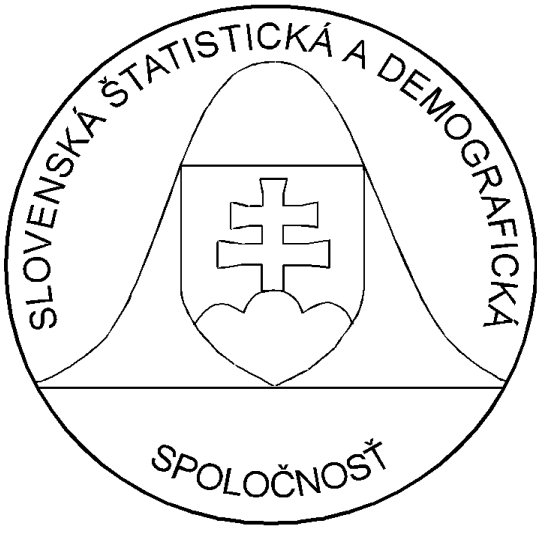
3/2005

FORUM STATISTICUM SLOVACUM



ISSN 1336-7420

Slovenská štatistická a demografická spoločnosť Miletičova 3, 824 67 Bratislava
www.ssds.sk

Naše najbližšie akcie:

(pozri tiež www.ssds.sk, blok Poriadané akcie)

Konferencia Pohľady na ekonomiku Slovenska 2006,
tematické zameranie: *Makroekonomický vývoj; daň z pridanej hodnoty*
4. 4. 2006, Bratislava

13. Slovenská štatistická konferencia,
tematické zameranie: *Štatistické metódy*
3. – 5. 5. 2006 alebo 20. – 22. 9. 2006 (podľa termínu konania volieb),
Malacky, Hotel ATRIUM

EKOMSTAT 2006 – jubilejná 20. škola štatistiky,
tematické zameranie: *Štatistické metódy v praxi*
21. – 26. 5. 2006, Trenčianske Teplice

PRASTAN 2006,
tematické zameranie: *Pravdepodobnosť, štatistika a numerika*
5. – 9. 6. 2006, Banskobystrický kraj

FERNSTAT 2006,
tematické zameranie: *Aplikovaná, demografická, matematická štatistika, štatistické riadenie kvality*
5. - 6. 10. 2006, Tajov pri Banskej Bystrici

11. Slovenská demografická konferencia,
tematické zameranie: *Migrácia*
3 dni, rok 2007, Košický kraj

Slávnostná konferencia 40 rokov SŠDS,
marec 2008, Bratislava

ÚVOD

Vážené kolegyne, vážení kolegovia

tretie nového časopisu, ktorý vydáva Slovenská štatistická a demografická spoločnosť (SŠDS) je zostavené z príspevkov, ktoré autori pripravili pre Soločnú Slovensko – Českú štatistickú konferenciu (učitelia štatistiky v praxi) PRASTAN 2005. Táto konferencia sa uskutočila v dňoch 10. až 15. júna v hoteli Lesák v Tajove. Tématické okruhy konferencie boli:

- problémy výučby štatistiky a pravdepodobnosti
- aplikácie matematickej štatistiky a pravdepodobnosti
- nové trendy v štatistike a pravdepodobnosti
- numerická matematika
- virtuálna univerzita, e-learning

Konferenciu organizovala Slovenská štatistická a demografická spoločnosť v spolupráci s Českou štatistickou spoločnosťou, Katedrou matematiky a deskriptívnej geometrie Stavebnej fakulty STU, Bratislava, Katedrou matematiky Fakulty prírodných vied UMB, Banská Bystrica a Fakultou managementu UK, Bratislava.

Programový výbor pracoval v zložení: Beloslav Riečan – predseda, členovia: Jaromír Antoch, Jozef Chajdiak, Martin Kalina, Jozef Komorník, Magda Komorníková, Ján Luha, Oľga Nánásiová, Roman Nedela, Jan Ámos Víšek, Gejza Wimmer. Organizačne k úspešnému priebehu konferencie prispeli Alžbeta Michalíková, Mária Minárová a Magda Renčová.

Na príprave a zostavení tohoto čísla participovali: Martin Kalina, Oľga Nánásiová a Mária Minárová.

Recenziu príspevkov zabezpečili: Mária Bohdalová, Tomáš Bacigál, Angela Handlovičová, Jana Kalická, Martin Kalina, Magda Komorníková, Pavol Kráľ, Alžbeta Michalíková, Mária Minárová, Oľga Nánásiová, Iveta Stankovičová, Jana Šiagiová, Katarína Trokanová, Peter Volauf, Viktor Witkovský, Ivan Žembery.

Výbor SŠDS

FORUM STATISTICUM SLOVACUM

Vydavateľ
Slovenská štatistická a demografická spoločnosť
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Redakcia
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Fax
02/63812565

e-mail
chajdiak@statis.biz
Jan.Luha@statistics.sk

Registráciu vykonalo
Ministerstvo kultúry Slovenskej republiky

Registračné číslo
3416/2005

Tematická skupina
B1

Dátum registrácie
22. 7. 2005

Objednávky
Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava 24
Slovenská republika
IČO: 178764
Číslo účtu: 0011469672/0900

ISSN 1336-7420

Ročník
I.

Číslo
3/2005

Cena výtlačku 500 SKK / 20 EUR
Ročné predplatné 1500 SKK / 60 EUR

Štatistické spracovanie biologických podmienok toku

Marek Ando, Andrej Škrinár

Ekologický stav toku je ovplyvnený mnohými faktormi, z ktorých k najdôležitejším patrí biotop fauny a flóry akvatickej oblasti toku, ďalej definovaný ako habitat. Štruktúra habitatu toku významnou mierou vplýva na organizáciu a štruktúru biologických spoločenstiev v ňom žijúcich. V snahe zabrániť zaplavovaniu územia sa používa niekoľko spôsobov úprav tokov. Porozumenie vplyvu dôsledkov ľudskej činnosti na štruktúru habitatu zostáva jednou z najzanedbanejších oblastí výskumu vo vodnom hospodárstve.
Jedným z dôležitých, ak vôbec nie najdôležitejších bioindikátorov povodia sú práve ryby. V spolupráci Katedry Vodného hospodárstva krajiny s Katedrou Matematiky sa pokúšame nájsť určitú homogenitu resp. závislosť výskytu rýb ako dôležitého charakterizačného faktora habitatu prostredia na hydrologických parametroch príslušného habitatu. Pre hodnotenie kvality habitatu akvatickej oblasti jednotlivých tokov bola použitá metodika IFIM Instream Flow Incremental Methodology – Prírastková metodika prúdenia v toku (IFIM) [1,2,3,4,]. Je to interdisciplinárny rozhodovací systém, ktorý vychádza z poznatku, že väčšina druhov rýb uprednostňuje isté kombinácie hĺbok, rýchlostí prúdenia, teploty vody a dnového materiálu. Ak sú tieto hodnoty pre rybie druhy v dotknutom úseku známe, dá sa pre každý určiť prognóza vplyvu zmien abiotických faktorov na biologické prostredie toku.

Výskum môžeme rozdeliť do dvoch etáp a síce na prácu v teréne a zber potrebných dát; a následné štatistické spracovanie nameraných údajov.

Zber dát spočíva vo výbere reprezentatívnych tokov na základe geologického podkladu prostredia. Ide vlastne o modelovanie kvality habitatu metodikou IFIM,ktorú možno rozdeliť do dvoch oblastí: 1. abiotická oblasť, ktorá predstavuje hlavne hydraulické modelovanie; 2..biotická oblasť do ktorej patrí predovšetkým ichtyologický výskum

V 1. dochádza ku komplexnej hydrologickej charakteristike vybraných tokov, čiže meraniu charakteristických hydraulických parametrov ako sú prietok, hĺbka a rýchlosť vody, šírka toku v hladine, hyd. polomer, drsnosť dna, atď. Meranie sa realizuje tak v intraviláne toku, čo je časť toku nachádzajúca sa priamo v obci a býva upravená, ako aj v extraviláne, čo je časť toku mimo obce, zachovávajúca si svoj pôvodný prirodzený charakter.

Po tomto môžeme pristúpiť k zberu údajov určených k štatistickému spracovaniu. To sa realizuje tak, že na danom toku sa spravidla trikrát za sebou urobí výlov rýb pomocou elektrického agregátu, pričom pri každom jednom ulovenom kuse sa zaznamenáva o aký druh ryby ide a v akej hĺbke a pri akej rýchlosti prúdenia vody bol ten ktorý druh chytený. Práve na základe týchto údajov sa vyhotovujú tzv. vhodnostné krivky výskytu pre jednotlivé druhy rýb zvlášť pre hĺbky a pre rýchlosti.

Jednou z dôležitých úloh štatistiky je skúmanie vzájomných vzťahov a súvislostí medzi jednotlivými javmi. Pri štatistickom skúmaní závislostí ide predovšetkým o príčinné (kauzálne) závislosti javov. O príčinnej závislosti sa dá hovoriť vtedy, keď jeden jav, alebo skupina javov (príčina, resp. komplex príčin) vyvoláva iný jav, resp. skupinu javov, pričom:

a) Vzťah medzi príčinou a účinkom môže byť jednostranný, to znamená že nemožno hovoriť o priamom spätnom pôsobení účinku na príčinu. V takom prípade hovoríme o jednostrannej závislosti.
b) Príčina účinok na seba v rámci určitých podmienok trvale vzájomne pôsobia a navzájom sa ovplyvňujú. V takomto prípade hovorírme o obojstrannej závislosti.

Popri kauzálnych vzťahoch, ktoré sú hlavným obsahom štatistického skúmania závislosti, sa štatistika zaoberá aj skúmaním iného typu závislostí. Nazvime ju združenosťou, keď ide o typ závislosti, pri ktorej nejde o príčinné vzťahy medzi danými javmi, pričom možno pozorovať, že určitej veľkosti, alebo obmene jedného javu spravidla zodpovedá nejaká veľkosť, alebo obmena iného javu.

Vonkajším prejavom jednotlivých javov, resp. ich vlastností sú štatistické znaky. Pri štatistickom skúmaní závislostí ide potom o meranie a kvantitatívne opísanie vzťahov medzi rôznymi štatistickými znakmi pomocou vhodných metód, pričom metódy, použiteľné pri meraní závislostí medzi kvantitatívnymi znakmi sa líšia od metód, ktoré sa používajú pri skúmaní závislostí medzi kvalitatívnymi znakmi.

Štatistické skúmanie závislosti medzi kvantitatívnymi znakmi sa často nazýva korelačný počet [5].

Sila závislosti medzi dvoma kvantitatívnymi premennými, čiže korelačný koeficient bola testovaná na deviatich reprezentatívnych tokoch flyšového charakteru pre viaceré druhy rýb. Testované parametre boli v tomto prípade hĺbka resp. rýchlosť v závislosti na kvázirovnomernej konštante M. Dôležitým faktorom bol aj výber vhodného reprezentatívneho druhu rýb, keďže nie všetky druhy sa vyskytovali na každom meranom toku, bolo nutné prihliadnuť aj na túto skutočnosť. Najpočetnejší výskytom na všetkých tokoch bol pstruh potočný. Koreláciou sme sa snažili zistiť či existuje nejaká určitá predpokladaná závislosť, alebo by sme skôr mohli hovoriť o vzťahu vhodnostných kriviek pre hĺbku (rýchlosť) a M, teda či na základe zmeny parametra M dokážeme odhadnúť ideálnu hĺbku (rýchlosť) výskytu danej ryby.

Korelačný koeficient bol vyhotovený zvlášť pre vzťah hĺbka – M a rýchlosť – M najskôr spoločne pre všetky toky, na ktorých sa zvolený druh (pstruh potočný) vyskytol. Ako prvé sme porovnali vrcholy vhodnostných kriviek, to znamená miesta s najväčším výskytom daného druhu v určitej hĺbke a pri určitej rýchlosti.

Korelačná tabuľka hĺbka – M

	hlbka	*pocet*	*M*
hlbka	1		
pocet	-0,3237	1	
M	0,135034	0,15279	1

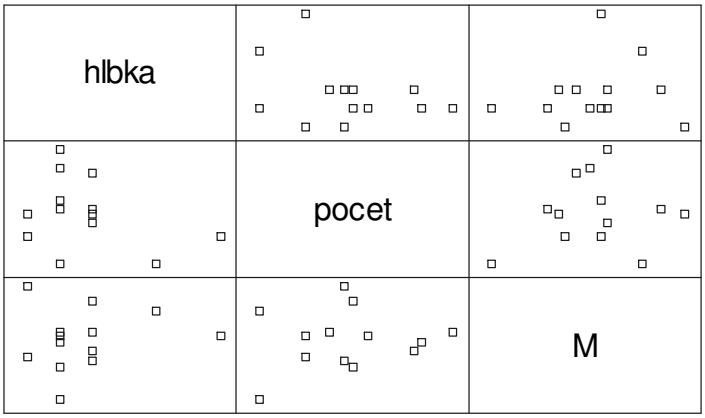
Korelačná tabuľka rýchlosť – M

	rychlost	*pocet*	*M*
rychlost	1		
pocet	0,379658	1	
M	0,113808	-0,05562	1

Korelačná tabuľka zobrazuje Pearsonov koeficient korelácie $r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{s_x s_y}$ medzi jednotlivými pármi premenných (hĺbka-počet, hĺbka-M; rýchlosť-počet, rýchlosť-M). Tieto korelačné koeficienty v rozsahu (-1, +1) vyjadrujú mieru závislosti lineárneho vzťahu medzi jednotlivými premennými.

Korelačný koeficient vzťahu hĺbka – M má hodnotu 0,135034 a vzťahu rýchlosť – M hodnotu 0,06225, čo môžeme považovať za závislosť celkom zanedbateľnú.

Grafické zobrazenie korelačnej matice: hĺbka - M



Korelačná tabuľka hĺbka – M

	hlbka	*pocet*	*M*
hlbka	1		
pocet	0,408248	1	
M	-0,2798	0,460057	1

Korelačná tabuľka rýchlosť – M

	rychlost	*pocet*	*M*
rychlost	1		
pocet	-0,19112	1	
M	0,644549	0,127323	1

Korelačný koeficient vzťahu hĺbka – M má hodnotu -0,2798, čo indikuje nepriamy vzťah a vzťahu rýchlosť – M hodnotu 0,644549, čo je vcelku veľká závislosť.

Prirodzené toky:

Korelačná tabuľka hĺbka – M

	hlbka	*pocet*	*M*
hlbka	1		
pocet	-0,5187	1	
M	0,524142	-0,20194	1

Korelačná tabuľka rýchlosť – M

	rychlost	*pocet*	*M*
rychlost	1		
pocet	0,295812	1	
M	0,125754	-0,59209	1

Korelačný koeficient vzťahu hĺbka – M má hodnotu 0,524142, čiže silu závislosti môžeme hodnotiť ako veľkú; a vzťahu rýchlosť – M hodnotu 0,125754, čo je malá závislosť.

Literatúra

[1] MACURA, V. – ČISTÝ, M.: Návrh optimálnych parametrov habitatu revitalizovaného toku. *Acta Horticulturae et regiotecturae*, 4, mim.číslo, 2001, 43 - 46.

[2] MACURA, V. - ŠKRINÁR, A.: Vplyv minimálnych prietokov na biologické prostredie akvatickej oblasti toku *Acta Horticulturae et regiotecturae*, Roč.4, č.2, 2004, 271 – 275.

[3] MACURA, V. - ŠKRINÁR, A, - ZAČKA, T.: Zovšeobecnenie vhodnostných kriviek rýb flyšových tokov na Slovensku a ich aplikácia pri modelovaní neovplyvnených úsekov tokov

v súlade s rámcovou smernicou EU. In: Vodní toky. Hradec Králové september 2004 ISBN 80-86386-55-4.

[4] MACURA, V. - ŠKRINÁR, A. – ZAČKA, T: Modelovanie biologických podmienok toku v súlade s rámcovou smernicou EÚ o vode. *In: Integrovaný manažment povodí a implementácia Rámcovej smernice EÚ o vode.* Bratislava apríl 2004 ISBN80-969136-0-3 s. 98-103,

[5] BAKYTOVÁ, H. – UGRON, M. – KONTŠEKOVÁ, O. : Základy štatistiky

Mgr. Marek Ando

Ing. Andrej Škrinár

Katedra vodného hospodárstva krajiny

Stavebná fakulta Slovenskej technickej univertity v Bratislave

Radlinského 11

813 68 Bratislava

E-mail: ando@.stuba.sk

andrej.skrinar@stuba.sk

Problémy spojené s inferenciou o variančnom komponente v zmiešaných lineárnych modeloch

Barbora Arendacká
Ústav merania SAV, Dúbravská cesta 9, 841 04 Bratislava
barendacka@gmail.com

1. Úvod

Pri hľadaní testov hypotéz a metód na konštrukciu konfidenčných intervalov pre variančný komponent v zmiešaných lineárnych modeloch sa obvykle osobitne uvažujú model s dvomi komponentami a model s viac komponentami. V prípade modelu s dvomi komponentami sú existujúce metódy početnejšie, pre špeciálny tvar hypotézy sú známe presné riešenia (takáto situácia vo všeobecnom prípade modelu s viac komponentami nenastáva). Okrem toho, že mnohé známe metódy(pre testovanie, či konštrukciu intervalov spoľahlivosti) sú približné, mnohé z nich sú použiteľné len v špeciálnych prípadoch zmiešaného lineárneho modelu(ako bude definovaný v nasledujúcej časti). V článku sa budeme podrobnejšie zaoberať práve príčinami popísaného stavu, ktorý je spôsobený problémami pri nachádzaní postačujúcich štatistík a prítomnosťou rušivých parametrov.

2. Zmiešaný lineárny model

Nech n-rozmerný náhodný vektor pozorovaní Y spĺňa nasledujúci model:

$$Y = X\beta + A_1\alpha_1 + A_2\alpha_2 + ... + A_k\alpha_k + \epsilon, \tag{1}$$

kde X je známa $n \times p$ matica, A_i, $i = 1, ..., k$ sú známe $n \times q_i$ matice, β je p-rozmerný vektor neznámych parametrov, α_i, $i = 1, ..., k$ sú navzájom nekorelované q_i-rozmerné normálne rozdelené náhodné vektory s nulovou strednou hodnotou a kovariančnou maticou $\sigma_i^2 I_{q_i}$, $\sigma_i^2 \geq 0$ (neznáme), a ϵ je n-rozmerný vektor náhodných chýb, $\epsilon \sim N_n(0, \sigma^2 I_n)$, $\sigma^2 > 0$(neznáme) a $cov(\alpha_i, \epsilon) = 0$, $i = 1, ..., k$. Za týchto predpokladov

$$Y \sim N_n(X\beta, \sum_{i=1}^{k} \sigma_i^2 V_i + \sigma^2 I_n), \tag{2}$$

kde $V_i = A_i A_i^T$. Ďalej predpokladáme, že $\mathcal{R}(A_i) \not\subseteq \mathcal{R}(X)$, $i = 1, ..., k$, kde $\mathcal{R}(A)$ označuje priestor generovaný stĺpcami matice A.

Za uvedených predpokladov sú náhodné vektory α_i vlastne charakterizované iba hodnotami parametrov σ_i^2, ktoré spolu s parametrom σ^2 nazývame variančné komponenty. Vektory α_i vnášajú do modelu variabilitu, ktorá sa do výsledkov experimentu dostáva výberom konkrétnych podmienok, za ktorých je experiment prevedený, z celej populácie možných podmienok. Napr. výberom konkrétnych objektov, na ktorých uskutočňujeme meranie. Skúmaním veľkosti variančných komponentov (ich testovaním, resp. konštruovaním informačne bohatších konfidenčných intervalov) je možné odhaľovať a posudzovať jednotlivé zdroje variability. Model (1) zahŕňa napr. ANOVA modely s náhodnými efektmi, zmiešanými efektmi(pevné efekty sú potom združené vo vektore β), ale tiež napr. lineárnu regresiu s chybami v regresných parametroch.

Keďže naším záujmom je testovať hypotézy o, resp. konštruovať intervaly spoľahlivosti pre niektorý z variančných komponentov $\sigma_1^2, ..., \sigma_k^2$ a v triede distribúcií

(2) nie sú tieto parametre ovplyvnené posunutím v strednej hodnote, t.j. grupou transformácií $\{y \longmapsto y + Xb, b \in R^p\}$, obvyklým postupom je redukcia modelu skonštruovaním maximálneho invariantu vzhľadom na spomínanú grupu posunutí, t.j. skonštruovaním vektora $Z = B_X^T Y$, kde, ak hodnosť matice X je p_1, B_X je $n \times (n - p_1)$ matica taká, že $B_X B_X^T = M_X = I_n - X(X^T X)^- X^T$ a $B_X^T B_X = I_{n-p_1}$. Pre náhodný vektor Z potom platí

$$Z \sim N_{n-p_1}(0, \sum_{i=1}^{k} \sigma_i^2 W_i + \sigma^2 I_{n-p_1}), \tag{3}$$

kde $W_i = B_X^T V_i B_X$. Ďalej budeme predpokladať, že $W_1, ..., W_k, I_{n-p_1}$ sú lineárne nezávislé.

Bez straty všeobecnosti môžme za parameter záujmu pokladať parameter σ_1^2 (jednotlivé variančné komponenty $\sigma_1^2, ..., \sigma_k^2$ môžme podľa potreby preznačiť), a teda podľa vyššie uvedeného, úlohou je nájsť testy, resp. metódy pre konštrukciu konfidenčných intervalov pre tento parameter v triede rozdelení (3). Obvyklým krokom v podobných situáciách je nájdenie minimálnych postačujúcich štatistík pre (3), na ktorých je potom založená všetka inferencia. Ako ďalej uvidíme, práve problémy pri konštruovaní postačujúcich štatistík spoluspôsobujú už spomínané delenie riešení na riešenia pre model s dvomi komponentami a pre model s viac komponentami, väčšiu rozpracovanosť riešení v prvej situácii a sú významnou prekážkou pri prechode od modelu s dvomi komponentami k všeobecnému prípadu modelu s viac komponentami.

3. Minimálne postačujúce štatistiky a ich vlastnosti

Označme $\bar{\sigma} = (\sigma_1^2, ..., \sigma_k^2, \sigma^2)$. Vektor Z v (3) má hustotu

$$f(z|\bar{\sigma}) = h(\bar{\sigma}) exp\{-\frac{1}{2} z^T (\sum_{i=1}^{k} \sigma_i^2 W_i + \sigma^2 I)^{-1} z\}. \tag{4}$$

Pri hľadaní postačujúcich štatistík pomocou vety o faktorizácii hustoty je teda potrebné nájsť inverziu kovariančnej matice $\Sigma_{\bar{\sigma}} = \sum_{i=1}^{k} \sigma_i^2 W_i + \sigma^2 I$. Seely [9] ukázal, že túto inverziu možno vyjadriť pomocou bázy kvadratického podpriestoru generovaného maticami $W_1, ..., W_k, I$.

Kvadratický podpriestor

Nech $\mathcal{A}$ označuje konečnorozmerný priestor reálnych symetrických matíc(príslušných rozmerov) so skalárnym súčinom definovaným $< A, B >= trAB, A, B \in \mathcal{A}$. Podpriestor $\mathcal{B}$ priestoru $\mathcal{A}$ sa nazýva kvadratický podpriestor, ak $B \in \mathcal{B}$ implikuje $B^2 \in \mathcal{B}$. (Podľa uvedenej definície je $\mathcal{A}$ tiež kvadratickým podpriestorom.) Dôležitou vlastnosťou kvadratického podpriestoru je, že ak matica $A \in \mathcal{B}$ a $A = \sum_{i=1}^{t} \delta_i P_i$ je jej spektrálny rozklad, kde $\delta_1, ..., \delta_t$ sú navzájom rôzne, nenulové vlastné čísla matice A a $P_1, ..., P_t$ sú im zodpovedajúce idempotentné, symetrické, po dvojiciach ortogonálne ($P_i P_j = 0, i \neq j$) matice, tak $P_1, ..., P_t$ tiež patria do $\mathcal{B}$. Z toho vylýva, že spolu s každou maticou $A \in \mathcal{B}$ do $\mathcal{B}$ patrí aj jej Moore-Penroseova pseudoinverzia A^+ a že v každom kvadratickom podpriestore existuje jeho báza zložená zo symetrických, idempotentných matíc. Ďalšie vlastnosti kvadratického podpriestoru sú popísané v [9]. Z pohľadu postačujúcich štatistík je dôležitá ešte jeho komutatívnosť. Kvadratický podpriestor je komutatívny, ak jeho prvky navzájom komutujú. Platí(pozri [9], lema 6), že kvadratický podpriestor je komutatívny práve vtedy, keď existuje jeho

báza, ktorú tvoria symetrické, idempotentné a po dvojiciach ortogonálne matice. Navyše, táto báza komutatívneho kvadratického podpriestoru je jediná. Ako ďalej uvidíme, práve komutatívnosť, resp. nekomutatívnosť kvadratického podpriestoru výrazne ovplyvňujú vlastnosti postačujúcich štatistík.

Postačujúce štatistiky

Vráťme sa k nášmu problému. Potrebujeme vyjadriť inverziu kovariančnej matice vektora Z. Ak označíme $\mathcal{B}_W = gen\{W_1, ..., W_k, I\}$ kvadratický podpriestor generovaný (reálnymi, symetrickými) maticami $W_1, .., W_k, I$, tak je zrejmé, že pre každú kombináciu variančných komponentov je kovariančná matica $\Sigma_{\bar{\sigma}}$ (ako lineárna kombinácia generujúcich prvkov) prvkom $\mathcal{B}_W$ a teda aj jej inverzia patrí do $\mathcal{B}_W$. Preto, ak $\{R_1, ..., R_r\}$ je báza $\mathcal{B}_W$, tak pre každé $\bar{\sigma}$

$$\left(\sum_{i=1}^{k} \sigma_i^2 W_i + \sigma^2 I\right)^{-1} = \sum_{i=1}^{r} f_i(\bar{\sigma}) R_i,$$

pre nejaké koeficienty $f_1(\bar{\sigma}), ..., f_r(\bar{\sigma})$. Po dosadení do (4) je zrejmé, že postačujúce štatistiky pre (3) sú $U_i = Z^T R_i Z, i = 1, ..., r$. Táto množina postačujúcich štatistík je zároveň minimálna, čo možno dokázať pomocou toho, že štatistika je minimálna postačujúca, ak sa jej hodnoty pre dva ľubovoľné body výberového priestoru z_1, z_2 rovnajú práve vtedy, keď podiel $f(z_1|\bar{\sigma})/f(z_2|\bar{\sigma})$ je konštanta ako funkcia $\bar{\sigma}$ ([3], str. 255) a z faktu, že lineárny priestor generovaný maticami $\{(I + \sum_{i=1}^{k} \theta_i W_i)^{-1}$, $0 \leq \theta_i \leq (k+1)^{-1}\}$ je najmenší kvadratický podpriestor obsahujúci lineárny priestor generovaný maticami $W_1, ..., W_k, I$ ([7], str.15). V prípade ak $r = k + 1$, Seely [9] dokázal, že U_i sú aj úplné. Vtedy vlastne bázu $\mathcal{B}_W$ tvoria matice $W_1, ..., W_k, I$.

Vlastnosti postačujúcich štatistík

Vlastnosti postačujúcich štatistík závisia na komutatívnosti modelu zodpovedajúceho kvadratického podpriestoru $\mathcal{B}_W$. Ak je $\mathcal{B}_W$ komutatívny, za $R_1, ..., R_r$ môžme zobrať matice tvoriace jeho symetrickú, idempotentnú, ortogonálnu bázu ($R_i R_j = 0$, $i \neq j$), ktorú vieme nájsť jednoduchým postupom, pomocou spektrálneho rozkladu jednotlivých matíc $W_1, ..., W_k, I$ a algoritmu popísaného v [9], lema 5 (pozri Prílohu A). Keďže $\Sigma_{\bar{\sigma}} \in \mathcal{B}_W$, $\Sigma_{\bar{\sigma}} = \sum_{i=1}^{k} \sigma_i^2 W_i + \sigma^2 I = \sum_{j=1}^{r} k_j(\bar{\sigma}) R_j$ a z vlastností rozdelenia kvadratických foriem normálne rozdeleného náhodného vektora je zrejmé, že

$$U_i \sim k_i(\bar{\sigma}) \chi^2_{\nu_i}, \quad \nu_i = tr(R_i), \quad i = 1, ..., r,$$

kde $k_i(\bar{\sigma}) = \sum_{j=1}^{k} \gamma_i^{W_j} \sigma_j^2 + \sigma^2$, kde $W_i = \sum_{j=1}^{r} \gamma_j^{W_i} R_j, i = 1, ..., k$ a $I = \sum_{i=1}^{r} R_i$ a štatistiky U_i sú navzájom nezávislé. (Modely, pre ktoré je $\mathcal{B}_W$ komutatívny a $dim(\mathcal{B}_W) = k + 1$, nazývame regulárne.)

V prípade nekomutatívnosti $\mathcal{B}_W$ je jednak vo všeobecnosti zložité nájsť bázu tohto kvadratického podpriestoru (riešenie v niektorých špeciálnych prípadoch je uvedené v [5], str. 92-93), jednak, keďže neexistuje jeho báza, ktorej prvky by boli súčasne idempotentné a ortogonálne, zostrojené postačujúce štatistiky nemusia byť vo všeobecnosti nezávislé a jednoduchosť ich rozdelenia sa tiež stráca.

4. Dôsledky uvedeného a rušivé parametre

Z uvedeného je zrejmé, že v prípade komutatívneho kvadratického podpriestoru zodpovedajúceho modelu nastáva pomerne priaznivá situácia pre hľadanie testovacích štatistík a metód pre konštrukciu konfidenčných intervalov pre variančný

komponent σ_1^2. Práve tento fakt je jednou z príčin v úvode spomínaného delenia známych metód na metódy pre model s dvomi komponentami a pre model s viac komponentami. Model s dvomi komponentami má totiž vždy zodpovedajúci kvadratický podpriestor $\mathcal{B}_W = gen\{W_1, I\}$ komutatívny, a teda vždy vieme nájsť množinu nezávislých, χ^2 rozdelených postačujúcich štatistík, na ktorej môžme založiť celú inferenciu. Bázu $\mathcal{B}_W$ v tomto prípade tvoria symetrické, idempotentné a po dvojiciach ortogonálne matice $E_t, t \in \tau$ zo spektrálneho rozkladu matice W_1 zodpovedajúce jej navzájom rôznym nenulovým vlastným číslam a matica $E = I - \sum_{t\in\tau} E_t$. Problémom, s ktorým sa zostáva potýkať je prítomnosť rušivého parametra σ^2.

V prípade modelov s viac ako dvomi variančnými komponentami sa stretávame s modelmi, pre ktoré je $\mathcal{B}_W$ komutatívny, aj s modelmi, pre ktoré je $\mathcal{B}_W$ nekomutatívny. Do prvej skupiny patria napríklad vyvážené ANOVA modely rôznych typov, ktoré sú väčšinou navyše aj regulárne. Postačujúce štatistiky U_i sa potom zhodujú so sumami štvorcov zo známej tabuľky analýzy rozptylu. Stačí však uvažovať nevyvážené ANOVA modely a vo všeobecnosti dostávame situáciu s nekomutatívnym modelu zodpovedajúcim kvadratickým podpriestorom (isté triedy tzv. čiastočne vyvážených modelov, ktoré sú ale regulárne, sú popísané v [12]). Nezanedbateľná je aj prítomnosť rušivých parametrov, ktorých je väčší počet a sú komplikovanejšie previazané ako v prípade modelu s dvomi komponentami. Preto napríklad, kým pre testovanie nulovosti σ_1^2 existujú v modeli s dvomi komponentami presné testy, v modeli s viac komponentami sú vo všeobecnosti známe iba testy približné, a to aj v prípade regulárnych modelov (pozri [4, 10]). Rušivé parametre sú napr. aj príčinou toho, že známe metódy v prípade konštrukcie konfidenčných intervalov sú len približné dokonca aj v modeloch s dvomi komponentami. (V modeloch s viac komponentami sú známe len pre špeciálny typ modelov - ANOVA modely (aj keď vyvážené aj nevyvážené)(pozri [2]).)

Ťažkosti spojené s inferenciou v modeloch s viac komponentami sa niekedy riešia redukciou modelu na model iba s dvomi komponentami, ktorý predstavuje jednoduchšiu situáciu. Ide o postupné uplatňovanie vhodných projekcií, ktoré vylučujú z modelu variančné komponenty, ktoré nás nezaujímajú. Podrobný opis tejto metódy možno nájsť napr. v [10]. Tento postup obchádza problém s postačujúcimi štatistikami a zároveň znižuje počet rušivých parametrov. Jeho nevýhodou ale je, že pri projekciách dochádza k znižovaniu dimenzie problému, takže stav s iba dvomi komponentami (a dimenziou väčšou ako 1) nemusí byť dosiahnuteľný. (Ako uvidíme, dimenzia závisí od hodnosti matíc vystupujúcich pri variančných komponentoch, ktorú jednotlivé projekcie môžu neželane znížiť.)

Pre ilustráciu vhodných projekcií uvedieme príklad redukcie v modeli (1) s tromi komponentami. Po redukcii na maximálny invariant dostávame:

$$Z \sim N_{n-h(X)}(0, \sigma_1^2 W_1 + \sigma_2^2 W_2 + \sigma^2 I),$$

kde $W_1 = B_X^T A_1 A_1^T B_X, W_2 = B_X^T A_2 A_2^T B_X$. Vylúčiť z modelu niektorý z rušivých parametrov môžme jedným z nasledujúcich dvoch krokov:

1. ak $\mathcal{R}(B_X^T A_1) \not\subseteq \mathcal{R}(B_X^T A_2)$, môžme použiť projektor $M_{B_X^T A_2} = I - B_X^T A_2 (A_2^T M_X A_2)^- A_2^T B_X$, resp. jeho rozklad $M_{B_X^T A_2} = B_2 B_2^T$, $B_2^T B_2 = I_d$, $d = n - h(X) - h(B_X^T A_2)$ a transformáciou $B_2^T Z$ vylúčime z modelu komponent σ_2^2, dimenzia sa zníži o hodnosť matice $B_X^T A_2$. ($Var(B_2^T Z) = \sigma_1^2 B_2^T W_1 B_2 + \sigma^2 I_d$.)

2. ak podmienka $\mathcal{R}(B_X^T A_1) \not\subseteq \mathcal{R}(B_X^T A_2)$ nie je splnená, nemožno uplatniť projekčnú maticu $M_{B_X^T A_2}$ bez straty komponentu σ_1^2 z modelu. Je ale možné použiť projekčnú maticu $P_{B_x^T A_2} = I - M_{B_X^T A_2}$, resp. jej rozklad, a k transformovanému vektoru Z ešte pripočítať vhodne upravený vektor rezíduí (podrobnosti viď [10]), čím sa odstráni komponent σ^2. Variančná matica tak bude tvorená zložkami $\sigma_1^2\tilde{W}_1$ a $(\sigma_2^2 + c\sigma^2)\tilde{W}_2$, resp. $\sigma_1^2\tilde{W}_1$ a $\tilde{\sigma}^2\tilde{W}_2$. Aby sme dospeli k modelu (3) s dvomi komponentami$(\sigma_1^2, \tilde{\sigma}^2)$, je nutné ešte uplatniť transformáciu, ktorá prevedie maticu $\tilde{W}_2$ na jednotkovú. Tou transformáciou je prenásobenie modelu maticou B^T, takou, že $B^T\tilde{W}_2 B = I_d$. d, čo je dimenzia problému po výslednej transformácii, je v tomto prípade rovné hodnosti matice $\tilde{W}_2$.

V prípade modelov s väčším počtom komponentov sa uvedené dva kroky viacnásobne opakujú. Vzhľadom na to, že hodnosť matíc W_i, resp. $\tilde{W}_i$, je obvykle malá, k najväčšej redukcii dimenzie dochádza pri uplatnení kroku 2.

Vráťme sa ešte k problému rušivých parametrov. Ako aj v iných úlohách, kde sa vyskytujú, ich prítomnosť sťažuje konštrukciu testovacích štatistík, resp. pivotov (v prípade konfidenčných intervalov), ktorých rozdelenie nezávisí na rušivých parametroch. Uplatnením princípov zovšeobecnenej inferencie však možno tieto obtiaže prekonať. Pojem zovšeobecnenej inferencie zaviedli Tsui a Weerahandi [11] a ďalej rozpracoval Weerahandi [13, 14]. Ide o inferenciu, ktorá sa robí podmienene na napozorovaných dátach, keďže testovacie štatistiky, či pivoty môžu od napozorovaných dát závisieť. Práve začlenenie napozorovaných dát do vyjadrenia týchto kvantít výrazne zjednodušuje konštrukciu výrazov s distribúciou, ktorá pri pevných dátach nezávisí na rušivých parametroch. Zovšeobecnené testy a konfidenčné intervaly sú potom odvodené z presných pravdepodobnostných tvrdení, ktoré sú však iba podmienené(napozorovanými dátami) a dodržanie predpísanej hladiny významnosti, resp. konfidenčnej úrovne sa musí overovať simulačne, keďže ich skutočná hladina (úroveň) nie je explicitne známa. V mnohých situáciách sa však javia ako vhodné riešenie. Aplikácie princípov zovšeobecnenej inferencie na úlohy o variančných komponentoch v zmiešaných lineárnych modeloch možno nájsť v [14, 8, 6, 1]. Tak ako sú uvedené v týchto prácach, zovšeobecnené testy a konfidenčné intervaly sa však opierajú o znalosť postačujúcich štatistík.

5. Zhrnutie

V predchádzajúcich riadkoch sme načrtli problémy, ktoré sťažujú inferenciu o variančnom komponente v zmiešaných lineárnych modeloch (1). Ide o ťažkosti pri nachádzaní postačujúcich štatistík a prítomnosť rušivých parametrov. Kým komplikácie spôsobené rušivými parametrami možno obísť pomocou zovšeobecnenej inferencie, problém s postačujúcimi štatistikami je závažnejší a komplikuje prechod od modelu s dvomi komponentami k modelu s viac komponentami. Testy alebo veličiny pre konštrukciu konfidenčných intervalov vo všeobecnom prípade zmiešaného lineárneho modelu bude preto asi treba založiť na nejakých iných, vhodných kvadratických formách vektora pozorovaní, prípadne na nejakej aproximácii modelu. V úplne všeobecnom prípade, bez použitia redukcie, sú doteraz známe len testy nulovosti σ_1^2 odvodené lokálne, pri konkrétnych hodnotách rušivých parametrov, metódy pre konštrukciu konfidenčných intervalov nie sú známe vôbec.

Poďakovanie

Práca bola podporená grantom VEGA 1/0264/03 Vedeckej grantovej agentúry

Slovenskej republiky.

A Ortogonálna idempotentná báza v $\mathcal{B}_W$

Ak je $\mathcal{B}_W$ komutatívny kvadratický podpriestor, vieme nájsť jeho ortogonálnu, idempotentnú bázu nasledujúcim postupom.

1. nájdeme spektrálny rozklad matíc $W_1, ..., W_k$: $W_i = \sum_{j=1}^{r_i} \lambda_j^{W_i} F_j^{W_i}$,

2. z množiny matíc $\{I, F_j^{W_i}, i = 1, ..., k, j = 1, ..., r_i\}$ a všetkých možných súčinov z nich utvorených dvojíc, trojíc,..., k-tic vyberieme lineárne nezávislé prvky, ktoré budú tvoriť idempotentnú bázu komutatívneho kvadratického podpriestoru $\mathcal{B}_W$,

3. idempotentnú bázu z predchádzajúceho kroku ortogonalizujeme algoritmom popísaným Seelym (1971), lema 5:

   Ak $R_1, ..., R_m, R$ sú nenulové, symetrické, idempotentné matice také, že $R_1, ..., R_m$ sú po dvojiciach ortogonálne, R s nimi komutuje a je s nimi lineárne nezávislá, potom matice

   $$P_i = RR_i, \quad i = 1, ..., m$$

   $$P_{m+i} = R_i - RR_i, \quad i = 1, ..., m$$

   $$P_{2m+1} = R - R(R_1 + ... + R_m)$$

   sú symetrické, idempotentné, po dvojiciach ortogonálne a existuje medzi nimi aspoň $m + 1$ lineárne nezávislých matíc. Tiež priestor generovaný maticami $R_1, ..., R_m, R$ je podpriestorom priestoru generovaného maticami $P_1, ..., P_{2m+1}$.

   Teda, ak $C_1, ..., C_r$ je idempotentná báza komutatívneho kvadratického podpriestoru $\mathcal{B}_W$, tak podľa predchádzajúceho vieme z matíc

   $$C_1C_2 \quad C_1 - C_2 \quad C_2 - C_1C_2 \quad C_3, ..., C_r$$

   vybrať bázu $D_1, ..., D_r$ pre $\mathcal{B}_W$ takú, že $D_1D_2 = 0$. Ďalej, z matíc

   $$D_1D_3, D_2D_3, D_1 - D_1D_3, D_2 - D_2D_3, D_3 - D_3(D_1 + D_2), D_4, ..., D_r$$

   môžme vybrať bázu $G_1, ..., G_r$ pre $\mathcal{B}_W$ takú, že $G_1G_2 = G_2G_3 = G_1G_3 = 0$. Pokračujúc podobne ďalej dostaneme nakoniec ortogonálnu, idempotentnú bázu pre $\mathcal{B}_W$.

Referencie

1. Arendacká, B.: Generalized confidence intervals on the variance component in mixed linear models with two variance components, Statistics 39(2005), pp. 275-286

2. Burdick, R.K., Graybill F.A.: Confidence Intervals On Variance Components, Marcel Dekker Inc., New York, 1992

3. Casella, G., Berger, R.L.: Statistical Inference, Duxbury Press, 1990

4. Fonseca, M., Mexia, J.T., Zmyślony, R.: Estimators and tests for variance components in cross nested orthogonal designs, Discussiones Mathematicae Probability and Statistics 23(2003), pp. 175-201

5. Gnot, S.: Estymacja komponentów wariancyjnych w modelach liniowych, Teoria i zastosowania, WNT, Warszava, 1991

6. Park, D.J., Burdick, R.K.: Performance of confidence intervals in regression models with unbalanced one-fol nested error structures, COMMUNICATIONS IN STATISTICS Simulation and Computation 32(2003), pp. 717-732

7. Rao, C.R., Kleffe, J.: Estimation of Variance Components and Applications, North-Holland series in statistics and probability, Amsterdam, 1988

8. Sárköziová, Z.: Confidence intervals for variance components in balanced ANOVA models, Tatra Mountains Mathematical Publications 22(2001), pp. 149-157

9. Seely, J.: Quadratic spaces and completeness, The Annals of Mathematical Statistics 42(1971), pp. 710-721

10. Šírková, L., Witkovský, V.: On testing variance components in unbalanced mixed linear model, Applications of Mathematics 46(2001), pp. 191-213

11. Tsui, K.W., Weerahandi, S.: Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters, Journal of the American Statistical Association 84(1989), pp. 602 - 607

12. Vanleeuwen, D.M., Birkes, D.S., Seely, J.F.: Balance and orthogonality in designs for mixed classification models, The Annals of Statistics 27(199), pp. 1927-1947

13. Weerahandi, S.: Generalized confidence intervals, Journal of the American Statistical Association 88(1993), pp. 899-905

14. Weerahandi, S.: Exact Statistical Methods for Data Analysis, Springer-Verlag, New York, 1995

15. Zhou, L., Mathew, T.: Some tests for variance components using generalized p-values, Technometrics 36(1994), pp. 394-402

Odhad časového zatížení vysokoškolského učitele dílčími činnostmi (problémy vyhodnocení dotazníkového šetření)

Jitka Bartošová
University of Economics, Czech Republic

Abstrakt: Cílem příspěvku je podat informaci o problematice odhadu parametru polohy zatížení vysokoškolského pracovníka – vědce jednotlivými činnostmi. Výběrový soubor, použitý k odhadu uvedeného parametru, vznikl agregací výsledků předběžného, hlavního a dodatečného dotazníkového šetření, které bylo provedeno řešiteli grantového projektu GAČR č. 402/03/1341 na dvou vysokoškolských pracovištích – na Vysoké škole ekonomické a na Jihočeské univerzitě v průběhu roku 2003.

Klíčová slova: bodový a intervalový odhad, dotazníkové šetření, parametr polohy.

Úvod

Pro zdárné vyřešení problému znormování výkonů vysokoškolského učitele je nezbytné určit jeho časové zatížení jednotlivými dílčími činnostmi tak, aby mohlo být určeno jeho relativní časové zatížení každou ze sledovaných položek. Požadovanou časovou charakteristiku získáme vhodným bodovým (intervalovým) odhadem parametru polohy rozdělení výběrových souborů, které obsahují údaje o časové náročnosti dílčích činností VŠ pedagoga. Výsledkem je jednoznačné přiřazení množiny časových hodnot k množině dílčích činností, jejíž prvky (činnosti) jsou předem dány požadavky grantového projektu. Charakteristiky polohy výběrových rozdělení mohou vystihovat buď centrální nebo extremální tendenci dat. Z hlediska požadavků grantového projektu se zde zaměříme pouze na určení vhodné „centrální“ míry polohy rozdělení. K „centrálním“ měrám obecně řadíme jak momentové typy charakteristik, tj. různé průměry (včetně robustních verzí – např. useknutého či winsorizovaného průměru), tak kvantilové typy charakteristik, tj. medián, BES odhad, Gastwirthův odhad, polosumu, pivotovou polosumu apod. K realizaci odhadu charakteristik lze použít rovněž více metod, např. metodu maximální věrohodnosti, metodu nejmenších čtverců, kvantilovou metodu. Výběr z uvedené nabídky charakteristik polohy a metod jejich odhadů je dán konkrétním charakterem datového souboru. Základním úkolem statistické analýzy daných výběrových souborů je tedy volba a realizace nejlepšího odhadu „centrální“ hodnoty časového zatížení vysokoškolského pedagoga jednotlivými dílčími činnostmi.

Metodika řešení

Problém bodového (intervalového) odhadu „centální“ hodnoty časového zatížení vysokoškolského pedagoga jednotlivými dílčími činnostmi je součástí obecné problematiky odhadu parametrů výběrového rozdělení. Pro charakterizaci polohy (popřípadě variability a tvaru) výběrového rozdělení bývají většinou používány momentové míry. Tyto míry jsou však vesměs velice citlivé na výskyt odlehlých hodnot. Proto, pokud jsou např. v datech identifikovány odlehlé nebo extrémní hodnoty, je vhodné požít k charakterizaci některou z robustních měr (viz např. ANTOCH a VORLÍČKOVÁ, 1992, BARTOŠOVÁ 2003 a, b, JUREČKOVÁ, 2001). Momentovou mírou polohy výběrového souboru, kterou lze vystihnout střední hodnotu časového zatížení VŠ pedagogů jednotlivými pedagogickými i nepedagogickými činnostmi, je výběrový průměr $\bar{x}$ (viz BARTOŠOVÁ a STEJSKALOVÁ, 2003, STEJSKALOVÁ a BARTOŠOVÁ, 2003, STEJSKALOVÁ, ROLÍNEK a BARTOŠOVÁ, 2003). Výběrový průměr je však nejlepším odhadem parametru polohy pouze za předpokladu, že výběr není „malý“ a je pořízen ze souboru, která mí symetrické rozdělení a není kontaminován odlehlými pozorováními. Výběrový průměr je odhadem, který není B-robustní, to znamená, že je vysoce citlivý na výskyt odlehlých hodnot. Jeho asymptotickým bodem zvratu je 0.

Velmi vhodná a na výpočty jednoduchá je třída kvantilových výběrových charakteristik (viz např. BLATNÁ, 1994). Jedná se o vysoce robustní míry, které můžeme použít i v případě, že rozdělení souboru je neznámé. Tyto míry však nemusí být nejlepšími výběrovými charakteristikami polohy (popřípadě variability a tvaru) výběrového rozdělení. Pokud se prokáže, že výběrový soubor pochází z některého známého rozdělení, pak nejlepší metodou odhadu parametrů tohoto rozdělení je maximálně věrohodný odhad. Z nabídky kvantilových charakteristik polohy, kterými lze vystihnout „centrální“ časové zatížení VŠ pedagogů jednotlivými dílčími činnostmi, můžeme jmenovat např. medián, BES-odhad, Gastwirthův odhad, polosumu, pivotovou polosumu a další (viz např. BLATNÁ, 1999). Medián je ze všech měr polohy nejvíce B-robustní, jeho

asymptotickým bodem zvratu je $\frac{1}{2}$. Avšak v případě, že výběrové rozdělení není symetrické, není již medián nejvhodnější kvantilovou výběrovou charakteristikou polohy. V takovém případě vystihují centrální polohu souboru lépe některé další kvantilové míry, jako jsou např. BES-odhad, Gastwirthův odhad, pivotová polosuma apod.

Vzhledem k tomu, že četnosti realizací jednotlivých položek, získaných z dotazníkového šetření, jsou převážně „malé“ (od 4 do 20 hodnot), je vhodné použít k určení parametru polohy a variability Hornův postup. Pro odhad parametru polohy bude tedy vhodné použít pivotovou polosumu P_L a pro odhad variability pivotové rozpětí R_L. Pivotová polosuma je kvantilová charkateristika polohy, která je tvořena lineární kombinací horního a dolního pivotu. Pro rozsah výběru od 4 do 20 hodnot jsou pro potřeby určení intervalového odhadu parametru polohy publikovány vybrané kvantily $t_{L(1-\frac{\alpha}{2})}(n)$ rozdělení statistiky $T_L = \frac{P_L}{R_L}$ v práci (MELOUN, MILITKÝ 2002, s. 147). Např. 95%-ní oboustranný interval spolehlivosti „centrální“ hodnoty časového vytížení VŠ pedagogů jednotlivými dílčími činnostmi, můžeme vyjádřit vztahem:

$$P_L - t_{L(0,975)}(n) \cdot R_L \le \mu \le P_L + t_{L(0,975)}(n) \cdot R_L .$$

Avšak pokud jsou četnosti realizací jednotlivých položek, získaných z dotazníkového šetření, „velmi malé“ ($n_i < 4$), nelze vyvozovat z výsledků šetření žádné obecné závěry (viz MELOUN, MILITKÝ 2002, s. 146).

Charakteristika datových soborů

V rámci grantového projektu GAČR č. 402/03/1341 byly na dvou veřejných vysokých školách v České republice – Jihočeské univerzitě a Vysoké škole ekonomické – realizovány v průběhu roku 2003 postupně dva dotazníkové průzkumy, týkající se rozdělení časového zatížení vysokoškolských pedagogů. Na základě těchto šetření byly získány soubory informací o časovém zatížení pedagogů jednak z pilotního výzkumu (12 dotazníků ze Zemědělské fakulty JU a 18 dotazníků z Fakulty informatiky a statistiky VŠE v Praze) a jednak z následného hlavního výzkumu (29 dotazníků ze Zemědělské fakulty JU a 22 dotazníků z Fakulty managementu VŠE). Uvedené výběrové soubory tvoří datovou základnu pro realizaci odhadů charakteristik polohy i dalších statistických analýz.

Pilotní a hlavní šetření bylo provedeno na dvou disjunktních podmnožinách množiny všech vysokoškolských pedagogů JU a VŠE a mělo náhodný charakter. Výsledky obou šetření proto mohly být pro účely statistické analýzy agregovány do jednoho souboru, obsahujícího 40 dotazníků z VŠE a 41 dotazníků z JU. Důvodem k agregaci dat byla snaha o maximální zvýšení četností realizací jednotlivých položek pro účely statistického vyhodnocování. Vzhledem k nedostatečným četnostem odpovědí v oblasti podávání a řešení grantů, bylo provedeno na Fakultě managementu VŠE ještě dodatečné dotazníkové šetření. Toto šetření bylo realizováno formou nenáhodného výběru z množiny všech pedagogů Fakulty managementu tak, aby všechny provedené výběry byly disjunktní. Na základě tohoto šetření byl soubor informací o časovém vytížení pedagogů při podávání a řešení grantových projektů doplněn o další hodnoty.

Komplexní statistická analýza obsahovala vyhodnocení celkem 73 dílčích činností vysokoškolských pedagogů, rozdělených do několika skupin podle typu činnosti na

1. pedagogické činnosti (celkem 42 dílčích činností),
   a) činnosti v denní formě studia (30 dílčích činností),
   b) činnosti v kombinované formě studia (12 dílčích činností),
2. publikační a vědecké činnosti (celkem 31 dílčích činností).¨

Tabulka 1: Procentuální zastoupení četností informací, získaných z dotazníkových šetřeních na VŠE a JU, které lze hodnotit jako „velmi malé“, „malé“ a „ostatní“.

Druh činností	podíl na VŠE	podíl na JU	podíl na VŠE	podíl na JU	podíl na VŠE	podíl na JU
	$n_i < 4$	$n_i < 4$	**$4 \le n_i \le 20$**	**$4 \le n_i \le 20$**	$n_i > 20$	$n_i > 20$
Pedagogické v PS	13,3%	20,0%	**53,4%**	**46,7%**	33,3%	33,3%
Pedagogické v KS	**58,3%**	**75,0%**	41,7%	25,0%	0,0%	0,0%
Nepedagogické	45,2%	35,5%	**54,8%**	**64,5%**	0,0%	0,0%
Celkem	34,2%	35,6%	**52,1%**	**50,7%**	13,7%	13,7%

Četnosti informací o dílčích pedagogických činnostech v prezenční formě studia (PS), které byly získány na VŠE (na JU), se pohybují od 1 do 34 (od 0 do 32) hodnot. V tabulce 1 jsou přehledně uspořádány informace o procentuálním zastoupení četností odpovědí, které lze hodnotit jako „velmi malé“ ($n_i < 4$), „malé“ ($n_i \in \langle 4,20 \rangle$) a „ostatní“ ($n_i > 20$). Z tabulky 1 vyplývá, že v intervalu $\langle 4,20 \rangle$ se nachází četnosti informací získaných na VŠE (na JU) v 53,4% případů (v 46,7% případů). Méně než 4 hodnoty byly získány na VŠE (na JU) v 13,3% případů (20,0% případů) a více než 20 hodnot bylo získáno na VŠE i JU shodně v 33,3% případů.

Četnosti informací o dílčích pedagogických činnostech v kombinované formě studia (KS), které byly získány na VŠE (na JU), se pohybují od 0 do 12 (od 0 do 8) hodnot. V intervalu $\langle 4,20 \rangle$ se nachází četnosti informací získaných na VŠE (na JU) v 41,7% případů (v 25,0% případů). Méně než 4 hodnoty byly získány na VŠE (na JU) v 58,3% případů (75,0% případů) a více než 20 hodnot nebylo získáno na žádně z těchto vysokých škol.

Četnosti informací o dílčích nepedagogických činnostech (o dílčích publikačních a vědeckých činnostech), které byly získány na VŠE (na JU), se pohybují od 0 do 20 (od 0 do 18) hodnot. V intervalu $\langle 4,20 \rangle$ se nachází četnosti informací získaných na VŠE (na JU) v 54,8% případů (v 64,5% případů). Méně než 4 hodnoty byly získány na VŠE (na JU) v 45,2% případů (35,5% případů) a více než 20 hodnot bylo získáno na VŠE (na JU) v 0,0% případů (0,0% případů).

To znamená, že celkem se v intervalu $\langle 4,20 \rangle$ nachází četnosti informací získaných na VŠE (na JU) v 52,1% případů (v 50,7% případů). Méně než 4 hodnoty byly získány na VŠE (na JU) v 34,2% případů (35,6% případů) a více než 20 hodnot bylo získáno na VŠE i JU shodně v 13,7%. Z uvedených výsledků (viz tabulka 1) plyne závěr, že v převážné většině dílčích činností se procentuální zastoupení získaných informací pohybuje v intervalu $\langle 4,20 \rangle$, což lze ohodnotit jako „malé“ zastoupení. Pouze v oblasti pedagogických činností v kombinované formě studia převažuje „velmi malé“ procentuální zastoupení.

Závěry

Z této analýzy vyplývá, že hlavním problémem komplexního a detailního zmapování časového zatížení vysokoškolského pedagoga – vědce, jsou „velmi malé“ četnosti odpovědí v některých oblastech dílčích činností, a to především v oblasti časového vytížení vysokoškolských učitelů pedagogickými činnostmi v kombinované formě studia a některými nepedagogickými činnostmi. Tyto „nedostatečné“ četnosti realizací některých položek nám nedovolují činit z dosažených výsledků žádné závěry.

Rovněž „malé“ četnosti odpovědí, které se vyskytují u nadpoloviční většiny všech sledovaných položek, vyžadují opatrnost při zobecňování závěrů provedené analýzy. „Malé“ četnosti odpovědí také nedovolují dále zvyšovat detailnost a komplexnost zmapování činností VŠ pedagoga. Další detailnější roztřídění stávajících položek by totiž vedlo převážně ke vzniku položek s „velmi malými“ četnostmi.

V návaznosti na uvedené skutečnosti byla za jednotný odhad parametru polohy rozdělení časového vytížení vysokoškolského pedagoga – vědce zvolena pivotová polosuma, která má optimální vlastnosti vzhledem ke konkrétním analyzovaným datovým souborům. Tato metoda umožňuje realizaci bodového i intervalového odhadu parametru polohy rozdělení na souborech s četnostmi realizací v intervalu $\langle 4,20 \rangle$, tj. v nadpoloviční většině všech sledovaných případů.

Idemtifikace odlehlých pozorování pomocí krabičkového diagramu prokázalo kontaminaci agregovaných datových souborů odlehlými hodnotami. Vesměs se jednalo o hodnoty odlehlé shora. Tyto hodnoty se mohou vyskytovat v souborech ze dvou odlišných důvodů. Jedná se buď o chyby, vzniklé špatným vyplněním dotazníků, anebo o hodnoty, které ukazující na skutečnou extrémní časovou zátěž pedagoga. Příčinou může být např. spocializace zkoumané problematiky, náročnost psaného textu, popřípadě složitost grafického zobrazení apod.

Na základě problémů, vzniklých při statistické analýze datových souborů, pocházejících z dotazníkového šetření provedeného v rámci projektu GAČR, lze formulovat doporučení, které by mělo přispět ke zlepšení výsledků dalších podobných dotazníkových šetření. Jedná se o změnu volné formy odpovědí na formu škálovou. Rozsah škály by měl vždy korespondovat s hodnotami příslušných dolních a horních kvantilů, určených z výsledků vyhodnocení tohoto dotazníkového šetření. Důsledkem uvedené změny by měla být jednak úplná eliminace kontaminant a jednak snížení časové náročnosti při vyplňování dotazníků. Zabráněním vzniku

kontaminant a především pak snížením náročnosti vypl%nování lze docílit zvýšení četností realizací jednotlivých položek, které lze zahrnout do statistické analýzy.

Literatura

1. ANTOCH, J., VORLÍČKOVÁ, D. (1992).*Vybrané metody statistické analýzy dat*. ACADEMIA, Praha, ISBN 80-200-0204-9.
2. BARTOŠOVÁ, J (2003 b). *Příjmové modely*. 12. medzinárodný semiár VÝPOČTOVÁ ŠTATISTIKA, Zborník príspevkov,, SŠDS, Bratislava, s.7-11, ISBN 80-88946-29-8.
3. BARTOŠOVÁ, J (2004 a). *Identifikace odlehlých pozorování v souboru hodnot časového vytížení vysokoškolského pedagoga dílčími činnostmi*. 13. medzinárodný semiár VÝPOČTOVÁ ŠTATISTIKA, Zborník príspevkov, SŠDS, Bratislava, s. 11-14, ISBN 8088946-38-7.
4. BARTOŠOVÁ, J (2004 b). *Problematika vhodného odhadu parametru polohy časového vytížení vysokoškolského pedagoga – vědce jednotlivými pedagogickými i nepedagogickými činnostmi*. ACTA UNIVERSITATIS Bohemiae Meridionales, Ročník VIII, JU, České Budějovice, s. , ISSN .
5. BARTOŠOVÁ, J (2004 c). *Některé výsledky průzkumu časového vytížení vysokoškolských pedagogů jednotlivými dílčími činnostmi*. FORUM METRICUM SLOVACUM, Tom VIII, SŠDS, Bratislava, s. 38-44, ISBN 80-88946-23-9.
6. BARTOŠOVÁ, J., STEJSKALOVÁ, I. (2003). *Statistická analýza pilotního průzkumu v rámci projektu Hodnocení pracovní činnosti vysokoškolského pedagoga (vědce) jako základ pro alokaci mzdových prostředků*, FORUM METRICUM SLOVACUM, Tom VII, SŠDS, Bratislava, s. 82-87, ISBN 80-88946-30-1.
7. BLATNÁ, D. (1999): *Neparametrické metody II. – Neparametrické odhady*. VŠE, Praha, ISBN 80-7079-694-4.
8. JUREČKOVÁ, J. (2001). *Robustní statistické metody*. Karolinum, Praha, ISBN 80-246-0259-8.
9. MELOUN, M., MILITKÝ, J. (2002). *Kompendium statistického zpracování dat*. ACADEMIA, Praha, ISBN 80-200-1008-4.
10. STEJSKALOVÁ, I., BARTOŠOVÁ, J. (2003). Výsledky pilotního výzkumu z grantu hodnocení pracovní činnosti na vysokých školách. VÝUKA A VÝZKUM V ODVĚTVOVÝCH EKONOMIKÁCH A PODNIKOVÉM MANAGEMENTU NA VYSOKÝCH ŠKOLÁCH, Sborník z vědecko-pedagogické konference s mezinárodní účastí, Univerzita Pardubice, s. 251-255, ISBN 80-7194-623-0.
11. STEJSKALOVÁ, I., ROLÍNEK, l., BARTOŠOVÁ, J. (2004). ACTA UNIVERSITATIS Bohemiae Meridionales, Ročník VII, JU České Budějovi

Maximum Likelihood Estimates of Parameters of Model of Households Income Distribution in the Czech Republic

Jitka Bartošová
Vysoká škola ekonomická v Praze
Fakulta managementu
Katedra managementu informací
Jarošovská 1117/II
377 01 Jindřichův Hradec
bartosov@fm.vse.cz

Abstract

While making a statistical model it is important both, to find a theoretical distribution function that would characterize empirical frequency distribution and to choose a suitable method of parameter estimate of such model. Recently used type of a theoretical model of income distribution (the log-normal distribution) was derived from the character of a particular feature and from a longtime experience with its behavior before revolution. The logarithmic-normal distribution, especially its variation with three parameters, has served a good approximation of income distribution of most of social classes. An important aspect of income distribution modeling is choice of suitable parameter estimate of the relevant logarithm-normal curves.

Keywords:
point estimate, method of maximal likelihood, parameter of model, income distribution.

1. Introduction

Economists' interest in income of the population in developed countries arises out of efforts to solve matters concerning the level of living standard of the population in an objective manner. Income models may be easily used to directly evaluate the level of standard or to compare the level of standard in different regions or nations. While making a statistical analysis of the level of living standard, we focus only on measurable elements of the level of living standard. In order to correctly quantify the element of the level of standard that directly depends on incomes, we need to characterize the level and structure of population income in their complexity, i.e. to find out suitable statistical models of income distributions both in different social classes and in the whole population (without regarding social classes). Knowledge of the current statistical model of income distribution, which is a simple approximation of sample distribution and knowledge of tendency of its parameters development may be used to predict behavior of the particular variable in the following period of time.

While making a statistical model it is important both, to find out a theoretical distribution function that would characterize empirical frequency distribution and to choose suitable methods to calculate parameters of the model. As a theoretical model of income distribution, logarithm-normal distribution has been used so far. Recently used type of theoretical model was derived from the character of a particular feature and from a longtime experience with its behavior before revolution. Especially the three-parameter logarithm-normal distribution has represented a good approximation of income distribution for most of social classes.

Before revolution, planned economy in the Czech Republic experienced high homogeneity of income in the population in all social classes. In presence, the transformation to market economic system has caused a significant change in income distribution. The variety of income sources and present process of differentiation of wages brings about discrepancies between empirical income distribution and the theoretical model. There are values of income that may be concerned as outliers and that causes contamination of the model (see [3], [11]). Distributions in some social classes correspond to commixture of several theoretical curves etc. These differences are still deepening and in a different rate and inertia they are progressively reflected in both income distribution of the whole population without regarding social classes and income distribution of some social classes (see [4] - [7]).

There is some inertia in income distribution, so its changes would noticeably display in a term of several years after revolution. The first sample survey focusing on distribution of income of population (Mikrocensus), in which we may anticipate significant changes in its results was carried out by Czech Bureau of Statistic in 1996. Therefore, one of the important tasks of the present time is to make statistical analysis of impact of economic changes on distribution of annual financial income per household (or per person) in 1996.

To make a model, it's important to choose a good method to estimate its parameters (see [1], [2]). Since we expect changes in the theoretical model of distribution (high variability of incomes, contamination of the model etc.), it's necessary to choose the method that will provide us with good results under real conditions.

2. Methods of solution

The choice of the method depends on needed quality. Quality of the estimate of $\hat{\tau}(x)$ of the parametrical function $\tau(\theta)$, where θ is an unknown parameter, may be assessed according to the distribution. To quantify this requirement there is so-called loss function, which measures the deviation of $\hat{\tau}(x)$ from $\tau(\theta)$, and the risk function, that is the expected value of the loss function. The best estimate would minimize the risk function in the whole parametrical space (i.e. for all θ).

Since such a perfect estimate in the class of all estimates and in the whole parametrical space is usually impossible to gain, it's necessary to add further requirements. We require the estimate to be both unbiased and sufficient.

A direct application of the method to construct the best unbiased estimate is in some cases very difficult. Under these adverse conditions there work the estimates can't fulfill requirements of unbiasement and minimal variation exactly, but asymptotically. Asymptotically optimal methods of estimates shall meet the requirements of consistence and efficiency.

In the cases, where the null model is contaminated with outliers, we must add one more requirement of robustness.

There are several possibilities how to estimate the vector of parameters $\vec{\theta} = (\theta_1, \dots \theta_k)$ of the theoretical model of the distribution. We can use the method of quintiles, the method of moments or one of asymptotical methods, e.g. the method of maximum likelihood or the method of minimal χ^2.

Sample moments $m_1, \dots, m_k$ are due to Chin chin's theorem consistent estimates of the theoretical moments $\mu_1, \dots, \mu_k$, but they aren't efficient. That's the reason why this method is usually used in the cases, where more precise methods are numerically too difficult or in the first approximation step. Since the moment statistics aren't robust, they aren't suitable for estimates in the contaminated models. Quin-

tile estimates are consistent, robust and easy, but they aren't efficient, so that they are used mainly in the first approximation step on sorted data sets.

The method of minimal χ^2 is based on the principal of the minimization of the statistics χ^2. It can be proven (see e.g. [2]) that for the vector of random quantity with multinomial distribution and for $n \to \infty$ there is the only solution that converges to the real value of $\vec{\theta}$.

Another asymptotical method is the method of maximum likelihood. From the form of the likelihood function we can see that maximum likelihood estimates are always the functions of sufficient statistics. The most important for the theory of estimates and its applications are asymptotical characteristics of maximum likelihood estimates. If the function $\ln f(x;\vec{\theta})$ has in the k-dimensional interval, where the vector of real values $\vec{\theta}_0 = (\theta_{0_1}, \ldots, \theta_{0_k})$ is included, the first differentiation with respect to all components θ_r, $r = 1, \ldots, k$, then for $n \to \infty$ there is a vector solution $\hat{\vec{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ of the system of likelihood equations. Components of the solution are consistent estimates of corresponding components of the vector of parameters $\vec{\theta} = (\theta_1, \ldots, \theta_k)$. These consistent solutions of the system of likelihood equations are efficient estimates of components of the vector $\vec{\theta}$ and they are of asymptotically k-dimensional distribution with the vector of expected values equal to the vector real values and with the covariant matrix $\frac{1}{n \cdot \tilde{J}(\vec{\theta})}$, where $\tilde{J}(\vec{\theta})$ is the information matrix incident to the distribution of
$f(x;\vec{\theta})$ (see [12]).

3. Results

¿From the facts above we can see that in our case, where we have sample sets of incomes of wide range, the point estimate of the parameters of the three-parameter logarithm-normal distribution should be made by one of asymptotical methods. A disadvantage of the asymptotical methods is their difficulty. In both cases above the estimates of the parameters of the three-parameter logarithm-normal distribution can be obtained only by the numerical way.

On bases of the analysis of the characteristics of the methods of estimates, we may consider the method of maximum likelihood as the optimal method for estimates of the parameters of the three-parameter logarithm-normal model of household incomes. Maximum likelihood estimate of the parameters $\vec{\theta} = (\mu, \sigma^2, \gamma)$ is the solutions of the system of likelihood equations

$$\frac{\partial \lambda(\mu, \sigma^2, \gamma|x)}{\partial \mu} = 0 \wedge \frac{\partial \lambda(\mu, \sigma^2, \gamma|x)}{\partial \sigma^2} = 0 \wedge \frac{\partial \lambda(\mu, \sigma^2, \gamma|x)}{\partial \gamma} = 0,$$

where $\lambda(\mu, \sigma^2, \gamma|x) = -\frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\ln(x_i - \gamma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\ln(x_i - \gamma) - \mu\right]^2$ is the logarithmic likelihood function for random quantity of size n of the three-parameter logarithm-normal distribution. The system can't be solved but numerically. Substitution of the maximum likelihood estimates of the parameters $\hat{\mu}$ and $\hat{\sigma}^2$ from the first two equations in the logarithmic likelihood function considerably simplifies the computation. We gain this function if the form (in means of parameter γ)

$$\lambda(\gamma) = -n\left[\hat{\mu}(\gamma) + \frac{1}{2}\ln\hat{\sigma}^2(\gamma)\right].$$

It means that we get the maximum likelihood estimates of the vector of parameters $\vec{\theta} = (\mu, \sigma^2, \gamma)$ by numerical calculation of the maximum of the function $\lambda(\gamma)$. It can be proven that there is just one solution of the task in the open interval $(-\infty, x_{min})$.

Social groups	Estimates			Log. likelihood function $-\ln L$
	μ	σ^2	γ	
Workers	12,1841	0,12416	-30125	98665
Self-employed	12,2498	0,33054	-1256	20445
Employees	12,2060	0,22624	-10753	79266
Self-employed farmers	12,0965	0,46730	6066	1535
Cooperative farmers	12,1635	0,14463	-14197	2183
Retired with EA members	11,7931	0,18621	34829	12661
Retired without EA members	11,0184	0,20975	11972	88564
Unemployed	11,3404	0,32690	-8185	2803
The others	11,0142	0,49000	3272	2515
All	11,7721	0,39144	2666	318161

Table 1: Maximum likelihood estimates of parameters of three-parameter logarithmic-normal model of empirical annual income distribution per household.

The estimated theoretical minimum of the model (i.e. the parameter γ) often in the models of income distribution acquire negative values. We can expect that the interval where there's the value is lower bounded. As we consider the character of a particular feature (net annual households' incomes) it's suitable to choose $-x_{min}$ as the lower bound of the interval. Numerical solution of the task above then consist in calculating the value of $\hat{\gamma}$, where the logarithmic likelihood function reaches its maximum in the interval $(-x_{min}, x_{min})$. The task was solved in the program Mathlab.

Data set in households' incomes in the Czech Republic in 1996 comes of sample survey Mikrocensus 1996. For statistical analysis following indices were chosen

- Social class of the head of the household,
- Number of members of the household,
- Net financial income of the household (in CZK).

The maximum likelihood estimates of the parameters of the three-parameter logarithm-normal models of distribution of net financial incomes per household in the social classes one by one and altogether are demonstrated in the table 1. The estimates of the parameters of the model of income distribution per one member of the household (per head) are in the table 2.

The maximum likelihood estimates in the tables were used in the construction of three-parameter logarithm-normal models of households' income distribution. On the bases of Person's χ^2 test, which was used to illustrate the measure of agreement of empirical and theoretical distribution, we can state that in the case of income distribution per household 1% significance level there's agreement in classes of self-employed, self-employed farmers, cooperative farmers, retired with economically active members, unemployed and the others. In the case of income distribution per head there's agreement of the empirical income distribution and the theoretical model in classes of self-employed, cooperative farmers, retired with economically active members and unemployed.

It doesn't mean that it's impossible to use this model in the other classes. The classes are of wide range, so that power of the test is big in these classes. Another

Social groups	Estimates			Log. likelihood function $-\ln L$
	μ	σ^2	γ	
Workers	10,8854	0,18693	4231	88975
Self-employed	11,0041	0,48396	7596	18601
Employees	11,0311	0,30152	9740	72135
Self-employed farmers	10,6747	0,84755	10535	1388
Cooperative farmers	11,0236	0,11049	-4379	1965
Retired with EA members	11,0320	0,09850	-391	11413
Retired without EA members	10,7734	0,05378	4897	80557
Unemployed	10,2194	0,30573	1959	2503
The others	9,9224	0,65831	6653	2292
All	10,9083	0,21206	4555	285220

Table 2: Maximum likelihood estimates of parameters of three-parameter logarithmic-normal model of empirical annual income distribution per head.

important factor, which influences the agreement of the empirical distribution and the chosen statistical model, is the presence of outliers. These high differences of incomes can cause adverse results of χ^2 test (see [8] - [10]).

4. Conclusion

The transformation of the market from planned economics to market economic system, which began more than 15 years ago, has caused changes of the height and the structure of income of the population. There is both a movement in the height and a significant differentiation of incomes. In some social classes there appear single households or groups of households that feature extremely high or low incomes and they may cause a contamination of the theoretical model. The factors above may be the cause of a distortion of some types of estimates of characters of incomes' distribution and decline their efficiency. This process may result in inaccurate and distorted estimates of parameters of the corresponding statistical models. Therefore the choice of the optimal method of estimation of model's parameters is an important step in a construction of the model. It's the only way to quality results. As far as we consider the range of sets and characters of income sets, it's needy to choose the only method for all social classes as well as for the whole income set, without regarding the class. The method must be primarily consistent and efficient. Executed analysis of single characters of the methods has shown that demanded estimate is the maximum likelihood estimate.

References

1. Anděl, J. (1993). Statistické metody. Matfyzpress MFF UK, Praha

2. Anděl, J. (2002). Základy matematické statistiky. Preprint MFF UK, Praha

3. Antoch, J., Vorlíčková, D. (1992). Vybrané metody statistické analýzy dat. ACADEMIA, Praha.

4. Bartošová, J. (2003a). Robustní metody odhadů. Oeconomica, Praha, 234-246

5. Bartošová, J. (2003b). Příjmové modely. Výpočtová štatistika, SŠFD, Bratislava, 7-11

6. Bartošová, J. (2004a). Statistic Model of Households' Annual Income Distribution in the Czech Republic. COMPSTAT 2004, Praha

7. Bartošová, J. (2004b) Příspěvek k analýze rozdělení příjmů domácností v ČR. ROBUST 2004, Třešť, 451-458

8. Bartošová, J. (2004c). Statistický model příjmových rozdělení. Acta Universitatis Nohemiae Meridionales (Vědecký časopis pro ekonomiku, řízení a obchod) VII., JU, České Budějovice, 39-46

9. Bartošová, J. (2004d). Contamination level estimate of household income distribution by distant observations in the Czech Republic. Forum Metricum Slovacum VIII., SŠDS, Bratislava, 74-78.

10. Bartošová, J. (2004e). Identification of distant observations in the sample of annual household incomes in the Czech Republic. Výpočtová štatistika, SŠDS, Bratislava, 6-10.

11. Jurečková, J. (2001). Robustní statistické metody. Karolinum, Praha.

12. Machek, J. (1980). Teorie odhadu. MFF UK, Praha.

Regresní analýza a neuronové sítě

Roman Biskup, Anna Čermáková
Jihočeská univerzita v Českých Budějovicích, Zemědělská fakulta, Katedra aplikované matematiky a informatiky, Studentská 13, 387 05 České Budějovice
biskup@zf.jcu.cz, annacer@pf.jcu.cz

Abstrakt

Obsahem tohoto článku je srovnání regresních možností neuronových sítí a „klasické“ regresní analýzy. Na dvojici konkrétních datových množin je představeno řešení získané pomocí zvolené regresní funkce, jejíž parametry byly odhadnuty pomocí metody nejmenších čtverců, a výsledek získaný pomocí neuronové sítě typu backpropagation s poměrně triviální architekturou a typickým učícím algoritmem. Článek pro celkové pochopení předpokládá jak znalost regresní analýzy, tak znalost základních principů neuronových sítí a příslušné terminologie.

1. Úvod

Co je člověk člověkem, snaží se přijít na kloub různým situacím. Netřeba rozebírat fakt, že většinu situací nejsme schopni popsat deterministicky. Příčinou tohoto faktu jsou jednak ne zcela známý komplex procesů děj ovlivňující a pak nepřesná měření. Dalo by se říci, že právě to bylo v počátku živnou půdou pro vznik matematiky náhody, později přejmenované na pravděpodobnost. Pravděpodobnost následně dala vzniknout dalším matematickým oborům jako jsou matematická statistika, či statistika jako taková. Jednou z úloh, kterou statistika řeší pomocí svého aparátu, je úloha regresní analýzy. Tedy postup, jak co nejvhodněji popsat typ závislosti mezi vysvětlujícími a vysvětlovanými proměnnými. Pomocí tohoto postupu pak lze na základě hodnot proměnných vysvětlujících odhadnout ty vysvětlované.

S rozvojem počítačů a další vědní disciplíny – matematického modelování vznikají poměrně cílevědomé projekty k simulování toho či onoho. Za poměrně ambiciózní projekt lze považovat úspěšné vytvoření umělé neuronové sítě schopné procesu učení. Taková neuronová síť, ať už biologická či umělá, skrývající v sobě informace, odvážněji znalost, může podobně jako regresní model, odhadovat.

2. Datové množiny

2.1. Předpoklady

Předpokládejme, že je třeba řešit situaci, která je charakterizována řadou proměnných, které z logického hlediska lze rozdělit na dvě skupiny. Rozdělením proměnných, jehož opodstatnění i vysvětlení lze nalézt v každých základech regresní analýzy, je rozuměno rozdělení na proměnné vysvětlující $x_1, x_2, \ldots, x_{n_0}$ a proměnné vysvětlované $y_1, y_2, \ldots, y_{n_p}$. Dále předpokládejme, že existuje dostatečně velká množina pozorování (značení bude zavedeno níže).

Klasická regresní analýza má poněkud omezenější třídu úloh, jež je schopna řešit, než neuronová síť proto předpokládejme, že proměnné jsou spojité náhodné veličiny sledující vícerozměrné $(n_0 + n_p)$ normální rozdělení. Potom metodou nejmenších čtverců získáme maximálně věrohodné odhady koeficientů zvolených regresních modelů.

Zpracování úlohy pomocí neuronové sítě typu backpropagation[1] klade podmínky jen na číselnou interpretaci vysvětlujících a vysvětlovaných proměnných. V souvislosti s terminologií běžnou na poli neuronových sítí se však častěji používá vstupní a výstupní proměnné. Vzhledem k řečenému není nelogické požadovat, aby množinu pozorování tvořily vektory $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, kde $\boldsymbol{x}_i = \big(x_{1i}, \ldots, x_{n_0 i}, y_{1i}, \ldots, y_{n_p i}\big)$, pro $i = 1, \ldots, n$. Pro úspěšný trénink neuronové sítě je však nutno dodat poměrně velkou datovou množinu.

V této práci bylo zpracování dat provedeno v statistickém softwaru Statistica 6.0 Cz,[2] Neurex 4.0,[3] a NetVisualiser.[4] V programu Statistica 6.0 Cz byl proveden výpočet koeficientů regresního modelu, program Neurex 4.0 sloužil k natrénování neuronové sítě pro aktivní dynamiku sítě a konečně program NetVisualiser posloužil pro grafické přiblížení jak procesu učení tak výsledného naučení neuronové sítě. S ohledem na trénink neuronové sítě a jeho grafický výstup byly zvoleny příklady s jednou vysvětlující a jednou vysvětlovanou proměnnou, tj. $\boldsymbol{x}_i = (x_i, y_i)$, pro $i = 1, \ldots, n$, kde x odpovídá vysvětlující a y vysvětlované proměnné.

Výsledky jsou prezentovány jak na množině reálných dat, získaných měřením délky jatečního trupu a výšky průměrného podkožního tuku (obrázek 1) tak vygenerovány tak, aby „zamotaly“ hlavu kubickému regresnímu modelu (obrázek 2). Pro potřeby programu NetVisualizer byly číselné hodnoty reálného modelu transformovány.

3. Výpočet modelů

3.1. Regresní analýza

Jak již bylo naznačeno pro odhady parametrů regresních modelů bylo zvoleno metody nejmenších čtverců a parametry byly vypočteny prostřednictvím statistického softwaru Statistica 6.0 Cz. V prvém případě byl po věcném zhodnocení závislosti a s ohledem na tvar korelačního pole zvolen hyperbolický regresní model s předpisem $y = a + b/x$. V případě druhém, přesně dle záměru, pak kubický regresní model ($y = a + bx + cx^2 + dx^3$). Protože vyjádření chyb modelu od dat je v případě programu Neurex 4.0 poněkud těžkopádné a veškeré výsledky jsou presentovány především na základě grafických výsledků, jsou i odhadnuté parametry ponechány jen v rámci obrázků, viz obrázek 1 a obrázek 2.

3.2. Neuronová síť

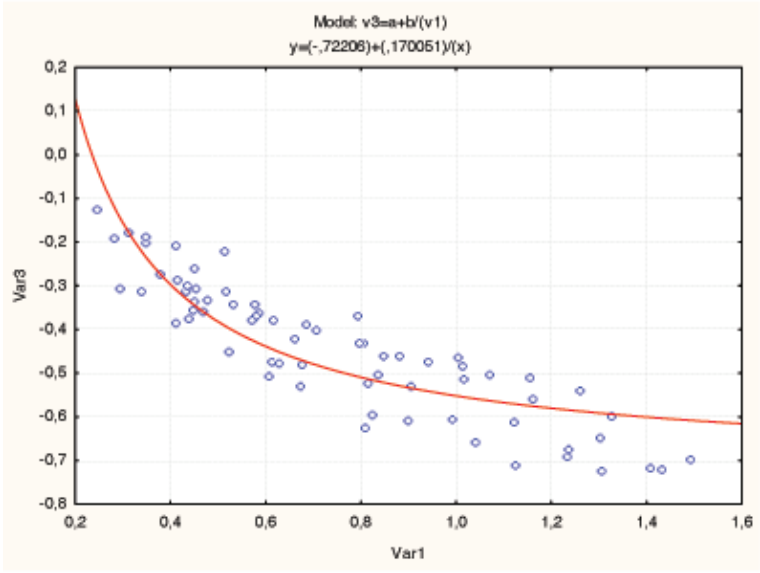
Bez ohledu na datové množiny byla využita síť backpropagation s jednoduchou architekturou typu 1–3–2–1. Je nabíledni, že jeden vstupní a výstupní neuron odpovídá charakteru problému a dvě skryté vrstvy po 3 a 2 neuronech tvoří výpočetní sílu neuronové sítě. Topologie sítě je jednoznačně dána architekturou a nastavením jednotlivých programů. Učícím algoritmem byl, jak pro vizualizaci tak pro výpočet, modifikovaný algoritmus se zpětným šířením chyby (backpropagation) obsahující vedle učící konstanty též moment. Více se laskavý čtenář může dozvědět v uživatel-

[1] Tato síť bývá také někdy nazývána vícevrstvý perceptron. Název backpropagation se vžil díky nejčastěji používanému postupu trénování neuronové sítě.

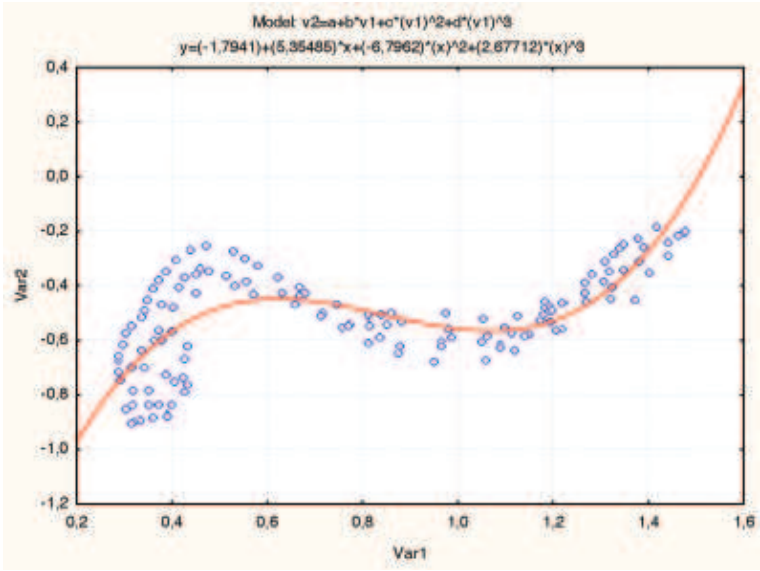
[2] Statistický software společnosti StatSoft, s lokací do češtiny, který po dokoupení modulu Neural Networks nabízí velmi výkonný systém pro tvorbu a trénink neuronových sítí nejrůznějších topologií.

[3] Neuronová síť typu backpropagation vytvořená Ivo Vondrákem primárně jako expertní systém. Více v [3].

[4] Vizualizační program nejen neuronové sítě typu backpropagation vytvořený Petrem Marťánem pro vizualizaci tréninkového procesu výsledků neuronových sítí. Více v [1].



Obrázek 1: Výsledky regresní analýzy v programu Statistica 6.0 Cz – hyperbolická regrese



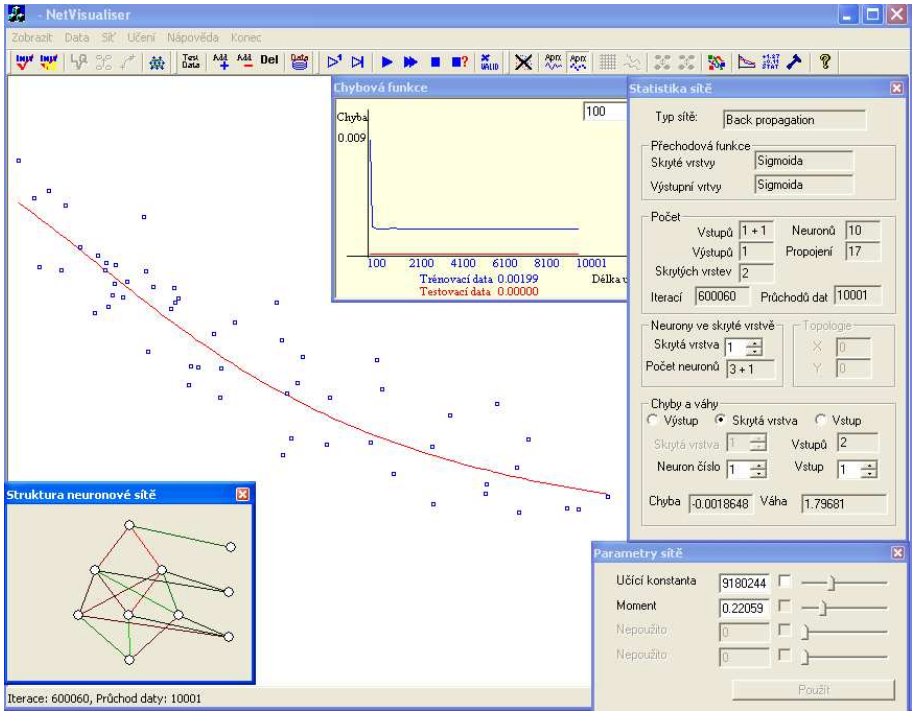
Obrázek 2: Výsledky regresní analýzy v programu Statistica 6.0 Cz– kubická regrese

ských manuálech k využitým programům volně dostupných na internetu [3, 1].

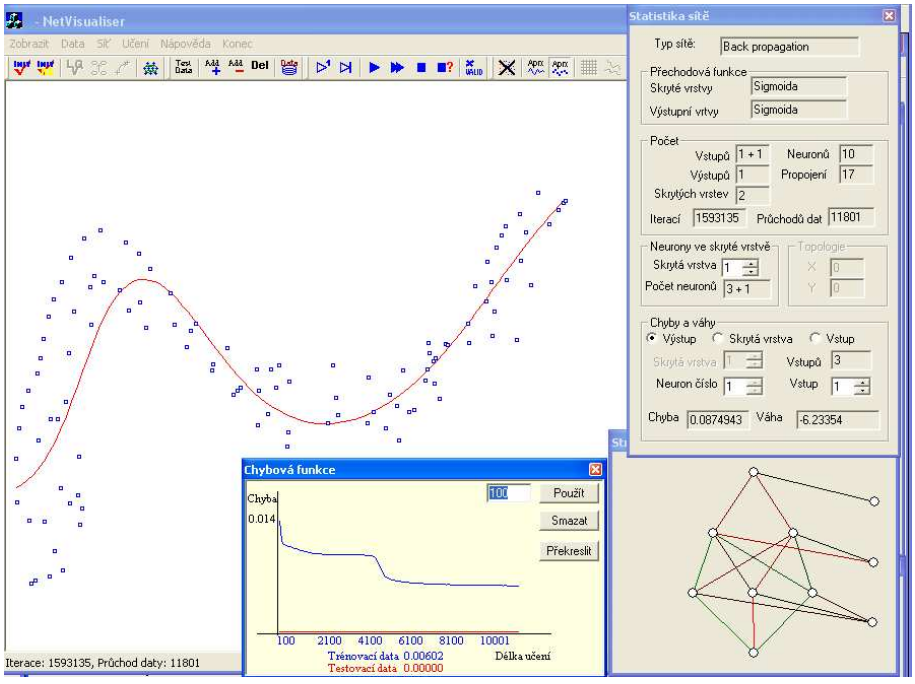
Prostředí programu NetVizualizer, jakož i zvolenou architekturu, průběh chyby učení a v neposlední řadě výsledek tréninku lze shlédnout na obrázcích 3 a 4. Obrázky dávají jasný pohled na část práce neuronové sítě označovaný jako adaptivní dynamika i na výsledek sítě (na obrázku „proložení“ korelačního pole). Této části práce neuronové sítě se říká aktivní dynamika.[5]

Parametry učící konstanty byly voleny experimentálně a modifikovány podle pra-

[5]Bias (práh) je zobrazen na obrázku pomocí izolovaného elektronu ovlivňující vždy jen jednu vrstvu. Rozdílnost biasů pro elektrony ve stejné vrstvě je zajištěna pomocí vah měnících se v průběhu učení.



Obrázek 3: Trénink neuronové sítě v programu NetVizualizer



Obrázek 4: Trénink neuronové sítě v programu NetVizualizer

videl navržených v [3, 1 a 2], tj. učící konstanta byla postupně zvyšována.

3.3. Srovnání

V prvním případě ještě z obrázku (Obrázek 5) není tolik patrné lepší „přimknutí“ křivky vykreslené za pomoci neuronové sítě proti hyperbole získané pomocí regresní

analýzy. Jen na okrajích korelačního pole lze vypozorovat odklon hyperboly od hodnot datové množiny. Srovnání několika hodnot odhadnutných na základě regrese a neuronové sítě shrnuje tabulka 1.

vstup	0,2	0,4	0,6	0,8	1	1,2	1,4	1,6
regrese	0,128	−0,297	−0,439	−0,510	−0,552	−0,580	−0,601	−0,616
síť	−0,145	−0,289	−0,407	−0,499	−0,568	−0,621	−0,661	−0,691

Tabulka 1: Srovnání několika odhadnutých hodnot

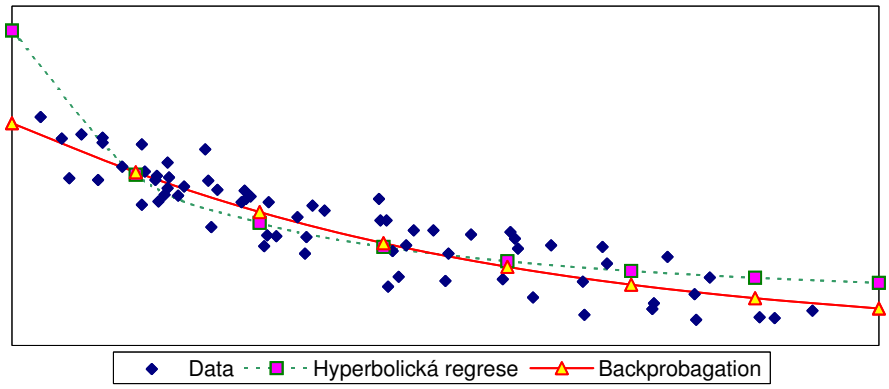
Pro zajímavost při testování významnosti regresních koeficientů byl absolutní člen označen za nevýznamný.

Data v druhém případě byla tvořena s jasným záměrem. Na jednu stranu korelační pole zřejmě připomíná svým tvarem kubickou parabolu, na stranu druhou je levá část tohoto pole sevřenější než ta druhá, proto proložení není příliš ideální (viz obrázek 6. Stále dost jednoduchá neuronová síť prokládá korelační pole přesvědčivěji. Ani ona si moc dobře neporadila s levou částí korelačního pole, kde směrnice ideální regresní funkce vzhledem k datům roste nad všechny meze.

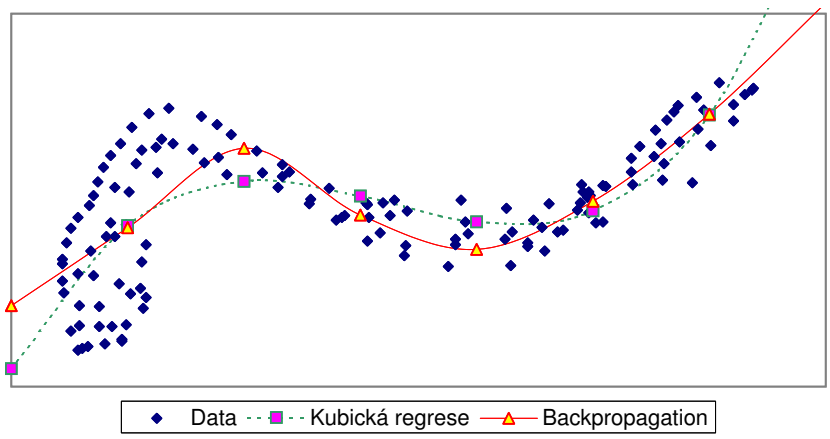
Teto problém lze samozřejmě řešit v obou případech. Buďto změnou regresního modelu, nebo přidáním neuronů, či celých vrstev. Pouhým zvýšením stupně regresního polynomu, nebylo kýženého výsledku dosaženo. Přidání neuronů zas zvýšilo náročnost trénovacího procesu natolik, že nebylo v rozumném čase získáno ani tak dobré řešení, jako při původní architektuře. Tento zádrhel lze vyřešit, lze-li si vybírat, změnou tréninkového procesu. Dobré výsledky slibuje použití genetických algoritmů v procesu učení. Srovnání několika hodnot odhadnutých na základě regrese a neuronové sítě opět shrnuje tabulka 2.

vstup	0,2	0,4	0,6	0,8	1	1,2	1,4	1,6
regrese	−0,952	−0,568	−0,450	−0,489	−0,558	−0,529	−0,272	0,341
síť	−0,783	−0,574	−0,361	−0,540	−0,632	−0,503	−0,270	0,028

Tabulka 2: Srovnání několika odhadnutých hodnot



Obrázek 5: Srovnání výsledků aproximace pomocí reg. analýzy a neuronové sítě – hyperbolická regrese



Obrázek 6: Srovnání výsledků aproximace pomocí reg. analýzy a neuronové sítě – kubická regrese

4. Závěr a diskuse

Přestože článek vyznívá ve prospěch neuronových sítí, není závěr úplně jednoznačný. Neuronové sítě nabízejí řadu výhod pro nezasvěcené uživatele. Oproti regresní analýze odpadá nutnost na začátku provádět analýzu, z níž by například měl vyplynout typ regresní rovnice. Předpoklad normality dat je u neuronových sítí irelevantním požadavkem. Na druhou stranu, je nutno zvolit vhodnou architekturu respektive topologii sítě a vhodný trénovací algoritmus. Značně výkonný počítač je naprostou nutností. Softwarové implementace neuronových sítí jsou poměrně drahou záležitostí, v čemž si nezadají s cenami kvalitních komerčních statistický balíků. Pokud však je uživatel ochotný investovat nemalé finanční prostředky do neuronových sítí například od společnosti StatSoft, dostane tak velice výkonný nástroj včetně rádce pro volbu vhodné sítě i jejího nastavení. Samostatná heuristika řešení problému je též součástí tohoto modulu.

To co se jeví na jedné straně jako výhodné, může svým způsobem uživatele okrádat o informace, jenž jsou vedle možnosti odhadu též důležité. Ze všech stojí za zmínku nemožnost odhadu chyby odhadu, standardního posouzení korelace či intervalový odhad na základě modelu. V neposlední řadě může být nevýhodná neznalost předpisu regresního modelu, tj. typu závislosti, která sama o sobě bývá užitečná a dává možnost interpretaci dané závislosti.

Literatura

1. Marťán, P.: Vizualizace vybraných metod strojového učení, Diplomová práce, Brno (2002), pp 8–19, 23–42.

2. Šíma, J., Neruda, R.: Teoretické otázky neuronových sítí, Matfyzpress, Praha, (1996)

3. Vondrák, I.: Neurex 4.0 – Expertní systém na bázi neuronových sítí, Ostrava (1993).

The optimal strategy in statistic games

Roman Biskup, Pavel Tlustý
Jihočeská univerzita v Českých Budějovicích, Zemědělská fakulta, Katedra aplikované matematiky a informatiky, Studentská 13, 387 05 České Budějovice
biskup@zf.jcu.cz, tlusty@pf.jcu.cz

Abstract

This work is based on description and analysis of not very common paradox in Probability Theory. It deals paradox in which is classical definition of probability is used. This problem is also complemented by simulation study as a motivation for students.

1. Introduction

The theory of probability is one of the most difficult subjects in curriculum at the Pedagogical faculty. Theory of probability as a part of Statistics at the Agriculture faculty is nearly a nightmare for students. Simulation study of statistical games seems to be a good opportunity how to expound probability and its serviceability to students.

2. Parrondo's paradox

Suppose that we are about to play two games as it is described in the following paragraph. Playing each of them, we will lose sooner or later. However, when playing the games in the alternating order, we can be sure to win.

The principle of the game can be outlined as a model of a staircase. At the beginning of the game, we stand in the middle of the staircase on a particular step which we denote with number 0. Our aim is to get to top of the staircase. The direction of our movement is given by heads or tails tossed on a coin. Tossing heads, we go one step upwards, tossing the opposite, we have go downwards.

2.1. Game number one

The first game is very simple, let's label it with letter S. The coin used for the tossing is slightly asymmetrical (49.5 % for tossing heads and 50.5 % for tossing tails). This game is by all means loosing for us and after some time we will end up at the bottom of the staircase.

2.2. Game number two

The rules of the second game are more complicated. This game is denoted with letter C. We use two different coins:

- Coin A is significantly asymmetrical (9.5 % for tossing heads and 90.5 % for tossing tails).
- Coin B is also asymmetrical (74.5 % for tossing heads and 25.5 % for tossing tails).

The use of each of the coins is given by the following rules: The coin A is tossed when we are standing on a step which number is divisible by three, the coin B if we are not. This game is losing for us too as it can be explained by this consideration:

At the beginning, we are staying on step 0, which is divisible by 3. Therefore we have to toss coin A. It is highly probable that we will have to go downwards ending up on step -1.

Then we have to toss coin B. After tossing this coin, we will probably return back on step 0. This situation will be repeating again for some time. However, one moment it will happen that tail is tossed on coin A and consecutively another tail is tossed on coin B. Undertaking all that we will end up on step -3 and the cycle will start again. Probability of this sequence is $90.5\,\% \cdot 25.5\,\% \cdot 25.5\,\% = 5.88\,\%$. On the other hand, the possible movement towards the top is given by the following tossing sequence (i.e. tossing head on coin A and tossing twice heads on coin B), which is less probable ($9.5\,\% \cdot 74.5\,\% \cdot 74.5\,\% = 5.27\,\%$). The movement towards the bottom is slightly more likely.

2.3. Combined Game

The combination of both game named as the Combined Game, is denoted with letter K. The rules are simple. Both games are alternated in a fixed cycle, e.g. two games S are followed by two games C and so on. We can start with any of the games and we do not change our position when changing of games. There is a surprising fact that there are many combinations how to get to the top of the staircase. However, there are several exceptions – when one game S is exchanged by one game C or when e.g. one game S is followed by 50 games C, all this is still losing for us.

3. Computer simulation

3.1. Game S

The first program simulates game S. This game was played on 201, 401, 1001 and 2001 steps of stairs with 100 repetitions. Program wrote number of taking bottom and top of the staircase. The average probability was evaluated from 10 trials. Any trial did not stay in never-ending cycle. For results see Table 1. It is evident that with increasing number of stairs decrease the probability of taking top of the staircase (i.e. we lose more).

Number of steps	P(Bottom)	P(Top)
201	0.513	0.487
401	0.688	0.312
1001	0.909	0.091
2001	0.994	0.006

Table 1: Outcome of simulation – Game S

3.2. Game C

Game C was also computer-simulated on 201, 401, 1001 and 2001 step staircase and the average probability was calculated in the same way as it was in case of game S. The time necessary for reaching the bottom is shorter than in S; taking 201 step staircase, the top was reached only in 19 % cases; taking 2001 steps, the probability of reaching the top equals 0. All results see in Table 2.

3.3. Game K

third program simulates combined game K. There are a lot of ways to play this game according to the chosen sequence of games S and C. For the following, pictures

Number of steps	P(Bottom)	P(Top)
201	0.807	0.193
401	0.944	0.056
1001	0.998	0.002
2001	1	0

Table 2: Outcome of simulation – Game C

and tables ss-cc mean that two games S a two games C are repeated in cycles in this sequence again and again. This denotation is evident.

Combination	P(Bottom)	P(Top)
s-c	0.717	0.283
ss-c	0.005	0.995
s-cc	0.005	0.995
ss-cc	0.046	0.954
sss-c	0.031	0.969
s-ccc	0.233	0.767
s-cccc	0.009	0.991
ss-ccc	0.216	0.784
sss-cc	0.009	0.991
ssss-c	0.177	0.823
sss-ccc	0.210	0.790
s-ccccc	0.101	0.899
ss-cccc	0.075	0.925
ssss-cc	0.033	0.967
sssss-c	0.234	0.766

Table 3: Outcome of simulation - Games K (201 steps)

To compare the two previous games we have chosen the cycle ss–cc. For this game as well as for games S and C it is true that if we play with a small number of steps it will end up on the top in 98 % (for 201 steps), but in case of greater number of steps is more losing (80 % for 2001 steps). Results of comparison represent Figure 1.

Number of steps	P(Bottom)	P(Top)
201	0.022	0.978
401	0.032	0.968
1001	0.093	0.907
2001	0.196	0.804

Table 4: Outcome of simulation - Game K (ss-cc)

The simulation has revealed the differences between successfulness of various sequences depending on the chosen cycles of games S and C. Some games have been simulated on a staircase with 201 and 1001 steps. The most successful games from this range on 201–steps–stair are games s-cc, s-cccc and sss-cc. In the second case, the most successful games from this range on 1001–steps–stair are games ss-c, ssss-c, sss-cc and ss-ccc. See the probability of ending up on the top of the staircase in diagram (Figure 2 and 3). In both cases combination s-c is disadvantageous.

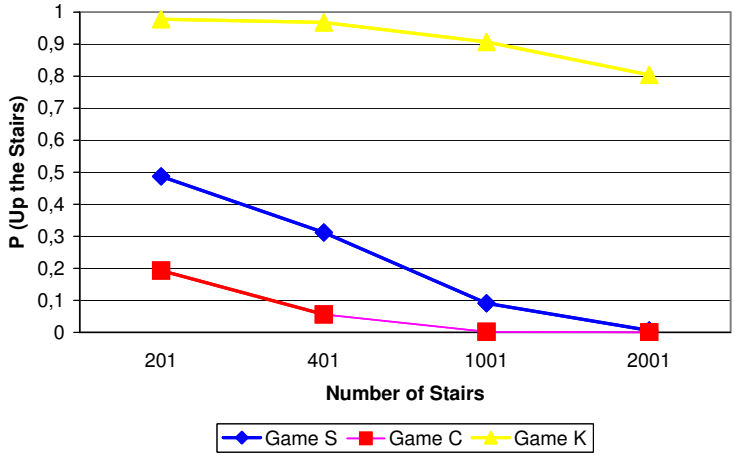


Figure 1: Probability of Games

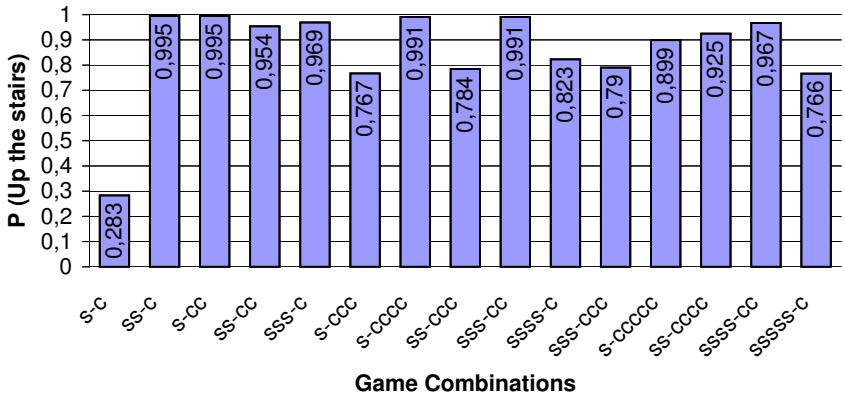


Figure 2: Game Combinations, 201 stairs

When comparing the successfulness of the combined games depending on the number of steps, we can come to a conclusion that the length of the staircase is rather irrelevant.

4. Conclusion

In this contribution we tried to show how to introduce Probability Theory to students in an interesting way. It's exciting to compare theoretical probabilities with likelihood acquired via simulation study. Compare results in Table 1 and 5.

Number of steps	P(Bottom)	P(Top)
201	0.881	0.119
401	0.982	0.117
1001	$\doteq 1$	$\doteq 0$
2001	$\doteq 1$	$\doteq 0$

Table 5: Theoretical probabilities – Game S

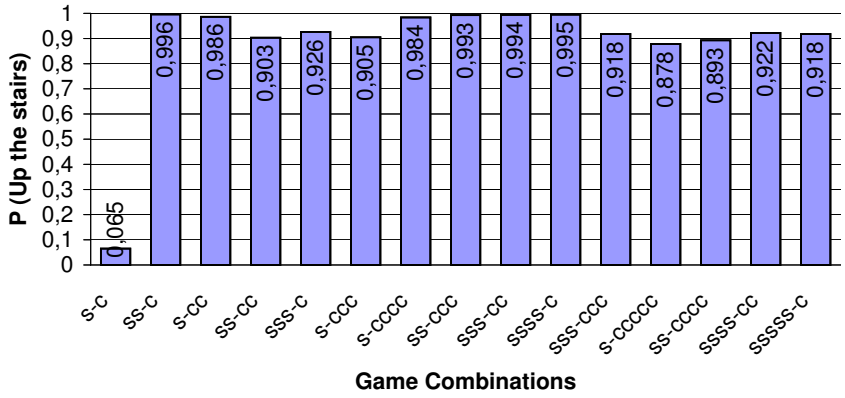


Figure 3: Game Combinations, 1001 stairs

References

1. Polách, E.: Programování v jazyku Turbo Pascal, PF JU v Českých Budějovicích, České Budějovice, (1993)

2. Székely, Gábor J.: Paradoxes in Probability Theory and Mathematical Statistics. Akadémiai Kiadó, Budapest (1986).

VÝUKA STATISTIKY PRO EKONOMY NA VŠE V PRAZE: SOUČASNOST A PŘIPRAVOVANÉ ZMĚNY

Blatná Dagmar
Fakulta informatiky a statistiky Vysoké školy ekonomické v Praze,
nám. W.Churchilla 4, Praha 3, Česká republika
Blatna@vse.cz

Statistické předměty jsou na VŠE v Praze vyučovány ve dvou úrovních. Jako předměty povinného studijního základu jsou zařazeny do studijních plánů pro studenty všech fakult VŠE ve formě dvou jednosemestrálních základních statistických kurzů, oba rozsahu 2/1. Tyto základní kurzy zahrnují popisnou statistiku, základy teorie pravděpodobnosti, matematickou statistiku (statistické odhady a testování statistických hypotéz) a základní metody statistické analýzy (analýzu rozptylu, analýzu kategoriálních dat, regresní analýzu, korelační analýzu, metody analýzy a prognózy časových řad a indexní analýzu). Pro studenty studijního oboru Statistické a pojistné inženýrství (SP) jsou určeny kurzy, které podrobněji a s vyšší mírou teorie obsahují látku jednotlivých oblastí a partií statistiky (pravděpodobnost a matematická statistika, regrese, analýza kategoriálních dat, terénní průzkumy, výběrová šetření, časové řady, vícerozměrné statistické metody, navrhování experimentů, nelineární regrese, neparametrické a robustní metody atd.)

Nejdříve se budu zabývat výukou statistiky pro ekonomy – nestatistiky.

V současnosti jsou statistické předměty vyučovány s využíváním systému STATGRAPHICS Plus verze 3.1 v systému WINDOWS.

STATGRAPHICS lze z hlediska výuky statistických metod charakterizovat jako poměrně jednoduchý nástroj, který z větší části pokrývá potřeby výuky základních kurzů statistiky a částečně i specializovaných kurzů. V souvislosti se zvyšující se gramotností studentů při práci s počítači, nečiní studentům práce se STATGRAPHICSem potíže a zvládnou ji i při nízkém počtu cvičení. Výhodou je i existence statistického rádce u každého výsledku, který lze využít pro pochopení dané metody a správné interpretace získaných výsledků (zejména pro studenty z nestatisticky zaměřených oborů studia). Na druhé straně velkou nevýhodou používání STATGRAPHICSu pro výuku statistiky je to, že studenti nemají možnost instalace tohoto produktu na vlastních počítačích doma, neexistuje možnost distribuovat studentům tzv. studentské verze. Další nevýhodou tohoto produktu je fakt, že se systémem STATGRAPHICS se studenti setkávají vesměs pouze na akademické půdě, ale rozšíření tohoto programu v praxi je poměrně sporadické. To ve svém důsledku znamená, že znalosti získané při výuce na VŠE jsou omezeně využitelné v praxi, na což studenti oprávněně poukazují.

V souvislosti s končící dobou licence používání systému STATGRAPHICS na VŠE a po posouzení předností a nevýhod jednotlivých dostupných systémů včetně posouzení podmínek jejich používání a finančních nákladů spojených s pořízením i používáním při výuce, bylo navrženo, zavést pro výuku statistiky na VŠE systém SAS.

Systém SAS je velice silným nástrojem pro správu dat a práci s nimi. Umožňuje provádět statistické analýzy od těch nejjednodušších až po analýzy velmi složité a speciální. Základem ovládání systému SAS je zadávání příkazů jazyka *SAS Language*. Prvky programovacího jazyka jsou příkazy, výrazy, funkce, volby a formáty. Úlohy, na

jejichž realizaci jsou využívány moduly SAS/ASSIST, SAS/CALC a SAS/QC lze zadávat pomocí nabídek základního menu. Pro práci s ostatními moduly je určen speciální modul SAS/ASSIST – nabídkově řízené uživatelské rozhraní. V okně SAS/ASSIST se pomocí tlačítek volí oblast požadovaných činností a v druhé hierarchické úrovni detailnější oblast (např. regresní analýza), potom následuje klasická nabídka činností.

Protože ovládání SASu je ve srovnání se systémem Statgraphics poměrně náročné a nelze předpokládat, že by jej všichni studenti základního kurzu zvládli bez problémů při malém počtu cvičení, přichází pro výuku v úvahu využití programové nadstavby SASu *Enterprise Guide*, která umožňuje využívat většinu procedur formou jednoduše ovládaného rozhraní a značně tak zefektivňuje a urychluje práci se systémem, neobsahuje však vše, co je v SASu.

Enterprise Guide zahrnuje omezené verze komponent SAS/GRAPH, SAS/STAT, SAS/QC, SAS/ETS. Zadávání úloh je možné dvěmi způsoby: buď ze seznamu *Tasks of Category* nabídkového okna *Task List* nebo z rozhraní *Data, Analysis* a *Graphs*.

Z látky probírané v základním kurzu *Enterprise Guide* nenabízí funkce pro výpočet hodnot distribuční či pravděpodobnostní funkce a výpočet je nutno realizovat pomocí zadání programového kódu a jeho spuštění z menu Code, Run on Local, což může zejména pro uživatele s menší zkušeností s prací na PC činit potíže.

Ve srovnání se STATGRAPHICSem je *Enterprise Guide* v některých partiích širší, někde naopak chudší, výstupy jsou ale méně přehledné a získání grafů je mnohdy složitější. Na základě prvních zkušeností se SASem lze tedy konstatovat, že STATGRAPHICS je uživatelsky přívětivější.

Jedna z výše uvedených předností SASu je existence programovacího jazyku SAS, který umožňuje psaní vlastního programového kódu a tvorbu vlastních procedur v situacích, kdy příslušné procedury nejsou v SASu k dispozici. Tato přednost zřejmě nebude pro výuku v základních kurzech využitelná, neboť se většinou netýká procedur, používaných pro výuku, nýbrž pokročilých analýz, takže tento rys ocení především učitelé a studenti při vědecké práci.

Ke zkvalitnění výuky statistických předmětů pomocí SASu by měla přispět i kniha „Statistika pro ekonomy – aplikace“, která bude využívána při výuce statistiky všemi studenty VŠE. Tato kniha je koncipována jako doplněk pro procvičení a hlubší pochopení metod a postupů popsaných v celostátně používané učebnici pro vysoké školy ekonomického zaměření Statistika pro ekonomy (autoři R. Hindls, S. Hronová, J.Seger). Uvedené postupy a metody a příklady jsou ilustrovány a řešeny s využitím systému SAS. Součástí knihy je vložené CD, které obsahuje výukovou verzi programu SAS a datové soubory, použité v příkladech. V obsažených příkladech jsou ukázky řešení v programu SAS a jeho nadstavbě Enterprise Guide. Uživatel si tedy může oba programy přímo nainstalovat na vlastním počítači a prakticky si řešení některých uvedených příkladů vyzkoušet. V příloze knihy jsou také popsány základy ovládání SASu a programu Enterprise Guide, jakož i stručný manuál pro výpočty z oblasti pravděpodobnosti a analýzy časových řad.

Studenti tak budou mít možnost instalovat uvedený produkt na vlastním počítači a pracovat se systémem SAS i mimo počítačové učebny VŠE. Tento fakt jistě napomůže studentům k lepšímu zvládnutí programu a usnadní jim tvorbu úkolů, které jsou ve výuce zadávány a jejichž váha v dále uvedeném hodnocení studentů se výrazně

zvýší. Hodnocení za samostatně zpracované úkoly bude součástí výsledného hodnocení studenta. Uvolní se rovněž pochopitelně i kapacita počítačových učeben, protože až dosud musí studenti většinu úkolů řešit v prostorách a učebnách VŠE, resp. na koleji, kam je zavedena školní počítačová síť.

Zavedení výuky statistiky s využitím systému SAS by mělo mít i další přínos pro studenty a to je praktické uplatnění získaných softwarových znalostí v praxi. Protože systém SAS je v praxi systémem poměrně rozšířeným, a to celosvětově, studenti budou moci získané znalosti uplatnit i v praxi mimo půdu VŠE. Na rozdíl od STATGRAPHICSu je pravděpodobné, že se budou po ukončení studia se SASem setkávat v bankách, měnových institucích, ve velkých společnostech atd. Protože systém SAS se používá i na velkém množství zahraničních univerzit a vysokých škol, budou moci studenti uplatnit znalosti tohoto produktu i v rámci svých studijních pobytů a stáží v zahraničí, což je významných faktorem zejména v souvislosti s přechodem na systém EC, o němž budu mluvit později.

V souvislosti s přípravou zavedení systému SAS do výuky statistiky byl získán grant FRVŠ jehož cílem bylo inovovat obsah i formu výuky statistických předmětů na VŠE tak, aby studenti byli více schopni získané teoretické znalosti z oblasti statistiky aplikovat při řešení konkrétních ekonomických problémů a úloh s využitím statistického softwaru, o němž se dá předpokládat, že s ním budou pracovat po skončení studia v praxi a to nejen u nás, ale i v zahraničí. Předpokládalo se rozšíření zejména v oblasti matematické statistiky a to o některé neparametrické a robustní odhady a testy a z analytických metod o rozšíření regresní analýzy o regresní diagnostiku, metody posouzení vhodnosti volby regresního modelu a zařazení dalších dosud nevyučovaných metod a testů (např. neparametrické analýzy rozptylu (Kruskall – Wallisův test).

Realita je ale zcela jiná. Systém SAS bude sice od září t.r. zaveden do výuky, ale k uvažovanému rozšíření obsahu nedojde, neboť na VŠE se chystají výrazné změny ve studijních programech v souvislosti s přechodem školy na evropský systém přenosu kreditů (European Credit Transfer and Accumulation System – ECTS).

Zavedení evropského systému přenosu kreditů (na Vysoké škole ekonomické by mělo usnadnit uznávání studijních oborů a udělovaných diplomů z VŠE v kontextu evropských vysokoškolských institucí a je předpokladem pro volný pohyb studentů a pedagogů VŠE v rámci evropského vzdělávacího prostoru. ECTS je postupně zaváděn na evropských vysokých školách od akademického roku 1989/90. Kredit je chápán jako jednotka studijního postupu a měl by odrážet množství práce spojené s daným předmětem či jinou studijní povinností, vztažené k celkovému objemu práce vynaložené na absolvování ročního studijního plánu daného oboru za celý akademický rok. ECTS je založen na celkové pracovní zátěži studenta a nejen na hodinách přímého kontaktu se školou.

VŠE v Praze je od roku 1998 plnoprávným členem CEMS (The Community of European Management Schools). V CEMS je ze 17 evropských států zařazena vždy jen jedna vysoká škola ekonomického zaměření, v ČR je to právě VŠE v Praze. Cílem škol sdružených v CEMS je vytvoření jednotné mezinárodní sítě navzájem spolupracujících univerzit a institucí, která by umožňovala bezproblémový přechod studentů z jedné školy CEMS na druhou.

Zásadou připravovaných změn studia na VŠE v Praze je, aby celoškolně povinný studijní základ bakalářského studia všech fakult pokryl základní znalosti (The Common Body of Knowlege) všech 18 domén studijního programu ekonomických škol zařazených v CEMS:

Domény CEMS

The Common Body of Knowledge (CBK)

CBK1 Introduction to Management
CBK2 Mathematics & Statistics
Objective: To provide the tools for theoretical modelling and empirical studies.
Minimum common topics: Calculus; linear algebra; probability theory; descriptive statistics; parametric and non-parametric statistics; introduction to methods of empirical research.
CBK3 Marketing
CBK4 Operation Management
CBK5 Organisation Theory
CBK6 Corporate Finance
CBK7 Accounting & Control
CBK8 Management of Information Systems
CBK9 Business Policy
CBK10 Economics (micro)
CBK11 Economics (macro)
CBK12 Economic Policy & Public Finance
CBK13 Financial Institutions and Markets
CBK14 International Economics
CBK15 Private, Commercial, Company and Labour Law (of a European Country)
CBK16 Constitutional and Administrative Environment of Business
CBK17 EU-Institutions and EU-Legislation
CBK18 Economic History

Pro nás, jako učitele exaktních předmětů je ale nevýhodné, že pouze jediná doména z 18 domén CEMS obsahuje exaktní předměty (doména Mathematics and Statistics). Uvedený cíl změny studia na VŠE tedy bude mít jednoznačně za následek snížení počtu výukových hodin jak matematiky, tak i statistiky. Matematika, která se dosud učí dva semestry v rozsahu 2/2 bude snížena na polovinu (jeden semestr 2/2) a statistika ze současných dvou semestrů 2/1 bude zredukována na jeden semestr rozsahu 2/2. VŠE v Praze se tak dostane na jednu z nejnižších příček hodinové dotace výuky exaktních předmětů mezi českými ekonomickými vysokými školami jak státními tak i soukromými.

Z toho je zřejmé, že uvažované rozšíření obsahu výuky statistiky je nerealizovatelné. Spíše se dostáváme do situace zvážit, které partie statistiky vypustit nebo alespoň redukovat. Domníváme se, že je do značné míry utopií představa, která je jedním ze základních pilířů přestavby studia, a to zvýšení samostatné práce studentů na úkor přímé výuky. Moc nevěříme, že studenti budou ochotni a schopni více studovat samostatně i partie, které jim nebudou odpřednášeny, neboť již nyní značná část studentů s obtížemi zvládá látku, která byla přednášena a navíc i procvičena na seminářích. Kromě toho se dá předpokládat, že se sníží i znalosti matematiky v souvislosti s poloviční hodinovou dotací výuky ve srovnání se současným stavem (a to neuvažuji obecnější případ, jak matematicky „erudovaní“ studenti budou na vysoké školy přicházet v souvislosti se zavedením nepovinné maturity z matematiky na středních školách.).

Nabízí se několik možností a směrů změny výuky statistiky na VŠE jak z hlediska obsahu zařazených partií statistiky, tak podrobností a hloubky jednotlivých partií. Otevřenou otázkou je i možnost vyučovat statistiku specializovaně pro jednotlivé fakulty. Konkretizace a řešení všech otevřených otázek je ve stadiu rozpracovanosti,

základní ucelená představa by měla být hotova v zimním semestru letošního roku, neboť na systém ECTS přejdou již od školního roku 2005/2006 dvě fakulty VŠE. Protože statistika bude u prvé z přecházejících fakult zařazena do druhého semestru 1.ročníku, budeme mít prvé konkrétní poznatky a zkušenosti z výuky nově koncipovaného redukovaného kurzu statistiky (a rovněž zkušenosti o tom, jak budou studenti tento kurz zvládat) až po skončení letního semestru 2005/2006. Zbývající fakulty VŠE na systém ECTS přejdou od školního roku 2006/2007.

Uvedené připravované změny se týkají výuky statistiky pro ekonomy – nestatistiky. Pro naši fakultu (informatiky a statistiky) se jeví jako nejpřijatelnější varianta vytvořit modifikovaný základní kurz statistiky s upraveným rozsahem a hlavně vyšší matematickou a počítačovou náročností výuky a zařazení dalšího povinného kurzu matematiky. Reálná by pak mohla být pro tuto fakultu úprava obsahu statistiky předpokládaná ve výše zmíněném grantu FRVŠ.

Nyní přejdu k programu výuky statistických předmětů pro statistiky. VŠE v Praze vedle MFF UK v Praze jsou jediné vysoké školy v ČR, které mají v programu statistiku jako ucelený studijní program. Dosavadní studijní program bakalářského oboru Statistika a ekonometrie a navazujícího magisterského studia Statisticko-pojistné inženýrství je následující:

Bakalářské studium studijní obor Statistika a ekonometrie

Celoškolsky povinný studijní základ 125 kreditů

	V tom	
STP201	Pravděpodobnost a statistika	2/1
STP202	Statistické metody	2/1
OBOROVĚ POVINNÉ PŘEDMĚTY 42 kreditů		
blok Statistické metody a pravděpodobnost		
STP314	Pravděpodobnost	2/2
STP401	Matematická statistika	2/2
Státní závěrečná zkouška: Statistika a pravděpodobnost		
blok Matematické metody v ekonomii		
EKO205	Lineární modely	2/2
EKO206	Ekonometrické modely	2/2
EKO309	Programy pro matematické modelování	2/2
blok Hospodářská a sociální statistika a demografie		
DEM201	Základy demografie	2/2
EST407	Hospodářská a sociální statistika	2/2
blok samostatných předmětů a jazyků		
JAZF	Jazyk odborný – úroveň F	0/2
STP102	Výzkumy veřejného mínění	2/0
STP103	Statistická data	1/1
EKO421	Stochastické modely	2/2
FP_303	Finanční analýza a plánování podniku	2/2
OBOROVĚ VOLITELNÉ PŘEDMĚTY - 10 kreditů		
Z nabídky katedry statistiky a pravděpodobnosti		
STP303	Analýza kategoriálních dat	1/1
STP305	Rozhodování podnikatelů při riziku a nejistotě	2/0
STP309	Úvod do finanční a pojistné matematiky	4/0
STP310	Statistické výpočetní prostředí	0/2
STP323	Dějiny statistiky	1/0

Z nabídky katedry ekonometrie

EKO304	Praktikum z operačního výzkumu	2/1
EKO401	Teorie her a ekonomické rozhodování	2/2

Z nabídky katedry ekonomické statistiky

EST506	Mezinárodní srovnávání	0/2

Z nabídky katedry demografie

DEM202	Demografická praktika	0/2
DEM203	Demografický seminář	0/2

Navazující magisterské studium Statisticko-pojistné inženýrství

POVINNÉ PŘEDMĚTY HLAVNÍ SPECIALIZACE 36 kreditů

DEM414	Aktuárská demografie	2/2
EKO402	Simulační modely	2/2
EKO424	Ekonometrie	2/2
EST409	Systém národního účetnictví	2/2
EST508	Statistické hospodářské rozbory	0/2
STP413	Teorie výběrových šetření	2/2
STP420	Životní pojištění	3/0
STP431	Časové řady	2/2
STP502	Vícerozměrné statistické metody	2/1
STP520	Věcné pojištění	2/2

VOLITELNÉ PŘEDMĚTY HLAVNÍ SPECIALIZACE 6 kreditů

STP408	Statistické metody v řízení jakosti a spolehlivosti	2/0
STP412	Subjektivní pravděpodobnost a Bayesovská statistika	2/0
STP419	Životní pojištění - seminář	0/2
STP440	Statistické metody a kapitálové trhy	1/1
STP503	Navrhování experimentů	2/0
STP504	Statistika a SPSS	1/1
STP507	Neparametrické a robustní metody	2/0
STP515	Statistika v SAS	1/1
STP530	Podnikání v pojišťovnictví	2/0
STP540	Pojišťovnictví II	2/0
EST506	Mezinárodní srovnávání	0/2
EKO421	Stochastické modely	2/2
EKO422	Teorie rozhodování	4/2
EKO423	Řízení projektů	2/2
DEM415	Ekonomická demografie	2/2
DEM416	Demografické modely	1/1

Nový program studia pro obor Statistika a ekonometrie a magisterský Statisticko pojistné inženýrství je ve stadiu příprav. Rozsah dosud vyučovaných statistických předmětů v žádném případě nechceme zužovat. Problémem ale zůstává, jak naložit s množstvím specializovaných nízkokreditových kurzů, které v novém systému nebudou mít možnost být zařazeny do studijních plánů, neboť jedna ze zásad přestavby studia omezuje počet zapsaných předmětů v semestru na 4 – 6. Bude tedy nutno některé specializované statistické kurzy, které jsou většinou rozsahu 2/2, 2/1 nebo 2/0 rozšířit, resp. je spojit do větších celků.

Na druhé straně se ale rýsuje možnost zvýšit podíl výuky statistických předmětů ve studiu tím, že bude omezena přebujelá možnost zapisování volitelných předmětů

ostatních oborů a specializací do studijního programů studentů oboru statistika. Tím se i sníží možnost studentů volit si místo relativně těžších exaktních předmětů (nejen statistiky, ale i matematiky, demografie, ekonometrie) a informatiky různé „únikové“ lehké předměty z oblasti obchodu, personalistiky, managementu, podnikání apod.

Literatura:

1. Studijní programy. Akademický rok 2004/2005. VŠE v Praze, fakulta informatiky a statistiky. http://fis.vse.cz.

2. Diskusní materiály VŠE k přípravě ECTS.

Lo a MacKinleyho test podielu rozptylov

RNDr. Mária Bohdalová
Univerzita Komenského v Bratislave, Fakulta Managementu

Abstrakt

Stock market efficiency has been debated by both academics and financial market practicioners. In this paper Lo and Mac Kinley variance ratios to test the random walk model of price behavior is used. Data for this paper follow from Center for research in Security Prices (CRSP) and daily equal weighted index is used to generate weekly price series. Time period used in this paper – September 6, 1962, through December 26, 1985 is that used in Lo and Mac Kinley [Lo88]. My goal is to illustrate how SAS procedures PROC MEANS a DATA can be used in this type of research.

Úvod

Problematika modelovania zmien cien cenných papierov je nanajvýš aktuálna a v konkurenčnej ekonomike často diskutovaná medzi akademickými a finančnými odborníkmi. V príspevku je uvedený menej známy Lo a Mac Kinleyho test podielu rozptylov (variance ratio test) pre otestovanie či je pre *q*-týždenné zmeny cien cenných papierov vhodný stochastický model náhodnej prechádzky. Údaje, ktoré sú použité v príspevku pochádzajú z Centra pre výskum cien cenných papierov (CRSP) a sú z obdobia 6.september 1962 až 26. december 1985, tak ako ich použili Lo a MacKinley vo svojom článku [Lo88]. Zmeny cien cenných papierov sú vypočítané na základe CRSP vážených indexov (CRSP equal weighted index). V príspevku sú uvažované týždenné, dvojtýždenné, štvortýždenné a osemtýždňové zmeny indexu cien pre CRSP equal weighted index. Lo a Mac Kinleyho test podielu rozptylov je naprogramovaný v softvérovom balíku SAS® V8, s použitím procedúry PROC MEANS a DATA kroku.

Ekonomické pozadie a Lo a MacKinleyho test podielu rozptylov

Ceny cenných papierov sa vytvárajú na trhu. Úlohou trhu v konkurenčnej ekonomike je rozdeliť vzácne zdroje medzi konkurenčné odvetvia spôsobom, ktorý povedie k čo možno najefektívnejšiemu využitiu týchto zdrojov. Hodnotenie efektivity trhu s cennými papiermi sa skladá z viacerých hľadísk. Jedno z nich je informačná efektivita (informationally efficient)[Bl95]. Trh je informačne efektívny práve vtedy, keď bežná, tržná cena (current market price) „neustále a úplne odráža všetky relevantné dostupné informácie“ [Bl95]. Tvrdenie, že tržné ceny „neustále a úplne odrážajú všetky relevantné dostupné informácie“ je známe ako tzv. *hypotéza efektívnych trhov* (EHM, efficient market hypothesis) [Bl95]. Ak je toto tvrdenie pravdivé, tak to znamená, že tržné ceny cenných papierov sa vždy rovnajú spravodlivým či fundamentálnym hodnotám týchto cenných papierov. Presnejšie, ak je EHM pravdivá, tak trhy cenných papierov sú v neustálej stochastickej rovnováhe.

Je zrejmé, že prichádzajúce nové informácie resp. novinky sú zo svojej podstaty nepredvídateľné, pretože by to neboli novinky. Dá sa preto očakávať, že cena cenných papierov sa mení v závislosti na nových informáciách tak, že smer a výška zmeny ceny sú opäť nepredvídateľné. Z toho zasa vyplýva, že najlepší odhad zajtrajšej ceny cenného papiera vychádza z dnešnej ceny. Hoci zajtrajšia cena sa celkom môže líšiť od dnešnej ceny, bude sa líšiť spôsobom, ktorý je opäť nepredvídateľný. Preto najlepším odhadom zajtrajšej ceny je cena dnešná. Z toho vyplýva, že pokiaľ je EHM pravdivá môžeme cenu cenných papierov matematicky opísať

stochastickým modelom náhodnej prechádzky (random walk) - známej aj pod názvom martingalský či Brownov pohyb (martingale, Brownian motion). Model náhodnej prechádzky hovorí, že príchod informácií je nepredvídateľný a preto najlepšia predpoveď ceny cenných papierov je ich dnešná (aktuálna) hodnota. Zajtrajšia cena cenných papierov sa rovná dnešnej cene plus čiastke závisiacej na novej informácii, ktorá sa objaví medzi dneškom a zajtrajškom. Tá je ale z hľadiska dnešného súboru informácií Ω_t nepredvídateľná. Základný model náhodnej prechádzky má nasledujúce matematické vyjadrenie:

$$P_t = P_{t-1} + \varepsilon_t \tag{1}$$

kde P_t je dnešná (aktuálna) cena cenného papiera, P_{t-1} je jeho včerajšia cena, resp. cena v predchádzajúcom období (predchádzajúcej perióde) a ε_t je t -ty člen vektora náhodných chýb. Každý člen náhodnej chyby predstavuje príchod novej informácie, ktorá ak má byť nepredvídateľná musí byť nezávislá od všetkých predchádzajúcich chýb. Za predpokladu, že náhodné chyby sú nezávislé a pochádzajú z normovaného normálneho rozdelenia dostaneme štatisticky významný výsledok. (pozri[Bl95])

Lo a MacKinlay [Lo88] vo svojom teste rozpracovali limitné rozdelenie pre odhad podielu rozptylov (variance ratio estimators) s a bez existencie heteroskedasticity údajov a ukázali, že ceny cenných papierov sa nemusia nevyhnutne správať podľa modelu náhodnej prechádzky. Pre svoje výpočty použili predpoklad, že rozptyl vektora náhodných chýb sa mení lineárne s časovým obdobím, v ktorom sú zaznamenané cenové indexy. Presnejšie povedané predpokladali, že rozptyl dvojtýždenných cenových zmien je dvojnásobok rozptylu týždenných cenových zmien; rozptyl mesačných cien sa môže zmeniť štvornásobne v porovnaní s týždennou zmenou cien atď. Táto myšlienka je základom ich testu [Lo88] (prípadne i v [Bo02]) pre otestovanie nulovej hypotézy: q -týždenné zmeny cien cenných papierov je možné modelovať pomocou modelu náhodnej prechádzky, t.j. spĺňajú vzťah (1). Ich alternatívna hypotéza tvrdí opak : q -týždenné zmeny cien cenných papierov nie je možné modelovať pomocou modelu náhodnej prechádzky.

Lo a MacKinley definovali nasledujúce vzťahy:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n}(P_k - P_{k-1}) \tag{2}$$

$$\bar{\sigma}_a^2 = \frac{1}{n-1}\sum_{k=1}^{n}(P_k - P_{k-1} - \hat{\mu})^2 \tag{3}$$

$$\begin{aligned} \bar{\sigma}_q^2 &= \frac{1}{m}\sum_{k=q}^{n}(P_k - P_{k-q} - \hat{\mu})^2 \\ m &= q(n-q+1)\left(1-\frac{q}{n}\right) \end{aligned} \tag{4}$$

Pomocou vzťahu (2) vyjadrili priemer týždenných zmien cien cenných papierov (P_k je cena cenného papiera v k-tom týždni), pričom n je počet uvažovaných týždňov. Vzťahom (3) vyjadrili odhad rozptylu pre týždenné zmeny cien cenných papierov. Vzťahom (4) odhadli rozptyl pre q-týždňové zmeny cien cenných papierov. Písmenom m je označený upravený menovateľ pre q-týždňový odhad rozptylu. Samotný podiel rozptylov definovali nasledovne:

$$\bar{M}_r = \frac{\bar{\sigma}_q}{\bar{\sigma}_a} - 1 \tag{5}$$

Za predpokladu heteroscedasticity, štandardizovaná testovacia štatistika z^* má asymptoticky normované normálne rozdelenie a je definovaná nasledovne:

$$z^* = \sqrt{n}\frac{\bar{M}_r}{\sqrt{\hat{\theta}}} \to N(0,1)$$

$$\text{kde } \hat{\theta} = \sum_{j=1}^{q-1}\left[\frac{2(q-j)}{q}\right]^2 \hat{\delta}(j) \qquad (6)$$

$$\text{a } \hat{\delta}(j) = \frac{n\sum_{j+1}^{n}(P_k - P_{k-1} - \hat{\mu})^2 (P_{k-j} - P_{k-j-1} - \hat{\mu})^2}{(P_k - P_{k-1} - \hat{\mu})^2}$$

Nulová hypotéza sa zamieta na hladine významnosti α ak z^* je väčšie ako $z_{1-\alpha}$ (čo je kritická - tabuľková hodnota normovaného normálneho rozdelenia).

Vzťahy (2) až (6) sú naprogramované pomocou SAS DATA kroku a procedúry PROC MEANS v softwarovom produkte SAS® V8 pre dvoj- a štvortýždenné zmeny cien [Bo02] a pre účely tohto príspevku som ich upravila aj pre 8-týždňové zmeny cien cenných papierov.

Záver

Ukážka údajov:

begwed	endwed	eindbeg	Eindend	ewhpr
5.9.1962	12.9.1962	19.1069	19.1822	-0.004357
12.9.1962	19.9.1962	19.1822	19.1341	0.000331
19.9.1962	26.9.1962	19.1341	18.2523	-0.048601
26.9.1962	3.10.1962	18.2523	18.1306	-0.018000

kde Begwed je dátum prvého pondelka v týždni v časovom súbore údajov,
Endwed je dátum nasledujúceho pondelka v týždni v časovom súbore údajov,
Eindbeg je CRSP equal weighted index ku dňu s dátumom begwed,
Eindend je CRSP equal weighted index ku dňu s dátumom endwed.

Nasledujúca tabuľka obsahuje pre premenné ehpr, ehpr2, ehpr4 a ehpr8, ktoré predstavujú týždenné, dvoj-, štvor- a osem-týždenné cenové zmeny pre CRSP equal weighted index, informácie o ich priemernej hodnote, smerodajnej odchýlke a rozptyle za odpovedajúce časové obdobie.

Variable	N	Mean	Std Dev	Variance
ehpr	1216	0.0034260	0.0221916	0.000492467
ehpr2	1215	0.0068504	0.0357010	0.001274600
ehpr4	1213	0.0137436	0.0567862	0.003224700
ehpr8	1209	0.0276382	0.0872040	0.007604500

Generated by the SAS System on 02SEP2005 at 12:14 PM

Z poslednej tabuľky vidíme, že testovacia štatistika z^* má pre dvojtýždňové zmeny cien cenných papierov hodnotu 7,51232 a preto nulovú hypotézu o možnosti modelovania zmien cien cenných papierov modelom náhodnej prechádzky zamietame na akejkoľvek hladine významnosti α. Podiel rozptylov (variance ratio) pre dvojtýždňové zmeny cien je 1,29512 čo znamená, že približne 30% zmien cien v dvojtýždňovom období môžeme vysvetliť zmenou cien v predchádzajúcich dvoch týždňoch. Na základe týchto výsledkov môžeme povedať, že cenové zmeny nie sú náhodné a teda obsahujú istý stupeň predvídateľnosti. Podobný výsledok sme získali aj pre štvortýždňové a osemtýždňové zmeny cien CRSP váženého indexu.

Number of Weekly Returns	Number q of Week	Variance Ratio for q- Week Returns	Heteroskedastic Robust Test Statistic z^*
1216	2	1.29512	7.51232
1216	4	1.64105	8.88444
1216	8	1.94141	8.49697

Generated by the SAS System on 02SEP2005 at 12:14 PM

Použité SAS moduly a nástroje:
SAS/STAT V8, PROC MEANS, DATA STEP.
Zdroj údajov:
CRSP equal weihted index v období od 6. septembra 1962 do 26. decembra 1985

Literatúra a referencie:

[Lo88]: Lo,Andrew, and Craig MacKinlay: *Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specifikation Test.* Review of Financial Studies 1,1988,41-66
[Bl95]: Blake D.: *Analýza finančných trhov.* 1.vyd .Praha:Grada Publishing,1995.80-7169-201-8
[Bo02]: Boehmer E., Broussard J.P. a Juha-Pekka Kallunki: *Using SAS® in Financial Research*.1.vyd .North Carolina:SAS Institute Inc.,2002.1-59047-039-7

Kontakt autor:
RNDr. Mária Bohdalová
Univerzita Komenského v Bratislave, Fakulta managementu, KIS, Odbojárov 10, Bratislava
E-mail: Maria.Bohdalova@fm.uniba.sk

Automatic Knot Detection in the Piecewise-Cubic Approximation
(Algorithm and MS .NET Components)

N.D. Dikoussar
Laboratory of Information Technologies, JINR, dnd@jinr.ru
Cs. Török
Technical University in Košice, Slovakia, torokcs@tuke.sk

Abstract

Approximation problems remain to be of wide interest in theoretical and experimental sciences, and the technical practice, too. To address these problems we offer a novel approach to data approximation based on a new type of 4-point transformation, the discrete projective transformation. The proposed auto-tracking piecewise cubic approximation divides the interval into subintervals of various lengths and provides for every segment integral cubic approximants. Finding the breakpoints in an auto-tracking mode and the iterative computation schemes are the two main features of the proposed method that uses a special approximation model.

1. Motivation The analysis of dependence between variables is one of the main tasks of technical and scientific research. Methods of approximations are used every day in data analysis and information gain process, and the associated problems are of wide interest in theoretical and experimental sciences. One of the main problems in data/signal denoising, analysis and forecasting is to find an optimal or good representation. Once it is achieved many other goals of drawing information from data become possible.

Piecewise polynomial methods and splines have been widely used. Various approaches and methods are proposed recently in this area. These include the *segment approximation problem* (or the free knots problem in the spline theory) [7, 8], smoothing spline methods and wavelet techniques [9, 10].

2. Problem Statement The segment approximation problem is closely related to the piecewise and spline approximation problems. Spline continuity conditions at the breakpoints *are dropped* in the case of segment approximations. A search interval is divided into *subintervals* and an approximation problem is solved over each of these subintervals. It is clear that different subdivisions into subintervals lead to qualitatively different results. The main goal is to find a subdivision where the errors over the subintervals are as small as possible. The effectiveness of a spline representation of data *depends critically on their number and positions* [7, 8, and 10]. Notice that free knots optimization is a very hard non-linear problem.

3. Automatic knot detection using APCA We suggested a new approach to the analysis of complex dependence with relatively small noise using the four point methodology [3]. The suggested algorithm LOCUSD [1] divides the interval/curve into subintervals/segments of various lengths, provides for every segment local cubic

estimations and gives a technique for obtaining integral cubic approximants. Finding the breakpoints in an *auto-tracking mode* and the *iterative computation schemes* are the two main features of the proposed method that uses a *special* approximation model [1]. MS Visual C# components for autotracking piecewise cubic approximation (APCA): a class library and a Windows-application have been developed too [2].

In our method neither the number of knots nor their placement are unknown. This is very important for applications in approximation and reduces real world data. The knots of the subintervals are detected *in auto- tracking mode* using a digitized curve (data points). A three-point cubic parametric model is used as a local approximant with three control (three different points at *x* axis), three fixed (ordinates on the curve) and one free (1/6 of a third derivative of the model) parameters. A free parameter θ is found in a line following mode, using either step-by-step averaging or the first order recursive least squares method (RLSM). A formula for expression of the free parameter via a length of the segment and values of a function and derivatives in the joining points is received. The C^1-smoothness depends on the accuracy of the θ- estimate.

Let

$$\left\{\mathcal{P}(x_m, \tilde{f}_m)\right\}_{m=\overline{1,N}},\ \tilde{f}_m = f(x_m) + e_m,\ x_m < x_{m+1},\ 4 << N,$$

be a given set of data points, where $e_m \sim \mathcal{N}(0, \sigma^2)$ and the first coordinates of the N (there are at least four) data points $[x_m, \tilde{f}_m]$ are ordered. We consider tetrads $\mathcal{T} = \{\mathcal{P}_\alpha, \mathcal{P}_\beta, \mathcal{P}_0, \mathcal{P}_m\}$, $\mathcal{T} \in \{\mathcal{P}\}$, $\mathcal{P}_i \neq \mathcal{P}_j;\ i \neq j,\ i, j \in (\alpha, \beta, 0, m)$. Three points $\mathcal{R} = (\mathcal{P}_\alpha, \mathcal{P}_\beta, \mathcal{P}_0) \in \mathcal{T}_n$ are called as *reference points* and the fourth one $\mathcal{P}_m$ is a variable point. To approximate a piece of a curve *f* (the segment) at interval $[\alpha, \beta]$ we use a parametric cubic model (Fig. 1)

$$f \approx \mathrm{S} = \Pi + \theta Q, \tag{1}$$

where Π is a quadratic parabola passing via reference points $\mathcal{R}$ and Q is a cubic parabola: $Q = \tau(\tau - \alpha)(\tau - \beta)$; θ is a free parameter. The ordinates of the reference points are used as fixed parameters: $\mathbf{r}_0 = [r_\alpha, r_\beta, r_0]^T$, where $r_* \approx f_*$ or $r_* \equiv f_*$ if $e_* = 0$. Parameters $\mathbf{a} = (\alpha, \beta, x_0)$ and τ_m are defined as $\alpha = x_\alpha - x_0$, $\beta = x_\beta - x_0$ and $\tau_m = x_m - x_0$. These parameters are used to evaluate weight homographic functions $\mathbf{d}_0 = [d_1, d_2, d_3]^T$ and Q. Fig. 1 shows a cubic arc S approximating a piece of a curve f on the subinterval $[\alpha, \beta]$. To obtain $\{d_i\}_{i=1}^3$ we use single-purpose cross-ratio functions [3]. In these terms the parabola equation for the nth segment is written as $\Pi(\tau; \mathbf{a}_n, \mathbf{r}_{0n}) = (\mathbf{r}_{0n}, \mathbf{d}_n)$. θ is the free parameter which is related to a third derivative of the model: $\theta = S'''/6$. Eq. (1) and weight functions construction yield some advantages in development of algorithms: *flexibility*, *control*, *stability*, *simple computations* and so on.

The stability of the method w.r.t. input errors is shown as well. The factors of error suppression (K_n and $d_{in}K_n$) are shown in Fig. 2. The key parameters of the

approximation are: *the parameters of the weight functions*, *the variance of the input errors*, and a *sampling step*.

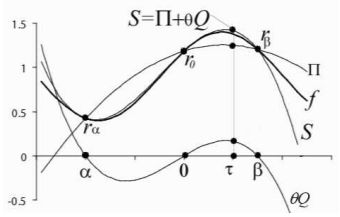


Fig. 1. The cubic model

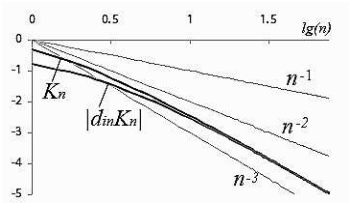


Fig. 2. $Log10\,K_n$ и $log10\left|d_{in}K_n\right|$

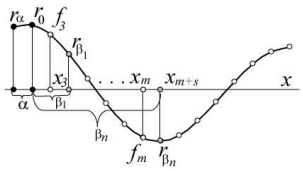


Fig. 3. The choice of tetrads

The constancy of the third derivative of the cubic model ($S'''=const$) is used as a criterion for knot detection in the dynamic mode. This value can be estimated via a local $\hat{\theta}$ using four points on the curve and Eq1. A choice of tetrads at every segment uses two *fixed* points ($\mathcal{P}_\alpha, \mathcal{P}_0$) and two *variable* points ($\mathcal{P}_m, \mathcal{P}_{m+1\equiv\beta_n}$) (Fig 3). To get the global estimate $\hat{\theta}$ at the whole segment we use a recurrent formula using the recursive least-squares method (RLSM):

$$\hat{\theta}_n = \hat{\theta}_{n-1} + K_n\left(\tilde{f}_m - \Pi_{nm} - \hat{\theta}_{n-1}Q_{nm}\right), \hat{\theta}_0 = 0, m \in \{1,...,N\}, n = 1,2,...,n_*, \qquad (2)$$

where $K_n = Q_{nm} / \sum_{j=1}^{n} Q_j^2$ (see Fig. 2). The number n_* is defined as $n_* = m$ under condition $\left|\delta_m\right| > \delta$, where $\delta_m = f_m - \Pi(\tau_m; \mathbf{a}_n, \mathbf{r}_{0n}) - \hat{\theta}_n Q(\tau_m, \alpha, \beta_n)$ and δ is the given control parameter.

The efficiency of the method is shown by numerical calculations on test examples (see Figs. 5 – 9) and real measurements.

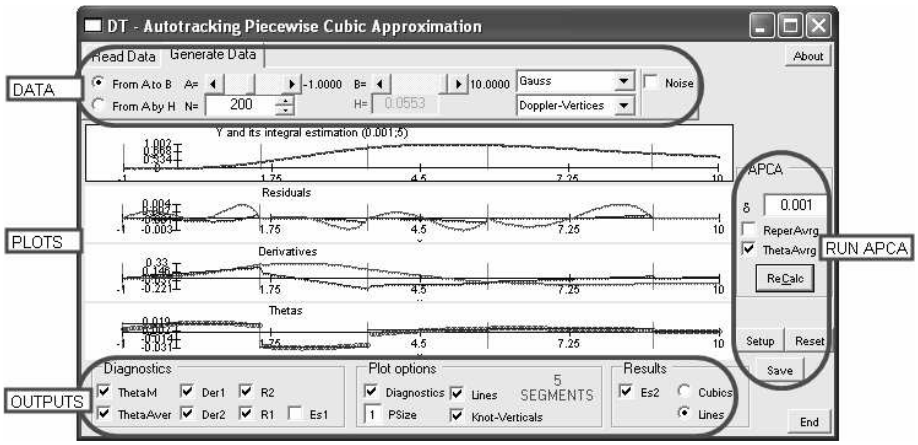


Fig. 4. The Windows application GUI

To perform this analysis and construct data approximants a Windows-application was built based on class components (Fig. 4). We introduced within a .NET framework namespace LinAlg[5] special vector and matrix types with a wide range of object and static numerical, statistical, database and visualization methods, properties and

components that enable to perform in Windows and Web environment not only exploratory data analysis, but also our approximation and compression techniques, and that are extensible and manageable.

The three main objects in APCA are the data points, segments and the interval (set of all segments). Based on them we designed an object-oriented implementation of APCA in MS Visual C# with three classes/components: `Point4`, `Segment`, `SegmentsAll`. Due to this architecture one can easily access a given segment and gain information about it. This feature may play a key role in the component's extension connected with the generalization and improvement of APCA in the future.

4. Examples Figure 5 shows the result of *knot detection* and piecewise-cubic approximation for 134 points situated on the test curve $f = 25/(x^2+25)-0.55Sin(x+2)/(x+2)$ using LOCUSD.

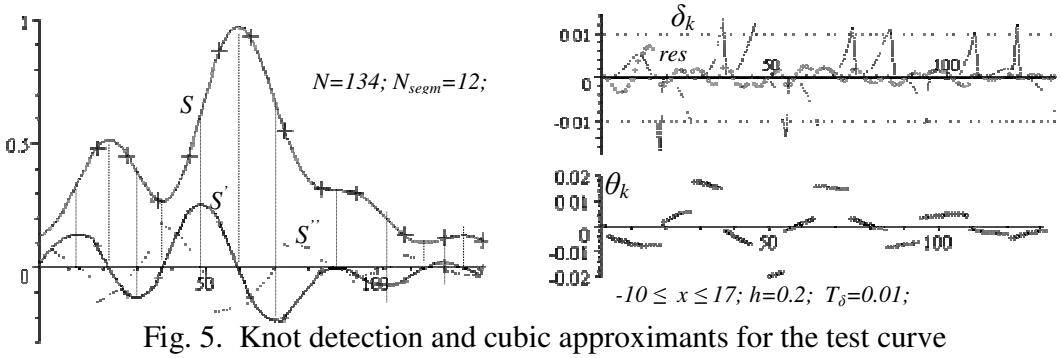


Fig. 5. Knot detection and cubic approximants for the test curve

The data of the following example were gained by numerical modeling of the electron thermal capacity (ETC) for D-acetone molecule and so they are practically without errors [6]. Fig. 6 display the data and their APCA approximants, fig. 7 the estimations of the derivations and fig. 8 the residuals. The quality of the automatically detected 25 approximants ($\delta = 0.005$) is satisfactory. The subintervals are shorter in the left part of the

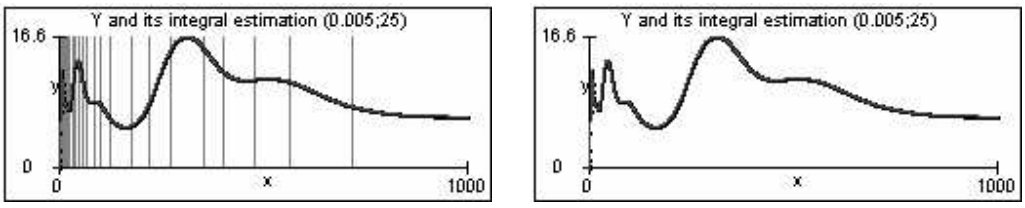


Fig. 6. ETC and its approximants [6]

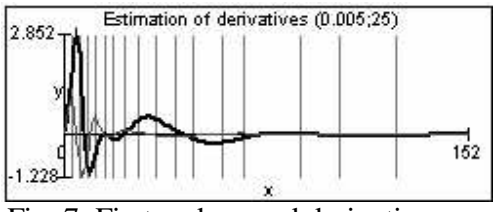


Fig. 7. First and second derivations

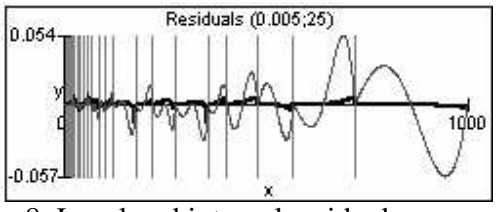


Fig. 8. Local and interval residuals

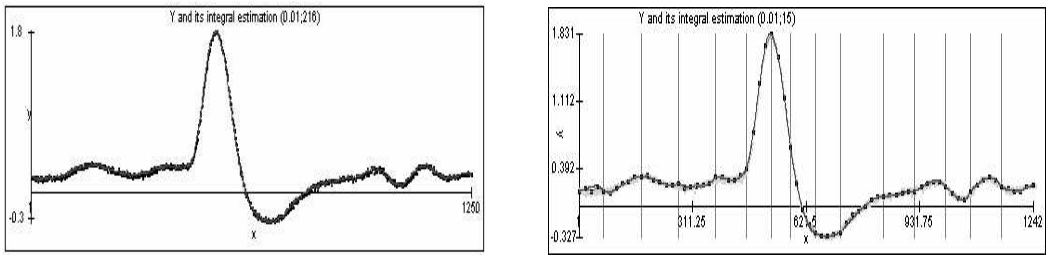


Fig. 9. Knot detection and cubic approximation for data with small noise

figures, where the graph is more dynamic. In fig. 9 both approximants are computed by δ = 0.01, the number of segments is 216 and 15. The right approximation was gained based on every seventeenth point.

5. Conclusion Let us summing up the results:

- an *automatic knot detection and a piecewise-cubic approximation method* are proposed;
- *algorithm LOCUSD*, *MS.NET components* and *Windows-application APCA* for segment approximation and analysis are developed;
- the continuity of the first derivatives of the approximants for functions presented by data without errors are acceptable;
- the goal is to find such δ that yields desirable approximation quality and an acceptable count of segments:
    - a smaller δ results in more segments with more precise approximants;
    - for noisy data it is advisable to choose a greater *sampling step* and δ;

Our plan is to develop methods, algorithms and tools for smoothing data point with a low signal to noise ratio.

Acknowledgement. The second author is thankful to VEGA SR for supporting the project 1/1006/04.

References

[1] *Дикусар Н.Д. Ч. Торок*. Кусочно-кубическая аппроксимация в режиме автоматического слежения, Сообщение ОИЯИ, P11-2004-187, Дубна, 2004; *http://www.jinr.ru/publish/Preprints/2004/187 / (P11-2004-187).pdf.*

Дикусар Н.Д. Ч. Торок. Автоматический поиск узлов для кусочно-кубической аппроксимации, 2005г. Статья направлена в журнал Математическое Моделирование.

[2] *Торок Ч., Дикусар Н.Д.* MS.NET компоненты для кусочно-кубической аппроксимации, Сообщение ОИЯИ, P10-2004-202, Дубна, 2004; *http://www.jinr.ru/publish/Preprints/2004/187(P10-2004-202).pdf.*

[3] *Дикусар Н.Д.* Методы 4-точечных преобразований в задачах аппроксимации и сглаживания кривых и поверхностей, 10-2002-155. Автореферат диссертации на соискание ученой степени доктора физико-математических наук, Дубна 2002; *http://www.jinr.ru/publish/Preprints/*2004/*155(10-2002-155).pdf*.

[4] *Kepič T., Török Cs., Dikoussar N.D.* Wavelet compression, 13. International Workshop on computational statistics, Bratislava (2004) (49-52).
[5] *Török Cs.* Visualization and Data analysis in the MS .NET Framework, JINR Commun., E10-2004-136, Dubna, 2004; *http://www.jinr.ru/publish/Preprints/*2004/ 136(e10-2004-136)*.pdf*.
[6] *Костенко Б.Ф., Прибыш Я*. Теплоемкость электронов в плазме, образующейся при схлопывании кавитационного пузырька в D-ацетоне, Сообщение ОИЯИ, Р11-2004-139, Дубна, 2004; *http://www.jinr.ru/publish/Preprints/*2004/*139(P11-2004-139).pdf*.
[7] *Wolters H. J*. A Newton-type Method for Computing Best Segment Approximations, Communications on Pure and Applied Analysis, 3(1) (2004) 133-149.
[8] *C. Conti at all*. Cubic spline data reduction choosing the knots from a third derivative criterion, Numerical Algorithms, 28 (2001) 45–61.
[9] *G. Farin*. NURBS from Projective Geometry to Practical Use (Second Edition). A K Peters Ltd, Natick, 1999.
[10] *H. Schwetlick and T. Schütze*, Least squares approximation by splines with free knots, BIT 35(1) (1995) 361–384.
[11] *Э. Гуннерсон*. Введение в C#, Издательский дом Питер, Санкт-Петербург, 2001.
[12] *Hastie T., Tibshirani R. and Friedman J.* The Elements of Statistical Learning. Data Mining, Inference and Prediction. New York: Springer, 2001.
[13] *Mao W. and Zhao L.H.* Free-knot polynomial splines with confidence intervals. J. R. Statist. Soc. B, **65**(2003) 901–919.
[14] *Ruppert D.* Selecting the number of knots for penalized splines. Journal of Computational and Graphical Statistics, 11(4) (2002) 735-757.

BIOMETRICKÉ METÓDY PRI RIEŠENÍ VEDECKO-VÝSKUMNÝCH ÚLOH ŽIVOČÍŠNEJ VÝROBY

Pavel Fľak
Výskumný ústav živočíšnej výroby, Hlohovská 2, 949 92 Nitra, Slovensko
flak@vuzv.sk

Abstract. Animal experimentation can be divided into the experiments in which there are used mathematical methods for describing the biological processes (i.e. biosystems) analysed by so-called biomathematical methods and into the typical experiments in which biological events are of stochastical nature, where biometrical methods are applied for analyses. The aim of paper is to review biometrical methods applied in animal experimentation.

Key words: biometrics, designs and evaluation of animal experiments, growth and development, feeding trials, feeds evaluation, physiology and animal health, production traits, ethology, ecology, genetics, biotechnology, econometrics, biosystems.

1. Úvod

Plánovanie a realizácia projektov výskumných úloh v živočíšnej výrobe sú charakteristické špecifickými podmienkami, ktoré podmieňujú nielen samotný plán pokusu ale aj jeho realizáciu. Špecifičnosť pokusov podmieňuje tiež výber vhodných matematicko-štatistických metód, pomocou ktorých je potrebné spoľahlivo hodnotiť získané experimentálne údaje a tým aj adekvátne a efficientne riešiť položenú výskumnú otázku. Cieľom príspevku je preto prehľad aplikovaných biometrických metód užiteľných v zootechnickom výskume podľa jeho vlastnej klasifikácie.

2. Sekvencia plánovania a analýz pokusov

Sekvenciu plánovania a analýz experimetov v zootechnickom výskume možno podobne ako aj v iných vedných oblastiach zhrnúť do bodov:

1. Formulácia experimentálneho objektu a výskumnej otázky.
2. Formulovanie hypotézy/hypotéz.
3. Plánovanie pokusu.
4. Uskutočnenie pokusu.
5. Zber údajov a ich verifikácia na správnosť- nenáhodné vplyvy, napr. odľahlé pozorovania.
6. Analýza experimentálnych údajov pomocou adekvátnych štatistických modelov, obyčajne pomocou lineárnych modelov.
7. Verifikácia postulovaných hypotéz.
8. Testovanie podmienok ukončenia experimentu/ov.
9. Prehodnotenie všetkých urobených krokov a praktické ukončenie pokusu.
10. Formulácia zodpovedajúcich syntetických záverov.

Sekvencia je charakteristická nasledovným reťazcom:

model→ deduktívna teória pravdepodobnosti→ výberová populácia→ induktívna štatistika→ model.

3. Experimentálne údaje a rozdelenie matematicko-štatistických metód v zootechnike

Experimentálne údaje možno získať na základe plánovaných pokusov a tzv. „poľných“ pokusov vhodným výberom. Charakter tzv. „poľných“ experimetálnych údajov v zootechnickom výskume je oproti pokusom v poľnom pokusníctve často charakterizovaný svojou *nevyváženosťou*, t.j. rôznym počtom pozorovaní v najmenších podtriedach $n_{ijkl...}$, ale tiež počtom úrovní jednotlivých faktorov a ich interakcií, ako aj *heterogénnymi výberovými rozptylmi* a tiež tým aj často *nenormalitou* ukazovateľov.

Vzhľadom na takýto charakter zootechnických údajov sa pre ich analýzu využívajú moderné štatistické metódy nevyvážených plánov pokusov, ktoré sa riešia najčastejšie pomocou riešenia zmiešaných lineárnych modelov, teda riešením tzv. systému rovníc *zmiešaného lineárneho modelu.* Aplikované matematicko-štatistické metódy možno rozdeliť na základe počtu ukazovateľov na *jednorozmerné a viacrozmerné*, a na základe poznania teoretických rozdelení ukazovateľov na *parametrické a neparametrické*.

Samotnú aplikáciu biometrických metód používaných pri analýzach zootechnických pokusov možno rozdeliť podľa existujúcej zootechnickej klasifikácie podobne, ako je tomu vo všeobecnej a špeciálnej zootechnike, približne do 10 tried, ktoré sa však navzájom dopĺňajú a prelínajú, a teda neexistujú medzi nimi ostré hranice. V nasledujúcej časti si ich preto stručne a prehľadne uvedieme s príslušnou, avšak nevyčerpávajúcou charakteristikou.

4. Základné typy pokusov v živočíšnej výrobe

4.1. Modely a metódy analýz rastu a vývinu zvierat

Rast a vývin zvierat tvoria prirodzený základ všetkých ostatných pokusov v zootechnických experimentoch. Rastové a vývinové zmeny sú determinované vo väčšine prípadov regresiou od času/veku, ktorý je chronologického charakteru, avšak je biologickej povahy, resp. regresiou od iného determinujúceho regresora. Charakteristiky rastu a vývinu zvierat sledujeme a analyzujeme obyčajne podľa typu experimentálnych údajov, ktoré možno rozdeliť na: *statické, pozdĺžné, priečne* a *kombináciu pozdĺžnych a priečnych údajov*.

Používané biometrické metódy rastu a vývinu zvierat teda úzko závisia na uvedených typoch údajov a z tohto hľadiska *regresné metódy analýz rastu a vývinu* možno rozdeliť na: *lineárne, nelineárne* a *regresné problémy odhadu parametrov nelinearizovateľných funkcií.*

Analýzy pozdĺžných rastových údajov sú veľmi často charakterizované autokoreláciou a autoregresiou. Inými problémami analýz dvoch a niekoľkých ukazovateľov je skutočnosť, že sú merané s vlastnými experimentálnymi chybami, takže vyžadujú aplikáciu regresných metód postihujúcich tento stav, známych pod pojmom *analýza funkčných/funkcionálnych vzťahov*.

Popis rastu zvierat pomocou rastových modelov alebo funkcií je obyčajne postulovaný na adekvátnej formulácií rastových procesov pomocou diferenciálnych rovníc, podobne ako je tomu pri popise biologických javov a procesov, vlastných biomatematickým metódam.

Pre popis rastu zvierat sa používajú najčastejšie nasledovné rastové funkcie: *Brodyho autoakceleračná a autoretardačná rastová funkcia, parabolická Schmalhausenová funkcia, Bertalanffyho rastová funkcia, Gompertzová funkcia, logistická funkcia, Richardsová funkcia, Lehmannová funkcia,* ako aj ďalšie menej známe, avšak v poslednom období využívané funkcie. Relatívny rast časti organizmu a celku sú základnými pojmami tzv. *alometrického rastu* zvierat. V poslednom desaťročí sú predmetom záujmu taktiež: *fenomologické a multifázové/ické rastové funkcie.*

Okrem uvedených typicky regresných metód pri analýze rastu a vývinu zvierat sa používajú taktiež nasledovné štatistické metódy: *analýzy rozptylu (AR) a kovariancie, analýzy rozptylu s opakovanými pozorovaniami*, napr. typu *split-plot* a *zložitejšie modely AR*, a to jednak j*ednorozmerné* a *viacrozmerné AR* a tiež *viacrozmerné štatistické metódy*, ako metóda *základných komponentov, faktorová analýza,* a pod.

Vývin zvierat sa v poslednom období hodnotí pomocou viacrozmerných štatistických metód na základe tzv. *lineárneho hodnotenia,* ktoré sa obyčajne vzťahuje k produkčným ukazovateľom hospodárskych zvierat.

4.2. Plány pokusov hodnotiacich odchov zvierat a porovnávacie výkrmové experimenty

Pokusy tohto druhu prakticky korešpondujú s analýzami rastu a vývinu, hodnotenia výkrmu, jatočnej hodnoty a kvality mäsa (aj pri in vivo hodnotení modernou prístrojovou technikou) hospodárskych zvierat. Pri výkrmových pokusoch výsledky sú závisle na individuálnom resp. skupinovom type výkrmu. Typickými aplikovanými biometrickými metódami sú tu okrem vyššie spomenutých biometrických metód analýz rastu a vývinu, najmä metódy: *bloková analýza rozptylu* (*AR*), *dvoj -, troj-* a *viacfaktorové nalýzy rozptylu, faktoriálne plány pokusov* s uvažovaním interakčných efektov hlavných faktorov *a zložitejšie plány experimentov* a teda aj *analýz kovariancií*, samozrejme zahŕňujúce tiež *sprievodné premenné*. Vhodnými sa javia tiež niektoré *viacrozmerné štatistické metódy,* ako viacrozmerná analýza rozptylu, diskriminačná analýza, analýza hlavných komponentov, faktorová analýza, kanonická korelácia či zhluková analýza.

4.3. Plány pokusov hodnotiacich krmivá a metódy hodnotenia výživnej hodnoty, účinnosti a využiteľnosti krmív

Základnými problémami týchto typov experimentov je štúdium rôznych typov krmív, ich charakteristík a popisu, napr. chemického zloženia a metód hodnotenia ich nutričných hodnôt in vivo a in vitro realizovanými pokusmi, a v súčasnosti tiež mobilných sáčkov účinnosti využitia krmív. Pokusy tohto druhu sú veľmi nákladné a obyčajne sú inštalované iba s limitujúcim počtom, t.j. minimálnym počtom experimetálnych jednotiek, zvierat. Vieryhodnosť týchto skúmaní je potrebné verifikovať pomocou vhodného počtu opakovaní pokusov. Základnými experimentálnymi plánmi a tým aj biometrickými metódami sú: *latinské* a *grécko-latinské štvorce, pokusnícke série, plány pokusov s experimentálnymi pozorovaniami na totožných jedincoch, split-plot, split-split plot,* a pod.

Stupeň degradovateľnosti a stráviteľnosti krmív a iné charakteristiky možno analyzovať pomocou: *regresných analýz (*hlavne *nelineárnych)* a špecifických väčšinou lineárnych (ale tiež nelineárnych) modelov popisujúcich sledované javy a procesy, napr. metabolizmus trávenia a pod.

4.4. Metódy hodnotenia fyziologických procesov reprodukcie a zdravia zvierat

K týmto metódam patria hodnotenie samčích a samičích pohlavných žliaz, analýzy, hodnotenie a konzervovanie semena, procesy tvorby vajíčok a ovulácie, faktory ovplyvňujúce zárodočné bunky, estrálny cyklus, prirodzené a umelé oplodnenie (inseminácia), obdobie a dĺžku gravidity, dlhovekosť, plodnosť samcov a samíc, atď. V súčasnosti sem zaraďujeme aj moderné biotechnologické reprodukčné metódy.

Z veterinárneho, ale tiež hospodárskeho a genetického hľadiska sem patrí štúdium výskytu rôznych chorôb a ich prevencia, ako aj kontrola dedičnosti hospodárskych zvierat. Vhodnými biometrickými metódami týchto oblastí sú: *lineárna a nelineárna regresia, analýzy lineárnych modelov* (*AR* a pod.), *hodnotenie výskytu* biologických javov, napr. *kontingenčné tabuľky,*

loglineárne modely a metódy riešenia *zmiešaných lineárnych modelov pomocou metódy BLUP* (najlepšej lineárnej nevychýlenej predpovede - NLNP) a *BLUP-Animal modelu*.

4.5. Metódy hodnotenia produkčných ukazovateľov a kvality živočíšnych produktov

Metódy hodnotenia ukazovateľov produkcie možno rozdeliť na podskupiny: mliekova produkcia a laktačné krivky, dlhovekosť, mäsová produkcia a ostatné produkty, ako sú vlna, vajcia, atď.

Analýzy produkčných ukazovateľov zvierat sú robené jednak na základe plánovaných pokusov a hlavne z tzv. „poľných pokusov“, pri ktorých sú údaje získavané napr. pri kontrole úžitkovosti produkcie a dedičnosti. Experimentálne údaje, ktoré získavame na kontrolných staniciach sú charakteristické určitým stupňom plánovitých experimentov. Biometrické metódy používané pri hodnotení produkčných ukazovateľoch, okrem mliekovej produkcie, sú obdobné ako tie, ktoré sa využívajú pri raste a vývine zvierat, výkrmových a reprodukčných ukazovateľoch.

Ukazovatele mliekovej produkcie dojníc, oviec a kôz sa hodnotia pomocou riešenia systémov *rovníc zmiešaných lineárnych modelov pomocou metódy BLUP* (NLNP - najlepšej lineárnej nevychýlenej predpovede), BLUP - *Animal modelu* (BLUP-AM) a *Test Day Modelu* (TDM).

Predikcia mliekovej produkcie dojníc, oviec a kôz, ale tiež produkcia vajec nosníc v závisloti na čase využíva hlavne *metódy nelineárnej regresnej analýzy* a tiež špecifické modely predpovede sólovo jedného ukazovateľa, ale tiež simultánne aj niekoľkých ukazovateľov. Biometrické hodnotenie ukazovateľov mäsovej produkcie je obdobné ako pri hodnotení rastu zvierat, odchove a výkrmu jatočných zvierat, pričom sa tiež využívajú metódy BLUP a BLUP-AM. Pri hodnotení produkcie vlny a jej kvality sa používajú špecifické metódy biometriky zohľadňujúce napr. opakovateľnosť nielen v čase, ale tiež v priestore. Je len samozrejmé, že pri hodnotení tak produkčných ukazovateľov, ako aj ukazovateľov kvality živočíšnych produktov majú svoje nezastupiteľné miesto moderné *viacrozmerné štatistické metódy*, napr. pri lineárnom hodnotení typu zvierat a produkčných ukazovateľov.

4.6. Metódy hodnotenia etologických a ekologických pokusov

Pri pokusoch behaviorálneho charakteru sa obyčajne hodnotia jednotlivé zvieratá pomocou údajov typu etogramu za určitých „prirodzených“ a technologických podmienok prostredia. Pri typicky skupinových pokusoch sa hodnotí napr. sociálna aktivita zvierat, sociálne poradie, agresivita, nadriadenosť a podriadenosť a podobne.

Pri etologických pokusoch sa využívajú tak *parametické* ako aj *neparametrické, jedno-* aj *viacrozmerné štatistické metódy.* Pre behaviorálne a sociálne aktivity sa v neposlednej miere používajú tiež *metódy kauzálnej analýzy, teórie informácie, metódy klasifikácie, úsekové* a *štrukturálne analýzy.* Boli skonštruované rôzne modely behaviorálnej aktivity, mobility, emigrácie a imigrácie indivíduí a zmien populačnej štruktúry. Pozornosť bola venovaná niektorým novým javom, ako je konfliktovosť v súvise s veľkosťou skupín, a pod. Pri sledovaní etologických a ekologických pokusov sa venuje zvláštna pozornosť hodnoteniu podmienok prostredia a jeho kvality, podmieňujúcich úroveň produkčných ukazovateľov. Prakticky všetky etologické a ekologické experimenty možno organizovať len v úzkej interakcii prostredia, zvierat a človeka.

4.7. Biometrické metódy populačnej a kvantitatívnej genetiky

V princípe biometricko-genetické hodnotenie je sústredené na *odhad genetických priemerov* a *genetických a negenetických komponentov fenotypovej premenlivosti* pri *odhade genetických parametrov,* ktoré sa využívajú pri *selekcii zvierat* a skupín zvierat za účelom zvyšovania produkcie hospodárskych zvierat. Pre tieto účely sa predpokladá, že analyzované ukazovatele zvierat možno popísať pomocou jednoduchých *lineárnych modelov* (modely Mendela, Fishera, Mathera a Wrighta, resp. tiež modelu ekologicko-genetickej kontroly). Genetické hodnotenie je postulované na definícii *teoretickej skutočnej genetickej hodnoty*, ako teoretického selekčného indexu predikcie skutočnej genetickej hodnoty pomocou lineárnej či nelineárnej kombinácie genetickej informácie známych úžitkovostí príbuzných zvierat pomocou vhodných váh z rodokmeňa.

Z hľadiska štatistického sa obyčajne predpokladá, že analyzované ukazovatele sú charakterizované viacrozmerným normálnym rozdelením, čo implikuje, že ukazovatele sú determinované nekonečným počtom aditívnych génov infinitesimálneho efektu lokusov, ktoré nie sú vo väzbe, teda sú charakterizované tzv. *infinitesimálnym modelom*.

Biometrické metódy odhadu realizovanej genetickej hodnoty, t.j. *plemennej hodnoty (PH)* v čistokrvnej plemenitbe je možno rozdeliť na metódy: *jednoduchých regresných modelov, metódy analýzy rozptylu* a *kovariancie, viacnásobné regresné modely* a *moderné metódy analýz zmiešaných lineárnych modelov.* Najstaršie metódy odhadu PH boli založené na vyjadrení produkčných ukazovateľov v odchýlkach od populačných či subpopulačných priemerov a korekciách na efekty stád, rokov, sezón, vekov a podobne. Do polovice sedemdesiatych až počiatku osemdesiatych rokov metódy odhadu PH boli založené v podstate na: *metóde najmenších štvorcov* a *regresovanej metóde najmenších štvorcov.* V súčasnosti pre odhad PH sa využíva hlavne metóda riešenia *zmiešaných lineárnych modelov* pomocou *metódy BLUP-AM* a *Test Day Model*. Odhady komponentov rozptylu a kovariancií a z nich odvodených genetických parametrov sa realizujú pomocou *ohraničenej*/*reštrikovanej metódy maximálnej vierohodnosti (REML)* a jej variet. Optimalizácia odhadov komponentov rozptylu a tým aj genetických parametrov sa uskutočňuje pomocou *Bayesovej metódy* a metódy *Gibss sampling.*

Genetické hodnotenie *hybridných populácií* je založené na *neaditívnych genetických efektoch*, hlavne dominancie a epistázy t.j. *heterózy*, alebo všeobecne na *odhade efektoch kríženia*. Najčastejšie sa pritom využíva *jedno-* alebo *dvojlokusový aditívno-dominatný model*, alebo *n-lokusový všeobecný lineárny model.* Odhady efektov kríženia sa pri všeobecnom lineárnom modeli robia pomocou: *metódy najmenších štvorcov, váženej metódy najmenších štvorcov* a *zovšeobecnenej metódy najmenších štvorcov*, podľa povahy experimentálnych dát.

Špecifickými metódami odhadu efektov kríženia sú metódy *plánovaných experimentov*, napr. *vrcholového kríženia*, rôzne druhy *dialelného kríženia*, *faktoriálne kríženie, plány multialelného kríženia* a ich kombinácie. Pre odhad efektov kríženia sa využívajú *metódy analýzy rozptylu* a *kovariancie.* V súčasnosti sa uskutočňuje simultánne hodnotenie čistých a hybridných populácií zvierat, pomocou kombinácie genetických/plemenných hodnôt vhodnými váhami.

Významnou oblasťou je hodnotenie *interakcie genotypu* a *prostredia* či už pri mliekovej a mäsovej úžitkovosti i u ostatných živočíšnych produktov, a to tak pri čistokrvnej plemenitbe ako aj pri krížení zvierat a samozrejme tiež pri ich simultánnom hodnotení vhodnými biometrickými metódami. Komplexné genetické hodnotenie populácií zvierat nie je mysliteľné bez aplikácií *viacrozmerných biometricko-genetických metód, odhadu selekčných indexov, zovšeobecneného genetického skupinového indexu* a *viacrozmerného* (viacznakového) *BLUPu* a *BLUP-Animal modelu.* Je však potrebné poukázať na úzky súvis medzi selekčným indexom a poslednými dvoma modernými metódami.

4.8. Biometrické metódy hodnotenia biotechnologických experimentov a genetického inžinierstva

Pre hodnotenie biotechnologických experimentov, reprodukčných procesov a genetického hodnotenia sa používajú metódy popísané v predchádzajúcom odstavci a v časti 4.4 nášho príspevku. Pomocou týchto metód hodnotíme: embryo transfér, delenie/klónovanie, riadenie pohlavia/sexovanie embryí a semena, androgénne pripárovanie a samooplodnenie, chiméry, polyploidiu, genetický polymorfizmus, hodnotenie major génov a genetických markérov (markérovo podporovaná selekcia, MAS), genetické mapovanie, transfér génov, ako aj ďalšie biotechnologické techniky a oblasti genetického inžinierstva.

4.9. Ekonometrické metódy pri genetickom hodnotení zvierat

Živočíšná výroba je silne determinovaná ekonomickými hodnotami vstupu a výstupu a hlavne vlastnou ziskovosťou produkcie. Významú úlohu pri biometricko-genetickom hodnotení populácií zvierat majú teda analýzy zisku a ziskové funkcie, ako aj stanovenie *ekonomických váh* hodnotených ukazovateľov. Významné sú tiež produkčné schopnosti hospodárskych zvierat hodnotené pomocou rôznych produkčných funkcií, ako aj pri genetickom hodnotení komplexných ukazovateľov napr. pomocou *selekčných indexov.* Pritom je potrebné tiež spomenúť metódy *lineárneho* aj *nelineárneho programovania* a špecifické ekonometrické metódy aplikovateľné v šľachtení hospodárskych zvierat, ako sú napr. metódy sledovania technologických zmien v súvise s nákladovými a ziskovými funkciami, modely úžitočnosti, všeobecná lineárna a Cobb-Douglasova produkčná funkcia, či rizikovosť ekonomických hodnôt, a pod.

4.10. Aplikované biometrické metódy modelovania a hodnotenia biologických a poľnohospodárskych systémov

Živočíšná výroba je veľmi zložitý systém. Pre zvládnutie tohto systému sú potrebné všeobecné a špecifické odborové aj medziodborové poznatky, zahŕňujúce najmä fyziológiu, výživu, genetiku, manažment ale i ekonomiku, atď. Základné poznatky z uvedených vedeckých odborov sú koncentrovane využívané pri modelovaní a hodnotení biologických systémov či systémov zvierat, všeobecne teda biosystémov.

Pre analýzy zložitých kvantitatívnych a dynamických interakcií medzi mikro- a makro-systémami zvierat sa využívajú moderné *metódy matematického modelovania, počítačovej simulácie* a samozrejme *vlastného programovania.*

Biometrické metódy tu majú svoje špecifické postavenie, keďže biosystémy zvierat sú veľmi dôležitou súčasťou všeobecne komplexných poľnohospodárskych systémov. Metodologické aspekty biologických systémov v rámci živočíšnej výroby a poľnohospodárskej vedy majú teda primárny význam z hľadiska účinného riešenia budúceho postavenia a významu živočíšnej výroby v poľnohospodárstve.

5. Záver

Zo stručného prehľadu jednotlivých typov pokusov, s ktorými sa stretávame pri riešení vedecko-výskumných otázok živočíšnej výroby, vyplýva ich široká paleta, pričom je potrebné tiež poznamenať, že v súčasnosti sa stretávame taktiež s problémami, pri ktorých je potrebné použiť moderné biomatematické metódy. Samotné riešenia sú nemysliteľné bez použitia modernej výpočtovej techniky a samozrejme bez využívania matematických a štatistických balíkov programov. Vzhľadom k uvedenému členovia Komisie biometriky pri Predsedníctve Slovenskej akadémie pôdohospodárskych vied vítajú organizovanie konferencii typu Stakan či Prastan, a sú presvedčení o užitočnosti spolupráce matematikov, štatistikov a

odborníkov z výpočtovej techniky s biológmi riešiacimi náročné problémy pôdohospodárskeho výskumu.

Odporúčaná literatúra

1. Anděl, J. Matematická statistika. SNTL, Alfa, Praha 1985, 352 s.
2. Batschelet, E.: Introduction to mathematics for life scientist. Springer Verlag, Berlin, Heidelberg, New York, 1971, 495 pp.
3. Bertalanffy, L., Von: General System Theory. Foundations. Development. Applications; Brazilier, New York,1968.
4. Bulmer, M. G.: The Mathematical Theory of Quantitative Genetics. Clarendon Press, Oxford, 1980, 255 pp.
5. Fľak, P.: Biometricko-genetické aspekty hodnotenia biotechnológii. In Zborník Biometrická genetika a systémová analýza v živočíšnej výrobe. Skalský dvůr, 29.-30. 11. 1988, 1988, s. 48-65.
6. Fľak, P.: Biologická a genetická determinácia rastu zvierat. Doktorská dizertačná práca, VÚŽV Nitra, 1990, 485 s.
7. Fľak, P.: Princípy modelovania a simulácie biosystémov a ich aplikácie v zootechnickom výskume. In Zborník Modelovanie biologických systémov, X. letná škola biometriky, Račkova dolina, 8. - 12. 6. 1992, 1992, s. 105-141.
8. Fľak, P.: Biometrical methods for genetic evaluation of livestock, (review). Sborník referátů XIII. letné školy biometriky, Cikháj, 21.- 25. září, 1998, 1998, s. 69-75.
9. Fľak, P.: Linear models in population genetics. Folia Fac. Sci. Nat. Univ. Masarykianae Brunensis, Mathematica 9, 2001, p. 29-52.
10. Fľak, P.: Biometrika a biomatematika v udržateľnej živočíšnej výrobe. In Zborník XVI. letná škola biometriky, Biometrické metódy a modely v pôdohospodárskej vede, výskume a výučbe. Konferencia s medzinárodnou účasťou. Račkova dolina, 21. -25. júna 2004, 2004, s. 11-18.
11. Gianola, D., Hammond, K. (Eds.): Advances in Statistical Methods for Genetic Improvement of Livestock. Springer Verlag. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, 199, 534 pp.
12. Graybill, F. A.: An Introduction to Linear Statistical Methods. Volume I. New York, Toronto, London, Mc Graw-Hill Book Company, Inc., 1961, 463 pp.
13. Grofík, R., Fľak, P.: Štatistické metódy v poľnohospodárstve. Príroda, Bratislava, 1990, 344 s.
14. Havránek, T.: Statistika pro biologické a lékařské vědy. Academia Praha, 1993, 480 s.
15. Hebák, P., Hustopecký, J.: Vícerozměrné statistické metody s aplikacemi. SNTL, Alfa, Praha, 1987, 466 s.
16. Henderson, C. R.: Design and analysis of animal science experiments. In Techniques and Procedures in Animal Science Research. Published by American Society of Animal Science, N. Y. 1969, 12210, p. 2-35.
17. Henderson, C. R.: Application of Linear Models in Animal Breeding. University Guelph., 1984, XXIII + 462 pp.
18. Hill, W. G.: Maintenance of quantitative genetic variation in animal breeding programmes. 49th Ann. Meet. of the EAAP, 24-27 August 1998, Warsaw, Poland, 1998, reprint, 8 pp.
19. Hollander, M., Wolfe, D. A.: Nonparametric Statistical Methods. John Wiley and Sons. New York, London, Sydney, Toronto, 1973, 503 pp.
20. Hušek, R., Walter, J.: Ekonometrie. SNTL, Praha, 1976, 246 s.
21. Parks, J. R.: A Teory of Feeding and Growth of Animals. Springer-Verlag, Berlin,

Heidelberg, New York, 1982, 322 pp.
22. Rao, C. R.: Lineární metódy statistické indukce a jejich aplikace. Academia, Praha, 1978, 666 s.
23. Riggs, D. S.: A Critical Primer The Mathematical Approach to Physiological Problems. The M.I.T. Press, Camdridge, Massachusetts, and London, England, 1963, 445 pp.
24. Snedecor, G. W., Cohran, W.: Statistical Methods. The Iowa State University Press, Ames Iowa, U.S.A., 7th ed., 1982, 507 pp.
25. Żuk, B.: Biometria stosowana. Państwowe Wydawnictwo Naukowe, Warsawa,1989, 425 s.

Možnosti predikcie variability finančných časových radov v podmienkach SR

Rudolf Gavliak
Fakulta financií, Cesta na amfiteáter 1, 974 01 Banská Bystrica
rudolf.gavliak@umb.sk

1. Úvod

V poslednom období sa na svetových finančných trhoch často objavujú veľké výkyvy bez zjavných dôvodov. Výkyvy na kapitálových trhoch sú poslednej dobe spájané s neistotou spôsobenou teroristickou hrozbou, ale aj účtovnými škandálmi v Spojených štátoch. Čoraz častejšie sa objavujú výrazné výkyvy aj na devízových trhoch, a to aj na najvýznamnejších menových pároch. Pri súčasnom objeme medzinárodného obchodu predstavujú tieto výkyvy vážny problém. Z tohto dôvodu je dôležité matematicky analyzovať, modelovať a predikovať výšku volatility vo finančných časových radoch.

2. Cieľ a metodika

Najčastejšie riešenou úlohou v štatistike je modelovanie podmienenej strednej hodnoty. V tomto článku sa ale zaoberáme využitím štatistickej metódy modelovania podmieneného rozptylu, alebo variability premennej. V ekonomických vedách existuje niekoľko dôvodov, prečo modelovať a predvídať volatilitu. V prvom rade je potrebné analyzovať riziko držby aktíva, alebo určiť cenu derivátu na spomínané aktívum. Druhou aplikáciou v ekonomických vedách je modelovanie rozptylu chyby modelu, vzhľadom na časovú nestabilitu intervalov spoľahlivosti. Treťou aplikáciou je zostrojenie robustnejších modelov v prípade heteroskedascitného charakteru chýb modelu. Modely autoregresnej podmienenej heteroskedasticity (ARCH) boli navrhnuté za účelom predpovede podmieneného rozptylu. Rozptyl závislej premennej (výnos) je modelovaný ako funkcia minulých hodnôt. ARCH modely boli zovšeobecnené do podoby všeobecných autoregresných heteroskedasticitných modelov (GARCH). Modely GARCH majú dva parametre, ktoré sa uvádzajú v zátvorke za označením modelu. Hodnota prvého parametra hovorí o hĺbke pamäte autoregresného procesu minulej variability. Hodnota druhého parametra definuje hĺbku pamäte procesu druhých mocnín reziduí modelu výnosu. Pre podmienený rozptyl platí vzťah:

$$\sigma_t^2 = \omega + \alpha_1 \left(y_{t-1} - \beta x_{t-1}^{'} \right)^2 + \alpha_2 \sigma_{t-1}^2 \qquad (1)$$

Tento vzťah považuje za odhad volatility súčet dlhodobého priemeru volatility a váženého priemeru predpovede volatility z predchádzajúceho obdobia, a skutočnej volatility nameranej v predchádzajúcom období. V prípade, že cena aktíva (underlying) neočakávane poklesla, alebo vzrástla, potom vzrastie odhad budúcej volatility. Tento model vyhovuje, často pozorovanému efektu vo finančných časových radoch, kedy obdobie vysokých výnosov je nasledované obdobím s ešte vyšším výnosom a naopak Tento efekt sa nazýva efektom zhlukovania volatility (*volatility clustering effect*). Dá sa dokázať, že vzťah pre podmienený rozptyl modelu GARCH (1,1) (1) je rekurzívnou substitúciou možné upraviť na tvar exponenciálneho váženého súčtu. Postup substitúcie je nasledujúci:

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\sigma_{t-1}^2 = \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\left(\omega + \alpha_1\varepsilon_{t-2}^2 + \alpha_2\sigma_{t-2}^2\right) = \\
&= \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\omega + \alpha_2\alpha_1\varepsilon_{t-2}^2 + \alpha_2^2\left(\omega + \alpha_1\varepsilon_{t-3}^2 + \alpha_2\sigma_{t-3}^2\right) = \\
&\ldots = \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\omega + \alpha_2\alpha_1\varepsilon_{t-2}^2 + \alpha_2^2\omega + \alpha_2^2\alpha_1\varepsilon_{t-3}^2 + \alpha_2^2\alpha_2\left(\ldots\right) = \\
&\ldots = \left(\omega + \alpha_2\omega + \alpha_2^2\omega + \ldots\alpha_2^{\infty}\omega\right) + \\
&\quad + \left(\alpha_1\varepsilon_{t-1}^2 + \alpha_2\alpha_1\varepsilon_{t-2}^2 + \alpha_2^2\alpha_1\varepsilon_{t-3}^2 + \ldots\alpha_2^{j-1}\alpha_1\varepsilon_{t-j}^2\right).
\end{aligned} \tag{2}$$

Zjednodušením vzťahu (2) dospejeme ku konečnému rekurzívnemu vzťahu pre určenie predpokladanej volatility:

$$\sigma_t^2 = \frac{\omega}{\left(1-\alpha_2\right)} + \alpha_1\sum_{j=1}^{n}\alpha_2^{j-1}\cdot\varepsilon_{t-j}^2 \tag{3}$$

Chyba predpovede volatility, ktorej sa dopúšťame je daná nasledujúcim vzťahom:

$$\upsilon_t = \varepsilon_t^2 - \sigma_t^2 \tag{4}$$

Po substitúcii tohto vzťahu do pôvodného vyjadrenia vzťahu pre výpočet predpovede volatility (1) môžeme vykonať nasledujúce úpravy:

$$\begin{aligned}
\sigma_t^2 + \varepsilon_t^2 - \sigma_t^2 &= \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\sigma_{t-1}^2 + \varepsilon_t^2 - \sigma_t^2 \\
\varepsilon_t^2 &= \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\sigma_{t-1}^2 + \upsilon_t \\
\varepsilon_t^2 &= \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\left(\varepsilon_{t-1}^2 - \upsilon_{t-1}\right) + \upsilon_t \\
\varepsilon_t^2 &= \omega + \left(\alpha_1 + \alpha_2\right)\varepsilon_{t-1}^2 - \alpha_2\upsilon_{t-1} + \upsilon_t.
\end{aligned} \tag{5}$$

Zo vzťahu (5) vyplýva, že štvorec chyby modelu je popísaný procesom ARMA (1,1). Súčasná volatilita je popísaná ako funkcia predošlej volatility a kĺzavého priemeru chýb predpovede volatility. Schopnosť zotrvania vplyvu predchádzajúcej chyby na súčasnú hodnotu závisí od súčtu parametrov α_1 a α_2. V ekonomických aplikáciách sa súčet hodnôt týchto parametrov blíži k jednej, čo znamená, že šok z predchádzajúceho obdobia má výrazný vplyv na očakávanú hodnotu nasledujúcej chyby predikcie.

Je zrejmé, že model GARCH (1,1) má tri parametre $\left(\alpha_1,\alpha_2,\omega\right)$, ktoré je potrebné odhadnúť. Existujú taktiež modifikácie modelu GARCH (1,1), pri ktorých je parameter ω nahradený výrazom $\omega = V\left(1-\alpha_1-\alpha_2\right)$, kde *V* je dlhodobý rozptyl. Táto modifikácia sa nazýva cielenie rozptylu (*variance targeting*) [7].

Vo všeobecnosti je model GARCH (*p,q*) pre predpoveď volatility o jedno obdobie dopredu vyjadrený pomocou *p* minulých chýb odhadu výnosu a *q* minulých predpovedí volatility. Tento model má $p+q+1$ parametrov a vzťah pre odhad volatility o jedno obdobie dopredu je nasledujúci:

$$\sigma_{t|t-1}^2 = \omega + \sum_{i=1}^{p}\alpha_{1i}\varepsilon_{t-i}^2 + \sum_{j=1}^{q}\alpha_{1j}\sigma_{t-j,t-j-1}^2. \tag{6}$$

Predpokladajme normalitu rozdelenia chýb ekonometrického modelu výnosu. Potom vierohodnostná funkcia pre odhad parametrov za predpokladu lineárneho regresného modelu výnosu $\left(\hat{y}_t = \beta_1 + \beta_2 x_t\right)$ má nasledujúci tvar:

$$f_y\left(\omega,\alpha_1,\alpha_2,\beta_1,\beta_2\right) = \prod\frac{1}{\sqrt{2\pi\sigma_{t|t-1}^2}}\cdot e^{-\frac{\left(y_t - x_t'\beta\right)^2}{2\sigma_{t|t-1}^2}}. \tag{7}$$

Logaritmus vierohodnostnej funkcie nadobúda nasledujúci tvar:

$$\log f_y^{'}(\omega,\alpha_1,\alpha_2,\beta_1,\beta_2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum\log\left(\sigma^2_{t|t-1}\right) - \frac{1}{2}\sum\frac{\left(y_t - x_t^{'}\beta\right)^2}{\sigma^2_{t|t-1}} =$$
$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum\log\left(\sigma^2_{t|t-1}\right) - \frac{1}{2}\sum\frac{\varepsilon_t^2}{\sigma^2_{t|t-1}}. \tag{8}$$

Keďže prvý člen je konštantou, stačí maximalizovať nasledujúci výraz:

$$f_y(\omega,\alpha_1,\alpha_2,\beta_1,\beta_2) = -\sum\log\left(\sigma^2_{t|t-1}\right) - \sum\frac{\varepsilon_t^2}{\sigma^2_{t|t-1}} =$$
$$= -\sum\log\left(\omega + \alpha_1\varepsilon^2_{t-1} + \alpha_2\sigma^2_{t-1|t-2}\right) - \sum\frac{\varepsilon_t^2}{\omega + \alpha_1\varepsilon^2_{t-1} + \alpha_2\sigma^2_{t-1|t-2}}. \tag{9}$$

Pre odhad hodnôt vektora parametrov $\vec{\theta} = (\omega,\alpha_1,\alpha_2,\beta_1,\beta_2)$, ktoré je možné odhadnúť z tohto vzťahu, platí tento vzťah:

$$\hat{\theta} = \underset{\theta=\hat{\theta}}{\arg\max}\left(-\sum_t\log\left(\sigma^2_{t|t-1}\right) - \sum\frac{\varepsilon_t^2}{\sigma^2_{t|t-1}}\right). \tag{10}$$

Uvedený problém je možné formulovať aj ako úlohu lineárneho programovania, kde účelovú funkciu predstavuje predchádzajúci vzťah za predpokladu, že parametre modelu spĺňajú nasledujúce kritériá:

$$\alpha_1 + \alpha_2 < 1,$$
$$\omega,\alpha_1,\alpha_2 \geq 0. \tag{11}$$

3. Model a dáta

Vstupné dáta predstavujú tempá rastu výmenného kurzu EUR/SKK od 5. januára 1999 až do 7. júna 2005, čo predstavuje 1 607 údajov. Z údajov výmenného kurzu P_t sme v prvom kroku vypočítali tempá rastu kurzu $R_t \approx \log(P_t / P_{t-1})$. Volatilitu výmenného kurzu (R_t^2) sme v prvom kroku modelovali pomocou modelu GARCH (1,1), pričom trend výnosu (tempa zmeny) výmenného kurzu bol modelovaný pomocou konštanty. Navrhnutý model má nasledujúci tvar:

$$R_t = -8\cdot10^{-5} + \varepsilon_t \tag{12}$$

$$\sigma_t^2 = 9.92\cdot10^{-8} + 0.24272\cdot\varepsilon_{t-1} + 0.741069\cdot\sigma^2_{t-1} \tag{13}$$

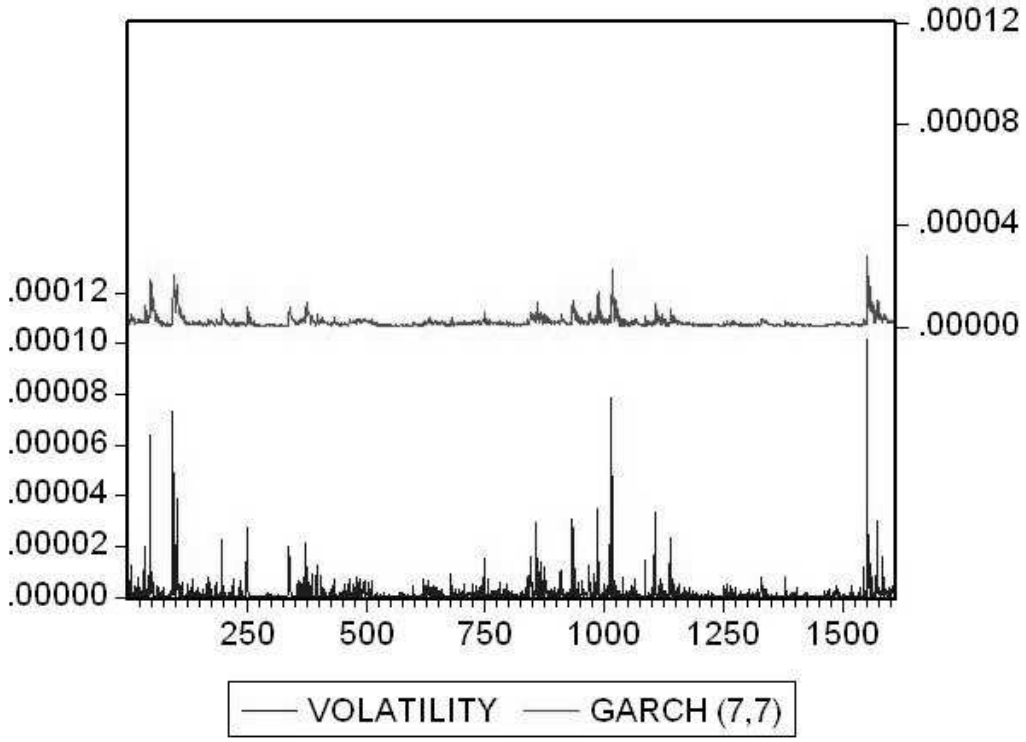
Koeficienty uvedené vo vzťahoch sú štatisticky významné, avšak hodnota Durbin – Watsonovej štatistiky na úrovni 1,8 poukazuje na zostatkovú heteroskedasticitu v reziduách. Na histograme štandardizovaných reziduí (e_t / σ_t) je viditeľná vysoká špicatosť rozdelenia. Jarque – Bera test zamietol hypotézu o normálnom rozdelení reziduí na všetkých bežných hladinách významnosti. Na základe hodnôt ARCH LM testov zamietame nulovú hypotézu o homoskedasticite reziduí. Na odstránenie podmienenej heteroskedasticity bolo potrebné zvýšiť stupeň modelu. Minimálny stupeň modelu, ktorý výrazne redukuje štatistickú významnosť zostatkovej autokorelácie je tvar GARCH (3,3). Hypotézu o rozptyle reziduí nezávislom na čase môžeme prijať až po zvýšení stupňa modelu na GARCH (7,7). Daňou za odstránenie autokorelácie reziduí je štatistická nevýznamnosť šiestich parametrov, keď na hladine významnosti $\alpha = 0.05$ nemôžeme zamietnuť nulovú hypotézu o nulovej hodnote parametrov. Konečný tvar pre modelovanie volatility je GARCH (3,5), t. z. že model obsahuje tri významné multiplikátory predchádzajúcich chýb predpovede a päť významných autoregresných koeficientov pre historické predpovede volatility. Tvar rovnice pre predpoveď tempa

rastu (poklesu) kurzu EUR/SKK (*mean equation*) a predpoveď volatility má nasledujúci tvar:

$$R_t = -6.23 \cdot 10^{-5} + \varepsilon_t \tag{14}$$

$$\begin{aligned}\sigma_t^2 = {} & 4.7 \cdot 10^{-7} + 0.256438 \cdot \varepsilon_{t-1} + 0.292404 \cdot \varepsilon_{t-2} + 0.317525 \cdot \varepsilon_{t-3} - 0.533773 \cdot \sigma_{t-1}^2 \\ & - 0.705318 \cdot \sigma_{t-2}^2 + 0.232341 \cdot \sigma_{t-3}^2 + 0.379112 \cdot \sigma_{t-5}^2 + 0.479138 \cdot \sigma_{t-7}^2\end{aligned} \tag{15}$$

Časový rad vypočítanej volatility a modelu GARCH (7,7) sú uvedené na obrázku 1.



Obr. 1 Porovnanie volatility výmenného kurzu EUR/SKK a modelu GARCH (7,7)

Z vizuálneho posúdenia je zrejmé, že uvedený model nie je schopný v plnej miere zachytiť extrémne hodnoty volatility. Na druhej strane správne identifikuje okamih a relatívny význam extrémnych hodnôt volatility. Verifikácia modelu potvrdila odstránenie autokorelácie reziduí a ARCH LM testy potvrdili, že ďalšie zvýšenie stupňa modelu neodstráni heteroskedasticitu reziduí. Špicatosť rozdelenia reziduí sa znížila z hodnoty 8.72 na 7.63. Jarque – Bera štatistika nepotvrdila normalitu rozdelenia reziduí.

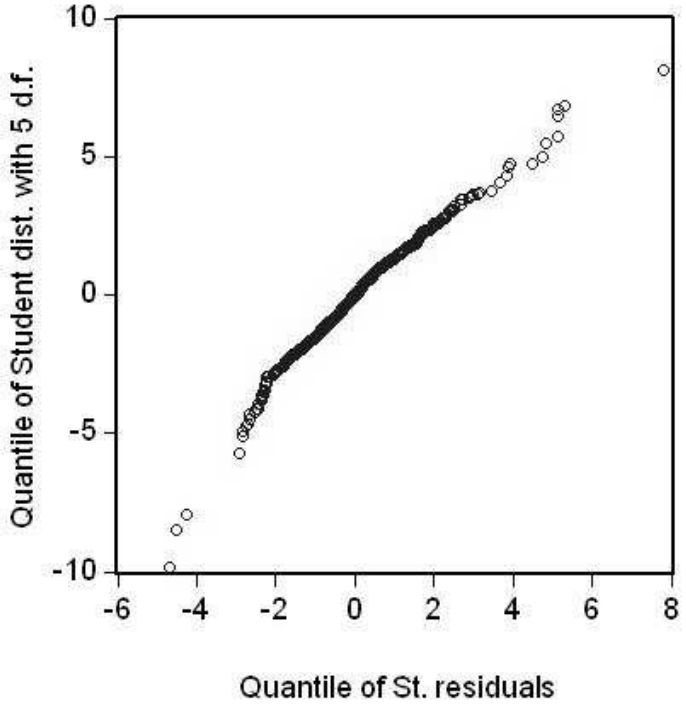
Vzhľadom na neuspokojivé výsledky normality rozdelenia reziduí, sme testovali možnosť Studentovho rozdelenia reziduí. Najvhodnejšie sa javí Studentovo rozdelenie s piatimi stupňami voľnosti. Na obrázku 2 je znázornená závislosť kvantilov Studentovho rozdelenia a kvantilov reziduí.

V prípade, že by sa rozdelenie reziduí riadilo testovaným rozdelením, bola by závislosť kvantilov rozdelenia reziduí a teoretického rozdelenia lineárna, t. z., že na grafe by bola znázornená „rovná čiara“. Posúdenie je diskutabilné, ale Studentovo rozdelenie s piatimi stupňami voľnosti zodpovedá pravepodobnostnému rozdeleniu reziduí lepšie, ako je tomu pri normálnom rozdelení.

4. Predpoveď rozptylu

Cieľom modelov volatility nie je iba popísanie modelom, ale aj predikcia volatility v horizonte *h*. Vo všeobecnosti predpokladajme model GARCH (*p, q*) v tvare:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_p \sigma_{t-1}^2. \tag{16}$$



Obr. 2 Závislosť kvantilov Studentovho rozdelenia a kvantilov reziduí

Potom hodnotu predpovede volatility s najmenšou štvorcovou chybou je možné zapísať v tvare:

$$\sigma_T^2(h) = \omega + \alpha_1 \varepsilon_{T+h-1}^2 + \ldots + \alpha_q \varepsilon_{T+h-q}^2 + \beta_1 \sigma_{T+h-1}^2 + \ldots + \beta_p \sigma_{T+h-p}^2. \tag{17}$$

V prípade predpovedi konštruovanej na základe modelu GARCH (1,1) pre podmienený rozptyl v čase $t = T + h$ má vzťah pre jeho výpočet nasledujúci tvar:

$$\sigma_{T+h}^2 = \omega + \alpha_1 \varepsilon_{T+h-1}^2 + \alpha_2 \sigma_{T+h-1}^2 \tag{18}$$

Po úprave je možné predpovedanú hodnotu budúceho rozptylu vyjadriť nasledujúcim vzťahom:

$$\begin{aligned} \sigma_T^2(h) &= \omega \sum_{j=0}^{h-1} (\alpha_1 + \alpha_2)^j + (\alpha_1 + \alpha_2)^{h-1} \alpha_1 \varepsilon_T^2 + (\alpha_1 + \alpha_2)^{h-1} \alpha_2 \sigma_T^2 = \\ &= \omega \sum_{j=0}^{h-2} (\alpha_1 + \alpha_2)^j + (\alpha_1 + \alpha_2)^{h-1} \sigma_{T+1}^2. \end{aligned} \tag{19}$$

V prípade, že je model GARCH (1,1) stacionárny v kovarianciách, musia parametre modelu spĺňať nasledujúce kritériá:

$$\begin{aligned} \alpha_1 + \alpha_2 &< 1 \\ \sigma_\varepsilon^2 &= \frac{\omega}{(1 - \alpha_1 - \alpha_2)} \end{aligned} \tag{20}$$

Potom vzťah pre odhad podmienenej volatility nadobudne nasledujúci tvar:

$$\sigma_T^2(h) = \sigma_\varepsilon^2 + (\alpha_1 + \alpha_2)^{h-1} (\sigma_{T+1}^2 - \sigma_\varepsilon^2). \tag{21}$$

Pomocou uvedených vzťahov odhadneme volatilitu o štyri obdobia dopredu ($\sigma^2_{t+4|t}$) pomocou modelu volatility GARCH (1,1) a modelu volatility GARCH (7,7) v dňoch od 8. júna 2005 do 13. júna 2005. Analyzovaný časový rad je tvorený kurzom NBS EUR/SKK vyhlásený v predchádzajúci obchodný deň o 12.00 hod. SEČ. Výsledné porovnanie je uvedené v Tab. 1.

Tab. 1 Porovnanie odhadovanej volatility so skutočnou volatilitou kurzu EUR/SKK

	Volatilita (σ2t+4,t)		Kurz (P_t)				Skutočnosť	
Dátum	GARCH(1,1)	GARCH(7,7)	GARCH(1,1)		GARCH(7,7)		Volatilita	Kurz (P_t)
8.6.2005	0,00021%	0,00005%	38,70	38,59	38,67	38,62	0,00029%	38,58
9.6.2005	0,00037%	0,00014%	38,77	38,51	38,72	38,57	0,00048%	38,66
10.6.2005	0,00041%	0,00045%	38,85	38,44	38,80	38,49	0,00023%	38,61
13.6.2005	0,00033%	0,00035%	38,92	38,37	38,87	38,42	0,00005%	38,58

5. Záver

Zvýšená volatilita menových kurzov prináša na jednej strane ziskové príležitosti, či už na spotovom, alebo termínovanom trhu, preto si matematické modely volatility našli uplatnenie pri oceňovaní derivátov. Väčšina ekonomických subjektov ale nemá prístup na kapitálový trh. Pre podniky predstavuje volatilita menového kurzu medzi domácou menou a menou krajiny dovozcu, resp. vývozcu riziko finančných strát. Pre podniky nie je preto dôležitá iba absolútna výška menového kurzu, ale aj intenzita a výška výkyvov menového kurzu. Pre spotrebiteľov je takisto vhodnejší stabilný vývoj menového kurzu, keďže potenciálne finančné straty podnikov sa premietnu do úrovne cenovej hladiny. Vyhlásenia typu „silná koruna – lacná dovolenka“ sú zavádzajúce. Krátkodobé zisky sa v dlhodobom horizonte navzájom vyrušia, ale v cenovej hladine zostáva započítané finančné riziko spôsobené pohybom menových kurzov.

V príspevku sme modelovali tempá rastu (výnosy) menového kurzu EUR/SKK pomocou modelov GARCH (Generalized Autoregressive Conditional Heteroskedasticity). Najjednoduchším modelom bol GARCH (1,1), ktorý dokázal správne identifikovať obdobia s vysokou volatilitou avšak nedokázal modelovať extrémne výkyvy. Štatistická verifikácia objavila autokoreláciu reziduí, výrazná bola najmä autokorelácia reziduí posunutých o sedem období. Napriek snahe, aby výsledný model bol GARCH model čo najnižšieho stupňa, museli sme nakoniec na reprezentáciu analyzovaného časového radu zvoliť model GARCH (7,7). ARCH LM testy zamietli hypotézu o zostatkovej autokorelácii reziduí výnosov. Problematická, ale zostala vysoká špicatosť rozdelenia reziduí, ktorá nezodpovedá normálnemu rozdeleniu reziduí. Testovali sme zhodu so Studentovým rozdelením a rozdelenie reziduí najlepšie zodpovedá Studentovmu rozdeleniu s piatimi stupňami voľnosti.

Ďalej sme určili predikcie volatility v krátkodobom horizonte štyroch dní ($h = 4$). Platí, že s rastúcim horizontom predikcie podmienený rozptyl monotónne konverguje k nepodmienenému rozptylu a význam súčasnej informácie sa stráca. Už GARCH (1,1) dáva v krátkodobom horizonte dobré výsledky. Teória ďalej hovorí, že na predikciu v dlhšom časovom horizonte je vhodný model vyššieho stupňa. Výsledné hodnoty predikovanej volatility modelom GARCH (1,1) a GARCH (7,7) sú podobné. Predikcie volatility získané modelom GARCH (1,1) sa ale prekvapivo viac blížia pozorovanej volatilite. Naopak hranice, v ktorých by sa mal pohybovať kurz EUR/SKK sú užšie v prípade modelu GARCH (7,7) (viď Tab. 1), čo znamená presnejšiu predstavu o vývoji budúceho kurzu EUR/SKK.

6. Literatúra

1. ARLT, J. – ARLTOVÁ, M. 2003. Finanční časové řady.Vlastnosti, metody modelování, příklady a aplikace. Praha : Grada Publishing, 2003. ISBN 80-247-0330-0.
2. ENGLE, R. F. 1983. Estimates of the Variance of U.S. Inflation based upon the ARCH model. In: Journal of Money, Credit and Banking. Volume 3, No. 15, 1983. s. 286 – 301.
3. ENGLE, R. F. – ROSENBERG, J. 1995. GARCH Gamma. Cambridge: NBER Working Paper, Working Paper No. 5128, 1995.
4. ENGLE, R. F. 2001. The Use of GARCH/ARCH Models in Applied Econometrics. In: Journal of Economic Perspectives. Volume 15, No. 4, 2001. s. 157 – 168.
5. GAZDA, V. – VÝROST, T. 2003. Application of GARCH Models in forecasting the volatility of the Slovak share index (SAX). In: Biatec, Volume XI, 2/2003. Bratislava : NBS, 2003.
6. [b.a.] EViews 4 Users's Guide. Irvine: Quantitative Micro Software, 2002. ISBN 1-880411-28-8.
7. ZMEŠKAL, Z. A KOL. 2004. Finanční modely. Praha : EKOPRESS, 2004. ISBN 80-86119-87-4.

Kontaktná adresa:
Ing. Rudolf Gavliak
Katedra kvantitatívnych metód Fakulty financií UMB Banská Bystrica
Cesta na amfiteáter 1
974 01 Banská Bystrica
rudolf.gavliak@umb.sk

Patrí Kalmanov filter do základného kurzu analýzy časových radov?

M. Grendár
Inštitút matematiky a informatiky, Severná 5, 974 01 Banská Bystrica
umergren@savba.sk

Úvodná poznámka

Nasledujúci text je pracovnou verziou kapitoly z pripravovaných skrípt k jednosemestrovej prednáške 'Úvod do ekonometrickej analýzy časových radov' pre študentov odboru Finančná matematika a štatistika na FPV UMB.

Dva dôvody prečo by študent, ktorého štátnicovým predmetom je štatistika, mal byť oboznámený s Kalmanovým filtrom (KF) sú uvedené v texte: KF sa používa 1) na výpočet hodnoty exaktnej vierohodnostnej funkcie ARMA modelu a 2) v štrukturálnom prístupe k analýze časových radov. Okrem toho, stavový model (1), (2) s ktorým je KF bezprostredne zviazaný, je dostatočne bohatý na to, aby v sebe zahŕňal aj klasický zmiešaný lineárny regresný model a longitudinálne modely. S nimi by sa študent štatistiky tiež mal raz stretnúť. Takže základná informácia o KF by nemala nikomu zaškodiť.

Pochopenie KF z bayesovskej strany si vyžaduje len znalosti zo základného kurzu pravdepodobnosti (Bayesova veta) a štatistiky (vlastnosti normálneho a združeného normálneho rozdelenia).

Kalman filter and likelihood evaluation *(Lecture notes)*

1. Introduction

Though the conditional likelihood estimation of a gaussian ARMA model is rather straightforward, it has undesirable properties (the estimate depends upon the conditioning first observation value, hence estimate obtained from reverted time series is in general different). Exact likelihood function uses all data points equivalently, however it is harder to calculate and maximize. One way how the exact likelihood function of a covariance stationary gaussian ARMA time series can be calculated is by means of Kalman filter.

2. Kalman filter

Kalman filter (KF) was devised for a purpose rather different than likelihood function evaluation.

2.1 States, observations, filtration problem

Let there be an object whose state is at any moment n characterized by a fixed set of variables (e.g. position, speed, acceleration), gathered into state vector x_n. Let the time evolution of state follows a continuous Markov process: $x_n = Ax_{n-1} + v_n$, where v is a zero-mean white noise. The state is not directly observable. Rather observable variables y are measured, which are linearly dependent on the state variables x and the dependence is distorted by a zero-mean white noise (measurement error) w, i.e., $y_n = Cx_n + w_n$.

Full technical specification of the model: x_n is a state vector of dimension m, y_n

is an observation vector of dimension d.

$$x_n = Ax_{n-1} + v_n \tag{1}$$
$$y_n = Cx_n + w_n \tag{2}$$

where the first equation is called state or system equation, the second one is commonly known as measurement or observation equation. v and w are zero-mean vector white noise processes. Q is covariance matrix of v, R is covariance matrix of w. Dimensions of matrices: A, Q are $m \times m$; C is $d \times m$ and R is $d \times d$. It is assumed that v_n, w_n are uncorrelated with the state x_n. This implies [fill-in-gap 1[1]] that: i) $E(x_n, w_l) = 0, \forall n, l$; ii) $E(x_n, v_l) = 0$ for $n \leq l$; iii) $E(y_n, w_l) = 0$ for $n \leq l-1$; iv) $E(y_n, v_l) = 0$ for $n \leq l$; v) $E(v_n, w_n) = 0, \forall n, l$; vi) $E(v_n, v_l) = 0$ for $n \neq l$ and $= Q$ for $n = l$; vii) $E(w_n, w_l) = 0$ for $n \neq l$ and $= R$ for $n = l$. The matrices A, C, Q, R are assumed to be known, as is also mean μ and covariance Σ of initial state x_0.

Given a time series of observations y_1^n, known transition matrices A, C as well as noise covariance matrices Q (for v_n), R (for w_n) and a characterization of the initial state x_0 it is desirable to estimate the sequence of states x_1^n. This task constitutes a filtration problem.

The filtration problem can be approached in two ways: i) the state vector x_1^n is estimated at once (say by ML method); when new observation y_{n+1} arrives the estimation should be done anew, or ii) sequentially/adaptively; previously estimated sequence is used as an input to calculation of the estimate of $n+1$-st state. In the latter case the filtration problem turns into a task of estimating x_n, sequentially. This can be done via Bayes filter.

2.2 Bayes filter

Instead of devising such a sequential estimator (filter) of $x_n|y_1^n$ and calculating its co-variance it appears more tractable to solve a seemingly harder problem of estimating the conditional distribution $p(x_n|y_1^n)$. A point characteristic of the estimated conditional distribution (e.g. mean, mode) can be then used as the desired estimator of x_n.

The filtration will be performed in two steps: a prediction and a correction. The steps will be repeated sequentially. A density $p(x_n|y_1^{n-1})$ resulting from prediction step will be used as an input to the correction step. Then the density $p(x_n|y_1^n)$ which results from the correction step will enter the next prediction step. And so on.

Prediction step Given the first $n-1$ observations and $p(x_{n-1}|y_1^{n-1})$ the probability $p(x_n|y_1^{n-1})$ is calculated.

The predicted density can be obtained via

$$p(x_n|y_1^{n-1}) = \int p(x_n|x_{n-1})\, p(x_{n-1}|y_1^{n-1})\, dx_{n-1}, \tag{3}$$

which results directly from the property [fig2]

$$p(x_n|y_1^{n-1}, x_{n-1}) = p(x_n|x_{n-1}). \tag{4}$$

Correction step The predicted pdf $p(x_n|y_1^{n-1})$ is combined with the measurement/observation y_n in order to get the distribution of $x_n|y_1^n$.

The conditional probability $p(x_n|y_1^n)$ can be obtained via Bayes formula. Equivalently, it can be constructed from definition as the ratio $p(x_n, y_1^n)/p(y_1^n)$. Since

[1] Abbreviated: fig1.

$p(x_n, y_1^n) = p(y_n|x_n, y_1^{n-1})p(x_n, y_1^{n-1})$ the joint probability $p(x_n|y_1^n) = p(y_n|x_n, y_1^{n-1}) \cdot p(x_n|y_1^{n-1})/p(y_n|y_1^{n-1})$. This can be simplified, by finding that [fig3]

$$p(y_n|x_n, y_1^{n-1}) = p(y_n|x_n) \tag{5}$$

into

$$p(x_n|y_1^n) = \frac{p(y_n|x_n)p(x_n|y_1^{n-1})}{p(y_n|y_1^{n-1})}. \tag{6}$$

Bayes filter is based on the prediction step (3) and correction step (6). In order to make it operational, the relevant distributions should be known/specified which amounts to specification of distribution of v_n, w_n and x_0.

2.3 Kalman Filter: gaussian Bayes filter

Assume that distribution of v is multivariate normal $n(0, Q)$, and distribution of w is as well gaussian $n(0, R)$. Let also the distribution of initial state be gaussian, $x_0 \sim n(\hat{x}_0, P_0)$. Gaussian Bayes filter is known as Kalman filter. In this special case the densities of prediction step (3) and correction step (6) can be calculated analytically.

Let $x_{n-1}|y_1^{n-1} \sim n(\hat{x}_{n-1}, P_{n-1})$ be the density which resulted from $(n-1)$-st correction step; [q1: how do we know that it is gaussian?]. Thanks to normality of the distribution $p(x_{n-1}|y_1^{n-1})$, the prediction step can be done without having to use (3). Indeed, since states evolve according to $x_n = Ax_{n-1} + v_n$, and thanks to the well-known properties of gaussian distribution (i.e., *i*) if $u \sim n(\mu, \Sigma)$ then $v = Du$ is $v \sim n(D\mu, D\Sigma D')$ and *ii*) $u + v \sim n(\mu + D\mu, \Sigma + D\Sigma D')$) the sought prediction step distribution is

$$x_n|y_1^{n-1} \sim n(A\hat{x}_{n-1}, AP_{n-1}A' + Q) \triangleq n(\hat{x}_n^-, P_n^-). \tag{7}$$

The prediction step density is then corrected via (6). Evaluation of the right-hand-side of (6) is not an easy task. Fortunately, the need to evaluate the RHS of (6) can be avoided, thanks to the following Theorem:

Thm. *Let X_1 and X_2 have a bivariate normal distribution with means μ_1 and μ_2 and a covariance matrix Σ. Then the conditional distribution of X_1 given X_2 is: $X_1|X_2 = x_2 \sim n(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$. And vice versa: if the conditional distribution has this form than X_1 and X_2 are the bivariate normal.*

The Theorem permits to find [fig4] the correction step distribution:

$$x_n|y_1^n \sim n(SC'R^{-1}y_n + S(P_n^-)^{-1}\hat{x}_n^-, S) \tag{8}$$

where $S^{-1} \triangleq C'R^{-1}C + (P_n^-)^{-1}$. Thanks to a matrix inversion formula: $(A^{-1} + B^{-1})^{-1} = A - A(A+B)^{-1}A$, mean and covariance of the density (8) can be simplified:

$$x_n|y_1^n \sim n(Ky_n + [I - KC]\hat{x}_n^-, [I - KC]P_n^-) \triangleq n(\hat{x}_n, P_n) \tag{9}$$

where $K \triangleq P_n^-C'[CP_n^-C' + R]^{-1}$ is $m \times d$ matrix known as Kalman gain (matrix).

In summary, gaussian Bayes filtering steps are:

Predict $n(\hat{x}_{n-1}, P_{n-1}) \Rightarrow n(\hat{x}_n^-, P_n^-)$:

1. $\hat{x}_n^- = A\hat{x}_{n-1}$

2. $P_n^- = AP_{n-1}A' + Q$

<u>*Correct*</u> $n(\hat{x}_n^-, P_n^-) \Rightarrow n(\hat{x}_n, P_n)$:

1. $K = P_n^- C'[CP_n^- C' + R]^{-1}$
2. $\hat{x}_n = \hat{x}_n^- + K[y_n - C\hat{x}_n^-]$
3. $P_n = [I - KC]P_n^-$

Mean $\hat{x}_n$ of $x_n|y_1^n$ is KF forecast of x_n, given y_1^n. It is a minMSE forecast. Covariance P_n is MSE of the forecast. Note that $C\hat{x}_n^-$ is KF forecast of y_n based on the past observations y_1^{n-1}. Hence, $y_n - C\hat{x}_n^-$ is y_n-forecast error.

2.4 Initialization

Yet, one issue remains unsettled: how to initialize KF? More precisely, how to select the zero-time-period parameters of the distribution of x_0: mean $\hat{x}_0$ and covariance P_0?

At $n = 0$ there is no x, y available. So, in order to start the filter, it is necessary to make a prediction of value of state vector x_1, given no observation. Hence $x_1|y_0$ should have a gaussian distribution with mean $\hat{x}_1^-$ given by unconditional expectation $E(x_1)$ of x_1, and covariance P_1^- equal to unconditional covariance $Cov(x_1)$ of x_1. This fact - in the case that the process x is covariance stationary - dictates how mean $\hat{x}_1^-$ and covariance P_1^- of the predicted $x_1|y_0$ should be set up.

Taking unconditional expectation of the left hand side and the right hand side of (1) produces $E(x_n) = AE(x_{n-1})$ equation. Let x be covariance stationary. Then, $E(x_n) \equiv E(x_{n-1})$, which implies that $E(x_n)$ should satisfy a requirement: $(I - A)E(x_n) = 0$. The requirement is satisfied solely by $E(x_n) = 0$ (since $I - A$ is nonsingular).

In similar manner it can be shown [fig5] that the unconditional covariance $Cov(x_n)$ should satisfy $Cov(x_n) = ACov(x_n)A + Q$ equation. Its solution is $vec(Cov(x_n)) = (I - (A \otimes A))^{-1} vec(Q)$. ($vec(M)$ transforms a matrix M into a vector, by stacking columns of M. $\otimes$ denotes Kronecker product.)

Thus, the KF has to be started with $\hat{x}_1^- = 0$ and P_1^- given by $vec(P_1^-) = (I_{m^2} - (A \otimes A))^{-1} vec(Q)$. As the first observation y_1 arrives, the prediction is corrected. The corrected mean and covariance $\hat{x}_1$, P_1 are then predicted. And so on.

Note that for covariance stationary process x thus initialization of KF does not depend on $\hat{x}_0$, P_0. If the process is not covariance stationary, the parameters should be chosen by an analyst, and projected ahead. In this case, likelihood function evaluated via KF is no more exact.

3. State-space representation of ARMA process

ARMA process can be cast into the form of model which is described by (1), (2). It is called state-space representation of the process. A process can have several state-space representations.

For example, an MA(1) process $y_t - \mu = \epsilon_t + \theta\epsilon_{t-1}$ can be in one way represented at the state-space form (1), (2), by means of the following mappings: $x_n = [\epsilon_t, \epsilon_{t-1}]$, $y_n = y_t - \mu$,

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad Q = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix},$$

$C = [1, \theta]$, $v_n = [\epsilon_t, 0]$, $w_n = 0$, $R = 0$; there σ^2 denotes the variance of ϵ_t. Note that this is a state-space representation of $y_t - \mu$ rather than of y_t process. In

order to turn this into a state-space representation of y_t process, another term – a deterministic input term Bz_n – has to be added to the observation equation (2):

$$y_n = Bz_n + Cx_n + w_n. \tag{2'}$$

There B is a matrix and z_n is a vector of predetermined exogenous variables. In the case of the MA(1) process $B = \mu$ and $z_n = 1$, for all n.

A general ARMA(p, q) process can be written as: $Y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_m(y_{t-m} - \mu) + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_{m-1}\epsilon_{t-m+1}$, where m (i.e., dimension of state vector) is $m = \max(p, q+1)$ and ϕ_i is set to be zero for $i > p$, $\theta_i = 0$ for $i > q$. Then ARMA process can be written in a convenient state-space form (1), (2'): $x_n = [y_t, \ldots, y_{t-m}]$, $y_n = y_t$, $B = \mu$, $z_n = 1$,

$$A = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_{r-1} & \phi_r \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}, \qquad Q = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix},$$

$C = [1, \theta_1, \ldots, \theta_{m-1}]$, $v_n = [\epsilon_t, 0, \ldots, 0]$, $w_n = 0$, $R = 0$.

Using the model (1), (2') instead of (1), (2) requires to modify Kalman Filter equations (7), (9) accordingly; [fig6].

Two things are worth observation: 1) at the state representation of ARMA the transition matrices A, C depend on parameters of the process (ϕ, θ). 2) Regardless of state representation of a process, the forecasts of $y_n|y_1^{n-1}$ generated by KF is identical; [q2: why?]. Of course, for different state representation KF forecasts of $x_n|y_1^{n-1}$ are in general different.

4. Exact likelihood calculation via KF

Exact likelihood function for a time series of length T generated by a gaussian ARMA process is a multivariate normal distribution with covariance matrix implied by the AR and MA dependence.

It is straightforward to see that joint probability density function (i.e., likelihood function) $p(y_1^T)$ can be expressed as a product of conditionals (so-called prediction-error decomposition):

$$p(y_1^T) = p(y_1) \prod_{n=2}^{T} p(y_n|y_1^{n-1}). \tag{10}$$

Moreover, Bayes formula implies that $p(y_1) \equiv p(y_1|y_0)$.

Given a time series y_1^T generated by an ARMA(p,q) process the exact likelihood function given by (10) can be calculated for chosen values of parameters ϕ, θ, σ^2 of the process by means of KF, since KF can be (mis-:)used for sequential calculation of the conditional probability density functions $p(y_n|y_1^{n-1})$. To see how, recall that the prediction step (Eq. 7) results in conditional distribution $x_n|y_1^{n-1} \sim n(\hat{x}_n^-, P_n^-)$. This, together with the observation equation (Eq. 2) implies that

$$y_n|y_1^{n-1} \sim n(C\hat{x}_n^-, CP_n^-C' + R). \tag{11}$$

Note, that for the model (1), (2') the conditional density $y_n|y_1^{n-1}$ will take a slightly different form; [fig7].

The calculated likelihood can be then used as an input to a likelihood maximization routine (e.g. EM algorithm).

5. Fill-in-gaps, questions, exercises

Fill-in gaps: [fig1]; [fig2] and [fig3] (Hint: first show that $P(A|B,C) = P(A|C) \cdot P(B|A,C)/P(B|C)$ and make use of it); [fig4] (Hint: first find pdf of $y_n|x_n, y_1^{n-1}$. Then use the 'vice-versa' part of the Thm to get pdf of $x_n, y_n|y_1^{n-1}$. Finally, use the 'direct part' of the Thm to get the desired pdf of $x_n|y_n, y_1^{n-1}$.); [fig5] (Hint: Use *iii*) of the model specification); [fig6: Write KF equations for model specified by (1),(2')]. [fig7: Write the conditional density $y_n|y_1^{n-1}$ for the model given by (1), (2')].

Questions: [q1], [q2].

Exercises:

1) Assume an MA(1) process. Propose a state-space representation of the process. KF can be used to find exact finite sample one-step ahead minimum mean squared error forecasts of y_t given y_1^{n-1}. Do you know how? (Hint: cf. Sect. 4, Eq. (11) and recall that conditional mean is the minMSE forecast). Initialize the filter, and find out formula for the forecast and its MSE.

2) KF can be used to calculate exact s-step-ahead-forecasts. Can you see how?

3) Consider a state-space model defined by state equation $x_n = \rho x_{n-1} + w_n$ and observation equation $y_n = z_n x_n + v_n$. There both x_n, y_n are scalars; z is a predetermined explanatory variable. Observation eq. defines a regression model where the regression coefficient x evolves in time according to an autoregression process specified by the state equation. Assume gaussian errors, pick up some values of ρ, Q, R, z and generate y. Use KF to find a minMSE estimate of the time series of states. Plot both true and KF estimated states.

In practice, ρ, Q, R are rarely known; analyst should select a guess values of them, say ρ_g, Q_g, R_g. Make a your choice and use KF to find an estimate of time series of states. Use computer to investigate how does performance of KF depends on guess values.

Use KF to get a series of predicted y's and calculate value of the exact likelihood function for both the true values of ρ, Q, R and guess values.

4) Consider a scalar gaussian ARMA(1,1) process $y_t - \mu = \phi(y_{t-1} - \mu) + \epsilon_t + \theta\epsilon_{t-1}$. Write the process into the state space form which was presented at Sect. 3. What is P_1^- here?

Choose some values of parameters $\mu, \phi, \theta, \sigma^2$ and generate $T = 30$ observations from the ARMA(1,1). Next, using the same values of parameters, initialize and run the KF, in order to evaluate likelihood function of the generated time series.

Do not forget to use the KF equations appropriate for the model of (1), (2') form.

6. Notes

Kalman filter was originally proposed as a method for sequential recovering of hidden state variables of a system from noisy measurements, by making use of knowledge of the system dynamics.

For filtration of systems with non-linear dynamics and/or non-gaussian errors an appropriate Bayes filter or other approaches have been and are being developed. Yet, KF is still used for navigation, missile tracking etc.

In econometrics, its use is twofold. Since economy is a system, KF is suitable for its evolution tracking. The other use is for calculation of likelihood function of

ARMA process. For the purpose of likelihood calculations it is even more suitable at the area of structural modeling of time series, since there the models are directly formulated at the state-space form.

KF was presented here as a special, gaussian, Bayes filter. Alternatively, KF can be motivated through recursive least squares.

7. Sources, further readings

Sources, and also some further readings:

Hamilton J. D., *Time series analysis*, PUP, Princeton, 1994.
Meinhold R. J. and Singpurwalla N. D., Understanding the Kalman filter, *Am. Stat.*, 37/2, pp. 123-127, 1983.
Meinhold R. J. and Singpurwalla N. D., Robustification of Kalman filter models, *JASA*, 84/406, pp. 479-486, 1989.
Rao C. R., A note on Kalman filter, *Proc. Natl. Acad. Sci.*, 98/19, pp. 10557-10559, 2001.
Schwardt L., The Kalman filter, Lecture notes, DSP813 Lecture 10, University of Stellenbosch, 2004.
Welling M., The Kalman filter, Lecture notes, CalTech.

This is a part of lecture notes (under preparation) based on an undergraduate introductory course on econometric time series analysis taught for students majoring in statistics. Comments welcome (m grendar, im&cs, banska bystrica, slovakia, umergren@savba.sk).

Jan 23-30, 2005

KLASIFIKÁCIA ERYTROPOETÍNOVÝCH OBRAZCOV METÓDOU OPORNÝCH BODOV (SUPPORT VECTOR MACHINES)

Klára Hornišová
Ústav merania SAV, Bratislava
umerhorn@savba.sk

1. Úvod. Odhaľovanie dopingu syntetickými náhradami erytropoetínu (EPO)

Tvorbu červených krviniek, a tým i prenos kyslíka krvou podporuje hormón obličiek erytropoetín. Jeho zvýšená hladina môže vo vytrvalostných športoch zvýšiť výkonnosť až o 10%. Na tento účel sa ako doping zneužívajú dve syntetické náhrady erytropoetínu *rEPO* a *NESP*, vyvinuté pôvodne ako liečivá. Odhaliť a odlíšiť od seba navzájom i od prirodzeného erytropoetínu sa dajú reakciou vzorky so špeciálnym substrátom s meniacimi sa hodnotami *pH* a následným zobrazením chemoluminiscenciou. Prirodzený *EPO*, *rEPO* a *NESP* sa líšia intervalmi hodnôt *pH*, v ktorých najintenzívnejšie reagujú so substrátom, čo sa prejavuje výskytom oválnych svetlých machuliek - porade ide o pásmo prostredných, vysokých a nízkych hodnôt *pH*.

Pri automatizovaní spracovania obrazcov vzoriek vzniká úloha odlíšiť machuľky zodpovedajúce skutočným prejavom reakcie so substrátom od artefaktov - falošných machuliek, ktoré môžu byť napríklad dôsledkom optických šumov pri zobrazovaní. Jednou z fáz algoritmu je segmentácia - transformácia pôvodného rastrového obrazu s 256−mi úrovňami sivej farby na $0-1$-ový obraz pozadia a machuliek. Na účely triedenia na pravé machuľky a artefakty sa $1-$ové ostrovčeky opísali 8−mi veličinami (niektoré zo zaužívaných charakteristík 2−rozmerných útvarov, implementované napr. v Matlabe): 1.) tvar $= \frac{\text{výška}}{\text{šírka}}$, 2.) výstrednosť v smere osi x $= |stred_x -$ ťažisko$_x|$, 3.) zložitosť hranice $= 2\frac{\text{šírka+výška}}{\text{obvod}}$, 4.) obdĺžnikovosť $= \frac{\text{obsah}}{\text{šírka}*\text{výška}}$, 5.) veľkosť $= \frac{\text{obsah}}{\text{obsah stĺpca}}$, 6.) výstrednosť $= \frac{c}{a}$, 7.) orientácia $= \theta$, 8.) konvexnosť $= \frac{\text{obsah}}{\text{obsah konvex. obalu}}$, kde šírka, výška = rozmery najmenšieho opísaného obdĺžnika so stranami rovnobežnými so súradnicovými osami; a, c, θ = dlhšia polos, polovica ohniskovej vzdialenosti a uhol v stupňoch x-ovej osi s dlhšou polosou elipsy, ktorá má rovnaké ťažisko a spektrálny rozklad ako množina stredov pixlov, z ktorých sa skladá machuľka.

Na klasifikáciu týchto 8−rozmerných vektorov sa na našom pracovisku použili metódy: Fisherov lineárny klasifikátor, metóda k najbližších susedov, metóda založená na vážených poradiach, neurónové siete, logistická regresia, metóda oporných bodov (support vector machine).

Na korektné vzájomné porovnanie ich výsledkov sa pre každý variant segmentácie - iný súbor škvŕn expertmi zatriedených do dvoch až štyroch tried: b) ozajstná machuľka, a) artefakt, ab) ozajstná machuľka zliata s artefaktom, bb) zliate 2 ozajstné machuľky
dohodol jednotný postup:

A) - každý súbor sa rozdelil na oficiálnu trénovaciu ($\doteq 60\%$) a testovaciu množinu. Pri hľadaní optimálnych parametrov každej z metód sa použila len trénovacia časť súboru. Chyba zovšeobecnenia sa potom odhadovala podielom zle zatriedených testovacích vzoriek.

B) - každý súbor sa 100-krát náhodne rozdelil na trénovaciu ($\doteq 60\%, 75\%, 90\%$) a testovaciu množinu. Chyby zovšobecnenia sa odhadovala priemerným podielom zle zatriedených testovacích vzoriek.

V ďalšom podrobnejšie vyložíme metódu oporných bodov založenú na Vapnikovej štatistickej teórii učenia sa (podľa [10], [7] a [1]).

2. Vapnikova štatistická teória učenia sa

V teórii sa hľadajú nutné a postačujúce podmienky zovšeobecnenia výsledkov z empirických metód použitých na daný náhodný výber na iné údaje z toho istého rozdelenia. Ako motiváciu najprv pripomeňme Glivenkovu vetu o rovnomernej konvergencii empirických distribučných funkcií k skutočnej (pre jednoduchosť iba pre jednorozmerné náhodné premenné):

Glivenkova veta. Nech P je pravdepodobnostná miera na $(\mathbb{R}, \mathcal{B})$, nech P_n je príslušná empirická miera založená na náhodnom výbere z P s rozsahom n. Potom

$$\sup_{t\in\mathbb{R}} |P_n((-\infty,t)) - P((-\infty,t))| \overset{\text{skoro isto}}{\longrightarrow} 0.$$

Ekvivalentne:

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \overset{\text{skoro isto}}{\longrightarrow} 0, \tag{1}$$

kde $\mathcal{A} = \{(-\infty, t); t \in \mathbb{R}\}$.

Vapnik a Červonenkis si položili otázku: Aké ďalšie systémy množín $\mathcal{A}$ spĺňajú (1)? Nutná a postačujúca podmienka sa dá sformulovať pomocou nasledujúceho pojmu:

Definícia. Systém množín $\mathcal{A}$ rozbíja množinu M na n častí, ak $\text{card}\{A \cap M; A \in \mathcal{A}\} = n$.

Definícia. Systém množín $\mathcal{A}$ je VČ trieda, ak $\exists m \in \mathbb{N}$ také, že pre $\forall M; \text{card}M = m$ platí, že $\mathcal{A}$ rozbíja M na menej ako 2^m častí.

Vapnikova - Červonenkisova (VČ) (zovšeobecnená Glivenkova) veta. Nech $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R})$. Potom (1) platí práve vtedy, ak $\mathcal{A}$ je VČ trieda.

(Výsledok možno rozšíriť na náhodné veličiny s hodnotami v poľských metrických priestoroch.)

Žiadalo by sa mať čo najväčšiu VČ triedu, vie sa však, že maximálna VČ trieda neexistuje.

Príklady VČ tried v $\mathbb{R}^n$:

- systém vš. gúľ
- systémy množín kladnosti polynómov s ohraničeným stupňom
- $\mathcal{B}$ nie je VČ trieda

VČ veta je limitnou vetou pre empirické náhodné veličiny $P_n(A)$, indexované množinami $A \in \mathcal{A}$. Keďže pravdepodobnosť na poľskom priestore $\mathcal{X}$ možno ekvivalentne (podľa Rieszovej vety o reprezentácii) definovať ako lineárny funkcionál na $C(\mathcal{X})$ (Radonovu mieru), možno uvažovať rôzne systémy $\mathcal{L}$ všeobecnejších náhodných veličín indexovaných funkciami $f \in \mathcal{F}$, kde $\mathcal{F}$ je nejaká množina funkcií na $\mathcal{X}$. Potom možno riešiť úlohy podobné tej zo zovšeobecnenej Glivenkovej vety: Aké sú nutné a postačujúce podmienky, t.j. vlastnosti systému $\mathcal{F}$, na rovnomernú konvergenciu istých postupností náhodných veličín z $L_n(f)$ z $\mathcal{L}$; $f \in \mathcal{F}$? V ďalšom sa obmedzíme na systémy náhodných veličín prirodzene súvisiacich s úlohou binárnej klasifikácie.

Úlohu riadenej binárnej klasifikácie možno popísať takto: Dané sú trénovacie vzorky

$$(x_1, y_1), \ldots, (x_\ell, y_\ell), \qquad x_i \in \mathbb{R}^N, y_i \in \{-1, +1\},$$

ktoré tvoria náhodný výber z neznámeho pravdepodobnostného rozdelenia $P(x, y)$. Treba určiť funkciu f^* z danej množiny $\mathcal{F}$ funkcií $f : \mathbb{R}^N \to \{-1, 1\}$, ktorá minimalizuje riziko

$$R(f) = \int L(f(x), y) dP(x, y) \; ; \; f \in \mathcal{F} \tag{2}$$

kde $L(.,.)$ je stratová funkcia, napr. $L(w, y) = \max\{1 - wy, 0\}$. Keďže $P(x, y)$ je neznáme, je prirodzené hľadať funkciu f^ℓ, ktorá minimalizuje *empirické riziko*

$$R_{emp}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(f(x_i), y_i) \quad ; \quad f \in \mathcal{F}. \tag{3}$$

Ak je však ℓ malé, takýto prístup vo všeobecnosti nedáva informáciu, či bude malé aj riziko (2).

Ak vezmeme do úvahy aj vlastnosti množiny $\mathcal{F}$, možno odvodiť nutné a postačujúce podmienky na konzistenciu (3) (t.j. na konvergenciu v pravdepodobnosti

$$R(f^\ell) \to R(f^*) \tag{4}$$

a

$$R_{emp}(f^\ell) \to R(f^*) \tag{5}$$

ak $\ell \to \infty$).

Ak existujú A, B také, že pre $\forall f \in \mathcal{F}$

$$A \leq \int L(f(x), y) dP(x, y) \leq B, \tag{6}$$

tak (4) a (5) sú ekvivalentné s podmienkou

$$\forall \varepsilon; \lim_{\ell \to \infty} P \left\{ \sup_{f \in \mathcal{F}} (R(f) - R_{emp}(f)) > \varepsilon \right\} = 0. \tag{7}$$

Na formuláciu ďalších ekvivalencií potrebujeme niekoľko charakteristík množiny funkcií $L(f(x), y); f \in \mathcal{F}$:

1. Entropia:
   entropia množiny indikátorových funkcií $L(f(x), y); f \in \mathcal{F}$: Pre náh. výber

$$(x, y) := (x_1, y_1), \ldots, (x_\ell, y_\ell)$$

nech $N^{\mathcal{F}}(x, y)$ označuje počet rozdelení tohoto výberu na dve časti pomocou funkcií $L(f(x), y); f \in \mathcal{F}$. $N^{\mathcal{F}}(x, y)$ sa takisto rovná počtu vrcholov $\ell-$rozmernej kocky $< 0; 1 >^\ell$, ktoré majú tvar

$$(L(f(x_1), y_1), \ldots, L(f(x_\ell), y_\ell)); f \in \mathcal{F}.$$

Ďalej definujme náhodnú entropiu

$$H^{\mathcal{F}}(x, y) := \ln N^{\mathcal{F}}(x, y)$$

a entropiu pre výbery s rozsahom ℓ

$$H^{\mathcal{F}}(\ell) := E H^{\mathcal{F}}(x, y),$$

(str. hodnota vzhľ. na $P(x_1, y_1). \ldots . P(x_\ell, y_\ell)$.

Definícia. Zocelená (annealed) VČ-entropia je

$$H^{\mathcal{F}}_{ann}(\ell) = \ln E N^{\mathcal{F}}(x, y)$$

a rastová funkcia je

$$G^{\mathcal{F}}(\ell) = \ln \sup_{(x,y)} N^{\mathcal{F}}(x, y).$$

Veta.

$$H^{\mathcal{F}}(\ell) \leq H^{\mathcal{F}}_{ann}(\ell) \leq G^{\mathcal{F}}(\ell).$$

Veta. (4) a (5) platia pre $\forall$ p. mieru $P(x, y) \iff$

$$\lim_{\ell \to \infty} \frac{G^{\mathcal{F}}(\ell)}{\ell} = 0.$$

Doteraz sme uvádzali iba asymptotické vlastnosti minimalizácie empirického rizika (MER). V ďalšom budeme potrebovať novú charakterizáciu kapacity - VČ dimenziu. (Kapacitou množiny $\{L(f(x), y); f \in \mathcal{F}\}$ sa rozumie jej schopnosť bezchybne sa naučiť akúkoľvek trénovaciu množinu.)

Veta. Buď

$$G^{\mathcal{F}}(\ell) = \ell \ln 2 \tag{8}$$

alebo

$$G^{\mathcal{F}}(\ell) < h(\ln \frac{\ell}{h} + 1), \tag{9}$$

kde $h \in \mathbb{Z}$ spĺňa

$$G^{\mathcal{F}}(h) = h \ln 2 \quad \text{a} \quad G^{\mathcal{F}}(h+1) \neq (h+1) \ln 2.$$

Definícia. VČ dimenzia množiny indikátorových funkcií $\{L(f(x), y); f \in \mathcal{F}\}$ je nekonečná, ak platí (8).

VČ dimenzia množiny indikátorových funkcií $\{L(f(x), y); f \in \mathcal{F}\}$ je konečná a rovná sa h, ak platí (9).

Ekvivalentne, VČ dimenzia množiny indikátorových funkcií $\{L(f(x), y); f \in \mathcal{F}\}$ je maximálny počet h vektorov $(x_1, y_1), \ldots, (x_h, y_h)$, ktoré možno oddeliť funkciami z tejto množiny všetkými 2^h spôsobmi.

Veta. MER na množine indikátorových funkcií $\{L(f(x),y); f \in \mathcal{F}\}$ je konzistentná pre $\forall$ $P(x,y)$ práve vtedy, ak je VČ dimenzia tejto množiny konečná.

Doterajšie asympt. výsledky o rovnomernej konvergencii nám teraz umožnia odvodiť neasymptotické ohraničenia rizika $R(f)$ založené na empirickom riziku $R_{emp}(f)$ a počte trénovacích vzoriek ℓ:

Veta. Ak platí (6) a $A = 0$, tak s pravdepodobnosťou aspoň $1 - \eta$ pre $\forall f \in \mathcal{F}$ platí

$$R(f) \leq R_{emp}(f) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(f)}{B\varepsilon}}\right), \tag{10}$$

kde

$$\varepsilon = 4\frac{h\left(\ln\frac{2\ell}{h} + 1\right) - \ln\eta}{\ell}.$$

(druhý člen na pravej strane (10) sa volá VČ konfidencia)

Princíp MER je vhodný pre veľké ℓ. (Rozsah výberu ℓ sa považuje za veľký, ak je pomer ℓ/h veľký, povedzme $\ell/h \geq 20$.) Vtedy je riziko R blízke empirickému riziku, ako to vyplýva z (10). Na základe týchto ohraničení Vapnik navrhol princíp minimalizácie rizikového funkcionálu, ktorý berie do úvahy rozsah ℓ trénovacej množiny, ak je pomer ℓ/h malý. To vedie k metódam vhodným pre daný rozsah výberu. Keďže pre malé ℓ/h ešte malé $R_{emp}(f^\ell)$ nezaručuje aj malé R, nový princíp by mal simultánne minimalizovať oba členy v (10). Pritom jeden z nich závisí od empirického rizika, druhý od VČ dimenzie množiny funkcií. Ďalej popíšeme tento nový princíp - minimalizáciu štrukturálneho rizika (MŠR):

Nech je na $S := \{L(f(x),y); f \in \mathcal{F}\}$ daná štruktúra podmnožín

$$S_1 \subset \cdots \subset S_n \ldots$$

taká, že

1) $S^* := \cup_k S_k$ je všade hustá v S.

2) VČ dimenzia h_k každej S_k je konečná.

3) Pre $\forall k\ \exists B_k \in \mathbb{R}$ také, že pre $\forall f \in \mathcal{F}_k$ platí $0 \leq L(f(x),y) \leq B_k$.

Podľa princípu MŠR by sa mala vybrať podmnožina S_n, kde $n = n(\ell)$ (presnejšie uvedieme ďalej) a z S_n vybrať funkcia f pre ktorú je zaručené riziko na pravej strane (10) minimálne, t.j. tým sa vyvážia kvalita aproximácie a komplexita aproximujúcej funkcie.

Veta. Pre $\forall$ $P(x,y)$ konverguje MŠR k minimu s pravdepodobnosťou 1.

V horeuvedených výsledkoch štatistickej teórie učenia sa podmienky zovšeobecňujúcej schopnosti a konzistencie MER formulovali pomocou vlastností priestoru hypotéz (t.j. rozhodovacích funkcií). Novšie sa dôraz presúva na vlastnosti samotného algoritmu učenia sa - takými podmienkami sú rôzne druhy jeho stability - ak z trénovacej množiny vynecháme jednu vzorku, výsledná hypotéza sa veľmi nezmení ([9]). Takéto nové podmienky môžu viesť k všeobecnejším algoritmom.

3. Učiace sa stroje využívajúce oporné body

Doteraz sme neuvažovali o konkrétnych stratových funkciách $L(.,.)$. Pri binárnej klasifikácii by bola najprirodzenejšia 0-1-ová stratová funkcia, ktorá dáva bayesovské optimálne rozhodnutie, tá však vedie k neúnosným výpočtom. Namiesto nej je nižšieopísaný algoritmus oporných bodov založený na štruktúre množiny nadrovín $\left\{x \in \mathbb{R}^N; < w, x > +b = 0\right\}$, kde $(w,b) \in \mathbb{R}^N \times \mathbb{R}$ spĺňa normovaciu podmienku

$$\min_{i=1,\ldots,\ell} | < w, x_i > +b| = 1,$$

ktorým zodpovedajú jednoduché rozhodovacie funkcie

$$f_{w,b} : B_{x_1,\ldots,x_\ell} \to \{\pm 1\}$$
$$f_{w,b} = sgn(< w, x > +b), \tag{11}$$

kde $B_{x_1,\ldots,x_\ell}$ je najmenšia guľa v $\mathbb{R}^N$ obsahujúca body $x_1, \ldots, x_\ell$. Vie sa, že pre VČ dimenziu h množiny $\{f_{w,b}; \|w\| \leq A\}$ platí

$$h \leq R^2 A^2,$$

kde R je polomer $B_{x_1,\ldots,x_\ell}$, takže na množine nadrovín možno zaviesť štruktúru s vlastnosťami 1),...,3).

MŠR možno potom približne uskutočniť riešením úlohy kvadratického programovania

$$\min_{w,\xi,b} < w,w > + C\sum_{i=1}^{\ell}\xi_i \tag{12}$$
$$y_i(< w,x_i > +b) \geq 1-\xi_i; i=1,\ldots,\ell$$
$$\xi_i \geq 0; i=1,\ldots,\ell.$$

Člen $< w,w >$ zodpovedá VČ konfidencii a člen $\sum_{i=1}^{\ell}\xi_i$ je horným ohraničením počtu nesprávnych zatriedení v trénovacej množine, takže od neho závisí člen zodpovedajúci empirickému riziku na pravej strane (10). Regularizačnú konštantu $C>0$, ktorá vyvažuje empirickú chybu a komplexitu, treba vhodne zvoliť. Oddeliteľný prípad ($\equiv$ s pevným rozhraním) (t.j. bez trénovacích chýb) nastáva, ak: $C=0$ a $\xi_i=0; i=1,\ldots,\ell$. Opakom je neoddeliteľný prípad ($\equiv$ s voľným rozhraním).

Lineárne rozhodovacie plochy nemusia byť dosť všestranné, metódu však možno zovšeobecniť: Vstupné vektory x sa vhodným zobrazením ϕ transformujú do nejakého mnohorozmerného (aj nekonečnorozmerného) priestoru $\mathcal{X}$ a lineárne oddelenie sa vykoná v $\mathcal{X}$. Zobrazenie $\phi(.)$ nemusí byť explicitne známe, keďže na riešenie úlohy (12), v ktorej sa body x_i nahradia bodmi $\phi(x_i)$, stačí poznať skalárne súčiny tvaru $<\phi(x),\phi(x_i)>$. Ak existuje jadro - funkcia $K(.,.)$ taká, že

$$K(x,x_i) =< \phi(x),\phi(x_i) >,$$

výpočty sa dajú veľmi zjednodušiť. Najbežnejsie sú jadrá:
a) polynomické

$$K(x_i,x) = (< x_i,x > +\theta)^d$$

b) štandardné gaussovské

$$K(x_i,x) = \exp(-\|x-x_i\|^2/\sigma^2)$$

c) všeobecné gaussovské

$$K(x,x') = \exp(-(x-x')^\top A(x-x')),$$

kde A je symetrická kladne definitná matica,
d) neurónovosieťové

$$K(x_i,x) = \tanh(\kappa. < x_i,x > -\theta).$$

Ak je jadro $K(.,.)$ symetrická, nezáporne definitná funkcia, t.j. ak pre každé $n\in\mathbb{N}$ a $\forall$ $a_1,\ldots,a_n\in\mathbb{R}$ a $x_1,\ldots,x_n\in X$

$$\sum_{i,j=1}^{n} a_i a_j K(x_i,x_j) \geq 0$$

a ak je navyše $K(.,.)$ spojitá a patrí do $L_2(X\times X)$, tak Hilbertov - Schmidtov operátor

$$A:\mathcal{L}_2(X)\to\mathcal{L}_2(X)$$
$$Af(x) := \int_X K(x,x')f(x')dx'$$

je pozitívny, kompaktný a samoadjungovaný, a teda A má ortonormálnu postupnosť spojitých vlastných funkcií $\Phi_1,\Phi_2,\cdots\in\mathcal{L}_2(X)$ a vlastné hodnoty $\lambda_1\geq\lambda_2\geq\cdots\geq 0$ a jadro možno vyjadriť v tvare

$$K(x,x') = \sum_{\nu=1}^{\infty}\lambda_\nu\Phi_\nu(x)\Phi_\nu(x') \tag{13}$$

(pozri napr. [6]). Z (13) vidno, že $\phi(x)$ možno definovať takto:

$$\phi(x) = (\sqrt{\lambda_1}\Phi_1(x),\sqrt{\lambda_2}\Phi_2(x),\ldots)\in\ell_2.$$

Podobné tvrdenia možno dostať, ak podmienku kladnej definitnosti nahradíme slabšími predpokladmi podmienene kladnej či nezápornej semidefinitnosti.

Namiesto úlohy (12) teda možno riešiť jej $\phi(.)-$obmenu

$$\begin{aligned} \min_{w,\xi,b} \frac{1}{2} < w, w > + C \sum_{i=1}^{\ell} \xi_i \\ y_i(< w, \phi(x_i) > + b) \geq 1 - \xi_i; i = 1, \ldots, \ell \\ \xi_i \geq 0; i = 1, \ldots, \ell. \end{aligned} \tag{14}$$

(14) je konvexná úloha kvadratického programovania. Z teórie matematického programovania je známe (napr. [8]), že namiesto (14) možno riešiť ľahšiu duálnu úlohu

$$\begin{aligned} \max_{\alpha} \quad & \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top Q \alpha \\ \text{pri podmienkach} \quad & 0 \leq \alpha_i \leq C; 1 = 1, \ldots, \ell, \\ & y^\top \alpha = 0, \end{aligned} \tag{15}$$

kde $Q_{ij} = y_i y_j K(x_i, x_j)$

Pre optimálne riešenie primárnej a duálnej úlohy platí

$$w = \sum_{i=1}^{\ell} \alpha_i y_i \phi(x_i)$$

$$\frac{1}{2} w^\top w + C \sum_{i=1}^{\ell} \xi_i = \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top Q \alpha$$

Rozhodovacia funkcia potom je

$$\begin{aligned} f(x) = sgn(< w, \phi(x) > + b) = \\ = sgn(\sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x) + b). \end{aligned}$$

Optimálne riešenie úlohy (15) navyše spĺňa Kuhnove-Tuckerove (KT) podmienky

$$\begin{aligned} \alpha_i = 0 \Rightarrow y_i sgn(< w, \phi(x_i) > + b) \geq 1 \text{ and } \xi_i = 0 \\ 0 < \alpha_i < C \Rightarrow y_i sgn(< w, \phi(x_i) > + b) = 1 \text{ and } \xi_i = 0 \\ \alpha_i = C \Rightarrow y_i sgn(< w, \phi(x_i) > + b) \leq 1 \text{ and } \xi_i \geq 0. \end{aligned} \tag{16}$$

Zo (16) vidno dôležitú vlastnosť metódy SVM - riedkosť optimálneho vektora α. Trénovacie vzorky x_i, pre ktoré $\alpha_i \neq 0$, sa volajú *oporné body*. Sú blízko rozhodovacej hranice medzi dvoma triedami. Všetky tréningové vzorky x_i s $\alpha_i = 0$ sa mohli na začiatku algoritmu vynechať bez straty informácie - sú hlboko vnútri oblasti jednej z dvoch tried.

Treba poznamenať, že optimálne riešenie α nie je jediné.

Duálna úloha (15) sa spravidla rieši iteračnou dekompozičnóu metódou. Vyjdúc z počiatočného riešenia, v každom kroku sa cieľová funkcia maximalizuje len na pracovnej podmnožine súradníc vektora α, kým ostatné sa nemenia. Jej prvky sa vyberajú tak, aby čo najviac porušovali KT podmienky. Pri sekvenčnej minimálnej optimalizácii (SMO) sa používajú $2-$prvkové pracovné podmnožiny, vďaka čomu sa optimalizačné úlohy v jednotlivých iteračných krokoch dajú riešiť analyticky.

Numericky stabilná hodnota prahu b je

$$b = \frac{1}{|I|} \sum_{i \in I} \left(y_i - \sum_{j=1}^{\ell} y_j \alpha_j K(x_i, x_j) \right),$$

kde $I = \{i; 0 < \alpha_i < C\}$ pre (14) a $|I|$ je počet prvkov I.

4. Obmeny metódy oporných bodov pri triedení na k tried, $k > 2$

Na riešenie úlohy triedenia na viacero skupín sa navrhlo niekoľko prístupov. Najstaršie z nich zostrojujú viactriedny klasifikátor ako kombináciu niekoľkých binárnych klasifikátorov. Nech k je počet tried. V metóde *jedna proti všetkým* sa zostrojí k modelov SVM. V i-tom modeli

majú všetky vzorky z i-tej triedy označenie $+1$, kým všetky ostatné vzorky majú označenie -1. Riešením i-tej úlohy je rozhodovacia funkcia

$$(w^i)^\top \phi(x) + b^i.$$

Vzorka x sa zaradí do triedy s najväčšou hodnotou rozhodovacej funkcie

$$\text{trieda pre } x \equiv \arg \max_{i=1,\dots,k} ((w^i)^\top \phi(x) + b^i).$$

V metóde *jedna proti jednej* sa zostrojí $k(k-1)/2$ klasifikátorov, ktoré zodpovedajú všetkým možným dvojiciam tried. V $(i,j)-$tom modeli sa využívajú iba trénovacie údaje z i-tej (s označením $+1$) a j-tej triedy (s označením -1). Pri testovaní sa uplatňuje hlasovacia stratégia: ak $sgn((w^{ij})^\top \phi(x) + b^{ij}) = +1$, tak sa počet hlasov za i-tu triedu zväčší o jeden. Vzorka x sa zaradí do triedy s najväčším počtom hlasov.

Niektoré metódy uvažujú všetky triedy naraz. Podobne, ako v metóde jedna proti jednej, sa zostrojí k rozhodovacích funkcií, ktoré sú však riešením jedinej úlohy, napríklad

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \sum_{m=1}^{k} w_m^\top w_m + C \sum_{i=1}^{\ell} \sum_{m \neq y_i} \xi_i^m \\ & w_{y_i}^\top \phi(x_i) + b_{y_i} \geq w_m^\top \phi(x_i) + b_m + 2 - \xi_i^m \\ & \xi_i^m \geq 0, \qquad i = 1, \dots, \ell, \qquad m \in \{1, \dots, k\} \backslash y_i. \end{aligned}$$

Rozhodovacia funkcia je potom

$$\arg \max_{m=1,\dots,k} (w_m^\top \phi(x) + b_m).$$

V inej podobnej metóde (pozri [2]) sa rieši úloha

$$\begin{aligned} \min_{w_m,\xi_i} \quad & \frac{1}{2} \sum_{m=1}^{k} w_m^\top w_m + C \sum_{i=1}^{\ell} \xi_i \\ & w_{y_i}^\top \phi(x_i) - w_m^\top \phi(x_i) \leq e_i^m - \xi_i, \qquad i = 1, \dots, \ell, \end{aligned}$$

kde $e_i^m := 1 - \delta_{y_i}^m$. Rozhodovacia funkcia potom je

$$\arg \max_{m=1,\dots,k} w_m^\top \phi(x).$$

Experimenty v [5] na malých množinách údajov naznačujú, že metóda $1:1$ je najrýchlejšia, a pritom dosť spoľahlivá.

5. Voľba modelu

Dosiaľ sme opísali trénovaciu fázu algoritmu oporných bodov, ktorá prebieha s vopred danou hodnotou parametra C a daným tvarom jadra s danými parametrami. Voľba týchto parametrov modelu však môže mať veľký vplyv na úspešnosť triedenia. Používa sa niekoľko metód voľby modelu, ktoré poskytujú predpovede chyby zovšeobecnenia.

Najpoužívanejšia, a pritom najzdĺhavejšia, je m-násobná crossvalidácia (CV). Pri nej sa pre každý bod z nejakej siete v priestore parametrov modelu trénovacia množina náhodne rozdelí na m podmnožín s približne rovnakým počtom prvkov. Potom sa postupne v i-tom kroku vyberie i-ta podmnožina ako testovacia a so zvyškom sa pracuje ako s trénovacou množinou. Takto sa klasifikátor m-krát natrénuje, a zakaždým sa vypočíta chyba klasifikácie pre i-tu podmnožinu. Vie sa, že priemer týchto m chýb je dosť dobrým odhadom chyby zovšeobecnenia.

Pre výpočtovú zložitosť je vyššie opísané prehľadávanie siete mrežových bodov v parametrickom priestore pri crossvalidácii únosné iba pri optimalizácii hodnôt veľmi malého poqv ctu parametrov. Inou možnosťou je použiť rôzne horné ohraničenia chyby zovšeobecnenia. Ak je takéto ohraničenie diferencovateľné, možno ho minimalizovať gradientnými metódami najväčšieho spádu a nájdený argument optima použiť ako aproximáciu hodnôt parametrov modelu, ktoré minimalizujú chybu zovšeobecnenia.

Pri veľkom počte parametrov modelu a pri optimalizácii nielen hodnôt parametrov, ale i tvaru jadra, viacerí autori experimentovali s evolučné stratégie a genetickým programovaním (napr. [3], [4]).

6. Programové vybavenie

Na výpočty sme použili program LIBSVM ([1]). Vykonáva všetky fázy algoritmu oporných bodov a jeho autori priebežne zdokonaľujú v ňom implementované optimalizačné metódy riešenia úlohy kvadratického programovania.

Program rieši rozšírenú verziu úlohy SVM klasifikácie na dve skupiny ((14)) tvaru

$$\min_{w,\xi,b} \frac{1}{2} < w, w > + Cw_{+1} \sum_{y_i=1} \xi_i + Cw_{-1} \sum_{y_i=-1} \xi_i \tag{17}$$
$$y_i(< w, \phi(x_i) > +b) \geq 1 - \xi_i; i = 1, \ldots, \ell$$
$$\xi_i \geq 0; i = 1, \ldots, \ell.$$

Pri triedení na viac ako dve skupiny používa metódu $1:1$. Parametre $w_1, \ldots, w_k$, kde k je počet tried, môžu vyvážiť skresľujúci vplyv nerovnakého počtu trénovacích vzoriek v rôznych triedach.

Vstupmi do programu sú trénovacia a testovacia množina, druh jadra ((a), b) alebo d)), parametre jadra, parametre C a w_i a toleranciu ε určujúcu kritérium ukončenia iteračnej metódy. Pred spustením vlastného algoritmu sa odporúča osobitne každý atribút x v tréningovej množine lineárne preškálovať do intervalu $< u, v >=< -1, 1 >$ alebo $< u, v >=< 0, 1 >$. Ak porade označíme ako m_{train} a M_{train} minimum a maximum x v trénovacej množine, tak hodnoty toho istého atribútu x' v testovacej množine sa potom preškálujú transformáciou

$$g(x') = u + \frac{x' - m_{\text{train}}}{M_{\text{train}} - m_{\text{train}}}(v - u).$$

Takouto transformáciou sa potlačí skreslenie spôsobené rozličnými jednotkami, v ktorých sa merajú jednotlivé atribúty, a urýchlia sa numerické výpočty.

Ako nástroj na voľbu modelu je v LIBSVM implementovaná $m-$násobná crossvalidácia. Na automatické prehľadávanie siete mrežových bodov v parametrickom priestore sa dá použiť iba pre štandardné gaussovské jadro.

Simulácie rozdelenia množiny všetkých vzoriek na trénovaciu a testovaciu časť sa robili v Matlabe s volaním matlabovských verzií funkcií programu LIBSVM.

7. Úspešnosť metódy pri erytropoetínových údajoch

V tejto časti uvedieme výsledky klasifikácie súboru škvŕn získanej zo záverečnej verzie segmentácie chemoluminiscenčných obrazcov EPO. Súbor sa skladal z 5977 vzoriek, z ktorých bolo 1789 pravých machuliek (trieda s označením 1), 4091 artefaktov (označenie 0), 69 predstavovalo artefakt zliaty s pravou machuľkou (označenie 2) a 28 predstavovalo dve zliate machuľky (označenie 3).

Pri triedení údajov sme uvažovali iba štandardné gaussovské jadro. Metaparametre C a $g = 1/\sigma^2$ sa hľadali ako body z mriežky $\{2^{-5}, 2^{-3}, \ldots, 2^{17}\} \times \{2^{-15}, 2^{-13}, \ldots, 2^3\}$ s najväčším podielom správne zatriedených vzoriek pri $5-$ alebo $10-$násobnej crossvalidácii na vopred určenej oficiálnej trénovacej množine.

Aby boli výpočty únosné, pre každú zo 100 simulácií rozdelenia danej údajovej množiny na trénovaciu a testovaciu časť sa trénovalo už iba s touto spoločnou hodnotou (C, g).

V nasledujúcich dvoch tabuľkách sú výsledky $5-$ a $10-$násobnej crossvalidácie na oficiálnej tréningovej množine na základe ktorých sa vybrali hodnoty metaparametrov C (regularizačná konštanta) a $g = \frac{1}{\sigma^2}$ (parameter štandardného gaussovského jadra); $\tilde{C} = \ln_2 C$, $\tilde{g} = \ln_2 g$; podiely crossvalidáciou správne zatriedených tréningových vzoriek (v percentách) sú pre vybrané hodnoty (C, g) vysádzané polotučne:

$\tilde{g}\backslash\tilde{C}$	-5	-3	-1	1	3	5	7	9	11	13	15	17
-15	68.4607	68.4607	68.4607	68.4607	68.4607	90.1562	93.5862	94.3112	94.4785	94.7016	94.9805	95.092
-13	68.4607	68.4607	68.4607	68.4607	90.1283	93.5862	94.3112	94.5344	94.7016	95.0084	95.1478	95.1757
-11	68.4607	68.4607	68.4607	90.1283	93.5862	94.3112	94.5622	94.6737	95.0363	95.2315	95.2593	95.3709
-9	68.4607	68.4607	90.0446	93.5862	94.3391	94.5622	94.7295	95.2315	95.4267	95.5382	95.4824	95.4267
-7	68.4607	89.6263	93.6698	94.367	94.618	95.0641	95.3709	95.5661	95.5103	95.4824	95.3988	95.4267
-5	88.8455	93.6977	94.5343	94.841	95.2593	95.4545	95.5661	95.4545	95.5661	95.5103	95.5103	95.5661
-3	94.0045	94.8689	95.2593	95.4267	95.4267	95.4545	95.594	95.5382	95.6219	95.5103	95.1478	94.367
-1	95.2593	95.4267	95.3988	95.5103	95.6497	95.6219	95.6219	95.3988	94.8968	94.5622	93.8371	93.5304
1	95.1757	95.4545	95.5103	95.6776	**95.8728**	95.6219	94.7853	94.367	93.3352	92.8332	92.7775	92.3592
3	92.6102	94.7016	95.6219	95.7334	95.7055	94.4228	93.7535	93.5862	93.4746	93.5025	93.4467	93.4189

$\tilde{g}\backslash\tilde{C}$	−5	−3	−1	1	3	5	7	9	11	13	15	17
−15	68.4607	68.4607	68.4607	68.4607	68.4607	92.2755	93.6977	94.367	94.5064	94.7016	95.0084	95.0084
−13	68.4607	68.4607	68.4607	68.4607	92.2755	93.6977	94.367	94.5064	94.7295	94.9805	95.092	95.1478
−11	68.4607	68.4607	68.4607	92.2755	93.6977	94.367	94.4785	94.7295	95.0363	95.1478	92.2593	95.5103
−9	68.4607	68.4607	92.2755	93.6977	94.367	94.5343	94.7574	95.2593	95.3709	95.5661	95.5103	95.4267
−7	68.4607	92.2755	93.6977	94.367	94.5901	94.9805	95.4545	95.5661	95.5661	95.5103	95.594	95.594
−5	92.1361	93.7535	94.5622	94.9526	95.2872	95.5103	95.6497	95.5661	95.6497	95.6776	95.7334	95.7334
−3	94.0881	94.8968	95.2315	95.3988	95.5382	95.6497	95.6776	95.6497	95.7334	95.7055	95.3988	95.0084
−1	95.2593	95.4545	95.4267	95.5382	95.6497	95.7334	95.7613	95.6219	95.0363	94.7016	93.9766	93.2794
1	95.1478	95.3988	95.5382	95.6497	95.7334	95.594	95.0084	94.5622	93.4746	92.7496	92.638	92.3034
3	93.0842	94.7574	95.6776	**95.7892**	95.6219	94.6458	93.6419	93.3631	93.2236	93.2236	93.2236	93.3352

Pre takto vybrané hodnoty (C, g) ďalej uvádzame podiely zle zatriedených vzoriek v celej testovacej množine, aj v rámci jej jednotlivých kategórií 1, 0, 2 ,3. A) Oficiálna trénovacia (3586 vzoriek $\doteq 60\%$ z 5977) a testovacia množina (2391 vzoriek):
preškálované dáta:

(C, g)	$5-$nás. CV	$10-$nás. CV	Chyba1	Chyba0	Chyba2	Chyba3	Chyba
$(2^3; 2^1)$	95.8728%	95.7334%	0.0475	0.0208	1	1	0.0448
$(2^1; 2^3)$	95.7334%	95.7892%	0.0461	0.0214	1	1	0.0448

B) Priemerné výsledky zo 100 simulácií náhodného rozdelenia celej údajovej množiny na trénovaciu (predstavuje 60%) a testovaciu množinu (jej doplnok):
preškálované dáta:

(C, g)	Chyba	jej smer. odch	Chyba1	Chyba0	Chyba2	Chyba3
$(2^3; 2^1)$	0.0452	0.0025	0.0533	0.0196	0.9716	1.0000
$(2^1; 2^3)$	0.0466	0.0025	0.0581	0.0193	0.9857	1.0000

Chybové matice pre tieto 2 hodnoty metaparametra sú ($(i, j)-$prvok tejto matice udáva podiel počtu vzoriek, ktoré sú v skutočnosti (t.j. podľa pôvodného zatriedenia experta) z $i-$tej kategórie a klasifikátorom boli zaradené do $j-$tej kategórie, a počtu vzoriek, ktoré sú v skutočnosti z $i-$tej kategórie. Poradie indexov tried je 0, 1, 2, 3.):

$$\begin{pmatrix} 0.9804 & 0.0189 & 0.0007 & 0.0000 \\ 0.0527 & 0.9467 & 0.0005 & 0 \\ 0.0702 & 0.9014 & 0.0284 & 0 \\ 0.1443 & 0.8552 & 0.0005 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0.9807 & 0.0190 & 0.0003 & 0 \\ 0.0576 & 0.9419 & 0.0005 & 0.0000 \\ 0.1147 & 0.8710 & 0.0143 & 0 \\ 0.1402 & 0.8549 & 0.0050 & 0 \end{pmatrix}.$$

Keďže podľa numerických výsledkov sa objekty z tried 2 a 3 takmer nedajú rozpoznať (a najčastejšie boli zaradené do triedy 1), a vzhľadom na to, že tieto dve triedy aj spolu zaberajú iba zanedbateľnú časť celého súboru, ďalej sa skúmali varianty postupov, ktoré objekty tried 2 a 3 vyradili z trénovacej množiny alebo ich priradili do triedy 1. Nimi sa celková chyba zmenšila asi na 3 percentá.

V porovnaní s inými klasifikačnými metódami patrila metóda oporných bodov medzi najlepšie, najmä pri rozpoznávaní najpočetnejšej triedy. Naopak, pri rozpoznávaní prvkov zriedkavých tried zaostávala aj za celkovo horšími metódami.

Na výskum prispel grant agentúry Vega č. 2/4026/04.

Literatúra

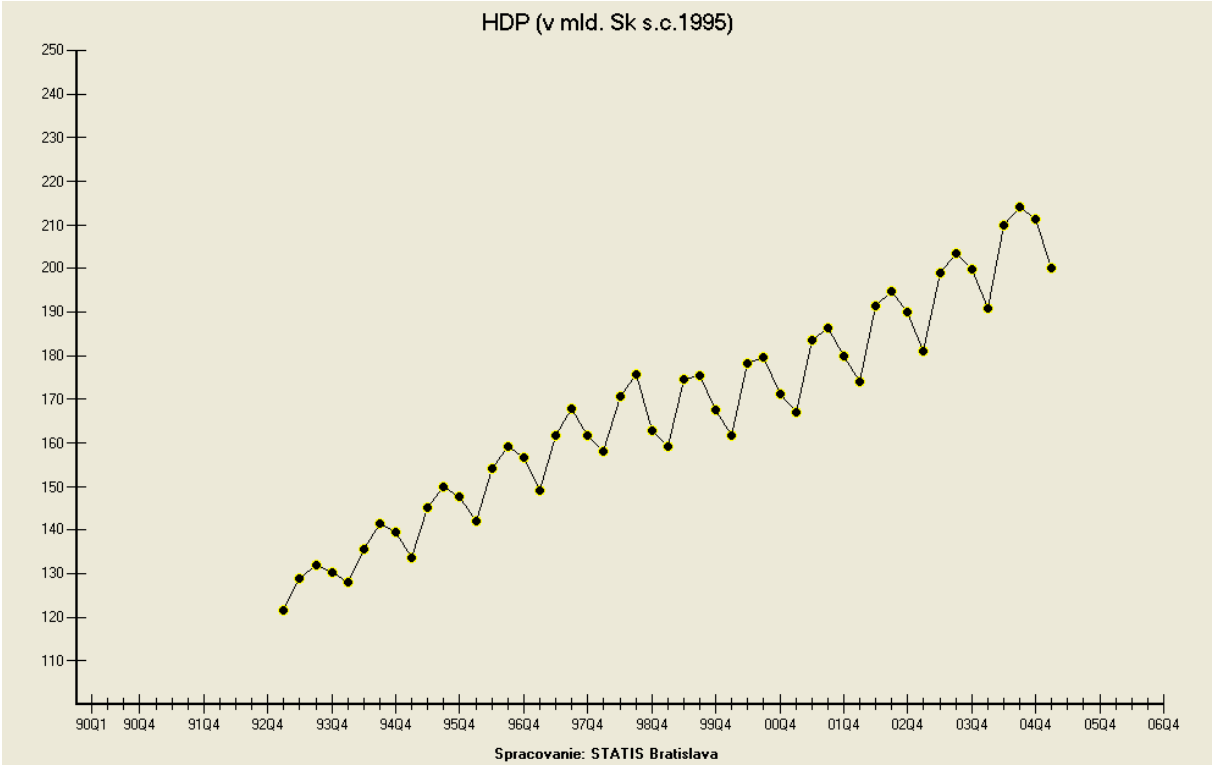
[1] Chang, Ch.-Ch., Lin, Ch.-J. (2001, posledná verzia 2005) *LIBSVM: a library for support vector machines.* National Taiwan University, Taipei, 1 – 28.
http://www.csie.ntu.edu.tw/∼cjlin/papers/libsvm.ps.gz
Software: http://www.csie.ntu.edu.tw/∼cjlin/libsvm
[2] Crammer, K., Singer, Y. (2001) On the algorithmic implementation of multiclass kernel-based vector machines, J. of Machine Learning Research 2, 265-292.
[3] Friedrichs, F., Igel, Ch. (2004) Evolutionary tuning of multiple SVM parameters, Preprint, 1-13.
[4] Givon, L. (2004) Searching for SVM kernels with genetic programming, internet. 1-4.

[5] Hsu, Ch.-W., Lin, Ch.-J. (2002) A comparison of methods for multi-class support vector machines, IEEE Trans. on Neural Networks 13, 2, 415-425.
[6] Kolmogorov, A. N., Fomin, S.V. (1975) Základy teorie funkcí a funkcionální analýzy, SNTL, Praha.
[7] Mizera, I. (1993) Vybrané kapitoly z matematickej štatistiky. Poznámky z prednášok, MFF UK, Bratislava.
[8] Plesník, J. (1987) Matematické programovanie po základoch lineárneho programovania, MFF UK, Bratislava.
[9] Poggio, T., Rifkin, R., Mukherjee, S., Niyogi, P. (2004) General conditions for predictivity in learning theory, Nature 428, 419-422.
[10] Vapnik, V. N. (1998) Statistical learning theory, J. Wiley, New York.

Aktuálny ekonomický vývoj SR do apríla 2005

Jozef Chajdiak

Vývoj štvrťročných objem HDP v stálych cenách roku 1995 je uvedený na obr. 1. Vidíme sínusoidu vývoja s rastúcim trendom ale tento obrázok je vhodný skôr pre špecialistov. Podstatne zrozumiteľnejšia je prezentácia vývoja HDP na obr. 2, na ktorom je znázornený medziročný vývoj (tempá prírastku v %) kĺzavých ročných objemov HDP v s.c. (t.j. rok 1994 k roku 1993; rok: 2. štvrťrok 1994 až 1. štvrťrok 1995 k roku: 2. štvrťrok 1993 až 1. štvrťrok 1994; rok: 3. štvrťrok 1994 až 2. štvrťrok 1995 k roku: 3. štvrťrok 1993 až 2. štvrťrok 1994, atď.). Vo vývoji vidieť jasne dve etapy. Prvá etapa od roku 1993 po 3. štvrťrok 1998 s postupným poklesom tempa prírastku z úrovní nad 6 % o zhruba dva percentuálne body, potom následné radikálne zníženie temp prírastku pod jedno percento (rok: 4. štvrťrok 1998 až 3. štvrťrok 1999 k roku: 4. štvrťrok 1997 až 3. štvrťrok 1998 – z hľadiska tempa prírastku najhorší výsledok). V druhej etape od roku 2000 doteraz vidíme postupný zrýchľujúci sa rast.



Obr. 1 Vývoj štvrťročných objemov HDP (v mld. Sk s. c. 1995)

Na obr. 3 a 4 je znázornený vývoj zahraničného obchodu tovarov. Na obr. 3 sú kĺzavé ročné saldá a mesačné saldá zahraničného obchodu rozpočtu. Vidíme postupné zvyšovanie salda. Na obr. 4 sú uvedené kĺzavé ročné hodnoty exportu a importu tovarov. Vidíme rast objemu zahraničného obchodu.

Na obr. 5 a 6 je znázornený vývoj štátneho rozpočtu. Na obr. 5 sú kĺzavé ročné saldá a mesačné saldá štátneho rozpočtu. V poslednom období môžeme vidieť náznak zlepšovania salda štátneho rozpočtu a na obr. 6 s náznakom zbližovania príjmov a výdavkov štátneho rozpočtu.

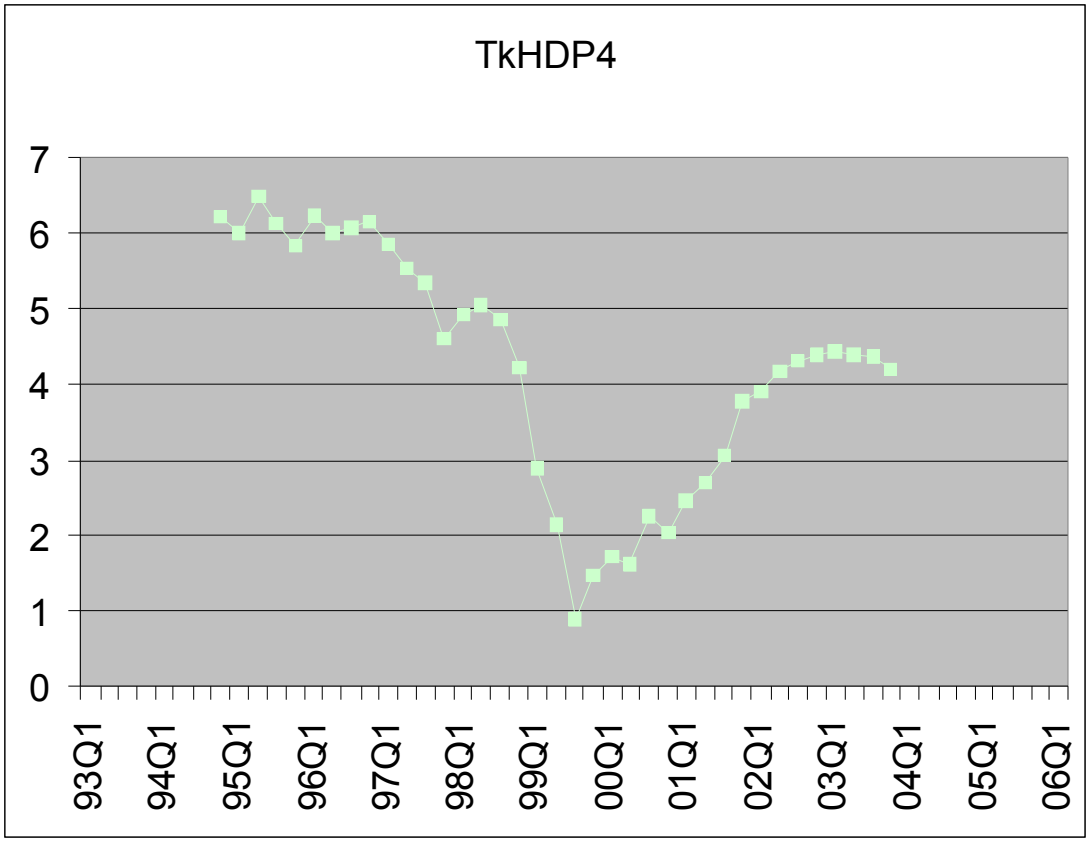
Na obr. 7 vidíme vývoj počtu nezamestnaných v SR. Kým v predchádzajúcich rokoch sa maximá pohybovali vyše 560 tisíc nezamestnaných, v poslednom roku je zreteľný pokles ich počtu: napriek tomu počet cez 300 tisíc je zdrvujúco vysoký.

Na obr. 8 je vývoj medziročnej inflácie. Po historickom minime vo veľkosti 2% v lete roku 2002 ho máme opäť v lete 2005.

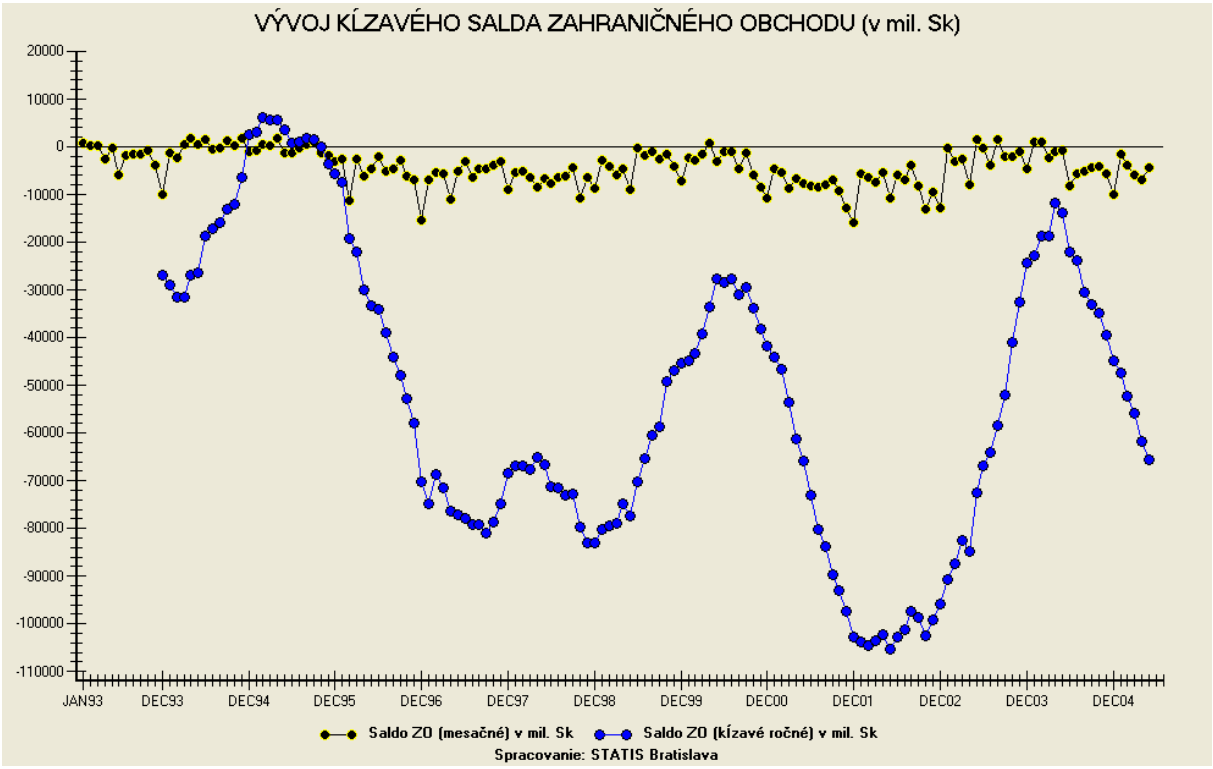
Na obr. 9 je znázornený vývoj podielu salda štátneho rozpočtu k objemu HDP. Cifry svedčia o bohatých rezervách v procese zlepšovania hospodárenia so štátnym rozpočtom.

Na obr. 10 je znázornený vývoj reálnych miezd v SR. Základ (čiara 1) predstavuje úroveň miezd v jednotlivých štvrťrokoch roku 1989. Vidíme, že sa zarába menej ako v roku 1989. Posledná cifra za 1. štvrťrok 2005 predstavuje 93.5 % zo 1. štvrťroku 1989.

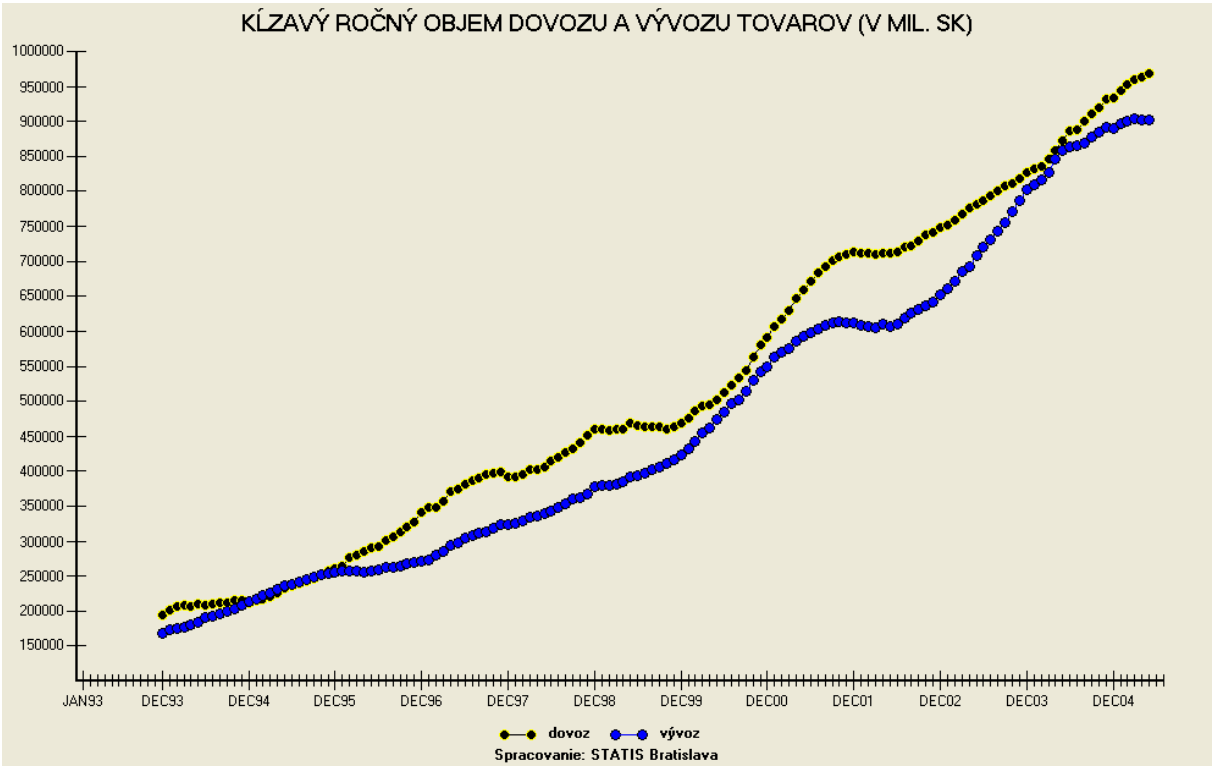
Na obr. 11 je znázornený vývoj koeficienta makroekonomického vývoja v SR. S výnimkou roku 1995 sa stále pohybujeme pod úrovňou nula, čo predstavuje negatívny rast (pokles) ekonomiky SR



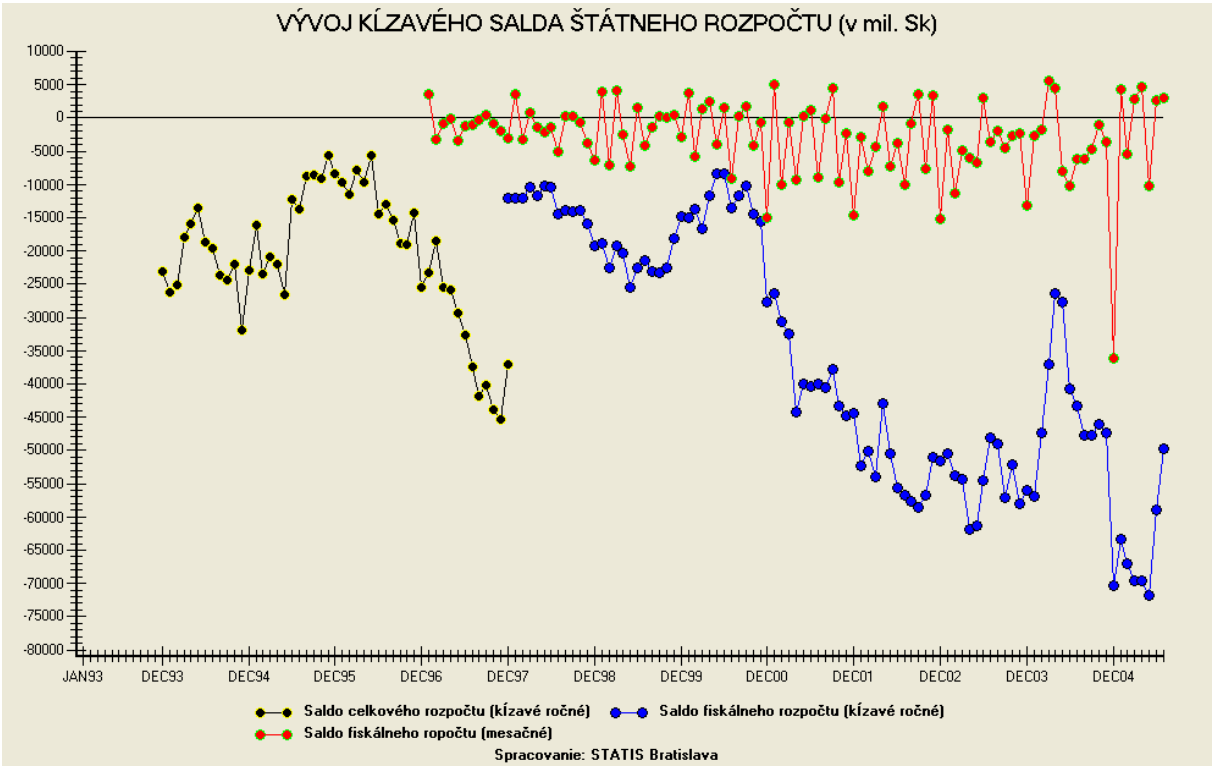
Obr. 2 Medziročný vývoj temp prírastku kĺzavých ročných objemov HDP


VÝVOJ KĹZAVÉHO SALDA ZAHRANIČNÉHO OBCHODU (v mil. Sk)
20000
10000
0
-10000
-20000
-30000
-40000
-50000
-60000
-70000
-80000
-90000
-100000
-110000
JAN93
DEC93
DEC94
DEC95
DEC96
DEC97
DEC98
DEC99
DEC00
DEC01
DEC02
DEC03
DEC04
Saldo ZO (mesačné) v mil. Sk
Saldo ZO (kĺzavé ročné) v mil. Sk
Spracovanie: STATIS Bratislava

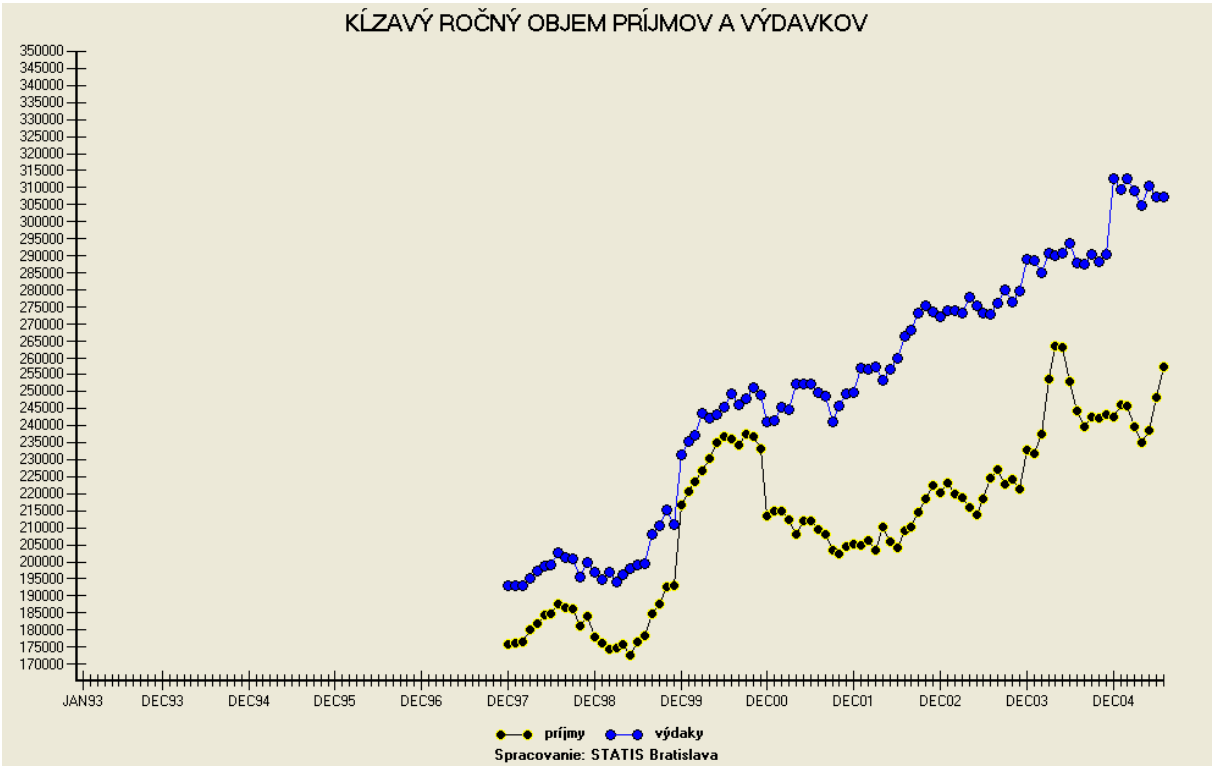
Obr. 3


KĹZAVÝ ROČNÝ OBJEM DOVOZU A VÝVOZU TOVAROV (V MIL. SK)
1000000
950000
900000
850000
800000
750000
700000
650000
600000
550000
500000
450000
400000
350000
300000
250000
200000
150000
JAN93
DEC93
DEC94
DEC95
DEC96
DEC97
DEC98
DEC99
DEC00
DEC01
DEC02
DEC03
DEC04
dovoz
vývoz
Spracovanie: STATIS Bratislava

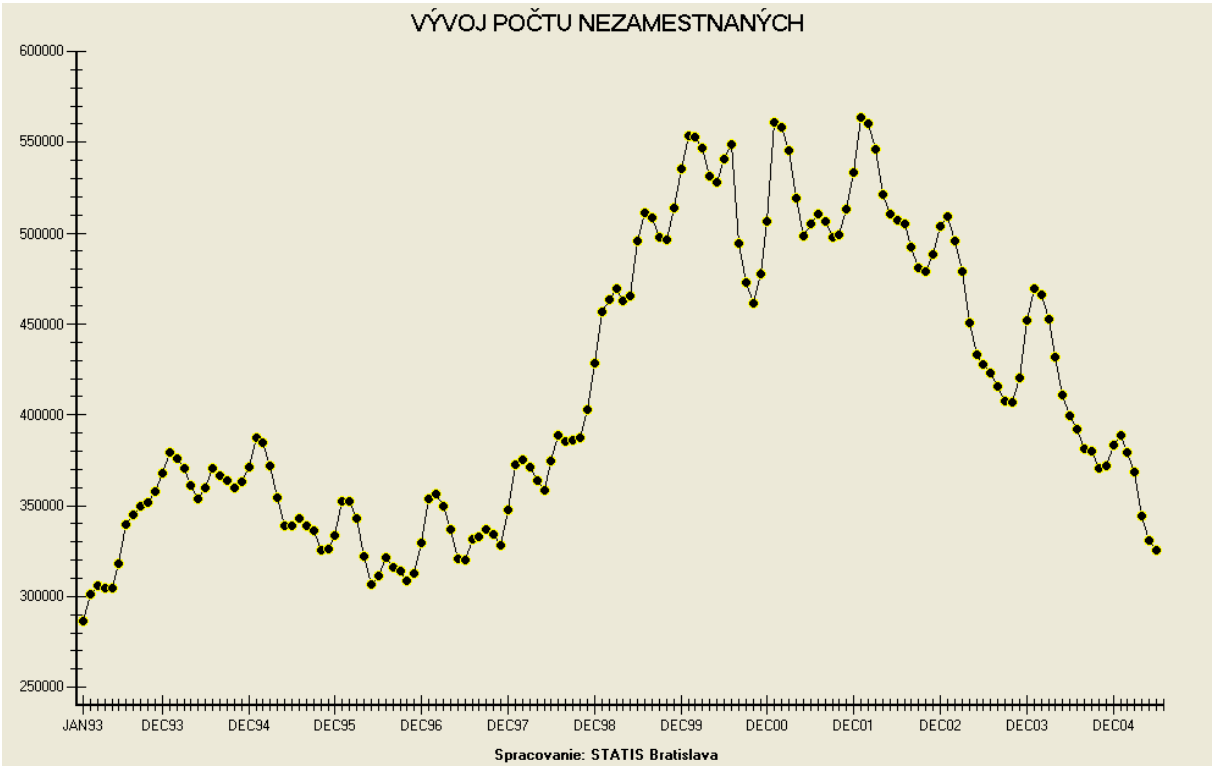
Obr. 4

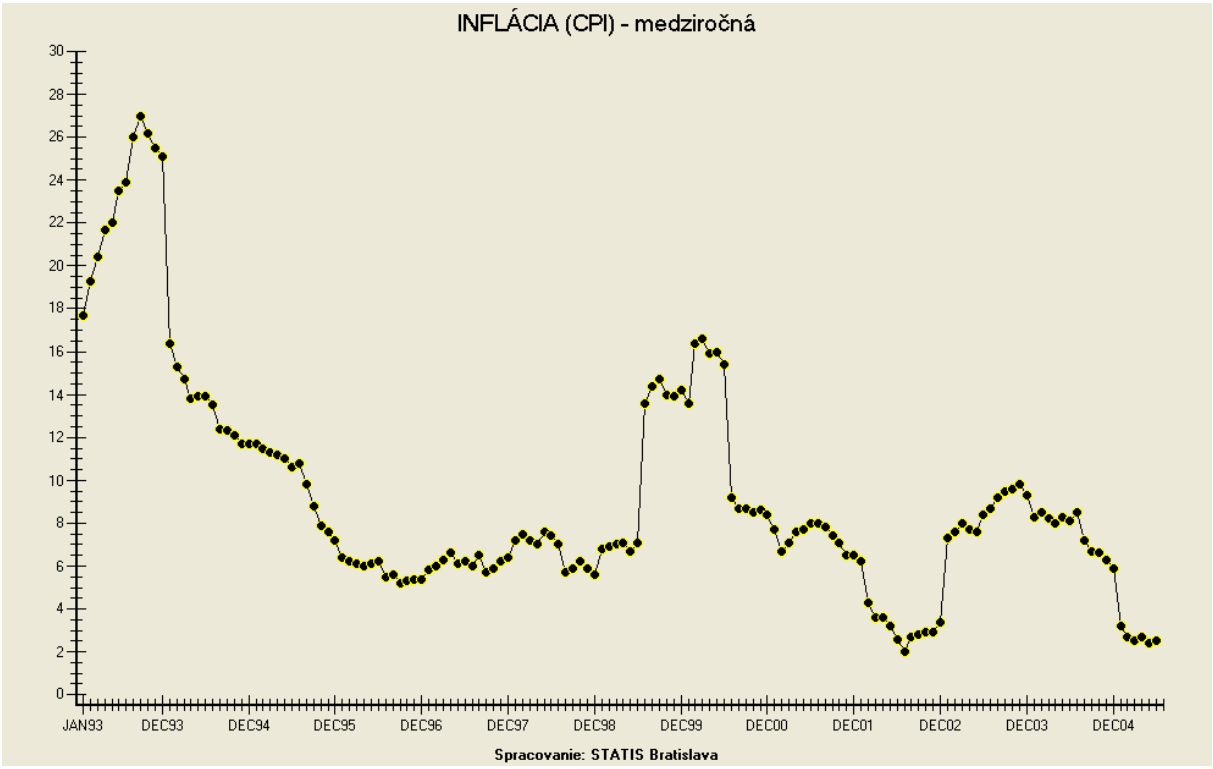

VÝVOJ KĹZAVÉHO SALDA ŠTÁTNEHO ROZPOČTU (v mil. Sk)
10000
5000
0
-5000
-10000
-15000
-20000
-25000
-30000
-35000
-40000
-45000
-50000
-55000
-60000
-65000
-70000
-75000
-80000
JAN93
DEC93
DEC94
DEC95
DEC96
DEC97
DEC98
DEC99
DEC00
DEC01
DEC02
DEC03
DEC04
Saldo celkového rozpočtu (kĺzavé ročné)
Saldo fiskálneho rozpočtu (kĺzavé ročné)
Saldo fiskálneho ropočtu (mesačné)
Spracovanie: STATIS Bratislava

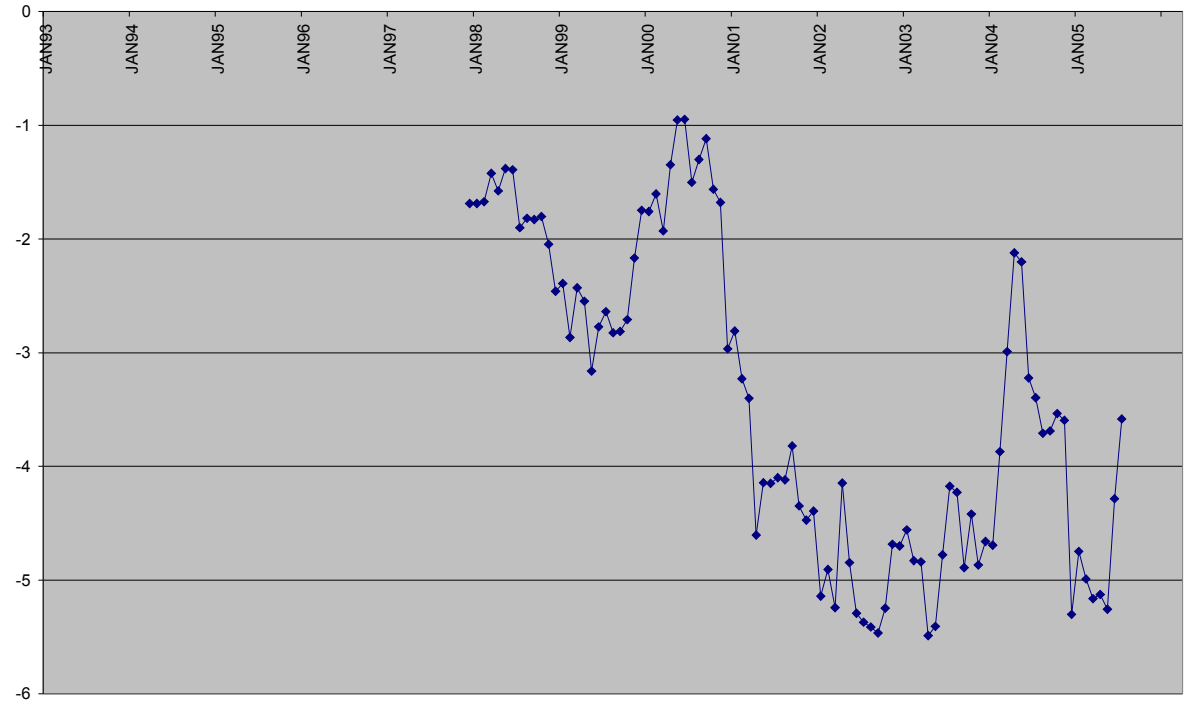
Obr. 5


KĹZAVÝ ROČNÝ OBJEM PRÍJMOV A VÝDAVKOV
350000
345000
340000
335000
330000
325000
320000
315000
310000
305000
300000
295000
290000
285000
280000
275000
270000
265000
260000
255000
250000
245000
240000
235000
230000
225000
220000
215000
210000
205000
200000
195000
190000
185000
180000
175000
170000
JAN93
DEC93
DEC94
DEC95
DEC96
DEC97
DEC98
DEC99
DEC00
DEC01
DEC02
DEC03
DEC04
príjmy
výdaky
Spracovanie: STATIS Bratislava

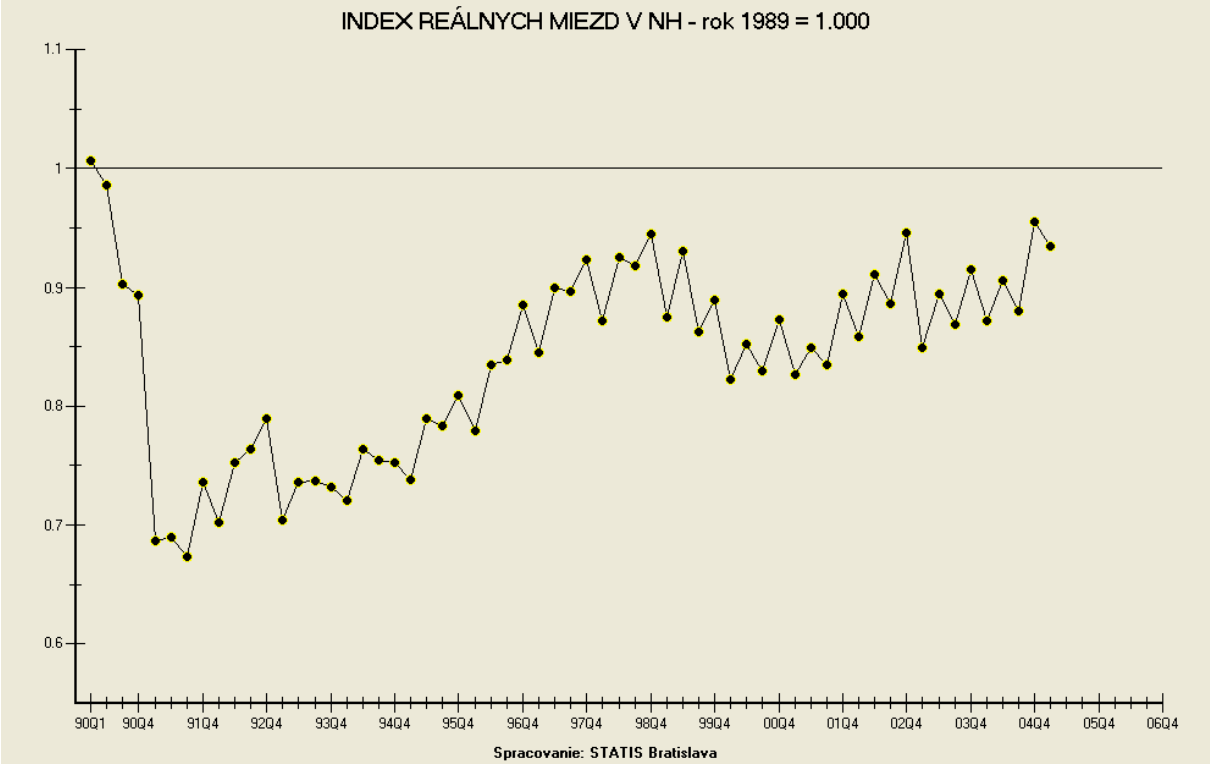
Obr. 6


VÝVOJ POČTU NEZAMESTNANÝCH
600000
550000
500000
450000
400000
350000
300000
250000
JAN93
DEC93
DEC94
DEC95
DEC96
DEC97
DEC98
DEC99
DEC00
DEC01
DEC02
DEC03
DEC04
Spracovanie: STATIS Bratislava
Obr. 7

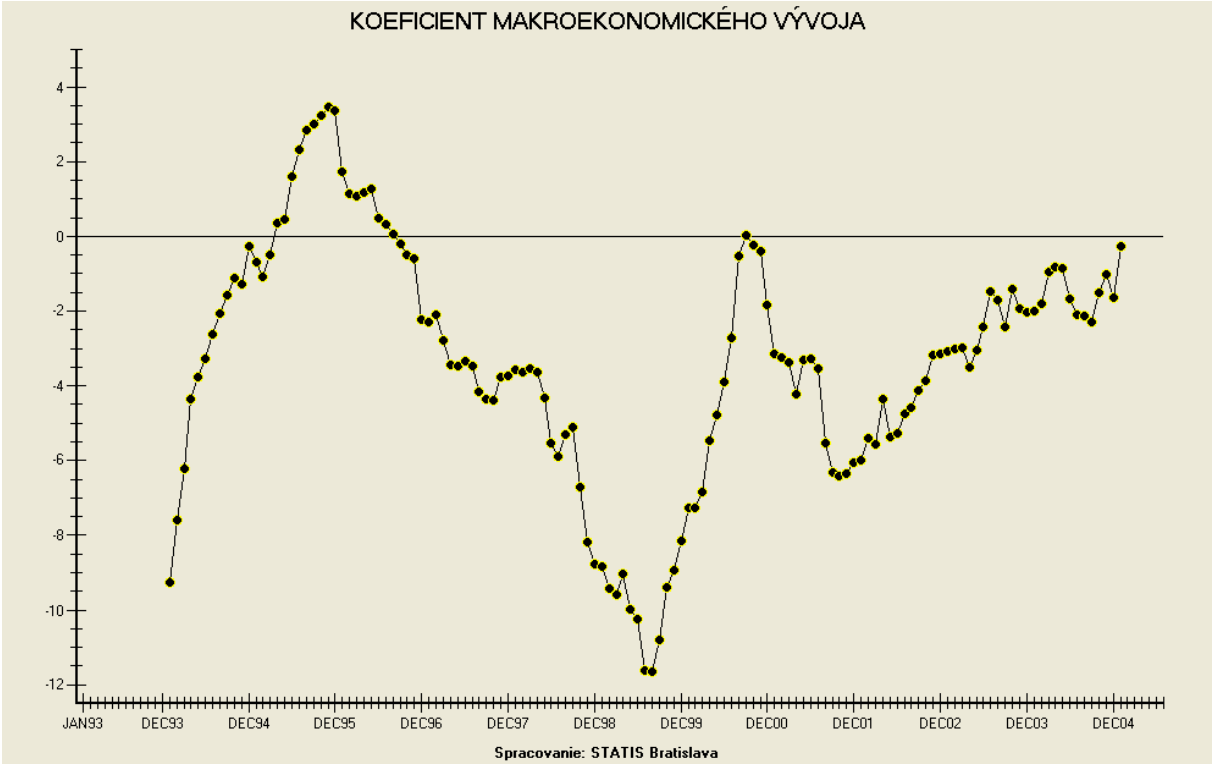

INFLÁCIA (CPI) - medziročná
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
JAN93
DEC93
DEC94
DEC95
DEC96
DEC97
DEC98
DEC99
DEC00
DEC01
DEC02
DEC03
DEC04
Spracovanie: STATIS Bratislava
Obr. 8



Obr. 9 Vývoj podielu salda štátneho rozpočtu na HDP (z kĺzavých ročných hodnôt)



Obr. 10



Obr. 11

Adresa autora:
chajdiak@statis.biz
chajdiak@euba.sk

Štatistika na EF UMB v Banskej Bystrici

A. Kaščáková, P. Kráľ[1], L. Kulčár, G. Nedelová
Ekonomická fakulta, Univerzita Mateja Bela
Katedra hospodárskej informatiky
Tajovského 10, 975 90 Banská Bystrica
Slovensko
alena.kascakova@umb.sk
pavol.kral@umb.sk
ladislav.kulcar@umb.sk
gabriela.nedelova@umb.sk

1. Úvod

Cieľom nášho článku je predstaviť vyučovanie štatistiky (presnejšie štatistických predmetov) na Ekonomickej fakulte Univerzity Mateja Bela v Banskej Bystrici (ďalej len EF), pričom sa zameriame predovšetkým na základný kurz štatistiky pre študentov denného štúdia. Začneme zaradením štatistických predmetov v študijných plánoch EF, hodinovou dotáciou a obsahom základného kurzu štatistiky. Ďalšiu časť článku tvoria ukážky materiálov z cvičení a písomiek a úspešnosť študentov jednotlivých odborov na skúške z predmetov Štatistika, Štatistika 1, Štatistika 2 v rokoch 2001–2003. V závere naznačíme problémy, s ktorými sa stretávame vo vyučovacom procese a prvotné návrhy ich riešenia, subjektívne názory študentov i vyučujúcich na štatistické predmety a predstavy vyučujúcich o ďalšom vývoji vyučovania štatistiky na EF.

2. Štatistické predmety na EF

Štatistické predmety vyučované môžeme rozdeliť na dve základné, na seba nadväzujúce skupiny:

1. základný kurz – prebieha v druhom roku denného štúdia vo všetkých odboroch v rozsahu 1 – 2 semestre, tvoria ho predmety Štatistika 1, Štatistika 2 a Štatistika.

2. voliteľné a povinne voliteľné predmety – sú umiestnené v treťom a štvrtom roku štúdia, patria sem predmety Matematicko štatistické metódy, Analytické metódy prieskumu trhu (Analýza časových radov) a Kvantitatívny manažment (Teória rozhodovania). Nebudeme sa nimi podrobnejšie zaoberať, len poznámenávame, že ich hlavným cieľom je prehĺbiť poznatky študentov získané počas základného kurzu a viesť ich k praktickému využitiu štatistických metód (napríklad pri písaní diplomovej práce).

2.1. Štatistika 1

Tento predmet absolvujú študenti denného štúdia odboru ERP(Ekonomika a riadenie podniku) v zimnom semestri druhého ročníka. Časová dotácia predmetu je prednáška 2h (80 min)/týždeň a cvičenie 2h/týždeň. Osnova predmetu Štatistika 1 je nasledujúca:

[1]korešpondujúci autor

1. Podstata štatistiky, spôsoby štatistického zisťovania a metódy triedenia štatistických údajov, jednorozmerné a dvojrozmerné triedenie, charakter štatistických znakov. Absolútne, relatívne a kumulatívne početnosti.

2. Znázorňovacie prostriedky v štatistike.

3. Meranie v štatistike, stredné hodnoty, klasifikácia stredných hodnôt, priemery aritmetický, geometrický, harmonický.

4. Stredné hodnoty polohy (modus a medián), kvantily. Využitie stredných hodnôt na určenie typu zošikmenia rozdelenia početností.

5. Variabilita, miery variability. Rozklad rozptylu. Šikmosť a špicatosť rozdelenia.

6. Kovariancia.

7. Meranie dynamiky ekonomických javov pomocou indexov. Klasifikácia indexov. Laspeyresova a Paascheho koncepcia určovania indexov. Vyjadrenie indexov relatívne a absolútne.

8. Časové rady, druhy časových radov a podmienky ich zostavenia. Chronologický priemer. Pomocné prostriedky rozboru časových radov. Zložky časového radu, analytické a neanalytické určovanie trendovej zložky časového radu. Prognóza v časových radoch. Určenie sezónnosti v časovom rade.

Študenti v priebehu semestra absolvujú dve 50 bodové písomné práce, pričom každá je tvorená príkladovou (30b) a teoretickou časťou (20 b). Minimálna hranica pre úspešné absolvovanie celého predmetu je 65b.

2.2. Štatistika 2

Tento predmet absolvujú študenti denného štúdia odboru ERP(Ekonomika a riadenie podniku) v letnom semestri druhého ročníka. Časová dotácia predmetu je prednáška 2h (80 min)/týždeň a cvičenie 2h/týždeň. Osnova predmetu Štatistika 2 je nasledujúca:

1. Pravdepodobnosť - základné pojmy a definície. Náhodná premenná, distribučná funkcia funkcia hustoty pravdepodobnosti. Zákon rozdelenia pravdepodobnosti diskrétnej a spojitej náhodnej premennej.

2. Výberové skúmanie - druhy a technika výberov. Bodový a intervalový odhad. Intervalový odhad strednej hodnoty, relatívnej početnosti a rozptylu základného súboru. Testovanie hypotéz o stredných hodnotách, relatívnych početnostiach a o rozptyle.

3. Testovanie hypotéz o zhode stredných hodnôt, relatívnych početností a rozptylov. Test zhody rozdelenia, test normality rozdelenia.

4. Analýza závislosti kvantitatívnych znakov. Regresná analýza, metóda najmenších štvorcov, párová lineárna, nelineárna a viacnásobná regresia. Odhady a testy regresných a korelačných charakteristík.

5. Korelačná analýza, druhý rozklad rozptylu na jeho zložky. Miery tesnosti závislosti. Viacnásobná a čiastková korelácia.

6. Triedenie štatistického súboru podľa kvalitatívnych znakov. Analýza závislosti kvalitatívnych znakov, test nezávislosti kvalitatívnych znakov, miery tesnosti závislosti kvalitatívnych znakov.

Spôsob hodnotenia je rovnaký ako v prípade predmetu Štatistika 1.

2.3. Štatistika

Tento predmet absolvujú študenti denného štúdia odborov CR (Cestovný ruch) a VES (Verejná ekonomika a správa) v zimnom semestri druhého ročníka. Časová dotácia predmetu je prednáška 2h (80 min)/týždeň a cvičenie 2h/týždeň. Osnova predmetu je v podstate zjednotením osnov predmetov Štatistika 1 a Štatistika 2 odboru ERP. Študenti v priebehu semestra absolvujú tri písomné práce, pričom každá je tvorená príkladovou a teoretickou časťou. Bodové hodnotenie jednotlivých písomných prác je nasledujúce:

1. písomná práca 35b (20b príklady, 15b teoretické otázky)
2. písomná práca 35b (20b príklady, 15b teoretické otázky)
3. písomná práca 30b (20b príklady, 10b teoretické otázky)

Minimálna hranica pre úspešné absolvovanie predmetu je 65b.

3. Ukážka úloh z cvičenia

V tejto ukážke použijeme úlohy z cvičenia zameraného na korelačnú a regresnú analýzu.

Príklad č.1

V piatich veľkých firmách sa skúmal vzťah produkcie a nákladov (v tis. Sk). Boli zistené nasledovné údaje:

Firma	Produkcia	Celkové náklady
1	2	9
2	4	12
3	6	15
4	5	14
5	3	10

Úlohy:

a) Odhadnite lineárnu nákladovú funkciu $C_i = \alpha + \beta \cdot q_i + U_i$, kde C_i sú celkové náklady, q_i je objem produkcie a U_i je náhodná zložka.

$$[C_i = 5,6 + 1,6q_i]$$

b) Na hladine významnosti 0,05 overte, či má daná nezávislá premenná vplyv na závislú náhodnú premennú.

c) Určte 95 %– ný konfidenčný interval pre β.

d) Na 2,5 %–nej hladine významnosti overte predpoklad, že β je menšia ako 1,8. Všetky získané výsledky interpretujte!

Príklad č.2

V 100 náhodne vybraných závodoch sa robil rozbor priemernej štvrťročnej nepodarkovosti výroby. Z previerky máte k dispozícii nasledovné údaje:

Priemerný počet nepodarkov za štvrťrok	Percento plnenia výroby tovaru za štvrťrok	Počet závodov
160	104	8
200	103	16
210	103	24
270	101	19
320	101	15
340	100	12
350	99	4
390	99	2

Úlohy:

a) Za predpokladu linearity vývoja uvedených javov odhadnite funkciu, modelujúcu závislosť plnenia plánu od priemerného štvrťročného počtu nepodarkov.

b) Pri dodržaní stabilných podmienok, vypočítajte predpokladané percento plnenia plánu výroby tovaru, ak bolo za štvrťrok vyrobených 400 nepodarkov.

c) Vypočítajte 95 %-ný interval spoľahlivosti pre regresný koeficient základného súboru.

Príklad č. 3

Majiteľ siete predajní s počítačovou technikou pravidelne sleduje svoje týždenné tržby a vynaložené náklady na reklamu. Keďže predpokladá závislosť medzi týmito ukazovateľmi, chce odhadnúť funkciu, modelujúcu túto závislosť s využitím pre extrapoláciu. V tabuľke sú napozorované hodnoty výšky tržieb a nákladov na reklamu za niekoľko mesiacov prevádzky siete predajní.

Týždeň	Tržby (tis. USD)	Náklady na reklamu (v stovkách USD)
1	1,1	3,9
2	1,7	4,9
3	2,6	7,6
4	2,4	6,8
5	2,3	5,9
6	2,9	9,1
7	0,4	3,4
8	3,2	11,6
9	3,3	14,1
10	3,1	14,9
11	3,2	10,5
12	3,0	9,9
13	3,7	17,1
14	3,3	14,4

Úlohy :

a) Skonštruujte bodový graf závislosti sledovaných ukazovateľov.

b) Za predpokladu lineárnej závislosti zistite, ako sa zmení objem tržieb pri dodatočnom vložení 100 USD do nákladov na reklamu.

c) Vyrovnajte uvedenú závislosť ukazovateľov parabolou.

d) Vypočítajte, aký objem tržieb je možné očakávať pri nákladoch na reklamu vo výške 2 000 USD. Využite oba modely a výsledky porovnajte.

e) Na základe porovnania príslušných mier tesnosti závislosti určte, ktorá z funkcií je vhodnejšia na posúdenie priebehu závislosti.

Príklad č.4

Výsledky skúšok nového pracieho prášku vo veľkokapacitnej automatickej práčke sú uvedené v nasledujúcej tabuľke :

Dávka pracieho prostriedku v g	Účinnosť pracieho prostriedku v %
100	23
200	56
300	51
400	82
500	72
600	100
700	59
800	54

Úloha: Vypočítajte účinnosť pracieho prostriedku pri dávke 900 gramov, ak predbežná analýza ukázala, že vývoj závislosti týchto dvoch javov je definovaný parabolou.

Príklad č.5

Na základe údajov z nasledujúcej tabuľky, pomocou exponenciálnej funkcie charakterizujte priebeh závislosti výdavkov za služby (y_i) od výšky príjmu domácnosti (x_i) v tis. Sk za štvrťrok v 10 náhodne vybraných domácnostiach.

x_i	23	27	26	29	32	35	28	34	37	29
y_i	6,8	7,6	7,2	8,7	10,8	13,4	7,8	12,3	14,5	8,9

Ďalej vypočítajte, aký príjem možno očakávať v domácnosti, ktorá vydala za služby 5 000 Sk za štvrťrok .

Príklad č.6

Spoločnosť, zaoberajúca sa predajom kvetov, sledovala týždenný objem svojich tržieb. Manažér tejto spoločnosti chcel zistiť, ako sa tento sledovaný ukazovateľ vyvíja pri zmenách Ďalších ukazovateľov. Uvažoval o tom, že na predaj kvetov vplýva ročné obdobie a tiež reklama. Preto náhodne vybral 8 týždňov a v nich sledoval týždenný počet reklám v televízii a v tlači a priemernú teplotu ovzdušia v sledovanom týždni.

Týždenný objem tržieb (tis. Sk)	Počet reklám za týždeň	Priemerná týždenná teplota (v ^0C)
37,5	3	5
45,6	5	30
39,3	4	1
39,9	4	8
42,6	5	17
34,8	3	-5
38,1	3	13
48,9	6	28

Úlohy:

a) Odhadnite funkciu, modelujúcu závislosť objemu tržieb od počtu odvysielaných reklám a od priemernej týždennej teploty.

b) Manažér vie, že na budúci týždeň sú objednané 4 reklamy a podľa predpovede má byť priemerná týždenná teplota asi 6 ^{0}C. Aký objem tržieb môže očakávať?

Príklad č.8

V rámci výskumu trhu, ktorý si objednal výrobca určitého výrobku v agentúre pre výskum trhu boli získané okrem iných, nasledovné informácie o tržbách za tento tovar v tis. Sk (y_i), nákladoch na reklamu v tis. Sk (x_i) a o podieli tohto výrobcu na celkovom predaji daného tovaru (z_i):

y_i	149	152	155,7	159	163,3	166	169	172	174,5	176,1	176,5	179
x_i	21	21,8	22,4	23	23,7	24,3	24,9	25,5	25,8	26,	26,2	26,3
z_i	42,5	43,7	44,8	46	47	47,9	49	49,9	50,3	50,9	50,8	51,1

Úlohy:

a) Vypočítajte, koľkými percentami by bolo možné vysvetliť variabilitu tržieb variabilitou podielu na celkovom predaji pri nezmenených nákladoch na reklamu, ak viete, že $r_{yz} = 0.997$ a $r_{zx} = 0.999$.

b) Odhadnite bodovým odhadom, na koľko percent sa dajú rozdiely vo výške tržieb vysvetliť ostatnými uvažovanými činiteľmi.

Príklad č.9

V mesiaci október sa sledoval počet majetkových trestných činov vo vybraných 27 – ich okresoch Slovenska a vplyv miery nezamestnanosti na tento sledovaný ukazovateľ. Výsledky zisťovania boli vytriedené do nasledujúcej korelačnej tabuľky:

x_i/y_i	10–20	–30	–40	–50	50+
5–10	1	2	1	–	–
–15	–	3	7	–	–
–20	–	–	1	5	1
20+	–	–	–	1	5

Úlohy:

a) Vhodnou mierou zistite, či je počet majetkových trestných činov korelovaný s mierou nezamestnanosti a do akej miery.

b) Nájdite bodový odhad regresnej priamky.

c) Vypočítajte výberový koeficient korelácie.

d) Na 5%-nej hladine významnosti overte predpoklad, podľa ktorého sú sledované dva javy úplne lineárne nezávislé.

e) Zostrojte 95 %-ný konfidenčný interval pre koeficient korelácie základného súboru

f) Na hladine významnosti 0,05 overte predpoklad, podľa ktorého má koeficient korelácie základného súboru hodnotu väčšiu ako 0,6.

Študenti po ukončení určitého celku dostanú vzorové riešenia príkladov. Riešenia sú k dispozícii v papierovej a v elektronickej podobe, ktorá je umiestnená na www stránke fakulty. Prípadné otázky môžu prediskutovat okrem samotných cvičení s vyučujúcim počas konzultačných hodín (180 min/týždeň). Zvyčajne je možné si dohodnúť konzultáciu aj mimo vymedzeného času. Vyučujúcim je možne klásť otázky aj prostredníctvom emailu a telefonicky. Túto možnosť ale využívajú najmä študenti diaľkového štúdia.

5. Písomná práca

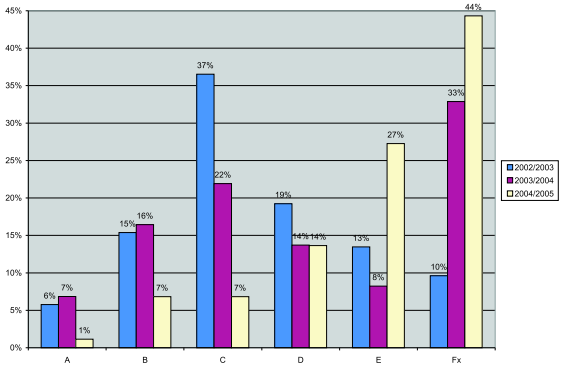
Ako už bolo uvedené, písomné práce pozostávajú z dvoch častí - príkladovej a teoretickej. Z bodového hodnotenia a z ukážky príkladov z cvičení je zrejmé, že väčšiu váhu sme priradili riešeniu praktických úloh, pretože nimi študent preukazuje schopnosť riešiť konkrétne problémy praxe s použitím štatistických metód. Tejto snahe sme dokonca prispôsobili aj teoretické otázky, ktoré často úzko súvisia s popisom konkrétnej metódy, s interpretáciou prípadných výsledkov použitia jednotlivých metód (napríklad interpretácia indexov) a hľadaním príkladov. Príkladová časť písomnej práce je tvorené úlohami, ktoré sú analogické úlohám z cvičení. Z pochopiteľných dôvodov tu neuvedieme konkrétne znenia príkladov (v prípade záujmu môžete kontaktovať jedného z autorov), ale len príklad teoretických otázok:

- vymenujte stredné hodnoty polohy, ktoré poznáte a definujte ich,
- napíšte vlastnosti aritmetického priemeru,
- napíšte vzťah medzi hodnotovým indexom a indexami fyzického objemu a ceny,
- aké zložky môže mať časový rad, popíšte ich,
- napíšte postup hľadania modusu v prípade intervalového triedenia.

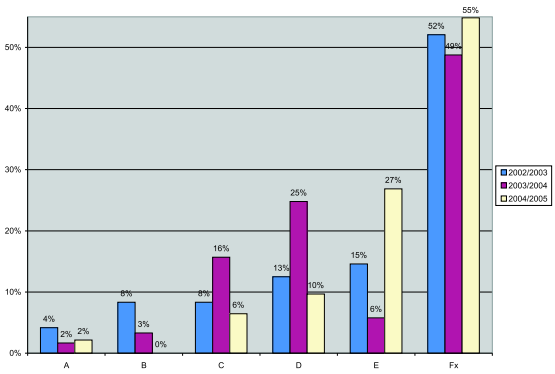
Písomka má teoretickú časť, ktorá nahrádza ústnu skúšku od akademického roka 2004/2005.

6. Úspešnosť študentov 2. ročníka DŠ

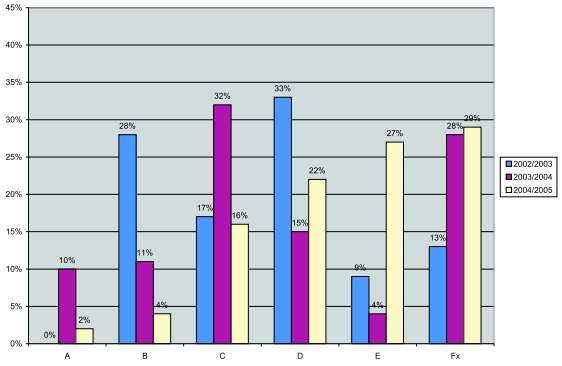
V kreditovom štúdiu na EF sa používa šesť stupňové hodnotenie:
A(100% - 94%), B(93%-87%), C(86%-80%), D(79%-73%), E(72-65%), FX (menej ako 64%). Úspešnosť študentov v akademických rokoch 2002/2003 - 2004/2005 sú v nasledujúcich grafoch:



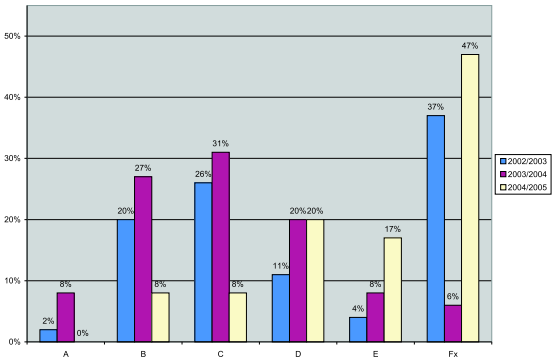
Predmet Štatistika 1, odbor ERP, Banská Bystrica



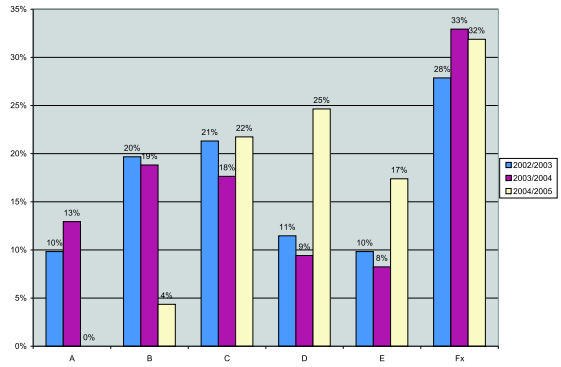
Predmet Štatistika 2, odbor ERP, Banská Bystrica



Predmet Štatistika 1, odbor ERP, Poprad



Predmet Štatistika 2, odbor ERP, Poprad



Predmet Štatistika, odbor CR, Banská Bystrica

7. Subjektívne zhodnotenie výsledkov

Tvrdenia, ktoré sú obsahom tejto kapitoly sú subjektívnou výpoveďou vyučujúcich (a čiastočne aj študentov) štatistických predmetov a na ich objektivizáciu by bolo potrebné uskutočnit rozsiahlejší výskum.

Na základe prvotnej analýzy uvedených grafov je možné povedať, že napriek tomu, že ústna skúška je nahradená teoretickou časťou písomky (čo považujeme za výrazné zníženie náročnosti), vedomosti študentov a ich výsledky sa zhoršujú (vo všetkých odboroch). Percento neúspešných študentov v predmete Štatistika 1 je menšie ako v predmete Štatistika 2 (s výnimkou akademického roka 2003/2004). Táto skutočnosť môže byt spôsobená vyššou náročnosťou učiva.

Ďalej sa ukazuje, že výsledky odboru ERP v Banskej Bystrici a na detašovanom pracovisku v Poprade sú veľmi podobné. V prospech tohoto predpokladu hovoria napríklad nasledujúce skutočnosti:
Zimný semester (predmet Štatistika 1)

- v zimnom semestri v r. 2003/04 je charakter zastúpenia známok podobný bimodálny s dvoma maximami - jednom pri známke C (22% v BB,22% v PP), druhé maximum pri známke FX (33% v BB, 28% v PP),
- v zimnom semestri 2004/05 vidieť rovnaký charakter rozdelenia známok v BB a v PP v tom zmysle, že percentuálne zastúpenie známok sa zvyšuje smerom od známky A po FX, kde nadobúda maximum (v BB 44% a v PP 29%),
- určitý, avšak nie podstatný rozdiel vidíme v rozdelení zastúpenia známok medzi BB a PP v akad. roku 2002/03 v tom zmysle, že v BB je výrazné maximum pri známke C (37 %), zatiaľ čo v PP je toto maximum pri známke D (33%). Tento rozdiel v maximách pri rôznych známkach v BB a PP nemusí byť taký podstatný z toho dôvodu, že obe známky C a D sa bodovým hodnotením výrazne neodlišujú.

Letný semester (predmet Štatistika 2):

- v r. 2002/03 v Poprade bimodálne rozdelenie s maximom pri C (26%) a FX (37%), pričom v BB je prítomný postupný nárast zastúpenia známok od A po FX - unimodálne rozdelenie s maximom pri FX 52%.
- v r. 2003/04 je v PP unimodálne rozdelenie známok s maximom pri C (31%) a s malým zastúpením známky FX (iba 6%), pričom v BB je vidieť dve maximá pri známke D (25%) a FX (48%).

- v r. 2004/05 opäť v PP bimodálne rozdelenie s maximami pri D (20%) a FX (47%), pričom v BB vidieť opäť rovnaký charakter postupného nárastu zastúpenia známok od A, resp. B (2, resp. 0%) až po FX (55%).

Podľa študentov patria štatistické predmety medzi náročné a od toho sa odvíjajú aj ich názory (v zátvorke sú uvedené komentáre vyučujúcich):

- je to zbytočný predmet (tento názor je len dôsledkom toho, že študenti majú problém s jeho absolvovaním),
- zadania úloh sú nezrozumiteľné (snaha naučiť sa kľúčové slová a podľa nich identifikovať problém),
- všetko by sa malo riešiť pomocou počítača (nie je to vhodné bez pochopenia základných princípov),
- rozsah učiva je priveľký (obsah predmetu je v podstate zhodný s obsahom predmetu vyučovaného na školách podobného zamerania).

8. Záver – plány do budúcnosti

V krátkosti sme predstavili vyučovanie štatistických predmetov (najmä základného kurzu štatistiky) a problémov, ktoré sú s ním spojené na EF UMB v Banskej Bystrici v akademickom roku 2004/2005.

Naším hlavným cielom do budúcnosti je neustála snaha o zlepšenie celkovej kvality vyučovania. Nie je možné ju dosiahnuť bez identifikácie možných faktorov podieľajúcich sa na neúspechu študentov, to si ale vyžaduje ďalší výskum presahujúci rámec tohto článku.

Ďalším cieľom je rozvinutie prepojenosti povinne voliteľných a voliteľných štatistických predmetov (nadväzujúcich na základný kurz) a ekonomickej praxe, čo v konečnom dôsledku môže zvýšiť kvalitu absolventov EF.

Aplikácia kváziperiodických časových radov v kardiológii

[1,2] **Stanislav Katina,** [1] **František Štulajter** a [2] **Eva Kellerová**
[1]Katedra aplikovanej matematiky a štatistiky, Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Bratislava
katina@fmph.uniba.sk
[2]ÚNPF SAV, Sienkiewiczova 1, Bratislava

1. Úvod

Matematická štatistika má široké uplatnenie v biomedicínskych vedách. Jedným z príkladov je dizajn experimentu s opakovanými meraniami v ekvidištantných časových bodoch. Práve vyššie spomenutý dizajn vedie k použitiu kváziperiodických časových radov.

S narastajúcimi možnosťami ambulantného monitorovania krvného tlaku (*TK*, *Ambulatory Blood Pressure Monitoring*) u dospelých sa potvrdzuje, že významnou zložkou intraindividuálnej variability *TK* je jeho periodické kolísanie. V literatúre najskôr a najviac (rôznymi metódami pre analýzu periodických procesov) dokumentovaným biologickým rytmom *TK* je jeho cirkadiánne (*CD*) kolísanie s nočným minimom. Najčastejšou štatistickou metódou jeho dôkazu a amplitúdovo-časovej kvantifikácie je Halbergom zavedená kosinorová analýza (Halberg et al. 1967). Svoje miesto v diagnostike hypertenzie, resp. v chronofarmakológii má hľadanie charakteru cca 24-hodinového periodického výkyvu jeho amplitúdy okolo celodenného priemeru, časovanie maxima a minima a taktiež prípadná neprítomnosť nočného poklesu *TK*.

Konštrukcia vlastného ultrazvukového prístroja na meranie *TK* u novorodencov (Kellerová et al. 1978) nám umožnila prioritný dôkaz existencie *CD* kolísania *TK* už u novorodencov v druhom postnatálnom dni (Kellerová a Kittová 1980, Kellerová 1981). Disperzia jednotlivých hodnôt *TK* okolo odhadu strednej hodnoty s 24 hodinovou periódou (s použitím len kosínusovej funkcie) si vyžiadala podrobnejšiu analýzu na prítomnosť superponovaných ultradiánnych (*UD*) periodicít. Táto ukázala, že kým *CD* rytmus *TK* a pulzovej frekvencie sa vyskytuje u 100% zdravých dospelých, približne po 2 pomalé *UD* rytmy skoro u každého a rýchle UD len asi z 30%, v porovnaní s tým u novorodencov vyšetrovaných v kontrolovanom režime novorodeneckého oddelenia, sa *CD* a pomalé *UD* cykly pozorovali v $60 - 57\%$ a rýchlych *UD* periód bolo približne 4-násobne viac (Kellerová et al. 1989).

Cieľom predkladaného príspevku je ukázať použitie priemerného periodogramu a periodogramu z priemerov sledovanej premennej v jednotlivých časoch na výpočet periód za celý súbor jedincov a naviac použitie Akaikeho informačného kritéria (AIC) pri výbere významných periód.

2. Materiál a metodika

2.1. Súbor jedincov

Súbor jedincov tvorilo $p = 20$ fyziologických novorodencov s postnatálnym vekom 45 ± 11 hodín, pôrodnou hmotnosťou 3551 ± 172 gramov (aritmetický priemer $\pm$ smerodajná odchýlka), vo voľnom režime *"rooming in"* s matkou. Na týchto jedincoch sme merali *TK* v hodinových intervaloch automatickým oscilometrickým tlakomerom *Nippon-Collin*. Z ďalšej analýzy sme vyradili 3, pretože mali jedno alebo viac chýbajúcich pozorovaní, teda $p = 17$.

2.2. Štatistické metódy

Analýza kváziperiodických časových radov je postavená na báze analýzy frekvencií asociovaných s cyklami, ktoré sa v dátach nachádzajú. Najčastejšie používanou množinou matematických funkcií je zmes kosínusov a sínusov.

Majme *nelineárny regresný model* (*NRM,* Štulajter 2002)

$$X(t) = \beta_1 + \beta_2 t + \sum_{j=1}^{k} \left[\beta_{1j} \cos(\lambda_j t) + \beta_{2j} \sin(\lambda_j t)\right] + \varepsilon(t), t \in T, \tag{1}$$

kde $\beta_{full} = (k, \beta, \lambda)^T = (k, \beta_1, \beta_2, \beta_{11}, ..., \beta_{1k}, \beta_{21}, ..., \beta_{2k}, \lambda_1, ..., \lambda_k)^T$ je neznámy vektor regresných parametrov (počet parametrov modelu (1) je $3k + 3 < n$, kde k je potrebné tiež odhadnúť), $\varepsilon(t)$ je *biely šum* (nekorelované náhodné premenné s nulovou strednou hodnotou a s rovnakou disperziou σ^2). Neznáme parametre modelu (1) odhadneme z dát $X(t); t = 1, 2, ..., n$ tak, že najprv vypočítame odhady lineárneho trendu $\widehat{\beta}_1$ a $\widehat{\beta}_2$ a to obyčajnou metódou najmenších štvorcov (*MNŠ*), potom vypočítame $\widehat{X}(t) = X(t) - \widehat{\beta}_1 - \widehat{\beta}_2 t; t = 1, 2, ..., n$ a tieto hodnoty použijeme na výpočet *periodogramu* podľa vzťahu (Christensen 2003)

$$I_n(\lambda) = \frac{1}{2\pi n}\left[\left(\sum_{t=1}^{n} \widehat{X}(t) \cos(\lambda t)\right)^2 + \left(\sum_{t=1}^{n} \widehat{X}(t) \sin(\lambda t)\right)^2\right], \lambda \in \langle 0, \pi\rangle, t \in T. \tag{2}$$

Pozn.: Periodogram sa počíta len v niektorých frekvenciách. Najčastejšie sa používajú, ak n je párne, *Furierove frekvencie* $\lambda_j = \frac{2\pi}{n} j; j = 1, 2, ..., n/2$. Frekvenciám λ zodpovedajú periódy $T_\lambda = \frac{2\pi}{\lambda}$.

Hodnotu k a frekvencie $\widehat{\lambda}_j$, odhady frekvencií $\lambda_j; j = 1, 2, ..., k$, určíme z periodogramu, je to počet všetkých lokálnych maxím funkcie (2), resp. frekvencie, v ktorých sú tieto lokálne maximá. Potom uvažujeme *lineárny regresný model (LRM)*

$$X(t) = \widehat{\beta}_1 + \widehat{\beta}_2 t + \sum_{j=1}^{k} \left[\beta_{1j} \cos\left(\widehat{\lambda}_j t\right) + \beta_{2j} \sin\left(\widehat{\lambda}_j t\right)\right] + \varepsilon(t), t \in T, \tag{3}$$

v ktorom odhadneme neznáme parametre $\beta_{11}, ..., \beta_{1k}, \beta_{21}, ..., \beta_{2k}$ pomocou obyčajnej *MNŠ*.

Model (3) je *aditívnym saturovaným modelom bez interakcií*. Vychádzajúc z tohoto modelu, pomocou *AIC* (Akaike 1974)

$$\begin{aligned} AIC\left(\widehat{\mathbf{F}\beta}, \widehat{\sigma}^2\right) &= -2l\left(\widehat{\mathbf{F}\beta}, \widehat{\sigma}^2\right) - 2k \\ &= -2\begin{bmatrix} -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left(\widehat{\sigma}^{2n}\right) \\ -\frac{1}{2\widehat{\sigma}^2}\left(X - \widehat{\mathbf{F}\beta}\right)^T\left(X - \widehat{\mathbf{F}\beta}\right) \end{bmatrix} - 2k, \end{aligned}$$

kde $l(\cdot)$ je maximum funkcie vierohodnosti, $\widehat{\mathbf{F}\beta}$ je maticový zápis odhadu strednej hodnoty, nájdeme *optimálny submodel* použitím *spätného krokového mechanizmu* (Vanables a Ripley 2002).

Teda iterovaním za predpokladu konvergencie vyraďujeme pomocou slučiek tie regresné parametre, ktoré spôsobujú pokles *AIC*. Iterácie zastanú vtedy, keď už nenastane ďalší pokles *AIC*.

Načrtnutý postup hľadania optimálneho submodelu zhrnieme v nasledovných bodoch:

1. najprv vyberieme všetky lokálne maximá funkcie $I_n(\cdot)$ v (2), ich počet označíme $k < n$,

2. potom frekvencie k nim prislúchajúce použijeme v *LRM* (3), kde regresné parametre odhadneme *MNŠ*,

3. ďalej po spätnej krokovej procedúre ostane v modeli (3) $k_1 \leq 2k$ regresných parametrov, kedy tento model budeme považovať za optimálny submodel.

Majme p jedincov v náhodnom výbere, na ktorých opakovane meriame n-krát nejakú premennú v čase za rovnakých podmienok. Pre každého i-teho jedinca máme realizácie $x_i(t); i = 1, 2, ..., p; t = 1, 2, ..., n$ a pre každého z nich uvažujeme *NRM* (1), periodogram $I_n^{(i)}(\lambda)$ (2) s $\lambda \in \langle 0, \pi \rangle$ a *LRM* (3). Hlavnou úlohou je identifikovať najčastejšie sa vyskytujúce periódy, za predpokladu, že realizácie $x_i(t)$ môžu byť teoreticky u každého jedinca posunuté o inú fázu a naviac výskyt hľadaných periód môže byť rôzny. Preto definujeme priemerný periodogram

$$I_n^{(mean)}(\lambda_j) = \frac{1}{p} \sum_{i=1}^{p} I_{n,i}(\lambda_j), \lambda \in \langle 0, \pi \rangle, j = 1, 2, ..., s. \tag{4}$$

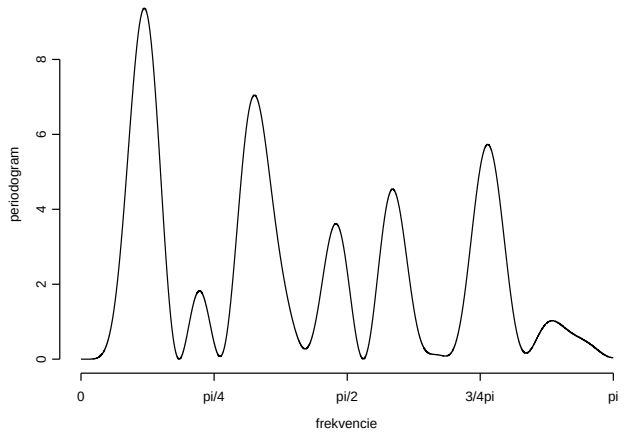
Potom zoradíme lokálne maximá priemerného periodogramu (4) podľa veľkosti a dostaneme nasledovný vektor nim zodpovedajúcich periód $T_{\lambda_j} = s/j, j = 1, 2, ..., s$ v tvare

$$T_{\widehat{\lambda}}^{(1)}, T_{\widehat{\lambda}}^{(2)}, ..., T_{\widehat{\lambda}}^{(s)},$$

kde najčastejšie sa vyskytujúca sa perióda bude $T_{\widehat{\lambda}}^{(s)}$ (v príspevku $s = 60$).

3. Výsledky a diskusia

V našom príklade meriame t v hodinách ($n = 24$), $t = 1$ je pre 7. hodinu ráno, a $t = 24$ je 6. hodina ráno nasledujúceho dňa. Na ilustráciu sme si vybrali časové zmeny diastolického krvného tlaku novorodenca č. 15, ktorého denný priemer bol $\overline{x}^{(15)} = 36.813$ *mm Hg*. Z periodogramu (graf 1) vyberieme všetky lokálne maximá, ktoré spolu s odhadnutým lineárnym trendom vstupujú do iteračných procedúr.



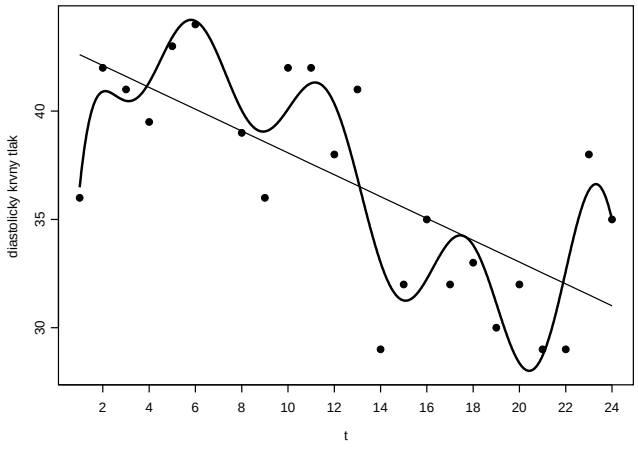
Graf 1: Periodogram (novorodenec č. 15)

Optimálny submodel bude mať potom nasledovný tvar

$$\begin{aligned}\widehat{x}\,(t) &= 43.114 - 0.504t - 91.642\cos(0.117\pi t) - 44.818\cos(0.217\pi t)\\ &\quad +43.593\cos(0.333\pi t) + 53.833\cos(0.483\pi t) + 11.637\cos(0.583\pi t)\\ &\quad +12.471\sin(0.117\pi t) + 94.708\sin(0.217\pi t) + 71.364\sin(0.333\pi t)\\ &\quad -20.745\sin(0.583\pi t) - 7.761\sin(0.767\pi t)\,,\end{aligned}$$

kde sú zahrnuté nasledovné frekvencie $\widehat{\lambda}_{(6)} = 0.117\pi$, $\widehat{\lambda}_{(1)} = 0.217\pi$, $\widehat{\lambda}_{(5)} = 0.333\pi$, $\widehat{\lambda}_{(2)} = 0.483\pi$, $\widehat{\lambda}_{(3)} = 0.583\pi$ a $\widehat{\lambda}_{(4)} = 0.767\pi$ a k nim prislúchajúce periódy $T^{(6)}_{0.117\pi} = 8.571$, $T^{(1)}_{0.217\pi} = 4.615$, $T^{(5)}_{0.333\pi} = 3.000$, $T^{(2)}_{0.483\pi} = 2.069$, $T^{(3)}_{0.583\pi} = 1.714$ a $T^{(4)}_{0.767\pi} = 1.304$ hodiny (tu $k_1 = 10$).

Odhadnutú strednú hodnotu optimálneho submodelu časového radu znázorňujeme ako spojitú funkciu času. Do grafu zakreslujeme aj skutočné (namerané) hodnoty sledovaného parametra v diskrétnych časoch a lineárny trend (graf 2).



Graf 2: Kváziperiodický časový rad diastolického krvného tlaku (novorodenec č. 15), odhadnutá stredná hodnota a lineárny trend

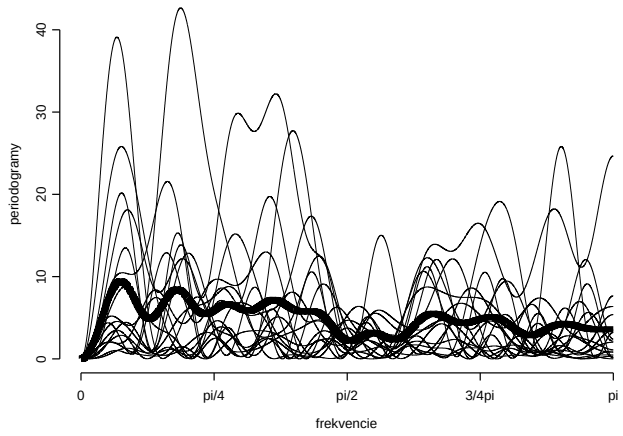
Globálny pohľad na periodické zmeny diastolického krvného tlaku novorodencov v nami sledovanom súbore môžeme vyjadriť pomocou priemerného periodogramu (graf 3).

Z priemerného periodogramu vyplýva, že sa podobá na periodogram bieleho šumu, ktorý je konštantný.

Z periodogramu (graf 4) vypočítaného z dát

$$\overline{x}\,(t) - \overline{x} = \frac{1}{17}\sum_{i=1}^{17} x_i\,(t) - \frac{1}{24}\sum_{t=1}^{24}\overline{x}\,(t)\,;t = 1,2,...,24$$

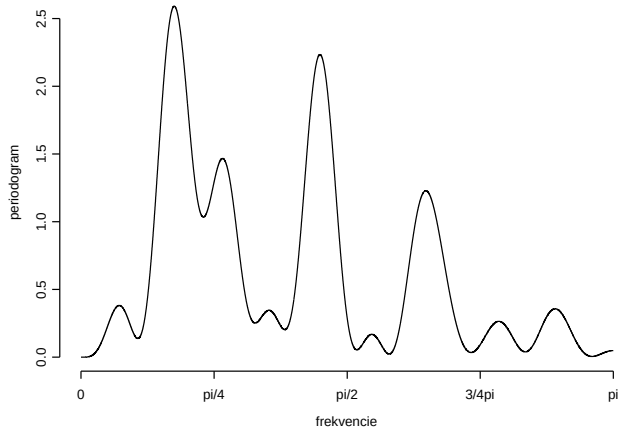
však vyplýva, že v diastolickom krvnom tlaku pozorovanej skupiny detí sú významné periódy.



Graf 3: Periodogramy všetkých $p = 17$ novorodencov (tenšie čiary) a priemerný periodogram (hrubšia čiara)

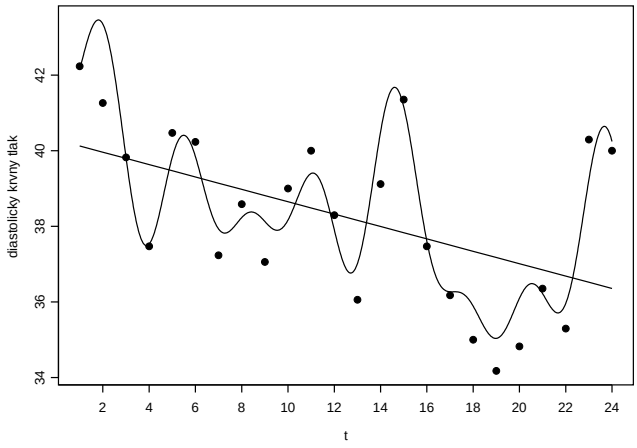
Optimálny submodel bude mať potom nasledovný tvar (graf 5)

$$\begin{aligned}\widehat{x}\,(t) &= 40.290 - 0.164t + 0.935\cos(0.267\pi t) - 0.961\cos(0.450\pi t) \\ &\quad +1.308\sin(0.183\pi t) + 1.003\sin(0.450\pi t) - 0.913\sin(0.650\pi t)\,.\end{aligned}$$



Graf 4: Periodogram z časového radu priemerného diastolického krvného tlaku sledovanej skupiny detí

Teda najvýznamnejšie frekvencie sú $\widehat{\lambda}_{(4)} = 0.183\pi$, $\widehat{\lambda}_{(2)} = 0.267\pi$, $\widehat{\lambda}_{(3)} = 0.450\pi$ a $\widehat{\lambda}_{(1)} = 0.650\pi$ a k nim prislúchajúce periódy $T^{(4)}_{0.183\pi} = 5.455$, $T^{(2)}_{0.267\pi} = 5.455$, $T^{(3)}_{0.450\pi} = 2.222$ a $T^{(1)}_{0.650\pi} = 1.538$ hodiny (tu $k_1 = 5$). U novorodencov, podobne ako aj u dospelých, sa prejavuje nočný pokles krvného tlaku, čo si vyžaduje (podobne ako aj denné kolísanie) ďalšiu štatistickú analýzu.



Graf 5: Kváziperiodický časový rad diastolického krvného tlaku sledovanej skupiny detí, odhadnutá stredná hodnota a lineárny trend

4. Záver

V praxi pri používaní kváziperiodických časových radov, ako *NRM*, odporúčame nepoužívať priemerný periodogram, ale periodogram z priemerov sledovanej premennej v jednotlivých časoch na výpočet periód za celý súbor a naviac *AIC* pri výbere významných periód spätným krokovým mechanizmom z lineárneho aditívneho saturovaného regresného modelu bez interakcií.

Literatúra

1. Akaike, H., 1974: A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **AU-19**: 716 - 722

2. Halberg F., Tong Y.L., Johnson E.A., 1967: Circadian system phase - an aspect of temporal morphology: Procedures and illustrative examples In: *The Cellular Aspects of Biorhythms*, Ed. von Mayersbach, H., Springer, New York, 20-48

3. Christensen, R., 2003: *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data: Nonparametric Regression and Response Surface Maximalization*. Springer, New York

4. Kellerová E., 1981: Physiological responses of blood pressure and heart rate in neonates and infants. *Adv.Physiol.Sci. Cardiovascular physiology.* Eds.: Kovach AGB, et al., Pergamon Press - Akad. Kiado, **9**: 367-375

5. Kellerová E., Kittová M., 1980: Diurnal periodicity of circulatory functions in man. *Activ.nerv.super*, **22** (**1**): 63-64

6. Kellerová E., Kováčik P., Kittová M., 1978: Neinvazívna metóda merania krvného tlaku u novorodencov, na princípe Dopplerovho fenoménu ultrazvuku. *Bratislavské lekárske listy*, **70**: 409-418

7. Kellerová E., Mikulecký M., Kubáček L., Andrasyová D., 1989: Circa- and ultradian blood pressure and heart rate rhythmicity in normal newborns. *Chronobiologia*, **16** (**2**): 150

8. Štulajter, F., 2002: *Prediction in Time Series Using Regression Models.* Springer, New York

9. Vanables, V.N., Ripley, B.D., 2002: *Modern Applied Statistics with S.* Springer, New York

Poďakovanie

Výskum bol podporený projektami 1/0272/03 a 2/3203/25 grantovej agentúry VEGA.

O výučbe štatistiky a jej využití v environmentalistike

Miriam Jadroňová, Zuzana Kimáková
Katedra aplikovanej matematiky SjF TU Košice
miriam.jadronova@tuke.sk, zuzana.kimakova@tuke.sk

1 O výučbe štatistiky

Katedra aplikovanej matematiky Strojníckej fakulty Technickej univerzity v Košiciach zabezpečuje výučbu matematiky na troch fakultách – Strojníckej fakulte, Hutníckej fakulte a Fakulte baníctva, ekológie, riadenia a geotechnológií (FBERG). V prvej časti príspevku sa zameriame na výučbu štatistiky na Strojníckej fakulte, pretože výučbu štatistických predmetov na Hutníckej fakulte a FBERG zabezpečujú finálne katedry materských fakúlt.

1.1 Akademický rok 1970/1971 – 1999/2000

Podľa archivovaných údajov o výučbe matematiky je možné urobiť prehľad o počte hodín venovaných práve výučbe štatistiky. Viac ako 30 rokov sa na našej katedre vyučoval predmet Matematika IV, ktorý bol súčasťou základného kurzu matematiky a bol povinný pre všetkých študentov inžinierskeho štúdia Strojníckej fakulty. Osnova predmetu zahrňovala základy teórie pravdepodobnosti, popisnú štatistiku, teóriu odhadu, testovanie hypotéz, korelačnú a regresnú analýzu. Údaje o rozsahu výučby daného predmetu sú uvedené v tabuľke 1.

Tab. 1: Výučba predmetu Matematika IV od akademického roku 1970/71 do roku 1999/2000 na Strojníckej fakulte TU v Košiciach

Predmet	**Akademický rok (počet hodín v týždni) prednáška / cvičenie**							
	1970/ 1971	**1975/ 1976**	**1980/ 1981**	**1985/ 1986**	**1990/ 1991**	**1995/ 1996**	**1998/ 1999**	**1999/ 2000**
Matematiky IV - **Pravdepodobnosť a matematická štatistika** 2. ročník	3/2 P, s	2/4 P, s	2/3 P, s	2/3 P, s	4/3 P, s	2/2 P, s	2/2 P, kz	2/2 V, kz

(Poznámka: kz – klasifikovaný zápočet, s – skúška, P – povinný predmet, V – voliteľný predmet)

V 4. ročníku prebiehala výučba predmetov Ekonomická štatistika (v odbore Manažment podniku, Ekonomika a riadenie stroj. výroby) a Štatistické metódy (odbor Kvalita produkcie a bezpečnosť technických systémov).

1.2 Akademický rok 2000/2001 – 2004/2005

Prelomovým a pre výučbu štatistiky zvlášť nepriaznivým rokom bol akademický rok 2000/2001. V tomto roku došlo k redukcii počtu hodín vo výučbe matematiky a povinný

predmet Matematika IV v 4. semestri bol nahradený voliteľným predmetom Pravdepodobnosť a matematická štatistika, ktorý končí klasifikovaným zápočtom.

Po roku 2000/2001 ostáva v 4. ročníku denného štúdia predmet - Štatistické metódy, s týždennou hodinovou dotáciou 2 +3, ktorý je povinný len pre jediný zo 16 inžinierskych študijných odborov: Kvalita produkcie a bezpečnosť technických systémov. Tým, že študentom chýbali základy matematickej štatistiky vybudované v predmete Matematika IV, bolo nutné upraviť obsahovú náplň tohto predmetu a zaviesť základný kurz štatistiky. Tabuľka 2 ukazuje prehľad výučby štatistických predmetov a počet študentov, ktorí dané predmety absolvovali od akademického roku 2000/2001.

Tab. 2: Výučba štatistiky od akademického roku 2000/2001

Predmet	**Akademický rok** (počet študentov v predmete/ celkový počet študentov)				
	00/01	**01/02**	**02/03**	**03/04**	**04/05**
Pravdepodobnosť a matematická štatistika 2.ročník DŠ, Ing., SjF, V, 2/2, kz Odbor: bez zamerania	N	5 (250)	7 (315)	7 (316)	6 (237)
Štatistické metódy 4. ročník DŠ, Ing, SjF, P, 2/3, z/s Odbor: Kvalita produkcie a bezpečnosť technických systémov	25	26	30	26	22
Matematická štatistika 2. ročník EŠ, Bc, SjF, P, 2/3, z/s Odbor: Kvalita produkcie a bezpečnosť technických systémov	-	-	-	14	28
Štatistické metódy v environmentalistike 5. ročník EŠ, Ing, SjF, V, 2/3, kz Odbor: Technika ochrany životného prostredia	-	-	-	28	-

(Poznámka: z – zápočet, s – skúška, kz – klasifikovaný zápočet, V – voliteľný predmet, P – povinný predmet, DŠ – denné štúdium, EŠ – externé štúdium, N –predmet sa nenachádza v študijnom programe)

Z uvedených údajov vyplýva, že len hrozivo malá časť absolventov, aj napriek ich technickému zameraniu štúdia, ovláda základy štatistiky (počet absolventov dennej a externej formy štúdia je približne 500 študentov ročne).

Počet tých študentov, ktorí si daný predmet zapísali a predovšetkým aj úspešne absolvovali sa radikálne znížil. Príčinu vidíme vo výbere, z pohľadu študentov, ľahšie zvládnuteľných predmetov (Počítačové konštruovanie, Tabuľkové procesory, Databázy a informačné siete).

Podľa nášho názoru, študent ťažko sám dokáže posúdiť dôležitosť a potrebnosť predmetu pre jeho ďalšie štúdium. Neuvedomuje si, že neznalosť štatistiky a štatistických metód negatívne ovplyvňuje štúdium predmetov jeho odboru a najmä uplatnenie v praxi.

1.3 Od akademického roku 2005/2006

Bakalárske štúdium je už na Slovensku realitou a od akademického roku 2005/2006 všetky slovenské vysoké školy nastupujú na trojstupňové štúdium. Študijné plány bakalárskych študijných programov, ktoré má SjF TU KE akreditované pre nasledujúce obdobie, obsahujú z pohľadu štatistiky tieto predmety: Štatistika pre environmentalistov (2/3, P, z/s, 2.roč., Environmentálne Manažérstvo; 2/2, P, z/s, 2.roč., Technika ochrany životného prostredia), Štatistické metódy (2/2, P, z/s, 2.roč., Priemyselné inžinierstvo), Ekonomická štatistika (2/2, P, z/s, 3. roč., Priemyselné inžinierstvo).

Ak si uvedomíme, že SjF TU KE má akreditovaných 8 bakalárskych študijných programov, tak výučba štatistiky sa týka skoro polovice z nich. Z predbežných údajov o počte uchádzačov o jednotlivé študijné programy vyplýva, že len 35% našich budúcich bakalárov absolvuje počas štúdia niektorý zo štatistických predmetov.

2 Štatistika v environmentalistike

Environmentalistika predstavuje „vednú disciplínu zaoberajúcu sa ochranou a tvorbou životného prostredia“. Vývoj prístupu k ochrane životného prostredia bol výrazne ovplyvnený Konferenciou o životnom prostredí a udržateľnom rozvoji v Rio de Janeiro v roku 1992 a Svetovým summitom OSN, ktorý sa konal v Johannesburgu v roku 2002. V súlade so závermi tejto konferencie je životné prostredie v súčasnosti chápané ako jeden z pilierov trvalo udržateľného rozvoja (tzv. environmentalistického piliera).

V poslednom období vzrástla potreba aplikácie matematických a štatistických metód aj v oblasti environmentálnej výchovy a environmentalistiky. Existuje mnoho aspektov environmentálnych problémov: ekonomické, politické, spoločenské, medicínske, psychologické, technologické a iné, v ktorých je možné využitie štatistiky (koncentrácia toxických látok v pôde, vo vode, emisie plynov a tuhých častíc do ovzdušia, recyklácia odpadov a pod.) Porozumenie týmto problémom a ich riešenie súvisí predovšetkým so získavaním a analyzovaním údajov. Podľa [1] je „analýza údajov čiastočne vedou, čiastočne šikovnosťou a čiastočne umením. Zručnosť a talent pomáha, skúsenosti majú cenu a štatistické nástroje sú nevyhnutné.“

Nevyhnutnosť aplikácie štatistických nástrojov pri analyzovaní štatistických problémov, návrhov na ich riešenie a overovaní účinnosti a kvality ich riešení ilustrujeme na niekoľkých príkladoch z praxe. Uvedené úlohy je možné použiť aj v samotnom vyučovacom procese.

<u>Príklad 1:</u> Emisie skleníkových plynov v SR

„Zmena globálnej klímy, spôsobená antropogénnou [1] emisiou skleníkových plynov je najvýznamnejší environmentálny problém v doterajšej histórii ľudstva.“ (podľa [4]). Kjótsky protokol, ktorý bol prijatý na tretej konferencii strán Rámcového dohovoru v Kjóte v decembri 1997, zosilnil medzinárodnú zodpovednosť za zmenu klímy. Slovenská republika a väčšina krajín strednej a východnej Európy musí znížiť do roku 2008 emisie kľúčových skleníkových plynov (oxid uhličitý CO_2, metán CH_4, oxid dusný N_2O , F – plyny) o 8 % oproti základnému roku 1990.

[1] Antropogénny – vzniknutý ľudskou činnosťou.

Návrh úlohy: Nájdite charakteristiky polohy a rozptylu, nakreslite histogram, polygón pre jednotlivé skleníkové plyny (tab.3).

Tab. 3 Celkové emisie skleníkových plynov v SR

plyny	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
CO_2	59,6	52,5	48,7	45,8	42,9	44,2	44,7	45,0	44,0	43,1	40,6	43,0	42,5
CH_4	6,5	6,0	5,6	5,2	5,1	5,2	5,3	5,0	4,7	4,6	4,5	4,6	4,7
N_2O	6,0	5,2	4,4	3,8	4,0	4,2	4,2	4,2	4,2	3,8	3,8	4,0	3,8
F-plyny	0,3	0,3	0,2	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1

Príklad 2: Emisie z automobilovej dopravy

Rozvoj cestnej dopravy, vrátane individuálneho motorizmu, patrí medzi strategické ciele v rezorte dopravy a hospodárstva SR. No okrem hospodárskeho, úžitkového a ekonomického významu negatívne pôsobí na životné prostredie. Súčasné dopravné prostriedky patria k veľkým spotrebiteľom fosílnych palív a produkované emisie ovplyvňujú zemskú atmosféru.

V roku 1999 bola v rámci diplomovej práce sledovaná vzorka 150 automobilov značky Škoda [2]. Skúmané údaje – rok výroby, počet najazdených kilometrov, obsah uhľovodíkov HC [ppm][2] a obsah oxidov uhlíka CO_x [%] boli získané z protokolu technickej kontroly.

Návrh úlohy: Určte závislosť obsahu uhľovodíkov HC a obsahu oxidov uhlíka CO_x od roku výroby vozidla a od počtu najazdených kilometrov (kvôli veľkému rozsahu údajov sú namerané hodnoty uvedené na www.tuke.sk/jadronova).

Pre porovnanie dosiahnutých výsledkov môže slúžiť zákon NR SR č. 127/94 Z.z. O posudzovaní vplyvov na životné prostredie, z ktorého vyplýva: „V rámci emisných skúšok vozidiel koncentrácia oxidu uhoľnatého CO a nespálených uhľovodíkov HC nesmie prekročiť limity

- 3,5 % CO a 800 ppm HC
- 4,5 % CO a 1200 ppm HC na vozidle vyrobenom pred rokom 1986.“

Príklad 3: Denitrifikácia (znižovanie emisií oxidov dusíka NO_x) v elektrárňach Vojany

Pôsobenie tepelných elektrární na životné prostredie je v súčasnosti celosvetovým problémom ľudstva. Ich vplyv v rámci globálnej biosféry nie je zanedbateľný - produkujú až 60 % všetkých exhalátov, ktoré sa dostávajú priemyselnou činnosťou do atmosféry. Spaľovaním fosílnych palív (uhlie, topný olej, zemný plyn) dochádza v dôsledku chemickej reakcie medzi vzdušným kyslíkom a dusíkom ku vzniku oxidov dusíka NO_x [3].

Návrh úlohy: Určte závislosť výstupnej koncentrácie oxidov dusíka NO_x

- na parnom výkone kotla
- obsahu kyslíka O_2 v spalinách v spaľovacej komore
- na teplote v spaľovacej komore

(kvôli veľkému rozsahu údajov sú namerané hodnoty uvedené na www.tuke.sk/jadronova).

[2] ppm - parts per million = miliónтina = 1 / 1000000

Príklad 4: Koncentrácia toxických látok (olova) v pitnej vode.

Jednou z látok, ktoré môžu ohroziť kvalitu a zdravotnú bezchybnosť vody je olovo, ktoré má toxické účinky na ľudský nervový systém. Chronická otrava olovom môže spôsobiť bolesti hlavy, nepokojnosť, dlhotrvajúcu únavu, problémy s pamäťou, črevné kŕče, depresiu, impotenciu a poškodenie obličiek. V tabuľke 4 sú uvedené namerané hodnoty olova v pitnej vode [μg/L] z vodovodného potrubia.

Tab. 4: Koncentrácia olova v pitnej vode

Pb [μg/L]	0-0,9	1-1,9	2-2,9	3-3,9	4-4,9	5-9,9	10-14,9	15-19,9	20-29,9	30-39,9	40-49,9	50-50,9	60-69,9	70-79,9
počet	20	16	32	11	13	27	7	4	6	1	1	1	0	1

Svetová zdravotnícka organizácia (WMO) určila limit koncentrácie olova 0,01 mg na liter vody. Mnohé krajiny majú ešte stále rozdielne limitné hodnoty, ale od roku 2003 nie je povolená norma viac ako 0,025 mg na liter. Na Slovensku je limitná hodnota 0,01 mg na liter pitnej vody (Z.z. č.491/2002), čím naša krajina už teraz spĺňa limit stanovený WMO 0,01 mg na liter s celosvetovou platnosťou od roku 2023.

Návrh úlohy: Nájdite 90% - ný obojstranný interval spoľahlivosti pre strednú hodnotu. Nájdite charakteristiky polohy a rozptylu.

Príklad 5: Koncentrácia toxických látok (kadmia) v pôde

Kvalita pôdy patrí medzi najvýznamnejšie faktory využívania a rozvoja územia a jej kontaminácia ťažkými kovmi (Cd, Cr, As, Pb, Hg, Ni, Cu) je pretrvávajúci problém. Kadmium je toxický, karcinogénny prvok, prirodzene sa vyskytujúci v pôde. K jeho zvýšenej koncentrácii prispievajú aj fosforečné hnojivá, pričom prirodzený obsah kadmia vo veľkej miere závisí od druhu a typu pôdy, od materskej horniny a intenzity zvetrávania. Limitná hodnota pre koncentráciu kadmia v poľnohospodárskej pôde je stanovená na 0,1 mg na kilogram pôdy (Z.z č.220/2004). Tabuľka 5 obsahuje 39 meraní koncentrácie kadmia [mg/kg] v pôde.

Tab. 5: Výskyt kadmia v pôde (zdroj údajov [1])

0,023	0,005	0,005	0,020	0,005	0,032	0,010	0,005	0,031	0,020	0,013	0,005	0,020
0,005	0,014	0,020	0,094	0,020	0,010	0,011	0,005	0,010	0,005	0,027	0,010	0,005
0,015	0,010	0,005	0,015	0,010	0,028	0,034	0,010	0,010	0,005	0,005	0,018	0,013

Príklad 6: Rizikové faktory v životnom prostredí – havarijné zhoršenie kvality vody

Na mimoriadnom zhoršení alebo ohrození kvality vody (MOV) tak povrchových, ako aj podzemných, sa predovšetkým podieľajú ropné látky, žieraviny, priemyselné hnojivá, odpadové látky a pod. Vývoj v počte MOV podľa druhu látok škodiacim vodám je prezentovaný v tab.6.

Tab. 6: Vývoj v počte MOV podľa druhu látok v rokoch 1993 – 2003 (zdroj údajov [5])

Druh látok škodiacich vodám	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**
Ropné látky	70	63	76	69	50	61	54	33	40	64	59
Žieraviny	5	3	3	5	10	3	5	2	2	5	3
Pesticídy	2	1	0	1	1	3	1	0	0	1	0
Exkrementy hospod. zvierat	8	9	11	14	8	3	7	5	4	9	21
Silážne šťavy	0	0	0	1	1	0	2	4	0	2	1
Priemyselné hnojivá	0	0	1	0	0	0	0	0	0	0	1
Iné toxické látky	5	5	5	1	5	0	6	12	5	3	3
Nerozpustné látky	11	4	6	4	8	7	1	5	2	6	11
Odpadové vody	8	6	1	6	11	17	6	10	10	17	35
Iné látky	4	13	10	9	6	6	4	2	1	3	7
Látky škodiace vodám, u ktorých sa šetrením nepodarilo zistiť druh	29	17	16	7	9	17	12	9	7	17	35

Návrh úlohy: Určte závislosť počtu mimoriadnych zhoršení a ohrození kvality vody od druhu látok škodiacich vodám.

3 Záver

Výučba štatistiky by mala čo najlepšie odpovedať potrebám katedier garantujúcich jednotlivé študijné programy a študenti by mali byť vedení k tomu, aby systematicky využívali získané matematické vedomosti v odborných predmetoch. Zároveň je veľmi dôležité a nevyhnutné prispôsobiť výber príkladov tematicky jednotlivým študijným programom. V tomto prípade veľkým zdrojom a inšpiráciou môžu byť správy a štatistické ročenky o stave životného prostredia alebo i záverečné práce študentov.

Jedným zo spôsobov skvalitnenia a zatraktívnenia výučby štatistiky, rovnako ako zintenzívnenia spolupráce s finálnymi katedrami v oblasti výskumu a riešenia úloh z praxe sa javí zaradenie štatistického softvéru do vyučovacieho procesu. Pre splnenie tohoto cieľa bude na Katedre aplikovanej matematiky SjF TU KE od budúceho semestra slúžiť počítačové laboratórium vybavené vhodným softvérovým produktom.

Literatúra

[1] Berhouex, P. M. – Brown, L. C.: Statistics for Environmental Engineers. Lewis Publishers, 2002.

[2] Somráky, F.: Optimalizácia emisných limitov z automobilovej dopravy. Diplomová práca 1999, TU SjF Košice.

[3] Steranková, A. – Optimalizácia procesov denitrifikácie v EVO Vojany. Diplomová práca 2005, TU SjF Košice.

[4] Správa o kvalite ovzdušia a podiele jednotlivých zdrojov na jeho znečistení v SR – 2003, www.sazp.sk/slovak/periodika/sprava/sprava2003/kapitoly/svk2003s_ovzd.pdf

[5] Správa o stave životného prostredia Slovenskej republiky v roku 2003, http://enviroportal.sk/spravy-zp/sprava-detail.php?stav=29

[6] www.enviro.gov.sk/servlets/page/166

Testovanie spoločných trendov v prietokoch slovenských riek

Danuša Szökeová
Magda Komorníková
Stavebná fakulta, Slovenská technická univerzita
Bratislava

Pri štúdiu rôznych časových radov získaných pozorovaním sa zistilo, že tieto časové rady obsahujú lineárny deterministický trend. V tejto práci prezentujeme výsledky testovania spoločných lineárnych trendov v prietokoch slovenských riek na základe priemerných mesačných prietokov, ktoré boli namerané v rokoch 1931 až 2001. Využili sme pritom tri štatistické testy. Spoločné trendy sme testovali na území celého Slovenska ako aj v rámci piatich regiónov, do ktorých sme zaradili pozorovacie stanice podľa základných povodí (Dolné a Horné Považie, Ponitrie, Pohronie, Bodrog - Hornád). Štatistické testy sme aplikovali na celkové časové rady ako aj na dvanásť časových radov, ktoré vznikli rozdelením podľa mesiacov. Vyhodnotením výsledkov sa ukázalo, že existujú regióny, v ktorých prietoky majú spoločný trend a regióny, v ktorých prietoky nemajú spoločný trend. Tieto trendy sa vyskytujú v celkových časových radoch ako aj mesačných radoch.
Kľúčové slová: mesačné priemerné prietoky, viacrozmerné deterministické trendy, testovanie hypotéz, F-test, Waldov test.

1. Úvod

Vplyv klimatických zmien za posledné desaťročia sa prejavuje aj zmenami v množstve zrážok a v objeme prietokov na vodných tokoch. Namerané údaje však samé osebe nevypovedajú o trendoch a vzájomných súvislostiach medzi tokmi v jednotlivých regiónoch. Preto je potrebné ich štatistické spracovanie a odborná interpretácia získaných výsledkov. Zaujíma nás, či dva alebo viac vodných tokov má spoločný stúpajúci alebo klesajúci trend a ako tieto spoločné trendy závisia od regionálneho a časového členenia.

V druhej a tretej kapitole popisujeme modely, testovacie štatistiky a testované časové rady. Záverečná kapitola obsahuje podrobné zhodnotenie výsledkov testovania z hľadiska hydrológie, t.j. v ktorých regiónoch a v ktorých časových obdobiach v rámci územia Slovenska majú prietoky spoločný trend a v ktorých nemajú.

2. Model a testovacie štatistiky

V tejto kapitole popíšeme model, odhad parametrov a použité testovacie štatistiky. Uvažujme m trendovo stacionárnych časových radov $y_{1,t}$, ..., $y_{m,t}$, t = 1, ..., n. Predpokladajme, že ich môžeme vyjadriť v tvare:

$$y_{1,t} = \mu_1 + \beta_1\, t + u_{1,t}$$

$$y_{2,t} = \mu_2 + \beta_2\, t + u_{2,t}$$

…

$$y_{m,t} = \mu_m + \beta_m\, t + u_{m,t}\ ,$$

čo sa dá vo vektorovom tvare zapísať

$$\mathbf{Y}_t = \boldsymbol{\mu} + \boldsymbol{\beta}\, t + \mathbf{U}_t$$

kde $\mathbf{Y}_t=(y_{1,t}, y_{2,t}, \ldots, y_{m,t})'$, $\mathbf{U}_t=(u_{1,t}, u_{2,t},\ldots, u_{m,t})'$, $\boldsymbol{\beta}=(\beta_1, \beta_2,\ldots, \beta_m)'$, $\boldsymbol{\mu}=(\mu_1, \mu_2,\ldots, \mu_m)'$.

Odhad $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\beta}}$ vektorov $\boldsymbol{\mu}$ a $\boldsymbol{\beta}$ získame klasickou metódou najmenších štvorcov. Hodnota m je závislá od počtu meracích staníc v regióne. Predpokladajme ďalej, že

$$\frac{1}{n}\sum_{t=1}^{[\tau n]}\mathbf{U}_t \Rightarrow \Lambda W_m(\tau)$$

kde $\Rightarrow$ označuje slabú konvergenciu, $W_m(\tau)$ je m x 1 rozmerný štandardný Wienerov proces a $[\tau\, n]$ je celá časť τ n. Označme Ω kovariančnú maticu $\mathbf{U}_t$, t. j.

$$\Omega = \Lambda\,\Lambda' = \sum_{j=-\infty}^{\infty}\Gamma_j\ ,\ \text{kde}\ \ \Gamma_j\ = \mathrm{Cov}(\mathbf{U}_t\ \mathbf{U}'_{t-j})\ .$$

Hypotéza, ktorú testujeme je:

$H_0 : R\ \boldsymbol{\beta} = \mathbf{r}$ $\qquad\qquad$ $H_1 : R\ \boldsymbol{\beta} \neq \mathbf{r}$,

kde R je matica typu q x m (zložená z prvkov 0 a 1), **r** je vektor typu q x 1, (q = m - 1). Zamietnutie nulovej hypotézy znamená, že neexistuje spoločný lineárny deterministický trend v skúmaných časových radoch.

Na odhad kovariančnej matice Ω použijeme tri rôzne konzistentné odhady (podrobnejšie napr. v [1, 2, 4]):

$$\hat{\Omega}_{HAC} = \hat{\Gamma}_0 + \sum_{j=1}^{n-1}\left(1-\frac{j}{\sqrt[3]{n}}\right)\left(\hat{\Gamma}_j + \hat{\Gamma}_j'\ \right),\qquad \hat{\Gamma}_j = \frac{1}{n}\sum_{t=j+1}^{n}\hat{\mathbf{U}}_t\ \hat{\mathbf{U}}'_{t-j}$$

$$\hat{\Omega}_n = \hat{\Gamma}_0 + \sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right)$$

$$\tilde{\Omega}_n = \tilde{\Gamma}_0 + \sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\left(\tilde{\Gamma}_j + \tilde{\Gamma}_j'\ \right),\quad \tilde{\Gamma}_j = \frac{1}{n}\sum_{t=j+1}^{n}\left[(t-\bar{t})\ \hat{\mathbf{U}}_t\right]\left[(t-j-\bar{t})\ \hat{\mathbf{U}}'_{t-j}\right],\quad \bar{t} = \frac{1}{n}\sum_{t=1}^{n} t\,.$$

Uvažujeme tri testovacie štatistiky vzhľadom na uvedené odhady Ω. Prvé dve štatistiky sú založené na **F - teste** (pre q > 1) :

$$F_1^* = n\left(R\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left[R \frac{n}{\sum_{t=1}^{n}\left(t-\bar{t}\right)^2} \tilde{\Omega}_n \frac{n}{\sum_{t=1}^{n}\left(t-\bar{t}\right)^2} R' \right]^{-1} \frac{\left(R\hat{\boldsymbol{\beta}} - \mathbf{r}\right)}{q}$$

$$F_2^* = \left(R\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left[R \frac{1}{\sum_{t=1}^{n}\left(t-\bar{t}\right)^2} \hat{\Omega}_n R' \right]^{-1} \frac{\left(R\hat{\boldsymbol{\beta}} - \mathbf{r}\right)}{q}$$

Tretím testom je **Waldov test**:

$$W_{HAC} = \left(R\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left[R \frac{1}{\sum_{t=1}^{n}\left(t-\bar{t}\right)^2} \hat{\Omega}_{HAC} R^T \right]^{-1} \left(R\hat{\boldsymbol{\beta}} - \mathbf{r}\right) \quad .$$

Podrobnú teóriu týkajúcu sa uvedených štatistických testov ako aj kritické hodnoty pre tieto testy možno nájsť v článku [4].

3. Dáta

Pri testovaní sme použili údaje o priemerných mesačných prietokoch získané z dlhodobých meraní v priebehu rokov 1931 až 2001 na 29 pozorovacích staniciach. Z každej pozorovacej stanice máme k dispozícii 852 údajov.

Na základe lineárnej regresie uvedených časových radov sa zistilo, že na niektorých tokoch sa v uvedenom období prejavuje stúpajúci, resp. klesajúci trend objemu vodných prietokov. Sledované toky sme rozdelili do nasledujúcich piatich skupín (ako je zrejmé z tabuľky 1).

Región	Rieka (Stanica)	Počet staníc
Horné Považie	Belá (Podbanské), Revúca (Podsuchá), Váh (Lipt. Mikuláš), Biely Váh (Východná), Boca (Kráľ.Lehota)	5
Dolné Považie	Kysuca (Čadca), Kysuca (K.N.Mesto), Ľubochnianka (Ľubochňa), Rajčianka (Poluvsie), Turiec (Martin),	5
Ponitrie	Nitra (N.Streda), Nitra (Chalmová), Bebrava (Biskupice)	3
Pohronie	Vajskovský p. (D.Lehota), Č.Hron (Hronec), Hron (Brehy), Hron (B.Bystrica), Štiavnička (Mýto p.Ďumb), Rimavica (Kokava.n.Rimavicou), Ipeľ (Holiša), Lietava (Plášťovce), Krupinica (Plášťovce), Štítnik (Štítnik), Dobšinský p. (Dobšiná)	11
Bogrog, Hornád, Poprad	Uh (Lekárovce), Topľa (Hanušovce), Torysa (Koš.Olšany), Poprad (Chmelnica), Poprad (Matejovce)	5

Tabuľka 1: Rozdelenie riek a pozorovacích staníc do regiónov podľa povodí

4. Výsledky

Testy popísané v kapitole 2 sme použili na analýzu celkových a mesačných spoločných trendov v prietokoch slovenských riek. Výpočty sme realizovali pomocou výpočtového systému *Mathematica*.

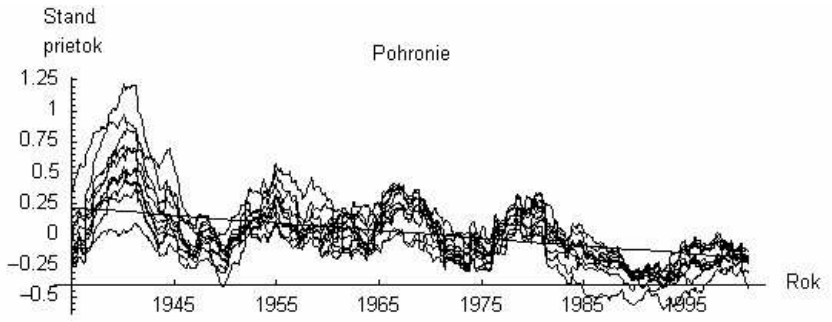
Pred testovaním boli všetky časové rady upravené do štandardizovaného tvaru podľa vzťahu: $z_t = \frac{Y_t - \overline{Y}}{\sigma_y}$, kde $\overline{Y}$ je stredná hodnota, σ_y smerodajná odchýlka.

4.1 Celkové trendy

Najskôr nás zaujímalo, či celkové rady, t.j rady obsahujúce 852 hodnôt priemerných mesačných prietokov, majú vzhľadom na kritické hodnoty spoločný trend. Hodnoty vyčíslených štatistík testov F1 a F2 sme porovnávali s kritickými hodnotami pre α=0.05.

Na základe výsledkov uvedených v tabuľke 2 pre celé obdobie 1931 až 2001 možno povedať, že :

1. v rámci celého územia Slovenska ani jeden z testov nepotvrdil existenciu spoločného trendu,
2. v regiónoch Horné a Dolné Považie sa nepotvrdila existencia spoločného trendu,
3. v regiónoch Ponitrie, Pohronie a Bodrog-Hornád možno predpokladať mierne klesajúci trend vyjadrený zápornou hodnotou smernice β (v tabuľke 2 sú hodnoty spoločných trendov označené hviezdičkou). Najväčší záporný celkový trend bol zistený v regióne Pohronie.



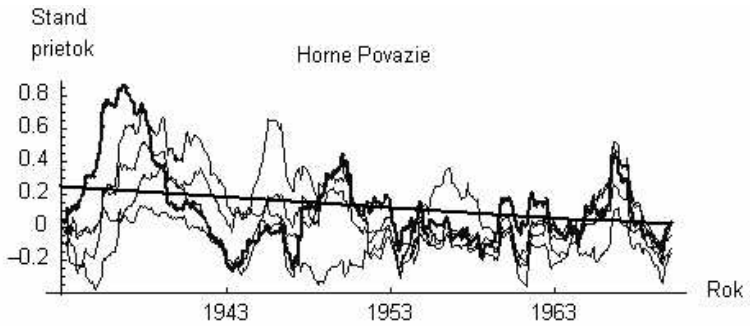
Obrázok 1: Štandardizované prietoky upravené kĺzavými priemermi a spoločný trend v regióne Pohronie, obdobie 1931–2001

Všetky tri štatistiky sme aplikovali aj na ďalšie rady, ktoré vznikli rozdelením pôvodných radov na kratšie časové obdobia. V niektorých prípadoch testy potvrdili

existenciu spoločného trendu (rastúceho alebo klesajúceho), aj keď celkový časový rad sa spoločným trendom nevyznačuje. Pretože obdobia, v ktorých sa striedajú rastúce a klesajúce trendy prietokov nie sú pravidelné, nezistili sme ich pri vyšetrovaní cyklickej zložky časových radov v rámci spektrálnej analýzy.

V tabuľke 3 sú výsledky testovania obdobia 1931 až 1970 (prvých 40 rokov). Na základe vypočítaných štatistík možno povedať, že:

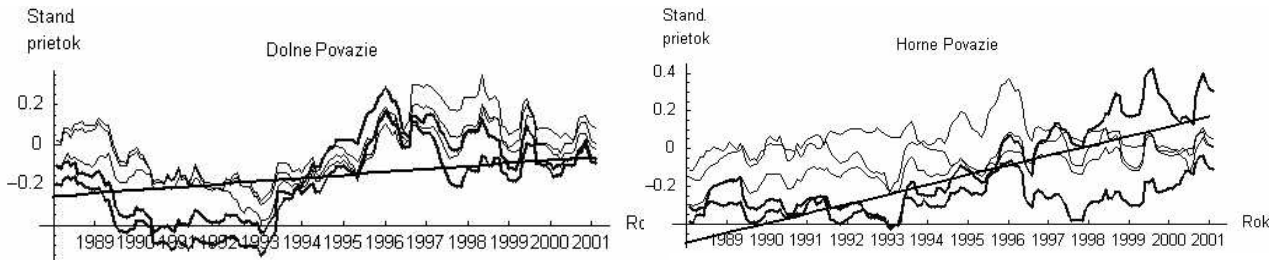
1. v rámci celého územia Slovenska ani jeden z testov nepotvrdil existenciu spoločného trendu,
2. v regióne Horné Považie a Bodrog-Hornád možno predpokladať existencia klesajúceho trendu,
3. v regióne Dolné Považie a Ponitrie možno predpokladať existencia stúpajúceho trendu,
4. v regióne Pohronie sa nepotvrdila existencia spoločného trendu.

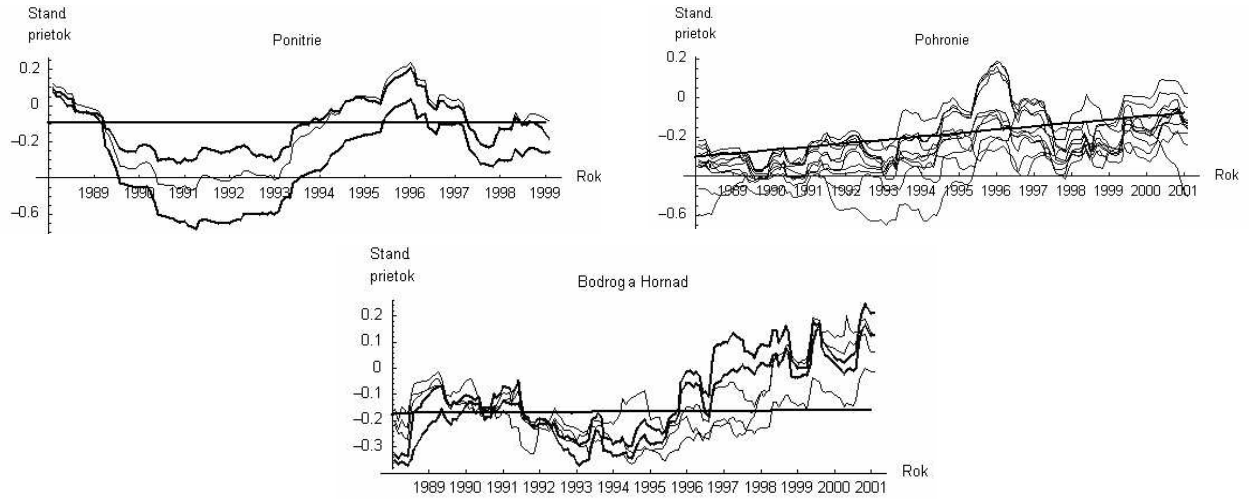


Obrázok 2: Štandardizované prietoky upravené kĺzavými priemermi a spoločný trend v regióne Horné Považie, obdobie 1931–1970

V tabuľke 4 sú výsledky testovania prietokov z obdobia 1986 až 2001 (posledných 16 rokov):

1. v rámci celého územia Slovenska ani jeden z testov nepotvrdil existenciu spoločného globálneho trendu,
2. v regiónoch Horné, Dolné Považie a Pohronie existuje stúpajúci trend,
3. v regiónoch Ponitrie, Bodrog-Hornád sú hodnoty prietokov za uvedené obdobie takmer konštantné.



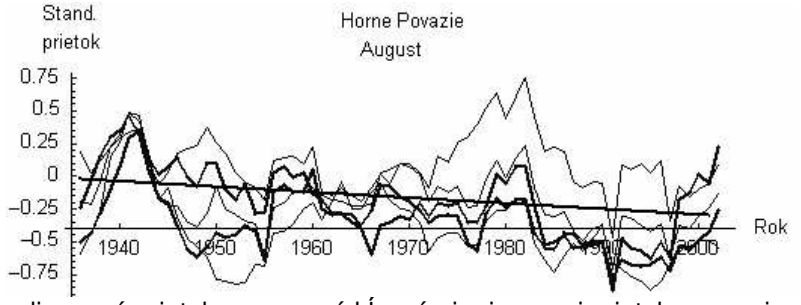


Obrázok 3: Štandardizované prietoky upravené kĺzavými priemermi a spoločný trend vo všetkých regiónoch, obdobie 1986–2001

4.2 Mesačné trendy

Hodnoty priemerných mesačných prietokov sme ďalej rozdelili do časových radov, ktoré odpovedajú jednotlivým mesiacom roka, čím vzniklo dvanásť radov po 71 údajov. Hodnoty vypočítaných štatistík na základe mesačných radov pre jednotlivé regióny sú uvedené v tabuľkách 5 až 10. V rámci celého územia Slovenska ani jeden z testov opäť nepotvrdil existenciu spoločného trendu v žiadnom mesiaci roka (tabuľka 5). V jednotlivých regiónoch možno predpokladať existenciu nasledujúcich trendov:

1. Horné Považie klesajúci trend v mesiacoch apríl, júl, august, september, október (tabuľka 6),
2. Dolné Považie klesajúci trend v mesiacoch február, august, september, október, november (tabuľka 7),
3. Ponitrie mierne klesajúci trendu v mesiacoch január, február, marec, jún, júl, november, december (tabuľka 8),
4. Pohronie klesajúci trendu v mesiacoch apríl (tabuľka 9),



Obrázok 4: Štandardizované prietoky upravené kĺzavými priemermi prietokov, mesiac august, a spoločný trend v regióne Horné Považie

5. Bodrog, Hornád, Poprad stúpajúci trend v mesiaci júl, v ostatných mesiacoch okrem novembra klesajúci trend (tabuľka 10).

Najväčšie hodnoty klesajúcich trendov boli zistené v regióne Považie v mesiaci august, v regióne Ponitrie v mesiaci marec, v Pohroní v mesiaci apríl a na východe Slovenska v mesiaci marec.

Testovali sme aj hypotézu existencie nulového trendu, ktorá sa však v žiadnom regióne nepotvrdila.

5. Záver

Aplikácia štatistických testov na hydrologické údaje potvrdila, že sa tieto testy dajú využiť pri overovaní hypotéz týkajúcich sa globálnych a lokálnych zmien v prietokoch a pri skúmaní týchto zmien podľa ročných sezón. Pochopenie a odhalenie príčin takýchto trendov môže byť zaujímavým obsahom ďalšieho skúmania.

Región	Počet tokov	q	Smernica trendu β	F1		F2		Waldov test	
				Štatistika	Kritická	Štatistika	Kritická	Štatistika	p-value
1.	5	4	–0,00033	80,98	**46,75**	66,71	**43,84**	2,97	0,56
2.	5	4	0,00007	52,75	**46,75**	46,95	**43,84**	2,09	0,71
3.	3	2	–0,0002*	4,94	**38,10**	12,33	**40,68**	0,28	0,87
4.	11	10	–0,00051*	39,48	**70,99**	30,71	**58,90**	3,42	0,97
5.	5	4	–0,00017*	3,39	**46,75**	5,96	**43,84**	0,27	0.99
Všetky	29	28	–0,00038	667,18	**136,3**	666,27	**106,8**	182,14	10^{-15}

Tabuľka 2: Hodnoty štatistík pre ročné trendy podľa regiónov, obdobie 1931–2001

Región	Počet tokov	q	Smernica trendu β	F1		F2		Waldov test	
				Štatistika	Kritická	Štatistika	Kritická	Štatistika	p-value
1.	5	4	–0,00054*	12,82	**46,75**	11,65	**43,84**	0,76	0,94
2.	5	4	0,00039*	3,48	**46,75**	4,63	**43,84**	0,30	0,99
3.	3	2	0,0001*	0,16	**38,10**	0,35	**40,68**	0,01	0,99
4.	11	10	–0,00033	68,09	**70,99**	68,60	**58,90**	11,19	0,34
5.	5	4	–0,00051*	19,79	**46,75**	33,25	**43,84**	2,57	0,63
Všetky	29	28	–0,00016	413,21	**136,3**	305,35	**106,8**	139,47	10^{-15}

Tabuľka 3: Hodnoty štatistík pre celkové trendy podľa regiónov, obdobie 1931 – 1970

Región	Počet tokov	q	Smernica Trendu β	F1		F2		Waldov test	
				Štatistika	Kritická	Štatistika	Kritická	Štatistika	p-value
1.	5	4	0,00109*	18,89	**46,75**	28,15	**43,84**	3,38	0,50
2.	5	4	0,00125*	1,85	**46,75**	1,10	**43,84**	0,13	0,99
3.	3	2	-10^{-7}	1,88	**38,10**	2,35	**40,68**	0,16	0,93
4.	11	10	0,00145*	24,24	**70,99**	19,43	**58,90**	5,84	0,83
5.	5	4	10^{-5}	16,69	**46,75**	23,08	**43,84**	2,77	0,60
Všetky	29	28	0,00084	284,45	**136,3**	286,69	**106,8**	241,20	10^{-15}

Tabuľka 4: Hodnoty štatistík pre ročné trendy podľa regiónov, obdobie 1986–2001

Celé Slovensko					
Počet riek: 29 q = 28	**Smernica trendu β**	**F1**	**F2**	**Waldov test**	
Mesiac		**Kritická: 136,3 Štatistika**	**Kritická: 106,8 Štatistika**	**Štatistika**	**p-value**
Január	−0,00021	447,50	310,73	489,30	10^{-14}
Február	0,0068	425,87	265,77	418,50	10^{-15}
Marec	−0,00107	253,77	232,44	366,01	10^{-15}
Apríl	−0,00022	297,03	193,66	304,95	10^{-15}
Máj	0,0089	231,30	172,48	27161	10^{-15}
Jún	0,0026	512,88	312,34	491,83	10^{-15}
Júl	−0,00017	450,75	349,88	550,95	10^{-15}
August	−0,00222	507,21	375,56	591,36	10^{-15}
September	0,00107	457,13	337,04	530,73	10^{-14}
Október	−0,00005	764,47	452,60	712,69	10^{-15}
November	−0,00495	733,88	631,11	993,79	10^{-14}
December	−0,00273	340,38	256,61	404,08	10^{-16}

Tabuľka 5: Hodnoty štatistík pre mesačné trendy, celé Slovensko, obdobie 1931-2001

Horné Považie					
Počet riek: 5 q = 4	**Smernica trendu β**	**F1**	**F2**	**Waldov test**	
Mesiac		**Kritická: 46,75 Štatistika**	**Kritická: 43,84 Štatistika**	**Štatistika**	**p-value**
Január	−0,0058	114,54	141,99	33,13	10^{-6}
Február	−0,0067	64,05	71,19	16,61	0,002
Marec	−0,001	60,55	89,73	20,93	10^{-9}
Apríl	−0,00022*	9,85	9,01	2,10	0,72
Máj	0,0089	16,88	26,00	6,07	0,19
Jún	0,0026	23,30	31,43	7,33	0,12
Júl	−0,0027*	7,88	8,81	2,06	0,73
August	−0,0043*	5,67	5,24	1,22	0,87
September	−0,0036*	7,01	8,20	1,91	0,75
Október	−0,0001*	5,49	4,29	1,00	0,91
November	−0,0107	179,26	83,37	19,45	10^{-5}
December	−0,0059	49,91	68,61	16,00	0,003

Tabuľka 6: Hodnoty štatistík pre mesačné trendy v regióne Horné Považie, obdobie 1931-2001

Región Dolné Považie					
Počet riek: 5 q = 4	**Smernica trendu β**	**F1**	**F2**	**Waldov test**	
Mesiac		**Kritická: 46,75 Štatistika**	**Kritická: 43,84 Štatistika**	**Štatistika**	**p-value**
Január	0,00688	107,89	86,26	20,12	10^{-4}
Február	−0,0025*	6,58	4,95	1,15	0,88
Marec	0,0025	51,99	61,41	14,32	10^{-4}
Apríl	−0,0011	41,32	36,51	8,51	0,07
Máj	0,0025	41,02	38,42	8,96	0,06
Jún	0,0006	105,73	95,56	22,29	10^{-4}

Júl	0,0068	50,16	41,54	9,69	0,04
August	–0,0056*	1,84	4,06	0,95	0,92
September	–0,00079*	5,65	2,90	0,68	0,95
Október	–0,0076*	6,33	9,18	2,14	0,71
November	–0,0037*	10,74	11,34	2,65	0,62
December	0,0027	51,19	58,23	13,58	0,01

Tabuľka 7: Hodnoty štatistík pre mesačné trendy v regióne dolné Považie, obdobie 1931-2001

Región Ponitrie					
Počet riek: 3 q = 2	**Smernica trendu β**	**F1**	**F2**	**Waldov test**	
Mesiac		**Kritická: 38,10 Štatistika**	**Kritická: 40,68 Štatistika**	**Štatistika**	**p-value**
Január	–0,0014*	0,28	0,55	0,06	0,97
Február	–0,0023*	0,66	1,52	0,18	0,92
Marec	–0,0176*	3,37	8,21	0,96	0,62
Apríl	–0,00131	26,12	29,40	3,43	0,18
Máj	0,00	1,23	2,50	0,30	0,86
Jún	0,00157*	2,80	3,88	0,45	0,79
Júl	–0,0038*	4,17	5,65	0,66	0,72
August	–0,0032	41,65	54,94	6,41	0,04
September	0,0004	27,50	28,88	3,37	0,18
Október	–0,00037	29,32	47,21	5,51	0,06
November	–0,00123*	2,79	7,15	0,83	0,66
December	–0,00687*	1,79	3,37	0,39	0,82

Tabuľka 8: Hodnoty štatistiky pre mesačné trendy v regióne Ponitrie, obdobie 1931-2001

Región Pohronie					
Počet riek: 11 q = 10	**Smernica trendu β**	**F1**	**F2**	**Waldov test**	
Mesiac		**Kritická: 70,99 α = 0.05**	**Kritická: 58,90 α = 0.05**	**Štatistika**	**p-value**
Január	–0,00915	88,41	72,13	42,06	10^{-4}
Február	–0,00183	144,21	62,27	38,31	10^{-6}
Marec	–0,0018	101,07	64,32	37,51	10^{-6}
Apríl	–0,0080*	14,61	11,61	6,77	0,75
Máj	–0,0039	45,53	38,72	22,58	0,01
Jún	0,0015	137,20	98,44	57,41	0,001
Júl	–0,0031	80,78	63,27	30,89	10^{-6}
August	–0,0017	93,08	49,07	28,62	0,001
September	–0,0082	156,47	102,89	60,00	10^{-9}
Október	0,00065	105,16	106,85	62,31	10^{-9}
November	–0,0016	90,90	98,13	57,23	10^{-9}
December	–0,017	182,16	107,53	62,71	10^{-9}

Tabuľka 9: Hodnoty štatistík pre mesačné trendy v regióne Pohronie, obdobie1931-2001

Región Bodrog, Hornád, Poprad					
Počet riek: 5 q = 4	Smernica trendu β	F1	F2	Waldov test	
Mesiac		Kritická: 46,75 Štatistika	Kritická: 43,84 Štatistika	Štatistika	p-value
Január	–0,0034*	3,49	3,27	0,76	0,94
Február	–0,0039*	4,04	5,49	1,28	0,86
Marec	–0,0116*	16,15	15,26	3,56	0,47
Apríl	–0,0083*	6,70	5,38	1,26	0,86
Máj	–0,0031*	7,65	8,41	1,96	0,74
Jún	–0,0020*	12,13	11,82	2,76	0,60
Júl	0,0043*	5,77	9,51	2,22	0,70
August	–0,0031*	5,38	8,07	1,88	0,76
September	–0,0023*	2,19	2,78	0,65	0,96
Október	–0,0029*	1,18	1,08	0,25	0,99
November	–0,0081	56,45	65,29	15,23	0,0042
December	–0,0047*	1,12	1,47	0,34	0,99

Tabuľka 10: Hodnoty štatistík pre mesačné trendy v regióne Bodrog, Hornád, Poprad, 1931-2001

Literatúra:

[1] Andrews, D. W. K. (1991): Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estomation. *Econometrica* 59, 817 – 854Fomby, B.T. and Vogelsand, T.J. (2003): Test of Common Deterministic Trend Slopes Applied to Quartely Global Temperature Data. *Working Papers*, Department of Economics Southern Methodist University DallasKiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2000): Simple Robust Testing of Regression Hypotheses. *Econometrica* 68, 695 – 714

[2] Newey, W. K. and West, K. D. (1987): A Simple Positive Semi-Definite, heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55, 703 - 708

[3] Pekárová,P.,Miklánek,P.: Abflusstrends slowakischer Flüsse und mögliche Zusammenhänge mit ENSO/NAO - Esscheinungen. In Östereichische Wasser und Abfallwirtschaft, Springer, 2004, 1-2, s. 17-25

[4] Vogelsand, T.J., Franses, P.H. (2001) : Testing for Common Deterministic Trend Slopes. Erasmus University Rotterdam, Econometric Institute in its series *Econometric Institute Report* with number 224

Tento článok bol vypracovaný s podporou projektu APVT-20-003204 a grantov VEGA 1/1145/04 a 1/2032/05.

Nový studijní obor na PřF UP v Olomouci
Statistické a počítačové modelování

Pavla Kunderová
Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta UP, Olomouc, Tomkova 40
e-mail: kunderov@inf.upol.cz

Katedra matematické analýzy a aplikací matematiky na Přírodovědecké fakultě UP garantuje výuku dvou studijních programů: Matematika a Aplikovaná matematika. V rámci druhého programu se již od roku 1996 vyuv cuje bakalářský obor „Matematika–ekonomie se zaměřením na bankovnictví“. Zavedení tohoto studia bylo iniciováno tehdejším ředitelem pobov cky Komerv cní banky v Olomouci, který poukazoval na to, v ze v regionu stv rední Moravy chybí nabídka vysokoškolského vzdělání pro pracovníky v bankovní sféře. Učební program studia byl sestaven ve spolupráci s pracovníky Komerční banky, kteří také několik let zajišťovali výuku ekonomických předmětů. O studium je stále zájem, hlásí se řádově stovky studentů, do prvního ročníku je přijímáno vždy 60–90 studentů, v letošním roce bude počet přijatých nejméně dvakrát vyšší. Absolventi tohoto oboru mohou pokračovat v navazujícím dvouletém studiu „Aplikace matematiky v ekonomii“.

Zdálo by se tedy, že katedra nemá důvod nic měnit. Bohužel o studium „čisté matematiky – rozuměj neučitelské“ v programech Matematika a Aplikovaná matematika zájem stále klesá, v některých skupinách se již jedná o téměř individuální vyučování. Proto katedra připravila dva nové studijní obory, které jsou v současnosti posuzovány akreditační komisí. Jsou to: „Matematika–ekonomie se zaměřením na pojišťovnictví“ a „Statistické a počítačové modelování“. Můj příspěvek bude věnován druhému z uvedených oborů. Nejprve krátce k historii, jak učební plán vznikal.

Na počátku bylo zadání: vytvořit dostatečně ucelené tříleté bakalářské studium, které by bylo orientované na matematickou statistiku a její aplikace. Představa byla taková, že absolvent by měl být schopen tvořivě zpracovávat experimentální data, pomáhat „klientům“ formulovat jejich pracovní hypotézy atd. To je samozřejmě obrovský úkol pro tři roky studia a jistě se shodneme, že úkol téměř neřešitelný. Přesto jsme se pokusili najít optimální variantu (pochopitelně s přihlédnutím k personálním možnostem katedry). Dlouho jsme také debatovali o názvu oboru, aby (podle hesla „obal prodává“) zájemce ke studiu přitahoval a neodrazoval je.

Pod vedením prof. dr. ing. Lubomíra Kubáčka, DrSc. se začala nepravidelně scházet pracovní skupina ve složení dr. Fišerová, doc. Kunderová, dr. Müller, Mgr. Marek, Mgr. Vrbková. Mnoho návrhů bylo učiněno v průběhu času, mnoho z nich bylo zamítnuto: co učit, v jakém rozsahu, na jaké teoretické úrovni (např. pravděpodobnost bez teorie míry a teorie integrálu), zda užívat jen hotový software, zda a jak důkladně učit studenty samostatně programovat atd. Základy učebního plánu vznikly v debatách kolektivním vyjednáváním, největší zásluhu na jeho definitivní podobě mají Mgr. Jaroslav Marek a Mgr. Jana Vrbková, kteří návrhy předmětů vyladili, rozdělili je do ročníků, prověřili jejich návaznosti a (co jak známo není vůbec jednoduché) vytvořili všechny podklady nutné pro akreditaci. Výsledný učební plán je následující.

Studijní program: B1103 Aplikovaná matematika Kreditní limit: 180 kr.

Studijní obor: **Statistické a počítačové modelování**

Etapa: první Kreditní limit: 162 kr.

Povinné předměty Počet kreditů: 139 kr.

Název předmětu	Počet kreditů	Rozsah výuky Př+Cv+Sem	Zakonč.	Dopor. Rok	Sem.
Lineární algebra 1	7	3+2+0	Zp,Zk	1	Z
Matematika 1	11	4+2+0	Zp,Zk	1	Z
Software pro matematiky 1	3	1+2+0	Zp	1	Z
Úvod do výpočetní techniky	3	1+3+0	Zp	1	Z
Angličtina 3 (odb.)	1	0+2+0	Zp	1	Z
Lineární algebra 2	7	2+2+0	Zp,Zk	1	L
Matematika 2	11	4+2+0	Zp,Zk	1	L
Popisná statistika	5	2+1+0	Zp,Zk	1	L
Statistický software 1	3	0+0+2	Zp	1	L
Úvod do pravděpodobnosti	6	2+2+0	Zp,Zk	1	L
Angličtina 4 (odb.)	3	0+2+0	Ko	1	L
Matematika 3	3	1+1+0	Zp,Zk	2	Z
Pravděpodobnost a statistika 1	6	2+2+0	Zp,Zk	2	Z
Statistický software 2	3	0+0+2	Zp	2	Z
Úvod do numerických metod	4	2+1+0	Zp,Zk	2	Z
Výběrová šetření	4	2+1+0	Zp,Zk	2	Z
Biometrie	4	2+1+0	Zp,Ko	2	L
Neparametrické metody	4	2+1+0	Zp,Zk	2	L
Pravděpodobnost a statistika 2	5	2+2+0	Zp,Zk	2	L
Statistická kontrola kvality	4	2+1+0	Zp,Zk	2	L
Statistické lineární modely	4	2+1+0	Zp,Zk	2	L
Statistický software 3	3	0+0+2	Zp	2	L
Časové řady	5	3+1+0	Zp,Zk	3	Z
Mnohorozměrná stat. analýza	4	3+1+0	Zp,Zk	3	Z
Psychometrie	4	2+1+0	Ko	3	Z
Statistický software 4	3	0+0+2	Zp	3	Z
Pravděpodobnost a statistika 4	3	2+1+0	Zp	3	L
Statistický software 5	3	0+0+2	Zp	3	L
Diplomová práce - bakalářská	13	0+0+0	Zp	3	L

Volitelné předměty Volba min.: 23 kr.

Název předmětu	Počet kreditů	Rozsah výuky Př+Cv+Sem	Zakonč.	Doporuč. Rok	Sem.
Úvod do matematiky	4	2+1+0	Zp	1	Z
Software pro matematiky 2	3	1+2+0	Zp	1	L
Logistika	3	2+1+0	Zp,Zk	2	Z
Software pro matematiky 3	3	1+0+2	Zp	2	Z
Databáze	2	1+2+0	Zp,Zk	2	L
Mat. teorie rozhodování 1	4	2+1+0	Zp	2	L
Finanční matematika	3	2+1+0	Zp,Zk	3	Z
Fuzzy množiny a jejich apl. 1	4	2+1+0	Zp	3	Z
Lineární programování	4	2+1+0	Ko	3	Z
Mat. teorie rozhodování 2	4	2+1+0	Zp,Zk	3	Z
Stat. teorie experimentu 1	4	2+1+0	Zp,Zk	3	Z
Soft Computing	4	2+1+0	Zk	3	ZS
Ekonometrie	4	2+1+0	Zp,Zk	3	L
Fuzzy množiny a jejich apl. 2	4	2+1+0	Zp,Zk	3	L
Stat. teorie experimentu 2	3	2+1+0	Zp,Zk	3	L
Teorie odhadu	3	2+0+0	Zk	3	L
Informační systémy	4	2+2+0	Zk	3	LS

Doplňující předměty

Název předmětu	Počet kred.	Rozsah výuky Př+Cv+Sem	Zakonč.	Doporuč. sem.
Bankovnictví a peněžní ekonomie 1	5	2+1+0	Zp	Z
Bankovnictví a peněžní ekonomie 3	4	2+1+0	Zp	Z
Funkce komplexní proměnné	5	2+2+0	Zp,Zk	Z
Nelineární programování	4	3+1+0	Zp,Zk	Z
Obchodní a bankovní právo 1	4	2+0+0	Ko	Z
Pojišťovnictví 1	5	2+1+0	Zp	Z
Pojistný trh	4	2+1+0	Zp	Z
Pojistné právo	4	2+0+0	Ko	ZS
TeX pro pokročilé Z	2	0+2+0	Zp	Z
Angličtina 1 (všeo.)	1	0+2+0	Zp	Z
Bankovnictví a peněžní ekonomie 2	4	2+1+0	Zp,Zk	L
Bankovnictví a peněžní ekonomie 4	4	2+1+0	Zp,Zk	L
Numerické metody optimalizace	3	2+1+0	Zp,Zk	L
Obchodní a bankovní právo 2	4	2+0+0	Ko	L
Psychologie obchodního jednání	3	0+0+2	Ko	L
Pojistná matematika	3	2+1+0	Zp	L
Pojištovnictví 2	4	2+1+0	Zp,Zk	L
Komerční pojišťovna	4	2+1+0	Zp,Zk	L
TeX pro pokročilé L	2	0+2+0	Zp	L
Angličtina 2 (všeo.)	3	0+2+0	Zp,Zk	L

Kolektiv autorů prosí všechny čtenáře, aby přispěli svými zkušenostmi zda a jak by bylo možné tento učební plán vylepšit. Samozřejmě, bude-li obor akreditován (jak doufáme), můžeme provádět jen dílčí úpravy učebního plánu, ale i tak budeme vděčni za každou radu a připomínku.

Potlačenie šumu v RDPT

Matejčiková Andrea, Török Csaba
KM SvF TU Košice
andrea.matejcikova@tuke.sk, csaba.torok@tuke.sk

Abstrakt

Článok sa zaoberá odhadom stupňa polynomiálnej regresie pomocou modifikovanej DPT, rozoberá výhody a nevýhody tejto modifikácie, poukazuje na možnosť jej špeciálnej definície, ktorá umožňuje potlačenie šumu.

1. Úvod

Uvažujme regresný model $y = a_0 + a_1 x + a_2 x^2 + ... + a_r x^r + \varepsilon$, kde koeficienty $a_0,...a_r$ a stupeň regresie r sú neznáme koeficienty, $\varepsilon \sim N(0,\sigma^2)$.

Príspevok bude hovoriť o nových možnostiach odhadu stupňa regresného polynómu. Hneď v druhej kapitole priblíži metódu diskrétnej projektívnej transformácie a jej modifikáciu RDPT. V ďalšej časti bude popísané štatistické kritérium použité na odhadovanie neznámeho stupňa polynómu. Sekcia 4 sa venuje problémom súvisiacim s výberom pivotov pri RDPT a nasledujúca kapitola ponúka nové riešenie. V šiestej časti sú odvodené niektoré teoretické výsledky pre rozdelenie veličín, s ktorými pracujeme. V závere sú zhrnuté výsledky dosiahnuté pomocou pohyblivých pivotov.

2. DPT a RDPT

Diskrétna projektívna transformácia (DPT) je operácia definovaná pre všetky diferencovateľné funkcie definované pomocou vzorca alebo pomocou tabuľky. Táto časť sa zaoberá stručným popisom DPT a modifikáciou RDPT.

Ak na grafe ľubovoľnej spojitej funkcie $f(x)$ fixujeme dva body (pivoty) $[\lambda, f(\lambda)]$ a $[L, f(L)]$, $\lambda \neq L$, potom DPT zobrazí ľubovoľný bod $[x, f(x)]$ (rôzny od predchádzajúcich) na zodpovedajúci bod $[x, h(x)]$ na krivke jednoduchšej geometrickej štruktúry.

Nech $x \neq \lambda \neq L$, $x, \lambda, L \in R$, potom definujeme funkcie p_i, $i = 1,...3$:

$$p_1 = p_1(x,\lambda,L) = \frac{xL}{(\lambda - x)(\lambda - L)}, p_2 = p_2(x,\lambda,L) = \frac{\lambda x}{(L - \lambda)(L - x)}, p_3 = p_3(x,\lambda,L) = \frac{\lambda L}{(x - \lambda)(x - L)}.$$

Tieto funkcie zohrávajú kľúčovú úlohu v definícii DPT (pozri podrobnejšie [1]):

Výskum bol podporený z grantového projektu VEGA 1/1006/04 9150 MŠ

Definícia 1:
DPT ľubovoľnej diferencovateľnej funkcie $f(x)$ *je definovaná pre každé* $x \neq \lambda \neq L$ *nasledovne:* $D[f(x)] = p_1 f(\lambda) + p_2 f(L) + p_3 f(x)$.

DPT transformácia má niekoľko dôležitých vlastností, vďaka ktorým ju môžeme použiť na odhad stupňa regresného polynómu.
Ak transformujeme mocninovú funkciu, tak môžeme ľahko dokázať, že:

$D[c] = c$ pre ľubovoľnú reálnu konštantu c,

$D[x] = D[x^2] = 0$

$D[x^n] = g(x), n \geq 3$, pričom $g(x)$ je polynóm stupňa $n-2$.

Viacnásobným použitím DPT na polynómy skôr či neskôr dostaneme konštantnú funkciu. Pri odhade stupňa regresie hľadáme *konštantnú transformáciu* a na základe počtu vykonaných transformácií spätne odhadneme stupeň regresného polynómu.

Vyššie popísaná transformácia má pri pivotoch vybraných z meraní isté nevýhody (strata najmenej dvoch bodov pri každej transformácii, zväčšovanie chyby – viď [5]), kvôli ktorým sme ju v [2] modifikovali.

Modifikácia spočíva vo využití regresie na výber pivotov. Nasleduje zmenená definícia regresnej DPT (resp. RDPT):

Definícia 2:
RDPT ľubovoľného polynómu $P_A(x)$ *je definovaná pre každé* $x \neq \lambda \neq L$ *nasledovne:* $D_B[P_A(x) + \varepsilon] = p_1 P_B(\lambda) + p_2 P_B(L) + p_3 (P_A(x) + \varepsilon)$, *kde* $\varepsilon \sim N(0, \sigma^2)$ *a* $P_B(.)$ *označuje regresný polynóm s vektorom odhadnutých koeficientov* B.

RDPT má podobné vlastnosti ako DPT, čo je dôležité pre jej použitie pri odhadovaní stupňa regresie. Definícia nám umožňuje vybrať za pivot ľubovoľný bod na $P_B(.)$, pričom je jednoduché ďalej počítať jeho transformácie.

Pozrime sa na to, ako to vyzerá s RDPT vyššieho stupňa. Zavedieme si nasledovné označenie: $\tilde{y}_i = a_0 + a_1 x_i + a_2 x_i^2 + ... + a_r x_i^r + \varepsilon_i$

$$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2 + ... + b_r x_i^r + \varepsilon_i, \; p_{3i} = p_3(x_i, \lambda, L)$$

Potom je možné ukázať, že: $D_B \tilde{y}_i = D\tilde{y}_i + p_{3i}(\tilde{y}_i - \hat{y}_i)$,

$$D_B^m \tilde{y}_i = D^m \tilde{y}_i + (p_{3i})^m (\tilde{y}_i - \hat{y}_i) \qquad (1)$$

Teda RDPT je súčtom klasickej DPT a reziduálneho člena. V ďalšom budeme využívať vektory $\tilde{Y}, \hat{Y}$ so zložkami $\tilde{y}_i, \hat{y}_i$ pre $i = 1,...,n$.

3. Analytické kritérium

Pri analýze výsledkov sme vychádzali z Akaike kritéria. Ako je známe, toto kritérium penalizuje rastúci počet parametrov. V [3] sa používa penalizačná funkcia

$$g_n(k) = s_k^2 (1 + q_{k,n}), k = 1,...,K, \qquad (2)$$

$s_k^2 = \frac{(\tilde{Y} - \hat{Y})'(\tilde{Y} - \hat{Y})}{n-k}, q_{k,n} = kn^{-0.25}$, K je maximálny uvažovaný stupeň polynómu.

Stupeň regresného polynómu r sa aproximuje číslom $r^* = k - 1$, kde k minimalizuje funkciu $g_n(k)$.

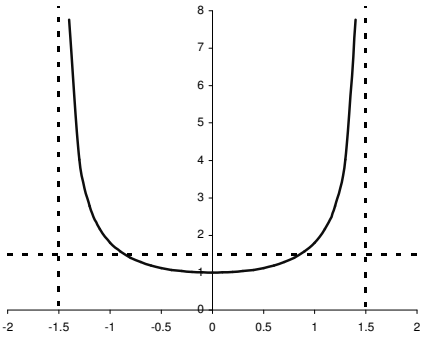
Penalizačnú funkciu modifikujeme nasledovne:

$$G_n(k) = S_k^2(1 + q_{k,n}), k = 1,..., K, q_{k,n} = kn^{-0.25}$$
$$S_k^2 = \frac{(D_B^m \tilde{Y} - b_0)'(D_B^m \tilde{Y} - b_0)}{\sum_{i=1}^{n} (p_{3i})^{2m}(1 - X_i'(X'X)^{-1}X_i)}, m = \left\lfloor \frac{k}{2} \right\rfloor \quad (3)$$

4. RDPT s konečnými a nekonečnými pivotmi

V tejto časti sa zaoberáme výberom pivotov a jeho vplyvom na chybu RDPT. Uvažujme konštantnú transformáciu. Potom z (1) vyplýva $D_B^m \tilde{y}_i = b_0 + (p_{3i})^m (\tilde{y}_i - \hat{y}_i)$.

Teraz sa budeme trochu hlbšie zaoberať týmto výrazom a dôsledkami, ktoré táto forma prináša. Keďže hodnota b_0 je pre nás známa (je to odhad absolútneho člena v použitej regresii) a neovplyvňuje presnosť výsledku, nebudeme sa ňou ďalej zaoberať. Čitateľ si iste všimol skutočnosť, že je to práve výraz p_{3i}, ktorý má najväčší vplyv na samotný výsledok, resp. na jeho presnosť. Zrejmý je aj fakt, že čím viac transformácií vykonáme, tým je výsledok viac ovplyvnený vďaka umocňovaniu p_{3i}. Nakoľko nás zaujíma zmenšovanie šumu, môžeme si ľahko overiť, že takáto funkcia nadobúda minimálnu hodnotu v bode $\lambda = -L$, čím dostaneme $p_{3i} = \frac{L^2}{L^2 - x_i^2}$. Keď si podrobnejšie pozrieme správanie sa funkcie p_{3i}, zistíme, že jej graf má nasledujúci priebeh:

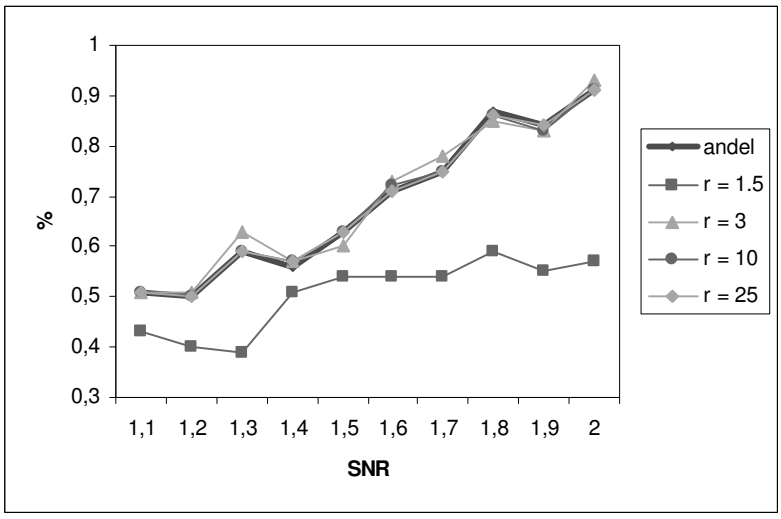


Obr. 1: Funkcia p_3

Pretože nám ide o odhad stupňa regresného polynómu, bez ujmy na všeobecnosti x-ové súradnice dát transformujeme (kvôli numerickej stabilite a aby boli chyby symetrické) do konečného symetrického intervalu $< -a, a >, a \in R$. Z grafu vidíme teda, že pre konečné L je $p_{3i} > 1$, čo vedie kvôli umocňovaniu k zväčšeniu reziduálneho člena. Ak sa zameriame na limitné vlastnosti p_{3i}, môžeme si všimnúť, že

pre $x \in < -a, a >$: $\lim_{L\to\infty} p_{3i} = \lim_{L\to\infty} \frac{L^2}{L^2 - x_i^2} = 1$.

To znamená, že keď chceme potlačiť zväčšenie reziduálneho člena, vyberieme pivoty tak, aby boli od meraní dostatočne vzdialené a aby bol vplyv mocnín p_{3i} čo najmenší. Túto skutočnosť nám potvrdzujú aj simulačné výsledky. Odhady na základe kritéria (3) s pevnými, ale malými pivotmi ($h = 1.5$ resp. $h = 3$, pričom $L = h \cdot a$) sa líšia od výsledkov dosiahnutých v [3], pri vyšších hodnotách sa prakticky rovnajú (viď Obr. 2). V ďalšej časti navrhujeme iné riešenie daného problému.



Obr. 2: Percentuálna úspešnosť kritéria s pevnými pivotmi pre $n = 30$

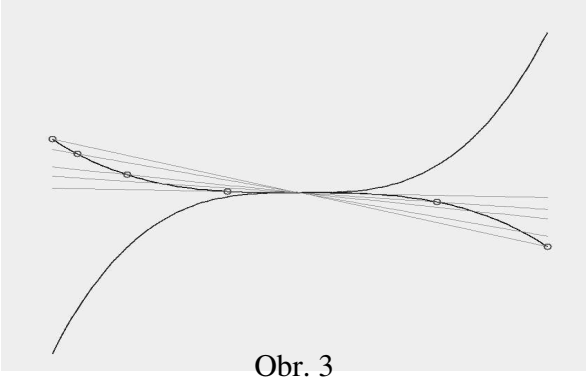
5. Pohyblivé pivoty

Táto časť obsahuje základný výsledok práce, ktorý nám umožní vyriešiť problém šumu bez limitného prechodu.

Mohli sme si všimnúť, že všetky problémy, s ktorými sme sa doteraz stretli, má na svedomí funkcia p_{3i}. Pokúsme sa s p_{3i} niečo spraviť. Je jasné, že každá zmena p_{3i} má vplyv na zmenu šumu vo výslednom RDPT. Zdá sa teda, že $p_{3i} = c$, c je konštanta, by mohlo vyriešiť problém so šumom. Priblížime si túto situáciu a pozrieme sa na dôsledky, ktoré nám vyplývajú z tejto podmienky.

Ak $p_{3i} = c$, tak $c = \frac{L^2}{L^2 - x_i^2}$. Budeme vychádzať z toho, že konštantu c si chceme zvoliť. x_i sú merania, ktoré máme dané, takže nám ostáva hodnota L, ktorá musí byť závislá od c a x_i. Po úprave dostaneme vzťah $L = |x_i| \sqrt{\frac{c}{c-1}}$.

Takáto voľba L pre doteraz fixné pivoty spôsobí, že sa zmenia na pohyblivé, nakoľko sa transformácie budú počítať pre každý bod zvlášť. Nezmení to však podstatné vlastnosti RDPT potrebné k odhadovaniu stupňa polynómu, teda každá transformácia zníži stupeň polynómu o 2 stupne. Podstatný rozdiel oproti statickej RDPT spočíva v tom, že pohyblivú RDPT nemôžeme využívať na vizuálny test. Zatiaľ čo doteraz sme mohli na grafoch jednotlivých transformácií pozorovať pokles stupňa polynómu, tento postup nám takéto vyhodnocovanie neumožňuje, nakoľko každý bod je transformovaný podľa iných pivotov a teda každý bod má po transformácii vlastný polynóm nižšieho stupňa, aj keď tento stupeň je pre všetky rovnaký. Konkrétne to môžeme vidieť na obr. 3. Statická RDPT transformuje polynóm x^3 na priamku. Na obrázku vidíme, že transformované body neležia na jednej priamke.



Obr. 3

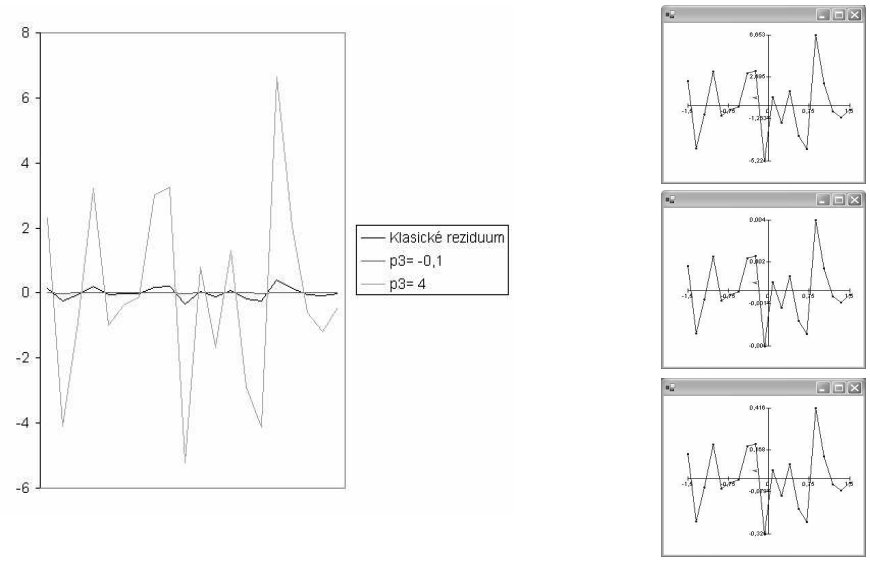
Keď sa na to pozrieme z teoretického hľadiska, kubická funkcia sa pri pevných pivotoch transformuje podľa vzťahu $Dx^3 = \lambda L x$. Pri pohyblivých pivotoch zvolíme $\lambda = -L, L = hx_i$. Každý transformovaný bod leží na inej priamke, ale keď uvažujeme tieto body ako celok, dostaneme kubickú parabolu $-h^2 x^3$. Napriek tomu vizuálny test nemusíme celkom odmietnuť. Môžeme vyhodnocovať konštantnú transformáciu, ktorá je rovnaká pre všetky body bez ohľadu na to, aké pivoty sa použili na transformácie.

Máme teda výraz $L = |x_i| \sqrt{\frac{c}{c-1}}$. Aby sme mohli určiť hodnotu L, musí byť samozrejme výraz $\frac{c}{c-1} > 0$ (nie $= 0 = 0$, lebo $L \neq \lambda = -L$), teda dostaneme, že $c \in (-\infty, 0) \cup (1, \infty)$. Pozrime sa na výraz $D_B^m \tilde{y}_i = b_0 + \underline{\underline{c^m}} (\tilde{y}_i - \hat{y}_i)$, ako ho ovplyvní

voľba konštanty c. Keďže pred $(\tilde{y}_i - \hat{y}_i)$ nie je len samotné c, ale c^m, tak budú pre nás zaujímavé dve množiny:

1. $c \in (-\infty,-1) \cup (1,\infty) \Rightarrow \left|c^m\right| > 1$
2. $c \in (-1,0) \Rightarrow \left|c^m\right| < 1$

Špecifický prípad nastane, ak $c = -1$, vtedy sa najviac priblížime k hodnote klasického rezídua, dokonca, ak m bude párne, tak $c^m(\tilde{y}_i - \hat{y}_i) = \tilde{y}_i - \hat{y}_i$.
Týmto spôsobom sme teda získali nástroj, ktorý nám umožňuje regulovať výraz $D_B^m \tilde{y}_i = b_0 + \underline{\underline{c^m}}(\tilde{y}_i - \hat{y}_i)$ prostredníctvom voľby konštanty c. Ďalší obrázok (obr. 4) nám však, žiaľ, ukazuje, že všetky rezídua majú rovnaký priebeh, rozdielny je len interval funkčných hodnôt, ktoré dosahujú (viď vľavo).



Obr. 4: Porovnanie regulovaného rezídua

Vidíme teda, že konštanta c má vplyv na amplitúdu rezídua (pozri Obr. 3) a tiež na rezíduálny súčet $(D_B^m \tilde{Y} - b_0)'(D_B^m \tilde{Y} - b_0)$.

Z rovnosti $(D_B^m \tilde{Y} - b_0)'(D_B^m \tilde{Y} - b_0) = c^{2m} \sum_{i=1}^{n} (\tilde{y}_i - \hat{y}_i)^2$ dostaneme

$$\frac{(D_B^m \tilde{Y} - b_0)'(D_B^m \tilde{Y} - b_0)}{c^{2m}\sigma^2} = \frac{(\tilde{Y} - \hat{Y})'(\tilde{Y} - \hat{Y})}{\sigma^2},$$

čo umožňuje spojiť naše výsledky s výsledkami J. Anděla v tejto oblasti. Keď sa spätne pozrieme na Akaike kritérium (3) zostrojené na základe RDPT pri pohyblivých pivotoch s konštantnou hodnotou p_3, môžeme konštatovať, že sa zhoduje s kritériom navrhnutým v [3], čo nám potvrdzujú aj simulácie.

6. Vplyv RDPT na rozdelenie výrazu s $(\tilde{y}_i - \hat{y}_i)$

Rozdelenie rezídua pri klasickej regresii je všeobecne známe:

$(\tilde{y}_i - \hat{y}_i) \sim N(0, \sigma^2(1 - X_i^{'}(X'X)^{-1}X_i))$, ako aj $\frac{(\tilde{Y} - \hat{Y})'(\tilde{Y} - \hat{Y})}{\sigma^2} \sim \chi^2_{n-k}$.

Ďalej budeme vychádzať z týchto známych faktov spolu s ďalšími z [4]. Ako to teda vyzerá pre RDPT s fixnými resp. pohyblivými pivotmi?

Pri statickej RDPT sa stretávame s výrazom $D_B^m \tilde{y}_i = b_0 + (p_{3i})^m(\tilde{y}_i - \hat{y}_i)$. V tomto prípade dostaneme $(p_{3i})^m(\tilde{y}_i - \hat{y}_i) \sim N(0, \underline{\underline{p_{3i}^{2m}}}\sigma^2(1 - X_i^{'}(X'X)^{-1}X_i))$, s odvodením rozdelenia pre $(D_B^m\tilde{Y} - b_0)'(D_B^m\tilde{Y} - b_0)$ je to trochu náročnejšie a zatiaľ otvorené (pretože dostaneme výraz $\sum_{i=1}^{n} p_{3i}^{2m}(\tilde{y}_i - \hat{y}_i)^2$, odkiaľ nemožno vybrať výraz p_{3i}).

Oveľa lepšia situácia je pri RDPT s pohyblivými pivotmi. Výraz $D_B^m \tilde{y}_i$ má tvar :

$D_B^m \tilde{y}_i = b_0 + c^m(\tilde{y}_i - \hat{y}_i)$, $c^m(\tilde{y}_i - \hat{y}_i) \sim N(0, \underline{\underline{c^{2m}}}\sigma^2(1 - X_i^{'}(X'X)^{-1}X_i))$.

Je zrejmé, že $\frac{(D_B^m\tilde{Y} - b_0)'(D_B^m\tilde{Y} - b_0)}{c^{2m}\sigma^2} \sim \chi^2_{n-k}$.

7. Záver

Článok obsahuje krátke zhrnutie výsledkov dosiahnutých pomocou regresnej DP transformácie polynómov. Prináša nový pohľad na problematiku – použitie pohyblivých pivotov. Pohyblivé pivoty sú špeciálne vyberané tak, aby sme mohli regulovať chybu transformácie. Táto myšlienka je rozpracovaná tak po teoretickej ako aj po praktickej stránke. Do budúcnosti by bolo vhodné nájsť kritérium, pre ktoré by bola škálovateľnosť reziduálneho člena väčšou výhodou.

Literatúra

1. Dikoussar N.D.: Function parametrization by using 4-point transforms, Computer Physics Communications 99 (1997), pp. 235-254
2. Török Cs., Matejčiková A.: Estimating the polynomial degree, Prastan 2004, pp. 71-77
3. Anděl J., Garrido M.T., Insua A.: Estimating the dimension of a linear model, Kybernetika 17 (1981), pp. 514-525
4. Anděl J.: Matematická statistika
5. Török Cs.:4-Point transforms and approximation, Computer Physics Communications 125 (2000), pp. 154-166

Pád guľky po Galtonovej doske

Iveta Molnárová
Katedra prírodných vied
Akadémia ozbrojených síl gen.M.R.Štefánika, Liptovský Mikuláš
imolnar@valm.sk

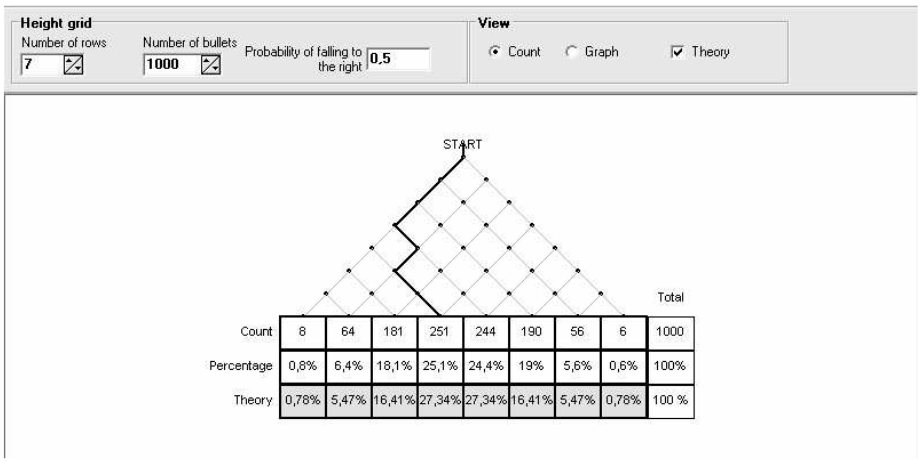
Abstract *The aim of this paper is to ilustrate possibilities how to use VUStat in the course of elementary Statistics.*

S našim súčasným životom sa spája širšie používanie informačných technológií. Ich využívanie sa stalo súčasťou vyučovania na všetkých typoch škôl, pri vyučovaní pravdepodobnosti a štatistiky sa to prejavilo hľadaním a využívaním rôznych softvérov. Pri ich nekritickom používaní hrozí nebezpečenstvo, že zo štatistiky sa môže stať zbierka návodov a postupov ako pracovať s daným softvérom, bez dôrazu na interpretáciu získaných výstupov a bez overenia vstupných predpokladov.

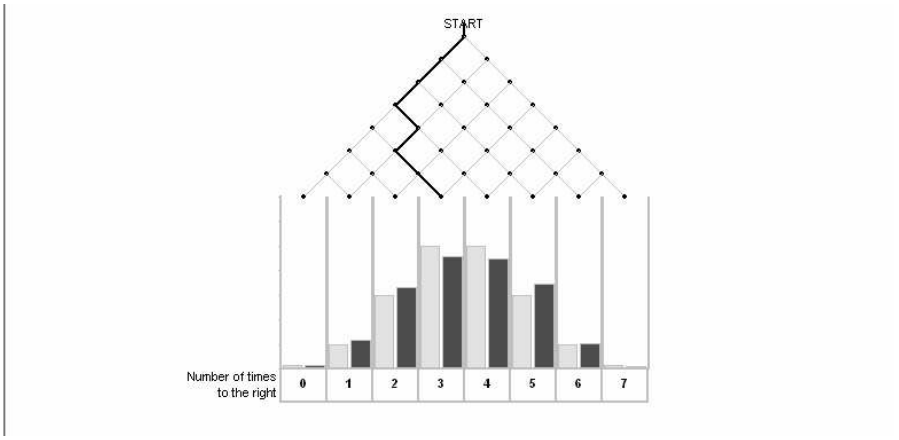
Ako doplnok k takýmto štatistickým produktom je možné použiť výučbový program VUStat, ktorý dobre ilustruje základné štatistické pojmy, umožňuje graficky i numericky spracovávať údaje, simulovať hromadné javy. Je to systém, ktorý sa v Holandsku používa na pravdepodobnostné a štatistické vzdelávanie učiteľov a žiakov vo veku 12-18 rokov. Demoverzia tohto programu sa nachádza na stránke www.vusoft2.nl . Jedna aplikácia tohto programu je obsahom príspevku.

Dnes známa Galtonova doska je losovací nástroj popísaný v roku 1889 lordom Francisom Galtonom. Pozostáva zo šikmej dosky, na ktorej je v siedmich radoch nabitých do tvaru rovnostranného trojuholníka postupne 1, 2, 3, ..., 7 kolíkov. Guľka po spustení po doske naráža na kolíky, ktoré ju odrazia vpravo alebo vľavo s pravdepodobnosťou 0,5. Guľka končí svoju dráhu v jednej z ôsmich priehradok očíslovaných 0, 1, 2, ..., 7 (obr.1).

Na ilustráciu aproximácie binomického rozdelenia normálnym použil Rényi v [1] Galtonovu dosku, kde po dostatočne veľkom množstve napr. $N = 1000$ spustených guliek po doske sa tieto roztriedili do stĺpcov, ktoré spolu vytvorili útvar, pripomínajúci Gaussovu krivku. Tento náhodný pokus sa dá simulovať aj v tomto prostredí, dá sa meniť počet radov kolíkov n, pravdepodobnosť p odrazenia guľky doprava, počet spustených guliek, výsledkom sú absolútne i relatívne početnosti guliek v priehradkách, ako aj teoretické pravdepodobnosti (sú vyjadrené v percentách). Na porovnanie empirického a teoretického rozdelenia poslúžia aj oba histogramy, ktoré sa dajú zobraziť (obr.2).



Obr.1



Obr.2

Pád guľky po Galtonovej doske môžeme použiť na:
- konštrukciu pravdepodobnostného modelu pádu guľky
- na ilustráciu zákona veľkých čísiel
- na overenie zhody empirického rozdelenia s daným binomickým rozdelením
- na overenie zhody empirického rozdelenia s normálnym alebo Poissonovým rozdelením

1. Pravdepodobnostný model pádu guľky

Ak označíme odrazenie guľky vpravo ako úspech (1) a vľavo ako neúspech (0), tak dráha guľky podľa obr.1 spočíva v 7-násobnom nezávislom opakovaní toho istého pokusu a dá sa zapísať pomocou sedemčlennej variácie množiny $\{0,1\}$. Napr. dráha 0001011 znamená, že guľka skončila v tretej priehradke. Pád guľky po Galtonovej doske je teda Bernoulliho schéma so siedmimi pokusmi. Ak pravdepodobnosť úspechu je p a neúspechu $1-p$, tak pravdepodobnosť toho, že guľka prejde dráhu 0001011sa vďaka nezávislosti rovná $p^3.(1-p)^4$. V *k*-tej priehradke $(k=0,1,...,7)$ končí $\binom{7}{k}$dráh. Preto pravdepodobnosť toho, že dráha guľky skončí v *k*- tej priehradke , je $\binom{7}{k}p^k(1-p)^{7-k}$. Teda pád guľky po klasickej Galtonovej doske má binomické rozdelenie Bi(7, *p*).

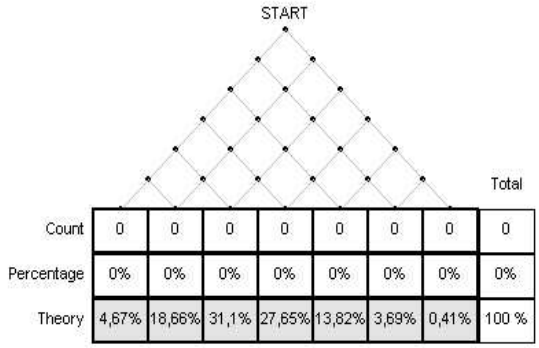
S využitím schémy pre Galtonovu dosku môžeme riešiť aj iné úlohy:

a) Minimálne koľko pokusov treba urobiť, aby úspech nastal aspoň raz s pravdepodobnosťou aspoň *P*, ak v jednom pokuse nastane s pravdepodobnosťou *p*.

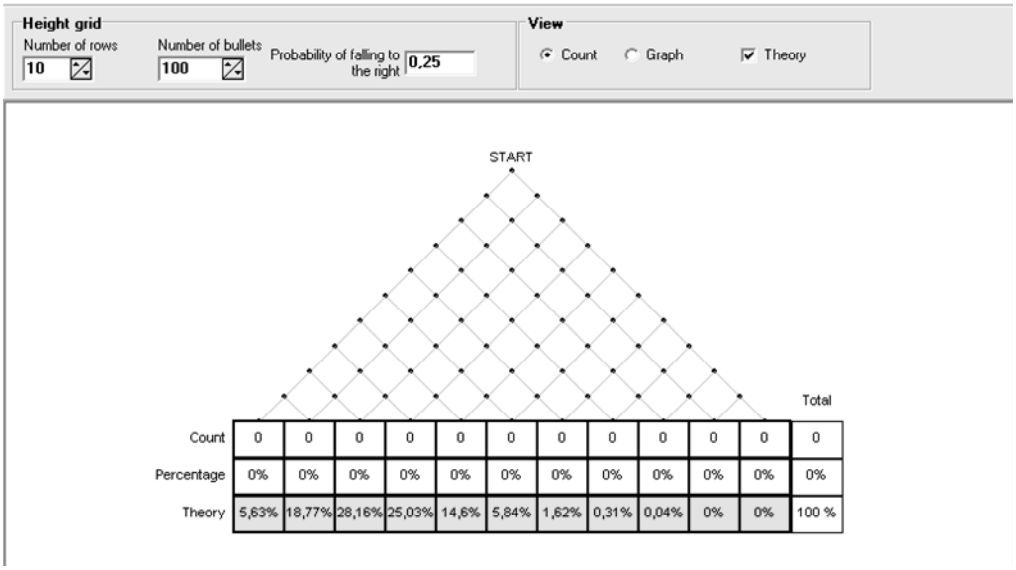
Ak si zvolíme napr. $p=0,4$ a $P=0,95$, tak postupným menením hodnoty $n=1,2,3,..$ zistíme, že najmenšie *n,* pre ktoré platí $P(X\geq 1)\geq 0,95$ je $n=6$. Z obrázka 3 vidieť, že $P(X\geq 1)=95,33\%$.

b) Aká je pravdepodobnosť, že žiak, ktorý nič nevie, odpovie v teste s *n* otázkami správne aspoň na *k* otázok, ak na každú otázku je ponúknutých *r* možných odpovedí.

Ak si zvolíme 10 otázkový test a 4 možné odpovede na každú otázku, potom pravdepodobnosť náhodného uhádnutia správnej odpovede na každú otázku je 0,25. Stačí v schéme pre Galtonovu dosku zadať tieto hodnoty a z obrázka určiť napr. $P(X\geq 4)=22,41\%$ (obr.4).



Obr.3



Obr.4

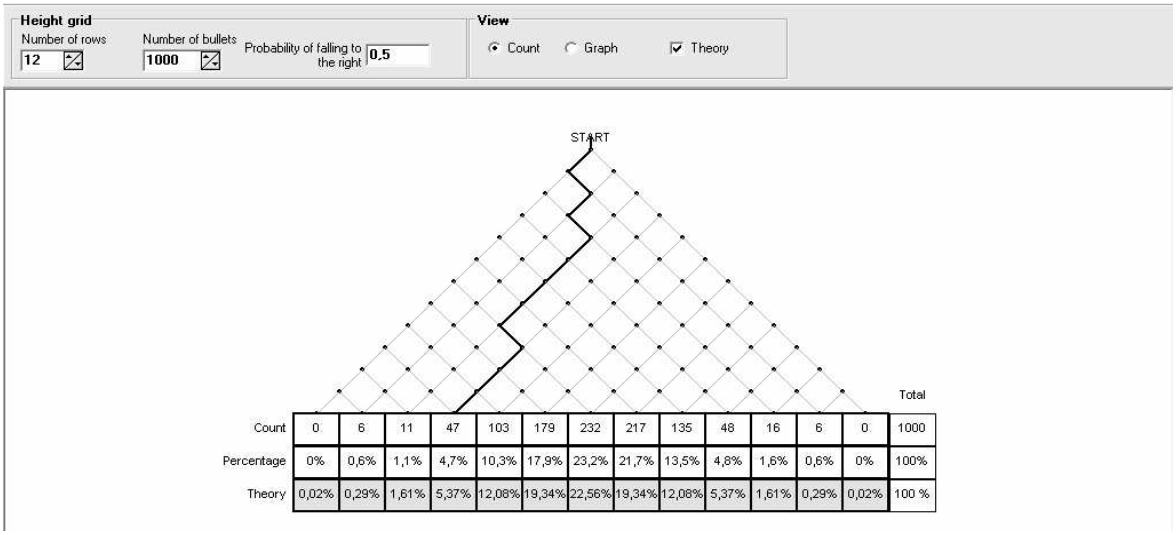
c) V rovnakom teste ako v úlohe b) potrebujeme určiť pravidlo pre vierohodné ohodnotenie žiaka. To znamená, aké si zvoliť kritérium pre úspešnosť v teste, teda určiť minimálny počet správnych odpovedí *k*. Požadujeme, aby pravdepodobnosť náhodného urobenia testu v prípade, že študent nič nevie, bola menšia ako napr. 0,05, t.j. aby $P(X \geq k) < 0{,}05$.

Využijeme predchádzajúci obr.4 a zistíme, že treba zvoliť $k = 6$, lebo $P(X \geq 6) = 0{,}0197$ (na obrázku v percentách 1,97%). Pre $k = 5$ nie je podmienka ešte splnená. Týmto príkladom sa dostávame k testovaniu štatistických hypotéz, kde H_0 : študent nič nevie, H_1 : študent niečo vie. V prípade platnosti nulovej hypotézy je vysoký počet správnych odpovedí málo pravdepodobný, preto v takomto prípade musíme nulovú hypotézu zamietnuť a prijať alternatívnu.

2. Ilustrácia zákona veľkých čísiel

Doteraz sme nevyužili podstatnú vlastnosť tohto programu, simulovať pád veľkého množstva guliek. Po dokončení simulácie pre $n = 12$ radov kolíkov, $N = 1000$ guliek a $p = 0{,}5$ môže mať výsledok podobu obrázka 5. Vzniknuté rozdiely medzi relatívnymi početnosťami a pravdepodobnosťami (vyjadrené v percentách) sú veľmi malé, čo nás v praxi oprávňuje po-

užívať relatívnu početnosť namiesto neznámej pravdepodobnosti pri veľkom počte pokusov. V ukážke je maximálny rozdiel 0,0236.

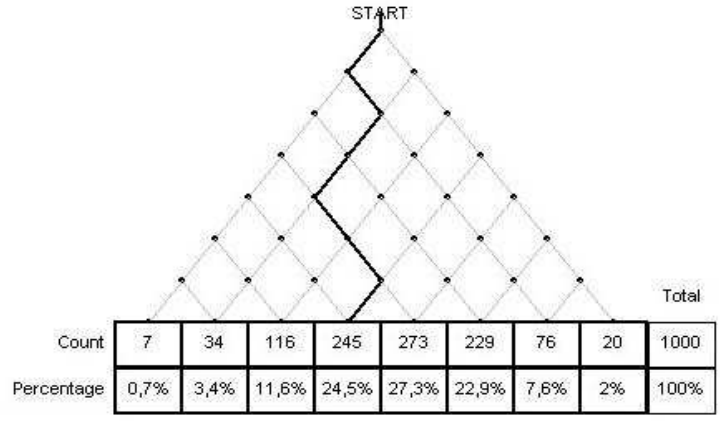


Obr.5

3. Zhoda empirického rozdelenia s daným binomickým rozdelením

Môže sa stať, že Galtonova doska ako losovací nástroj bude položená tak, že pravdepodobnosti odrazenia guľky vpravo a vľavo nebudú rovnaké, to znamená $p \neq 0{,}5$. Máme k dispozícii náhodný výber - výsledok simulácie pádu 1000 guliek (obr.6) a zdá sa nám, že guľka padala častejšie do pravých priečinkov. Na základe náhodného výberu máme testovať hypotézu H_0 proti alternatíve H_1:

H_0: Náhodný výber pochádza z $Bi(7; 0{,}5)$.
H_1: Náhodný výber nepochádza z $Bi(7; 0{,}5)$.



Obr.6

Na posúdenie zhody empirického a nulovou hypotézou predpokladaného binomického rozdelenia $Bi(7;0{,}5)$ použijeme Chí-kvadrát test dobrej zhody. Výpočet hodnoty testovacej štatistiky, p-hodnoty i záver sú v tab.1.

x_i	n_i	p_i	$n.p_i$	$\frac{(n_i - np_i)^2}{np_i}$	
0	7	0,0078	7,8	0,082051282	
1	34	0,0547	54,7	7,83345521	
2	116	0,1641	164,1	14,09878123	
3	245	0,2734	273,4	2,950109729	
4	273	0,2734	273,4	0,000585223	
5	229	0,1641	164,1	25,66733699	
6	76	0,0547	54,7	8,294149909	
7	20	0,0078	7,8	19,08205128	
		$\chi^2 = \sum_{i=0}^{7} \frac{(n_i - np_i)^2}{np_i}$		78,00852086	
		d.f.=7			
		p-hodnota		3,50645E-14	
Na všetkých bežných hladinách významnosti zamietame nulovú hypotézu.					

Tab.1

4. Zhoda empirického rozdelenia s normálnym alebo Poissonovým rozdelením

Binomické rozdelenie $Bi(n;p)$ konverguje k normálnemu rozdeleniu $N(np;npq)$ pre $n \to \infty$ a hodnoty p nie blízke jednotke alebo nule. Na Galtonovej doske sa dá meniť počet radov kolíkov len od 1,...12. V tomto prípade zhodu empirických údajov z obr.5 pre $n = 12, p = 0{,}5$ s teoretickým modelom $N(6;3)$ môžeme tiež overiť Chí- kvadrát testom dobrej zhody. Výhodne pre určenie p-hodnoty použijeme vo VUState graf distribučnej funkcie χ^2 – rozdelenia.

Na príklade simulácie pádu guľky po Galtonovej doske v programe VUStat chcem vyzdvihnúť prednosti, ktoré tlačené učebnice nemajú, dynamickosť celého procesu, možnosť priameho zasahovania do voľby parametrov a úpravy obrázkov.

Literatúra

[1] Rényi A.:*Teorie pravděpodobnosti,* Academia Praha, 1972
[2] Plocki A.: Pravděpodobnost kolem nás, Ústí nad Labem,2001
[3] www.vusoft2.nl, program vustatengdemo.zip

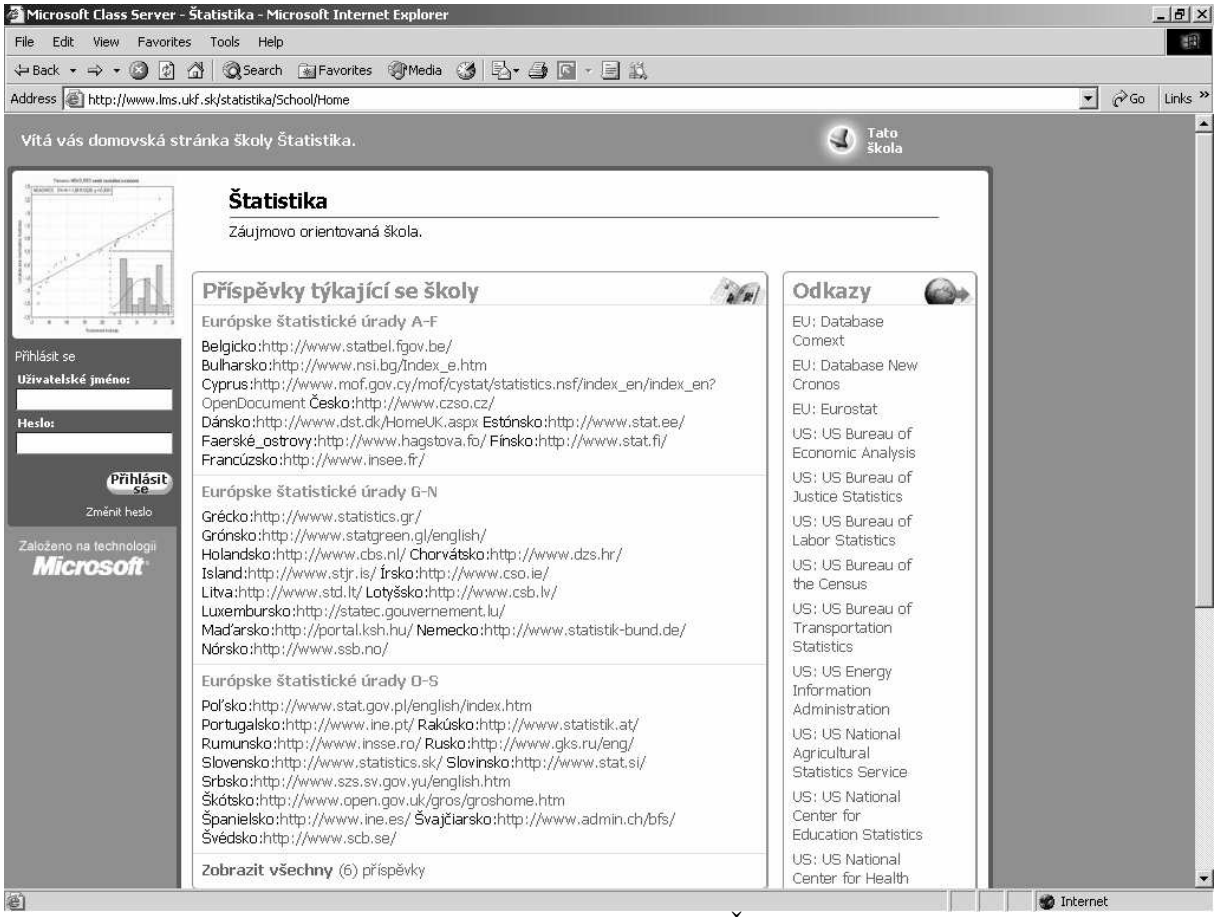
Virtuálna škola „Štatistika“

Michal Munk
Ústav technológie vzdelávania, PF UKF v Nitre
Tr. A. Hlinku 1, 949 01 Nitra
mmunk@ukf.sk

Jozef Kapusta
Katedra informatiky, FPV UKF v Nitre
Tr. A. Hlinku 1, 949 01 Nitra
jkapusta@ukf.sk

Článok popisuje vytvorenú virtuálnu školu a jeden z vybraných kurzov, ktorý je jej súčasťou. Škola má charakter HelpDesku a kurzy elektronických učebných pomôcok.

Virtuálna škola „Štatistika“ slúži ako HelpDesk k spracovaniu rôznych výskumných problémov. Obsahuje odkazy na literatúru, inštitúcie a verejné databázy, ktoré môžu pre používateľa predstavovať cenný zdroj dát a informácií. Informuje používateľov o udalostiach (konferenciách, školeniach a pod.) súvisiacich s touto problematikou. Vytvorené kurzy, ktoré sú jej súčasťou majú prevažne charakter elektronických pomôcok. Elektronické učebné pomôcky majú uplatnenie hlavne tam, kde sa očakáva samostatná práca študentov. Napríklad riešenie projektov, výskumných problémov k záverečným prácam a pod.



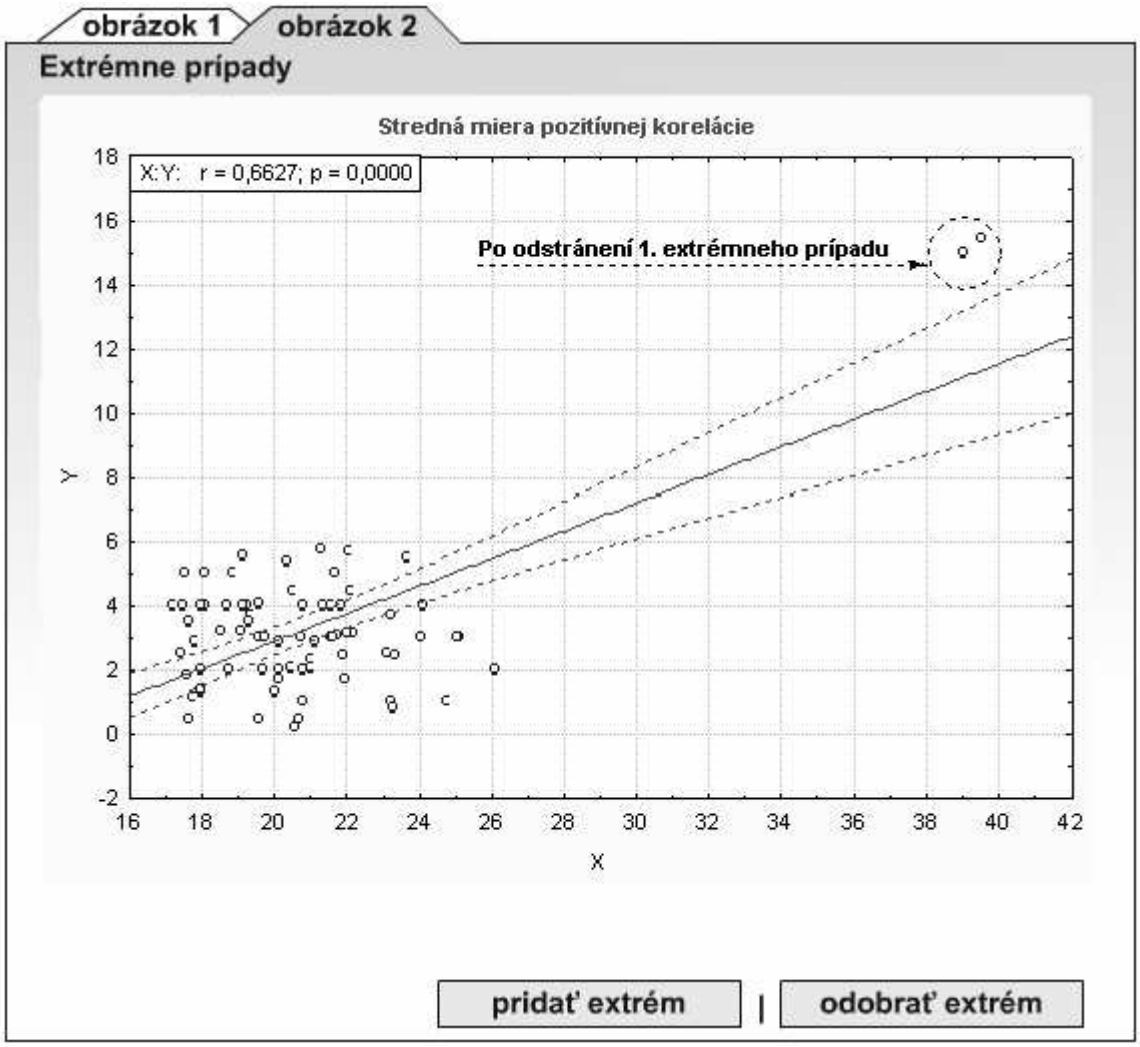
Obrázok 1 Virtuálna škola „Štatistika“

Príkladom je kurz „Štatistické spracovanie experimentu“. Kurz predstavuje elektronickú učebnú pomôcku k štatistickému spracovaniu experimentu. Používateľ sa nemusí

učiť obsah kurzu naspamäť, stačí, aby vedel pracovať s týmto materiálom a používal ho ako pomôcku pri štatistickom spracovávaní experimentu. Veľký dôraz sme kládli na vizualizáciu problematiky, ktorou sa kurz zaoberá.

Štruktúra kurzu:

Hlavné kapitoly kurzu tvoria jednotlivé fázy experimentu. Každá fáza je ilustrovaná na príklade. Každá kapitola obsahuje grafické prvky (animácie, schémy, grafy, tabuľky), ktoré zásadným spôsobom sprehľadňujú danú problematiku.



Obrázok 2 Extrémne prípady

Kurz sprehľadňujú použité symboly.

 Kľúčové slová

 Cieľ kurzu

 Učebný text

 Príklad k doplneniu učebného textu

 Pozri tabuľku/obrázok

 On-line štatistická kalkulačka, umožňuje generovanie náhodných čísel, prevod testovacích štatistík na p-hodnotu a pod.

 Stromový graf analytických metód, pomôcka pri výbere správnej metódy

 Zhrnutie

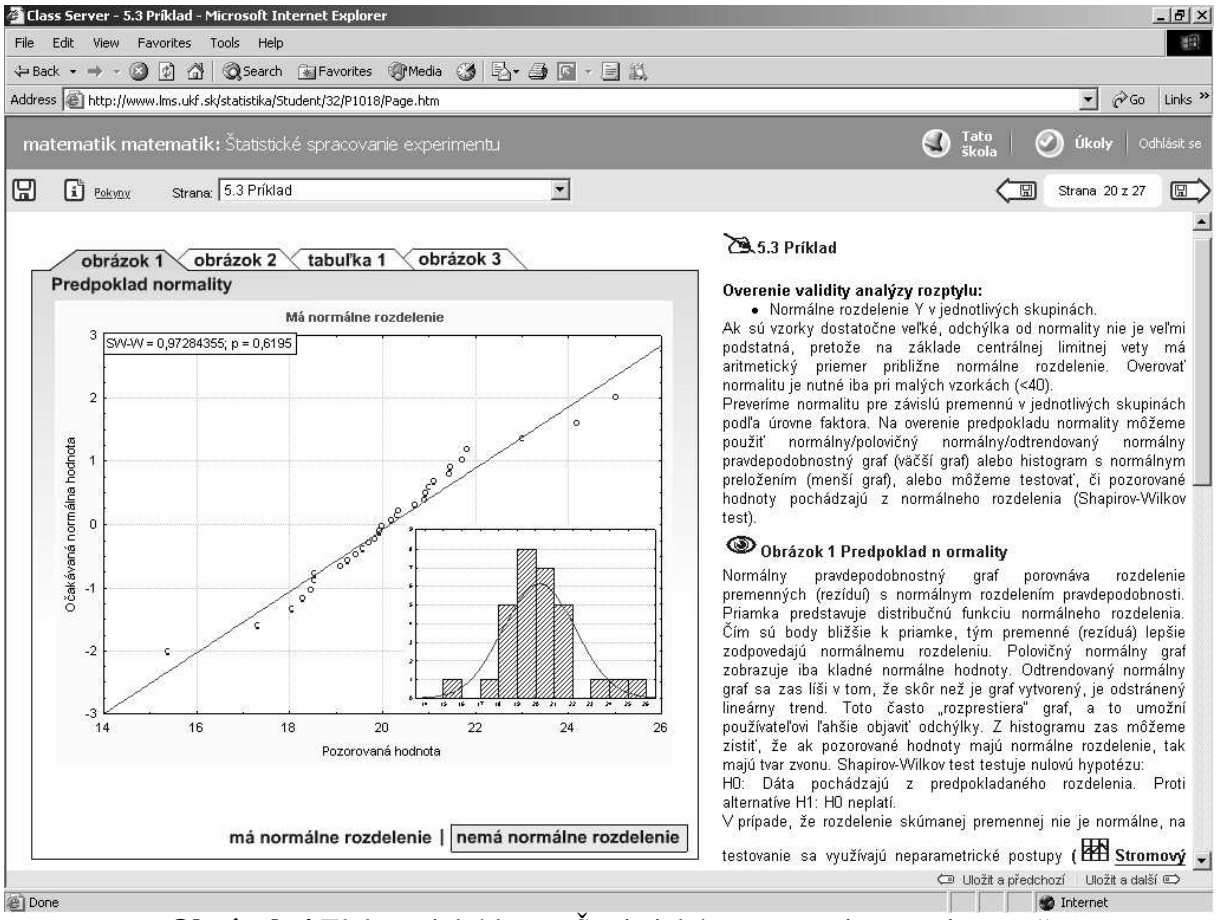
 Doporučená literatúra a zdroje

 Kontakt na realizačný tím

Obrázok 3 Použité symboly

Kurz oboznamuje používateľa o postupe pri spracovaní experimentu od vytvorenia experimentálnej a kontrolnej skupiny až po interpretáciu výsledkov. Kladie dôraz na potrebu vytvorenia kvalitných meracích procedúr – v našom prípade didaktických testov. Udáva viacero metód ako overiť reliabilitu a validitu vytvorených didaktických testov. Ponúka používateľovi k realizácii viacero experimentálnych plánov. Upozorňuje na potrebu porozumieť dátam, t.j. vypočítať popisné charakteristiky a intervaly spoľahlivosti a následne ich vizualizovať a na základe týchto výsledkov postaviť nulovú štatistickú hypotézu. K testovaniu stanovenej hypotézy kurz ponúka dve metódy – analýzu rozptylu a analýzu kovariancie. Okrem popisu týchto metód a postupu ako overiť predpoklady ich použitia udáva aj riešenia prípadných porušení predpokladov validity.

Vytvorený elektronický kurz je kombináciou html kódu a flashu. Je rozvrhnutý tak, že v ľavej časti obrazovky sú zobrazené grafické prvky (animácie, schémy, grafy a tabuľky), pravá časť obsahuje text a matematické vzťahy.



Obrázok 4 Elektronický kurz „Štatistické spracovanie experimentu“

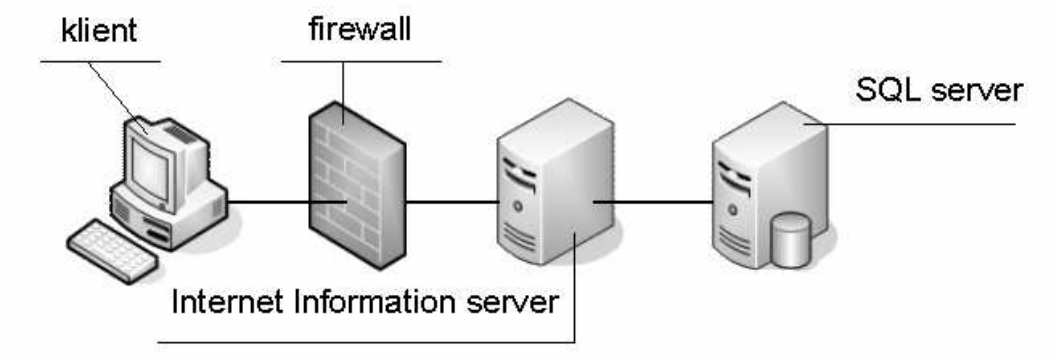
Technické požiadavky:

1. Pre prístup k virtuálnej škole je možný akýkoľvek počítač s prehliadačom Internet Explorer 5.0, či kompatibilným (napríklad Netscape 7.1, či IE 5.0 pre iMac). Pre dostatočne kvalitné zobrazenie kurzu doporučujeme Internet Explorer 6.0.
2. Doporučené grafické rozlíšenie 1024x768.

Ako riadiaci LMS sa používa produkt MS Class Server 3.0, ktorý umožňuje pohodlnú administratívu virtuálnych škôl. Serverový produkt spoločnosti Microsoft – Class Server pracuje nad serverovou platformou Windows 2000 a novšou.

MS Class Server pre svoj chod vyžaduje:

- Ø služby **Internet Information Serveru** (softvérové služby, ktoré okrem ďalších funkcií siete Internet podporujú vytváranie, konfigurovanie a správu webových stránok), ktorý je súčasťou sieťových operačných systémov,
- Ø služby **SQL Serveru** alebo u menších inštalácií produktu **MSDE** (bezplatný produkt), ktorý sa využíva ako dátový sklad pre ukladanie študijných materiálov, informácií o študentoch a pod.



Obrázok 5 MS Class Server

Záver

Ako sa očakávalo, virtuálna škola „Štatistika“ je prevažne využívaná študentmi pri riešení výskumných problémov k záverečným prácam, preto sa aj témy kurzov najčastejšie týkajú problémov vyskytujúcich sa v oblasti spracovania a analýzy dát.

Literatúra

1. http://www.lms.ukf.sk/statistika

Moodle - virtuálna univerzita?

Oľga Nánásiová, Katarína Trokanová

KMDG SvF STU, Radlinského 11, 813 68 Bratislava

Pri slovnom spojení virtuálna univerzita nás môže napadnúť viacero významov s dosť širokým významovým spektrom. Od pejoratívneho významu až po futuristické predstavy. Napriek tomu sme si toto slovné spojenie zvolili pred niekoľkými rokmi ako názov projektu KEGA, v ktorom sme sa chceli zaoberať touto problematikou. V dnešných časov sa v tejto súvislosti častejšie spomína výraz e-learning. Napriek tomu, že výhody a nevýhody takejto formy vzdelávania sú dostatočne známe, diskusia k tejto téme tak skoro neskončí. Vo virtuálnom konflikte je osobnosť učiteľa vo vzdelávacom procese a technika. Pravda je však taká, že ani najdokonalejšia „vzdelávacia technika“ nemôže nahradiť osobný kontakt učiteľa a študenta.

Slovné spojenie virtuálna univerzita ani v zmysle e-learningu nemá jednotný význam. Snáď najmodernejší, respektíve najdokonalejšie prevedenie virtuálnej univerzity do praxe je výroba digitálnych záznamov celých prednášok, ktoré si študent môže buď zakúpiť (platené štúdium), alebo môže byť poskytované zadarmo ako súčasť výukových programov. Takéto pokusy sa v histórii už objavili. V sedemdesiatych rokoch minulého storočia boli spracované jednotné televízne prednášky pre diaľkové štúdium z rôznych oborov /konkrétne aj matematika/ na BBC a aj v ruskej televízii. Zdá sa, že tieto projekty zlyhali asi na tom, že sa snažili byť jednotné pre všetky skupiny poslucháčov a na absencii osobnosti učiteľa. Aj k takej vede ako je matematika Je vôbec reálna nasledujúca predstava?

Študent si kliknutím vyberie tému a pred neho sa postaví trojrozmerný obraz učiteľa, ktorý mu vysvetlí danú tému, dokonca vie odpovedať na jednoduché otázky. Problém je, že učiteľ už dávno nežije.

Takáto forma by zrejme viedla k veľkej stagnácii, pretože aj prístup k úplne najzákladnejším poznatkom vedy je časom neustále mení. To iste pozná každý pedagóg na vysokej škole. Ak by sme takého programy mali vyrábať každý rok, alebo aspoň raz za 5 rokov, potom by sa vzdelávanie nezmyselne predražilo.

Článok bol podporený grantom KEGA číslo 3/0038/02

Predstava o uniformite všetkých materiálov pre študentov na našej katedre nemala šancu . Rozhodli sme sa pre zmiešaný model. Z ponúkaných programov pre virtuálnu univerzitu sme si vybrali program MOODLE. Výber mal niekoľko dôvodov. Tie najdôležitejšie: dá sa získať zadarmo, ak nie je využívaný komerčne, je ľahko ovládateľný a má pomerne slušnú verziu v slovenčine, dajú sa v ňom používať príkazy LaTeX-u na editovanie matematických textov. Po krátkom školení je schopný zapĺňať program aj človek, ktorý nemá žiadne poznatky o tvorbe www-stránok.

Moodle stránka KMaDG

Ste pripojený ako Nánásiová Oľga (Odhlásiť)
Slovenčina (sk)

Hlavné menu

Miestne správy

Administratíva

Kurzy

Kurzy

Matematika I.
Základy štatistiky
Štatistické metódy
Matematika II.
Statistics and Probability
Matematika 4
Teória spracovania meraní
Matematická štatistika
Ako na to
Všetky kurzy...

Miestne správy Odhlásiť z tohoto fóra

(Žiadne novinky neboli zaslané)

Toto je MOODLE stránka Katedry matematiky a deskriptívnej geometrie Slovenskej Technickej univerzity v Bratislave. Táto stránka vznikla v rámci projektu "Virtuálna univerzita" (e-learning) KEGA číslo 3/0038/02 a zatiaľ je len v experimentálnom štádiu vývoja.

Calendar

<< október 2005 >>

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						1
2	3	**4**	5	6	7	8
9	10	11	12	**13**	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Upcoming Events

prihlasenie
Tomorrow (09:10)

Go to calendar...

Článok bol podporený grantom KEGA číslo 3/0038/02

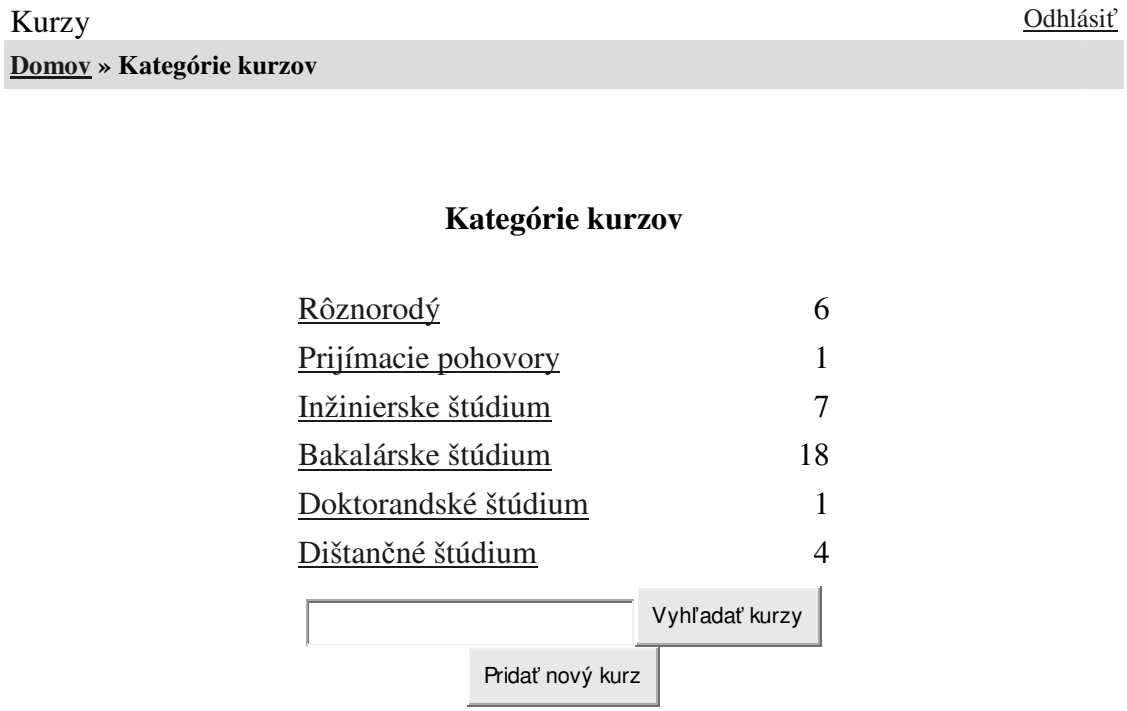
Program Moodle má základné vlastnosti, ktoré sa vyžadujú od podobných programov:

- ochrana obsahu kurzu – do kurzu sa vstupuje s kľúčom , ktorý sa zadá iba požadovanému skupine poslucháčov,
- rozdelenie veľkého počtu študentov na skupiny, práca s jednotlivými skupinami,
- komunikácia so všetkými, jednotlivými alebo skupinami študentov pomocou e-mailov,
- on line konzultácia – výhoda hlavne pri dištančnom štúdiu,
- oznamovanie výsledkov písomiek , skúšok s ochranou osobných údajov, t.z. že študent vidí len svoje výsledky a učiteľ ich môže aj komentovať, čo je zvlášť dôležité pri externom štúdiu,
- možnosť kontroly študenta, pomocou prístupov k jednotlivým zložkám kurzu, napr. k testom, študijnym materiálom atď.,
- vytváranie testov. Testy sú významný podporný nástroj výuky. Ich tvorba v týchto programoch je pomerne jednoduchá s radou výhod, ktoré by sa inak museli pracne programovať napr. zámena poradia otázok a odpovedí, náhodný výber z databáz príkladov atď.,
- prenos materiálov zo starších www-stránok je pomerne jednoduchý,
- otvorenosť týchto systémov. Dá sa naprogramovať modul aký potrebujete, pre konkrétny problém,

Na KMaDG sme sa rozhodli pracovať s týmto programom www.math.sk/moodle.

Článok bol podporený grantom KEGA číslo 3/0038/02

Na nasledujúcich obrázkoch môžete vidieť prostredie MOODLE.



Článok bol podporený grantom KEGA číslo 3/0038/02

Ľudia

- Účastníci
- Groups
- Upraviť profil

Aktivity

- Chat-y
- Fóra
- Testy
- Zadania
- Zdroje

Administratíva

- Zapnúť upravovanie
- Nastavenia...
- Učitelia...
- Študenti...
- Zálohovanie...
- Obnoviť zo zálohy...
- Stupnice...
- Známky...
- Prihlásenia...
- Súbory...
- Pomoc...
- Učiteľské fórum

Týždenný prehľad

VÍTAME VÁS V PREDMETE MATEMATIKA 2

- PRAVIDLÁ PRE ZÍSKANIE ZÁPOČTU A SKÚŠKY Z PREDMETU MATEMATIKA
- Fórum noviniek
- SYLABY
- SKÚŠKY 2002/2003
- SKÚŠKY 2003/2004
- KONZULTAČNÉ HODINY - KAMENNÉ - PREDNÁŠAJÚCI
- KONZULTAČNÉ HODINY - ON-LINE - STREDA 12.-13. HOD.
- VZORCE
- VÝMENA NA CVIČENIACH
- DOLPŇUJÚCI KURZ MATEMATIKA 2
- VYHODNOTENIE SKÚŠKY MATEMATIKA 1, 1-paralelka
- VYHODNOTENIE SKÚŠKY MATEMATIKA 1, 2-paralelka
- OPRAVY V SKRIPTÁCH MATEMATIKA I,II

1 14 február - 20 február

NEURČITÝ INTEGRÁL - ZÁKLADNÉ POJMY

- PREDNÁŠKA
- PREDNÁŠKOVÉ CVIČENIE
- CVIČENIE
- DOMÁCA ÚLOHA
- TEST INTEGRÁLY

Weeks

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Jump to current week

Článok bol podporený grantom KEGA číslo 3/0038/02

Quantian jako vědecké počítačové prostředí

Radim Remeš; Michael Rost; Roman Biskup
Zemědělská fakulta, Jihočeská univerzita, České Budějovice
inrem@zf.jcu.cz; rost@zf.jcu.cz; biskup@zf.jcu.cz

1. Živé linuxové distribuce

Živé nebo také bootovatelné distribuce linuxu se vyznačují automatickou detekcí hardwaru při zavádění systému. Většinou jsou distribuovány na CD nebo DVD médiích, ale nejsou výjimkou ani ZIP média nebo USB flash disky. Taková distribuce linuxu nevyžaduje instalovat cokoliv na pevný disk. Lze ji dokonce používat na bezdiskových stanicích. Všechny soubory, které během práce vytvoříte, se ukládají v operační paměti.

1.1. Příprava prostředí

Než je připraveno pracovní prostředí, provádí se určité iniciační akce. Tyto akce jsou prováděny ve třech krocích:

1. zavaděč systému nahraje jádro systému a dochází k iniciaci RAM disku.

2. dochází k autodetekci SCSI disků, vyhlédání a připojení CD mechanik, vytvoření symbolických linků a předání kontroly hlavnímu řídícímu procesu.

3. příprava komplexní hardwarové konfigurace (detekce hardwaru, nahrání příslušných modulů, vytvoření odkazů na zařízeních, generování konfiguračních souborů), otestování grafického zařízení a nakonec spuštění grafického uživatelského rozhraní (GUI).

Celý tento proces netrvá déle, než několik desítek sekund. Nyní je k dispozici pracovní prostředí obsahující množství aplikací pokrývající běžné pracovní spektrum, včetně rozličných nástrojů pro kancelářskou práci (od textového procesoru, přes prezentační aplikace až po profesionální grafické nástroje).

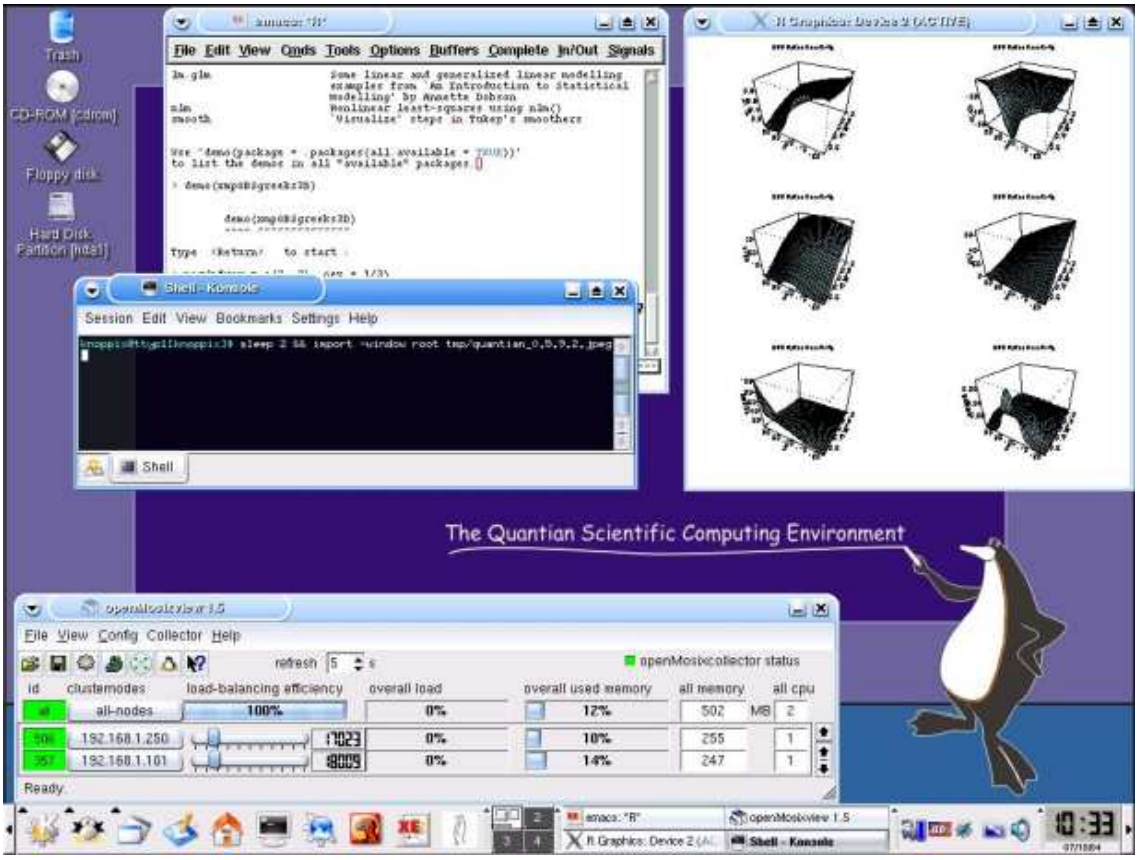
1.2. Knoppix

Knoppix je živá linuxová distribuce na CD. Jeho autor, Dipl. Ing. Klaus Knopper, jej představil v první verzi na 4. konferenci Annual Linux Showcase 2000 v Atlantě (Georgia, USA). Na 700 MB CD je velké množství programů. Veškeré programy jsou na CD zkomprimovány, při požadavku na zpřístupnění se v reálném čase dekomprimují a transparentně nahrají do paměti.

1.3. Quantian

Quantian je často označovaný jako sofistikované vědecké počítačové prostředí. Je to varianta Knoppixu konstruována speciálně pro využití s numerickou a kvantitativní analýzou. Autorem této úpravy je Dirk Eddelbuettel. Jeho poslední verze jsou derivací clusterKnoppixu, který obsahuje podporu OpenMosixu. OpenMosix je rozšíření linuxového jádra pro samostatné systémy, které mohou být zřetězeny a umožnit tak obyčejným počítačům prostřednictvím sítě vytvořit superpočítač.

V distribuci Quantian je díky kompresi uloženo na DVD téměř 5 GB aplikací, z čehož přes 3 GB dat jsou právě vědecké aplikace. Hlavní část těchto programů je zaměřena na aplikované nebo teoretické použití.



Quantian — pracovní prostředí

Z velkého souboru kvantitativních, numerických či vědeckých programů se zaměříme na statistický programovací jazyk a prostředí GNU `R`. V distribuci se nachází společně s vizualizačním programem Ggobi.

2. Prostředí R

Výhody a nevýhody tohoto prostředí do značné míry souvisí s úhlem pohledu. S prostředím `R` se komunikuje prostřednictvím příkazové řádky. To sice klade poněkud vyšší nároky na uživatele, ale křivka učení je mnohem strmější, než u převážné většiny statistických paketů. Pokud je počáteční bariéra překonána, poskytuje „*Erko*" v podstatě neomezené možnosti v oblasti statistického zpracování dat. Obrovskou výhodou je volná dostupnost tohoto prostředí, za splnění podmínek GNU Licence. Oproti tradičnímu statistickému software (Statistica, SPSS, SAS) zahrnuje `R` nejmodernější (cutting-edge) statistické metody a techniky. Zájemce si může rovněž velmi jednoduše naprogramovat vlastní metodu a okamžitě ji začít využívat. V průběhu tvorby programu lze použít princip objektově orientovaného programování. `R` disponuje velmi kvalitní grafikou a dosti podrobnou nápovědou s odkazy na literaturu.

Podívejme se na jeden školní příklad. Pokusme se, prostřednictvím metody maximální věrohodnosti, odhadnout parametry gama rozdělení z napozorovaných dat. Úkol řešme pomocí „*Erka*". Logaritmus věrohodnostní funkce můžeme v případě gama rozdělení zapsat jako:

$$l(x_1; x_2; \cdots; x_n; \alpha; \lambda) = n\alpha \ln(\lambda) - n \ln(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} x_i - \lambda \sum_{i=1}^{n} x_i$$

V programovacím prostředí existuje několik možností jak úlohu vyřešit. Například, lze využít optimalizační funkci `optim()`, nebo funkci `fitdistr()` obsaženou v knihovně autorů [4]. Další možností je využití knihovny `stats4` a vestavěné funkce `mle()`. Použijeme tedy posledně zmíněnou funkci `mle()`. Syntaxe v jazyce `R` pak může vypadat následujícím způsobem:

```
library(stats4)
verohod<-function(lambda,alfa,rozsah){
n<-rozsah
x<-data -n*alfa*log(lambda)+n*log(gamma(alfa))
-(alfa1)*sum(log(x))+lambda*sum(x)}

odhad<-mle(minuslog=verohod,start=list(lambda=2,alfa=1))
```

V průběhu iteračního procesu využívá funkce `mle()` metodu BFGS. Ta je známá rovněž pod názvem Quasi–Newtonova metoda. Výsledky získané iterací jsou uloženy do objektu s názvem `odhad`. Jako počáteční odhady parametrů byly nastaveny hodnoty $\alpha = 1$ a $\lambda = 2$. Finální hodnoty odhadů parametrů α a λ lze získat pomocí funkce `summary()`, tj. zapsáním do příkazového řádku `summary(odhad)`. Pokud bychom měli například 200 údajů, o kterých bychom předpokládali že sledují gama rozdělení, mohli bychom získat následující výstup:

```
Maximum likelihood estimation

Call:
mle(minuslogl = verohod, start = list(lambda = 2, alfa = 1))

Coefficients:
        Estimate Std. Error
lambda 0.5583953 0.05708424
alfa   4.2926668 0.41366065

-2 log L: 1059.119
```

Z výstupu je patrné, že odhady parametrů činí pro $\lambda = 0,5583953$ a $\alpha = 4,222668$. Všimněme si však také standardních chyb odhadů obou parametrů. V případě parametru λ je standardní chyba odhadu relativně malá. Zcela jiná je však situace u parametru α. Standardní chyba je v tomto případě mnohem vyšší.

3. Závěr

Prostředí `R` lze považovat za velmi flexibilní nástroj využitelný pro výuku statistiky. Svou povahou je spíše předurčen pro pokročilejší kurzy statistiky, doktorandy a výzkumné pracovníky. Využití při výuce základních kurzů však není vyloučené. V současné době zahrnuje `R` přes 300 knihoven, které umožňují využít široké množství nejmodernějších statistických metod pro různé aplikace (ekonomie, populační genetika, analýza DNA, geoinformatika, atd...). Nespornou výhodou je jeho freewarová povaha, která nezatěžuje již tak napjaté rozpočty vzdělávacích institucí v ČR.

Poděkování

Tento příspěvek vznikl za finanční podpory grantu MSM 6007665806.

Reference

1. Chambers, J., M.: Programming with Data: A guide to the S language., Springer-Verlag, New York, 1998

2. Cribari-Neto, F., Zarkos, S., G.: R: Yet econometric programming environment., Journal of applied econometric, 14: 319-329, 1999

3. Gentelman, R., Ihaka, R.: R: a language for data analysis and graphics., Journal of Computational and Graphical Statistics, 5, 299-314, 1996

4. Venables, W. N., Ripley, B.D.: Modern Applied Statistics with S-plus. 3rd edition, Springer, New York, 2000

5. R Development Core Team: R: A language and environment for statistical computing [online], c3.8.2004, [cit. 20.6.2005].
   URL: <http://www.r-project.org/>

6. Knopper, K.: Knoppix — Live Linux Filesystem on CD [online], c30.8.2001, [cit. 10.6.2005].
   URL: <http://www.knopper.net/knoppix/>.

7. Eddelbuettel, D.: The Quantian Scientific Computing Environment [online], c15.12.2004, [cit. 10.6.2005].
   URL: <http://dirk.eddelbuettel.com/quantian.html>.

Pravdepodobnosť a štatistika na základných školách

Magdaléna Renčová

Univerzita Mateja Bela,
Tajovského 40, 974 01 Banská Bystrica
rencova@fpv.umb.sk

1. Úvod

V súčasnej dobe, ktorú charakterizuje celoplošná informatizácia spoločnosti, sú na matematiku kladené úplne nové požiadavky. „Spojitá“ matematika, budovaná na reálnej osi a geometrických predstavách o objektoch, stráca svoje výsadné postavenie a rastie význam diskrétnej matematiky, teórie pravdepodobnosti, štatistiky, numerickej matematiky, logiky a teórie čísel, podľa [2].

Kombinatorika podporuje rozvoj pozornosti, flexibility, kreativity, divergentného, logického a kombinačného myslenia. Mnohé z úloh zaradených do tejto „kombinatorickej“ kategórie, majú niekoľko riešení, teda rozvíjajú divergentné rozumové operácie, ktoré súvisia s kreatívnou činnosťou. Úlohy z kombinatoriky majú veľkú motivačnú hodnotu a približujú matematiku k životu. Kombinatorika poskytuje možnosť žiakom ukázať, že aj matematika môže byť hrou. Skúsenosti detí získané pri „kombinatorickej hre“ zaiste obohatia ich vnútorný svet a snáď aj prispejú k zlepšeniu ich vzťahu k matematike, podľa [3]. Za podobné, divergentné, logické a kreatívne myslenie rozvíjajúce úlohy považujeme aj tie, ktoré patria medzi úlohy zamerané na určenie pravdepodobnosti a takisto štatistické úlohy.

Tento článok je venovaný problematike výučby kombinatoriky, pravdepodobnosti a štatistiky na základných školách. Obsahuje informácie o sledovanej problematike, počnúc obsahom pedagogických dokumentov, ako sú učebné osnovy, časovo – tematické plány a štandardy, ktoré určujú obsahovú náplň vzdelávania, končiac prieskumom skutočných zručnosti, ktoré žiak po absolvovaní základnej školy získal.

2. Tematické plány

Tematický plán učiva je školský dokument, v ktorom sa učivo vyučovacieho predmetu, predpísané učebnou osnovou, rozpisuje v časovej postupnosti na jednotlivé vyučovacie jednotky, podľa [1]. S problematikou pravdepodobnosti a štatistiky sa v časovo-tematických plánoch učiva matematiky na základných školách stretávame až v 8. a 9. ročníku. Správnemu vyriešeniu pravdepodobnostných a štatistických úloh však predchádza schopnosť riešiť kombinačné úlohy. Z tohto dôvodu je dôležité sledovať okrem tematických celkov *Kombinatorika a pravdepodobnosť, Kombinatorika, štatistika a pravdepodobnosť* aj tematické celky *Kombinatorika* a *Kombinatorika v úlohách.*

Nasledujúca tabuľka prináša výňatok z časovo - tematických plánov v jednotlivých ročníkoch druhého stupňa základných škôl venovaných skúmanej problematike:

OBSAH UČIVA	POČ.H.	VZDELÁVACÍ CIEĽ	POŽIADAVKY NA VEDOMOSTI A ZRUČNOSTI
6. ročník			
8. Kombinatorika v úlohách	**10h**		
8.1. Všetky možné usporiadania daného počtu prvkov	5h	Systematicky usporiadať daný počet prvkov všetkými možnými spôsobmi.	Vedieť systematicky usporiadať daný počet prvkov (cifier, písmen) všetkými možnými spôsobmi.
8.2. Výber a usporiadanie prvkov	5h	Vedieť z daného počtu prvkov vybrať menší počet prvkov ako daný a vybrané prvky usp.. Z daného počtu prvkov vybrať usporiadanú skupinu prvkov menšiu ako daný, určiť počet takýchto skupín prvkov.	Vedieť z daného počtu prvkov vybrať menší počet prvkov a usporiadať ich. Vedieť vypočítať kombinatorické úlohy podľa pravidla súčinu aspoň pomocou názorneho zobrazenia.
7. ročník			
11. Kombinatorika	**6h**		
11.1. Výber prvkov bez ich usporiadania	3h	Vedieť pokračovať v systéme vypisovania vš. prípadov. V rôznych úlohách nájsť spoločnú mat. podstatu.	Vedieť systematicky usporiadať daný počet (menší ako 6) prvkov (cifier, písmen, ...) všetkými možnými spôsobmi.
11.2. Ďalšie úlohy z kombinatoriky	3h	V jednotlivých úlohách objaviť spôsob tvorenia možných riešení.	Vedieť z daného počtu prvkov (menší ako 6) vybrať menší počet prvkov ako je daný počet a tieto vybraté prvky usporiadať. Určiť počet takýchto vybratých a usporiadaných prvkov
		Systematicky vytvárať všetky možné riešenia.	Vedieť z daného počtu prvkov (menší ako 6) vybrať usporiadanú skupinu prvkov menšiu ako je daný počet . Určiť počet takýchto skupín prvkov.
		Riešiť rôzne primerané kombinatorické úlohy.	
8. ročník			
9. Kombinatorika a pravdepodobnosť	**7h**		
9.1 Náhodné pokusy	1h	Získať skúsenosti v pozorovaní udalostí	Získať skúsenosti v porovnávaní rôznych udalostí z hľadiska miery ich pravdepodobnosti
9.2. Relatívna početnosť udalosti a jej výpočet	2h	Rozoznať isté, možné, neisté a nemožné udalosti	Rozoznať isté, možné, nemožné udalosti. Vedieť vypočítať relatívnu početnosť určitej udalosti
9.3. Pravdepodobnosť udalosti a jej výpočet	4h	Odhad pravdepodobnosti udalosti	Odhadnúť pravdepodobnosti udalosti

9. ročník			
8. Kombinatorika, štatistika a pravdepodobnosť	**13h**		
8.1. Riešenie rôznych kombinatorických úloh	3 h	Opakovanie kombinatorických pojmov	Vedieť systematicky usporiadať daný počet prvkov rôznymi spôsobmi, vedieť z daného počtu prvkov vybrať menší počet ako daný a usporiadať ich, vedieť z daného počtu prvkov vybrať usporiadanú skupinu prvkov.
8.2. Riešenie rôznych štatistických úloh	3 h	Opakovanie pravdepodobnostných pojmov	Získať skúsenosti v porovnávaní rôznych udalostí z hľadiska miery ich pravdepodobnosti. Rozoznať isté, možné a nemožné udalosti. Vedieť vypočítať rel. početnosť náhodnej udalosti.
8.3. Štatistický súbor, jednotka, znak, početnosť javu, aritm. priem, relatívna početnosť	3 h	Zavedenie štatistických pojmov	Vedieť zaznamenať a usporiadať údaje získané z praxe, vedieť uskutočniť jednoduché štatistické zisťovanie a výsledky zaznamenať formou tabuľky. Čítať tabuľky a vedieť interpretovať v praxi.
8.4. Riešenie úloh s pravdepodobnostnou tematikou	4 h	Precvičovanie práce s pravdepod. pojmami	Vedieť vypočítať aritmetický priemer

Celkovo je problematike kombinatorika, pravdepodobnosť a štatistika na druhom stupni základných škôl venovaný časový priestor 36 hodín z celkového počtu 792 hodín, čo predstavuje približne 4,5%.

3. Štandardy

Pod pojmom štandard rozumieme „stupeň dokonalosti požadovaný pre určitý účel alebo akceptovaný či odsúhlasený model (vzor, norma, miera), s ktorou sú reálne objekty a procesy rovnakého druhu porovnávané alebo merané“. Vzdelávací štandard predstavuje súbor požiadaviek na žiakov, úspešné zvládnutie ktorých im umožní postúpiť na ďalší stupeň vzdelávania, podľa [1].

Príklad výstupných štandardov zameraných na tému *Štatistika* predstavuje súbor požiadaviek na žiakov 9. ročníka, doplnený konkrétnymi úlohami. Žiaci 9. ročníka základnej školy by mali po ukončení 8. tematického celku v časovej dotácii od 7 do 13 hodín vedieť:

1.1 Zaznamenať a usporiadať údaje získané z praxe.

1. Zaznamenajte čo najvhodnejšie výsledky z vašich pozorovaní napríklad z premávky osobných, nákladných, dodávkových áut a bicyklov pred vašim domom alebo pred školou počas 1 hodiny.

2. Zaznamenajte výsledky 50 hodov kockou, pomocou sčítacích čiar.

1.2 Uskutočniť jednoduché štatistické zisťovanie a výsledky zapísať formou tabuľky alebo zaznamenávať stĺpcovým diagramom.

1. Spracujte učebné výsledky žiakov z vašej triedy dosiahnuté z matematiky na konci minulého roka.

1.3 Čítať tabuľky a grafy a vedieť ich interpretovať v praxi.

1. V obchodnom dome kupujúci občania nakúpili:

Nakúpili	**spotrebného tovaru za Sk**	**potravín za Sk**
1.týždeň	8 735 973	12 499
2.týždeň	20 689 750	21 763 840

a) Zistite, v ktorom týždni v mesiaci bol príjem za predaj spotrebného tovaru najvyšší? V ktorom týždni najnižší?

b) Podobné zistenia preveďte aj na príjem za predaj potravín.

2. Údaje uvedené v tabuľke znázornite stĺpcovým diagramom.

4. Monitor

V predchádzajúcich častiach sme venovali pozornosť pedagogickým dokumentom, ktoré určujú, aké zručnosti pri riešení úloh z kombinatoriky, pravdepodobnosti a štatistiky by mali získať žiaci po absolvovaní základných škôl. Jedna z foriem preverovania získaných zručností, okrem výstupných písomných prác, didaktických testov, testov ministerstva školstva preverovaných pri inšpekciách a pod. predstavujú testy z matematiky vydávané Štátnym pedagogickým ústavom v Bratislave – *MONITOR9.*

Prvý krát boli pokusne zavedené v roku 2003 a boli určené len pre špecifickú skupinu žiakov, ktorí mali podané prihlášky na štúdium na gymnáziá. Podobne boli celoštátne testovaní len žiaci, ktorí si podali prihlášky na gymnáziá aj v nasledujúcom roku. Výraznejšia zmena nastala v roku 2005, kedy sa testu zúčastnili všetci žiaci. Úlohy boli formulované tak, aby výsledky z týchto testov mohli byť zohľadnené ako jedno z kritérií, ktoré určujú stredné školy pri prijímacích pohovoroch. Z tohto dôvodu, podľa typu školy kam sa hlásili, si žiaci mohli voliť sadu úloh z dvoch variant, kde variant II predstavoval súbor testových úloh s vyššou obtiažnosťou.

Počas testovania žiakov roku 2003 sa z celkového počtu 20 úloh testovali zručnosti žiakov riešiť 3 úlohy so zameraním na kombinatoriku, pravdepodobnosť a štatistiku a boli to nasledujúce úlohy:

Úloha 1: Eva si vždy oblieka blúzku so sukňou alebo pulóver s nohavicami. Má štyri blúzky a sedem sukní, pričom každá sukňa sa hodí ku všetkým blúzkam. Má tri pulóvre a dvoje nohavíc, pričom každé nohavice sa hodia ku všetkým pulóvrom. Koľkými rôznymi spôsobmi sa môže Eva obliecť?

A. 16 **B.** 28 **C.** 34 **D.** 55 **E.** 168

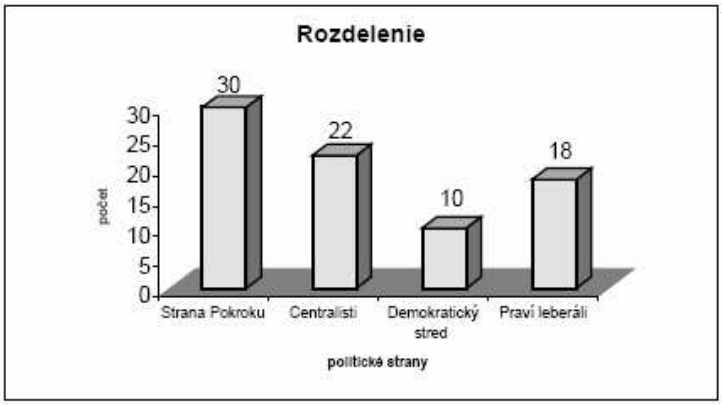
Úloha 2: V tabuľke vidíme, ako v 4.C dopadla písomná práca z matematiky. Aká je pravdepodobnosť, že náhodne vybraný chlapec zo 4.C nemá z tejto písomnej práce horšiu známku ako 2?

Známka	Počet dievčat	Počet chlapcov
1	6	7
2	2	3
3	4	2
4	1	2

A. $\frac{7}{27}$ **B.** $\frac{5}{7}$ **C.** $\frac{1}{2}$ **D.** $\frac{2}{3}$ **E.** $\frac{10}{27}$

Úloha 3: Stĺpcový diagram znázorňuje rozdelenie kresiel v 80-člennom parlamente krajiny Demoland medzi 4 politické strany. Novinár chce toto rozdelenie znázorniť kruhovým diagramom. Aká bude v tomto diagrame veľkosť uhla, ktorý bude prislúchať Strane pokroku?

Politická strana	Strana pokroku	Centralisti	Demokratický stred	Praví liberáli
Počet kresiel	30	22	10	18

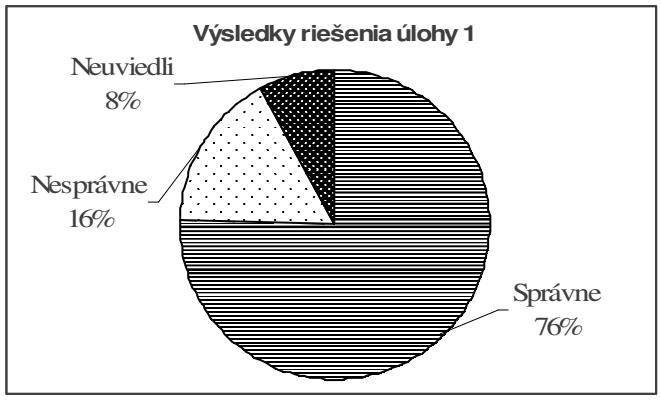


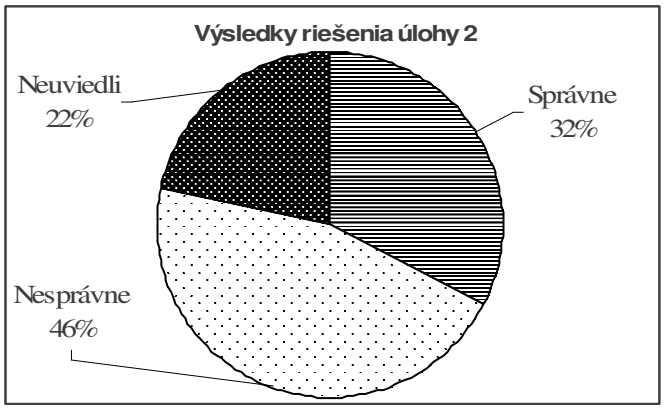
A. 135° **B.** 120° **C.** 150° **D.** 140° **E.** 125°

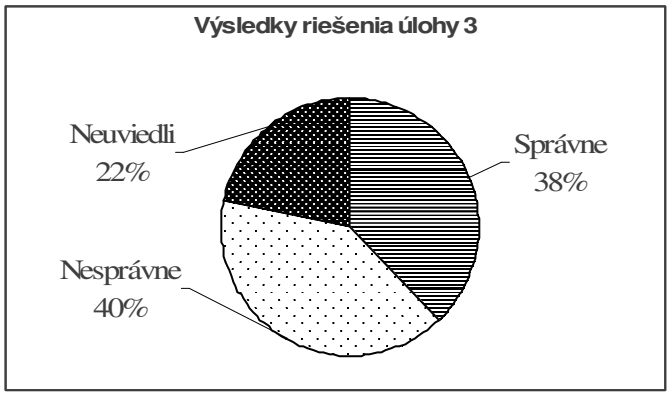
Tohto testovania sa zúčastnili aj tri triedy 9. ročníka zo Základnej školy Spojová 14, Banská Bystrica. Z celkového počtu 68 žiakov sa ho zúčastnilo 37 budúcich nádejných gymnazistov. Výsledky riešenia hore uvedených úloh prináša nasledujúca tabuľka:

	úloha 1				**úloha 2**				**úloha 3**			
trieda	**A**	**B**	**C**	**Σ**	**A**	**B**	**C**	**Σ**	**A**	**B**	**C**	**Σ**
správne	9	7	12	**28**	2	6	4	**12**	3	6	5	**14**
nesprávne	1	1	4	**6**	5	1	11	**17**	5	3	7	**15**
neuviedli	1	2	0	**3**	4	3	1	**8**	3	1	4	**8**

Kvôli názornosti uvedieme výsledky jednotlivých úloh znázornené aj kruhovými diagramami:







Na základe výsledkov pokusného zavedenia Monitoru 9 v roku 2003 je zrejmé, že nadpolovičnú úspešnosť má len úloha z kombinatoriky, úloha na výpočet pravdepodobnosti a čítanie zo štatistickej tabuľky alebo grafu má dokonca menej ako 40% úspešnosť.

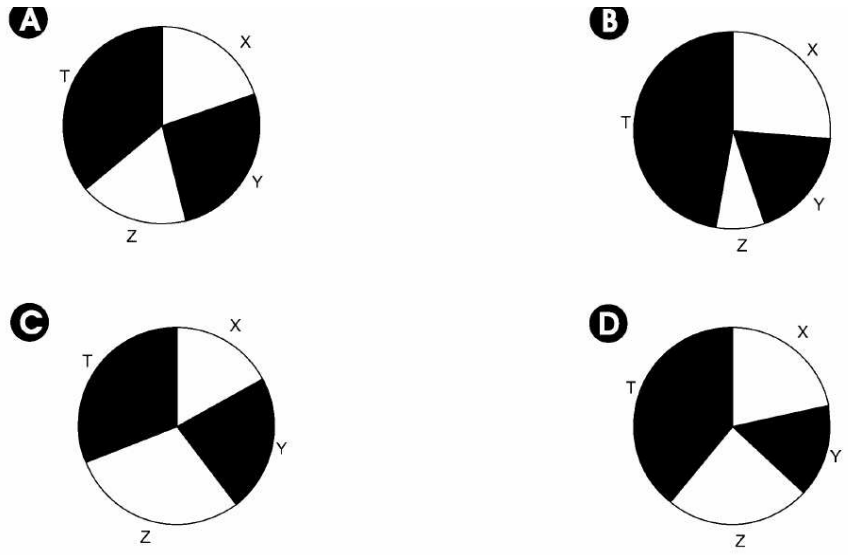
V roku 2005 riešili všetci žiaci 9. ročníka testové úlohy MONITOR 9. Testovanie žiakov bolo realizované Štátnym pedagogickým ústavom v Bratislave za pomoci Európskeho sociálneho fondu. Žiaci si vyberali z dvoch variantov, kde variant II. bol náročnejšieho charakteru. Testy pozostávali z 30 úloh, vyskytli sa v nich dve úlohy so zameraním na kombinatoriku a štatistiku, absentovala úloha pravdepodobnostného charakteru.

Spomínané úlohy sledovaného charakteru boli:
Variant I:

Úloha 1: V záhrade chceme do radu vysadiť 5 ovocných stromov, z ktorých sú tri jablone a dve hrušky. Koľkými spôsobmi ich môžeme usporiadať?

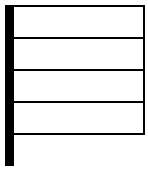
A. 6 **B.** 7 **C.** 10 **D.** 8

Úloha 2: Firma X vyrobila 6 miliónov áut. Firma Y vyrobila o 1,8 milióna kusov áut viac ako firma X. Firma Z vyrobila o 2,4 milióna kusov áut menej ako firma Y. Firma T vyrobila 10,8 miliónov kusov áut.
V ktorej z možností kruhový diagram graficky znázorňuje počet kusov áut vyrobených v týchto štyroch firmách?



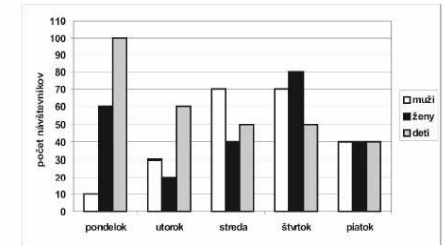
Variant II (vyššia obtiažnosť):

Úloha 1: Vlajku na obrázku tvoria 4 vodorovné pruhy. Na farebné prevedenie vlajky môžeme použiť 2-krát zelený pruh, 1-krát modrý pruh a 1-krát biely pruh. Dva zelené pruhy musia byť stále vedľa seba. Koľko je takýchto rôznych vlajok?



A. 3 **B.** 4 **C.** 6 **D.** 12

Úloha 2: Graf znázorňuje návštevnosť kina Úsmev vo vybraných dňoch. Určte rozdiel v počte návštevníkov medzi dňom s najväčšou a dňom s najmenšou návštevnosťou.



A. 30 **B.** 60 **C.** 80 **D.** 90

Je zaujímavé, že pokiaľ sledovanej problematike je v učebných plánoch vyčlenený časový priestor predstavujúci ani nie 5%, v testových úlohách MONITOR 9 bola tomuto

typu úloh venovaná až priveľká pozornosť, keď v roku 2003 tvorili 15% úloh, v roku 2004 15% a v roku 2005 bol už primeranejší priestor a to takmer 7%.

5. Záver

Jedným z hlavných cieľov súčasnej pedagogickej činnosti je uskutočňovanie krokov vedúcich k zlepšeniu pozitívneho vnímania matematiky v spoločnosti, jej zatraktívnenie a najmä dôraz na prepájanie teoretických poznatkov s využitím aplikácií v praxi. Toto je potrebné už na základných školách. Tradičné témy školskej matematiky ako aritmetika, algebra a geometria sú samozrejme dôležité a je takisto veľmi potrebné získať v nich dôležité základy. Ale iné časti matematiky, ako diskrétna matematika a štatistika si v súčasnom svete, vďaka informatizácii spoločnosti, budujú dôležité postavenie, ktoré je nutné zohľadniť pri vzdelávaní. Atraktivita týchto oblastí je v tom, že už na elementárnej školskej úrovni možno predviesť a preukázať ich užitočnosť v aplikácii reálneho života. Teda žiakom dáva možnosť skúmať procesy, ktoré majú praktický význam, čím sa učivo stáva pochopiteľnejším a atraktívnejším.

Neúspech pri riešení testových úloh z diskrétnej matematiky a štatistiky, počas MONITORu 9 v roku 2003 je možné odôvodniť nasledujúcimi faktami:

Vzhľadom k tomu, že diskrétna matematika a štatistika bola do učebných osnov zaradená v roku 1997, má veľká časť učiteľov nedostatočnú skúsenosť s vyučovaním týchto oblastí matematiky. Učitelia cítia potrebu doplniť si odborné vedomosti, ale hlavne metodiku výučby diskrétnej matematiky a štatistiky, podľa [2].

Učivo je vo všetkých ročníkoch zaradené síce ako základné učivo, ale až na konci školského roku, kedy je pozornosť žiakov už prirodzene orientovaná aj na iné oblasti. Ďalším problémom je nepostačujúci časový priestor, pripomeňme, že je mu venovaných necelých 5% z celkového času.

Ako problém môžeme uviesť, že sa z pohľadu mnohých učiteľov a následne aj žiakov javí toto učivo „menej dôležité“ a to najmä vďaka procesuálnemu spôsobu vyučovania a chápania, ktoré je odlišné od klasického konceptuálneho charakteru.

V neposlednom rade, je aj neúmerná náročnosť vybraných úloh na test z matematiky MONITOR 9, kde najmä úloha 3 bola nadštandardne náročná a dovolím si tvrdiť, že nejeden absolvent gymnázia by s jej vyriešením mal problémy.

Určite existuje ešte niekoľko možných vysvetlení, avšak aj z tých, ktoré sme uviedli je zrejmé, že úlohy z diskrétnej matematiky a štatistiky majú svoje miesto v učive matematiky základných škôl.

Literatúra

1. Turek, I.: *Tvorba a výber učiva. Vzdelávacie štandardy.* Metodické centrum Banská Bystrica (1996), s.22-26, 39-43.

2. Scholtzová, I.: *Integrácia kombinatoriky do vyučovania matematiky na základnej škole.* Metodicko-pedagogické centrum, Prešov (2004), s.3.

3. Pringerová, G.: *Kombinatorika v 6. ročníku základnej školy z hľadiska poznávacieho procesu.* In: Matematika v škole dnes a zajtra (2004), s.2.

4. *Učebné osnovy, Matematika pre 5. – 9. ročník základnej školy.* Ministerstvo školstva SR (1997).

5. *Testy MONITOR 9*. Štátny pedagogický ústav, Bratislava (2003, 2004, 2005).

Distribuované počítanie v MS .NET Framework

Martina Révayová, Csaba Török
Technická univerzita Košice, Stavebná fakulta,
Vysokoškolská 4, 042 00 Košice
Martina.Revayova@tuke.sk, Csaba.Torok@tuke.sk

Abstrakt

Článok sa venuje úlohe, ktorá využíva viaceré počítače. Komunikácia medzi počítačmi beží na platforme MS .NET. Na vytvorenie spojenia medzi počítačmi používame .NET Remoting ako jeden z nástrojov na programovanie sieťových aplikácií v .NET Framework. Prezentovaná úloha je implementovaná v MS Visual C#. V úlohe, v ktorej klient - zapisovateľ posiela na server kritickú hodnotu, a ten informuje všetkých klientov - čitateľov o sledovanej hodnote, bolo potrebné zrealizovať volanie čitateľov zo servera. Je to riešené cez spätné volanie klienta zo servera pomocou udalostí.

1. Úvod

V poslednom čase je bežné budovať aplikácie ako množinu komponentov, ktoré sú distribuované počítačovou sieťou a pracujú súčasne ako časť jedného programu. Distribuovaná aplikácia vyžaduje technológiu založenú na komponentoch a objektoch. Takou technológiou je u firmy Microsoft® DCOM (Distributed Component Object Model), u skupiny OMG (Object Management Group) je CORBA (Common Object Request Broker Architecture), alebo u firmy Sun je RMI (Remote Method Invocation) [1]. Tieto technológie, založené na komponentoch, pracujú veľmi dobre v prostredí Intranetu, ale ich použitie na Internete predstavuje značný problém. Pozrime sa teraz stručne na DCOM.

V minulosti medziprocesná komunikácia medzi aplikáciami bola riadená MS cez DCOM. DCOM dobre pracuje a výkonnosť je dostačujúca pri aplikáciách bežiacich na počítačoch podobného typu v tej istej sieti. Avšak, nedostatkom DCOM je spoliehanie sa na binárny protokol, ktorý nepodporujú všetky objektové modely. DCOM má problém pracovať cez firewall. DCOM chce komunikovať s portami, ktoré blokuje firewall. Je to možné dosiahnuť vypnutím firewall. .NET Remoting odstraňuje ťažkosti DCOM podporou rôznych formátov transportných a komunikačných protokolov.

V článku uvádzame a popíšeme kód, ktorý sme implementovali na riešenie problému, v ktorom beží komunikácia medzi jedným serverom a rôznym počtom klientov dvoch typov.

V nasledujúcej časti článku je definovaný pojem distribuované počítanie a sú popísané dva typy distribuovaných aplikácií podľa počtu klientov a serverov. Tretia časť sa venuje prostriedkom .NET Framework, ktoré umožňujú programovať distribuované aplikácie. Podrobnejšie o jednotlivých súčastiach .NET Frameworku je písané v štvrtej časti. Ďalšia časť rozoberá dva typy vzdialených objektov. Predposledná, najdôležitejšia časť popisuje riešenie úlohy s jedným serverom a viacerými klientmi. V poslednej časti sú uvedené odporúčania pre testovanie.

2. Distribuované počítanie

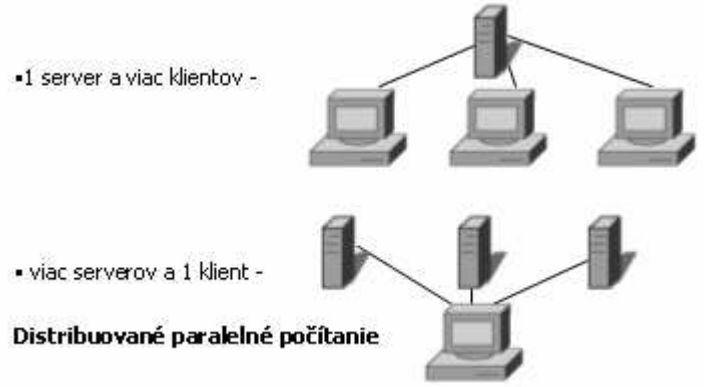
Pod pojmom distribuované počítanie rozumieme počítanie na viacerých počítačoch. Pri časovo náročných výpočtoch chceme využiť viaceré počítače, ktoré máme k dispozícii. Počítač, ktorý posiela úlohu inému počítaču nazývame klientom a počítač, ktorý úlohu počíta

Článok bol podporený grantovým projektom VEGA 1/1006/04 9150/MŠ

pre klienta nazývame serverom. Najjednoduchšia komunikácia je medzi dvoma počítačmi, keď máme jedného klienta a jeden server. Distribuované aplikácie sú teda aplikáciami typu klient – server. Podľa počtu klientov a serverov môžeme rozlíšiť dva typy úloh distribuovaného počítania:

- viac klientov a jeden server,
- jeden klient a viac serverov – tento prípad už zaraďujeme k distribuovanému paralelnému počítaniu.

Tento článok sa sústreďuje na prvý typ, lebo z hľadiska implementácie pomocou udalostí vyžaduje viac úsilia. Kým pri distribuovanom paralelnom počítaní na strane servera nemusíme používať udalosti, v nami riešenej úlohe (prvý typ) použitie udalostí na strane servera poskytuje isté výhody: menšie zaťaženie siete a klienta.



3. Distribuované aplikácie

.NET Framework poskytuje tri prostriedky na programovanie sieťových aplikácií: TCP/IP, Web Services, Remoting. TCP/IP umožňuje programovanie distribuovaných aplikácií na najnižšej úrovni pomocou posielania dát. Web Services a Remoting poskytuje programovanie distribuovaných aplikácií na vyššej úrovni. Web Services aj Remoting majú svoje výhody a nevýhody. Výhodou Web Services v porovnaní s Remoting je nezávislosť od platformy, OS. Na druhej strane pomocou Remoting môžeme využívať udalosti, rôzne formáty a protokoly. .NET Remoting poskytuje infraštruktúru pre distribuované objekty.

4. Súčasti .NET Frameworku

Pri tvorbe distribuovaných aplikácií v .NET Frameworku sú potrebné poznatky o vláknach, delegátoch a udalostiach.

4.1. Vlákna - thready

Na vykonanie aplikácie sa využíva jedno hlavné vlákno. Pri časovo náročných výpočtoch je možné uvoľniť hlavné vlákno aplikácie a výpočet robiť na pozadí v inom vlákne [2]. Vytvorenie a spustenie nového vlákna ilustruje kód:

```
// testThread.cs
class MyClass
{
   public void MyMethod() {...}
   static void Main() {
       Thread myThread = new Thread(new ThreadStart(MyMethod));
       myThread.Start();
       ...
   }
}
```

Tu metóda `MyMethod` nesmie mať žiadny vstupný argument, lebo toto štandardné použitie vlákien nedovoľuje odovzdať argumenty metóde.

4.2. Delegáti

Ak potrebujeme podať vstupné argumenty metóde, je možné namiesto vlákien použiť delegátov [2]. Delegát zastupuje metódy alebo funkcie istého typu a na delegát sa môžeme pozrieť ako na prototyp metód a funkcií. V nasledujúcom príklade delegát `ComputeDlgt` zodpovedá metódam s jedným vstupným parametrom. Inštancia `myDlgt` zastupuje metódu `MyMethod`.

```
// testDelegate.cs
delegate void ComputeDlgt(int n);
class MyClass
{
   public void MyMethod(int i) {...}
   static void Main() {
       ComputeDlgt myDlgt = new ComputeDlgt(MyMethod);
       ...
   }
}
```

Metóda `MyMethod` má jeden vstupný argument, a preto aj delegát `ComputeDlgt` je deklarovaný s jedným argumentom. Vyvolanie metódy `MyMethod` prostredníctvom delegáta môže byť synchrónne alebo asynchrónne. Pri synchrónnom spustení metódy delegáta sa metóda vykonáva sekvenčne v hlavnom vlákne aplikácie. Asynchrónne spustenie delegáta vyvolá metódu v novom vlákne.

```
...
myDlgt(5);  // <=> MyMethod(5);        synchrónne vyvolanie metódy MyMethod
myDlgt.BeginInvoke(5, null, null); //  asynchrónne vyvolanie metódy MyMethod
```

V poslednom kódovom riadku `Thread` nevystupuje – skonštruuje ho v pozadí `BeginInvoke`.

4.3. Udalosti

Udalosť je špeciálny typ delegáta. Pri práci s udalosťou rozlišujeme dve štádia: definícia a volanie udalosti na strane servera a jeho využitie na strane klienta. Pri definícii udalosti najprv sa deklaruje typ delegáta a udalosť, a potom nasleduje spätné volanie udalosti (klientovej metódy ošetrenia udalosti). Ďalej ilustrujeme dané dve štádia.
Definícia:

```
public delegate void CancelDlgt(object sender, CancelEventArgs aea);
static public event CancelDlgt Cancel;
```

Vyvolanie udalosti `Cancel`:

```
...
if (Cancel != null)
Cancel(this, new CancelEventArgs(true));      // callback
...
```

Využitie udalosti zahŕňa dva kroky: registráciu metódy na udalosť a implementáciu metódy.
Registrácia:

```
Cancel += new CancelDlgt(On_Cancel);              // CancelDlgt event registration
```

Implementácia metódy:

```
private void On_Cancel(object sender, CancelEventArgs aea)
{...}
```

5. Vzdialené objekty v Remoting

Každá aplikácia v .NET beží vo svojej aplikačnej doméne a štandardne nemôže pristupovať ku kódu/objektom aplikácie v inej aplikačnej doméne. Objekty môžu komunikovať len v rámci jednej aplikačnej domény. Vzdialený objekt je objekt nachádzajúci sa mimo aplikačnej domény buď v rámci jedného počítača alebo na rôznych počítačoch. Na komunikáciu medzi objektmi dvoch aplikačných domén môžeme použiť .NET Remoting.

Pri komunikácii medzi aplikačnými doménami rozlišujeme dva typy objektov [2]:

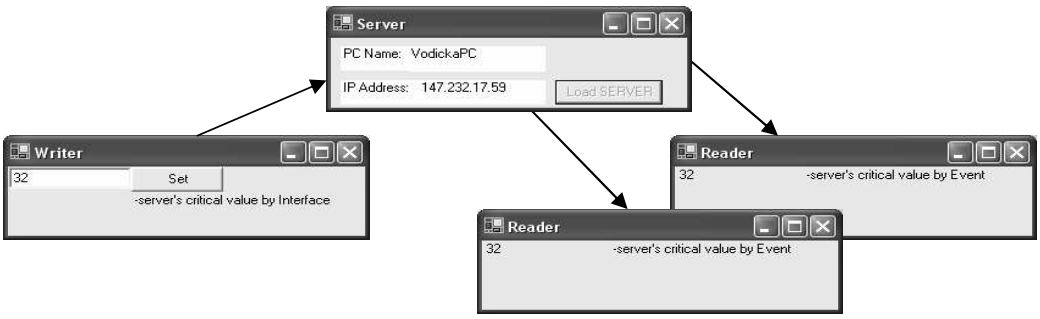
- objekty podávané odkazom
  Ak chceme pristupovať k vzdialenému objektu, objekt musí byť odvodený od triedy `System.MarshalByRefObject`, ktorá umožňuje pristupovať k objektu z inej aplikačnej domény. Aktivácia `MarshalByRefObject` objektu môže byť dvoma spôsobmi:
  - serverom aktivované objekty (SAO). Serverom aktivované objekty sú vytvorené serverom až keď sú potrebné, keď klient prvý krát volá metódu objektu.
  - klientom aktivované objekty (CAO). Klientom aktivované objekty sú vytvorené na serveri keď klient vytvorí objekt cez `new` alebo `Activator.CreateInstance`.
- objekty podávané hodnotou
  Využívame ich pri odovzdávaní parametrov metód. Tento typ objektu sa prenáša celý, vytvára sa kópia objektu. Ak chceme predať kópiu, objekt musí byť serializovaný. Štandardné typy (`int, float, double, String...`) sú serializované. Užívateľské typy je možné serializovať pridaním atribútu [Serializable] pred názov typu pri jeho implementácii. Vytvorenie serializovanej triedy vyzerá nasledovne:

```
[Serializable]
public class MyObject {
    public int n = 0;
    public String str = null;
}
```

Pri komunikácii klient (jedna aplikačná doména) pristupuje k objektom mimo aplikačnej domény – na server (druhá aplikačná doména)

6. Aplikácia v .NET Remoting

Zaoberali sme sa úlohou, keď máme jeden server a klientov dvoch rôznych typov. Prvý typ klienta writer (zapisovateľ) posiela na server sledovanú hodnotu (kritickú hodnotu - CV): najnovšiu informáciu napr. o počasí, o stave na burze. Druhý typ klienta tieto informácie zo servera získava alebo číta, nazývame ho reader (čitateľ). Túto úlohu je možné riešiť dvoma spôsobmi: buď bude iniciatívny klient alebo server. V prvom prípade klient v istých časových intervaloch žiada od servera aktuálnu hodnotu, ale nevýhodou je neustále kontrolovanie servera (pomocou nekonečného cyklu na strane klienta) a zaťaženie siete. V druhom prípade iniciátorom sa stáva server, hneď ako dostane novú hodnotu rozošle zaregistrovaným čitateľom túto hodnotu. Druhý prípad, keď iniciatívny je server, je riešený pomocou udalosti takým spôsobom, že server v momente, keď dostáva najnovšiu informáciu, pošle všetkým pripojeným čitateľom najnovšiu sledovanú hodnotu. Na tejto aplikácii ukážeme, ako sa vytvoria klienti a server, ktorí budú medzi sebou komunikovať. Pre skrátenie vysvetľovania, nepoužijeme konfiguračné súbory ako je to bežné v riešeniach dostupných na Internete.



Obrázok ilustruje komunikáciu medzi serverom, zapisovateľom a čitateľmi v našej Windows aplikácii.

Táto aplikácia bude používať dva rôzne vzdialené `MarshalByRefObject` objekty typu:

- `CVServer` – writer keď chce na serveri aktualizovať kritickú hodnotu pristupuje ku vzdialenému objektu tohto typu. Typ `CVServer` je definovaný aj inicializovaný na serveri (je odvodený od `ICriticalValue` rozhrania definovanom v Common.cs)
- `CVChanged_Wrp` - server informuje všetkých čiteteľov o nástupe novej `CVChanged` udalosti cez vzdialený objekt `CVChanged_Wrp`, definovaný v Common.cs a inicializovaný u čitateľov.

V nasledujúcich dvoch častiach popíšeme definíciu týchto dvoch `MarshalByRefObject` typov a v ďalších troch ich registráciu, inicializáciu a použitie.

6.1. Vytvorenie vzdialeného objektu `CVServer` (SAO)

Predtým ako vytvoríme vzdialený objekt `CVServer`, ktorý bude umiestnený na serveri, aby k nemu mohli pristupovať klienti, potrebujeme rozhranie `ICriticalValue`, ktorý obsahuje dve metódy a udalosť `CVChanged`. Pomocou nej server zistí všetkých zaregistrovaných čitateľov.

```
// Common.cs
using System;
namespace Common
{
    public delegate void CVChangedDlgt(String cv);
    public interface ICriticalValue
    {
        event CVChangedDlgt CVChanged;
        void setValue(String text);
        String getValue();
    }
```

Server implementuje vzdialený typ `CVServer`, ktorý je odvodený od triedy `MarshalByRefObject` a rozhrania `ICriticalValue`.

```
// Server.cs
public class CVServer: MarshalByRefObject, ICriticalValue
{
    public event Common.CVChangedDlgt CVChanged;
    String _text = "";

    public CVServer()
    {
        MessageBox.Show("ServerVector activated");
    }
    public void setValue(String text)
    {
        _text = text;
        SafeInvokeEvent(text);
    }
    ...
    public String getValue()
    {
        return _text;
    }
}
```

6.2. Vytvorenie vzdialeného objektu `CVChanged_Wrp` (CAO)

V nasledujúcich riadkoch kódu definujeme vzdialený objekt CVChanged_Wrp, ktorý je tiež odvodený od triedy `MarshalByRefObject`. Tento vzdialený objekt má udalosť `CVChanged2` a metódu `CVChanged2_Trigger`, ktorá vyvolá metódu zaregistrovanú na udalosť `CVChanged2`. Tento vzdialený objekt bude inicializovaný u čiteteľa.

```
// Common.cs
using System;
namespace Common
{
    public delegate void CVChangedDlgt(String cv);
    ...
```

```
    public class CVChanged_Wrp: MarshalByRefObject
    {
        public event CVChangedDlgt CVChanged2;
        public void CVChanged2_Trigger (String cv)
        {
            CVChanged2(cv);                    // 2. callback
        }
        public override Object InitializeLifetimeService()
        {
            return null;
        }
    }
}
```

Táto trieda `CVChanged_Wrp` je pomocná, je iba obal. Sprostredkuje komunikáciu medzi čitateľom a serverom. Najdôležitejším momentom je tu spätné volanie metódy na ošetrenie udalosti `CVChanged2` príkazom `CVChanged2(cv)`.

6.3. Zavedenie `CVServer` objektu

V predchádzajúcich dvoch častiach sme definovali `MarshalByRefObject` typy: `CVServer` a `CVChanged_Wrp`. Táto časť stručne popíše zavedenie `CVServer` objektu. Pred zavedením vzdialeného objektu na strane servera, je potrebné vytvoriť a zaregistrovať komunikačný kanál, na ktorom bude prebiehať komunikácia medzi serverom a klientmi (napr. číslo 6789 určuje voľné číslo portu).

```
// Server.cs
...
IDictionary prop = new Hashtable();
prop["port"] = 6789;
sProvider = new BinaryServerFormatterSinkProvider();
cProvider = new BinaryClientFormatterSinkProvider();
sProvider.TypeFilterLevel = TypeFilterLevel.Full;

TcpChannel chan = new TcpChannel(prop, sProvider, cProvider);
ChannelServices.RegisterChannel(chan);
```

Zavedenie a registrovanie vzdialeného objektu typu CVServer ako `WellKnownServiceType` uskutočníme príkazom:

```
RemotingConfiguration.RegisterWellKnownServiceType(
   typeof(CVServer), "CVServer.soap", WellKnownObjectMode.Singleton);
```

Prvý parameter určuje typ vzdialeného objektu, druhý parameter určuje ľubovoľné meno vzdialeného objektu (použitie prípony .soap odporúča MS), pomocou ktorého je objekt jednoznačne identifikovaný v sieti. Tretí parameter nastavuje spôsob aktivácie objektu, či objekt je vytvorený len raz pre všetky volania (`Singleton`), alebo viackrát pre každé volanie objektu (`SingleCall`).

6.4. Použitie vzdialeného objektu `CVServer` u zapisovateľa

Ak chceme u klienta používať vzdialený objekt, je potrebná registrácia komunikačného kanála aj na strane klienta. Všimnime si, že tu nie je nutné udávať číslo portu:

```
// Writer.cs
TcpChannel chan = new TcpChannel();
ChannelServices.RegisterChannel(chan);
```

Na vytvorenie inštancie použijeme `Activator.GetObject`, pričom druhý parameter udáva presné miesto, kde sa nachádza objekt s menom objektu:

```
ICriticalValue _remObj = (ICriticalValue)Activator.GetObject(
typeof(ICriticalValue), "tcp://KM_pc01.zamestnanci.svf.tuke.sk:6789/CVServer.soap");
```

Volanie metódy s aktiváciou vzdialeného objektu `CVServer` na strane klienta zapisovateľa:

```
String t = textBox1.Text;
_remObj.setValue(t);
label1.Text = _remObj.getValue();
```

Napriek tomu, že sa pristupuje ku vzdialenému objektu, dva posledné príkazy sa vykonajú synchrónne, lebo sú v jednom vlákne, nenastáva predbiehanie.

6.5. Použitie dvoch vzdialených objektov a udalosti u čitateľa

Registrácia komunikačného kanála a vytvorenie inštancie sa uskutočňuje kódom:

```
// Reader.cs
TcpChannel chan = new TcpChannel(0);
ChannelServices.RegisterChannel(chan);
_RemObj = (ICriticalValue)Activator.GetObject(typeof(ICriticalValue),
"tcp://KM_pc01.zamestnanci.svf.tuke.sk:6789/CVServer.soap");
```

Kód prezentovaný v tejto časti je snáď najdôležitejší (ale určite najzložitejší). Narába sa s dvoma rôznymi vzdialenými objektmi, dvoma udalosťami a asynchrónnym volaním metódy. Aby server vedel koľkí čitatelia sú naňho pripojení a mohol spätne vyvolať metódu klienta `On_CVChanged`, vytvoríme u čitateľa inštanciu vzdialeného objektu typu `CVChanged_Wrp`, ktorý obsahuje metódu `CVChanged2_Trigger` a udalosť `CVChanged2`. Na udalosť `CVChanged` vzdialeného objektu `CVServer` sa zaregistruje metóda `On_CVChanged` na ošetrenie udalosti `CVChanged2` vzdialeného objektu `CVChanged_Wrp`.

```
// Reader.cs
_evWrp = new CVChanged_Wrp();
_evWrp.CVChanged2 += new CVChangedDlgt(On_CVChanged2);             // see Callback 2
_remObj.CVChanged += new CVChangedDlgt(_evWrp.CVChanged2_Trigger); // see Callback 1

private void SetLabelText(String text)
{
    label1.Text = text;
    label1.Refresh();
}
private delegate void SetLabelDlgt(String s);
private void On_CVChanged2(String s)
{
    this.BeginInvoke(new SetLabelDlgt(SetLabelText), new object[]{s});
}
```

Pretože metóda `CVChanged2_Trigger` využíva udalosť `CVChanged2` treba ju zaregistrovať pred registráciou udalosti `CVChanged`. Kým metóda `On_CVChanged2` na ošetrenie udalosti `CVChanged2` je implementovaná u čitateľa, metóda `CVChanged2_Trigger` na ošetrenie udalosti `CVChanged` je implementovaná v Common.cs. Metóda `_evWrp.CVChanged2_Trigger` volá `On_CVChanged2`. V metóde `On_CVChanged2` je nebezpečné používať kód [3]:

```
label1.Text = text;
```

lebo je súčasťou vlákna, ktoré inicializoval server a riadiace prvky sú ošetrené hlavným GUI vláknom. Konštrukcia

```
this.BeginInvoke(new SetLabelDlgt(SetLabelText), new object[]{s});
```

zabezpečí volanie metódy v hlavnom GUI vlákne.

Na ukončenie celého príbehu zostalo nám ukázať kde nastáva spätné volanie u čitateľov zaregistrovaných metód na ošetrenie udalosti `CVChanged`.

6.6. Vyvolanie udalosti `CVChanged` u servera

Čiteteľa zaujímajú nové hodnoty sledovanej veličiny na serveri. Aby bol promptne informovaný o každej zmene hodnoty, mal by sa zaregistrovať na udalosť `CVChanged` (zmena kritickej hodnoty CV). Nemôže to urobiť priamo, musí to realizovať sprostredkovane cez pomocnú triedu `CVChanged_Wrp`. V časti 6.2 sme videli ako tento pomocný vzdialený objekt volá spätne metódu čitateľa. Ostalo nám iba ukázať ako server volá spätne metódu pomocného objektu.

```
// Server.cs
private void SafeInvokeEvent(String cv)
{
    if (CVChanged != null)
    {
        CVChangedDlgt delCV = null;
        foreach (Delegate del in CVChanged.GetInvocationList())
        {
            try
            {
                delCV = (CVChangedDlgt) del;
                delCV(cv);                  // 1. Callback
            }
            catch (Exception ex)
            {
                MessageBox.Show("Exception occured " + ex.Message);
                CVChanged -= delCV;
            }
        }
    }
}
```

Cez udalosť `CVChanged` sa zistia všetci zaregistrovaní čitatelia. Zoznam čitateľov získame `CVChanged.GetInvocationList()`. Vyvolanie metódy čitateľa uskutoční `delCV(cv);`

7. Skúsenosti pri testovaní

Pri zadávaní adresy vzdialeného objektu, adresa počítača musí byť jednoznačná. Zlá adresa vzdialeného objektu: `" tcp://KM_pc01:6789/CVServer.soap".`
Vhodné adresy: `"tcp://147.232.177.159:6789/CVServer.soap",`
`"tcp://KM_pc01.zamestnanci.svf.tuke.sk:6789/CVServer.soap".`
V našich podmienkach pri istých sieťových nastaveniach bolo naviac potrebné vypnúť firewall, čo pri riešení distribuovaných úloh bez udalosti, definovanej na serveri, nie je potrebné.

8. Záver

V práci sme sa venovali komunikácii medzi jedným serverom a viacerými klientami (prvý typ distribuovaného počítania). V budúcnosti plánujeme venovať sa druhému typu distribuovaného počítania – keď máme viac serverov a jedeného klienta. Tento typ je možné použiť na urýchlenie výpočtu pri časovo náročných výpočtoch: neurónové siete [4], autokorelačná funkcia pre viac tisíc dát, MIF (mutual information function – zahŕňa výpočet viacrozmerného integrálu).

Literatúra:

1. ASP.NET Web Services or .NET Remoting: How to Choose, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnbda/html/bdadotnetarch16.asp
2. Rammer I., Szpuszta M.: Advanced .NET Remoting, Apress, April 5, 2002
3. Chris Sells: Windows Forms Programming in C#, Addison Wesley, August 27, 2003.
4. Návrat P. a kol.: Umelá inteligencia. STU, Bratislava, 2002.

Niektoré konfidenčné intervaly pre spoločný stred[1]

Alexander Savin
Ústav merania Slovenskej akadémie vied
Dúbravská cesta 9, 841 04 Bratislava, Slovenská Republika
alexander.savin@gmail.com

1. Úvod

V medzilaboratórnych štúdiach sa uvažujú merania, ktoré sa vykonávajú na rovnakom objekte záujmu v rôznych laboratóriách. Laboratória môžu vykazovať rôzne medzilaboratórne chyby (heteroskedasticita). Budeme predpokladať normálne rozdelenie jednotlivých meraní a počet laboratórií budeme uvažovať za pevne daný.

Teda uvažujme model jednoduchého triedenia s pevnými efektmi:

$$Y_{ij} = \mu + \varepsilon_{ij} \tag{1}$$

s navzájom nezávislými chybami, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$, pre $i = 1, \ldots, k$ a $j = 1, \ldots, n_i$. (Neznáme) variančné komponenty σ_i^2 sú rušivé parametre: medzilaboratórne variancie. Parameter záujmu je (neznáma) stredná hodnota μ — spoločný stred.

Cieľom tohto textu je zosumarizovať a porovnať doteraz známe metódy, a odporučiť ako postupovať pri hľadaní *konfidenčného intervalu* pre spoločný stred μ s čo najlepším pokrytím a najkratšou dĺžkou.

2. Metódy

Aby sme mohli robiť štatistickú inferenciu — testovanie, konfidenčné intervaly apod., potrebujeme nájsť odhad pre spoločný stred μ. Budeme sa snažiť hľadať najlepšie lineárne nevychýlené odhady (BLUE), resp. ich odhady alebo priblíženia. V nasledujúcich častiach budeme popisovať jednotlivé metódy založené na odhadoch spoločného stredu μ za alebo bez predpokladu variančnej homogenity — homoskesticita alebo heteroskedasticita.

2.1. Homoskedastický variant

Homogenita variancií nám upravuje model nasledovne:

$$Y_{ij} = \mu + \varepsilon_{ij}, \tag{2}$$

kde $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, pre $i = 1, \ldots, k$ a $j = 1, \ldots, n_i$ sú navzájom nezávislé.

LS odhad Za predpokladu variančnej homogenity — homoskedasticity, najlepším lineárnym nevychýleným odhadom (BLUE) pre strednú hodnotu μ je odhad pomocou metódy najmeších štvorcov (LS):

$$\hat{\mu}_{LS} = \bar{Y}_n = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{Y}_{i.}, \tag{3}$$

kde $\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ a $N = \sum_{i=1}^{k} n_i$, s varianciou

$$Var\left(\bar{Y}_n\right) = \frac{\sigma^2}{N}. \tag{4}$$

[1]Článok bol podporený grantom z Vedeckej grantovej agentúry Slovenskej Republiky VEGA 1/0264/03.

Keďže varianciu odhadu nepoznáme, môžeme použiť jej odhad:

$$\widehat{Var\left(\bar{Y}_n\right)} = \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{k}\sum_{i=1}^{n_i}\left(Y_{ij}-\bar{Y}_n\right)^2$$
$$= \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{k}\left[\left(n_i-1\right)S_i^2 + n_i\left(\bar{Y}_{i.}-\bar{Y}_n\right)^2\right], \tag{5}$$

kde $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n}\left(Y_{ij}-\bar{Y}_{i.}\right)^2$. Potom štatistika T_n

$$T_n = \frac{\bar{Y}_n-\mu}{\sqrt{\widehat{Var\left(\bar{Y}_n\right)}}} \sim t_{N-1} \tag{6}$$

má t rozdelenie s $N-1$ stupňami voľnosti. Konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ je

$$CI_n = \left[\bar{Y}_n - t_{N-1;1-\frac{\alpha}{2}}\sqrt{\widehat{Var\left(\bar{Y}_n\right)}}\ ,\quad \bar{Y}_n + t_{N-1;1-\frac{\alpha}{2}}\sqrt{\widehat{Var\left(\bar{Y}_n\right)}}\right], \tag{7}$$

kde $t_{N-1;1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$ kvantil t náhodnej premennej s $N-1$ stupňami voľnosti.

2.1. Heteroskedastický variant

Porušením predpokladu o homogenite variancií, teda platí: $Var\left(\varepsilon_{ij}\right) = \sigma_i^2$, $\sigma_i^2 \neq \sigma_j^2$ pre $i \neq j$, odhad $\bar{Y}_n$ prestáva byť BLUE. Aby sme dosiahli BLUE, potrebujeme uvažovať metódu zovšeobecnených najmenších štvorcov (GLS).

GLS odhad. Vďaka porušeniu homogenity variancií, sa model (2) mení na:

$$Y_{ij} = \mu + \varepsilon_{ij}, \tag{8}$$

kde $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_i^2\right)$, pre $i = 1, \ldots, k$ a $j = 1, \ldots, n_i$ sú navzájom nezávislé.

Najlepším lineárnym nevychýleným odhadom (BLUE) je odhad založený na metóde zovšeobecnených najmenších štvorcov (GLSE):

$$\hat{\mu}_{GLS} = \bar{Y}_\omega = \left(\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\right)^{-1}\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\bar{Y}_{i.} = \frac{\sum_{i=1}^{k}\omega_i\bar{Y}_{i.}}{\sum_{i=1}^{k}\omega_i} \tag{9}$$

s varianciou

$$Var\left(\bar{Y}_\omega\right) = \left(\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\right)^{-1} = \frac{1}{\sum_{i=1}^{k}\omega_i} = \Phi. \tag{10}$$

Ak by sme poznali variančné komponenty σ_i^2, potom pivot T_ω založený na GLSE

$$T_\omega = \frac{\bar{Y}_\omega-\mu}{\sqrt{\left(\sum_{i=1}^{k}\omega_i\right)^{-1}}} \sim \mathcal{N}\left(0,1\right), \tag{11}$$

má štandardizované rozdelenie a $(1-\alpha)\,100\%$ konfidenčný interval spoľahlivosti pre spoločnú strednú hodnotu μ založený na tomto pivote

$$CI_\omega = \left[\bar{Y}_\omega - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\left(\sum_{i=1}^{k}\omega_i\right)}}\ ,\quad \bar{Y}_\omega + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\left(\sum_{i=1}^{k}\omega_i\right)}}\right], \tag{12}$$

kde $u_{1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$ kvantil štandardizovanej normálnej náhodnej premennej, je najlepší možný, nakoľko $\bar{Y}_\omega$ je v tomto prípade BLUE, teda má najmenšiu možnú varianciu.

Graybillov-Dealov odhad Lenže v našom prípade σ_i^2 nepoznáme. Dosadením nevychýlených odhadov σ_i^2 — S_i^2, získame odhad BLUE, teda približný odhad μ — Graybillov a Dealov odhad (pozri [3]):

$$\hat{\mu}_{GD} = \bar{Y}_W = \left(\sum_{i=1}^{k} \frac{n_i}{S_i^2}\right)^{-1} \sum_{i=1}^{k} \frac{n_i}{S_i^2} \bar{Y}_{i.} = \frac{\sum_{i=1}^{k} W_i \bar{Y}_{i.}}{\sum_{i=1}^{k} W_i} \tag{13}$$

s odhadom variancie

$$\widehat{Var\left(\bar{Y}_\omega\right)} = \left(\sum_{i=1}^{k} \frac{n_i}{S_i^2}\right)^{-1} = \frac{1}{\sum_{i=1}^{k} W_i}, \tag{14}$$

ktorý BLUE už nie je.

Rozdelenie štatistiky T_W

$$T_W = \frac{\bar{Y}_W - \mu}{\sqrt{\left(\sum_{i=1}^{k} W_i\right)^{-1}}} \overset{prib.}{\sim} \mathcal{N}(0,1) \tag{15}$$

nepoznáme, ale vieme, že asymptoticky konverguje k normálnemu rozdeleniu. Na základne tohto poznatku vieme zostrojiť približný konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ:

$$CI_{GD} = \left[\bar{Y}_W - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\left(\sum_{i=1}^{k} W_i\right)}} \leq \mu \leq \bar{Y}_W + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\left(\sum_{i=1}^{k} W_i\right)}}\right], \tag{16}$$

kde $u_{1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$ kvantil štandardizovanej normálnej náhodnej premennej.

Vďaka neznalosti presného rozdelenia štatistiky $\bar{Y}_W$, vzniklo veľa metód, ktoré nám ho aproximujú rôznym spôsobom. V nasledujúcich paragrafoch sa bližšie pozrieme na niektoré z nich.

Metóda Kackara a Harvillea. Kackar a Harville (1984) v [6] aproximovali varianciu pre Graybillov-Dealov odhad $\bar{Y}_W$ rozdelením na dve časti, menovite $\bar{Y}_W - \mu = \left(\bar{Y}_W - \bar{Y}_\omega\right) + \left(\bar{Y}_\omega - \mu\right)$. Vďaka nezávislosti $\bar{\boldsymbol{Y}} = \left(\bar{Y}_{1.} \ldots \bar{Y}_{k.}\right)'$ a $\boldsymbol{S^2} = (S_1^2, \ldots, S_k^2)'$ vidno, že $\bar{Y}_W - \bar{Y}_\omega$ a $\bar{Y}_\omega - \mu$ sú tiež štatisticky nezávislé. Preto platí:

$$\begin{aligned} Var\left(\bar{Y}_W\right) &= E\left(\bar{Y}_W - \mu\right)^2, \\ &= E\left(\bar{Y}_\omega - \mu\right)^2 + E\left(\bar{Y}_W - \bar{Y}_\omega\right)^2, \\ &= \Phi + E\left(\bar{Y}_W - \bar{Y}_\omega\right)^2. \end{aligned} \tag{17}$$

Vidíme, že Φ nám podhodnocuje variacniu $\bar{Y}_W$ o $E\left(\bar{Y}_W - \bar{Y}_\omega\right)^2$. Rozpísaním $\bar{Y}_W$ do Taylorovho radu v $S^2 = (S_1^2, \ldots, S_k^2)'$ okolo bodu $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)'$ dostaneme:

$$\bar{Y}_W = \bar{Y}_\omega + \sum_{i=1}^{k} \left(S_i^2 - \sigma_i^2\right) \frac{\partial \bar{Y}_\omega}{\partial \sigma_i^2} + \frac{1}{2} \sum_{i=1}^{k} \left(S_i^2 - \sigma_i^2\right)^2 \frac{\partial^2 \bar{Y}_\omega}{\partial \sigma_i^4}.$$

Použitím iba lineárneho členu v horeuvedenom rozvoji a úpravou, umocnením a zobratím strednej hodnoty dostaneme nasledujúcu aproximáciu:

$$E\left(\bar{Y}_W - \bar{Y}_\omega\right)^2 \approx E\left[\sum_{i=1}^{k}\left(S_i^2 - \sigma_i^2\right)\frac{\partial \bar{Y}_\omega}{\partial \sigma_i^2}\right]^2.$$

Vieme ukázať, že platí:

$$\frac{\partial \bar{Y}_\omega}{\partial \sigma^2} = \frac{\omega_i^2}{n_i \sum_{j=1}^{k}\omega_j}\left(\bar{Y}_\omega - \bar{Y}_{i.}\right),$$

a

$$E\left(\frac{\partial \bar{Y}_\omega}{\partial \sigma^2}\right)^2 = \frac{\omega_i^4}{n_i^2\left(\sum_{j=1}^{k}\omega_j\right)^2}\left(\frac{\sigma_i^2}{n_i} - \frac{1}{\sum_{j=1}^{k}\omega_j}\right).$$

Vďaka nezávislosti všetkých členov $\bar{\boldsymbol{Y}}$ a $\boldsymbol{S^2}$,

$$\begin{aligned} E\left(\bar{Y}_W - \bar{Y}_\omega\right)^2 &= \sum_{i=1}^{k} E\left(S_i^2 - \sigma_i^2\right)^2 E\left(\frac{\partial \bar{Y}_\omega}{\partial \sigma^2}\right)^2 \\ &\approx \sum_{i=1}^{k} Var\left(S_i^2\right)\frac{\omega_i^4}{n_i^2\left(\sum_{j=1}^{k}\omega_j\right)^2}\left(\frac{\sigma_i^2}{n_i} - \frac{1}{\sum_{j=1}^{k}\omega_j}\right) \end{aligned}$$

Inými slovami, vieme vyjadriť $E\left(\bar{Y}_W - \mu\right)^2$ približne ako

$$Var\left(\bar{Y}_W\right) \approx \frac{1}{\sum_{j=1}^{k}\omega_j} + \sum_{i=1}^{k} Var\left(S_i^2\right)\frac{\omega_i^3}{n_i^2\left(\sum_{j=1}^{k}\omega_j\right)^2}\left(1 - \frac{\omega_i}{\sum_{j=1}^{k}\omega_j}\right),$$

s prirodzeným odhadom vzniknutým dosadením odhadov S_i^2 za neznáme variančné komponenty σ_i^2:

$$\widehat{Var\left(\bar{Y}_W\right)} \approx \frac{1}{\sum_{j=1}^{k} W_j} + \sum_{i=1}^{k} \widehat{Var\left(S_i^2\right)}\frac{W_i^3}{n_i^2\left(\sum_{j=1}^{k} W_j\right)^2}\left(1 - \frac{W_i}{\sum_{j=1}^{k} W_j}\right).$$

Teraz môžeme písať nasledovnú štatistiku, ktorá asymptoticky konverguje k štandardizovanému normálnemu rozdeleniu:

$$T_{KH} = \frac{\bar{Y}_W - \mu}{\sqrt{\widehat{Var\left(\bar{Y}_W\right)}_{HK}}} \overset{prib.}{\sim} \mathcal{N}\left(0,1\right), \tag{18}$$

s

$$\widehat{Var\left(\bar{Y}_W\right)}_{KH} = \left(\sum_{i=1}^{k}\frac{W_i}{c_i}\right)^{-1} + \frac{2}{\left(\sum_{i=1}^{k} W_i\right)^2}\sum_{i=1}^{k}\frac{W_i}{f_i}\left[1 - \frac{W_i}{\sum_{i=1}^{k} W_i}\right], \tag{19}$$

kde $\widehat{Var\left(\bar{Y}_W\right)}_{KH}$ je výsledný odhad variancie s použitím nevychýleného odhadu Φ — $\left(\sum_{i=1}^{k}\frac{n_i}{c_i S_i^2}\right)^{-1}$ uvedenej Böeckenhoffovou a Hartungungom (1998) v [1], kde

$c_i = f_i/\left(f_i - 2\right)$, pre $i = 1, \ldots, k$. Štatistiku T_{KH} aproximuje pomocou t-rozdelenia so stupňami voľnosti odhadovanými podľa Sattehwaitovej metódy:

$$T_{KH} = \frac{\bar{Y}_W - \mu}{\sqrt{\widehat{Var}\left(\bar{Y}_W\right)_{KH}}} \overset{prib.}{\sim} t_{\hat{\nu}_W}, \tag{20}$$

kde

$$\hat{\nu}_W = \frac{\left(\sum_{i=1}^k W_i\right)^2}{\sum_{i=1}^k W_i^2/f_i}. \tag{21}$$

Konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ je

$$CI_{KH} = \left[\bar{Y}_W - t_{\hat{\nu}_W;1-\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\bar{Y}_W\right)_{KH}} \leq \mu \leq \bar{Y}_W + t_{\hat{\nu}_W;1-\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\bar{Y}_W\right)_{KH}}\right], \tag{22}$$

kde $t_{\hat{\nu}_W;1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$ kvantil t-náhodnej premennej s $\hat{\nu}_W$ stupňami voľnosti.

Metóda Kenwarda a Rogera. Z metódy odvodenej Kackarom a Harvilleom (1984) (v [6]) vychádzali aj Kenward a Roger (1997) (pozri [7]), kde vychýlený odhad pre Φ odhadovali pomocou Taylorovho rozvoja $\hat{\Phi}$ v $S^2 = (S_1^2, \ldots, S_k^2)'$ okolo bodu $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)'$ v strednej hodnote:

$$\begin{aligned}\frac{1}{\sum_{i=1}^k W_i} \approx \frac{1}{\sum_{i=1}^k \omega_i} &+ \sum_{i=1}^k \left(S_i^2 - \sigma_i^2\right) \frac{\partial \left(\sum_{i=1}^k W_i\right)^{-1}}{\partial \sigma_i^2} \\ &+ \frac{1}{2}\sum_{i=1}^k \left(S_i^2 - \sigma_i^2\right)^2 \frac{\partial^2 \left(\sum_{i=1}^k W_i\right)^{-1}}{\partial \sigma_i^4}.\end{aligned}$$

Teda

$$\begin{aligned}E\left[\frac{1}{\sum_{i=1}^k W_i}\right] &\approx \frac{1}{\sum_{i=1}^k \omega_i} + \frac{1}{2}\sum_{i=1}^k Var\left(S^2\right) \frac{\partial^2 \left(\sum_{i=1}^k \omega_i\right)^{-1}}{\partial \sigma_i^4}, \\ &= \frac{1}{\sum_{i=1}^k \omega_i} + \sum_{i=1}^k Var\left(S^2\right) \frac{1}{\left(\sum_{i=1}^k \omega_i\right)^2} \frac{n_i}{\sigma_i^6}\left[\frac{\omega_i}{\sum_{i=1}^k \omega_i} - 1\right].\end{aligned} \tag{23}$$

Dosadením odhadov neznámych variancií máme odhad tejto strednej hodnoty:

$$\begin{aligned}\tilde{\Phi} = E\left[\widehat{\frac{1}{\sum_{i=1}^k W_i}}\right] &\approx \frac{1}{\sum_{i=1}^k W_i} + \sum_{i=1}^k \widehat{Var}\left(S^2\right) \frac{1}{\left(\sum_{i=1}^k W_i\right)^2} \frac{n_i}{S_i^6}\left[\frac{W_i}{\sum_{i=1}^k W_i} - 1\right] \\ &= \frac{1}{\sum_{i=1}^k W_i} + \frac{2}{\left(\sum_{i=1}^k W_i\right)^2} \sum_{i=1}^k \frac{W_i}{f_i}\left[1 - \frac{W_i}{\sum_{i=1}^k W_i}\right].\end{aligned}$$

Sčítaním odhadu $\tilde{\Phi}$ a penalizácie odvodenej Kackarom a Harvilleom v [6] dostaneme nasledovnú aproximáciu variancie $\bar{Y}_W$:

$$\widehat{Var}\left(\bar{Y}_W\right)_{KR} = \frac{1}{\sum_{i=1}^k W_i} + \frac{4}{\left(\sum_{i=1}^k W_i\right)^2} \sum_{i=1}^k \frac{W_i}{f_i}\left[1 - \frac{W_i}{\sum_{i=1}^k W_i}\right]. \tag{24}$$

Nasledovnú štatistiku

$$T_{KR} = \frac{\bar{Y}_W - \mu}{\sqrt{\widehat{Var}\left(\bar{Y}_W\right)_{KR}}} \overset{prib.}{\sim} t_{\hat{\nu}_W} \tag{25}$$

aproximujeme, podobne ako v predchadzajúcej sekcii, pomocou t-rozdelenia so stupňami voľnosti odhadovanými podľa Sattehwaitovej metódy, kde

$$\hat{\nu}_W = \frac{\left(\sum_{i=1}^{k} W_i\right)^2}{\sum_{i=1}^{k} W_i^2/f_i}. \tag{26}$$

K tejto štatistike s približným t-rozdelením, dospel pre tetno špeciálny prípad aj Meier (1953) (pozri [8]). Konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ je

$$CI_{KR} = \left[\bar{Y}_W - t_{\hat{\nu}_W;1-\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\bar{Y}_W\right)_{KR}}\,,\quad \bar{Y}_W + t_{\hat{\nu}_W;1-\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\bar{Y}_W\right)_{KR}}\right], \tag{27}$$

kde $t_{\hat{\nu}_W;1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$ kvantil t-náhodnej premennej s $\hat{\nu}_W$ stupňami voľnosti.

Metódy založené na *F*-štatistikách. Hartung a Makambi (2002) v [5] postupovali iným spôsobom, neodvádzali konfidenčné intervaly aproximáciou rozdelenia odhadu $\bar{Y}_W$ ale aproximovaním rozdelenia pivotálnej štatistiky T_W^2,

$$T_W^2 = \frac{\left(\bar{Y}_W - \mu\right)^2}{\left(\sum_{i=1}^{k} W_i\right)^{-1}} \sim F_{1,\nu_{T_W^2}},$$

F-rozdelením so stupňami voľnosti 1 a $\nu_{T_W^2}$, kde stupne voľnosti $\nu_{T_W^2}$ odhadujeme momentovou metódou:

$$\nu_{T_W^2} = 2\frac{E\left(T_W^2\right)}{E\left(T_W^2\right) - 1}.$$

Keďže strednú hodnotu T_W^2 nepoznáme, odhadujeme ju. Ak rozvinieme $\sum_{i=1}^{k} W_i$ do Taylorovho radu, máme

$$\sum_{i=1}^{k} W_i \approx \sum_{i=1}^{k} \omega_i + \frac{1}{2}E\left[\sum_{i=1}^{k}\left(S_i^2 - \sigma_i^2\right)^2 \frac{\partial^2 \sum_{i=1}^{k}\omega_i}{\partial \sigma_i^4}\right].$$

Použitím vzťahu $E(X) = E_Y\left[E(X|Y)\right]$, rozvinutím a ignorovaním závislosti medzi $\hat{Y}_W$ a $\left(\sum_{i=1}^{k} W_i\right)^{-1}$ dostaneme aproximaciu strednej hodnoty T_W^2:

$$E\left(T_W^2\right) \approx 1 + \frac{1}{\sum_{i=1}^{k}\omega_i}\sum_{i=1}^{k} Var\left(S_i^2\right)\frac{\omega_i}{\sigma_i^4}.$$

Teda, štatistika T_W^2 je distribuovaná ako F-rozdelenie s 1 a $\nu_{T_W^2}$ stupňami voľnosti, kde stupne voľnosti ohdadujeme momentovou metódou:

$$\nu_{T_W^2} = 2\frac{E\left(T_W^2\right)}{E\left(T_W^2\right) - 1}.$$

Odhad $E(T_W^2)$ môžeme odhadovať pomocou ďalšieho Tayolorovho rozvoja štatistiky T_W^2 v stupňoch voľnosti opísaným v [5] s použitím nevychýleného odhadu $\left(\sum_{i=1}^{k} \frac{W_i}{c_i}\right)^{-1}$ podľa [1], kde $c_i = f_i/\left(f_i - 2\right)$:

$$\widehat{E(T_W^2)} = 1 + \left(\sum_{i=1}^{k} \frac{W_i}{c_i}\right)^{-2} \sum_{i=1}^{k} \frac{2W_i}{f_i} \left(2\sum_{i=1}^{k} W_i - W_i\right).$$

Potom odhadované stupne voľnosti sú:

$$\hat{\nu}_{HMF} = 2\frac{\widehat{E(T_W^2)}}{\widehat{E(T_W^2)} - 1}, \tag{28}$$

Konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ je

$$CI_{HMF} = \left[\bar{Y}_W - \sqrt{\frac{F_{1,\hat{\nu}_{HMF};1-\alpha}}{\sum_{i=1}^{k} W_i}}\ ,\ \ \bar{Y}_W + \sqrt{\frac{F_{1,\hat{\nu}_{HMF};1-\alpha}}{\sum_{i=1}^{k} W_i}}\right], \tag{29}$$

kde $F_{1,\hat{\nu}_{HMF};1-\alpha}$ je $(1-\alpha)$ kvantil F-náhodnej premennej s 1 a $\hat{\nu}_{HMF}$ stupňami voľnosti.

Ďalšie priblíženia — násobky *F*-rozdelení Keďže odhad $\hat{\Phi}$ nám podhodnocuje varianciu $\bar{Y}_W$, priame použitie odhadovania distribúcie T_W^2 dáva väčšie hodnoty, ako by sme očákavali. Hartung a Makambi (2002) v [5] pokračovali v ďalšej aproximácie distribúcie štatistiky T_W^2 ako násoboku F-rozdelenia:

$$\epsilon T_W^2 \sim F_{1,\nu_\epsilon}$$

Porovnaním momentov F-rozdelenia s momentmi T_W^2 dostaneme nasledujúce odhady násobku a stupňov voľnosti:

$$\epsilon = \frac{\nu_\epsilon}{\left(\nu_\epsilon - 2\right) E\left(T_W^2\right)}$$
$$\nu_\epsilon = 4 + \frac{6E^2\left(T_W^2\right)}{Var\left(T_W^2\right) - 2E^2\left(T_W^2\right)}$$

Opäť, podobným postupom ako v predhádzajúcej sekcii odhadujeme varianciu T_W^2 $Var\left(T_W^2\right)$ a jej ohdad:

$$\widehat{Var\left(T_W^2\right)} = 2\left(\sum_{i=1}^{k} \frac{W_i}{c_i}\right)^{-2} \left\{\sum_{i=1}^{k} W_i^2 + 2\left(\sum_{i=1}^{k} \frac{W_i}{c_i}\right)^{-2} \sum_{i=1}^{k} \frac{W_i^2}{f_i} \left[10\left(\sum_{i=1}^{k} W_i\right)^2 + \left(3 - 2\frac{\sum_{i=1}^{k} W_i}{W_i}\right)\sum_{i=1}^{k} W_i^2 - 8W_i \sum_{i=1}^{k} W_i\right]\right\}.$$

Odhadované stupne voľnosti sú:

$$\hat{\nu}_{HMsF} = 4 + \frac{6\widehat{E^2\left(T_W^2\right)}}{\widehat{Var\left(T_W^2\right)} - 2\widehat{E^2\left(T_W^2\right)}}, \tag{30}$$

kde $\widehat{E(T_W^2)}$ je odhad strednej hodnoty štatistiky T_W^2 uvedený vyššie. Hodnota násobiacej konštanty je

$$\hat{\epsilon} = \frac{\hat{\nu}_{HMsF}}{(\hat{\nu}_{HMsF} - 2)\,\widehat{E(T_W^2)}_2}$$

Konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ je

$$CI_{HMsF} = \left[\bar{Y}_W - \sqrt{\frac{F_{1,\hat{\nu}_{HMsF};1-\alpha}}{\hat{\epsilon}_1 \sum_{i=1}^{k} W_i}}\ ,\ \ \bar{Y}_W + \sqrt{\frac{F_{1,\hat{\nu}_{HMsF};1-\alpha}}{\hat{\epsilon}_1 \sum_{i=1}^{k} W_i}}\right], \tag{31}$$

kde $F_{1,\hat{\nu}_{HMsF};1-\alpha}$ je $(1-\alpha)$ kvantil F-náhodnej premennej s 1 a $\hat{\nu}_{HMsF}$ stupňami voľnosti.

Fairweatherova metóda. Fairweather (1972) odvodil v [2] svoju exaktnú metódu na rozdeleniach lineárnych kombinácií nezávislých t-rozdelení. V jednotlivých triedach štatistiky t_i majú

$$t_i = \frac{\bar{Y}_{i.} - \mu}{S_i}\sqrt{n_i} \sim t_{f_i}$$

t-rozdelenia s $f_i = n_i - 1$ stupňami voľnosti. Ak zoberieme lineárnú kombináciu týchto t-štatistík

$$T_F = \sum_{i=1}^{k} d_i \frac{\bar{Y}_{i.} - \mu}{S_i}\sqrt{n_i}, \tag{32}$$

kde d_i sú nenulové váhy, $\sum_{i=1}^{k} d_i = 1$, môžeme dostať po znormovaní odhad μ:

$$\hat{\mu}_{T_F} = \frac{\sum_{i=1}^{k} d_i \bar{Y}_{i.}\sqrt{W_i}}{\sum_{i=1}^{k} d_i \sqrt{W_i}}. \tag{33}$$

Konfidenčný $(1-\alpha)\,100\%$ interval spoľahlivosti pre spoločnú strednú hodnotu μ je

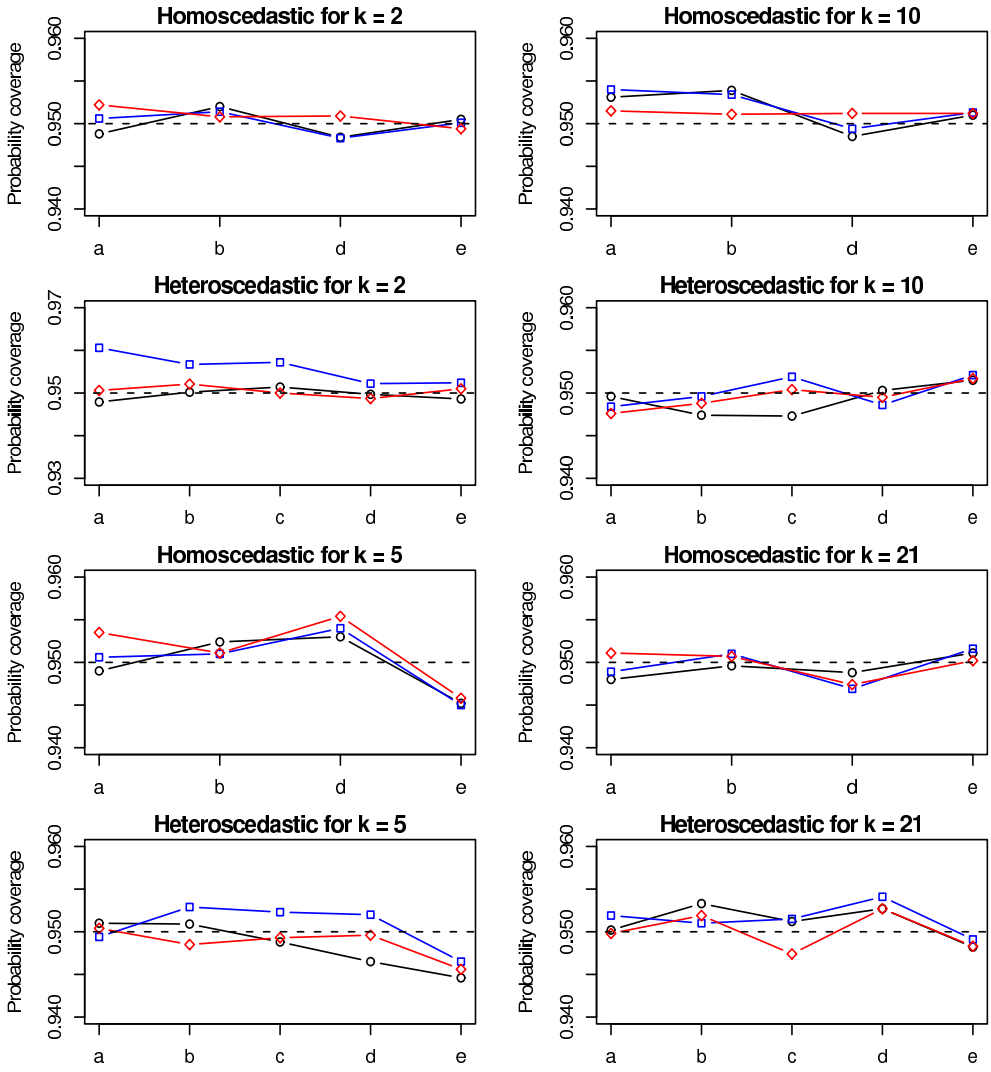
$$CI_F = \left[\hat{\mu}_{T_F} - \frac{tt_{1-\frac{\alpha}{2}}}{\sum_{i=1}^{k} d_i\sqrt{W_i}}\ ,\ \ \hat{\mu}_{T_F} + \frac{tt_{1-\frac{\alpha}{2}}}{\sum_{i=1}^{k} d_i\sqrt{W_i}}\right], \tag{34}$$

kde $tt_{\hat{\nu}_W;1-\frac{\alpha}{2}}$ je $(1-\alpha/2)$ kvantil lineárnej kombinácie nezávislých t-náhodných premenných s f_i stupňami voľnosti a váhami d_i. Presnosť lineárnej kombinácie T_F je maximálna ak sa váhy d_i zvolia inverzne k varianciam t_i:

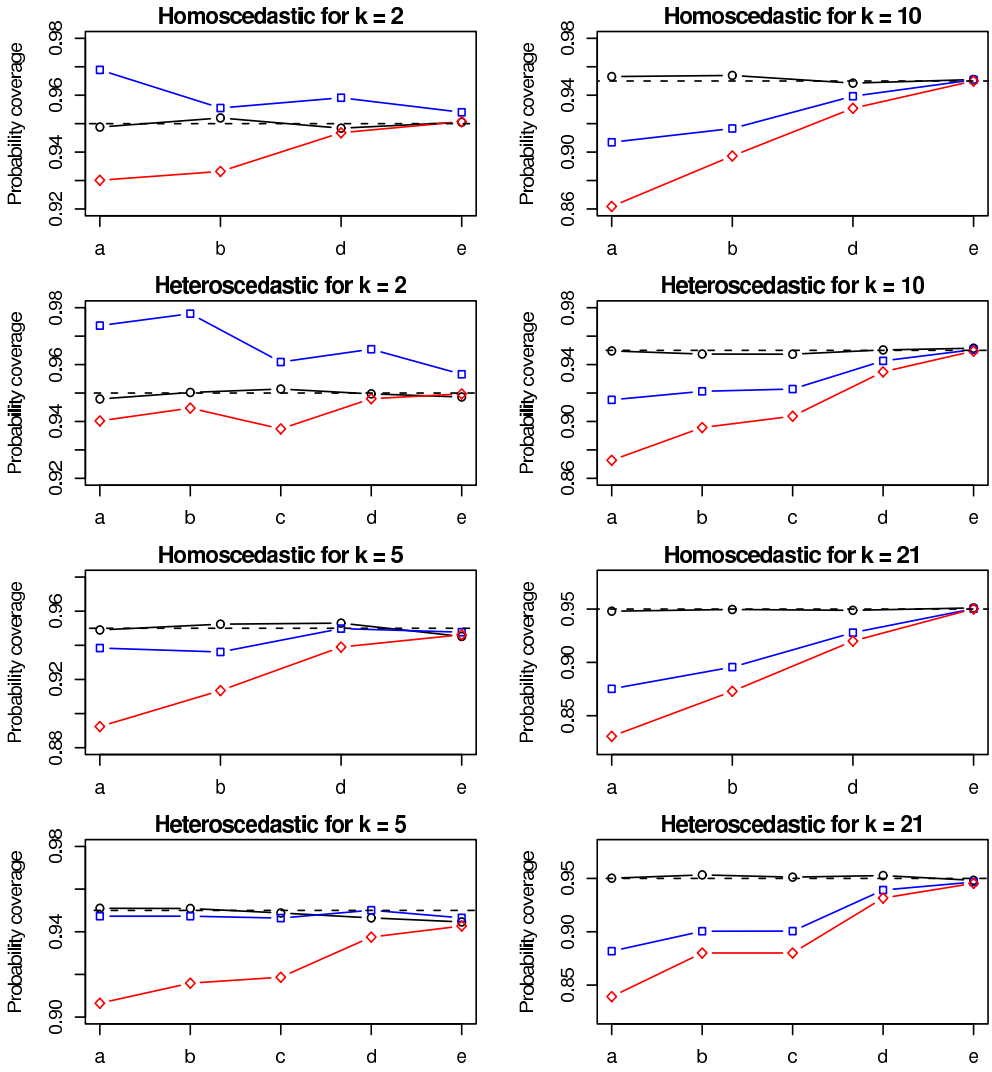
$$d_i = \frac{f_i - 2}{f_i}\left(\sum_{i=1}^{k} \frac{f_i - 2}{f_i}\right)^{-1}.$$

5. Záver

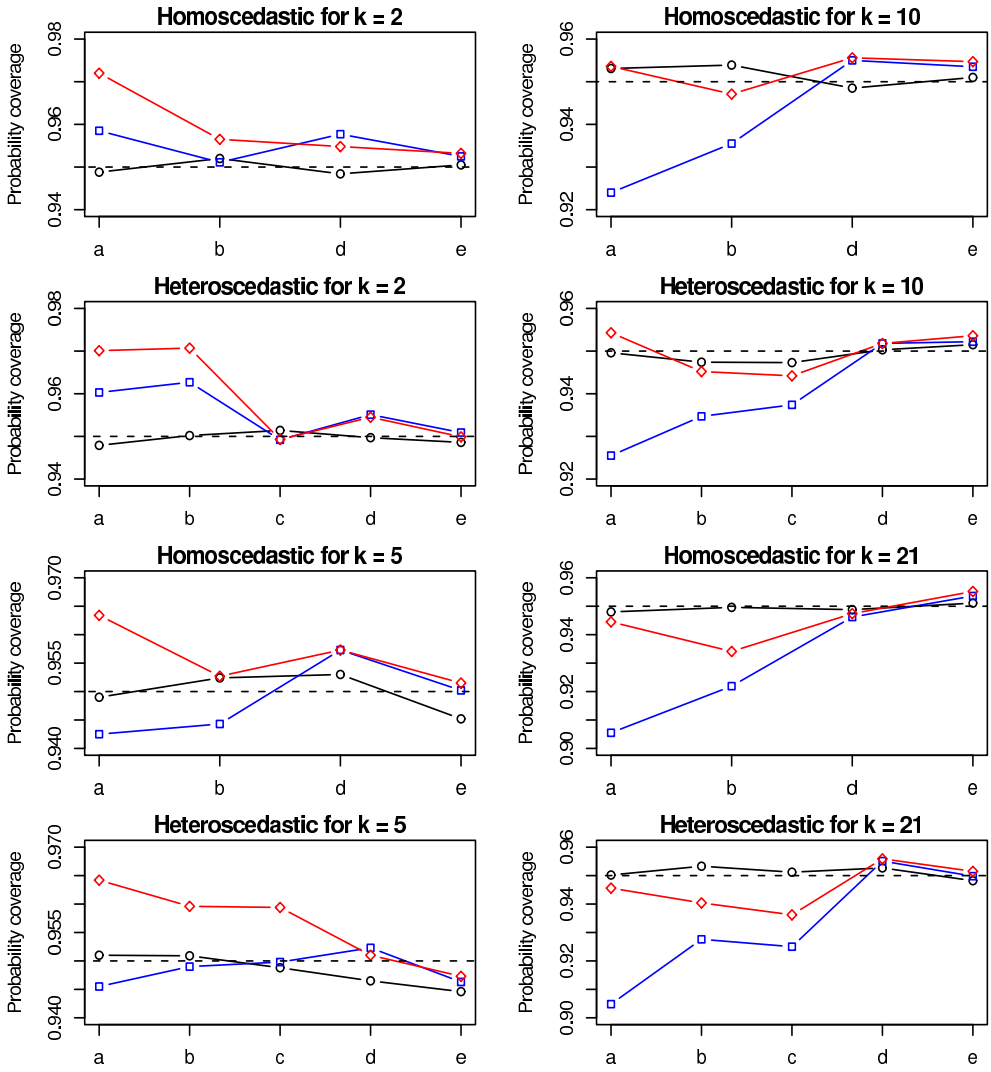
V texte sme uvažovali konfidenčné intervaly pre spoločný stred μ v medzilaboratórnych meraniach založené na rôznych metódach. V prezentovanej simulačnej štúdii sme zistili nasledujúce výsledky:



Obrázok 1: Empirické pravdepodobnosti pokrytia 95% *intervalov spoľahlivosti* založených na 10 000 Monte-Carlo opakovaniach pre každý špecifický dizajn, kde metódy sú BLUE pri známych varianciách (CI_ω) $\circ$, jednoduchá t-štatistika (CI_n) $\square$ a Fairweatherova metóda (CI_F) $\diamond$. Písmeno "a" reprezentuje, že počet laboratórii je $n_i = 5$, "b" reprezentuje $n_i = 5, 10$, "c" reprezentuje $n_i = 5, 10$ s alternatívnym typom variancií len v heteroskedastickom prípade, "d" reprezentuje $n_i = 10$ a "e" $n_i = 30$.



Obrázok 2: Empirické pravdepodobnosti pokrytia 95% *intervalov spoľahlivosti* založených na 10 000 Monte-Carlo opakovaniach pre každý špecifický dizajn, kde metódy sú BLUE pri známych varianciách (CI_ω) $\circ$, Kackarova a Hartungova metóda (CI_{KH}) $\square$, Kenwardova a Rogerova metóda (CI_{KR}) $\diamond$. Písmeno "a" reprezentuje, že počet laboratórii je $n_i = 5$, "b" reprezentuje $n_i = 5, 10$, "c" reprezentuje $n_i = 5, 10$ s alternatívnym typom variancií len v heteroskedastickom prípade, "d" reprezentuje $n_i = 10$ a "e" $n_i = 30$.



Obrázok 3: Empirické pravdepodobnosti pokrytia 95% *intervalov spoľahlivosti* založených na 10 000 Monte-Carlo opakovaniach pre každý špecifický dizajn, kde metódy sú BLUE pri známych varianciách ∘, Hartungova a Makambiho metóda založená na F-štatistike (CI_{HMF}) □ a Hartungova a Makambiho metóda založená na násobku F-štatistiky (CI_{HMsF}) ◇. Písmeno "a" reprezentuje, že počet laboratórii je $n_i = 5$, "b" reprezentuje $n_i = 5, 10$, "c" reprezentuje $n_i = 5, 10$ s alternatívnym typom variancií len v heteroskedastickom prípade, "d" reprezentuje $n_i = 10$ a "e" $n_i = 30$.

- uvažované konfidenčné intervaly založené na štatistike T_n distribuovanej jednoduchým t-rozdelením (CI_n) má veľmi dobré empirické pravdepodobnosti pokrytia pre všetky prípady. Jedinou nevýhodou, je veľká dĺžka intervalov v heteroskedastických prípadoch oproti ideálnemu konfidenčnému intervalu pri známych varančných komponentoch. Je to dané tým, lebo v týchto prípadoch aritmetický priemer $\bar{Y}_n$, na ktorom je založená štatistika T_n, prestáva byť najlepším lineárnym nevychýleným odhadom (BLUE) pre spoločný stred μ. Konfidenčné intervaly založené na lineárnej kombinácii nezávislých t-rozdelení — Fraiweatherová metóda (CI_F) má veľmi dobré empirické pravdepodobnosti pokrytia pre všetky prípady. Pozri obrázok 1.

- Intervaly založené na Kackarovej-Harvilleovej (CI_{HK}) a Kenwadovej-Rogerovej metóde (CI_{KR}) majú pre malé počty pozorovaní v triedach a pre narastajúci počet tried pravdepodobnostné pokrytie nižšie ako požadované pokrytie. Je to dané tým, že pri malých počtoch opakovaní, vygenerované dáta sa chovajú z nadmodelu, ktorý v tomto predstavuje model jednoduchej klasifikácie s náhodným efektom:
$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$
kde chyby sú navzájom nezávisle, $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$ a $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$, pre $i = 1, \ldots, k$ a $j = 1, \ldots, n_i$.
Pozri obrázok 2.

- konfidenčné intervaly založené Hartungovej a Makambiho metóde založenej F-štatistikách (CI_{HMF}, CI_{HMsF}) majú dobré empirické pravdepodobnosti pokrytia, okrem prípadov s malým počtom opakovaní v jednotlivých triedach a malým počtom tried, kde pokrytie je vyššie ako požadované. Pozri obrázok 3.

Pri uvažovaní nevyváženého a heteroskedasticikého modelu, intervaly spoľahlivosti zostrojené na rôznych priblíženiach rozdelenia GLS odhadu (Kenward&Roger, Hartung&Makambi) majú dĺžku intervalov v heteroskedastických prípadoch oproti ideálnemu konfidenčnému intervalu pri známych varančných komponentoch najnižšie, ale pravdepodobnostné pokrytie je pre narastajúci počet tried a nízky počet opakovaní v jednotlivých triedach nižší (Kenward&Roger) ako požadované pokrytie. Výborné pokrytia s pomerne dobrými dĺžkami majú intervaly zostrojene pomocou Fairweatherovej metódy aj pri rôzne nízkych počtoch opakovaní v jednotlivých triedach.

Referencie

1. Boeckenhoff, A., Hartung, J.: Some corrections of significance level in meta-analysis. Biometrical Journal, 40, (1998), pp. 937–947.

2. Fairweather, W. R.: A method of obtaining an exact confidence interval for the common mean of several normal populations. Applied Statistics, 21, (1972), pp. 229–233.

3. Graybill, F. A., Deal R. B.: Combining unbiased estimators. Biometrics, 15, (1959), pp. 543–550.

4. Hartung, J., Makambi, K. H.: Simple t-distribution based tests for meta-analysis. Preprint, (1999). Department of Statistics, University of Dortmund, `www.sfb475.uni-dortmund.de/berichte`.

5. Hartung, J., Makambi, K. H.: Alternative test procedures and confidence intervals on the common mean in the fixed effects model for meta-analysis. Preprint, (2002). Department of Statistics, University of Dortmund, www.sfb475.uni-dortmund.de/berichte.

6. Kackar, A. N., Harville, D. A.: Approximation for standart errors of estimators of fixed and random effects in linear models. Journal of the American Statistical Association, 79, (1984), pp. 853–862.

7. Kenward, M. G. and Roger, J. H.: Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 53, (1997), pp. 983–997.

8. Meier, P.: Variance of a weighted mean. Biometrics, 9, (1953), pp. 59–73.

9. Satterthwaite, E. F.: Synthesis of variance. Psychometrika, 6, (1941), pp. 309–316.

10. Savin, A., Wimmer, G. and Witkovský V.: On Kenward–Roger confidence intervals for common mean in interlaboratory trials. Measurement Science Review, Theoretical Problems Of Measurement, Vol. 3, (2003), pp. 53–56, www.measurement.sk.

11. Savin, A.: Confidence intervals for common mean in one-way classification model with fixed effects. Measurement 2005, submitted, www.measurement.sk/M2005.

12. Witkovský V.: Comparison of some Exact and Approximate Interval Estimators for Common Mean. Measurement 2005, submitted, www.measurement.sk/M2005.

De Moivreova rozdělení

Ivan Saxl[1]
Matematický ústav AV ČR, Žitná 25, 115 67 Praha 1
saxl@math.cas.cz

1. Úvod

Abraham De Moivre (1667 - 1754), od jehož smrti v loňském roce uplynulo 250 let (27. listopadu), je obecně znám především důležitým goniometrickým vztahem $(\cos\alpha + i\sin\alpha)^n = \cos n\alpha + i\sin n\alpha$ po něm pojmenovaným. Pojmenování je však demonstrací Stiglerova zákona [20], podle nějž žádný vědecký objev není pojmenován po svém původním objeviteli[2]. De Moivre tento vztah ve svých pracích sice používal a rozšířil tak jeho známost, ve výše uvedené formě jej však nikdy nenapsal. Slovní formulaci vztahu podal jako první François Viète (1540 - 1603) ve své práci *Angulares Sectiones* z roku 1570, spočetl odpovídající tabulky pro $n = 2, 3, 4, 5$ a používal jej i v dalších pracích (podrobněji viz [17]). Nynější forma vztahu i jeho důkaz jsou přičítány Rogeru Cotesovi (1682 - 1716), exponenciální formu $e^{i\alpha} = \cos\alpha + i\sin\alpha$, z níž De Moivreův vztah plyne automaticky, zavedl Leonhard Euler (*Introductio in Analysin Infinitorum*, Lausanne 1648). Stiglerův zákon se však na De Moivreovi uplatňuje ještě dvakrát, neboť by po něm správně měla být nazvána dvě významná pravděpodobnostní rozdělení: Gaussovo (normální nebo také Gaussovo-Laplaceovo) a Poissonovo, která první odvodil jako limitní případy rozdělení binomického, popsal jejich vlastnosti a používal je[3].

De Moivre se do svých čtyřiceti let zabýval převážně teorií nekonečných řad, jejich součtů a součinů, dále pak geometrií a algebrou [5, 12]. Své výsledky zveřejnil v 15 pracích publikovaných v *Philosophical Transactions of the Royal Society*. V dalších letech se z existenčních důvodů[4] rozhodl věnovat teorii pravděpodobnosti, zvláště jejímu uplatnění v hazardních hrách a pojišťovnictví. Jeho hlavními díly jsou proto knihy *The Doctrine of Chances* a *Annuities on Lives* určené širší veřejnosti; upravovány a rozšiřovány vyšly první v letech 1718, 1738 a posmrtně 1756, druhá v letech 1724, 1731 v Dublinu, 1743 a 1750. Další kniha, *Miscellanea Analytica de Seriebus et Quadraturis* z roku 1730, shrnuje jeho matematické výsledky především z teorie nekonečných řad. Široce je ovšem využíval i v pracích z oblasti pravděpodobnosti a objevují se v pozdějších vydáních *The Doctrine of Chances*.

[1]Práce byla podpořena grantem GAČR č. 201/03/0946.

[2]Stigler připouští, že zákon se vztahuje i na něj, protože za vše s jeho objevem spjaté dluží R.K. Mertonovi.

[3]V [20] je metodou nenáhodného výběru z literatury publikované v letech 1816 až 1976 zkoumán výskyt eponymického názvu normálního rozdělení s následujícími výsledky (symbol $[X]_{Y/Z}$ značí, že v rozmezí let X bylo eponymické označení v Y zkoumaných knihách nalezeno Z-krát): $[1816-1884]_{17/2}$, $[1884-1917]_{21/8}$, $[1919-1939]_{19/9}$, $[1947-1976]_{19/9}$. Ve všech těchto případech se vyskytoval název Gaussovo nebo Gaussovo-Laplaceovo rozdělení. Laplaceovo jméno se objevilo poprve v italské práci v roce 1920, poté převažuje v pracích francouzských.

[4]Abraham Moivre pocházel z francouzské hugenotské rodiny. Když po zrušení ediktu nanteského povolujícího protestantskou víru odmítl přestoupit na víru katolickou, byl tři roky vězněn. Po propuštění v roce 1688 okamžitě uprchl do Anglie, kde strávil zbývajících téměř 70 let svého života (*De* před jménem si začal psát až v Anglii). Jako cizinec však neměl právo vyučovat na universitě. Živil se proto jako soukromý učitel, docházející za svými šlechtickými žáky do jejich domovů, což bylo namahavé, časově náročné a nepříliš výnosné. V pozdějších letech si díky výsledkům své vědecké práce přilepšoval vydáváním knih a poskytováním rad hazardním hráčům a soukromým pojišťovatelům. Podrobně viz [7, 22].

Posmrtné vydání z roku 1756 obsahuje také poslední verzi knihy *Annuities on Lives* a je dnes obecně dostupné díky několikanásobnému reprintování v posledních letech.

De Moivreovy výsledky a knihy dovršily první fázi rozvoje teorie pravděpodobnosti reprezentované pracemi B. Pascala a P. Fermata (korespondence a souborné posmrtné vydání Pascalových spisů z roku 1665), Ch. Huyghense (*De ratiociniis in ludo aleae* [O výpočtech v hazardní hře], 1657), Jakoba Bernoulli (*Ars Conjectandi* [Umění otázky resp. předpokladu][5]) a P. Rémonda de Montmort (*Essai d'analyse sur le Jeux de Hasard*, 1708, rozšířené vydání 1714). Rozvoj demografie a pojišťovnictví podnítila kniha J. Graunta (*Natural and Political Observations*, 1662, úmrtnostní tabulky z údajů získaných pro Londýn) a práce E. Halleye (*An estimate of the Degrees of the Mortality of Mankind*, 1693, úmrtnostní tabulky podle matrik města Vratislavi).

2. Poissonovo rozdělení

Již v první De Moivreově práci věnované pravděpodobnostním úlohám v teorii her *De mensura sortis seu de probabilite eventuum in ludis casu fortuito pendentibus* [O míře náhody nebo o pravděpodobnosti jevu v hrách na náhodě založených], publikované ve *Philosophical Transactions* v roce 1711 (překlad do angličtiny viz Hald [10]) se vyskytují úlohy[6], v nichž De Moivre jako první dospěl k Poissonovu rozdělení $\mathrm{Po}(\lambda)$. To je limitním případem rozdělení binomického $\mathrm{Bi}(n,p)$ (n je počet opakování alternativního pokusu a p pravděpodobnost vybrané alternativy; pravděpodobnost alternativy opačné značíme q) pro případ, že $p \to 0, n \to \infty$ a $np = \lambda > 0$ je konstanta. Pravděpodobnost, že náhodná veličina X mající rozdělení $\mathrm{Po}(\lambda)$ nabude hodnoty k, je potom $\mathrm{P}_\lambda[X = k] = \frac{\lambda^k}{k!}e^{-\lambda}$.

De Moivrem řešený problém se objevuje již u Jakoba Bernoulli v *Ars Conjectandi* (to však v roce 1711 ještě nevyšlo) a zní následovně:

Mějme jev, jehož pravděpodobnost je p. *Kolik pokusů je třeba provést, aby pravděpodobnost, že jej realizujeme nejméně* c*-krát, byla rovna* $\frac{1}{2}$*?*

Triviální ilustrací je úloha hození aspoň jedné šestky opakovanými hody kostkou: musí platit $(\frac{5}{6})^n \le \frac{1}{2}$, tj. $n \ge \log 2/(-\log \frac{5}{6}) = 3.80$. Obecné úloze zřejmě odpovídá podmínka

$$\mathrm{P}[x \ge c] = \mathrm{P}[x \le c-1] = \sum_{x=0}^{c-1} \binom{n}{x} p^x q^{n-x} = \frac{1}{2},$$

kde x je počet úspěšných pokusů při n realizacích. De Moivre zavedl $r = q/p$, tj. $q^{-1} = 1 + \frac{1}{r}$ a řeší případ pro $p \to 0$, takže $r \to \infty$, přičemž zřejmě také $n \to \infty$. Pak provede následující úpravu, v níž zavede $m = \frac{n}{r} = \frac{np}{q}$ a v závěru klade $n - x \approx n$ pro $x = 1, 2, \ldots, c-1$:

$$\frac{1}{2} = \sum_{x=0}^{c-1} \binom{n}{x} p^x q^{n-x} = q^n \sum_{x=0}^{c-1} \binom{n}{x} \frac{1}{r^x} = \left(1+\frac{1}{r}\right)^{-n} \sum_{x=0}^{c-1} \binom{n}{x} \frac{1}{r^x} \doteq \left(1+\frac{m}{n}\right)^{-n} \sum_{x=0}^{c-1} \frac{m^x}{x!}.$$

V moderním zápisu bychom člen před sumací aproximovali jeho limitou pro $n \to \infty$ a napsali

$$\frac{1}{2} = \sum_{x=0}^{c-1} \mathrm{P}_m[X = x],$$

[5]Kniha byla vydána až posmrtně roku 1713, ale její rukopis pochází z počátku devadesátých let XVII. století. Její název připomíná port-royalskou Arnaudovu knihu o logickém myšlení *Ars Cogitandi* [Umění myslet] vydanou roku 1677.

[6]Čísla 5, 6 a 7, stejně ve vydání z r. 1713, jako úlohy 3, 4, 5 ve vydáních pozdějších.

kde $P_m[X = x]$ je pravděpodobnost, že náhodná veličina s rozdělením Po(m) nabývá hodnoty x. De Moivre ovšem hledá hodnoty $m = \frac{np}{q}$. Proto konstatuje, že $\left(1+\frac{m}{n}\right)^{-n}$ pro $n \to \infty$ je veličina mající hyperbolický (tj. přirozený) logaritmus $-m$, zápis typu e^{-m} tehdy ještě nebyl zaveden, a vztah zapíše ve tvaru

$$m = \ln 2 + ln\left[1 + m + \frac{m^2}{2!} + \cdots + \frac{m^{c-1}}{(c-1)!}\right].$$

Z něj pak počítal m pro $c = 1, 2, \ldots, 6$ a dostal hodnoty 0.693, 1.678 [1.67835], 2.675 [2.67405], 3.6719 [3.67206], 4.67 [4.67091] a 5.668 [5.67016] (čísla v [] jsou výsledky získané softwarem *Mathematica*). Např. pro hod tří šestek třemi kostkami dostaneme odhad počtu takových hodů pro rovnou šanci jako násobek $n = m \times \frac{q}{p} = 0.693 \times 215 \approx 149$ (DeMoivre násobí 0.7 a dostává výsledek mezi 150 a 151). Chceme-li však, aby nám tři šestky padly dvakrát s pravděpodobností $\frac{1}{2}$, potřebujeme pokusů $215 \times 1.678 \approx 361$. Lze tedy konstatovat, že De Moivre Poissonovskou aproximaci binomického rozdělení jako první odvodil i použil. Poissonův podobný postup z roku 1837 je nastíněn v [11].

3. Rovnoměrné rozdělení

Diskrétní rovnoměrné rozdělení se u De Moivrea vyskytuje implicitně v úloze o součtu n náhodně vybraných přirozených čísel z intervalu 1 až f. Jedná se o historickou úlohu řešenou po staletí, konkrétně pro počty $f = 6$ a $n = 2, 3$, které odpovídají hodům dvěma a třemi kostkami. Řešen je počet možností, jimiž lze realizovat dané celé číslo s jako součet bodů na n kostkách resp. jako součet bodů po n hodech jednou kostkou. Pro tři kostky se šesti stěnami podal řešení jako první Richard de Fournival (1190 - 1260), humanista a kancléř katedrály v Amiens, a to v básni *De vetula* [O stařence], obsahující správné kombinatorické zdůvodnění a v pozdějších tištěných vydáních básně i tabulku. Dalšími řešiteli jsou Girolamo Cardano, Galileo Galilei, Christiaan Huyghens a také Jakob Bernoulli v *Ars Conjectandi*. Obecné kombinatorické řešení podal také P. Renard de Montmort v druhém vydání své výše zmíněné knihy.

V *De Mensura Sortis* je uveden pouze algoritmus pro řešení úlohy, teprve v *Miscellanea Analytica* je důkaz, který je pak přetištěn i v dalších vydáních *The Doctrine of Chances*. De Moivreův postup je velmi důmyslný a využívá jeho poznatků o práci s mocninami polynomů; poprvé je zde zavedena později Laplacem užívaná a pojmenovaná *vytvořující funkce posloupnosti*[7] - opět jedna z De Moivreových priorit. Protože v De Moivreově době ještě nebyla užívána kombinační čísla, je důkaz poměrně dlouhý a proto jej zde reprodukuji v moderním zápise a jen nepodstatně upravený podle Halda [11]. De Moivre předpokládá, že kostka má t stěn označených 1, t^2 stěn označených 2 atd., až do t^f stěn označených f. Celkový počet stěn je tedy $t+t^2+\cdots+t^f$. Provedeme-li n hodů, budou podle pravidla o násobení pravděpodobností nezávislých jevů všechny možné případy zahrnuty v mocnině řady $g(t) = (t+t^2+\cdots+t^f)^n$. Po roznásobení dostaneme členy typu $N(s; n, f)t^s$, $s = n, n+1, \ldots, nf$, kde $N(s; n, f)$

[7] Nechť $\{p\} = \{p_0, p_1, \ldots\}$ je posloupnost reálných čísel a $g(t) = \sum_{i=0}^{\infty} p_x t^x$. Pokud $g(t)$ konverguje pro t z nějakého intervalu, nazývá se *vytvořující funkce posloupnosti* $\{p\}$; pokud je posloupnost $\{p\}$ konečná, konverguje $g(t)$ pro libovolné omezené t. Např. pro přirozené x je $g(t) = (1+t)^x$ generující funkcí posloupnosti binomických koeficientů. V teorii pravděpodobnosti se používá momentová vytvořující funkce $g(t) = \mathbf{E}e^{Xt}$, která je vytvořující funkcí posloupnosti $\{\mu'_r/r!\}$, kde μ'_r jsou obecné momenty náhodné proměnné X.

jsou hledané počty možností vytvoření bodového součtu s. De Moivre proto sečte geometrickou posloupnost součtu stěn, tu umocní a dále upraví následovně:

$$(t+t^2+\dots t^f)^n = t^n\left(\frac{1-t^f}{1-t}\right)^n = t^n\sum_{i=0}^{n}(-1)^i\binom{n}{i}t^{if}\sum_{j=0}^{\infty}\binom{n+j-1}{j}t^j,$$

kde čitatele rozepíše podle běžné binomické relace a pro jmenovatele použije její tvar pro záporný exponent $(1-x)^{-n} = \sum_{i=n-1}^{\infty}\binom{i}{n-1}x^{i-n+1}$ (viz [1], str. 822 v 7. vydání z roku 1968), v němž zavede $j = i-n+1$. Potom $s = n+if+j$, takže když $j=0$, pak i je maximální: $i = [(n-s)/f]$. Pro každé menší i je pak z nekonečného součtu vybráno jediné $j = s-n-if$, takže koeficient u t^s je

$$N(s;n,f) = \sum_{i=0}^{[(n-s)/f]}(-1)^i\binom{n}{i}\binom{s-if-1}{n-1},$$

což je hledaný počet možných součtů s. Snadno se přesvědčíme, že pro $f=6, n=3$ a $s=11,14$ dostaneme např. $N(11;3,6)=27$ a $N(14;3,6)=15$. V De Moivreově postupu bychom ovšem správně měli pracovat nikoliv s absolutními četnostmi stejně označených stěn t^i, ale s jejich četnostmi relativními, aby konvergence byla zaručena.

Součty s lze chápat jako součty nezávislých rovnoměrně rozložených náhodných veličin a musí se tedy na ně vztahovat centrální limitní teorém. Pro velká n tedy funkce $N(s;n,f)$ musí konvergovat k normálnímu rozdělení. Grafická demonstrace této skutečnosti je v knize [1] (str. 210): „normálnost“ $N(s;n,6)$ je docela dobře patrná již při $n=4$ a 8.

Jako první použil De Moivre také spojitého rovnoměrného rozložení pravděpodobnosti, a to v knize *Annuities on Lives* v příkladu 20 (vydání z roku 1756 spolu s *The Doctrine of Chances*). Tamtéž v příkladu 21 vystupují i další spojitá rozdělení pravděpodobnosti (De Moivre pro ně počítá první momenty) - viz [8], podrobně Hald v [11].

4. XX. století objevuje De Moivrea

Karl Pearson ve svých přednáškách [12] o historii statistiky v XVII. a XVIII. století na londýnské University College v roce 1922 a poté v článku [13] seznamuje veřejnost se svým zjištěním, že normální rozdělení poprvé nezavedl ani Gauss, ani Laplace, ale De Moivre. Zjistil to při studiu *Miscellanea Analytica* (1730) ze svazku ve vlastnictví University College Library, který měl dva Dodatky. První z nich obsahoval čtrnáctimístné tabulky logaritmů faktoriálů pro $n = 10(10)900$, druhým - datovaným 12. 11. 1733 - byla pak práce nazvaná "Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi Autore A.D.M.R.S.S." (písmena jsou zkratkou Abraham De Moivre, Royal Society Socius). V ní De Moivre dokazuje, že limitou binomického rozdělení pro velký počet pokusů je rozdělení, kterému dnes říkáme Gaussovo, Laplaceovo-Gaussovo či normální. Zavádí směrodatnou odchylku $\sigma = \sqrt{pqn}$, kde $p = 1-q$ je pravděpodobnost sledovaného jevu a n je počet sledování, a jako první tedy odhaluje závislost $\sigma \propto \sqrt{n}$. Sumaci pravděpodobností jednotlivých jevů nahrazuje integrály limitního rozdělení, které pak přibližně počítá pro dva konkrétní případy a dospívá k dnes dobře známému výsledku, že v rozmezích $\pm 2\sigma$ a $\pm 3\sigma$ leží 0.95428 a 0.99874 jevů (přesné hodnoty jsou 0.95450 a 0.99730). Dále se pak věnuje zpřesnění centrální limitní věty v návaznosti na výsledek Jakoba Bernoulli. Při svých výpočtech používá odhad faktoriálu velkých čísel ve tvaru

$m! = B\sqrt{m}e^{-m}m^m$, kde B počítá z rozvoje $B = 1 - \frac{1}{12} + \frac{1}{260} + \frac{1}{1260} + \frac{1}{1650} = 2.5074$, který známe jako vzorec Stirlingův s konstantou úměrnosti $B = \sqrt{2\pi} = 2.5066$.

Pearson závěrem svého příspěvku konstatuje, že jednak Laplaceovo-Gaussovo rozdělení objevil a prozkoumal De Moivre, že má prvenství také v případě Stirlingovy formule[8], a dále, že jeho centrální limitní věta značně zlepšuje formuli získanou Jakobem Bernoulli v *Ars Conjectandi*. V těchto souvislostech kritizuje I. Todhuntera, že ve své *Historii* [21] nedocenil De Moivreův přínos dostatečně a jeho objev normálního rozdělení přehlédl. Pearsonovo zjišťování přítomnosti druhého Dodatku v dalších dostupných exemplářích *Miscellanea Analytica* ukázalo, že z dvanácti prozkoumaných výtisků uložených v různých britských knihovnách byl první Dodatek v sedmi a druhý Dodatek pouze v jediném, což lze vysvětlit tím, že se mohl objevit pouze ve výtiscích prodaných po roce 1733.

V dalším článku [14] Pearson provádí podrobné srovnání obou přístupů k centrálnímu limitnímu teorému a ukazuje, že De Moivreovy odhady počtu pokusů pro dosažení požadované přesnosti odhadu pravděpodobnosti jevu jsou pro $p = 0.6$ zhruba třikrát nižší než odhady Jakoba Bernoulli. Článek pak uzavírá konstatováním, že čtvrtá kapitola *Ars Conjectandi* není tak významná, za jakou je považována[9].

Zásadní nedostatek Pearsonova vystoupení postřehl R. C. Archibald v diskusi [2, 15], která proběhla v Nature, a poté rozebral v samostatném článku [3]. K získání Pearsonem uvedených poznatků o De Moivreových objevech totiž nebylo nutné najít druhý Dodatek k *Miscellanea Analytica*, ale stačilo si jej samotným De Moivrem doslovně přeložený a o něco rozšířený přečíst ve druhém vydání *The Doctrine of Chances* z roku 1738, což ostatně Pearson ve svém příspěvku z roku 1925 uvádí a konstatuje, že s touto knihou pracoval. Ve třetím vydání téže knihy z roku 1756, je Dodatek opět zařazen a místo původních šesti a půl strany jich má jedenáct. Navíc Archibald zpochybňuje tvrzení, že se jedná o Dodatek k *Miscellanea Analytica*, protože jej De Moivre při publikaci v roce 1738 uvádí konstatováním: „Zde překládám svůj článek, který byl vytištěn 12. 11. 1733 a dán na vědomí některým přátelům, nikdy však publikován, vyhrazuje si právo své myšlenky rozšířit, jak se to bude hodit“[10]. V přeloženém a snad i v latinském textu ještě uvádí, že jsou to výsledky staré nějakých 12 let, takže je lze datovat do roku 1721.

Archibald se rovněž zastává Todhuntera a konstatuje, že De Moivreovy výsledky z řádně vytištěných Dodatků *The Doctrine of Chances* uvádí správně, a dále spolu se svým příspěvkem v Isis publikuje fotokopii faksimile latinské verze, kterou pro něj zhotovila University College Library v Londýně. Konečně v poznámce pod čarou žádá o patřičnou informaci všechny, kdo by našli další kopie této latinské verze. Ve své části diskuse připouští Pearson [16], že tvrzení o druhém Dodatku bylo možná nepatřičné, ale že pokud budou hledány další jeho kopie, pak stejně doporučuje začít s *Miscellanea Analytica*. Posledním příspěvkem do debaty je Archibaldova zpráva z Berlína [4], v níž Pearsonovi a ostatním zájemcům oznamuje nález další kopie *Approximatio* svázané s *Miscellanea Analytica* v Pruské státní knihovně.

[8] Jak upozorňuje Sheynin [18], tento názor sdílel i A.A. Markov.

[9] "The contributions of the Bernoullis to mathematical science are considerable, but they have been in more than one instance greatly exaggerated. The *Pars Quarta* of the *Ars Conjectandi* has not the importance which has often been attributed to it."

[10] "I shall here translate a Paper of mine which was printed *November* 12, 1733, and communicated to some Friends, but never made public, reserving to myself the right of enlarging my own Thoughts, as occasion shall require."

Objevení údajného latinského Dodatku k *Miscellanea Analytica*, obsahujícího zkrácenou verzi Dodatků následně řádně publikovaných a tedy teoreticky dobře známých, tak Pearsonovi vlastně jen posloužilo ke implicitnímu konstatování, že téměř dvěstě let všichni slavnou De Moivreovu knihu buď nečetli nebo četli nepozorně nebo nepochopili, co čtou (Todhunterův text, který je přesným přepisem textu De Moivreova včetně značení, tento dojem vzbuzuje). Patrně proto, že mezi *všechny* by bylo třeba zařadit i jeho samotného, zvolil Pearson tuto poněkud neobratnou formu. Za ni jej Archibald kritizoval, podstatu sdělení o De Moivreových prioritách však nijak nezpochybnil. V publikacích o historii pravděpodobnosti a statistiky jsou od té doby De Moivreovy zásluhy zmiňovány, do obecného povědomí však pronikají jen pomalu. Tím by se zdálo, že je příběh ukončen, ale není tomu tak.

Po téměř padesáti letech R. H. Daw a E. S. Pearson [8] publikují příspěvek vracející se k této problematice. V první části ji stručně, ale s větší přesností rekapitulují (např. se dozvíme, že mezi *Miscellanea Analytica* a *Approximatio* byl vevázán skutečný Dodatek nazvaný *Miscellaneis Analyticis Supplementum* s několika poznámkami patrně reagujícími na připomínky Jamese Stirlinga poslané De Moivreovi bezprostředně po vydání knihy v roce 1730).

V další části autoři popisují výsledky svého hledání dalších kopií *Approximatio*. Ještě jedna se našla v londýnské University College Library, tentokrát však svázaná s italskou zeměpisnou knihou z roku 1789; v katalogu původně nebyla uvedena a našla se patrně náhodou. Na její zadní osmé (nepotištěné) stránce byl nápis:

for Mr. Stirling

The above is an autograph of A. de Moivre

Podpis De Moivrea však chyběl a dalo se usoudit, že patrně byl odříznut při vazbě. Naproti tomu Archibaldova kopie v Pruské státní knihovně již není a nepodařilo se ji nalézt v žádné berlínské knihovně; patrně zmizela nebo byla zničena za II. světové války. Zato se podařilo najít tři další kopie, vždy svázané společně s *Miscellanea* a jejich *Supplementum* v pevném výše popsaném pořadí. První z nich v Basileji nalezl již I. Schneider [17], další se našly v Moskvě a v Petrohradě. A tak se kuriózně zdá, že přece jenom *Approximatio* prakticky - možná z rozhodnutí vydavatele - tvořilo dodatek k *Miscellanea Analytica*.

K tématu se vztahuje také práce O.B. Sheynina [18], podle autora publikovaná k třístému výročí narození De Moivrea. Je ovšem, jak z názvu vyplývá, poněkud širšího zaměření, avšak obsahuje i několik zajímavých dílčích poznatků. Výše zmíněné tabulky faktoriálů Sheynin porovnal s tabulkami současnými a zjistil, že jsou správné do jedenáctého až dvanáctého desetinného místa s jedinou výjimkou u $n = 380$, kde chyba byla na pátém desetinném místě. Volba $\sqrt{n}$ jako míry přesnosti měla původně čistě formální význam; vzdálenost od maxima (modu) rozdělení $\sigma = \sqrt{n}/2$ byla hranicí pro dva způsoby přibližné integrace funkce hustoty pravděpodobnosti

$$\int_0^\ell e^{-\frac{2x^2}{n}} dx.$$

Pro $\ell \leq \sqrt{n}/2$ používal De Moivre mocninné řady, mimo tento interval pak metodu Newtonovu-Cotesovu. Sheynin dále vyjadřuje názor, že stať *Approximatio* byla psána především s ohledem na praktické aplikace, tj. testování konkrétních experimentálně realizovaných náhodných jevů, jak dosvědčuje text předcházející Dodatek v *The Doctrine of Chances*. V něm De Moivre slibuje dále, tj. v *Approximatio*, dát návod pro případ, kdy statistický výsledek testu je v rozporu s apriorním předpok-

ladem, a je třeba rozhodnout, zda jej máme zamítnout.

Sheynin také upozorňuje na to, že prvním, kdo si všiml normálního rozdělení v De Moivreových pracích nebyl K. Pearson v roce 1924, ale v roce 1894 švýcarský statistik J. Eggenberger [9], který je málo známý, nicméně citovaný v roce 1899 nejen svým současníkem Czuberem [6], ale i o sto let později Haldem [12].

5. Normální rozdělení

Aproximaci binomického rozdělení $\mathrm{Bi}(n,p) = \binom{n}{k} p^k q^{n-k}$ pro velké n se De Moivre zabýval dlouho. V řadě úloh, zvláště tam, kde se měl spočítat počet pokusů nutných pro dosažení požadovaného rozmezí pravděpodobností, byl totiž výpočet binomických koeficientů obtížný resp. s tehdejšími prostředky prakticky nemožný. Prvním krokem je nalezení výšky jeho maxima, které pro symetrické rozdělení $p = \frac{1}{2}$ a sudé $n = 2m$ nastává při hodnotě $k = m$. Pak pro $\mathrm{Bi}(2m, \frac{1}{2})$ je

$$M(m) = \mathrm{P}[X = m] = \binom{2m}{m} \frac{1}{2^{2m}}.$$

Nejprve používal odhad založený na Stirlingově formuli odhadující $n!$ pro velká n, v roce 1730 však využívá Wallisovy formule[11]

$$\frac{\pi}{2} = \lim_{m\to\infty} \frac{[2 \times 4 \times \cdots \times (2m)]^2}{[1 \times 3 \times 5 \times \cdots \times (2m-1)]^2} \frac{1}{2m+1}$$

k získání odhadu přesného. Přepsáním $M(m)$ do tvaru

$$M(m) = \frac{2m(2m-1)\cdots\times 1}{m!m!}\frac{1}{2^m} = \frac{2m(2m-1)\cdots\times 1}{[2m(2m-2)\ldots 2]^2} = \frac{(2m-1)(2m-3)\cdots\times 1}{2m(2m-2)\ldots 2}$$

a zápisem Wallisovy formule jako $\lim_{m\to\infty} \frac{1}{M(m)^2}\frac{1}{2m+1} = \frac{\pi}{2}$, dostaneme $M(m) \approx \sqrt{\frac{2}{\pi n}}$.

Dalším nezbytným krokem k nalezení normální aproximace binomického rozdělení byl odhad poměru maxima $M(m)$ a pravděpodobnosti $\mathrm{P}[X = m+\ell] \equiv Q(\ell)$ odpovídající hodnotě X vzdálené od maxima o ℓ. Lze psát

$$\frac{M(m)}{Q(\ell)} = \frac{(m+\ell)!(m-\ell)!}{(m!)^2} = \frac{m+\ell}{m}\prod_{i=1}^{\ell-1}\frac{m+i}{m-i}.$$

Postupem založeným na údajně Newtonem odvozeném rozvoji

$$\ln\frac{1+x}{1-x} = 2\sum_{k=1}^{\infty}\frac{x^{2k-1}}{2k-1}$$

a na vztahu Jakoba Bernoulli z *Ars Conjectandi* pro konečný součet celočíselných mocnin celých čísel

$$\sum_{i=1}^{u} i^v = u^v\left(\frac{u}{v+1}+\frac{1}{2}\right) + \sum_{j=1}^{J}\frac{1}{2j}\binom{v}{2j-1}B_{2j}u^{v+1-2j},\ J = \max(j|v-2j>0),$$

[11] John Wallis (1616 - 1703), anglický polyhistor, duchovní a jeden z prvních kryptografů, ač samouk, nejvlivnější anglický matematik před Newtonem, profesor geometrie v Oxfordu od roku 1649 a jeden ze zakladatelů Royal Society. Přívrženec Cromwellův, avšak protestoval proti popravě Karla I., což mu nejen zachovalo profesorské místo, ale byl Karlem II. jmenován královským kaplanem. V *Arithmetica infinitorum* z roku 1656 je uvedena jeho slavná formule pro π, kterou odvodil při řešení integrálu z $\sqrt{1-x^2}$ od 0 do 1 (výpočet plochy čtvrtkruhu) interpolací (slovo jím zavedené) podle Cavalieriho.

kde B_{2j} jsou čísla Bernoulli rovná $\frac{1}{6}, -\frac{1}{30}, -\frac{1}{42}, \ldots$ pro $j = 1, 2, 3, \ldots$ atd., dostaneme

$$\ln \frac{M(m)}{Q(\ell)} \approx (m+\ell-\frac{1}{2}) \ln (m+\ell-1)+(m-\ell+\frac{1}{2}) \ln (m-\ell+1)-2m \ln m+\ln \frac{m+\ell}{m}.$$

Po vydání *Miscellanea Analytica* publikuje De Moivre ještě v roce 1730 dvaadvacetistránkové *Miscellaneis Analytici Supplementum*, v němž své odhady s ohledem na svůj a Stirlingův vzorec pro $n!$ ještě zpřesňuje. Získané aproximace jsou sice velmi přesné, ale pro praktické výpočty stále jen obtížně použitelné. Teprve v *Approximatio* z roku 1733 se daří De Moivreovi najít odhad jednoduchý. Při velkém m a $\ell \ll m$ totiž oba výrazy $m \pm \ell \pm \frac{1}{2}, m \pm \ell \pm 1$ konvergují k $m \pm \ell$ a po rozvinutí logaritmů v řadu s ponecháním prvních dvou členů dostane (s připomenutím historické notace)

$$\ln \frac{M(m)}{Q(\ell)} \approx (m+\ell)\left(\frac{\ell}{m} - \frac{\ell\ell}{2mm}\right) + (m-\ell)\left(-\frac{\ell}{m} - \frac{\ell\ell}{2mm}\right) = \frac{\ell\ell}{m} = \frac{2\ell\ell}{n},$$

odkud v současné notaci

$$Q(\ell) = \sqrt{\frac{2}{\pi n}} \exp\left(-\frac{2\ell^2}{n}\right) + O(1/n^2).$$

De Moivre nyní využívá jednoduchého tvaru získané aproximace a počítá celkovou pravděpodobnost soustředěnou v jistém intervalu kolem maxima m a konstatuje, že rozhodující veličinou bude $\sqrt{n}$, kterou nazývá *modulem*. Dostává

$$\mathrm{P}_\ell(t) = \sum_{|m-x|\le\ell} \mathrm{P}[X = x] = \frac{4}{\sqrt{2\pi n}} \int_0^\ell e^{-2x^2/n} dx = \frac{4}{\sqrt{2\pi}} \int_0^t e^{-2y^2} dy \qquad (*)$$

pro $\ell = t\sqrt{n}, t = \frac{1}{2}, 1, \frac{3}{2}$. Exponenciálu rozvede v řadu, integruje a dostává hodnoty 0.682688, 0.95428 a 0.99874. Přesné dnešní tabulkové hodnoty jsou 0.682689, 0.95450 a 0.99730 (větší rozdíl pro $k = \frac{3}{2}$ je zřejmě důsledek jeho početní chyby, neboť při numerické integraci provedené stejným postupem lze dostat 0.99710). To tedy znamená, že kolem maximální hodnoty $\frac{n}{2}$ je zhruba 68% výsledků pokusu soustředěno v rozmezí $\pm\sqrt{n}/2$ a 99.7% výsledků leží v rozmezí $\pm\frac{3}{2}\sqrt{n}/2$. Přesnost výsledku náhodného pokusu je tedy úměrná $\sqrt{n}$ a nikoliv počtu pokusů n. Dodatek *Approximatio* uveřejněný v následujících vydáních *The Doctrine of Chances* má ještě jednostránkový komentář (Corollary 10 začínající posledním odstavcem z *Approximatio* ve vydání 1738 a Remark I ve vydání 1756), v němž De Moivre diskutuje důsledky rovnice (*).

Bez důkazu uvádí v *Approximatio* i výsledek pro asymetrické binomické rozdělení s $p \neq \frac{1}{2}$. Označíme-li npq mající význam variance[12] σ^2 a střední hodnotu binomického rozdělení np jako μ, lze jeho výsledek v současné notaci psát

$$\lim_{n\to\infty} \mathrm{P}[X = x] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Je jistě podivné, že asymetrické rozdělení je aproximováno rozdělením symetrickým, ale když opakuje odvození se zachováním členů vyšších řadů, dostává pro $\mathrm{P}[X = x]$ navíc člen řádu $1/\sqrt{n}$ popisující asymetrii $\mathrm{P}[X = x]$ při $p \neq \frac{1}{2}$.

[12]Směrodatná odchylka a variance v dnešním významu se neužívaly nejen v době De Moivreově, ale ani v Laplaceově.

6. Závěr

Svého životního výsledku dosáhl De Moivre ve svých 66 letech a nelze se divit, že jej již dále podstatně nerozvinul. Ostatně jeho smyslem bylo řešení obtížně počítatelných úloh z teorie her a toho bylo normální aproximací vyhovujícím způsoben dosaženo. Jeho práce v této oblasti byly obecně uznávány a měly velkou popularitu. Podle Stiglera [19] jej citovala většina encyklopedií vydaných v XVIII. století včetně Encyclopedia Britannica. Pokud se jedná o jeho výsledky týkající se rozdělení dnes známého jako normální, zdá se, že odpovídající pozornost nevzbudily. Důvodem je patrně to, že aktuálním se stal inversní problém, k němuž neměly jak přispět (opět podle [19]). Jeho výsledek lze sice považovat za první úspěšnou formulaci centrálního limitního teorému, ovšem pouze pro procesy charakterizovatelné binomickým rozdělením. O jiné procesy se De Moivre nezajímal, i když, jak bylo ukázáno, jich dovedl použít, pokud to považoval za vhodné.

V dalších vydáních *The Doctrine of Chances* rozšiřoval okruh řešených úloh a dopisoval filosofické úvahy náboženské povahy, podle nichž právě získaná aproximace ukazuje, jak Boží vůlí je náhoda spoutána pevným zákonem. Ve třetím (posmrtném) vydání z roku 1756 se tak v *Approximatio* objevuje ještě Remark II, se slavnou a hojně citovanou pasáží: „Stejně jako lze ukázat, že z povahy věcí vyplývají jisté zákony, podle nichž se události dějí, tak z pozorování také vyplývá, že tyto zákony slouží moudrým a prospěšným cílům: zachovat pevný řád vesmíru, rozmnožovat některé druhy bytostí a dát jim tolik pocitu štěstí, kolik je pro jejich stav vhodné. Tyto zákony stejně jako jejich původní návrh musejí přicházet z *vnějšku*: setrvačnost věcí a povaha všeho stvoření znemožňují, aby cokoliv bylo samo schopno změnit svou vlastní podstatu. ... A odtud, pokud se nedáme oslepit metafyzickým prachem, jsme vedeni krátkou a zřejmou cestou k uznání velkého Tvůrce a Vladaře všeho, vševědoucího, všemohoucího a dobrého.“

References

1. Abramowitz, M. a Stegun, I.A.: *Handbook of Mathematical Functions.* National Bureau of Standards, Washington, D. C., 1964.

2. Archibald, R. C.: Abraham de Moivre (Letter). Nature 117 (1926), p. 551.

3. Archibald, R. C.: A Rare Pamphlet of Moivre and some of his Discoveries. Isis 8, (1926), pp. 671-683.

4. Archibald, R. C.: Abraham de Moivre (Letter). Nature 117 (1926), p. 894.

5. Cantor, M.: *Vorlesungen über Geschichte der Mathematik.* Band 3. Teubner Verlag, Leipzig, 1898, pp. 82-84.

6. Czuber, E.: Die Entwicklung der Wahrscheinlichkeits Theorie und ihrer Anwendungen. Jahresber. deutsch. Math.-Vereinigung, Vol. 7., no. 2. Teubner, Leipzig, 1899.

7. David, F. N.: *Games, Gods and Gambling, A history of probability and statistical ideas.* Dover Publ., Inc., Mineola (N.Y.), 1998 (reprint knihy z roku 1962).

8. Daw, R. H. a Pearson, E. S.: Abraham De Moivre's 1733 derivation of the normal curve: A bibliographical note. Biometrika 59 (1972), pp. 677-680.

9. Eggenberger, J.: Beiträge zur Darstellung des Bernoullisches Theorems, des Gammafunktion und des Laplaceschen Integrals. Mitt. Naturforsch. Ges. Bern 50 (1894), pp. 110-182.

10. Hald, A.: A. de Moivre: "De Mensura Sortis" or "On the Measurement of Chance". Inter. Stat. Sci. 52 (1984), pp. 229-262.

11. Hald, A.: *History of Probability and Statistics and Their Applications before 1750.* Wiley-Interscience, Haboken, New Jersey, 2003.

12. Hald, A.: *A History of Mathematical Statistics from 1750 to 1930.* John Wiley & Sons, Inc., New York, 1998.

13. Pearson, K.: Historical Note on the Origin of the Normal Curve of errors. Biometrika 16 (1924), pp. 402-404.

14. Pearson, K.: James Bernoulli's Theorem. Biometrika 17 (1925), pp. 201-210.

15. Pearson, K.: Abraham de Moivre. Reply to Professor Archibald (Letter). Nature 117 (1926), pp. 551-552.

16. Pearson, K.: *The history of Statistics in the 17th and 18th centuries*, ed. E. S. Pearson. [Přednášky K. Pearsona na University College v Londýně v letech 1921 až 1933.] Griffin, London, 1978.

17. Schneider, I.: Der Mathematiker Abraham de Moivre. Arch. Hist. Ex. Sci. 5 (1968), pp. 177-317.

18. Sheynin, O.B.: On the early history of the law of large numbers. Biometrika 55 (1968), pp. 459-467.

19. Stigler, S. M.: *The History of Statistics.* Harvard University Press, Cambridge (Mass.), 1986.

20. Stigler, S. M.: *Statistics on the Table.* Harvard University Press, Cambridge (Mass.), 1999.

21. Todhunter, I.: *A History of the Mathematical Theory of Probability.* MacMillan and Co., Cambridge and London, 1865.

22. Walker, H. M.: Abraham De Moivre, Scripta Mathematica II (1934), pp. 316-333. (Práce je systematicky přetiskována v současných reprintech třetího vydání *The Doctrine of Chances* z roku 1756.)

Chaotic Time Series

Jiří Trešl

University of Economics Prague
W.Churchill Sq.4, 130 67 Prague 3
tresl@vse.cz

1. Basic characteristics of nonlinear systems

Generally, many systems can be described with the use of first-order ordinary homogeneous differential equations [1]

(01) $$x_1^{\cdot} = f_1(x_1, x_2, \ldots, x_n) \qquad x_2^{\cdot} = f_2(x_1, x_2, \ldots, x_n) \ldots\ldots x_n^{\cdot} = f_n(x_1, x_2, \ldots, x_n)$$

where f_i are real functions of n real variables. In vector notation

(02) $$\mathbf{x}^{\cdot} = \mathbf{f(x)} \qquad \mathbf{x}(t) = \left[x_1(t), x_2(t), \ldots, x_n(t)\right] \in R^n$$

The set of equations (01) or (02) can be considered as a mathematical model of a dynamical system and **x**(t) as a system state. The space R^n is called **state space** or **phase space** of a system. To each point, a vector **f**(**x**) is assigned, starting from this point and giving the velocity of state change **x´=**d**x**/dt (**phase velocity**). The mapping $\varphi(t)$ gives the movement of a point **x**(t) and his depiction in state space is **state trajectory** or **phase trajectory.** Hereat it holds

(03) $$\varphi^{\cdot}(t) = f\left[\varphi(t)\right]$$

For example, let us consider damped harmonic oscillations

(04) $$x(t) = A\cos(\omega t + \delta) \qquad x^{\cdot}(t) = v(t) = -A\omega\sin(\omega t + \delta)$$

Eliminating time, we obtain an ellipse as state trajectory

(05) $$\frac{x^2}{A^2} + \frac{v^2}{(A\omega)^2} = 1$$

To each value of amplitude A, certain trajectory is assigned and neighbouring trajectories do not intersect.

From the point of view of terms, we can distinguish:

$\mathbf{x}^{\cdot} = \mathbf{f}(\mathbf{x})$	autonomous system	(not excited, time invariable)
$\mathbf{x}^{\cdot} = \mathbf{f}(t, \mathbf{x})$	non-autonomous system	(not excited, time variable)
$\mathbf{x}^{\cdot} = \mathbf{f}(\mathbf{x}, \mathbf{u})$	non-autonomous system	(excited, time invariable)
$\mathbf{x}^{\cdot} = \mathbf{f}(t, \mathbf{x}, \mathbf{u})$	non-autonomous system	(excited, time variable)

A set of all solutions **x**(t) for all possible initial conditions can be expressed using the notion **phase flux Φ** of differential equation $\mathbf{x}^{\cdot} = \mathbf{f}(\mathbf{x})$, i.e. a mapping assigning to initial state $\mathbf{x_0}$ and to given time a solution **x**(t). The following relations hold

(06) $$\Phi_{\mathbf{t}} = \mathbf{x}(t) \qquad \Phi_{t2}\left[\Phi_{t1}(\mathbf{x_0})\right] = \Phi_{t1+t2}(\mathbf{x_0})$$

If we know the solution $\mathbf{x}(t_1)$ with initial state $\mathbf{x}(t_0) = \mathbf{x_0}$, then the solution $\mathbf{x}(t_2)$ with initial state $\mathbf{x}(t_1)$ is identical with the solution $\mathbf{x}(t_1 + t_2)$ with initial state $\mathbf{x}(t_0)$. Trajectories generated by flux $\mathbf{\Phi}$ do not intersect. For example, in the case of linear autonomous system

(07) $\qquad x^{\cdot} + kx = 0 \qquad k \rangle 0 \qquad x(0) = x_0 \quad \Rightarrow \quad x(t) = x_0 \exp(-kt)$

it clearly holds

(08) $\qquad \Phi_t\, x_0 = x_0 \exp(-kt) \quad \Rightarrow \quad \Phi_t = \exp(-kt)$

Let $\mathbf{x}^{\cdot} = \mathbf{f(x)}$ is a set of differential equations. Then the point satisfying the condition $\mathbf{f}(\bar{\mathbf{x}}) = \mathbf{0}$ is called **fixed point** or **stationary point**. Periodic trajectory with basic period T is a solution with the property

(09) $\qquad \mathbf{x}(t) = \mathbf{x}(t + T) \qquad \mathbf{x}(t) \neq \mathbf{x}(0) \qquad$ for 0<t<T

Closed set A is called **attractor**, if there is such neighbourhood U, that for all $\mathbf{x} \in U$ and for $\Phi_{\mathbf{t}}(\mathbf{x}) \in U \Rightarrow \lim_{t\to\infty} \Phi_{\mathbf{t}}(\mathbf{x}) = A$. A set U is then the **area of attraction of the attractor** A. **Repellor** is a set having all properties the same as attractor, but trajectories are repelled.

Let us consider one-dimensional state space with corresponding dynamical equation

(10) $\qquad x^{\cdot} = f(x)$

According to definition, the position of fixed points is determined from the relation

(11) $\qquad x^{\cdot}\big|_{x=\bar{x}} = f(\bar{x}) = 0$

At one-dimensional case, three types of fixed points can arise, and namely

- **attractors** (stable fixed points attracting neighbouring trajectories);
- **repellors** (unstable fixed points repelling neighbouring trajectories);
- **saddles** (trajectories are attracted on the one side and repelled on the second side).

With a view to (11), Taylor expansion round a fixed point will be

(12) $\qquad f(x) = \frac{df}{dx}(x - \bar{x}) + \frac{1}{2}\frac{d^2 f}{dx^2}(x - \bar{x})^2 + \ldots$

where all derivatives are computed at fixed point. We introduce a new variable $u = x - \bar{x}$, representing the distance of a trajectory from fixed point. In first approximation, it will hold

(13) $\qquad u^{\cdot} = u\frac{df}{dx}\Big|_{x=\bar{x}} = \lambda u \qquad \Rightarrow \qquad u(t) = u(0)\exp(\lambda t)$

Lyapunov exponent

(14) $\qquad \lambda = \frac{df(x)}{dx}\Big|_{x=\bar{x}}$

is a local characteristic of the stability of fixed point. If $\lambda < 0$, then fixed point is stable, whereas for $\lambda > 0$ it is unstable.

As an example, let us consider logistic differential equation

(15) $\qquad x^{\cdot} = Ax(1 - x) \qquad A \rangle 0 \qquad \Rightarrow \qquad x(t) = \frac{x_0}{x_0 - (x_0 - 1)\exp(-At)}$

which has two fixed points $\bar{x}_1 = 0, \bar{x}_2 = 1$. The corresponding Lyapunov exponents are

(16) $$\lambda_1 = A(1-2x)\big|_{x=0} = A \rangle 0 \qquad \lambda_2 = A(1-2x)\big|_{x=1} = -A \langle 0$$

and the point $\overline{x_1} = 0$ is unstable, whereas the point $\overline{x_2} = 1$ is stable.

In two-dimensional state space, dynamical equations will have form

(17) $$x_1^{\cdot} = f_1(x_1, x_2) \qquad x_2^{\cdot} = f_2(x_1, x_2)$$

and fixed points are determined from the relations

(18) $$f_1\left(\overline{x_1}, \overline{x_2}\right) = 0 \qquad f_2\left(\overline{x_1}, \overline{x_2}\right) = 0$$

Taylor´s development in the surroundings of fixed points will be

(19) $$x_1^{\cdot} = f_1(x_1, x_2) = f_{11}\frac{\partial f_1}{\partial x_1} + f_{12}\frac{\partial f_1}{\partial x_2} + \ldots \qquad f_{ij} = \frac{\partial f_i}{\partial x_j}$$
$$x_2^{\cdot} = f_2(x_1, x_2) = f_{21}\frac{\partial f_2}{\partial x_1} + f_{22}\frac{\partial f_2}{\partial x_2} + \ldots$$

Keeping only first-order terms, we obtain

(20) $$u_1^{\cdot} = f_{11}u_1 + f_{12}u_2 \qquad u_1 = x_1 - \overline{x_1}$$
$$u_2^{\cdot} = f_{21}u_1 + f_{22}u_2 \qquad u_2 = x_2 - \overline{x_2}$$

This system can be replaced by single equation of the second order

(21) $$u_1^{\cdot\cdot} = (f_{11} + f_{22})u_1^{\cdot} + (f_{12}f_{21} - f_{11}f_{22})u_1$$

Seeking a solution in the form $u_1(t) = C\exp(\lambda t)$, we get characteristic equation for the determination of Lyapunov exponents and its solution in the form

(22) $$\lambda^2 - (f_{11} + f_{22})\lambda + (f_{11}f_{22} - f_{12}f_{21}) = 0$$
$$\lambda_{1,2} = \frac{1}{2}\left(f_{11} + f_{22} \pm \sqrt{(f_{11} + f_{22})^2 - 4(f_{11}f_{22} - f_{12}f_{21})}\right)$$

Clearly, the nature of a fixed point is determined by the value of $f_{11} + f_{22} \langle 0$ (repellor) or $f_{11} + f_{22} \rangle 0$ (attractor). In the case of negative discriminant of characteristic equation, periodic trajectories can arise as well.

2. Deterministic chaos

Even some simple **non-linear deterministic systems** can under certain conditions pass to chaotic states [1]. Chaotic behavior is bounded, not periodic and similar to random one. It is highly sensitive to small change of initial conditions and cannot be predicted for long time. It is called **deterministic chaos**.

In the case of autonomous systems with continuous time, chaotic behavior can appear in third order systems, in the case of non-autonomous already in second order systems. Some systems with discrete time (described by difference equations) are capable of producing of chaotic behaviour already in one-dimensional case.

As an simple example [2], let us consider discrete system described by **logistic difference equation**

(23) $$x_{n+1} = A x_n (1 - x_n) = f(x_n) \qquad 0 < A \leq 4$$

where A is control parameter. For $0 < A \leq 4$, values from the interval $< 0,1>$ will be mapped also into this interval. This system has been originally used for the modelling of the growth of a population at limited territory providing that individual generations do not overlap. State variable x_n gives the number of objects in n-th generation. Quadratic term prevents the population from unlimited growth, parameter A describes the influence of surroundings.

The function $f(x_n)$ is so-called **iterative function**. The position of fixed points can be determined from the relation

(24) $$\bar{x} = A\bar{x}(1 - \bar{x}) \qquad \Rightarrow \qquad \bar{x}_1 = 0 \qquad \bar{x}_2 = 1 - \frac{1}{A}$$

and their stability from the behavior of first derivatives

(25) $$\frac{df}{dx}\Big|_{x=0} = A \qquad \frac{df}{dx}\Big|_{x=1-1/A} = 2 - A$$

For $A \leq 1$, there is only one fixed point $\bar{x}_1 = 0$, which is stable (attractor). A sequence of values $x_0, x_1, x_2, \ldots$ tends to converge to zero (a population becomes extinct).

In the range $1 < A < 3$, two fixed points exist. Now, the point $\bar{x}_1 = 0$ is unstable (repellor) and $\bar{x}_2 = 1 - (1/A)$ is stable (attractor). Trajectories from arbitrary initial condition converge to one-point attractor (a population reach stationary state).

For $A=3$, the first **bifurcation** occurs and the second fixed point will be unstable as well. In the case of further increase of control parameter A, the function f^2 will have four fixed points. Two of them correspond to unstable points of function f, another two ones are stable and correspond to periodic attractor of function f with period 2. For $A=3.45$, the trajectory with period two 2 becomes unstable and stable trajectory with period 4 arises. This doubling of the period is repeated always at the growth of A and stable trajectories with periods 8,16,32,...arise.

Let A_1 is the value of control parameter, which gives rise to period 2, A_2 is the value at which period 4 arises and A_n is the value at which period 2^n arises. Let us denote

(26) $$\delta_n = \frac{A_n - A_{n-1}}{A_{n+1} - A_n}$$

Feigenbaum has shown, this ratio is roughly the same for all n and it approaches to the limit

(27) $$\delta = \lim_{n \to \infty} \delta_n = 4.66920161\ldots$$

Thus, the intervals of the parameter A, at which bifurcation occurs, are still contracted in this ratio. The sequence of bifurcation values A_n, at which stable trajectories with period 2^n occur, is convergent and tends to the limit

(28) $$A_\infty = \lim_{n \to \infty} A_n = 3.569946\ldots$$

Within interval $A_\infty < A \leq 4$, the system behaviour becomes very complex. There is infinitely many intervals of the parameter A (periodic windows) with stable periodic trajectories. On the other hand, there are certain parameter values leading to chaotic behavior. In the limit case $A=4$, also analytical solution exists in the form

(29) $$x_n = \sin^2\left(2^n \arcsin\sqrt{x_0}\right)$$

Clearly, from this solution, obvious extreme sensitivity to very small changes of initial value x_0 is seen.

Thus, the solution of logistic difference equation (23) leads in the case of increase of control parameter A from periodic solution with period 2 through bifurcation cascade of period doubling to chaotic behavior.

Now we mention briefly another way of the emergence of a chaotic state. It is the case of the class of **by part linear mapping**. For example, to this class belongs **symmetric roof mapping**

$$\begin{aligned} x_{n+1} &= 2Ax_n \qquad for \quad 0 \le x_n \le 0.5 \\ x_{n+1} &= 2A\left(1-x_n\right) \qquad for \quad 0.5 \le x_n \le 1 \end{aligned} \tag{30}$$

with control parameter $0 < A \le 1$. This function is continuous, but it has not the derivative at the point $x=0.5$. In the case $A < 0.5$, only one fixed point exists and namely $\bar{x} = 0$; this point is stable, because $2A < 1$. For $A > 0.5$, there are two fixed points

$$\bar{x}_1 = 0 \qquad \bar{x}_2 = \frac{2A}{2A+1} \tag{31}$$

and it holds

$$\left|\frac{df}{dx}\left(x=\bar{x}_1\right)\right| = \left|\frac{df}{dx}\left(x=\bar{x}_2\right)\right| = 2A > 1 \tag{32}$$

Both fixed points are unstable and trajectories are chaotic. In this case, chaotic behavior arises suddenly for $A > 0.5$ and no bifurcations occur.

3. Quantification of chaotic behavior

The reasons for the construction of quantitative characteristics of chaotic behavior are the following:

- quantificators can help to distinguish deterministic chaos from „noisy“ behavior, produced by the action of external random influences
- quantificators can help to determine minimum number of variables needed for the construction of a dynamical model of the system
- quantificators can help to classify systems according to universally valid regularities
- changes of quantificators may signalise changes in qualitative behavior of a system

On principle, we can use two kinds of description. The first type stresses the dynamics of chaotic behavior and corresponding quantificators describe the system evolution and mutual position of neighbouring trajectories. The second type is based on the geometry of trajectories in state space. Here we shall confine ourselves to the first kind of the description.

One of characteristic features of chaotic behavior is the divergence of neighbouring trajectories. Let us consider two close points $x_0, x_0 + \varepsilon$. Applying certain iterative function n-times, the distance between these points will be

$$d_n = \left| f^{(n)}\left(x_0 + \varepsilon\right) - f^{(n)}\left(x_0\right) \right| \tag{33}$$

In the case of chaotic behavior, we expect exponential growth of this distance with the number of iterations

(34) $$\frac{d_n}{\varepsilon} = \frac{\left|f^{(n)}(x_0+\varepsilon) - f^{(n)}(x_0)\right|}{\varepsilon} = \exp(\lambda n)$$

and thus

(35) $$\lambda = \frac{1}{n}\ln\left(\frac{\left|f^{(n)}(x_0+\varepsilon) - f^{(n)}(x_0)\right|}{\varepsilon}\right)$$

In limit approaching $\varepsilon \to 0$ we obtain after small algebra

(36) $$\lambda = \frac{1}{n}\ln\left(\left|f'(x_0)\right|\left|f'(x_1)\right|\ldots\left|f'(x_n)\right|\right)$$

or in alternative expression

(37) $$\lambda = \frac{1}{n}\left(\ln\left|f'(x_0)\right| + \ln\left|f'(x_1)\right| + \ldots + \ln\left|f'(x_n)\right|\right)$$

This **Lyapunov exponent** characterizes the speed of the divergence of neighbouring trajectoriesí. It is given as averaged natural logarithm of absolute value of the derivatives of iterative function at individual points of a trajectory. One-dimensional iterative function has **chaotic trajectories**, if average Lyapunov exponent is positive (the condition of divergence).

Having a time series of equidistant values $x_0, x_1, x_2, \ldots, x_n$, we can determine Lyapunov exponent using the following approach. Let us take two trajectories with starting points x_i, x_j and let us create the sequence of differences

(38) $$d_0 = \left|x_j - x_i\right| \qquad d_1 = \left|x_{j+1} - x_{i+1}\right| \qquad \ldots \qquad d_n = \left|x_{j+n} - x_{i+n}\right|$$

We shall assume time evolution in the form

(39) $$d_n = d_0 \exp(\lambda n) \qquad \Rightarrow \qquad \lambda = \frac{1}{n}\ln\frac{d_n}{d_0}$$

For $\lambda > 0$, the behavior of trajectories is chaotic. However, generally, λ depends on the choice of x_i. Therefore, the more reliable is to compute **average Lyapunov coefficient** from large number N initial values regularly distributed over whole attractor. Then we get

(40) $$\overline{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda(x_i)$$

In the case of logistic function with control parameter A=4 is average Lyapunov exponent given by analytical expression

(41) $$\overline{\lambda} = \int_0^1 \frac{\ln\left|4(1-2x)\right|}{\sqrt{x(1-x)}}\,dx = \ln 2$$

Grassberger and Procaccia introduced the characteristic called **correlation dimension**, based on the behavior of so-called correlation sum [3]. For the computation of correlation dimension, we need data about the evolution of a trajectory (in sum n values). For each i-th point of the trajectory, we seek relative frequency $p_i(r)$ of trajectory points, lying at distance less than r from the point i (except i-th point)

(42) $$p_i\ (r) = \frac{n_i}{n-1}$$

Correlation sum is then computed as average ralative frequency

(43) $$C_1\,(r) = \frac{1}{n}\sum_{i=1}^{n} p_i\ (r)$$

Obviously $C_1(r)=0$, if r is less than minimal distance among the points of a trajectory. On contrary $C_1(r)=1$ means, the distances among individual points do not exceed r . Minimal possible non-zero value $C_1(r)=2/(n(n-1))$ occurs in the case, only one distance is less than r.

Relative frequency can be formally expressed using Heaviside step function

(44) $$H(x) = 0 \qquad for \quad x < 0$$
$$H(x) = 1 \qquad for \quad x \geq 0$$

(45) $$p_i\ (r) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} H\left(r - \left|x_i - x_j\right|\right)$$

Similarly, correlation sum can be written as

(46) $$C_1\,(r) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} H\left(r - \left|x_i - x_j\right|\right)$$

In limit case $n\to\infty$, correlation sum is melted into **correlation integral**. Correlation dimension is then given by formula

(47) $$D_1 = \lim_{r\to 0}\frac{\log C_1(r)}{\log r}$$

A time series of single variable can be often sufficient for the determination of important characteristics of a multidimensional dynamical system. The groups of values

(48) $$x_{t+1}, x_{t+2}, \ldots, x_{t+d} \qquad t = 0,1,2,\ldots,(n-d)$$

gives the coordinates of a point in d-dimensional space. Then the sequence of these groups describes the time evolution of a system in d-dimensional **embedding space.**

In this case, correlation sum can be written as

(49) $$C_d\,(r) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} H\left(r - \left|\mathbf{x_i} - \mathbf{x_j}\right|\right)$$

because it depends on embedded dimesion d. Vector $\mathbf{x_i}$ of dimension d has components

(50) $$\mathbf{x_i} = \left(x_i, x_{i+L}, x_{i+2L}, \ldots, x_{i+(d-1)L}\right)$$

where L is time lag between neighbouring values. The length of the difference of two vectors is mostly calculated as Euclid distance

(51) $$\left|\mathbf{x_i} - \mathbf{x_j}\right| = \sqrt{\sum_{k=0}^{d-1}\left(x_{i+kL} - x_{j+kL}\right)^2}$$

Sometimes also Tchebychef distance is used

(52) $$\left|\mathbf{x_i} - \mathbf{x_j}\right| = \max_k \left| x_{i+kL} - x_{j+kL} \right|$$

Then it holds for correlation dimension

(53) $$D_d = \lim_{r\to 0} \frac{\log C_d(r)}{\log r}$$

Grassberger and Procaccia have studied the behavior of these characteristic. In the case of i.i.d. (**i**ndependent **i**dentically **d**istributed) process with regular distribution is

(54) $$D_1 = \lim_{r\to 0} \frac{\log C_1(r)}{\log r} = \lim_{r\to 0} \frac{\log 2 + \log r}{\log r} = 1$$

(55) $$D_2 = \lim_{r\to 0} \frac{\log C_2(r)}{\log r} = \lim_{r\to 0} \frac{\log 4 + 2\log r}{\log r} = 2$$

and generally $D_d = d$. On the contrary, in the case of non-linear deterministic process is the behavior of correlation sum quite different. For example, in the case of roof mapping is $D_1 = D_2 = \ldots = D_d = 1$.

Brock, Dechert and Scheinkman showed, for a finite *r* and i.i.d. process the following relation is valid [4]

(56) $$C_d(r) = \left[C_1(r)\right]^d$$

and they suggested test statistic

(57) $$T_d(r,n) = \frac{C_d(r,n) - \left[C_1(r,n)\right]^d}{s_d(r,n)}$$

where $C_d(r,n), C_1(r,n)$ are sample correlation sums and $s_d(r,n)$ is the estimate of the standard deviation of the difference. This statistic has asymptotically standard normal distribution N(0,1) providing the validity of null hypothesis (i.i.d. process).

Hsieh suggested another type of non-linearity testing with the use of third standardized moment [5]

(58) $$\varphi(i,j) = \frac{E\left(x_t x_{t-i} x_{t-j}\right)}{\left[E\left(x_t^2\right)\right]^{3/2}}$$

For an IID process, $\varphi(i,j) = 0$ for all *i, j* > 0. The estimate of $\varphi(i,j)$ is the statistic

(59) $$f(i,j) = \frac{(1/n)\sum_t x_t x_{t-i} x_{t-j}}{\left[(1/n)\sum_t x_t^2\right]^{3/2}}$$

Under validity of null hypothesis $\varphi(i,j) = 0$ has the statistic $f(i,j)\sqrt{n}$ asymptotically normal distribution, whose variance can be estimated as follows

(60) $$s^2 = \frac{(1/n)\sum x_t^2 x_{t-i}^2 x_{t-j}^2}{\left((1/n)\sum x_t^2\right)^3}$$

This test is based only on third-order moments, but in general, different moments can be used simultaneously. Another possibility is the use of autoregressive model in the form

(61) $$x_t = \sum_{i=1} a_i x_{t-i} + \sum_{i=1} \sum_{j=i} b_{ij} x_{t-i} x_{t-j} + \sum_{i=1} \sum_{j=1} \sum_{k=j} c_{ijk} x_{t-i} x_{t-j} x_{t-k} + \ldots$$

and testing the significance of individual non-linear terms.

Acknowledgement

This study was performed with financial support from the Grant Agency of the Czech Republic, contract number 402/05/0128.

References

1. Alligood, K., Sauer, T., Yorke, J.: Chaos: An introduction to dynamical systems. Springer 2000, 603 p.

2. May, R.: Simple mathematical models with very complicated dynamics. Nature 261 (1976), pp.459-467.

3. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica 9D (1983), pp.189-208.

4. Campbell, J. et al.: The econometrics of financial markets. Princeton 1997, 611 pp.

5. Hsieh, D.: Trstiny for nonlinear dependence in daily foreign exchange rates. J.of Business 62 (1989), pp.339-368.

Predikcie v neurónových sieťach a logistickej regresii

Štefan Varga[a,*], Peter Grančič[a], Svetozár Katuščák[b] a Attila Szitás[b]
[a]Katedra matematiky, FChPT STU Bratislava
[b]Katedra chemickej technológie dreva, celulózy a papiera, FChPT STU Bratislava
stefan.varga@stuba.sk

1. Úvod

Riešil sa problém rozlíšiteľnosti niekoľkých druhov drevín (buk, dub, čerešňa, jaseň a javor) na základe hodnôt jednoduchých deskriptorov. Tento problém sa riešil použitím neurónovej siete a metódami logistickej regresie.

2.1 Neurónová sieť

Neurónová sieť (NS), je paralelný distribuovaný systém schopný ukladať a opätovne používať skúsenosti. Skúsenosti sú získavané v procese učenia a uložené v parametroch NS – vo váhových w_{ij} a prahových ϑ_i koeficientoch.

Formálne možno NS považovať za súvislý orientovaný ohodnotený graf. Neuróny NS sú vrcholmi grafu a prepojenia medzi neurónmi sú ohodnotené orientované hrany v grafe.

Signál sa NS šíri v smere orientovaných hrán. Neuróny, prijímajúce signál z externého prostredia, sa nazývajú vstupné neuróny. Naopak, neuróny, z ktorých je signál odovzdávaný prostrediu, sa nazývajú výstupné neuróny. Vo všeobecnosti môžu byť neuróny oboch uvedených typov zároveň. Neuróny, ktoré nie sú ani vstupné ani výstupné, sa nazývajú skryté neuróny.

Aktívna fáza NS je proces, pri ktorom sa postupne šíri signál od vstupných neurónov k výstupným neurónom. Aktivity skrytých a výstupných neurónov sú dané pomocou rekurentných vzťahov

$$x_i = \sum_j w_{ij} x_j - \vartheta_i \tag{1}$$

$$x_i = \sigma(x_i) \tag{2}$$

kde x_i, x_j sú aktivity i-teho, resp. j-teho neurónu[1]; w_{ij} je váhový koeficient (ohodnotenie hrany, spájajúcej j-ty neurón s i-tym neurónom); ϑ_i je prahový koeficient i-teho neurónu a $\sigma(x)$ je sigmoidálna prechodová funkcia

$$\sigma(x) = \frac{b + ae^{-\lambda x}}{1 + e^{-\lambda x}} \tag{3}$$

zobrazujúca $\mathbb{R} \to (a, b)$, kde λ je koeficient strmosti. Pre $\lambda > 0$ je $\sigma(x)$ monotónne rastúca funkcia, vyhovujúca dvom nasledujúcim asymptotickým podmienkam: $\sigma(x) \to a$ pre $x \to -\infty$ a $\sigma(x) \to b$ pre $x \to \infty$.

Pri kontrolovanom učení je tréningová množina tvorená usporiadanými dvojicami k vektorov $\boldsymbol{x}^{(k)}$ a $\boldsymbol{y}_{req}^{(k)}$, kde $\boldsymbol{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_N^{(k)})$ je vektor N vstupov a $\boldsymbol{y}_{req}^{(k)} = (y_{1,req}^{(k)}, y_{2,req}^{(k)}, \ldots, y_{M,req}^{(k)})$ je vektor M požadovaných výstupov.

Definujme účelovú funkciu ako sumu štvorcov rozdielu medzi k-tym vypočítaným $\boldsymbol{y}^{(k)}$ a požadovaným $\boldsymbol{y}_{req}^{(k)}$ výstupom siete

$$E^{(k)}(w, \vartheta) = \sum_{i=1}^{M} (y_i^{(k)} - y_{i,req}^{(k)})^2 \tag{4}$$

[1]Je konvenciou, že spojenie vedie z j-teho do i-teho neurónu.

a celkovú účelovú funkciu $E(w, \vartheta)$ pre všetky dvojice tréningovej množiny

$$E(w, \vartheta) = \sum_k E^{(k)}(w, \vartheta) \tag{5}$$

Proces učenia spočíva v nájdení takých optimálnych hodnôt prahových ϑ_{opt} a váhových w_{opt} koeficientov, aby účelová funkcia $E(w, \vartheta)$ bola minimálna. Ide teda o optimalizačný problém v tvare

$$w_{opt}, \vartheta_{opt} = \arg\min_{w,\vartheta \in \mathbb{R}} E(w, \vartheta) \tag{6}$$

Najčastejšie používanou optimalizačnou metódou je gradientová metóda najprudšieho spádu. Váhové a prahové koeficienty sú iteračne upravované podľa vzťahov

$$w_{i+1} = w_i - \alpha \frac{\partial E}{\partial w_i} \tag{7}$$

$$\vartheta_{i+1} = \vartheta_i - \alpha \frac{\partial E}{\partial \vartheta_i} \tag{8}$$

kde koeficient α je *rýchlosť učenia*, $\alpha \in \langle 0, 1 \rangle$. Niekedy sa k rovniciam 7 a 8 zvykne pridávať tzv. momentový člen, zohľadňujúci rozdiel v hodnotách z posledných dvoch iterácií[2].

Parciálne derivácie z rovníc 7 a 8 možno numericky aproximovať pomocou vzťahov

$$\frac{\partial E}{\partial w} \approx \frac{E(w + \varepsilon) - E(w - \varepsilon)}{2\varepsilon} \tag{9}$$

$$\frac{\partial E}{\partial \vartheta} \approx \frac{E(\vartheta + \varepsilon) - E(\vartheta - \varepsilon)}{2\varepsilon} \tag{10}$$

kde ε je dostatočne malé číslo. Presnosť uvedenej aproximácie možno charakterizovať kvadratickou chybou $\mathcal{O}(\varepsilon^2)$.

Procedúra beží pokiaľ nie je splnená nejaká zastavovacia podmienka, napríklad pokiaľ dĺžka gradientu nie je menšia ako nejaká vopred zvolená hodnota δ.

2.2 Logistická regresia

Predpokladajme, že jednotlivé hodnoty výstupnej veličiny majú alternatívne rozdelenie

$$P(y_i = 1) = p, \quad P(y_i = 0) = 1 - p \tag{11}$$

a skúmajme závislosť pravdepodobnosti $P(y_i = 1) = p$ od vektora prediktorov $\boldsymbol{x} = (x_1, \ldots, x_k)^T$. Regresný model by mohol byť v tvare

$$p = a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_m f_m(x_i) \tag{12}$$

ale problém by nastal, keby pre niektorú kombináciu odhadu regresných koeficientov a funkčných hodnôt prediktorov vyšla pravdepodobnosť mimo intervalu $\langle 0, 1 \rangle$. Preto sa v logistickej regresii uvažuje model

$$\log \frac{p}{1 - p} = a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_m f_m(x_i) \tag{13}$$

[2]Momentový člen má za úlohu zabezpečiť, aby minimalizačná procedúra neuviazla v lokálnom minime, jeho tvar je $\mu \Delta w_i$, kde $\Delta w_i = w_i - w_{i-1}$ je rozdiel v hodnotách váhového koeficientu v ostatných iteračných krokoch. Koeficient μ je *momentum*, $\mu \in \langle 0, 1 \rangle$. Momentový člen pre prahový koeficient možno vyjadriť analogicky, $\mu \Delta \vartheta_i = \mu(\vartheta_i - \vartheta_{i-1})$.

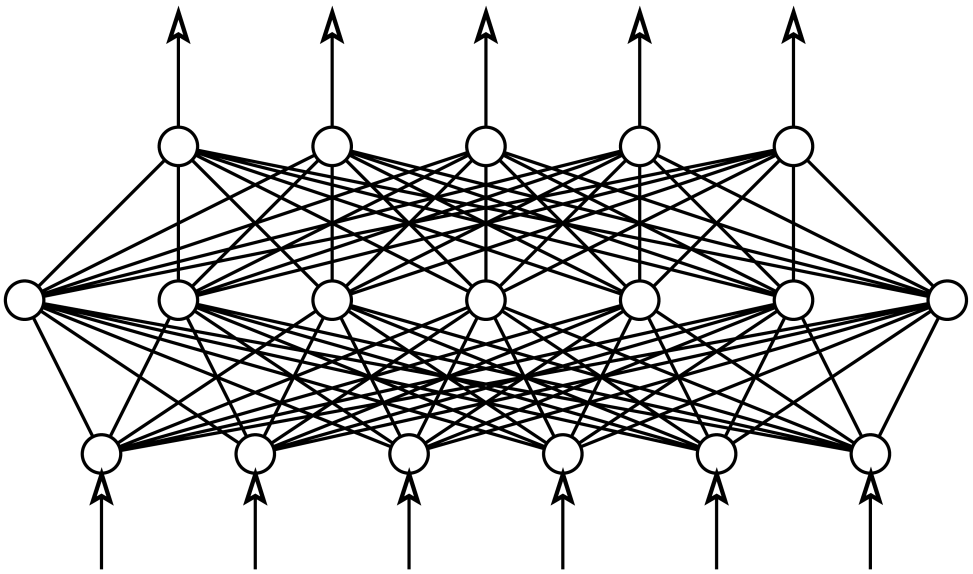
Výstupná veličina už nadobúda hodnoty z $\mathbb{R} = (-\infty, \infty)$ a pravdepodobnosť $P(y_i = 1) = p$ sa dá spätne prepočítať. Výraz

$$\frac{p}{1-p} \in \langle 0, \infty) \tag{14}$$

sa zvykne nazývať šancou toho, že $y_i = 1$. V logistickej regresii sa vlastne predpovedá pravdepodobnosť p toho, že $y_i = 1$, pri daných (zvolených) hodnotách prediktorov $x = (x_1, \ldots, x_k)^T$. Ďalšou z úloh logistickej regresie je identifikovať prediktory, ktoré významne ovplyvňujú pravdepodobnosť $P(y_i = 1) = p$. Testuje sa významnosť jednotlivých regresných koeficientov ($H_0 \colon a_i = 0$, Waldov test).

3. Formulovanie problému

Riešený problém možno charakterizovať ako klasický klasifikačno-predikčný problém. Úlohou je na základe jednoduchých popisných charakteristík (deskriptorov) použitím metód neurónovej siete a logistickej regresie rozlíšiť a správne zaradiť dreviny. Východiskom boli záznamy farebných bodov drevených dosiek v priečnom smere v súradniciach farebného priestoru RGB a z nich vypočítané jednoduché popisno-štatistické charakteristiky ($\mu(R)$ – stredná hodnota R, $\mu(G)$ – stredná hodnota G, $\mu(B)$ – stredná hodnota B, $\sigma(R)$ – smerodajná odchýlka R, $\sigma(G)$ – smerodajná odchýlka G a $\sigma(B)$ – smerodajná odchýlka B). Každá z použitých metód však vyžadovala rozdielny prístup.



Obr. 1: Neurónová sieť s dopredným šírením signálu.

3.1 Formulovanie problému pre neurónovú sieť

Použitá trojvrstvová neurónová sieť s dopredným šírením signálu (obrázok 1) pozostávala zo šiestich vstupných (6 deskriptorov $\mu(R)$, $\mu(G)$, $\mu(B)$, $\sigma(R)$, $\sigma(G)$, $\sigma(B)$), niekoľkých skrytých a $N = 5$ výstupných neurónov. Nech N je počet uvažovaných druhov drevín. Každú drevinu potom možno jednoznačne popísať jedným vektorom v ortonormálnej báze N-rozmerného vektorového priestoru.

Cieľom je tiež navrhnúť optimálnu architektúru neurónovej siete (optimálny počet skrytých neurónov) a optimálny počet adaptačných krokov v procese učenia (adaptácie).

3.2 Formulovanie problému pre logistickú regresiu

V klasickom regresnom modeli

$$y = a_1 f_1(x) + a_2 f_2(x) + \ldots + a_m f_m(x) \tag{15}$$

kde $f_i(x)$ $(i = 1, 2 \ldots, m)$ sú známe funkcie vstupného vektora prediktorov $\boldsymbol{x} = (x_1, \ldots, x_k)^T$, y je výstupná veličina (odozva) a $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)^T$ je vektor neznámych regresných koeficientov, sa predikcie veličiny y robia pomocou bodových odhadov vektora regresných koeficientov $\boldsymbol{a}$

$$est_{LS} a = \arg\min_{a \in \mathbb{R}^m} \sum_{i=1}^{n} (y_i - est\, y_i)^2 \tag{16}$$

Za predpokladu, že jednotlivé pozorovania $(i = 1, 2 \ldots, m)$

$$y_i = a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_m f_m(x_i) + e_i \tag{17}$$

sú nezávislé, rovnako rozdelené náhodné veličiny so strednými hodnotami (e_i je chyba i-teho pozorovania)

$$E(y_i) = a_1 f_1(x_i) + a_2 f_2(x_i) + \ldots + a_m f_m(x_i) \tag{18}$$

$$E(e_i) = 0 \tag{19}$$

a disperziami

$$D(y_i) = D(e_i) = \sigma^2 \tag{20}$$

kvalita uvedeného odhadu vektora regresných koeficientov nezávisí od typu rozdelenia výstupnej veličiny y. Zväčša sa predpokladá, že jednotlivé pozorovania majú normálne rozdelenie. V prípade klasifikácie druhu drevín je vhodné predpokladať, že jednotlivé pozorovania majú alternatívne rozdelenie. Úlohou je totiž, pomocou hodnôt niekoľkých prediktorov, rozhodnúť, či uvedená drevina je napríklad buk ($y_i = 1$) alebo nie je ($y_i = 0$). Tento problém rieši aj logistická regresia.

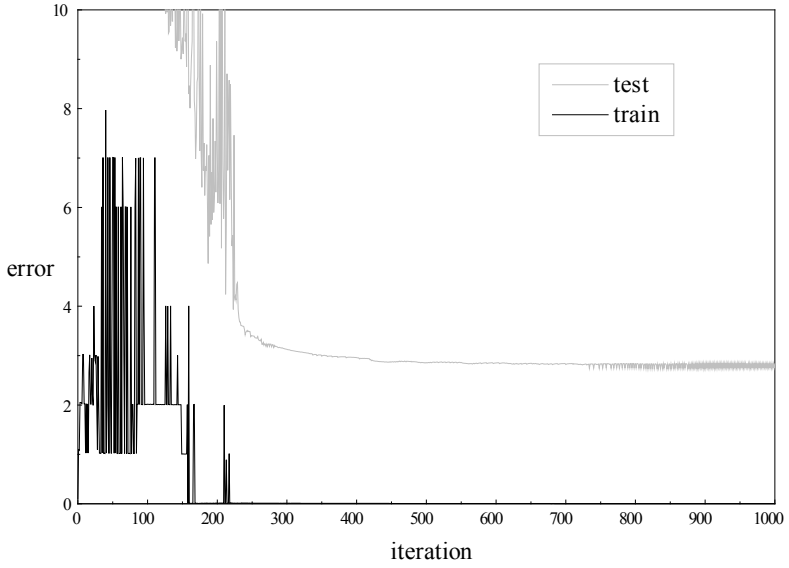
druh	train	test
dub	100%	90%
buk	100%	100%
javor	100%	100%
čerešňa	100%	90%
jaseň	100%	100%
celkom	100%	96%

Tabuľka 1: Celkový výsledok pre neurónovú sieť.

4. Výsledky experimentu

4.1 Výsledky experimentu pre neurónovú sieť

Množina všetkých drevín (5 druhov, pre každý druh 25 pozorovaní optických veličín) bola rozdelená na dve disjunktné podmnožiny – na trénovaciu a testovaciu množinu (pre každý druh 15 trénovacích a 10 testovacích vzoriek). Trénovacie vzorky boli použité na hľadanie optimálnych hodnôt parametrov (váhových w_{ij} a



Obr. 2: Priebeh chybovej funkcie.

prahových ϑ_i koeficientov) neurónovej siete. Testovacie vzorky boli použité na overenie predikčnej schopnosti už natrénovanej neurónovej siete.

Najlepšie výsledky boli získané pre neurónovú sieť s 8 skrytými neurónmi pri kontrolovanom sekvenčnom učení metódou *backprop with momentum* s náhodným usporiadaním trénovacích vzoriek po 1000 iteráciach ($\alpha = 0.6$, $\mu = 0.1$). Úspešnosť neurónovej siete pre jednotlivé druhy drevín vyjadruje tabuľka 1. Podrobnejšie výsledky sú uvedené v tabuľke 2.

Typický priebeh chybovej funkcie E pre trénovaciu a testovaciu množinu naznačuje obrázok 2. Optimálny počet skrytých neurónov (6 až 8) bol stanovený experimentálne.

4.2 Výsledky experimentu pre logistickú regresiu

Pre 25 meraní optických veličín x_1, x_2, x_3, x_4, x_5, x_6, zodpovedajúcim 5 druhom dreva (dub, buk, javor, čerešňa a jaseň), spolu teda 125 šestíc meraní s kvalitatívnou veličinou druh dreva, bolo úlohou

1) rozhodnúť ktoré veličiny štatisticky významne vplývajú na spoznávanie druhov dreva;

2) spoznať jednotlivé druhy dreva z meraní šiestich (prípadne menej) optických veličín.

Ukázalo sa, že logistická regresia dokázala na 100% spoznať buk, na 98,4% javor, na 98,4% čerešňu, na 88,8% dub a najhoršie, na 62,4% jaseň. Nasledujúca rovnica predstavuje výstup pre buk, kde v prvej etape boli uvažované všetky optické veličiny

$$\log \frac{p}{1-p} = 243.64 + 0.68x_1 - 2.06x_2 + 0.30x_3 - 2.74x_4 + 2.31x_5 - 2.31x_6 \quad (21)$$

Po testovaní významnosti jednotlivých veličín ostal redukovaný model

$$\log \frac{p}{1-p} = 357.28 - 1.61x_2 - 4.76x_4 + 4.91x_5 - 3.31x_6 \quad (22)$$

požadovaný výstup					vypočítaný výstup				
y_1^{req}	y_2^{req}	y_3^{req}	y_4^{req}	y_5^{req}	y_1	y_2	y_3	y_4	y_5
tréningová množina									
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	0.9999	0.0000	0.0002	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0025	0.9966	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	0.9981	0.0000	0.0036	0.0000
0	1	0	0	0	0.0000	0.9988	0.0000	0.0000	0.0008
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0007
0	1	0	0	0	0.0001	0.9994	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0001	0.0000
0	1	0	0	0	0.0013	0.9995	0.0000	0.0018	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0002	0.9991	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
testovacia množina									
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	0.9999	0.0000	0.0001	0.0000
0	1	0	0	0	0.0155	0.9817	0.0000	0.0000	0.0014
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	0.9998	0.0000	0.0001	0.0000
0	1	0	0	0	0.0000	1.0000	0.0000	0.0000	0.0000
0	1	0	0	0	0.0000	0.9987	0.0000	0.0000	0.0067

Tabuľka 2: Výsledky neurónovej siete pre *buk*.

z ktorého sa počítala pravdepodobnosť spoznania buka. Táto pravdepodobnosť, ako aj hodnoty jednotlivých optických veličín sú uvedené v tabuľke 3. Zo 125 riadkov sú tam len riadky s poradovými číslami 21 - 54, lebo v riadkoch 25 - 49 sú hodnoty optických veličín merané na buku.

5. Záver

Práca ukázala, že pomocou jednoduchých popisno-štatistických charakteristík (deskriptorov) je možné dostatočne popísať a charakterizovať uvedené druhy drevín.

Kvalitatívne veličiny – deskriptory $\mu(R)$, $\mu(G)$, $\mu(B)$, $\sigma(R)$, $\sigma(G)$, $\sigma(B)$, boli použité ako vstupy trojvrstvovej neurónovej siete a nezávislé premenné v logistickej regresii za účelom predikcie. Celková úspešnosť predikcie (98.4%-ná pre neurónovú sieť a 89.6%-ná pre logistickú regresiu) predstavovala uspokojivý výsledok a potvrdila tak hypotézu, podľa ktorej je možné taký komplexný problém, akým je strojové rozpoznávanie druhov drevín na základe hodnôt optických veličín, výrazne zredukovať.

n	x_1	x_2	x_3	x_4	x_5	x_6	druh	p
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	239.83	209.29	132.69	18.47	24.23	25.51	0	8.87E-12
22	247.74	233.37	174.84	12.52	16.69	21.56	0	9.51E-27
23	241.57	203.23	119.18	16.94	23.47	29.59	0	6.89E-12
24	237.92	202.98	131.80	22.52	25.93	31.81	0	3.43E-21
25	236.90	176.01	89.06	10.81	11.87	14.62	1	1.0
26	222.42	168.42	79.66	12.79	15.36	24.33	1	1.0
27	232.35	167.39	77.42	13.01	14.56	22.92	1	1.0
28	221.66	169.23	93.15	16.08	18.37	23.41	1	1.0
29	218.45	167.53	92.57	19.23	19.34	24.26	1	1.0
30	244.35	188.24	106.50	9.73	12.30	12.98	1	1.0
31	235.11	186.99	108.71	12.48	15.15	18.22	1	1.0
32	238.18	177.95	94.99	11.39	13.11	17.47	1	1.0
33	224.54	170.24	96.26	17.17	19.28	22.69	1	1.0
34	218.73	164.46	91.74	19.25	19.45	23.70	1	1.0
35	239.69	177.72	89.75	9.13	10.91	15.28	1	1.0
36	219.25	163.70	71.44	14.15	17.16	28.50	1	1.0
37	234.15	170.76	78.36	12.99	14.69	23.62	1	1.0
38	214.27	161.51	86.42	17.84	20.01	25.89	1	1.0
39	214.96	166.27	91.53	19.11	20.49	25.87	1	1.0
40	232.29	169.54	79.27	9.76	11.67	18.07	1	1.0
41	216.29	160.48	69.06	13.17	16.31	26.99	1	1.0
42	228.15	161.83	65.05	14.02	15.59	25.15	1	1.0
43	212.32	159.84	84.88	18.50	20.23	26.60	1	1.0
44	214.94	167.16	95.65	18.62	19.55	24.23	1	1.0
45	233.41	169.96	81.44	10.03	11.03	15.98	1	1.0
46	221.65	164.57	76.79	12.29	15.25	25.82	1	1.0
47	230.24	162.18	70.09	15.25	17.64	27.48	1	1.0
48	216.43	159.67	86.27	17.53	19.37	24.66	1	1.0
49	215.28	161.89	91.68	18.52	18.40	22.48	1	1.0
50	248.22	215.95	153.37	10.39	13.93	16.91	0	2.19E-09
51	242.81	217.90	159.88	9.52	11.11	15.67	0	3.43E-13
52	237.43	209.60	144.32	9.75	11.94	16.01	0	1.31E-06
53	236.76	214.29	157.12	16.27	19.47	24.11	0	6.58E-19
54	236.32	218.57	164.89	17.91	21.68	26.19	0	1.55E-23
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabuľka 3: Pravdepodobnosti a hodnoty jednotlivých optických veličín pre logistickú regresiu.

Literatúra

1. Hristev R. M.: *The ANN Book.* GNU Public License, 1998.

2. Kvasnička V., Beňušková Ľ., Pospíchal J., Farkaš I., Tiňo P., Kráľ A.: *Úvod do teórie neurónových sietí.* IRIS, Bratislava, 1997.

3. Mařík V., Štěpánková O., Lažanský J. a kol.: *Umělá inteligence 1.* Academia, Praha, 2001.

4. Mařík V., Štěpánková O., Lažanský J. a kol.: *Umělá inteligence 4.* Academia, Praha, 2003.

5. Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P.: *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, New York, 2002.

6. Everitt B. S.: *Statistical Analyses Using S-PLUS.*, Chapman & Hall, 1995.

7. Venables W. N., Ripley B. D.: *Modern Applied Statistics with S-PLUS.* Springer-Verlag, 1994.

8. Hosmer D. W., Lameshow S.: *Applied Logistic Regression.* Wiley, New York, 2000.

O asociácii náhodných veličín a kopuliach.

Peter Volauf

Katedra matematiky FEI STU, Ilkovičova 3, 812 19 Bratislava

peter.volauf@stuba.sk

Abstrakt. V prvej časti príspevku hovoríme o rôznych, hlavne kvalitatívnych (snáď menej známych) mierach závislosti náhodných veličín. V druhej časti krátko predstavíme pojem kopuly a ukážeme, ako kopuly umožňujú formulovať a vyšetrovať rôzne typy závislosti náhodných veličín. Na záver poukážeme na to, že s pomocou kopúl je možné zostrojiť náhodný vektor, ktorého rozdelenie má predpísané vlastnosti.

1 Asociácia náhodných veličín

Termín *asociácia* sa používa v dvoch významoch. V širšom zmysle asociácia znamená súvislosť medzi zložkami náhodného vektora, t.j. stochastickú väzbu medzi náhodnými veličinami. V užšom zmysle je asociovanosť konkrétny pojem, ktorý sa radí medzi viaceré tie pojmy, ktoré *kvalitatívne* vypovedajú o stochastickej väzbe medzi veličinami.

Keď sa v základnom kurze pravdepodobnosti a štatistiky na technických univerzitách hovorí o mierach závislosti medzi náhodnými veličinami, tak sa v prvom rade hovorí o Pearsonovom korelačnom koeficiente ako o kvantitatívnej charakteristike, ktorá zachytáva lineárnu väzbu medzi zložkami náhodného vektora. Žiaľ, z nedostatku času, často nemáme možnosť hovoriť o oboch verziách, to znamená o teoretickej a aj o výberovej verzii tohoto koeficientu. Ďalšími kvantitatívnymi mierami závislosti sú napr. Spearmanov korelačný koeficient, alebo korelačný pomer. Kým o prvom môžeme povedať, že detekuje monotónny vzťah medzi veličinami, korelačný pomer detekuje funkcionálny vzťah medzi veličinami, pretože platí: $K_X(Y) = 1$ práve vtedy, ak $Y = g(X)$, pre nejakú borelovsky merateľnú funkciu g (viď kap. 5 Rényiho knihy [7]). Rényi diskutuje aj ďalšie kvantitatívne miery závislosti, napr. kontingenciu, alebo maximálnu koreláciu.

Druhá možnosť ako hovoriť o stochastickej väzbe, t.j. o závislosti náhodných veličín, je hovoriť o *kvalitatívnych* mierach závislosti. Napríklad, hovorme, že X a Y sú *pozitívne (nezáporne) korelované*, ak $cov(X, Y) \geq 0$. Teraz nás netrápi, že kovariancia má známu vlastnosť: $cov(aX, bY) = ab.cov(X, Y)$, vďaka ktorej sa pri *kvantitatívnom* vyšetrovaní dáva prednosť Pearsonovmu momentovému korelačnému koeficientu. Teraz nás samotná hodnota kovariancie nezaujíma. Ide len o to, či je nezáporná. Reláciu (R1) nezápornej korelovanosti zosilnime takto:

X, Y sú v relácii (R2), práve ak $cov(f(X), g(Y)) \geq 0$, pre všetky neklesajúce f, g

pre ktoré kovariancia existuje. Zrejme ak X(R2)Y, tak X(R1)Y a na relatívne jednoduchom príklade je možné ukázať, že relácie R1, R2 nie sú (vo všeobecnosti) identické. Ďalšie zosilnenie relácie (R2) prináša pojem *asociovanosti* (asociovanosť v užšom zmysle):

X, Y nazývame **asociované,** ak $cov(f(X, Y), g(X, Y)) \geq 0$, pre všetky neklesajúce f, g

pre ktoré kovariancia existuje (f, g sú reálne funkcie *dvoch* premenných – funkciu dvoch premenných nazývame neklesajúca, ak je neklesajúcou v každej premennej).

Pojem asociovanosti zaviedli Esary, Proschan a Walkup v článku [1]. Pomocou tohoto pojmu (a odhalených súvislostí s inými pojmami) dosiahli zjednodušenie niektorých skutočností (napr. dôkazov nerovností, ako ilustruje nasledujúca nerovnosť:

Ak $X_1, X_2, ... X_n$ sú nezávislé a S_i je postupnosť ich čiastočných súčtov, tak platí:

$$P(S_1 < s_1, S_2 < s_2, ..., S_n < s_n) \geq P(S_1 < s_1).P(S_2 < s_2)...P(S_n < s_n)$$

Bez asociovanosti ju dokázal Robbins [5]. Ak však využijeme asociovanosť a súvislosti z nej vyplývajúce, nerovnosť sa dá dokázať podstatne jednoduchšie (viď [1]).

Predtým než sformulujeme základné vlastnosti a súvislosti, uvedieme dve technické lemy, ktoré v ďalšom nájdu časté uplatnenie.

Prvou je Čebyševova "druhá" veta:

Pre každú náhodnú veličinu platí: $cov(f(X), g(X)) \geq 0$, pre neklesajúce f, g.

Dôkazový trik spočíva v uvažovaní nezávislej kópie Y veličiny X, teda vo využití dvojice X, Y, kde Y je nezávislá s X, majúca to isté rozdelenie. Potom zrejme z neklesajúcosti f a g

$$(f(X) - f(Y)).(g(X) - g(Y)) \geq 0$$

a preto $$E[(f(X) - f(Y)).(g(X) - g(Y))] \geq 0$$

z čoho po roznásobení a využití predpokladu o X, resp. Y, rezultuje záver. Skutočnosť, že $cov(f(X), g(X)) \geq 0$, pre neklesajúce f, g znamená, že singleton $\{ X \}$ je asociovaná množina.

Druhou (a rozhodujúcou) technickou pomôckou pre nasledujúce úvahy je **Hoeffdingova lema** (r. 1940):

$$\operatorname{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H(x, y) - F(x)G(y)) \, dx \, dy$$

Dôkaz opäť vyžaduje uvažovanie "nezávislej kópie", ale tentoraz vektora. Nech (X_1, Y_1), (X_2, Y_2) sú nezávislé vektory, majúce rovnakú distribúciu $H(x, y)$. Ďalej sa uplatní trochu netypická práca s integrálmi charakteristických funkcií, teda s indikátormi. Napr. platí

$$(X_1 - X_2)(\omega) = \int_{-\infty}^{\infty} (I_x(X_2(\omega)) - I_x(X_1(\omega))) \, dx$$

kde I_x znamená indikátor intervalu $(-\infty, x)$. Zrejme z uvedených predpokladov vyplýva:

$$2.cov(X_1, Y_1) = 2.(E(X_1 Y_1) - E(X_1).E(Y_1)) = E[(X_1 - X_2).(Y_1 - Y_2)] =$$

$$= 2 E (\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (I_x(X_2) - I_x(X_1)).(I_x(Y_2) - I_x(Y_1)) \, dx \, dy)$$

Ak teraz prejdeme operátorom strednej hodnoty za znak integrálu a uvážime, že platí napr.:

$$E(I_x(X_1).I_y(Y_1)) = P(X_1 < x, Y_1 < y) = H(x, y),$$

dostávame žiadaný vzťah.

Hore uvedená definícia asociovanosti dvoch veličín sa zovšeobeňuje na konečnú množinu veličín $X_1, ..., X_n$ celkom prirodzene:

$X_1, .., X_n$ sú **asociované,** ak $cov(f(X_1, .., X_n), g(X_1, .., X_n)) \geq 0$,

pre všetky neklesajúce funkcie f, g (pre ktoré existuje kovariancia).

Pod neklesajúcosťou funkcie *n* premenných rozumieme opäť neklesajúcosť funkcie v každej premennej. Základné fakty a vlastnosti pojmu asociovanosť:

1) Ak $X_1, X_2, .., X_n$ sú asociované, tak každá ich podmnožina je množinou asociovaných veličín.
2) Ak $X_1, X_2, .., X_n$ sú asociované, tak $-X_1, -X_2, .., -X_n$ sú asociované.
3) Ak $X_1, .., X_n$ sú asociované, $Y_j = f_j(X_1, .., X_n)$, kde f_j sú neklesajúce, tak $Y_1, .., Y_m$ sú asociované.
4) Ak $X_1, ..., X_n$ sú asociované, $Y_1, ..., Y_m$ sú asociované, a pritom vektory $(X_1, ..., X_n)$, $(Y_1, ..., Y_m)$ sú nezávislé, potom $X_1, ..., X_n, Y_1, ..., Y_m$ sú asociované.

Dôkaz bodov 1, 2, 3 nerobí problém. Dôkaz bodu 4 je náročnejší a vyžaduje netriviálnu prácu s podmieňovaním (viď [1]).

Dôsledok bodu (3):
Ak $X_1, X_2, .., X_n$ sú asociované, tak pre neklesajúce funkcie h_i, sú veličiny $h_1(X_1), h_2(X_2), ..., h_n(X_n)$ asociované, čo znamená, že asociovanosť je invariantná na neklesajúce transformácie (to sa označuje slovami, že asociovanosť je "scale-free dependence concept").

Dôsledok bodu (4):
Nezávislé veličiny $X_1, ..., X_n$ sú asociované (a to je možno prekvapujúce).

Ako príklady asociovaných veličín môžu slúžiť napr.:

1) Ak $Y = u(X)$, kde $u(.)$ je neklesajúca, tak X, Y sú asociované.
2) Ak $X_1, ..., X_n$ je náhodný výber, tak poriadkové štatistiky $X_{(1)}, ..., X_{(n)}$ sú asociované (zrejme každá $X_{(i)}$ je neklesajúcou funkciou náhodného výberu $X_1, ..., X_n$).
3) Ak $X_1, ..., X_n$ sú nezávislé, tak *n*-tica čiastočných súčtov $S_1, S_2, ..., S_n$ je asociovaná (ide zrejme o dôsledok bodu 3).
4) Ak $(X_1, ..., X_n)$ je gaussovský vektor, tak ide o asociovanú *n*-ticu práve vtedy, ak $cov(X_i, X_j) \geq 0$, pre všetky i, j (viď [6]).

Ešte pred zrodom pojmu *asociovanosť* boli známe aj iné pojmy, ktoré kvalitatívne popisujú závislosť. Niektoré z nich skúmal Lehman v [3], medzi nimi *pozitívnu kvadrantnú závislosť* – PQD (positive quadrant dependence).

Hovoríme, že X, Y sú pozitívne kvadrantne závislé (PQD), ak pre všetky reálne x, y platí

$$P(X < x, Y < y) \geq P(X < x).P(Y < y).$$

Zo spojitosti pravdepodobnosti vyplýva, že znaky nerovností v uvažovaných udalostiach môžu byť neostré (musia mať rovnaký smer na oboch stranách nerovnosti). Ďalej, znaky nerovností (vyznačujúce udalosti) môžu mať opačný smer (avšak, potom všetky). Napríklad, hore uvedená podmienka je ekvivalentná podmienke

$$P(X > x, Y > y) \geq P(X > x).P(Y > y).$$

Je podstatné, že pojem PQD je invariantný vzhľadom na neklesajúce transformácie, to znamená, že platí (a pomerne ľahko sa dokáže):

Ak X, Y sú PQD, tak pre všetky neklesajúce f, g platí, že $f(X), g(Y)$ sú PQD.

Hoeffdingova lema umožňuje formulovať PQD cez kovarianciu:

X, Y sú PQD práve vtedy, keď $cov(f(X), g(Y)) \geq 0$, pre všetky neklesajúce f, g.

Dôkaz. Nech X, Y sú PQD. Potom aj $f(X)$ a $g(Y)$ sú PQD a nerovnosť $cov(f(X), g(Y)) \geq 0$ vyplýva z Hoeffdingovej lemy, pretože integrál z nezáporného integrandu je nezáporný. Teraz naopak, nech $cov(f(X), g(Y)) \geq 0$, pre všetky neklesajúce f, g. Potom pre všetky x, y máme

$$0 \leq \mathrm{cov}(1 - \mathbf{I}_{[x, \inf)}(X), 1 - \mathbf{I}_{[y, \inf)}(Y)) = \mathrm{cov}(\mathbf{I}_{(-\inf, x)}(X), \mathbf{I}_{(-\inf, y)}(Y)) =$$

$$= E(\mathbf{I}_{(-\inf, x)}(X).\mathbf{I}_{(-\inf, y)}(Y)) - E(\mathbf{I}_{(-\inf, x)}(X)).E(\mathbf{I}_{(-\inf, y)}(Y)) =$$

$$= P(X < x, Y < y) - P(X < x).P(Y < y).$$

Vzťahy medzi uvedenými pojmami určujú nasledujúce implikácie (ASS je asociovanosť, NC je nezáporná korelovanosť):

$$\mathrm{ASS} \Rightarrow \mathrm{PQD} \Rightarrow \mathrm{NC}$$

pričom žiadna dvojica pojmov (vo všeobecnosti) nie je ekvivalentná. Je zaujímavé, že pre binárne náhodné veličiny sú pojmy ASS, PQD a NC totožné.

Lehman je autor nasledujúcej zaujímavej súvislosti:

Ak $X(\mathrm{PQD})Y$, a pritom $cov(X, Y) = 0$, tak X, Y sú nezávislé.

Dôkaz tohoto tvrdenia ilustruje zásadný význam Hoeffdingovej lemy:

$$0 = \mathrm{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (P(X < x, Y < y) - P(X < x).P(Y < y))\, dx\, dy$$

Vďaka PQD je integrand nezáporná funkcia. Ak ale integrál sa rovná nule, tak integrand $H(x, y) - F(x).G(y)$ musí byť funkcia rovná nule skoro všade. Vďaka spojitosti zľava distribučných funkcií to ale znamená, že rovnosť $H(x, y) = F(x).G(y)$ platí všade, čo bolo treba ukázať.

Neskoršieho dáta je viacrozmerná analógia tohoto tvrdenia:

Ak $X_1, ..., X_n$ sú asociované a $cov(X_i, X_j) = 0$, pre všetky i, j $(i \neq j)$, potom $X_1, ..., X_n$ sú nezávislé.

(Dôkaz je podstatne zložitejší, viď [2], resp. [8]).

2 Kopule

Pojem kopule sa zrodil v článku [9] avšak prebehli tri desaťročia, kým sa objavil napr. v Encyclopedia of Statistics. Pritom (ako za chvíľu ukážeme) pojem kopuly umožňuje dať odpoveď na celý rad zaujímavých otázok teórie pravdepodobnosti. Systematický výklad problematiky predstavuje knižka [4]. Naším cieľom je predstaviť pojem a naznačiť jeho uplatnenie. Neformálne je možné pojem kopuly definovať pomocou pojmu náhodný vektor takto:
Nech (X, Y) je náhodný vektor s distribučnou funkciou $H(x, y)$, pričom nech $F(x)$ a $G(y)$ sú distribučné funkcie zložiek X a Y. Predpokladajme, že F, G sú spojité (ak za ich definičný

obor berieme [-∞ ; ∞], tak za uvedeného predpokladu spojitosti je obor ich hodnôt interval [0, 1]). Za týchto predpokladov je možné uvažovať zobrazenie z $[-\infty, \infty]^2$ do $[0, 1]^3$, ktoré bodu (x, y) priradí bod $(F(x), G(y), H(x, y))$.

Kopula je zobrazenie, ktoré dvojici $(F(x), G(y))$ priraďuje $H(x, y)$, to znamená, že pre všetky $x, y \in [-\infty, \infty]$ platí:

$$C(F(x), G(y)) = H(x, y)$$

Ako vidíme, kopula $C(.,.)$ "spája" hodnoty marginálnych distribúcií a z nich vytvorí hodnotu združenej distribučnej funkcie. Formálna definícia sa však neopiera o pojem náhodného vektora.

Definícia. 2-dimenzionálna kopula je funkcia C: $[0, 1]^2 \to [0, 1]$ s vlastnosťami:

(c1) $C(0, u) = C(u, 0) = 0$, $C(1, u) = C(u, 1) = u$

(c2) C je "2-increasing", t.j. ak $a < b$, $c < d$, tak

$$C(b, d) - C(a, d) - C(b, c) + C(a, c) \geq 0$$

Dôsledky vlastností (c1) a (c2) sú napr.:

(c3) $C(.,.)$ je neklesajúca v každej premennej.

(c4) $C(.,.)$ je Lipschitzovská (a preto rovnomerne spojitá), t.j.

$$| C(b, d) - C(a, c) | \leq | b - a | + | d - c |$$

Zásadný význam má **Sklarova veta** ([9], resp. [4]): Ak $F(.)$, $G(.)$ sú distribučné funkcie a $C(.,.)$ je copula, tak funkcia H definovaná vzťahom

$$H(x, y) = C(F(x), G(y))$$

je distribučnou funkciou, ktorej marginálami sú práve funkcie F a G. Naopak, ak $H(.,.)$ je združená distribučná funkcia a F, G sú jej marginálnymi funkciami, tak existuje copula $C(.,.)$, že platí

$$H(x, y) = C(F(x), G(y))$$

(ak F, G sú spojité, tak $C(.,.)$ je jednoznačná).

Ukážeme, že 2-dimenzionálnu kopulu možno chápať ako distribučnú funkciu dvojrozmerného náhodného vektora. Nech (X, Y) je náhodný vektor so združenou distribučnou funkciou $H(. , .)$ a nech marginálne distribučné funkcie F, G sú spojité.

Ak $C(.,.)$ je kopula vektora (X, Y), tak $C(.,.)$ môžeme chápať ako distribučnú funkciu vektora (U, V), kde $U = F(X)$, $V = G(Y)$. Je dobre známe, že rozdelenie veličiny U a tiež rozdelenie V je rovnomerné rozdelenie na intervale [0, 1] (vďaka spojitosti funkcií F a G). Dôkaz je celkom jednoduchý (symboly $F^{(-1)}$ a $G^{(-1)}$ sú kvantilové funkcie, teda pseudoinverzné funkcie ku F a G). Platí

$$F_{UV}(u, v) = P(F(X) < u, G(Y) < v) = P(X < F^{(-1)}(u), Y < G^{(-1)}(v)) =$$

$$= H(F^{(-1)}(u), G^{(-1)}(v)) = C(F(F^{(-1)}(u)), G(G^{(-1)}(v))) = C(u, v)$$

Teraz ukážme, že kopula je invariantná k neklesajúcim transformáciám, to znamená, že platí:

Ak α, β sú neklesajúce funkcie, tak kopula vektora $(\alpha(X), \beta(Y))$ je tá istá, ako kopula vektora (X, Y).

Dôkaz. Pseudoinverzné funkcie ku α, resp. β označme α^* , resp. β^*. Platí

$H_{\alpha X\, \beta Y}(u, v) = P(\alpha X < u,\ \beta Y < v) = P(X < \alpha^*(u), Y < \beta^*(v)) = H_{X\,Y}(\alpha^*(u), \beta^*(v))$

avšak

$H_{\alpha X\, \beta Y}(u, v) = C_{\alpha X\ \beta Y}(F_{\alpha X}(u), G_{\beta Y}(v)) = C_{\alpha X\ \beta Y}(F_X(\alpha^*(u)), G_Y(\beta^*(v)))$

a na druhej strane

$H_{X\,Y}(\alpha^*(u), \beta^*(v)) = C_{X\,Y}(F(\alpha^*(u)), G(\beta^*(v)))$

Pretože F a G sú spojité, kopula je jednoznačná a rovnosť $C_{X\,Y}(u, v) = C_{\alpha X\, \beta Y}(u, v)$ je dokázaná.

Je prirodzené pýtať sa, ako k danému 2-rozmernému rozdeleniu nájsť copulu. Podľa Sklarovej vety

$$H(x, y) = C(F(x), G(y))$$

a preto

$$C(u,v) = H(F^{(-1)}(u), G^{(-1)}(v))$$

Napríklad, nech distribučná funkcia vektora (X, Y) je daná vzťahom

$$H(x, y) = x\,y(x + y)/2 \quad \text{na } [0, 1]^2.$$

To znamená, že vektor má hustotu $f(x, y) = x + y$ na $[0, 1]^2$ (mimo štvorca je hustota nulová) a marginálnymi distribučnými funkciami sú funkcie

$F(x) = x(1 + x)/2$ na $[0, 1]$, $G = F$ a pre inverziu máme $F^{(-1)}(u) = (-1 + \mathrm{sqrt}(1+8u))$

Preto kopulou vektora (X, Y) je funkcia

$$C(u, v) = (1/16)(-1 + \mathrm{sqrt}(1 + 8u)).(-1 + \mathrm{sqrt}(-1 + 8v)).(-2 + \mathrm{sqrt}(1 + 8u) + \mathrm{sqrt}(1 + 8v))$$

Pojem kopuly umožňuje zachytiť, resp. popísať závislosť náhodných veličín. Ako vieme z prvej časti príspevku, relácia pozitívnej kvadrantnej závislosti znamená, že platí

$$P(X < x, Y < y) \geq P(X < x).P(Y < y)$$

a teda $$C_{XY}(F(x), G(y)) \geq F(x).G(y)$$

čo znamená, že nerovnosť

$$C_{XY}(u, v) \geq u.v$$

platí pre všetky body jednotkového štvorca. Takto jednoducho kopula vektora (X, Y) popisuje reláciu PQD veličín X, Y.

Dá sa ľahko ukázať, že pre akúkoľvek kopulu platí *Frechet-Hoeffdingova* nerovnosť:

$$max(u + v - 1; 0) \leq C(u, v) \leq min(u, v)$$

a podľa uvedeného X, Y sú v relácii PQD práve vtedy, keď

$$u.v \leq C(u, v) \leq min(u, v)$$

V nasledujúcom ukážeme, ako kopule umožňujú získať dvojrozmerné rozdelenie požadovaných vlastností. Predpokladajme, že chceme navrhnúť triedu dvojrozmerných rozdelení, ktoré majú predpísané marginálne distribučné funkcie *F, G,* pričom korelácia zložiek má byť funkciou parametra tak, aby ju bolo možné zmenou parametra meniť.

Uvažujme triedu kopúl FGM (*Farlie-Gumbel-Morgenstern*), ktorá je daná vzťahom

$$C(u, v; \theta) = u.v + \theta u.v(1 - u)(1 - v), \quad -1 \le \theta \le 1$$

Ak kopulu $C(.,.; \theta)$ chápeme ako dvojrozmernú distribučnú funkciu rovnomerne rozdelených veličín U, V, ľahko sa vypočíta, že Pearsonov korelačný koeficient zložiek má hodnotu $\theta/3$. Keďže pre spojité náhodné veličiny X, Y, ich Spearmanov korelačný koeficient sa rovná Pearsonovmu koeficientu veličín $F(X)$, $G(Y)$, tak konštrukcia dvojrozmernej distribučnej funkcie prostredníctvom kopuly $C(u, v; \theta)$ umožňuje získať rozdelenie vektora (X, Y) s predpísanými marginálnymi rozdeleniami a so Spearmanovým koeficientom, ktorého hodnota môže byť ľubovoľné číslo z intervalu [-1/3, 1/3].

Pre konkrétnosť, predpíšme marginálne distribúcie napr. takto:

$F(x) = x$, pre $x \in [0, 1]$, resp. $G(y) = y^2$, pre $y \in [0, 1]$ (definovanie mimo [0, 1] je zrejmé). Podľa Sklarovej vety

$$H(x, y) = C(F(x), G(y))$$

a preto

$$H(x, y) = x.y^2 + \theta xy^2(1 - x)(1 - y^2)$$

je distribučná funkcia s predpísanými marginálnymi rozdeleniami, a pritom Spearmanov korelačný koeficient vektora (X, Y) s touto distribučnou funkciou $H(.,.)$ sa rovná $\theta/3$, čo znamená, že môže nadobúdať ľubovoľnú hodnotu z intervalu [-1/3, 1/3].

Ako vidíme, trieda kopúl FGM umožňuje získať združené rozdelenie, ktorého zložky majú relatívne slabú stochastickú väzbu (pokiaľ ju meriame Spearmanovým korelačným koeficientom). Nasledujúca konštrukcia sa opiera o dvojrozmerné rozdelenie predstavené v práci [10] ako *diagonal band distribution*. Ide o jednoparametrické rozdelenie na jednotkovom štvorci, ktorého zložky majú rovnomerné rozdelenie, pričom korelačný koeficient (závisí na parametri α) môže nadobúdať ľubovoľnú hodnotu z intervalu [-1; 1]. Rozdelenie bude definované hustotou $b(u, v; \alpha)$, ktorá na štvorci $[0, 1]^2$ nadobúda len tri hodnoty.

Nech $\alpha \in (0, 1)$. Položme $a = 1 - \alpha$ a definujme hustotu $b(u, v; \alpha)$ v bodoch trojuholníka určeného vrcholmi [0, 0], [a, 0], [0, a] hodnotou $1/a$. Rovnakú hodnotu nech má hustota $b(u, v; \alpha)$ v bodoch trojuholníka s vrcholmi [α, 1], [1, α], [1, 1]. V bodoch obdĺžnika určeného bodmi [0, a], [a, 0], [1, α] a [α, 1] nech je $b(u, v; \alpha)$ definovaná hodnotou $1/(2a)$. V ostatných bodoch štvorca má hustota $b(u, v; \alpha)$ hodnotu 0. Pre $\alpha = 0$ hodnota hustoty $b(u, v; \alpha)$ sa rovná 1 vo všetkých bodoch štvorca (zrejme zložky v tomto prípade sú stochasticky nezávislé a korelačný koeficient sa rovná nule). Pre $\alpha = 1$ uvažované rozdelenie hustotu nemá a za limitné rozdelenie (pre $\alpha \to 1$) považujeme singulárne rovnomerné rozdelenie na diagonále štvorca, pričom platí: P(X = Y) = 1 (a zrejme korelačný koeficient zložiek sa rovná 1).

Štandardným (aj keď zdĺhavým) výpočtom sa dá ukázať, že Pearsonov korelačný koeficient zložiek rozdelenia určeného hustotou $b(u, v; \alpha)$ má hodnotu $-\alpha^3 + \alpha^2 + \alpha$. Ak α prebieha interval [0, 1], hodnoty tohoto výrazu prebiehajú celý interval [0, 1]. To znamená, že *distribučná* funkcia s hustotou $b(u, v; \alpha)$ predstavuje kopulu, ktorá umožňuje konštrukciu náhodného vektora s predpísanými spojitými magrinálnymi distribúciami a predpísanou hodnotou Spearmanovho korelačného koeficientu.

Poznamenávame, že pre hodnoty parametra α z intervalu [-1, 0] je potrebné hustotu definovať symetricky vzhľadom na doteraz popísanú konštrukciu $b(u, v; \alpha)$, a to symetricky podľa priamky y = 0,5 (hodnoty korelácie budú prebiehať interval [-1, 0].

Popísaná kopula umožňuje definovať náhodný vektor s predpísanými marginálnymi distribučnými funkciami a predpísanou hodnotou Spearmanovho korelačného koeficientu, ktorá môže byť ľubovoľnou hodnotou intervalu [-1, 1].

Literatúra:

1. Esary, J., Proschan, F., Walkup, D.: Association of random variables with applications. Ann. Math. Statist. 38, (1967), pp. 1466 – 1474.
2. Lebowitz, J.: Bounds on the correlations and analyticy properties of ferromagnetic Ising spin systems. Comm. Math. Phys. 28, (1972), pp. 313 – 321.
3. Lehmann, E. L.: Some concepts of dependence. Ann. Math. Statist. 37, (1966), pp. 1137 – 1153.
4. Nelsen, R. B.: An Introduction to Copulas. Lecture Notes in Statistics. Springer, New York, 1999.
5. Robbins, H.: A remark on the joint distribution of cumulative sums. Ann. Math. Statist. 25, (1954), pp. 614 – 616.
6. Pitt, L. D.: Positively correlated normal variables are associated. Ann. Probability, 10, (1982), pp. 496 – 499.
7. Rényi, A.: Teorie pravděpodobnosti. Academia, Praha 1972.
8. Volauf, M.: Asociované náhodné premenné. Diplomová práca, MFF UK, 1998.
9. Sklar, A.: Fonctions de répartition a n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, (1959), pp. 229 – 231.
10. Waij, R., Cook, R.M.: Monte Carlo sampling for generalized knowledge dependence with application to human reliability. Risk Analysis, 6, (1986), pp. 335 – 343.

Príspevok vznikol s podporou grantu VEGA 1/0085/03.