 4 priečky. Postavenie Slovenska sa nezmenilo.

Komisia stanovila pre vybrané ukazovatele zamestnanosti aj cieľové hodnoty, ktoré by mali krajiny EÚ dosiahnuť v roku 2005 a 2010. Hodnoty cieľov sú vytýčené nasledovne:

- pre celkovú zamestnanosť (Zam): r. 2005 – 67%, r. 2010 – 70%,
- pre zamestnanosť žien (Zam_zeny): r. 2005 – 57%, r. 2010 – 60%,
- pre zamestnanosť starších ľudí (Zam_starsi): r. 2005 – neurčené, r. 2010 – 50%.

Cieľové hodnoty mier zamestnanosti pre rok 2010 použijeme aj v našej nasledujúcej analýze. Na zistenie vzdialenosti ukazovateľov zamestnanosti jednotlivých krajín od vytýčených cieľov aplikujeme metódu poradí na základe vzdialenosti od fiktívneho objektu. Za „fiktívne“ hodnoty pre vybrané ukazovatele použijeme cieľové hodnoty pre rok 2010. Pre jednotlivé krajiny

Tabuľka 1: Popisné štatistiky ukazovateľov zamestnanosti EÚ 27 (2003 a 2006)

	Zam03	Zam_zeny03	Zam_starsi03	Zam06	Zam_zeny06	Zam_starsi06
N	27,0	27,0	27,0	27,0	27,0	27,0
Mean	63,0	55,7	40,5	65,0	58,0	44,4
Std Dev	6,4	8,8	11,6	6,0	8,4	11,0
Minimum	51,2	33,6	23,5	54,5	34,9	28,1
Maximum	75,1	71,5	68,6	77,4	73,4	69,6
Median	62,6	57,3	40,7	65,3	59,3	44,1

Tabuľka 2: Korelačná matica (rok 2003)

Correlation Matrix			
	Zam03	Zam_zeny03	Zam_starsi03
Zam03		1	0,7348
Zam_zeny03	0,8855		1
Zam_starsi03	0,7348	0,6803	

Tabuľka 3: Korelačná matica (rok 2006)

Correlation Matrix			
	Zam06	Zam_zeny06	Zam_starsi06
Zam06		1	0,7789
Zam_zeny06	0,9085		1
Zam_starsi06	0,7789	0,767	

Tabuľka 4: Vlastné čísla z korelačnej matice (2003)

Eigenvalues of the Correlation Matrix			
	Eigenvalue	Difference	Proportion Cumulative
1	2,537	2,184	0,846
2	0,353	0,243	0,118
3	0,110		0,037
			1,000

Tabuľka 5: Vlastné čísla z korelačnej matice (2006)

Eigenvalues of the Correlation Matrix			
	Eigenvalue	Difference	Proportion Cumulative
1	2,638	2,367	0,879
2	0,271	0,180	0,090
3	0,091		0,030
			1,000

Tabuľka 6: Korelácia medzi PRIN1 a ukazovateľmi zamestnanosti (2003)

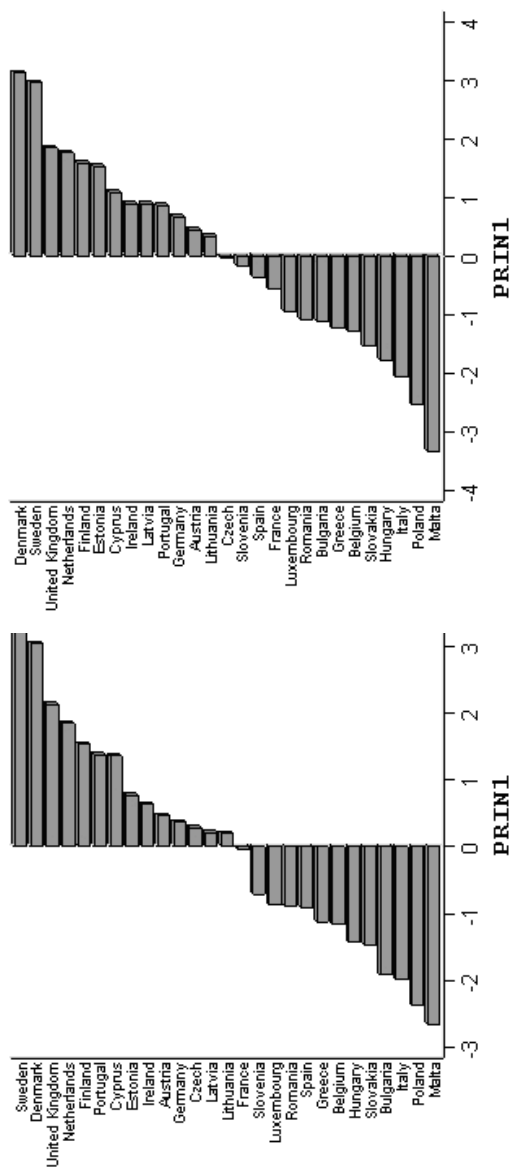
Pearson Correlation Coefficients, N = 27			
		Prob > r under H0 : Rho = 0	
PRIN1	Zam03	Zam_zeny03	Zam_starsi03
	0,954	0,934	0,869
	< 0,001	< 0,001	< 0,001

Tabuľka 7: Korelácia medzi PRIN1 a ukazovateľmi zamestnanosti (2006)

Pearson Correlation Coefficients, N = 27			
		Prob > r under $H_0 : \rho = 0$	
PRIN1	Zam06	Zam_zeny06	Zam_starsi06
	0,957	0,953	0,902
	< 0,001	< 0,001	< 0,001

Tabuľka 8: Vlastné čísla prvého hlavného komponentu PRIN1 (2003 a 2006)

Eigenvectors		
	PRIN1_03	PRIN1_06
Zam	0,599	0,589
Zam_zeny	0,586	0,587
Zam_starsi	0,546	0,555



Obr. 1: Poradie krajín EÚ 27 na základe PRIN1 (rok 2003)

Obr. 2: Poradie krajín EÚ 27 na základe PRIN1 (rok 2006)

vypočítame vzdialenosti d_{ij} od zvoleného cieľa x_{0j} pre všetkých p ukazovateľov zamestnanosti zvlášť a nakoniec zistíme priemernú vzdialenosť d_i pre krajinu podľa vzťahu:

$$d_i = \frac{1}{p} \sum_{j=1}^p (x_{ij} - x_{0j}) \quad \text{pre } i = 1, 2, \dots, n \quad \text{a } j = 1, 2, \dots, p$$

kde i označuje počet štatistických jednotiek ($n = 27$ krajín) a j označuje počet ukazovateľov hodnotenia, ktorých je p (v našom prípade $p = 3$). Konečné poradie určíme tak, že za najlepší štát v hodnotení budeme považovať ten, ktorý má najvyšiu kladnú priemernú vzdialenosť od cieľovej hodnoty. Kladné hodnoty vzdialeností znamenajú, že krajina už v roku 2003, resp. 2006 dosiahla ciele stanovené komisiou na rok 2010 a sú to najlepšie krajiny v oblasti štrukturálnych ukazovateľov zamestnanosti v rámci EÚ 27 na základe našich výsledkov (Tabuľka 10). Sú to opäť štáty Švédsko, Dánsko a Veľká Británia. Záporné hodnoty znamenajú, že krajiny boli v roku 2003, resp. 2006 ešte ďaleko od vytýčených cieľov. Patria sem krajiny Malta, Poľsko, Taliansko a aj Slovensko. Česká republika bola už v roku 2003 veľmi blízko stanoveným cieľom pre rok 2010 vo všetkých troch ukazovateľoch zamestnanosti (Tabuľka 9).

Výsledky oboch použitých metód viackriteriálneho hodnotenia zamestnanosti nám priniesli rovnaké výsledky. Na základe výsledkov metódy vzdialenosti od fiktívneho objektu sme sa okrem zistenia, ktoré krajiny sú na ktorej strane rebríčka dozvedeli, ako bola v roku 2003, resp. 2006 tá ktorá krajina ďaleko od vytýčených cieľov pre rok 2010.

Tabuľka 9: Hodnoty vybraných mier zamestnanosti za roky 2003 a 2006 pre štáty EÚ 27

Pc.	Skratka	Krajina	Zam03	Zam06	Zam_zeny03	Zam_zeny06	Zam_stars03	Zam_stars06
1	MT	Malta	54,2	54,8	33,6	34,9	32,5	30,0
2	PL	Poland	51,2	54,5	46,0	48,2	26,9	28,1
3	IT	Italy	56,1	58,4	42,7	46,3	30,3	32,5
4	HU	Hungary	57,0	57,3	50,9	51,1	28,9	33,6
5	SK	Slovakia	57,7	59,4	52,2	51,9	24,6	33,1
6	BE	Belgium	59,6	61,0	51,8	54,0	28,1	32,0
7	GR	Greece	58,7	61,0	44,3	47,4	41,3	42,3
8	LU	Luxembourg	62,2	63,6	50,9	54,6	30,3	33,2
9	BG	Bulgaria	52,5	58,6	49,0	54,6	30,0	39,6
10	RO	Romania	57,6	58,8	51,5	53,0	38,1	41,7
11	FR	France	63,3	63,0	57,3	57,7	36,8	37,6
12	SL	Slovenia	62,6	66,6	57,6	61,8	23,5	32,6
13	ES	Spain	59,8	64,8	46,3	53,2	40,7	44,1
14	CZ	Czech	64,7	65,3	56,3	56,8	42,3	45,2
15	AT	Austria	68,9	70,2	61,6	63,5	30,3	35,5
16	LT	Lithuania	61,1	63,6	58,4	61,0	44,7	49,6
17	DE	Germany	65,0	67,2	58,9	61,5	39,9	48,4
18	PT	Portugal	68,1	67,9	61,4	62,0	51,6	50,1
19	IE	Ireland	65,5	68,6	55,7	59,3	49,0	53,1
20	LV	Latvia	61,8	66,3	57,9	62,4	44,1	53,3
21	CY	Cyprus	69,2	69,6	60,4	60,3	50,4	53,6
22	NL	Netherlands	73,6	74,3	66,0	67,7	44,3	47,7
23	FI	Finland	67,7	69,3	65,7	67,3	49,6	54,5

Tabuľka 9: Pokračovanie tabuľky...

Pc.	Skratka	Krajina	Zam03	Zam06	Zam_zeny03	Zam_zeny06	Zam_starsi03	Zam_starsi06
24	EE	Estonia	62,9	68,1	59,0	65,3	52,3	58,5
25	UK	United Kingdom	71,5	71,5	65,3	65,8	55,4	57,4
26	DK	Denmark	75,1	77,4	70,5	73,4	60,2	50,7
27	SE	Sweden	72,9	73,1	71,5	70,7	68,6	69,6

Tabuľka 10: Vzďialenosti mier zamestnanosti od cieľov stanovených pre rok 2010 (2003, 2006) – zoradené podľa priemernej vzďialenosti v r. 2006 (vzostupne)

Pc.	Skratka	Vzdialenosti od cieľa 2010 (r. 2003)				Vzdialenosti od cieľa 2010 (r. 2006)			
		Zam	Zam_zeny	Zam_starsi	Priemer	Zam	Zam_zeny	Zam_starsi	Priemer
1	MT	-15,8	-26,4	-17,5	-19,9	-15,2	-25,1	-20,0	-20,1
2	PL	-18,8	-14,0	-23,1	-18,6	-15,5	-11,8	-21,9	-16,4
3	IT	-13,9	-17,3	-19,7	-17,0	-11,6	-13,7	-17,5	-14,3
4	HU	-13,0	-9,1	-21,1	-14,4	-12,7	-8,9	-16,4	-12,7
5	SK	-12,3	-7,8	-25,4	-15,2	-10,6	-8,1	-16,9	-11,9
6	BE	-10,4	-8,2	-21,9	-13,5	-9,0	-6,0	-18,0	-11,0
7	GR	-11,3	-15,7	-8,7	-11,9	-9,0	-12,6	-7,7	-9,8
8	LU	-7,8	-9,1	-19,7	-12,2	-6,4	-5,4	-16,8	-9,5
9	BG	-17,5	-11,0	-20,0	-20,0	-11,4	-5,4	-10,4	-9,1
10	RO	-12,4	-8,5	-11,9	-10,9	-11,2	-7,0	-8,3	-8,8
11	FR	-6,7	-2,7	-13,2	-7,5	-7,0	-2,3	-12,4	-7,2
12	SL	-7,4	-2,4	-26,5	-12,1	-3,4	1,8	-17,4	-6,3

Tabuľka 10: Pokračovanie tabuľky...

Pc.	Skratka	Vzdialenosti od cieľa 2010 (r. 2003)				Vzdialenosti od cieľa 2010 (r. 2006)			
		Zam	Zam_zeny	Zam_starsi	Priemer	Zam	Zam_zeny	Zam_starsi	Priemer
13	ES	-10,2	-13,7	-9,3	-11,1	-5,2	-6,8	-5,9	-6,0
14	CZ	-5,3	-3,7	-7,7	-5,6	-4,7	-3,2	-4,8	-4,2
15	AT	-1,1	1,6	-19,7	-6,4	0,2	3,5	-14,5	-3,6
16	LT	-8,9	-1,6	-5,3	-5,3	-6,4	1,0	-0,4	-1,9
17	DE	-5,0	-1,1	-10,1	-5,4	-2,8	1,5	-1,6	-1,0
18	PT	-1,9	1,4	1,6	0,4	-2,1	2,0	0,1	0,0
19	IE	-4,5	-4,3	-1,0	-3,3	-1,4	-0,7	3,1	0,3
20	LV	-8,2	-2,1	-5,9	-5,4	-3,7	2,4	3,3	0,7
21	CY	-0,8	0,4	0,4	0,0	-0,4	0,3	3,6	1,2
22	NL	3,6	6,0	-5,7	1,3	4,3	7,7	-2,3	-3,2
23	FI	-2,3	5,7	-0,4	-1,0	-0,7	7,3	4,5	3,7
24	EE	-7,1	-1,0	2,3	-1,9	-1,9	5,3	8,5	4,0
25	UK	1,5	5,3	5,4	4,1	1,5	5,8	7,4	4,9
26	DK	5,1	10,5	10,2	8,6	7,4	13,4	10,7	10,5
27	SE	2,9	11,5	18,6	11,0	3,1	10,7	19,6	11,1

5. Záver

Uvedený článok je zameraný na aplikáciu metód viacrozmernej štatistickej analýzy pri hodnotení členských krajín EÚ na základe ukazovateľov Lisabonskej stratégie, konkrétne išlo o analýzu zamestnanosti. Na základe výsledkov viacrozmerných metód sa javia ako najlepšie v tejto oblasti 2 krajiny, a to Švédsko a Dánsko. Hodnoty ich integrálnych ukazovateľov sú výrazne najvyššie. Na opačnej strane rebríčka sa nachádzajú krajiny ako Malta, Poľsko, Taliansko, Maďarsko, Slovensko, ale aj Belgicko.

Dnes je zrejmé, že pomalý pokrok EÚ má na svedomí sama Európska únia. Kľúčovú úlohu zohráva hlavne nedostatok rozhodného politického tlaku. Lisabonská stratégia je dnes ešte naliehavejšia, pretože dochádza k prehľbovaniu rozdielov v raste medzi Severnou Amerikou, Áziou a Európou, pričom sa Európa musí vyrovnávať s kombináciou problémov nízkeho rastu populácie a jej starnutím.

Literatúra

- [1] BARTOŠOVÁ, JITKA: Základy statistiky pro manažery. Nakladatelství VŠE v Prahe Oeconomica, Praha 2006. ISBN 80-245-1019-7.
- [2] EUROPEAN COMMISSION: Facing the challenge. The Lisbon strategy from growth and employment. Luxembourg: Office for Official Publications of the EC. November 2004. ISBN 92-894-7054-2. http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf
- [3] HEBÁK, PETR A KOLEKTÍV: Vícerozmerné statistické metody (3). Informatorium. Praha 2005. ISBN 80-7333-039-3.
- [4] HÚSEK, DUŠAN, ŘEZÁNKOVÁ, HANA, SNÁŠEL, VÁCLAV: Shluková analýza dat. Praha, Professional Publishing, 2007. ISBN 978-80-86946-26-9.
- [5] SHARMA, SUBHASH: Applied Multivariate Techniques. New York, John Wiley & Sons, Inc., 1996.
- [6] STANKOVIČOVÁ, IVETA: Viacrozmerná analýza rentability poisťovní SR pomocou Enterprise Guide. In: 10. medzinárodný seminár Výpočtová štatistika. – Bratislava: SŠDS 2001. – ISBN 80-88946-14-X.
- [7] VOJTKOVÁ, MÁRIA: Hľadanie podobnosti krajín EÚ podľa vybraných ukazovateľov Lisabonskej stratégie, In: Zborník z medzinárod-

- nej konferencie STAKAN 2007. Česká statistická společnost 2007.
<http://www.statspol.cz/stakan/index.htm>
- [8] VOJTKOVÁ, MÁRIA: Viackriteriálne hodnotenie podnikov priemyslu Slovenskej republiky. Roč. 2003, č. 3, s. 320-331, Ekonomické rozhľady. ISSN 0323-262X.
- [9] EUROPEAN COMMISSION: Facing the challenge. The Lisbon strategy from growth and employment. Luxembourg: Office for Official Publications of the EC. November 2004. ISBN 92-894-7054-2.
http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf
- [10] EUROSTAT HOME PAGE:
http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1090,30070682,1090_30298591&_dad=portal&_schema=PORTAL

Adresa:

Ing. Iveta Stankovičová, PhD.
Katedra informačných systémov
Fakulta managementu UK v Bratislave
Odbojárov 10, P. O. Box 95
820 05 Bratislava 25

E-mail: iveta.stankovicova@fm.uniba.sk

Poděkování: Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/4586/07 „Modelovanie sociálnej situácie obyvateľstva a domácností v Slovenskej republike a jej regionálne a medzinárodné porovnania“.

CLASSIFICATION TREES IN SOFTWARE RELIABILITY

Jan A. Strouhal, CQR

Key words: Classification tree, software reliability.

1. Introduction

Identifying problematic software modules early in the development lifecycle is a difficult task. According to the "80:20 rule", approximately 20 percent of software *modules* are responsible for 80 percent of its errors, costs, and rework. Software modules are parts (components) of a software system and developers save their resources when they concentrate their effort on error-prone modules only, not on whole software systems. Focusing developers on this fact is intended to help improve quality efficiency. Metric based classification trees take advantage of this rule and provide an empirically guided approach for identifying various classes of high-risk software components throughout the software lifecycle. Classification trees guide developers by focusing the application of specialised software analysis, testing, construction techniques and tools.

2. Classification Tree Construction

2.1. Differentiation to classes

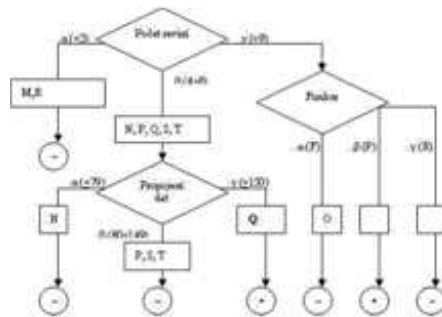
Firstly and most important in classification tree construction is the differentiation of classes. The basic classification range is binary only, that means there are only two different classes of results. There are positive classes and negative classes. The meaning of the positive class is that the module is prone to errors (error-prone), the negative class is not. It is recommended to be careful; this method does not exactly predict the number of errors in a module. This is a basic classification range of results. We can extend the basic classification range with different categories, e.g. five range classification:

- (++) double positive
- (+) positive
- (o) neutral
- (-) negative
- (--) double negative

The modules for tree construction are all tested, developed and the tree helps in making a decision about which class it belongs to, for example by error importancy for the development process.

2.2. Leafs

Leaf is the ending of a classification tree. There are no metrics left to do the next splitting, or it is not necessary. When all classes in node are the same, the node is *homogenous leaf*. If the node has no classes (no modules) at the time of the tree construction it results in *blind leaf*. Blind leaf appears, when there are more splitting points than are necessary. At the time of a module classification it can happen, that a module drops down into this blind leaf and then it's class cannot be decided. That is the reason why the method of floating and splitting a point is used.



2.3. Software Metrics

Secondly, but not less important, we must define software *metrics* for the classification of tree node splitting. Software metrics are simply measures of software components. It could be an abstract (subjective – i.e. "Is user friendly"), but it is strongly recommended to use any simple measurable criteria (i.e. "Number of code lines, type of module, etc."). It is hard to put a best list of metrics; it depends on the basic properties of the software modules in (particular) concrete cases.

2.4. Splitting Points

Thirdly, for the selected metrics, it is necessary to define suitable splitting points for node splitting. Basically, binary splitting is used where we need

only one fixed point (fixed by an expert). For example take 2897 module code lines. Modules which have less code lines are separated to the left branch of splitting, equal or more goes to the right branch. We extend this construction by accepting splitting to more branches using more splitting points. Another extension, we involve, consists in using flexible splitting points. Looking for optimal flexible metrics splitting points is a difficult problem. From the number of splitting points and their values we are able to know the metric selection in all tree nodes.

2.5. Metric Selection Function

Next there must be a defined way, by which we can decide, which metric is best for use in node splitting – metric selection function. In our case, the metric selection function is based on an evaluation, function F , to measure the homogeneity of a single subset. For binary splitting a metric selection function is defined as:

$$F(p_i; n_i) = -\frac{p_i}{p_i + n_i} \log_2 \frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i} \log_2 \frac{n_i}{p_i + n_i}$$

Where p is the amount of positive modules, n is the amount of negative modules. All is in i -th split. F is in $[0, 1]$ interval. It is also necessary to use weight $weight_i = \frac{p_i + n_i}{|C|}$ Metric A , where is $E(C, A) = \sum_{i=1}^v [weight_i * F(p_i + n_i)]$

A minimal is used as a new root for the next splitting. All this must be calculated simultaneously for all metric and only one metric, A , is used for one final node.

2.6. Metric Selection Function for Five Range Classification

For a five range splitting a metric selection function is defined as:

$$\begin{aligned} F(a_i; b_i; c_i; d_i; e_i) = \\ = -\frac{a_i}{s_i} \log_5 \frac{a_i}{s_i} - \frac{b_i}{s_i} \log_5 \frac{b_i}{s_i} - \frac{c_i}{s_i} \log_5 \frac{c_i}{s_i} - \frac{d_i}{s_i} \log_5 \frac{d_i}{s_i} - \frac{e_i}{s_i} \log_5 \frac{e_i}{s_i}, \end{aligned}$$

2.7. Termination Criteria

Lastly, we need some termination criteria. This is a condition, under which we shall stop tree splitting and declare a set of modules as the terminal leaf.

For example, termination criteria: "class is homogenous", this means that all modules have double plus class. Not necessarily only one, termination criteria.

A five range classification usually cannot be achieved with a homogenous leaf, so it can be defined with a more mild criteria, i.e. half of a leaf module is in one class.

3. Conclusion

The construction of the tree can be fully automated then. The algorithm will go through all the metrics and all splitting points looking for the optimals until the termination criteria is fulfilled.

References

- [1] Strouhal, J. A.: *Klasifikační stromy ve spolehlivosti software (in Czech)*. Sborník příspěvků konference ROBUST 2006, JČMF Praha, 2006.

VOLUNTARY UNIVERSITY COURSE: COMPUTERISED DATA PROCESSING

Pavel Stříž

Abstract: The article briefly introduces a new concept of Computerised Data Processing course at Tomas Bata University in Zlín.

Key words: Voluntary university course, basics of statistics, visualisation, programming, teaching the mathematical and statistical software.

Abstrakt: Článek informuje o plánech a ideích povinně volitelného předmětu Počítačové zpracování dat, který je vyučován na Univerzitě Tomáše Bati ve Zlíně. Příspěvek je přetištěn se souhlasem ediční rady FSS z čísla 5/2007.

Klíčová slova: Volitelný univerzitní předmět, základy statistiky, vizualizace, programování, výuka programů matematických a statistických.

1. Starting Position

We hope that you will find this article inspiring, because animations saved in Flash files *via screen recorders* (.swf) give us a tool to improve our classes. Further, students have self-study materials which *are not limited to* the class time.

We think that lecturers of such a new subject, or similar one, should teach students *following* successful passes in two or three semesters of mathematics, two semesters of probability and mathematical statistics.

At the Faculty of Management and Economics, we call these preceding subjects Mathematics I, II and III, Statistical Analysis Methods (Statistics A or 1) and Applied Statistics (Statistics B or 2).

2. Course Characteristics

Course type: Voluntary.

Time: 1 hour of lectures and 2 hours of seminars per week plus self-study.

Number of credits: 3; 3 hours a week in PC laboratory.

Semester: Summer / once a year.

Preceding subjects: None.

Best way: 3 semesters of mathematics and 2 semesters of statistics.

Parallel teaching in Czech: daily and combined students; RIUS project.

Parallel teaching in English: ERASMUS and EVENE projects.

The subject focuses on general mathematical and statistical software products, visualisation, and basics of programming. We try to use software under GNU licences or trial versions of commercial products available from the official websites for students, or any other person, for study purposes.

First we need *either* a previously bought university licence *or prior* written or *email permission* from the author(s) to use *the trial version*. That is the beginning of the development in our course outline.

3. Course Outline

We recommend to the students a lot of study materials in English, starting with summary and formula cards. Please, check this Formula Card out¹. We try to support classes with colourful graphs and animations to catch students' interest. We carefully choose problems, applied methods, assessments and examination topics from the content of all presented Flash animations and other study materials.

The outline should look like, sorted in semester weeks:

1. Downloading data: FTP, Web, DC++, eMule, BitTorrent, iTunes.
Web server package AEOnServ, FTP server FileZilla, RealVNC.
XLStatistics and XLMathematics refreshments.
Typesetting: T_EX, Microsoft Word, MathType and OO.org Writer.
2. EuroStat: European data – ESDS.
Electronic forms: AEOnServ via PHP and OO.org Writer.
Image, sound and video processing: Gimp2, TS-Midi, VirtualDub.
Data processing: VBA in Microsoft Word and Microsoft Excel.
3. Simulation: partly MuPAD, Extend 5LT and ARENA².
4. Minitab, Statgraphics, JMP, Origin.
5. Exact procedures in StatXact³. Artificial neural networks: SNNS.
6. List of difficult, tricky and unsolvable problems from real life.
7. First exam: General knowledge of working with statistical software; editing and building own ARENA models.
8. Statistica 1: ANOVA, nonparametric tests.
9. Statistica 2: Regression and correlation analysis, time series analysis.
10. Extend, Evolver, Mathematica, Maple.

¹bcs.wiley.com/he-bcs/Books?action=index&bcsId=2850&itemId=0471755303

²<http://www.arenasimulation.com/>

³<http://www.cytel.com/>

11. MATLAB plus complex example on taking a vacation decision.
12. GIS: Grass, OpenDX. Visualisation tools: ViSta, VisiCube.
 $\mathbb{R}$ +PHP, $\mathbb{R}$ +Rpad, graphic tools in $\mathbb{R}^4$.
 Quantian:⁵ Maxima, Scilab, Octave, Mayavi, GnuPlot, Ggobi, ...
 $\text{\TeX}$ and three related wysiwyg editors: $\text{\TeX}$ macs, LyX and kile.
13. Expedition to planetarium: selected lecture and seeing the night sky.
 Selected lecture series⁶.
14. Final *self-evaluation* exam on time series analysis:
 Decomposition, ARIMA, Fourier analysis, ANN, technical analysis.

4. Future Plans

We will animate and voice comment software products using *Wink*⁷ and *InstantDemo*⁸. It will give us a new perspective on teaching mathematics, statistics and programming. What to teach is a weekly decision depending on students' interests and the teacher's time.

The advantage is that we may use these files again next year.

Initial experimental server without animations:

<<http://study.uis.fame.utb.cz/courses/>>

Forthcoming EXPerimental server with animations in Flash files:

{Running at full capacity approximately from September 2008.}

<<http://exp.uis.fame.utb.cz/>>, username: student, password: exp.

5. A Final Note

Flash examples in Czech may be found under [01] to [12f].⁹ An inspiring example of the whole course in English may be found under Video Lectures.¹⁰ The best site known to the author is Total Training.¹¹

Address: Pavel Stríž, Faculty of Management and Economics, Tomas Bata University in Zlín, Mostní 5139, 760 01 Zlín, The Czech Republic

E-mail: striz@fame.utb.cz

⁴<<http://bg9.imslab.co.jp/Rhelp/servlet/getImage>>

⁵<<http://dirk.eddelbuettel.com/quantian.html>>

⁶<<http://www.avc-cvut.cz/>>

⁷<<http://www.debugmode.com/wink/>>

⁸<<http://www.instant-demo.com/>>

⁹<<http://study.uis.fame.utb.cz/courses/pdar/>>

¹⁰<<http://www.math.tamu.edu/~mpilant/math696/>>

¹¹<<http://www.totaltraining.com/>>

GRAFY A TABULKY VE STATISTICE (ANEBO NA CO VE VÝUCE OBVYKLE NENÍ ČAS)

Josef Tvrdlík

Adresa: Katedra informatiky, Přírodovědecká fakulta Ostravské university

Abstrakt: V článku jsou uvedeny některé jednoduché zásady a doporučení pro vhodnou prezentaci statistických výsledků, zejména tabulek a grafů. Tyto zásady a doporučení vycházejí z literatury a ze zkušeností z aplikací statistiky v různých oborech. Některé chyby v prezentaci výsledků jsou podrobně diskutovány a je také doporučeno vhodnější řešení.

Klíčová slova: Aplikace statistiky, prezentace výsledků, tabulky, statistické grafy.

Abstract:

Graphs and Tables in Statistics

Josef Tvrdlík

Department of Computer Science, University of Ostrava, CZ

This paper deals with some simple rules for proper presentation of statistical results, mainly for graphs and tables. These rules are based on the recommendations published in literature and on the experience from application of statistical data analysis in various fields of research. Some mistakes in presentation are discussed in detail and better form of presentation is recommended.

Key words:

Applied statistics, presentation of results, tables, statistical graphs.

Úvod

Zkušenost s pracemi studentů, články a zprávami výzkumníků různých oborů i s některými publikacemi ukazují, že prezentace výsledků statistických analýz je často zbytečně nepřehledná a obtížně čitelná. Mnohdy jako by autoři odmítali užívat zdravý rozum. Při prezentaci výsledků vůbec nemyslí na čtenáře a na to, aby jejich výsledky mohl někdo pokud možno snadno a správně vnímat a jejich badatelské úsilí také docenit. Právě zdravý rozum a ohled na čtenáře jsou důležité zásady, které by měly pomoci k přehledné prezentaci

výsledků, ale zřejmě takto vágní a různými lidmi různě vnímané doporučení nestačí. Přesnější formulace univerzálních zásad přehledné prezentace výsledků však není patrně možná.

Řadu dílčích doporučení, která mohou být užitečná, uvádí van Belle [5], kde zásadám správné prezentace statistických výsledků je věnována kapitola 7, nazvaná „Words, Tables, and Graphs“. Jako ilustraci toho, že na formě prezentace výsledků záleží, je na začátku této kapitoly uveden následující příklad tří způsobů prezentace téhož jednoduchého výsledku:

- The blood type in the population of the United States is approximately 40%, 11%, 4% and 45% A, B, AB, and O, respectively.
- The blood type in the population of the United States is approximately 40% A, 11% B, 4% AB and 45% O.

The blood type in the population of the United States is approximately,

O	45%
A	40%
B	11%
AB	4%

Rozdíly ve snadnosti či obtížnosti vnímání tohoto jednoduchého výsledku nepotřebují žádné další vysvětlování a snad jsou dostatečným argumentem pro to, že na způsobu prezentace výsledků záleží a že bychom se nad tím měli důkladně zamýšlet. V následujících třech odstavcích jsou uvedeny příklady a doporučení pro prezentaci tabulek a grafů a také pro interpretaci výsledků podle [5].

Prezentace tabulek

Některé chyby ukazuje tabulka 1, ve které jsou uvedeny počty pracovníků v různých zdravotnických profesích v USA roku 1988, názvy kategorií jsou ponechány v angličtině. Tabulka je nedokonalá nejméně ve dvou ohledech:

- Číselné údaje jsou téměř jistě zatíženy různou nepřesností. Zatímco u lékařů, sester, dentistů a optiků to jsou hodnoty získané z příslušných registrů, u některých jiných kategorií jako řečových, fyzických a pracovních terapeutů nebo pedikérů (podiatrists) jde jen o odhad v tisících., údaje v tabulce však vyvolávají dojem, že všechna čísla jsou přesná,
- van Belle jako chybu uvádí i to, že řádky tabulky jsou seřazeny podle abecedního pořadí názvů profesí, ne podle číselných hodnot. Možná se nám tato výhrada zdá neoprávněná, jsme asi zkaženi návyky jak z místních publikací, tak i většinou statistického softwaru, kde je četnostní

Tabulka 1: Počet aktivních zdravotníků v USA v roce 1980 (ze zprávy National Center for Health Statistics, 2000)

Occupation	1980
Chiropractors	25 600
Dentists	121 240
Nutritionists/Dieticians	32 000
Nurses, registered	1 272 900
Occupational Therapists	25 000
Optometrists	22 330
Pharmacists	142 780
Physical Therapists	50 000
Physicians	427 122
Podiatrists	7 000
Speech Therapists	50 000

Tabulka 2: Údaje z tabulky 1 seřazené podle počtu, zaokrouhleno na tisíce.

Occupation in 1000's	1980
Nurses, registered	1 273
Physicians	427
Pharmacists	143
Dentists	121
Physical Therapists	50
Speech Therapists	50
Nutritionists/Dieticians	32
Chiropractors	26
Occupational Therapists	25
Optometrists	22
Podiatrists	7

tabulka seřazena podle názvů kategorií nebo jejich číselných kódů. Ale argument, že pořadí řádků by nemělo záviset na tom, v jakém jazyku publikujeme, nelze jen tak vyvrátit.

Podle [5] by tabulka měla mít formu uvedenou v tabulce 2, tj. číselné údaje zaokrouhlené na tisíce a řádky seřazeny sestupně podle číselných hodnot.

Van Belle dále doporučuje užívat rozumný počet významných číslic a zejména v tabulkách brát ohled na tzv. „efektivní číslice“. To jsou ty číslice, jejichž hodnoty nejsou konstantní, ale mění se. Např. šestimístná čísla 354 691, 357 234, 356 991 mají jen čtyři efektivní číslice. Pokud bychom chtěli je prezentovat přijatelněji, pak bychom měli odečíst od těchto hodnot 350 000 a uvádět tento výsledný rozdíl. V tabulkách ovšem mají být pokud možno nejvýše dvě efektivní číslice, neboť tři a více efektivních číslic člověk obtížně vnímá.

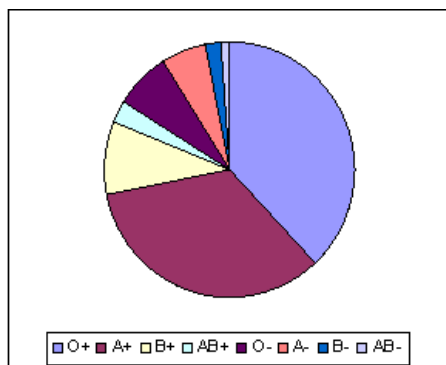
Grafy nebo tabulky?

Všeobecně hlášaná zásada, že grafy místo číselných údajů jsou lepší, není vždy správná. Někdy je tabulka vhodnější než graf, zejména když zvolený typ grafu neodpovídá struktuře dat. Jedním z doporučení [5] je **neužívat výsečové grafy**. Van Belle odkazuje citát: „*Jediná věc je horší než výsečový graf – několik nebo dokonce mnoho výsečových grafů*“.

Výsečové (koláčové) grafy ignorují strukturu dat, čtenář si musí propojovat legendu s výsečemi. Další van Bellův argument proti výsečovým grafům působí na první pohled úsměvně – při tisku výsečových grafů se spotřebuje moc inkoustu. Ale pokud se nad tím zamyslíme, je oprávněný. Porovnáme-li spotřebu inkoustu na bodový graf závislosti hodnot dvou veličin, kdy při malé spotřebě inkoust získáme náhled na tuto závislost se spotřebou na výsečové grafy, kdy při velké spotřebě nezískáme nic (viz příklad, obr. 1), pak závažnost argumentu musíme uznat.

Z výsečového grafu na obr. 1 se opravdu mnoho nedozvíme, struktura grafu neodpovídá struktuře dat, propojování legendy a výsečí je zbytečně namáhavé a spotřeba inkoustu velká. Tabulka 3 prezentuje stejný výsledek daleko přehledněji a srozumitelněji.

V diskusi na konferenci Stakan připomněl Jaromír Běláček, že u číselných údajů v procentech by znak % měl být uveden za každou číselnou hodnotou. Domnívám se, že tato zásada nemá univerzální platnost a autor by měl vždy zvážit, kdy znak % srozumitelnosti pomáhá a kdy je nadbytečný. Ale procenta, ať se znakem % či bez, jsou čitelnější než relativní četnosti jako desetinná čísla z intervalu [0,1]. U takto uváděných relativních četností přibude neefektivní číslice 0, neefektivní je i znak desetinné tečky nebo čárky a vytratil



Obrázek 1: Relativní četnosti (v %) krevních skupin a Rh faktoru v populaci USA

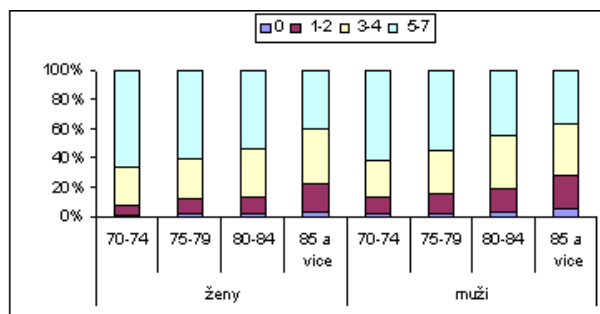
Tabulka 3: Relativní četnosti (v %) krevních skupin a Rh faktoru v populaci USA

Blood Type	Rh+	Rh-	Total
O	38	7	45
A	34	6	40
B	9	2	11
AB	3	1	4
Total	84	16	100

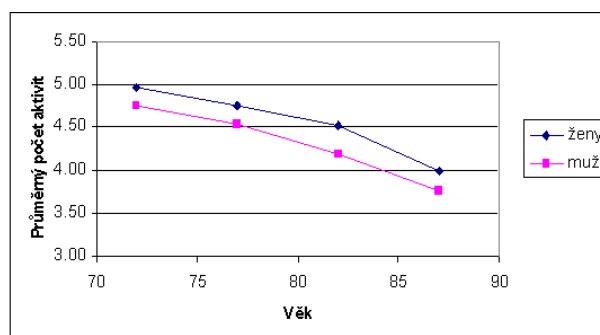
se grafická informace o počtu efektivních číslí ve sloupcích a řádcích, čímž se zhorší přehlednost a srozumitelnost, jak se můžeme přesvědčit porovnáním tabulek 3 a 4.

Další zásada podle [5] je **neužívat sloupcové skládané sloupcové grafy**. Skládané (kumulované, stackbar) sloupcové grafy jsou hůře čitelné než jednoduché sloupcové grafy a často lze najít efektivnější možnost, jak nahlédnout do struktury dat. Van Belle to ilustruje na následujícím příkladu.

Souhrnná zdrojová data z průzkumu počtu aktivit provozovaných seniory v průběhu dvou týdnů jsou uvedena v tabulce 5. Ve zprávě Státního centra pro zdravotní statistiku byly tyto údaje prezentovány formou skládaného sloupcového grafu (obr. 2), což ke vnímání jejich obsahu nijak nepřispělo, spíše naopak. Prezentace by měla usnadňovat odpovědi na následující jedno-



Obrázek 2: Počet aktivit v průběhu dvou týdnů – četnosti v % (Kramarov et al., zpráva National Center for Health Statistics, 1999).



Obrázek 3: Průměrný počet aktivit podle věku a pohlaví

Tabulka 4: Relativní četnosti krevních skupin a Rh faktoru v populaci USA

Blood Type	Rh+	Rh-	Total
O	0,38	0,07	0,45
A	0,34	0,06	0,40
B	0,09	0,02	0,11
AB	0,03	0,01	0,04
Total	0,84	0,16	1,00

duché a přirozené otázky: Mají více aktivit muži nebo ženy? Jak mění počet aktivit s věkem? Liší se tyto změny u mužů a žen?. To ovšem spojovaný sloupcový graf rozhodně neusnadňuje.

Tabulka 5: Počet aktivit seniorů v průběhu dvou týdnů – četnosti v %

	Počet aktivit	70-74	75-79	80-84	85 a více
Ženy	0	1	1.3	2.1	3.1
	1-2	6.8	10.5	11.9	19.2
	3-4	26.8	27.5	32.5	38.3
	5-7	65.4	60.7	53.5	39.4
Muži	0	1.9	1.7	2.9	5.3
	1-2	10.5	13.3	15.9	23
	3-4	26.3	30.3	36.7	35.9
	5-7	61.2	54.7	44.5	35.9

Přitom docela jednoduchý přepočet a grafické zobrazení průměrných hodnot aktivit pro muže a ženy podle věkových kategorií (obr. 3) vypovídá, že ženy jsou o trochu aktivnější, počet aktivit s věkem klesá a rychlost tohoto poklesu je u obou pohlaví zhruba stejná.

Opatrně při interpretaci výsledků statistické analýzy

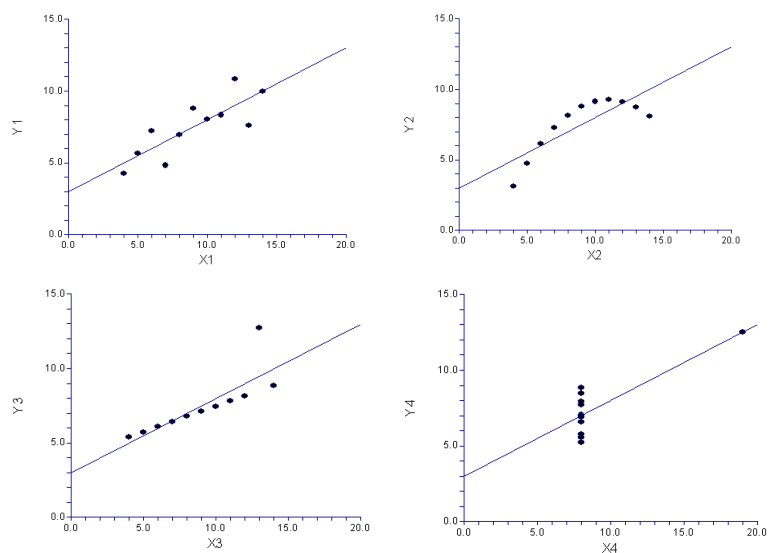
To je další doporučení, kterému se van Belle věnuje dosti podrobně. Neměli bychom přeceňovat (zdánlivě) podrobné numerické výstupy a při interpretaci se držet dobré zásady „Důvěřuj, ale prověřuj“. Ošidnost interpretace „věrohodného“ numerického výstupu je ilustrována známým příkladem (Anscombe [1]), který je užit i jako ilustrace na titulní straně knihy [5]. Číselné výsledky

Tabulka 6: Popisné statistiky veličin

Variable	n	průměr	sm. odch
x	11	9,00	3,32
Y	11	7,50	2,03

Tabulka 7: Výsledky lineární regrese –
závislost Y na x

	b(i)	s(bi)	T	p
Intercept	3,00	1,125	2,67	0,026
x	0,50	0,118	4,24	0,002
Rovnice regresní přímky:	$Y = 3,00 + 0,50x$			
Index determinace (R^2):	0,67			
Koeficient korelace:	0,82			
Směrodatná odchylka reziduí:	1,24			



Obrázek 4: Závislosti Y na x k příkladu z Anscombe [1]

analýzy závislosti dvou veličin z tohoto příkladu jsou uvedeny v následujících tabulkách.

Na obr. 4 jsou grafy čtyř závislostí veličiny Y na x . Patří výsledky z tabulek 6 a 7 k některé závislosti z tohoto obrázku? Jen k té první? Možná vás překvapí správná odpověď, že ke všem čtyřem. Tento příklad je dosti populární, data a různé komentáře lze najít i na webu, např. [2,3]. Zde ke správné interpretaci výsledků pomůže jednoduchý graf. Ale ne vždy lze nesprávnou interpretaci podrobných numerických výsledků odhalit tak snadno, viz např. nějaký složitější zobecněný lineární model ap.

Jak vyjít vstříc čtenáři?

Na to nelze dát všeobecné doporučení, ale některé možnosti naznačí následující příklad. Je to výsek výsledků z článku [4], kde jsou porovnávány stochastické algoritmy pro odhady parametrů nelineárních regresních modelů. Jednou ze sledovaných veličin charakterizující časovou náročnost algoritmů je počet vyhodnocení účelové funkce, tj. součtu residuálních čtverců, potřebných k dosažení podmínky konvergence. V tabulce 8 jsou výsledky porovnání tří algoritmů ve formě, která se v prezentacích popisných statistik užívá často zcela automaticky, jako by to bylo dáno normou. Pro každý algoritmus a pro každou z osmi testovaných úloh je uveden vždy průměrný počet vyhodnocení funkce (ne) a směrodatná odchylka (s) počtu vyhodnocení.

Tabulka 8: Porovnání časové náročnosti tří algoritmů na osmi úlohách – průměry a směrodatné odchylky.

Úloha	Algoritmus1		Algoritmus2		Algoritmus3	
	ne	s	ne	s	ne	s
bennett5	41335	14054	47122	16493	36788	14347
boxbod	1308	92	1478	118	824	74
eckerle4	2629	105	2944	147	1709	103
mggh09	10422	834	11777	1060	8859	974
mggh10	20761	2076	24290	972	20969	1677
rat42	2942	177	3354	168	1912	115
rat43	4807	192	5432	217	2932	147
thurber	13915	417	15446	309	9741	292

Jelikož jde o porovnání časové náročnosti algoritmů (a variability této časové náročnosti) na několika úlohách, můžeme porovnávání čtenáři usnadnit, viz tabulka 9. Především můžeme místo směrodatné odchylky užít koeficient

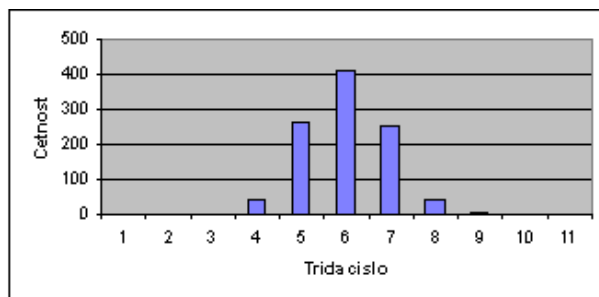
variace vyjádřený v procentech průměrné hodnoty (sloupce vc). Tím oproti tabulce 8 podstatně snížíme počet efektivních čísel charakterizujících variabilitu. Dále je možné průměrnou časovou náročnost druhého a třetího algoritmu uvádět jako relativní změnu (v procentech) vzhledem k algoritmu1. Potom budou záporné hodnoty znamenat zrychlení proti algoritmu1, kladné hodnoty naopak větší časovou náročnost. Navíc bezprostředně vidíme, zda je tato změna časové náročnosti věcně významná. Pokud by nás zajímala i statistická významnost, lze ji z údajů v tabulce vyhodnotit. Tabulky 8 a 9 jsou co do informačního obsahu shodné, jednu z druhé lze snadno přepočítat, ale tabulka 9 je pro porovnávání algoritmů výrazně pohodlnější.

Tabulka 9: Porovnání časové náročnosti tří algoritmů na osmi úlohách

Úloha	Algoritmus1		Algoritmus2		Algoritmus3	
	ne	vc	ne	vc	ne	vc
bennett5	41335	34	14	35	-11	39
boxbod	1308	7	13	8	-37	9
eckerle4	2629	4	12	5	-35	6
mggh09	10422	8	13	9	-15	11
mggh10	20761	10	17	4	1	8
rat42	2942	6	14	5	-35	6
rat43	4807	4	13	4	-39	5
thurber	13915	3	11	2	-30	3

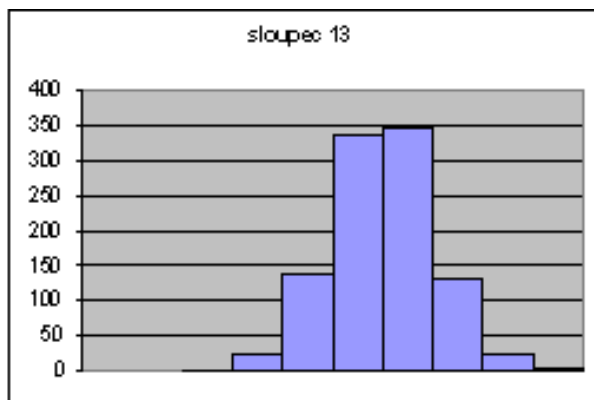
Příklady chyb v prezentaci výsledků

V tomto odstavci jsou komentovány chyby z korespondenčních úloh studentů v kombinované formě bakalářského studia aplikované informatiky v předmětu Analýza dat, který mají studenti ve třetím ročníku studia na Přírodovědecké fakultě Ostravské univerzity. Ukázky jsou z letního semestru roku 2007, komentáře jsou psány kurzivou.



Obrázek 5: Histogram – častá chyba z naprosté nedbalosti

Histogram na obr. 5 je prezentován tak, jak ho nabízí Excel, zdravý rozum si vybral dovolenou, ohled na čtenáře žádný. Ponechány mezery mezi sloupci, nevhodně zvolené měřítko vodorovné osy (pět tříd s nulovou četností), nic nevypovídající popis vodorovné osy.



Obrázek 6: Histogram – další častá chyba způsobená nedbalostí

V histogramu na obr. 6 chybí popis os, zbytečný je nic neříkající nadpis histogramu ,opět nevhodně zvolené měřítko vodorovné osy.

$H_0: \mu = 6$

průměr $x = 5,959409417$

$s = 0,99046792$

hodnota testového kritéria: -1,29593994

Typická ukázka nesprávného a nepřehledného prezentování číselných výsledků s nadbytečným počtem platných číslic.

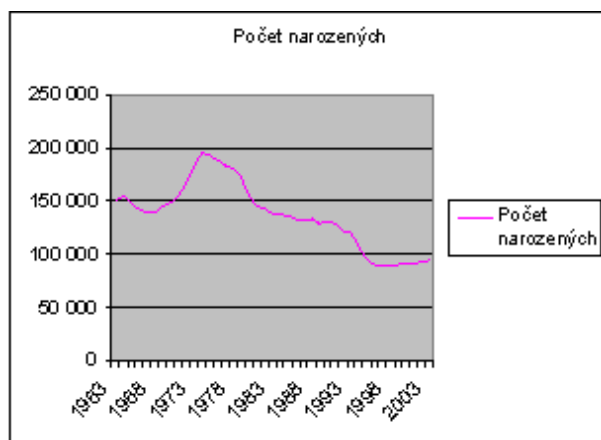
$$b_1 = 0,90711042$$

$$b_0 = 17,0189542$$

$$Se = \sum (Y_i - b_0 - b_1 x_i)^2 = 423,839904$$

$$s^2 = Se / (n - 2) = 26,489994$$

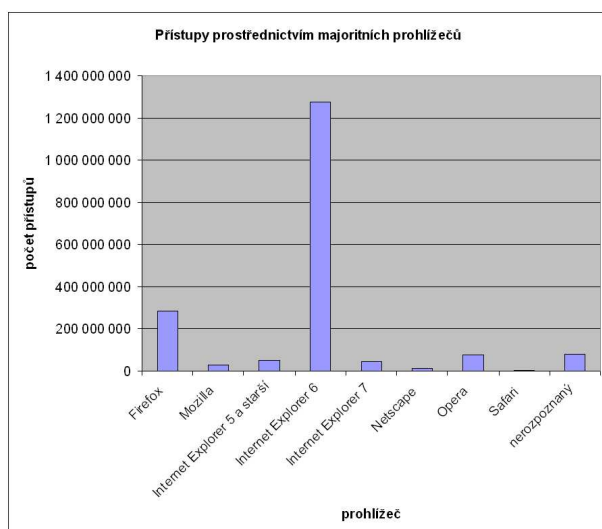
Podobné chyby jako v předchozí ukázce, tady navíc i neobratný a nepřesný zápis symbolů a vzorců.



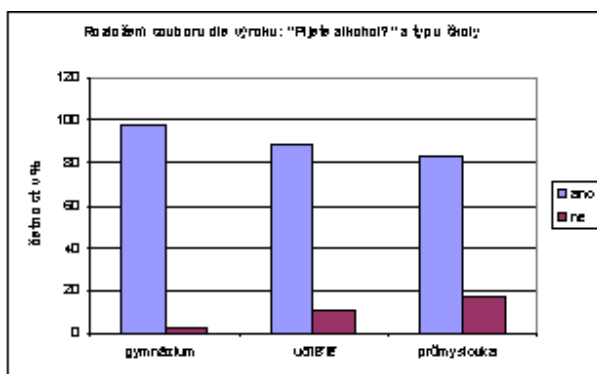
Obrázek 7: Časový průběh počtu narozených

Na obr. 7 chybí popis os grafu, nevhodné jednotky na svislé ose (tři neefektivní nuly, počet narozených měl být v tisících), legenda je nadbytečná a zbytečně zabírá značnou část kreslicí plochy, význam čáry nejasný (bylo užito nějaké vyhlazování?), časová řada by měla být nakreslena jako body, případně se spojnicemi.

Na obr. 8 jsou užity nevhodné jednotky na svislé ose sloupcového grafu (8 neefektivních číslic), vhodnější by bylo uvádět počet přístupů v milionech nebo lépe ve stovkách milionů. Zobrazení devíti značně odlišných četností formou sloupcového grafu není nejvhodnější způsob prezentace tohoto výsledku, tabulka by vypovídala o struktuře a obsahu dat lépe.

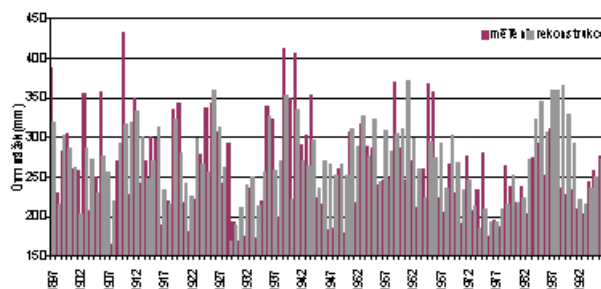


Obrázek 8: Nevhodný sloupkový graf



Obrázek 9: Další nesprávný sloupkový graf

Na první pohled (pomineme-li neobratnou formulaci nadpisu) sloupcový graf na obr. 9 vypadá uspokojivě. Ale jaký je význam druhých sloupečků? Jsou to doplňky do 100%, takže jsou nadbytečné stejně jako legenda. Tři zjištěné relativní četnosti stačilo uvést jako tabulku, zabralo by to méně místa a vypovídalo jasně.



Obrázek 10: Nevhodně užitý typ grafu

Na obr. 10 je nevhodně zvolený typ grafu pro zobrazení dvou časových řad do jednoho obrázku, takže výsledek je nepoužitelný pro naprostou nečitelnost. Pro takové závislosti jsou vhodné bodové grafy, případně se spojnicemi bodů.

Závěr

Van Belle [5] požaduje, aby se v prezentaci výsledků statistických analýz věda spojovala s uměním. Možná je to požadavek až příliš náročný, ale rozhodně bychom měli dbát alespoň na dobrou řemeslnou úroveň, využívat základní prezentační dovednosti, při prezentaci výsledků statistických analýz užívat zdravý rozum, přihlížet k možnostem vnímání čtenáře, mít ke čtenáři respekt a snažit se o co největší přehlednost a srozumitelnost výsledků. Otázkou je, jak to učit. Existuje nějaká metodika takové výuky? Může být taková metodika účinná? Lze dobré „mravy“ prezentace statistických výsledků někoho naučit ve studijním kurzu? Není pro dobrou prezentaci statistických výsledků nejdůležitější poctivá snaha a praxí a omyly získané zkušenosti? Tyto otázky bohužel zůstávají dosud nezodpovězeny.

Reference

- [1] Anscombe F.J. Graphs in statistical analysis. The American Statistician 27, 17-21, (1973).
- [2] Diggle P.J., www.maths.lancs.ac.uk/~diggle/gsse401/anscombe.dat, last modified October 30, 2001.
- [3] Scatterplots, Anscombe's datasets, <http://exploringdata.cqu.edu.au/anscomb1.htm>.
- [4] Tvrdík J., Křivý I., Mišík L., Adaptive population-based search: Application to estimation of nonlinear regression parameters, Computational Statistics & Data Analysis. 52(2), 713-724, 2007. Available online 9 November 2006, (<http://www.sciencedirect.com/science/article/B6V8V-4M9HWXT-3/2/e4fa1077aa80c154b130396b3c486286>).
- [5] van Belle G.: *Statistical Rules of Thumb*, John Wiley & Sons, 2002

Poděkování: Tato práce byla zčásti podporována grantem 201/05/0284 Grantové agentury České republiky.

VYUŽITIE ŠTATISTIKY V POISTNEJ MATEMATIKE

Marta Urbaníková

Abstrakt: Cieľom príspevku je stručne informovať o niektorých možnostiach aplikácie pravdepodobnosti a matematickej štatistiky v poistení osôb a poistení majetku.

1. Úvod

Teória pravdepodobnosti a matematickej štatistiky tvorí dôležitú zložku poistných vied. Vyplýva to zo samotnej podstaty poistenia. Poistenie predstavuje nástroj na elimináciu dôsledkov poistných udalostí, ktoré sú predmetom poistenia. Každá poistná udalosť má charakter náhodnej udalosti o ktorej nevieme dopredu povedať či nastane a kedy nastane. Niet pochýb, že základom poisťovníctva sú pravdepodobnostné zákonitosti výskytu týchto náhodných udalostí. Na poistenie sa možno pozeráť ako na ochranu proti rizikám. Poistený prenáša svoje riziká na poisťovňu, ktorá pri dostatočne veľkom súbore rizík podobného charakteru je schopná tieto riziká zvládať, nakoľko s rastom počtu uzavretých poistných zmlúv sa poistno-technické riziko znižuje.

2. Životné poistenie

Základným nástrojom využívaným v životnom poistení sú úmrtnostné tabuľky. Úmrtnostné tabuľky predstavujú model úmrtnosti. V jednotlivých stĺpcoch úmrtnostnej tabuľky sú nasledujúce údaje:

- x – vek osoby
 $x = 0, 1, 2, \dots, \omega$; ω je maximálny sledovaný vek v tabuľke
- l_x – počet osôb dožívajúcich sa veku x
počet jedincov z koreňa l_0 , ktorí sa dožijú veku x , $\{l_i\}_{i=0}^{\omega}$ je nerastúca postupnosť
- d_x – počet zomretých vo veku x
počet jedincov z koreňa, ktorý zomrú vo veku x
 $d_x = l_x - l_{x+1}$,
- q_x – pravdepodobnosť úmrtia vo veku x
pravdepodobnosť toho, že jedinec, ktorý je nažive vo veku x , zomrie

pred dosiahnutím veku $x + 1$.

$$q_x = \frac{l_x - l_{x+1}}{l_x},$$

- p_x – pravdepodobnosť dožitia veku $x + 1$
pravdepodobnosť toho, že jedinec, ktorý je nažive vo veku x , sa dožije veku $x + 1$

$$p_x = \frac{l_{x+1}}{l_x}, p_x + q_x = 1$$

- ${}_n p_x$ – pravdepodobnosť žitia vo vekovom intervale $(x, x + n)$
pravdepodobnosť toho, že jedinec, ktorý je nažive vo veku x sa dožije veku $x + n$

$${}_n p_x = \frac{l_{x+n}}{l_x} \quad q_x = 1 - {}_n p_x = \frac{l_x - l_{x+n}}{l_x}$$

Komutačné čísla sú pomocné hodnoty, ktoré vznikajú diskontovaním hodnôt z úmrtnostných tabuliek. Umožňujú ľahší a rýchlejší výpočet všetkých hodnôt používaných v poistení. Sú tabelované pri najčastejšie používaných úrokových mierach. Najčastejšie sa používa šesť komutačných čísel:

- D_x – diskontovaný počet dožívajúcich sa veku x

$$D_x = l_x \cdot v^x$$

pričom $v = (1 + i)^{-1}$.

- C_x – diskontovaný počet zomrelých vo veku x

$$C_x = d_x \cdot v^{x+1}$$

- N_x – je súčet D_{x+k} od veku x až po koniec úmrtnostnej tabuľky

$$N_x = \sum_{k=0}^{\omega-x} D_{x+k}$$

- S_x – je súčet N_{x+k} od veku x až po koniec úmrtnostnej tabuľky

$$S_x = \sum_{k=0}^{\omega-x} N_{x+k}$$

- M_x – je súčet C_{x+k} od veku x až po koniec úmrtnostnej tabuľky

$$M_x = \sum_{k=0}^{\omega-x} C_{x+k}$$

- R_x – je súčet M_{x+k} od veku x až po koniec úmrtnostnej tabuľky

$$R_x = \sum_{k=0}^{\omega-x} M_{x+k}$$

Výpočet poistného

Pri výpočte poistných sadziieb – brutto poistného pre jednotlivé produkty poisťovní postupuje tak, že najskôr vypočíta pre daný produkt netto poistné, ku ktorému potom pripočíta správne náklady poisťovne a bezpečnostnú prirážku.

Netto poistné je počítané tak, aby v priemere pokrylo poistné plnenie poisťovne. Všetky výpočty spojené s poistením osôb vychádzajú z dvoch základných princípov:

- *Princíp fiktívneho súboru* vychádza z predpokladu, že počet osôb uzatvárajúcich vo veku x určitý typ poistnej zmluvy sa rovná hodnote l_x z používanej úmrtnostnej tabuľky.
- *Princíp ekvivalencie* vyjadruje tú základnú požiadavku, že pri uzatváraní súboru poistných zmlúv rovnakého typu musia byť v rámci tohto súboru všetky príjmy poisťovne v rovnováhe s jej výdavkami, pričom sa príjmy a výdavky diskontujú k spoločnej časovej základni.

Tieto predpoklady neodzrkadľujú plne realitu, ale podstatne zjednodušujú všetky úvahy a vedú k správnym výsledkom.

Nakoľko poistné sa môže platiť buď jednorazovo pri podpísaní zmluvy, alebo v pravidelných splátkach, rozlišujeme

- jednorazové poistné pre jednotlivé produkty kapitálového životného poistenia,
- bežné poistné pre jednotlivé produkty kapitálového životného poistenia.

Netto poistné sa počíta vždy pre jednotkovú poistnú sumu, t. j. pre poistné plnenie vo výške 1 Sk.

Poisťovňa oceňuje svoje budúce príjmy (prijaté poistné) a výdaje (poistné plnenia) pomocou očakávaných počiatočných hodnôt. Na základe princípu ekvivalencie platí

$$\begin{array}{ccc} \textit{očakávaná počiatočná} & = & \textit{očakávaná počiatočná hodnota} \\ \textit{hodnota poistného} & & \textit{poistného plnenia} \end{array}$$

Ako príklad uvedieme výpočet jednorazového netto poistného pre poistenie pre prípad dožitia.

Predpokladajme, že osoba vo veku x uzavrie tento typ poistenia na dobu n rokov. Poisťovňa je povinná vyplatiť poistnú sumu, ak sa poistený dožije konca dojednanej poistnej doby, t. j. veku $x + n$. Ak poistený zomrie pred uplynutím poistnej doby, poistenie zaniká bez náhrady. Nech ${}_nE_x$ je suma, ktorú poistník musí zaplatiť, aby mu na konci poistnej doby poisťovňa vyplátila 1 Sk. Podľa princípu ekvivalencie, kde príjmy poisťovne sa musia rovnať výdajom platí:

$${}_nE_x \cdot l_x \cdot v^x = l_{x+n} \cdot v^{x+n}$$

t. j. hodnota poistného vynásobená počtom poistených v čase x sa má rovnať súčtu všetkých poistných plnení vyplatených poisťovňou osobám, ktoré žijú v čase $x + n$, pričom obidve strany sú diskontované k tomu istému okamihu, v tomto prípade k okamihu narodenia jedincov sledovaného súboru. Potom platí:

$${}_nE_x = \frac{D_{x+n}}{D_x}$$

Výpočet poistného sa dá formulovať aj ako výpočet strednej hodnoty vhodne zvolenej náhodnej premennej. Nech náhodná premenná Z je definovaná nasledovne:

$$Z = \begin{cases} v^n & \text{s pravdepodobnosťou } {}_np_x \\ 0 & \text{s pravdepodobnosťou } {}_nq_x \end{cases}$$

$${}_nE_x = E(Z) = {}_np_x \cdot v^n + {}_nq_x \cdot 0 = {}_np_x \cdot v^n = v^n \cdot \frac{l_{x+n}}{l_x} = \frac{D_{x+n}}{D_x}$$

Skutočne realizované hodnoty sa môžu výrazne líšiť od strednej hodnoty. Treba uvažovať s poistno-technickým rizikom poisťovne. Pre ocenenie takéhoto rizika by sa okrem výpočtu stredných hodnôt mali skúmať aj ich pravdepodobnostné rozdelenia. V praxi sa však prevažne počítajú iba smerodajné odchýlky.

Presnosť stanovenia poistného, t. j. smerodajná odchýlka náhodnej premennej Z :

$$\sqrt{\text{var}(Z)} = \sqrt{{}_nE_x^2 - ({}_nE_x)^2} = \sqrt{v^{2n} \cdot {}_np_x - {}_np_x^2}$$

3. Neživotné poistenie

Na rozdiel od poistenia osôb v neživotnom poistení býva výška škody a zodpovedajúce poistné plnenie zvyčajne menšie než poistná suma. Teda aj keď poisťovňa pozná rozsah prevzatých záväzkov, nevie aké budú jej výdavky. Tie môže odhadnúť zo získaných štatistických údajov. Zdrojmi údajov sú väčšinou vlastné dáta poisťovne z minulosti, zaistovateľa, štatistiky, alebo iné údaje z trhu. Najlepšie sú vlastné údaje poisťovne, ale tie pri novom type rizika často nie sú k dispozícii.

Pri stanovení poistného používajú poisťovne okrem škodových ukazovateľov aj škodové tabuľky alebo výlukový poriadok zo škodového stavu. *Škodové tabuľky* sú analógiou úmrtnostných tabuliek v životnom poistení. Používajú sa na určenie škodového stupňa, sú konštruované na základe skutočných dát pre hypotetický súbor škôd. Škodová tabuľka je zjednodušene povedané tabuľkou rozdelenia početnosti výšky škôd. *Výlukový poriadok zo škodového stavu* nahrádza škodovú tabuľku v tých poistných produktoch, kde výška škody závisí od doby trvania jej následkov. Napr. pri úrazovom zdravotnom poistení, kde poistné plnenie bude vyplácané po dobu nevyhnutného liečenia, pri poistení ušlej mzdy pri práceneschopnosti a iné.

Škodovú frekvenciu a škodový stupeň možno určiť na základe pravdepodobnostných rozdelení. Každá poistná udalosť je náhodnou udalosťou. Pravdepodobnostné rozdelenia sú vhodné na modelovanie počtu a výšky poistných plnení pri rôznych poistných produktoch.

Počet škôd, teda počet poistných plnení n má najčastejšie niektoré z nasledujúcich rozdelení:

- binomické,
- Poissonovo,
- negatívne binomické rozdelenie.

V prípade Poissonovho rozdelenia sa jeho parameter λ rovná škodovej frekvencii q_1 .

Výber vhodného rozdelenia počtu poistných plnení (škôd) možno uskutočniť na základe vzťahu medzi strednou hodnotou a disperziou náhodnej premennej takto:

$$\begin{aligned}
E(n) > D(n) &\rightarrow \text{binomické rozdelenie,} \\
E(n) = D(n) &\rightarrow \text{Poissonovo rozdelenie,} \\
E(n) < D(n) &\rightarrow \text{negatívne binomické rozdelenie.}
\end{aligned}$$

Na základe Centrálnej limitnej vety môžeme dané diskkrétne rozdelenia za splnenia istých podmienok aproximovať normálnym rozdelením.

Výška škody X je tiež náhodná premenná. Na opis výšky škôd možno použiť niektoré z nasledujúcich rozdelení:

- Lognormálne rozdelenie
Je vhodné na modelovanie výšky škôd v havarijnom poistení, v požiar-
nom poistení, v poistení proti víchriciam a v úrazovom poistení.
- Paretovo rozdelenie
Je vhodné na modelovanie škôd, ktoré môžu nadobúdať extrémne hod-
noty, napr. v nemocenskom poistení a v poistení proti požiarom.
- Gama rozdelenie
Je vhodné na modelovanie výšky škôd pri poistení motorových vozidiel.
- Beta rozdelenie
Používa sa na modelovanie výšky škôd v požiar-
nom poistení, kde často
nastávajú buď veľmi malé alebo veľmi veľké škody.

Všetky uvedené rozdelenie závisia od jedného, alebo viacerých paramet-
rov. V praxi poisťovní je treba tieto parametre rozdelení odhadnúť na základe
výberových údajov. Na odhad týchto parametrov možno využiť metódu ma-
ximálnej vierohodnosti. Takéto odhady majú vynikajúce asymptotické vlast-
nosti.

Ďalšia oblasť kde možno aplikovať teóriu pravdepodobnosti a matematic-
kej štatistiky je oblasť kolektívneho rizika.

Modely kolektívneho rizika

Riziko poisťovateľa spočíva v nebezpečenstve, že prijaté poistné nebude po-
stačovať na vyplatenie všetkých poistných plnení. Pri riešení týchto problé-
mov sa využívajú zložené rozdelenia celkových poistných plnení, založené na
kombinácii rozdelenia počtu aj výšky poistných plnení.

Teória krachu

Zaoberá modelmi kolektívneho rizika pre dlhšie časové obdobie. Vychádza
z analýzy stochastických modelov poistných rezerv pri neživotnom poistení.

Teória kredibility

Predstavuje súbor postupov a techník na výpočet poistného pri krátkodobých kontraktoch a na systematickú úpravu poistných sadzieb tak ako sú zaznamenávané nové údaje o škodovom priebehu. Kredibilné poistné sa určuje ako lineárna kombinácia poistného odhadnutého pomocou údajov z vlastného portfólia a poistného odhadnutého z cudzích, porovnateľných rizík. Základom týchto metód je moderná bayesovská štatistika.

Literatúra

- [1] Cipra, T.: *Pojistná matematika*. Ekopress, Praha 1999.
- [2] Cipra, T.: *Zajištění a přenos rizik v pojišťovnictví*. Grada 2004.
- [3] Markechová, D. – Tirpáková, A.: *O matematických metódach ako interdisciplinárnych metódach vo vedecko-výskumnej praxi*. Zborník z medzinárodného vedeckého sympózia „Kultúra – priestor interdisciplinárneho myslenia“ konaného v dňoch 21.–22. septembra 2004 na UKF v Nitre (2004), 92-97, ISBN 80-8050-837-2.
- [4] Pacáková, V.: *Aplikovaná poistná štatistika*. Elita, Bratislava 2000.
- [5] Pacáková, V. – Bohdalová, M.: *Simulácia kolektívneho rizika metódou Monte Carlo*. In: SAS Letná škola 2006, Bratislava, 7. 6. 2006, s. 1-16.

Adresa: RNDr. Marta Urbaníková, CSc.
UMIT FPV UKF v Nitre, Slovenská republika
E-mail: murbanikova@ukf.sks

ZKUŠENOSTI S VYUŽITÍM EXCELU PŘI VÝUCE APLIKOVANÉ STATISTIKY

Vladimíra Vlčková, Otakar Machač

Abstract: This paper deals with experiences in spreadsheet Excel utilization for teaching subject applied statistics on Faculty of Chemical Technology, University of Pardubice. The reasons which led authors to choose Excel are described together with summarization of problems which authors met during practical education with Excel. Authors see as major advantage its broad accessibility and its relative simplicity, easy import into text editor Word, possibility of project education, possibility to carry out simultaneously calculations and exploit many of Excel functions, mainly contingent tables and charts. As imperfection of Czech version authors see mainly inconsistent and very often even incorrect translation of some functions titles, their non systemic sorting and often incomprehensible help to particular functions.

Úvod

Nově koncipovaný předmět aplikovaná statistika byl akreditován na chemicko-technologické fakultě Univerzity Pardubice v roce 2003. Výuka probíhá od školního roku 2005/6, a to jak v prezenční tak i v kombinované formě ve druhém ročníku dvou studijních programů bakalářského studia: Chemie a technologie potravin, studijní obor Hodnocení a analýza potravin jako povinný předmět a Chemie a technická chemie ve všech studijních oborech jako povinně volitelný předmět. Od příštího roku bude v této podobě vyučován také ve studijním programu Polygrafie jako povinný předmět v 1. ročníku magisterského studia. Rozsah výuky je 2 hodiny přednášek a 2 hodiny semináře po dobu jednoho semestru. Tématicky na něj navazují v 1. ročníku magisterského studia podle typu studijního oboru předměty: ekonomická statistika, výpočetní technika v životním prostředí, chemometrie, statistika, a to buď ve formě povinného a nebo povinně volitelného předmětu. Pro některé studenty je to tak jediný předmět zabývající se statistikou, protože není mnoho studentů, kteří si vyberou některý z uvedených nadstavbových předmětů pokud je pouze volitelný. Alespoň základní znalosti statistiky však studenti potřebují při řešení bakalářské, případně diplomové práce, o další vědecké či manažerské praxi ani nemluvě.

Pro nás, jako garanty tohoto předmětu, nebylo proto jednoduché vytvořit koncepci takto široce pojatého předmětu s jednosemestrální dotací a rozhodnout se pro vhodný způsob počítačové podpory. Statistické softwary nemalou

mirou napomáhají ke změnám ve způsobu výuky statistiky. Dnes je zřejmá tendence přecházet od výkladu a také zkoušení důkazů a vzorců spíše k praktickým aplikacím, tedy skutečně k aplikované statistice. Specializované statistické programy (pakety) sice často přinášejí větší komfort při analýze dat a výpočtu požadovaných charakteristik, ale mají na druhé straně i své nevýhody. Statistických paketů je nabízeno velké množství, což také znamená, že nejsou všude a vždy dostupné, jsou na různé úrovni, nejsou vždy plně kompatibilní s Excelem, vyžadují mnohdy speciální školení a ty lepší jsou zpravidla neúnosně nákladné pro výuku. Zřejmou nevýhodou jejich používání ve výuce v bakalářských programech je, že studenti, kteří se statistikou teprve začínají, se nenaučí hledat souvislosti, přemýšlet nad daným problémem a formulovat předpoklady pro platnost závěrů z vybraných statistických charakteristik. V krajním případě může používání takovýchto specifických softwarů sklouznout až k pouhému naučenému sledu příkazů, s cílem získání určitých charakteristik bez znalosti jejich významu a předpokladů pro jejich použití. To se pak projeví v zavádějících nebo až chybných interpretacích výsledků. Proto jsme se rozhodli pro používání Excelu při výuce aplikované statistiky. Za jeho přednost považujeme především jeho relativní jednoduchost a dostupnost. Lze ho tak snadno použít i pro projektovou výuku [5], [7].

Pro statistické analýzy a výpočty lze využívat mnoho excelovských funkcí, např. matematických, logických, ale především databázových a statistických [1]. Jsme přesvědčeni, že Excel plně postačuje pro řešení většiny základních úloh aplikované statistiky a že je pro výuku v bakalářských programech názornější a vhodnější. Také na zahraničních univerzitách, jak jsem měla možnost se přesvědčit na College of Business and Economics, Lehigh University, Bethlehem, USA, je Excel považován za základní, vhodný a dostačující prostředek pro statistické analýzy [2], [4]. Při praktické výuce s používáním Excelu jsme se však také setkali s problémy, které jsme se pokusili v tomto článku shrnout.

1. Statistické funkce v Excelu

Jako nedostatek české verze statistických funkcí v Excelu vidíme nedůsledné a často i nesprávné překlady názvů některých funkcí, jejich nesystémové třídění (zařazování) a také mnohdy nesrozumitelné nápovědy k jednotlivým funkcím. Je škoda, že autoři české verze nekonzultovali názvy a označení jednotlivých funkcí se statistiky a při volbě názvů zcela nesystémově kombinují názvy anglické s českými. Tak máme např. směrodatnou odchylku, ale varianci (VAR) místo rozptylu, odmocninu, ale power místo mocniny atd. Takových příkladů lze najít mnohem více.

Některé názvy funkcí jsou nedůsledné a ve své podstatě přímo chybné. Např. počet permutací je zde definován jako:

$$P_{k,n} = \frac{n!}{(n-k)!},$$

což je však počet variací, které nejsou v Excelu definovány, resp. jsou, ale pod chybným názvem permutace. Kombinace pak nelze nalézt, na rozdíl od permutací, ve statistických funkcích, ale ve funkcích matematických.

Dalším problémem je určování kvantilů v Excelu. Jsou zde definovány odlišně od běžně používaných tabulek. Hledáme-li v tabulkách např. kvantil $\chi^2_{0,975(2)}$ rozdělení chí-kvadrát, zadáme v Excelu ve funkcích CHIINV pro pravděpodobnost hodnotu 0,025, tj. hledáme $\chi^2_{0,025(2)}$ a získáme hodnotu 7,377758908. Obdobné problémy jsou také s kvantily jiných, v Excelu definovaných rozdělení pravděpodobností, jak můžeme vidět v tab. 1.

Tabulka 1: Příklady určování kvantilů některých pravděpodobnostních rozdělení

Pravděpod. rozdělení	Kvantil v tabulkách	Kvantil v Excelu	
chí-kvadrát rozdělení	$1 - \alpha/2$	CHIINV	$\alpha/2$
	$\alpha/2$	CHIINV	$1 - \alpha/2$
Studentovo rozdělení	$1 - \alpha/2$	TINV	α
	$1 - \alpha$	TINV	2α

Doporučujeme proto být při určování kvantilů v Excelu obezřetní a pečlivě si přečíst nápovědu k dané funkci, případně hodnoty, alespoň ze začátku, zkonfrontovat s odpovídajícími tabulkami. Za určitý přínos lze pak považovat skutečnost, že studenti získají zkušenost a snad i návyk, automaticky nepřebírat výsledky z počítačových programů. Uvědomí si, že je nutné si ověřit v manuálech nebo v nápovědě co skutečně bylo spočítáno a kontrolovat důsledně zjištěné výsledky během zpracovávání dat.

2. Kontingenční a korelační tabulka

Korelační a kontingenční tabulky patří ve statistických aplikacích k často používaným nástrojům zřehlednění rozsáhlých dvourozměrných statistických souborů při zkoumání závislostí mezi dvěma znaky (proměnnými) [6]. Starší z autorů ještě vzpomíná na úmorné sestavování korelační tabulky „čárkovací metodou“, což byla u rozsáhlých souborů práce velmi lopotná a nikterak

inspirativní. Tím více dnes oceňuje možnosti, které poskytuje při tvorbě korelační tabulky ze základních dat právě funkce Kontingenční tabulka a graf v hlavní nabídce Data v rámci programu Excel. Tuto funkci považujeme z hlediska metodiky výuky základů aplikované statistiky za jednu z nejsilnějších nástrojů Excelu. Také studenti řadí tuto funkci mezi nejužitečnější a nejefektivnější nástroje použitelné v jejich další práci.

Přestože tato funkce je v Excelu označena jako **Kontingenční tabulka**, její využití je mnohem širší než jako klasická kontingenční tabulka, která je dnes ve statistické literatuře obvykle definována pro zkoumání závislosti dvou kategoriálních znaků (Srovn. [3] str. 102 a další.). Dvourozměrná tabulka dvou numerických proměnných se standardně nazývá **korelační tabulka**. Určitou mezeru v názvu jsme ve statistické literatuře zjistili pro označení tabulky, kde závisle proměnná je numerická a nezávisle proměnná kategoriální znak.

Kontingenční tabulka v Excelu umožňuje pracovat s kategoriálními i numerickými proměnnými, a to jak s diskrétními, tak i spojitými, které lze seskupovat do libovolně velkých intervalů. Může tedy plnit funkci jak klasické kontingenční tabulky, tak i tabulky korelační.

Dále se soustředíme na případy, kdy závisle proměnná y je numerický (kvantitativní) znak a závislost na proměnné x lze zkoumat na základě analýzy rozptylu S_y^2 této závisle proměnné veličiny y . Při další analýze je třeba rozlišovat minimálně dvě rozdílné situace z hlediska charakteru nezávisle proměnné x :

- Jestliže nezávisle proměnnou je kategoriální znak nebo kvantitativní diskrétní veličina, která nabývá malého počtu obměn (např. počet členů v rodině), pak je relevantní a jednoznačnou mírou těsnosti statistické závislosti y na x tzv. *poměr determinace*. Tuto charakteristiku však Excel neumí určit přímo jako funkci. Musí se proto spočítat na základě známého rozkladu celkového rozptylu závisle proměnné y na dvě složky, a to rozptyl podmíněných průměrů a průměr z podmíněných rozptylů. Jako míra těsnosti se používá i odmocnina poměru determinace, tj. korelační poměr.
- V případě, že nezávisle proměnná je kvantitativní veličina, srovnává se často poměr determinace s indexem determinace, který vyjadřuje těsnost závislosti vyjádřené zvolenou regresní funkcí.

$$\text{Pro tento typ korelační tabulky platí vztah: } \eta_{yx}^2 \geq I_{yx}^2 \quad (1)$$

Z velikosti rozdílu $\eta_{yx}^2 - I_{yx}^2$ lze za jistých podmínek usuzovat na vhodnost (výstižnost) zvolené regresní funkce pro popis zkoumané závislosti.

Vztah (1) však nemusí platit pro případ, že nezávisle proměnná je spojitá nebo diskrétní veličina, která může nabývat velkého počtu obměn a je pro účely konstrukce korelační tabulky seskupována do určitých skupin neboli intervalů.

Prozkoumejme nyní podrobněji situaci, kdy nezávisle proměnnou veličinou x je také kvantitativní veličina, která může nabývat velkého počtu obměn a je třeba ji pro přehlednost korelační tabulky seskupit do určitých zvolených intervalů. V těchto případech se pro zjednodušení výpočtů nebo v situaci, kdy už původní data, z nichž tabulka vznikla, nejsou k dispozici, doporučují výpočty, v nichž proměnné x a y v jednotlivých intervalech zastupují středy intervalů. Je zřejmé, že celkový rozptyl S_y^2 je pak vždy podhodnocen, neboť v něm chybí složka, která představuje rozptyly v jednotlivých políčkách (buňkách) tabulky. Protože korelační tabulka v Excelu je konstruována z původních dat, lze pro výpočet podmíněných průměrů a rozptylů použít přesnější postup: Místo středu intervalů se použijí průměrné hodnoty a dílčí rozptyly, které lze v excelovské tabulce snadno získat pro každé políčko tabulky.

Pro podmíněné průměry v každé j -té skupině nezávisle proměnné x platí jednoduchý vztah:

$$\bar{y}_j = \frac{\sum_{i=1}^m \bar{y}_{ij} n_{ij}}{\sum_{i=1}^m n_{ij}},$$

kde $\bar{y}_{ij}$ je průměrná hodnota proměnné y v každém jednotlivém políčku tabulky, tj. pro $i = 1, 2, \dots, m$ řádků a $j = 1, 2, \dots, n$ sloupců.

Složitější situace nastává při stanovení podmíněných rozptylů. Pokud chceme získat přesné hodnoty podmíněných rozptylů S_j^2 , musíme je vyjádřit jako součet dvou složek, a to průměru z dílčích rozptylů S_{ij}^2 v jednotlivých buňkách tabulky a rozptylů dílčích průměrů kolem podmíněných průměrů.

Přesný výpočet dostaneme tak, že v každém políčku tabulky použijeme dílčí průměry a rozptyly podle vztahů:

$$s_j^2 = \frac{\sum_{j=1}^m y_{ij}^2 n_{ij}}{\sum_{j=1}^m n_{ij}} - \bar{y}_j = \overline{s_j^2} + s_{\bar{y}_j}^2 \quad (1)$$

kde první složka je průměr z dílčích rozptylů uvnitř buněk (i,j) pro každý j-tý sloupec

$$\overline{s_j^2} = \frac{\sum_{i=1}^m s_{ij}^2 n_{ij}}{\sum_{i=1}^m n_{ij}} \quad (2)$$

a druhá složka je rozptyl dílčích průměrů v j-tém sloupci kolem j-tého podmíněného průměru:

$$s_{\bar{y}_j}^2 = \frac{\sum_{i=1}^m (\bar{y}_{ij} - \bar{y}_j)^2 n_{ij}}{\sum_{i=1}^m n_{ij}} \quad (3)$$

Přestože podmíněné rozptyly pro jednotlivé i-té skupiny nezávisle proměnné x lze také z excelovské kontingenční tabulky získat, použijeme dále pro výpočet poměru determinace rozptyl podmíněných průměrů a vyčíslíme jej dle vztahu:

$$\eta_{yx}^2 = \frac{s_{\bar{y}}^2}{s_y^2} \quad (4)$$

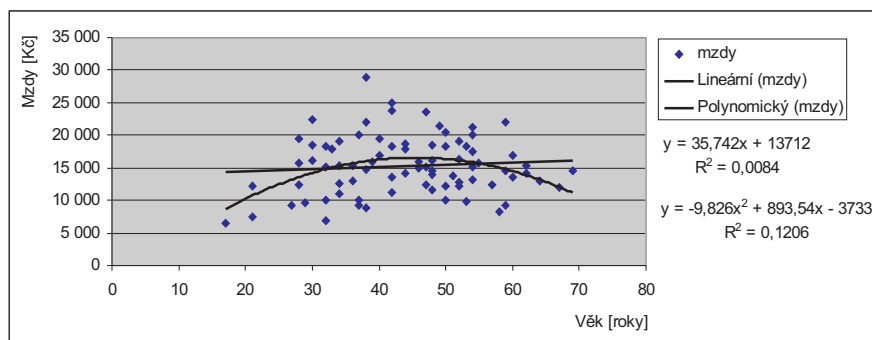
Pro názornost ukážeme postup řešení na příkladu jak je řeší studenti na seminářích.

Jako vstupní data byly zadány údaje o mzdách v Kč (y) a stáří (v letech) osmdesáti zaměstnanců jistého podniku. Základní představu o zkoumané závislosti výše mzdy na stáří zaměstnanců poskytl bodový diagram, viz obr. 1.

Zadáme-li v bodovém grafu funkci spojnice trendu, můžeme dále zvolit různé typy regresních funkcí i s výpočtem indexu determinace. Ten je však v Excelu nazýván jako „hodnota spolehlivosti“ a je označován symbolem R^2 , který je obvykle ve statistické literatuře vyhrazen pro koeficient determinace. To ukazuje na další nedůslednost a nepřesnost českého označování charakteristik v Excelu a je potřeba na tuto skutečnost studenty upozornit.

Zadáme-li v bodovém grafu funkci spojnice trendu, můžeme dále zvolit různé typy regresních funkcí i s výpočtem indexu determinace. Ten je však v Excelu nazýván jako „hodnota spolehlivosti“ a je označován symbolem R^2 , který je obvykle ve statistické literatuře vyhrazen pro koeficient determinace. To ukazuje na další nedůslednost a nepřesnost českého označování charakteristik v Excelu a je potřeba na tuto skutečnost studenty upozornit.

Obrázek 1: Bodový diagram závislosti mezd na stáří zaměstnanců



Tabulka 2: Rozdělení četností pro zkoumání závislosti mezd na stáří

Počet z mzdy	Věk [roky]						
Mzdy [Kč]	15-24	25-34	35-44	45-54	55-64	65-74	Celkem
6000 – 8999	2	1	1		1		5
9000 – 11999		4	3	3	1		11
12000 – 14999	1	2	4	8	5	2	22
15000 – 17999		5	4	7	3		19
18000 – 20999		4	4	6			14
21000 – 23999		1	2	3	1		7
24000 – 26999			1				1
27000 – 29999			1				1
Celkový součet	3	17	20	27	11	2	80

S tabulkou lze v Excelu dále pracovat. Např. pomocí funkce „Nastavení pole“ lze v celé tabulce získat průměry, rozptyly, součty a další charakteristiky. Nastavíme-li pole na „průměry“, zobrazí se v každém políčku průměrná hodnota mzdy a v součtovém řádku se zobrazí podmíněné průměrné mzdy pro každý sloupec nastaveného intervalu stáří zaměstnanců. Pomocí těchto podmíněných průměrů a celkové průměrné mzdy pak vyjádříme rozptyl podmíněných průměrů a jeho podíl na celkovém rozptylu mezd dle (5) je 0,1375.

To znamená, že zhruba 13,75 % rozptylu mezd lze vysvětlit závislostí na stáří lidí, kdežto zbytek, tj. 86,25 % rozptylu lze vysvětlit jinými proměnnými a náhodnými vlivy.

Nastavíme-li pole na „rozptyly“ získáme jednak dílčí rozptyly v jednotlivých políčkách tabulky a podmíněné rozptyly jako okrajové (součtové) hodnoty.

Dále excelovská kontingenční tabulka umožňuje jednoduše zjistit, jaký vliv bude mít změna nastavení délky intervalu nezávisle proměnné veličiny na podmíněné průměry a poměr determinace. Tak např. při nastavení délky intervalů věku na 5 let jsme získali hodnotu poměru determinace 0,1668. Část rozptylu, vysvětlená závislostí mezd se tedy při zmenšení intervalů na polovinu zvětšila zhruba o 3 %.

Výhodou excelovské kontingenční tabulky tedy je, že můžeme snadno experimentovat s délkou intervalů a zjišťovat vliv délky intervalů na poměr determinace. Obecně platné závěry však při těchto experimentech nelze přijímat. Takovéto postupné hledání řešení, při měnících se vstupních podmínkách pak, na rozdíl od přímého použití sofistikovaných statistických softwarů, pomáhá rozvíjet myšlení studentů a lépe chápat statistické procedury a zjištěné výsledky.

Závěr

I přes výše uvedená úskalí používání Excelu se domníváme, že je pro výuku základů aplikované statistiky velmi vhodným nástrojem, a to z několika důvodů. Mezi hlavní přednosti patří jeho široká dostupnost a tudíž i předpoklad dalšího používání v praxi. Univerzitní příprava studentů je tak více konzistentní s pracovním prostředím, pro které jsou studenti vzděláváni.

Možnost současně provádět výpočty a využívat excelovských funkcí pomáhá studentům lépe porozumět statistickým charakteristikám a tím i kvalifikovaněji a přesněji interpretovat zjištěné charakteristiky.

Snadná ovladatelnost a předpoklad poměrně dobré znalosti studentů práce s programy firmy Microsoft, by měla studentům usnadnit a zejména zefektivnit statistické zpracování dat proti klasickému zpracování kalkulátorem. V praxi však ještě stále často není tento předpoklad, proti našemu očekávání, splněn. To sice ztěžuje práci nám vyučujícím a na některé studenty to klade zpočátku vyšší nároky, ale tím spíš se domníváme, že je nezbytné tyto studenty práci s Excelem naučit. Proto je vhodné zařadit do požadavků ke zkoušce z předmětu aplikovaná statistika také odevzdání studentem vypracovaného projektu, ve kterém řeší statistickou úlohu z praxe.

Pohodlný import z tabulkového editoru Excel do textového editoru Word pak zjednodušuje psaní zpráv z laboratoří (protokolů), později bakalářských a diplomových prací, čímž je vytvářen předpoklad pro jeho další používání a zdokonalování se.

Reference

- [1] Bartošová, J.: *Základy statistiky pro manažery*. Oeconomica, Praha, 2006, 198 stran. ISBN 80-245-1019-7.
- [2] Black, K.: *Business Statistics for Contemporary Decision Making*. 4th edition, Wiley, USA, 2004, ISBN 0-471-42983-X.
- [3] Hindls, R., Hronová, S., Novák, I.: *Analýza dat v manažerském rozhodování*. Grada Publishing, 1999, ISBN 80-7169-255-7.
- [4] Levine, M., D., Stephan, D., Krehbiel, T. C., Berenson, M., L.: *Statistics for Managers using Microsoft Excel*. 3rd edition Prentice Hall, New Jersey, USA, 2002, ISBN 0-13-060016-4.
- [5] Stankovičová, I., Ďurčíková, A.: *Projektové vyučovanie predmetu Základy štatistiky na FM UK*, In: Matematická štatistika a numerická matematika a ich aplikácie. Stavebná fakulta STU, Bratislava, 1999.
- [6] Vojtková, M.: *Riešenie problému závislosti medzi vybranými ukazovateľmi efektívnosti*. In: Zborník príspevkov, Výpočtová štatistika 2002. SŠDS, Bratislava, 2002.
- [7] Vlčková, V.: *Inovace výuky statistiky na KEMCH Univerzity Pardubice*. In: Sborník příspěvků, Výuka a výzkum v odvětvových ekonomikách a podnikovém managementu na technických vysokých školách. Univerzita Pardubice, 2000, ISBN 80-7194-301-0.

Adresa: Vladimíra Vlčková a Otakar Machač

Univerzita Pardubice, Studentská 84, 532 10 Pardubice

E-mail: vladimira.vlckova@upce.cz, otakar.machac@upce.cz

HĽADANIE PODOBNOSTI KRAJÍN EÚ PODĽA VYBRANÝCH UKAZOVATEĽOV LISABONSKEJ STRATÉGIE

Mária Vojtková

Abstrakt: This article is interested in identifying groups of European countries that have similar 14 selected indicators of the Lisbon Strategy. These indicators are focus on the European economies and they were establishment by the European Council in March 2004. Cluster analysis is a useful technique for such a purpose. However, the selected indicators are correlated among themselves. In such a case principal components analysis should be performed.

Key words: the Lisbon Strategy, 14 Lisbon indicators, principal component analysis, hierarchical clustering, Ward's method.

1. Úvod

Cieľom tohto príspevku je klasifikácia krajín Európskej únie charakterizovaných vybranými štruktúrnymi ukazovateľmi Lisabonskej stratégie za rok 2003. Pre tento druh analýz bola vyvinutá široká paleta metód a postupov, ktoré sa používajú pri riešení problémov typológie objektov a ich klasifikácie, pričom ich spoločný názov je zhuková analýza.

Pôvodný vstupný súbor obsahoval 15 štruktúrnych ukazovateľov Lisabonskej stratégie za 25 krajín Európskej únie. Podrobný popis jednotlivých ukazovateľov obsahuje iný článok v tomto zborníku¹, pričom uvedená analýza je jeho voľným pokračovaním. Vzhľadom k tomu, že za Maltu nie sú k dispozícii všetky skutočné hodnoty daných ukazovateľov nebudeme s ňou v ďalšej analýze uvažovať. Podobne hodnoty ukazovateľa regionálny rozptyl miery nezamestnanosti nie sú vo všetkých krajinách EÚ dôsledne vykazované, čiže tento ukazovateľ takisto vylúčime z analýzy.

¹Od strany 162 v tomto zborníku. Iveta Stankovičová: Viackriteriálne hodnotenie zamestnanosti členských krajín EÚ na základe vybraných ukazovateľov Lisabonskej stratégie. Zborník príspevkov z medzinárodnej konferencie STAKAN 2007.

2. Podstata vybraných viacrozmerných štatistických metód

V príspevku bola aplikovaná jedna z najpoužívanějších hierarchických metód zhlukovania t.j. Wardova metóda. Pri tejto metóde ide o optimalizačný postup, založený na minimalizácii vnútrozhlukového súčtu štvorcov odchýlok od priemeru zhluku. Pre porovnanie kvality zhlukovania boli použité i ďalšie hierarchické metódy, napr. centroidná metóda a metóda priemernej väzby, avšak pre podobnosť získaných výsledkov v cieľovej analýze už budeme hovoriť iba o Wardovej metóde.

Pri charakterizovaní podobnosti jednotlivých krajín EÚ bola aplikovaná jedna z najbežnejších mier vzdialenosti, konkrétne Euklidovská vzdialenosť. Jej použitie si vyžaduje vzájomnú nekorelovanosť premenných, pričom v našom prípade vzhľadom na prepojenosť konštrukcie jednotlivých štruktúrnych ukazovateľov tento predpoklad nebol dodržaný (korelačnú maticu vzhľadom k rozsahu v článku neuvádzame). Dôsledkom nesplnenia tohto predpokladu je duplicitnosť analyzovaných informácií obsiahnutých vo vstupných ukazovateľoch, ktorá môže viesť ku značnému skresleniu výsledkov.

Tento problém je možné riešiť vynechaním štatisticky významných dvojíc ukazovateľov, čo v prípade ich dôležitosti nie je možné. Z toho dôvodu sme uvedený problém riešili transformáciou vstupných ukazovateľov na vzájomne nezávislé hlavné komponenty aplikáciou metódy hlavných komponentov.

Metóda hlavných komponentov spočíva v transformácii k -rozmerného vektora premenných X_j na q -rozmerný vektor hlavných komponentov $F_h (q \leq k)$ tak, aby jednotlivé hlavné komponenty boli navzájom ortogonálne a vyčerpávali maximum celkového rozptylu:

$$F_h = a_{1h}X_1 + a_{2h} + \dots + a_{qh}X_q,$$

kde a_{jh} sú komponentné saturácie alebo váhy, $h = 1, 2, \dots, q$.

Nové (skryté, latentné) premenné musia spĺňať nasledovné vlastnosti:

- výberové hlavné komponenty (HK) sú lineárnou kombináciou pôvodných štandardizovaných premenných X_j ,
- maximálne možno vytvoriť rovnaký počet HK ako pôvodných premenných,
- nové hlavné komponenty sú vzájomne nekorelované (nezávislé, ortogonálne).

3. Analýza podobnosti členských krajín EÚ

Vhodnosť vstupných štruktúrnych ukazovateľov pre analýzu hlavných komponentov možno overiť pomocou Kaiser-Meyer-Olkinovej miery (KMO). Ide o index porovnávajúci veľkosť korelačných koeficientov voči veľkosti parciálnych korelačných koeficientov. Jej doporučované hodnoty sú nad 0,5, pričom vyššia hodnota signalizuje vhodnejšie použitie príslušného ukazovateľa (maximálna hodnota je 1). V našom prípade priemerná hodnota KMO dosahuje veľkosť 0,641, čo možno považovať za dostatočnú hodnotu.

Tabuľka 1: Vlastné čísla korelačnej matice ukazovateľov Lisabonskej stratégie

Eigenvalues of the Correlation Matrix: Total = 14 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.0208	3.5849	0.4301	0.4301
2	2.4359	0.3281	0.1740	0.6041
3	2.1078	1.1843	0.1506	0.7546
4	0.9234	0.2466	0.0660	0.8206
5	0.6768	0.0193	0.0483	0.8689
6	0.6575	0.2333	0.0470	0.9159
7	0.4242	0.1092	0.0303	0.9462
8	0.3150	0.1691	0.0225	0.9687
9	0.1459	0.0046	0.0104	0.9791
10	0.1412	0.0667	0.0101	0.9892
11	0.0746	0.0383	0.0053	0.9945
12	0.0363	0.0097	0.0026	0.9971
13	0.0265	0.0125	0.0019	0.9990
14	0.0141		0.0010	1.0000

Štruktúrne ukazovatele Lisabonskej stratégie sú teda vhodné pre použitie metód zníženia dimenzie a je možné uskutočniť ich transformáciu na nezávislé hlavné komponenty, ktoré vystihujú maximálne množstvo informácie o pôvodnom súbore. V Tabuľka 1 sú vlastné čísla (eigenvalues), vyjadrujúce absolútnu hodnotu rozptylu vysvetlenú jednotlivými hlavnými komponentmi. Podľa ich veľkosti môžeme uvažovať o počte štatisticky významných hlavných komponentov. Na základe Kaiserovho pravidla prvé tri hlavné komponenty majú vlastné číslo väčšie ako 1, ale vysvetľujú iba 75,46 % variability pôvodných premenných. V záujme zachovania, čo najväčšieho množstva pôvodných

informácií a tiež v dôsledku najlepšie interpretovateľného riešenia budeme uvažovať s 5 hlavnými komponentami, ktoré vysvetľujú 86,89 % variability.

Výsledkom aplikácie metódy hlavných komponentov je rotovaná komponentná matica obsahujúca komponentné saturácie pre jednotlivé štruktúrne ukazovatele a komponenty (Tabuľka 2). V prípade ortogonálnej rotácie (bola použitá rotácia Equamax) ide o párové koeficienty korelácie príslušnej premennej a daného hlavného komponenta. Za významné komponentné saturácie sa považujú váhy väčšie ako 0,5. V našom prípade sme na základe ich hodnôt pomenovali nasledovné hypotetické premenné takto:

1. **hlavný komponent** – zamestnanosť verzus nezamestnanosť,
2. **hlavný komponent** – makroekonomické ukazovatele a výskum,
3. **hlavný komponent** – životné prostredie verzus spotreba energie,
4. **hlavný komponent** – súkromný sektor a nákladná preprava,
5. **hlavný komponent** – vzdelanie verzus chudoba.

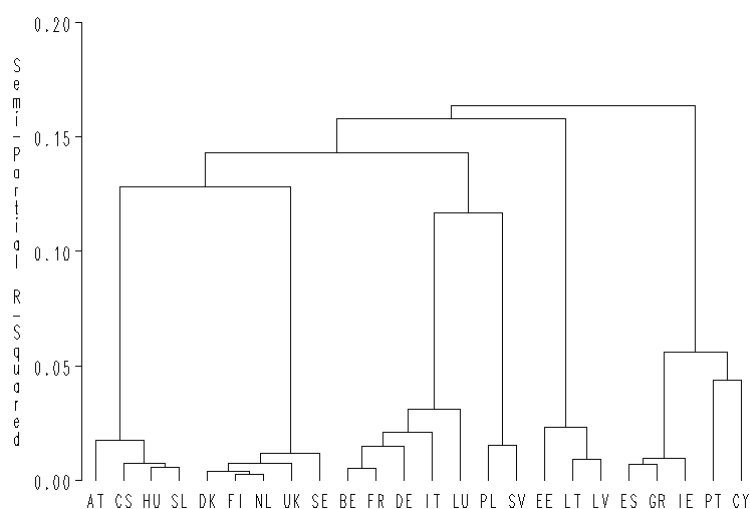
Tabuľka 2: Matica komponentných saturácií ukazovateľov Lisabonskej stratégie

Rotovaná komponentná matica					
	HK1	HK2	HK3	HK4	HK5
Zam_zeny	0.9223	0.1518	-0.0514	-0.1968	0.1481
Zam	0.8956	0.2507	0.2378	-0.1025	0.0379
Zam_starsi	0.8596	-0.0655	0.0255	-0.0929	-0.3146
Dlhodob_nezam	-0.6318	-0.4740	-0.4334	-0.1491	0.0322
HDP_obyv	0.0654	0.8253	0.3951	-0.1905	-0.1085
Prod_prac	-0.0417	0.7358	0.5676	-0.2268	-0.1126
Vydaje_vyskum	0.3891	0.6313	-0.0420	-0.4637	0.1703
Porovnavie_cien	0.3872	0.6278	0.4682	-0.3300	-0.1182
Sklen_plyny	0.0466	0.0114	0.9349	-0.0617	-0.2032
Spotr_energie	-0.0746	-0.4999	-0.7388	0.2958	0.1411
Investicie	-0.1715	-0.2809	-0.1700	0.8334	0.0777
Nakl_preprava	0.1238	0.1056	-0.0769	0.7826	-0.4992
Stup_vzdel	-0.1091	-0.1499	-0.1437	-0.0420	0.8429
Risk_chudoby	-0.3300	-0.4933	0.1901	0.2092	-0.5883

V ďalšej analýze už budeme pracovať s uvedenými piatimi hlavnými komponentami namiesto pôvodných štrnástich premenných. Miesto štruktúrnych ukazovateľov budeme teda analyzovať hodnoty komponentných skóre. Ide

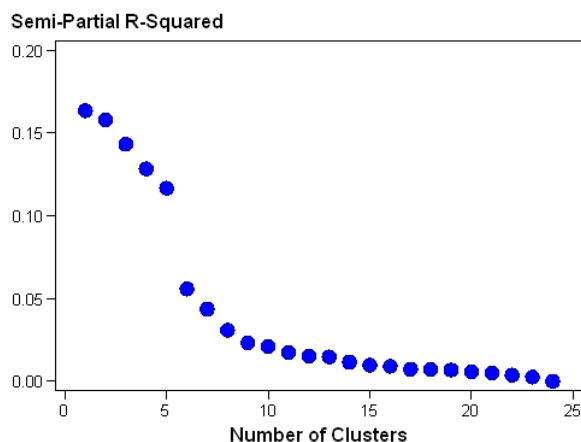
o kombinovanú mieru každého komponentu vyčíslenú pre jednotlivé krajiny. Výsledné komponenty sú vzájomne nezávislé, čo umožňuje ich ďalšie použitie pri klasifikácii krajín EÚ pomocou metód zhlukovej analýzy.

Grafické znázornenie zhlukovania krajín EÚ na jednotlivých úrovniach prehľadne a podrobne poskytuje hierarchický strom – dendrogram (Obrázok 1). Na osi X-ovej sú skratky krajín EÚ, ich plné znenie nie je vhodné použiť vzhľadom k dĺžke jednotlivých názvov, čo by malo za následok neprehľadnosť spomínaného grafu. Ak si rovnobežne s horizontálnou osou preložíme myšlenú čiaru, môžeme pozorovať krajiny patriace do jednotlivých zhlukov podľa rôznych úrovní zhlukovania.



Obr. 1: Dendrogram zhlukovania krajín EÚ podľa ukazovateľov Lisabonskej stratégie pomocou Wardovej metódy

Semiparciálny koeficient determinácie je charakteristikou, ktorú použijeme na zistenie počtu významných zhlukov (jeho hodnoty podľa jednotlivých úrovní sú znázornené na Y-ovej osi Obrázok 1). Našou snahou je zistiť, na ktorom stupni zhlukovania nadobúda tento koeficient minimálnu hodnotu, pri čo najnižšom prírastku pre ďalšie úrovne zhlukovania. Na posúdenie tohto výsledku použijeme jednak absolútne hodnoty semiparciálneho koeficienta determinácie vypočítané pre rôzne stupne zhlukovania a tiež názorný pohľad pomocou grafu znázorňujúceho vývoj tejto charakteristiky v závislosti od počtu zhlukov (Obrázok 2).



Obr. 2: Vývoj semiparciálneho koeficienta determinácie v závislosti od počtu vytvorených zhlukov podľa Wardovej hierarchickej metódy zhľukovania

K výraznému poklesu semiparciálneho koeficienta determinácie dochádza na 6. úrovni zhľukovania, pričom jeho hodnota na tejto úrovni je 0,0558. Hodnota tohto koeficienta blízka k nule poukazuje na významnosť počtu vytvorených zhlukov. Zaujímavá sa javí aj 8. úroveň zhľukovania, pri ktorej charakteristika semiparciálneho koeficienta determinácie nadobúda hodnotu 0,0309. Na ostatných úrovniach dochádza už iba k minimálnemu úbytku tejto charakteristiky.

Na základe týchto výsledkov sme sa rozhodli krajiny EÚ rozdeliť do šiestich zhlukov, pretože vzhľadom k počtu analyzovaných krajín sa nám javí táto hodnota vhodnejšia. Príslušnosť jednotlivých krajín k zhľukom obsahuje Tabuľka 3.

Posledný krok je zameraný na interpretáciu štatisticky významných zhlukov. I keď samotná zhľuková analýza bola uskutočnená s využitím hypotetických nezávislých hlavných komponentov, interpretácia vytvorených zhlukov musí byť zameraná na pôvodné ukazovatele Lisabonskej stratégie, čiže na záver vypočítame zhľukové centroidy pôvodných štruktúrnych ukazovateľov.

Keďže pôvodný vstupný súbor obsahuje pomerne veľký počet vstupných premenných, ich samotná interpretácia podľa jednotlivých zhlukov, ktorých je v našej analýze 6, sa stáva dosť rozsiahla. Zjednodušenie interpretácie nám

Tabuľka 3: Rozdelenie krajín EÚ do zhlukov pomocou Wardovej zhlukovacej metódy

CLUSTER=1		CLUSTER=2		CLUSTER=3	
Row number	Krajina	Row number	Krajina	Row number	Krajina
1	FINSKO	6	BELGICKO	11	MADARSKO
2	HOLANDSKO	7	FRANCUZSKO	12	SLOVINSKO
3	DANSKO	8	NEMECKO	13	CESKA REP.
4	V. BRITANIA	9	ITALIANSKO	14	RAKOUSKO
5	SVEDSKO	10	LUXEMBURSKO		

CLUSTER=4		CLUSTER=5		CLUSTER=6	
Row number	Krajina	Row number	Krajina	Row number	Krajina
15	SPANIELSKO	20	LOTYSSKO	23	POLSKO
16	GRECKO	21	LITVA	24	SLOVENSKO
17	IRSKO	22	ESTONSKO		
18	PORTUGALSKO				
19	CYPRUS				

umožňuje využitie niektorej z metód viacrozmerného hodnotenia, pričom jej použitie si vyžaduje dôslednú znalosť typu (žiaduceho smeru vývoja) vstupného ukazovateľa. V našom prípade je stanovenie typu niektorých štruktúrnych ukazovateľov dosť problematické (napr. ukazovateľ celková vnútroštátna spotreba energie alebo množstvo nákladnej prepravy), takže sa obmedzíme iba na hodnotenie jednotlivých zhlukov z hľadiska veľkosti ich zhlukových centroidov (Tabuľka 4).

Hodnoty zhlukových centroidov podľa jednotlivých zhlukov za štruktúrne ukazovatele Lisabonskej stratégie obsahuje Tabuľka 4. Výsledky zhlukovania možno zhrnúť:

1. zhluk – obsahuje skupinu krajín, ktoré nadobúdajú najvyššie priemerné hodnoty za štruktúrne ukazovatele miery zamestnanosti, porovnania cenovej úrovne s úrovňou EÚ 15 a ukazovateľa výdaje na výskum a rozvoj ako % HDP. Tieto krajiny sú tiež charakteristické najnižšou priemernou dlhodobou mierou nezamestnanosti a najnižšími priemernými investíciami v súkromnom sektore. Pomerne nízke priemerné hodnoty dosahujú i ukazovatele celkovej vnútroštátnej spotreby energie, množstva nákladnej prepravy a miery riziku chudoby. Ostatné hodnoty priemerných ukazovateľov nadobúdajú vyššie hodnoty.

2. zhluk – zahŕňa krajiny, ktoré nadobúdajú najlepšie priemerné hodnoty ukazovateľov HDP a produktivity práce a súčasne najnižšie priemerné hodnoty ukazovateľa celková vnútroštátna spotreba energie. Priemerné ukazovatele zamestnanosti sú na nízkej úrovni, čiže na druhej strane priemerná miera nezamestnanosti dosahuje vyššiu hodnotu. Z hľadiska priemerných výdavkov na výskum a rozvoj a tiež porovnania cenovej úrovne ide o krajiny s druhými najvyššími hodnotami.

3. zhluk – ide o krajiny s najnižšou priemernou mierou riziku chudoby a s vyššou priemernou hodnotou ukazovateľa investícií v súkromnom sektore a dosiahnutého stupňa vzdelania. Pomerne nízke hodnoty dosahuje i ukazovateľ priemernej nezamestnanosti, avšak i zamestnanosti starších pracovníkov. Ostatné štruktúrne ukazovatele nadobúdajú stredné hodnoty zhlukových centroidov.

4. zhluk – obsahuje skupinu krajín s najvyššou priemernou hodnotou ukazovateľa riziku chudoby a ukazovateľa emisie skleníkových plynov, s ktorým súvisí i vyššia hodnota priemerného množstva nákladnej prepravy. Tento zhluk je tiež typický najnižšou priemernou hodnotou dosiahnutého stupňa

vzdelania a pomerne nízke sú priemerné výdaje na výskum a rozvoj ako % HDP. Priemerná celková miera zamestnanosti nadobúda vyššie hodnoty.

5. zhluk – ide o krajiny s najnižšou priemernou hodnotou HDP, produktivity práce a teda i nízkou priemernou hodnotou ukazovateľa porovnanie cenovej úrovne s úrovňou EÚ 15. Z pohľadu životného prostredia ide o krajiny s najnižšou priemernou hodnotou emisie skleníkových plynov, čo sa javí v protiklade s najvyššou priemernou hodnotou nákladnej prepravy a tiež najvyššou priemernou vnútroštátnou spotrebou energie. Zaujímavé priemerné hodnoty dosahuje tento zhluk pri sledovaní zamestnanosti. Celková priemerná zamestnanosť je na strednej úrovni, avšak miera zamestnanosti žien a starších pracovníkov dosahuje vyššie priemerné hodnoty.

6. zhluk – zahŕňa dve krajiny s problémami mier zamestnanosti. Priemerné hodnoty týchto ukazovateľov dosahujú najnižšie výsledky, čo je v súlade s najvyššou priemernou hodnotou dlhodobej miery nezamestnanosti. Pomerne nízke sú tiež priemerné hodnoty HDP, produktivity práce a porovnania cenovej úrovne s úrovňou EÚ 15. Ide o krajiny s najnižším priemerným množstvom nákladnej prepravy a súčasne s druhým najnižším priemerným ukazovateľom emisie skleníkových plynov. Na druhej strane tieto krajiny dosahujú najvyššie priemerné hodnoty dosiahnutého stupňa vzdelania a pomerne vysoké priemerné hodnoty riziku chudoby a vnútroštátnej spotreby energie.

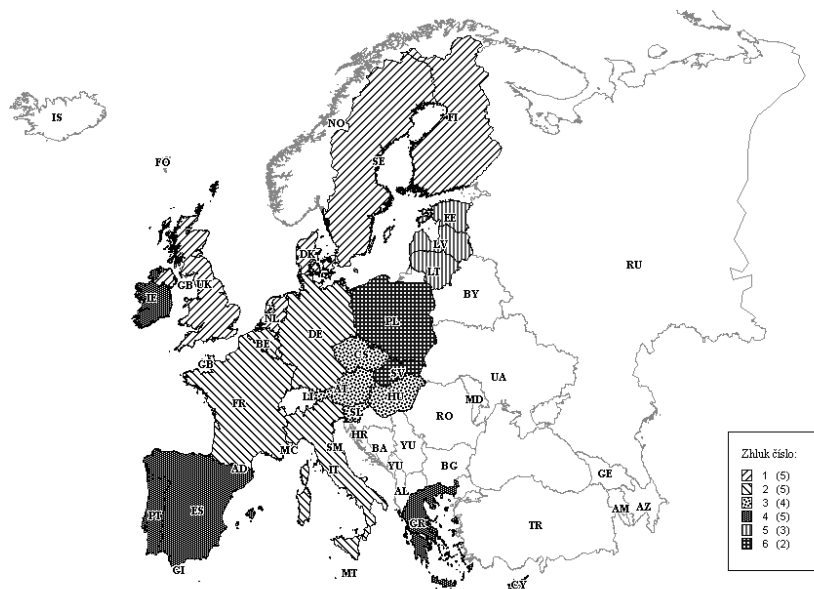
Tabuľka 4: Zhlukové centroidy ukazovateľov Lisabonskej stratégie krajín EÚ

Zhluk	Ukazovateľ	Centroid	Zhluk	Ukazovateľ	Centroid	Zhluk	Ukazovateľ	Centroid
1	HDP_obyv	107,80	3	HDP_obyv	76,25	5	HDP_obyv	41,33
	Prod_prac	98,20		Prod_prac	72,50		Prod_prac	42,33
	Zam	72,40		Zam	63,50		Zam	62,00
	Zam_zeny	68,00		Zam_zeny	57,00		Zam_zeny	58,33
	Zam_starsi	56,00		Zam_starsi	31,25		Zam_starsi	47,00
	Stup_vzdel	79,20		Stup_vzdel	88,00		Stup_vzdel	79,00
	Vydaje_vyskum	2,80		Vydaje_vyskum	1,50		Vydaje_vyskum	0,67
	Investicie	15,60		Investicie	20,75		Investicie	22,00
	Porovnavie_cien	116,20		Porovnavie_cien	70,75		Porovnavie_cien	55,33
	Risk_chudoby	11,60		Risk_chudoby	10,25		Risk_chudoby	17,00
	Dlhodob_nezam	1,20		Dlhodob_nezam	2,50		Dlhodob_nezam	5,00
	Sklen_plyny	97,60		Sklen_plyny	87,75		Sklen_plyny	40,67
	Spotr_energie	206,60		Spotr_energie	493,50		Spotr_energie	1062,67
	Nakl_preprava	90,60		Nakl_preprava	100,75		Nakl_preprava	139,67

Tabuľka 4: Pokračovanie tabuľky. . .

Zhluk	Ukazovateľ	Centroid	Zhluk	Ukazovateľ	Centroid	Zhluk	Ukazovateľ	Centroid
2	HDP_obyv	120,60	4	HDP_obyv	85,20	6	HDP_obyv	44,50
	Prod_prac	112,40		Prod_prac	89,20		Prod_prac	52,00
	Zam	61,40		Zam	63,80		Zam	54,50
	Zam_zeny	52,60		Zam_zeny	53,40		Zam_zeny	49,00
	Zam_starsi	33,00		Zam_starsi	46,60		Zam_starsi	26,00
	Stup_vzdel	75,00		Stup_vzdel	72,20		Stup_vzdel	91,50
	Vydaje_vyskum	2,00		Vydaje_vyskum	0,80		Vydaje_vyskum	1,00
	Investicie	16,40		Investicie	19,40		Investicie	19,00
	Porovnavie_cien	99,60		Porovnavie_cien	87,40		Porovnavie_cien	51,00
	Risk_chudoby	14,00		Risk_chudoby	19,20		Risk_chudoby	18,00
	Dlhodob_nezam	3,80		Dlhodob_nezam	2,80		Dlhodob_nezam	11,00
	Sklen_plyny	95,00		Sklen_plyny	137,20		Sklen_plyny	70,00
	Spotr_energie	189,60		Spotr_energie	237,00		Spotr_energie	807,00
	Nakl_preprava	102,20		Nakl_preprava	123,20		Nakl_preprava	66,00

Prehľadné znázornenie začlenenia krajín Európskej únie na základe podobnosti štruktúrnych ukazovateľov Lisabonskej stratégie do jednotlivých zhlukov je možné pozorovať v danej mape (Obrázok 3).



Obr. 3: Mapa začlenenia krajín EÚ do jednotlivých zhlukov podľa ukazovateľov Lisabonskej stratégie

Výsledné riešenie začlenenia jednotlivých krajín Európskej únie do zhlukov môžeme vo všeobecnosti charakterizovať ako prijateľné a interpretácie schopné. Pre úplnosť je možné dodať, že v prípade rozčlenenia krajín do 8 zhlukov dochádza k zmene v štvrtom zhluku, kde sa samostatne odčlenia krajiny Portugalsko a Cyprus. K pôvodným šiestim zhlukom teda pribudnú dva jednoprvkové zhluky. Naším cieľom v rámci klasifikácie však nie je tvorba izolovaných zhlukov, čiže v danom prípade považujeme analýzu so šiestimi zhlukmi za vhodnejšiu. Uvedená klasifikácia môže viesť k hľadaniu riešenia na zlepšenie nepriaznivého vývoja sledovaných štruktúrnych ukazovateľov Lisabonskej stratégie podľa začlenenia krajín do skupín.

4. Záver

Uvedený článok bol zameraný na aplikáciu vybraných metód viacrozmernej štatistickej analýzy pri hodnotení členských krajín EÚ na základe ukazovateľov Lisabonskej stratégie. Autorka článku sa zaoberá výučbou týchto metód na Ekonomickej univerzite. Vzhľadom k dostupnosti rôznych štatistických softvérov je ich aplikácia pomerne jednoduchá, avšak problémy nastávajú pri interpretácii výsledkov. Často krát je to spôsobené predovšetkým nedôsledným overením prípadne nedodržaním podmienok pre použitie spomínaných metód.

Metódy viacrozmernej štatistickej analýzy sa mnohokrát znevažujú vzhľadom k veľkej subjektivite, ktorú poskytujú. Na druhej strane práve táto subjektivita umožňuje skúsenému analytikovi vybrať to najvhodnejšie riešenie. Na záver by som ešte raz chcela zdôrazniť, že pri interpretácii výsledkov analýzy získaných práve metódami viacrozmernej štatistiky treba byť veľmi opatrný. Prvotné výsledky analýzy je nutné skonfrontovať s viacerými možnými prístupmi a nakoniec vybrať to najpriateľnejšie riešenie.

Literatúra

- [1] EUROPEAN COMMISSION: Facing the challenge. The Lisbon strategy from growth and employment. Luxembourg: Office for Official Publications of the EC. November 2004. ISBN 92-894-7054-2. http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf
- [2] HEBÁK, PETR A KOLEKTÍV: Vícerozměrné statistické metody (3). Praha: Informatorium, 2005. ISBN 80-7333-039-3
- [3] HÚSEK, DUŠAN, ŘEZÁNKOVÁ, HANA, SNÁŠEL, VÁCLAV: Shluková analýza dat. Praha: Professional Publishing, 2007. ISBN 978-80-86946-26-9
- [4] PAŽITNÁ, MÁRIA, LABUDOVÁ, VIERA: Metódy štatistického porovnávania. Kapitola 8, s. 148-173. Bratislava: ES EU, 2007.
- [5] SHARMA, SUBHASH: Applied Multivariate Techniques. New York: John Wiley & Sons, Inc., 1996.
- [6] STANKOVIČOVÁ, IVETA: Viacrozmerná analýza rentability poisťovní SR pomocou Enterprise Guide. In: 10. medzinárodný seminár Výpočtová štatistika. Bratislava: SŠDS, 2001. ISBN 80-88946-14-X

- [7] VOJTKOVÁ, MÁRIA: Viackriteriálne hodnotenie podnikov priemyslu Slovenskej republiky. Roč. 2003, č. 3, s. 320-331, Ekonomické rozhľady. ISSN 0323-262X
- [8] VLČKOVÁ, VLADIMÍRA: Zkušenosti s e-learningem v předmětu Aplikovaná statistika. Forum Statisticum Slovaca 3, s. 191-201, Bratislava: SŠDS, 2006. ISSN 1336-7420
- [9] EUROPEAN COMMISSION: Facing the challenge. The Lisbon strategy from growth and employment. Luxembourg: Office for Official Publications of the EC. November 2004. ISBN 92-894-7054-2. http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf
- [10] EUROSTAT HOME PAGE:
http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1090,30070682,1090_30298591&_dad=portal&_schema=PORTAL

Adresa:

Ing. Mária Vojtková, PhD.
 Katedra štatistiky, FHI, Ekonomická univerzita
 Dolnozemska 1/b
 852 35 Bratislava

E-mail: vojtkova@euba.sk

Poděkování: Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA 1/4586/07 „Modelovanie sociálnej situácie obyvateľstva a domácností v Slovenskej republike a jej regionálne a medzinárodné porovnania“.

MATEMATICKÉ MODELOVANIE V JAZYKOVEDE

MATHEMATICAL MODELLING IN LINGUISTICS

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ЛИНГВИСТИКЕ

Gejza Wimmer

FPV UMB, Banská Bystrica; MÚ SAV, Bratislava; ÚMS PrF MU Brno¹

Abstract: The contribution is focused

- (i) on mathematical modelling in linguistic by the use of discrete probability distributions (morphological productivity of stems in languages, semantic productivity of the language, theory of word lengths) and
- (ii) on an attempt toward a unified derivation of some linguistic laws.

Абстракт: Статья сосредоточена на

- (i) математическое моделирование в лингвистике при помощи дискретных вероятностных распределений (морфологическая производительность основы, семантическая производительность языка, теория длины слов) и
- (ii) попытку об единственной производной большего класса лингвистических законов.

Úvod

Príspevok je ukážkou

- a) *modelovania v jazykovede pomocou diskrétnych pravdepodobnostných rozdelení* a
- b) *pokusu o jednotné odvodenie (veľkej) triedy jazykovedných zákonov*.

Práca vznikla v spolupráci s mnohými jazykovedcami. Výrazná bola spolupráca s prof. G. Altmannom.

¹Pozvaná prednáška na konferencii STAKAN 2007. Podporené grantom VEGA, 1/3016/06 a projektom MŠMT ČR č. LC06024.

1. Jazykovedné zákony a diskkrétne rozdelenia pravdepodobnosti

V jazykovede rovnako ako aj v iných vedných disciplínach sa pokúšame hľadať a odhaľovať zákony (zákonitosti). Podobne ako v iných empirických vedách toto hľadanie nie je jednoduché a môže s realizovať mnohými spôsobmi. Niekedy sa hľadajú analógie s postupmi v iných vedách (fyzika, chémia), niekedy špekulatívnou, deduktívnou cestou vyjdúc z nejakej teórie prideme k hypotéze, ktorá sa po praktickom preverení neskôr môže považovať za zákonitosť, zákon. Často sa pokúšame získať zákonitosť z nameraných údajov, ale táto cesta vo väčšine prípadov stroskotá. Z nameraných údajov obyčajne môžeme (v najlepšom prípade) dostať nejakú funkciu, ktorá dostatočne fituje (vhodne napasuje) získané údaje, dobrú predikciu alebo interpoláciu. Ale veľmi zriedkavo zákon vo všeobecnosti. Pozrime sa na tri problémy.

Predpokladajme, že všetky procesy (samozrejme aj v jazykovede) podliehajú určitým zákonitostiam.

1. Hľadáme zákon *morfologickej produktivity slovných kmeňov* v jazyku. To znamená „Ako môže byť matematicky formulovaný (modelovaný) vznik nových slov a zánik starých slov v jazyku zapríčinený morfologickými zmenami.“
2. Analogicky nás môže zaujímať sémantická produktivita jazyka (vznik nových významov slova resp. strácanie rôznych starých významov).
3. Pýtame sa, akej zákonitosti podlieha (podľa akej zákonitosti sa správa) distribúcia slov podľa ich dĺžky v jazyku (v slovníku jazyka, v textoch jednotlivého autora, atď.).

Takisto nás môže zaujímať zákonitosť výskytu resp. tvorby viet podľa ich dĺžky (v texte), atď. Dĺžkou slova, vety rozumieme počet skladajúcich jazykových jednotiek (napr. slabík, slov, atď.).

Vo všetkých vyšetrovaných analyzovaných prípadoch predpokladáme, že namerané údaje sú dané vo forme frekvenčnej tabuľky. V prvom prípade táto tabuľka vyzerá nasledovne.

kmeň, ktorý vytvára x nových slov morfologickými zmenami	f_x – relatívna frekvencia týchto kmeňov
$x = 0$	f_0
$x = 1$	f_1
$\vdots$	$\vdots$
$x \geq 50$	f_{50}

Vo všetkých skúmaných prípadoch popíšeme hľadajú zákonitosť *diskrétnym pravdepodobnostným rozdelením*. Ako sa k nemu dostaneme? Základné možnosti sú

- použijeme urnovú schému
- kombinatorické úvahy
- iné cesty.

Morfologická produktivita slovných kmeňov v jazyku

V prvom probléme (morfologická produktivita slovných kmeňov v jazyku) z teórie (predchádzajúcich úvah a usudzovaní) je známe, že slová z hľadiska morfolologickej produktivity podliehajú dvom procesom:

- (i) nové slovo vznikne kreativitou (invenciou, improvizáciou) tvorcu (hovoriaceho, redaktora, spisovateľa),
- (ii) nové slovo zanikne
 - (a) naraz (napr. v jednom roku v časopise Spiegel bolo zistených 8000 nových slov, dovtedy nevyskytujúcich sa v žiadnom nemeckom slovníku, ktoré vznikli „len tak“)
 - (b) v priebehu času (keď prestane byť aktuálne, keď sa nájde nový, priliehavejší výraz).

Zánik slova je zapríčinený

- (α) príjemca (poslucháč) rozumie novému slovu, ale ho nikdy nepoužije
- (β) jazyk ako taký „má nastavenú“ hornú hranicu pre tvorbu nových slov z existujúcich slovných kmeňov (keby tomu tak nebolo, bolo by možné z jedného slovného kmeňa vytvoriť morfológickými pochodmi všetky slová a všetky ostatné slovné kmene by sa mohli eliminovať).

Prípád (i) popisuje *proces vzniku (birth process)*, prípad (ii) zase *proces zániku (death process)* definované v teórii pravdepodobnosti.

- (i) V triede slovných kmeňov s vlastnosťou, že z každého kmeňa (v tejto triede) je možné vytvoriť x nových slov morfológickými zmenami je intenzita procesu vzniku

$$\lambda_x = \frac{f(x)}{g(x)} = \frac{a + xc}{b + (n - x - 1)c}, \quad x = 1, 2, \dots, n - 1. \quad (1.1)$$

(ii) Intenzita procesu zániku pre x -tú triedu je

$$\mu_x = \frac{x}{n-x+1}, \quad x = 1, 2, \dots, n. \quad (1.2)$$

V jazyku existuje rovnovážny stav. To znamená, že nové slová vznikajú a tiež slová zanikajú, ale nikdy neprichádza „ku katastrofe“ v zmysle komunikačnej schopnosti jazyka. Matematicky to vyjadrujeme rovnicami pre rovnovážny stav (steady state equations)

$$\begin{aligned} \lambda_0 P_0 &= \mu_1 P_1 \\ (\lambda_i + \mu_i) P_i &= \lambda_{i-1} P_{i-1} + \mu_{i+1} P_{i+1}, \quad i = 1, 2, \dots, n-1 \\ \lambda_{n-1} P_{n-1} &= \mu_n P_n. \end{aligned} \quad (1.3)$$

(P_x je pravdepodobnosť, že slovný kmeň patrí do x -tej triedy, t.j. môže sa z neho vytvoriť morfológickými procesmi x nových slov.)

Keď vyjdeme z predpokladu, že $\{P_0, P_1, \dots, P_n\}$ je pravdepodobnostná funkcia (určuje istú distribúciu), tak použitím niektorých kombinatorických identít dostaneme

$$P_x = \frac{\binom{\frac{a}{c}+x-1}{x} \binom{\frac{b}{c}+n-x-1}{n-x}}{\binom{\frac{a+b}{c}+n-1}{n}}, \quad x = 0, 1, \dots, n,$$

čo je dobre známa pravdepodobnostná funkcia diskkrétnej náhodnej veličiny s *Pólyovým rozdelením pravdepodobnosti*.

Vskutku, keď testujeme testom dobrej zhody údaje získané z mnohých jazykov, Pólyovo rozdelenie dáva vynikajúce výsledky, t.j. test „prakticky vždy“ nezamieta, že údaje sú realizáciami náhodnej veličiny s Pólyovým rozdelením pravdepodobnosti. Toto podporuje hypotézu, že morfológická produktivita jazyka sa môže matematicky modelovať (popísať) ako proces vzniku a zániku s intenzitami (1.1) a (1.2). Zákonitosť (zákon) je vyjadrená pomocou Pólyovho rozdelenia pravdepodobnosti. Intenzity majú veľmi rozumnú lingvistickú interpretáciu. Čitateľ procesu vzniku vyjadruje vytváranie (kreáciu) nových slovných konštruktov (hovoriacim, redaktorom, spisovateľom). Menovateľ vyjadruje „brzdiacu silu“ počúvajúceho (príjemcu, poslucháča, čitateľa) ako aj samotného jazyka (možnou hornou hranicou vyjadrenou číslom n). Odhadnuté parametre a , b , c , n prinášajú informáciu o jazyku, o autorovi, o čase vzniku textu, atď. Podrobnejší výklad pozri v článku Wimmer, Altmann (1995).

Sémantická produktivita jazyka

V druhom probléme (sémantická produktivita jazyka) sú v hre (sú navrhnuté) dve cesty (dve riešenia).

Prvá cesta vedie k procesu vzniku a zániku s intenzitou vzniku

$$\lambda_x = a + x, \quad x = 0, 1, 2, \dots$$

a intenzitou zániku

$$\mu_x = a + b + x, \quad x = 0, 1, 2, \dots$$

čo vedie na *Waringovo rozdelenie pravdepodobnosti*

$$P_x = \frac{b}{a+b} \frac{a^{(x)}}{(a+b+1)^{(x)}}, \quad x = 0, 1, 2, \dots$$

pre rovnovážny stav v komunikácii.

Druhá cesta vedie k *Bissingerovmu geometrickému rozdeleniu*, ktoré patrí do triedy rozdelení čiastočných súčtov (partial-sums distributions), pričom

$$P_x = \frac{p}{1-p} \sum_{j=x}^{\infty} \frac{(1-p)^j}{j}, \quad x = 0, 1, 2, \dots, 0 < p < 1.$$

Obidva modely boli testované na údajoch z maorského jazyka. Druhá cesta sa zdá byť výhodnejšia (priaznivejšia, lepšia) lebo automaticky zahŕňa (zaraďuje) tento výsledok do širšej (všeobecnejšej) teórie, potvrdzuje (podopiera) ju a táto teória sama dostáva deduktívnu podporu pomocou nej. Spomenutá teória bude prezentovaná neskôr v mojom príspevku. Viac o sémantickej produktivite v maorskom jazyku pozri v článku Wimmer, Altmann (1999).

Teória slovných dĺžok

Pokúsme sa analyzovať tretí problém (tvorba slov v jazyku podľa ich dĺžok). Slovná dĺžka je ovplyvnená obrovským množstvom faktorov. Na základe lingvistických analýz týchto faktorov by sa mala slovná dĺžka správať „ako biely šum“. Ale opak je pravdou. Príčiny tohoto sú:

- (i) dĺžka slova je ovplyvnená mnohými faktormi, ale ona sama (slovná dĺžka) ovplyvňuje mnohé ďalšie vlastnosti jazyka. V tomto zmysle vplýva na rôzne atribúty (vlastnosti, črty) jazyka, „riadi“ mnohé zákonitosti. Toto spôsobuje samoreguláciu (self-regulation) v jazyku.

- (ii) každé novovytvorené alebo zaniknuté slovo vnáša poruchy (disturban-
cie) aspoň do jednej zákonitosti jazyka. Keď je táto porucha malá, za-
účinkuje samo-regulačný proces. Ak je porucha veľká, spôsobuje spon-
tánnu samoreguláciu k novému rovnovážnemu stavu (k novému atrak-
toru, k novému pravdepodobnostnému rozdeleniu). Tieto lingvistické
úvahy môžu byť matematicky vyjadrené ako

$$P_x (= \text{pravdepodobnosť, že slovo má } x \text{ slabík}) \sim P_{x-1}. \quad (1.4)$$

Nech

$$P_x = g(x)P_{x-1}. \quad (1.5)$$

Keď zvolíme $g(x) = ax^{-b}$, $a > 0$, $b > 0$ (čo je vhodná funkcia vyho-
vujúca tzv. Menzerathovému zákonu), dostaneme *Conwayovo-Maxwellovo-
Poissonovo rozdelenie*

$$P_x = \frac{a^x}{(x!)^b} P_0, \quad x = 0, 1, 2, \dots,$$

Toto rozdelenie pravdepodobnosti sa našlo v slovenčine, kórejšine, maďarčine,
poľštine. Analýza veľkého množstva textov v mnohých jazykoch ukázala, že
tvorba slov podľa ich dĺžok je stále len modifikácia základného modelu (1.5),
konkrétne

$$b = 0 \quad (\text{a rozšírenia})$$

$$g(x) = a \quad - \text{dostávame z (1.5) } \textit{geometrické rozdelenie}$$

$$P_x = (1 - a)a^x, \quad x = 0, 1, 2, \dots$$

Toto rozdelenie pravdepodobnosti sa doteraz neobjavilo v praxi (u slovných
dĺžok).

$$g(x) = a(R - x + 1) \quad - \text{dostávame z (1.5) } \textit{Palmovo-Poissonovo rozdelenie}$$

$$P_x = \frac{a^x R(R-1) \dots (R-x+1)}{\sum_{j=0}^R R(R-1) \dots (R-j+1) a^j}$$

Toto rozdelenie sa našlo v taliančine.

$$b = 1 \quad (\text{a rozšírenia})$$

$$g(x) = \frac{a}{x} \quad - \text{ dostávame z (1.5) Poissonovo rozdelenie}$$

$$P_x = \frac{e^{-a} a^x}{x!}, \quad x = 0, 1, \dots$$

Toto rozdelenie sa našlo v nemčine, ruštine, poľštine, maďarčine.

$$g(x) = \frac{a + cx}{x} \quad - \text{ dostávame z (1.5) negatívne binomické rozdelenie}$$

$$P_x = \binom{\frac{a}{c} + x}{x} (1 - c)^{\frac{a}{c} + 1} c^x, \quad x = 0, 1, \dots$$

Toto rozdelenie sa našlo v nemčine, dánštine, nórštine.

$$g(x) = \frac{a - cx}{x} \quad - \text{ dostávame z (1.5) binomické rozdelenie}$$

$$P_x = \binom{\frac{a}{c} - 1}{x} (1 + c)^{-\frac{a}{c} + 1} c^x, \quad x = 0, 1, \dots, \frac{a}{c} - 1$$

Toto rozdelenie sa našlo v češtine, turečtine, latinčine, poľštine, estónčine.

$$g(x) = \frac{a}{c + x} \quad - \text{ dostávame z (1.5) hyperpoissonovské rozdelenie}$$

$$P_x = \frac{a^x}{c^{(x)} \sum_{j=0}^{\infty} \frac{a^j}{c^{(j)} j!}}, \quad x = 0, 1, \dots$$

Toto rozdelenie sa našlo v slovenčine, kórejšine, nemčine, staroislandštine, gréčtine, estónčine, starohebrejšine.

$$g(x) = \frac{a + cx}{d + ex} \quad - \text{ dostávame z (1.5) hyperpascalovské rozdelenie}$$

$$P_x = \frac{\binom{\frac{a}{e} + x}{x}}{\binom{\frac{d}{e} + x}{x}} \left(\frac{c}{e}\right)^x P_0, \quad x = 0, 1, \dots$$

Toto rozdelenie sa našlo vo fínštine.

Rozšírením vzťahov (1.4) a (1.5) na

$$P_x = g(x) \sum_{j=1}^x h(j) P_{x-j}, \quad (1.6)$$

pričom v (1.6) uvažujeme $g(x)$ a $h(j) = j\Pi_j(\{\Pi_0, \Pi_1, \dots\})$ je nejaká pravdepodobnostná funkcia), dostávame zovšeobecnené *Poissonovo rozdelenie*. V špeciálnom prípade, keď uvažujeme $\Pi_1 = a$, $\Pi_2 = 1 - a$, $0 < a < 1$ (*Bernoulliho rozdelenie*), dostávame z (1.6) *Hiratovo-Poissonovo rozdelenie*

$$P_x = \sum_{i=0}^{[x/2]} \binom{x-1}{i} \frac{e^{-a} a^{x-i}}{(x-i)!} (1-a)^{x-2i} a^i, \quad x = 0, 1, \dots$$

Toto rozdelenie sa našlo vo francúzštine a nemčine.

Ak

$$\Pi_j = \frac{e^{-bj} b^j j^{j-1}}{(j-1)!}, \quad j = 1, 2, \dots$$

(*Borelovo rozdelenie*), dostávame z (1.6) *Consulovo-Jainovo-Poissonovo rozdelenie*

$$P_x = \frac{a(a+bx)^{x-1} e^{-(a+bx)}}{x!}, \quad x = 0, 1, \dots$$

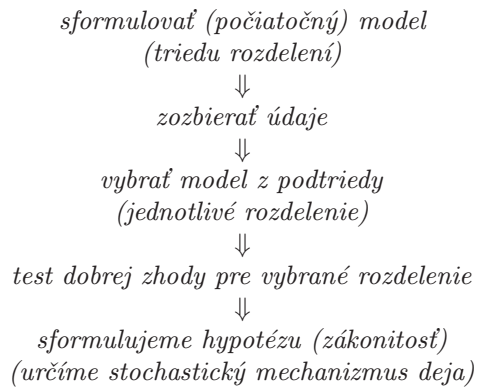
Toto rozdelenie sa našlo skoro v každom jazyku.

Pri hľadaní zákonitostí sme postupovali opačne. Použili sme špeciálny softvér tzv. „Altmann Fitter“, ktorý fituje namerané údaje (vo forme frekvenčnej tabuľky) na viac ako 200 pravdepodobnostných rozdelení. Podľa vhodnosti fitu (napasovania), vždy niektoré rozdelenia boli zvolené na ďalšiu analýzu. Podľa zhody (vybraných) matematických vlastností týchto rozdelení s možnými lingvistickými interpretáciami týchto vlastností sme sa pokúšali sformulovať zákonitosť (teda vybrať vhodné rozdelenie pravdepodobnosti pre skúmanú zákonitosť).

Keď zhrnieme doterajšie úvahy, môžeme skonštatovať, že namiesto tradičnej cesty objavovania určitej zákonitosti, ktorá používa metódu

$$\text{postaviť hypotézu} \implies \text{zozbierať údaje} \implies \text{testovať hypotézu}$$

sme použili



Modelovanie pomocou diskretných pravdepodobnostných rozdelení vo všeobecnosti v empirických vedách môže prebiehať dvomi spôsobmi:

- (i) pomocou modelovania vytvárajúcich mechanizmov, ktoré sa historicky vyvíjali a prinášajú informácie o vzniku dát,
- (ii) pomocou usporiadania nameraných údajov a hľadania zákonitostí v tomto usporiadaní.

Vyššie spomenuté výskumy patrili k spôsobu (i). Spôsobom (ii) odhalíme tiež veľa lingvistických modelov – zákonitostí (spomenieme napr. *Zipfov zákon* s jeho mnohými modifikáciami). Veľmi široká trieda diskretných pravdepodobnostných modelov, ktorá sem tiež patrí sú vyššie spomenuté rozdelenia *čiasťových súčtov*.

Podíme teraz k druhej téme nášho príspevku.

2. Jednotné odvedenie (veľkej) triedy jazykovedných zákonov

V každej vednej oblasti začína výskum roztrúsene, teda ako membra disiecta, lebo neexistuje žiadna teória, ktorá by systematizovala vedomosti a pomocou ktorej by sa mohli formulovať hypotézy.

Sami tí, čo robia výskum, majú rozličné vedecké záujmy a najmä v začiatkoch skúmajú úzke výseky reality. Neskôr sa spájajú krok za krokom nesúrodé oblasti výskumu (napr. jednotná reprezentácia všetkých druhov pohybu v makrosвете pomocou Newtonovej teórie). Staršie teórie sa obyčajne stávajú špeciálnymi prípadmi nových teórií. Hovorí sa o epistemickej integrácii (týkajúcej sa poznávania alebo vedenia) (Bunge 1983: „Integrácia

prístupov, dát, hypotéz, teórií, ba dokonca celých oblastí výskumu je potrebná nielen na vysvetlenie vecí, ktoré silne interagujú so svojim okolím. Epistemická integrácia je potrebná všade, lebo neexistujú úplne izolované veci. Každá vlastnosť súvisí s inými vlastnosťami. Každá vec je systém alebo časť systému... Teda tak isto, ako rôznorodosť sveta (skutočnosti) si vyžaduje veľké množstvo disciplín, ktoré realitu skúmajú, ich integrácia je nutná z hľadiska jednotnosti sveta.“

Kvantitatívna lingvistika stojí na začiatku takéhoto vývoja. Existujú dva „veľké“ integrujúce „cezhraničné“ prístupy – Köhlerova (1986) jazyková synergetika a Hřebíčková (1997) teória textov. Ďalej sú tu niektoré „menšie“, ktoré spájajú rôzne jazykovedné javy. Medzi ne patria napríklad:

- a) Baayen (1989), Chitashvili a Baayen (1993), Zörnig a Boroda (1992), Balasubrahmanyam a Naranan (1997) ukazujú, že distribúcie, ktoré dostaneme usporiadaním frekvencií môžu byť transformované a vyšetované ako ne-rankované, čo bolo neformálne naznačené už Rapoportom (1982).
- b) Altmann (1990) ukázal, že Bühlerova „teória“ je len špeciálnym prípadom Zipfovej teórie (1949), ktorý videl „princíp najmenšej námahy (sily)“ za každým ľudským javom (úkazom, činnosťou).
- c) Viac integrujúci je Menzerathov zákon, ktorého účinky môžeme pozorovať nielen v rôznych oblastiach lingvistiky, ale aj v molekulárnej biológii, sociológii a psychológii (Altmann, Schwibbe (1989)). Je to paralela k alometrickému zákonu a môžeme ho nájsť aj v teórii chaosu (Schroeder (1990), Hřebíček (1997)) alebo v muzikológii (Boroda, Altmann (1991)).
- d) Orlov, Boroda a Nadarejshvili (1982) hľadali spoločné črty vyskytujúce sa v jazykovede, hudbe aj vo výtvarnom umení. Objavili platnosť Zipfovho-Mandelbrotovej zákona.
- e) Krylov, Naranan a Balasubrahmanyam, všetko fyzici, prišli zhodne a nezávisle k poznatku, že princípom maximalizácie entropie sa dajú vynikajúco odvodiť distribúcie niektorých lingvistických entít.

Mohli by sme dlho pokračovať vo vymenovaní príkladov „zjednocovania oblastí“. Vyššie sme uviedli len niekoľko príkladov. Na každom prípade môžeme vidieť spoločný poznatok, že v podstate všetko vedie k teórii systémov. Všetky veci sú systémy. Spájame dva oblasti ak nachádzame izomorfizmy, podobnosti, paralely medzi príslušnými systémami alebo ak zistíme, že ony sú špeciálnymi prípadmi všeobecnejšieho systému. Z času na čas treba spraviť takúto integráciu, aby sme dostali jednotnejšie teórie a lepšie utriedenie vedomostí o skúmanom objekte. V tomto príspevku chceme ukázať prístup

ktorý zjednocuje mnohé známe lingvistické hypotézy, ľahko sa dá zovšeobecniť a je veľmi jednoduchý – aj keď jednoduchosť nepatrí k nutným cnostiam vedy (pozri. Bunge (1963)). Tento prístup je logické rozšírenie „synergetického“ prístupu (pozri. Wimmer, Köhler, Grotjahn, Altmann (1994), Wimmer, Altmann (1996), Altmann, Köhler (1996)). Jednotlivé hypotézy patriace k tomuto systému boli skôr sformulované ako empirické krivky (funkcie), ktoré dobre fitujú určité úkazy (javy) alebo boli odvodené z rôznych iných prístupov.

Spojité prístup

V jazykovede sa môžeme stretnúť so spojitými premennými predovšetkým vo fonetike, ale si musíme uvedomiť, že „premenná“ je iba konštrukt (koncept) nášho nástroja – matematiky, ktorým sa snažíme vystihnúť (zachytiť) stupne skutočných vlastností vecí. Mnohokrát ich transformujeme z „diskrétnych“ na „spojité“ (napr. prímer) alebo naopak (napr. rozdelením spojitých stupnice na intervaly), podľa toho, ako to potrebujeme. Toto nie je nič neobvyklé vo vede. V tomto zmysle neurobíme nič zlého, ak modelujeme spojitý jav použitím diskrétnych modelov alebo naopak. „Spojitý“ a „diskrétny“ sú vlastnosti našich koncepcií, prvé aproximácie nášho epistemického snaženia (snaženia poznávať, vedieť).

Začneme z dvoch predpokladov, ktoré sú v jazykovede veľmi rozšírené a akceptované. Najprv spojitý prípad:

- (i) Nech y je spojitá premenná. Zmena dy tejto premennej je regulovaná (ovládaná) priamo jej veľkosťou (závisí od jej veľkosti), lebo každá lingvistická premenná je konečná a je časťou samoregulujúceho sa systému. Preto pri modelovaní môžeme vždy použiť relatívny pomer zmeny dy/y .
- (ii) Každá lingvistická premenná y je spojená s najmenej jednou inou premennou x ktorá ovplyvňuje jej správanie sa a ktorú budeme považovať v danom prípade za nezávislú premennú. Nezávislá premenná x ovplyvňuje závislú premennú aj cez jej zmenu dx , ktorá je spätne riadená (ovplyvnená) rôznymi mocninami hodnôt x , ktoré sú prepojené s rôznymi inými faktormi (silami, vplyvmi, atď.).

Predpokladajme, že x a y sú rôzne škálované a preto dva vyššie uvedené predpoklady môžeme formálne napísať ako

$$\frac{dy}{y-d} = \left(a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{(x-b_{1i})^{c_1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{(x-b_{2i})^{c_2}} + \dots \right) dx \quad (2.1)$$

s $c_i \neq c_j$, $i \neq j$. (Poznamenávame len, že pre $k_s = 0$ je $\sum_{i=1}^{k_s} \frac{a_{si}}{(x-b_{si})^{c_s}} = 0$.) Konštanty a_{ij} musia byť v každom jednotlivom prípade rozlične interpretované. Reprezentujú vlastnosti, „sily“, „príkazové (radiace) parametre“, požiadavky systému, atď., ktoré sa aktívne zúčastňujú pri prepojení premenných x a y (pozri Köhler (1986, 1987, 1989, 1990)), ale zostávajú konštantné pre podmienku *ceteris paribus* (modelujeme za predpokladu, že ostatné premenné sú konštantné). V diferenciálnej rovnici (2.1) sú premenné už separované. Riešenie (2.1) je

$$y = C e^{a_0 x} \prod_{i=1}^{k_1} (x - b_{1i})^{a_{1i}} e^{\left(\sum_{j \geq 2} \sum_{i=1}^{k_j} \frac{a_{ji}}{(1-c)(x-b_{ji})^{c_j-1}} \right)} + d \quad (2.2)$$

Najčastejšie a najznámejšie riešenia tohto prístupu sú

- (a) počet všetkých slov – počet odlišných slov v texte, tzv. type-token krivky,
- (b) Menzerathov zákon, pri ktorom ide o zákonitosť vzťahu veľkosti konštruktu a konštituent,
- (c) Piotrowského-Bektajeva-Piotrowskej zákon o raste slovníka,
- (d) Narananov-Balasubrahmanyánov model výskytu slov,
- (e) Geršícovo-Altmannovo rozdelenie pre trvanie samohlások, ktoré je identické s jedným Narananovým-Subrahmanyánovým rozdelením
- (f) Jobov-Altmannov model zmeny jednotlivých foném, modelujúci pravdepodobnosť zmeny foném v závislosti na ich norme a komplexite
- (g) Tuldavov zákon polysémie,
- (h) Uhlírovej zákon vyjadrujúci závislosť výskytu podstatných mien na danej pozícii vo vete,
- (i) spojená varianta Zipfovho-Mandelbrotovej zákona a jej špeciálne prípady (pozri napr. Zörnig-Altmann(1993)) atď. atď.

Spojité dvojrozmerný prístup

Doterajší prístup samozrejme nie je vo všeobecnosti postačujúci. V synergetickej lingvistike existuje množstvo vzťahov, ktoré nemožno vystihnúť pomocou funkcie jednej premennej, keď ostatné premenné „zatajíme“ podmienkou „*ceteris paribus*“ (považujeme ich za konštantné). Niekedy ich musíme vziať do úvahy.

Najprv uvažujme jednoduchý špeciálny prípad vzorca (2.1), a síce

$$\frac{dy}{y} = \left(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \dots \right) dx \quad (2.3)$$

ktorého riešenie je

$$y = C e^{a_0 x} x^{a_1} e^{-a_2/x}. \quad (2.4)$$

Vyjadruje napr. Geršicov-Altmanov model trvania samohlásky. V (2.3) predpokladáme, že všetky ostatné faktory (okrem x) sú slabšie ako x a môžu byť považované za konštanty vzhľadom k mocninám x (napr. k a_2/x^2 , a_3/x^3 atď.). Ale v synergetickej lingvistike toto nie je pravidlom. V mnohých modeloch sa ukázalo, že treba uvažovať závislosť jednej premennej od mnohých iných premenných. Dostávame v prvom priblížení sústavu

$$\frac{\partial y}{\partial x} = y \left(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \dots \right); \quad \frac{\partial y}{\partial z} = y \left(b_0 + \frac{b_1}{z} + \frac{b_2}{z^2} + \dots \right), \quad (2.5)$$

ktorej riešenie je

$$y = C e^{a_0 x + b_0 z} x^{a_1} z^{b_1} \exp \left(- \sum_{i=1}^{\infty} \frac{a_{i+1}}{i x^i} - \sum_{i=1}^{\infty} \frac{b_{i+1}}{i z^i} \right). \quad (2.6)$$

Špeciálne prípady (2.6) sa často vyskytujú v synergetickej lingvistike, pričom sa uvažuje o viac ako dvoch premenných. Takýto zovšeobecnený systém, ktorý môže obsahovať ľubovoľné (konečné) množstvo premenných, môže obsiahnuť v podstate celú synergetickú lingvistiku a je aplikovateľný na veľmi zložité systémy. Niektoré dobre známe prípady zo synergetickej lingvistiky sú

$$y = C x^a z^b \quad (2.7)$$

$$y = C e^{ax+bz} \quad (2.8)$$

$$y = C e^{ax+bz} x^a z^b \quad (2.9)$$

atď.

Diskrétny prístup

Ak X je diskretná premenná (čo zvyčajne v lingvistike býva), potom použijeme namiesto infinitezimálneho prírastku dx diferenciu $\Delta x = x - (x-1) = 1$. Obyčajne sa jedná o celočíselné nezáporné náhodné premenné s pravdepodobnostnou funkciou $\{P_0, P_1, \dots\}$, preto uvažujeme relatívnu zmenu

$$\frac{\Delta P_{x-1}}{P_{x-1}} = \frac{P_x - P_{x-1}}{P_{x-1}} \quad (2.10)$$

a dostávame diskretnú analógiu vzťahu (2.1), síce

$$\frac{\Delta P_{x-1}}{P_{x-1}} = a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{(x-b_{1i})^{c_1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{(x-b_{2i})^{c_2}} + \dots \quad (2.11)$$

Ak $k_1 = k_2 = \dots = 1$, $d = b_{11} = b_{21} = \dots = 0$, $c_i = i$, $a_{i1} = a_i$, $i = 1, 2, \dots$, ekvivalentná forma (2.11) je

$$P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \dots\right) P_{x-1}. \quad (2.12)$$

Systém, ktorý sa najviac používa v lingvistike je

$$P_x = \left(1 + a_0 + \frac{a_1}{x-b_1} + \frac{a_2}{x-b_2}\right) P_{x-1}. \quad (2.13)$$

Jeho riešenie je

$$P_x = \frac{(1+a_0)^x \binom{C-B+x}{x} \binom{D-B+x}{x}}{\binom{-b_1+x}{x} \binom{-b_2+x}{x}} {}_3F_2^{-1}(1, C-B+1, D-B+1; -b_1+1, -b_2+1; 1+a_0), \quad x = 0, 1, 2, \dots, \quad (2.14)$$

kde

$$B = \frac{b_1 + b_2}{2},$$

$$C = \frac{a_1 + a_2 - \left(2(1+a_0)^2(b_1-b_2)^2 - 2(1+a_0)(a_1-a_2)(b_1-b_2) + (a_1+a_2)^2\right)^{1/2}}{2(1+a_0)},$$

$$D = \frac{a_1 + a_2 - \left(2(1+a_0)^2(b_1-b_2)^2 - 2(1+a_0)(a_1-a_2)(b_1-b_2) + (a_1+a_2)^2\right)^{1/2}}{2(1+a_0)}.$$

Z rekurentných vzťahov (2.12) a (2.13) dostávame množstvo známych rozdelení, ktoré sa používa v lingvistike, napr. geometrické rozdelenie, Katzovu

triedu rozdelení, diverzifikačné rozdelenia, rozdelenia usporiadaných frekvencií, rozdelenia vzdialeností, Poissonovo rozdelenie, negatívne binomické rozdelenie, hyperpoissonovo rozdelenie, hyperpascalovo rozdelenie, Yuleovo rozdelenie, Simonovo rozdelenie, Waringovo rozdelenie, Johnsonovo-Kotzovo rozdelenie, negatívne hypergeometrické rozdelenie, Conwayovo-Maxwellovo-Poissonovo rozdelenie, atď., atď.

Zákony (zákonitosti), ktoré sa dajú formulovať pomocou tohto systému rozdelení sú napr. Frumkinin zákon, rôzne zákony rozdelenia dĺžok slabík, slov a viet, niektoré formy Zipfovho zákona, zákony usporiadania, rozdelenia syntaktických vlastností, Krylovov sémantický zákon, atď., atď.

Diskrétny dvojrozmerný prístup

Rovnakým spôsobom ako v spojitom prípade, môžeme zovšeobecniť aj diskretný prípad na niekoľko premenných. Pretože doterajšia lingvistická analýza v tomto smere je málo početná (článok Uhlírovej a Wimmera a článok o slabikovej štruktúre od Zörniga and Altmanna), ukážeme len metódu.

V jednorozmernom diskretnom prístupe sme mali rekurentné vzorce (napr. (2.12) alebo (2.13)) ktoré sa môžu písať ako

$$P_x = g(x)P_{x-1}, \quad (2.15)$$

kde $g(x)$ bola časť nekonečnej postupnosti. Pretože tentokrát máme dve premenné, môžeme sformulovať model nasledovne

$$P_{i,j} = g(i,j)P_{i,j-1}, \quad P_{i,j} = h(i,j)P_{i-1,j}, \quad (2.16)$$

kde $g(i,j)$ a $h(i,j)$ sú rôzne funkcie i a j . Rovnice sa musia riešiť súčasne. Výsledok samozrejme závisí od zvolených funkcií $g(i,j)$ a $h(i,j)$. Takto Uhlířová a Wimmer dostali dvojrozmerné binomické rozdelenie, kým Zörnig a Altmann dostali dvojrozmerné Conwayovo-Maxwellovo-Poissonovo rozdelenie.

Záver

Skutočnosť, že týmto spôsobom môžeme integrovať rôzne hypotézy, má niekoľko dôsledkov:

- (i) Ukazuje sa, že v pozadí mnohých dejov v lingvistike existuje jednotný mechanizmus – reprezentovaný vzťahmi (2.1), (2.5), (2.11), (2.16). V rámci tohto mechanizmu môžeme kombinovať premenné a „sily“.
- (ii) Vzorce (2.1), (2.5), (2.11), (2.14) predstavujú systémy ktoré môžu obsahovať aj mimosystémové faktory.

- (iii) Tento prístup dovoľuje indukzívne testovať nové, doteraz neznáme vzťahy a systemizovať ich do teórie s korektnou interpretáciou faktorov; toto obyčajne nie je možné ak postupujeme indukzívne. Exploratívna časť práce sa preto môže urýchliť vhodným softvérom. Nedá sa predpokladať, že použitím tohto prístupu budeme môcť všetko v jazyku vysvetliť, ale môžeme uspokojivo zjednotiť a interpretovať aposteriórne mnoho rôznorodých javov.

Literatúra

- [1] Altmann, G. (1990). *Bühler or Zipf? A re-interpretation*. In: Koch, W.A. (Hrsg.), *Aspekte einer Kultursemiotik*: 1-6. Bochum: Brockmeyer.
- [2] Altmann, G., Köhler, R. (1996). „*Language Forces*“ and *synergetic modeling of language phenomena*. *Glottometrika* 15, 62-76.
- [3] Baayen, R.H. (1989). *A corpus-based approach to morphological productivity*. Amsterdam: Centrum voor Wiskunde en Informatica.
- [4] Balasubrahmanyam, V.K., Naranan, S. (1997). *Quantitative linguistics and complex system studies*. *Journal of Quantitative Linguistics* 3, 177-228.
- [5] Boroda, M.G., Altmann, G. (1991). *Menzerath's law in musical texts*. *Musikometrika* 3, 1-13. Bunge, M. (1983). *Understanding the world*. Dordrecht: Reidel.
- [6] Chitashvili, R.J., Baayen, R.H. (1993). *Word frequency distributions of texts and corpora as large number of rare event distributions*. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative Text Analysis*: 54-135. Trier: WVT.
- [7] Geršić, S., Altmann, G. (1988). *Ein Modell für die Variabilität der Vokaldauer*. *Glottometrika* 9, 49-58.
- [8] Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.
- [9] Köhler, R. (1986). *Zur linguistischen Synergetik*. Struktur und Dynamik der Lexik. Bochum: Brockmeyer.
- [10] Köhler, R. (1987). *Systems theoretical linguistics*. *Theoretical Linguistics* 14, 241-257.

- [11] Köhler, R. (1989). *Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation*. Glottometrika 11, 1-18.
- [12] Köhler, R. (1990). *Elemente der synergetischen Linguistik*. Glottometrika 12, 179-187.
- [13] Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen*. Bochum: Brockmeyer
- [14] Rapoport, A. (1982). *Zipf's law re-visited*. In: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf's Law*: 1-28. Bochum: Brockmeyer.
- [15] Schroeder, M. (1990). *Fractals, chaos, power laws. Minutes from an infinite paradise*. New York: Freeman.
- [16] Uhlířová, L., Wimmer, G. (2003). *A contribution to word length theory*. In: Festschrift für Werner Lehfeldt zum 60. Geburtstag (Kempgen, S., Schweier, V., Berger, T. (Eds.)). München: Verlag Otto Sagner, 524-530.
- [17] Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). *Towards a theory of word length distribution*. Journal of Quantitative Linguistics 1, 98-106.
- [18] Wimmer, G., Altmann, G. (1995) *A model of morphological productivity*. Journal of Quantitative Linguistics 2, 212-216.
- [19] Wimmer, G., Altmann, G. (1996). *The theory of word length: Some results and generalizations*. Glottometrika 15, 112-133.
- [20] Wimmer, G., Altmann, G. (1999) *Rozdelenie polysemie v maorijčine* (Distribution of Polysemy in Maori), Pange Lingua; Genzor, J., Ondrejovič, S. (Eds.). Bratislava: Veda, 17-25.
- [21] Wimmer, G., Altmann, G. (2005). *Unified derivation of some linguistic laws*. In: Quantitative Linguistics. An International Handbook. (Köhler, R., Altmann, G., Piotrowski, R., G. (Eds.)). Berlin: Walter de Gruyter, 791-807.
- [22] Zörnig, P., Altmann, G. (1993). *A model for the distribution of syllable types*. Glottometrika 14, 190-196.
- [23] Zörnig, P., Boroda, M.G. *The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent texts*. Glottometrika 13, 205-218.

METODA LATINSKÝCH ČTVERCŮ VE SPOLEHLIVOSTNÍCH ÚLOHÁCH

Jan Záruba a Daniela Jarušková

Abstrakt: V mnoha úlohách z oblasti spolehlivosti je možno odhadnout pravděpodobnost poruchy pomocí metod Monte Carlo. Cílem tohoto příspěvku je porovnat rozptyly odhadů získaných pomocí prosté metody Monte Carlo s rozptyly odhadů získaných pomocí metody latinských čtverců.

Klíčová slova: Metody Monte Carlo, metoda latinských čtverců, odhad pravděpodobnosti poruchy.

Abstract: The Monte Carlo methods may be applied to get an estimate of probability of failure in many reliability problems. The aim of this paper is to compare variances of estimates of failure probability obtained by the brute Monte Carlo method and by the Latin hypercube sampling.

1. Úvod

Jednou z nejčastějších úloh, která se řeší metodou Monte Carlo je odhad střední hodnoty veličiny Z , která je známou funkcí jiných náhodných veličin $X_1, \dots, X_k$, jejichž rozdělení je rovněž známé. V následujícím budeme předpokládat, že veličiny $X_1, \dots, X_k$ jsou nezávislé. Vzhledem k tomu, že distribuční funkce veličin $X_1, \dots, X_k$ jsou známé, můžeme $X_1, \dots, X_k$ transformovat na veličiny, které mají rovnoměrné rozdělení na intervalu $(0, 1)$. Bez újmy na obecnosti budeme proto dále předpokládat, že veličiny $X_1, \dots, X_k$ mají rovnoměrné rozdělení na intervalu $(0, 1)$.

Při použití prosté metody Monte Carlo (brute Monte Carlo) se generují nezávislé realizace vektoru $\{(X_{i1}, \dots, X_{ik}), i = 1, \dots, n\}$, pro jednotlivé realizace vstupů se počítají hodnoty výstupu $\{Z_i = h(X_{i1}, \dots, X_{ik}), i = 1, \dots, n\}$, (které jsou samozřejmě vzájemně nezávislé) a nakonec se odhadne střední hodnota

$$\mu = EZ = E(h(X_1, \dots, X_k))$$

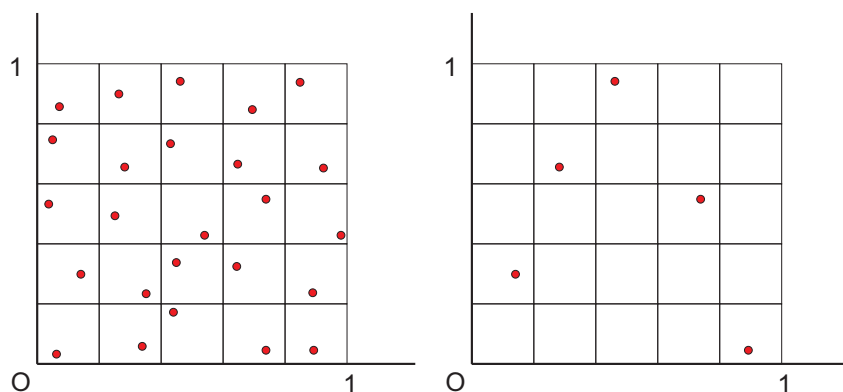
pomocí průměru

$$\hat{\mu} = \bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i.$$

Pokud je počet obdržených výstupů n nízký, může být získaný odhad značně nepřesný; jinými slovy odhad $\hat{\mu}$ má velký rozptyl. Taková situace může nastat

například tehdy, když je výpočet funkce $h(x_1, \dots, x_k)$ časově náročný, takže můžeme získat jen malý počet výstupů $\{Z_i\}$.

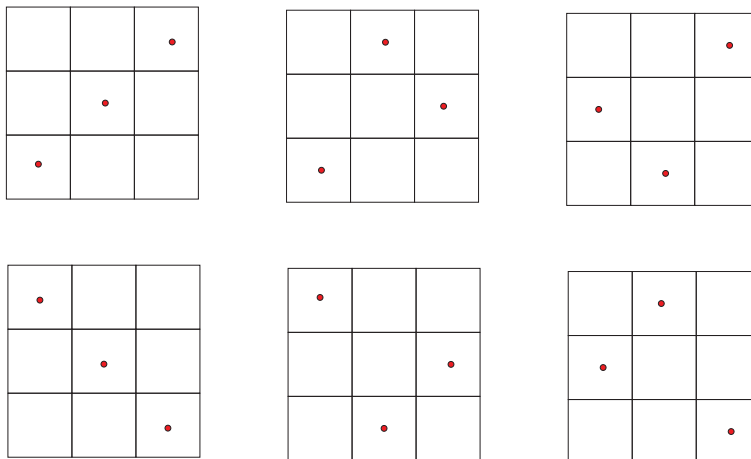
Metoda latinských čtverců je jednou z metod, které pomáhají rozptyl snižovat. Vychází z metody stratifikovaného výběru (metody vrstev). Metoda vrstev spočívá v rozdělení každého intervalu $(0, 1)$, $i = 1, \dots, k$ do N podintervalů: $(0, \frac{1}{N}), \dots, (\frac{N-1}{N}, 1)$. Zřejmě $P(X_i \in (\frac{i-1}{N}, \frac{i}{N})) = \frac{1}{N}$ pro $i = 1, \dots, k$ a $j = 1, \dots, N$. Nadkrychle $(0, 1) \times \dots \times (0, 1)$ je tak rozdělena do N^k nadkrychliček. Zřejmě $P(X_1 \in (\frac{j_1-1}{N}, \frac{j_1}{N}), \dots, X_k \in (\frac{j_k-1}{N}, \frac{j_k}{N})) = \frac{1}{N^k}$ pro $j_1 = 1, \dots, N, \dots, j_k = 1, \dots, N$. Realizace náhodných veličin jsou pak generovány z jednotlivých nadkrychliček, a to buď z každé nadkrychličky jedna realizace, nebo může být rozsah výběru snížen tím, že vybíráme jen z menšího počtu nadkrychliček. Metoda latinských čtverců spočívá ve speciálním výběru nadkrychliček. Nadkrychličky zde nevolíme zcela náhodně, nýbrž tak, že v každé vrstvě budeme vybírat pouze jednu. Rozdíl mezi generováním pomocí metody vrstev a pomocí latinských čtverců ilustrují následující obrázky, odpovídající situaci se dvěma vstupními náhodnými veličinami.



Všimněme si toho, že zatímco v případě metody vrstev vybíráme z 25 čtverečků, v případě latinských čtverců z 5 čtverečků. Náhoda zde vystupuje:

- ve výběru, ve kterých čtverečkách budeme generovat,
- v rovnoměrném (náhodném) výběru hodnoty uvnitř vybraného čtverečku.

Pro ilustraci ukažme, jaké jsou možnosti, jestliže chceme získat tři realizace vektoru o dvou složkách.



Praktický postup pro generování dvou proměnných:

- Rozhodneme, kolik chceme získat simulací. (Chceme-li získat například 5 simulací, rozdělíme čtverec $(0, 1) \times (0, 1)$ na 25 čtverečků, které můžeme označit například následovně: $(1, 1), (1, 2), \dots, (5, 5)$.)
- Vygenerujeme dvě náhodné permutace čísel $\{1, 2, 3, 4, 5\}$. Jestliže například dostaneme:

2 4
3 5
4 3
1 2
5 1,

pak budeme vybírat ze čtverečků: $(2, 4), (3, 5), (4, 3), (1, 2), (5, 1)$.

- Vybereme rovnoměrně náhodně hodnotu (x, y) uvnitř vybraných pěti čtverečků.

Je zřejmé, že v případě k veličin je třeba generovat k náhodných permutací. (Přesněji řečeno stačí generovat $k - 1$ náhodných permutací.)

2. Vlastnosti metody latinských čtverců

Metoda latinských čtverců byla navržena autory McKayem, Beckmanem a Conoverem. Ve svém článku, viz McKay a spol. (1979), autoři ukazují, že odhad pomocí metody latinských čtverců má menší rozptyl v případě, že funkce $h(x_1, \dots, x_k)$ je monotónní v každé proměnné. Tvrzení je důsledkem toho, že v případě monotónnosti funkce h jsou hodnoty Z_i^{LHS} , $i = 1, \dots, n$ získané použitím metody latinských čtverců záporně zkorelované, a tudíž je rozptyl jejich průměrů menší než v případě nezávislých hodnot Z_i^{BMC} , $i = 1, \dots, n$, které získáváme, generujeme-li prostou metodou Monte Carlo. V důkazu se používají výsledky získané Lehmannem (1966) o různých druzích závislostí (kvadrantová závislost a korelovanost).

Stein (1987) ukázal, že pokud $\int h^2(x_1, \dots, x_k) dF(x_1, \dots, x_k) < \infty$, pak lze funkci h rozložit následovně:

$$h(x_1, \dots, x_k) = \mu + \sum_{j=1}^k \alpha_j(x_j) + r(x_1, \dots, x_k),$$

kde $\mu = \int h(x_1, \dots, x_k) dF(x_1, \dots, x_k)$ je hledaný integrál,

$$\alpha_j(x_j) = \int (h(x_1, \dots, x_k) - \mu) dF_{-j}(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k)$$

je hlavní efekt a $r(x_1, \dots, x_k)$ je reziduální člen. (V předchozím jsme použili značení $F(x_1, \dots, x_k) = \prod_{i=1}^k F_i(x_i)$, $F_{-j}(x_1, \dots, x_k) = \prod_{i \neq j} F_i(x_i)$, přičemž $F_i(x_i)$ je distribuční funkce náhodné veličiny X_i). Stein (1987) ukazuje, že pro $n \rightarrow \infty$ pro metodu latinských čtverců platí:

$$var_{LHS} \bar{Z} = \frac{1}{n} \int (r(x_1, \dots, x_k))^2 dF(x_1, \dots, x_k) + o\left(\frac{1}{n}\right),$$

zatímco pro prostou metodu Monte Carlo platí:

$$var_{BMC} \bar{Z} = \frac{1}{n} \int (r(x_1, \dots, x_k))^2 dF(x_1, \dots, x_k) + \frac{1}{n} \sum_{j=1}^k \int (\alpha_j(x_j))^2 dF_j(x_j).$$

Steinovo tvrzení umožňuje v některých případech odhadnout, o kolik procent je metoda latinských čtverců asymptoticky vydatnější, tj. má menší rozptyl.

3. Metoda latinských čtverců v problémech určování spolehlivosti

V následující kapitole se chceme zabývat problémem, zda a do jaké míry je metoda latinských čtverců vhodná pro úlohy z teorie spolehlivosti.

Teorie spolehlivosti (např. stavebních konstrukcí) vychází ze základního vztahu účinku zatížení konstrukce a její schopnosti odolávat takovému zatížení. Označme Q celkový účinek zatížení na konstrukci a R odolnost konstrukce. Pro bezpečný stav konstrukce musí platit

$$Q < R .$$

Kritický rovnovážný stav nastane právě, když

$$R - Q = 0 .$$

Cílem je stanovit pravděpodobnost poruchy konstrukce p_p ze vztahu

$$p_p = P(Q \geq R) .$$

V případě, že celkový účinek zatížení konstrukce Q i odolnost R jsou nezávislé veličiny s normálním rozdělením, lze pravděpodobnost poruchy explicitně vypočítat, protože rezerva spolehlivosti W , definovaná rozdílem $W = R - Q$, má také normální rozdělení s parametry:

$$\mu_W = \mu_R - \mu_Q ,$$

$$\sigma_W^2 = \sigma_R^2 + \sigma_Q^2 .$$

Pravděpodobnost poruchy je zde dána vztahem

$$p_p = P(Q \geq R) = (W \leq 0) = \Phi \left(\frac{\mu_R - \mu_Q}{\sqrt{\sigma_R^2 + \sigma_Q^2}} \right) .$$

Následující jednoduchý příklad může sloužit k provnání prosté metody Monte Carlo s metodou latinských čtverců. Předpokládejme, že odolnost $R \sim N(55, 6^2)[-]$, a účinek zatížení konstrukce $Q \sim N(40, 5^2)[-]$. Nejdříve spočtíme analytické řešení:

$$\mu_W = \mu_R - \mu_Q = 55 - 40 = 15 ,$$

$$\sigma_W^2 = \sigma_R^2 + \sigma_Q^2 = 6^2 + 5^2 = 61 .$$

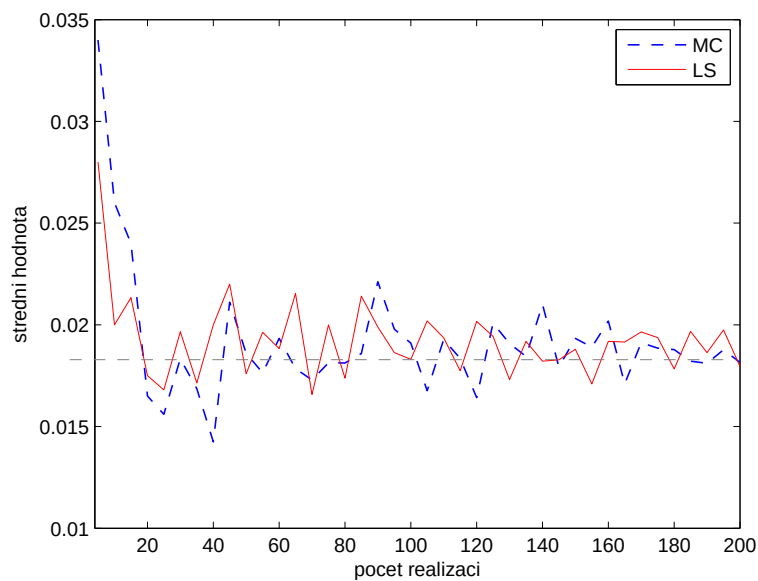
Pravděpodobnost poruchy je

$$p_p = P(Q \geq R) = (W \geq 0) = \Phi_W(0) = 0.0188 .$$

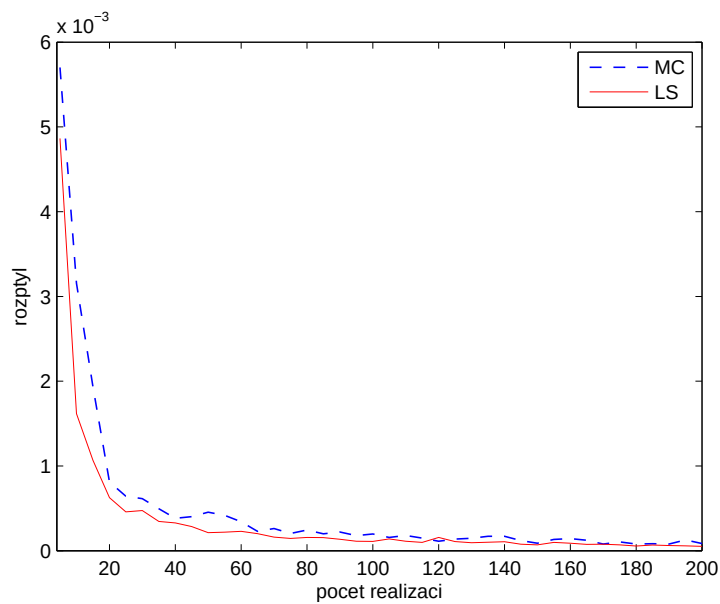
Nyní budeme stejnou úlohu řešit prostou metodou Monte Carlo a metodou latinských čtverců. Především nás bude zajímat vliv počtu realizací n na přesnost odhadu poruchy.

Pro daný pevný počet simulací n jsme opakovaně počítali odhady pravděpodobnosti poruchy jak prostou metodou Monte Carlo, tak i metodou latinských čtverců. Poté, co jsme z použitím každé z těchto metod získali soubor o sta odhadech, spočetli jsme z nich průměr a rozptyl.

Následující obrázek ukazuje rychlost konvergence odhadů (průměrů) ke správné hodnotě.



Vliv počtu realizací na rozptyl odhadu je ilustrován následujícím obrázkem.



Následující dvě úlohy ilustrují, do jaké míry se sníží rozptyl, použijeme-li metodu latinských čtverců.

Prostý nosník

Předpokládejme, že máme k dispozici nosník zatížený uprostřed koncentrovanou silou P [kN]. Délka nosníku je L [m]. Ohybová momentová kapacita nosníku je $W \cdot T$, kde W [m^3] je plastický průřezový modul a T [kPa] je napětí na mezi plasticity. Všechny veličiny jsou nezávislé, normálně rozdělené, přičemž:

$$P \sim N(10, 4), \quad L \sim N(8, 0.01), \quad W \sim N(100 \cdot 10^{-6}, 400 \cdot 10^{-12}), \\ T \sim N(600 \cdot 10^3, 10 \cdot 10^9).$$

K porušení nosníku dojde v situaci, kdy platí:

$$\frac{P \cdot L}{4} > W \cdot T.$$

Hodnotu pravděpodobnosti, s jakou dojde k poruše nosníku, odhadneme ze 100, 1000 a 10000 simulací. Experiment 100× zopakujeme a z obdržených

hodnot vypočteme střední hodnotu a rozptyl odhadu. Pro simulování použijeme prostou metodu Monte Carlo a metodu latinských čtverců. Výsledné hodnoty porovnáme.

Pro 100 realizací:

<i>metoda</i>	μ	σ^2	$\frac{\sigma}{\mu}$
<i>Monte Carlo</i>	0.0024	$2.4485 \cdot 10^{-5}$	2.0618
<i>LS</i>	0.0017	$1.4253 \cdot 10^{-5}$	2.2207

Pro 1000 realizací:

<i>metoda</i>	μ	σ^2	$\frac{\sigma}{\mu}$
<i>Monte Carlo</i>	0.00235	$2.5126 \cdot 10^{-6}$	0.6745
<i>LS</i>	0.00222	$1.5067 \cdot 10^{-6}$	0.5529

Pro 10000 realizací:

<i>metoda</i>	μ	σ^2	$\frac{\sigma}{\mu}$
<i>Monte Carlo</i>	0.002196	$1.9918 \cdot 10^{-7}$	0.2032
<i>LS</i>	0.002173	$1.5351 \cdot 10^{-7}$	0.1803

Metodou latinských čtverců se nám podařilo snížit rozptyl, a to následovně:

realizace	snížení σ^2 o:
100	41.8%
1000	40.0%
10000	22.9%

Železo – betonový nosník (průvlak)

Je známo, že k poruše železo – betonového nosníku dojde právě když platí $M < Z_1$,

kde Z_1 je ohybový moment definovaný $Z_1 \sim N(0.01, 0.003^2)[MNm]$.

Proti němu působí moment únosnosti průřezu M , který je definován:

$$M = Z_2 \cdot Z_3 \cdot Z_4 \cdot - \frac{Z_5 \cdot Z_3^2 \cdot Z_4^2}{Z_6 \cdot Z_7} ,$$

kde Z_2 je efektivní umístění výztuže $\sim N(0.30, 0.015^2)[m]$,

Z_3 je napětí na mezi plasticity $\sim N(360, 36^2)[MPa]$,

Z_4 je plocha průřezu výztuže $\sim N(226 \cdot 10^{-6}, 11.3^2 \cdot 10^{-12})[m^2]$,
 Z_5 je faktor vztažený k pracovnímu diagramu $\sim N(0.5, 0.05^2)[-]$,
 Z_6 je šířka nosníku $\sim N(0.12, 0.006^2)[m]$,
 Z_7 je maximální napětí betonu v tlaku $\sim N(40, 6^2)[MPa]$.

Opět jako v předchozím příkladu experiment $100 \times$ zopakujeme pro 100, 1000 a 10000 realizací výpočtu.

Výsledky: Pro 100 realizací:

<i>metoda</i>	μ	σ^2	$\frac{\sigma}{\mu}$
<i>Monte Carlo</i>	0.0007	$6.5758 \cdot 10^{-6}$	3.6633
<i>LS</i>	0.0005	$4.7980 \cdot 10^{-6}$	4.3809

Pro 1000 realizací:

<i>metoda</i>	μ	σ^2	$\frac{\sigma}{\mu}$
<i>Monte Carlo</i>	0.00052	$7.7737 \cdot 10^{-7}$	1.6956
<i>LS</i>	0.00047	$4.7380 \cdot 10^{-7}$	1.4646

Pro 10000 realizací:

<i>metoda</i>	μ	σ^2	$\frac{\sigma}{\mu}$
<i>Monte Carlo</i>	0.000475	$5.3409 \cdot 10^{-8}$	0.4865
<i>LS</i>	0.000455	$4.7551 \cdot 10^{-8}$	0.4793

Metodou latinských čtverců se nám podařilo snížit rozptyl a to:

realizace	snížení σ^2 o:
100	27.0%
1000	39.1%
10000	11.0%

4. Odhad vydatnosti pro některé jednoduché problémy

Na základě simulací, případně úvah založených na Steinově tvrzení, bychom rádi uvedli několik závěrů týkajících se vydatnosti (eficience) metody latinských čtverců. Pod pojmem vydatnost (eficience) zde rozumíme poměr mezi rozptylem odhadu získaného metodou latinských čtverců a rozptylem odhadu získaného prostou metodou Monte Carlo. Všechny odhady vydatnosti v následujících tabulkách byly odhadnuty na základě 100 000 simulací.

Odhadujeme-li pravděpodobnost $P(X + Y > C)$ pro X a Y se standardním normálním rozdělením, pak je metoda latinských čtverců tím méně eficientní, čím je hodnota C dále od 0. Následující tabulka uvádí pro ilustraci vydatnosti pro odhad pravděpodobnosti $P(X + Y > \sqrt{2}C)$ pro různé hodnoty C .

C	0	0.4	0.8	1.0	1.4	1.8	2.0	2.4	2.8	3.0
$\frac{var_{LHS}}{var_{BMC}}$	0.333	0.352	0.407	0.442	0.533	0.635	0.690	0.780	0.851	0.902

Vydatnost odhadu $P(X + Y > \sqrt{2}C)$ metodou latinských čtverců.

Odhadujeme-li pravděpodobnost $P\left(\frac{X_1 + \dots + X_k}{\sqrt{k}} > C\right)$, kde $X_i, i = 1, \dots, k$ jsou nezávislé náhodné veličiny rozdělené podle standardního normálního rozdělení, pak při daném C a daném počtu simulací n se eficeience metody latinských čtverců snižuje s počtem proměnných k . Následující tabulky ukazují, jak se mění vydatnost s rostoucím k pro $C = 1, 2, 3$.

k	2	4	6	8	10
$\frac{var_{LHS}}{var_{BMC}}$	0.44	0.51	0.53	0.54	0.54

Vydatnost odhadu $P\left(\frac{X_1 + \dots + X_k}{\sqrt{k}} > 1\right)$ metodou latinských čtverců.

k	2	4	6	8	10
$\frac{var_{LHS}}{var_{BMC}}$	0.68	0.78	0.81	0.83	0.84

Vydatnost odhadu $P\left(\frac{X_1 + \dots + X_k}{\sqrt{k}} > 2\right)$ metodou latinských čtverců.

k	2	4	6	8	10
$\frac{var_{LHS}}{var_{BMC}}$	0.90	0.96	0.97	0.98	0.99

Vydatnost odhadu $P\left(\frac{X_1 + \dots + X_k}{\sqrt{k}} > 3\right)$ metodou latinských čtverců.

Odhadujeme-li pravděpodobnost $P(aX_1 + X_2 > C)$, kde X_1 a X_2 jsou nezávislé veličiny rozdělené podle standardního normálního rozdělení, pak je metoda latinských čtverců při daném C a daném n tím vydatnější, čím je a větší. Následující tabulky uvádějí vydatnosti pro odhad pravděpodobnosti $P(aX_1 + X_2 > C\sqrt{a^2 + 1})$ pro $C = 1$ a $C = 2$ pro různé hodnoty a .

a	1	2	3	4	5
$\frac{var_{LHS}}{var_{BMC}}$	0.67	0.62	0.53	0.46	0.42

Vydatnost odhadu $P(aX + Y > \sqrt{a^2 + 1})$ metodou latinských čtverců.

a	1	2	3	4	5
$\frac{var_{LHS}}{var_{BMC}}$	0.84	0.74	0.64	0.57	0.50

Vydatnost odhadu $P(aX + Y > 2\sqrt{a^2 + 1})$ metodou latinských čtverců.

5. Závěr

Obecně platí, že pokud je pravděpodobnost poruchy monotónní funkcí sledovaných veličin, pak metoda latinských čtverců poskytuje odhad s nižším rozptylem, a tedy je přesnější. Rozdíl v rozptylech odhadů se však zmenšuje se vzrůstajícím počtem simulací. Vzhledem k tomu, že pro rozumný odhad malých pravděpodobností poruchy je třeba provést poměrně značný počet simulací, pak zde není metoda latinských čtverců tak výhodná jako je při odhadování jiných charakteristik. Generování náhodných čísel pomocí metody latinských čtverců je totiž časově náročnější než prostou metodou Monte Carlo. Metoda je vhodná tedy především tam, kde je výpočet hodnot Z_i , $i = 1, \dots, n$ obtížný, a tedy časově náročný.

Reference

- [1] E. L. Lehmann (1966): Some concepts of dependence. *Ann. Math. Statist.*, **37**, No 5, 1137–1153.
- [2] M. D. McKay, R. J. Beckman and W. J. Conover (1979): A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, No 2, 239–245.
- [3] A. B. Owen (1992): A central limit theorem for latin hypercube sampling. *Jour. of Royal Stat. Soc., Series B.* **54**, No 2, 541–551.
- [4] M. Stein (1999): Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, **29**, 143–151.

JAK NA ROZHODOVACÍ STROMY

Marta Žambochová

Abstract: The tree structure is a popular instrument of the information presentation in many spheres of common life. It finds its use in data analysis on account of its simplicity and its clarity. The expanded group of trees in data modelling and simulation is the group of assorted decision trees. We can solve the classification and prediction tasks by means of decision trees. The paper deals with a comparison of some algorithms for a creation of decision trees. This article shows a way how and why to include decisions trees in an education on faculties of economics. It shows using the trees for finding the resolution of some economic problems.

Key words: The decision tree, CART, QUEST, an education on faculties of economics, facultative subjects.

Úvod

Velmi rozšířenou skupinou stromů, kterých se využívá v datových modelech, jsou různé typy rozhodovacích stromů. Rozhodovací stromy jsou struktury, které rekurzivně rozdělují zkoumaná data dle určitých rozhodovacích kritérií. Kořen stromu reprezentuje celý populační soubor. Vnitřní uzly stromu reprezentují podmnožiny populačního souboru. V listech stromu můžeme vyčíst hodnoty vysvětlované proměnné.

Rozhodovací strom se vytváří rekurzivně dělením prostoru hodnot prediktorů. Máme-li strom s jedním listem, hledáme otázku (podmínku větvení), která nejlépe rozděljuje prostor zkoumaných dat do podmnožin. Takto nám vznikne strom s více listy. Nyní pro každý nový list hledáme otázku, která množinu dat náležící tomuto listu co nejlépe dělí do podmnožin.

Proces dělení se zastaví, pokud bude splněno kritérium pro zastavení. Omezení obsažená v kritériu pro zastavení mohou být např. „hloubka“ stromu, počet listů stromu, stupeň homogenosti množin dat v listech, ...

Dalším krokem algoritmů je prořezávání stromu (prunning). Je nutno určit „správnou“ velikost stromu (příliš malé stromy dostatečně nevystihují všechny zákonitosti v datech, příliš velké stromy zahrnují do popisu i nahodilé vlastnosti dat). Vygenerují se podstromy stromu vzniklého algoritmem a porovnává se jejich kvalita generalizace (jak dobře vystihují data).

Postup může být takový, že se rozhodovací stromy nejdříve vytváří na tzv. trénovacích datech a poté se jejich kvalita ověří na tzv. testovacích datech.

Jiným způsobem je křížová validace (cross validation), kdy k vytváření stromu a jeho podstromů použijí všechna data. Poté se data rozdělí na několik disjunktních, přibližně stejně velkých částí a postupně se vždy jedna část dat ze souboru vyjme. Pomocí vzniklých souborů dat se ověřuje kvalita stromu a jeho podstromů.

Vybere se takový podstrom, který má nejnižší odhad skutečné chyby. Pokud existuje více podstromů se srovnatelným odhadem skutečné chyby, vybírá se ten nejmenší.

Jednotlivé algoritmy vytváření rozhodovacích stromů se liší následnými charakteristikami:

- pravidlo dělení (splitting rule)
- kritérium pro zastavení (stopping rule)
- typ podmínek větvení
 - multivariantní (testuje se několik prediktorů)
 - univariantní (v daném kroku se testuje pouze jeden z prediktorů)
- způsob větvení
 - binární (každý z uzlů, kromě listů, se dělí na dva následníky)
 - k-ární (některý z uzlů se dělí na více než dvě části)
- typ výsledného stromu, popis obsahu listů
 - klasifikační stromy (v každém listu je přiřazení třídy)
 - regresní stromy (v každém listu je přiřazení konstanty – odhad hodnoty závislé proměnné)
- typ prediktorů
 - kategoriální
 - ordinální

Algoritmy pro vytváření rozhodovacích stromů

Pro vytváření rozhodovacích stromů bylo vyvinuto velké množství algoritmů. Nejvíce používané jsou CART (L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, 1984), ID3 (J. R. Quinlan, 1975), C4.5 (J. R. Quinlan, 1993), AID (J. N. Morgan a J. A. Sonquist, 1963), CHAID (G. V. Kass, 1980) a QUEST (W. Y. Loh and Y. S. Shih, 1997).

Ve článku budou stručně zmíněny dva ze jmenovaných algoritmů, CART a QUEST, které jsou základem algoritmů v SW produktu STATISTICA, jenž jsou použity ke zpracování dat v motivačním příkladu uvedeném ve článku.

Algoritmus CART

Algoritmus poprvé popsali autoři L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone v roce 1984 ve článku „Classification and Regression trees“.

Algoritmus je použitelný v případě, že máme jednu nebo více nezávislých proměnných. Tyto proměnné mohou být buď spojité nebo kategoriální (ordinální i nominální). Dále máme jednu závislou proměnnou, která také může být kategoriální (nominální i ordinální) nebo spojitá.

Výsledkem jsou binární stromy, protože jsou zde přípustné pouze otázky (podmínky dělení), na které je možno odpovědět ano/ne (Je věk menší než 30 let? Je pohlaví mužské? ...).

Algoritmus dělení je různý pro klasifikační stromy a pro stromy regresní.

Klasifikační stromy používáme v případě, že je závislá proměnná kategoriální. To znamená, že se soubor původních dat snažíme v závislosti na nezávislých proměnných rozdělit do skupin, přičemž, v ideálním případě, každá skupina má přiřazení ke stejné kategorii závislé proměnné.

Homogenita uzlů-potomků je měřena pomocí tzv. funkce znečištění (impurity function) $i(t)$. Maximální homogenita vzniklých dvou potomků je počítána jako maximální změna (snížení) znečištění $\Delta i(t)$.

Algoritmus CART řeší pro každý uzel maximalizační problém pro funkci $\Delta i(t)$ přes všechna možná dělení uzlu, to znamená, že hledá dělení, které přináší maximální zlepšení homogenity dat.

Regresní stromy se používají v případě, že závislá proměnná není kategoriální. Každá její hodnota může být v obecnosti různá.

V tomto případě algoritmus hledá nejlepší dělení na základě minimalizace součtu rozptylů v rámci jednotlivých dvou vzniklých uzlů-potomků. Algoritmus pracuje na základě algoritmu minimalizace součtu čtverců.

Algoritmus QUEST

Metoda je popsána ve článku z roku 1997 W. Y. Loh and Y. S. Shih: „Split selection methods for classification trees“. Algoritmus je použitelný pouze pro nominální závislou proměnnou. Obdobně, jako v případě CART, jsou vytvářeny binární stromy. Na rozdíl od metody CART, provádí metoda QUEST v průběhu budování stromu odděleně výběr proměnné pro štěpení uzlu a výběr dělicího bodu. Metoda QUEST (for Quick, Unbiased, Efficient, Statistical Tree) odstraňuje některé nevýhody algoritmů používajících vyčerpávající hledání (např. CART), jako je náročnost zpracování, snížení obecnosti výsledku, a podobně.

Tato metoda je vylepšením algoritmu FACT, který popsali v roce 1988 autoři W. Y. Loh a N. Vanichsetakul. V prvním kroku algoritmus převede všechny kategoriální nezávislé proměnné na „ordinální“ pomocí CRIMCO-ORD transformace.

Dále v každém listovém uzlu, je pro každou proměnnou prováděn ANOVA F-test. Pokud největší ze vzniklých F-statistik je větší než předem daná hodnota F_0 , pak příslušná proměnná je vybrána pro dělení uzlu. Pokud tomu tak není, je pro všechny proměnné proveden Levenův F-test. Pokud je největší Levenova F-statistika větší než F_0 , pak je pro dělení uzlu vybrána tato proměnná. Jinak (tzn. není žádná ANOVA F-statistika ani Levenova F-statistika větší než F_0) je pro dělení vybrána proměnná s největší ANOVA F-statistikou. Pro dělení uzlu je tedy vybrána ten prediktor, který je se závislou proměnnou nejvíce asociován.

Pro hledání dělícího bodu pro vybranou nezávislou proměnnou je využívána metoda Kvadratické diskriminační analýzy (QDA), na rozdíl od algoritmu FACT, kde je využívána metoda Lineární diskriminační analýzy (LDA).

Postup je rekurzivně opakován až do zastavení (na základě kritéria pro zastavení).

SW pro zpracování rozhodovacích stromů

- velké statistické SW balíky
 - výhody
 - * přehlednost
 - * dobrá dostupnost
 - * relativně dobrá dokumentace
 - nevýhody
 - * vysoké pořizovací náklady
 - * nutno kupovat několik modulů
 - * velké nároky na HW
 - * nepřesný popis použitých metod
- ostatní (komerční a nekomerční)
 - výhody
 - * malé (resp. žádné) pořizovací náklady
 - * je možno koupit samostatně modul na tvorbu rozhodovacích stromů

- * relativně malé nároky na HW
- nevýhody
 - * mnohdy nedostatečná dokumentace
 - * většinou chybí jakákoliv zmínka o použitých metodách

Motivační příklad

Studovali jsme vzorek studentů střední a vysoké školy a zkoumali jsme, zda lze z určitých hledisek životního stylu vyvodit váhovou kategorii získanou na základě BMI (Body Mass Index) vypočítaného ze zjištěné váhy a výšky osoby.

Ze sledovaných položek jsme za nezávislé proměnné vybrali počet hodin denně strávených u počítače a u televize, průměrný počet hodin sportu za týden, průměrný počet hodin spánku, průměrný počet jídel během dne, převažující druh stravování (fast food, stravování v jídelně resp. restauraci, domácí strava, studená kuchyně).

Jako závislou proměnnou jsme zvolili kategoriální proměnnou nabývající hodnot „podváha“, „normální váha“, „nadváha“ a „obezita“.

Ve statistickém SW STATISTICA jsme použili různé možnosti sestavení rozhodovacího stromu, vypovídajícího o struktuře sledovaného vzorku studentů. Jednak jsme vytvořili klasifikační strom pomocí algoritmu C&RT vyčerpávajícího prohledávání (viz obr. 1, tab. 1), jednak pomocí metody založené na principu QUEST (viz obr. 2, tab. 2). Dále jsme vytvořili strom pomocí standardní metody C&RT z modulu Data-Mining, včetně V-fold Crossvalidation – metody na výběr neoptimálnějšího stromu (viz obr. 3, tab. 3).

- Global CV cost = 0,13636; s.d. CV cost = 0,03272 (vyčerpávající C&RT)
- Global CV cost = 0,22727; s.d. CV cost = 0,03996 (QUEST)
- Global CV cost = 0,081818; s.d. CV cost = 0,026133 (standard C&RT)

Z tohoto hlediska se tedy jeví jako neoptimálnější strom vytvořený posledním způsobem.

Pokud rozhodovací strom převedeme na pravidla, pak se dostáváme k závěru, že největší vliv (ze sledovaných hledisek) na váhovou kategorii má druh stravy. Nejvíce ohroženy jsou osoby stravující se ve „fast foodech“ a jídelnách, resp. restauracích.

Ve skupině osob ohrožených vysokou hmotností se oddělují tři skupiny, a to na základě množství sportu. Sport je tedy druhým faktorem ovlivňujícím váhu osob. Osoby trpící obezitou se projevují velkým nedostatkem sportu.

Osoby s nadváhou sportují pouze málo a osoby s normální vahou mají nejvyšší intenzitu sportování.

Skupina stravujících se jiným způsobem se dělí na dvě skupiny odlišné množstvím spánku. Obě tyto skupiny se dále dělí na základě počtu denních jídel. Délka spánku a počet denních jídel tedy také ovlivňují váhovou kategorii.

Ze struktury stromu je zřejmé, jak se výše zmíněné faktory projevují na zařazení osob do váhové kategorie.

Zařazení problematiky rozhodovacích stromů do výuky

V posledních letech se většina ekonomicky zaměřených vysokých škol potýká s problémem snižování hodinových dotací tzv. kvantitativních předmětů (matematika, statistika, ...). Náhrada povinných předmětů nepovinnými však není vždy jednoduchá.

Jedním z problémů je zajištění potřebné návaznosti jednotlivých předmětů. Zavedení výběrových seminářů s matematickou či statistickou tematikou se také potýká s malým zájmem ze strany studentů, kteří mají většinou z obdobných předmětů strach.

Pomoci by mohlo zavedení předmětů, které studenty na první pohled neodradí. Částečným řešením může být i pouze vhodná volba názvu předmětu. To ale není dlouhodobě příliš účinné. Účinnější možností je zavádění předmětů, které názorně ukazují řešení reálných situací z oborů blízkých studijnímu zaměření studentů za skrytého použití matematických či statistických metod. Pokud se podaří studenty zaujmout problematikou, budou více přístupni přijmout vysvětlení metod, na kterých je řešení založeno, i kdyby se jednalo o metody matematické či statistické.

Výklad látky pak neprobíhá standardním způsobem hierarchicky od nejjednoduššího postupně stále ke složitějšímu, ale naopak se začne cílovým stupněm, to znamená nejsložitějším, postupně se pak osvětlují potřebné informace na nižším stupni složitosti a to až do úplného porozumění problematice.

Rozhodovací stromy jsou problematikou, která splňuje předchozí požadavky. Na první pohled jsou rozhodovací stromy velmi názorné, i laik se velmi rychle zorientuje v grafice stromové struktury a je schopen vyčíst potřebné informace. Využití rozhodovacích stromů je v ekonomickém oboru velice široké, takže je možno studenty seznámit s konkrétními příklady z ekonomického života.

V teorii rozhodovacích stromů jsou jednak využity základní poznatky z oblasti teorie grafů a jednak poněkud širší poznatky ze statistiky. Studenti se

učí jednotlivým informacím s vědomím jejich využitelnosti a důležitosti pro tvorbu rozhodovacích stromů. Tím odpadá potřeba přesvědčovat studenty o potřebě daného učiva.

Celkově si myslím, že zavedení tématu rozhodovacích stromů (a jiných podobných témat) do výuky ve formě výběrových seminářů, je velmi dobrou cestou k rozšíření látky s matematickým zaměřením.

Reference

- [1] Antoch J., Klasifikace a regresní stromy. Sborník ROBUST 88.
- [2] Bentley, J. L.: Multidimensional Binary Search Trees Used for Associative Searching. Comm. ACM, vol. 18, pp. 509-517, 1975.
- [3] Berikov, V., Litvinenko, A.: Methods for statistical data analysis with decision trees,
<http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm>
- [4] Loh, W.-Y. and Shih, Y.-S., Split selection methods for classification trees, Statistica Sinica, vol. 7, 815-840., 1997.
- [5] Savický, P., Klaschka, J., a Antoch J.: Optimální klasifikační stromy. Sborník ROBUST 2000.
- [6] SPSS-white paper-AnswerTree Algorithm Summary.
- [7] Timofeev R.: Classification and Regression Trees (CART) Theory and Applications, CASE – Center of Applied Statistics and Economics, Humboldt University, Berlin, 2004.
- [8] Wilkinson, L.: Tree Structured Data Analysis: AID, CHAID and CART – Sun Valley, ID, Sawtooth/SYSTAT Joint Software Conference, 1992.
- [9] Žambochová M.: Použití stromů ve statistice – Sborník, 2006, Ústí n. L., ISBN 80-7044-795-8.
- [10] Classification Trees: <http://www.statsoft.com/textbook/stclatre.html>.
- [11] Classification and Regression Trees (C&RT): <http://www.fmi.uni-sofia.bg/fmi/statist/education/textbook/ENG/stcart.html>.

Adresa: RNDr. Marta Žambochová, Univerzita J. E. Purkyně v Ústí n. L., Fakulta sociálně ekonomická, Katedra matematiky a statistiky

E-mail: zambochova@FSE.UJEP.cz

Obrazové a tabulkové přílohy

Node	Node branch	Left branch	Right normální	n in cls podváha	n in cls nadváha	n in cls obezita	n in cls Predict.	Split	Split	Split	Split
1	2	3	51	15	31	13	normální		strava	2	1
2	4	5	2	0	28	13	nadváha	-0,5	sport		
3	6	7	49	15	3	0	normální	-6,5	spánek		
4			0	0	0	11	obezita				
5			2	0	28	2	nadváha				
6	8	9	4	13	0	0	podváha	-3,5	počet		
7			45	2	3	0	normální				
8			1	13	0	0	podváha				
9			3	0	0	0	normální				

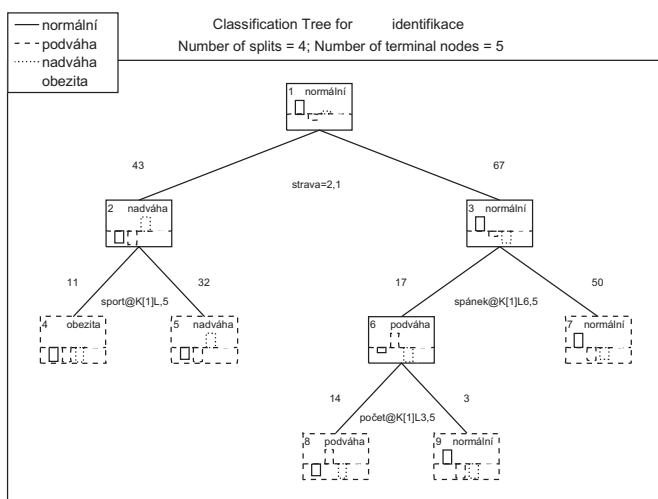
Tabulka 1: Popis uzlů klasifikačního stromu vytvořeného pomocí algoritmu C&RT vyčerpávajícího prohledávání

	Node	Node branch	Left branch	Right normální	n in cls podváha	n in cls nadváha	n in cls obezita	n in cls Predict.	Split	Split	Split	Split
1	2		3	51	15	31	13	normální	-3,57773	televize		
2	4		5	48	10	7	0	normální	-5,90627	spánek		
3	6		7	3	5	24	13	nadváha	-4,32758	sport		
4				0	4	0	0	podváha				
5	8		9	48	6	7	0	normální		strava	3	4
6	10		11	2	3	24	13	nadváha	-3,49525	sport		
7				1	2	0	0	podváha				
8	12		13	47	6	2	0	normální	-2,33213	počet		
9				1	0	1	0	normální				
10	14		15	1	3	23	13	nadváha	-2,05128	sport		
11				1	0	1	0	normální				
12				0	3	1	0	podváha				
13				47	3	1	0	normální				
14	16		17	0	3	21	13	nadváha	-0,634035	sport		
15				1	0	2	0	nadváha				
16				0	1	0	11	obezita				
17	18		19	0	2	21	2	nadváha		strava	1	2
18				0	0	21	2	nadváha				
19				0	2	0	0	podváha				

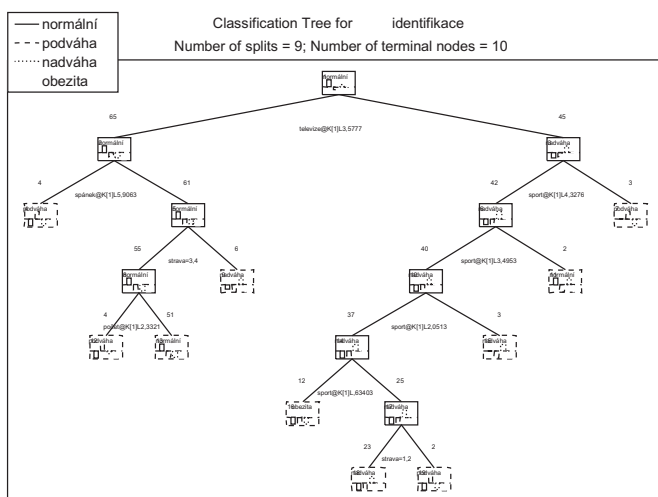
Tabulka 2: Popis uzlů klasifikačního stromu vytvořeného na základě algoritmu QUEST

Node	Left branch	Right branch	Size of node	N in class normální	N in class podváha	N in class nadváha	N in class obezita	Selected	Split	Split	Split	Split
1	2	3	110	51	15	31	13	normální	strava		4	3
2	4	5	67	49	15	3	0	normální	spánek	6,5		
4	6	7	17	4	13	0	0	podváha	počet	3,5		
6			14	1	13	0	0	podváha				
7			3	3	0	0	0	normální				
5	10		11	50	45	2	3	normální	počet	2,5		
5	10		11	50	45	2	3	normální	počet	2,5		
10			4	0	2	2	0	podváha				
11			46	45	0	1	0	normální				
3	14	15	43	2	0	28	13	nadváha	sport	0,5		
14			11	0	0	0	11	obezita				
15	16	17	32	2	0	28	2	nadváha	sport	3,5		
16			30	0	0	28	2	nadváha				
17			2	2	0	0	0	normální				

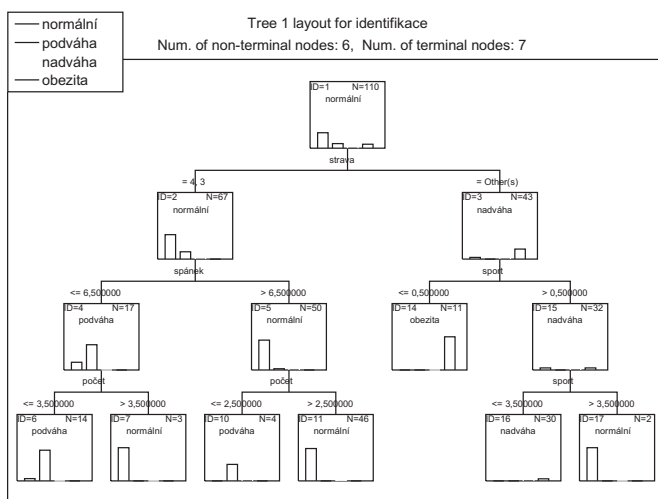
Tabulka 3: Popis uzlů klasifikačního stromu vytvořeného na základě standardní metody C&RT z modulu Data Mining



Obrázek 1: Klasifikační strom vytvořený pomocí algoritmu C&RT vyčerpávajícího prohledávání



Obrázek 2: Klasifikační strom vytvořený na základě algoritmu QUEST



Obrázek 3: Strom vytvořený pomocí standardní metody C&RT z modulu Data-Mining, včetně V-fold Crossvalidation – metody na výběr nejoptimálnějšího stromu

STATISTICKÉ VÝPOČETNÍ PROSTŘEDÍ 2007

STATISTICAL COMPUTING ENVIROMENT 2007

Jiří Žváček

Abstrakt: V článku jsou popsány současné tendence ve vývoji software a jeho využívání, které poněkud rozšiřuje pojetí Statistického výpočetního prostředí a je navrhován termín Statistické informační technologie. Ve druhé části je popsán současný stav v oblasti statistického softwaru.

Abstract: New tendencies in the development of information technology are described in the paper and the term Statistical information technology instead of Statistical computing enviroment is suggested.

In the second part of the paper the situation in the statistical computing software is described.

Před dávnými časy (1987) jsme na VŠE spolu s kolegy ze Slovenska začali zavádět předměty z oblasti výpočetní statistiky. Cílem bylo zavést kurzy, které by učily statistiky zpracovávat statistické úlohy na současné výpočetní technice.

V Čechách jsme zavedli předmět **Výpočetní statistika**, ve které jsme probírali širší spektrum oblastí od programování a speciálních numerických metod až po ovládání softwarových produktů, zatímco kolegové se zaměřili spíše přímo na **Statistické pakety**.

Vzhledem k dynamice vývoje softwaru vznikla i u nás lidovější varianta kursu Výpočetní statistika, který tvořilo ovládání softwarových prostředků pro statistické výpočty a který jsme nazvali **Statistické výpočetní prostředí**. Zahrnoval zejména práci se statistickými pakety, tabulkovými procesory a práci s databázemi. Vzhledem k tomu, že zejména tabulkový procesor často pro jednodušší statistické výpočty a zejména přípravu dat postačuje a že prostředí je třeba ještě rozšířit přinejmenším o prostředky publikování, převažují dnes spíše takovéto kurzy.

V tomto přehledu se pokusím shrnout to, co se podle mého názoru podstatného událo v oblasti statistického výpočetního prostředí za tento rok.

1. Zásadní změna prostředí

Duchu doby by lépe odpovídal název **statistické informační technologie**, protože dochází k zásadním změnám.

Statistikové totiž kromě vlastního počítání potřebují výpočetní prostředky například i ke

komunikaci kam lze zařadit třeba vyhledávání, sběr a výměnu informací,

prezentaci tedy jakým způsobem publikovat statistiku tiskem a na internetu moderními technologiemi,

výpočtům v širším slova smyslu včetně interaktivní analýzy a stále důležitější grafické prezentace,

podpoře výuky včetně online komunikace a multimediálních interaktivních učebnic vytvářených pomocí matematického a statistického softwaru.

Přirozeně už nejde pouze o počítače (a tím méně o stolní počítače), ale o všechny technické prostředky, které statistik či student při své práci využívá. Sledovat je tedy třeba širší okruh prostředků, softwaru a zejména a v první řadě internet.

V tomto směru bude tedy třeba doplnit své vzdělání (a příslušný kurs) a naučit se tato zařízení ovládat a programovat.

2. Vývoj v oblasti informačních technologií

Málokterá oblast lidské činnosti vykazuje tak dlouhé období dynamického vývoje.

2.1. Hardware

Posledních třicet let exponenciálně rostou technické možnosti a za nimi pádí naše schopnost je využívat. Konec platnosti **Moorova zákona**¹ je stále v nedohlednu.

Základní tendence jsou

- jsme v oblasti **nanotechnologií**, procesory i další zařízení jsou stále menší,
- mění se architektura, procesory jsou 64bitové a vícejádrové,
- ceny hardwaru dramaticky klesají,

¹http://en.wikipedia.org/wiki/Moore's_law

- procesory se vyskytují v mnoha zařízeních a produktech (platební karty, RFID identifikace, ...).

Důsledkem je, že z mnoha dalších zařízení se postupně stávají specializované počítače (obsahují kromě procesoru i paměti, operační systémy a komunikaci). Zřejmě již zanedlouho budeme obklopeni miniaturními chipy, po zvířatech a zboží dostáváme i my postupně svůj **RFID**².

I počítače budou vypadat úplně jinak. Přinejmenším již dnes začínají převažovat levné notebooky a nastupují ultramobilní **UMPC**³.

Počítat se bude zejména na mobilních zařízeních a i na jiných zařízeních než jsou počítače. Pro statistiky jsou aktuální zejména:

- **Kalkulačky** což jsou dnes specializované vědeckotechnické počítače, které mají stejné schopnosti jako matematické pakety a jsou navrhovány speciálně pro výuku a mobilní výpočty. (Viz třeba **TI-Nspire**⁴, který umí i symbolické výpočty.)
- **Mobily** dnes mají operační systémy, prohlížeče a přístup k internetu (třeba Nokia má Symbian **S60**⁵) a existuje řada programů pro podporu matematiky a statistiky na mobilech (viz např. **matematika**⁶). Programování mobilů se příliš neliší od programování počítačů a potenciálně mají stejné možnosti.
- **PDA** jsou malé počítače do ruky, které dnes už mohou totéž co stolní počítače. Pokud mají operační systém Windows Mobile, tak mají i trochu redukováný EXCEL a WORD. Existují již i první statistické pakety pro mobilní zařízení (třeba **Statgraphics Mobile**⁷), na němž lze počítat i spolupracovat s počítačem. Nejnovější (např. **Fooleo**⁸) již pracují s plnohodnotným prohlížečem (Opera 9) a mohou tedy využívat všech služeb internetu.

²<http://www.pcworld.cz/pcw.nsf/redakcni-blok/-cipovani_lidi_vstoupilo_do_dalsiho_kola>

³<http://www.itnews.sk/buxus_dev/generate_page.php?page_id=43273&buxus_itnews=71bce3993137526cb5fd0752fd0fb461>

⁴<<http://en.wikipedia.org/wiki/TI-Nspire>>

⁵<<http://www.s60.cz/c/s60-cz/>>

⁶<<http://milanl.wz.cz/java/matematika/index.htm>>

⁷<http://www.statgraphics.com/statgraphics_mobile.htm>

⁸<http://digiweb.ihned.cz/c4-10050430-21293980-i00000_d-palm-uvadi-na-trh-foleo-zarizeni-podobne-mininotebooku>

2.2. Internet

Masová dostupnost internetu spolu s rostoucí kvalitou připojení již má za následek, že mnoho činností expanduje na internet. Kromě informací a komunikace jsou to zejména ekonomické činnosti, který již je významnější než mnohé klasické ekonomické činnosti. Zájem významných ekonomických subjektů se stává **všeobecná dostupnost internetu** a internet se postupně stává **primárním zdrojem informací**. Postupně se na něj přenáší i většina písemných a vizuálních dat.

Mění se dokonce hovorový jazyk, i čeština přejímá krátká anglická slova jako jsou **web**, **blog**, **chat**, **web2**, **wiki** atd. (viz třeba [Slovníček](#)⁹). Ten kdo je mimo oblast informatiky (lama) může mít trochu problémy při hovoru s mladou generací.

Širokopásmový rychlý internet umožňuje využívat i prostředky, které mění náš pohled na výpočetní prostředky (multimedialita, interaktivita).

Internet se stává nejenom nejvýznamnějším motorem vývoje v mnoha oblastech ale i prostředím, ve kterém jsme každodenně.

2.3. Software

Výpočetní prostředky se paradoxně stávají mnohem jednodušší – hardware je velmi podobný a většina rozdílů je v softwaru.

Dnešní software většiny moderních přístrojů si již nelze představit bez internetu. I ve statistice každý program či paket (dokonce i výrobek) musí mít internetovou stránku, která umožňuje realizovat celou řadu služeb. Úspěšnost softwaru je přitom stále více závislá na kvalitě poskytovaných služeb.

Novinkou poslední doby je zejména **Web 2**. Není druhá generace webu ale nové anglické slovo. V anglickém jazyce je to foneticky „web k“, tedy něco co je poskytováno z webu uživateli nebo naopak od uživatele na web. Týká se to celého postoje k internetu, ale projevuje se to zejména u softwaru.

2.3.1. Uživatelská podpora Snad ke každému úspěšnějšímu softwaru si prodávající vytváří možnosti pro kumulaci zkušeností a znalostí uživatelů. Kromě diskusního fóra existuje mnoho dalších aktivit jako je nabídka maker, tutoriálů, webinářů (seminářů na internetu) atd.

Velmi se osvědčilo **umožnění tvorby uživatelských pluginů**, které umožňují operativně doplňovat do základní aplikace jednoduchým způsobem další činnosti. Umožňují to ty nejpokročilejší programy právě proto, aby zbytečně neobtěžovaly a přitáhly pozornost hravé odborné veřejnosti (snad každý

⁹[<http://www.root.cz/slovnicek/>](http://www.root.cz/slovnicek/)

je zná z Total Commanderu, všech prohlížečů atd.). Podmínkou je, aby implementace byla i pro průměrně zdatného uživatele snadná.

2.3.2. Stahovatelný software Velká většina dnešního softwaru je k dispozici na internetu a i komerční software lze přinejmenším na zkušební dobu použít. Kromě známých komerčních programů a služeb hraje stále důležitější roli software zdarma, který má více forem (viz pěkný [graf vztahů](#)¹⁰), založených zejména na vlastní aktivitě uživatelů.

Zřejmě nejdynamičtější oblastí vývoje software je [opensource](#)¹¹ software. Model otevřené spolupráce je tak úspěšný, že často vzniká i na základě komerčního programu, na který již nemá autor či firma dost sil pro adekvátní podporu. Kolem skupiny nadšenců se vytvoří okruh přispívajících a produkt se rychle vyvíjí. Příklady je mnoho (Linux, Firefox, Open Office, PHP) a okruh se stále rozšiřuje (přibyla Java). Přehled je na stránce [SourceForge.net](#)¹², která eviduje více než 132 000 projektů a poskytuje základní služby.

2.3.3. Webové aplikace Webová aplikace, též **online software** je software, který ovládáme z internetu (program je mimo počítač uživatele).

Vznikla řada internetových služeb, které umožňují přenést mnohé činnosti a data na internet a jistou formou je sdílet s dalšími uživateli. Tímto směrem vrhly i velké softwarové firmy Microsoft, Google a AOL.

Je to také reakce na přechod k mobilním zařízením a současné práci na více počítačích.

2.3.4. Wiki Fenomén internetové encyklopedie [Wikipedia](#)¹³ vytvářené uživateli internetu inicioval vytvoření nového mezinárodního slova a zavedl nový způsob práce na webu. **Wiki** je internetový obsah vytvářený a editovatelný návštěvníky.

Wikipedia je realizována jako [opensource software MediaWiki](#)¹⁴, takže ji může provozovat každý a existuje řada serverů, které umožňují hostování wiki stránek.

¹⁰<http://www.gnu.org/philosophy/category.png>

¹¹http://cs.wikipedia.org/wiki/Open_source_software

¹²<http://sourceforge.net/index.php>

¹³<http://www.wikipedia.com/>

¹⁴<http://www.mediawiki.org/wiki/MediaWiki>

Vzniká mnoho nejružnějších wiki stránek a stránek s obdobnou filozofií. Jsou to například [Wikibooks](#)¹⁵ což jsou kolektivně upravované knihy, obdobně [Wiktionery](#)¹⁶ jsou slovníky a třeba [Wikinews](#)¹⁷ jsou novinky.

Stránky typu wiki jsou vhodné pro dokumenty, na kterých pracuje více autorů a kde záleží na rychlosti aktualizace.

3. Statistický software

I na statistický software je vhodné se podívat z hlediska současných tendencí.

3.1. Software zdarma

Stále více produktů je k dispozici zdarma, zejména pro nekomerční účely. Je toho mnoho a musí se to umět najít a naučit ovládat. Důležitou roli hrají přehledy

- [Free Statistical Software](#)¹⁸ je rozsáhlý přehled Pezullův,
- [FreeBSD/math](#)¹⁹ je přehled matematického softwaru zdarma.

Software poskytovaný zdarma má nejružnější formy, které se liší zejména způsobem přístupu k vlastnickým právům.

3.1.1. Opensource Je i hodně statistických opensource projektů (řádově 900 jich souvisí se statistikou). Mezi nejvýznamnější patří

The R Project [<http://www.r-project.org/>](http://www.r-project.org/) Jazyk **R** je opensource klon komerčního paketu **S+** s velmi širokou škálou navazujících produktů a dynamickým vývojem. Nové verze by měly vycházet vždy 1.4. a 1.9., lze se zúčastnit vývoje na beta verzích a objednat si novinky. Současná verze je 2.5.0, vše podstatné je na specializované [wiki stránce](#)²⁰. R má wikibook [Statistical Analysis using R](#).²¹

OCTAVE [<http://octave.sourceforge.net/>](http://octave.sourceforge.net/) je open source varianta paketu MATLAB. Viz [wiki/GNU_Octave](#)²². Časté inovace. V 3/2007 reorganizován do paketového systému. Vznikla česká podpůrná stránka [Octave](#)²³.

¹⁵[<http://wikibooks.org/>](http://wikibooks.org/)

¹⁶[<http://wiktionery.org/>](http://wiktionery.org/)

¹⁷http://en.wikinews.org/wiki/Main_Page

¹⁸<http://statpages.org/javasta2.html#Freebies>

¹⁹<http://www.freebsdsoftware.org/math/>

²⁰<http://wiki.r-project.org/rwiki/doku.php>

²¹http://en.wikibooks.org/wiki/Statistical_Analysis:_an_Introduction_using_R

²²http://en.wikipedia.org/wiki/GNU_Octave

²³<http://www.octave.cz/>

MATLAB klonů je více, např.

SciLab <<http://www.scilab.com/>> je podobný MATLABu (popis viz [Scilab-Wikipedia](#)²⁴) a má nyní verzi 4.1.

Z mnoha dalších jsou „živé“ například

- [Gretl](#)²⁵ zajímavý opensource pro časové řady a ekonometrii (s možným výstupem do TeXu).
- [OpenEpi](#)²⁶ epidemiologická statistika.
- [Gnumeric](#)²⁷ spreadsheet s mnoha funkcemi.
- [Tanagra](#)²⁸ je pro datamining.
- [Slovak Math Ubuntu](#)²⁹ je speciální implementace Linuxu obsahující některé OS matematické programy.
- [PAST](#)³⁰ je pro statistiku v paleontologii.

3.1.2. Online software Prakticky vše lze dnes spočítat online. Například

- [Interactive Statistical Calculation Pages](#)³¹ přehled (Pezullo).
- Interaktivní prostředí pro R (opensource [R commander](#)³²) lze instalovat na vlastní stránky.
Funkční verze tohoto prostředí je na [R Online](#)³³.
- [Wessa](#)³⁴ je další verze pokročilého interaktivního rozhraní pro R, kde je mnoho hotových interaktivních statistických výpočtů. Zde je možno publikovat i vlastní algoritmy (postarají se o úpravy při změnách verzí R).
- [Fyzikální generátor náhodných čísel online](#)³⁵.

²⁴<<http://en.wikipedia.org/wiki/Scilab>>

²⁵<<http://en.wikipedia.org/wiki/Gretl>>

²⁶<<http://www.openepi.com/Menu/OpenEpiMenu.htm>>

²⁷<<http://www.gnome.org/projects/gnumeric/>>

²⁸<<http://chirouble.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>>

²⁹<<http://people.tuke.sk/peter.mann/ubuntu/slovak-math-ubuntu-livecd.html>>

³⁰<<http://folk.uio.no/ohammer/past/>>

³¹<<http://statpages.org/>>

³²<http://en.wikipedia.org/wiki/R_Commander>

³³<<http://user.cs.tu-berlin.de/~ulfi/cgi-bin/r-online/r-online.cgi>>

³⁴<<http://www.wessa.net/stat.wasp>>

³⁵<<http://www.randomnumbers.info/>>

- [WebMathematica](#)³⁶ obsahuje mnoho online matematických výpočtů, např. [eFunda: Exponential Curve Fitting](#)³⁷.

3.1.3. Statistické wiki Existují rozšíření wiki pro matematiku, které umožňují implementovat spousty užitečných vlastností. Třeba textové zadávání grafů ([Graphviz](#)³⁸, grafika [Graph](#)³⁹, [interaktivní grafy TeX editor](#)⁴⁰, [hodnoty matematických a statistických funkcí](#)⁴¹, [výpočty](#)⁴² atd.

Speciálně pro statistiku jsou již aktivní anglické wikiknihy

- [Handbook of Descriptive Statistics – Wikibooks, collection of open-content textbooks.](#)⁴³
- [Probability – Wikibooks, collection of open-content textbooks.](#)⁴⁴
- [Statistics – Wikibooks, collection of open-content textbooks.](#)⁴⁵

3.2. Novinky v oblasti paketů

Statistické pakety si stále udržují svou zdánlivou nezbytnost.

Rozšiřuje se okruh služeb poskytovaných k paketu. Jsou to např. videozáznamy přednášek (podcast, viz [vysvětlení termínu](#)⁴⁶, [webcasty](#)⁴⁷), internetové vysílání kursů (Webcasts, [webcast](#)⁴⁸), odpovídání na technické dotazy, internetový deník (blog)), výměna souborů, analýz a skriptů. Řada novinek a zejména chyb statistického software je na stránkách [IASC](#)⁴⁹.

Za největší letošní novinku pokládám vznik seznamu statistických paketů [List](#)⁵⁰ a popisů jednotlivých produktů na internetové encyklopedii Wikipedia. Paráda je, že můžeme prakticky na jednom místě sledovat celou problematiku paketů.

³⁶http://www.efunda.com/webM/home/webM_home.cfm

³⁷<http://www.efunda.com/webM/numerical/curvefitexp.cfm>

³⁸<http://graphviz.org/>

³⁹http://meta.wikimedia.org/wiki/Graph_extension

⁴⁰http://meta.wikimedia.org/wiki/Touchgraph_Extension

⁴¹<http://meta.wikimedia.org/wiki/MathStatFunctions>

⁴²<http://meta.wikimedia.org/wiki/Help:Calculation>

⁴³http://en.wikibooks.org/wiki/Handbook_of_Descriptive_Statistics

⁴⁴<http://en.wikibooks.org/wiki/Probability>

⁴⁵<http://en.wikibooks.org/wiki/Statistics>

⁴⁶<http://www.lupa.cz/clanky/podcast-revoluce-v-internetovem-vysilani/>

⁴⁷<http://www.akamonitor.cz/webcasting.htm>

⁴⁸<http://www.akamonitor.cz/eseminar.htm>

⁴⁹<http://www.csdassn.org/softlist.cfm>

⁵⁰http://en.wikipedia.org/wiki/List_of_statistical_packages

Nejnovější stav můžeme sledovat na stránce [Comparison of statistical packages](http://en.wikipedia.org/wiki/Comparison_of_statistical_packages)⁵¹.

Věcně za nejzajímavější pokládám rozšiřující pluginy v Pythonu a vůbec vývoj u SPSS a verze pro PDA u STATGRAPHICSu a WINKSe.

Významnější inovace nastaly zejména u následujících produktů (subjektivně v pořadí užitečnosti změn)

JMP <<http://www.jmp.com/>> Paket původně pro Apple koupený firmou SAS. Nyní verze 7. Online kursy (Webcasts). Má emailové novinky i RSS. Také rostoucí sbírku skriptů na nejrůznější témata.

SPSS <<http://www.spss.com/>> U nás spss.cz⁵². Jeden z nejstarších a nejrozsáhlejších paketů vypustil verzi **SPSS 15**⁵³ s větší podporou pdf. Jde cestou uživatelských pluginů a moderních jazyků (Python). Pro Vistu potřebuje doinstalovat plugin. Připravuje se verze 16 s rozhraním pro projekt R.

STATGRAPHICS Plus <<http://www.statgraphics.com/>> Inovoval na verzi 15.2 a zejména vznikla verze **Statgraphics Mobile** pro handheldy s operačním systémem Windows Mobile.

SYSTAT <<http://www.systat.com/>> Oblíbený paket, nyní ve verzi 12.

BMDP <<http://www.statsol.ie/bmdp/bmdp.htm>> Nyní verze BMDP 2007.

STATISTICA <<http://www.statsoftinc.com/>> inovovala na verzi **STATISTICA 8**⁵⁴. České zastoupení⁵⁵ o tom mlčí a ceny pouze na optání, tedy asi individuální. Zaujala mne propagace pomocí lákavých titulků: **STATISTICA je skutečný lídr mezi statistickými balíky**.

Xplore <<http://www.xplore-stat.de/>> Humboldtovy university je nyní ve verzi 4.7. A ve vykleštěné podobě pro akademické účely zdarma.

GAUSS <<http://www.aptech.com/>> Inovoval na verzi 8.

MINITAB <<http://www.minitab.com/>> Inovoval na verzi 15. **ActivStats** for MINITAB je interaktivní multimediální statistický text, ve kterém jsou animace, dynamické grafy a videoklipy.

S-PLUS <<http://www.insightful.com/>> U nás podporuje tento paket firma [TriloByte](http://www.trilobyte.cz)⁵⁶, **Veliký pokrok**. Verze 7 pracuje s obrovskými soubory, implementuje pokročilou statistiku atd. Nyní už verze 8 umožňuje snadné doplňování dalších funkcí. Vhodný pro profesionály, lze v něm napsat vlastní aplikace.

⁵¹<http://en.wikipedia.org/wiki/Comparison_of_statistical_packages>

⁵²<<http://spss.cz/>>

⁵³<<http://spss.cz/sw-spss-whatsnew.htm>>

⁵⁴<<http://www.statsoft.com/list.htm>>

⁵⁵<<http://www.statsoft.cz/>>

⁵⁶<<http://www.trilobyte.cz/>>

WINKS <<http://www.texasoft.com/>> Malý, jednoduchý a levný (od 99 dolarů i méně) paket, naprosto postačující pro výuku. Nyní verze 6 a verze pro PDA. On-line manual, tutoriály metod atd.

NCSS <<http://www.ncss.com/>> Rozsáhlý paket je ve verzi NCSS 2007, přibyla řada procedur, makra atd. Studentská verze 6 je pro výuku zdarma.

GENSTAT <<http://www.vsn-intl.com/>> Britský paket s dobrou podporou, inovoval na verzi 9. Je **zdarma pro 85 rozvojových zemí**.

StatPlus 2007 <<http://www.analystsoft.com/en/>> Jednoduchý a levný paket (120 \$).

QCexpert <<http://www.trilobyte.cz/extras/qcexpert.3.pdf>> Jediný větší český statistický paket, má verzi 3.

Na rozumný levný paket pro výuku milionů studentů stále čekáme.

3.3. Tabulkové procesory

Prakticky zbyl pouze EXCEL, který je nyní ve verzi EXCEL 2007. Podrobný přehled ze statistického hlediska je na stránce [Chyby, problémy a opravy](#)⁵⁷, včetně základních numerických problémů a oprav.

K EXCELU existuje mnoho statistických nadstavb (viz přehled [EXCEL Add-Ins](#)⁵⁸). Konkrétně třeba [UNISTAT](#)⁵⁹, který je dokonce lokalizován do češtiny (už pracuje i s Windows Vista a Office 2007), ale i mnohem levnější [Winstat](#)⁶⁰, [Sigmazone](#)⁶¹, [Lumeaut](#)⁶² je pro studenty a učitele na rok zdarma, [statistiXL](#)⁶³ stojí na rok 40 \$ a australský [XLStat](#)⁶⁴ je zcela zdarma. I u nás se objevují hotové statistické skripty pro EXCEL (letos zejména [Klára Mrázová](#)⁶⁵), objevila se i učebnice [Mat-MAPLE](#)⁶⁶ s návody pro EXCEL a MATLAB.

Počítat v EXCELU lze interaktivně i na webu [XL2Web](#)⁶⁷.

⁵⁷<<http://www.daheiser.info/EXCEL/frontpage.html>>

⁵⁸<<http://dmoz.org/Computers/Software/Spreadsheets/EXCEL/Add-Ins/>>

⁵⁹<<http://www.unistat.com/>>

⁶⁰<<http://www.winstat.cz/>>

⁶¹<<http://www.sigmazone.com/>>

⁶²<<http://www.lumenaut.com/statistics.htm>>

⁶³<<http://www.statistixl.com/>>

⁶⁴<<http://www.deakin.edu.au/~rodneyc/XLSTATS.HTM>>

⁶⁵<<http://www.pcsvet.cz/art/author.php?id=429&page=2%20>>

⁶⁶<http://www.vscht.cz/mat/matstat/stat_m_e/Stat_MATLAB_EXCEL.html>

⁶⁷<<http://www.itksoftware.com/home.jsp>>

3.4. Matematické systémy

Matematické systémy začínají dnes být vážným konkurentem statistických paketů v oblasti výuky. Obchodně se zaměřují na pozdější úspěch, takže jejich znalost lze očekávat u lepších středoškoláků. Pracují i se symbolickou matematikou, takže jsou dobře použitelné i v oblasti teorie.

Nejvhodnější jsou podle mne nyní:

MATHEMATICA <<http://www.wolfram.com/>> (u nás [Elkan](#)⁶⁸).

Popis nejlépe na [Wiki Mathematica](#)⁶⁹.

Nová, údajně revoluční verze **Mathematica 6** přináší dynamické a interaktivní grafy a výpočty. Pro statistiku jsou zajímavé zejména symbolické statistické výpočty a zpracování dat.

Vzhledem k tomu, že existuje zdarma [Mathematica Player](#)⁷⁰, který po instalaci umožňuje prohlížení dokumentů vytvořených v Mathematice, je tedy možno psát výukové stránky pro internet (viz třeba [pravděpodobnost](#)⁷¹ či [statistika](#)⁷²).

Mathematica umožňuje také publikovat interaktivní výpočty na stránce. Na firemní stránce je mnoho aplikací zdarma od uživatelů, v 17. 6. 2007 to bylo [268](#)⁷³ statistických úloh. Problém je v tom, že jsou publikovány zejména ve firemním časopise [The Mathematica Journal](#)⁷⁴ a je třeba zaplatit za přístup. Zdarma jsou [wiki Mathematica](#)⁷⁵ např. [eFunda](#)⁷⁶ ale jsou jich tisíce z mnoha oborů.

MAPLE <<http://de.wikipedia.org/wiki/MAPLE>> popis na [wikiMAPLE](#)⁷⁷.

Je vzhledem k cenám častější (na internetu Evropané nadávají, že MATHEMATICA je v Evropě o 70 % dražší než v USA). Verze **MAPLE 11** podporuje **interaktivní dokumenty** a **pracuje se na prohlížeči MAPLE dokumentů zdarma**. Už MAPLE 10 umožňovala

⁶⁸<<http://www.elkan.cz/>>

⁶⁹<<http://en.wikipedia.org/wiki/Mathematica>>

⁷⁰<<http://www.wolfram.com/products/player/download.cgi>>

⁷¹<<http://demonstrations.wolfram.com/topic.html?topic=Probability&limit=20>>

⁷²<<http://demonstrations.wolfram.com/topic.html?topic=Statistics&limit=100>>

⁷³<http://library.wolfram.com/infocenter/BySubject/Mathematics/ProbabilityStatistics/?page=1;pages_count=100000>

⁷⁴<<http://www.mathematica-journal.com/issue/v10i2/>>

⁷⁵<http://www.mathematica-users.org/webMathematica/wiki/wiki.jsp?pageName=Main_Page>

⁷⁶<http://www.efunda.com/webM/home/webM_home.cfm>

⁷⁷<[http://en.wikipedia.org/wiki/MAPLE_\(software\)](http://en.wikipedia.org/wiki/MAPLE_(software))>

publikovat do HTML dynamické grafy (viz [výukové listy k pravděpodobnosti](#)⁷⁸).

MATHCAD <<http://www.mathcad.com/products/mathcad/>> též [Mathsoft.cz](#)⁷⁹ a informace na [Wiki/Mathcad](#)⁸⁰. Každoroční inovace je 14, demo 13.1 pro studenty a učitele je prodlouženo na 4 měsíce. Spousty řešení statistických úloh na [uživatelském fóru](#)⁸¹.

MATLAB <<http://www.mathworks.com/>> [Wiki/MATLAB](#)⁸² (speciálně [programování](#)⁸³). Statistic Toolbox inovoval na verzi 5.3. MATLAB je populární zejména u inženýrů a má výbornou podporu uživatelů.

Matematické systémy jsou velmi vhodné pro psaní dynamických interaktivních výukových textů.

3.5. Specializované programy

Sem zařazujeme zpravidla **dílčí statistické systémy**, které pokrývají pouze určité statistické metody. Je jich obrovská spousta pro nejrůznější úlohy a často to jsou astronomicky drahé speciální programy pro bohaté firmy a úřady. Mnoho statistických firem si vytváří vlastní software.

Vystopovat lze zejména některé okruhy šířeji použitelných metod, které mají vlastní statistické programy a pakety. Dynamický vývoj je zejména ve dvou oblastech:

Datamining Datamining je tak trochu paběrkování na datových smetištích, kde vyslovovat klasické předpoklady o náhodných veličinách se zdá být neadekvátní, takže klasické statistické usuzování jde trochu stranou. Metody jsou spíše heuristické, fundamentalističtí statistikové trochu ohrnují nos, takže se převážně zařazují do informatiky. Metody bývají jako samostatné programy u všech významnějších firem. Asi nejpokročilejší je asi nová SPSS Clementine 10.

Statistické grafy Význam grafiky obecně stoupá, možnosti počítačů a požadavky na estetické ztvárnění se zvyšují.

Toho využily spousty firem specializovaných na statistickou grafiku. Novinek je příliš mnoho, bylo by to na několik samostatných článků.

⁷⁸<<http://phi.kenyon.edu/personal/hartlaub/MellonProject/mellon.html>>

⁷⁹<<http://www.mathsoft.cz/>>

⁸⁰<<http://en.wikipedia.org/wiki/MathCad>>

⁸¹<<http://collab.mathsoft.com/>>

⁸²<<http://en.wikipedia.org/wiki/MATLAB>>

⁸³<<http://en.wikibooks.org/wiki/Programming:MATLAB>>

Z ostatních novinek mne zaujaly zejména:

AM <<http://am.air.org/>> Analýza složitých výběrů.

JMP Genomics 3.0 ⁸⁴ Je určen pro statistické analýzy v genetice.

Interactive Neural Network Book ⁸⁵ Je interaktivní učebnice teorie odhadu pomocí neuronových sítí (funkční demo zdarma).

SSI <<http://www.ssicentral.com/index.html>> Má novou stránku a nejslavnější program LISREL (strukturální modely) je ve verzi 8.8. Starší verze pro výuku zdarma.

Xtremes Webpage <<http://www.xtremes.math.uni-siegen.de/xtremes/>> Stránka ke knize Extremální rozdělení a specializovanému programu (starší verze pro výuku zdarma). Nová kniha (3. vydání, vyjde tuto zimu), program 4.0, StatPascal.

Resampling Stats <<http://www.resample.com/>> Samotný program už ve verzi 5.0, ad-ony k EXCELU a MATLABu.

Specializované programy sice odebírají paketům významnou část trhu, ale zejména v oblasti výuky je plně nahradit nemohou.

3.6. Výukové stránky

Výukové stránky pomalu vznikají na stránkách kateder, někde i oddělené od katedrální stránky (VŠE) a sem tam některý učitel dává konečně alespoň vzorce na web. Jsou projekty ve výhledu, vše co víme je na stránkách kateder na [Statspol](http://www.statapol.cz)⁸⁶.

4. Závěr?

Z pohledu informačních technologií vidím tyto tendence:

- Klasický model výuky statistiky v počítačové učebně skomírá. Studenti mají mobilní hardware, mnoho interaktivity lze nahradit internetem a stejně není na přímou výuku čas.
- Kvůli ceně se ani při výuce nepoužívají jednotným způsobem statistické pakety (když vůbec). Chybí jednoduchý rozšířený statistický paket, jakási minimální norma znalostí a nejhorší je, že u statistiků mizí tradiční vynalézavost a kutilství.

⁸⁴<<http://blogs.sas.com/jmp/index.php?/archives/-11-JMP-Genomics-3.0-is-into-production!.html>>

⁸⁵<<http://www.neurosolutions.com/products/nsbook/>>

⁸⁶<<http://www.statapol.cz/>>

- Aktivita uživatelů začíná převažovat nad možnostmi specializovaných firem. Řada firem dokonce vzdává vlastní vývoj a svůj software převádí na opensource. Kdysi to udělal Netscape, dnes SONY a rozumný software bez uživatelské podpory nenajdete. Na wiki stránce k produktu se dozvíme více než než na stránce výrobce a v širších souvislostech. I ve statistice je spousta diskutovaných a řešených problémů uživatelů ba i jejich softwaru na uživatelských stránkách.

4.1. Co by šlo udělat

To je hlavně na Vás, to nemůže udělat jeden.

Opensource minipaket Ve spolupráci s informatiky, obsahující pouze skutečně potřebné metody pro základní výuku a umožňující pluginy. Možnost se nabízí – diplomka u informatiků jako základ **opensource projektu**.

V Čechách i na Slovensku je dost programátorů, kteří s tím mají zkušenosti. Určitě to umí studenti informatiky a měli by to umět statistické, kteří se specializují na Výpočetní statistiku.

Statistické wiki stránky Šlo by vytvořit statistický wikislovník, wikibook atd. Pokud by chtěla alespoň malá skupina lidí spolupracovat, lze to založit. Přinejmenším by se omezila rostoucí terminologická džungle. Určitý návrh podal student (jak symptomatické) pod značkou [Glivi](#)⁸⁷ podle fungujícího [WikiProject Mathematics](#)⁸⁸.

Publikovat na webu Spousta článků má dočasný charakter, přibývají odkazy, barevné a dynamické obrázky, videa. To vše je možno snadno realizovat na internetu. Je třeba se jenom dohodnout, bylo by dobré, kdyby to zájemce našel vše co nejdříve na určitém místě.

Kdo není na internetu, jako by nebyl.

Z akcí České statistické společnosti se pravidelně vytvářejí CD, které obsahují texty referátů, prezentace atd. S určitým zpožděním se objevují i na stránce [www.statspol.cz](#)⁸⁹ (jsou zde všechny dosavadní sborníky z akcí společnosti a kompletní texty statistických bulletinů v pdf).

Bylo by vhodné, aby takto postupovali i další organizátoři, lze to zařídit i na stránce společnosti.

⁸⁷http://cs.wikipedia.org/wiki/Wikipedista:Glivi/Návrh_projektu_matematika

⁸⁸http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Mathematics

⁸⁹<http://www.statspol.cz/>

Výpočetní statistika Měl by být zaveden jiný způsob výuky. Výpočetní statistiky a navázána těsnější spolupráce s informatiky, aby statistika navázala na dynamiku vývoje. Statistikové musí znovu začít pracovat s daty, programovat a publikovat na současné úrovni a nebýt pouze mačkači knoflíků.

Řekl bych, že zejména současné pokolení studentů, zvyklé na počítače, hry a dynamiku, musí současná výuka Výpočetní statistiky spíše odrazovat (alespoň podle toho, co čtu na internetu).

Internetovská, průběžně inovovaná verze bude na adrese:
<http://www.stahroun.me.cz/eseje/stakan2007/index.htm>.

Adresa: Jiří Žváček, V úvalu 84, Nem. Motol, LDN, 7. stanice

E-mail: jzvacek@seznam.cz

SEZNAM AUTORŮ

BARTOŠOVÁ, Jitka	17
BĚLÁČEK, Jaromír	9
BOHDALOVÁ, Mária	27
DOLEJŠOVÁ, Miroslava	43
HEBÁK, Petr	55
HRBÁČEK, Pavel	74
CHAJDIÁK, Jozef	79
JARUŠKOVÁ, Daniela	247
KLÍMEK, Petr	82, 88, 95
KRÁL, Pavol	12, 108
MACHAČ, Otakar	207
MALÝ, Marek	115
MINÁROVÁ, Mária	142
MOHAMMED, Ahmad	142
MUNK, Michal	124
MURAKAMI, Hidetoshi	134
NÁNÁSIOVÁ, Olga	142
NEDELOVÁ, Gabriela	108
ŘEZANKOVÁ, Hana	149
STANKOVIČOVÁ, Iveta	162
STROUHAL, Jan	178
STŘÍŽ, Pavel	74, 182
TVRDÍK, Josef	185
URBANÍKOVÁ, Marta	200
VLČKOVÁ, Vladimíra	207
VOJTKOVÁ, Mária	216
VRÁBELOVÁ, Marta	124
WIMMER, Gejza	230
ZÁRUBA, Jan	247
ŽAMBOCHOVÁ, Marta	258
ŽVÁČEK, Jiří	270

OBSAH SBORNÍKU

PŘEDMLUVA	5
ROBUST 2008	6
 STAKAN 2007 V DROBNOHLEDU DVOU ČLÁNKŮ	
STAKAN 2007 – Dojmy a reflexe	
BĚLÁČEK, Jaromír	9
 STAKAN 2007	
KRÁL a kolektiv	12
 STAKAN 2007 PŘÍSPĚVKY	
Statistika na Fakultě managementu VŠE	
BARTOŠOVÁ, Jitka	17
 Dynamická versus klasická simulační metoda Monte Carlo	
BOHDALOVÁ, Mária	27
 Zařazení geografických informačních systémů do výuky předmětu Informatika ve veřejné správě	
DOLEJŠOVÁ, Miroslava	43
 Výuka statistiky 2007	
HEBÁK, Petr	55
 Spolupráce mezi ČSÚ a Univerzitou Tomáše Bati ve Zlíně	
HRBÁČEK, Pavel; STŘÍŽ, Pavel	74
 Niekoľko poznámok k výučbe základného kurzu statistiky	
CHAJDIK, Jozef	79

Data Mining and Software Tools	
KLÍMEK, Petr	82
Historie a současnost výuky statistiky na FaME, UTB ve Zlíně	
KLÍMEK, Petr	88
XLStatistics for Teaching Statistics in FaME, TBU in Zlín	
KLÍMEK, Petr	95
O vyučování viacrozmerých Štatistických metód na Školách ekonomického zamerania	
KRÁL, Pavol; NEDELOVÁ, Gabriela	108
K otázkám výuky statistických konzultací	
MALÝ, Marek	115
Vyučovanie štatistiky v nematematických odboroch	
MUNK, Michal; VRÁBELOVÁ, Marta	124
A Multivariate Two-sample Test in Nonparametric Methods	
MURAKAMI, Hidetoshi	134
Probability and Quantum Logic	
NÁNÁSIOVÁ, Olga; MINÁROVÁ, Mária; MOHAMMED, Ahmad	142
Výuka jednorozměrné a dvourozměrné analýzy kategorických dat	
ŘEZANKOVÁ, Hana	149
Viackriteriálne hodnotenie zamestnanosti členských krajín EÚ na základe vybraných ukazovateľov Lisabonskej stratégie	
STANKOVIČOVÁ, Iveta	162
Classification Trees in Software Reliability	
STROUHAL, Jan	178

Voluntary University Course: Computerised Data Processing	
STŘÍŽ, Pavel	182
Využitie štatistiky v poistnej matematike	
URBANÍKOVÁ, Marta	200
Grafy a tabulky ve statistice	
(aneb Na co ve výuce obvykle není čas)	
TVRDÍK, Josef	185
Zkušenosti s využitím Excelu při výuce Aplikované statistiky	
VLČKOVÁ, Vladimíra; MACHAČ, Otakar	207
Hľadanie podobnosti krajín EÚ podľa vybraných ukazovateľov	
Lisabonskej stratégie	
VOJTKOVÁ, Mária	216
Matematické modelovanie v jazykovede	
WIMMER, Gejza	230
Metoda latinských čtverců ve spolehlivostních úlohách	
ZÁRUBA, Jan; JARUŠKOVÁ, Daniela	247
Jak na rozhodovací stromy	
ŽAMBOCHOVÁ, Marta	258
Statistické výpočetní prostředí 2007	
ŽVÁČEK, Jiří	270

Stakan 2007 – Sborník příspěvků

Rusava, Česká republika, 25. – 27. května 2007

Editoři: Pavel STŘÍŽ, Gejza DOHNAL, Jaromír ANTOCH

Vydalo v XII/2007 nakladatelství Martin STŘÍŽ, Bučovice.

Počet stran publikace je 290. 1. vydání.

ISBN 978-80-87106-07-5 (DVD-ROM)

