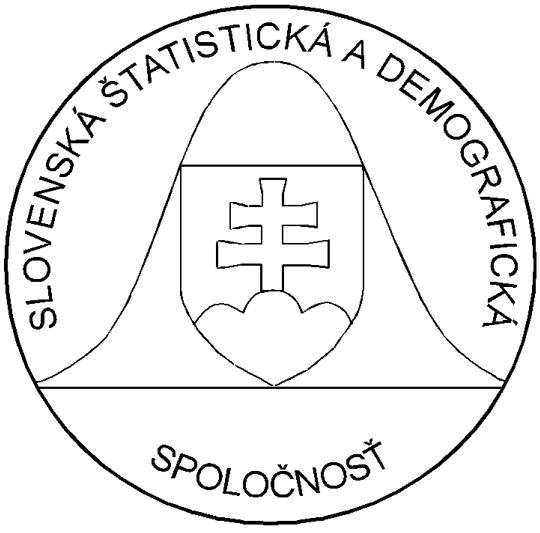
3/2009

# FORUM STATISTICUM SLOVACUM




ISSN 1336-7420

# ÚVOD

Vážené kolegyne, vážení kolegovia,

tretie číslo piateho ročníka časopisu, ktorý vydáva Slovenská štatistická a demografická spoločnosť (SŠDS) je zostavené z príspevkov, ktoré autori pripravili pre konferenciu PRASTAN 2009, ktorá sa uskutočnila v dňoch 10. 6. - 12. 6. 2009 v Kočovciach, v učebno-výcvikovom zariadení Stavebnej fakulty STU v Bratislave. Konferenciu organizovala Slovenská štatistická a demografická spoločnosť v spolupráci so Stavebnou fakultou STU, Fakultou managementu UK a Štatistickým úradom SR. Na konferencii odzneli dve pozvané prednášky - Marián Grendár: *Bayesovská štatistika*, Ján Luha: *Matematicko-štatistické aspekty spracovania dotazníkových výskumov*.

Programový a organizačný výbor pracoval v zložení: Martin Kalina, Oľga Nánásiová – predsedovia, členovia: Mária Bohdalová, Michal Greguš, Angela Handlovičová, Jozef Chajdiak, Jozef Komorník, Magda Komorníková, Ján Luha, Mária Minárová, Iveta Stankovičová, Jiří Vala, Jan Ámos Víšek, Gejza Wimmer.

Na príprave a zostavení tohto čísla participovali: Mária Bohdalová, Martin Kalina, Mária Minárová, Oľga Nánásiová, Iveta Stankovičová,

Editori tohto čísla ďakujú recenzentom za rýchlu a kvalitnú prácu.

Výbor SŠDS

# Výučba štatistiky na Strojníckej fakulte Technickej univerzity Košice
# Teaching of Statistics at The Faculty of Mechanical Engineering of Technical University Košice

Andrejiová Miriam

**Abstract:** Department of applied mathematics of the Faculty of Mechanical Engineering at Technical University Košice provides mathematics teaching at three faculties - Faculty of Mechanical Engineering, Faculty of Metallurgy, Faculty of Mining, Ecology, Process Control and Geotechnologies. The aim of this paper is to present the statistics teaching at the Faculty of Mechanical Engineering in the academic year 2008/2009.

**Key words:** statistics, teaching statistics.

**Kľúčové slová:** štatistika, výučba štatistiky.

## 1. Úvod

Viac ako 30 rokov sa na našej katedre vyučoval predmet Matematika IV, ktorý zahŕňal základy teórie pravdepodobnosti, popisnú štatistiku, teóriu odhadu, testovanie hypotéz, korelačnú a regresnú analýzu. Matematika IV bola súčasťou základného kurzu matematiky a bola povinná pre všetkých študentov inžinierskeho štúdia Strojníckej fakulty.

Prelomovým a pre výučbu štatistiky zvlášť nepriaznivým rokom bol akademický rok 2000/2001. V tomto roku došlo k redukcii počtu hodín a povinný predmet Matematika IV bol nahradený voliteľným predmetom Pravdepodobnosť a matematická štatistika, ktorý končí klasifikovaným zápočtom. V 4. ročníku denného štúdia naďalej ostal predmet Štatistické metódy, ktorý bol povinný len pre jediný zo 16 inžinierskych študijných odborov: Kvalita produkcie a bezpečnosť technických systémov.

Tabuľka 1 ukazuje prehľad výučby štatistických predmetov a počet študentov, ktorí dané predmety absolvovali od akademického roku 2000/2001 do 2005/2006.

***Tabuľka 1: Výučba štatistiky v rokoch 2000 - 2006***

| **Predmet** | **Počet študentov** | | | | | |
|---|---|---|---|---|---|---|
| | **00/01** | **01/02** | **02/03** | **03/04** | **04/05** | **05/06** |
| **Pravdepodobnosť a mat. štatistika**<br>2.ročník DŠ, Ing., SjF, V, 2/2, kz<br>Odbor: bez zamerania | N | 5 | 7 | 7 | 6 | 11 |
| **Štatistické metódy**<br>4. ročník DŠ, Ing, SjF, P, 2/3, z/s<br>Odbor: Kvalita produkcie a bezpečnosť technických systémov | 25 | 26 | 30 | 26 | 22 | 21 |
| **Matematická štatistika**<br>2. ročník EŠ, Bc, SjF, P, 2/3, z/s<br>Odbor: Kvalita produkcie a bezpečnosť technických systémov | × | × | × | 14 | 28 | 36 |
| **Štat. metódy v environmentalistike**<br>4. ročník EŠ, Ing, SjF, V, 2/3, kz<br>Odbor: Technika ochrany životného prostredia | × | × | × | × | 28 | 25 |

(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)

## 2. Štatistika - 1.stupeň vysokoškolského štúdia

Od akademického roku 2005/2006 nastúpila naša fakulta na trojstupňové štúdium. Študijné plány bakalárskych študijných programov, ktoré má fakulta akreditované, sú pre výučbu štatistiky v porovnaní s predchádzajúcimi rokmi omnoho priaznivejšie.

Tri študijné programy (Environmentálne manažérstvo, Priemyselné inžinierstvo, Technika ochrany životného prostredia) majú vo svojich študijných plánoch povinný predmet, ktorého osnova zahrňuje základy teórie pravdepodobnosti, popisnú štatistiku, teóriu odhadu, testovanie hypotéz, korelačnú a regresnú analýzu (tabuľka 2). Najrozsiahlejší kurz z matematickej štatistiky a pravdepodobnosti v bakalárskom štúdiu má študijný program Kvalita produkcie (KP), v ktorom je v priebehu 3-ročného bakalárskeho štúdia naplánovaných týždenne 8 hodín prednášok a 6 hodín cvičení.

***Tabuľka 2: Povinné štatistické predmety – 1. stupeň***

| **Študijný program /Predmet** |
|---|
| **Environmentálne manažérstvo (EM)** |
| Štatistika pre environmentalistov (2.ročník, ZS, 2/2, z/s) |
| **Priemyselné inžinierstvo (PI)** |
| Štatistické metódy (2.ročník, LS, 2/2, z/s) |
| **Technika ochrany životného prostredia (TOŽP)** |
| Štatistika pre environmentalistov (2.ročník, ZS, 2/2, z/s) |
| **Kvalita produkcie (KP)** |
| Teória pravdepodobnosti (2.ročník, ZS, 2/2, z/s) |
| Štatistické metódy manažérstva kvality I (2.ročník, LS, 3/2, z/s) |
| Štatistické metódy manažérstva kvality II (3.ročník, LS, 3/2, z/s) |

So skutočne základnými pojmami a minimálnym štatistickým kurzom sa stretnú aj študenti študujúci v študijnom programe Automobilová výroba, Bezpečnosť a ochrana zdravia pri práci, Počítačová podpora strojárskej výroby a Prototika a ortotika (tabuľka 3), a to vďaka tomu, že obsahová náplň niektorých povinných predmetov 1. ročníka je doplnená práve o štatistické minimum (popisná štatistika, základné pojmy z pravdepodobnosti). Zvyčajne ide len o 6 hodín prednášok a 6 hodín cvičení počas semestra.

***Tabuľka 3: Predmety s doplneným obsahom štatistiky – 1. stupeň***

| **Študijný program / Predmet** |
|---|
| **Automobilová výroba /** Matematika II (1.ročník, LS, 3/2, z/s) |
| **Bezpečnosť a ochrana zdravia pri práci /**Aplikovaná matematika (1.ročník, LS, 2/2, z/s) |
| **Počítačová podpora strojárskej výroby /** Aplikovaná matematika (1.ročník, LS, 2/2, z/s) |
| **Prototika a ortotika /** Aplikovaná matematika (1.ročník, LS, 2/2, z/s) |

Možnosť vybrať si predmet priamo súvisiaci so štatistikou ponúkajú v 2. ročníku dva študijné programy – Počítačová podpora a strojárska výroba (PPSV) a Prototika a ortotika (PaO). Študenti majú možnosť vybrať si povinne voliteľný predmet Štatistické metódy, ktorý končí klasifikovaným zápočtom (tabuľka 4). Počet tých študentov, ktorí si daný predmet zapíšu a predovšetkým aj úspešne absolvujú, je každoročne veľmi nízky.

Aj napriek našej snahe rozšíriť výučbu štatistky vo všetkých študijných programoch bakalárskeho štúdia, máme tri študijné programy (Mechatronika, Prevádzka a údržba strojov, Všeobecné strojárstvo), v ktorých študenti sa počas celého svojho štúdia nestretnú so základnými pojmami pravdepodobnosti a štatistiky.

*Tabuľka 4: Povinne voliteľné štatistické predmety – 1. stupeň*

| **Študijný program/Povinne voliteľný predmet** |
|---|
| **Počítačová podpora strojárskej výroby (PPSV) /** Štatistické metódy (2.ročník, ZS, 2/2, kz) |
| **Prototika a ortotika (PaO) /** Štatistické metódy (2.ročník, ZS, 2/2, kz) |

## 3. Štatistika - 2.stupeň vysokoškolského štúdia

Pokračovanie základného kurzu matematickej štatistiky v 2.stupni vysokoškolského štúdia má len veľmi malá časť z 22 akreditovaných študijných programov inžinierskeho štúdia (tabuľka 5).

*Tabuľka 5: Povinné štatistické predmety – 2. stupeň*

| **Študijný program /Predmet** |
|---|
| **Bezpečnosť technických systémov (BTS)** |
| Štatistické metódy (1.ročník, ZS, 2/3, z/s) |
| **Environmentálne manažérstvo (EM)** |
| Štatistika pre environmentalistiku (1.ročník, ZS, 2/2, z/s) |
| Štatistické spracovanie environmentálnych informácií (2.ročník, ZS, 1/2, kz) |
| **Inžinierstvo kvality produkcie (IKT)** |
| Štatistické metódy (1.ročník, ZS, 2/3, z/s) |

Aj v 2. stupni vysokoškolského štúdia máme dva študijné programy (Počítačová podpora strojárskej výroby, Strojárske technológie), v ktorých študenti majú možnosť rozšíriť si svoje vedomosti zo štatistiky a vybrať si príslušný - povinne voliteľný predmet končiaci klasifikovaným zápočtom (tabuľka 6). Nevýhodou voľby pre učiteľa a študentov je tá, že daný predmet môžu absolvovať aj tí študenti, ktorí počas 1. stupňa bakalárskeho štúdia neabsolvovali základný kurz štatistiky. Ale aj tu môžeme skonštatovať, že reálny počet študentov, ktorí si daný predmet zapísali a zapíšu je veľmi nízky.

*Tabuľka 6: Povinne voliteľné štatistické predmety – 2. stupeň*

| **Študijný program/Povinne voliteľný predmet** |
|---|
| **Počítač. podpora strojárskej výroby (PPSV) /**Štat. metódy vo výrobe (2.ročník, ZS, 2/1, kz) |
| **Strojárske technológie (ST)/** Inžinierska štatistika (1.ročník, ZS, 1/2, kz) |

Študenti ďalších dvoch odborov (Robotická technika, Automatizácia a riadenie v strojárstve) sa stretávajú so základným kurzom matematickej štatistiky v 1. ročníku v predmete Aplikovaná matematika a Matematické metódy v automatizácii.

## 4. Výučba štatistických predmetov

Po roku 2006 musela naša katedra pristúpiť k vytvoreniu a doplneniu učebných osnov niekoľkých nových štatistických predmetov podľa požiadaviek a potrieb finálnych odborných katedier. Jedným zo spôsobov skvalitnenia a zatraktívnenia výučby štatistiky, rovnako ako zintenzívnenia spolupráce s odbornými katedrami v oblasti výskumu a riešenia úloh z praxe, sa ako najlepšie riešenie ukazovalo vybudovanie vlastnej počítačovej miestnosti s vhodným softvérom. V akademickom roku 2007/2008 boli konečne všetky práce (počnúc búracích a murárskych prác až po zakúpenie počítačov a nainštalovania príslušných programov) na laboratóriu štatistických metód ukončené. Výsledkom snaženia mnohých zainteresovaných je laboratórium s 11 počítačovými terminálmi.

Pri výbere zodpovedajúcich a vhodných softvérových nástrojov sme zvolili predovšetkým OSS programy ako Maxima, Octave a Gnuplot. Tieto programy sú doplnené

Excelom a Matlabom, ktorý sa využíva vo výučbe niektorých odborných predmetov na odborných katedrách. Na katedre máme zakúpenú aj jednu licenciu programu STATISTICA a v súčasnej dobe prebieha realizácia zakúpenia multilicencie.



***Obrázok 1: Laboratórium štatistických metód***

Počet študentov, ktorí absolvujú výučbu štatistických predmetov alebo prejdú základným kurzom, je počas jedného semestra veľký. Len v zimnom semestri akademického roku 2008/2009 to bolo viac ako 400 študentov denného a externého štúdia. Z kapacitných dôvodov laboratória sme museli pristúpiť k rozdeleniu výučby štatistických predmetov na dve časti. V bakalárskom stupni štúdia prebieha výučba štatistických predmetov klasickými vyučovacími metódami bez počítačov, s využitím kalkulačiek, tabuliek a vzorcov. V tomto prípade je veľmi dôležitý aj vhodný výber úloh. Neskôr, v inžinierskom stupni štúdia, výučba prebieha pomocou počítačov a vhodne zvoleného softvéru. Aj v tomto prípade je dôležité a nevyhnutné prispôsobiť výber príkladov tematicky jednotlivým študijným programom.

## 5. Záver

Z archívnych údajov vyplýva, že len hrozivo malá časť absolventov, aj napriek ich technickému zameraniu štúdia, ovládala po roku 2000/2001 základy štatistiky. Po akademickom roku 2005/2006 nastali pre výučbu štatistiky priaznivejšie podmienky, i keď aj v súčasnosti určité nezanedbateľné percento našich budúcich bakalárov a inžinierov neabsolvuje počas štúdia niektorý zo štatistických predmetov.

Výučba štatistiky by mala čo najlepšie odpovedať potrebám katedier garantujúcich jednotlivé študijné programy. Cieľom výučby na našej katedre nie je len poskytnúť študentom základné vedomosti zo štatistiky, ale našou snahou je predovšetkým viesť študentov k ich aplikovaniu pri analyzovaní a riešení rôznych štatistických problémov z praxe.

## 6. Literatúra

[1] JADROŇOVÁ, M. – KIMÁKOVÁ, Z. 2005. Statistics teaching and its use in Environmentalistic. In: FORUM STATISTICUM SLOVACUM, č. 3, 2005, s. 20 – 25.

## Adresa autora(-ov):

Andrejiová Miriam, RNDr., PhD.
Strojnícka fakulta TU Košice
Katedra aplikovanej matematiky
Letná 9
040 00 Košice
miriam.andrejiova@tuke.sk

# Modelovanie škálovacích exponentov zrážok interpolačnými metódami[1]
# Scaling exponent of rainfall modeling by interpolation methods

Bohdal Róbert, Bohdalová Mária

**Abstract:** The goal of this work is the modelling of spatial and temporal scalling exponent of rainfall over a range of scales. The interpolating spline methods are applied to the scalling exponent of rainfall. Three different interpolation methods are employed, and examples of the results are given. All three modeling approaches are used to predict the rainfall intensity over the all places in Slovakia. These model approaches gives acceptable forecasts. Its give consistently smaller prediction errors compared to triangular methods. The models can be used to predict in real time the spatial rainfall.

**Key words:** thin plate spline, Shepard's method, rainfall, scaling exponent
**Kľúčové slová:** tenkostenný splajn, Shepardova metóda, zrážky, škálovací exponent

## 1. Úvod

Cieľom tohto článku je modelovanie škálovacích koeficientov zrážok, ktoré boli získané metódou škálovania zrážok (pozri [5], [6]). Množstvo zrážok (ich hodnoty) poskytujú zrážkomerné prístroje. Väčšina zrážkomerných prístrojov dáva informácie o jednodenných zrážkových úhrnoch, avšak pre vodohospodárske účely sú často potrebné aj údaje s väčším časovým rozlíšením. Metóda škálovania zrážok umožňuje určiť návrhové hodnoty dažďov pre zvolenú dobu opakovania pre trvania kratšie ako jeden deň, s využitím denných zrážkomerných záznamov. Zrážkomerné prístroje sú nainštalované len v niekoľkých miestach na Slovensku, avšak pre vodohospodárske účely je potrebné poznať hodnoty škálovacích koeficientov aj v miestach, kde zrážkomerné prístroje nie sú. Z tohto dôvodu sme sa rozhodli použiť interpolačné metódy, ktoré umožňujú vytvoriť model plochy reprezentujúcej zrážkovú činnosť na celom území Slovenska. Z množstva známych interpolačných metód sme vybrali tie, ktoré sú použiteľné pre ľubovoľne rozmiestnené meracie stanice (čiže také, ktoré nie sú rozmiestnené v pravouhlej mriežke).

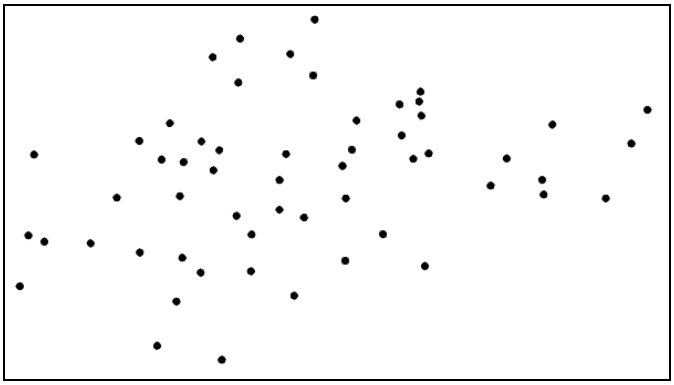
## 2. Interpolačné metódy

Majme dané geodetické súradnice $\boldsymbol{p}_i[x_i, y_i] \in \mathbb{E}^2$; $i = 1, \ldots, n$, a v nich zadané reálne hodnoty $v_i \in \mathbb{R}$, resp. namerané množstvo zrážok v jednotlivých meracích staniciach Slovenska. Našou úlohou je nájsť takú interpolačnú funkciu $f$: $\mathbb{E}^2 \to \mathbb{R}$, pre ktorú platí $f(\boldsymbol{p}_i) = v_i$, pre $i = 1, 2, \ldots, n$.

Pri výbere interpolačnej plochy treba vziať do úvahy fakt, že merné stanice sú na Slovensku rozložené v nepravidelnej mriežke (pozri Obrázok 1).

[1] Článok vznikol za podpory grantu agentúry VEGA-1/4024/07

***Obrázok 1: Rozmiestnenie zrážkomerných staníc na Slovensku***



Z vyššie uvedeného dôvodu, sme z viacerých interpolačných metód vybrali interpoláciu pomocou radiálnych bázických funkcií, známu aj pod názvom tenkostenný splajn a Shepardovu metódu a jej modifikáciu. Ich stručný opis uvádzame v nasledujúcich častiach článku.

## 2.1. Tenkostenný splajn

Interpolácia pomocou tenkostenných splajnov (interpolácia pomocou *radiálnych bázických funkcií*) je jednou z najpoužívanejších metód určených na interpoláciu nerovnomerne rozložených dát.

Metóda interpolácie tenkostenným splajnom je špeciálnym prípadom všeobecných interpolačných metód pomocou radiálnych bázických funkcií. Interpolačná funkcia $f(\boldsymbol{x})$ týchto metód má v priestore $\mathbb{E}^d$ nasledujúce vyjadrenie ([1]):

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i R_{d,k}\left(\|\boldsymbol{x} - \boldsymbol{p}_i\|\right) + \sum_{|\boldsymbol{\alpha}|<k} \boldsymbol{c}_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}\,, \tag{1}$$

kde $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdot \cdots \cdot x_d^{\alpha_d}$ , $(\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, $k \in \mathbb{N}$,

a $R_{d,k}(r)$ je trieda radiálnych bázických funkcií v tvare:

$$R_{d,k}(r) = \begin{cases} r^{2k-d}, & \text{ak } d \text{ je nepárne} \\ r^{2k-d}\log(r), & \text{ak } d \text{ je párne} \end{cases} \qquad \text{pre } 2k > d.$$

V prípade tenkostenných splajnov je $d = 2$ a $k = 2$, z čoho dostaneme $R_{2,2}(r) = r^2\log(r)$. Tenkostenný splajn má potom vyjadrenie:

$$f(x,y) = c_1 + c_2 x + c_3 y + \frac{1}{2}\sum_{i=1}^{n} \lambda_i r_i^2 \log\left(r_i^2\right), \text{ pre } [x,y] \in \mathbb{E}^2, \tag{1}$$

pričom $r_i^2 = (x - x_i)^2 + (y - y_i)^2$ a $c_1$, $c_2$, $c_3$, $\lambda_i$ sú neznáme. Parametre $\lambda_i$, $i = 1, \ldots, n$ musia spĺňať podmienky:

$$\sum_{i=1}^{n} \lambda_i = 0 \quad \text{a} \quad \sum_{i=1}^{n} \lambda_i \boldsymbol{p}_i = \boldsymbol{0}\,. \tag{2}$$

Aplikovaním interpolačných podmienok $f(\boldsymbol{p}_i) = v_i$, kde $i = 1, 2, \ldots, n$ spolu s podmienkami (3) vypočítame neznáme zo systému rovníc:

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 0 & x_1 & x_2 & \cdots & x_n \\ 0 & 0 & 0 & y_1 & y_2 & \cdots & y_n \\ 1 & x_1 & y_1 & 0 & r_{21}^2\log\left(r_{21}^2\right) & \cdots & r_{n1}^2\log\left(r_{n1}^2\right) \\ 1 & x_2 & y_2 & r_{12}^2\log\left(r_{12}^2\right) & 0 & \cdots & r_{n2}^2\log\left(r_{n2}^2\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & y_n & r_{1n}^2\log\left(r_{1n}^2\right) & r_{2n}^2\log\left(r_{2n}^2\right) & \cdots & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \lambda_1/2 \\ \lambda_2/2 \\ \vdots \\ \lambda_n/2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}, \tag{3}$$

kde $r_{ij}^2 = r_{ji}^2 = \left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2$ .

## 2.2. Shepardova metóda interpolácie

Shepardova metóda patrí k najjednoduchším prístupom k interpolácii nerovnomerne rozložených dát [2].

Analogicky ako v predchádzajúcej časti budeme hľadať interpolačnú funkciu vyhovujúcu podmienkam $f(\boldsymbol{p}_i) = v_i$, pre $i = 1, 2, \ldots, n$.

Shepard navrhol interpolačnú funkciu v tvare váženého priemeru hodnôt $v_i$.

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i(\boldsymbol{x}) v_i \ . \tag{5}$$

Váhovaciu funkciu $\omega_i(\boldsymbol{x})$ zo vzťahu (5) môžeme vyjadriť v tvare:

$$\omega_i(\boldsymbol{x}) = \frac{\sigma_i(\boldsymbol{x})}{\sum_{j=1}^{n} \sigma_j(\boldsymbol{x})}, \tag{6}$$

kde $\sigma_i(\boldsymbol{x}) = \left\|\boldsymbol{x} - \boldsymbol{p}_i\right\|^{-\mu_i}$, pre $\mu_i > 0$. Parameter $\mu_i$ umožňuje ovplyvňovať tvar výslednej plochy v okolí interpolovaných bodov.

V literatúre (napr. v [3]) sa často vyskytuje zovšeobecnená Shepardova metóda používajúca lokálne interpolanty, ktoré nahrádzajú hodnoty $v_i$ lokálnymi interpolačnými funkciami $L_i(\boldsymbol{x})$ s vlastnosťou $L_i(\boldsymbol{p}_i) = v_i$. Potom funkcia $f(\boldsymbol{x})$ zo vzťahu (5) bude mať tvar:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i(\boldsymbol{x}) L_i(\boldsymbol{x}). \tag{7}$$

Ak interpolačné funkcie budú kvadratické, tak dostaneme dostatočne hladkú plochu s relatívne malou výpočtovou náročnosťou.

Modifikovaná kvadratická Shepardova metóda má tvar (pozri [4]):

$$f(\boldsymbol{x}) = f(x, y) = \sum_{i=1}^{n} \omega_i(x, y) Q_i(x, y), \tag{8}$$

pričom lokálny kvadratický interpolant $Q_i(x, y)$ je určený predpisom:

$$\begin{aligned} Q_i(x, y) = {} & c_{i,1}(x - x_i)^2 + c_{i,2}(x - x_i)(y - y_i) + \\ & c_{i,3}(y - y_i)^2 + c_{i,4}(x - x_i) + c_{i,5}(y - y_i) + v_i \ . \end{aligned} \tag{9}$$

Neznáme koeficienty $c_{i,j}$ v $Q_i(x, y)$ sú vypočítané metódou najmenších štvorcov s použitím podmienok:

$$\sum_{k=1,\, k\neq i}^{n} \omega_k(x_i, y_i)\left[c_{i,1}(x_k - x_i)^2 + \cdots + c_{i,5}(y_k - y_i) + f_i - f_k\right]^2 \to min, \tag{10}$$

kde $\omega_k(x,y) = \left( \frac{R_q - d_k(x,y)}{R_q d_k(x,y)} \right)_+^2$ a $R_q$ je polomer vplyvu okolia bodu $\boldsymbol{p}_i[x_i, y_i]$.

## 3. Praktická časť

Vstupné údaje tvorili maximálne intenzity zrážok z 56 zrážkomerných staníc z celého územia Slovenska, spracované postupom podľa Šamaja a Valoviča ([8]), pre trvania 5 až 180 min, doplnené o denné údaje. V týchto merných staniciach sme použili metódu jednoduchého škálovania na určenie návrhových dažďových intenzít na Slovensku (pozri [5]). Metóda jednoduchého škálovania umožňuje určiť návrhové hodnoty zrážok pre trvania dažďov kratších než jeden deň a pre zvolenú dobu opakovania využíva denné záznamy o úhrnoch zrážok. Metóda bola aplikovaná vo viacerých oblastiach Európy, ako i zámoria, a osvedčila sa ako vhodná na vyjadrenie vzťahov medzi intenzitou, trvaním a periodicitou zrážok ([6], [9], [7]).

Našou úlohou bolo nameranými hodnotami preložiť dostatočne vhodnú interpolačnú splajnovú plochu s využitím rôznych metód a overiť ich presnosť pomocou známych štatistických mier. Overovanie modelu sme uskutočnili pomocou známej metódy, v ktorej sa postupne vylúči vždy jedno meranie (z nameraných hodnôt) a následne sa spočíta chyba interpolácie pre vylúčené meranie. Takto sme postupovali u všetkých interpolačný metód. Grafické výstupy interpolačných plôch získaných jednotlivými metódami uvádzame na obrázku 2, 3 a 4.

Pre jednotlivé metódy sme analyzovali ich chyby. V nasledujúcej tabuľke uvádzame prehľad popisných štatistík podľa jednotlivých metód:

***Tabuľka 1: Porovnanie interpolačných splajnových modelov***

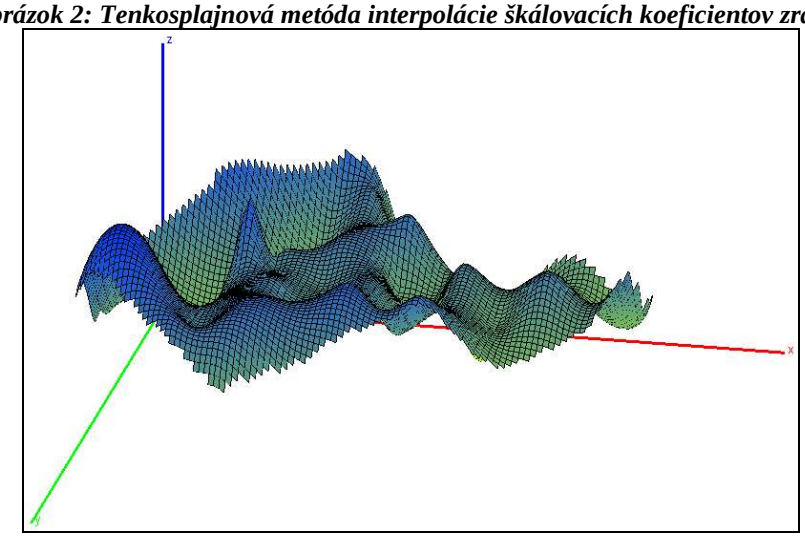
| | ***Tenkosplajnová metóda*** | ***Shepardova metóda*** | ***Kvadratická Shepardova metóda*** |
|---|---|---|---|
| **Priemer** | 0,0007 | -0,0004 | -0,0006 |
| **Štandardná chyba** | 0,0047 | 0,0038 | 0,0071 |
| **Medián** | -0,0009 | 0,0017 | 0,0020 |
| **Smerodajná odchýlka** | 0,0354 | 0,287 | 0,0531 |
| **Rozptyl výberu** | 0,0013 | 0,0008 | 0,0028 |
| **Strmosť** | -0,1010 | -0,2412 | 3,9433 |
| **Šikmosť** | -0,1816 | -0,0560 | -0,6079 |
| **Rozsah** | 0,1619 | 0,1240 | 0,3423 |
| **Minimum** | -0,0834 | -0,0724 | -0,2000 |
| **Maximum** | 0,0785 | 0,0516 | 0,1423 |

Pre chyby jednotlivých modelov sme otestovali hypotézu: $H_0$: priemerná chyba sa rovná nule, oproti alternatíve, že priemerná chyba nie je rovná nule. Pre všetky metódy Studentov *t*-test nezamietol nulovú hypotézu na ľubovoľnej hladine významnosti α. Ďalej sme overili normalitu chýb. Pomocou *Kolmogorovho-Smirnovho* testu, sme zistili, že tenkosplajnová a Shepardova metóda majú normálne rozdelené chyby pre $\alpha = 0{,}05$ ale kvadratická Shepardova metóda má normálne rozdelené chyby len pre $\alpha = 0{,}01$. Výsledky týchto hypotéz uvádzame v nasledujúcej tabuľke:

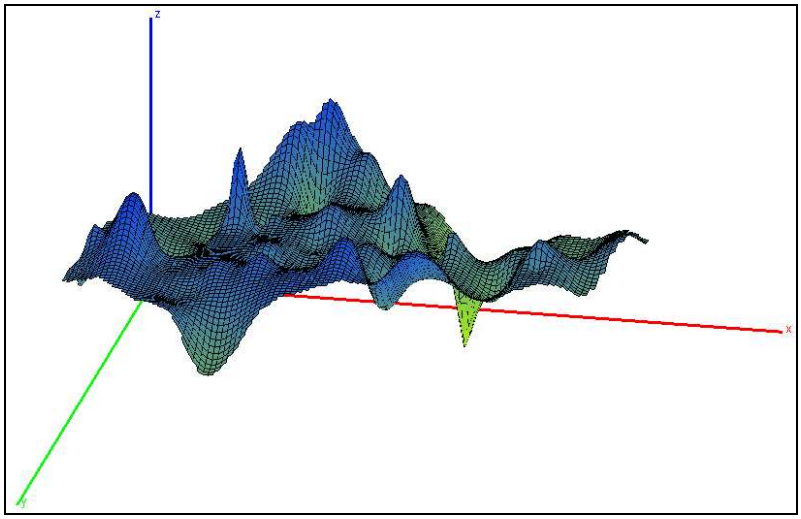
***Tabuľka 2: Výsledky testu: priemerná chyba je nulová a testu normality***

| ***Tenkosplajnová metóda*** | | ***Shepardova metóda*** | | ***Kvadratická Shepardova metóda*** | |
|---|---|---|---|---|---|
| Student *t*-test (*p*-value) | test normality | Student *t*-test (*p*-value) | test normality | Student *t*-test (*p*-value) | test normality |
| 0,8866 | 0,1500 | 0,9128 | 0,1500 | 0,9280 | 0,0306 |

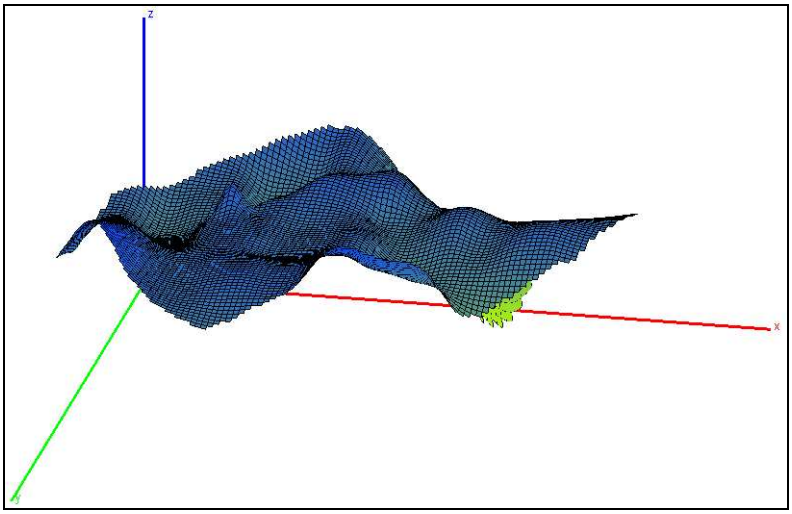
*Obrázok 2: Tenkosplajnová metóda interpolácie škálovacích koeficientov zrážok*



*Obrázok 3: Shepardova metóda interpolácie škálovacích koeficientov zrážok*



*Obrázok 4: Kvadratická Shepardova metóda interpolácie škálovacích koeficientov zrážok*



## 4. Záver

Pre modelovanie škálovacích exponentov zrážok sme okrem vyššie uvedených troch interpolačných metód použili aj interpolačné metódy založené na triangulániach (*Cloughovu-Tocherovu*, *Powellovu-Sabinovu* metódu a iné), avšak ich výsledky boli menej presné a preto ich v článku neuvádzame. Výsledky získané tenkosplajnovou metódou, Shepardovou metódou a kvadratickou Shepardovou metódou budú ďalej podrobené analýzam pre vodohospodárske účely, prípadne budeme ďalej uvažovať o využití interpolačných metód založených na radiálnych bázických funkciách s lokálnym nosičom ([10]).

## 5. Literatúra


[1] DUCHON, J. 1977. Lecture Notes in Mathematics 571. Springer-Verlag, Berlin, 1977, s. 85–100.

[2] SHEPARD, D. 1968. A two dimensional interpolation function for irregular spaced data. Proceedings 23rd ACM. National Conference, 1968, s. 517–524.

[3] FRANKE, R., NIELSON, G. 1980. Smooth interpolation of large sets of scattered data. In International Journal for Numerical Methods in Engineering, vol. 15, no. 11, 1980, s. 1691–1704.

[4] RENKA, R. 1988. Multivariate interpolation of large sets of scattered data. In: ACM Transactions on Mathematical Software, vol. 14, no. 2, 1988, s. 139–148.

[5] BARA, M., GAÁL, L., KOHNOVÁ, S., SZOLGAY, J.,HLAVČOVÁ, K. 2008. Simple scaling of extreme rainfall in Slovakia: a case study. In: Meteorological Journal. ISSN 1335–339X. 2008, 11, č.4, s. 153–157.

[6] MENABDE, M. – SEED, A. – PEGRAM, G. 1999. A simple scaling model for extreme rainfall. Water Resour. Res., 35 (1), 1999, s. 335–339

[7] MOLNAR,P. - BURLANDO, P. 2005. Preservation of rainfall properties in stochastic disaggregation by a simple random cascade model. Atmospheric Research, 77, 2005, s. 137–151

[8] ŠAMAJ, F. – VALOVIČ, Š. 1973, Intensities of short-term rainfall in Slovakia. Proceedings of works of HMI, Nr. 5. SPN Bratislava. (in Slovak)

[9] YU, P.SH. – YANG, T.CH. – LIN, CH.SH. 2004, Regional rainfall intensity formulas based on scaling property of rainfall. In: Journal of Hydrology, 295 (1-4), 2004 s. 108–123

[10] FORNEFETT, M. – ROHR,.K. – STIEHL, H. 1999 Elastic Registration of Medical Images Using Radial Basis Functions with Compact Support. Computer Vision and Pattern Recognition, 1999, s. 402–407.


## Adresy autorov:


Bohdal Róbert, RNDr., PhD.
FMFI UK, Mlynská dolina
842 48 Bratislava
Robert.bohdal@fmph.uniba.sk

Bohdalová Mária, RNDr., PhD.
FM UK, Odbojárov 10
825 05 Bratislava
maria.bohdalova@fm.uniba.sk

# Financial time series and chaos
# Finančné časové rady a chaos

Bohdalová Mária, Greguš Michal

**Abstract:** Over the last few years, it has become clear that chaos theory and fractals are a subset of a much larger universe of discourse: complexity theory. In this paper we introduce the fractal market analysis. Fractal structure accepts global determinism and local randomness of the behavior of the financial time series. We will use R/S analysis in this paper. R/S analysis can distinguish fractals from other types of time series, revealing the self-similar statistical structure.

**Abstrakt:** Teória chaosu sa zaoberá nelineárnymi systémami, ktoré majú skrytý nejaký vnútorný poriadok aj keď sa navonok zdá, že sa jedná o náhodné procesy. Článok je úvodom do fraktálnej analýzy finančných trhov. Fraktálna štruktúra akceptuje globálny determinizmus a lokálnu náhodnosť správania sa finančných časových radov. V článku sa zaoberáme R/S analýzou, ktorá umožňuje odlíšiť fraktál od iných časových radov.

**Key words:** financial time series, fractal, R/S analysis

**Kľúčové slová:** finančné časové rady, fraktál, R/S analýza

## 1. Introduction to Fractals and the Fractal dimensions

The development of fractal geometry has been one of the 20-th century's most useful and fascinating discoveries in mathematics ([2], p.45). Fractals give structure to complexity, and beauty to chaos. Most natural shapes, and time series, are best described by fractals. Fractals are self-referential, or self-similar. Fractal shapes show self-similarity with respect to space. Fractal time series are random fractals, which have more in common with natural objects than the pure mathematical fractals we will cover initially. We will be concerned primarily with fractal time series, but fractal shapes give a good intuitive base for what "self-similarity" actually means. Figure 1 shows daily, weekly and monthly Bank of America Corporation prices[1] for consecutive observations from march 2007 to may 2009. With no scale on the *X* and *Y* axes, we are not able to determine which graph is which. Figure 1 illustrates self-similarity in a time series.

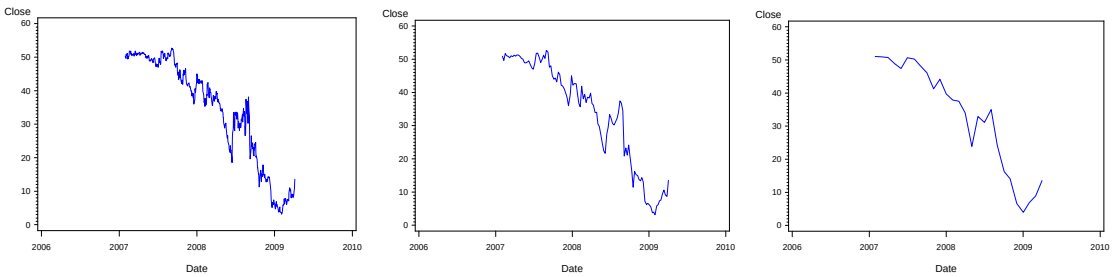



**Figure 1: Daily, weekly and monthly prices of stock Bank of America Corp.**

Fractal shapes can be generated in many ways. The simplest way is to take a generating rule and iterate it over and over again. Random fractals are combination of generating rules chosen at random for different scales. Combination of randomness coupled with deterministic generation rules, or "causality", can make fractals useful in capital market analysis. Random

[1] Datas follow from www.yahoo.com

fractals ([2], p.51) do not necessarily have pieces that look like pieces of the whole. Instead, they may be qualitatively related. In the case of time series, we will find that fractal time series are qualitatively self similar in that, at different scales, the series have similar statistical characteristics. If we would like to understand the underlying causality of the structure of time series, then classical geometry offers little help. May be, time series is a random walk – a system so complex that the prediction becomes impossible. In statistical term, the number of degrees of freedom or factors influencing the system is very large. These systems are not well-described by standard Gaussian statistics. Standard statistical analysis begins by assuming that the system under study is primarily random; that is, the causal process that created the time series has many component parts, or degree of freedom, and the interaction of those components is so complex that deterministic explanation is not possible ([1], p.53). Only probabilities can help us to understand and take advantage of the process. The underlying philosophy implies that randomness and determinism cannot coexist. In order to study the statistics of these systems and create a more general analytical framework, we need a probability theory that is nonparametric. In this paper we introduce nonparametric methodology that was discovered by H.E. Hurst[2].

In advance, we introduce the term: fractal dimension. The fractal dimension describes how a time series fills its space, is the product of all factors influencing the system that produces time series ([2], p.57). Fractal time series can have fractional dimensions. The fractal dimension of a time series measures how jagged the time series is ([1], p.16). As would be expected, a straight line has a fractal dimension of 1. Time series is only random when it is influenced by a large number of events that are equally likely to occur. In statistical term, it has a high number of degree of freedom. A random series would have no correlation with previous points. Nothing would keep the points in the same vicinity, to preserve their dimensionality. Instead, they will fill up whatever space they are placed in. A nonrandom time series will reflect the nonrandom nature of its influences. The data will clump together, to reflect the correlations inherent in its influences. In other words, the time series will be fractal. To determine the fractal dimension, we must measure how the object clumps together in its space. However, a random walk has 50-50 chance of rising or falling, hence, its fractal dimension is 1.50. The fractal dimension of a time series is important because it recognizes that process can be somewhere between deterministic (a line with fractal dimension of 1) and random (a fractal dimension of 1.50). In fact, the fractal dimension of a line can range from 1 to 2. The normal distribution has an integer dimension of 2, which many of characteristics of the time series. At values $1.50<d<2$, a time series is more jagged than a random series.

They are many ways of calculating fractal dimensions. We introduce methodology of the Hurst exponent *H*, and we convert it into the fractal dimension *d* in this paper.

## 2. R/S analysis and Hurst exponent

Hurst was aware of Einstein`s[3] work of Brownian motion. Brownian motion became the primary model for a random walk process. Einstein found that the distance that a random particle covers increases with the square root of time used to measure it, or:

$$R=T^{0.50}, \quad (1)$$

where $R$ is the distance covered and $T$ is a time index.

Equation (1) is called the $T$ to *the one-half rule*, and it is commonly used in statistics. Financial economists use it to annualize volatility or standard deviation. To standardize the measure over time, Hurst decided to create a dimensionless ratio by dividing the range by the standard deviation of the observations. Hence, the analysis is called rescaled range analysis

[2] Hurst, H.E. 1951. The Long-Term Storage Capacity of Reservoirs. In: Transaction of the American Society of Civil Engineers 116.

[3] Einstein, A. 1908. Uber die von der molekularkinetischen Theorie der Warme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annals of Physics 322.

(R/S analysis). Hurst found that most natural phenomena follow a "biased random walk" – a trend with noise. The strength of the trend and the level of noise could be measured by how the rescaled range scales with time, that is, by how high $H$ is above 0.50. Peters ([1] p.56) reformulated Hurst`s work for a general time series as follows.

We begin with a time series, $X=\{x_1, x_2,\ldots, x_n\}$, to represent $n$ consecutive values. For markets, it can be the daily changes in price of a stock index. We divide this period into $m$ contiguous subperiods of length $q$, such that $mq=n$. The rescaled range was calculated by first rescaling or "normalizing" the data by subtracting the sample mean $x_m$:

$$Z_r=(x_r-x_m), \qquad r=1,2,\ldots,n \tag{2}$$

The resulting series, $Z$, now has a mean of zero. The next step creates a cumulative time series Y:

$$Y_1=(Z_1+Z_r), \qquad r=2,3,\ldots,n \tag{3}$$

Note that, by definition, the last value of $Y$ ($Y_n$) will always be zero because $Z$ has a mean of zero. The adjusted range, $R_n$, is the maximum minus minimum value of the $Y_r$:

$$R_n=max(Y_1, Y_2,\ldots, Y_n)-min(Y_1, Y_2,\ldots, Y_n). \tag{4}$$

The subscript, $n$, for $R_n$ now signifies that this is the adjusted range for $x_1, x_2, \ldots, x_n$. Because $Y$ has been adjusted to a mean of zero, the maximum value of $Y$ will always be greater than or equal to zero, and the minimum will always be less than or equal to zero. Hence, the adjusted range, $R_n$, will always be nonnegative. This adjusted range, $R_n$, is the distance that the system travels for time index $n$. If we set $n=T$, we can apply equation (1), provided that the time series, $X$, is independent for increasing values of $n$. However, equation (1) applies only to time series that are in Brownian motion (they have zero mean, and variance equal to one). To apply this concept to time series that are not in Brownian motion, we need to generalized equation (1) and take into account systems that are not independent. Hurst found that the following was a more general form of equation (1):

$$(R/S)_n=c.n^H \tag{5}$$

The subscript, $n$, for $(R/S)_n$ refers to the R/S value for $x_1, x_2, \ldots, x_n$ and $c$ is a constant.

The R/S value of equation (5) is referred to as the *rescaled range* because it has zero mean and is expressed in terms of local standard deviation. In general, the R/S value scaled as we increase the time increment, $n$, by a power–law value equal to $H$, generally called the *Hurst exponent* ($n$ is an integer value).

Rescaling allows us to compare periods of time that may be many apart. In comparing stock returns of the 1920s with those of the 1980, prices present a problem because of inflationary growth. Rescaling minimize this problem, by rescaling the data to zero mean and standard deviation of one, to allow diverse phenomena and time periods to be compared. Rescaled range analysis can also describe time series that have no characteristic scale. This is a characteristic of fractals.

The Hurst exponent can be approximated by plotting the $\log(R/S_n)$ versus the $\log(n)$ and solving for the slope through an ordinary least squares regression:

$$log(R/S_n)=log(c)+H.log(n) \tag{6}$$

If a system is independently distributed, then $H$=0.50. When $H$ differed from 0.50, the observations are not independent. Each observation carried a „memory" of all the events that preceded it. What happens today influences the future. Where we are now is a result of where we have been in the past. Time is important. The impact of the present on the future can be expressed as a correlation:

$$C=2^{(2H-1)}-1, \tag{7}$$

where $C$ is correlation measure and $H$ is Hurst exponent.

There are three distinct classifications for the Hurst exponent ([2], p.64):

1. $H$=0.50: time series is random, events are random and uncorrelated. Equation (7) equals zero. The present does not influence the future. Its probability density function can be normal curve, but it does not have to be. R/S analysis can classify an independent series, no mater what the shape of the underlying distribution.

2. $0 \le H < 0.50$: time series is antipersistent, or ergodic. If the time series has been up in the previous period, it is more likely to be down in the next period. Conversely, if it was down before, it is more likely to be up in the next period. The strength of this antipersistent behavior depends on how close $H$ is to zero. The closer it is to zero, the closer $C$ in equation (7) moves toward –0.50, or negative correlation. This time series is more volatile than a random series.
3. $0.50 \le H < 1.00$: time series have a persistent or trend-reinforcing character. If the series has been up (down) in the last period, then the chances are that it will continue to be positive (negative) in the next period. Trend is apparent. The strength of the trend-reinforcing behavior, or persistence, increases as $H$ approaches 1.0. The closes $H$ is to 0.5, the noisier it will be, and the less defined its trends will be. Persistent series are fractional Brownian motion, or biased random walk[4]. The strength of the bias depends on how far $H$ is above 0.50.

## 3. Testing R/S analysis

To evaluate the significance of R/S analysis, we calculate expected value of the R/S statistics and the Hurst exponent. We compare the behavior of our process, described by R/S analysis with an independent and random system and gauge its significance.

We will test this null hypothesis: "The process is independent, identically distributed and is characterized by a random walk"[5].

To verify this hypothesis, we calculate expected value of the adjusted range[6] $E(R/S_n)$ and its variance[7] $Var(E(R/S_n))$.

$$E(R/S_n) = \frac{n-0.5}{n} \cdot \left( n \cdot \frac{\pi}{2} \right)^{-0.5} \sum_{r=1}^{n-1} \sqrt{\frac{(n-r)}{r}} \tag{8}$$

$$Var(E(R/S_n)) = \left( \frac{\pi^2}{6} - \frac{\pi}{2} \right) \cdot n \,. \tag{9}$$

Using the results of equation (8) we can generate expected values of the Hurst exponent. The expected Hurst exponent will vary depending on the values of $n$ we use to run the regression. Any range will be appropriate as long as the system under study and the $E(R/S_n)$ series cover to the same values of $n$. For financial purpose, we will begin with $n$=10. The final value of $n$ will depend on the system under study.

R/S values are random variables, normally distributed and therefore we would expect that the values of $H$ would also normally distributed (see Peters [1], p.72):

$$Var(E(H_n)) = \frac{1}{T}, \tag{10}$$

where T is total number of observations in the sample. Note that the $Var(H_n)$ does not depend on $n$ or $H$, but depends on the total sample size $T$.

Now $t$-statistics will be used to verify of the significance of the null hypothesis.

## 4. Empirical study

We apply R/S analysis to the daily, weekly and monthly closing stock prices Bank of America from 29.05.1986 to 7.5.2009 and the data follow from www.yahoo.finance.com (Figure 2). R/S analysis needs a long time intervals. We use 5775 observations for daily

[4] Biased random walks were extensively studied by Hurst in the 1940s and again by Mandelbrot in the 1960s and 1970s. Mandelbrot called them fractional brownian motions.([2], p.61)

[5] This process has Gaussian structure (see [1], p.66).

[6] This formula was derrived by Anis and Lloyd (1976), ([1], p.71)

[7] Variance was calculated by Feller (1951) ([1], p.66)

frequency (only trading days), 2880 observations for weekly frequency and 276 observations for monthly frequency.

When analyzing markets, we use logarithmic returns, defined as follows:

$$S_t = \ln \frac{P_t}{P_{t-1}}, \tag{8}$$

where $S_t$ is logarithmic return at time $t$ and $P_t$ is stock price at time $t$.

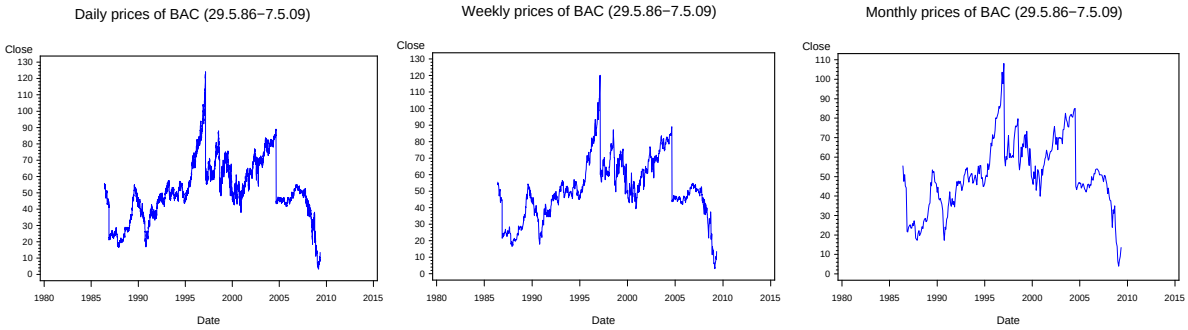



**Figure 2: Daily, weekly and monthly prices of Bank of America Corp. from 29.5.86 to 7.5.09**

For R/S analysis, logarithmic returns are more appropriate than the more commonly used percentage change in prices. The range used in R/S analysis is the cumulative deviation form the average, and logarithmic returns sum to cumulative return, while percentage changes do not. ([2], p.83).

We will examine the behavior of $H$ over different time increments, from daily to monthly returns of stock Bank of America Corp. (BAC).

Table 1–Table 3 show both the $R/S_n$ and $E(R/S_n)$ values. Figure 3 shows the *log R/S* plot for daily return data for $T$=5775 observations. Also plotted is $E(R/S_n)$ (calculated using equation (8)) as a comparison against the null hypothesis that the system is an independent process. There is clearly a systematic deviation from the expected values. Figure 4 shows the *log R/S* plot with $E(R/S)$ plot for weekly return data for $T$=2880 observations and Figure 5 shows the *log R/S* plot with $E(R/S)$ plot for monthly return data for $T$=276 observations.

The regression yielded $H$=0.53540 and $E(H)$=0.56213 for daily returns (see Table 4). The variance of $E(H)$, as shown in equation (10) is 0.0002, for Gaussian random variables. The standard deviation of $E(H)$ is 0.0132. The $H$ value for daily returns is –2.0313 standard deviation bellow its expected value, a highly significant result. The regression yielded $H$=0.53520 and $E(H)$=0.56952 for weekly returns (see Table 5). The variance of $E(H)$ is 0.0003 and standard deviation of $E(H)$ is 0.0132. The $H$ value for daily returns is –1.8418 standard deviation bellow its expected value, a non significant result for confidence level $\alpha$=0.05. The regression yielded $H$=0.59589 and $E(H)$=0.6097 for monthly returns (see Table 6). The variance of $E(H)$ is 0.0036 and standard deviation of $E(H)$ is 0.0602. The $H$ value for daily returns is –0.2301 standard deviation bellow its expected value, a non significant result for confidence level $\alpha$=0.05. Significance is confirmed only for daily returns and this may be caused by insufficient number observations.

Our stock Bank of America has $H$ greater than 0.5, it has persistence character, it is fractal and application of standard statistical analysis becomes of questionable value.

| N | log N | log R/S | R/S | E(R/S) |
|---|---|---|---|---|
| 10 | 2,302585 | 1,122582 | 3,072779 | 2,650278 |
| 17 | 2,833213 | 1,463863 | 4,322625 | 3,879877 |
| 20 | 2,995732 | 1,546395 | 4,694515 | 4,324742 |
| 34 | 3,526361 | 1,852481 | 6,37562 | 6,050077 |
| 68 | 4,219508 | 2,231924 | 9,317779 | 9,101265 |
| 85 | 4,442651 | 2,325834 | 10,23522 | 10,32771 |
| 170 | 5,135798 | 2,701515 | 14,90229 | 15,13091 |
| 289 | 5,666427 | 2,978472 | 19,65775 | 20,10602 |
| 340 | 5,828946 | 3,057879 | 21,28237 | 21,94454 |
| 578 | 6,359574 | 3,347838 | 28,44118 | 28,96639 |
| 1156 | 7,052721 | 3,68908 | 40,00803 | 41,44743 |
| 1445 | 7,275865 | 3,796379 | 44,5396 | 46,47719 |
| 2890 | 7,969012 | 4,307041 | 74,22053 | 66,21137 |

**Table 1: R/S analysis, Bank of America: daily returns**

| N | log N | log R/S | R/S | E(R/S) |
|---|---|---|---|---|
| 11 | 2,397895 | 1,166956 | 3,212199 | 2,848343 |
| 12 | 2,484907 | 1,186805 | 3,276594 | 3,037391 |
| 18 | 2,890372 | 1,433343 | 4,192693 | 4,032329 |
| 22 | 3,091042 | 1,532168 | 4,628198 | 4,602551 |
| 27 | 3,295837 | 1,63842 | 5,147031 | 5,245172 |
| 33 | 3,496508 | 1,73145 | 5,648838 | 5,940635 |
| 36 | 3,583519 | 1,818525 | 6,162762 | 6,264156 |
| 44 | 3,78419 | 1,890857 | 6,625041 | 7,065256 |
| 54 | 3,988984 | 2,07206 | 7,941168 | 7,968737 |
| 66 | 4,189655 | 2,188393 | 8,920869 | 8,947224 |
| 99 | 4,59512 | 2,373262 | 10,73235 | 11,2472 |
| 108 | 4,682131 | 2,371679 | 10,71537 | 11,80396 |
| 132 | 4,882802 | 2,551656 | 12,82832 | 13,18354 |
| 198 | 5,288267 | 2,868894 | 17,61753 | 16,4285 |
| 297 | 5,693732 | 2,975899 | 19,60723 | 20,39936 |
| 396 | 5,981414 | 3,195528 | 24,42308 | 23,77524 |
| 594 | 6,386879 | 3,502434 | 33,19614 | 29,3806 |

**Table 2: R/S analysis, Bank of America: weekly returns**

| N | log N | log R/S | R/S | E(R/S) |
|---|---|---|---|---|
| 12 | 2,484907 | 1,245205 | 3,473647 | 3,037391 |
| 23 | 3,135494 | 1,610174 | 5,003683 | 4,736613 |
| 46 | 3,828641 | 2,072317 | 7,943204 | 7,253682 |
| 69 | 4,234107 | 2,208782 | 9,104621 | 9,177427 |
| 92 | 4,521789 | 2,399576 | 11,01851 | 10,79628 |
| 138 | 4,927254 | 2,750749 | 15,65435 | 13,5082 |

**Table 3: R/S analysis, Bank of America: monthly returns**

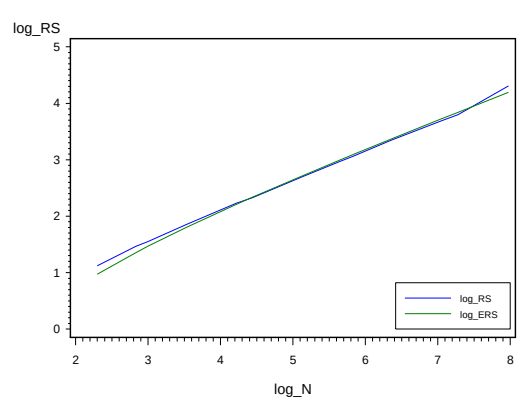



**Figure 3: R/S analysis, Bank of America, daily returns**

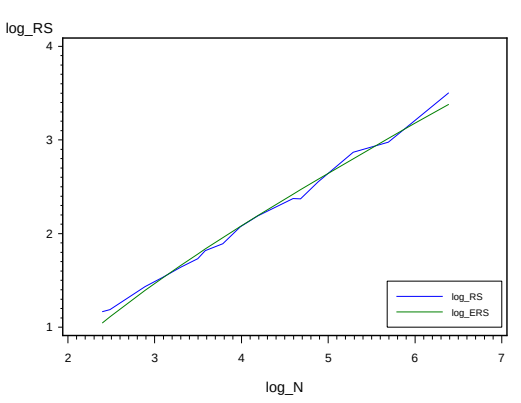



**Figure 4: R/S analysis, Bank of America, weekly returns**

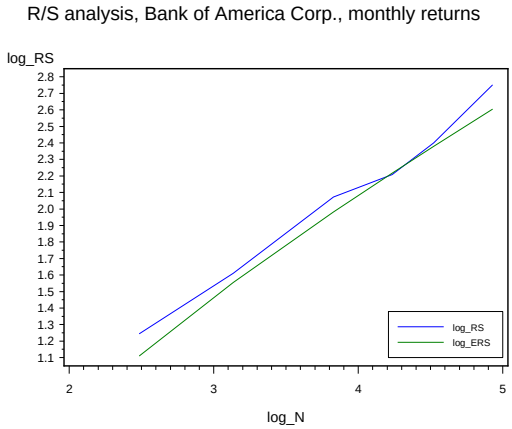



**Figure 5: R/S analysis, Bank of America: monthly returns**

| Parameter Estimates for daily returns of BAC | | | | | |
|---|---|---|---|---|---|
| | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | -0,05695 | 0,01963 | -2,90 | 0,0099 |
| **Hurst exponent *H*** | 1 | 0,53540 | 0,00393 | 136,16 | <.0001 |
| **R-Square for *H*** | 0,9991 | | | | |
| **Adj R-Sq for *H*** | 0,9990 | | | | |
| | | | | | |
| **Expected Intercept** | 1 | -0,20089 | 0,03168 | -6,34 | <.0001 |
| **expected Hurst exponent *E(H)*** | 1 | 0,56213 | 0,00635 | 88,58 | <.0001 |
| **R-Square for *E(H)*** | 0,9978 | | | | |
| **Adj R-Sq for *E(H)*** | 0,9977 | | | | |
| **Number of Observations** | 19 | | | | |
| **Var(*E*(*H*))** | 0,0002 | | | | |
| s(***E***(H)) | 0,0132 | | | | |
| **significance** | -2,0313 | | | | |

**Table 4: Hurst exponent for R/S analysis, Bank of America: daily returns**

| Parameter Estimates for weekly returns of BAC | | | | | |
|---|---|---|---|---|---|
| | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| **Intercept** | 1 | -0,02633 | 0,03035 | -0,87 | 0,3922 |
| **Hurst exponent *H*** | 1 | 0,53520 | 0,00647 | 82,72 | <.0001 |
| **R-Square for *H*** | 0,9955 | | | | |
| **Adj R-Sq for *H*** | 0,9953 | | | | |
| | | | | | |
| **Expected Intercept** | 1 | -0,23026 | 0,02350 | -9,80 | <.0001 |
| **expected Hurst exponent *E(H)*** | 1 | 0,56952 | 0,00501 | 113,66 | <.0001 |
| **R-Square for *E(H)*** | 0,9976 | | | | |
| **Adj R-Sq for *E(H)*** | 0,9975 | | | | |
| **Number of Observations** | 33 | | | | |
| **Var(*E*(*H*))** | 0,0003 | | | | |
| **s(*E*(H))** | 0,0186 | | | | |
| **significance** | -1,8418 | | | | |

**Table 5: : Hurst exponent for R/S analysis, Bank of America: weekly returns**

| Parameter Estimates for monthly returns of BAC | | | | | |
|---|---|---|---|---|---|
| | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| **Intercept** | 1 | -0,2496 | 0,10745 | -2,32 | 0,0809 |
| **Hurst exponent *H*** | 1 | 0,59589 | 0,02725 | 21,87 | <.0001 |
| **R-Square for *H*** | 0,9917 | | | | |
| **Adj R-Sq for *H*** | 0,9896 | | | | |
| | | | | | |
| **Expected Intercept** | 1,0000 | -0,3763 | 0,0481 | -7,8300 | 0,0014 |
| **expected Hurst exponent *E(H)*** | 1,0000 | 0,6097 | 0,0122 | 50,0400 | <.0001 |
| **R-Square for *E(H)*** | 0,9984 | | | | |
| **Adj R-Sq for *E(H)*** | 0,9980 | | | | |
| **Number of Observations** | 6 | | | | |
| **Var(*E*(*H*))** | 0,0036 | | | | |
| **s(*E*(H))** | 0,0602 | | | | |
| **significance** | -0,2301 | | | | |

**Table 6: : Hurst exponent for R/S analysis, Bank of America: monthly returns**

## 5. Conclusion

In this paper we showed how it is possible to measure the impact of information on the time series by using Hurst exponent *H*([2], p.102). *H*=0.50 implies a random walk. Yesterday`s events do not impact today. Today`s events do not impact tomorrow. The events are uncorrelated. Old news has already been absorbed and discounted by the market. *H* greater than 0.50 implies that today`s events do impact tomorrow. Information received today continues to be discounted by the market after it has been received. This is not simply serial correlation, it is a longer memory function. Information can impact the future for very long periods, and it goes across time scales.

## 6. Acknowledgement

The work on this paper has been supported by Science and Technology Assistance Agency under the contract No. APVV-0375-06, and by the VEGA grant agency, grant numbers 1/0500/09, 1/4024/07 and 1/0373/08.

## 7. Bibliography

[1] PETERS, E.E. 1994. Fractal market analysis. New York: John Wiley & Sons, Inc., 1994. 315 s. ISBN 0-471-58524-6.

[2] PETERS, E.E. 1996. Chaos and order in the capital markets. New York: John Wiley & Sons, Inc., 1996. 274 s. ISBN 0-471-13938-6.

[3] TREŠL, J. 2003. Statistical methods and capital markets. Praha: Oeconomica, 2003. 110 s. ISBN 80-245-0598-3.

## Addresses of authors:

Bohdalová Mária, RNDr., PhD.
Odbojárov 10
820 05 Bratislava
maria.bohdalova@fm.uniba.sk

Greguš Michal, Doc., RNDr., PhD.
Odbojárov 10
820 05 Bratislava
michal.gregus@fm.uniba.sk

# Bayesovská štatistika

M. Grendár
Katedra matematiky FPV UMB, Tajovského 40, 974 01 Banská Bystrica
Inštitút matematiky a informatiky, SAV a UMB, Banská Bystrica
Ústav merania SAV, Bratislava
*marian.grendar@savba.sk*

## 1. Úvod

Bayesiáncom je každý, kto robí štatistické úvahy pomocou Bayesovej vety. Bayesovská štatistika je koncepčne veľmi jednoduchá. Na príklade hádzania mincou sa pokúsime ilustrovať jej základné prvky. Posudzovanie modelu a bayesovské priemerovanie modelov budú predstavené prostredníctvom regresného modelu. Výpočtová stránka a niektoré ďalšie aspekty bayesovskej štatistiky sú spomenuté len okrajovo.

Začnime s nebayesovskou analýzou výsledkov hádzania mincou.

## 2. Nebayesovská štatistika: analýza hodov mincou

Modelujme výsledok hodu mincou pomocou náhodnej premennej $X$ ktorá nadobúda hodnoty z dvojprvkovej množiny $\mathcal{X} = \{0, 1\}$ s pravdepodobnosťou $\mathrm{P}(X; \theta) = \theta^x(1-\theta)^{1-x}$, kde parameter $\theta$ je pravdepodobnosť úspechu $\mathrm{P}(X = 1)$. Rozdelenie $\mathrm{P}(X; \theta)$ je bernoulliovské a náhodný výber $X_1^n = X_1, X_2, \ldots, X_n$ je známy aj ako bernoulliovská schéma. Nech sa v $n$-tici hodov mincou vyskytoval úspech $n_1$ krát. Na základe tejto informácie chceme v rámci nebayesovskej štatistiky urobiť bodový a intervalový odhad hodnoty parametra $\theta$, test hypotézy o $\theta$ a predpoveď.

**2.1 Odhad** V nebayesovskej štatistike existuje veľké množstvo metód na odhadovanie parametrov: metóda najväčšej vierohodnosti, momentová metóda, zovšeobecnená momentová metóda, metóda empirickej vierohodnosti, robustné odhadovanie, atď. Najčastejšie používaná je asi metóda najväčšej vierohodnosti (angl., *Maximum Likelihood*, ML), v ktorej sa za odhad $\hat{\theta}_{\mathrm{ML}}$ parametra $\theta$ berie tá hodnotu parametra, pri ktorej by daná realizácia výberu z danej parametrickej triedy rozdelení mala najväčšiu hodnotu vierohodnosti:

$$\hat{\theta}_{\mathrm{ML}} = \arg\sup_{\theta\in\Theta} L(\theta; x_1^n),$$

kde, v tomto prípade, vierohodnostná funkcia (čiže pravdepodobnosť realizácie výberu, chápaná ako funkcia parametra) $L(\theta; x_1^n) = \binom{n}{n_1}\theta^{n_1}(1-\theta)^{n-n_1}$ a parametrický priestor $\Theta = [0, 1]$. Ľahko sa zistí, že $\hat{\theta}_{\mathrm{ML}} = \frac{n_1}{n}$. V prípade tohto jednoduchého zadania by asi všetky nebayesovské odhadovacie metódy viedli na tento odhad.

*Príklad*: Nech $n = 20$, $n_1 = 5$. Potom $\hat{\theta}_{\mathrm{ML}} = \frac{n_1}{n} = \frac{5}{20} = 0.25$. Teda, na základe 20-tich hodov mincou, v ktorých sa strana identifikovaná s $X = 1$ vyskytovala 5 krát, zoberieme ako odhad skutočnej, nám neznámej hodnoty parametra $\theta$ hodnotu 0.25. Tento bodový odhad by teda naznačoval, že minca je nevyvážená.

Ako prípravu na bayesovské úvahy poznamenajme, že mince zvyknú mať homogénne zloženie, bývajú vyvážené, takže by 0 a 1 mali padať s rovnakou pravdepodobnosťou.

Otázkou je, ako zahrnúť túto mimodátovú informáciu do štatistických úvah, konkrétne do odhadovania parametra $\theta$? Nebayesovská štatistika na to nemá odpoveď. $\diamond$

**2.2 Inferencie** Nebayesovský bodový odhad je v prípade bernoulliovskej schémy veľmi jednoduchý. To isté sa už nedá povedať o inferenciách, teda o konfidenčných intervaloch a testoch. Tvorba konfidenčných intervalov pre $\theta$ stále zamestnáva nebayesovských štatistikov, ako o tom svedčí napr. prehľadový článok [21]. Jeden z najobľúbenejších konfidenčných intervalov pre $\theta$ je waldovský 95%-ný interval: $\hat{\theta}_{\mathrm{ML}} \pm 1.96\sqrt{\hat{\theta}_{\mathrm{ML}}(1-\hat{\theta}_{\mathrm{ML}})/n}$, ktorý je založený na Centrálnej limitnej vete (alebo, v tomto prípade, ekvivalentne, asymptotickej normalite ML odhadu). Jeho pravdepodobnosť pokrytia je u výberov o veľkosti 20 iba 80%-ná[1]. Preto bolo navrhutých viacero korekcií waldovského intervalu. Iný populárny interval je Clopperov Pearsonov interval, zviazaný dualitou s rovnomenným testom.

*Príklad (pokr.)*: Clopperov Pearsonov test dáva pre naše dáta p-hodnotu 0.0414, teda v teste nulovej hypotézy $H_0 : \theta = 0.5$ oproti alternatíve $H_1 : \theta \neq 0.5$, nulovú hypotézu, na konvenčnej 5%-tnej hladine významnosti zamietame. A 95%-ný konfidenčný interval je $(0.086, 0.491)$. Asymptotický 95%-ný konfidenčný interval je $(0.060, 0.440)$[2]. Tieto nebayesovské inferencie vedú k záveru, že minca je nevyvážená. $\diamond$

Konfidenčný interval je jeden z mnohých *keby*-konceptov (angl. *pre-data concept*) nebayesovskej štatistiky. Keby sme realizovali veľký (nekonečný) počet 20-tíc hodov peniazom, tak by daný $100(1-\alpha)$%ný konfidenčný interval pokryl neznámu skutočnú hodnotu parametra v $100(1-\alpha)$%-ách prípadov. Či v prípade konkrétnej realizácie 20-tich hodov mincou daný interval pokryl alebo nepokryl skutočnú hodnotu parametra sa ale nedozvieme: buď pokryl, alebo nepokryl. Nie vždy je toto keby-uvažovanie zmysluplné, nevraviac o tom, že v praxi je obyčajne dôležitejšie *keď*-uvažovanie (angl., *post-data reasoning*): keď máme túto konkrétnu realizáciu, aká je pravdepodobnosť, že skutočná hodnota parametra leží v tomto konkrétnom intervale? Táto otázka má zmysel v bayesovskej štatistike.

Testovanie hypotéz je notoricky problematické, viď napr. [2], [24], [29]. Zo všetkých výhrad proti testovaniu hypotéz pripomeňme tú Lindleyho: 'O nulových hypotézach neexistencie rozdielu sa vie už vopred, že sú neplatné. Potom zamietnutie alebo nezamietnutie takejto hypotézy neodráža nič iné len veľkosť výberu a silu testu, a to je sotva nejakým príspevkom do vedy'.

**2.3 Predikcie** V prípade náhodného výberu je predpovedanie veľmi jednoduché.

*Príklad (pokr.)*: Pravdepodobnosť $\mathrm{P}(X = 1 \mid x_1^n)$, že výsledok budúceho hodu mincou bude $X = 1$ je rovná $\theta$. Aby sme ju odhadli, je prirodzené nahradiť neznáme $\theta$ jeho odhadom $\hat{\theta}_{\mathrm{ML}}$, čo je v tomto prípade 0.25. Teda, podľa nášho odhadu padne v nasledujúcom hode $X = 1$ s pravdepodobnosťou 1/4. $\diamond$

## 3. Bayesovský rámec

Rámec, v ktorom sa deje bayesovská štatistika je daný vetou Bayesa[3].

[1] Ďalšou črtou tohto intervalu je, že ak $\hat{\theta}_{\mathrm{ML}} = 0$, tak interval obsahuje, bez ohladu na $n$, jediný bod: nulu.

[2] Vypočítané v `R`. Kód: `libray(Hmisc); binconf(5, 20, method = 'all')`. `R`-kovský zdrojový kód ku všetkým výpočtom z tohto článku sa dá nájsť na www.savbb.sk/∼grendar.

[3] Zaujímavá diskusia Bayesovej vety, ako aj základná charakterizácia odlišnosti bayesovskej a

**3.1 Bayesova veta** Bayesovský štatistik musí mimodátovú informáciu o parametri $\theta \in \Theta$ dáta-generujúceho rozdelenia pravdepodobnosti $p(X_1^n \,|\, \theta)$ vyjadriť vo forme apriórneho rozdelenia pravdepodobnosti $p(\theta)$. Prior, ako sa apriórne rozdelenie stručnejšie nazýva, vyjadruje štatistikovu neistotu o hodnote parametra $\theta$. Po tom ako je získaná realizácia $x_1^n$ výberu $X_1^n$, modifikuje bayesiánec svoju apriórnu informáciu pomocou Bayesovej vety

$$p(\theta \,|\, x_1^n) = \frac{p(x_1^n \,|\, \theta)p(\theta)}{\int_\Theta p(x_1^n \,|\, \theta, )p(\theta)\, d\theta}, \tag{1}$$

a získa tak aposteriórne rozdelenie $p(\theta \,|\, x_1^n)$ parametra $\theta$, pri daných dátach $x_1^n$. Posteriórne rozdelenie (alebo posterior) vyjadruje štatistikovu neistotu o $\theta$ po tom, ako boli pozorované dáta. Posterior samozrejme závisí aj od modelu – teda od apriórneho rozdelenia $p(\theta)$ a dáta-generujúceho rozdelenia $p(x_1^n \,|\, \theta)$.
Niektoré bayesovské úvahy sa dajú robiť aj bez menovateľa v Bayesovej vete, ktorý sa zvykne nazývať evidencia (angl., *evidence*), alebo aj marginálna vierohodnosť. V takom prípade je zvykom písať Bayesovu vetu v stručnejšom tvare:

$$p(\theta \,|\, x_1^n) \propto p(x_1^n \,|\, \theta)p(\theta),$$

alebo, neformálne, posterior $\propto$ vierohodnosť $\times$ prior.

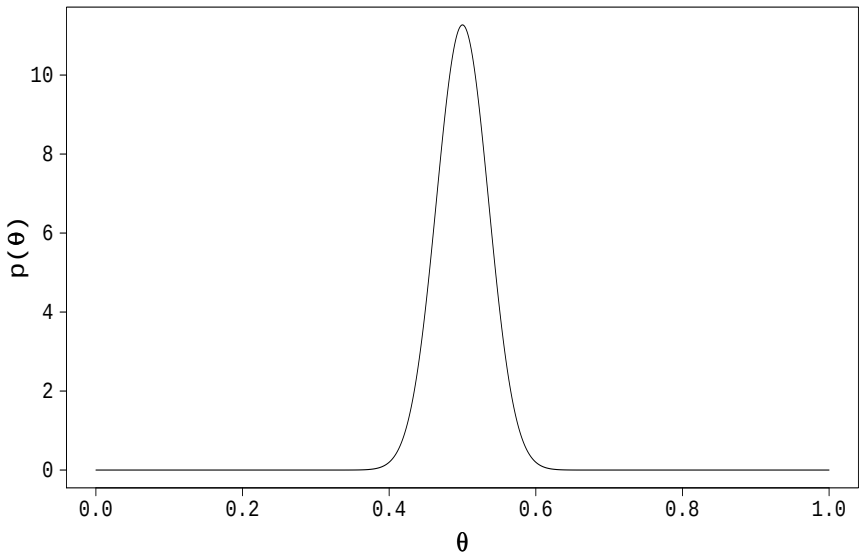
*Príklad (pokr.)*: Neistota o parametre $\theta$ je v našom prípade dosť malá: mince sú väčšinou vyvážené. Ináč povedané, je dosť pravdepodobné, že náhodne vybratá minca, ktorou ideme hádzať, je vyvážená. Mieru neistoty o $\theta$ môžeme kvantifikovať napríklad nasledovne: $\mathrm{P}(0.44 < \theta < 0.56) = 0.9$. Teda, apriórne sme si takmer istí (naša miera presvedčenia je 0.9), že $\theta$ (t.j., pravdepodobnosť padnutia tej strany mince, ktorú sme zviazali s hodnotou náhodnej premennej $X = 1$), je v intervale $(0.44, 0.56)$ – teda, že minca sa nevelmi odlišuje od vyváženej. Ako sme spomínali, bayesiánec musí svoje apriórne presvedčenie vyjadriť pomocou prioru. Jeden možný prior, ktorý súhlasí s týmto apriórnym presvedčením je $\mathrm{Beta}(\alpha = 100, \beta = 100)$.
Pripomeňme:

$$\mathrm{Beta}(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

pričom $\alpha, \beta > 0$, $\theta \in [0, 1]$. Momenty: $\mathrm{E}\theta = \frac{\alpha}{\alpha+\beta}$, $\mathrm{Var}\theta = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. A mód$(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$.
Skombinujeme

- $\mathrm{Beta}(\alpha = 100, \beta = 100)$ prior $p(\theta)$
- s vierohodnosťou $p(n_1 \,|\, \theta) \propto \theta^{n_1}(1-\theta)^{n-n_1}$ dát $x_1^n$ v ktorých sa $X = 1$ vyskytovalo $n_1$ krát,
- pomocou Bayesovej vety,
- a dostaneme posterior $p(x_1^n \,|\, \theta)$. Dá sa uvidieť, že posteriórne rozdelenie je $\mathrm{Beta}(\alpha + n_1 = 105, \beta + n - n_1 = 115)$.

nebayesovskej štatistiky je v práci [30], ktorej znalosť predpokladáme.




Obrázok 1: Beta(100, 100) prior.

Posterior ktorý sme obdržali patrí do tej istej triedy rozdelení ako prior. Prior môže mať akýkoľvek tvar. Prior, ktorý po bayesovskom skombinovaní s vierohodnostnou funkciou vedie na posterior patriaci do tej istej triedy rozdelení čo prior, sa nazýva konjugovaný (angl., *conjugate prior*). Pretože sú bayesovské výpočty s konjugovanými priormi ovela ľahšie, zvyknú sa takéto priory tiež nazývať pohodlnými priormi (angl., *convenience priors*). Mimo pohodlnosti, neexistujú vo všeobecnosti dôvody prečo by mala byť apriórna informácia vyjadrená konjugovaným priorom.
Vierohodnostná funkcia, prior a posterior sú zobrazené[4] na obr. 2. Maximum vierohodnosti sa dosahuje v 0.25 čo je $\hat{\theta}_{\mathrm{ML}}$. Prior je koncentrovaný okolo bodu 0.5. Dát je málo ($n = 20$), preto je aj posterior nimi len minimálne ovplyvnený. $\diamond$

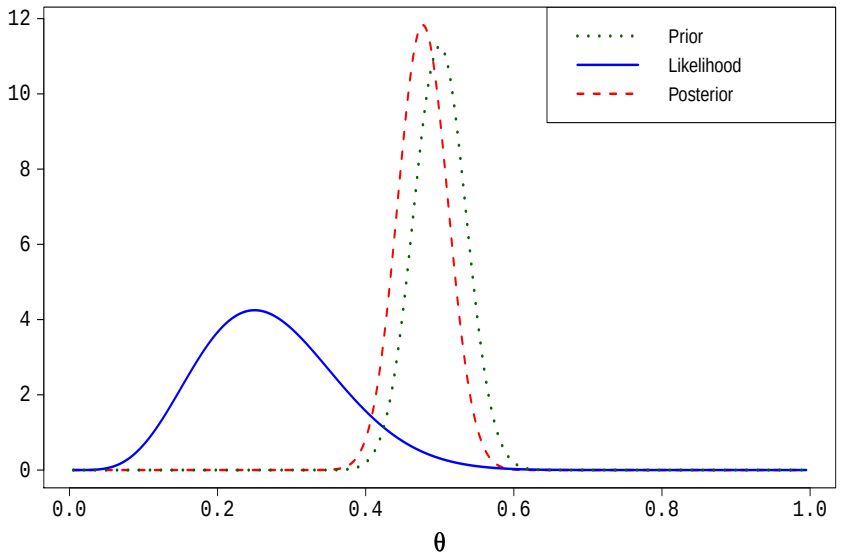
**3.2 Bodové charakteristiky posterioru** Posteriórne rozdelenie obsahuje všetku dostupnú informáciu: sumarizuje model a dáta.
V prípade, že je ale nutné zhrnúť posterior do jediného 'bodu', dá sa tak urobiť v rámci bayesovskej teórie rozhodovania. Rozhodovanie sa deje za neurčitosti. Nesprávne rozhodnutie má za následok straty. Je nutné špecifikovať stratovú funkciu $L(\theta, a)$, ktorá udáva závislosť straty od toho ako sa líši bayesovský odhad $a \in \Theta$ od $\theta$. Po tom ako bola špecifikovaná stratová funkcia je prirodzené hľadať také bodové zhrnutie $a$ posterióru (tzv. bayesovský estimátor/odhad), ktoré minimalizuje priemernú posteriórnu stratu

$$\int_{\Theta} L(\theta, a) p(\theta \mid x_1^n)\, d\theta.$$

Niektoré základné stratové funkcie a príslušné bayesovské odhady:

- posteriórna stredná hodnota, minimalizuje kvadratickú stratovú funkciu $L(\theta, a) = (a - \theta)^2$,

[4] `library(LearnBayes); prior = c(100, 100) # params of Beta prior; data = c(5, 15) # 5 successes and 15 failures; triplot(prior, data)`




Obrázok 2: Triplot pre Beta(100, 100) prior.

- posteriórny medián, minimalizuje absolútnu stratovú funkciu $L(\theta, a) = |a - \theta|$,
- posteriórny mód, minimalizuje všetko-alebo-nič (angl., *zero-one*) stratovú funkciu.

Nie vždy je zmysluplné hľadať bodovú charakteristiku v rámci bayesovskej teórie rozhodovania. V mnohých prípadoch sa za bodový odhad berie posteriórna stredná hodnota. Ak má ale posterior viac než jeden mód, je zrejmé, že v tomto prípade posteriórna stredná hodnota nie je dobrou bodovou charakteristikou.

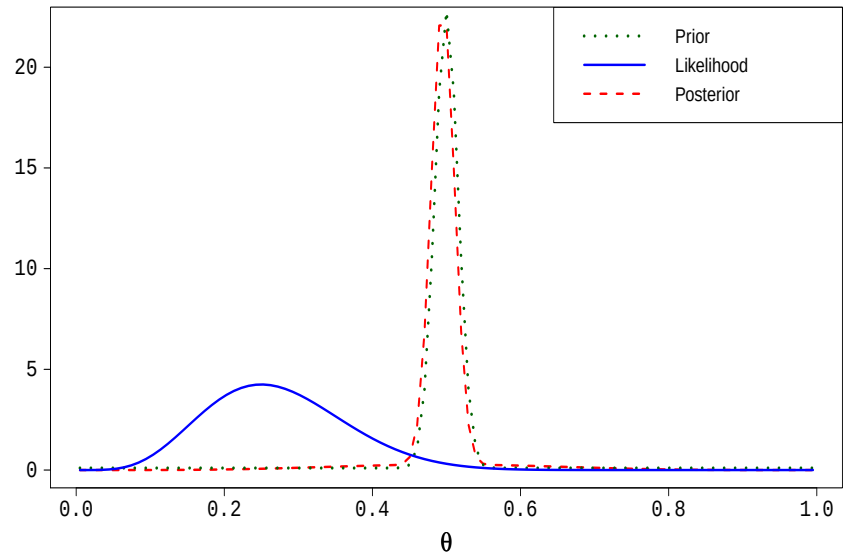
*Príklad (pokr.)*: Nakoľko posterior je Beta(105, 115), tak posteriórna stredná hodnota je $\frac{105}{105+115} = 0.47727$. Medián posterioru je[5] 0.47720, a posteriórny mód je 0.47706. $\diamond$

**3.3 Bayesovská robustnosť** Do akej miery závisí posterior na voľbe prioru? Tejto otázke je v bayesovských výskumoch venovaná veľká pozornosť. Tvoria náplň toho, čo sa zvykne nazývať bayesovská robustnosť alebo analýza citlivosti (angl., *sensitivity analysis*), viď [17]. Navrhnuté boli viaceré techniky na zisťovanie citlivosti posteriórnych úvah na priore, ako aj kvantitatívne miery robustnosti, viď napr. [2].

*Príklad (pokr.)*: Obmedzíme sa len na veľmi primitívnu formu analýzy citlivosti. Skúsime zobrať iný prior, ktorý by vyhovoval nášmu apriórnemu presvedčeniu $\mathrm{P}(0.44 < \theta < 0.56) = 0.9$. Napríklad prior daný ako nasledovná zmes hustôt $p(\theta) = 0.9\,\mathrm{Beta}(500, 500) + 0.1\,\mathrm{Beta}(1, 1)$ je taký. Aj posterior je potom zmes, a to $p(\theta \mid x_1^n) = 0.9\,\mathrm{Beta}(505, 515) + 0.1\,\mathrm{Beta}(6, 16)$. Na obrázku 3 je triplot[6] prioru, vierohodnostnej funkcie a posterioru.

[5] `qbeta(0.5, 105, 115)`

[6] `x = seq(0,1,0.001); curve(0.9*dbeta(x,505, 515) + 0.1*dbeta(x, 6,6), col = 'red') # posterior; curve(0.9*dbeta(x,500, 500) + 0.1*dbeta(x, 1,1), col = 'darkgreen', add = TRUE) # prior; curve(dbeta(x, 6, 16), col = 'blue', add = TRUE) # like; legend('topright', c("Prior", "Likelihood", "Posterior"), col = c("darkgreen", "blue", "red"))`




Obrázok 3: Triplot pre zmesový prior.

Posteriórna stredná hodnota je 0.472861, a teda je len o čosi menšia než hodnota 0.4772727, ktorú sme dostali pri Beta(100, 100) priore. ⋄

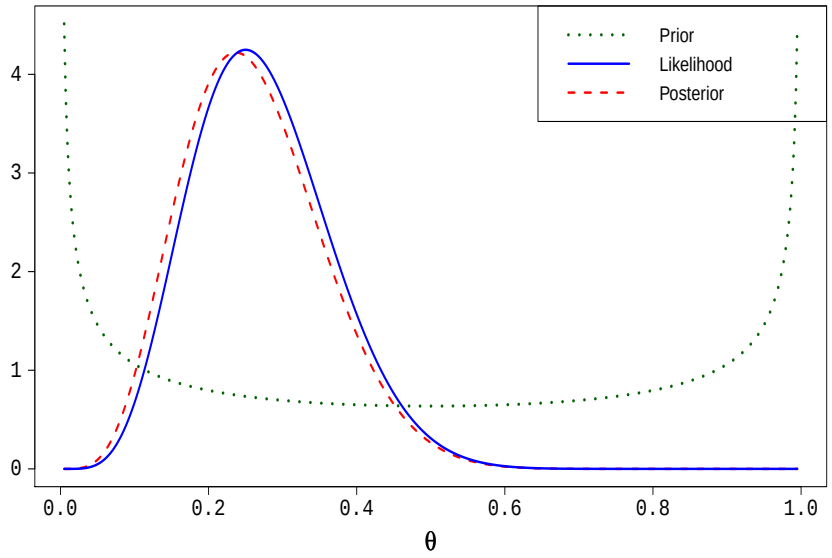
**3.4 Objektívne priory** Nie vždy je k dispozícii taká jasná apriórna informácia ako v prípade hádzania mincou. Značná časť bayesovských výskumov je venovaná tomu, ako pretransformovať žiadne alebo veľmi slabé apriórne presvedčenie o skutočnej hodnote parametra do podoby apriórneho rozdelenia. Tento problém je taký zásadný, že rozdeľuje bayesiáncov na dve veľké skupiny: subjektivistov a objektivistov. So značnou dávkou zjednodušenia sa dá povedať, že subjektivisti sú presvedčení, že nemá zmysel hľadať akési automatické (nesubjektívne) metódy na konštrukciu priorov. Naopak, objektivisti aktívne navrhujú metódy na prekladanie neznalosti (angl., *ignorance*) do podoby apriórneho rozdelenia. Jeden z prominentných predstaviteľov objektívnej bayesovskej štatistiky, James Berger, vraví: "V objektívnej bayesovskej analýze sa priory volia tak, aby predstavovali 'neutrálne' znalosti o neznámych". Väčšina bayesiáncov zastáva pragmatický postoj: v prípade modelu s mnohými parametrami sú pre nepodstatné parametre a parametre o ktorých hodnotách sa apriórne toho veľa nevie použité neinformatívne priory, a pre ostatné parametre sa subjektívne špecifikujú informatívne priory.

Existuje veľké množstvo techník na konštrukciu objektívnych priorov. Asi najznámejšou je technika navrhnutá Haroldom Jeffreysom. Jeffreysov prior má tvar $p(\theta) \propto \sqrt{I(\theta)}$, kde $I(\theta)$ je Fisherova informácia pre parameter $\theta$. Takto skonštruovaný prior spĺňa požiadavku invariantnosti[7]: pri transformácii $\theta$ na $g(\theta)$ sa Jeffreysova apriórna hustota transformuje tak, ako sa má hustota pravdepodobnosti transformovať, viď napr. [9]. Jefreysovské priory sú zvyčajne neznormalizovateľné, nazývajú sa preto pseudo-priory (angl., *improper priors*). Vie sa, že pre mnohé jeffreysovské priory býva výsledný posterior už

[7]Nebayesovské štatistické metódy nie sú, vo všeobecnosti, invariantné na reparametrizáciu. Jedinou výnimkou sú metódy založené na vierohodnostnej funkcii.

normalizovateľný, teda riadny (angl., *proper prior*). Neznormalizovateľnosť mnohých jeffreysovských priorov ale spôsobuje problémy pri bayesovskom porovnávaní hypotéz (pozri 3.7). Iné veľmi populárne objektívne priory sú tzv. referenčné priory, viď [19]. V posledných rokoch sa rozbehlo štúdium a aplikácie tzv. slabo informatívnych priorov (angl., *weakly informative priors*).

*Príklad (pokr.)*: Predpokladajme, že nemáme k dispozícii žiadnu apriórnu informáciu o minciach, netušíme, že mince sa razia strojovo zo značne homogénnych plechov kovu. Ináč povedané, apriórne nevieme o parametre $\theta$ bernoulliovského rozdelenia nič. Vyjadrime našu neznalosť jeffreysovským priorom: $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, čo je Beta$(0.5, 0.5)$ prior. Posterior je potom Beta$(5.5, 15.5)$. Z obrázku 4 je vidieť, že jeffreysovský prior naozaj vnáša do úvah minimum mimodátovej informácie.




Obrázok 4: Triplot pre Jeffreysov prior.

Stredná hodnota posterioru je 0.262. Mód posterioru je 0.237. Obe hodnoty sú blízke k ML odhadu. $\diamond$

**3.5 Intervalové charakteristiky posterioru** Existuje viacero možných intervalových charakteristík posterioru. V prípade, že posterior je približne symetrický, je rozumnou charakteristikou jeho rozpätia $100(1-\alpha)\%$-ný posteriórny interval, mimo ktorého leží pod každým chvostom posterioru $100(\alpha/2)\%$ masy rozdelenia. Takýto bayesovský konfidenčný interval sa nazýva (rovnako)chvostový (angl., *equal-tail*). V opačnom prípade je lepšou intervalovou charakteristikou vrcholový posteriórny interval (angl., *highest posterior density interval*, HPD), ktorý pozostáva z takej oblasti $s(x_1^n)$ parametrického priestoru $\Theta$, pre ktorú

$$1)\ \mathrm{P}(\theta \in s(x_1^n) \,|\, X_1^n = x_1^n) = 1-\alpha,$$
$$2)\ \text{ak } \theta_a \in s(x_1^n) \text{ a } \theta_b \notin s(x_1^n),\ \text{potom } p(\theta_a \,|\, X_1^n = x_1^n) > p(\theta_b \,|\, X_1^n = x_1^n).$$

Samozrejme, možné sú aj iné konštrukcie intervalových charakteristík posterioru.

*Príklad (pokr.)*: V prípade Beta$(100, 100)$ prioru a našich dát je posterior (t.j., Beta$(105, 115)$) v podstate symetrický. 95%-ný chvostový posteriórny interval je $(0.412, 0.543)$[8]. Pri danom priore a dátach môžeme tvrdiť, že s 95%-nou pravdepodobnosťou leží neznáma hodnota parametra $\theta$ v intervale $(0.412, 0.543)$.
Laplace, už v roku 1774 analyzoval bernoulliovskú schému pomocou rovnomerného prioru (ktorý je totožný s Beta$(1, 1)$). Posterior je v takom prípade Beta$(n_1 + 1, n - n_1 + 1)$. Pre naše dáta by v takomto prípade 95%-ný posteriórny chvostový interval bol $(0.132, 0.437)$.
$\diamond$

**3.6 Testovanie v teórii rozhodovania** Testovanie hypotéz je možné robiť v rámci bayesovskej teórie rozhodovania. Formuluje sa nulová hypotéza $H_0 : \theta \in \Theta_0$ a alternatívna hypotéza $H_1 : \theta \in \Theta_0^C$. Možné sú dve akcie: akcia $a_0$ – prijatie $H_0$, akcia $a_1$ – prijatie $H_1$. Je nutné špecifikovať stratovú funkciu, ktorá vyjadruje veľkosť straty v prípade chyby prvého a chyby druhého druhu, keď sa vykoná akcia $a$.
Najbežnejšou stratovou funkciou je zovšeobecnená 0-1 strata:

$$L(\theta, a_0) = \begin{cases} c_{\mathrm{II}} & \text{ak } \theta \in \Theta_0^c, \\ 0 & \text{inák}, \end{cases}$$

$$L(\theta, a_1) = \begin{cases} c_{\mathrm{I}} & \text{ak } \theta \in \Theta_0, \\ 0 & \text{inák}. \end{cases}$$

Zovšeobecnená je v tom, že strata $c_{\mathrm{II}}$ v prípade chyby druhého druhu nemusí byť rovnaká ako strata $c_{\mathrm{I}}$ v prípade chyby prvého druhu. Bayesovský test spočíva v tom, že sa vyberie tá z hypotéz, ktorá aposteriórne spôsobí v priemere menšiu stratu. Posteriórna priemerná strata sa zvykne tiež nazývať riziko (angl., *risk*). Riziko má v prípade zovšeobecnenej 0-1 straty tvar:

$$r(a_0 \,|\, x_1^n) = c_{\mathrm{II}}[1 - \mathrm{P}(H_0 \text{ true} \,|\, x_1^n)],$$

$$r(a_1 \,|\, x_1^n) = c_{\mathrm{I}}\mathrm{P}(H_0 \text{ true} \,|\, x_1^n).$$

Aby sme zamietli nulovú hypotézu musí byť teda $r(a_1 \,|\, x_1^n) < r(a_0 \,|\, x_1^n)$. To je ekvivalentné tvrdeniu, že

$$\text{zamietni } H_0 : \theta \in \Theta_0, \quad \text{ak} \quad \mathrm{P}(\theta \in \Theta_0 \,|\, x_1^n) < \frac{c_{\mathrm{II}}}{c_{\mathrm{I}} + c_{\mathrm{II}}}.$$

V opačnom prípade sa $H_0$ prijíma. Ak sú straty u oboch chybných rozhodnutí rovnaké, tak test zamieta nulovú hypotézu ak je jej posteriórna pravdepodobnosť menšia než $1/2$.
*Príklad (pokr.)*: Chceme testovať hypotézu, že $\theta$ je z $\Theta_0 = (0.47, 0.53)$. Použijeme Beta$(100, 100)$ prior $p(\theta)$. Prior určuje apriórnu pravdepodobnosť nulovej hypotézy $p_0 = \int_{\Theta_0} p(\theta)\, d\theta$, čo je $0.604$[9]. Apriórne sme teda o čosi viac presvedčení o platnosti nulovej hypotézy, než o alternatíve. Predpokladajme že straty sú rovnaké u oboch chybných rozhodnutí. Nakoľko $\mathrm{P}(\theta \in \Theta_0 \,|\, x_1^n) = 0.526$[10], tak nulovú hypotézu nezamietame. Všimnime si, že aposteriórne, po tom ako sme obdržali dáta, naše presvedčenie o platnosti nulovej hypotézy, v porovnaní s apriórnym presvedčením, kleslo. $\diamond$

[8] `qbeta(0.025, 105, 115); qbeta(0.975, 105, 115)`
[9] `u = pbeta(0.53, 100, 100); l = pbeta(0.47, 100, 100); p0 = u-l`
[10] `up = pbeta(0.53, 105, 115); lp = pbeta(0.47, 105, 115); pn0 = up-lp`

Za zmienku stojí, že ak je nulová hypotéza bodová, potom je bayesovské testovanie problematické, pretože si vyžaduje priradenie pravdepodobnostnej miery jednoprvkovej množine. Keď už bayesiánci musia testovať bodovú nulovú hypotézu, robia tak pomocou prioru obsahujúceho diracovskú mieru vo vyšetrovanom bode.

**3.7 Porovnávanie hypotéz** Testovanie v rámci teórie rozhodovania je síce principiálne, ale zato veľakrát príliš zväzujúce. V prípadoch, keď nie je možné alebo žiadúce formulovať stratovú funkciu, má bayesiánec možnosť namiesto testovania, hypotézy $H_0$ a $H_1$ porovnať. Robí sa to pomocou veľmi prirodzeného nástroja, ktorý sa nazýva bayesovský faktor (angl., *Bayes factor*). Ak je prior riadny, potom je možné definovať apriórny pomer šancí (angl., *prior odds ratio*):

$$\frac{p_0}{p_1} = \frac{\int_{\Theta_0} p(\theta)\, d\theta}{\int_{\Theta_1} p(\theta)\, d\theta},$$

a aposteriórny pomer šancí (angl., *posterior odds ratio*):

$$\frac{p_{n,0}}{p_{n,1}} = \frac{\int_{\Theta_0} p(\theta \mid x_1^n)\, d\theta}{\int_{\Theta_1} p(\theta \mid x_1^n)\, d\theta}.$$

*Bayesovský faktor* v prospech $H_0$ je

$$\mathrm{BF}_{01} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p_{n,0}/p_{n,1}}{p_0/p_1}.$$

Nakoľko nás skôr zaujíma či nulovú hypotézu zamietame, než jej potvrdenie, zvykne sa robiť s bayesovským faktorom $\mathrm{BF}_{10}$ proti $H_0$. Už Jeffreys navrhol kalibráciu $\mathrm{BF}_{10}$ podľa ktorej sa určuje, aký silný je dôkaz o neplatnosti $H_0$. Pred nedávnom bola táto kalibrácia mierne modifikovaná v práci [27]. Uvedená je v Tabuľke 1.

| $\mathrm{BF}_{10}$ | dôkaz o neplatnosti $H_0$ |
|---|---|
| 1 až 3 | takmer žiadny |
| 3 až 20 | mierny |
| 20 až 150 | silný |
| $> 150$ | veľmi silný |

Tabuľka 1: Kalibrácia bayesovského faktora $\mathrm{BF}_{10}$.

*Príklad (pokr.)*: Prior nech je Beta$(100, 100)$; Potom $\mathrm{BF}_{10}$ proti $H_0 : \theta \in (0.47, 0.53)$ je $1.374$[11], teda nie je takmer žiadny dôkaz toho, že by $H_0$ bola neplatná. $\diamond$

Mnohí bayesovskí štatistici považujú hypotézy za neužitočnú formu štatistického usudzovania. Napr. v [3] sa hypotézy spomínajú len v súvislosti s kritikou nebayesovskej štatistiky.

[11] `p1 = 1 - p0; pn1 = 1 - pn0; BF10 = pn1/pn0/(p1/p0)`

**3.8 Predikcie** Modelovanie sa často robí za účelom predpovedania, robenia inferencií o nových pozorovaniach $\tilde{x}$, teda, prediktívnych inferencií. Bayesiáncom na to slúži posteriórna prediktívna hustota (angl., *posterior predictive density/distribution*)

$$p(\tilde{x} \mid x_1^n) = \int p(\tilde{x} \mid \theta, x_1^n) p(\theta \mid x_1^n)\, d\theta. \tag{2}$$

Tento dôležitý koncept sa nepoužíva len na predikcie ale aj na posúdenie (angl., *validation*) modelu (t.j., prioru a dáta-generujúceho rozdelenia).

Podobne sa zavádza aj apriórna prediktívna hustota (angl. *prior predictive distribution*)

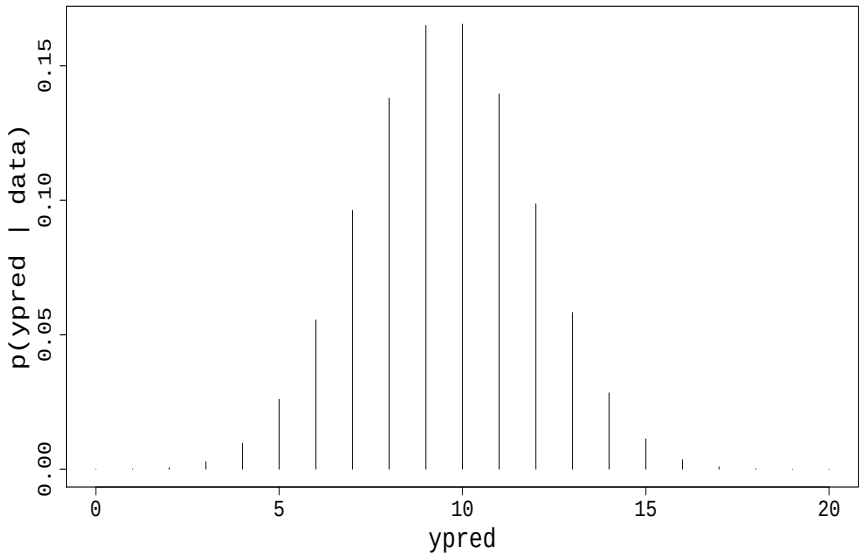
$$p(\tilde{x}) = \int p(\theta) p(\tilde{x} \mid \theta)\, d\theta,$$

nazývaná tiež marginálna hustota (angl., *marginal density*).

*Príklad (pokr.)*: Posteriórna prediktívna pravdepodonosť toho, že v budúcom výbere o veľkosti $m$ bude $\tilde{y}$ výskytov $X = 1$ je (viď [1])

$$p(\tilde{y} \mid x_1^n) = \binom{m}{\tilde{y}} \frac{B(\alpha + \tilde{y}, \beta + m - \tilde{y})}{B(\alpha, \beta)}, \tag{3}$$

kde $\tilde{y} = 0, 1, \ldots, m$ a $B(\cdot, \cdot)$ je beta funkcia; $\alpha$, $\beta$ sú parametre posterioru.
Prior nech je Beta$(100, 100)$, a nech $m = 20$. Posteriórne prediktívne rozdelenie[12] je na obr. 5.




Obrázok 5: Posteriórne prediktívne rozdelenie $p(\tilde{y} \mid x_1^n)$.

[12] `library(LearnBayes); ab = c(105, 115) # posterior; m = 20; ys = 0:20; pred = pbetap(ab, m, ys); plot(ys, pred, type = 'h')`

92%-ný posteriórny prediktívny interval pre $\tilde{y}$, pri daných dátach a priore je[13] $[6, 13]$. Teda, s 92%-nou pravdepodobnosťou bude v budúcich 20-tich hodoch mincou počet úspechov ($X = 1$) niekde medzi šesť až trinásť.
V prípade, že by nás zaujímala pravdepodobnosť toho, že v budúcom jedinom hode nastane úspech, potom by sme podľa (3) dostali $p(X = 1 \,|\, x_1^n) = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta}$, čo je priemer posterioru. V prípade rovnomerného prioru by sme dostali známe Laplaceovo pravidlo postupu (angl., *rule of succession*): $p(X = 1 \,|\, x_1^n) = \frac{n_1+1}{n+2}$.
Pre Beta$(100, 100)$ prior je pravdepodobnosť, že po sekvencii 20-tich hodov, v ktorých došlo k úspechu 5 krát, nastane ďalší úspech, rovná 0.477 – priemer posterioru. Pripomeňme, že nebayesiánec by túto pravdepodobnosť odhadol ML odhadom, ktorý je 0.25. Pre rovnomerný prior je hľadaná pravdepodobnosť rovná 0.273. $\diamond$

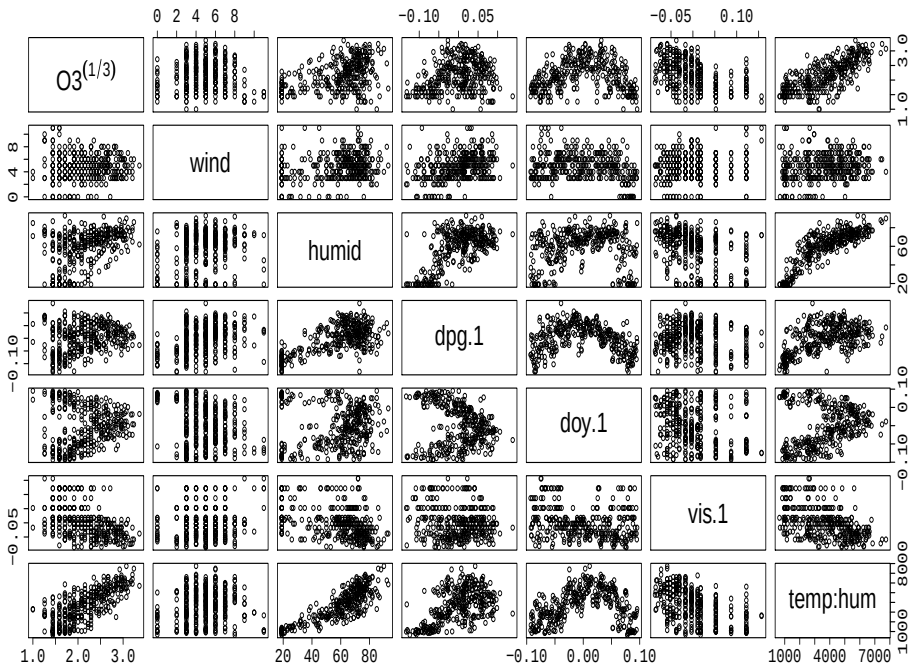
Ako už bolo spomenuté, posteriórna prediktívna pravdepodobnosť sa používa aj na posteriórnu prediktívnu kontrolu modelu, na posúdenie, zhodnotenie modelu. Voľne povedané, ak je pozorovaný údaj v strede posteriórneho prediktívneho rozdelenia, potom je v súlade s fitom modelu. Ak ale nameraná hodnota leží na chvoste prediktívneho posterioru, potom ju model nevystihuje. Na formálne vyhodnotenie kvality modelu sa používa posteriórna prediktívna p-hodnota, viď [3].

*Príklad (pokr.)*: Posteriórna prediktívna pravdepodonosť $p(\tilde{y} \leq 5 \,|x_1^n)$ toho, že v nasledovnej 20-tici hodov sa bude $X = 1$ vyskytovať 5-krát alebo menej je[14] 0.0392. To by naznačovalo, že náš model (t.j., prior a dáta-generujúce rozdelenie) nedostatočne vystihuje pozorované dáta, viď tiež obr. 5. Nakoľko dáta-generujúce rozdelenie je v našom prípade sotva spochybniteľné, mýliť sa môžme len v priore. Keďže máme ale dočinenia s malým výberom ($n = 20$) a apriórna informácia je dostatočne silná, nie je v tomto prípade dôvod meniť model. $\diamond$

Použitie prediktívneho posterioru na posúdenie modelu si zaslúži ešte jednu ilustráciu.

*Príklad*: V `R`-kovej knižnici `faraway` sa nachádzajú dáta `ozone` obsahujúce merania hladiny ozónu ako aj deviatich ďalších premenných. Vyberieme z nich šesť (wind, humidity, temp, dpg, vis, doy), a s ich pomocou budeme chcieť v rámci regresného modelu modelovať hladinu ozónu O3. Po fáze budovania regresného modelu, skončíme napríklad pri modeli $\text{O3}^{1/3} \sim$ wind+humidity+poly(dpg,2)+poly(doy, 2)+poly(vis,2)+temp:humidity. Párový graf je na obr. 6.
Na parametre modelu $\beta$, $\sigma^2$ položíme jeffreysovský prior. Nájdeme posterior: $\beta \,|\, \sigma^2, y \sim n(\hat{\beta}, \sigma^2 V_\beta)$ a $\sigma^2 \,|\, y \sim$ Inv-$\chi^2(n-k, s^2)$, kde $\hat{\beta} = (X'X)^{-1}X'y$, $V_\beta = (X'X)^{-1}$, $k$ je počet prediktorov, $n$ počet pozorovaní, $s^2$ odhad variancie chýb. Hoci je v tomto prípade aj posteriórna prediktívna hustota $p(\tilde{y}_1^n \,|\, y_1^n)$ vyjadriteľná analyticky (viď napr. [3]), stojí za zmienku, že predstava o prediktívnom posteriore $p(\tilde{y}_1^n \,|\, y_1^n)$ budúcich dát $\tilde{y}_1^n$, sa dá získať aj bez rátania integrálu v (2). Stačí ak vieme generovať nové $\tilde{y}$ z rozdelenia $p(\tilde{y}, \theta \,|\, y_1^n)$. Odhad hľadaného prediktívneho posterioru sa dá dostať napríklad jadrovým vyhladením vygenerovaných dát $\tilde{y}$. Nechajme si z posterioru vygenerovať 1000 nových dvojíc $(\beta, \sigma^2)$, a pre každú následne vygenerujeme pri danej matici plánu $X$, $n$-ticu nových $\tilde{y}$. Jadrovo vyhladíme všetkých 1000 $n$-tíc a do výsledného odhadu prediktívneho posterioru ešte za-

[13] `discint(cbind(ys, pred), 0.9)`
[14] `sum(pred[1:6])`




Obrázok 6: Párový graf dát.

kreslíme pomocou zvislých čiar pozície nameraných $y_1^n$. Výsledok[15] je na obr. 7. Z obrázka je zrejmé, že model v hrubých rysoch vystihuje dáta, ale zďaleka nie uspokojivo, nakoľko sa väčšina pozorovaných dát nachádza na pravom chvoste prediktívneho posterioru, a nie v strede.

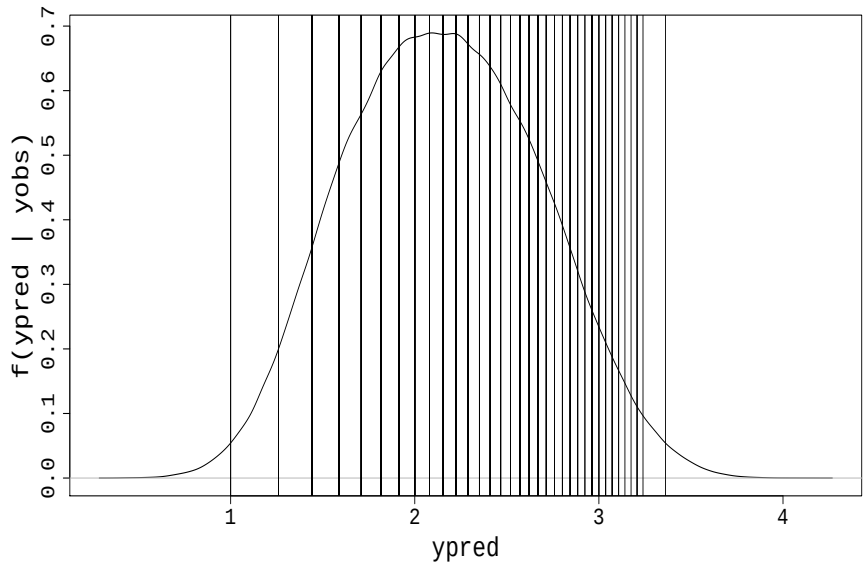
V inej forme je tá ista vada modelu viditeľná na obr. 8, zobrazujúcom graf závislosti priemerov z $\tilde{y}$-ov nasimulovaných z prediktívneho posterioru oproti pozorovaným $y_1^n$, ako aj referenčnú, 45° čiaru[16].

Z oboch obrázkov vyplýva, že by bolo potrebné modifikovať model tak, aby sa stred prediktívneho posterioru posunul doprava. ⋄

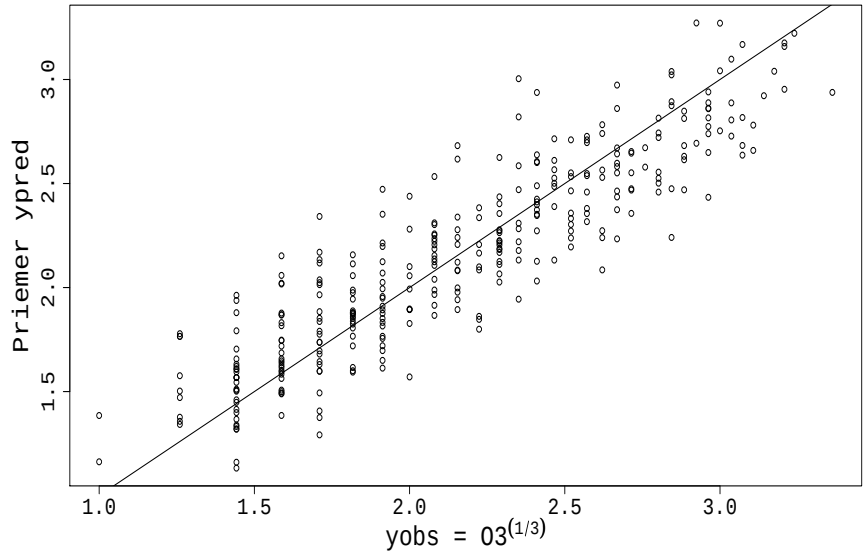
**3.9 Priemerovanie modelov** Aby sme mohli popísať ďalší užitočný bayesovský koncept – priemerovanie modelov, budeme musieť opustiť hádzanie mincou. Ideálna na to bude opäť regresia. Predpokladajme, že modelujeme regresným modelom náhodnú premennú $y$ pomocou nejakej množiny prediktorov. Býva zvykom hľadať najlepší podmodel, teda podmnožinu množiny prediktorov, ktorá by v nejakom zmysle najlepšie vystihovala chovanie závislej premennej. Obyčajne sa kvalita posudzuje nejakým kritériom (napr. Akaikeho kritériom, viď [6] alebo [8]), ktoré zaručuje parsimóniu, teda rovnováhu medzi počtom prediktorov a tesnosťou fitu. V bayesovskej štatistike na tento účel slúži Bayesovské informačné kritérium (angl., *Bayesian Information Criterion*, BIC) (viď [6]),

[15] `parsim = blinreg((oz$O3)^(1/3), as.matrix(Xm), m = 1000); predy = blinregpred(as.matrix(Xm), parsim); plot(density(predy)) for (i in 1:330){abline(v = ((oz$O3)[i])^(1/3))}`

[16] `plot((oz$O3)^(1/3), colMeans(predy)); abline(0,1)`




Obrázok 7: Prediktívna aposteriórna hustota $p(\tilde{y}_1^n \mid y_1^n)$ a pozície nameraných $y_1^n$.




Obrázok 8: Graf priemerov z $\tilde{y}$-ov nasimulovaných z prediktívneho posterioru oproti pozorovaným $y_1^n$.

ktoré je aproximáciou bayesovského faktora.

Snaha o výber 'naj' modelu je ale problematická, pretože výberom jediného modelu sa ignoruje neistota, ktorú máme o jednotlivých modeloch (angl., *model uncertainty*), a táto neistota veľmi často prevažuje všetky ostatné zdroje neistoty. Bayesovské priemerovanie modelov (angl., *Bayesian Model Avergaing*, BMA) poskytuje koherentný spôsob ako zobrať neistotu o modeli do úvahy. Vychádza sa pri tom z apriórneho (zvyčajne rovnomerného)

rozdelenia $p(\cdot)$ na množine modelov[17] $M = \{M_1, \ldots, M_k\}$. Prior sa následne bayesovsky koriguje dátami $X_1^n$ a obdrží sa posterior

$$p(M_j \mid x_1^n) = \frac{p(x_1^n \mid M_j)p(M_j)}{\sum_{l=1}^{k} p(x_1^n \mid M_l)p(M_l)}, \tag{4}$$

kde $p(x_1^n \mid M_j)$ je integrovaná vierohodnosť (angl. *integrated likelihood*)

$$p(x_1^n \mid M_j) = \int p(x_1^n \mid \theta_j, M_j)p(\theta_j \mid M_j)\, d\theta_j, \tag{5}$$

a $\theta_j$ je parameter modelu $M_j$. Posteriórne inferencie a predikcie sú teda založené na množine modelov, a nie na jedinom, akokoľvek optimálnom modeli.
Nech $\phi$ je parameter, ktorý nás zaujíma. Posteriórna modelovo-spriemerovaná hustota $p(\phi \mid x_1^n)$ je

$$p(\phi \mid x_1^n) = \sum_{j=1}^{k} p(\phi \mid x_1^n, M_j)p(M_j \mid x_1^n).$$

No a jej stredná hodnota a variancia sú modelovo-spriemernené posteriórne charakteristiky parametra $\phi$.

*Príklad*: V `R`-kovej knižnici `MASS` sa nachádzajú dáta `UScrime` obsahujúce údaje o 47 amerických štátoch, za rok 1960. Údaje boli zhromaždené za cieľom modelovania kriminality. Prediktorov je 15. Po logaritmickej transformácii prediktorov ako aj závislej premennej použijeme knižnicu `BMA`[18] (viď [31]) na nájdenie modelovo-spriemerovaných posteriórnych hustôt parametrov modelu[19].
Z nasledovného numerického zhrnutia sa môžeme dozvedieť, že 5 modelov s najvyššou posteriórnou pravdepodobnosťou má kumulatívnu pravdepodobnosť iba 0.3. Stĺpec `p!=0` obsahuje údaje, pre štyri z prediktorov, o tom, aká je (percentuálna) pravdepodobnosť, že daný prediktor je v modeli. V stĺpci `EV` sa nachádzajú stredné hodnoty modelovo-spriemernenej posteriórnej hustoty, a `SD` obsahuje smerodatné odchýlky, pre každú premennú. V ďalších stĺpcoch sú bayesovské odhady parametrov, v piatich najlepších modeloch.

```
  51  models were selected

 Best  5  models (cumulative posterior probability =  0.3 ):

    p!=0     EV     SD   model1  model2  model3  model4  model5
M   97.5   1.4014  0.532  1.478   1.514   1.605   1.268   1.461
```

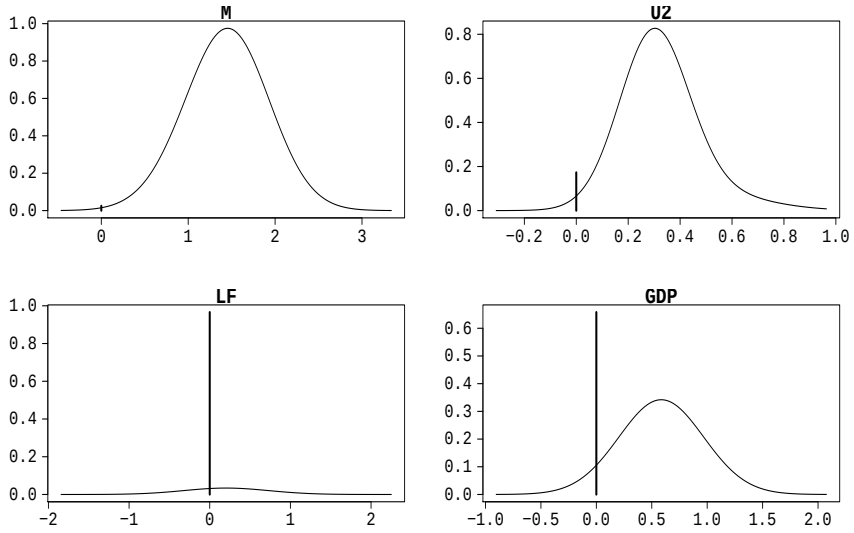
[17] Množina $M$ môže byť vo všeobecnosti spojitá.

[18] V knižnici je implementovaná aproximácia integrovanej vierohodnosti pomocou BICu, čo značne znižuje komplexitu výpočtov. V knižnici `BAS` je na výpočet použité samplovanie bez navracania, z posteriórnej hustoty.

[19] `library(MASS); library(BMA); data(UScrime); xcrime = UScrime[,-16]; xcrime = log(xcrime[,-2]); ycrime = log(UScrime[,16]); breg = bicreg(xcrime, ycrime); summary(breg, digits = 2); plot(breg, mfrow = c(2,2), include = c(1,10,5,11), include.inter = F); imageplot.bma(breg)`

```
U2  82.7   0.2709  0.194  0.289   0.322   0.274   0.281   0.330
LF   3.4   0.0069  0.103    .       .       .       .       .
GDP 34.2   0.2006  0.357    .       .     0.541     .       .
```

Na obr. 9 sú modelovo-spriemernené posteriórne hustoty zobrazené pre 4 vybrané prediktory. Veľkosť Diracovho impulzu v bode 0 udáva relatívny počet podmodelov (z celkového počtu 51 podmodelov s najvyššou posteriórnou pravdepodobnosťou), v ktorých sa daný prediktor nevyskytoval. Napríklad u prediktora M (percento mužov vo veku 14-24 rokov) je táto pravdepodobnosť (relatívna početnosť) veľmi malá. Krivka zobrazuje modelovo-spriemernenú posteriórnu hustotu, cez podmodely v ktorých sa daný prediktor nachádza.

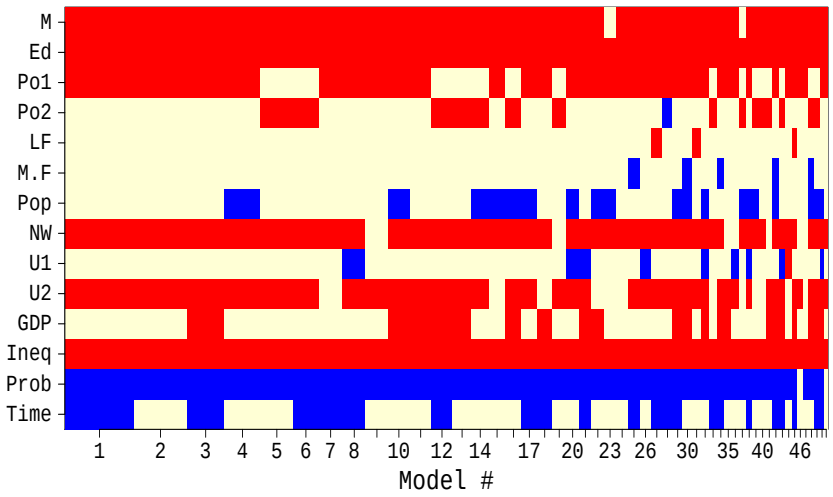



Obrázok 9: Modelovo-spriemernený posterior pre štyri parametre modelu.

Predstava o tom, ktorý prediktor je ako významný sa dá rýchlo získať aj z BMA koberca (angl., *image plot*). Na obr. 10 sú na $x$-ovej osi zoradené modely, zostupne, podľa posteriórnej pravdepodobnosti. Šírka binu vyjadruje veľkosť posteriórnej pravdepodobnosti daného modelu. Farebne je indikovaná posteriórna pravdepodobnosť (v %), že prediktor (na osi $y$) je v modeli. Červená[20] označuje hodnoty vyššie než 90%, modrá hodnoty[21] v intervale $(80, 90)$%.
Pre porovnanie[22], AIC (angl., *Akaike Information Criterion*) by do najlepšieho podmodelu vybralo prediktory M, Ed, Po1, M.F, Pop, NW, U2, GDP, Ineq, Prob, Time. Mnohé z nich sa v bayesovsky spriemerovanom modeli nenachádzajú. Naj-AIC model ani nepatrí medzi päť aposteriórne najpravdepodobnejších modelov. ⋄

Dôležitosť priemerovania modelov si vďaka bayesiáncom uvedomili aj nebayesiánci, viď [6].

[20] V tlačenej podobe tmavo sivá.
[21] V tlačenej podobe čierna.
[22] `d = cbind(ycrime, xcrime); m = lm(ycrime ~ ., data = d)`; summary(m); stepAIC(m)




Obrázok 10: BMA koberec

**3.10 A ďalšie** 1) Inferencie – založené na (1), predikcie a posúdenie modelu – založené na (2), priemerovanie modelov – založené na (4) a (5), ako aj ďalšie bayesovské operácie si vyžadujú výpočet integrálov. Bayesiánci sú spolu so štatistickými fyzikmi, od ktorých sa pred pár desaťročiami naučili integrovať mnohorozmerné integrály pomocou markovovského monte carlo simulovania (angl., *Markov chain Monte Carlo*, McMC), známi ako 'agresívni' výpočtári. Hoci sú základné McMC algoritmy (Metropolisov, Metropolisov-Hastingsov algoritmus, Gibbsov generátor (angl., *Gibbs sampler*)) pomerne ľahko popísateľné (viď napr. [9]), ich praktické zvládnutie si vyžaduje značnú dávku kumštu.

2) Bayesiánci sú majstri v modelovaní pomocou hierarchických modelov. S týmito modelmi úzko súvisí aj zaujímavá herézia, známa ako empirický bayes. Jednoduchým príkladom hierarchického bayesovského modelu je: $X_i \mid \theta_i \sim n(\theta_i, \sigma^2)$, kde $X_i$ $(i = 1, 2, \ldots, n)$ sú nezávislé; a $\theta_i \sim n(\mu, \tau^2)$, kde $\theta_i$ $(i = 1, 2, \ldots, n)$ sú zameniteľné. Dá sa ukázať, že marginálne rozdelenie $X_i$ je $n(\mu, \sigma^2 + \tau^2)$. Ak je $\sigma^2$ známe, môžme teda parametre $\mu$, $\tau^2$ prioru (nazývajú sa tiež hyperparametrami) odhadnúť z dát. Vďaka marginálnemu rozdeleniu je tak možné určiť hyperparametre prioru z dát. V takomto empirickom bayesovskom prístupe sú teda dáta použité dva razy: na určenie hyperparametrov prioru a aj na jeho aktualizáciu (angl., *updating*). V hierarchickom bayesovskom modelovaní je podstatný predpoklad zameniteľnosti[23] (angl., *exchangeability*) parametrov $\theta_i$. Parametre $\theta_1, \theta_2, \ldots, \theta_k$ sú zameniteľné ak ich rozdelenie je invariantné na ich permutáciu. So zameniteľnosťou úzko súvisí aj zásadná de Finettiho veta.

3) Odhliadnuc od fundamentálnej odlišnosti bayesovského prístupu od všetkých nebayesovských prístupov, je možné chápať bayesovanie ako metódu, pomocou ktorej sa dajú (princípiálnym spôsobom) regularizovať ML odhady. Napríklad známy hrebeňový odhad (angl., *ridge estimator*) parametrov gaussovského regresného modelu, používaný nebayesiánmi v prípade že matica plánu je takmer singulárna, sa dá bayesovsky obdržať pomocou gau-

[23] Zaujímavá diskusia o nezávislosti a zameniteľnosti je v [30].

ssovského prioru, viď napr. [9].

4) V posledných rokoch sa prudko rozvíja neparametrická bayesovská štatistika. Pravdepodobnostné modely parametrizované konečným počtom parametrov sú veľakrát príliš neflexibilné. V neparametrickom bayesovaní (viď napr. [5]) sa kladie prior na množinu všetkých rozdelení.

5) ABC (angl., *Approximate Bayesian Computation*) je zaujímavá technika, pomocou ktorej sa dajú robiť bayesovské úvahy v prípade, že vierohodnostná funkcia je numericky nezvládnuteľná. Na to, aby sa získala vzorka hodnôt z posterioru sa v ABC postupuje podľa tohto receptu:

1) zvoľte pracovnú (angl. *proposal*) apriórnu hustotu $p(\theta)$,

2) vygenerujte kandidáta $\theta^*$ z $p(\theta)$,

3) nasimulujte dáta $Y_1^n$ z $p(Y_1^n \mid \theta^*)$,

4) vypočítajte vhodnú sumárnu štatistiku $S(y_1^n)$ pre nasimulované dáta, ako aj pre napozorované dáta $S(x_1^n)$,

5) akceptujte $\theta^*$ v prípade, že vhodná miera vzdialenosti $\delta(\cdot,\cdot)$ spĺňa
$\delta(S(x_1^n), S(y_1^n)) < \epsilon$.

ABC sa objavilo v populačnej genetike a rozšírilo sa aj do ďalších disciplín.

## 4. Pêle-mêle

Pre mnohých nebayesovských štatistikov je bayesovský rámec príliš zväzujúci, rigidný. Z pohľadu bayesovskej štatistiky sa zas nebayesovská štatistika javí ako súbor 'ad hoc' procedúr, 'keby'-uvažovania, a paradoxov. Tento desaťročia sa vyvíjajúci bayesovsko-nebayesovský ping-pong vniesol mnoho poznania do základov štatistiky, aj do praktickej analýzy dát, ale schizma trvá. Viacerí štatistici hľadajú kompromis, viď napr. nedávnu prácu [29].

Vnútrobayesovským sporom (subjektivisti vs. objektivisti) je venovaná veľká časť tretieho zväzku prvého ročníka časopisu *Bayesian analysis*. V nedávnom čísle (Vol. 3, No. 3, 2008) toho istého časopisu je zas zaujímavá debata, vyprovokovaná Gelmanovým článkom, ktorý zhŕňa výhrady voči bayesovskej štatistike.

Thomas Bayes žil v rokoch 1702 až 1761. Jeho slávna 'Esej o riešení jedného problému v rámci doktríny šancí' vyšla v roku 1763. Pomenovanie 'bayesiánci' vyvoláva v mnohých dojem akejsi sekty, viď [32]. Keby sa bayesiánci volali napríklad posterioristi hneď by to znelo menej sektársky!

K posterioristom majú spomedzi tých iných najbližšie asi stúpenci prístupu k štatistike, ktorý je založený na vierohodnostnej funkcii[24]. Mimo zásadného rozdielu medzi likelihood-wallahovským 'keby' a posteriórskym 'keď' uvažovaním, sa oba prístupy líšia aj po technickej stránke: posterioristi integrujú, wallahovia maximalizujú.

Dva citáty, ako pripomienka na časy 'studenej vojny'. Maurice Kendall: 'Život by bol oveľa jednoduchší keby bayesiánci nasledovali svojho majstra a publikovali svoje práce posmrtne'. Dennis V. Lindley: 'Vo vnútri každého nebayesiánca je ukrytý bayesiánec, snažiaci sa predrať von.' Keď sa dvaja bijú, môže zmocnieť niekto tretí. Dolovanie v dátach (angl., *data mining*), strojové učenie (angl., *machine learning*) sú prístupy k analýze dát, v ktorých sa preferuje algoritmické získavanie informácie z dát (viď [20]),

[24] Basu, v poučnej práci [15] ich označuje slovom 'likelihood-wallah'.

oproti štatistickému prístupu, ktorý je – nebayesovský, rovnako ako bayesovský – založený na pravdepodobnostnom modeli.

**Referencie**


Učebnice bayesovskej štatistiky

1. Albert, J. (2009). *Bayesian computation with R*. New Yeork:Springer-Verlag. 2nd ed.
2. Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New Yeork:Springer-Verlag.
3. Gelman, A., Karlin, B., Stern, H. and Rubin, D. (2004). *Bayesian Data Analysis*. London:Chapman& Hall. 2nd ed.
4. Hoff, P. (2009). *A First Course in Bayesin Statistical Methods*. New Yeork:Springer-Verlag.
5. Ghosh, J. K., and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York:Springer-Verlag.
6. Hjort, N. L. and Claeskens, G. (2008). *Model Selection and Model Averaging*. Cambridge:CUP.
7. Koop, G. (2003). *Bayesian Econometrics*. Chichester:Wiley.
8. Mittelhammer, R., Judge, G. and Miller, D. (2000). *Econometric Foundations*. Cambridge:CUP.
9. Pázman, A. (2003). *Bayesovská štatistika*. Bratislava:Univerzita Komenského.
10. Roberts, Ch. (2001). *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*. New York:Springer-Verlag. 2001, 2nd ed.

Výpočtová stránka bayesovskej štatistiky

11. Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. New Yorks:Chapman & Hall.
12. I. Ntzoufras (2009). *Bayesian modeling using WinBUGS*. New Jersey:Wiley.
13. Marin, J.-M. and Robert, Ch. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York:Springer-Verlag.

Články

14. Aldrich, J. (2008). R. A. Fisher on Bayes and Bayes' Theorem. *Bayesian Analysis*. 3/1:161-170.
15. Basu, D. (1975). Statistical information and likelihood. *Sankhya*. 37/1:1-71.

16. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Royal Soc.* 53: 370-41.
Facsimile available at http://www.york.ac.uk/depts/maths/histstat/essay.pdf.

17. Berger, J. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference*, 25:303-328.

18. Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1/3:385-402. With discussion.

19. Berger, J., Bernardo, J. and Sun, D. (2009). The formal definition of reference priors. *Ann. Stat.*, 37:905-938.

20. Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16/3:199-231.

21. Brown, L. D., Cai, T. T. and Dasgupta, A. (2001). Interval Estimation for a Binomial Proportion. *Stat. Sci.*, 16:101-133.

22. Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Am. Stat.*, 49:327-225.

23. Clide, B. A. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics 6*, J. Bernardo et al. (eds.), pp. 157-185.

24. Christensen, R. (2005). Testing Fisher, Neyman, Pearson and Bayes, *Am. Stat.*, 59:121-126.

25. Gelman, A, (2008). Objections to Bayesian statistics. *Bayesian Analysis* 3/3:445-467. With discussion by J. M. Bernardo, J. B. Kadane, S. Senn, and L. Wasserman.

26. Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, 1/3:403-420. With discussion.

27. Kass, R. and Reftery, A. (1995). Bayes factors. *J. Am. Stat. Assoc.*, 90:773-795.

28. Kass, R. E. and Wasserman, L. A. (1996). The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.* 91:1343-1370.

29. Little, R. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Am. Stat.*, 60/3:1-11.

30. Pázman, A. (2004). Bayesovská štatistika (Ako alternatíva pre stredné školy). *Obzory mat. fyz. info.* 33/1:1-14.

31. Raftery, A. E., Painter, I. S. and Volinsky, Ch. T. (2005). BMA: An R package for Bayesian model averaging. *R news.* 5/2:2-8.

32. Robert, Ch. (2008). Misconceptions on Bayesianism. *The ISBA Bulletin*, 15/4:2-3.

# Štatistické riadenie kvality – Plány výberového skúmania. Skupinové náhodné vyberanie
# Statistical Quality Control - Sampling designs. Cluster random sampling

**Ľubica Hrnčiarová**

**Abstract:** Sampling is an integral part of statistical quality control. The author's attention in this paper will be focused on cluster random sampling**.**

**Key words:** Sample survey, sampling design, finite population, a sample, the point estimate and confidence interval for population mean population total, determination of sample size.

**Kľúčové slová:** Výberové skúmanie, plán výberového skúmania, konečný základný súbor, výberový súbor, bodový a intervalový odhad strednej hodnoty a úhrnu v základnom súbore, stanovenie rozsahu výberu.

## 1 Úvod

Primárnym cieľom vyberania sú úsudky o parametroch základného súboru (strednej hodnote, úhrnu, podielu a rozptylu) na základe informácií z výberového súboru vybratého zo základného súboru. Pri odhadoch parametrov je potrebné zhromaždiť údaje zo základného súboru, ktorý sa stáva predmetom nášho skúmania. Každý jeden údaj (jednotka) vo výberovom súbore poskytne informáciu o skúmanom parametri základného súboru. Príliš malý výberový súbor nemusí poskytovať dostatok informácií pre získanie presných odhadov a príliš rozsiahly výber môže mať za následok mrhanie finančnými prostriedkami. Pri výberovom skúmaní je veľmi dôležité stanoviť aj spôsob vyberania jednotiek, t.j. určiť plán výberového skúmania (sample design). Rozsah súboru je obyčajne stanovený stupňom požadovanej presnosti odhadov a finančnými obmedzeniami.

Najviac používaným plánom výberového skúmanie je *jednoduché náhodné vyberanie*. Tento plán pozostáva z vyberania *n* (rozsah výberu) výberových jednotiek takým spôsobom, že každá jednotka má rovnakú šancu (pravdepodobnosť) byť vybratá. Ak základný súbor je konečný rozsahu *N*, potom jednoduché náhodné vyberanie môže byť definované ako vyberanie *n* výberových jednotiek, pričom každý výber rozsahu *n* vybratý z $\binom{N}{n}$ z možných výberov má rovnakú pravdepodobnosť byť vybratý.

Predpokladajme, že chceme vytvoriť výberový súbor zo súčiastok vyrobených v danom podniku počas jednej zmeny. V tomto prípade najvhodnejším výberovým plánom bude jednoduché náhodné vyberanie.

Ďalším plánom výberového skúmania je *stratifikované náhodné vyberanie*. Týmto plánom výberového skúmania možno pri rovnakých nákladoch ako pri jednoduchom náhodnom vyberaní získať presnejšie odhady. Stratifikovaný náhodný výber je ale vhodnejší vtedy, keď základný súbor môžeme rozdeliť do viacerých nehomogénnych skupín (strat) tak, že výberové jednotky v každom state sú podobné, ale za jednotlivé statá sa medzi sebou líšia. Každé stratum vystupuje samostatne ako základný súbor (v rámci celku ako podsúbor) a jednoduchý náhodný výber je robený z týchto podsúborov (strat).V hospodárskej praxi nie je ničím výnimočným, že pri porovnávaní hodnôt skúmaného znaku medzi jednotlivými štátmi, regiónmi, podnikmi a pod. sa zistia rozdiely. Napríklad, zaujímala by nás životnosť určitého výrobku vyrábaného vo viacerých podnikoch. Životnosť výrobkov môže byť

rozdielna. Mohlo by to byť spôsobené napríklad vstupnými surovinami, odbornou úrovňou pracovníkov, technickou úrovňou strojového parku. V takýchto prípadoch základný súbor je nehomogénny (heterogénny) a presnejšie odhady parametrov základného súboru získame, keď použijeme plán výberového skúmania s názvom stratifikované náhodné vyberanie. V tomto prípade jednotlivé podniky predstavujú stratá. Iný príklad stratifikovaného náhodného vyberania je vo výrobe, keď výberové súbory tvoria výrobky vyrobené v rozdielnych dávkach . V tomto prípade výrobky vyrobené v rozdielnych dávkach predstavujú stratá.

Výhody stratifikovaného náhodného vyberania oproti jednoduchému náhodnému vyberaniu sú napríklad nasledujúce:

1. Pri rovnakom rozsahu výberu poskytuje presnejší odhad parametrov základného súboru.
2. Stratifikovaný výber je hospodárnejší, čo môže mať za následok zníženie nákladov na vyberanie.
3. V každom strate, ktoré sú homogénne, sa realizuje náhodné vyberanie. Tieto výbery sa preto môžu vhodne využiť na skúmanie každého strata samostatne, bez vynaloženia extra nákladov.

Pri *skupinovom náhodnom vyberaní* je výberovou jednotkou skupina (cluster) niekoľkých pôvodných výberových jednotiek(čiastkových jednotiek) zo základného súboru. Pri skupinovom náhodnom vyberaní vyberieme jednoduchým náhodným výberom skupiny. Môžeme povedať, že pri skupinovom vyberaní sa vyberá v „balíkoch", ktoré budeme nazývať skupiny (clusters). V prostredí výroby tento spôsob vyberania je obzvlášť vhodný, pretože nie je veľmi jednoduché zostaviť prehľad (zoznam) jednotlivých pôvodných výberových jednotiek základného súboru. Na druhej strane môže byť jednoduchšie zostaviť prehľad skupín, kde každá skupina obsahuje väčší počet čiastkových jednotiek. Skupinové náhodné vyberanie je vlastne jednoduché náhodné vyberanie týchto skupín. Výhodou skupinového náhodného vyberania je, že vyberaním jednoduchým náhodným výberom iba málo skupín môžeme v skutočnosti získať celkom rozsiahly výber čiastkových jednotiek za tieto skupiny, a to s minimálnymi nákladmi. Skupinové vyberanie nie je iba cenovo výhodné, ale tiež časovo výhodnejšie pretože zbieranie údajov susediacich jednotiek je lacnejšie, jednoduchšie a rýchlejšie ako jednotiek, ktoré sú navzájom vzdialené. Napríklad, môže byť oveľa jednoduchšie a lacnejšie, keď sa náhodne vyberú „balíky“ súčiastok a nie jednotlivé súčiastky.

Štvrtou metódou je *systematické náhodné vyberanie*. Tento spôsob vyberania je najjednoduchší a hlavne vhodný vo výrobných procesoch, keď vyberanie je realizované v reálnom čase (on line). Pri tejto metóde je prvá jednotka vybratá náhodne a od tejto jednotky potom každá *k*- tá jednotka ($k = N/n$), až kým dosiahneme výber požadovaného rozsahu *n*. Systematické náhodné vyberanie nie je iba jednoduché, čo sa týka vyberania, ale za istých podmienok presnejšie ako jednoduché náhodné vyberanie.

V ďalšom texte príspevku je popísaný postup skupinového náhodného vyberania, bodový a intervalový odhad strednej hodnoty, úhrnu v základnom súbore a určenie minimálneho rozsahu výberu. V závere príspevku je uvedený príklad odhadu strednej hodnoty a úhrnu v základnom súbore.

## 2 Skupinové náhodné vyberanie

Vytvoriť dobrú výberovú bázu je niekedy veľmi obtiažne. Buď pre tento druh výdavku nie je dostatok peňažných prostriedkov alebo v niektorých prípadoch výberové jednotky môžu byť roztrúsené, a tak zisťovanie každej výberovej jednotky je nielen finančne, ale aj časovo náročné. V týchto prípadoch zostavíme výberovú bázu z rozsiahlejších výberových jednotiek (skupín) tak, že každá skupina sa skladá z niekoľkých pôvodných výberových jednotiek

(čiastkových jednotiek) základného súboru. Potom vyberieme jednoduchým náhodným výberom skupiny.

Môžeme povedať, že pri skupinovom vyberaní sa vyberá v „balíkoch", ktoré budeme nazývať skupiny (clusters). Táto technika vyberania je známa pod názvom *plán skupinového náhodného vyberania* alebo jednoducho *plán skupinového vyberania*.

Poznáme jednostupňový a dvojstupňový skupinový výber. V *jednostupňovom skupinovom výbere* skúmame všetky výberové jednotky vo vybratých skupinách. Pri *dvojstupňovom skupinovom vyberaní* skúmame iba časť výberových jednotiek, ktoré sú vybraté z každej vybratej skupiny. Okrem toho pri skupinovom vyberaní rozsah skupín môže, ale nemusí byť rovnaký. Spravidla, výbery v rámci územného celku nie sú realizovateľné za skupiny rovnakého rozsahu. Napríklad, pri výberovom skúmaní rozsiahlej veľkomestskej aglomerácie môžeme mestské časti považovať za skupiny. Ak výberovými čiastkovými jednotkami sú domácnosti alebo osoby, potom obyčajne v jednotlivých mestských častiach ich počet bude rozdielny. Naopak, pri výberoch v rámci priemyselnej výroby môžeme vždy mať skupiny rovnakého rozsahu. Napríklad, za skupiny sa môžu považovať balíky, ktoré obsahujú rovnaký počet súčiastok.

## 2.1 Odhad strednej hodnoty a úhrnu v konečnom základnom súbore

Nech $N$ je rozsah skupín vo výberovom základnom súbore, pričom $i$- tá skupina má $m_i$ výberových čiastkových jednotiek. Máme jednoduchý náhodný výber $n$ skupín. Nech $x_{ij}$ sú zistené hodnoty skúmanej premennej $j$- tej čiastkovej jednotky v $i$- tej skupine, $j$=1,2,..., $m_i$; $i$=1, 2,…,$n$. Dostaneme nasledujúce výrazy:

suma všetkých hodnôt v $i$- tej skupine

$$t_i = \sum_{j=1}^{m_i} x_{ij} \ ,$$

celkový počet čiastkových jednotiek vo výbere

$$m = \sum_{i=1}^{n} m_i \ ,$$

priemerný rozsah skupiny vo výbere

$$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i = \frac{m}{n},$$

celkový počet čiastkových jednotiek v základnom súbore

$$M = \sum_{i=1}^{N} m_i$$

a priemerný rozsah skupiny v základnom súbore

$$\overline{M} = \frac{1}{N}\sum_{i=1}^{N} m_i = \frac{M}{N}.$$

Potom **odhadom strednej hodnoty základného súboru $\mu_K$ je**

$$\overline{X}_{sk} = \frac{\sum_{i=1}^{n} T_i}{m}. \qquad (1)$$

Odhadom rozptylu $\overline{X}_{sk}$ je

$$\hat{D}\left(\overline{X}_{sk}\right) = \left(\frac{N-n}{Nn\overline{M}^2}\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(t_i - \overline{X}_{sk}\cdot m_i\right)^2\right) \qquad (2)$$

Ak priemerný rozsah skupiny v základnom súbore $\overline{M}$ nepoznáme, nahradíme ho odhadom $\overline{m}$. Potom

$$\hat{D}\left(\overline{X}_{sk}\right)=\left(\frac{N-n}{Nn\overline{m}^2}\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(t_i-\overline{X}_{sk}\cdot m_i\right)^2\right) \tag{3}$$

**Odhadom úhrnu $\tau_K$ základného súboru je**

$$\hat{\tau}_{sk}=M\overline{X}_{sk}=\frac{M}{m}\sum_{i=1}^{n}T_i\,. \tag{4}$$

Odhadom rozptylu $\hat{\tau}_{sk}$ je

$$\hat{D}(\hat{\tau}_{sk})=M^2\hat{D}\left(\overline{X}_{sk}\right)=N^2\overline{M}^2\hat{D}\left(\overline{X}_{sk}\right) \tag{5}$$

Keď dosadíme hodnotu $\hat{D}\left(\overline{X}_{sk}\right)$ z výrazu (3) dostaneme

$$\hat{D}(\hat{\tau}_{sk})=N\left(\frac{N}{n}-1\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(T_i-\overline{X}_{sk}\cdot m_i\right)^2\right) \tag{6}$$

S pravdepodobnosťou $(1-\alpha)$ prípustná chyba

- odhadu strednej hodnoty v základnom súbore je

$$\pm u_{1-\alpha/2}\sqrt{\hat{D}\left(\overline{X}_{sk}\right)}=\pm u_{1-\alpha/2}\sqrt{\left(\frac{N-n}{Nn\overline{m}^2}\right)\left(\frac{1}{n-1}\left(\sum_{i=1}^{n}T_i-\overline{X}_{sk}\cdot m_i\right)^2\right)} \tag{7}$$

- odhadu úhrnu v základnom súbore je

$$\pm u_{1-\alpha/2}\sqrt{\hat{D}(\hat{\tau}_{sk})}=\pm u_{1-\alpha/2}\sqrt{N\left(\frac{N}{n}-1\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(T_i-\overline{X}_{sk}\cdot m_i\right)^2\right)} \tag{8}$$

Interval spoľahlivosti pre strednú hodnotu v základnom súbore je

$$\overline{X}_{sk}\pm u_{1-\alpha/2}\sqrt{\left(\frac{N-n}{Nn\overline{m}^2}\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(t_i-\overline{X}_{sk}\cdot m_i\right)^2\right)} \tag{9}$$

Interval spoľahlivosti pre úhrn v základnom súbore je

$$\hat{\tau}_{sk}\pm u_{1-\alpha/2}\sqrt{N\left(\frac{N}{n}-1\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(t_i-\overline{X}_{sk}\cdot m_i\right)^2\right)} \tag{10}$$

V ďalšom texte uvediem príklad odhadu strednej hodnoty a úhrnu nákladov vynaložených na opravu hydraulického čerpadla.

Manažéra kvality spoločnosti, ktorá vyrába hydraulické čerpadlá zaujímajú ročné náklady na záručné opravy u daného typu hydraulického čerpadla. Spoločnosť inštaluje daný typ čerpadla v 6- tich aplikáciách, a to v prevádzkach poskytujúcich potravinárke služby (automaty na nápoje), prevádzkach mliekarní, prevádzkach na plnenie fliaš s nealkoholickými nápojmi, prevádzkach pivovarov, pri spracovaní odpadových vôd a pri odvoze výkalov zo žúmp. Manažér kvality vie zistiť za jednotlivé prevádzky celkový počet čerpadiel daného typu, ale nevie zistiť náklady na záručné opravy za každé čerpadlo. Za jednotlivé prevádzky, prostredníctvom údajov o škodových udalostiach však vie zistiť pre daný typ čerpadla celkové ročné náklady na záručné opravy. V tomto prípade sa manažér kvality rozhodne pre skupinové vyberanie, pričom skupinou bude prevádzka. Náhodne vyberie $n$ = 10 z $N$ = 120 prevádzok, ktoré používajú sledovaný typ hydraulického čerpadla. V tabuľke 1 sú za

jednotlivé prevádzky údaje o nákladoch na opravy za jeden sledovaný rok počas záručnej doby a počet čerpadiel, ktoré vlastnia.

**Tabuľka 1** Počet hydraulických čerpadiel a celkové ročné náklady spoločnosti počas záručnej doby

| Výber<br>$i$ | Počet čerpadiel (ks)<br>$m_i$ (v ks) | Celkové ročné náklady počas záručnej doby (EUR)<br>$\sum_{j=1}^{m_i} x_{ij} = t_i$ | $\sum_{i=1}^{n} \left(t_i - \bar{x}_{sk} \cdot m_i\right)^2$ |
|---|---|---|---|
| 1 | 6 | 338 | 1523,23 |
| 2 | 9 | 489 | 1643,724 |
| 3 | 6 | 245 | 2912,914 |
| 4 | 3 | 180 | 931,1219 |
| 5 | 9 | 440 | 71,52305 |
| 6 | 11 | 519 | 847,6407 |
| 7 | 5 | 310 | 3703,593 |
| 8 | 8 | 403 | 19,10949 |
| 9 | 9 | 321 | 16245,32 |
| 10 | 4 | 243 | 1908,442 |
| $\sum_{i=1}^{n}$ | 70 | 3488 | 29806,62 |

Na základe údajov z tab. 1. dostaneme nasledujúce výpočty:

$$m = \sum_{i=1}^{n} m_i = 70 \qquad \bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i = \frac{m}{n} = \frac{70}{10} = 7$$

Za sledovaný rok spoločnosť vyrábajúca hydraulické čerpadlá vynaložila priemerne na opravu jedného čerpadla daného typu

$$\bar{X}_{sk} = \frac{\sum_{i=1}^{n} t_i}{m} = \frac{3488}{70} = 49{,}82857 = 49{,}83 \;\; EUR .$$

Pretože nemá informácie o celkovom počte čiastkových jednotiek $M$ v základnom súbore, odhadne ich počet

$$\hat{M} = N \cdot \hat{\bar{M}} = N \cdot \bar{m} = 120 \cdot 7 = 840 .$$

Celkové náklady spoločnosti na opravu daného typu čerpadla počas jedného sledovaného roku záručnej doby boli

$$\hat{\tau}_{sk} = \hat{M}\,\bar{X}_{sk} = 840 \times 49{,}82857 = 41\;856\; EUR .$$

95% interval spoľahlivosti pre strednú hodnotu nákladov v základnom súbore je

$$\bar{X}_{sk} \pm u_{1-\alpha/2}\sqrt{\left(\frac{N-n}{Nn\bar{m}^2}\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(t_i - \bar{X}_{sk}\cdot m_i\right)^2\right)} = 49{,}82857 \pm 1{,}96\sqrt{\left(\frac{120-10}{120\times 10\times 7^2}\right)\left(\frac{1}{9}29806{,}62\right)} =$$

$$= 49{,}82857 \pm 4{,}87864 = \left(44{,}95\; EUR;\;\; 54{,}71\; EUR\right)$$

95% interval spoľahlivosti pre celkové náklady v základnom súbore je

$$\hat{\tau}_{sk} \pm u_{1-\alpha/2}\sqrt{N\left(\frac{N}{n}-1\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(t_i-\overline{X}_{sk}m_i\right)^2\right)} = 4\ 1856 \pm 1{,}96\sqrt{120\left(\frac{120}{10}-1\right)\left(\frac{1}{9}29806{,}62\right)} =$$
$$= 41\ 856 \pm 4\ 098{,}05844 = \left(37\ 757{,}94\ EUR;\quad 45\ 954{,}06\ EUR\right)$$

**2.2 Určenie rozsahu výberu**

Rozsah výberu *n* potrebný na $(1-\alpha)$% odhad strednej hodnoty základného súboru s maximálnou prípustnou chybou odhadu $\Delta$ je

$$n \geq \frac{u_{1-\alpha/2}Ns^2}{N\overline{M}^2\Delta^2 + u_{1-\alpha/2}^2 s^2}, \qquad (11)$$

kde $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(t_i - \overline{X}_{sk}\cdot m_i\right)^2\right)$.

Rozsah výberu potrebný na $(1-\alpha)$% odhad úhrnu základného súboru s maximálnou prípustnou chybou odhadu $\Delta$ sa vypočíta

$$n \geq \frac{u_{1-\alpha/2}Ns^2}{N\overline{M}^2\Delta^2/M^2 + u_{1-\alpha/2}^2 s^2} \qquad (12)$$

## 3 Záver

*Skupinové vyberanie* nie je iba finančne, ale tiež časovo výhodnejšie pretože zhromažďovanie údajov za susediace jednotky je lacnejšie, jednoduchšie a rýchlejšie. Pri realizácii skupinového výberu je vhodné, aby skupiny neboli príliš veľké, aby ich bol dostatočný počet a jednotky, ktoré obsahujú, boli z hľadiska skúmaného znaku čo najviac heterogénne. Vtedy sa hovorí, že skupiny sú účinné.

Tento príspevok vznikol s prispením grantovej agentúry VEGA v rámci projektu číslo 1/0437/08 Kvantatitatívne metódy v stratégii šesť sigma.

## Literatúra

[1] GUPTA,B.C. – WALKER, H. F. 2007. Statistical Quality Control for the Six Sigma Green Belt. Milwaukee 53203:ASQ, Quality Press, 2007, 340s.,H 1277.
[2] LOHR, S. L. 1999: Sampling. Desing and Analysis. USA.: Brooks/Cole Publishing Company, A division od International Thomson Publishing Inc., 1999, 494s., ISBN 0-534-35361-4.
[3] ČERMÁK, V.- VRABEC, M. 2008. Teorie výběrových šetření, Část 3. Praha: Vysoká škola ekonomická, 116s.,1999, ISBN 80-245-0003-5.
[4] TEREK M. - HRNČIAROVÁ Ľ. 2008. Výberové skúmanie. Bratislava: Ekonóm, 2008, 108s., ISBN 978-80-225-2440-7.

## Adresa autora:

Ľubica Hrnčiarová, doc. Ing. PhD.
Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky
Dolnozemská cesta 1/a,
852 35 Bratislava
lubica.hrnciarova@euba.sk

# Štatistika v Exceli verzie 2007
# Statistics in Excel version 2007

Jozef Chajdiak

**Abstract:** Program system Excel version 2007 present syntactically new version tabular data processor type Excel. Allowance includes information about new form jobs and decision solution statistical work in Excel version 2007.

Programový systém Excel verzie 2007 predstavuje syntakticky novú verziu tabuľkových procesorov typu Excel. Príspevok obsahuje informáciu o nových formách práce a možnostiach riešenia štatistických úloh v Exceli verzie 2007.

**Kľúčové slová:** Excel, štatistické metódy, Kontingenčná tabuľka (Pivot Table), štatistické funkcie, štatistické nástroje, použitie makrojazyka

**Key words:** Excel, statistical methods, Pivot Table, statistical functions, statistical tools, using macro language

## Úvod

Podrobnejšiu informáciu možno získať v autorovej publikácii Chajdiak J. (2009): Štatistika v Exceli 2007. STATIS, Bratislava, ISBN 978-80-85659-49-8, 304 strán A5. Kniha nadväzuje na predchádzajúce publikácie: Chajdiak, J. (2005): Štatistické úlohy a ich riešenie v Exceli. STATIS, Bratislava, ISBN 80- 85659-39-5 resp. Chajdiak, J. (2002): Štatistika v Exceli. STATIS, Bratislava, ISBN 80- 85659-27-1. Dôležitým aspektom práce s uvedenými publikáciami je možnosť stiahnuť si používané súbory údajov v ilustratívnych príkladoch v knihách z webovskej stránky vydavateľa knižiek (podrobnosti - str. 34 knihy).

## 1 VŠEOBECNÉ POZNÁMKY

Od čitateľa sa vyžaduje základná znalosť Excelu ako je zápis hodnôt do políčok, ich kopírovanie, špeciálne kopírovanie, presúvanie, úprava na požadovaný tvar a pod. Výklad je sústredený na päť hlavných spôsobov štatistickej práce:

- použitie funkcií,
- použitie nástrojov,
- použitie podsystému kontingenčnej tabuľky (PivotTable),
- použitie vlastných formúl,
- použitie makier.

### 1.1 Príprava údajov na štatistickú analýzu

Riešenie štatistických úloh v Exceli nie je zložité. Vypočítať výsledky úloh vyžaduje:

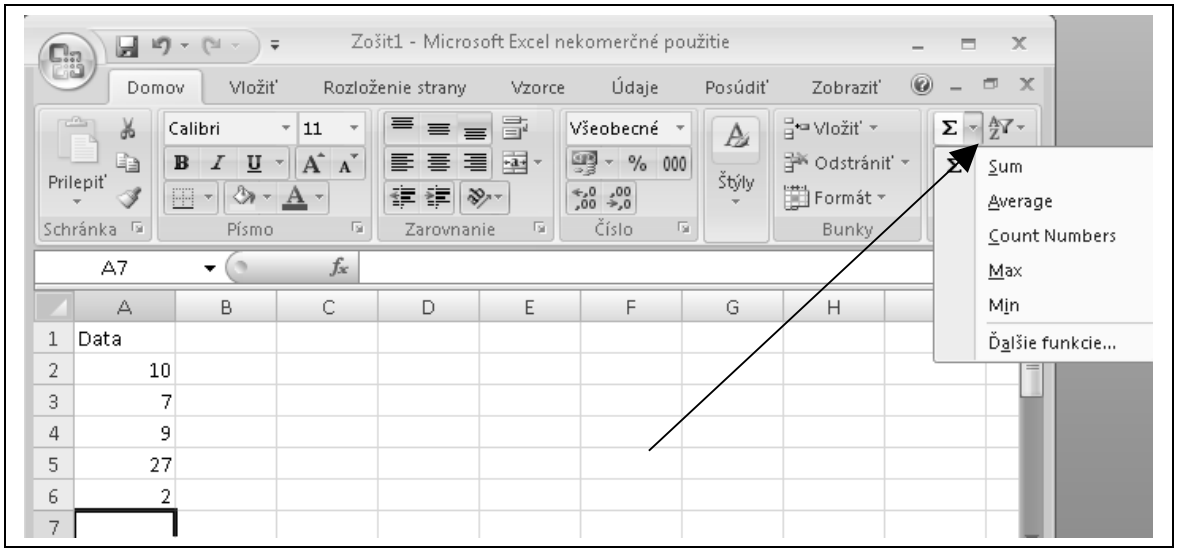
- sformulovať tieto úlohy,
- zozbierať potrebné údaje,
- využiť výpočtový štatistický aparát Excelu,
- vedieť prečítať výsledky na výstupe,
- vedieť interpretovať a využiť získané výsledky.

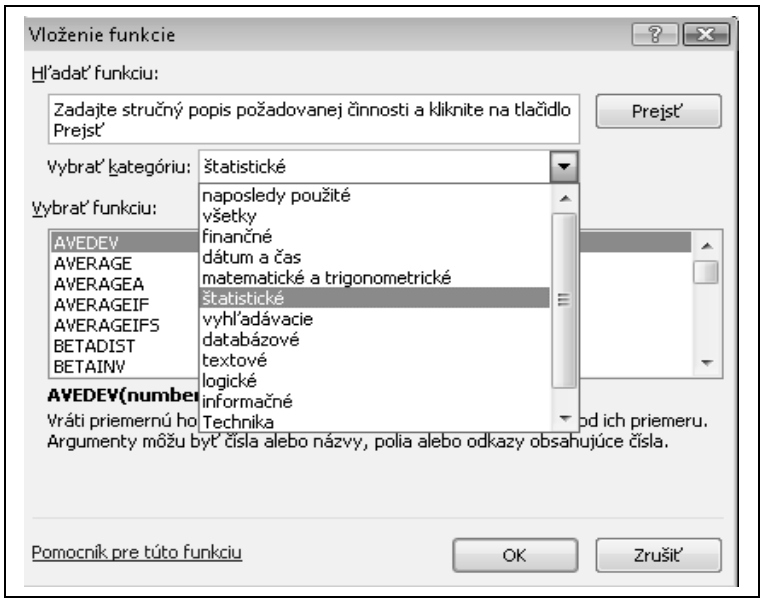
Opíšeme si všeobecné použitie štatistického aparátu.

### 1.1.1 Okno funkcie

Excel obsahuje bohatú množinu funkcií. Funkcie sú členené do skupín podľa ponuky okienka *Kategória funkcie:* Na štatistickú analýzu sa využíva hlavne kategória *štatistické*, v časti prípadov možno využiť kategóriu *matematické* a vo všeobecnej práci v Exceli hociktorú ďalšiu kategóriu.

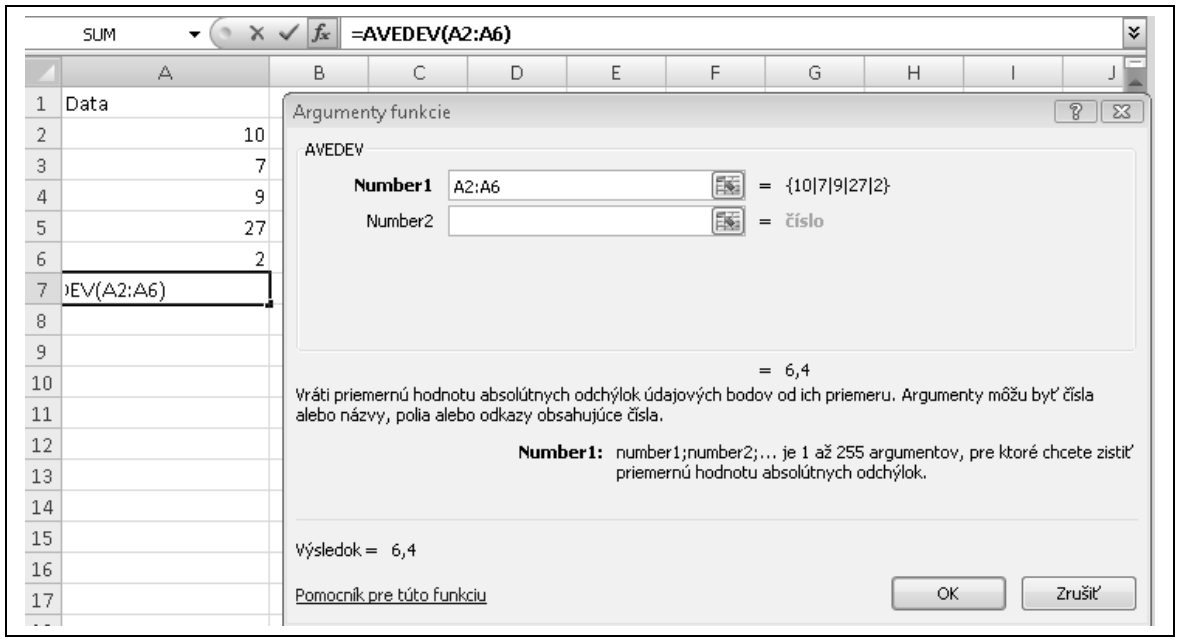
Funkciu môžeme zapísať priamo do políčka tabuľky alebo môžeme kurzor nastaviť na požadované políčko tabuľky a v časti *Domov/Úpravy* ťuknúť na ikonu Σ čím sa aktivuje okno *Prilepiť funkciu*. Ťuknutím na časť trojuholníka sa aktivuje ponuka funkcií (ikona Σ je v hornom riadku ikon v pravej časti na prvom obrázku):



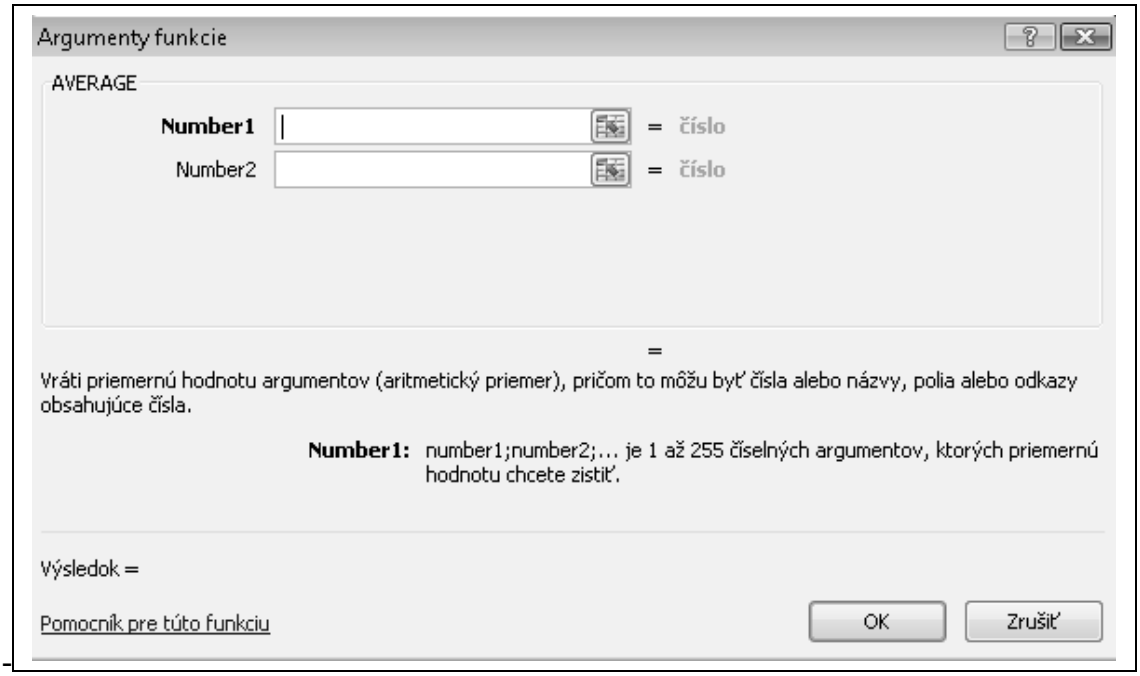




Vysvietime požadovanú funkciu priamo (úhrn, priemer, počet, maximum alebo minimum) alebo cez ťuknutie na *Ďalšie funkcie* (*More Functions...*) zvolíme kategóriu a požadovaný názov funkcie. V slovenskej verzii sú kategórie funkcií preložené do slovenčiny, konkrétne názvy funkcií sa neprekladali a sú zhodné z anglickými názvami (v českej verzii sú preložené do češtiny aj názvy funkcii). V okne *Vloženie funkcie* okrem okienka *Vybrať kategóriu:* a okienka *Vybrať funkciu* je v spodnej časti aj tvar všeobecného zápisu konkrétnej funkcie, ktorá je vysvietená v okienkach *Vybrať kategóriu:* a *Vybrať*

*funkciu:*. Ďalej je tu stručný opis čo daná funkcia robí. V spodnom riadku v rámčeku ťuknutím na *Pomocník pre túto funkciu* môžeme aktivovať pomoc (help) pre danú časť. Okno obsahuje tiež tlačidla *OK* a *Zrušiť* (*Cancel*). Po ťuknutí na tlačidlo *OK* sa aktivuje okno príslušnej vysvietenej funkcie.




Tvar každého okna Argumenty funkcie je dosť podobný.

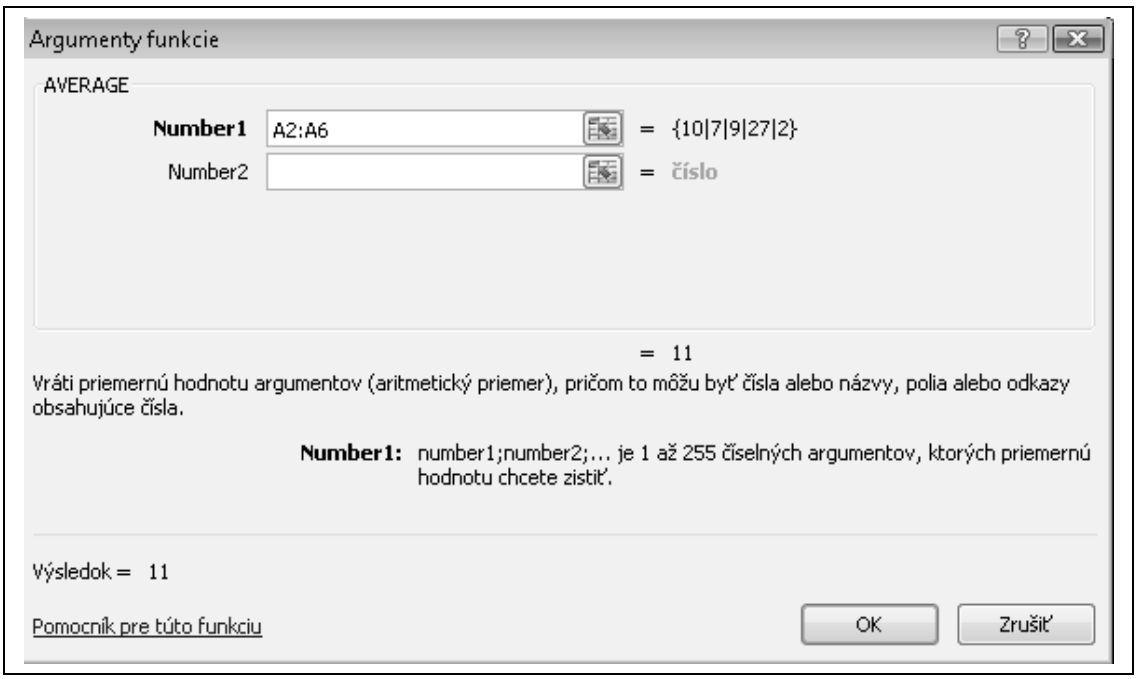



V rámčeku v hornej časti, v ľavom hornom roku je zobrazený názov funkcie. Na ukážke aktivovanej funkcie priemeru je to *AVERAGE*. V rámčeku sa špecifikujú oblasti údajov, ktoré sa použijú na výpočet funkcie a ďalšie skutočnosti.

Zápis týchto oblastí je dvojakým spôsobom. V prvom sa do okienka priamo napíše oblasť, kde sa údaje nachádzajú (napríklad: *A2.A6*). Druhý spôsob predstavuje ťuknúť na štvorček v pravom okraji okienka. Zobrazí sa pomocné okienko, do ktorého môžeme priamo písať identifikáciu oblasti s požadovanými údajmi alebo myšou vyznačiť oblasť priamo v tabuľke (automaticky sa vypíše v pomocnom okienku) a stlačiť *Enter*. Po zaplnení poľa *Number1* sa na konci riadku za okienkom zobrazí zoznam čísiel z oblasti (resp. prvá časť

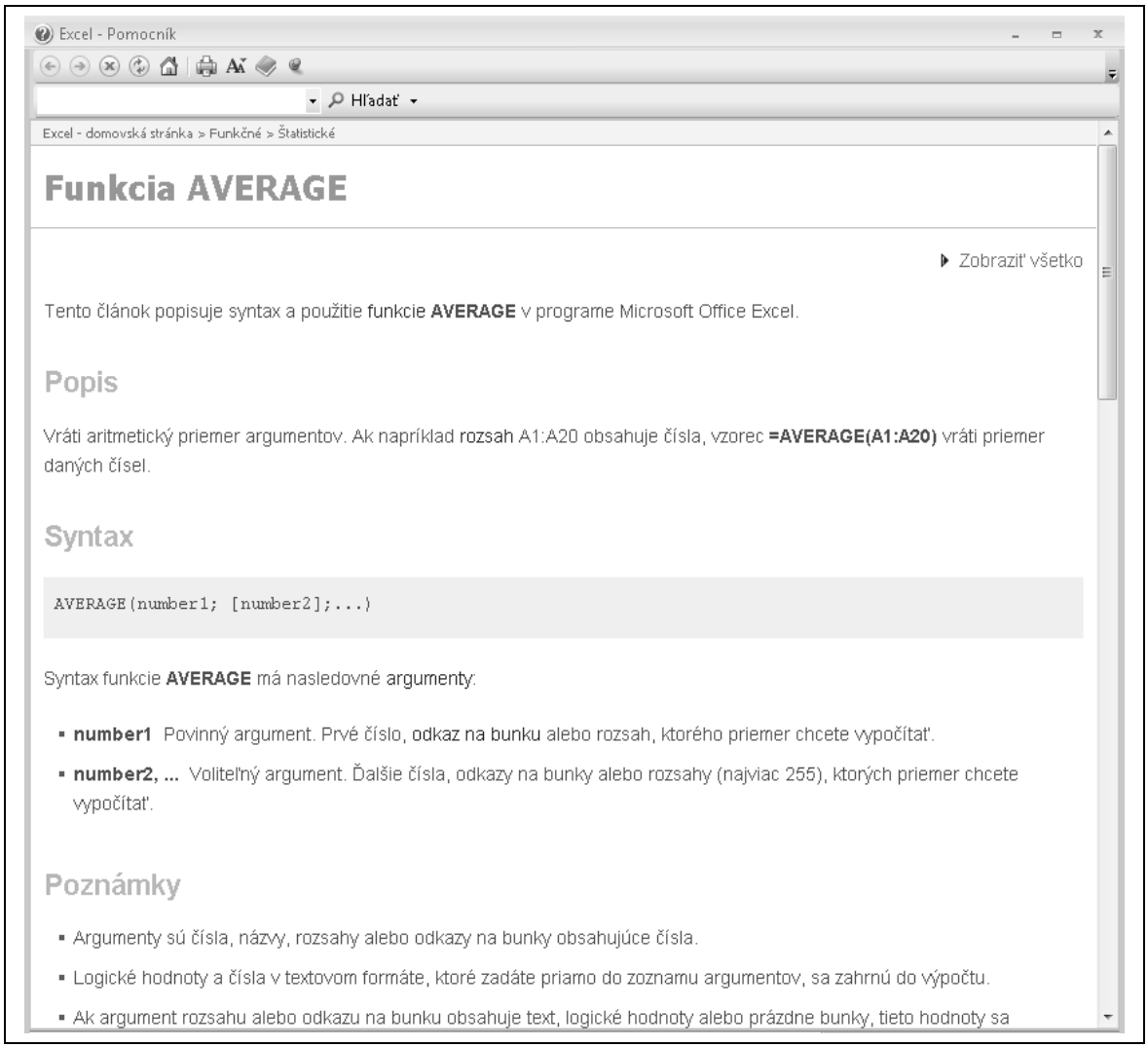
týchto čísiel až do konca rámčeka. Vo všeobecnosti môžu byť v takomto hornom rámčeku aj iné okienka z ďalšími parametrami potrebnými na výpočet funkcie. To okienko, ktoré je aktívne, má v dolnej časti vypísaný text, čo má obsahovať.

Hneď pod rámčekom je text opisujúci čo aktivovaná funkcia robí resp. vypočíta. Priebežný výsledok funkcie je zobrazený v spodnej časti za textom *Výsledok* = (*Formula result* =) a hneď pod rámčekom za znakom = .




Máme zobrazené dva rámčeky *AVERAGE* (priemer). Prvý je prázdny a v druhom je špecifikovaná oblasť A2 až A6, v ktorej je množina čísiel. Ich priemer sa rovná 11.

V okne funkcie máme tiež *Pomocník pre túto funkciu* (*Help on this function*). Po ťuknutí na tento text sa v pravej časti obrazovky zobrazí podrobný štatistický opis príslušnej funkcie. Napríklad pre funkciu *AVERAGE* (priemer) je to nasledujúci text (tu je zobrazená len horná časť):



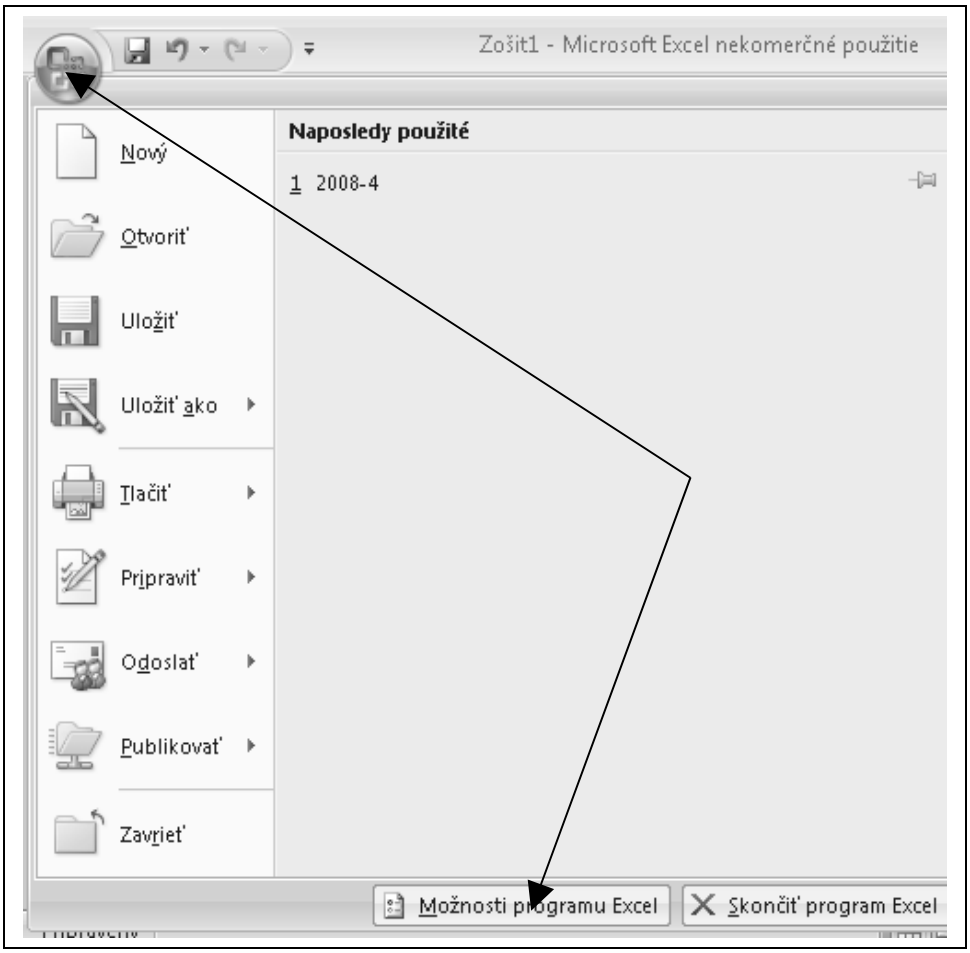

Po vyplnení všetkých povinných okienok ťukneme na tlačidlo *OK* a Excel do políčka, v ktorom je daná funkcia prenesie jej výsledok (tu je to *11*).

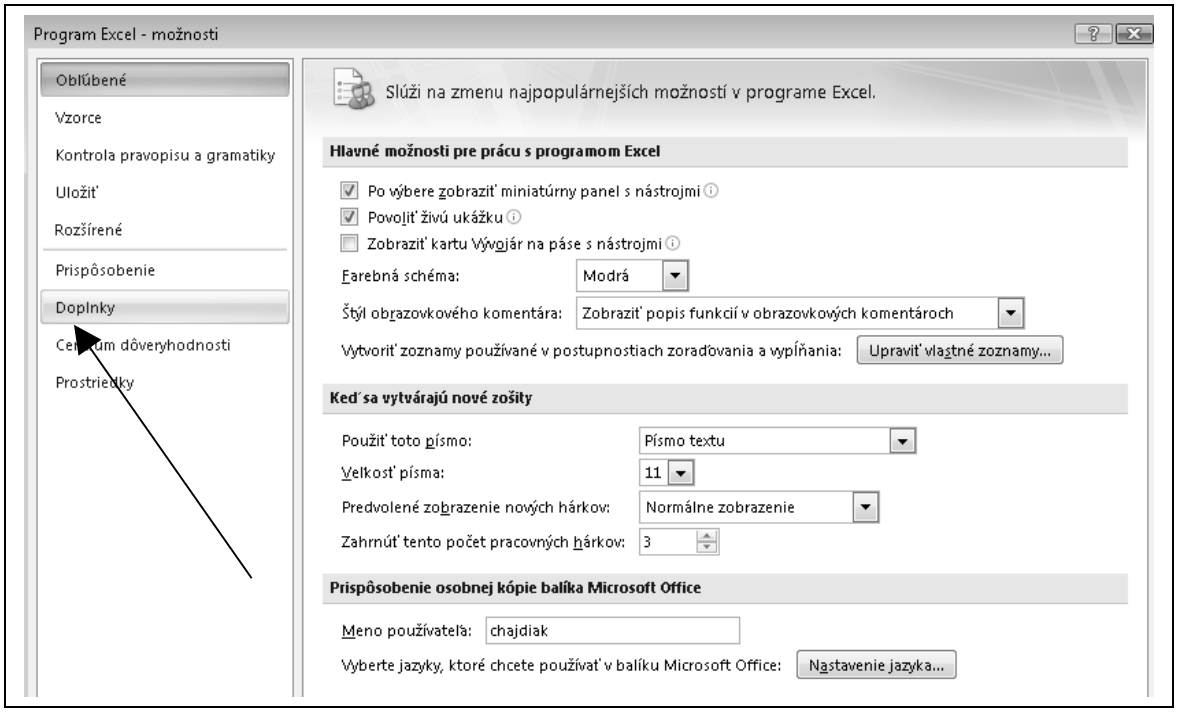
Časť okienok nemusíme vyplniť a Excel predsa vypočíta príslušnú funkciu. Čitateľovi odporúčame pozrieť, akú konkrétnu hodnotu systém dosadil.

### 1.1.2 Okno nástroja

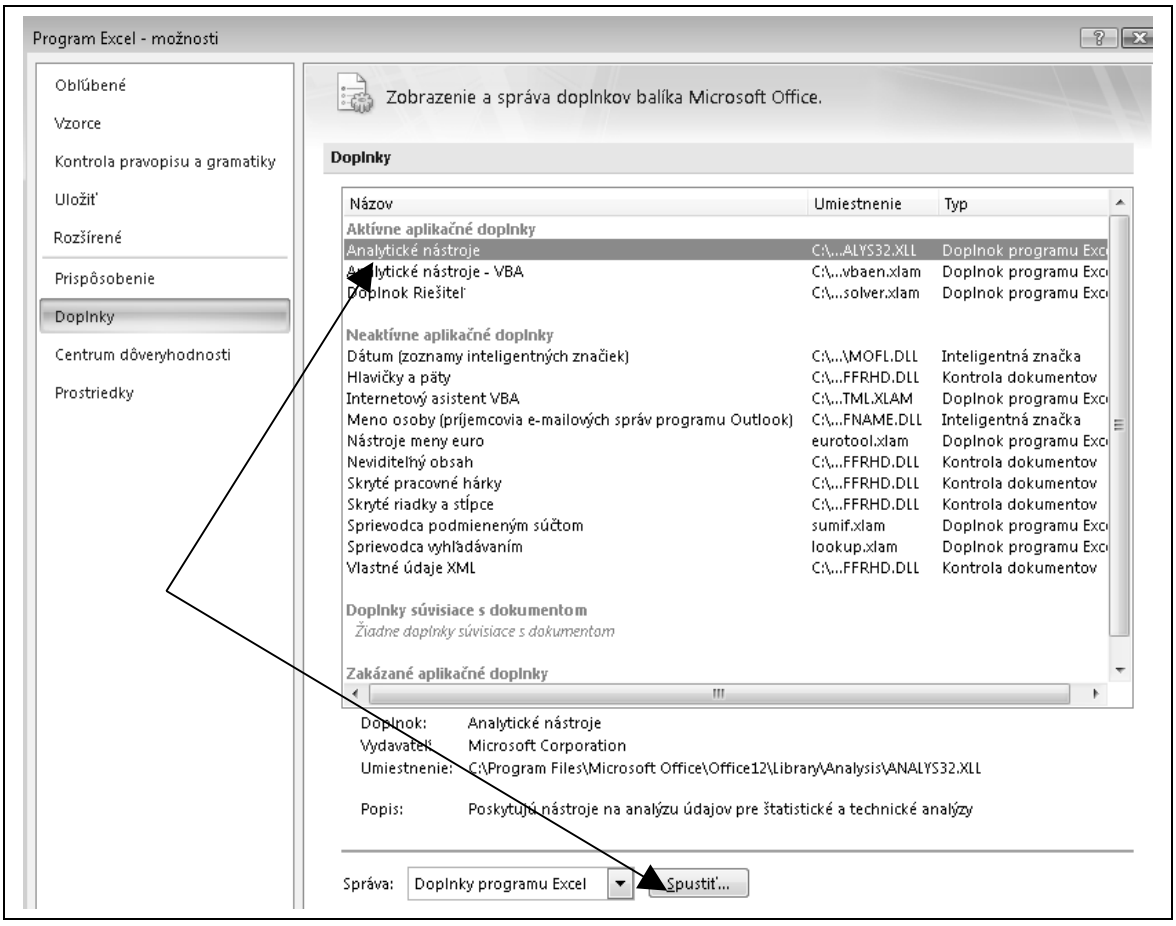
Nástroje na analýzu údajov sú špeciálnou časťou Excelu a pred prvým použitím ich treba doinštalovať. V závislosti od toho, kde sú údaje z príslušného CD, je vhodné mať toto CD s inštalačným programom Excel k dispozícii.

Na doinštalovanie nástroja *Analýza údajov* v ponuke ťukneme ústredné *Diamantové tlačidlo (Oficce)*. Objaví sa ústredná ponuka s možnosťami voľby *Nový (New)*, *Otvoriť (Open)* a ďalšími voľbami. V spodnom riadku okna je tlačidlo *Možnosti programu Excel* (*Excel Options*), na ktoré ťukneme myšou.

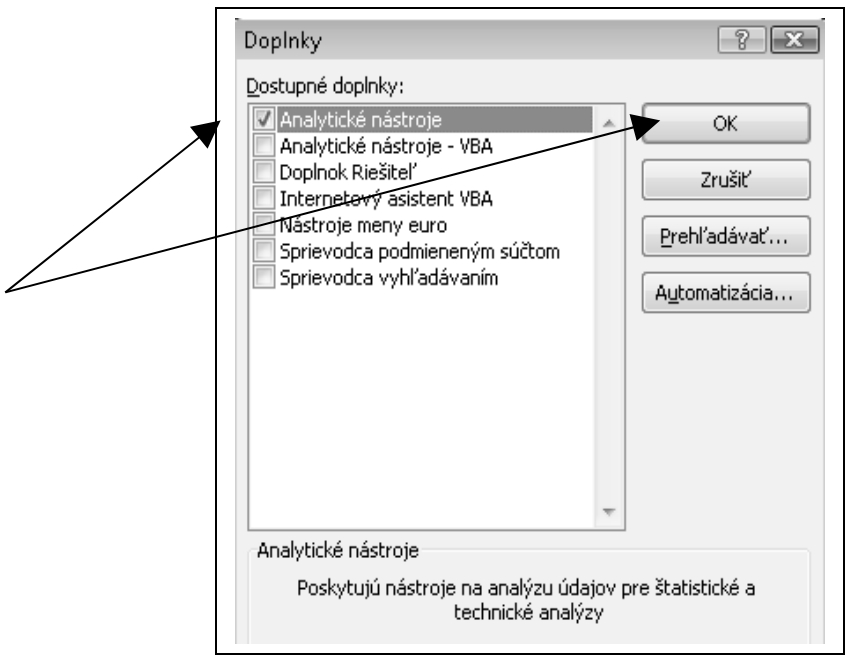



V okne *Program Excel – možnosti* ťukneme na možnosť *Doplnky* (*Add-Ins*) a v okne sa aktivuje blok Doplnky.

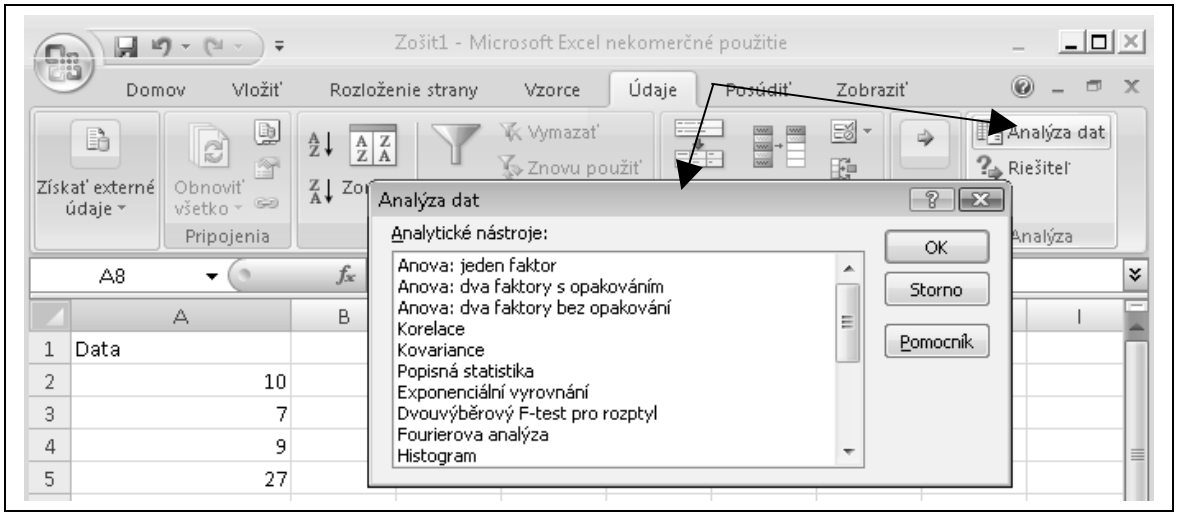



V časti *Neaktívne aplikačné doplnky* (*Inactive Applications Add-Ins*), vysvietime, ktoré doplnky chceme aktivovať: *Analytické nástroje* (*Analysis ToolPak*) a ťukneme na tlačidlo *Spustiť* (*Go...*). Objaví sa okno *Doplnky* ( *Add Ins available.*). V tomto okne odfajkujeme v ponuke *Analytické nástroje* (*Analysis ToolPak*) a ťukneme na tlačidlo *OK*.

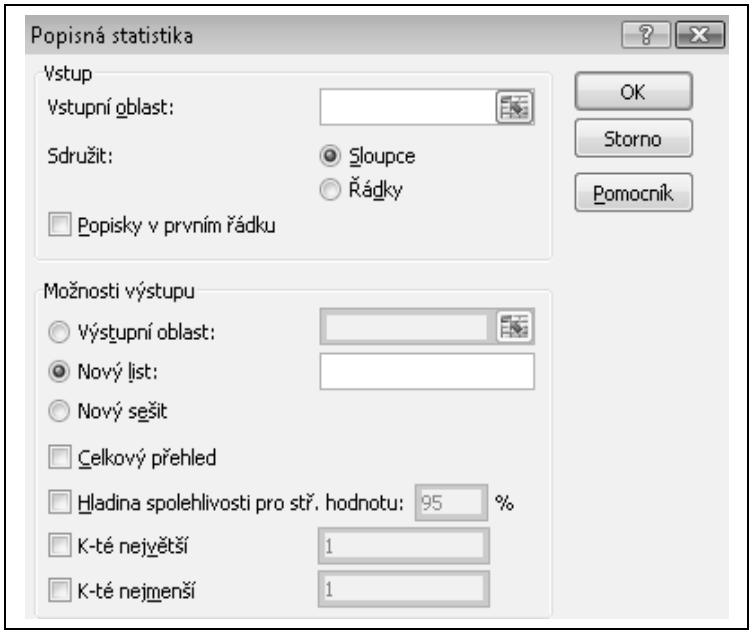


Excel samostatne doinštaluje časť *Analytické nástroje* (v niektorých prípadoch na to môže vyžadovať inštalačné CD s Excelom). Po doinštalovaní je nástroj *Analytické nástroje* trvale k dispozícii.

Aktivácia jednotlivých nástrojov z *Analytických nástrojov* sa realizuje postupnosťou príkazov: *Údaje / Analýza údajov* (*Data /Data Analysis ...*). Po aktivovaní nástroja *Analýza dat...* (*Data Analysis...*) sa objaví okno *Analýza dát* s ponukou jednotlivých *Analytických nástrojov*:




Vysvietime požadovaný z nich a ťukneme na tlačidlo *OK*. Nech sme vysvietili nástroj *Popisná statistika*. Jeho aktivované okno je nasledujúce:




V okne vyplníme potrebné údaje. Vždy musíme vyplniť okienko vstupných údajov *Vstupní oblast* (*Input Range:*). Veľmi často je potrebné špecifikovať, či údaje sú organizované po *stĺpcoch* (*Columns*) alebo *riadkoch* (*Rows*). Odfajkovaním okienka *Popisky v prvním řádku* (*Labels in First Row*) dávame Excelu na vedomie, že prvé políčko vstupnej oblasti obsahuje názov premennej (či iný text).

Druhú časť okna nástroja tvorí obsah spodného rámčeka *Možnosti výstupu* (*Output Options*). V nej, okrem iného, špecifikujeme kam chceme uložiť výstup realizácie príslušného nástroja. Automaticky sa predpokladá *Nový list:* (*New WorkSheet Ply:*) s uložením výstupu na nový hárok. Výstup môžeme tiež uložiť do špecifikovanej oblasti v aktuálnom hárku *Výstupní oblast:* (*Output Range*) alebo do nového súboru *Nový sešit* (*New WorkBook*).

V tretej časti okna máme možnosti požadovať (odfajkovaním) doplňujúce výpočty k základnému balíku výpočtov príslušného nástroja.

Keď máme všetky potrebné okienka vyplnené, ťukneme na realizáciu príslušného nástroja na tlačidlo *OK*.

### 1.1.3 Podsystém Kontingenčná tabuľka (PivotTable)[1]

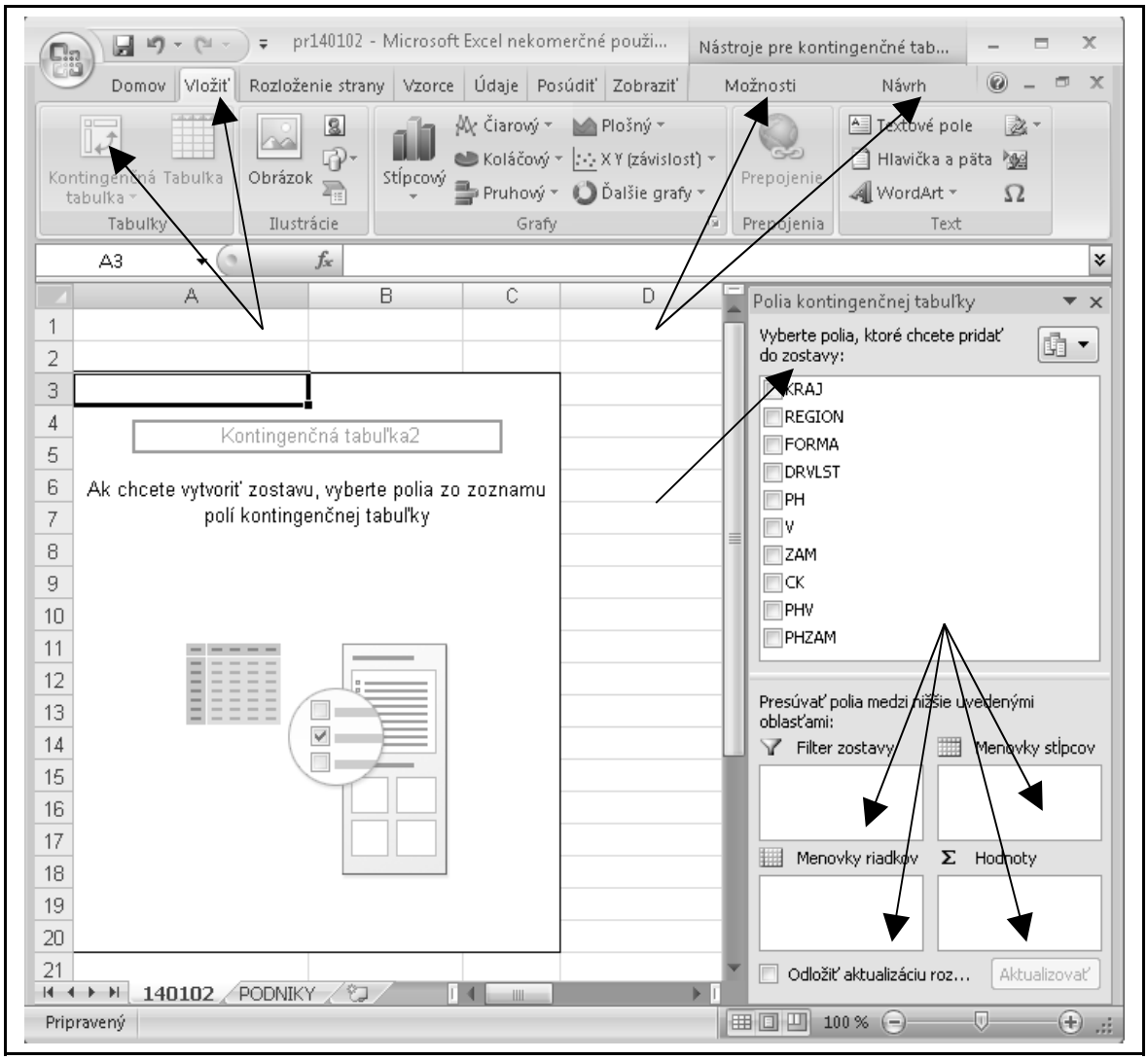
Zvyčajne je výsledkom zisťovania niekoľko premenných a „niekoľko“ môže znamenať aj desiatky, stovky, tisíce, či viac premenných (štatistických znakov) a „niekoľko“ riadkov má v Exceli 2007 limit v jednom milióne riadkov (pozorovaní za jednotlivé štatistické jednotky). Excel poskytuje na praktickú analýzu takýchto súborov pomerne jednoduchý, vysoko produktívny podsystém Kontingenčná tabuľka (Pivot Table). Každá z verzií Excelu (1995, 1998, 2000, 2003, 2007) poskytuje postupne vyššiu kvalitu práce podsystému Kontingenčná tabuľka. Verzia 2007 má navyše nový syntax usporiadania jednotlivých prvkov podsystému.

Klasická verzia práce vyžaduje „čistý“ súbor údajov, t.j. údaje sú usporiadané v tabuľke, pričom v prvom riadku sú uvedené mená premenných a v ďalších riadkoch napozorované hodnoty jednotlivých premenných – jeden riadok = jedno pozorovanie za jednu štatistickú jednotku. Tento súbor údajov je vhodné mať na samostatnom hárku a mať ešte aj záložnú kópiu v záložnom súbore.

Prácu podsystému Kontingenčná tabuľka (PivotTable) aktivujeme postupnosťou krokov: kurzor je v políčku tabuľky; ťukneme tlačidlo *Vložiť* (*Insert*) a ťukneme *Kontingenčná tabuľka* (*PivotTable*). Objaví sa okno *Vytvorenie kontingenčnej tabuľky* (*Choose the data that you want to analyze*). Parametre nastavenia ponecháme v excelom nastavenom tvare a ťukneme na tlačidlo *OK*. Na novom hárku sa objaví pracovná verzia Kontingenčnej tabuľky.

V hlavnej ponuke Excelu sa pod tlačidlom *Nástroje pre kontingenčnú tabuľku* vyskytujú vetvy *Možnosti* a *Návrh*. Vetva *Možnosti* má 8 podmožnosti pre špeciálnejšiu prácu a vetva *Návrh* má 3 podnávrhy.

[1] Originál v angličtine používa označenie *PivotTable*, ktoré bolo do slovenčiny nie plne korektne preložené ako *Kontingenčná tabuľka*. Keď si čitateľ pozrie 14. kapitolu zistí, že PivotTable síce produkuje aj kontingenčné tabuľky, ale tie nepredstavujú ani desatinu jeho možností.




Ťažiskovým pre prácu je okno *Polia kontingenčnej tabuľky* (*Pivot Table Field List),* ktoré v hornej časti *Vyberte polia, ktoré chcete pridať do zostavy*: (*Choose fields to add to report:*) obsahuje zoznam premenných (prvý riadok v úvode špecifikovaného bloku údajov, ktoré chceme analyzovať). Nimi sa podľa potreby úloh analýzy zapĺňajú štyri okienka v spodnej časti *Presúvať polia medzi nižšie uvedenými oblasťami:* (*Drag fields between areas below*:), konkrétne okienka *Filter zostavy* (*Report Filter*), *Menovky riadkov* (*Row Labels), Menovky stĺpcov* (*Column Labels*) a *Hodnoty* (*Values*). Okienko *Hodnoty* (*Values*) musí byť **určite** zaplnené, ostatne len podľa charakteru riešenej úlohy.

V ľavej časti výstupu je priestor na umiestnenie vypočítanej Kontingenčnej tabuľky a prípadné presuny v nej.

Podsystém Kontingenčná tabuľka obsahuje aj časť *Kontingenčný graf* (*PivotChart*).

Podrobnejšie si prácu s podsystémom Kontingenčná tabuľka čitateľ môže pozrieť v 14. Kapitole knihy.

### 1.1.4 Použitie formúl

Pôvodným určením tabuľkových kalkulátorov, a teda aj Excelu ako v súčasnosti najrozšírenejšieho kalkulátora, je realizovať veľké množstvo jednoduchých výpočtov. Jednoduché výpočty sa realizujú pomocou formúl (vzorcov) a „veľké“ množstvo pomocou kopírovania. Zápis formuly vždy začína operátorom „=“. Za nim nasleduje text formuly. Formula môže obsahovať číselnú konštantu, operátory:

+ - súčet,

– - rozdiel,

* - násobenie,

/ - delenie,

^ - umocňovanie.

Vo formulách môžeme použiť ľavú a pravú okrúhlu zátvorku *(*, *)*. Ďalej sa vo formule môže použiť ľubovoľná funkcia či konštanta.

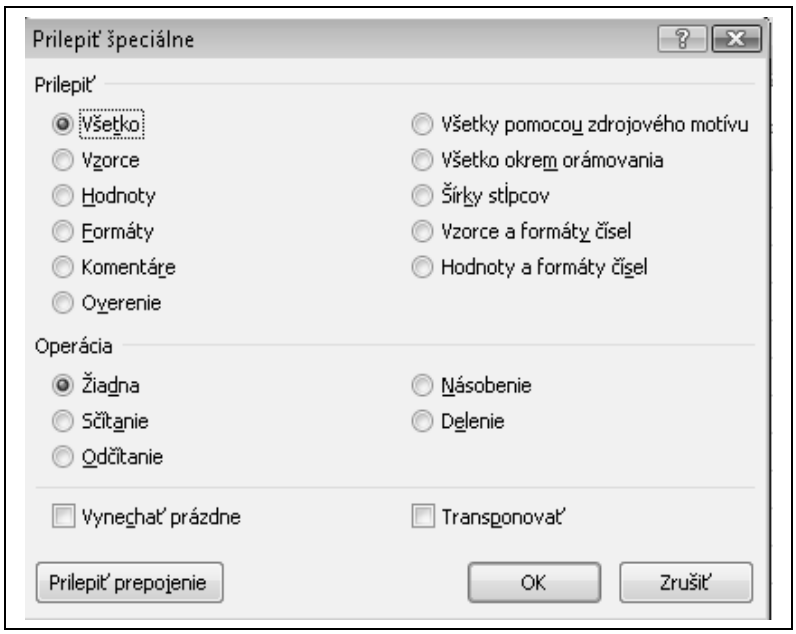
Realizáciu výpočtu zabezpečíme tým, že kým začneme zapisovať text formuly, nastavíme kurzor na políčko v tabuľke, v ktorom chceme mať výsledok formuly. Potom napíšeme text formuly a nakoniec stlačíme *Enter* resp. ťukneme na tlačidlo *OK*.

Veľkú časť formúl zapisujeme nielen s cieľom vypočítať ich hodnotu, ale aj s cieľom skopírovať ich zápis do ďalších políčok tabuľky a ihneď vypočítať hodnoty formuly aj v týchto nových políčkach. Snáď najdôležitejšou schopnosťou Excelu je schopnosť skopírovať formuly do ďalších políčok tabuľky tak, že sa automaticky mení v obsahu kopírovanej formuly identifikácia použitých políčok úmerne vzdialenosti cieľového políčka od kopírovania od zdrojového políčka kopírovania. Nech napríklad v políčku *C1* máme formulu *=A1+B1*. Po skopírovaní obsahu políčka *C1* do políčka *C2* (posunuli sme sa o jeden riadok dole) sa obsah formuly automaticky zmení na *=A2+B2* (identifikácia políčok vo formule sa posune o jeden riadok dole). Ak by sme chceli skopírovať obsah políčka *C1* do políčka *D2* (posuv o jeden riadok a zároveň aj o jeden stĺpec) obsah formuly sa zmení na *=B2+C2* (identifikácia políčok vo formule sa posunie o jeden riadok a jeden stĺpec).

Časté sú prípady, keď pri kopírovaní obsahu formuly identifikáciu niektorého z políčok nechceme meniť. Napríklad v políčku *A11* nech máme vypočítanú hodnotu priemeru z hodnôt v políčkach *A1* až *A10*. V políčkach *B1* až *B10* chceme vypočítať odchýlky zapísaných hodnôt od priemernej hodnoty v políčku *A11*. Ak by sme do políčka *B1* zapísali formulu *=A11-A1* a potom ju skopírovali do políčok *B2* až *B10*, identifikácia políčok vo formule sa bude meniť o jeden riadok, t. j. napríklad v políčku *B2* bude formula *=A12-A2* a v políčku *B10* bude formula *=A20-A10*, čiže niečo iné než odchýlka hodnoty od priemeru (hodnota je identifikovaná správne ale priemer stále máme v políčku *A11* a nie v políčkach *A12* až *A20*. Na zafixovanie polohy políčka vo formule sa používa znak dolára *$*. Zafixovať môžeme polohu riadku alebo polohu stĺpca, prípadne oboch. V našom prípade stačí zafixovať polohu riadku, t. j. formulu v políčku *B1* zapísať v tvare *=A$11-A1*. Po jej skopírovaní do políčok *B2* až *B10*, v políčku *B2* budeme mať formulu *=A$11-A2* a v políčku *B10* budeme mať formulu *=A$11-A10*, t. j. odchýlky hodnôt v políčkach *A1* až *A10* od priemeru v políčku *A11*. V štatistických úlohách je použitie znaku dolára na fixáciu polohy políčka vo formule veľmi užitočná schopnosť Excelu. Na fixáciu môžeme tiež použiť opakované stlačenie klávesu F4.

Vlastné kopírovanie možno realizovať stlačením pravého tlačítka myši. Vysvietime v tabuľke oblasť, ktorú chceme skopírovať, ťukneme pravé tlačítko myši, ťukneme na príkaz *Prilepiť* (*Copy*). Potom v tabuľke vysvietime oblasť kam chceme skopírovať východiskovú oblasť a ťukneme na príkaz *Prilepiť* resp. *Prilepiť špeciálne* (*Paste* resp. *Paste Special*). Príkaz *Prilepiť* (*Paste*) zabezpečí skopírovanie obsahu políčka (vysvietenej oblasti políčok)

do určeného priestoru políčok len so zmenou identifikácie políčok vo formule o príslušný posun pri kopírovaní.




Často sa stáva, že vo výslednej tabuľke nepotrebujeme v políčkach pôvodné formuly ale len výsledné číselné hodnoty. V takýchto prípadoch miesto príkazu *Prilepiť* (*Paste*) použijeme príkaz *Prilepiť špeciálne…* (*Paste Special…)*. V prípade, že ponecháme voľbu *Všetko* (*All*) v časti *Prilepiť* a *None* v časti *Operácie* (*Operation*), kopírovanie je zhodné z obyčajným kopírovaním. Pri riešení štatistických úloh však v niektorých prípadoch môže byť vhodné namiesto prenesenia obsahu formuly preniesť len vypočítanú hodnotu formuly. Takúto požiadavku zabezpečíme ťuknutím na voľbu *Hodnota* (*Value*). Iný prípad môže byť požiadavka transponovania hodnôt – zabezpečíme odfajkovaním okienka *Transponovať* (*Transpose*).

Na vlastné kopírovanie môžeme tiež použiť klávesnicu. Stlačením a držaním klávesu *Shift* kurzorovými klávesmi vysvietime oblasť, ktorú chceme skopírovať. Súčasne stlačíme klávesy *Ctrl* a *C*. Pomocou klávesov *Shift* a kurzorových klávesov vysvietime oblasť kam chceme skopírovať zdrojovú oblasť. Súčasne stlačíme klávesy *Ctrl* a *V*. Skopírovanie je zrealizované.

Tretí spôsob kopírovania predstavuje vysvietiť políčko (oblasť políčok) obsah ktorého chceme skopírovať. V pravom spodnom rohu rámčeka vysvietenej oblasti je malý štvorček (ukazovátko myši nastavíme na tento štvorček, zo šípky sa zmení na krížik). Stlačíme ľavé tlačidlo myši a krížikom vysvecujeme oblasť kam chceme skopírovať zdrojovú oblasť. Po pustení tlačidla myši je kopírovanie ukončené.

### 1.1.5 Použitie makier

Vyššou formou práce v Exceli je práca s využitím makier. Na jej efektívnu aplikáciu sa vyžadujú rutinne sa opakujúce alebo nie zrovna triviálne úlohy. Aj používateľ musí postúpiť do vyššej kategórie používateľov schopných pracovať s makrami. Môžeme ich rozdeliť do dvoch skupín a to na skupinu, ktorá dokáže hotové makro spustiť a tým zadanú úlohu vyriešiť a skupinu, ktorá dokáže makro aj naprogramovať. Kto dokáže makro

naprogramovať, ten ho vie aj spustiť. Kto dokáže makro spustiť, má určité predpoklady makro naprogramovať, len nie vždy má pre neho zmysel ho programovať – makro je už hotové alebo v rámci deľby práce je k dispozícii špecialista, ktorý ho naprogramuje zvyčajne rýchlejšie, efektívnejšie a v programátorskom aspekte spoľahlivejšie. Na druhej strane znalosť vecnej stránky obsahu makra a schopnosť si naprogramovať makro samostatne znamená, že špecialistovi na makrá nemusíme vysvetľovať podstatu riešeného problému a kontrolovať, či špecialista na makrá správne pochopil postup riešenia zadaného problému.

Viac podrobností k problematike práce s makrami je v poslednej, 17. kapitole.

Mladší čitateľ knihy by mal problematiku práce s makrami zvládnuť – určite sa v jeho živote vyskytne situácia, že použitie makra na jej riešenie je vhodné (to sú prakticky všetky situácie) ale aj dostatočne efektívne (čo už nie je až tak častá situácia), Starší čitateľ knihy zváži svoj potenciálny vzťah k makrám, má skúsenosti, ktoré mu dávajú podstatne jasnejšiu predstavu o efektívnosti použitia makra. A aj keď príde k záveru, že problematiku makier už nemusí zvládnuť, je skoro isté, že si 17 kapitolu minimálne prelistuje.

## 2. Záver

Excel poskytuje bohaté možnosti k riešeniu štatistických úloh. Použitie funkcií je bežné, hoci znalosť obsahu niektorých zložitejších funkcií až taká bežná nie je. Použitie nástrojov je dosť zriedkavé. Jednak ich treba aktivovať a používateľ by mal vedieť „čo chce“. Samotné použitie nástrojov je jednoduché, hoci čítanie výsledkov môže robiť časti používateľov problémy. Podsystém Pivot Table (Kontingenčná tabuľka) je priamo dostupný a vysoko efektívny nástroj analýzy rozsiahlych súborov údajov – kto ho ešte nepoužil, autor odporúča čitateľovi určite vyskúšať prácu s Pivot Table (Kontingenčná tabuľka); raz za čas sa určite vyskytne úloha, ktorú bude vhodné riešiť podsystémom Kontingenčná tabuľka. Použitie makier predstavuje profesionálnejšiu úroveň práce.

Hoci Excel nie je profesionálny štatistický softvér, pre absolútnu väčšinu bežných používateľov Excelu rozsah štatistických metód zabudovaných do Excelu výrazne presahuje ich vzdelanostnú úroveň z oblasti štatistiky.

## 3. Literatúra

Chajdiak J. (2009): Štatistika v Exceli 2007. STATIS, Bratislava, ISBN 978 – 80 - 85659-49-8.

Chajdiak J. (2005): Štatistické úlohy a ich riešenie v Exceli. STATIS, Bratislava, ISBN 80 - 85659-39-5

Chajdiak J. (2002): Štatistika v Exceli. STATIS, Bratislava, ISBN 80 - 85659-27-1

**Adresa autora:**

Doc. Ing. Jozef Chajdiak, CSc.

Ústav manažmentu STU Bratislava

chajdiak@statis.biz

# K algoritmizaci manažerských úloh využívajících statistické nástroje
# On the algorithmization of the managerial problems using statistical tools

Janová Jitka

**Abstrakt:** V příspěvku je diskutován problém začlenění statistických metod do praktické výuky manažerských předmětů. Na příkladě optimálního rozvržení reklamy s využitím stochastického programování jsou vysvětleny základní problémy studentů s pochopením a aplikací statistických metod v praxi (převod zadání do podoby matematického modelu, identifikace potřebných metod- matematického programování a metod statistických- a zvolení správného postupu k řešení). Je navržen postup využívající algoritmizovaných úloh, který usnadňuje studentům lépe pochopit a následně prakticky využívat metody probírané v předmětech aplikované matematiky a statistiky. Díky algoritmizovanému řešení úloh je přístup vhodný jak pro prezenční formu výuky tak pro tvorbu e-learningových opor manažerských předmětů, které využívají matematické a statistické nástroje.

**Key words:** optimization, marketing, stochastic programming,
**Kľúčové slová:** optimalizace, marketing, stochastické programování

## 1. Úvod

Optimalizační úlohy jsou dnes již běžnou součástí manažerského rozhodování a metody optimalizace jsou standardně obsaženy v univerzitních kurzech zabývajících se aplikovanou matematikou (ekonomicko-matematické metody, operační výzkum, atd.). Studenti ekonomických a manažerských oborů během studia procházejí nejprve předměty matematického základu na něž navazují teoretické statistické předměty. V posledním ročníku bakalářského studia případně v následném magisterském studiu pak na tyto předměty navazují kurzy aplikované: u statistiky se jedná zejména o ekonometrii, zatímco aplikace matematiky jsou podrobně probírány v operačním výzkumu nebo ekonomicko-matematických metodách. Široké spektrum kvantitativních metod vyučovaných ve vysokoškolských ekonomických oborech odráží fakt, že zaměstnavatelé při výběrových řízeních řadí mezi klíčové schopnosti úspěšného adepta schopnost analytického a logického myšlení společně s užíváním exaktních metod a příslušného SW. Často proto vidíme, že na manažerské pozice firmy hledají adepta s vysokoškolským technickým *nebo* ekonomickým vzděláním. Studenti ekonomických a manažerských oborů většinou nevnímají studium kvantitativních nástrojů jako prioritu a často nemají ani přirozeně dané analytické a logické uvažování. Přesto nastupují do manažerských oborů s cílem stát se manažery, kteří však zejména analyzují a řeší operační a strategické problémy. Aplikované matematické kurzy na ekonomických vysokých školách by měly kromě samotných informací o způsobu řešení konkrétních úloh také klást důraz na metodologii řešení logických a analytických problémů a vzdělávat tak studenty v oblasti, ve které mají studenti technického zaměření většinou náskok.

Možností, jak mohou studenti během řešení konkrétních problémů získat dovednost formulovat, analyzovat a nakonec vyřešit problém (nejen matematický), je přistupovat ve výuce k příkladům algoritmizovaně. Udržováním ustáleného postupu při řešení matematických úloh si studenti mohou vštípit zásady správné rozhodovací praxe. Nástrojem pro takový způsob výuky jsou tzv. typové úlohy (viz [4]), které nabízejí následující ucelený standardizovaný postup při řešení úloh z operačního výzkumu (popsaný postup je využíván při výuce předmětu Operační výzkum na Provozně ekonomické fakultě MZLU v Brně):

A. **Zadání typové úlohy** obsahuje konkrétní problém z praxe včetně číselných hodnot. V této části úlohy si studenti osvojí formulaci problému a zařadí problém

do jedné z kategorií problémů, které umí řešit (vícekriteriální rozhodování, optimalizace, …).

B. **Výběr vhodných matematicko ekonomických metod.** V případě dostatečné znalosti potřebných matematických metod v této části typové úlohy pouze vysvětlíme výběr metody a shrnujeme její podstatu. V opačném případě je nutno použitou metodu také vysvětlit, přičemž důraz nemá být kladen na podrobný matematický výklad, nýbrž na praktické užití metody. Dochází typicky k výběru a vysvětlení jednoho konkrétního způsobu řešení, ačkoliv daná metoda jich nabízí více a podobně.
C. **Řešení úlohy** krok za krokem vysvětluje postup řešení s konkrétními hodnotami.
D. **Diskuze výsledků úlohy a jejich možného využití v praxi**.
E. **Charakteristika skupiny úloh,** na níž tento typ řešení lze použít. Důraz je kladen na specifikum vybrané typové úlohy a z něj vyplývající podmínky, které musí být splněny, aby bylo možno použít algoritmus uvedený v typové úloze.

Výše uvedený postup popisuje výklad učitele. Stejný postup je však nutné vyžadovat po studentech při řešení úloh ve cvičení, aby si osvojili rutinu v přístupu k řešení problémů. Z hlediska matematického řešení úlohy je ve výše popsaném výčtu nejproblematičtější bod B, ve kterém jsou studenti nuceni zvolit vhodné matematické metody. Příčinou absence této dovednosti je způsob, jakým jsou standardně vyučovány matematické předměty. K probrané teoretické látce totiž zpravidla přísluší soubor příkladů, pro jejichž řešení jsou nutné postupy probírané v dané kapitole a žádné jiné. Studenti se proto naučí postup a pouze jej aplikují na příklady. V reálných rozhodovacích problémech je však téměř nejdůležitější součástí řešení identifikovat na základě formulace problému typ úlohy, o který jde, a návazně volit vhodné matematické a SW nástroje k řešení. Navíc schopnost rozpoznání vhodných matematických metod je u studentů ještě nižší, pokud je třeba pro vyřešení problému kombinovat více různých matematických oblastí.

Mluvíme-li o operačním výzkumu a optimalizacích, nastává tato situace v případech, kdy je třeba při hledání optimálního řešení úlohy využít znalosti ze statistických předmětů. Konkrétně se problémy dostavují při řešení úloh stochastického programování řešících optimalizační problémy, ve kterých se vyskytují náhodné veličiny. Standardní postup řešení úlohy je reformulace optimalizačního modelu takovým způsobem, že účelová funkce ani omezující podmínky neobsahují náhodné veličiny, a optimální řešení modelu přitom dobře vystihuje skutečné hledané optimum. V následující části uvedeme konkrétní typovou úlohu, která vyžaduje ke svému řešení syntézu znalostí z operačního výzkumu a statistiky. Popíšeme jednotlivé kroky řešení s důrazem na partie, které jsou pro studenty obtížně řešitelné, a nastíníme možnosti odstranění těchto úzkých míst.

## 2. Příklad typové úlohy: Optimalizace rozvržení reklamy

Zadání úlohy: Uvažujme následující optimalizační problém: Firma vyrábějící dětské hračky se rozhoduje o optimálním rozložení reklamních spotů v televizi během všedního dne. Sledovanost dětmi do 12-ti let $s_i$ v $i$-tém období dne je náhodná veličina se střední hodnotou $S_i$ a rozptylem $\sigma_i^2$. Cena za spot v $i$-tém časovém období je $p_i$ (viz tabulka 1). Kritériem optima je celková sledovanost spotů, přičemž je třeba maximalizovat sledovanost při daném rozpočtovém omezení 30 000Kč.

Ze zadání vyplývá, že půjde o optimalizační úlohu, neboť chceme maximalizovat sledovanost při daném rozpočtovém omezení, využijeme tedy metod matematického programování, konkrétně programování stochastického, neboť v úloze vystupují náhodné veličiny v podobě sledovaností. Jsou dány průměry a rozptyly, nikoliv však typ rozdělení náhodných veličin.

***Tabulka 1: Sledovanost televizní stanice dětmi do 12-ti let a ceny za reklamní spot***

| $i$ | hodiny | $S_i$ [%] | $\sigma_i$ | $\sigma_i^2$ | $p_i$ [Kč] |
|---|---|---|---|---|---|
| 1 | 0h-2h | 3 | 2.50 | 6.25 | 1000 |
| 2 | 2h-4h | 2.5 | 2.5 | 6.25 | 1000 |
| 3 | 4h-6h | 2 | 2.5 | 6.25 | 1000 |
| 4 | 6h-8h | 6 | 1.5 | 2.25 | 3900 |
| 5 | 8h-10h | 10 | 3 | 9.00 | 5900 |
| 6 | 10h-12h | 10 | 3.5 | 12.25 | 5900 |
| 7 | 12η–14η | 13 | 4 | 16.00 | 5900 |
| 8 | 14h-16h | 14 | 4.5 | 20.25 | 7900 |
| 9 | 16h-18h | 20 | 8 | 64.00 | 7900 |
| 10 | 18h-20h | 45 | 10 | 100.00 | 7900 |
| 11 | 20h-22h | 55 | 15 | 225.00 | 7900 |
| 12 | 22h-24h | 15 | 13 | 169.00 | 4900 |

Vybrané přístupy k řešení úlohy stochastického programování**:** Mějme obecně zadaný problém stochastického programování

$$z^* = \max c^T x \qquad (1)$$
$$Ax \leq b,$$
$$x \geq 0,$$

kde $c$ je vektor náhodných proměnných, $A$ je matice známých parametrů a $b$ je vektor známých disponibilních zdrojů. Hledáme tedy optimální řešení úlohy lineárního stochastického programování s náhodnými veličinami pouze v účelové funkci. Standardním předpokladem postupů vyučovaných v předmětech zahrnujících operační výzkum na magisterském stupni je, že všechny uvažované náhodné veličiny mají normální rozdělení a jsou vzájemně nezávislé.

V tomto bodě studenti většinou přímo přecházejí k formulaci úlohy stochastického programování, jak je uvedeno níže, aniž by si uvědomili, že tyto postupy lze použít pouze po ověření nezávislosti veličin a řádném otestování normality příslušných rozdělení. Studenti zapsaní na kurz operačního výzkumu, který bývá společně s ekonometrií nejpokročilejším matematickým kurzem magisterského studia, absolvovali přednášky o neparametrických testech a korelační analýze v předchozích statistických předmětech, nicméně téměř bez výjimky tyto znalosti nejsou schopni aplikovat, když narazí na jejich potřebu v jiném předmětu a jiném kontextu.

Shrňme tedy postup předkládaný studentům: V případě úlohy stochastického programování můžeme řešit úlohy, kde náhodné veličiny mají normální rozdělení (studenti jistě znají Kolmogorovův-Smirnovův test a $\chi^2$ –test dobré shody) a jsou nezávislé (studenti využijí korelační analýzu). Teprve po pozitivním výsledku můžeme pokračovat optimalizací. V případě opominutí statistické části úlohy mohou být řešení dosažená optimalizací zcela chybná a pro praxi nepřípustná či neoptimální. Důležitou součástí je upozornit na nutnost kvalitního sběru a vyhodnocení dat, na které studenti často zapomínají. V dané úloze se to projevuje tak, že studenti vyhledávají data pro testování rozdělení sledovaností v jednotlivých obdobích dne v zadané tabulce 1. Teprve po učitelem řízené diskuzi o tom, jaká data bychom měli mít pro testování k dispozici, docházejí k názoru, že nám chybí soubory sledovaností v každém období dne, ze kterých je možné teprve rozhodnout, zda jednotlivé veličiny $S_i$ – *sledovanost v i-tém období dne* mají normální rozdělení, či ne.

Je zřejmé, že uvedený zevrubný přístup k řešení příkladů je časově velmi náročný a vzhledem k hodinové dotaci Operačního výzkumu není možné důkladně provádět potřebné statistické analýzy. Proto v typové úloze pouze shrnujeme potřebné přípravné kroky před vlastním řešením optimalizační úlohy a opakujeme vhodné metody k jejich řešení, které jsou studentům známé.

Existuje řada variant, jak přistupovat k řešení úlohy stochastického programování. Pro potřeby magisterského kurzu jsou standardně uváděny tři vybrané přístupy k převodu stochastické úlohy na úlohu deterministického programování (viz např. [1], [5]):

*Kritérium střední hodnoty* představuje intuitivní přístup, ve kterém jsou náhodné veličiny v modelu (1) nahrazeny příslušnými středními hodnotami:

$$z^* = \max \sum_{i=1}^{12} x_i S_i \qquad (2)$$
$$\sum_{i=1}^{12} x_i p_i \le 30000,$$
$$x_i \ge 0.$$

Tento přístup ale opomíjí všechny informace o náhodnosti veličin a je proto vhodný v případech, kdy rozhodovatel je schopen odhalit nepřípustnost dosaženého „optimálního řešení".

*Kritérium minimálního rozptylu* zaměňuje požadavek maximalizace účelové funkce (1) za požadavek minimálního rozptylu této účelové funkce, který je doplněn požadavkem minimální přípustné hodnoty očekávaného zisku. Pro naši úlohu stanovujeme minimální přípustnou hranici očekávané sledovanosti na 180 jednotek. Tyto požadavky společně s přihlédnutím k nezávislosti uvažovaných náhodných veličin lze zapsat modelem kvadratického programování:

$$z^* = \min \sum_{i=1}^{12} x_i^2 \sigma_i^2 \qquad (3)$$
$$\sum_{i=1}^{12} x_i p_i \le 30000,$$
$$\sum_{i=1}^{12} x_i S_i \ge 180,$$
$$x_i \ge 0.$$

*Pravděpodobnostní kritérium* minimalizuje pravděpodobnost, že účelová funkce (1) klesne pod jistou úroveň (opět volíme 180 jednotek). V tomto případě je stochastický model (1) nahrazen deterministickým modelem nelineárního programování:

$$z^* = \min \frac{180 - \sum_{i=1}^{12} x_i S_i}{\sqrt{\sum_{i=1}^{12} x_i^2 \sigma_i^2}} \qquad (4)$$
$$\sum_{i=1}^{12} x_i p_i \le 30000,$$
$$x_i \ge 0.$$

Řešení úlohy: Všechny modely jsou řešitelné v Excelu pomocí nástroje Řešitel, studenti samostatně úlohu pro všechny zvolené účelové funkce řeší a seznamují se s možnostmi a postupy, které Excel pro optimalizační úlohy nabízí (viz např.[2], [3]). Získaná optimální řešení vidíme v tabulce 2.

***Tabulka 2: Optimální rozmístění reklamních spotů při různých kritériích optimality***

| *i* | hodiny | kritérium (2) | kritérium (3) | kritérium (4) |
|---|---|---|---|---|
| 1 | 0h-2h | 6 | 2 | 3 |
| 2 | 2h-4h | 0 | 2 | 1 |
| 3 | 4h-6h | 0 | 2 | 0 |
| 4 | 6h-8h | 0 | 0 | 0 |
| 5 | 8h-10h | 0 | 0 | 0 |
| 6 | 10h-12h | 0 | 0 | 0 |
| 7 | 12h-14h | 0 | 0 | 0 |
| 8 | 14h-16h | 0 | 0 | 0 |
| 9 | 16h-18h | 0 | 0 | 0 |
| 10 | 18h-20h | 0 | 0 | 0 |
| 11 | 20h-22h | 3 | 3 | 3 |
| 12 | 22h-24h | 0 | 0 | 1 |

Diskuze výsledků úlohy a jejich možného využití v praxi:

V úlohách, kde se vyskytují náhodné veličiny, mají studenti standardně problémy s interpretací výsledků. Je proto nutné jim zdůrazňovat, že hodnota celkové sledovanosti v optimu je očekávanou hodnotou, nikoliv přesnou hodnotou, kterou s jistotou docílíme v případě zvolení rozvržení reklamy podle optimálního řešení. Navíc studenti hůře interpretují nově zvolenou účelovou funkci v modelech (3) a (4). Zatímco u modelu (2) používají hodnocení z deterministického matematického programování: tj. účelová funkce je přímo očekávanou sledovaností, není jim u modelů (3) a (4) jasné, že očekávaná sledovanost je zakomponována do jedné z podmínek a tam je ji ve výsledkové zprávě Řešitele také nutno hledat.

V tabulce 2 vidíme, že optimální řešení dosažené nástroji matematického programování je silně závislé na zvoleném kritériu optimality. Nelze jednoznačně říci, které kritérium je nejlepší, protože pro každý problém a každého rozhodovatele bude vhodné jiné kritérium optima. V případě tří kritérií, které jsme v příspěvku zvolili, můžeme shrnout, že kritérium střední hodnoty eliminující informace o náhodné sledovanosti na její střední hodnotu zcela pomíjí náhodnost sledovanosti a realizace zvoleného řešení může vést k zásadně odlišným-nižším hodnotám účelové funkce, než jakou očekáváme, protože rozptyl účelové funkce při daném řešení může být značný. V případě, že rozhodovatel má averzi k riziku, může zvolit kritérium minimálního rozptylu, které se spokojuje zpravidla s nižší hodnotou očekávané sledovanosti, avšak s velkou pravděpodobností při optimálním rozložení spotů podle tohoto kritéria bude skutečná sledovanost blízká té očekávané. Pravděpodobnostní kritérium je pak rovněž vhodné pro rozhodovatele s averzí k riziku. Při realizaci tohoto optimálního řešení dosahujeme minimální pravděpodobnost, že sledovanost bude nižší než 180 jednotek.

Je také nutno studentům zdůraznit, že optimální výsledky podle prvního a ostatních kritérií se neliší zásadně (všechna kritéria umisťují reklamu do brzkých ranních nebo pozdních večerních hodin), což není způsobeno dobrou vypovídací schopností prvního kritéria, nýbrž volbou minimální přípustné hranice očekávaného zisku $z_0$=180 u druhých dvou kritérií, což je řádově srovnatelná hodnota s očekávanou sledovaností v optimu podle kritéria střední hodnoty.

Charakteristika skupiny úloh, na které lze postup použít:

Uvedený postup je předpřipraveným návodem pro řešení jednoduchých úloh stochastického programování. Studenti však musí dodržovat následující algoritmus:

1. Zápis modelu matematického programování
2. Identifikace náhodných veličin

3. Ověření nezávislosti náhodných veličin (využít poznatky ze statistiky: korelační analýza)
4. Test normality (využít poznatky ze statistiky: Kolmogorovův-Smirnovův test, $\chi^2$ –test dobré shody)
5. V případě, že náhodné veličiny jsou nezávislé a mají normální rozdělení rozhodovatel volí nejvhodnější kritérium optima. Vybírá z (2), (3), (4) případně konstruuje vlastní kritérium vystihující jeho postoj k riziku a sledované cíle optimalizace.
6. Řešení pomocí vhodného SW (např. Excel).
7. Interpretace výstupů. Nelze stanovit přesnou hodnotu celkové sledovanosti, která bude realizována při volbě vypočítaného optimálního rozložení reklamy.

## 3. Závěr

Studenti ekonomických vysokých škol mají obecně malou schopnost využívat matematické poznatky v praxi. Tato schopnost bývá většinou ještě snížena pokud je nezbytné kombinovat pro úspěšné vyřešení problému více matematických metod anebo metod z různých oblastí matematiky. V reálných rozhodovacích úlohách jsou však právě takové kombinace hojně využívány a často neschopnost jejich aplikování znemožňuje celkové úspěšné a přesné vyřešení problému. Možností, jak suplovat nižší analytické schopnosti je vštěpovat studentům algoritmizované postupy pro řešení standardních např. optimalizačních úloh. S tímto typem výuky v Operačním výzkumu je samozřejmě spjata zvýšená časová dotace na řešení mnoha úloh podobného typu na úkor probírání teoretických podkladových matematických struktur. V případě studentů ekonomických oborů je však tento trend poměrně žádoucí, protože jednak zvyšuje reálné využití teoretických matematických postupů v praxi a navíc studenti s pomocí algoritmizovaného řešení typových úloh získávají praxi v metodologii řešení problémů jako takových. Vštěpují si, že kvantitativní metody jsou pouze nástrojem k získání číselných výsledků získaných řešením přibližných modelů a teprve rozhodovatel, který musí přihlížet k počátečním zjednodušením a náhodnosti proměnných, dokáže správně a pro praxi přínosně interpretovat výsledek.

## 4. Literatura


[1] BIRGE, J.R.-LOUVEAUX 1997. Introduction to Stochastic programming. Springer, 1997. 448 s. ISBN 978-0387982175.

[2] GROS, I. 2003. Kvantitativní metody v manažerském rozhodování. Grada publishing. Praha, 2003. 432 s. ISBN 80-247-0421-8.

[3] JABLONSKÝ, J. 2002. Operační výzkum. Professional publishing. Praha, 2002. 320 s. ISBN 80-86419-42-8.

[4] JANOVÁ, J., 2007, Moderní matematicko-ekonomické metody pro vojenskou praxi. In: Vojenské rozhledy, zvláštní číslo, s. 60 – 67.

[5] KALL, P.-WALLACE, S.W. 1994. Stochastic programming. John Wiley & Sons, 1997. 320 s. ISBN 978-0471951087.


## Adresa autora:


Janová Jitka, Mgr. Ph.D.
Zemědělská 1
61300 Brno
janova@mendelu.cz

# Statistická kontrola procesu při výrobě v malých sériích

Jarošová Eva

**Abstract:** The paper deals with the statistical process control (SPC) during short production runs. Some of best known approaches to the quality control of short-run production are described in the paper. Features of Shewhart control charts expressed by means of the average run length and the risk of false alarm are recapitulated and some disadvantages of alternative methods are commented. Only the control charts for variables are considered. Two examples illustrate the calculation of nominal and standardized control charts. One of them uses subgroups, the other individual values without subgrouping.

**Key words:** Shewhart control chart, nominal and standardized control charts, average run length, risk of false alarm

**Klíčová slova:** Shewhartův regulační diagram, nominální a standardizovaný regulační diagram, průměrný počet výběrů vedoucí k signálu, riziko falešného signálu

## 1. Úvod

Statistická kontrola procesu (SPC) je v současné době předmětem značného zájmu. Ačkoli se v praxi stále používají především klasické Shewhartovy regulační digramy, byla navržena řada dalších postupů. Některé z nich se objevily jen o několik desítek let později než první regulační diagram, jiné jsou relativně nové. Důvodů pro hledání alternativ je více, např. snaha o zlepšení schopnosti detekce změny parametrů procesu nebo zohlednění faktu, že základní předpoklady, tj. normalita a nezávislost hodnot regulované veličiny nejsou v praxi často splněny. Je třeba si uvědomit, že mnohé procesy, na které se má statistická kontrola aplikovat, se podstatně liší od těch, pro něž byla původní metoda navržena. Např. použití klasického regulačního diagramu na proces s vysokou způsobilostí může vést k častým signálům naznačujícím existenci vymezitelné příčiny, ačkoli jde ve skutečnosti o příčiny, které jsou inherentní složkou procesu. Jiným příkladem je výroba malých sérií, která souvisí s uplatňováním moderního přístupu k zásobování, tzv. metodou *just-in-time*, jejímž cílem jsou nulové zásoby a stoprocentní kvalita. Aplikaci klasického postupu SPC brání nedostatečný počet měření pro určení regulačních mezí. Regulační diagramy pro kontrolu procesu při výrobě malých sérií jsou předmětem tohoto článku. Kromě postupů, které jsou nejčastěji zmiňovány v literatuře, např. [2], [3], [6] a jsou také implementovány v některých statistických softwarových produktech, jako např. Statistica a v omezené míře Minitab, budou uvedeny i některé další, beroucí v úvahu porušení předpokladu nezávislosti v důsledku modifikace klasického Shewhartova postupu. Omezíme se přitom na regulaci měřením s větším důrazem na kontrolu úrovně procesu.

## 2. Vlastnosti Shewhartova regulačního diagramu

SPC při kontrole měřením spočívá v pravidelném monitorování procesu prostřednictvím malých výběrů jednotek, v nichž se určují charakteristiky úrovně a variability, nejčastěji průměr a rozpětí. Používají se dva grafy, jeden pro kontrolu úrovně, druhý pro kontrolu variability regulované veličiny. Hodnoty charakteristik se vynášejí do grafu proti pořadovému číslu výběru (podskupiny). Graf obsahuje centrální přímku (CL), horní regulační mez (UCL) a dolní regulační mez (LCL). Centrální přímka je umístěna v referenční hodnotě znázorňované charakteristiky, regulační meze jsou ve vzdálenosti trojnásobku směrodatné odchylky zobrazované výběrové charakteristiky.

Ve fázi hodnocení, zda proces je či není ve statisticky zvládnutém stavu, kdy je variabilita způsobena jen náhodnými příčinami (první etapa regulace), se za referenční hodnotu považuje obvykle průměr znázorňovaných charakteristik. Při vlastní regulaci procesu (druhá etapa) se referenční hodnota stanoví buď na základě minulé zkušenosti s procesem nebo na základě technického zadání [8]. Volbou referenční hodnoty v diagramu pro variabilitu je ovlivněna nejen centrální přímka a regulační meze v tomto diagramu, ale rovněž regulační meze v diagramu pro kontrolu úrovně. Odvozujeme-li referenční hodnotu na základě minulých dat, probíhá konstrukce diagramu ve dvou fázích. Na základě dostatečně velkého počtu hodnot (doporučuje se alespoň 100, viz dále) se odhadnou parametry rozdělení a určí se předběžné regulační meze, které se aplikují retrospektivně. Vyskytují-li se některé body mimo tyto meze, příslušné podskupiny se vyloučí a meze se přepočítají. Výsledné meze se potom používají pro kontrolu pokračujícího procesu.

Vlastnosti regulačních diagramů se nejčastěji posuzují prostřednictvím rizika falešného signálu, tj. pravděpodobnosti, že vynesená hodnota charakteristiky překročí regulační meze, i když se parametry procesu nezměnily, nebo podle průměrného počtu výběrů vedoucích k signálu (nejčastěji výskytu bodu mimo regulační meze) *ARL*. Počet výběrů má geometrické rozdělení s parametrem *p*, který značí pravděpodobnost výskytu signálu u každého výběru. Za předpokladu normálního rozdělení regulované veličiny a při stanovených referenčních hodnotách, tedy známých parametrech normálního rozdělení, je pravděpodobnost překročení regulačních mezí u procesu pod kontrolou rovna 0,0027 a odpovídající *ARL* je rovno střední hodnotě geometrického rozdělení 1/*p*, tedy přibližně 370. Při posunu střední hodnoty je pravděpodobnost signálu vyšší a *ARL* naopak nižší. Shewhartův diagram má dobrou schopnost detekce větších odchylek střední hodnoty od referenční, např. při odchylce velikosti $\sigma/\sqrt{n}$ je *ARL* rovno 43, při dvojnásobné odchylce je *ARL* rovno 5. Pro detekci menších odchylek je vhodnější CUSUM diagram, viz např. [7].

Parametry normálního rozdělení jsou často, alespoň v první etapě kontroly procesu, odhadovány. K odhadu střední hodnoty se používá průměr z výběrových průměrů, k odhadu směrodatné odchylky průměr z výběrových rozpětí, směrodatných odchylek nebo rozptylů. Příslušné vzorce lze nalézt např. v [7] či [8]. V tomto případě nelze *ARL* určit analyticky, protože regulační meze jsou funkcí stejných realizací náhodné veličiny a odchylky vynášené charakteristiky od regulačních mezí nejsou již nezávislé. *ARL* je větší než v případě známých parametrů, viz např. [6]. Pomocí simulací bylo zjištěno, že pokud je odhad směrodatné odchylky založen alespoň na 100 pozorováních, jsou vlastnosti diagramu podobné případu se známými parametry.

## 3. Modifikace Shewhartova regulačního diagramu

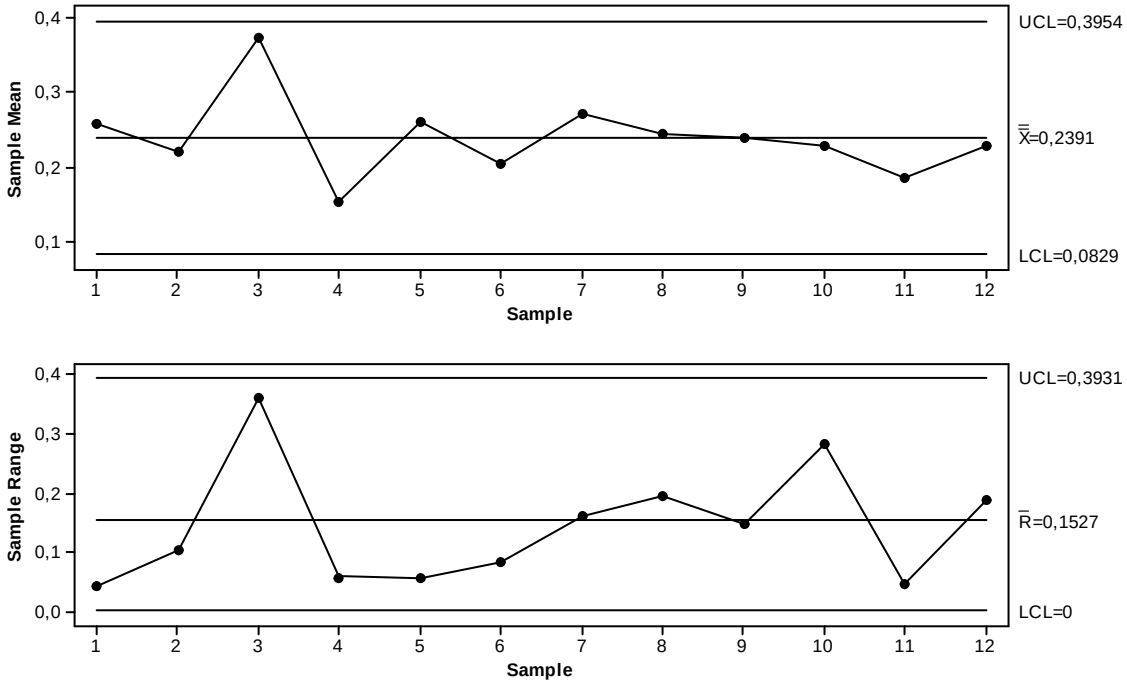
Snaha vyřešit problém nedostatečného počtu hodnot pro konstrukci regulačních mezí vedla k myšlence volby takové regulované veličiny, která umožní sloučit jednotlivé procesy představující výrobu malých sérií do jednoho procesu, pro nějž se zkonstruují společné regulační meze. Za autora je často považován Bothe, viz např. [4]. Tento postup přichází v úvahu především tehdy, probíhá-li výroba za podobných podmínek, např. vyrábí-li se na stejném stroji v malých sériích podobné díly lišící se jen jmenovitým rozměrem. Potom má smysl zaměřit se na kontrolu procesu v širším slova smyslu a nikoli na kontrolu jednotlivých sérií. Tento případ se označuje jako produkce s vyšším stupněm opakovatelnosti (*repetitive manufacturing*, viz např. [3]). Pokud jsou u jednotlivých procesů hodnoty $\mu_0$ a $\sigma_0^2$ stanoveny a u m-tého dílu můžeme předpokládat $X_m \sqcup \mathrm{N}(\mu_{0m}, \sigma_{0m}^2)$, vytvoří se transformované veličiny $Y_m = X_m - \mu_{0m}$ resp. $Y_m = (X_m - \mu_{0m})/\sigma_{0m}$, jejichž hodnoty se mohou „spojit“ a na výslednou veličinu se aplikuje klasický postup. Obvykle jsou však parametry $\sigma_m^2$, často i $\mu_m$,

odhadovány. Je otázkou, zda lze v případě uvažovaných procesů vůbec mluvit o druhé etapě regulace. U prvního typu diagramů, kdy je referenční hodnotou jmenovitá hodnota regulované veličiny (*nominal chart*) nebo je stanovena požadovaná hodnota (*target chart*), ale neznáme směrodatnou odchylku, potřebujeme ke konstrukci předběžných regulačních mezí podle výše zmíněného doporučení alespoň 100 hodnot veličiny *Y*. Tento postup navíc předpokládá, že náhodné kolísání je u různých výrobků stejné. U druhého typu diagramů (*standardized chart*) bychom k dosažení obvyklých vlastností regulačního diagramu potřebovali 100 hodnot od každého výrobku. Zatímco v souvislosti se Shewhartovým diagramem se nutnost dostatečného počtu měření většinou zdůrazňuje, u modifikovaného postupu se tento předpoklad obvykle neuvádí. Přitom vlastnosti regulačního diagramu jsou při malém počtu naměřených hodnot přinejmenším stejně nejisté jako v případě Shewhartova diagramu.

## 4. Příklady

Následující ukázky byly převzaty z [2] a jsou použity jen pro ilustraci postupu:

1) *Nominální regulační diagram*. U tří dílů a, b, c jsou zadány nominální hodnoty T. Z procesu se vybírají podskupiny s rozsahem $n = 3$, vzhledem k různým délkám sérií byly z první série odebrány dvě podskupiny, z druhé série čtyři podskupiny a ze třetí série šest podskupin. Podklady pro konstrukci diagramu jsou uvedeny v tabulce 1.



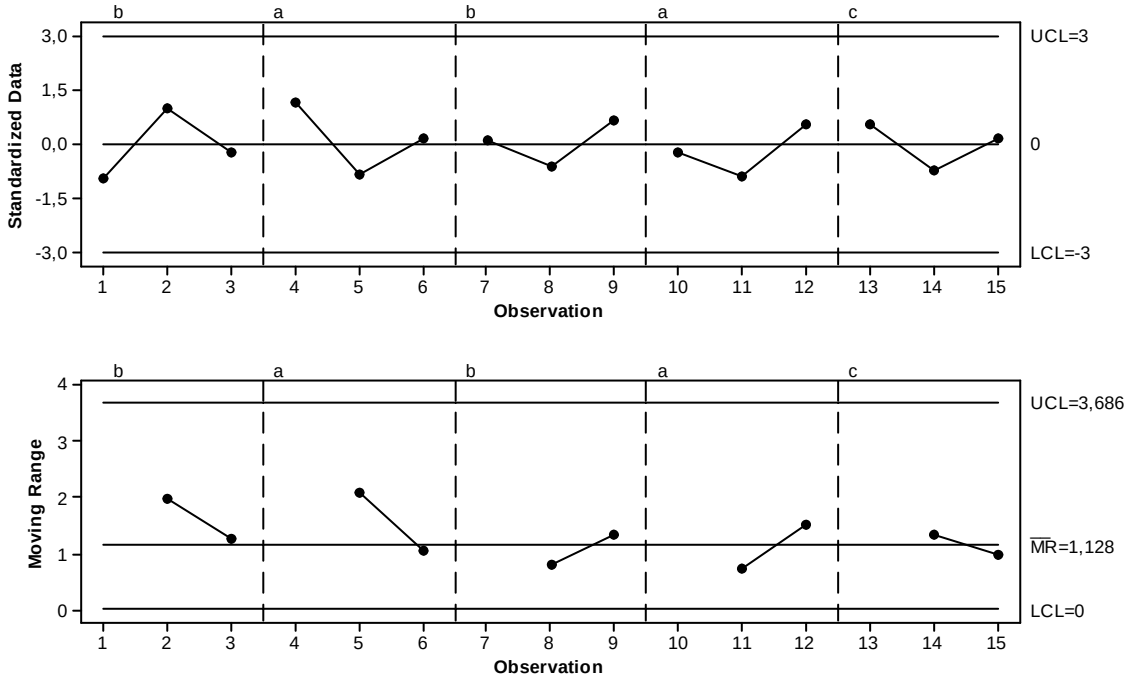

***Obrázek 1: Nominální regulační diagram (Minitab, Xbar-R )***

K zakreslení diagramů pro regulovanou veličinu $Y_m = X_m - \mu_{0m}$ lze použít jakýkoli program, který obsahuje proceduru pro klasické Shewhartovy diagramy; místo původních hodnot obsažených ve sloupcích M1 až M3 se použijí rozdíly uvedené ve sloupcích D1 až D3. Dvojice diagramů je znázorněna na obr. 1. Centrální přímky a regulační meze v diagramu pro průměr i pro rozpětí se určí podle známých vzorců, viz např. [7] nebo [8].

***Tabulka 1: Naměřené a transformované hodnoty regulované veličiny***

| Díl | T | M1 | M2 | M3 | D1 | D2 | D3 |
|---|---|---|---|---|---|---|---|
| a | 3,25 | 3,493 | 3,496 | 3,533 | 0,243 | 0,246 | 0,283 |
| a | 3,25 | 3,450 | 3,431 | 3,533 | 0,200 | 0,181 | 0,283 |
| b | 5,50 | 6,028 | 5,668 | 5,922 | 0,528 | 0,168 | 0,422 |
| b | 5,50 | 5,639 | 5,690 | 5,634 | 0,139 | 0,190 | 0,134 |
| b | 5,50 | 5,790 | 5,757 | 5,735 | 0,290 | 0,257 | 0,235 |
| b | 5,50 | 5,709 | 5,743 | 5,661 | 0,209 | 0,243 | 0,161 |
| c | 7,75 | 8,115 | 7,992 | 7,956 | 0,365 | 0,242 | 0,206 |
| c | 7,75 | 7,885 | 8,023 | 8,077 | 0,135 | 0,273 | 0,327 |
| c | 7,75 | 7,932 | 8,079 | 7,958 | 0,182 | 0,329 | 0,208 |
| c | 7,75 | 8,142 | 7,860 | 7,934 | 0,392 | 0,110 | 0,184 |
| c | 7,75 | 7,907 | 7,951 | 7,947 | 0,157 | 0,201 | 0,197 |
| c | 7,75 | 7,905 | 7,943 | 8,091 | 0,155 | 0,193 | 0,341 |

2) *Standardizovaný regulační diagram*. Ke kontrole tloušťky papíru různých druhů a, b, c, které se vyrábějí v malých sériích, byl použit standardizovaný regulační diagram. Individuální hodnoty z pěti sérií (tři hodnoty v každé sérii) jsou uvedeny v tabulce 2, diagramy pro individuální (standardizované) hodnoty a pro klouzavá rozpětí jsou na obr. 2.




***Obrázek 2: Standardizovaný regulační diagram (Minitab, Z-MR)***

Průměry uvedené ve třetím sloupci tabulky 2 jsou vypočteny vždy ze všech hodnot náležejících stejnému dílu (u dílů *a* a *b* ze šesti hodnot, u dílu *c* jen ze tří hodnot). Podobně je tomu u odhadu směrodatné odchylky ve čtvrtém sloupci. Odhad byl vypočítán pomocí klouzavých rozpětí, viz např. [7] nebo [8]. Hodnoty standardizované veličiny *Z* z pátého sloupce jsou vyneseny v horním diagramu na obr. 2, v posledním sloupci a v dolním diagramu jsou klouzavá rozpětí sousedních hodnot standardizované veličiny, zde se však uvažuje každá série zvlášť, jak je patrné z vynechaných políček v tabulce.

***Tabulka 2: Výpočty pro standardizovaný diagram***

| Druh | X | Průměr | Sigma | Z | Rozpětí |
|---|---|---|---|---|---|
| b | 1,435 | 1,502 | 0,070 | -0,954 | * |
| b | 1,572 | 1,502 | 0,070 | 1,012 | 1,966 |
| b | 1,486 | 1,502 | 0,070 | -0,222 | 1,234 |
| a | 1,883 | 1,785 | 0,082 | 1,203 | * |
| a | 1,715 | 1,785 | 0,082 | -0,852 | 2,055 |
| a | 1,799 | 1,785 | 0,082 | 0,175 | 1,028 |
| b | 1,511 | 1,502 | 0,070 | 0,136 | * |
| b | 1,457 | 1,502 | 0,070 | -0,639 | 0,775 |
| b | 1,548 | 1,502 | 0,070 | 0,667 | 1,306 |
| a | 1,768 | 1,785 | 0,082 | -0,204 | * |
| a | 1,711 | 1,785 | 0,082 | -0,901 | 0,697 |
| a | 1,832 | 1,785 | 0,082 | 0,579 | 1,480 |
| c | 1,427 | 1,392 | 0,063 | 0,557 | * |
| c | 1,344 | 1,392 | 0,063 | -0,752 | 1,309 |
| c | 1,404 | 1,392 | 0,063 | 0,195 | 0,947 |

## 5. Speciální regulační diagramy pro výrobu s nízkým stupněm opakovatelnosti

Pokud je třeba při přechodu na novou výrobní sérii zásadněji přizpůsobovat výrobní zařízení a slučování po sobě následujících sérií není možné, jde o produkci s nízkým stupněm opakovatelnosti (*non-repetitive manufacturing*). Třebaže se i v těchto případech používá výše uvedená transformace regulované veličiny, lze očekávat, že vlastnosti výsledného regulačního diagramu budou nedostatečným počtem hodnot ještě více ovlivněny. Hodnoty (charakteristiky či individuální hodnoty transformované regulační veličiny) vynášené do diagramu jsou závislé a riziko falešného signálu je vyšší, než by odpovídalo předpokladu nezávislosti. Proto byly navrženy různé alternativní metody. Jednou z nich je Hillierova metoda (viz např. [3]), která udržuje riziko falešného signálu na požadované úrovni $\alpha$ bez ohledu na počet podskupin. Uvažuje se tedy původní regulovaná veličina $X$ a v diagramu pro průměr jsou meze dány vztahem

$$\bar{\bar{x}} \pm \sqrt{\frac{k+1}{kn}} t_{1-\alpha/2,k(n-1)} s_p \quad \text{resp.} \quad \bar{\bar{x}} \pm \sqrt{\frac{k-1}{kn}} t_{1-\alpha/2,k(n-1)} s_p \qquad (1)$$

kde $\bar{\bar{x}}$ je celkový průměr, $s_p^2 = \frac{1}{k(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ průměrný vnitroskupinový rozptyl, $k$ je počet podskupin, $n$ jejich rozsah, $t_{1-\alpha/2,k(n-1)}$ je $1-\alpha/2$ kvantil t-rozdělení s $k(n-1)$ stupni volnosti. První vztah platí pro počáteční retrospektivní kontrolu, druhý pro další regulaci procesu.

Quesenberry navrhnul pro známou střední hodnotu a neznámý rozptyl Q-diagram, v němž se po vyhodnocení t výběrů ($t$ = 1, ..., $k$) vynášejí hodnoty (viz např. [3])

$$Q_t = \Phi^{-1}\left( F_{t-1}\left( \frac{x_t - \mu}{\hat{\sigma}_t} \right) \right) \text{ pro } n = 1 \quad \text{nebo} \quad Q_t = \Phi^{-1}\left( F_{tn}\left( \frac{\bar{x}_t - \mu}{\hat{\sigma}_t / \sqrt{n}} \right) \right) \text{ pro } n > 1, \qquad (2)$$

kde $F_{t-1}$ resp. $F_{tn}$ značí distribuční funkci t-rozdělení s $t - 1$ resp. $tn$ stupni volnosti a $\hat{\sigma}_t$ je odhad směrodatné odchylky získaný na základě všech $t$ hodnot, tj.

$$s_t^2 = \frac{1}{t}\sum_{j=1}^{t}(x_j - \mu)^2 \text{ pro } n = 1 \quad \text{nebo} \quad s_t^2 = \frac{1}{tn}\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{ij} - \mu)^2 \text{ pro } n > 1.$$

Regulační meze ve vzdálenosti ±3 jsou určeny hned od začátku kontroly procesu, vynášené hodnoty necharakterizují jednotlivé výběry, ale soubor obsahující všechny naměřené hodnoty až do okamžiku aktuálního výběru. I když obě uvedené metody zaručují požadované malé riziko falešného signálu, schopnost detekce změny parametrů procesu je špatná, jak se uvádí v [3], zvláště v případě, kdy změna parametrů nastane brzy.

## 6. Závěr

Modifikace Shewhartova diagramu spočívající v transformaci regulované veličiny vede při náhradě neznámých parametrů jejich odhady k závislosti hodnot vynášených do diagramu a ke zvýšení rizika falešného signálu. Alternativní metody uvedené v článku zaručují požadovanou velikost tohoto rizika, ale mají horší schopnost detekce změny parametrů procesu. I když lze v literatuře najít metody s lepšími vlastnostmi založené na Kalmanově filtru, původní myšlenka Shewharta, spočívající v jednoduchosti, se ztrácí. Jak se uvádí v kapitole o podstatě regulačních diagramů v normě, Shewhartův regulační diagram „by neměl být uvažován ve smyslu testu hypotézy. Shewhart zdůraznil empirickou užitečnost regulačního diagramu pro rozpoznání odchylek od stavu, kdy výrobní proces je statisticky zvládnutým, a snížil důraz na pravděpodobnostní interpretaci.“ V tomto smyslu není tedy nutné metody založené na modifikaci Shewhartova diagramu zcela odmítat, při aplikaci a především při interpretaci je však třeba jisté opatrnosti.

## 7. Literatura


[1] BISSELL, D. 1994: Statistical Methods for SPC and TQM. London: Chapman & Hall, 1994. 373 s. ISBN 0-412-39440-5.

[2] BREYFOGLE, F.W. 2003: Implementing Six Sigma: Smarter Solutions Using Statistical Methods. New Jersey: J. Wiley & Sons, 2003. 1187 s. ISBN 0-471-26572-1.

[3] DEL CASTILLO, E. – GRAYSON, J.M. – MONTGOMERY, D.C. – RUNGER, G.C. 1996: A review of statistical process control techniques for short run manufacturing systems. In: Communications in Statistics – Theory and Methods č. 11, 1996: s. 2723 – 2737.

[4] MAGUIRE, M., ed. 1999: Statistical gymnastics revisited: A debate on one approach to short-run control charts. In: Quality Progress č. 2, 1999: s. 84 – 94.

[5] QUESENBERRY, CH. 1998: Statistical Gymnastics. In: Quality Progress č. 9, 1998: s. 77 – 79

[6] RYAN, T.P. 2000: Statistical Methods for Quality Improvement. New York: J. Wiley & Sons, 2000. 555 s. ISBN 0-471-19775-0.

[7] TEREK, M. – HRNČIAROVÁ, L. 2004: Štatistické riadenie kvality. Bratislava: IURA EDITION, 2004. 234 s. ISBN 80-89047-97-1.

[8] ČSN ISO 8258, ČNI 1993


## Adresa autora:


Jarošová Eva, doc., Ing, CSc
Tř. Václava Klimenta 869
293 60 Mladá Boleslav
jarosova@vse.cz

# Adaptive finite volume scheme for 2D embryogenesis image filtering

Z. Krivá* K. Mikula * N. Peyriéras †

**Abstract**

In this paper we explore the effect of the adaptive scheme for image filtering applied to the data representing early stages of zebrafish embryogenesis. The method is based on solution of the regularized Perona–Malik equation and on its discretization by the finite volume method. The adaptive aproach is based on a coarsening strategy based on difference in image intensities and on a quadtree representation of the image. Because, due to filtering, the image intensity tends to be flat in large subregions of the image, it is not necessary to consider same fine resolution in the whole spatial domain and, consequently, our adaptive approach reduces computational effort considerably. In this paper we present 2D algorithms dealing with slices of 3D volume acquired by the multi-photon laser scanning microscopy, a fully 3D approach will be an objective of further study.

## 1 Introduction

An image is usually represented by a real function $u_0(x)$ representing values of greylevel intensity, defined in some rectangular subdomain $\Omega \subset I\!R^d$ (in our case $d = 2$). Then, the image multiscale analysis [1, 3, 11] associates with the initial image $u_0(x) = u(0, x)$ a sequence of images $u(t, x)$, depending on an abstract parameter $t > 0$ called scale. In many practical tasks of image processing, $u(t, x)$ is a solution of a specific, usually second order, nonlinear partial differential equation (PDE). The scale parameter $t$ then can be interpreted as a time in such evolutionary process. The well known

*Department of Mathematics, Slovak University of Technology, Radlinského 11, 813 68 Bratislava, Slovak Republic, kriva,mikula@math.sk

†CNRS-DEPSN, Institut de Neurobiologie Alfred Fessard, Batiment 32-33, Avenue de la Terrasse, 91198 Gif sur Yvette, France, nadine.peyrieras@iaf.cnrs-gif.fr

examples are nonlinear diffusion equations of Perona-Malik type [14, 5] and generalized mean curvature flow equations [2, 12, 8].

In this paper we are dealing with numerical solution to the regularized Perona-Malik problem suggested by Catté, Lions, Morel and Coll in the following form

$$\partial_t u - \nabla.(g(|\nabla G_\sigma * u|)\nabla u) = 0 \quad \text{in } Q_T \equiv I \times \Omega, \tag{1}$$

$$\partial_\nu u = 0 \quad \text{on } I \times \partial\Omega, \tag{2}$$

$$u(0,\cdot) = u_0 \quad \text{in } \Omega, \tag{3}$$

where $\Omega \subset I\!R^d$ is a rectangular domain, $I = [0,T]$ is a time interval, and

(4) $g(s)$ is a decreasing function, $g(0) = 1, 0 < g(s) \to 0$ for $s \to \infty$,

(5) $G_\sigma \in C^\infty(I\!R^d)$ is a smoothing kernel with $\int_{I\!R^d} G_\sigma(x)dx = 1$ and $G_\sigma(x) \to \delta_x$ for $\sigma \to 0$, $\delta_x$ - Dirac function at point $x$,

(6) $u_0 \in L^2(\Omega)$.

The diffusion process of (1) is governed by the shape of function $g$ and by its dependence on $\nabla G_\sigma * u$, an edge indicator. It diffuses image strongly outside edges while the diffusion is supressed across edges.

For the numerical solution of (1)-(3), we use the semi-implicit finite volume method suggested and analysed in [13] and its adaptive version given in [9, 10]. Semi-implicitness of the method means that nonlinearity of the equation is treated from the previous discrete scale step, i.e. the scheme is linear and leads to a solution of sparse linear systems in each discrete scale step of the algorithm.

The success of adaptivity in image processing follows from the observation that solution tends to be flat in large subregions of the image while filtering time is increasing. Due to that fact, we can improve considerably the efficiency of the method using non-uniform grids with decreasing number of finite volumes. Since the whole information about the image is contained in the initial grid and there is no spatial movement of the edges, no refinement is needed and we work just with grids, elements of which are obtained by merging of pixels. Such process is called grid coarsening in numerical methods for solving PDEs and it was introduced to image processing applications in [4].

The adaptive finite volume schemes for image filtering were introduced in [9, 10]. Here, the method is applied to specific type of images which are

given by multiphoton laser scanning microscopy and which represents early stages of the zebrafish embryogenesis. For this type of data, the filtering properties and computational efficiency of the various PDE models have been studied in [15]. In this paper we show that adaptive grid strategy brings further increase of efficiency of computations without any deteriorating of the results.

The rest of paper is organized as follows. In Section 2 we present the idea of coarsening. Section 3 is devoted to finite volume method on non-uniform grids based on coarsening. Section 4 describes application of the adaptive finite volume method to embryogenesis images filtering.

## 2 Coarsening strategy based on quadtrees

In this section we describe how to generate adaptively coarsened grids which are used in discrete scale steps of the computational method. The initial image is given as a set of discrete grey values on pixels of the uniform grid. At the beginning and especially with the increasing scale, we can merge cells using some coarsening criterion and instead on the regular grid we can work on the irregular adaptive structure. For its construction we chose an approach based on quadtrees, where the adaptive grid is represented by the leaves of quadtree. However, instead of organizing resulting structure into a tree (which is known as being inconvenient when access to neighbours is needed) we use a procedural approach and maintain the field of *indicators* which enable us to find out easily whether a given cell or its neighbour can be merged or not. Traversing this structure we stop on a higher level of hierarchy (i.e. on a coarser grid) if the following coarsening criterion is fulfilled. *The cells are merged if difference in intensities is below a prescribed tolerance* $\varepsilon$.

After creating the structure by setting the indicator field we calculate diffusion coefficients by its recursive traversing. In such way we create system of linear equations which is then solved using iterative method with low memory requirements. In order to simplify creating of the matrix of the linear system we require that the ratio of sides of two neighboring squares is $1:1$, $1:2$ or $2:1$. Later, such structure is called *balanced*.

### 2.1 Creating the adaptive grid

Without lost of generality, let us have an image with $2^n \times 2^n$ pixels. It is exactly the situation arising in the embryogenesis 2D image acquisition.

Then the indicator field has dimension $(2^n+1)\times(2^n+1)$ (see Figure 1). To set the values of the indicator field at the beginning we start on the lowest level of the structure (i.e. on the pixel structure of the image). We try to merge cells into $2\times 2$ cells according to the coarsening criterion. In order to perform merging we can use two stencils. One stencil $(2\times 2, 4\times 4, \ldots)$ is moving across the image and other $(3\times 3, 5\times 5, \ldots)$ stencil is moving across the indicator field. Every $2^j\times 2^j$ image stencil has the corresponding $(2^j+1)\times(2^j+1)$ stencil in the indicator field. While neighboring image stencils are not overlapping, their corresponding indicator stencils share the side. With the help of stencils the values in the indicator field are set in such way that

1. they indicate whether the inspected cell on the higher level contains quadruple suitable or not suitable for merging. If intensities in the quadruple are not within the range of $\varepsilon$ then the position in the center of the indicator stencil is set to 1 otherwise it is left 0;

2. they help to maintain the structure balanced. More precisely, after finding out that the inspected quadruple can not be merged, not only the central node of the indicator stencil is set to 1, but this value is set also to the corners of the stencil. Because one of the four corners on lower level becomes middle point of side of the stencil on the higher level we can control merging of the neighboring cells and thus to keep the structure balanced. E.g., structure like in the right part of Figure 2 is not created in the coarsening process.

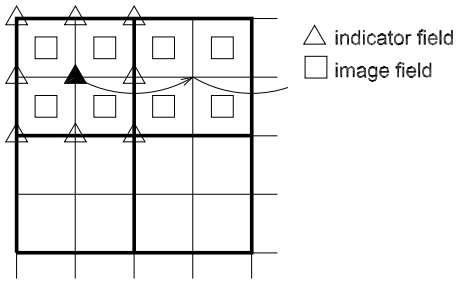



Figure 1: Image field and indicator field together with image stencil and corresponding stencil in the indicator field

If four cells are merged into a larger one then new value, given by the average of old values, is stored in the left lower corner of image stencil corresponding to the cell. This becomes the value representing intensity of merged pixels. Moreover, we remember maximal and minimal values for

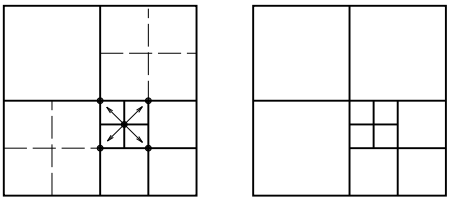


Figure 2: The role of indicator field in balancing of the structure. The grid on the left is balanced, the grid on the right is not balanced and cannot be created in the coarsening process.

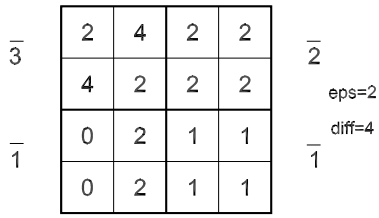



Figure 3: Possible cumulation of the errors, which is supressed in the coarsening process. For $\varepsilon = 2$, the difference in $4 \times 4$ volume is 4.

the merged quadruple in auxiliary fields (they can be free after creating the structure). Testing just intensity differences in the coarsening criterion using recursive process could cause cumulating of errors and in special cases resulting difference could be greater than $\varepsilon$. Such a situation is depicted in Figure 3. All $2 \times 2$ cells fulfill coarsening criterion for $\varepsilon = 2$ and their new values are set to the average. If we test just these new values processing the higher level, the coarsening criterion is fulfilled again. However, the intensity difference of original pixels is 4, twice behind the tolerance. Thus working on higher levels we calculate minimum of all minimal values and maximum of all maximal values for given cells and test their difference.

## 3 Finite volume scheme on adaptive grid

Now we introduce the finite volume computational scheme for solving (1)-(3) on adaptive grid obtained by means of coarsening algorithm described in the previous section. To that goal we adjust the finite volume method given in [13].

Let $\tau_h$ be a uniform mesh of $\Omega$ with cells $p$ of measure $m(p)$ (we assume

rectangular cells here). For every cell $p$ we consider set of neighbours $N(p)$ consisting of all cells $q \in \tau_h$ for which common interface of $p$ and $q$, denoted by $e_{pq}$, is of non-zero measure $m(e_{pq})$.

In the numerical scheme we will provide computations in the series of scale steps starting with $\overline{u}_p^0$, $p \in \tau_h$, corresponding to given intensities on the pixel structure of initial discrete image. We assume

$$\overline{u}_p^0 = \frac{1}{m(p)} \int_p u_0(x)\, dx, \ p \in \tau_h, \tag{7}$$

i.e, the discrete image intensity represents average cell value of the continuous intensity function $u_0(x)$. In the finite volume method, in every subsequent discrete scale step we get again piecewise constant approximations $\overline{u}_p^n$, $p \in \tau_h$, $n = 1, 2, \ldots$ of continuous solution (with possibly the same interpretation as cell averages of continuous solution). Convergence of such approximations to a weak solution of (1)-(3) provided the length of scale step and size of pixel tends to zero is given in [13]. In [13], it is assumed that for every $p$, there exists representative point $x_p \in p$, such that for every pair $p, q, q \in N(p)$, the vector $\frac{x_q - x_p}{|x_q - x_p|}$ is equal to unit vector $n_{pq}$ which is normal to $e_{pq}$ and oriented from $p$ to $q$ (Let us note, that this assumption is not fulfilled for adaptive grids given by the coarsening algorithm). In simple case of uniform grid we can take $x_p$ just as center of the pixel. Then, let $x_{pq}$ be the point of $e_{pq}$ intersecting the segment $\overline{x_p x_q}$. Then we define coefficients

$$T_{pq} := \frac{m(e_{pq})}{|x_q - x_p|} \tag{8}$$

$$g_{pq}^{\sigma,n} := g\left(|\nabla G_\sigma * \tilde{u}\,(x_{pq})|\right) \tag{9}$$

where $\tilde{u}$ is a periodic extension of discrete image computed in $n$-th scale step. The finite volume scheme on uniform grid is then written as follows:
*Let $0 = t_0 \le t_1 \le \ldots \le t_{N_{\max}} = T$ denote the scale discretization steps with $t_n = t_{n-1} + k$, where $k$ is the discrete scale step. For $n = 0, \ldots, N_{\max} - 1$ we look for $\overline{u}_p^{n+1}$, $p \in \tau_h$, satisfying the system of linear equations*

$$\left(\frac{m(p)}{k} + \sum_{q \in N(p)} g_{pq}^{\sigma,n} T_{pq}\right) \overline{u}_p^{n+1} - \sum_{q \in N(p)} g_{pq}^{\sigma,n} T_{pq} \overline{u}_q^{n+1} = \frac{m(p)}{k} \overline{u}_p^n. \tag{10}$$

The scheme (10) is linear **semi-implicit** in scale, since scale derivative is replace by backward difference and nonlinear terms of equation (1) are

treated from the previous scale step while the linear terms are discretized on the current scale level. After such scale discretization, (10) is derived by integrating corresponding elliptic equation over the cell, applying divergence theorem and approximating normal derivative on the boundary of cell by $\frac{u_q - u_p}{|x_q - x_p|}$.

In the scheme (10) we must compute term (9), i.e. the vector

$$\nabla G_\sigma * \tilde{u}\,(x_{pq}) = \left( \frac{\partial (G_\sigma * \tilde{u})}{\partial x}\,(x_{pq})\,, \frac{\partial (G_\sigma * \tilde{u})}{\partial y}\,(x_{pq}) \right),$$

which is an input of the Perona-Malik function $g$. For that goal, we use the following property of convolution

$$\frac{\partial (G_\sigma * \tilde{u})}{\partial x}\,(x_{pq}) = (\frac{\partial G_\sigma}{\partial x} * \tilde{u})\,(x_{pq})\,.$$

Then we get

$$(\frac{\partial G_\sigma}{\partial x} * \tilde{u})(x_{pq}) = \int_{I\!R^d} \frac{\partial G_\sigma}{\partial x}(x_{pq} - s)\tilde{u}\,(s)\,ds = \sum_r \overline{u}_r^n \int_r \frac{\partial G_\sigma}{\partial x}(x_{pq} - s)ds \tag{11}$$

and thus

$$\nabla G_\sigma * \tilde{u}\,(x_{pq}) = \sum_r \overline{u}_r^n \int_r \nabla G_\sigma\,(x_{pq} - s)\,ds \tag{12}$$

where the sum is restricted to control volumes $r$ inside $B_\sigma\,(x_{pq})$, the ball centered at $x_{pq}$ with radius $\sigma$. The ball $B_\sigma$ is given either by a support of compactly supported smoothing kernel or it can represent a "numerical support" of the Gauss function (a domain in which values of the Gauss function are above some treshold given e.g. by a computer precision). In any case just a finite sum in (12) is evaluated and coefficients of this sum, namely $\int_r \nabla G_\sigma\,(x_{pq} - s)\,ds$ can be precomputed in advance using a computer algebra system, e.g. Mathematica. It is worth noting that such approach for evaluation of diffusion coefficient $g_{pq}^{\sigma,n}$ avoids explicit computation of gradients. We use this fact also in adaptive scheme where computation of gradients on non-uniform grid with the so-called "hanging nodes" could cause some difficulties.

It is not possible to apply previous scheme straightforwardly to adaptive non-uniform grids obtained by coarsening algorithm, however, it is possible to modify it. For that goal, we will change a meaning of $x_{pq}$ in (9) and definition (8) of $T_{pq}$. Let in the sequel $x_{pq}$ be the middle point of the common boundary of two neighboring cells (with possibly non-equal measures). The

definition of $g_{pq}^{\sigma,n}$ will then remain the same. The only practical difference will be that the sum in (12) can be evaluated over non-equal control volumes. However, one can precompute all possible coefficients of the sum again in advance for every candidate larger cell on higher levels of hierarchy.

In the definition of $T_{pq}$ in (8), the value $|x_p - x_q|$ represents the distance used for approximation of the normal derivative $\frac{u_q - u_p}{|x_q - x_p|}$. Of course, in case of uniform rectangular grid with unite size of cells, $T_{pq}$ is equal 1. In case of non-uniform rectangular grids, we set this parameter to

$$T_{pq} = min\{l_p, l_q\} \tag{13}$$

where $l_p$ and $l_q$ are lengths of sides of two adjacent cells $p, q$ (of possibly non-equal measure). It is like we assume exchange of intensity between neighbouring cells just in a strip of unit thickness along a boundary of cell.

As our adaptive finite volume schemes we will consider system (10) where $x_{pq}$ represents the middle point of common boundary of two neighboring cells and $T_{pq}$ is given by (13). In every discrete scale step, the scheme gives linear system which is symmetric and strictly diagonally dominant (with positive diagonal and negative numbers out of diagonal) which guarantee existence of its unique solution, for which also $L_\infty$ stability can be easily proved.

# 4 Discussion on numerical experiments dealing with embryogenesis filtering

The adaptive approach is especially efficient if we have large areas of constant intensity in the image domain. The nature of the processed embryomics data and the size of its slices $2^n \times 2^n$ make applying of 2D adaptive algorithms slice by slice very efficient, because many of the slices obtained by microscopy contain only small regions of image information (see Fig.4).

In the slices with only small amount of image information, the *adaptive* grid contains significantly smaller amount of elements than the *nonadaptive grid* - such slice is processed much faster, because the number of unknowns in the linear system (10) is very low. The Fig.5 shows examples of adaptive grids for various slices of the data.

To study the work of the adaptive algorithm we chose a data volume of the size $256 \times 256 \times 30$. The large grid elements, which can be observed especially on the slices with very local image information, correspond to areas with constant intensity, or intensity values within a small tolerance.

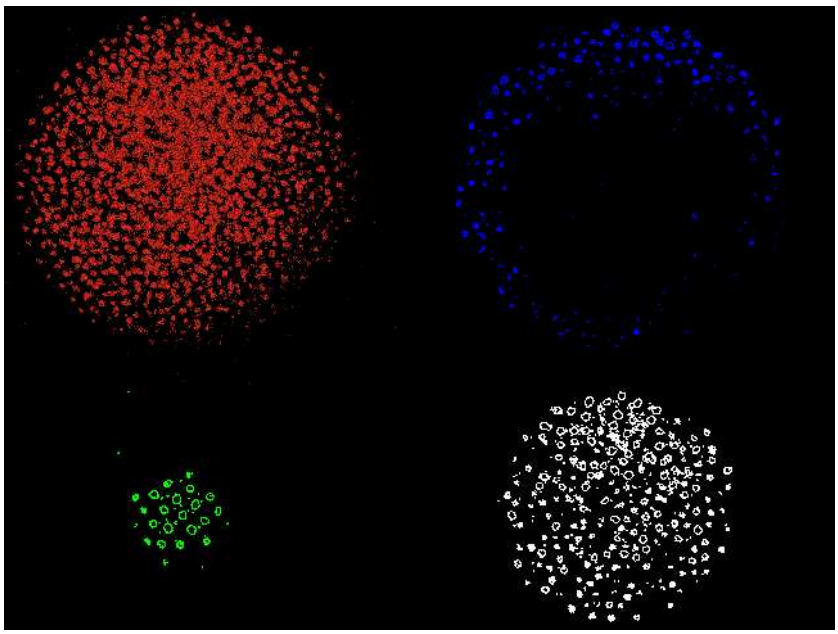


Figure 4: On the top left (in red) the original data viewed from the top. The nuclei of the zebrafish embryo are represented by the isosurfaces of value 40. The rest of pictures represent cuts of the nuclei by three different cutting planes. The green and blue ones are examples of slices where the adaptivity creates large grid elements in areas of constant intesity obtained during smoothing.

In the table depicted in Fig.6 we show an example of decrease of elements during adaptive computation on the data of thesize $256 \times 256 \times 30$. The original number of elements on every slice is 65 536. We performed 10 scale steps of adaptive algorithm and chose 7 slices to demonstrate the decrease of number of of elements during particular scale steps. The slice numbered as 0 is the top slice with mostly localized information, the slice numbered as 29 is the slice with the largest amount of embryo cells. The last column shows, how many percents of the original number of unknowns were used during the solving of linear system at the tenth scale step on every slice. From the practical point of view, we can study the behavior of our adaptive schemes by visual inspection. First, we deal with the data mentioned in the above paragraph. The cuts of the original noisy data are shown in Fig.7. We can observe strong noise and rather bad "separability" in $xy$ and $xz$ directions.

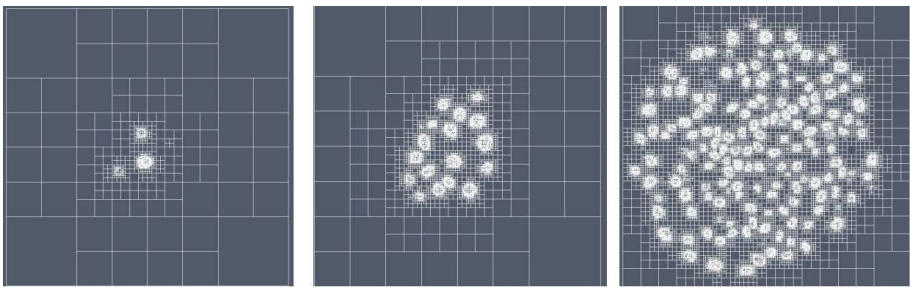


Figure 5: Examples of adaptive grids built on the slices of the 3D data.

Initial number of elements: 65536

| scale step / slice | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. | 9. | 10. | % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 23896 | 15469 | 8170 | 3103 | 1435 | 709 | 301 | 232 | 148 | 148 | 0,23% |
| 5 | 26278 | 17371 | 8926 | 3409 | 1363 | 694 | 415 | 358 | 358 | 331 | 0,51% |
| 10 | 29617 | 19864 | 10486 | 4984 | 2404 | 1489 | 973 | 727 | 640 | 397 | 0,61% |
| 15 | 35974 | 27313 | 17206 | 10351 | 7045 | 5254 | 4294 | 3832 | 3517 | 3226 | 4,92% |
| 20 | 43072 | 35263 | 25711 | 18109 | 13564 | 10939 | 9385 | 8467 | 7933 | 7519 | 11,47% |
| 25 | 49771 | 43021 | 33634 | 25501 | 20197 | 16864 | 14755 | 13321 | 12424 | 11770 | 17,96% |
| 29 | 53947 | 48103 | 38899 | 30106 | 23821 | 19708 | 17176 | 15322 | 14161 | 13354 | 20,38% |

Figure 6: Table showing the decrease of number of elements.

For the adaptive algorithm we used following settings of parameters: $h = 1, K = 200, \sigma = 0.5, \varepsilon = 0.02, k = 0.3, N = 10$, where $N$ is the number of scale steps. The Fig. 8 shows effect of the adaptive algorithm on isosurface smoothing in isosurface representation: we can observe that the finger like noise is removed and the cells surface is smoothed. Also, the separation of cells seems to be good. The data is clipped by a ball of a small radius - that is the reason, why some of the cells are depicted only partially.

The collection of images in Fig.9 tries to do the same using cuts of the selected parts of the 3D data in various directions. The cuts are displayed in pixel level: it means, that no interpolation of intensities is performed. While a couple of slices in $xy$ direction only demonstrates removing of the "small" noise, three couples of cuts in $xz$ and $xy$ directions show, that the smoothed cells, which were originally "connected" by noise are clearly separated by filtering.

To process the piece of the data mentioned above, we needed 76.53s for the nonadaptive version of the Perona-Malik algorithm and 8.95s for

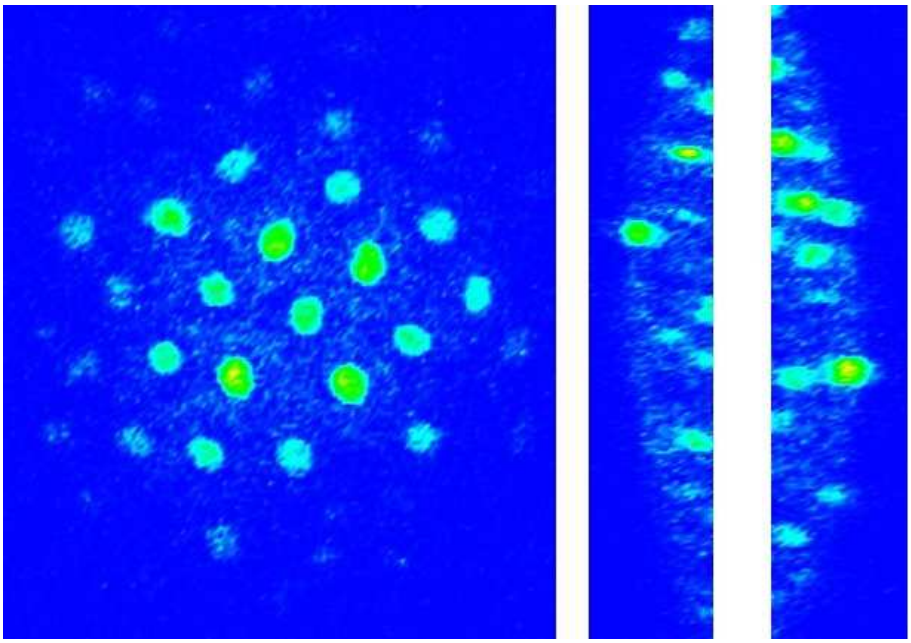


Figure 7: The $xy$,$yz$ and $xz$ 2D cuts through the data of the size $256\times256\times30$ (the intensities are interpolated.)

the adaptive algorithm. The best result for removing noise are usually obtained by the geodesic mean curvature flow filter ([15]), which is slower than Perona-Malik algorithm, so the time speed up is even greater. The question is, if the result of the adaptive algorithm is a good input for further image processing methods, e.g. for algorithms for detecting nuclei centers [6, 7]. Such experiments made on embryogenesis data show that the center detection was equally good and again faster. The pictures (fig. 10) show small box clips of the same data with slightly shifted boxes. They depict the original noisy cells and their centers found on the data filtered by the adaptive algorithm.

## References

[1] L.Alvarez, F.Guichard, P.L.Lions, J.M.Morel, Axioms and Fundamental Equations of Image Processing, *Arch. Rat. Mech. Anal.* 123 (1993) pp. 200-257

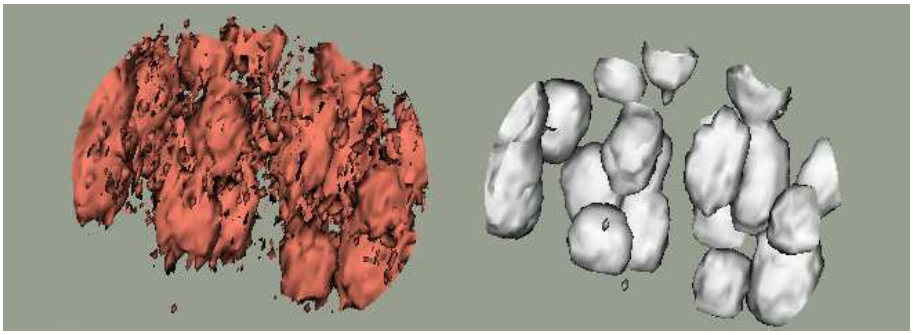


Figure 8: The detail of the data. The noisy data is on the left, the smoothed data is on the right. The value for isosurface is set to 40.

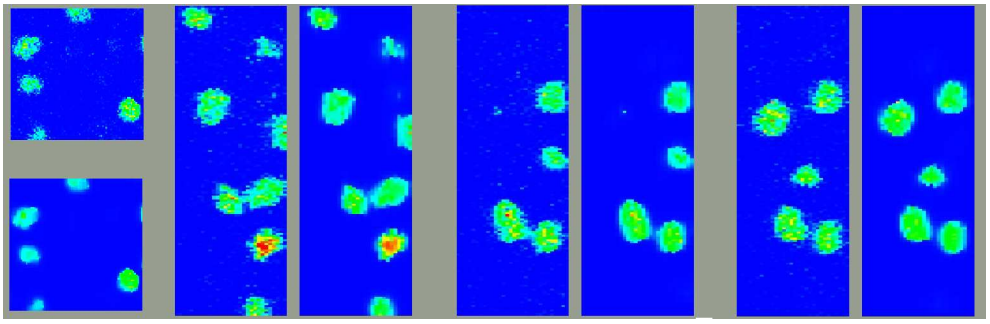


Figure 9: Slices demonstrating removing of the "small" noise and separation of the cells.

[2] L.Alvarez, P.L.Lions, J.M.Morel, Image selective smoothing and edge detection by nonlinear diffusion II, *SIAM J. Numer. Anal.* 29 (1992) pp. 845-866

[3] L.Alvarez, J.M.Morel, Formalization and computational aspects of image analysis, *Acta Numerica* (1994) pp. 1-59

[4] E.Bänsch, K.Mikula, A coarsening finite element strategy in image selective smoothing, *Computing and Visualization in Science*, Vol.1, No.1 (1997) pp. 53-61

[5] F.Catté, P.L.Lions, J.M.Morel, T.Coll, Image selective smoothing and edge detection by nonlinear diffusion, *SIAM J.Numer.Anal.* 29 (1992) pp. 182-193

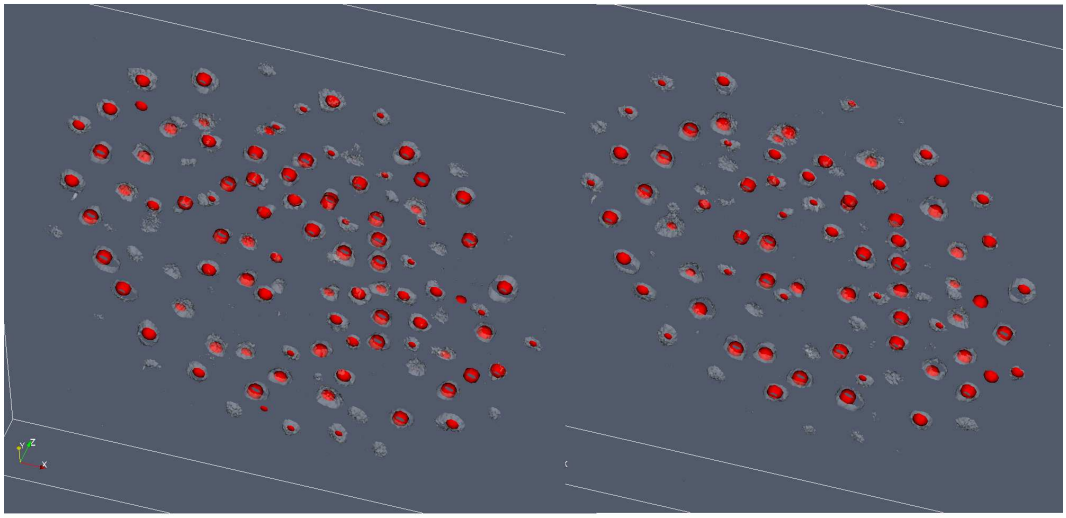


Figure 10: The small clips (slightly shifted) showing the original data and nuclei centers found on the data processed by the adaptive algorithm.

[6] O.Drblikova, M.Komornikova, M.Remešikova, P.Bourgine, K.Mikula, N.Peyrieras, A.Sarti,, Estimate of the cell number growth rate using PDE methods of image processing and time series analysis, *Journal of Electrical Engineering* Vol. 58, No 7/s (2007) pp. 86-92

[7] P.Frolkovič, K.Mikula, N.Peyrieras, A.Sarti, A counting number of cells and cell segmentation using advection-diffusion equations, *Kybernetika*, Vol. 43, No. 6(2007) pp. 817-829

[8] A.Handlovičová, K.Mikula, A.Sarti, Numerical solution of parabolic equations related to level set formulation of mean curvature flow, *Computing and Visualization in Science*, Vol.1, No.3 (1998) pp. 179-182

[9] Z.Krivá, K.Mikula, An adaptive finite volume scheme in processing of color images, *in: Proc. ALGORITMY 2000*, Conf. on Scientific Computing, Podbanské,M Slovakia (2000) pp. 174-188

[10] Z.Krivá, K.Mikula, An adaptive finite volume scheme for solving nonlinear diffusionM equations in image processing, *J. Visual Communication and Image Representation*, Vol. 13 (2002) pp. 22-35

[11] P.L.Lions, Axiomatic derivation of image processing models, *Mathematical Models and Methods in Applied Sciences* 4 (1994) pp. 467-475

[12] R.Malladi, J.Sethian, B.Vemuri, Shape modeling with front propagation: a level set approach, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.17, No.2 (1995) pp. 158-174

[13] K.Mikula, N.Ramarosy, Semi-implicit finite volume scheme for solving nonlinear diffusion equations in image processing, to appear in Numerische Mathematik

[14] P.Perona, J.Malik, Scale space and edge detection using anisotropic diffusion, *Proc. IEEE Computer Society Workshop on Computer Vision* (1987)

[15] B.Rizzi, M. Campagna, C.Zanella, C.Melani, R. Čunderlk, ,Z.Krivá, P.Bourgine, K.Mikula, N.Peyrié ras, A.Sarti, 3D zebrafish embryo image filtering by nonlinear partial diferential equations, *29th IEEE Annual International conference 2008*, No. 1., pp. 6251-6254

# Matematicko-štatistické aspekty spracovania dotazníkových výskumov
# Mathematic-statistical aspects processing of questionnaire researchs

RNDr. Ján Luha,CSc.

**Abstract:** This article deals with Mathematic-statistical aspects in processing of questionnaire researchs. We describe some general principles concerning questionnaire research including methodological aspects.

**Key words:** Mathematic-statistical aspects, processing of questionnaire researchs, general principles of questionnaire research, methodological aspects.

**Kľúčové slová:** Matematicko-štatistické aspekty, spracovanie dotazníkových výskumov, všeobecné princípy dotazníkových výskumov, metodologické aspekty.

## 1. Úvod

Dotazníkové výskumy sa používajú v mnohých oblastiach ako napríklad výskumy verejnej mienky, marketingové prieskumy, rozličné ankety. Štatistické metódy aplikované pri týchto prieskumoch majú v mnohých ohľadoch rovnaký charakter, preto sa v príspevku budeme venovať aplikácii na výskumy verejnej mienky.

Pri príprave, realizácii a vyhodnotení výskumov verejnej mienky sa využíva prakticky celá škála štatistických metód. Vzhľadom na to, že otázky v dotazníkoch pri výskumoch verejnej mienky obsahujú prevažne kvalitatívne štatistické znaky má súbor používaných štatistických metód určité špecifiká. Tieto špecifiká taktiež súvisia s tým, že realizácia výskumov verejnej mienky priamo vychádza z poznatkov výberových štatistických metód.

**Budeme ilustrovať využitie štatistických metód v jednotlivých fázach procesu výskumu verejnej mienky.** Pri ich popise nie je možné urobiť úplný zoznam. Obmedzíme sa na určitý výber metód. Využitie štatistických metód je determinované úlohami, ktoré pomocou týchto metód riešime.

Na štatistické spracovanie sú potrebné dáta, ktoré sú získavané v terénnej fáze výskumu. Najdôležitejšou úlohou terénnej fázy výskumu je získanie kvalitných dát, ktoré spĺňajú také atribúty ako objektivita, validita, reliabilita, reprezentativita a konzistencia. Po získaní empirických dát sú testované hore uvedené atribúty. Tu nachádzajú uplatnenie rôzne obmeny Chí- kvadrát testu, prípadne Fisherovho exaktného testu. Na overovanie relability pri položkovej analýze sa používajú koeficienty reliability ako napr. Cronbachovo alfa.

Na riešenie problému chýbajúcich údajov sa využívajú techniky váženia a imputácie.

Pri tvorbe dotazníka je potrebné vytvoriť na meranie zisťovaných javov vhodné škály. V dotazníku sú reprezentované predpísanými odpoveďami v otázkach. Základné typy škál sú: nominálna, ordinálna a kvantitatívna (numerická). Je dôležité aby škála znaku bola zvolená tak, aby dobre rozdeľovala. Nemá napr. zmysel zaradiť do dotazníka takú otázku na ktorú je možná jediná odpoveď. Pri konštrukcii zložitejších typov najmä ordinálnych škál sa využíva škálogramová analýza, napr. Likertova škála a Likertov koeficient diferenciácie.

Otázky dotazníka (znaky) ďalej delíme na zatvorené, otvorené, polootvorené a otázky s viacerými možnosťami odpovede "**multiresponse**". Otázky na ktoré je dovolené uviesť viac než jednu odpoveď "multiresponse" nie sú štatistické znaky v zmysle definície (štatistický znak je jednoznačná transformácia) a pri ich spracovaní treba využiť adekvátne techniky.

Na štatistické spracovanie dát môžeme použiť všetky dostupné metódy v závislosti od typov škál jednotlivých znakov. Pri stanovovaní základných štatistických charakteristík popisnej štatistiky sú využívané najmä frekvenčné tabuľky, kde je zvláštnosťou spracovanie "multiresponse" otázok. Najprepracovanejšie spracovanie tohoto typu otázok ponúka štatistický software SPSS. Početnosti, či už relatívne alebo absolútne sú počitané na dva základy - počet respondentov, resp. počet odpovedí.

Typické úlohy testovanie hypotéz sú spojené s podielovými veličinami. Napríklad testovanie rozdielu početností, porovnávanie početností, testovanie hypotézy o veľkosti relatívnej početnosti, test dobrej zhody, test nezávislosti, testovanie zhody poradí, overovanie hypotézy o zmene názoru a i.

Pri skúmaní závislostí sa používajú najmä kontingenčné tabuľky a príslušné testy medzi ktorými dominuje Chí-kvadrát test. Najznámejšie štatistické programy ponúkajú aj možnosť výpočtu exakných testov akým je napr. Fisherov exatný test. Môžeme uviesť modul Exact tests od SPSS.

Na analýzu kvantitatívych znakov sa využíva celá škala mnohorozmerných štatistických metód, na analýzu kvalitatívnych znakov sa využívajú postupy ako napríklad korešpondenčná analýza, analýza homogenity, ktorá je určitým analógom faktorovej analýzy pre kvantitatívne znaky, loglineárne modely. Pri konštrukcii modelov mnohorozmernej analýzy kvalitatívnych znakov je jeden z možných prístupov založený na z 0-1 reprezentácie týchto znakov (dichotomizácia, indikátorová premenná).

Dichotomizáciou zobrazíme znak Z pomocou m dichotomických znakov

$Z_i(o)=1$ , ak $Z(o)=i$ , pre $o \in O$,

$Z_i(o)=0$ , ak $Z(o) \neq i$ , i=1, ... ,m.

Dáta za súbor O pre znak Z možno zostaviť do matice X núl a jednotiek typu nxm. Vyjadrenie kvalitatívnych znakov pomocou dichotomizácie umožňuje využívať určité algebraické operácie. Platí:

$1'_n.X=(n_1, ...., n_m)$, kde $1'_n=(1,1, ..., 1)$ je vektor rádu n zostavený z jednotiek. (Výsledok je rozdelenie početností skúmaného znaku.)

$X.1_m=1_n$, kde $1_m$ je vektor rádu m zostavený z jednotiek.

## 2. O výskumoch verejnej mienky

Na korektnú realizáciu výskumov verejnej mienky je potrebné poznať základné zákonitosti, ktoré umožňujú získanie empirických dát, ich štatistickú analýzu a objektívnu interpretáciu. Tieto úlohy rieši metodológia výskumov verejnej mienky, čo je súbor pravidiel a postupov merania spoločenských javov a procesov.

Metódy na získanie empirických dát, ako aj techniky ich realizácie v teréne sú rozličné (pomocou dotazníka, telefonicky, poštová anketa, ...). Medzi najpoužívanejšie techniky výskumov verejnej mienky patrí metóda štandardizovaného rozhovoru pomocou dotazníka realizovaná školenými anketármi (face to face interview). S rozvojom počítačových a informačných technológií sú tieto metódy "vylepšované" elektronickými prostriedkami.

Medzi základné otázky metodológie výskumov verejnej mienky patria otázky navrhovania (tvorby) dotazníkov, metódy zberu empirických dát, metódy matematickoštatistickej analýzy dotazníkových výskumov, objektívna interpretácia a ich prezentácia.

Niektoré postupy, ktoré pripomínajú dnešný výskum verejnej mienky sú známe už z antiky. Napríklad v Aténach okolo roku 500 pred n. l. sa zisťovali názory o tom, či je, alebo nie je ohrozená demokracia a kto ju prípadne ohrozuje. Záujem o empirické poznávanie verejnej mienky sa objavuje začiatkom devätnásteho storočia. Najznámejšie sú pokusy s cieľom predpovedať výsledky amerických prezidentských volieb. Ku skutočnej explózii výskumov došlo v 30. rokoch 20. storočia. Vtedy vznikol najslávnejší inštitút v tomto odbore

American Institute of Public Opinion vedený Georgom Gallupom (1934). **Gallup presadzoval myšlienku reprezentatívneho výskumu založenom na kvótovom výbere a dotazovaní pomocou štandardizovaného rozhovoru.** Rozvoj metód výskumov verejnej mienky súvisí s rozvojom nových poznatkov v oblasti matematickej štatistiky a špeciálne v oblasti teórie výberových metód.

Najčastejšie používané metódy pri výskumoch verejnej mienky sú metóda kvótového a metóda náhodného výberu. Je dôležité, aby bola zabezpečená **reprezentatívnosť** výberovho súboru, teda zhoda s parametrami sledovaného základného súboru. Princíp reprezentatívnosti možno jednoducho parafrázovať výrokom Georga Gallupa: **Ak kuchárka dobre zamieša obsah hrnca, tak jej stačí nabrať a ochutnať iba za lyžicu polievky, aby zistila, ako to vyzerá s obsahom celého hrnca.**

Základné vlastnosti kladené na informácie z empirických výskumov sú objektívnosť, reliabilita (spoľahlivosť), validita (platnosť) a reprezentatívnosť. Pod objektívnosťou informácií rozumieme vlastnosť pravdivo odrážať alebo zachytávať skúmané javy resp. ich znaky. Postup skúmania je objektívny vtedy, keď získané fakty podávajú verný obraz o skúmanom jave. Objektivita je zabezpečená takým postupom, keď sú hodnoty znaku určené jednoznačne a výsledky nie sú závislé od osoby vyhodnocovateľa. Výskumná metóda je reliabilná vtedy, keď zisťuje (meria) tak, že pri opätovnom použití tejto metódy aj pri zisťovaní rôznymi výskumníkmi za inak rovnakých podmienok dostaneme v podstate tie isté výsledky. Validita je zabezpečená keď napr. otázky v dotazníku zisťujú a merajú tie vlastnosti, ktoré skutočne chceme zisťovať. Reprezentatívnosť, už spominaná prv, je zhoda parametrov výberového a základného súboru.

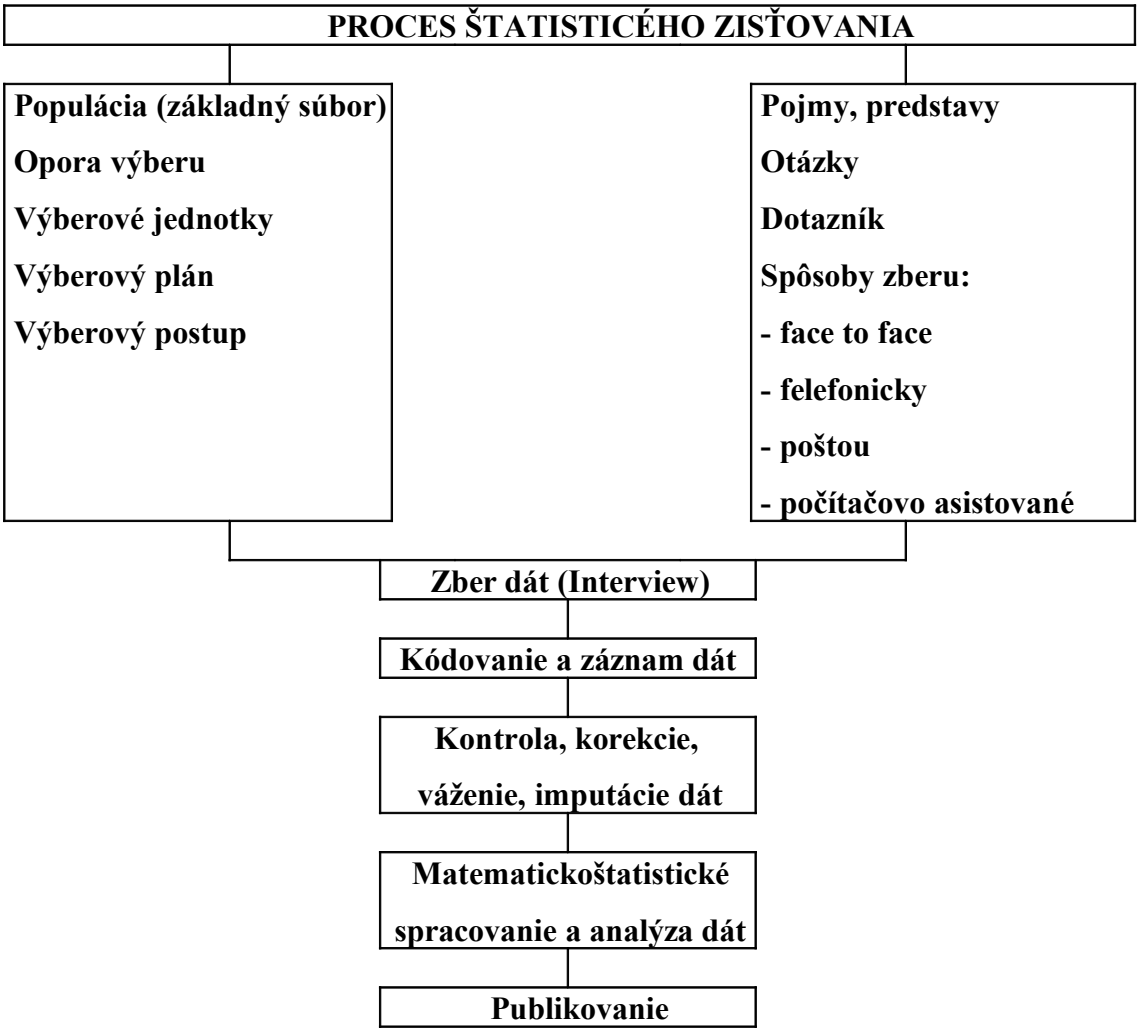
Celý proces prípravy a realizácie výskumu verejnej mienky rozdelíme na jeho základné fázy:

**Základné etapy procesu prípravy a realizácie výskumu verejnej mienky:**

- 1. fáza: Špecifikácia výskumného zámeru a príprava projektu
- 2. fáza: Príprava výskumného dotazníka
- 3. fáza: Projektovanie terénnej fázy
- 4. fáza: Terénna fáza
- 5. fáza: Matematickoštatistická analýza dát
- 6. fáza: Prezentácia výsledkov, publikovanie.

Na základe uvedenej schémy popíšeme najpoužívanejšie štatistické metódy v procese prípravy a realizácie výskumu verejnej mienky.

Inak možno tento proces možno schématicky zobraziť nasledovne:




## 3. Špecifikácia výskumného zámeru a príprava projektu

Ide o zložitú postupnosť. Na základe požiadaviek zadávateľov a mapovania aktuálnych spoločenských, politických, ekonomických udalostí, po dôkladných diskusiách odborníkov je vypracovaný ***projekt výskumu***, ktorý obsahuje: ciele výskumu v rámci ktorých je definovaný záujmový základný súbor (napríklad dospelá populácia SR, mládaž Bratislavy vo veku 15 až 29 rokov), z ktorého budeme vyberať vzorku, opora výberu, výberové jednotky (napr. osoby, domácnosti), výberová vzorka, rozsah výberu, výberový postup (napr. viacstupňový náhodný výber, kvótový výber), estimátor, ktorý treba počítať, metóda zberu dát (napríklad štandardizovaný rozhovor pomocou dotazníka), kontrola práce anketárov a pomoc pri práce v teréne, obdobie zberu dát (terénna fáza), konštruuje sa dotazník.

Na ilustráciu niektorých štatistických metód budeme uvažovať malú krajinu S, v ktorej žije 5 398 657 obyvateľov (na základe bilancie pohybu obyvateľstva - stav ku 1.1.2000). Za základný úbor zvolíme dospelú populáciu občanov tejto krajiny (tj. vo veku 18 a viac rokov). Na základe uvedených údajov súbor dospelej populácie našej krajiny S má N=4 069 739 prvkov, z ktorých budeme vyberať našu vzorku. Základný súbor má nasledovné charakteristiky:

| pohl | počet | % |
|---|---|---|
| **muži** | **1941804** | 47,80 |
| **ženy** | **2120935** | 52,20 |
| **spolu** | **4062739** | **100,00** |

| vso | počet | % |
|---|---|---|
| **do 2** | 1238683 | 30,49 |
| **2-10** | 808300 | 19,90 |
| **10-50** | 984352 | 24,23 |
| **50-100** | 490360 | 12,07 |
| **nad 100** | 541044 | 13,32 |
| **spolu** | **4062739** | **100,00** |

| kraj | počet | % |
|---|---|---|
| **Bratislavský** | 487124 | 11,99 |
| ***z toho Ba*** | 357301 | 8,79 |
| **Trnavský** | 421767 | 10,38 |
| **Trenčiansky** | 464123 | 11,42 |
| **Nitriansky** | 552270 | 13,59 |
| **Žilinský** | 511014 | 12,58 |
| **Banskobystrický** | 504771 | 12,42 |
| **Prešovský** | 556866 | 13,71 |
| **Košický** | 564804 | 13,90 |
| ***z toho Ke*** | 183743 | 4,52 |
| **spolu** | **4062739** | **100,00** |

| kraj | v1824 | v2529 | v3039 | v4049 | v5059 | v60pl | spolu |
|---|---|---|---|---|---|---|---|
| **Bratislavský** | 73893 | 46381 | 86604 | 107471 | 74502 | 98273 | 487124 |
| *z toho Ba* | *52859* | *32463* | *63056* | *81186* | *55556* | *72181* | *357301* |
| **Trnavský** | 68387 | 44212 | 78097 | 85654 | 60610 | 84807 | 421767 |
| **Trenčiansky** | 74757 | 46601 | 86252 | 93622 | 65747 | 97144 | 464123 |
| **Nitriansky** | 84696 | 55202 | 100333 | 110889 | 78985 | 122165 | 552270 |
| **Žilinský** | 85603 | 55425 | 97191 | 102823 | 68475 | 101497 | 511014 |
| **Banskobystrický** | 78659 | 49780 | 94004 | 102421 | 71959 | 107948 | 504771 |
| **Prešovský** | 97469 | 61476 | 110297 | 108623 | 71519 | 107482 | 556866 |
| **Košický** | 93202 | 60143 | 107895 | 114199 | 77790 | 111575 | 564804 |
| *z toho Ke* | *29212* | *20274* | *35856* | *38900* | *28363* | *31138* | *183743* |
| **spolu** | **656666** | **419220** | **760673** | **825702** | **569587** | **830891** | **4062739** |

Na ilustráciu sme uviedli charakteristiky vybraných znakov základného súbor dospelej populácie, ktoré tvoria súčasť výskumného zámeru a na základe ktorých stanovujeme parametre výskumu.

## 4. Príprava výskumného dotazníka

;Nebudeme rozoberať tvorbu dotazníka – poukážeme na štatistické aspekty tvorby otázok. Dotazník dostávajú anketári aby pomocou neho v prípade metódy štandardizovaného rozhovoru realizovali rozhovory, alebo pri „poštovom zisťovaní“ tento dotazník vyplnili respondenti, alebo slúži ako podklad pre záznamový formulár pri telefonickom dotazovaní atď. Tvorba dotazníka je náročná a zložitá práca.

Čo je to dobrá otázka? Dobrá otázka dáva odpovede, ktoré sú objektívne, reliabilné a validné.

### Štatistické kritériá tvorby otázok

Štatistický súbor sa skladá zo štatistických jednotiek. Na týchto jednotkách pozorujeme (meriame) hodnoty štatistických znakov. Zisťovanie hodnôt štatistických znakov nazývame tiež **meranie**.

Napr. pri dotazníkových výskumoch je dôležité, aby anketári postupovali rovnakým spôsobom - aby výsledky neboli ovplyvnené ich osobným prístupom (metodické pokyny na realizáciu zberu dát, školenia anketárov, (kalibrácia siete), ...).

Pre metódy zberu dát je tiež dôležité, aké typy štatistických znakov zisťujeme.

Rozlišujeme rôzne typy znakov a **rôzne typy škál** ich merania:
Nominálny: - škála merania nominálna napr. farba očí, pohlavie, profesia, národnosť a i.
Ordinálny: - škála je usporiadaná množina hodnôt. Kvalitatívny ordinálny: hodnosť v armáde. Kvantitatívny ordinálny: prospech.
Intervalový kvantitatívny znak: - škála je invariantná na lineárnu transformáciu: y=ax+b, b<>0. Príklad: teplota v °C, výška, hmotnosť, obvod hrudníka a i. (v závislosti od merných jednotiek).
Pomerový kvantitatívny znak: - škála je invariantná na lineárnu transformáciu: y=ax. Má absolútnu nulu. Napr. teplota v °K, vek, dĺžka školskej dochádzky. Možno počítať podiely.
Absolútna: - počet prvkov s danou vlastnosťou. Napr. počet opravených áut v danej opravovni.

Z definície štatistického znaku, ako jednoznačnej transformácie:
O $\Rightarrow$ H, kde O označuje množinu objektov (respondentov) a H označuje množinu odpovedí („hodnôt“) na danú otázku, dostávame **základné (formálne) štatistické požiadavky, ktoré je potrebné rešpektovať pri konštrukcii škál otázok dotazníka:**

- jednoznačnosť (zaručuje disjunktný rozklad množiny objektov),
- úplnosť (zaručuje zahrnutie všetkých možných hodnôt).

Obmedzíme sa ďalej na nominálne a ordinálne znaky - čo sú najfrekventovanejšie možnosti otázok dotazníka v sociologických výskumoch.
**Úplnosť:** škála znaku (otázky) musí byť konštruovaná tak, aby zahŕňala všetky možnosti. Majme napr. znak vzdelanie s kategóriami: základné, stredoškolské bez maturity, stredoškolské s maturitou a vysokoškolské. Ak by sme v dotazníku vynechali čo len jeden variant nemáme úplnú škálu.
**Jednoznačnosť (disjunktnosť)**: Jeden objekt nemôže mať priradenú viac než jednu hodnotu.
Poznámka: V sociologických výskumoch sa často stretávame s otázkami, na ktoré možno odpovedať viacerými možnosťami. takéto otázky nie sú štatistické znaky v zmysle definície štatistického znaku. Takéto otázky majú špecifické metódy spracovania.

**Formálna úprava** dotazníka musí byť prehľadná - aby sa minimalizovali chyby pri kódovaní dotazníkov, ako aj pri ich zázname do počítača. Okrem takejto úpravy musí dotazník obsahovať **identifikačné znaky**, ktoré sú pomôckou pri kontrolách: číslo dotazníka, číslo anketára. Dotazník musí, okrem **meritórnych znakov** (otázok) obsahovať patričné **demografické znaky** podľa skúmanej témy. Napr. pohlavie, vek, vzdelanie, národnosť, socio-profesijná skupina, veľkostná skupina obce, kraj (oblasť) a pod.

Je dobré v jednej otázke zachytávať jedno hľadisko. Pri kombinácii hľadísk je potrebné kombinácie variantov znakov "násobiť". Príklad:

| | | | | |
|---|---|---|---|---|
| | 2 | | | Ovplyvnila alebo neovplyvnila prebiehajúca verejná diskusia o vstupe Slovenska do NATO Váš postoj k referendu? |
| | | 1 | - | ovplyvnila, nechcel som sa ho zúčastniť, ale zúčastním sa ho |
| | | 2 | - | ovplyvnila, chcel som sa ho zúčastniť, ale nezúčastním sa ho |
| | | 3 | - | neovplyvnila, chcel som sa ho zúčastniť a zúčastním sa ho |
| | | 4 | - | neovplyvnila, nechcel som sa ho zúčastniťa nezúčastním sa ho |
| | | 5 | - | neviem posúdiť |

Niekedy môže byť výhodné dva alebo viac znakov kombinovať do jedného bez straty informácie. Napr.:

X1 { 0 v rodine nemajú deti
1 v rodine majú deti

X2 vek najstaršieho dieťaťa.
Možno kombinovať do:
Y1 vek najstaršieho dieťaťa, pritom kód 0 znamená, že nemajú deti. Redukuje sa počet otázok a eliminujú chýbajúce dáta v X2.
V dotazníku sa môžu vyskytnúť otázky, na ktoré neodpovedajú všetci respondenti na takúto otázku aplikujeme **filter**.
Otázky ďalej členíme na uzavreté, otvorené, polootvorené.

**Príklady:**

**Uzavretá otázka:**

| | | | |
|---|---|---|---|
| | | | Ak by ste sa zúčastnili na večierku, alebo v spoločnosti, kde by vám ponúkli drogu, vzali by ste si? |
| | 01 | - | určite áno |
| | 02 | - | asi áno |
| | 03 | - | asi nie |
| | 04 | - | určite nie |
| | 05 | - | neviem |

**Otvorená otázka:**
Uvedieme príklad otvorenej multiresponese otázky – pre túto je potrebné pre spracovanie pripraviť kódovací kľúč, tým ju vlastne „uzatvoríme“ a môžeme zaznamenať dáta:

| | |
|---|---|
| | Ktorý z politikov pôsobiacich na Slovensku má v súčasnosti vašu dôveru ?<br>**Možno uviesť najviac tri osobnosti.**<br>.....<br>.....<br>..... |

**Polootvorená otázka**:

| | | | |
|---|---|---|---|
| | | | **Národnosť:** |
| | 01 | - | slovenská |
| | 02 | - | maďarská |
| | 03 | - | iná (uveďte) ..... |

Okrem toho môžu byť do dotazníka zaradené aj **numerické** otázky, napríklad o veku, príjme, o počte členov domácnosti a pod.
Špecifikom sú otázky s možnosťou viac odpovedí „multiresponse“- príklad:

| | | | |
|---|---|---|---|
| | | | Povedzte, prosím, máte alebo nemáte v domácnosti, v ktorej žijete človeka, o ktorom sa dá povedať, že je závislý od niektorej z nasledujúcich „drog“? **Možno uviesť viac odpovedí**. |
| | 01 | - | od hracích automatov |
| | 02 | - | od Internetu |
| | 03 | - | od aktívneho športovania |
| | 04 | - | od mobilného telefónu |

| | | | |
|---|---|---|---|
| | 05 | - | od jedla (sladkostí) |
| | 06 | - | od televízie |
| | 07 | - | od alkoholu |
| | 08 | - | od cigariet, tabaku |
| | 09 | - | od iného (uveďte)..... |
| | 10 | - | nemáme takého v domácnosti |

Je dôležité aby škála znaku bola zvolená tak, **aby dobre rozdeľovala**. Nemá napr. zmysel zaradiť do dotazníka takú otázku na ktorú je možná jediná odpoveď. Definujeme pojem **zle rozdeľujúci znak**: Znak Z zle rozdeľuje súbor dát, ak existuje jediná hodnota j, taká, že hypotézu $H_0^j$: $p_j=0$ zamietame.
V prípade keď kategórie znaku majú malú početnosť zlúčime ich podľa logiky problému.

## 5. Projektovanie terénnej fázy

V tejto fáze sa stanovia (na základe cieľov výskumu) výberová vzorka, metóda zberu dát a inštrukcie na realizáciu terénnej fázy. Cieľom tejto fázy je získanie objektívnych dát o skúmanej problematike. Základné vlastnosti kladené na informácie z empirických výskumov, ako sme to spomenuli už prv, sú objektívnosť, reliabilita (spoľahlivosť), validita (platnosť), reprezentatívnosť a konzistentnosť.

**Výberová vzorka**

Najčastejšie používané metódy získania výberového súboru pri výskumoch verejnej mienky sú metóda kvótového a metóda náhodného výberu. Je dôležité, aby bola zabezpečená reprezentatívnosť výberovho súboru, teda zhoda s parametrami sledovaného základného súboru. Princíp **reprezentatívnosti** možno jednoducho parafrázovať výrokom Georga Gallupa: Ak kuchárka dobre zamieša obsah hrnca, tak jej stačí nabrať a ochutnať iba za lyžicu polievky, aby zistila, ako to vyzerá s obsahom celého hrnca.

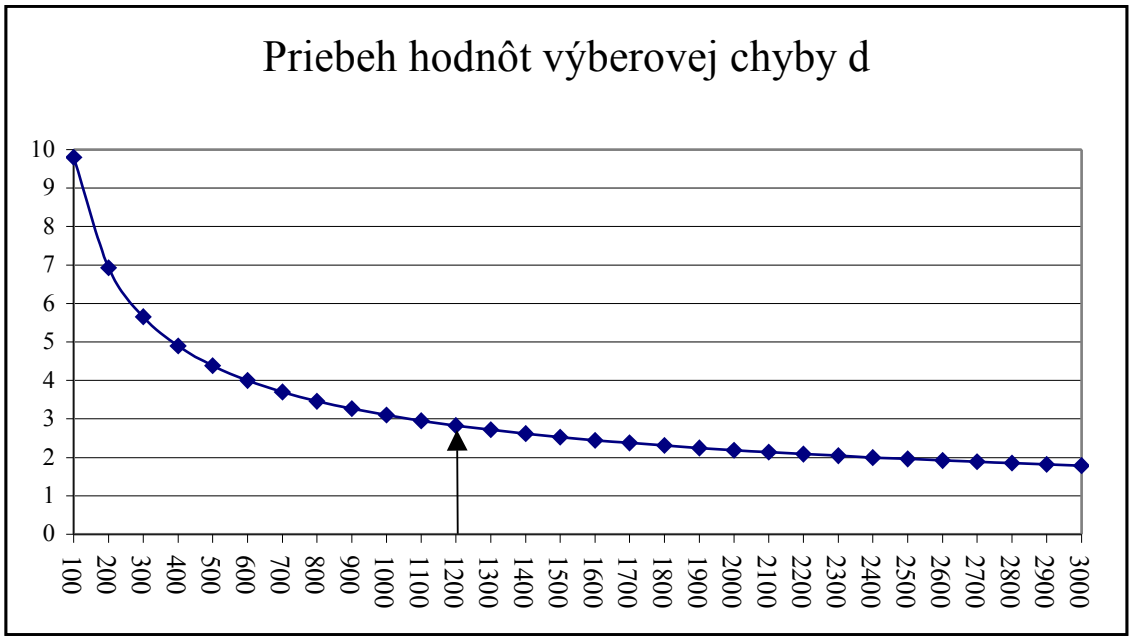
Na základe aktuálnych údajov zo sčítania, prípadne z bilancie pohybu obyvateľstva zistíme štruktúru základného súboru podľa demografických znakov: pohlavie, vek, národnosť, vzdelanie, veľkostná skupina obce a oblasť SR. Pre metódu kvótoveho výberu tak získame podklady na výpočet kvót výberovej vzorky, podľa ktorej sa potom spracuje rozpis počtu rozhovorov podľa jednotlivých kategórií demografických znakov do Smernice pre výskum (príklad uvádzame v 3. kapitole). Podľa tejto Smernice potom anketári vyhľadávajú respondentov. Pri metóde náhodného výberu slúžia údaje o štruktúre základného súboru na stanovenie oblastí, spravidla podľa veľkostných skupín obcí a oblastí (krajov) SR, v ktorých sa náhodne vyberajú určené počty respondentov.

**Výberová vzorka je ako tá lyžica polievky, ktorú ochutnáva kuchárka. Obsah hrnca je skúmaný základný súbor. Správne stanovenie kvót a ich dodržanie anketármi predstavuje varechu na dobré premiešanie obsahu hrnca - nášho základného súboru.**

Po dobrom "zamiešaní" získavame spoľahlivé výsledky. Presnosť výsledkov meriame výberovou chybou. Výberová chyba, pri dobre realizovanom zbere dát, závisí od rozsahu výberovej vzorky. Problematika merania spoľahlivosti je zložitá. Ilustrujeme ju jednoduchším spôsobom určenia výberovej chyby cez interval spoľahlivosti podielu (percenta) variantu štatistického znaku. Ak uvažujeme koeficient spoľahlivosti 95%, tak interval spoľahlivosti (výberovú chybu) možno aproximačne počítať pomocou vzťahu:

$$d= +/- 100*1,96\sqrt{p*(1-p)/n}.$$

Pri rozsahu výberovej vzorky nad 1000 osôb je výberová chyba menšia než 3%. Priebeh veľkosti výberovej chyby v závislosti od rozsahu výberu je znázornený na grafe.




Pri výskumoch verejnej mienky sa uspokojujeme s výberovou chybou +/- 3% a preto obvykle volíme rozsah výberu n=1400 respondentov. V tejto súvislosti často dostávame otázky:

1. **Prečo sa mňa nikto nepýta na moju mienku?**
2. **Ako je možné, že taká malá vzorka môže byť obrazom veľkého celku?**

Na prvú otázku možno ľahko odpovedať takto: Ak skúmame základný súbor dospelých obyvateľov SR, ktorých je 4 062 739, tak jedneho respondenta vyberáme približne z 2902 ľudí. Ak by sme realizovali mesačne 2 výskumy, tak za predpokladu, že sa niekto nedostane do výberu opakovane, je šanca byť vybraný ako respondent ÚVVM zhruba raz za 120 rokov. Na druhú otázku dáva odpoveď teória výberových metód. Zjednodušene na túto otázku možno odpovedať Gallupovým výrokom, ktorý sme už citovali.

**Výberové metódy**

Metódy získavania empirických dát sú determinované rozsahom skúmaného záujmového základného súboru. Pri základnom súbore malého rozsahu je možné realizovať vyčerpávajúce zisťovanie za všetky štatistické jednotky a nie je potrebný náhodný výber. Obvykle však musíme pristúpiť k výberu jednotiek, za ktoré budeme získavať údaje.

Základnými pojmami ako základný súbor, výberový súbor, opora výberu a i. sa nebudeme zaoberať.

Stručne sa zmienime o metódach zostavovania výberového súboru: panely, itineráre, metóda typickej jednotky, náhodný výber, kvótový výber.

Pri serióznych sociologických (empirických) výskumoch je potrebné, aby výberový súbor bol reprezentatívny. Náhodný výber je reprezentatívny na základe teoretických vlastností. Pri rozsiahlom základnom súbore  však náhode treba pomôcť a preto sa konštruujú zložitejšie typy náhodných výberov. Napr. oblastný, viacstupňový, výber skupín atď. Pri iných typoch výberu sa reprezentatívnosť dosahuje rozličnými spôsobmi. Pri kvótových výberoch reprezentatívnosť výberového súboru (definovanú ako zhodu parametrov kvótových znakov výberového súboru s parametrami základného súboru) zabezpečujeme rozpisom kvót počtu rozhovorov.

Zber dát je nejdelikátnejšou a najnákladnejšou operáciou a ani najpremyslenejšie metódy matematickej štatistiky naznamenajú nič, ak sa uplatňujú na informáciách zaťažených veľkými chybami. Teória výberových zisťovaní má snahu poskytnúť metodologickú oporu pre nájdenie najprimeranejších prostriedkov zberu informácií.

Rozlišujeme dve skupiny výberových metód:

- empirické
- náhodné.

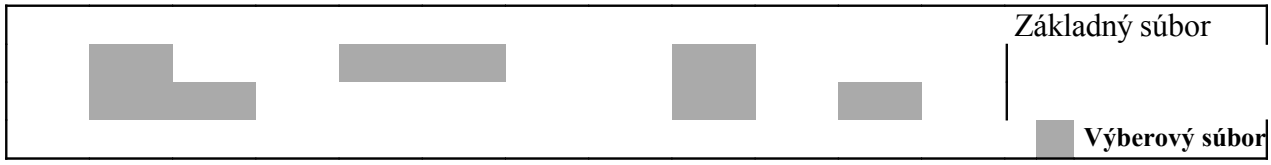
Empirické metódy neuvažujú oporu výberu priamo, ale využívajú čiastkové informácie známe vopred. Patria sem kvótový výber, metóda typických jednotiek, itineráre a i. Ich výhodou je jednoduchosť a nízke náklady.

**Metódy náhodného výberu** sa nazývajú tak preto, lebo sa predpokladá, že každá jednotka má nenulovú a známu pravdepodobnosť zahrnutia do výberu. Delíme ich na jednoduché - plán výberu sa týka skúmaného základného súboru ako celku - a viacstupňové - základný súbor je najprv rozdelený na podsúbory, ktoré sa nazývajú primárne jednotky. Vyberajú sa najprv primárne jednotky a v nich potom sekundárne jednotky atď. Zvláštne prípady viacstupňových výberov:

- Oblastný výber - do výberu sú zahrnuté všetky primárne jednotky.
- Výber skupín - vyberú sa náhodne primárne jednotky a v týchto vybraných primárnych jednotkách sa vyberú všetky sekundárne jednotky.

Schématické znázornenie základných typov náhodných výberov:

Jednoduchý náhodný výber




Dvostupňový výber Primárne jednotky


Oblastný výber Primárne jednotky


Výber skupín Primárne jednotky


Spôsoby získania náhodného výberu: Generovanie náhodných čísel, náhodné prechádzky, systematický výber jednotiek ... .

**Empirické výberové metódy** používame, keď nemáme k dispozícii oporu výberu, ale len niektoré informácie o skúmanom záujmovom základnom súbore. Vychádza sa z predpokladu, že ak je určitá charakteristika základného súboru známa a získame výberový súbor, ktorý je reprezentatívny pre túto charakteristiku, tak bude reprezentatívny aj pre charakteristiku, ktorú skúmame. Opierame sa o tieto informácie, aby sme získali reprezentetívny výberový súbor.

Empirické metódy vyžadujú veľké skúsenosti, zdravý rozum a objektivitu. Najdôležitejšiou z empirických výberových metód je kvótový výber.

**Kvótový výber**. Používa sa veľmi často pri výskume trhu a výskume verejnej mienky. Princíp: Je známe rozdelenie sledovanej populácie (napr. dospelí občania SR) pre demografické znaky (pohlavie, vek, vzdelanie, národnosť, veľkosť obce, kraj). Ak je výberový súbor reprezentatívny podľa vybraných kontrolovaných znakov, tak predpokladáme, že bude reprezentatívny aj pre zisťované meritórne znaky.

Výberová vzorka sa vypočíta podľa štruktúry kotrolovaných znakov v základnom súbore.

Výberový súbor obvykle zostavíme metódou voľného kvótového výberu. Táto metóda zabezpečuje reprezentatívnosť marginálne za jednotlivé kontrolované znaky. Teoreticky si možno predstavovať, že navzájom zviazané (krížové) kvóty sú vhodnejšie, ale:

-problém je v raste počtu kombinácií. Ak uvažujeme 6 kontrolovaných znakov s počtom variantov: 2,6,4,3,5,8, tak máme 5760 kombinácií. Ak stanovíme rozsah výberu n=1400, tak v krížových kvótach nemôžeme zabezpežiť všetky kombinácie.

- práca s marginálnymi kvótami je rýchlejšia.

- veľmi často nie sú krížové rozdelenia k dispozícii, marginálne však áno.

- experimentálne štúdie ukázali, že krížové kvóty neposkytujú v porovnaní s marginálnymi podstatný prírastok presnosti.

Bežne používané kvótové znaky sú:

pohlavie (01=muž, 02=žena),

vek (01=18-24, 02=25-29, 03=30-39, 04=40-49, 05=50-59, 06=60 a viac rokov),

vzdelanie (01=základné, 02=stredné bez maturity, 03=stredoškolské s maturitou, 04=vysokoškolské),

národnosť (01=slovenská, 02=maďarská, 03=iná),

veľkostná skupina obce (01=do 1999 obyv., 02=2000- 9999, 03=10000-49999, 04=50000-9999 05=nad 100000),

kraj (01=Bratislavský, 02=Trnavský, 03=Trenčiansky, 04=Nitriansky, 05=Žilinský, 06=Banskobystrický, 07=Prešovský, 08=Košický).

Príklad: Kvóty počtu rozhovorov podľa krajov a pohlavia. Rozsah vzorky n=1400. Základný súbor dospelá populácia SR.

| Kraj | Počet rozhovorov | Počet dospelých obyvateľov | % dospelých obyvateľov |
|---|---|---|---|
| Bratislavský | 167 | 464809 | 11,95 |
| Trnavský | 145 | 402542 | 10,35 |
| Trenčiansky | 160 | 444527 | 11,43 |
| Nitriansky | 193 | 534562 | 13,75 |
| Žilinský | 175 | 487175 | 12,53 |
| Banskobystrický | 176 | 488554 | 12,56 |

| | | | |
|---|---|---|---|
| Prešovský | 190 | 526745 | 13,54 |
| Košický | 194 | 540026 | 13,89 |
| SR | 1400 | 3888940 | 100,00 |

| Pohlavie | Počet rozhovorov | Počet dospelých | % dospelých |
|---|---|---|---|
| muž | 669 | 1858464 | 47,79 |
| žena | 731 | 2030476 | 52,21 |
| SR | 1400 | 3888940 | 100,00 |

Po výpočte výberovej vzorky sa vypracuje rozpis počtu rozhovorov pre jednotlivých anketárov.

**Príklad: Rozpis počtu rozhovorov pre jednotlivých anketárov.**

| ANKET | VSO | POCET | MUZI | ZENY | 1824 | 2529 | 3039 | 4049 | 5059 | 60+ | SLOV | MAD | INA | ZAKL | UCN | STRSM | VYS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 7 | 3 | 4 | 1 | 0 | 2 | 1 | 1 | 2 | 7 | 0 | 0 | 1 | 2 | 2 | 2 |
| 2 | 5 | 7 | 4 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 6 | 0 | 1 | 2 | 1 | 2 | 2 |
| 3 | 5 | 7 | 3 | 4 | 1 | 1 | 2 | 1 | 1 | 1 | 6 | 1 | 0 | 1 | 2 | 3 | 1 |
| . | | | | | | | . | | | | | | | | | | . |
| . | | | | | | | . | | | | | | | | | | . |
| 17 | 5 | 8 | 4 | 4 | 1 | 1 | 2 | 2 | 1 | 1 | 7 | 1 | 0 | 1 | 2 | 3 | 2 |
| 18 | 1 | 7 | 3 | 4 | 1 | 1 | 2 | 1 | 1 | 1 | 6 | 0 | 1 | 2 | 2 | 2 | 1 |
| 19 | 2 | 7 | 4 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 7 | 0 | 0 | 3 | 2 | 1 | 1 |
| 20 | 2 | 7 | 3 | 4 | 1 | 1 | 1 | 1 | 1 | 2 | 6 | 0 | 1 | 3 | 2 | 2 | 0 |
| 21 | 3 | 7 | 4 | 3 | 1 | 0 | 1 | 2 | 1 | 2 | 7 | 0 | 0 | 3 | 2 | 1 | 1 |
| 22 | 3 | 7 | 4 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 6 | 0 | 1 | 3 | 2 | 1 | 1 |
| 23 | 3 | 7 | 3 | 4 | 1 | 0 | 1 | 2 | 1 | 2 | 4 | 3 | 0 | 3 | 2 | 2 | 0 |
| 24 | 1 | 7 | 4 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 7 | 0 | 3 | 2 | 2 | 0 |
| 25 | 1 | 7 | 4 | 3 | 1 | 0 | 1 | 2 | 1 | 2 | 1 | 6 | 0 | 3 | 2 | 2 | 0 |
| . | | | | | | | . | | | | | | | | | | . |
| . | | | | | | | | | | | | | | | | | |

Anketári sú alokovaní po celom území SR tak, aby bola implicitne zabezpečená reprezentatívnosť za znaky veľkostná skupina obce a kraj.

Počty rozhovorov podľa ostatných kontrolovaných anketárov sú na základe rozpisu oznámené anketárom v Smernici k danej výskumnej úlohe.

**Príklad smernice pre anketára:**

Anketár: XXX YYYYY
Číslo anketára: 12345

**S M E R N I C E**
***pre výskum č.E01/2002***

Podľa všeobecných pokynov uvedených v Príručke pre anketára, vyhľadajte **7** osôb tak, aby medzi nimi bolo:

| **Počet** | **Miesto pre čiarky** | **Miesto pre krížiky** |
|---|---|---|

| | | (uskutočnené rozhovory) | (odmietnutia) |
|---|---|---|---|
| Mužov | **4** | ////..... | +++..... |
| Žien | **3** | ///..... | ++..... |
| ***Vo veku*** | | | |
| 18 až 24 rokov | **1** | /..... | + ..... |
| 25 až 29 rokov | **1** | /..... | ..... |
| 30 až 39 rokov | **2** | //..... | ++ ..... |
| 40 až 49 rokov | **1** | /..... | +..... |
| 50 až 59 rokov | **1** | /..... | . ..... |
| 60 rokov a viac | **1** | /..... | ..... |
| ***Národnosti*** | | | |
| slovenskej | **5** | ..... | ++..... |
| maďarskej | **1** | ..... | ++..... |
| inej | **1** | ..... | ..... |
| ***So vzdelaním*** | | | |
| základným | **2** | ..... . | + +..... |
| stredoškolským bez maturity | **2** | ..... . | +..... .. |
| stredoškolským s maturitou | **2** | ..... . | +..... .. |
| vysokoškolským | **1** | ..... . | +..... ... |

**Stanovenie rozsahu výberového súboru**

Rozsah výberového súboru stanovujeme z analýzy výberovej chyby, požadovanej presnosti odhadov a ďalších podmienok ako napr. cena výskumu.

**Rozsah výberového súboru** závisí od cieľov výskumu, od rozdelenia (i keď neznámeho) pravdepodobnosti skúmaných znakov a akú presnosť odhadu charakteritík budeme vyžadovať. Dôležité je aké triediace hľadiská budeme využívať pri spracovaní a interpretácii - do akého stupňa podrobnosti chceme realizovať triedenia podľa demografických znakov. Ak nás zaujíma reprezentatívnosť za celý skúmaný súbor, tak je potrebný rozsah výberového súboru menší.

Ak uvažujeme napríklad najjednoduchší znak typu áno / nie (podobne napr. pohlavie), tak vzorec na určenie rozsahu výberu s opakovaním je, pri úrovni významnosti 5% (za predpokladu npq > 9):

$$n=(1,96/d)^2p.q\ ,$$

kde q=1-p a d je predpokladaná presnosť odhadu podielu p výskytu hodnoty áno (alebo napr. muž v znaku pohlavie).

Ak uvažujeme výber bez opakovania zo základného súboru o veľkosti N, tak vzorec na výpočet rozsahu výberu má tvar:

$$n=(1,96/d)^2pq\ (1 - f),$$

kde f=n/N je výberový pomer.

**Príklad**: Nech d=0,03 a p=0,5, tak rozsah výberu pri výbere s opakovaním je $n=(1,96/0,03)^20,5.0,5=1067$.
Pre výber bez opakovania a konečnú veľkosť základného súboru napr. N=4.1067 bude potrebný rozsah výberu
$n=(1,96/0,03)^20,5.0,5.(1-n/N)=1067.(1-n/N)$. Vyriešime rovnosť a máme n=1067(1+1/4)=854.
**Iný príklad**: d=0,04, p=0,5, tak n=600 a pri výbere bez opakovania a N ako v predošlom príklade máme n=450.

**Konštrukcia výberových súborov**
Výberový súbor môžeme konštruovať mnohými spôsobmi. Môžeme aplikovať napríklad:
a - kvótový výber - ak poznáme štruktúru výberového súboru podľa rozhodujúcich znakov vypočítame kvóty, ktoré potrebujeme "naplniť" a podľa nich rozpíšeme počty rozhovorov.
b - oblastný náhodný výber - základný súbor rozdelíme na oblasti a náhodnýcm výberom v každej oblasti vyberieme respondentov.
c - skupinový výber - vyberieme určité skupiny sekundárnych jednotiek a z nich vyberáme primárne jednotky.
d - kombinácia výberových postupov a iné.
Výber postupu závisí od cieľov zisťovania, od finančných možností a iných kapacít (napr. ľudských - čo majú výskum realizovať).

**Príklad 1**: Viacstupňový skupinový výber:
Základný súbor študentov 2.st. gymnázií obsahuje X gymnázií, kde študuje XXX študentov. Výberový súbor zostavíme nasledovne:
1.st.: vyberieme x gymnázií v jednotlivých krajoch SR úmerne ich počtu v krajoch.
2. st. : v každej vybranej škole vyberieme v každom ročníku 1 triedu (ak máme viac tried v ročníku musíme zvážiť počet vybraných tried). Pri tomto príklade uvažujeme v každom ročníku s jednou triedou, potom máme vo vývere x*4 tried.
3. st.: respondenti budú všetci študenti vybraných tried - nech priemerný počet študentov je 25 - potom sme získali výberový súbor o rozsahu n=x*4.25. Alternatívne môžeme vyberať v týchto triedach napr. 10 študentov – v závislosti od stanoveného rozsahu výberu.

**Príklad 2**: Proporcionálny oblastný výber. Ak by sme mali register všetkých študentov uvažovaných škól, tak môžeme definovať oblasti nasledovne: (školy)x(ročníky). Počet oblastí bude K=x*4.
Proporcinálne ku počtu študentov v oblasti ($N_i$) vyberieme náhodným mechanizmom ni študentov tejto oblasti. Výberový súbor bude mať rozsah $n=\Sigma n_i$.

## 6. Terénna fáza

Zber dát je nejdelikátnejšou a najnákladnejšou operáciou a ani najpremyslenejšie metódy matematickej štatistiky naznamenajú nič, ak sa uplatňujú na informáciách zaťažených veľkými chybami. Teória výberových zisťovaní má snahu poskytnúť metodologickú oporu pre nájdenie najprimeranejších prostriedkov zberu informácií.

Terénnu fázu spravidla uskutočňujú anketári. Pri rozsahu výberovej vzorky n=1400 rozhovorov vyžaduje jeden výskum spravidla prácu 200 anketárov. Pri terénnej práci sa pri niektorých výskumoch **pracuje s adresami respondentov.** Pri metóde kvótového výberu sa v niektorých výskumoch anketári zaznamenávajú adresy respondentov s ktorými uskutočnili rozhovor a pri metóde náhodného výberu tieto adresy dostanú priamo.

**Metódy terénneho zisťovania**
Štandardizované osobné rozhovory realizované školenými externými spolupracovníkmi - anketármi.

Dotazníkové zisťovania distribuované prostredníctvom pošty.

Telefonické dotazovania.

Nové technológie využívajúce počítače a nové prostriedy spojovej techniky

Kvalitatívne techniky.

**Pre anketárov**
- Pre anketárov musia byť pripravené inštrukcie a dodatočné definície
- Uľahči prácu anketárom a zváž použitie kartičiek odpovedí (face-to-face interview)

Po zbere empirických dát nasleduje ich kontrola a taktiež sa kontroluje terénna práca anketárov.

## 7. Matematickoštatistická analýza dát

V tejto fáze sa kontrolujú, kódujú dotazníky, zaznamenávajú do počítača, testujú sa dáta a realizujú sa štatistické analýzy.
**Testovanie reprezentatívnosti:** Ide o posúdenie zhody štruktúry výberového a základného súboru podľa základných demografických znakov.

**Reprezentatívnosť výberového súboru populácie SR vo veku 18 a viac rokov**

| **Znak** | Základný súbor | Výberový súbor | Rozdiel % | CHI2 Test HV 5 % |
|---|---|---|---|---|
| **Pohlavie:** | | | | rozdiely nie sú |
| muž | 47,80 | 47,57 | -0,23 | štatisticky |
| žena | 52,20 | 52,43 | +0,23 | významné |
| **Vek:** | | | | rozdiely nie sú |
| 18-24 | 16,16 | 16,49 | +0,33 | štatisticky |
| 25-29 | 10,32 | 10,04 | -0,28 | významné |
| 30-39 | 18,72 | 18,74 | +0,02 | |
| 40-49 | 20,33 | 19,83 | -0,50 | |

| | | | | |
|---|---|---|---|---|
| 50-59 | 14,02 | 14,14 | +0,12 | |
| 60 a viac | 20,45 | 20,75 | +0,30 | |
| **Vzdelanie:** | | | | rozdiely nie sú |
| základné | 25,00 | 25,46 | +0,46 | štatisticky |
| stredné bez maturity | 33,00 | 33,42 | +0,42 | významné |
| stredné s maturitou | 33,00 | 33,00 | 0,00 | |
| vysokoškolské | 9,00 | 8,12 | -0,88 | |
| **Národnosť:** | | | | rozdiely nie sú |
| slovenská | 84,81 | 85,09 | +0,28 | štatisticky |
| maďarská | 11,53 | 11,56 | +0,03 | významné |
| iná | 3,66 | 3,35 | -0,31 | |
| **Veľkostná skupina obce:** | | | | rozdiely nie sú |
| do 1999 | 30,49 | 28,87 | -1,62 | štatisticky |
| 2000 -9999 | 19,89 | 22,68 | +2,79 | významné |
| 10000 - 49999 | 24,23 | 23,26 | -0,97 | |
| 50000 - 99999 | 12,07 | 13,05 | +0,98 | |
| 100 000 a viac | 13,32 | 12,13 | -1,19 | |
| **Kraj:** | | | | rozdiely nie sú |
| Bratislavský | 11,99 | 12,89 | +0,90 | štatisticky |
| Trnavský | 10,38 | 10,04 | -0,34 | významné |
| Trenčiansky | 11,42 | 11,30 | -0,12 | |
| Nitriansky | 13,59 | 14,73 | +1,14 | |
| Žilinský | 12,58 | 12,72 | +0,14 | |
| Banskobystrický | 12,43 | 11,30 | -1,13 | |
| Prešovský | 13,71 | 14,23 | +0,52 | |
| Košický | 13,90 | 12,80 | -1,10 | |

**Testovanie reliability**

Štatistickú presnosť výsledkov výskumu hodnotíme pomocou intervalov spoľahlivosti. **Kvóli jednoduchosti výpočtu sa obvykle používajú aproximačné hranice ( i ), hoci správnejšie je požívanie exaktných medzí ( d, h ). V tabuľke sú uvedené medze exaktných ( d=dolná, h= horná) a aproximačných ( i) intervalov spoľahlivosti pre vybrané hodnoty p a n, pre spoľahlivosť 1-$\alpha$ =95%.**
Príslušné intevaly spoľahlivosti vypočítame podľa vzťahov:
exaktné hranice: pd=p-d, ph=p+h a aproximačné pd=p-i, ph=p+i.

| | | % | % | % | % | % | % | % | % | % |
|---|---|---|---|---|---|---|---|---|---|---|
| | **p** | **3** | **5** | **10** | **15** | **20** | **25** | **30** | **40** | **50** |
| **n** | **Hranica** | | | | | | | | | |
| **100** | **d** | **2,38** | **3,36** | **5,10** | **6,35** | **7,33** | **8,12** | **8,76** | **9,67** | **10,17** |
| | **h** | **5,52** | **6,28** | **7,62** | **8,53** | **9,18** | **9,66** | **9,98** | **10,28** | **10,17** |
| | **i** | **3,41** | **4,36** | **6,00** | **7,14** | **8,00** | **8,66** | **9,17** | **9,80** | **10,00** |
| **200** | **d** | **1,89** | **2,58** | **3,78** | **4,65** | **5,31** | **5,84** | **6,26** | **6,85** | **7,13** |
| | **h** | **3,42** | **4,00** | **5,02** | **5,72** | **6,22** | **6,60** | **6,86** | **7,15** | **7,13** |
| | **i** | **2,41** | **3,08** | **4,24** | **5,05** | **5,66** | **6,12** | **6,48** | **6,93** | **7,07** |
| **500** | **d** | **1,31** | **1,74** | **2,49** | **3,02** | **3,42** | **3,74** | **3,99** | **4,32** | **4,47** |
| | **h** | **1,90** | **2,29** | **2,97** | **3,44** | **3,78** | **4,04** | **4,23** | **4,44** | **4,47** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **i** | **1,53** | **1,95** | **2,68** | **3,19** | **3,58** | **3,87** | **4,10** | **4,38** | **4,47** |
| **1000** | **d** | **0,97** | **1,27** | **1,79** | **2,16** | **2,44** | **2,66** | **2,83** | **3,05** | **3,15** |
| | **h** | **1,26** | **1,54** | **2,03** | **2,37** | **2,62** | **2,81** | **2,95** | **3,11** | **3,15** |
| | **i** | **1,08** | **1,38** | **1,90** | **2,26** | **2,53** | **2,74** | **2,90** | **3,10** | **3,16** |
| **1250** | **d** | **0,87** | **1,14** | **1,61** | **1,94** | **2,18** | **2,38** | **2,53** | **2,73** | **2,81** |
| | **h** | **1,10** | **1,36** | **1,80** | **2,10** | **2,33** | **2,50** | **2,63** | **2,78** | **2,81** |
| | **i** | **0,96** | **1,23** | **1,70** | **2,02** | **2,26** | **2,45** | **2,59** | **2,77** | **2,83** |
| **2000** | **d** | **0,70** | **0,91** | **1,28** | **1,54** | **1,73** | **1,88** | **2,00** | **2,16** | **2,21** |
| | **h** | **0,84** | **1,05** | **1,40** | **1,64** | **1,82** | **1,96** | **2,06** | **2,19** | **2,21** |
| | **i** | **0,76** | **0,97** | **1,34** | **1,60** | **1,79** | **1,94** | **2,05** | **2,19** | **2,24** |
| **2500** | **d** | **0,63** | **0,82** | **1,15** | **1,38** | **1,55** | **1,69** | **1,79** | **1,93** | **1,98** |
| | **h** | **0,75** | **0,93** | **1,24** | **1,46** | **1,62** | **1,75** | **1,84** | **1,95** | **1,98** |
| | **i** | **0,68** | **0,87** | **1,20** | **1,43** | **1,60** | **1,73** | **1,83** | **1,96** | **2,00** |
| **5000** | **d** | **0,46** | **0,59** | **0,82** | **0,98** | **1,10** | **1,20** | **1,27** | **1,36** | **1,40** |
| | **h** | **0,51** | **0,64** | **0,87** | **1,02** | **1,14** | **1,22** | **1,29** | **1,37** | **1,40** |
| | **i** | **0,48** | **0,62** | **0,85** | **1,01** | **1,13** | **1,22** | **1,30** | **1,39** | **1,41** |
| **7500** | **d** | **0,37** | **0,48** | **0,67** | **0,80** | **0,90** | **0,98** | **1,04** | **1,11** | **1,14** |
| | **h** | **0,41** | **0,52** | **0,70** | **0,83** | **0,92** | **1,00** | **1,05** | **1,12** | **1,14** |
| | **i** | **0,39** | **0,50** | **0,69** | **0,82** | **0,92** | **1,00** | **1,06** | **1,13** | **1,15** |
| **10000** | **d** | **0,33** | **0,42** | **0,58** | **0,69** | **0,78** | **0,85** | **0,90** | **0,96** | **0,98** |
| | **h** | **0,35** | **0,45** | **0,60** | **0,72** | **0,80** | **0,86** | **0,91** | **0,97** | **0,98** |
| | **i** | **0,34** | **0,44** | **0,60** | **0,71** | **0,80** | **0,87** | **0,92** | **0,98** | **1,00** |

Vzťahy pre exaktné hranice (dolnú p**d** resp. hornú p**h)** intervalov spoľahlivosti:

$$p\mathbf{d}(p,n)=x/[\ x + (n-x+1)F\mathbf{1}],$$

kde F**1** je 1- $\alpha$**1** kvantil F rozdelenia s počtom stupňov voľnosti: 2(n-x+1), 2x

$$p\mathbf{h}(p,n)=[\ (x+1)F\mathbf{2}]/[n-x +(x+1)F\mathbf{2}],$$

kde F**2** je 1- $\alpha$**2** kvantil F rozdelenia s počtom stupňov voľnosti: 2(x+1), 2(n-x).

Obvykle volíme $\alpha$**1**= $\alpha$**2**=$\alpha$/2.

**Na testovanie validity dát** používame napr. techniku rozdelenia výberového súboru náhodným spôsobom na dva približne rovnaké posúbory a testujeme zhodu štruktúry odpovedí.

## Základné metódy matematickoštatistickej analýzy dát z dotazníkových výskumov

Základom pre matematickoštatistickú analýzu dotazníkových výskumov sú metódy analýzy kvalitatívnych znakov (nominálnych , ordinálnych):

- frekvenčné tabuľky, vrátane frekvenčných tabuliek pre otázky s viacerými možnými odpoveďami,
- kontingenčné tabuľky a testy kontingenčných tabuliek,
- mnohorozmerné metódy ( faktorová analýza, modifikovaná metóda hlavných komponentov pre nominálne znaky (HOMALS), pre ordinálne znaky (PRINCALS) a pre kombinácie znakov (OVERALS).
- analýza indexov,

- ak sú v dotazníku zaradené numerické otázky môžeme pri ich štatistickom spracovaní využiť metódy analýzy kvantitatívnych znakov, ale tiež po ich kategorizácii aj metódy analýzy kvalitatívnych znakov a i.

Pri štatistických testoch, ktoré sú založené na aproximácii rozdelení testových štatistík normálnym rozdelením je potrebné sledovať dodržanie podmienok použiteľmosti týchto aproximácií.

**Príklady štatistických analýz:**

**Poznámka: Dostupné ilustratívne príklady sú realizované v staršej verzii štatistického softvéru SPSS.**

## 7. 1. Frekvenčné tabuľky

**O1 Sledujete diskusiu o vstupe do NATO?**

| Value Label | Value | Frequency | Percent | Valid Percent | Cum Percent |
|---|---|---|---|---|---|
| sledujem | 1 | 331 | 26,2 | 26,2 | 26,2 |
| sledujem obcas | 2 | 559 | 44,2 | 44,2 | 70,4 |
| nesledujem | 3 | 329 | 26,0 | 26,0 | 96,4 |
| neviem o nej | 4 | 45 | 3,6 | 3,6 | 100,0 |
| | 0 | 1 | ,1 | Missing | |
| | | ------- | ------- | ------- | |
| | Total | 1265 | 100,0 | 100,0 | |

Valid cases 1264 Missing cases 1

## 7. 2. Frekvenčné tabuľky pre otázku z viacerými možnosťami odpovede

**Otázka: Odkiaľ máte informácie o otázkach nášho vstupu do NATO? Možno uviasť viac odpovedí.**

Group $O3

| Category label | Code | Count | Pct of Responses | Pct of Cases |
|---|---|---|---|---|
| z televízie | 1 | 876 | 31,7 | 69,3 |
| z rozhlasu | 2 | 571 | 20,6 | 45,2 |
| z dennej tlace | 3 | 644 | 23,3 | 50,9 |
| z týXdenníkov,mesacníkov | 4 | 137 | 5,0 | 10,8 |

| | | | | |
|---|---|---|---|---|
| zo +peciálnych letákov, broXúr | 5 | 14 | ,5 | 1,1 |
| od priatelov, známych,rodiny | 6 | 274 | 9,9 | 21,7 |
| z iného zdroja | 7 | 13 | ,5 | 1,0 |
| nesledujem,nezáujem | 8 | 238 | 8,6 | 18,8 |
| | Total responses | 2767 | 100,0 | 218,9 |

1 missing cases; 1.264 valid cases

## 7. 3. Kontingenčné tabuľky

O32 pohlavie by O1 Sledujete diskusiu o vstupe do NATO?

O1 Page 1 of 1

| Count / Row Pct / Col Pct / Adj Res | 1 sledujem obcas | 2 sledujem em | 3 neslediuj nej | 4 neviem o | Row Total |
|---|---|---|---|---|---|
| O32 1 MUŽ | 222 | 266 | 106 | 10 | 604 |
| | 36,8% | 44,0% | 17,5% | 1,7% | 47,9% |
| | 67,5% | 47,6% | 32,3% | 22,2% | |
| | 8,3 | -,2 | -6,6 | -3,5 | |
| 2 ŽENA | 107 | 293 | 222 | 35 | 657 |
| | 16,3% | 44,6% | 33,8% | 5,3% | 52,1% |
| | 32,5% | 52,4% | 67,7% | 77,8% | |
| | -8,3 | ,2 | 6,6 | 3,5 | |
| Column Total | 329 26,1% | 559 44,3% | 328 26,0% | 45 3,6% | 1261 100,0% |

| Chi-Square | Value | DF | Significance |
|---|---|---|---|
| Pearson | 94,35404 | 3 | ,00000 |
| Likelihood Ratio | 96,77050 | 3 | ,00000 |
| Mantel-Haenszel test for linear association | 92,62789 | 1 | ,00000 |

Minimum Expected Frequency - 21,554

Number of Missing Observations: 4

## 7.4. Kontingenčné tabuľky. O3 má viac možností odpovede.

### 7. 4.a. Percentá - základ: počet respondentov

$O3 (group)
by O35 vzdelanie

| | | O35 | | | | |
|---|---|---|---|---|---|---|
| | Count | základné | stredn, bez matu rity | stredosk olské s matu | vysokosk olské | Row Total |
| | Row pct | | | | | |
| | Col pct | | | | | |
| | | 1 | 2 | 3 | 4 | |
| $O3 | | | | | | |
| | 1 | 270 | 260 | 256 | 90 | 876 |
| z televízie | | 30,8 | 29,7 | 29,2 | 10,3 | 69,4 |
| | | 62,6 | 69,0 | 74,2 | 81,8 | |
| | 2 | 173 | 171 | 172 | 55 | 571 |
| z rozhlasu | | 30,3 | 29,9 | 30,1 | 9,6 | 45,2 |
| | | 40,1 | 45,4 | 49,9 | 50,0 | |
| | 3 | 151 | 194 | 212 | 87 | 644 |
| z dennej tlace | | 23,4 | 30,1 | 32,9 | 13,5 | 51,0 |
| | | 35,0 | 51,5 | 61,4 | 79,1 | |
| | 4 | 35 | 32 | 50 | 20 | 137 |
| z týXdenníkov,mesacn | | 25,5 | 23,4 | 36,5 | 14,6 | 10,8 |
| | | 8,1 | 8,5 | 14,5 | 18,2 | |
| | 5 | 1 | 2 | 7 | 4 | 14 |
| zo +peciálnych leták | | 7,1 | 14,3 | 50,0 | 28,6 | 1,1 |
| | | ,2 | ,5 | 2,0 | 3,6 | |
| | 6 | 78 | 90 | 76 | 30 | 274 |
| od priatelov, známyc | | 28,5 | 32,8 | 27,7 | 10,9 | 21,7 |
| | | 18,1 | 23,9 | 22,0 | 27,3 | |
| | 7 | 4 | 2 | 5 | 2 | 13 |
| z iného zdroja | | 30,8 | 15,4 | 38,5 | 15,4 | 1,0 |
| | | ,9 | ,5 | 1,4 | 1,8 | |
| | 8 | 122 | 66 | 45 | 4 | 237 |
| nesledujem,nezáujem | | 51,5 | 27,8 | 19,0 | 1,7 | 18,8 |
| | | 28,3 | 17,5 | 13,0 | 3,6 | |
| | Column Total | 431 | 377 | 345 | 110 | 1263 |
| | | 34,1 | 29,8 | 27,3 | 8,7 | 100,0 |

Percents and totals based on respondents
1.263 valid cases; 2 missing cases

**7. 4.b. Percentá - základ: počet odpovedí**

$O3 (group)
by O35 vzdelanie

| | | O35 | | | | |
|---|---|---|---|---|---|---|
| | Count | základné | stredn, bez matu rity | stredosk olské s matu | vysokosk olské | Row Total |
| | Row pct | | | | | |
| | Col pct | | | | | |
| | | 1 | 2 | 3 | 4 | |
| $O3 | | | | | | |
| 1 | | 270 | 260 | 256 | 90 | 876 |
| z televízie | | 30,8 | 29,7 | 29,2 | 10,3 | 31,7 |
| | | 32,4 | 31,8 | 31,1 | 30,8 | |
| 2 | | 173 | 171 | 172 | 55 | 571 |
| z rozhlasu | | 30,3 | 29,9 | 30,1 | 9,6 | 20,6 |
| | | 20,7 | 20,9 | 20,9 | 18,8 | |
| 3 | | 151 | 194 | 212 | 87 | 644 |
| z dennej tlace | | 23,4 | 30,1 | 32,9 | 13,5 | 23,3 |
| | | 18,1 | 23,7 | 25,8 | 29,8 | |
| 4 | | 35 | 32 | 50 | 20 | 137 |
| z týXdenníkov,mesacn | | 25,5 | 23,4 | 36,5 | 14,6 | 5,0 |
| | | 4,2 | 3,9 | 6,1 | 6,8 | |
| 5 | | 1 | 2 | 7 | 4 | 14 |
| zo +peciálnych leták | | 7,1 | 14,3 | 50,0 | 28,6 | ,5 |
| | | ,1 | ,2 | ,9 | 1,4 | |
| 6 | | 78 | 90 | 76 | 30 | 274 |
| od priatelov, známyc | | 28,5 | 32,8 | 27,7 | 10,9 | 9,9 |
| | | 9,4 | 11,0 | 9,2 | 10,3 | |
| 7 | | 4 | 2 | 5 | 2 | 13 |
| z iného zdroja | | 30,8 | 15,4 | 38,5 | 15,4 | ,5 |
| | | ,5 | ,2 | ,6 | ,7 | |
| 8 | | 122 | 66 | 45 | 4 | 237 |
| nesledujem,nezáujem | | 51,5 | 27,8 | 19,0 | 1,7 | 8,6 |
| | | 14,6 | 8,1 | 5,5 | 1,4 | |
| | Column Total | 834 | 817 | 823 | 292 | 2766 |
| | | 30,2 | 29,5 | 29,8 | 10,6 | 100,0 |

Percents and totals based on responses
1.263 valid cases; 2 missing cases

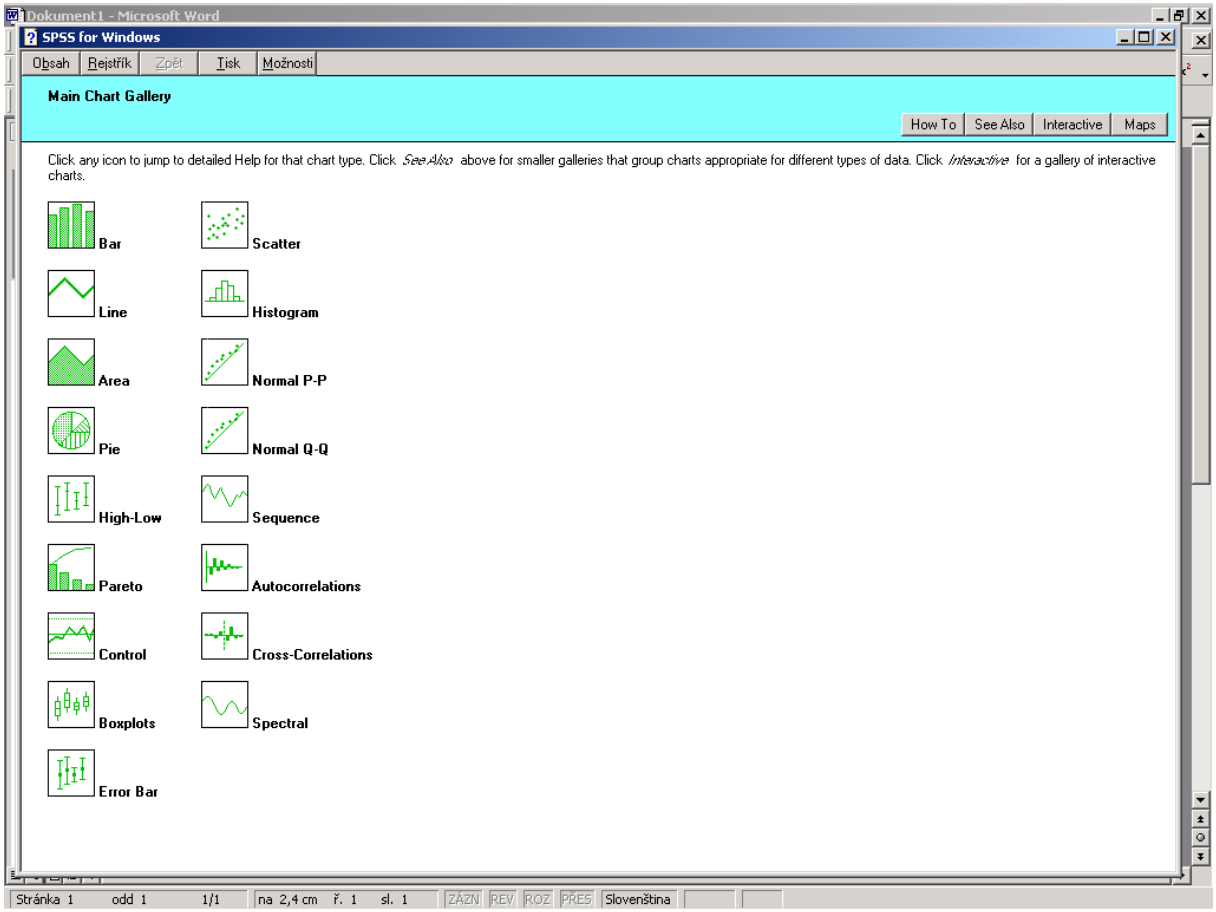
**8. Prezentácia výsledkov, publikovanie**

Záverečná fáza realizácie výskumu verejnej mienky je

**6. fáza: Vypracovanie správy z výskumu a publikácií na základe analýzy výsledkov.**

Tu využívame štatistické tabuľky a štatistické grafy a i.

Ako príklad uvedieme prehľad grafov, ktoré ponúka SPSS:




## 9. Software na štatistickú analýzu dát z výskumov verejnej mienky

Naznámejšie profesionálne programové produkty sú SPSS, SAS, SYSTAT, STATGRAPHICS, STATISTICA, BMDP, GENSTAT, SOLO, ADSTAT.
Existujú špeciálne programové produkty, prípadne moduly známych programov, ktoré sú orientované na riešenie úloh spojených s výberovými metódami napr. SUDAAN, STATPAC moduly SAS a SPSS, prípadne s výskumami verejnej mienky a marketingovými prieskumami – moduly SAS, moduly a programy rozširované SSPS WesVar Compex Samples, SaplePower, SurveyCraft.

Ukážky analýz boli vypracované v SPSS for WINDOWS, ktorý autor tejto štúdie uprednostňuje pri analýzach dát z výskumov verejnej mienky.

## 10. Literatúra


[1] Anděl, J.: Matematická statistika. SNTL/Alfa Praha 1978.

[2] Bárta, V., Bártová, H.(1991): Marketingový výskum trhu. Knižnica Hosp. novín. Economia, Praha 1991.

[3] Bishop, Y.M.M., Fienberg,S.E., Holland,P.W.: Discrete Multivariate Distributions: Theory and Practice. MIT Press Cambridge, Massachusetts and London, England 1975.

[4] Bradburn, Norman M. and Sudman, Seymour:"The Current Status of Questionnaire Design" in Paul, P. Biemer et al (Ed.s) Measurement Errors in Surveys. John Wiley & Sons. Inc. 1991.

[5] Clauss, G., Ebner, H.: Základy štatistiky pre psychológov pedagógov a sociológov. SPN, Bratislava 1988.

[6] Čermák, V.: Výběrové statistické zjišťování. SNTL/ALFA Praha 1980.

[7] Deming, W.E.: Sample Design in Business Research. Wiley, New York 1960.

Fowler, Floyd J.: Improving Survey Questions Design and Evaluation. Sage Publications, 1995.

[8] Gallup, G.: Pruvodce po vyzkumu verějného mínění. Orbis, Praha 1976.

[9] Grosbras, J.-M.: Méthodes Statistiques des Sondages. Paris 1990.

[10] Hamilton, J.: Co je výskum trhu. ESOMAR, GfK Praha 1989.

[11] Herzmann, J., Novák, I., Pecáková, I.: Výzkumy veřejného mínění. Skriptum VŠE Praha, Fakulta informatiky a statistiky, 1995.

[12] Chajdiak, J.(2003): Štatistika jednoducho. Statis Bratislava 2003, ISBN 808565928-X.

[13] Chajdiak J.(2005): Štatistické úlohy a ich riešenie v Exceli. STATIS, Bratislava, ISBN 80-85659-39-5.

[14] Johnson, N.L., Kotz S.: Distributions in Statistics. Vol. 1: Discrete Distributions. Wiley, New York 1969.

[15] Kanderová, M. – Úradníček, V.(2007): Štatistika a pravdepodobnosť pre ekonómov. 1. časť. OZ Financ, Banská Bystrica 2007, ISBN 978-80-969535-5-4.

[16] Kanderová, M. – Úradníček, V.(2007): Štatistika a pravdepodobnosť pre ekonómov. 2. časť. OZ Financ, Banská Bystrica 2007, ISBN 987-80-696535-6-1.

[17] Kendall,M.G., Stuart,A.: Statističeskije vyvody I svjazi. Nauka, Moskva 1973.

[18] Luha, J., Kevická, R.: Unifikácia indexov pomocou pravdepodobnostných modelov. Sociológia 1/1990.

[19] Luha, J. a kol.: Metódy štatistickej analýzy kvalitatívnych znakov I. ÚVT VŠ, Bratislava 1983.

[20] Luha, J. a kol.: Metódy štatistickej analýzy kvalitatívnych znakov II. ÚVT VŠ, Bratislava 1985.

[21] Luha, J. Testovanie štatistických hypotéz pri analýze súborov charakterizovaných kvalitatívnymi znakmi. STV Bratislava 1985.

[22] Luha, J.(1994): Meranie spoľahlivosti výsledkov výskumu verejnej mienky. Ekomstat 1994.

[23] Luha, J.: Štatistické a metodologické aspekty tvorby a analýzy dotazníkov. SŠDS, EKOMSTAT´97, Trenčianske Teplice 1997.

[24] Luha, J.: Výberové štatistické zisťovania. SŠDS, EKOMSTAT´99, Trenčianske Teplice 1999.

[25] Luha, J.: Analýza nominálnych a ordinálnych znakov. SŠDS, EKOMSTAT´2000, Trenčianske Teplice 2000.

[26] Luha J.(2003): Skúmanie súboru kvalitatívnych dát. EKOMSTAT 2003. SŠDS Bratislava 2003.

[27] Luha J.(2003): Matematickoštatistické aspekty spracovania dotazníkových výskumov. Štatistické metódy vo vedecko-výskumnej práci 2003, SŠDS, Bratislava 2003. ISBN 80-88946-32-8.

[28] Luha J.(2007): Kvótový výber. FORUM STATISTICUM SLOVACUM 1/2007. SŠDS Bratislava 2007. ISSN 1336-7420.

[29] Luha J.: Korelácia vnorených javov. FORUM STATISTICUM SLOVACUM 6/2008. SŠDS Bratislava 2008. ISSN 1336-7420.

[30] Luha J.: Korelácia disjunktných a komplementárnych javov. FORUM STATISTICUM SLOVACUM 6/2008. SŠDS Bratislava 2008. ISSN 1336-7420.

[31] Luha J.: Metodologické aspekty zberu a záznamu dát otázok s možnosťou viac odpovedí. FORUM STATISTICUM SLOVACUM 7/2008. SŠDS Bratislava 2008. ISSN 1336-7420.
[32] Luha J.: Korelácie medzi politikmi a stranami. FORUM STATISTICUM SLOVACUM 7/2008. SŠDS Bratislava 2008. ISSN 1336-7420.
[33] Mirkin, B.G.: Analiz kačetvennych priznakov i struktur. Statistika, Moskva 1980.
[34] Pecáková I.: Statistika v terénních průzkumech. Proffessional Publishing, Praha 2008. ISBN 978-80-86946-74-0.
[35] Řehák J., Řeháková B. (1986): Analýza kategorizovaných dat v sociologii. Academia Praha 1986.
[36] Řehák J., Bártová I.: Statistika pro výzkum trhu a marketing. Statistické konzultace a výpočty, SC&C Praha 1997.
[37] Řezanková A.(2007): Analýza dat z dotazníkových šetření. Proffessional Publishing, Praha 2007. ISBN 978-80-86946-49-8.
[38] Stehlíková, B.: Určovanie rozsahu výberového súboru. Okresné oddelenie ŠÚ SR Nové Zámky 1994.
[39] Stankovičová I., Vojtková M.(2007): Viacrozmerné štatistické metódy s aplikáciami. IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.
[40] Manuály štatistického software SPSS.

**11. Niektoré „zaujímavé“ www stránky:**

- www.surveysystem.com
- www.amstat.org
- www.casro.org
- www.aapor.org
- www.spss.com
- www.spss.cz
- www.scac.cz
- www.gallup.com
- www.gfk.com (aj cz a sk)
- www.aisa.com
- http://trochim.human.cornell.edu/kb/survey.htm
- ww.statsoft.com

**Adresa autora:**

RNDr. Ján Luha, CSc.
Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a FNsP Bratislava
jan.luha@fmed.uniba.sk

# Testovanie hypotézy o obrobenom povrchu
# Testing hypothesis about cutting surface

Macurová Anna, Macura Dušan

**Abstract:** The cutting surface dependency on many parameters in technical practice. The dependency quality of the cutting surface on the cutting speed is object of the experimentation. The Mann – Whitneyho $U$ test enables of the validity hypothesis of the experiment.

**Key words:** cutting surface**,** experimental values**,** hypothesis, Mann-Whitney $U$ test.

**Kľúčové slová:** obrobený povrch, experimentálne hodnoty, hypotéza, Mann-Whitneyho $U$ test.

**Úvod**

Obrábanie je časť výrobného procesu, pri ktorom polovýrobok dostáva požadovaný tvar a rozmer strojovej súčiastky odoberaním čiastočiek materiálu z povrchovej vrstvy. Funkciou obrábania je dať materiálom alebo polovýrobkom tzv. funkčnú presnosť, charakterizovanú rozmermi a stavom obrobených povrchov. Obrábaná plocha je plocha polovýrobku, ktorú treba v procese obrábania odstrániť a nahradiť novovzniknutou plochou. Rezná plocha sa vytvára hneď za reznou hranou nástroja. Obrobená plocha je výsledkom obrábania a tvoria ju zvyšky reznej plochy. Stav obrobeného povrchu, a tým aj tvorba povrchu pri rezaní závisí od použitej metódy obrábania. Rezanie môže byť vykonané geometricky určitou a geometricky neurčitou reznou hranou alebo progresívnymi metódami obrábania. Vlastnosti obrobeného povrchu sú ovplyvnené obrábaným materiálom, reznými podmienkami, stupňom opotrebenia reznej hrany, a statickou a dynamickou tuhosťou systému stroj – nástroj – obrobok.

## 1. Výsledky experimentu a aproximované hodnoty

***Tabuľka 1: Experimentálne a aproximované hodnoty nerovností obrobeného povrchu***

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| experimentálne hodnoty | 5,78 | 8 | 8,67 | 9 | 10,2 | 15 | 15 | 17,5 | 19 | 20 | 25 | 28,2 | 35 | 48 |
| aproximované hodnoty | 7,05 | 8,42 | 9,9 | 10,8 | 11,2 | 13,6 | 14,3 | 16 | 16,9 | 17,7 | 23,6 | 29,8 | 38,8 | 46 |

***Tabuľka 2: Usporiadanie hodnôt podľa veľkosti pre Mann-Whitneyho U-test***

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| experimentálne hodnoty | 5,78 | | 8 | | 8,67 | 9 | | 10,2 | | | | | |
| aproximované hodnoty | | 7,05 | | 8,42 | | | 9,9 | | 10,8 | 11,2 | 13,6 | 14,3 | |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | | | 17,5 | | 19 | 20 | | 25 | 28,2 | | 35 | | | 48 |
| | 16 | 16,9 | | 17,7 | | | 23,6 | | | 29,8 | | 38,8 | 46 | |

***Tabuľka 3: Poradie hodnôt po prisúdení poradového čísla pre Mann-Whitneyho U test***

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| experimentálne hodnoty | 1 | 3 | 5 | 6 | 8 | 13,5 | 17 | 19 | 20 | 22 | 23 | 25 | 28 | |
| aproximované hodnoty | 2 | 4 | 7 | 9 | 10 | 11 | 12 | 15 | 16 | 18 | 21 | 24 | 26 | 27 |

## 2. Formulácia hypotézy o obrobenom povrchu pre Mann-Whitneyho U test

Rozsah výberového súboru nameraných hodnôt je menší, preto tento súbor bude považovaný za prvý súbor. V tomto prípade je rozsah že $n_1 = 13$. Suma poradí v tomto súbore je $R_1 = 190{,}5$. Rozsah súboru $n_2 = 14$, suma poradí je $202$. Celkový rozsah súboru je 26. Hodnota testovacej štatistiky U je $8{,}5$. Testujeme hypotézu o tom, že hodnoty získané experimentom a aproximáciou sú spoľahlivé a vyhovujú požiadavkám o drsnosti obrobeného povrchu.

$$H_0 : Rzex = Rzap$$

$$H_1 : Rzex \neq Rzap$$

Na hladine významnosti α = 0,05 je oblasť zamietnutia určená kvantilom U rozdelenia $U_{1-\alpha} = U_{0,95} = 35$ $U_{1-\alpha} = U_{0,95} = 35$. Na hladine významnosti 0,05 nezamietame nulovú hypotézu.

## 3. Záver

Experimentálne hodnoty boli získané obrábaním (frézovaním) daného povrchu kovu, royhodujúcim parametrom pre získanie experimentálnej závislosti bola rezná rýchlosť nástroja. V tomto experimente Mann-Whitneyho U test potvrdil, že výsledky dosiahnuté nameraním hodnôt drsnomerom (experimentálne) zodpovedajú požadovanej drsnosti obrobeného povrchu. Porovnávané boli hodnoty získané aproximáciou, ktorá v tomto prípade vyjadrovala experiment spoľahlivo, čo je možné tvrdiť na základe záverov vyplývajúcich použitím Mann – Whitneyho U testu.

## 4. Literatúra


[1] HINES, W. W.- MONTGOMERY, D.C. 1990. Probability and Statistics in Engineering and Management Science. USA: John Wiley @Sons, 1990. 732 s. ISBN 0-471-60090-3.

[2] POTOCKÝ, R. – KALAS, J.- KOMORNÍK, J.- LAMOŠ, F. 1991. Zbierka úloh z pravdepodobnosti a matematickej štatistiky. Bratislava: Alfa, 1991. 392 s. ISBN 80-05-00524-5.

**Adresa autora (-ov):**

Macurová Anna, PaedDr., PhD.
FVT TU
Bayerova 1
080 01 Prešov
anna.macurova@tuke.sk

Macura Dušan, Mgr., PhD.
PU FHPV
Ul. 17 Novembra 1Ulica2
080 01 Prešov
macura@unipo.sk

# $\sigma$-additivity of s-maps

*Ivica Marinová, Ľubica Valášková*

**Abstract:** s-maps have been introduced as a tool for measurement of random events, which are not simultaneously measureable. Any s-map is additive in each coordinate. In this paper we deal with $\sigma$-additivity of s-maps. Further we study properties similar to $\sigma$-additivity of j-maps and d-maps.

**Key words:** Orthomodular lattice, s-map, j-map, d-map, $\sigma$-additivity, semicontinuoity

## 1. Introduction and premilinaries

In qantum mechanics the relation between two variables is difficult to study because of the influence of measurement of one variable to the measure of the other one.

In 2003 Oľga Nánásiová introduced s-maps as a way to omit measure of intersection of two events (the mathematical analogue of a simultaneous measurement). s-map is a function on an orthomodular lattice, which is additive in each coordinate. So a natural question arise: how is it with $\sigma$-additivity. In this paper we study $\sigma$-additivity of s-maps and similar properties of q-maps and d-maps.

At first we recall basic notions (see e.g. [2],[5],[6]).

**Definition 1. 1** *Let $L$ be a lattice (a nonempty set endowed with a partial ordering $\leq$, the lattice operations supremum $\vee$ and infimum $\wedge$) with the greatest element $I$ and the smallest element $O$. Let $\perp\colon L \to L$ be a unary operation on $L$ with the following properties:*

1. *$\forall a \in L\ \exists! a^{\perp} \in L$ such that $(a^{\perp})^{\perp} = a$ and $a \vee a^{\perp} = I$*
2. *If $a, b \in L$ and $a \leq b$ then $b^{\perp} \leq a^{\perp}$*
3. *If $a, b \in L$ and $a \leq b$ then $b = a \vee (a^{\perp} \wedge b)$ (orthomodular law)*

*Then $\mathcal{L} = (L, O, I, \vee, \wedge, \perp)$ is an orthomodular lattice (briefly an OML).*

**Definition 1. 2** *Let $\mathcal{L}$ be an OML. Then elements $a, b \in L$ are*

1. *orthogonal ($a \perp b$) if $a \leq b^{\perp}$;*
2. *compatible ($a \leftrightarrow b$) if $a = (a \wedge b) \vee (a \wedge b^{\perp})$ and $b = (a \wedge b) \vee (a^{\perp} \wedge b)$.*

The following three maps have been introduced in [1],[3],[4].

**Definition 1. 3** *Let $\mathcal{L}$ be an OML. A map $p : L^2 \to [0, 1]$ is an s-map if the following conditions hold:*

*(s1)* $p(I, I) = 1$;

*(s2) if $a \perp b$ then $p(a, b) = 0$;*

*(s3) if $a \perp b$ then for each $c \in L$*

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

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

**Definition 1. 4** *Let $\mathcal{L}$ be an OML. A map $q : L^2 \to [0, 1]$ is a join map (j-map) if the following conditions hold:*

*(q1) $q(I, I) = 1$;*

*(q2) if $a \perp b$ then $q(a, b) = q(a, a) + q(b, b)$;*

*(q3) if $a \perp b$ then for each $c \in L$*

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

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

**Definition 1. 5** *Let $\mathcal{L}$ be an OML. A map $d : L^2 \to [0, 1]$ is a difference map (d-map) if the following conditions hold:*

*(d1) $d(a, a) = 0$ and $d(O, I) = d(I, O) = 1$;*

*(d2) if $a \perp b$ then $d(a, b) = d(a, O) + d(O, b)$;*

*(d3) if $a \perp b$ then for each $c \in L$*

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

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

## 2. $\sigma$-properties of s-maps, j-maps and d-maps

**Definition 2. 1** *Let $\mathcal{L}$ be an OML. A map $f : L^2 \to [0, 1]$ is semicontinuous from below if for each nondecreasing sequence $\{c_n\}_{n=1}^{\infty}$ of elements of $L$ such that $\bigvee_{n=1}^{\infty} c_n \in L$ it holds $f(a, \bigvee_{n=1}^{\infty} c_n) = \lim_{n\to\infty} f(a, c_n)$ and $f(\bigvee_{n=1}^{\infty} c_n, a) = \lim_{n\to\infty} f(c_n, a)$ for all $a \in L$.*

The properties of s-maps, j-maps and d-maps listed in the following lemma result directly from the definitions. Naturally, they hold when we change the order of coordinates.

**Lemma 2. 1** *Let $\mathcal{L}$ be an OML. Let $p : L^2 \to [0, 1]$ be an s-map, $q : L^2 \to [0, 1]$ be a j-map and $d : L^2 \to [0, 1]$ be a d-map. Let $a \in L$ and $b_1, b_2, ..., b_n$ are pairvise orthogonal elements of $L$. Then for any $n \in N$ it holds:*

$$p(a, \bigvee_{i=1}^{n} b_i) = \sum_{i=1}^{n} p(a, b_i);$$

$$q(a, \bigvee_{i=1}^{n} b_i) = q(a, a) + \sum_{i=1}^{n} (q(a, b_i) - q(a, a));$$

$$d(a, \bigvee_{i=1}^{n} b_i) = d(a, O) + \sum_{i=1}^{n} (d(a, b_i) - d(a, O)).$$

Remind that $\sigma$-additivity on an OML means: $p$ is $\sigma$-additive in each coordinate if $p(a, \bigvee_{i=1}^{\infty} b_i) = \sum_{i=1}^{\infty} p(a, b_i)$ and $p(\bigvee_{i=1}^{\infty} b_i, a) = \sum_{i=1}^{\infty} p(b_i, a)$ for any $a \in L$ and any sequence $\{b_i\}_{i=1}^{\infty}$ of pairvise orthogonal elements of $L$ such that $\bigvee_{n=1}^{\infty} b_n \in L$.

**Proposition 2. 1** *Let $\mathcal{L}$ be an OML. Let $p : L^2 \to [0, 1]$ be an s-map semicontinuous from below. Then $p$ is $\sigma$-additive in each coordinate.*
*Proof.*

The proof is similar to the proof of the corresponding theorem in the measure theory. Let $\{b_i\}_{i=1}^{\infty}$ be a sequence of pairvise orthogonal elements of L such that $\bigvee_{n=1}^{\infty} b_n \in L$, $a \in L$. Then the sequence $\{c_n\}_{n=1}^{\infty}, c_n = \bigvee_{i=1}^{n} b_i$ is nondecreasing and $\bigvee_{i=1}^{\infty} c_i = \bigvee_{i=1}^{\infty} b_i$.

From Lemma 2.1 and the semicontinuoity of $p$ we get

$$\begin{aligned} p(a, \bigvee_{i=1}^{\infty} b_i) &= p(a, \bigvee_{i=1}^{\infty} c_i) \\ &= \lim_{n\to\infty} p(a, c_n) \\ &= \lim_{n\to\infty} p(a, \bigvee_{i=1}^{n} b_i) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} p(a, b_i) \\ &= \sum_{i=1}^{\infty} p(a, b_i). \end{aligned}$$

Analogously $p(\bigvee_{i=1}^{\infty} b_i, a) = \sum_{i=1}^{\infty} p(b_i, a)$.

There is one to one correspondence between s-maps and j-maps on a given OML. If an s-map $p$ is given, then $q(a, b) = p(a, a) + p(b, b) - p(a, b)$ is the corresponding j-map and vice versa $p(a, b) = q(a, a) + q(b, b) - q(a, b)$ is the corresponding s-map for any j-map q. The difference of corresponding maps $q$ and $p$ gives a d-map $d$ [4].

**Proposition 2. 2** *Let $\mathcal{L}$ be an OML and $p : L^2 \to [0, 1]$ be an s-map semicontinuous from below. Let $\{b_i\}_{i=1}^{\infty}$ be a sequence of pairvise orthogonal elements of L such that $\bigvee_{n=1}^{\infty} b_n \in L$, $a \in L$. Then the corresponging j-map $q$ is semicontinuous from below and*

$$q(a, \bigvee_{i=1}^{\infty} b_i) = q(a, a) + \sum_{i=1}^{\infty} (q(a, b_i) - q(a, a))$$

$$q(\bigvee_{i=1}^{\infty} b_i, a) = q(a, a) + \sum_{i=1}^{\infty} (q(b_i, a) - q(a, a)).$$

*Proof.*

Let $\{c_i\}_{i=1}^{\infty}$ be a nondecreasing sequence of elements of $L$ such that $\bigvee_{i=1}^{\infty} c_i \in L$. Then

$$q(a, \bigvee_{i=1}^{\infty} c_i) = p(a, a) + p(\bigvee_{i=1}^{\infty} c_i, \bigvee_{i=1}^{\infty} c_i) - p(a, \bigvee_{i=1}^{\infty} c_i)$$

$$
\begin{aligned}
&= p(a,a) + \lim_{i\to\infty} p(c_i, \bigvee_{j=1}^{\infty} c_j) - \lim_{i\to\infty} p(a,c_i) \\
&= p(a,a) + \lim_{i\to\infty} \lim_{j\to\infty} p(c_i,c_j) - \lim_{i\to\infty} p(a,c_i) \\
&= p(a,a) + \lim_{i\to\infty} \lim_{j\to\infty} p(c_i,c_i) - \lim_{i\to\infty} p(a,c_i) \\
&= p(a,a) + \lim_{i\to\infty} p(c_i,c_i) - \lim_{i\to\infty} p(a,c_i) \\
&= \lim_{i\to\infty} (p(a,a) + p(c_i,c_i) - p(a,c_i)) \\
&= \lim_{i\to\infty} q(a,c_i).
\end{aligned}
$$

Analogously $q(\bigvee_{i=1}^{\infty} c_i, a) = \lim_{i\to\infty} q(c_i,a)$, therefore $q$ is semicontinuous from below.

$$
\begin{aligned}
q(a, \bigvee_{i=1}^{\infty} b_i) &= p(a,a) + p(\bigvee_{i=1}^{\infty} b_i, \bigvee_{i=1}^{\infty} b_i) - p(a, \bigvee_{i=1}^{\infty} b_i) \\
&= q(a,a) + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p(b_i,b_j) - \sum_{i=1}^{\infty} p(a,b_i) \\
&= q(a,a) + \sum_{i=1}^{\infty} p(b_i,b_i) - \sum_{i=1}^{\infty} p(a,b_i) \\
&= q(a,a) + \sum_{i=1}^{\infty} q(b_i,b_i) - \sum_{i=1}^{\infty} (q(a,a) + q(b_i,b_i) - q(a,b_i)) \\
&= q(a,a) + \sum_{i=1}^{\infty} (q(a,b_i) - q(a,a)).
\end{aligned}
$$

The proof of the second statement is analogous.

**Proposition 2. 3** *Let $\mathcal{L}$ be an OML. Let $d : L^2 \to [0,1]$ be a d-map semicontinuous from below. Let $\{b_i\}_{i=1}^{\infty}$ be a sequence of pairvise orthogonal elements of $L$ such that $\bigvee_{n=1}^{\infty} b_n \in L$, $a \in L$. Then*

$$
\begin{aligned}
d(a, \bigvee_{i=1}^{\infty} b_i) &= d(a,O) + \sum_{i=1}^{\infty} (d(a,b_i) - d(a,O)) \\
d(\bigvee_{i=1}^{\infty} b_i, a) &= d(a,O) + \sum_{i=1}^{\infty} (d(b_i,a) - d(a,O)).
\end{aligned}
$$

*Proof.*

Sequence $\{c_n\}_{n=1}^{\infty}, c_n = \bigvee_{i=1}^{n} b_i$ is nondecreasing and $\bigvee_{i=1}^{\infty} c_i = \bigvee_{i=1}^{\infty} b_i$. From this and from the semicontinuoity of $d$ we get

$$
\begin{aligned}
d(a, \bigvee_{i=1}^{\infty} b_i) &= d(a, \bigvee_{i=1}^{\infty} c_i) \\
&= \lim_{n\to\infty} d(a,c_n) \\
&= \lim_{n\to\infty} d(a, \bigvee_{i=1}^{n} b_i) \\
&= \lim_{n\to\infty} (d(a,O) + \sum_{i=1}^{n} (d(a,b_i) - d(a,O))) \\
&= d(a,O) + \sum_{i=1}^{\infty} (d(a,b_i) - d(a,O)).
\end{aligned}
$$

The proof of the second statement is analogous.

**Corollary 2. 1** *Let $\mathcal{L}$ be an OML and $\{b_i\}_{i=1}^{\infty}$ be a sequence of pairvise orthogonal elements of $L$ such that $\bigvee_{n=1}^{\infty} b_n \in L$. Let $f$ be an s-map, j-map or d-map semicontinuous from below. Then*

$$f(\bigvee_{i=1}^{\infty} b_i, \bigvee_{i=1}^{\infty} b_i) = \sum_{i=1}^{\infty} f(b_i, b_i)$$

*Proof.*

The property for s-maps has been proved in Proposition 2.2, for d-maps it is trivial. Let $q$ be j-map. Then

$$\begin{aligned} q(\bigvee_{i=1}^{\infty} b_i, \bigvee_{i=1}^{\infty} b_i) &= q(\bigvee_{i=1}^{\infty} b_i, O) \\ &= q(O,O) + \sum_{i=1}^{\infty}(q(b_i,O) - q(O,O)) \\ &= \sum_{i=1}^{\infty}(q(b_i,O) \\ &= \sum_{i=1}^{\infty} q(b_i,b_i). \end{aligned}$$

**Acknowledgment** This work was supported by Science and Technology Assistance Agency under the contract No. APVV-0375-06, VEGA-1/0373/08.

# References

[1] Bohdalová M., Minárová M., Nánásiová O.,: A note to algebraic approach to uncertainty, F. Stat. Slov. *3*, (2006), pp. 31-39.

[2] Dvurečenskij, A., Pulmannová, S.: New Trends in Quantum Structures. Kluwer Acad. Publ., (2000).

[3] Nánásiová, O.: Map for Simultaneous Measurements for a Quantum Logic., Int. J. of Theor. Physics, *42*, (2003), pp. 1889-1903.

[4] Nánásiová, O., Minárová, M., Mohammed, A.: Measure of "symmetric difference". Proc. Magia 2006, ISBN 978-80-227-2583-5, pp. 55-60.

[5] Pták, P., Pulmannová S.: Quantum Logics. Kluwer Acad. Press, Bratislava, (1991).

[6] Varadarajan, V.: Geometry of quantum theory. Princeton, New Jersey, D. Van Nostrand, (1968).

**Autori (Authors):**

RNDr. Ivica Marinová, PhD.
Katedra matematiky, Fakulta elektrotechniky a informatiky STU
Ilkovičova 3
812 19 Bratislava
e-mail: ivica.marinova@stuba.sk

RNDr. Ľubica Valášková, PhD.
Katedra matematiky a deskriptívnej geometrie, Stavebná fakulta STU
Radlinského 11
813 68 Bratislava
e-mail: valaskova@math.sk

# Optimal Insulating Parameters Stipulation for an Old Family House

Minárová Mária

**Abstract:** The paper deals with the thermal and hygric analysis of building construction. The first part quantifies moisture parameters on the inside surfaces of the building construction, introduces the physical model. The second one discusses the problems joint with windows changing. It suggests the procedure, considerations and calculations that should go before to intended retrofitting.

**Key words:** boundary value problems, temperature field, moisture, vapor pressure against a wall, relative humidity, absolute humidity, insulation

## 1. Introduction

Sudden indoor climate change, dampness, inside plaster degradation, molds appearing – these are often the consequences of unprofessional windows exchanging in dwellings. The solution of such situation afterwards is obviously more complicated and costly.

The example is an about on hundred years old house with heavy walls made of stones and clay. The window changing in effort of energy saving caused mold appearance, later wet blots on the plaster, with salt (chemical reaction indicating) borders afterwards.

Whereas the old wooden windows guaranteed the sufficient air venting inevitable for dampness offtake, the new ones does not provide it and the moisture accumulates in the inside air.

## 2. Moisture in the air

There are two sorts of air humidity quantifiers:

***absolute humidity***

$c$ water vapor concentration [kg.m$^{-3}$] (amount of water in 1m$^3$ of air)

$x$ water vapor and dry air mass ratio [kg.kg$^{-1}$]

***relative humidity***

$p_d$ water vapor partial pressure in the air [Pa] (joined with temperature)

$p_{sat}$ saturation water vapor partial pressure in the air [Pa]

$\varphi$ relative humidity [%] (partial pressure and saturation partial pressure ratio),

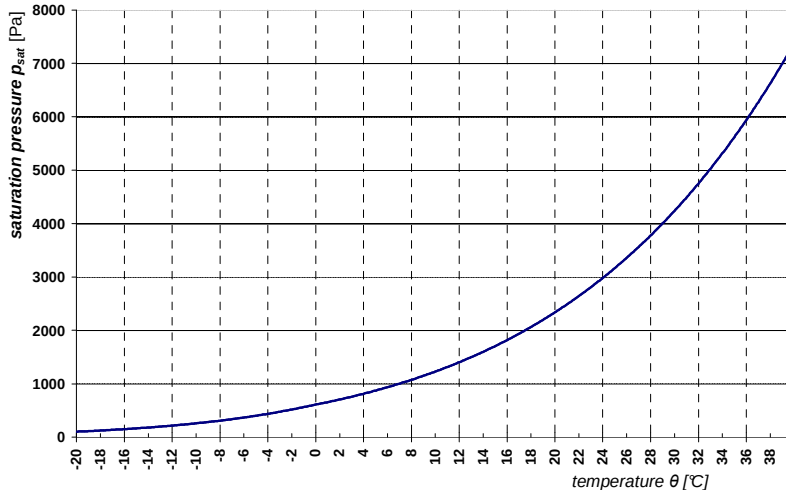



***Figure 2.1: Saturation pressure dependence on the temperature***

Base interrelations among air humidity quantifiers:

$$c = \frac{p_d}{R.T} \tag{1}$$

$$x = 0.621 \,.\, \frac{p_d}{p_{atm} - p_d} \tag{2}$$

$$\varphi = \frac{p_d}{p_{sat}(\theta)} \tag{3}$$

with

$p_{atm}$ atmospheric pressure [Pa]

$p_{sat}\,(\theta)$ saturation water vapor at the temperature $\theta$ [Pa]

$R$ gas constant for water vapor [J.(kg.K)$^{-1}$]

$T$ air temperature [K]

$\theta$ air temperature [°C]

The normative temperatures $\boldsymbol{\theta_{si}}$ **= 9.3°C** / $\boldsymbol{\theta_{si,80}}$ **= 12,6 °C** preserves condensation avoiding / mould occurrence avoiding [STN].

## 2.1. Empiric law and mass balance in the volume

In the case of both gasses (dry air and water vapour) being ideal and the internal and external air being homogenous) partial vapour pressure in the air dependence on the temperature is described by *Gay-Lussac's law:*

$$\boldsymbol{p \,.\, V = m \,.\, R \,.\, T} \tag{4}$$

with

$p$ partial pressure of the gas [Pa]

$V$ volume [m$^3$]

$m$ weight [kg]

Combining it with *mass balance in the space*, we obtain a dependence of partial vapour pressure on the several influencing parameters:

$$\boldsymbol{G + \frac{n.V\,.p_e}{R.T_i} = \frac{n.V\,.p_i}{R.T_i} + \frac{V}{R.T_i}.\frac{dp}{dt}} \tag{5}$$

with

$p_i$ , $p_e$ indoor, outdoor particular vapour pressure [Pa]

$V$ volume of the room [m$^3$]

$m$ weight [kg]

$R$ gas constant, for water vapor $R = 462$ [J/(kg.K)]

$T$, $T_i$ temperature, inside temperature [K]

$G$ vapour production [kg/h]

$n$ ventilation rate [h$^{-1}$]

t time [s]

In steady state ($\frac{dp_{di}}{dt} = 0$) the solution of (5) is

$$p_i = p_e + \frac{G.R.T_i}{n.V} \tag{6}$$

As the moisture problems coheres with saturation and saturation depends on the temperature, it is important to know the temperature distribution in the construction, and to control the most cold places on the inside surface. For this purpose we use the transient *heat conduction equation*

For the sake of the temperature field - temperature distribution in the investigated domain (fragment of building construction) we use 2D transient heat equation with Newton's boundary conditions:

$$\rho.c.\frac{\partial T}{\partial t} = -\, div\,(\ \lambda\ .\ \nabla T) \tag{7}$$

$$q_{ci} = h_{ci}\,(T - T_{ai}) \tag{8}$$

$$q_{ce} = h_{ce}\,(T - T_{ae})$$

and initial condition (initial temperature distribution in the domain):

$$T(x,y,0) = T_0(x,y) \tag{9}$$

Numerical calculation is realized by using the FEM software.

## 3. Situation

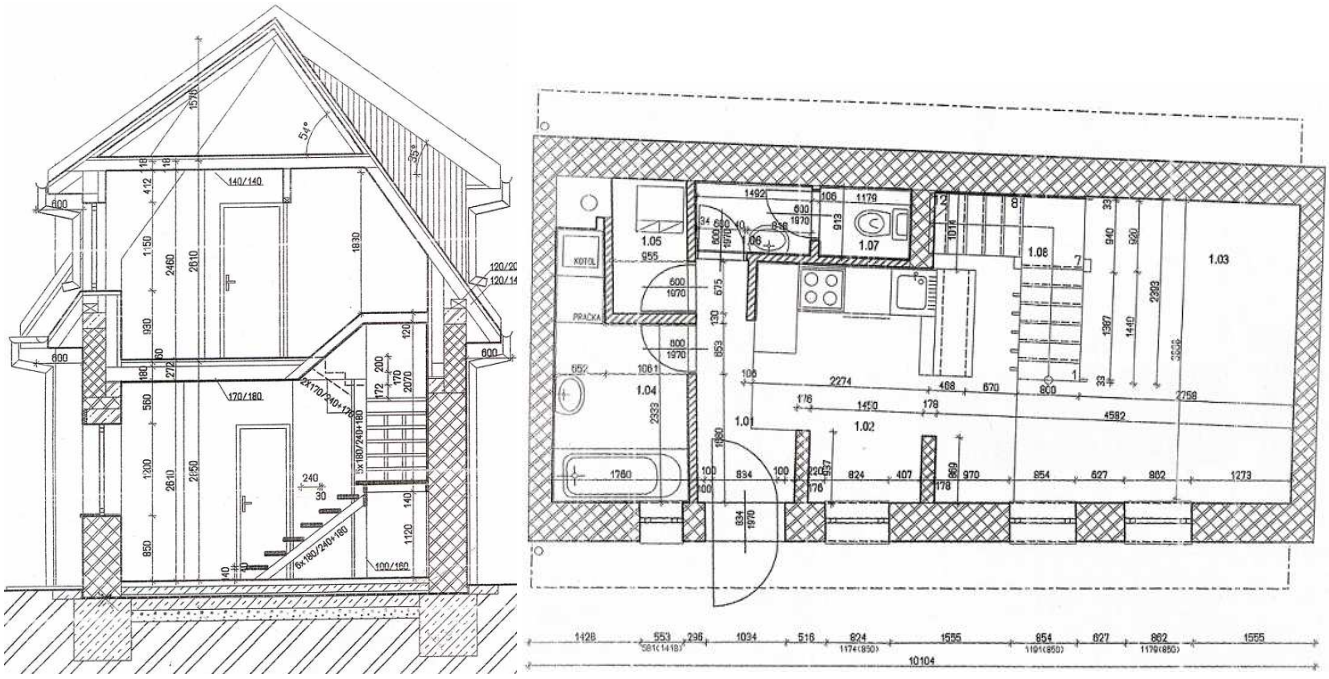



***Figure 3.1: The old house to be retrofitted. Vertical and horizontal (ground floor) section***

After the reconstruction including windows changing the indoor climate of the old family house (see figure 3.1) with stony and clue walls suddenly changed. The dampness increased so that the wet plaster splashes, molds occurred on the inside surface of the walls and furniture. Purposeful improving of the ventilation controlled by indwellers improved, but did not solve the situation. The owners decided for the general thermal insulation. To make the insulation optimal, the previous computations are useful.

## 4. Computer implementation

It is necessary to monitor all critical places (places with minimal surface temperature) of the house, see figure 4.1, choose the type and magnitude of the retrofitting material properly. The computation done for the actual stage showed the cause of excessive moisture – too low temperature, see figure 4.1

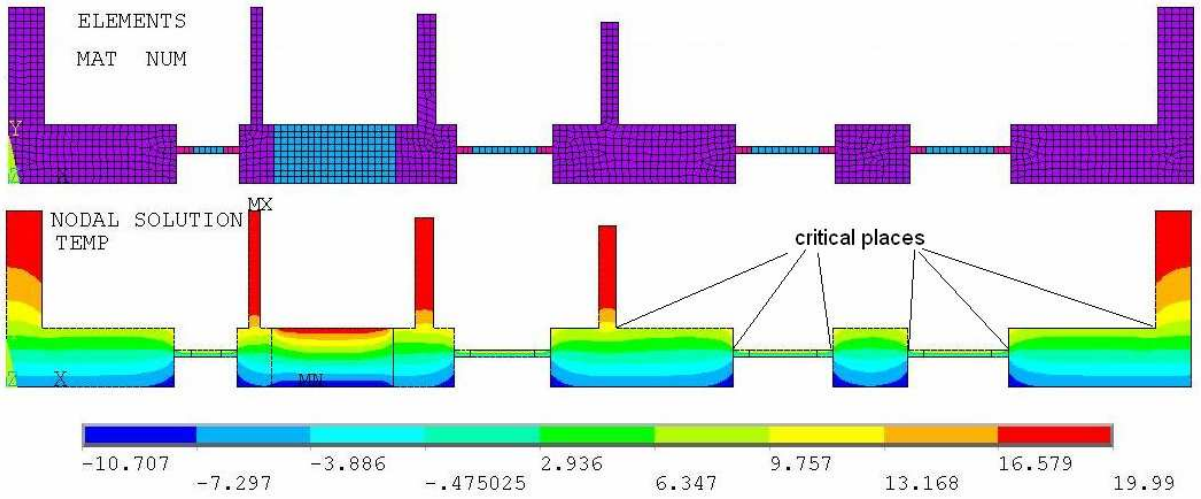



***Figure 4.1: Fragment of the horizontal section (meshed domain/temperature distribution)***

The temperatures in 16 critical points are too low, all values are below the normative values for moisture problems avoiding close to such cold places, see chapter 2.

By the retrofitting, the thermal resistance of the construction will increase and the temperature on the critical places will be sufficient for moisture problems avoiding. See figure 2.1, saturation line. It illustrates the dependence of saturation and the temperature.

Optimal parameters of the materials (which kind of the material should suit better, and what magnitude; also price should be included into the consideration) intended to be used for insulation can be stipulated by previous calculations

Variable List

| Node | Result Item | Minimum |
|---|---|---|
| | Time | 1 |
| 62 | Nodal Temperature | 9.28851 |
| 57 | Nodal Temperature | 4.23695 |
| 360 | Nodal Temperature | 5.03004 |
| 341 | Nodal Temperature | 12.6081 |
| 384 | Nodal Temperature | 12.2721 |
| 677 | Nodal Temperature | 11.3878 |
| 655 | Nodal Temperature | 11.5517 |
| 650 | Nodal Temperature | 4.83617 |
| 894 | Nodal Temperature | 4.45738 |
| 796 | Nodal Temperature | 10.8616 |
| 821 | Nodal Temperature | 10.8438 |
| 853 | Nodal Temperature | 4.25046 |
| 1157 | Nodal Temperature | 4.17706 |
| 1141 | Nodal Temperature | 4.16306 |
| 1255 | Nodal Temperature | 4.22761 |
| 1260 | Nodal Temperature | 9.29047 |

***Figure 4.2: Numerical output from the FEM software, nodal temperature in critical places***

## 4.1. Retrofitting concept

After some computations, comparisons, optimizations, the rock wool material of the 5 cm magnitude was chosen as the suitable thermo isolating layer on all building envelope outer surface of the house, see figure 4.3.

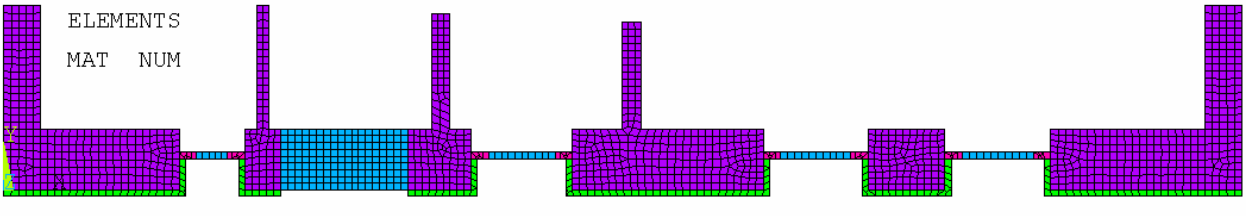



***Figure 4.3: Retrofitting concept, horizontal section. Meshed domain.***

New calculations – after supposed insulating of the house - predicts, that indoor climate of the house improves. I.e. in the 16 points of our interest (critical points of the inner surface) the temperature increases essentially, see figure 4.4. Consequently the risk of water vapor condensation on the surfaces will become negligible.

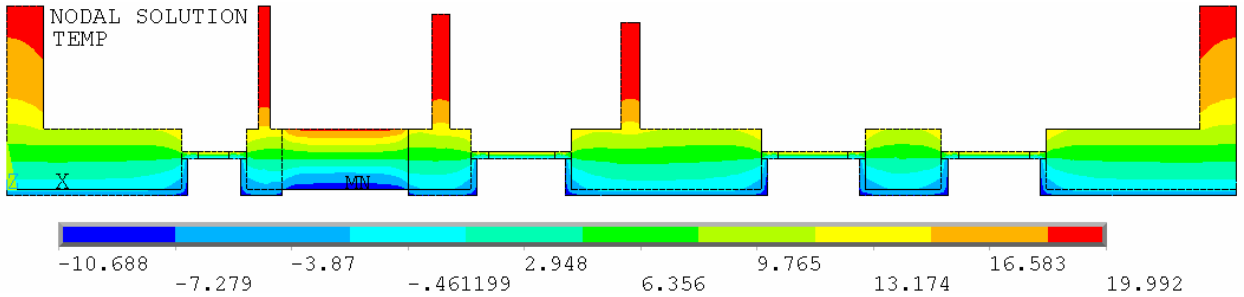



***Figure 4.4: Retrofitting concept, horizontal section. Isotherms***

Variable List

| Node | Result Item | Minimum |
|---|---|---|
| | Time | 1 |
| 62 | Nodal Temperature | 11.7325 |
| 57 | Nodal Temperature | 8.15291 |
| 360 | Nodal Temperature | 8.43613 |
| 341 | Nodal Temperature | 14.1319 |
| 384 | Nodal Temperature | 13.8227 |
| 677 | Nodal Temperature | 13.0959 |
| 655 | Nodal Temperature | 13.3281 |
| 650 | Nodal Temperature | 8.48343 |
| 894 | Nodal Temperature | 8.40816 |
| 796 | Nodal Temperature | 12.9739 |
| 821 | Nodal Temperature | 12.9485 |
| 853 | Nodal Temperature | 8.17926 |
| 1157 | Nodal Temperature | 8.0117 |
| 1141 | Nodal Temperature | 8.005 |
| 1255 | Nodal Temperature | 8.12746 |
| 1260 | Nodal Temperature | 11.7397 |

***Figure 4.5: Nodal temperature in the places with locally minimal temperature after retrofitting***

## 5. Conclusion

Some changes on the building construction – reconstruction, windows changing, retrofitting, etc, needs previous consultations with an expert even before their realizations. If needed, also the calculations with the aim of problems prediction are worthwhile.

## 6. References


[1] HALAHYJA M., CHMÚRNY I., STERNOVÁ Z.: Thermal Engineering of Buildings. Thermal Protection of Buildings. Jaga Group, 1998

[2] CHMÚRNY I., MINÁROVÁ M.: Risk of Mould Growth in Buildings, Journal of Civil Engineering 4/04

[3] MINÁROVÁ M.: Deformed temperature fields. Proceedings from PRASTAN Conference, Kočovce May 2004

[4] REKTORYS,K.: The Method of Discretization in Time and Partial Differential Equations, Prague 1982

[5] Report Annex XIV, Volume 1 – Sourcebook, International Energy Agency – Energy Conservation in building and community systems, 1991

[6] STN 73 05 40-1 Thermal Properties of Building Structures and Buildings.

[7] www.ansys.com

[8] www.nafems.com


## Address of author:


Minárová Mária, RNDr., PhD.
SvF STU
Radlinského 11
81368 Bratislava
minarova@math.sk

# Cena kvality

Fabian Oropeza

## Abstrakt

Základním úkolem managementu by měla být neustálá snaha o zvýšení kvality produkce. K tomu, aby tohoto cíle mohlo být dosaženo, musí management zajistit, aby se všechny operace ve společnosti řídily základním principem, tzn. že náklady a kvalita jsou komplementární a nemají rozporné cíle. Dříve zněla manažerská doporučení tak, že je nutné zvolit mezi náklady a kvalitou (tzv. trade-off rozhodnutí), protože lepší kvalita bude zkrátka stát více a bude spojena s většími obtížemi při realizaci produkce. Experimenty po celém světě však dokázaly, že toto schéma nemusí vždy platit a nyní už to začínají chápat i manažeři společností. Dobrá kvalita vede k vyšší produktivitě, ke snížení nákladů na kvalitu a postupně ke zvýšení prodeje, a průniku na nové trhy a zvýšení zisků.

Abychom mohli více rozvinout koncept nákladů na kvalitu, je nutné přesně stanovit rozdíl mezi náklady na kvalitu a náklady na organizaci kvality. Je důležité nebrat náklady na kvality jako pouhé výdaje, které se s růstem kvality zvyšují. V zásadě lze říci, že pokaždé, kdy je potřeba něco přepracovat, náklady na kvalitu rostou. Jako zřejmý příklad může sloužit přepracování průmyslově vyrobeného zboží, přestavování pracovních nástrojů či oprava chyb v bankovním výpise.

## Úvod

Cilem analyzý nákladů na kvalitu je poskytnout managementu nástroj pro realizaci programu sledování kvality a činností vedoucích ke zvýšení kvality. Vysledky těchto analysis mohou být vhodným způsobem využity k diagnostice silných i slabých stránek systému sledování kvality. Příslušné týmy je mohou využít k popisu pozitivních dopadů (vyjádřeno v penězích) a k vysvětlení zamýšlených změn.

## 1. Cíle systému kvalita cena

Cílem jakéhokoli systému nákladů na kvalitu je umožnit kvalitativní vazby; strategie pro použití nákladů na kvalitu je poměrně jednoduchá. (1) za měřit se s velkým důrazen na naklady selhání ve snaze snížit je na nulu; (2) investovat do "správné" preventivní činnosti, což by s sebou mělo přinést zlepšení; (3) snížit náklady na oceňování v závislosti na dosažených výsledcích, a (4) neustále vyhodnocovat a usměrňovat úsilí o prevenci vad za účelem dosažení dalšího zlepšení.

## 2. Management nákladů na kvalitu

Program nákladů na kvalitu by měl vždy být představen pozitivním způsobem. Pokud se tak nestane, je pravděpodobné, že se nesetká s velmi dobrým ohlasem, protože bude (najednou) vykazovat velké množství chyb, plýtvání a výdajů, které jsou nadbytečné v každé firmě s vysokou úrovní kvality produkce. Z tohoto důvodu je velmi důležité, aby všichni dotčení pracovníci byli pečlivě informováni a pochopili, že tento krok ve skutečnosti zlepšuje ekonomiku výrobního procesu. Není až tak důležité, jaké jsou výchozí údaje. Je však nutné, aby byl program

jednoduchý a praktický; jen tak je možné dosáhnout zlepšení kvality na všech stupních výroby. Zavádění tohoto programu je tedy potřeba velmi pečlivě naplánovat tak, aby bylo dosaženo zamýšlených cílů.

## 3. Kategorie nákladů na kvalitu

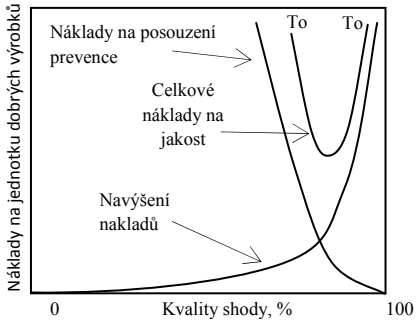
Náklady způsobené selháním můžeme rozdělit na náklady vnitřní a vnější.

**3.1 Náklady na prevenci** – náklady na to, aby se zabránilo nedostatečné kvalitě výrobků nebo služeb. Definice těchto nákladů je často chápána špatně, a to zejména tehdy, pokud je aplikace výše uvedené definice nejasná. Dodatečné náklady na prevenci mohou vznikat v souvislosti s testováním již porušeného výrobku/služby tak, aby bylo zamezeno ještě vyšším nákladům na případné selhání (např. dodatečné kontroly a opravy, které mají zabránit nově objeveným chybám ještě předtím, než se zboží/služba dostane k zákazníkovi). Tyto dodatečné náklady na identifikaci problému (na opravy či na analýzy příčin selhání) mohou být chápány jako snaha o prevenci budoucího problému. V praxi není až tak důležité, do které kategorie náklady spadají; důležité je, aby byla zachována konzistence.

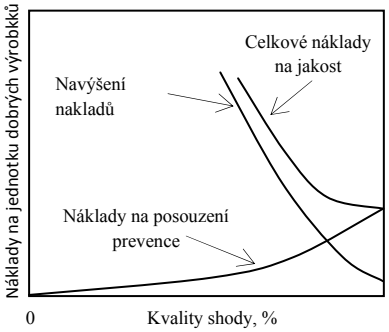
**3.2 Náklady na ocenění** – náklady spojené s měřením, hodnocením a auditem produktů/služeb za účelem splnění požadavků na kvalitu a výkon produktů/služeb.

**Příklad:**

Klasický model optimálních nákladů na kvalitu je na obr. 1. Náklady na prevenci a ocenění byly dříve znázorňovány jako asymptoticky rostoucí, dokud nebylo dosaženo úrovně dokonalé kvality. Nové technologie by měly snížit míru selhání materiálu a produkce, robotika a jiné formy automatizace snížily míru selhání lidského faktoru, jak ukazuje obra 2.

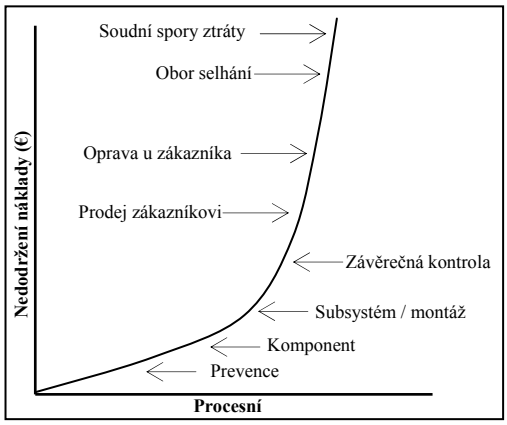



**Obrázek 1** Klasický model kvality nákladů.




**Obrázek 2** Nový optimální model kvality

Čím později je během výrobního procesu objevena chyba, tj. čím blíže je vadný produkt k užívání spotřebitelem, tím dražší je problém napravit. obr. 3 je příkladem tohoto pojetí převzatý z výroby, ale stejný princip platí i pro služby. Poznatky získané z tohoto zlepšení pak mohou být prostřednictvím prevence použity na všechny nové práce.




**Graf 3** Nedodržení náklady jako funkce detekce bodu v procesu.

**SPC – statistické řízení procesů**
SPC pomáhá při optimálním řízení procesů v průmyslu i v nevýrobních aplikacích – v bankovnictví, telekomunikacích či zdravotnictví. Pro řešení těchto procesů si vyzkoušíme Ishikawův diagram, Paretovu analýzu, histogram i regulační diagramy různých typů.

**3.3 Náklady vzniklé selháním** – náklady, které vzniknou, pokud výrobek či služba nesplní požadavky na kvalitu či potřeby zákazníků a spotřebitele, tzn. náklady plynoucí z nedostatečné kvality.

*Vnitřní náklady způsobené selháním* – náklady, které vzniknou předtím, než je produkt předán zákazníkovi. Jako příklady mohou sloužit tzv. zmetky, dodatečné testování, dodatečná kontrola, opravy či kontrola materiálu.

*Vnější náklady způsobené selháním* – objevují se poté, co byl produkt doručen zákazníkovi. Jako příklady uveďme vyřizování reklamací, navrácení výrobku zákazníkem, záruční servis či stažení výrobku z prodeje z důvodu nedostatečné kvality.

## 4. Analýza trendů a pokroku

Pokud chceme náklady na kvalitu smysluplně využít, je potřeba s nimi pracovat tak, aby byla umožněna jejich analýza. Jak již bylo řečeno, můžeme toho dosáhnout například tak, že na ně budeme pohlížet jako na podíl celkových nákladů. Pokud definujeme tento podíl, následující logickou fází je analyzovat dosažený výsledek tak, abychom mohli určit, zda se situace v čase zlepšuje či naopak. To je základem pro plánování realistických cílů zvyšování kvality do budoucího období. Jak již bylo řečeno, je nutné rozlišovat krátké a dlouhé období. V dlouhodobém hodnocení obvykle posuzujeme celkové náklady na kvalitu během dlouhého časového úseku. Toto hodnocení je používáno nejčastěji pro strategické plánování a manažerské sledování celkového pokroku. Krátkodobé trendové tabulky se potom připravují pro každou

oblast činnosti společnosti, kde je potřeba stanovit individuální cíle pro zlepšení nákladů na kvalitu. Jedním z přístupů ke krátkodobým trendům je určit jeden cíl pro každou oblast výroby, další možností je vypracovat tak detailní analýzu, jak jen to daný systém umožňuje.
Efektivní program nákladů na kvalitu má tyto součásti: (a) stanovení systému měření nákladů na kvalitu a vývoj vhodné trendové analýzy v dlouhém období (b) stanovení každoročních cílů celkového zvyšování kvality při současném snižování kvality (c) vývoj krátkodobé trendové analýzy s dílčími cíli, které společně vytvářejí požadavek na každoroční cíle

## 5. Implementace programu nákladů na kvalitu

Jakmile určíme úroveň nákladů na kvalitu, ihned by měly být navrženy možnosti ke zlepšení situace. Výsledky analýzy by měly dostatečně jasně poukázat na to, že je potřeba program zavést. Počáteční odhady nákladů na kvalitu se běžně pohybují i na úrovni 20 % hodnoty prodeje. Ačkoliv nelze provádět přímá srovnání, údaje z některých firem, které mají výraznější zkušenosti s rozsáhlým programem zdokonalení kvality a snížením nákladů na kvalitu, ukazují, že celkové náklady na kvalitu mohou být sníženy až na 2 – 4 % hodnoty prodeje. O toto snížení se poté zvýší zisk společnosti.
Nyní je tedy možné připravit zevrubný plán a časový rozvrh implementace programu nákladů na kvalitu. Jeho základní prvky by měly být následující:

- Prezentace pro management: měla by obecně představit možnosti, které jsou k dispozici, na příkladech ukázat, jakým způsobem lze dosáhnout uvedených pozitivních efektů, a získat podporu managementu pro program.
- Naplánování pilotního programu
- Vyškolení všech zainteresovaných pracovníků tak, aby se seznámili s plánovaným programem a aby se do něj aktivně zapojili.
- Vytvoření postupu vnitřního účetního zpracování nákladů na kvalitu.
- Celkový sběr a analýza údajů o nákladech na kvalitu.
- Podávání zpráv o nákladech na kvalitu (ve vztahu k systému řízení nákladů na kvalitu a programu na zlepšení kvality)

## 6. Pilotní program

Důvody pro zavedení pilotního programu

- Ověří fun€nost systému a jeho schopnost produkovat takové výstupy, které ušetří náklady na kvalitu.
- Znovu přesvědčí management o nutnosti zavedení programu nákladů na kvalitu.
- Má omezenou působnost co do počtu zapojených osob i výrobních provozů
- Umožní “vychytat chyby” předtím, než bude zavedena plná implementace

### 6.1 Vyškolení pracovníků

Cílem tohoto předávání informací je získat si podporu u těchto pracovníků pro program samotný a vysvětlit jim očekávané kladné výsledky. Bez jejich spolupráce bude implementace značně obtížná.

Jednotlivá oddělení by měla dostat příležitost vyjádřit se k plánovanému programu jako celku a ujasnit si v něm svou pozici. Je velmi důležité podporovat odborná oddělení v tom, aby k programu vznášely připomínky ze svého pohledu. Mohou si například připravit seznam aktivit v rámci svého oddělení, kde vidí příležitost ke zlepšení kvality nebo činnosti, které by se podle jejich názoru nemusely provádět, kdyby kvalita produkce byla perfektní.

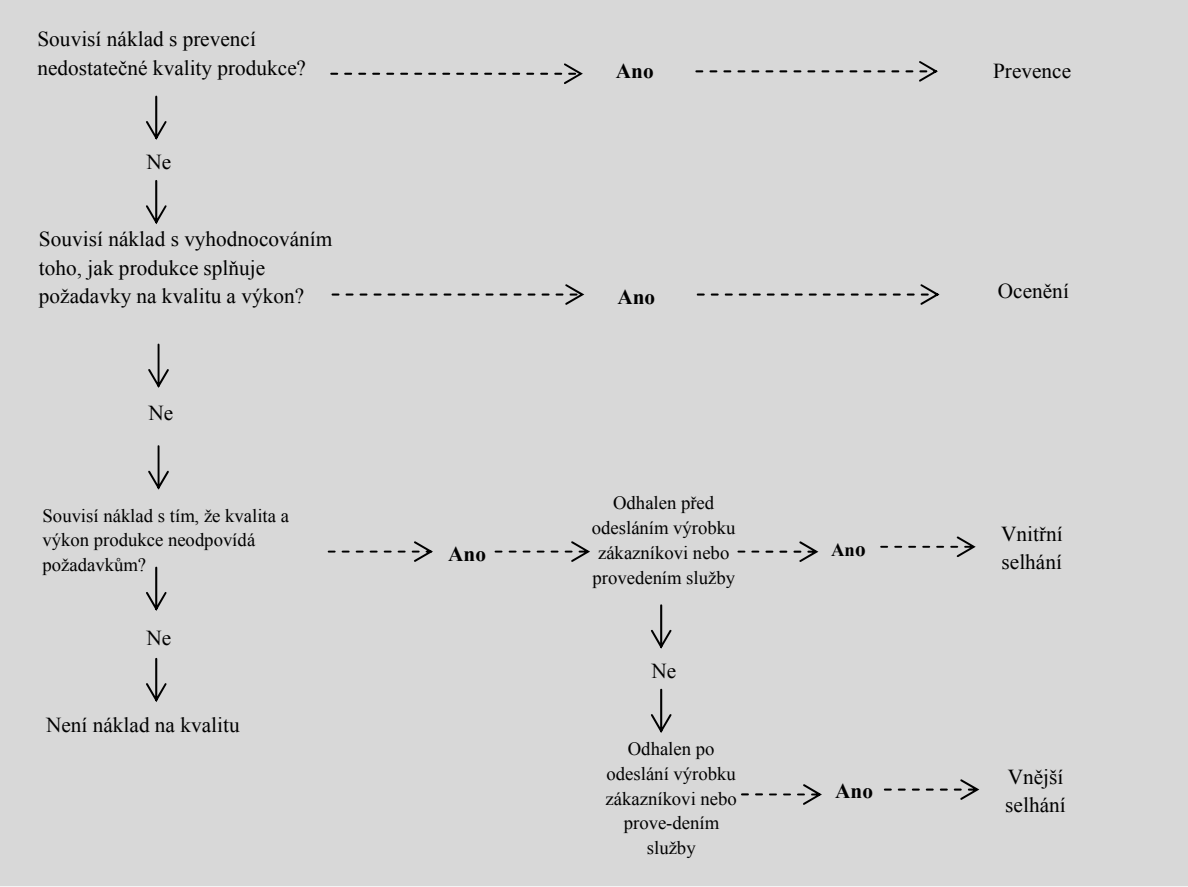



**Schéma 1.1** Zařazení jednotlivých nákladů na kvalitu do kategorií.

## 6.2 Zpracování nákladů na kvalitu v rámci firmy

Ve vnitřním zpracování nákladů na kvalitu je nutné popsat každý prvek nákladů, který ve firmě figuruje, a definovat, jak a kdy budou sbírány či odhadovány reálné náklady na kvalitu. V zájmu vyšší spolehlivosti dat, které souvisí s reálnými náklady společnosti, by pří postupu zpracování nákladů mělo byt definováno, jak se účtování výrobních nákladů přizpůsobí novému konceptu sledování nákladů kvality. Tento postup zpracovaní by měl také určit, kdo je zodpovědný za vykonávání programu a jak se budou vykazovat výsledky (jako příklad viz tabulku 1.2 i 1.5)

| Popis / Element | | Ucetnicví | Administrativa | Strojirenství | Odhád | Terénní služby | Výrobní inženýrství | Terénní služby | Marketing | Produkce | Výrobní kontrola | Kvalita | Příjem | doprava | Úhrny |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.1.1. | Marketing uživatel | | | | | | | | | | | | | | |
| 1.1.2. | Zadávácí dokumentace recenzi | | | | | | | | | | | | | | |
| 1.1.3. | Značkovací kvality pokrok recenze | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| 1.2.1. | Kvalitu designu pokrok recenze | | | | | | | | | | | | | | |
| 1.2.2. | Značkovací podpůmé činnosti | | | | | | | | | | | | | | |
| 1.2.3. | Produktový design kvalifikační zkouška | | | | | | | | | | | | | | |
| 1.2.4. | Služba design | | | | | | | | | | | | | | |
| 1.2.5. | Polních pokusů. | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| 1.3.1. | Dodavatel recenzi | | | | | | | | | | | | | | |
| 1.3.2. | Dodavatel rating | | | | | | | | | | | | | | |
| 1.3.3. | Nákup pořadí TECH DATA recenzi | | | | | | | | | | | | | | |

**Tabulka 1.** Souhrnné údaje týkající se nákladů na kvalitu

## 7. Od klíčových nákladů k jejich příčinám

Cílem tzv. Activity Based Costing (ABC) je vylepšit celkovou efektivitu vynaložených nákladů tím, že se budeme soustředit pouze na jejich klíčové prvky. Metodologie nákladů na kvalitu se snaží přiřadit náklady související s kvalitou specifickým aktivitám, produktům, procesům či oddělením. Tím je možné se na tyto náklady snadněji zaměřit. Díky použití metody ABC je tedy možné snadněji nalézt a adresovat tyto náklady.
Vzorovým příkladem je metody ABC, kde výstupem jsou grafy, tabulky a kalkulace. Data, které zde budu používat jsem získal od společnosti Valeo. VALEO Compresor Europe s.r.o. Vyrábí kompresory do klimatizačních jednotek osobních automobilů. Zákazníky jsou VW group, Renault, Nissan, Volvo, Fiat a TPCA Kolín.V tomto případě použiji metodu ABC, kde byl identifikován roční náklad na vnitřní selhání ve výši 19,385 €.- , který byl spojen s výrobou dva komponent hlava valce a volant s pedály. Schéma 1.2 znázorňuje, jak jsou náklady přiřazeny těmto dvěma produktům a zároveň příčinám způsobené selháním.

Díky tomu, že známe detailní rozdělení nákladů, je možné odhadnout, kolik bude stát odstranění připadneho problému. V tomto příkladě jsou náklady na "poškození" všech produktů € 8546,-. Pokud bychom byli schopni odstranit 75 % poškození tím, že identifikujeme a napravíme původní příčiny problému, ušetřili bychom ročně € 6,397.-. Odhadované úspory je poté možné porovnat s cenou investice. Pokud je například nutná investice ve výši € 3,230.- je doba návratnosti investice 12 měsicu x € 3,230.- / € 6,397.- , tzn. přibližně 6 měsíců. Detailní zpracování nákladů s použitím metody ABC nám tedy umožňuje přesnou analýzu nákladů

a výnosů a je základem pro kvalifikované rozhodnutí o investicích do zvyšování kvality produkce.

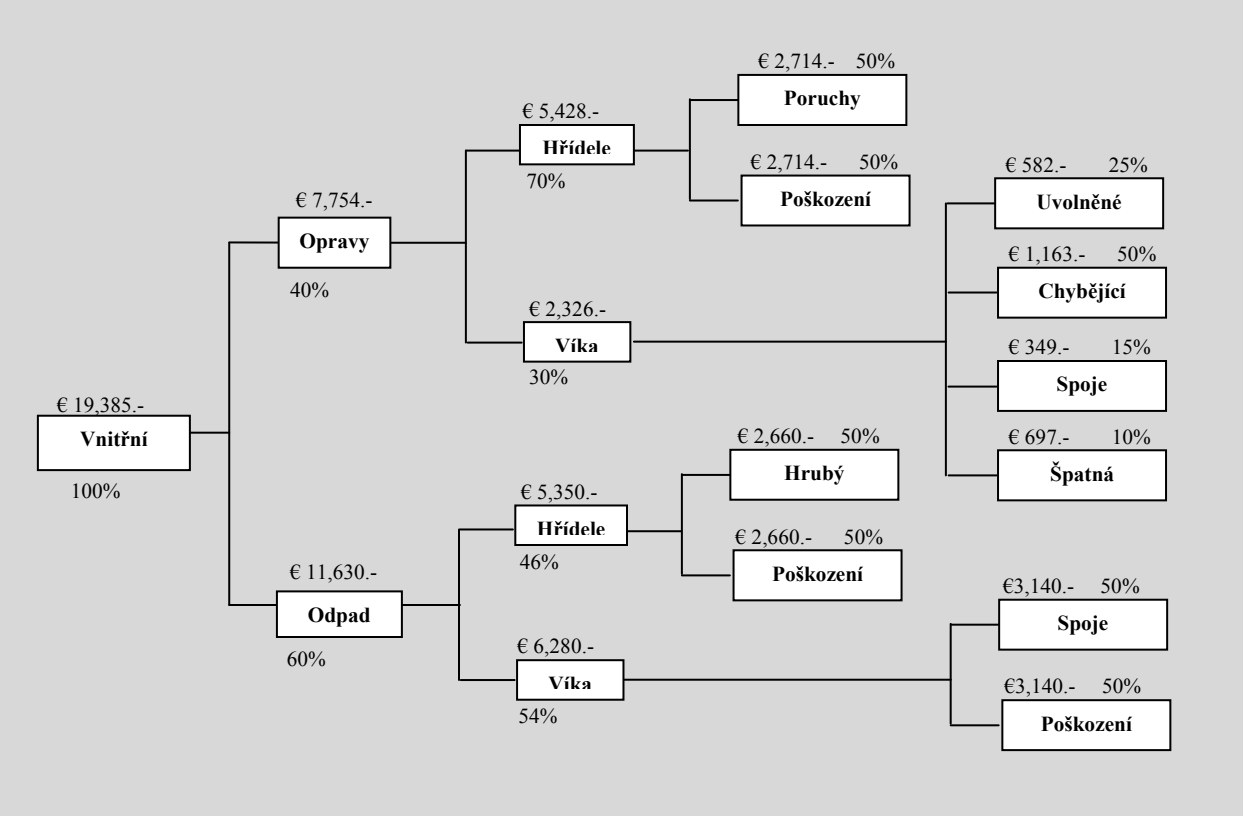



**Schéma 1.2** Členění nákladů na

## 8. Použití kvalitativních nákladů

### 8.1 Aplikace nákladů na kvalitu na kontrolu dodavatelů

Kupující, aby omezil kvalitativní náklady týkající se dodavatele, musí nejprve určit, jaké náklady jsou důležité. Porovnání relativního významu kvalitativních nákladů podle kategorie a základního prvku by měl být první krok. Dalším krokem je provést Paretovu analýzu ke zjištění, kteří dodavatelé působí problém. Nicméně, jestli dodavatel zjistí, že kupující používá takový program, nejviditelnější náklady pravděpodobně snížení. Budou-li tyto náklady sníženy, skryté náklady vydané jednak kupujícím jednak několika důležitými dodavateli by měly být rovněž sníženy. Výsledkem bude, že kvalita produktu/služby dodavatele stejně jako produktu/služby kupujícího se zlepší.

### 8.2 Výdaje na dosažení požadované kvality

Možná aplikace kvalitativních nákladů v dodavatelsko-hodnotícím programu je provedena u činnosti jednoho výrobce u pneumatických kompresorů produktů.

Na příkladu teto továrny, která vyrábí mimo jiné i tyto dva produkty (hlava valce a volant s pedály), si ukažme, jak budou zpracovány nepřímé náklady (konkrétně zpracování materiálu) s využitím běžných účetních metod a s využitím ABC.

| | | |
|---|---|---|
| Doba práce | = | 1000 |
| Celkové náklady na práci | = | € 15,000.- |
| Pracovní náročnost výroby Hlava valce a Volant s pedály | = | 0.25 hodin |
| Počet zpracovávaných jednotek | = | 1200 |
| Náklad na zpracování materiálu 260,-€ za jednotku | = | €12,308.- |
| **Požadavky na zpracování materiálu** | | |
| Hlava valce | = | 2 Jednotky na 100 ks |
| Volant s pedaly | = | 4 Jednotky na 100 ks |

Každá hodina práce s sebou dále nese nepřímý náklad:
[€ 15,000.- / 1,000 hodin = € 1,50.- za hodinu]
[0.25 hodin x € 1.5.- za hodinu = € 0.38.- (hřídele nebo víka)]
Vzhledem k tomu, že náklad na zpracování materiálu na výrobu vík je € 10.- za jednotku, tzn. € 369.-, na jednu jednotku vík pak připadá náklad € 0.4.-. Analogicky připadá na jednu jednotku hřídelí náklad € 0.2 .-.

**Aplikace ABC**

Využití ABC ve zmíněném příkladu:
[€ 12,000.- / 1200 jednotek = € 10.- za jednotku]
[€ 40.- / 100 Hlav valců = € 0.4.- **za hlavů valce**]
[€ 20.- / 100 Volantů s pedály = € 0.2.- **za Volant s pedály**]

V uvedeném příkladu je klíčovým nákladem zpracování materiálu, konkrétně počet vypravených jednotek. Ve skutečnosti obvykle existuje mnoho klíčových nákladů, např. na objednávku, na nastavení strojů, na přezkum stížností zákazníků, údržbu atd. Pokud jsou náklady na vyřízení jedné stížnosti € 650.- a společnost přijme 10 stížností měsíčně, potom je dle principů ABC potřeba zaúčtovat € 6,500.- na účet, náklady na vyřízení stížností zákazníků“ za dané období.

Pro tento vzorový případ je ke každému typu kompresoru přiřazený jeden dodavatel.

$$QCPI = \frac{VDRK + PN}{VDRK}$$

QCPI : Quality Conformance Preliminary Inspection
VDRK : Výdaje na dosažení požadované kvality
PN : Pořizovací náklady (Výrobní náklady)

| Náklady | Koncept | Přiklad |
|---|---|---|
| Odmítnutí při přejímce | Počet odmítnutých položek dodavatele násobené nákladem. | 20 [odmítnutý/lot] * € 808.- [lots/odmítnutý] = € 16,154.- |
| Vyřizování reklamací | Byla odhadnuta doba na vyřizování pro dodavatele, takže se vynásobí průměrná hodinová mzda. | 200 Rozsah usporý času * € 16.- = € 3,200.- |
| Kontrola kvality | Z průměrné hodinové mzdy inspektora a počtu kontrolovaných jednotek. | 1.00 Std. hod/lot * 0.62. - €/hod * 80 Lot = € 49.- |
| Vadný výrobek zjištěný inspekcí | Byla odhadnuta pro každého dodavatele vynásobením počtu vadných dílů zjištěné inspekcí původní pořizovací cenou. | 100 odmítnutý. kus * € 3.- nakup. cena/kus = € 300.- |
| **dodavatel kvalitních výdajů** | **Součet výše nákladů** | **€ 19,703.-** |

## 9. Klasifikace dodavatelů je součástí kvality nákladů

| Zákazníci | Typ kompresor | Dodavatelé | Výdaje na dosažení požadované kvality | Pořizovací náklady (€) (Vyrobní naklady) | Index (QCPI) |
|---|---|---|---|---|---|
| VW group | SD6V12 | A | 135 | 39,446 | 1.006 |
| Renault | SD7V16/SD6V12 | B | 750 | 67,846 | 1.232 |
| Nissan | DKS-16H | C | 1,668 | 121062 | 1.289 |
| Volvo | DKS16/15CH | D | 404 | 20,956 | 1.405 |
| Fiat | TV12SC | E | 2,261 | 112,800 | 1.421 |

Příklad výpočtu indexu za dodavateli

| Index (QCPI) | Výklad |
|---|---|
| 1.000 – 1.009 | Výborný |
| 1.010 – 1.039 | Dobrý |
| 1.040 – 1.069 | Vhodný |
| 1.070 – 1.099 | Dostatčný |
| 1.100 + | Vyžaduje okamžite nápravné opatření |

$$QCPI = \frac{VDRK + PN}{VDRK}$$

$$QCPI = \frac{€135 + €39{,}446}{€39{,}446} = 1.006$$

V idealním případě by měl být tento index roven:

$$QCPI = \frac{VDRK + PN}{VDRK} = \frac{€0 + €39{,}446}{€39{,}446} = 1$$

Při používání tohoto hodnocení bylo první prioritou pro vyrobce prověst okamžite korektivní opatření vzhledem k dodavatelum D a E.

### 9.1 Analýza výnosnosti vložených nákladů na kvalitu u dodavatelů

Cilém je vyřešit problém podniku, urychlit snížení začátečních obtíží. Při tomto rozhodování byl použit koncept návratnosti investic (return on investment - ROI):

$$ROI = \frac{Úspory * 100}{Investice} = \frac{€\,462 \text{ x } 100}{€\,15{,}000.-} = 3.8\%$$

Pro tuto situaci bylo potenciální snížení kvalitativních nákladů o € 462.- odhadnuto na investici € 15,000.- pro zajištění pomoci dodavateli E. Pokud cíle dosaženy, je tato investice efektivní.

### 9.2 Vliv kvality na výnos z prodeje

Důležitý výdaj, který není obvykle posouzen, je ušlá tržba v důsledku špatné kvality. Tato ušlá tržba je způsobena zákazníkovou nespokojeností se zbožím nebo poskytovanou službou.Tato nespokojenost může vyústit do ztráty současných zákazníků, „zákaznického zběhnutí“ a neschopnosti přitáhnout nového zákazníka kvůli poskvrněné pověsti kvality. V každém případě, vliv kvality na výnos z prodeje by měl být brán v úvahu, alespoň určením rozsahu zákazníkovy nespokojenosti a přijetím opatření, jak zlepšit udržení nynějších zákazníků a vytvořit okruh nových.

## 10. Závěr

Nalezení hlavních zdrojů nakladů na udržení kvalitní výrobý a odstranění zásadních příčin skýtá firmě výbornou možnost zajistit zákazníkům levnější a kvalitnější výrobky. V současnosti všechny firmy prohlašují, že mají produkty „nejlepší kvality“, ale toto tvrzení může být reálné jen tehdy, pokud užívají úspory v nákladech na kvalitu (dosažené úsilím kvalitativního zlepšení) k:

1) financování zlepšování vlastností produktu bez zvyšování ceny nebo
2) nižším cenám produktu s existujícími vlastnostmi.

Jakákoli cesta ke zvýšení hodnoty samozřejmě povede pro firmu k vyšším příjmům z prodeje.

## 11. Literatura

**[1]** CAMPANELLA, Jack.. Principles of Quality Costs, *Principy, Implementace a používání.* (Cap. 1 - 4) [z ang. orig. přeložil Fabian Oropeza]. 2008. Praha. nový trend v současné kvalitě.
**[2]** Valeo, [online] strana naposledy 2009-04-25. Dostupný z WWW:
http://www.valeohumpolec.cz/
**[3]** Valeo, Údržba a oprava pro klimatizace R 134a. [online]. Strana naposledy 2009-04-29. Dostupný z WWW:
http://www.autoclim.cz/dokumenty/7%20-%20Technick%C3%A9%20informace.pdf

**Adresa autora:**

Fabian Oropeza, Ing.
České vysoké učení technické v Praze,
Fakulta strojní, Ústav technické matematiky,
Karlovo nám. 13, 121 35 Praha 2
israelfabian.oropezapena@fs.cvut.cz

# Aplikácia štatistických metód v hydrológii
# Application of statistical methods in hydrology

Pastuchová Elena, Václavíková Štefánia

**Abstract:** The paper describes two ways of average monthly flow volume evaluations focused on the creation of the similar objects groups. The analysis and comparison of hydrological data is based on two methods of cluster analysis: K-means and fuzzy cluster analysis

**Key words:** Cluster analysis, K-means, fuzzy cluster analysis, mean monthly streamflows
**Kľúčové slová:** Zhluková analýza, K-means, fuzzy zhluková analýza, priemerné mesačné prietoky

## 1. Úvod

Určovanie priemerných návrhových prietokov pre malé a stredné povodia bez priamych pozorovaní patrí k častým úlohám inžinierskej hydrológie.

Ak sú dostupné reálne hodnoty charakteristík, je možné použitím vhodnej štatistickej metódy získať analýzu daného vodného toku a vyvodiť predpoveď. Územie Slovenska je charakteristické svojou členitosťou a rôznorodosťou, je potrebný odhad premenlivých charakteristík aj na časti vodných tokov bez meracích staníc. Snahou hydrológov je realizovať predpovede správania sa vodných tokov aj na oblasti, v ktorých neprebiehajú merania.

Z matematických metód je pri spracovaní hydrologických dát v prvej fáze riešenia úloh vhodné aplikovať zhlukovú analýzu. Táto metóda umožňuje spájať objekty s podobnými hydrologickými vlastnosťami a vytvoriť z nich skupiny s čo najväčšou mierou podobnosti vo vnútri skupín a s čo najväčšou mierou heterogenity medzi skupinami. Prvky takto vytvorených skupín nemusia tvoriť uzavretý geografický región.

V tomto príspevku sme spracovávali údaje o dlhodobých mesačných prietokoch nameraných na 209 malých a stredných povodiach Slovenska v rokoch . Cieľom je porovnanie metódy K-means a fuzzy zhlukovania. Dáta boli čerpané zo zdroja Slovenského hydrometeorologického ústavu.

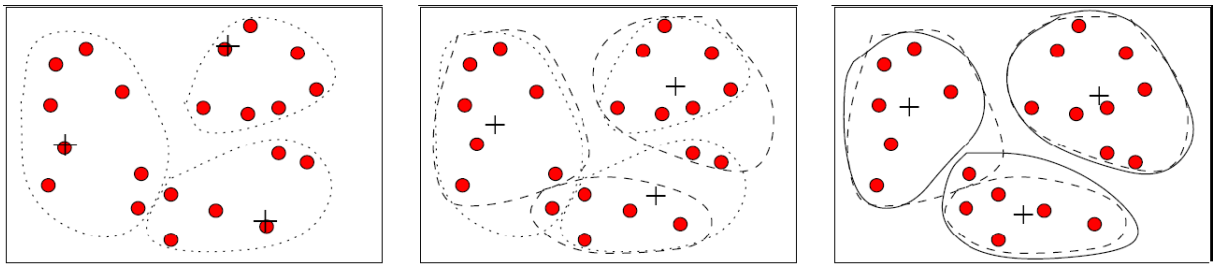
## 2. Metóda K-means

V súčasnosti k najčastejšie používaným nástrojom na vytvorenie zhlukov patrí metóda K-means. Táto nehierarchická metóda využíva K-means algorimus. [Meloun, Militký, Hill]

Ide o iteračný algoritmus, ktorý rozdeľuje analyzované objekty do *k* zhlukov tak, že vzdialenosť medzi všetkými objektmi a centrami zhlukov je minimalizovaná.

Obrázok 1. schematicky znázorňuje tri iterácie pri zhlukovaní pomocou K-means.

1. Z množiny všetkých objektov je náhodne vybraných k objektov, ktoré budú reprezentovať dočasné centrá zhlukov, označené „+“.V nasledujúcom kroku sa k vybraným centrám zhlukov priradia jednotlivé objekty tak, aby ich vzdialenosť k príslušným centrám bola čo najmenšia. Obr.1.a)
2. Po vytvorení prvotných zhlukov sa prepočítajú pozície všetkých objektov, vytvoria sa nové centrá, prepočítajú sa vzdialenosti medzi objektmi a centrami a vytvoria nové zhluky. Obr.1.b,c)
3. Algoritmus sa ukončí v momente, keď nenastane zmena rozloženia zhlukov, alebo je splnená ukončovacia podmienka n interácií.



a) b) c)

***Obrázok 1: Zhlukovanie pomocou K-means***

Subjektivita K-means spočíva v stanovení počtu zhlukov, ktoré chceme dosiahnuť a neexistuje všeobecné teoretické riešenie na nájdenie optimálneho počtu.

Vo všeobecnosti sa zhluková analýza líši od iných metód klasifikácie objektov tým, že typy objektov (zhluky) nie sú a priori známe. Zhluky sú v priebehu samotnej analýzy ovplyvňované počtom, charakteristikou dát, a najmä počiatočnou voľbou centra- ťažiska.

Vzhľadom na to, že v reálnych aplikáciách vzniká veľké množstvo neistôt, resp. nejednoznačností, je vhodné použiť viacero alternatívnych metód ako i zohľadniť možnosť, že hranica medzi zhlukmi nie je ostrá. Keďže zhlukovanie je empirická metóda, kde uplatnenie rôznych postupov vedie k rôznym typológiám, k rôznej interpretácii dát, je užitočné použiť k metóde K-means inú vhodnú alternatívu.

## 3. Fuzzy zhlukovanie

Na vytvorenie skupín dát s podobnou charakteristikou bol použitý software PERSIMPLEX, založený na algoritme fuzzy zhlukovej analýzy, ktorá využíva ako mieru podobnosti medzi zhlukovanými objektami podobnosť tvaru kriviek, resp. podobnosť blízkosti kriviek v grafickom vyjadrení.

Fuzzy zhlukovanie, na rozdiel od ostatných zhlukovacích metód, umožňuje čiastočné zaradenie objektu do viacerých zhlukov a to pomocou pravdepodobnosti. Cieľom je zabrániť skresleniu zhlukovania kvôli prítomnosti nezaraditeľných objektov. Takýto objekt sa nepriradí ku žiadnemu zhluku (od každého sa príliš odlišuje), ale priradia sa mu pravdepodobnosti s ktorými sa bude nachádzať v jednotlivých zhlukoch.

Vo všeobecnosti, vzhľadom na vnútorné a vonkajšie vlastnosti priestoru , zhluky nemusia mať ostré hranice. Inak povedané, jeden objekt môže mať rôzny stupeň príslušnosti ku všetkým existujúcim zhlukom a jeho hodnota charakteristickej funkcie môže nadobúdať ľubovoľné hodnoty od 0 po 1, nielen 0 alebo1 ako je to pri bežnom zhlukovaní.

Existuje viacero metód na zistenie podobnosti vzťahov v množinách a jeden z dôležitých prístupov k fuzzy zhlukovaniu je založený na minimalizovaní objektívnej funkcie.

Nech $d_{ik}$ predstavuje vzdialenosť medzi objektom $x_k$ a centrom zhluku $a_i$ vyjadrenú vzťahom

$$\text{napr.} \qquad d_{ik} = d(x_k, a_i) = \| x_k - a_i \|^2 \qquad (2.1)$$

Potom fuzzy objektívna funkcia je definovaná

$$J_m = \sum\sum u_{ij} \| x_k - a_i \|^2 \ , \quad 1 < m < \infty \qquad (2.2)$$

kde $m$ je ľubovoľné reálne číslo väčšie ako1, $u_{ij}$ je stupeň príslušnosti $x_i$ v zhluku $j$, $x_i$ je $i$-tý člen z n-rozmerných merateľných dát, $a_i$ je n -rozmerné centrum zhluku a (2.1) je ľubovoľná miera vyjadrujúca podobnosť medzi ľubovoľnými merateľnými dátami a centrami.

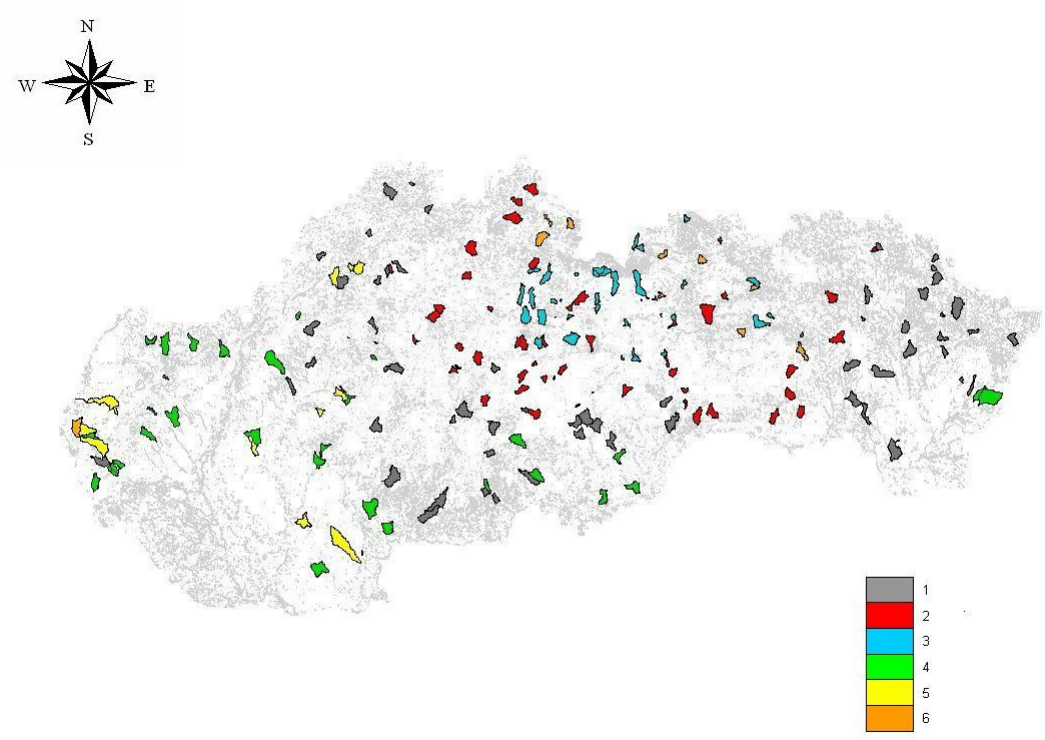
Dobrým spôsobom na určenie počtu zhlukov je použitie Dunnovho, alebo Kaufmannovho rozdeľovacieho koeficientu. (Meloun, Militký, Hill)

## 4. Porovnanie výsledkov analýzy

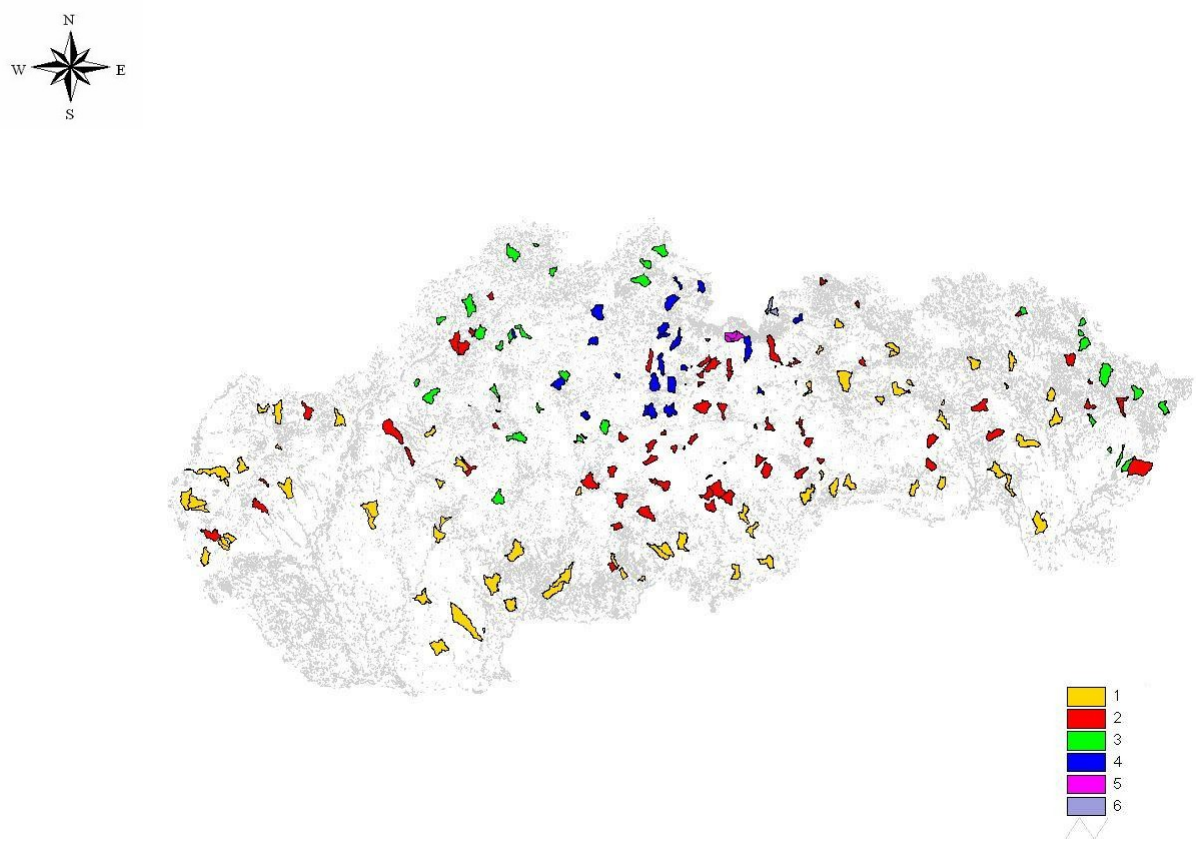
Vstupné údaje, ktoré sme analyzovali, pochádzajú z 209 vodomerných staníc. Pretože pri zhlukovaní dochádza k závislosti zhlukov od zmeny mierky, údaje sme štandardizovali a pre porovnanie sme na posúdenie vybrali šesť zhlukov vytvorených metódou K-means a fuzzy metódou, ktoré sa líšia počtom a rozložením objektov.

Pre názornosť je vhodné najskôr obidve riešenia vizualizovať priestorovým zobrazením (Obr3.1), kde je lepšie viditeľný rozdiel v zaradení objektov do zhlukov.

Južná časť Slovenska je fuzzy prístupom prekrytá viacerými zhlukmi, v porovnaní s K-means, kde prevažuje zhluk č.1. Podrobnejšia analýza a rozhodnutie o tom, ktoré z riešení je vhodné k ďalšej aplikácii zostáva na špecialistoch z odboru hydrológie.




***Obrázok .3.1a: Znázornenie zhlukov metódou fuzzy***




***Obrázok 3.1b: Znázornenie zhlukov metódou K-means***

Nasledujúce tabuľky udávajú mieru totožnosti objektov v jednotlivých zhlukoch. Pre prehľadnejšie porovnanie je tabuľke 4.1b vyjadrená percentuálna zhoda povodí, ktoré sa súčasne vyskytujú v zhlukoch vytvorených metódami K-means a fuzzy. Z porovnania je zrejmé, že v niektorých prípadoch je zhoda minimálna, zastúpenia objektov v zhlukoch sú výrazne odlišné. Môžeme skonštatovať, že regióny vytvorené uvedenými metódami predstavujú rôzne riešenia. Viditeľné je to v prípade fuzzy zhluku č.3, z ktorého časť objektov tvorí dva samostatné K-means zhlukyč.5a č.6.

***Tabuľka 3.1a: Porovnanie zastúpenia objektov v zhlukoch***

| | | Kmeans zhluky | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Fuzzy zhluky | | 1 | 2 | 3 | 4 | 5 | 6 | ∑ |
| | 1 | 18 | 21 | 24 | 0 | 0 | 0 | 63 |
| | 2 | 9 | 25 | 7 | 11 | 0 | 0 | 52 |
| | 3 | 6 | 12 | 0 | 12 | 4 | 2 | 36 |
| | 4 | 21 | 8 | 3 | 0 | 0 | 0 | 32 |
| | 5 | 13 | 2 | 1 | 0 | 0 | 0 | 16 |
| | 6 | 5 | 1 | 0 | 4 | 0 | 0 | 10 |
| | ∑ | 72 | 69 | 35 | 27 | 4 | 2 | 209 |

*Tabuľka 3.1b: Porovnanie zastúpenia objektov v zhlukoch vyjadrené v percentách*

| | | Kmeans zhluky | | | | | |
|---|---|---|---|---|---|---|---|
| Fuzzy zhluky | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | 25.00% | 30.43% | 68.57% | 0.00% | 0.00% | 0.00% |
| | 2 | 12.50% | 36.23% | 20.00% | 40.74% | 0.00% | 0.00% |
| | 3 | 8.33% | 17.39% | 0.00% | 44.44% | 100.00% | 100.00% |
| | 4 | 29.17% | 11.59% | 8.57% | 0.00% | 0.00% | 0.00% |
| | 5 | 18.06% | 2.90% | 2.86% | 0.00% | 0.00% | 0.00% |
| | 6 | 6.94% | 1.45% | 0.00% | 14.81% | 0.00% | 0.00% |
| | | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |

## 5. Záver

Zhluková analýza je vhodný nástroj na riešenie aplikačných problémov navzájom podobného typu. Vytvorenie zhlukov predstavuje oblasti s podobným odtokovým režimom a môže byť smerodajným prvkom k možnej predikcii pre územia bez meracích staníc. Z porovnania výsledkov metód použitých v tomto článku je zrejmé, že ide o dve odlišné riešenia, pričom nie je možné jednoznačne potvrdiť, alebo poprieť konkrétnu metódu a na komplexné vyhodnotenie sú nutné ďalšie matematické a štatistické analýzy. V tomto prípade dochádza k prekrytiu objektov s najväčšou mierou podobnosti v oblasti juhozápadného Slovenska- fuzzy zhluk č.5 a K-means .

Príspevok bol realizovaný s podporou grantu VEGA 1/4024/07,VEGA 1/0373/08.

## 6. Literatúra

.

[1] MELOUN,M. – MILITKÝ,J.-HILL,M 2005. Počítačová analýza vícerozměrných dat v příkladech, 2005. 150s.ISBN 80-200-1335-0

[2] HAN,J.- KAMBER,M. 2000. Data Mining,2000. 550s.ISBN-10 1558604898.

[3] DOROODCHI,M. - REZA, ALI M. Nonlinear Smoothing of Signals by Applying Fuzzy Clustering to Local Points. Symposium on Applied Computing: Philadelphia, Pennsylvania, United States,1996, s.595 - 599, ISBN:0-89791-820-7

[4] http://rimarcik.com/

[5] Kohnova, S.,Szolgay, J.Solin,L.(2006):Regional methods for prediction in ungauged basins, Key publishing, Brno.

## Adresa autorov

Pastuchová Elena, RNDr., PhD
Katedra matematiky
FEI STU
Ilkovičova 3
812 19 Bratislava
elena.pastuchova@stuba.sk

Václavíková Štefánia, Mgr.
Katedra matematiky a deskriptívnej geometrie,SvF STU
Radlinského 11
813 68 Bratislava
vaclavikova@is.stuba.sk

# Volba vyhlazovacího parametru Hodrick-Prescottova filtru
# Choice of smoothing parameter of Hodrick-Prescott filter

Jitka Poměnková

**Abstract:**

Presented paper is focused on impact of smoothing parameter $\lambda$ on estimate of the business cycle trend using Hodrick-Prescott filter. For this purpose trend estimate with generally recommended value $\lambda=1600$ for quarterly data is compared with trend estimate with optimized value of smoothing parameter. Optimization is done using rule derived on the basis of variance proportion. Available data are quarterly values of Gross Domestic Product in 1996/Q1 – 2008/Q4 in the Czech Republic.

**Key words:** smoothing parameter, business cycle, Hodrick-Prescott filter

**Klíčová slova:** vyhlazovací parametr, hospodářský cyklus, Hodrick-Prescottův filter

## 1. Úvod

Ekonomická teorie pracuje se dvěma definicemi hospodářského cyklu. Klasickou definicí podle Burnse a Mitchella (1947), kteří definují hospodářský cyklus jako opakující se fluktuace kolem ekonomické aktivity vykazující stejné tendence v odlišných sektorech ekonomiky. A růstovou definicí podle Lucase (1977) jako opakující se fluktuace časové řady makroekonomické proměnné okolo svého trendu. Růstový cyklus je tak založen na dekompozici časové řady zvoleného makroekonomického ukazatele na trendovou, cyklickou popřípadě sezónní a nepravidelnou složku. V souvislosti s analýzou klasického hospodářského cyklu se vyskytuje problém nerozlišování trendové a cyklické složky časové řady. Může se tak stát, že trendová složka, při dlouhodobé růstové tendenci ekonomiky, ovlivní cyklickou složku. Tento problém je částečně eliminován v případě růstového cyklu, který je na dlouhodobý trend méně citlivý. Pak základním bodem analýzy je výběr vhodné detrendovací metody, která se tak stává významným determinantem provedené analýzy cyklické komponenty a optimalizace jejich parametrů.

Podle Canovy (1998) můžeme provést dělení soudobých filtrační techniky na statistické a ekonomické. Mezi statistické techniky lze zařadit techniku prvních diferencí, deterministické modely (regresní přímka, polynom apd.), proceduru Beveridge a Nelsona, model nepozorované komponenty nebo eliminaci trendu předpokládající přímou nepozorovatelnost trendové nebo cyklické složky, k jejíž identifikaci však používají rozdílné statistické předpoklady. Ekonomický přístup říká, že trend je diktován ekonomickým modelem, preferencemi výzkumníka nebo formulovaným a řešeným problémem a chápe trendovou a cyklickou složku za neoddělitelné, podléhající stejným vlivům a vývoji v čase. Při analýze pracuje růstový cyklus s očištěnými detrendovanými hodnotami proměnných. Smyslem je pak identifikace fluktuace cyklické složky časové řady proměnné kolem dlouhodobého trendu. Mezi tyto techniky řadíme modely obvyklých deterministických a stochastických trendů vycházejících z ekonomické teorie jako je např. Hodrick-Prescottův (HP) filtr, Baxterův Kingův (BK) filtr.

Cílem předkládaného příspěvku je analyzovat vliv hodnoty vyhlazovacího parametru $\lambda$ Hodrick-Prescottova filtru na výsledný odhad trendu vývoje získaného pomocí zmíněného filtru, a to jak z pohledu využití v literatuře obecně doporučované hodnoty, tak z hlediska

odvození vlastní optimální hodnoty. Pro empirickou analýzu byly zvoleny hodnoty HDP v ČR v letech 1996/Q1 - 2008Q4 jako ukazatele hospodářského vývoje země.

## 2. Metodika

Velmi často používanou a populární metodou při analýze hospodářského cyklu je pro detrendování Hodrick-Prescottův (HP) filtr. Základní myšlenka spočívá v rozkladu nestacionární časové řady $Y$, která je k dispozici v období t=1,..,T na trendovou a cyklickou složku. HP filtr je přitom navržen tak, aby při extrakci trendu bral v úvahu dvě kritéria, a to velikost reziduí a míru vyhlazení trendu. Upravme zápis regresního modelu pro potřeby HP filtru následovně

$$Y_t = g_t + c_t,\ t=1,...,n,$$

kde $g_t$ označuje růstovou komponentu a $c_t$ cyklickou komponentu. Omezení růstové komponenty plyne z řešení následujícího problému, a to

$$\min_{\{g_t\}_{t=1}^{T}} \sum_{t=1}^{T}(Y_t - g_t)^2 + \lambda \sum_{t=1}^{T}\left[(g_{t+1} - g_t) - (g_t - g_{t-1})\right]^2 ,$$

kde cyklická komponenta $c_t = Y_t - g_t$ představuje odchylky od dlouhodobého trendu a její hladkost je měřena prostřednictvím kvadrátu druhých diferencí. Parametr $\lambda$ je kladné číslo, které penalizuje variabilitu růstové složky. Je-li $\lambda$=0, pak řešení výše uvedené minimalizační úlohy vede k tomu, že trendová složka je shodná s původní řadou. Pokud naopak $\lambda \to \infty$, je kladena větší váha na druhý člen a $g_t$ se přibližuje k lineárnímu trendu.

Jak ukázal Harvey a Jager (1993) v nekonečné verzi lze HP filtr interpretovat jako optimální lineární filtr trendové komponenty. King a Rebelo (1993) uvádějí, že komponenta hospodářského cyklu a druhá diference růstové komponenty musí mít stejnou reprezentaci klouzavých průměrů pro HP filtr, aby byl lineárním filtrem, který minimalizuje střední kvadratickou chybu. Tedy, aby tento filtr minimalizoval chybu

$$MSE = (1/T)\sum_{t=1}^{T}(ets(c_t) - c_t)^2 , \qquad (4)$$

kde $c_t$ je skutečná cyklická složka a $est(c_t)$ je její odhad. Hodrick and Prescot zjistili, že jestliže jsou cyklická komponenta ($c_t$) a druhá diference růstové komponenty ($\Delta^2 g_t$) identicky nezávisle normálně rozložené typu $c_t \sim N(0,\sigma_c^2)$, $\Delta^2 g_t \sim N(0,\sigma_g^2)$, pak nejlepší výběr vyhlazovacího parametr $\lambda$, ve smyslu minimalizace MSE, je $\lambda = \sigma_c^2/\sigma_g^2$. V mnoha studiích je doporučovanou hodnotou $\lambda$=1600 pro čtvrtletní periodicitu dat, například Ahumada, Garegnani (1990), Guy (1997), Hodrick-Prescott (1980) a další. Na základě výše uvedeného formulujme vlastní pomocné kritérium pro nalezení optimální hodnoty vyhlazovacího parametru $\lambda$.

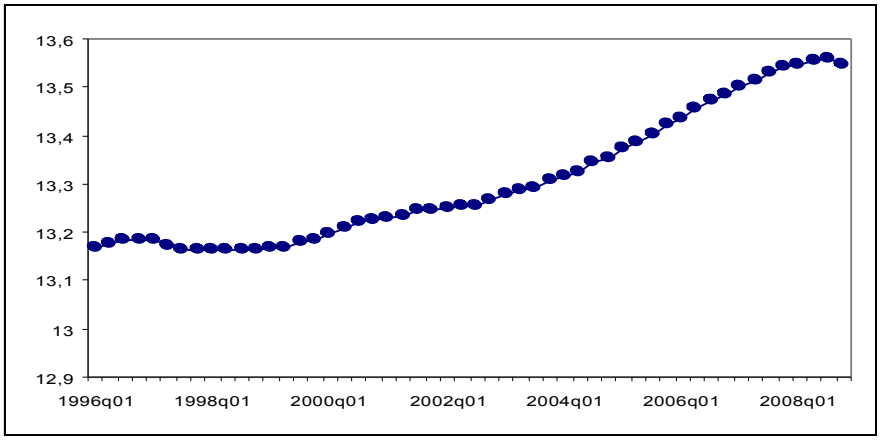
Mějme vstupní časovou řadu pozitivních hodnot $Y_t$, $t=1,...,n$, množinu hodnot vyhlazovacího parametru $\lambda$ označenou jako $L = [100/n; 100 \cdot n]$ a indexovou množinu $I = [1,...,k]$, kde $k$ je počet prvků množiny $L$. Pak pro každé $L_i \in L$, $i \in I$ můžeme vypočítat Hodrick – Prescottův odhad růstové a cyklické složky časové řady $Y_t$. Nechť cyklická složka $c_t$ a druhá diference růstové složky $g_t$ ($\Delta^2 g_t$) jsou identicky nezávisle normálně rozložené proměnné, $c_t \sim N(0,\sigma_c^2)$, $\Delta^2 g_t \sim N(0,\sigma_g^2)$. Pak optimální hodnota vyhlazovacího parametru $\lambda$ je ta, pro kterou platí

$$\lambda_{opt} = \sigma_c^2/\sigma_g^2 \Leftrightarrow L_i \approx \sigma_c^2/\sigma_g^2, \quad pro\ \forall\ i \in I$$

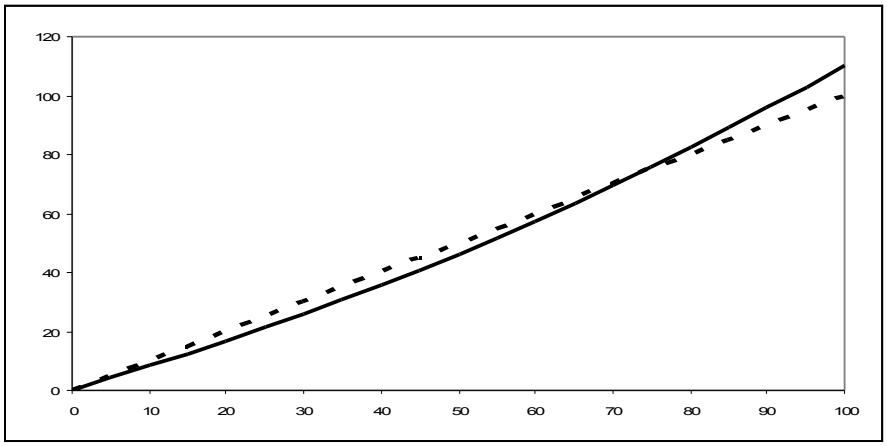
kde rozdíl $d = \left|L_i - \sigma_c^2/\sigma_g^2\right| \le 100/n$.

## 3. Empirická část

Pro empirickou analýzu byly zvoleny sezónně očištěné absolutní hodnoty HDP ve čtvrtletní frekvenci v období 1996/Q1 - 2008Q4, které budou před analýzou transformovány přirozeným logaritmem, neboť růstové charakteristiky lépe vystihují povahy cyklů. Vstupní hodnoty jsou zobrazeny na obr. 1. Trend vývoje hodnot HDP a průmyslu bude nejprve popsán HP filtrem s hodnotou vyhlazovacího parametru $\lambda$=1600. Poté bude proveden odhad optimální hodnoty vyhlazovacího parametru na základě pravidla odvozeného v předchozí části příspěvku a opět odhadnut trend vývoje zvoleného ukazatele pomocí HP filtru s optimální hodnotou $\lambda_{opt}$. Výsledné odhady trendu vývoje budou v závěru graficky zobrazeny.

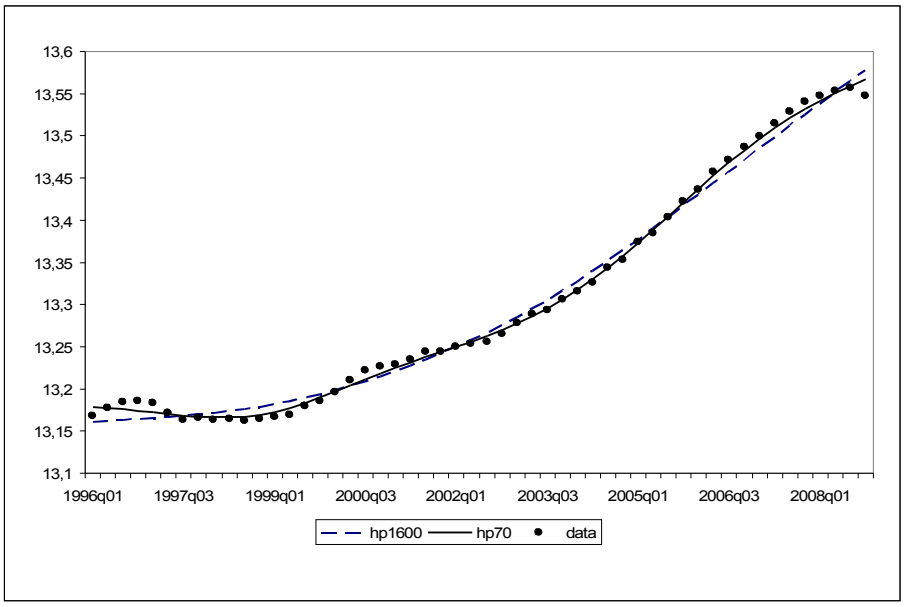



***Obrázek 1: HDP v ČR***




***Obrázek 2: Odhad $\lambda_{opt}$ pro HDP***

V případě HDP byla odhadnuta optimální hodnota vyhlazovacího parametru $\lambda_{opt}=70$ (obr. 2). Pro zjištěnou optimální hodnotu parametru byl následně proveden odhad trendu vývoje HDP pomocí HP filtru. Výsledné odhady trendu vývoje HDP pro hodnoty $\lambda=1600$ a $\lambda_{opt}=70$ jsou zachyceny na obr. 3.




***Obrázek 3: Odhad trendu vývoje HDP s hodnotami $\lambda$=1600* (hp1600) a $\lambda_{opt}$=70 (hp70)***

V případě hodnot HDP lze říci (obr. 3), že výsledný odhad s optimalizovanou hodnotou $\lambda_{opt}$ se oproti odhadu s hodnotou převzatou z literatury ($\lambda$=1600) liší. Jak je z obr. 3 patrné, v období 1996/Q1 – 2000/Q1 odhad s optimalizovanou hodnotou zachytil fázi mírného poklesu následovanou fází mírného růstu hodnot HDP. Rovněž v období 2002/Q1 – 2008/Q1 odhad s optimalizovanou hodnotou reflektoval na mírný konvexní růst následovaný po roce 2005/Q2 mírným konkávním růstem. Odhad s převzatou hodnotou toto chování nepostihl. Přestože jsou rozdíly v obou odhadech na první pohled malé, můžeme říci, že dynamika datového souboru je lépe zachycena odhadem s optimalizovanou hodnotou, zatímco hodnota převzatá z literatury může mít tendence k přehlazení výsledného odhadu. Získaný růstový cyklus detrendováním bez optimalizace vyhlazovacího parametru by tak mohl vést ke zdánlivé cykličnost. Mechanické převzetí hodnoty vyhlazovacího parametru se tak jeví jako nedostačující.

## 4. Závěr

Předkládaný příspěvek se zabýval vlivem vyhlazovacího parametru $\lambda$ Hodrick-Prescottova filtru na výsledný odhad trendu vývoje získaného pomocí zmíněného filtru, a to jak z pohledu využití v literatuře obecně doporučované hodnoty, tak z hlediska odvození vlastní optimální hodnoty. Pro empirickou analýzu byly zvoleny hodnoty HDP a průmyslové výroby v ČR v letech 1996/Q1 - 2008Q4 jako ukazatele hospodářského vývoje země.

Pro porovnání vlivu parametru byl proveden odhad trendu vývoje HDP s doporučovanou hodnotou $\lambda$=1600. Následně byly pomocí pravidla odvozeného na základě podílů rozptylů jednotlivých složek HP filtru odhadnuty optimální hodnoty parametru pro HDP $\lambda_{opt}$=70. Zjištěná optimální hodnota se významným způsobem lišila od doporučované hodnoty $\lambda$=1600. V případě HDP empirická analýza dospívá k závěru, že trend vývoje popsaný pomocí HP filtru s optimalizovanou hodnotou vystihuje přesněji charakter dat.

Na základě provedené empirické analýzy lze konstatovat, že mechanické převzetí vyhlazovacího parametru $\lambda$=1600 Hodrick-Prescottova filtru pro čtvrtletní data může způsobit přehlazení výsledného odhadu, čímž může dojít k potlačení informací obsažených v datovém souboru. V porovnání s tímto navrhované kritérium optimalizace umožňuje najít

hodnotu vyhlazovacího parametru, pro který výsledný odhad trendu zpravidla lépe vystihuje charakter datového souboru a umožňuje získání přesnějších hodnot pro další analýzu.

Předkládaný příspěvek vznikl za podpory výzkumného záměru „Česká ekonomika v procesech integrace a globalizace a vývoj agrárního sektoru a sektoru služeb v nových podmínkách evropského integrovaného trhu“.

## 5. Literatura

[1]AHUMADA, H., GAREGNANI, M. L. 1999. Hodrick-Prescott Filter in practice, UNLP

[2]BURNS, A.F., MITCHELL, W.C., 1946. Measuing Business Cycles. New York, National Bureau of Economic Research, pp.590, ISBN: 0-870-14085-3;

[3]CANOVA, F. 1998. De-trending and business cycle facts, *Journal of monetary Economic*, vol. 41, pp. 533-540

[4]GUAY,A. ST-AMANT, P. Do the Hodrick-Prescott and Baxter-king Filters Provide a Good Approximation of Business Cycles? Université a Québec á Montreál, Working paper No. 53, 1997

[5]HARVEY, A.C., JAEGER, A. (1993): De-trending, Stylized FActs and the Business Cycle. *Journal of Applied Econometrics 8*: pp. 231-47

[6]HODRICK, R.J., PRESCOTT, E.C. Post-war U.S.1980. Business Cycles: An Empirical Investigation, mimeo, Carnegie-Mellou University, Pitsburgh, PA. , 24pp.

[7]KING, R. G., REBELO, S. T. 1993. Low frequency Filtering and Real Business Cycles. *Journal of Economic Dynamics and Control*. vol. 17, p.207-231

[8]LUCAS, R.E. 1977. Understanding Business Cycles. In BRUNNER, K., MELTZER, A.H. (eds.): *Stabilization Domestic and International Economy*. Carnegie-Rochester Conference Series on Public Policy, vol. 5, pp. 7-29.

**Adresa autora:**

Jitka Poměnková, Ph.D., RNDr.

Ústav financí PEF MZLU v Brně

Zemědělská l

Česká republika

613 00 Brno

pomenka@mendelu.cz

# Princip analýzy rozptylu a jeho další využití
# Principle of analysis of variance and its further utilization

Řezanková Hana

**Abstract:** The paper focuses on the utilization of R-square measure and F statistic known from analysis of variance. The principle of R-square measure is applied in case of asymmetric dependency not only for quantitative continuous dependent variable but also for other types of dependent variables (quantitative discrete, nominal, ordinal). However, the suitable variability measures should be used. For ordinal variable, the principle of analysis of variance is applied when mean ranks of values in groups created according to categories of explanatory variable are investigated. Further, modified analysis of variance is used in regression analysis. Moreover, for evaluation of quality of clusters as results of cluster analysis methods, R-square measure, pseudo F statistic and some other measures based on sums of squares (between clusters and within clusters) are applied. It serves for cluster number determination.

**Key words:** R-square measure, F statistic, asymmetric dependency measures, variability of nominal variable, Kruskal-Wallis test, regression analysis, cluster number determination.
**Klíčová slova:** R-kvadrát míra, F statistika, míry asymetrické závislosti, variabilita nominální proměnné, Kruskalův-Wallisův test, regresní analýza, určování počtu shluků.

## 1. Úvod

V souvislosti s redukcí výuky kvantitativních metod na některých vysokých školách v ČR jsou stále častější diskuze na téma, které statistické metody jsou základní, o kterých by studenti měli slyšet (a v ideálním případě se příslušnou látku také naučit), a které naopak jsou zbytečných „přepychem“ a jen studenty „zatěžují“. Taková je i situace na Vysoké škole ekonomické v Praze, kdy byla výuka statistiky v tzv. celoškolním základu zredukována na pouhý jeden semestr z původních dvou semestrů. Úspěšní studenti se vyjadřují, že látka není těžká, ale že je jí na jeden semestr mnoho. Proto každoročně probíhají na katedře statistiky a pravděpodobnosti diskuze, která část z vyučované látky by se mohla vynechat, aniž by to bylo na úkor navazujících problematik. V letošním roce byla „na pořadu“ analýza rozptylu, což bylo inspirací pro tento příspěvek. V dalším textu bude pozornost věnována úlohám s jedním faktorem (v celoškolně povinném předmětu je zařazena pouze jednorozměrná úloha).

## 2. Jednorozměrná analýza rozptylu s jedním faktorem

Analýza rozptylu je v programových systémech známa jako metoda ANOVA (*Analysis of Variance*). Uvažujme skupiny určitého ukazatele, například tržby v prodejnách obchodního řetězce (za stanovené období) roztříděné podle krajů. Zajímá nás, zda jsou průměrné tržby v jednotlivých krajích přibližně stejné, nebo zda se liší. Předpokládáme, že údaje nejsou v centrální databázi a prodejen je mnoho, proto provedeme jejich náhodný výběr o rozsahu $n$. K dispozici jsou tedy hodnoty proměnné *tržby* (dále $Y$) roztříděné podle krajů.

Chceme provést test o shodě středních hodnot. Pro jednoduchost předpokládejme, že jsou splněny předpoklady pro použití analýzy rozptylu, tj. hodnoty ve skupinách jsou výběry z normálního rozdělení se stejným rozptylem (jedním z argumentů pro vynechání analýzy rozptylu z výuky byl právě řídký výskyt takové ideální situace v reálných ekonomických úlohách). Nulovou hypotézu zapíšeme ve tvaru $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$, kde $\mu_i$ je očekávaná (střední) hodnota tržby v $i$-tém kraji a $k$ je počet krajů. Alternativní hypotéza je její negací, tj. $H_1$: non $H_0$.

K testu se používá statistika $F$, která má za předpokladu platnosti nulové hypotézy F rozdělení s počty stupňů volnosti $(k - 1)$ a $(n - k)$. Zapišme si tuto statistiku ve tvaru

$$F = \frac{\frac{S(Y)_M}{k-1}}{\frac{S(Y)_V}{n-k}} = \frac{\frac{\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_i}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n-k}}, \quad (1)$$

kde $S(Y)_M$ je meziskupinový součet čtverců a $S(Y)_V$ je vnitroskupinový součet čtverců, $y_{ij}$ je tržba zjištěná u $j$-tého objektu v $i$-tém kraji, $\bar{y}_i$ je průměrná tržba prodejen v $i$-tém kraji a $\bar{y}$ je průměrná tržba všech šetřených prodejen.

## 3. Zkoumání asymetrické závislosti

Na základě zjištěných tržeb si můžeme položit otázku, zda tržby závisí na kraji. Proměnná *kraj* (dále $X$) je v tomto případě vysvětlující (neboli *faktor*) a proměnná *tržby* ($Y$) vysvětlovaná. Jde o závislost asymetrickou (jednosměrnou, jednostrannou), neboť opačný vztah nemá v daném kontextu smysl. Proměnná $X$ statisticky působí na $Y$, jestliže se změnami hodnot proměnné $X$ se mění statistické vlastnosti rozdělení hodnot proměnné $Y$ (viz [3]). To znamená, že zamítneme-li nulovou hypotézu o shodě středních hodnot, svědčí to o závislosti proměnné $Y$ na proměnné $X$.

Součástí zkoumání závislostí bývá ohodnocení její intenzity. Při analýze rozptylu se vychází z rozložení celkového součtu čtverců $S(Y)$, tj. součtu druhých mocnin odchylek jednotlivých hodnot od celkového průměru, na součet meziskupinového a vnitroskupinového součtu čtverců, tj. $S(Y) = S(Y)_M + S(Y)_V$. Intenzita závislosti se posuzuje na základě *poměru determinace*, který vyjadřuje podíl meziskupinové variability na celkové variabilitě, tj.

$$R^2 = \frac{S(Y)_M}{S(Y)} = \frac{S(Y) - S(Y)_V}{S(Y)}. \quad (2)$$

Buď se používá přímo tato míra R-kvadrát, nebo její odmocnina, viz též [1] a [7].

Proti návrhu nezařazovat analýzu rozptylu do výuky základního kurzu z důvodu obtížné splnitelnosti předpokladů pro F test lze uvést, že v praxi se míry závislosti kvantitativní proměnné na proměnné kategoriální používají bez ohledu na to, zda lze provést samotný test. Příkladem je zařazení druhé odmocniny z poměru determinace pod názvem *koeficient* $\eta$ jako jedné z měr závislosti počítaných na základě četností v kontingenční tabulce v některých programových systémech. Vyjdeme-li z vyjádření variability pomocí rozptylu $s^2(Y)$, pak s použitím sdružených četností $n_{ij}$ a marginálních četností $n_{i+}$ a $n_{+j}$ můžeme vzorec zapsat jako

$$\eta_{Y|X} = \sqrt{\frac{s^2(Y) - \sum_{i=1}^{r} \frac{n_{i+}}{n} s^2(Y|x_i)}{s^2(Y)}} = \sqrt{\frac{\sum_{i=1}^{r} \frac{1}{n_{i+}} \left( \sum_{j=1}^{s} n_{ij} y_j \right)^2 - \frac{1}{n} \left( \sum_{j=1}^{s} n_{+j} y_j \right)^2}{\sum_{j=1}^{s} n_{+j} y_j^2 - \frac{1}{n} \left( \sum_{j=1}^{s} n_{+j} y_j \right)^2}}, \quad (3)$$

kde $r$ je počet kategorií řádkové proměnné (tj. vysvětlující $X$) a $s$ je počet kategorií sloupcové proměnné (tj. vysvětlované $Y$); odvození vzorce viz [4].

Výše uvedený princip měření intenzity asymetrické závislosti se využívá i v případě, pokud je vysvětlovaná proměnná $Y$ nominální (její kategorie nelze uspořádat podle velikosti). V tomto případě se ovšem její variabilita (mutabilita) vyjadřuje pomocí speciálních měr. Obecně lze vzorec (2) zapsat jako

$$S_{Y|X} = \frac{var(Y,X)}{var(Y)} = \frac{var(Y) - var(Y|X)}{var(Y)}. \quad (4)$$

Uvažujme tři míry variability podle [3]. První z nich je *variační poměr* ve tvaru

$$v(Y) = 1 - n_{\mathrm{Mo}}/n, \quad (5)$$

kde $n_{\text{Mo}}$ je četnost modální kategorie. Dosazením do vzorce (4) s využitím symboliky pro kontingenční tabulku dostáváme *koeficient* $\lambda$ ve tvaru

$$\lambda_{Y|X} = \frac{v(Y) - \sum_{i=1}^{r} \frac{n_{i+}}{n} v(Y|x_i)}{v(Y)} = \frac{\sum_{i=1}^{r} n_{i\text{Mo}} - n_{+\text{Mo}}}{n - n_{+\text{Mo}}}, \tag{6}$$

kde $n_{i\text{Mo}}$ je četnost modální kategorie v *i*-tém řádku (odpovídající *i*-té kategorii proměnné *X*), obdobně $n_{+\text{Mo}}$ je četnost modální kategorie zjištěné na základě všech hodnot znaku *Y*.

Druhou mírou variability je *Giniho míra mutability*, nazvaná v [3] jako *nominální variance* (*nomvar*). Počítá se podle vzorce

$$nomvar(Y) = 1 - \sum_{i=1}^{k} \left( \frac{n_i}{n} \right)^2 = \frac{n^2 - \sum_{i=1}^{k} n_i^2}{n^2}, \tag{7}$$

kde *k* je počet kategorií sledované proměnné. Dosazením do vzorce (4) s využitím symboliky pro kontingenční tabulku dostáváme *koeficient* $\tau$ ve tvaru

$$\tau_{Y|X} = \frac{nomvar(Y) - \sum_{i=1}^{r} \frac{n_{i+}}{n} nomvar(Y|x_i)}{nomvar(Y)} = \frac{n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{n_{i+}}}{n^2 - \sum_{j=1}^{s} n_{+j}^2}, \tag{8}$$

kde $m_{ij}$ je očekávaná četnost v případě nezávislosti, počítaná jako $(n_{i+} \cdot n_{+j})/n$.

Třetí mírou je *entropie* vyjadřovaná ve tvaru

$$H(Y) = -\sum_{i=1}^{k} \frac{n_i}{n} \ln \frac{n_i}{n}. \tag{9}$$

Dosazením do vzorce (4) s využitím symboliky pro kontingenční tabulku dostáváme *koeficient nejistoty* (*neurčitosti*), neboli *informační koeficient* (viz [3]) ve tvaru

$$U_{Y|X} = \frac{H(Y) - \sum_{i=1}^{r} \frac{n_{i+}}{n} H(Y|x_i)}{H(Y)} = \frac{-\sum_{i=1}^{r} \frac{n_{i+}}{n} \ln \frac{n_{i+}}{n} - \sum_{j=1}^{s} \frac{n_{+j}}{n} \ln \frac{n_{+j}}{n} + \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}}{n} \ln \frac{n_{ij}}{n}}{-\sum_{j=1}^{s} \frac{n_{+j}}{n} \ln \frac{n_{+j}}{n}}. \tag{10}$$

Uvedený princip vyjadřování intenzity závislosti se používá také pro vzájemnou závislost. V případě koeficientu lambda a koeficientu nejistoty lze na základě dvou asymetrických měr odvodit symetrickou míru. Odvození výše uvedených vzorců, včetně symetrických měr, je uvedeno v [2] a [4]. Symetrické míry se používají jako míry podobnosti v metodách vícerozměrné analýzy, jako je shluková analýza nebo vícerozměrné škálování.

Porovnání některých statistických programových systémů z hlediska implementace výše uvedených měr je uvedeno v [5]. Výpočet koeficientu $\lambda$ je implementován v systémech SAS, SPSS a STATGRAPHICS, koeficientu $\tau$ v systému SPSS, koeficientu nejistoty v systémech SAS, SPSS, STATGRAPHICS a STATISTICA a výpočet koeficientu $\eta$ v systémech SPSS a STATGRAPHICS.

Je-li vysvětlovaná proměnná ordinální, pak sice nepoužijeme F test o shodě středních hodnot, ale můžeme použít některý z *neparametrických testů*, například *Kruskalův-Wallisův*. Stejně jako se při analýze rozptylu roztřídí hodnoty vysvětlované proměnné do skupin podle hodnot proměnné vysvětlující, tak se při použití Kruskalova-Wallisova testu roztřídí pořadí hodnot vysvětlované proměnné a zkoumají se průměrná pořadí ve skupinách. Tento test lze využít i pro kvantitativní vysvětlovanou proměnnou, pokud nejsou splněny předpoklady pro použití F testu v analýze rozptylu.

Intenzita závislosti v případě ordinální vysvětlované proměnné může být vyjádřena opět na základě vztahu (4). Jako míru variability lze použít *diskrétní ordinální varianci* (*dorvar*), viz [3], která je založena na kumulativních relativních četnostech $P_i = \sum_{j=1}^{i} \frac{n_j}{n}$, tj.

$$dorvar = 2\sum_{i=1}^{k} (P_i(1 - P_i)) . \tag{11}$$

Výsledkem dosazením do vzorce (4) je koeficient β ve tvaru

$$\beta_{Y|X} = \frac{dorvar(Y) - \sum_{i=1}^{r} \frac{n_{i+}}{n} dorvar(Y|x_i)}{dorvar(Y)} = 1 - \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{i+}}{n} P_{j|i}(1 - P_{j|i})}{\sum_{j=1}^{s} P_{+j}(1 - P_{+j})} . \tag{12}$$

S dalším využitím principu analýzy rozptylu se setkáváme v *regresní analýze*, která rovněž zkoumá asymetrickou závislost. Analogicky s testováním vztahu a měřením intenzity závislosti uvedenými výše se v rámci regresní analýzy používá F test na současnou nulovost všech regresních koeficientů (v případě přímky test, že přímka je rovnoběžná s osou *X*, tedy shoda očekávaných hodnot) a míra R-kvadrát (koeficient determinace) pro hodnocení kvality modelu. Při vyjádření celkového součtu čtverců jako součtu teoretického a reziduálního součtu čtverců, tj. $S(Y) = S(Y)_T + S(Y)_R$, se konstrukce obou statistik provádí podle stejného schématu s tím, že při určování počtu stupňů volnosti se místo počtu kategorií vysvětlující proměnné zohledňuje počet parametrů modelu.

## 4. Aplikace na hodnocení kvality shluků

Kromě regresní analýzy se princip analýzy rozptylu aplikuje v rámci metod shlukové analýzy, a to jednak obecně pro hodnocení kvality shluků, jednak pro stanovení jejich optimálního počtu. Zařadíme-li objekty charakterizované vícerozměrnými vektory pozorování do určitého počtu shluků, pak lze vytvořit proměnnou obsahující identifikace těchto shluků. K dispozici je tak skupina vysvětlovaných proměnných a kategoriální vysvětlující proměnná, což odpovídá vícerozměrné úloze analýzy rozptylu s jedním faktorem.

Na kvalitu shluků lze usuzovat například pomocí *R-kvadrát indexu*, který se konstruuje analogicky jako poměr determinace v jednorozměrné analýze rozptylu. Jde o podíl meziskupinové variability (charakterizované součtem čtverců $S_M$) na celkové variabilitě (vyjadřované součtem čtverců $S_C$). Vzhledem k tomu, že platí vztah $S_C = S_M + S_V$, kde $S_V$ je součet čtverců, který charakterizuje vnitroskupinovou variabilitu, a tedy $S_M = S_C - S_V$, může být R-kvadrát index (RSQ) vyjádřen pomocí vnitroskupinové variability, tj.

$$I_{\mathrm{RSQ}} = \frac{S_C - S_V}{S_C} = \frac{\sum_{i=1}^{n}\sum_{l=1}^{m} (y_{il} - \bar{y}_l)^2 - \sum_{h=1}^{k}\sum_{j=1}^{n_h}\sum_{l=1}^{m} (y_{jl} - \bar{y}_{hl})^2}{\sum_{i=1}^{n}\sum_{l=1}^{m} (y_{il} - \bar{y}_l)^2} , \tag{13}$$

kde $y_{il}$ je hodnota *l*-té proměnné zjištěná u *i*-tého objektu (obdobně $y_{jl}$), $\bar{y}_l$ je průměrná hodnota *l*-té proměnné počítaná pro všechny objekty, $\bar{y}_{hl}$ je průměrná hodnota *l*-té proměnné pro objekty v *h*-tém shluku a $n_h$ je počet objektů v *h*-tém shluku.

R-kvadrát index se používá například při porovnání různých postupů při hierarchické shlukové analýze, jak je tomu v systému SAS, viz [8]. Pokud jde o problematiku počtu shluků, pak platí, že čím je počet shluků větší, tím jsou shluky více homogenní, a tedy hodnota poměru determinace je vyšší. Z toho důvodu se pro stanovení optimálního počtu shluků využívají jiná kritéria.

Jedním z nich je *SPRSQ* (*semipartial R-squared*) *index*, který se používá pro stanovení vhodného počtu shluků při hierarchické shlukové analýze. Tento index reprezentuje pokles hodnoty R-kvadrát indexu, způsobený spojením dvou shluků, tj.

$$I_{\mathrm{SPRSQ}}(k) = I_{\mathrm{RSQ}}(k+1) - I_{\mathrm{RSQ}}(k) . \tag{14}$$

Čím je hodnota SPRSQ indexu nižší, tím menší je změna meziskupinové variability (a tudíž i vnitroskupinové). Malá změna indikuje, že spojením dvou shluků se vnitroskupinová variabilita zvýšila málo, a že je tudíž menší počet shluků lepším výsledkem. Ze stanoveného intervalu počtu shluků je nejvhodnější takový počet, pro který hodnota SPRSQ indexu nabývá minima.

Dalším indexem inspirovaným analýzou rozptylu je *pseudo F* (*PSF*) *index*, založený na analogii s *F* statistikou. Podle jeho autorů je označován též jako *CHF* (*Calinského-Habaraszův F*) *index*. Počítá se podle vzorce

$$I_{\mathrm{CHF}} = \frac{\dfrac{S_M}{k-1}}{\dfrac{S_V}{n-k}} = \frac{(n-k)\cdot S_M}{(k-1)\cdot S_V} , \tag{15}$$

v němž použité symboly mají stejný význam jako u RSQ indexu. Vysoké hodnoty CHF indexu indikují dobře oddělené shluky; hledá se tedy maximum v rámci zadaného intervalu počtu shluků.

Jiným indexem, který se v některých programových systémech používá pro stanovení počtu shluků v rámci hierarchické shlukové analýzy, je *index PTS* (*pseudo T-square*), resp. *PST2*. Tato *pseudo T-kvadrát statistika* hodnotí spojení dvou shluků do jednoho na určité úrovni shlukování. Spojení *h*-tého a *h'*-tého shluku je hodnoceno pomocí vztahu

$$I_{\mathrm{PTS}} = \frac{S_{M,hh'}}{\dfrac{S_{V,h} + S_{V,h'}}{n_h + n_{h'} - 2}} , \tag{16}$$

kde $S_{M,hh'}$ je mezishlukový součet čtverců a $S_{V,h}$ a $S_{V,h'}$ jsou vnitroshlukové součty čtverců. Graficky lze zobrazit závislost hodnot tohoto indexu na počtu shluků. PTS index kvantifikuje odlišnosti mezi dvěma shluky, které jsou v daném kroku spojovány. Je-li pro $k$ shluků hodnota indexu současně vyšší než pro $(k-1)$ a $(k+1)$ shluků (na křivce je zřejmý „skok“), pak vhodný počet shluků je $(k+1)$.

Pouze na vnitroskupinové variabilitě je založen *RMSSTD* (*Root Mean Square STandard Deviation*) *index*, který měří homogenitu výsledných shluků. Vzorec je

$$I_{\mathrm{RMSSTD}} = \sqrt{\frac{S_V}{m(n-k)}} , \tag{17}$$

kde $S_V$ je součet čtverců, který charakterizuje vnitroskupinovou variabilitu (viz RSQ index) a $m$ je počet proměnných. Nižší hodnoty indexu tedy indikují lepší rozdělení objektů do shluků. Při grafickém zobrazení uživatel usuzuje na optimální počet shluků podle toho, kde se křivka „láme“. Vnitroskupinová variabilita se uplatňuje i v některých dalších indexech.

Pokud jde o implementaci výše uvedených indexů v programových systémech, pak se tyto indexy počítají v rámci metod hierarchické shlukové analýzy pro počty shluků od jednoho (resp. dvou – podle kontextu) do zadaného maxima. Zobrazují se jednak konkrétní hodnoty pro počty shluků od zadaného maxima do jednoho (SAS), jednak graf závislosti hodnot indexu na počtu shluků (SAS, SYSTAT). RSQ index se počítá v systému SAS, stejně jako SPRSQ index. Pseudo F index a pseudo T-kvadrát index se počítá v obou systémech (SAS i SYSTAT), RMSSTD index v systému SYSTAT. Pseudo F statistika se v systému SAS počítá pro zadaný počet shluků také v rámci metody *k*-průměrů, podrobněji viz [6]. Kromě toho se v systému SAS při metodě *k*-průměrů hodnotí závislost každé z proměnných na nově vzniklé proměnné s *k* kategoriemi pomocí poměru determinace a dalších charakteristik.

## 5. Závěr

Analýza rozptylu je jedna ze základních metod zkoumání asymetrické závislosti. Jednorozměrná analýza rozptylu s jedním faktorem je založena na vyjádření variability vysvětlované proměnné pomocí variability uvnitř skupin a mezi skupinami, které jsou vytvořeny na základě kategorií vysvětlující proměnné. Lze jednak testovat shodu očekávaných hodnot ve skupinách, jednak ohodnotit vztah proměnných pomocí poměru determinace. V praxi se vztah proměnných hodnotí i tehdy, nelze-li provést test z důvodu, že nejsou splněny předpoklady pro jeho použití. Analogicky lze míru závislosti konstruovat pro nominální i ordinální proměnnou s tím, že se používají speciální míry variability. O významu některých měr svědčí jejich implementace v komerčních programových systémech.

Princip vícerozměrné analýzy rozptylu s jedním faktorem se využívá ke konstrukci různých indexů, pomocí nichž se určuje kvalita rozdělení objektů (charakterizovaných vícerozměrnými vektory hodnot) do shluků (skupin) jako výsledku některé z metod shlukové analýzy. Toto vyjádření kvality rozdělení pak slouží jednak k porovnání různých metod, jednak ke stanovení optimálního počtu shluků. Některé z indexů navrhovaných v literatuře jsou rovněž implementovány v komerčních programových systémech.

## 6. Literatura


[1] Löster. T. – Řezanková, H. – Langhamrová, J. 2008. Statistické metody a demografie. Praha : VŠEM, 2008. 252 s. ISBN 978-80-86730-40-0.

[2] Pecáková, I. 2008. Statistika v terénních průzkumech. Praha: Professional Publishing, 2008. 231 s. ISBN 978-80-86946-74-0.

[3] Řehák, J. – Řeháková, B. 1986. Analýza kategorizovaných dat v sociologii. Praha : Academia, 1986. 397 s.

[4] Řezanková, H. 2007. Analýza dat z dotazníkových šetření. Praha : Professional Publishing, 2007. 212 s. ISBN 978-80-86946-49-8.

[5] Řezanková, H. 2008. Výuka jednorozměrné a dvourozměrné analýzy kategoriálních dat. In: Informační Bulletin České statistické společnosti, č. 2, 2008, s. 18 – 30.

[6] Řezanková, H. – Húsek, D. – Snášel, V. 2009. Shluková analýza dat. 2. rozšířené vydání. Praha : Professional Publishing, 2009. 218 s. ISBN 978-80-86946-81-8.

[7] Řezanková, H. – Löster. T. 2009. Úvod do statistiky. Praha : Oeconomica, 2009. 111 s. ISBN 978-80-245-1514-4.

[8] Stankovičová, I. – Vojtková, M. 2007. Viacrozmerné štatistické metódy s aplikáciami. Bratislava : Iura Edition, 2007. 261 s. ISBN 978-80-8078-152-1.


## Adresa autorky:


Řezanková Hana, prof. Ing. CSc.
VŠE – KSTP
nám. W. Churchilla 4
130 67 Praha 3
Česká republika
hana.rezankova@vse.cz

# Príspevok k analýze subjektívnej chudoby v SR a ČR
# Contribution to the Analysis of Subjective Poverty in Slovak and Czech Republic

Iveta Stankovičová, Jitka Bartošová

**Abstract:** The aim of this paper is to present ones of views of subjective poverty in Slovak and Czech Republic. We used data from statistical survey EU SILC. The goal of this survey is to obtain information on the distribution of income, the level and structure of poverty and the social exclusion. The survey is being done in accordance to the Regulation No. 1177/2003 of the European Parliament and the Council of Europe on the Community statistics on income and living conditions. It is a harmonized survey, which has been carried out from 2005 in all 25 (27) EU member countries.

This survey treats the poverty as a multi-dimensional phenomenon. It traces the poverty in relation to work, education and health; it notes the subjective opinions of households and individuals and relates the poverty to the social exclusion. Thus, for the purposes of the survey of basic dimensions of social exclusion, the data on primary and secondary needs of households, on housing and expenditures of households have been inquired. These serve for the calculation of indicators on the monetary and non-monetary poverty.

**Key words:** subjective poverty, statistical survey EU SILC 2007, Slovak Republic, Czech Republic, system SAS

**Kľúčové slová:** subjektívna chudoba, štatistické zisťovanie EU SILC 2007, Slovenská republika, Česká republika, systém SAS

## 1. Úvod

Príspevok vznikol v procese zoznamovania sa s obsahom súborov údajov výberového zisťovania o príjmoch a životných podmienkach domácností EU SILC 2007. Tieto súbory dát za Slovenskú a Českú republiku tvoria údajovú základňu pre riešenie dvoch výskumných projektov:

- projektu VEGA 1/4586/07 s názvom *Modelovanie sociálnej situácie obyvateľstva a domácností v Slovenskej republike a jej regionálne a medzinárodné porovnania*, ktorý je riešený na Katedre štatistiky FHI EU v Bratislave pod vedením prof. V. Pacákovej,
- projektu GAČR 402/09/0515 s názvom *Analýza a modelování finančního potenciálu českých (slovenských) domácností*, ktorý je riešený na Katedre managementu informácií FM VŠE v Jindřichovom Hradci pod vedením Dr. J. Bartošovej.

## 2. Subjektívna chudoba

Na otázku čo je chudoba existujú rôzne odpovede. Rovnako dôležité, ako povedať, čo chudoba je, je odpovedať aj na otázku, čo chudoba nie je. Veľmi často sa totiž za chudobu vydáva aj to, čo ňou v skutočnosti nie je.

V zásade možno rozlíšiť dva prístupy k vymedzeniu chudoby – objektívny a subjektívny. Objektívny prístup definuje chudobu prostredníctvom určitých kritérií, týkajúcich sa väčšinou príjmu alebo majetku človeka. Subjektívny prístup zisťuje, či sa človek sám cíti chudobný, pociťuje príznaky chudoby, respektíve sám seba zaraďuje do kategórie chudobných. Treba povedať, že subjektívny prístup z hľadiska skutočnej chudoby

nič nerieši. Iste, je dôležité, ako sa človek cíti a vníma svoju realitu. Určite je sociologicky zaujímavé to skúmať, zisťovať a porovnávať, ale nič nám to nepovie o skutočnej chudobe.

V rámci objektívneho prístupu k zisťovaniu chudoby treba rozlišovať prístupy absolútne a relatívne. Absolútne prístupy definujú chudobu cez určitú pevne stanovenú hodnotu. Princíp absolútnej chudoby ju popisuje v termínoch prežitia, a odvoláva sa teda na nevyhnutné podmienky, ktoré zabezpečujú, aby človek nezomrel. Podľa OSN (Kodaň 2005) chudobný je ten človek, ktorý trpí hladom a podvýživou, nemá prístup k pitnej vode, hygienickým zariadeniam ani k zdravotnej starostlivosti, má obmedzený alebo nijaký prístup k vzdelaniu a informáciám. Okrem toho býva v neadekvátnych podmienkach, navyše v nezdravom životnom prostredí a v rámci jeho sociálnej skupiny rastie úmrtnosť.

Relatívne prístupy definujú chudobu cez pomer k niečomu inému – priemernému príjmu, mediánu príjmu, rozloženiu príjmových skupín. Tento druhý koncept hovorí o relatívnej chudobe, ktorú najlepšie vystihuje profesor Peter Townsend z London School of Economics. Hovorí, že jedincov, rodiny a skupiny v populácii možno považovať za chudobných, ak im chýbajú zdroje na zabezpečenie niektorých druhov stravy, životných podmienok a výdobytkov, ktoré sú zvyčajné v spoločnostiach, do ktorých patria. Pri takomto prístupe sa teda berie do úvahy aj stupeň rozvoja spoločnosti a pomery, ktoré v nej prevládajú. Význam sociálneho kontextu na tematizovanie chudoby zdôrazňuje definícia prijatá Európskou komisiou v roku 1984. Podľa nej za chudobných možno považovať osoby, rodiny a skupiny osôb, ktorých zdroje (materiálne, kultúrne a sociálne) sú také obmedzené, že ich vylučujú z minimálne akceptovateľného spôsobu života členských štátov, v ktorých žijú.

Zvolené nástroje na meranie chudoby vyplývajú z toho, akým spôsobom ju definujeme. Možno ju pritom merať prostredníctvom výšky príjmov, výdavkov na spotrebu, životného či existenčného minima alebo aj pomocou subjektívnych výpovedí. Pri meraní chudoby je preto dôležité poznať prečo, ako, čo a kto meria. Dôvody, ktoré vedú k meraniu chudoby, do veľkej miery určujú aj priebeh merania. Rôzne ciele totiž môžu viesť k rôznym metódam merania, teda aj k rôznym výsledkom. Výber konkrétnej metódy merania chudoby môže mať vážne morálne a politické dôsledky.

Všeobecne sa pri meraní chudoby uplatňujú tri prístupy. Profesionálni experti často vytvárajú tzv. budget standards a definujú nevyhnutný okruh tovarov a služieb pre rôzne typy domácností. Konsenzuálny – „demokratický“ – prístup sa opiera o názory celej populácie, nielen o názory expertov. Participatívny prístup vychádza z presvedčenia, že ľudia zažívajúci chudobu, sú sami najlepšími odborníkmi na tento problém a ich názory by sa mali zohľadňovať v každej fáze merania či výskumu o chudobe; nemali by teda byť len objektmi, ale aj subjektmi výskumného procesu.

V našom príspevku pôjde o určitý participatívny prístup k chudobe. Pokúsime sa analyzovať pocit chudoby na základe subjektívnych odpovedí respondentov výberového zisťovania EU SILC na Slovensku a v Čechách.

## 3. Údaje EU SILC 2007

Ako vstupné dáta pre analýzu a vizualizáciu pocitu subjektívnej chudoby sme použili údaje výberového Zisťovania o príjmoch a životných podmienkach domácností EU SILC 2007. Zisťovanie je realizované v zmysle Nariadenia č. 1177/2003 Európskeho parlamentu a Rady EÚ a doplňujúceho Nariadenia 1553/2005. EU SILC je ročné výberové zisťovanie, ktorého cieľom je získať informácie o rozdelení príjmov, o úrovni a štruktúre chudoby a o sociálnom vylúčení v sledovanej krajine. K dispozícii sme mali údaje EU SILC 2007 zo Slovenskej republiky (SR) aj Českej republiky (ČR).

Jednotkami výberu v EU SILC sú hospodáriace domácnosti a jej súčasní členovia. Hospodáriace domácnosti sú tiež referenčnou jednotkou zisťovania. Hospodáriace domácnosti sú definované ako súkromné domácnosti tvorené osobami v byte, ktoré spoločne žijú a

spoločne hospodária, vrátane spoločného zabezpečovania životných potrieb. Za znak spoločného hospodárenia sa považuje spoločná úhrada základných výdavkov domácnosti (strava, úhrada nákladov na bývanie, elektrina, plyn a pod.). Príjmové referenčné obdobie je kalendárny rok, ktorý predchádza roku zisťovania t. j. rok 2006.

Štatistické zisťovanie EU SILC 2007 SR sa uskutočnilo na Slovensku v apríli 2007. Do výberu bolo zaradených 5840 domácností a 12763 osôb. Miera návratnosti bola takmer 85%. V súbore o domácnostiach (HFILE) sa aktuálne nachádza 4941 záznamových viet, čiže vyšetrených domácností. Zberu údajov sa zúčastnilo vyše 400 opytovateľov, ktorí navštívili domácnosti v 308 obciach Slovenska. Pre EU SILC 2007 bol použitý jednostupňový stratifikovaný výber. Domácnosti sa vyberali proporcionálne jednoduchým náhodným výberom. Oporou výberu boli údaje zo Sčítania obyvateľov, domov a bytov 2001. Pri aktualizácii opory výberu sa použili informácie o úbytku, resp. prírastku (novopostavených a skolaudovaných) trvale obývaných domov a bytov v krajoch v období rokov 2001-2004 a 2004 -2006.

Štatistické zisťovanie EU SILC 2007 ČR sa uskutočnilo v Čechách od 17. februára do 29. apríla 2007. Do výberu bolo zaradených 11611 hospodáriacich domácností, ktoré žili v 11496 bytoch (čiže viac domácností žije v jednom byte). Celkovo bolo vyšetrených 9675 domácností, čiže miera návratnosti bola 83,3%. Výberový plán bol založený na náhodnom dvojstupňovom výbere pre každý kraj nezávisle tak, aby celkový počet vybraných bytov bol úmerný veľkosti jednotlivých krajov. Na prvom stupni boli na základe „Registru sčítacích obvodů“ náhodne vybrané sčítacie obvody, vo vybraných sčítacích obvodoch bolo následne na druhom stupni vyberaných 10 bytov. Pred výberom sčítacích obvodov bolo nutné oporu výberu upraviť tak, aby mohli byť do šetrenia zaradené aj sčítacie obvody s malým počtom bytov a tak sa dosiahlo požadovaného pokrytia celého územia ČR.

Definície premenných v EU SILC a niektorých štatistických mier (ukazovateľov) z údajov počítaných sú nasledovné:

*Celkový disponibilný príjem domácnosti* predstavuje príjem vypočítaný ako suma zložiek hrubého osobného príjmu všetkých členov domácnosti plus zložky hrubého príjmu na úrovni domácnosti (napr. príjem z prenájmu majetku, prijaté transfery od iných domácností) mínus pravidelné dane z majetku, pravidelné platené transfery medzi domácnosťami (napr. výživné, pravidelná peňažná pomoc od iných domácností), daň z príjmu a príspevky na sociálne poistenie.

*Ekvivalentná škála* je škála koeficientov použitá na výpočet indikátorov chudoby v súlade s metodikou Eurostatu, tzv. modifikovaná škála OECD, kde koeficient 1 sa použije pre prvého dospelého člena domácnosti, 0,5 pre druhého a každého dospelého člena domácnosti, 0,5 pre 14-ročných a starších a 0,3 pre každé dieťa mladšie ako 14 rokov. Používa sa na výpočet ekvivalentnej veľkosti domácnosti.

*Ekvivalentný disponibilný príjem* je disponibilný príjem domácnosti vydelený ekvivalentnou veľkosťou domácnosti. Tento príjem je potom priradený každému členovi domácnosti. Podľa výsledkov EU SILC 2007 SR bol ekvivalentný disponibilný príjem domácnosti na osobu a na rok 163 tis. Sk (t.j. na mesiac 14 096 Sk). V porovnaní s rokom 2005 ekvivalentný disponibilný príjem na osobu a mesiac vzrástol v roku 2006 o 2 334 Sk. Priemerný čistý príjem na osobu v ČR dosiahol 118 tis. Kč (ročne).

*Medián ekvivalentného disponibilného príjmu* je hodnota ekvivalentného disponibilného príjmu, ktorá rozdeľuje súbor podľa výšky príjmu na dve rovnako početné časti podľa počtu osôb.

*Hranica rizika chudoby* je hodnota hranice rizika chudoby (60% mediánu národného ekvivalentného príjmu) v Sk (resp. Kč), v prepočte na paritu kúpnej sily a na Euro. V SR je to asi 7 tis. Sk na mesiac a v ČR je to zhruba čiastka 8 tis. Kč mesačne.

*Miera rizika chudoby* predstavuje podiel osôb s ekvivalentným disponibilným príjmom pod hranicou 60 % národného mediánu ekvivalentného príjmu. Podľa výsledkov EU SILC 2007 SR bolo v roku 2006 ohrozených rizikom peňažnej (monetárnej) chudoby 10,7 % obyvateľov Slovenska. Oproti roku 2004 sa miera rizika chudoby znížila o 2,6 percentuálneho bodu a v porovnaní s rokom 2005 poklesla o 0,9 percentuálneho bodu. V ČR peňažnou chudobou bolo ohrozených celkom 995 tis. osôb (t.j. 9,76 % zo všetkých osôb) a v podstate sa tento podiel oproti roku 2005 nezmenil.

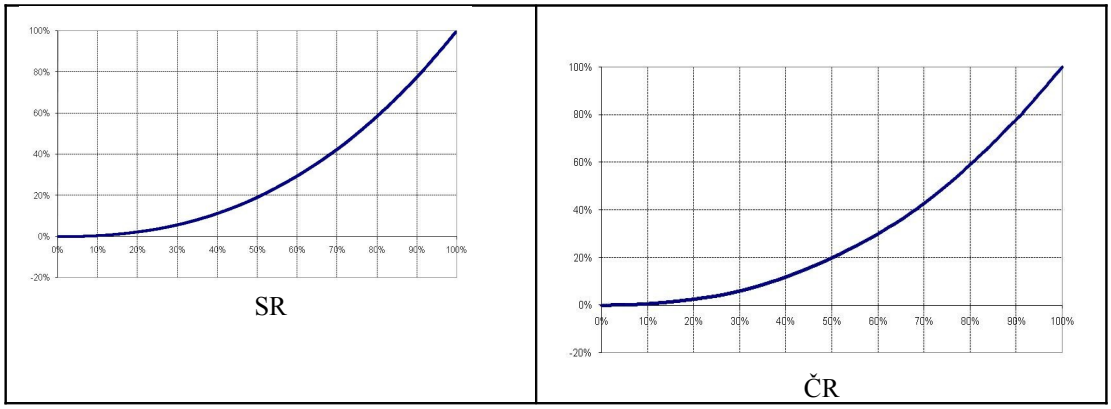
*Rozptyl okolo hranice rizika chudoby* je percentuálne vyjadrenie počtu osôb z celej populácie, ktorých ekvivalentný disponibilný príjem je nižší ako 40%, 50% a 70% mediánu ekvivalentného disponibilného príjmu.

*Pomer príjmov horného a dolného kvintilu (S80/S20)* je podiel sumy príjmov 20% osôb z populácie s najvyššími príjmami (horný kvintil) k podielu sumy príjmov 20% osôb z populácie s najnižšími príjmami (dolný kvintil).

*Gini koeficient* je súhrnná miera kumulatívneho podielu ekvivalentných disponibilných príjmov zodpovedajúceho kumulatívnemu percentu počtu jedincov. Vyjadruje nerovnomernosť príjmového rozdelenia. Na Slovensku bola hodnota Giniho indexu 24% (r. 2006) a v Čechách je to podobné číslo, 25% (r. 2005). Naše krajiny patria medzi štáty s pomerne rovnomerným rozdelením príjmov (viď Lorenzove krivky, Obrázok 1). Vysoké hodnoty tohto indexu mali napr. USA (45, r. 2007), UK (34, r. 2005), ale aj Bosna a Hercegovina (56, r. 2007). Naopak nízke hodnoty dosahujú napr. Švédsko (23, r. 2005), Dánsko (24, r. 2005) ale aj Chorvátsko (29, r. 2008)a Slovinsko (24, r. 2005).

*Relatívny pomer mediánu príjmov* osôb vo veku 65+ je definovaný ako pomer mediánu ekvivalentného disponibilného príjmu osôb vo veku 65+ k mediánu ekvivalentného disponibilného príjmu osôb vo veku 0-64.

*Agregovaný pomer kompenzácie* je pomer mediánu osobného príjmu z dôchodkov osôb v dôchodkovom veku, t. j. vo veku 65-74 rokov k mediánu osobného príjmu zo zamestnania, resp. podnikania osôb vo veku 50-59 rokov.




***Obrázok 1*** *Lorenzove krivky celkových disponibilných príjmov domácností v SR a ČR*

## 4. Použité premenné z EU SILC 2007 pre analýzu pocitu chudoby

Pre vizualizáciu a analýzu pocitu chudoby na Slovensku a v Čechách sme použili vybrané premenné z EU SILC 2007 v SR a ČR. Na výpočty a tvorbu grafov sme použili systém SAS Enterprise Guide verziu 4.1 a tiež SAS Enterprise Miner 5.3.

**Premenné z EU SILC 2007 SR a ich definície:**

- **HS120** - schopnosť vystačiť s peniazmi. Ide o odhad respondenta, na akej úrovni sa nachádza jeho domácnosť v súvislosti so schopnosťou vystačiť s peniazmi. V súvislosti

s celkovým mesačným príjmom domácnosti bolo potrebné posúdiť, s akým stupňom ťažkostí sa domácnosť vyrovnáva s platením zvyčajných výdavkov. Pod zvyčajnými výdavkami sa rozumeli výdavky domácnosti na stravu, nájomné, úhradu za elektrinu, plyn, atď.. (V dátach ČR premenná VYCHAZELA.)

***Tabuľka 1*** *Kódovanie hodnôt premenných HS120 a VYCHAZELA*

| Kód hodnoty | Popis hodnôt HS120 | Popis hodnôt VYCHAZELA |
|---|---|---|
| 1 | s veľkými ťažkosťami | s velkými obtížemi |
| 2 | s ťažkosťami | s obtížemi |
| 3 | s určitými ťažkosťami | s menšími obtížemi |
| 4 | pomerne ľahko | docela snadno |
| 5 | ľahko | snadno |
| 6 | veľmi ľahko | velmi snadno |

- **HS130** - najnižší mesačný príjem postačujúci na vyžitie (Sk za mesiac). Respondent v domácnosti mal odhadnúť sumu podľa vlastného chápania pojmu „vystačiť s peniazmi“. Otázka sa zodpovedala v súvislosti so súčasným zložením domácnosti a so súčasnými výdavkami. (V dátach ČR premenná MIN_PRIJ.)
- **HY010** - celkový hrubý príjem domácnosti (Sk za rok).
- **HY020** - celkový disponibilný príjem domácnosti (Sk za rok).
- **HY022** - celkový disponibilný príjem domácnosti pred sociálnymi transfermi inými ako starobnými dávkami a dávkami pre pozostalých (Sk za rok).
- **HY023** - celkový disponibilný príjem domácnosti pred sociálnymi transfermi vrátane starobných dávok a dávok pre pozostalých (Sk za rok).
- **HX050** - ekvivalentná veľkosť domácnosti. Pre výpočet ekvivalentnej veľkosti domácnosti sa použila OECD modifikovaná škála (váhy 1 pre prvého dospelého člena domácnosti, 0,5 pre každého ďalšieho dospelého člena a 0,3 pre dieťa mladšie ako 14 rokov). (V dátach ČR premenná EJ.)
- **HX100** - ekvivalentný disponibilný príjem domácnosti. Ekvivalentný disponibilný príjem sa vypočíta tak, že disponibilný príjem domácnosti (HX023) sa vydelí ekvivalentnou veľkosťou domácnosti (HX050) (v Sk za rok). Týmto postupom získame aj záporné hodnoty, lebo premenná HY023 môže nadobúdať aj nekladné hodnoty. Pre ďalšie analýzy by bolo vhodné záporné hodnoty vynechať (je to len 8 domácností v celom súbore HFILE, t.j. po prepočítaní na populáciu to predstavuje 0,2% domácností SR). Je možné tiež nahradiť záporné čísla hodnotou 0, ktorá by znamenala, že tieto domácnosti sú v hmotnej núdzi. (V dátach ČR 2005 premenná EU_PRIJ, v dátach ČR 2007 sa priamo nenachádza.)
- **DB090** - prierezové váhy domácnosti. Používajú sa na vyvodenie záverov týkajúcich sa základného súboru súkromných domácností na národnej a európskej úrovni. (V dátach ČR premenná PKOEF.)

**<u>Premenné z EU SILC 2007 ČR a ich definície:</u>**

- **VYCHAZELA** domácnost vycházela s příjmy. (V dátach SR premenná HS120. Kódovanie hodnôt je rovnaké (viď Tabuľka 1)).
- **MIN_PRIJ** - subjektivní minimální příjem - vlastní odhad (v Kč za měsíc). (V dátach SR premenná HS130.)
- **HPRIJMY** - hrubé peněžní příjmy (v Kč za rok). (V dátach SR premenná HY010.)
- **CP_PRIJ** - čistý peněžní příjem domácnosti (v Kč za rok).
- **EU_PRIJ** - čistý disponibilní příjem domácnosti dle def. EU (v Kč za rok). (V dátach SR pravdepodobne analogická premenná HX100. *Poznámka:* Táto premenná sa vyskytovala v českých dátach EU SILC len v roku 2005 a v roku 2007 už nie. Pre našu

analýzu sme túto premennú vytvorili ako podiel čistého peňažného príjmu (CP_PRIJ) a počtu spotrebných jednotiek podľa def. EU (EJ)).

- **PKOEF** - přepočítací koeficient pro přepočet výsledků z výběru na celý soubor v populaci.
- **SJ** - počet spotřebních jednotek - def. OECD. Součet za jednotlivé osoby v domácnosti; výše spotřební jednotky pro jednotlivé osoby závisí na složení domácnosti a věku dětí: 1,0 - osoba v čele domácnosti, 0,5 - děti ve věku 0 až 13 let, 0,7 - ostatní děti a osoby.
- **EJ** - počet spotřebních jednotek - def. EU. Součet za jednotlivé osoby v domácnosti; výše spotřební jednotky pro jednotlivé osoby závisí na složení domácnosti a věku dětí: 1,0 - osoba v čele domácnosti. 0,3 - děti ve věku 0 až 13 let, 0,5 - ostatní děti a osoby. (V dátach SR premenná HX050.)

## 5. Výsledky porovnania subjektívnej chudoby v SR a ČR a ich vizualizácia

Na základe popisných štatistík (Tabuľka 2) je zrejmé, že v SR máme vyše 1,9 mil. domácností a v ČR je to viac ako 4 mil. domácností. Zistili sme, že na Slovensku sú domácnosti početnejšie ako v Čechách. Premenná HX50 – ekvivalentná veľkosť domácnosti podľa metodiky EÚ v SR nadobúdala hodnoty 1 až 6,8 (priemer 1,8) a v ČR porovnateľná premenná EJ len hodnoty 1 až 4,9 (priemer 1,7). Neprepočítané počty osôb v domácnostiach sa pohybovali v intervale 1 až 15 osôb na Slovensku (HX070, priemer 2,8) a 1 až 10 osôb v Čechách (OSOB, priemer 2,5). V Čechách majú rodiny aj nižší počet detí, maximálne 8 (DETI - priemer 0,6 a DETI_EU – priemer 0,6), kým na Slovensku maximálne až 9 (HX090, priemer 0,7).

Základné charakteristiky príjmov domácností v SR a ČR uvádzame v prehľadnej tabuľke (Tabuľka 2) a tak si čitateľ môže urobiť prehľad sám. Aby sme odstránili určité skreslenia popisných štatistík z dát SR, tak sme zo súboru odfiltrovali záporné čísla premennej HX100 (8 hodnôt, t.j. 0,2% domácností SR) a tiež hodnoty nad 400 tis. Sk (94 hodnôt, t.j. 2,1% domácností SR). Tieto extrémne vysoké príjmy v súbore domácností SR dosahovali priemernú hodnotu 584 tis. Sk. Po filtrácii dát uvedeným postupom premenná HX100 - ekvivalentný disponibilný príjem domácnosti nadobúdal hodnoty z intervalu 1600 až 397762 Sk (Tabuľka 3).

***Tabuľka 2*** *Popisné štatistiky vybraných premenných o príjmoch domácností v ČR a SR*

| Štát | Variable | Mean | Std Dev | Min | Max | N | Lower Quartile | Median | Upper Quartile | Coeff of Variation |
|---|---|---|---|---|---|---|---|---|---|---|
| ČR | HPRIJMY | 355117 | 289950 | 16277 | 8062352 | 4038483 | 182400 | 291552 | 452508 | 81.6 |
| ČR | CP_PRIJ | 297627 | 208846 | 16254 | 5569100 | 4038483 | 170630 | 253446 | 374469 | 70.2 |
| ČR | EU_PRIJ | 173215 | 105745 | 11689 | 2651952 | 4038483 | 117600 | 149383 | 202873 | 61.0 |
| ČR | MIN_PRIJ | 19954 | 9892 | 3000 | 99999 | 4038483 | 13000 | 19000 | 25000 | 49.6 |
| SR | HY010 | 339032 | 241752 | 0 | 2643441 | 1907168 | 166080 | 285360 | 447100 | 71.3 |
| SR | HY020 | 291515 | 190884 | -1870 | 2158560 | 1907168 | 152637 | 251226 | 378755 | 65.5 |
| SR | HY022 | 270654 | 189797 | -25000 | 2130600 | 1907168 | 129500 | 225187 | 356780 | 70.1 |
| SR | HY023 | 206611 | 210192 | -35000 | 2130600 | 1907168 | 0 | 173506 | 327810 | 101.7 |
| SR | HX100 | 160203 | 93191 | -7480 | 2117669 | 1907168 | 110190 | 140872 | 186234 | 58.2 |
| SR | HS130 | 29318 | 17009 | 2000 | 400000 | 1902045 | 18000 | 25000 | 40000 | 58.0 |

***Tabuľka 3*** *Popisné štatistiky vybraných premenných o príjmoch domácností SR po úpravách súboru*

| Štát | Variable | Mean | Std Dev | Min | Max | N | Lower Quartile | Median | Upper Quartile | Coeff of Variation |
|---|---|---|---|---|---|---|---|---|---|---|
| **SR** | **HY010** | 324725 | 208164 | 1600 | 1540700 | 1863416 | 164182 | 280234 | 437000 | 64 |
| **SR** | **HY020** | 280466 | 165306 | 1600 | 1224395 | 1863416 | 151160 | 248014 | 370988 | 59 |
| **SR** | **HY022** | 259958 | 164715 | -25000 | 1224395 | 1863416 | 128664 | 222117 | 350221 | 63 |
| **SR** | **HY023** | 195944 | 187584 | -35000 | 1180408 | 1863416 | 0 | 169738 | 321115 | 96 |
| **SR** | **HX100** | 151463 | 60930 | 1600 | 397762 | 1863416 | 109795 | 139919 | 182175 | 40 |
| **SR** | **HS130** | 29096 | 16101 | 2000 | 330000 | 1859462 | 18000 | 25000 | 38000 | 55 |

Rozdelenie domácností v SR a ČR podľa otázky, ktorá zachytáva subjektívny pocit chudoby (Tabuľka 4, premenné HS120 a VYCHAZELA), nebolo významne rozdielne na hladine významnosti 0,1. Výsledná p-hodnota chí-kvadrát testu bola 0,07. V súbore domácností SR bolo 8 neudaných odpovedí na túto otázku, preto sa úhrnný počet domácností (suma stĺpca N) v tabuľke 6 líši od celkového počtu domácností (N) v tabuľkách 3 a 4. Zaujímavé je, že v súbore českých domácností sa chýbajúce odpovede nevyskytovali.

Zdá sa nám ďalej vhodné prekódovať počet možných odpovedí na otázku o pocite chudoby zo 6-tich odpovedí na dve, a tak vytvoriť binárnu premennú. Je dobré zlúčiť prvé tri odpovede a vytvoriť kategóriu 1 „subjektívne chudobných“ domácností a zlúčiť odpovede 4 až 6 a vytvoriť kategóriu 0 „subjektívne nechudobných“ domácností. Zisťujeme, že na Slovensku má pocit chudoby až 77,4% a v Čechách „len“ 63,3% domácností (Tabuľka 4).

Výrazné rozdiely sú aj v odpovediach o výške potrebnej sumy príjmu na mesiac pre domácnosť (HS130, MIN_PRIJ). Na Slovensku sa hodnoty pohybovali v rozmedzí 2 až 400 tis. Sk, resp. maximálne do 330 tis. Sk v odfiltrovaných dátach a priemerná hodnota bola zhruba 29 tis. Sk (Tabuľka 2 a Tabuľka 3, premenná HS130). V Čechách sa odpoveď na otázku o minimálnom príjme pohybovala medzi 3 až 100 tis. Kč s priemernou hodnotou skoro 20 tis. Kč (Tabuľka 2, MIN_PRIJ). Variabilita odpovedí na túto otázku je vysoká (viď Obrázok 2). Na Slovensku sú domácnosti, ktoré majú vysoké príjmy a ešte ich členovia majú pocit, že im ich príjmy nestačia na pokrytie ich vysokých nárokov. Tieto rozdiely by sme mohli vysvetliť rôznymi subjektívnymi aj objektívnymi príčinami, ale to nie je cieľom tohto príspevku.

***Tabuľka 4*** *Rozdelenie domácností (v %) podľa pocitu chudoby (HS120 a VYCHAZELA)*

| Pocit subjektívnej chudoby | Početnosť v % SR | Početnosť v % ČR | Kum. početnosť v % SR | Kum. početnosť v % ČR |
|---|---|---|---|---|
| **1** | 11.2 | 7.5 | 11.2 | 7.5 |
| **2** | 21.4 | 19.3 | 32.6 | 26.8 |
| **3** | 44.8 | 36.5 | **77.4** | **63.3** |
| **4** | 18.0 | 25.5 | 95.4 | 88.8 |
| **5** | 3.9 | 9.8 | 99.2 | 98.6 |
| **6** | 0.8 | 1.4 | 100.0 | 100.0 |

Nasledujúce tabuľky a grafy popisujú a vizualizujú pocit chudoby v domácnostiach SR a ČR podľa vyjadrení sa respondentov ako vychádzajú so svojími príjmami a koľko by mesačne potrebovali peňazí. Analyzujeme vybrané premenné podľa 6-tich možných odpovedí (HS120 a VYCHAZELA).

***Tabuľka 5*** *Subjektívny minimálny mesačný príjem (HS130) podľa pocitu chudoby (HS120) v SR*

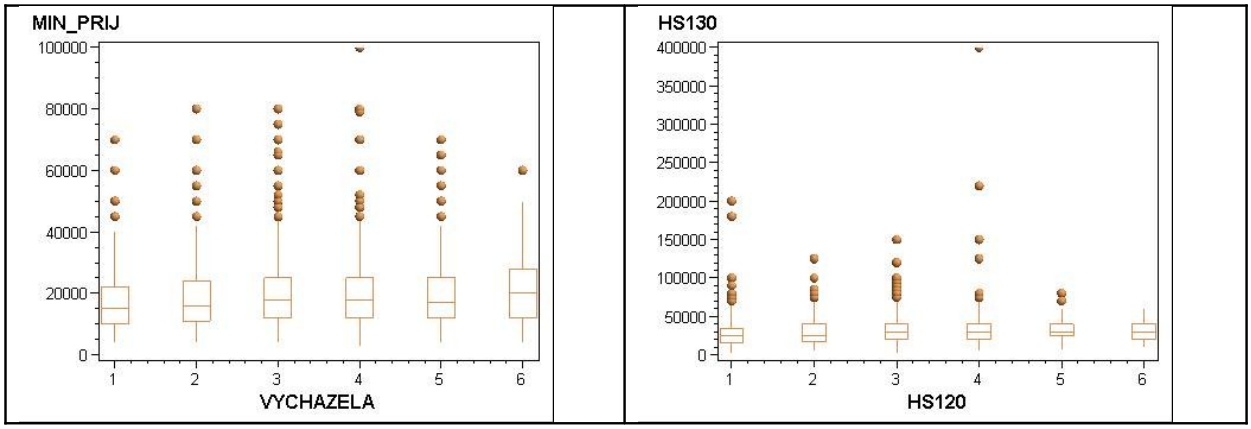
| HS120 | Mean HS130 | Std Dev | Min | Max | N | Lower Quartile | Median | Upper Quartile | Coeff of Variation |
|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 26023.0 | 16819.5 | 3000 | 200000 | 212352 | 15000 | 22000 | 30000 | 64.6 |
| **2** | 27931.6 | 15410.2 | 5500 | 125000 | 405914 | 15000 | 25000 | 38000 | 55.2 |
| **3** | 29411.3 | 14906.1 | 2000 | 150000 | 851283 | 20000 | 26000 | 40000 | 50.7 |
| **4** | 31898.6 | 20679.6 | 5000 | 400000 | 340970 | 20000 | 30000 | 40000 | 64.8 |
| **5** | 31897.4 | 13932.1 | 7000 | 80000 | 72903 | 20000 | 30000 | 40000 | 43.7 |
| **6** | 27711.5 | 11527.1 | 10000 | 60000 | 14601 | 20000 | 25000 | 35000 | 41.6 |

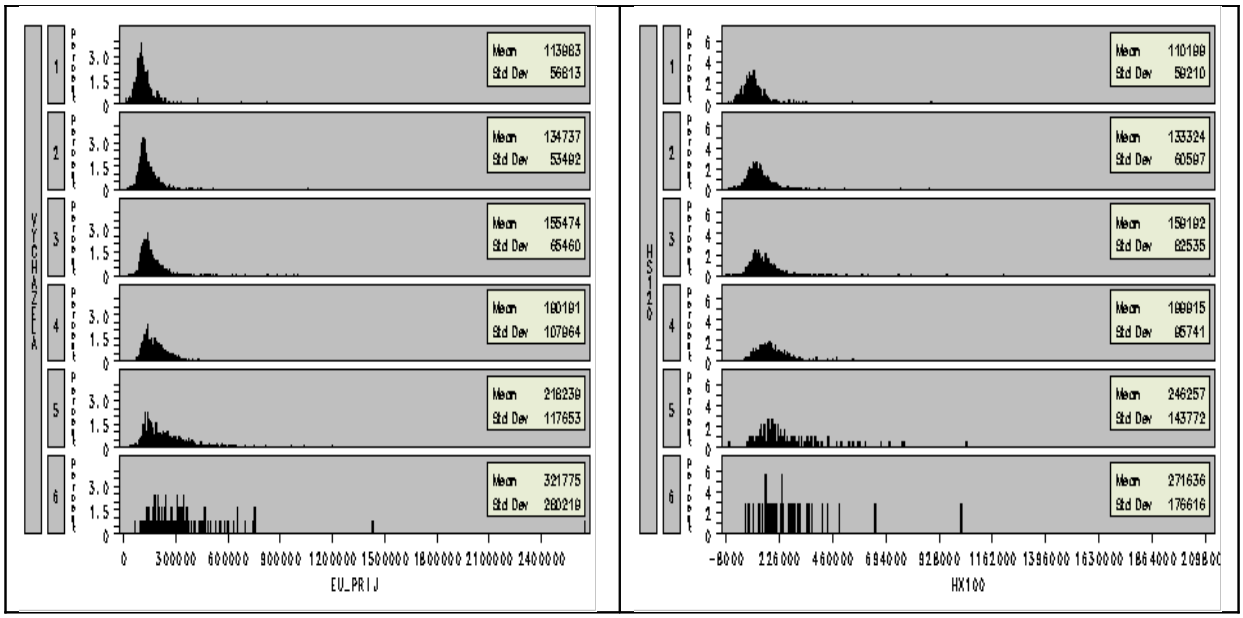
***Tabuľka 6*** *Subjektívny minimálny mesačný príjem (MIN_PRIJ) podľa pocitu chudoby (VYCHAZELA) v ČR*

| VYCHAZELA | Mean MIN_PRIJ | Std Dev | Min | Max | N | Lower Quartile | Median | Upper Quartile | Coeff of Variation |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 17922.9 | 9241.3 | 4000 | 70000 | 301839 | 10000 | 15000 | 24000 | 51.6 |
| **2** | 19326.1 | 9477.3 | 4000 | 80000 | 778990 | 12000 | 18000 | 25000 | 49.0 |
| **3** | 20308.5 | 9868.3 | 4000 | 80000 | 1474973 | 14000 | 20000 | 25000 | 48.6 |
| **4** | 20275.7 | 9917.5 | 3000 | 99999 | 1028854 | 14000 | 20000 | 25000 | 48.9 |
| **5** | 20149.8 | 10427.9 | 4000 | 70000 | 395776 | 13000 | 20000 | 25000 | 51.8 |
| **6** | 22924.7 | 12587.7 | 4000 | 60000 | 58051 | 15000 | 20000 | 30000 | 54.9 |

Z tabuliek a obrázkov uvádzaných nižšie je zrejmé, že čo sa týka potrebného minimálneho príjmu na mesiac pre vyžitie domácnosti, tak priemer udávaných hodnôt stúpa v ČR aj v SR v závislosti od odpovede na otázku o pocite subjektívnej chudoby. Podľa rôznych testov neparametrickej analýzy rozptylu sú tieto rozdiely štatisticky významné (p-hodnoty boli vždy nižšie ako 0,0001). Potvrdil sa známy fakt, že čím má človek viac, tak aj viac potrebuje na zabezpečenie svojich subjektívnych potrieb.

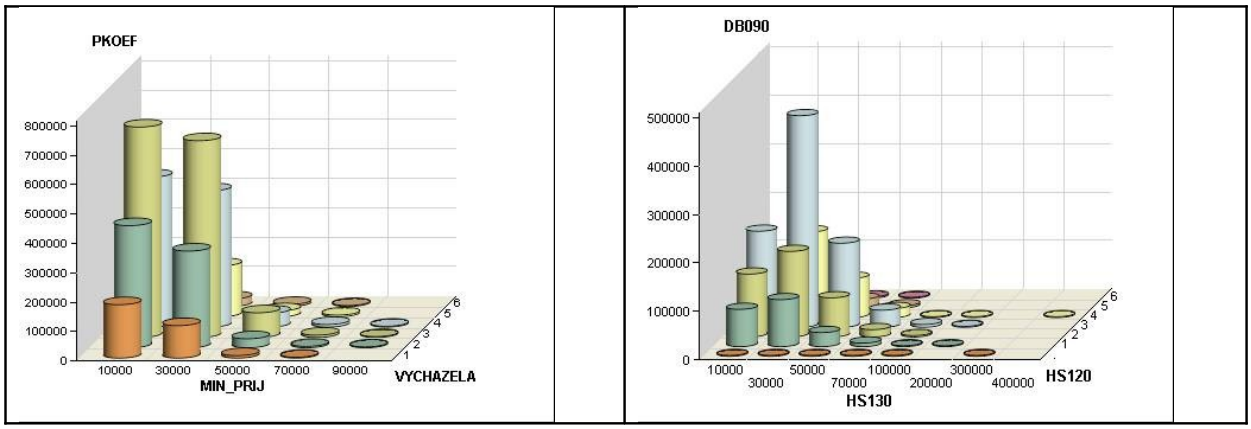



***Obrázok 2*** *Box-plot grafy pre odpovede o minimálnom príjme podľa schopnosti vystačiť s peniazmi v ČR a SR*

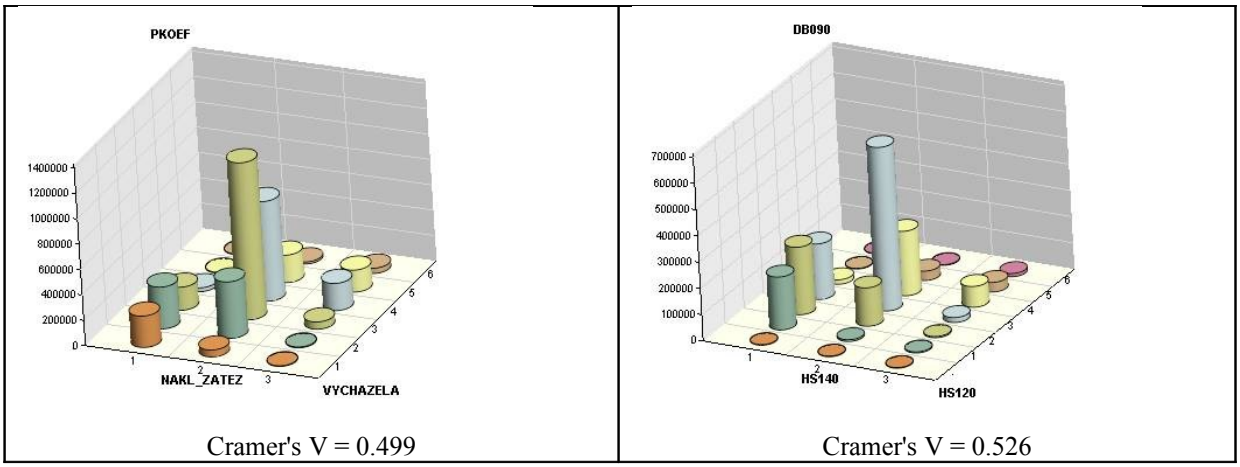



***Obrázok 3*** *Rozdelenia ekvivalentných disponibilných príjmov podľa schopnosti vystačiť s peniazmi v ČR a SR*

Odpovede na otázku ako domácnosť vychádzala s príjmami súvisia s odpoveďami na ostatné subjektívne otázky typu: „Ako domácnosť zaťažujú finančné náklady na bývanie?“, resp. „Ako domácnosť zaťažujú finančné náklady na elektrinu?“ a podobne. Analýza pomocou rozhodovacích stromov nám ukázala, že najviac s analyzovanou otázkou súvisí práve otázka o subjektívnom pocite, že domácnosť zaťažujú náklady na bývanie (HS140 - celkové náklady na bývanie ako finančná záťaž pre domácnosť: 1-veľmi zaťažuje, 2-trochu zaťažuje, 3-vôbec nezaťažuje. NAKL_ZATEZ - náklady na bydlení finanční zátěží: 1-velkou zátěží, 2-určitou zátěží, 3-žádnou zátěží.) Chí-kvadrát test potvrdil štatistickú významnosť tejto asociácie (viď Obrázok 5).

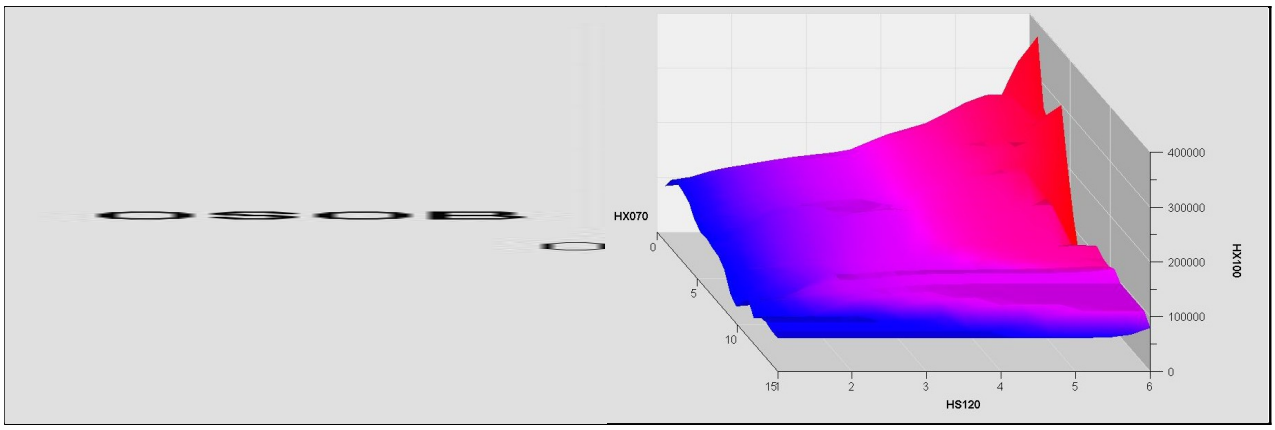



***Obrázok 4*** *Rozdelenie odpovedí o minimálnom príjme podľa schopnosti vystačiť s peniazmi v ČR a SR*




Cramer's V = 0.499 Cramer's V = 0.526

***Obrázok 5*** *Rozdelenie odpovedí o pocite chudoby podľa záťaže s finančnými nákladmi na bývanie v ČR a SR*




***Obrázok 6*** *Rozdelenie ekvivalentných disponibiln7ch príjmov podľa pocitu chudoby v závislosti od počtu osôb domácnosti v ČR a SR*

## 6. Záver

Analýza subjektívnej chudoby na základe údajov z výberového zisťovania EU SILC potvrdzuje, že sa o skutočnej chudobe nedozvieme veľa. Zistili sme len, že vysoký podiel domácností v SR aj ČR (viac ako 60%) nie je spokojných so svojím príjmom a má pocit, že na pokrytie svojich výdavkov, by potrebovali viac prostriedkov ako majú práve k dispozícii. Predstava o minimálnom príjme domácností je variabilná viac na Slovensku ako v Čechách. Zdá sa, že ľudia v Čechách v tejto oblasti uvažujú v reálnejších a nižších dimenziách ako na Slovensku.

Na záver treba ešte povedať, že výsledky sme prezentovali v národných menách oboch štátov, takže nie sú bezprostredne porovnateľné. Problém je, akými kurzami a v akej mene by bolo vhodné výsledky vyjadriť. Je niekoľko možností, ktoré nechávame na čitateľa, aby sa rozhodol a naše čísla si prepočítal.

## 7. Literatúra

[1]Bartošová, J.: Analysis and Modelling of Financial Power of Czech Households. In Aplimat – Journal of Applied Mathematics, Vol. 2, Nr. 3, Slovak Technical University, Bratislava, 2009, s.31-36, ISSN 1337-6365.

[2]Dostál, O.: Chudoba a jej príčiny. In Týždeň 49/2005 zo dňa 5. 12. 2005. Skrátená verzia prednášky z konferencie "Chudoba v slovenskej spoločnosti a vzťah slovenskej spoločnosti k chudobe". Bratislava, 14. 11. 2005. Konzervatívny inštitút M.R.Štefánika. Dostupné na internete (8. 5. 2009): <http://www.konzervativizmus.sk/article.php?723>

[3]Labudová, V.: Analýza monetárnej chudoby na Slovensku. SAS Forum 2008. Bratislava, október 2008.

[4]Stankovičová, I., Vojtková, M.: Viacrozmerné štatistické metódy s aplikáciami. Bratislava : Iura Edition, 2007. 261 s. ISBN 978-80-8078-152-1.

[5]The 2008 World FactBook. ISSN 1553-8133. Dostupné na internete (8. 5. 2009): <https://www.cia.gov/library/publications/the-world-factbook/index.html>

[6]Želinský, T., Hudec, O.: Odhad subjektívnej chudoby na Slovensku založený na distribučnej funkcii rozdelenia príjmov. In Forum Statisticum Slovacum 7/2008. SŠDS, Bratislava, s. 152-157. ISSN 1336-7420.

[7]Materiály zo Štatistického úradu SR, web stránka dostupná na : http://portal.statistics.sk

[8]Materiály z Českého štatistického úradu, web stránka dostupná na: http://www.czso.cz

[9]Životní podmínky (EU-SILC). Dostupné na (9. 5. 2009):
<http://www.czso.cz/csu/redakce.nsf/i/zivotni_podminky_(eu_silc)>

## 8. Zdroje údajov

[10] ŠÚ SR, SILC 2007 UDB verzia 20/08/08[1]

[11] ČSÚ, SILC 2007

## Adresy autoriek:

Ing. Iveta Stankovičová, Ph.D.
Katedra informačných systémov
Fakulta managementu UK v Bratislave
Odbojárov 10, P. O. Box 95
820 05 Bratislava 25
iveta.stankovicova@fm.uniba.sk

RNDr. Jitka Bartošová, Ph.D.
Katedra managementu informací
Fakulta managementu VŠE
Jarošovská 1117/II
377 01 Jindřichův Hradec
bartosov@fm.vse.cz

[i] Príspevok bol vytvorený ako súčasť riešenia projektu VEGA 1/4586/07: *Modelovanie sociálnej situácie obyvateľstva a domácností v Slovenskej republike a jej regionálne a medzinárodné porovnania* a projektu GAČR 402/09/0515: *Analýza a modelování finančního potenciálu českých (slovenských) domácností*.

# Aproximace Langevinovou funkcí

Jaroslav Marek, Michaela Tučková, Pavel Tuček

**Klíčová slova:** Nelineární regresní modely, linearizované regresní modely, BLUE, nanomateriály, Langevinova funkce, hysterézní smyčka.

**Abstract:** Owing to their wide application potential, magnetic nanomaterials are of great interest for the scientific community. Their magnetic properties can be deduced by measuring their magnetization, being the fundamental magnetic quantity of an arbitrary material, whose external magnetic field dependence at a given temperature gives a hysteresis loop that can be well-described by so-called the Langevin function. The Langevin function involves unknown parameters (physical constants) that unambigeously characterizes the material under investigation. Knowing these parameters makes possible to decide whether the investigated material is suitable or not for a particular application. The aim of this contribution is to present two possible approaches how to estimate the unknown parameters of the Langevin function by exploitation of special regression models. Both proposed algorithms are compared with each other and quantification of suitability of their usage is evaluated.

## 1. Úvod

Magnetické materiály díky svému velkému aplikačnímu potenciálu vzbuzují značný zájem odborné veřejnosti. Základní charakteristikou každého magnetického materiálu či nanomateriálu je magnetizace. Výstupem měření magnetizace je tzv. polní závislost magnetizace, která bývá obvykle označována jako hysterezní smyčka. Nanometrové magnetické materiály vykazují jeden ze zvláštních fenoménů, který je právě stěžejní pro jejich praktické využití. Tento jev se nazývá superparamagnetismus a jeho hysterezní smyčka prochází počátkem.

V roce 1905 francouzský fyzik Pierre Langevin odvodil pro hysterezní smyčku v superparamagnetickém stavu závislost danou tímto předpisem

$$y = l_1 \cdot \coth(l_2 \cdot x) - \frac{l_1}{l_2 \cdot x}, \tag{1}$$

kde parametry $l_1$ a $l_2$ jsou fyzikální konstanty charakterizující jednoznačně zkoumaný nanomateriál.

Naším cílem je tedy najít odhady parametrů této funkce pro data získaná z fyzikálního experimentu. K aproximaci použijeme regresní model, když nejprve nelineární funkci budeme aproximovat lineárním členem Taylorova rozvoje. V numerické části budeme aplikovat navržené algoritmy na datový soubor z měření nanočástic Gamma formy oxidu železitého, jejichž střední rozměr nabývá hodnoty přibližně 15 nm $N(= 450)$. Experimentální měření bylo provedeno v $n$ $(= 150)$ bodech vybraných z intervalu $-70000$ Oe až 70000 Oe. Každý bod byl měřen s replikacemi $r_1, \ldots, r_n$, kde $r_1 + \ldots + r_n = 450$. Vlastní měření uskutečnilo Nanocentrum Univerzity Palackého v Olomouci.

## 2. Linearizované regresní modely

Experiment budeme modelovat linearními regresní modely a to v prvním algoritmu modelem bez podmínek a ve druhém algoritmu modelem se systémem podmínek. Získané výsledky porovnáme.

### 2.1. Algoritmus 1 – regresní model nepřímého měření vektorového parametru bez podmínek

Experiment měření lze popsat pomocí nelineárního regresního modelu nepřímého měření vektorového parametru, ve tvaru

$$\mathbf{Y} \sim [\boldsymbol{\phi}(\boldsymbol{\Theta}), \boldsymbol{\Sigma}], \tag{2}$$

kde $\boldsymbol{\phi}(\boldsymbol{\Theta})$ je známá (v našem případě nelineární) funkce.

Abychom mohli přistoupit k výpočtu hodnot odhadů parametrů Langevinovy funkce, je potřebné provést linearizaci modelu (2). Je-li $\boldsymbol{\phi}(\boldsymbol{\Theta^0})$ známý vektor, lze model rozvinout do Taylorovy řady, v níž zanedbáme členy druhého a vyšších řádů. Po provedení linearizace budeme model psát ve tvaru

$$\overline{\mathbf{Y}} \sim_n \left(\mathbf{F}\boldsymbol{\Theta}, \sigma^2\boldsymbol{\Lambda}^{-1}\right), \tag{3}$$

kde $\boldsymbol{\Theta} = [l_1, l_2]'$ je vektor neznámých parametrů, $\sigma^2\boldsymbol{\Lambda}^{-1}$ je uvažovaná varianční matice a

$$\{\mathbf{F}\}_{i\cdot} = \frac{\partial\boldsymbol{\phi}(x_i, \boldsymbol{\Theta}^0)}{\partial\boldsymbol{\Theta}'} = \left(\frac{\partial\phi_i(x_i, \boldsymbol{\Theta}^0)}{\partial l_1}, \frac{\partial\phi_i(x_i, \boldsymbol{\Theta}^0)}{\partial l_2}\right), \tag{4}$$

je známá matice plánu, přičemž

$$\begin{aligned}
\frac{\partial\phi_i}{\partial l_1} &= \coth\left(l_2 \cdot x_i\right) - \frac{1}{l_2 \cdot x_i}, \\
\frac{\partial\phi_i}{\partial l_2} &= \frac{-l_1 \cdot x_i}{\sinh\left(l_2 \cdot x_i\right)^2} + \frac{l_1}{l_2^2 \cdot x_i}.
\end{aligned}$$

Nechť

$$\overline{\mathbf{Y}} = \begin{pmatrix} \frac{1}{r_1}\sum_{i=1}^{r_1} Y_i \\ \vdots \\ \frac{1}{r_n}\sum_{i=1}^{r_n} Y_i \end{pmatrix} \tag{5}$$

je observační vektor s varianční maticí ve tvaru

$$\mathrm{Var}\left(\overline{\mathbf{Y}}\right) = \sigma^2 \begin{pmatrix} \frac{1}{r_{i_1}} & & \\ & \ddots & \\ & & \frac{1}{r_{i_n}} \end{pmatrix} = \sigma^2\boldsymbol{\Lambda}^{-1}. \tag{6}$$

**Věta 1.** *Nechť je dán linearizovaný regresní model nepřímého měření vektorového parametru (3). Pak BLUE parametru* $\boldsymbol{\Theta}$ *je dán ve tvaru*

$$\widehat{\boldsymbol{\Theta}} = \left(\mathbf{F}'\boldsymbol{\Lambda}\mathbf{F}\right)^{-1}\mathbf{F}'\boldsymbol{\Lambda}\overline{\mathbf{Y}}, \tag{7}$$

*s varianční maticí*

$$\mathrm{Var}(\widehat{\boldsymbol{\Theta}}) = \sigma^2\left(\mathbf{F}'\boldsymbol{\Lambda}\mathbf{F}\right)^{-1}. \tag{8}$$

**Důkaz:** Viz [1].

Uvedeným postupem jsme získali následující odhady parametrů $l_1$ a $l_2$ Langevinovy funkce:

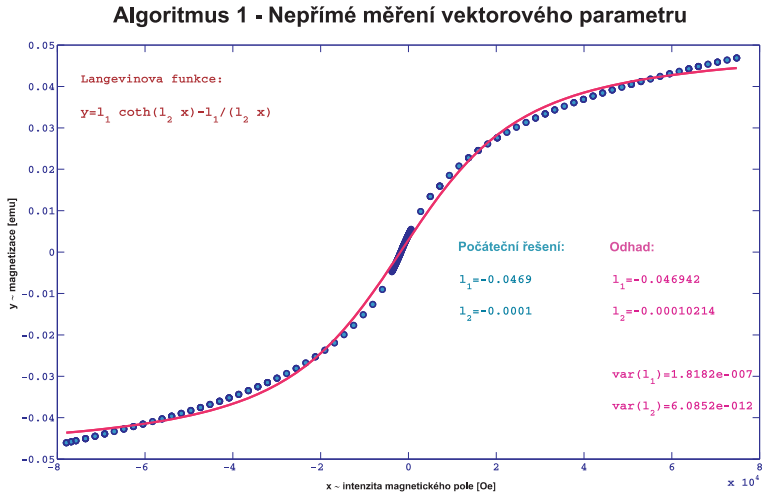
$$\widehat{\boldsymbol{\Theta}} = \begin{pmatrix} -0{,}04694215 \\ -0{,}00010214 \end{pmatrix}.$$

Variance těchto odhadů je potom

$$\mathrm{Var}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} (4{,}2640 \cdot 10^{-4})^2, & -9{,}2596 \cdot 10^{-10} \\ -9{,}2596 \cdot 10^{-10}, & (2{,}4668 \cdot 10^{-6})^2 \end{pmatrix}.$$

Pro výpočet bylo nutné znát jistý přibližný počáteční odhad, ten jsme získali pomocí Levenbergova–Marquardtova algoritmu.

Aproximace určená tímto algoritmem je znázorněna na obrázku.




## 2.2. Algoritmus 2 – regresní model přímého měření vektorového parametru se systémem podmínek typu II

Mějme měření hodnot $(Y_1, Y_2, \ldots Y_N)'$ vektorového parametru $\boldsymbol{\beta}$ v bodech $(x_1, x_2, \ldots x_n)'$ určených deterministicky. Hodnoty $x_i$, $i = 1, \ldots, n$ reprezentují stanovenou sílu vnějšího magnetického pole a hodnoty $Y_i$, $i = 1, \ldots, N$ představují měření odpovídající magnetizace materiálu. Hodnoty závislé proměnné mohou být vyjádřeny pomocí Langevinovy funkce takto:

$$Y_i = l_1 \cdot \coth(l_2 \cdot x_i) - \frac{l_1}{l_2 \cdot x_i}, \; i = 1, \ldots, n. \tag{9}$$

Nelineární model měření lze tedy popsat modelem

$$\overline{\mathbf{Y}} \sim N_n \left[ \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \boldsymbol{\Sigma} \right], \tag{10}$$

přičemž musí platit

$$\begin{aligned} g_i(\boldsymbol{\beta}, \boldsymbol{\Theta}) &= l_1 \cdot \coth(l_2 \cdot x_i) - \frac{l_1}{l_2 \cdot x_i} - \beta_i = 0\,, \\ i &= 1, \ldots, n, \end{aligned} \tag{11}$$

kde

$$\boldsymbol{\Theta} = [l_1, l_2]'\,.$$

Cílem bude nalézt odhady $\widehat{\widehat{\boldsymbol{\beta}}}$ skutečných hodnot $\boldsymbol{\beta}$ a dále odhady parametrů $\widehat{\widehat{\boldsymbol{\Theta}}} = [l_1,\, l_2]'$ vystupujících v Langevinově funkci.

Stejně, jako tomu bylo u modelu nepřímého měření vektorového parametru bez podmínek, je potřebné i zde provést linearizaci modelu pomocí Taylorova rozvoje.

Nelineární podmínky $\mathbf{g}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = (g_1(\boldsymbol{\beta}, \boldsymbol{\Theta}), \ldots, g_n(\boldsymbol{\beta}, \boldsymbol{\Theta}))' = 0$ lze po linearizaci a zanedbání vyšších členů psát pomocí Taylorova rozvoje v lineárním tvaru $\mathbf{B}\delta\boldsymbol{\beta} + \mathbf{G}\delta\boldsymbol{\Theta} + \mathbf{b} = \mathbf{0}$, kde $\mathbf{B} = \frac{\partial \mathbf{g}(\boldsymbol{\beta}^0, \boldsymbol{\Theta}^0)}{\partial \boldsymbol{\beta}'}$, $\mathbf{G} = \frac{\partial \mathbf{g}(\boldsymbol{\beta}^0, \boldsymbol{\Theta}^0)}{\partial \boldsymbol{\Theta}'}$, a $\mathbf{b} = \mathbf{g}(\boldsymbol{\beta}^0, \boldsymbol{\Theta}^0)$ v přibližném bodě $(\boldsymbol{\beta}^0, \boldsymbol{\Theta}^0)$.

Po určení příslušných parciálních derivací získáme matici $\mathbf{B}$ ve tvaru

$$\mathbf{B} = \begin{pmatrix} b_{11} & 0 & \ldots & 0 & 0 \\ 0 & b_{22} & \ldots & 0 & 0 \\ & & \ldots & & \\ 0 & 0 & \ldots & 0 & b_{nn} \end{pmatrix}, \text{ kde } b_{ii} = -1\,. \tag{12}$$

$$\mathbf{G} = \frac{\partial \mathbf{g}(\boldsymbol{\beta}^0, \boldsymbol{\Theta}^0)}{\partial \boldsymbol{\Theta}'} = \begin{pmatrix} \frac{\partial g_1}{\partial l_1} & \frac{\partial g_1}{\partial l_2} \\ \vdots & \vdots \\ \frac{\partial g_i}{\partial l_1} & \frac{\partial g_i}{\partial l_2} \\ \vdots & \vdots \\ \frac{\partial g_n}{\partial l_1} & \frac{\partial g_n}{\partial l_2} \end{pmatrix}, \tag{13}$$

kde $i = 1, \ldots, n$.

Nyní najdeme odhad parametrů v linearizovaném modelu. K tomuto účelu využijeme teorii lineárních statistických modelů uvedenou v [2], konkrétně model přímého měření vektorového parametru se systémem podmínek typu II.

**Definice 1.** *Model přímého měření s podmínkou II. typu na parametry 1. řádu má tvar*

$$\mathbf{Y} \sim_n (\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Lambda}^{-1}), \tag{14}$$

$$\mathbf{b} + \mathbf{B}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\Theta} = \mathbf{0}\,, \tag{15}$$

*kde $\boldsymbol{\beta} \in R^{k_1}, \boldsymbol{\Theta} \in R^{k_2}$ jsou neznámé.*
*Jestliže $h(\mathbf{X}_{(n,k)}) = k_1 < n$, $h(\mathbf{B}_{(q,k_1)}, \mathbf{G}_{(q,k_2)}) = q < k_1 + k_2, h(\mathbf{G}) = k_2 < q$ a matice $\boldsymbol{\Lambda}^{-1}$ je pozitivně definitní, potom model nazýváme regulárním.*

Dále budeme uvažovat pouze regulární model.

**Věta 2.** *BLUE vektoru* $\begin{pmatrix} \widehat{\widehat{\boldsymbol{\beta}}} \\ \widehat{\widehat{\boldsymbol{\Theta}}} \end{pmatrix}$ *jsou dány vztahy*

$$\widehat{\widehat{\boldsymbol{\beta}}} = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{B}'\left[\mathbf{T}^{-1} - \mathbf{T}^{-1}\mathbf{G}(\mathbf{G}'\mathbf{T}^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{T}^{-1}\right](\mathbf{b} + \mathbf{B}\widehat{\boldsymbol{\beta}}) \quad (16)$$

$$\widehat{\widehat{\boldsymbol{\Theta}}} = -(\mathbf{G}'\mathbf{T}^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{T}^{-1}(\mathbf{b} + \mathbf{B}\widehat{\boldsymbol{\beta}}), \quad (17)$$

*kde*

$$\mathbf{T} = \mathbf{B}(\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{B}' + \mathbf{G}\mathbf{G}', \quad (18)$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Lambda}\mathbf{Y}. \quad (19)$$

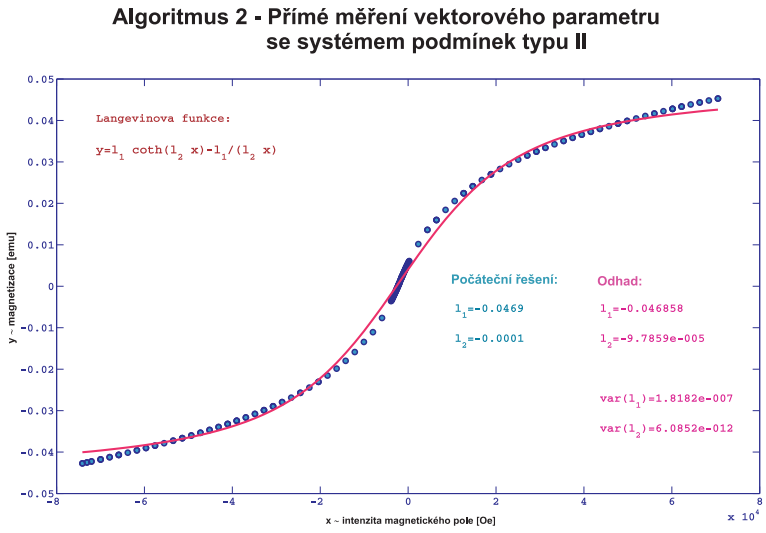
*($\widehat{\boldsymbol{\beta}}$ je odhad nerespektující podmínku týkající se parametrů $\boldsymbol{\beta}, \boldsymbol{\Theta}$).*

**Důkaz:** Viz [3], ekvivalentní odhady v maticovém zápisu také v [2].

**Věta 3.** *Kovarianční matice odhadů $\widehat{\widehat{\boldsymbol{\beta}}}$ a $\widehat{\widehat{\boldsymbol{\Theta}}}$ jsou*

$$\begin{aligned}\mathrm{Var}(\widehat{\widehat{\boldsymbol{\beta}}}) &= \sigma^2\left\{\mathbf{I} - (\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{B}'\left[\mathbf{T}^{-1} - \mathbf{T}^{-1}\mathbf{G}(\mathbf{G}'\mathbf{T}^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{T}^{-1}\right]\mathbf{B}\right\}\cdot \\ &\quad (\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}\left\{\mathbf{I} - (\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{B}'\cdot\right. \\ &\quad \left.\cdot\left[\mathbf{T}^{-1} - \mathbf{T}^{-1}\mathbf{G}(\mathbf{G}'\mathbf{T}^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{T}^{-1}\right]\mathbf{B}'\right\}, \\ \mathrm{Var}(\widehat{\widehat{\boldsymbol{\Theta}}}) &= \sigma^2\left\{(\mathbf{G}'\mathbf{T}^{-1}\mathbf{G})^{-1} - \mathbf{I}\right\}.\end{aligned}$$

**Důkaz:** Viz [3]




Na obrázku vidíme výsledky dosažené pomocí algoritmu 2 dle uvedeného modelu. Při výpočtu bylo třeba mít k dispozici vhodné počáteční řešení. To jsme získali pomocí Levenbergova–Marquardtova algoritmu.

Numericky určené odhady parametrů $l_1$ a $l_2$ jsou:

$$\widehat{\widehat{\boldsymbol{\Theta}}} = \begin{pmatrix} -0{,}043567 \\ -0{,}00010697 \end{pmatrix}$$

a variační matice těchto odhadů je

$$\mathrm{Var}(\widehat{\widehat{\boldsymbol{\Theta}}}) \;=\; \begin{pmatrix} (4{,}2640\cdot 10^{-4})^2 & -9{,}2596\cdot 10^{-10} \\ -9{,}2596\cdot 10^{-10} & (2{,}4668\cdot 10^{-6})^2 \end{pmatrix}.$$

## 3. Porovnání vhodnosti modelů

Pro porovnání kvality určených odhadů ze dvou studovaných nelineárních regresních modelů použijeme jako kritérium reziduální součet čtverců RSČ. Z tabulky vyplývá, že je vhodnější ten model, pro nějž nabývá tato statistika nižší hodnoty.

Vzhledem k tomu, že počet regresních parametrů významně ovlivňuje výslednou hodnotu reziduálního součtu čtverců byl pro srovnání použít i tzv. reziduální rozptyl $s_e^2$ a index determinace $I^2$.

| Parametr | Model bez podmínky | Model se systémem podm. typu II |
|---|---|---|
| Odhad $l_1$ | $-0{,}04694215$ | $-0{,}04685784$ |
| Odhad $l_2$ | $-0{,}00010214$ | $-0{,}00009785$ |
| Var $(l_1)$ | $(4{,}2640\cdot 10^{-4})^2$ | $(4{,}2640\cdot 10^{-4})^2$ |
| Var $(l_2)$ | $(2{,}4668\cdot 10^{-6})^2$ | $(2{,}4668\cdot 10^{-6})^2$ |
| RSČ | $5{,}9834\cdot 10^{-4}$ | $6{,}6413\cdot 10^{-4}$ |
| $s_e^2$ | $4{,}0429\cdot 10^{-6}$ | $4{,}4874\cdot 10^{-6}$ |
| $I^2$ | 0,997083 | 0,966263 |

Tabulka 1: Aproximace Langevinovy funkce — srovnání výsledků pro oba regresní modely

## 4. Závěr

V práci jsme se zabývali problematikou hledání odhadů neznámých parametrů Langevinovy funkce. Výsledky dosažené studovanými regresními modely zcela vystihují charakter měřených dat. V konkrétním případě model bez podmínky o něco lépe aproximoval daná data pomocí zvoleného funkčního předpisu. Tento závěr však nelze brát jako pravidlo, neboť pro jiný vzorek by mohla dopadnout situace zcela naopak. Shoda výsledků při použití obou dvou různých lineárních regresních modelů signalizuje, že proces linearizace ovlivnil výsledky pouze minimálně.

## 5. Literatura


1. Jiří Anděl: Základy matematické statistiky, Univerzita Karlova v Praze, Praha 2002
2. Lubomír Kubáček, Ludmila Kubáčková: Statistika a metrologie, UP Olomouc, 2000
3. Jaroslav Marek, Pavel Tuček: Statistické algoritmy pro aproximaci Lorentzovy funkce, KMAaAM PřF UP Olomouc, Preprint 9/2009, http://mant.upol.cz/cs/preprinty.asp



**Adresa korespondujícího autora**
Jaroslav Marek
Katedra matematické analýzy a aplikací matematiky, PřF UP Olomouc
Tomkova 40, 779 00 Olomouc, tel. (+420)585634606
*marek@inf.upol.cz*

# Numerical prediction of strains and stresses in early-age massive concrete structures

Vala Jiří, Šťastník Stanislav, Kozák Vladislav

**Abstract:** Strains and stresses in massive concrete structures, in addition to those caused by exterior mechanical loads, are results of rather complicated non-deterministic physical and chemical processes in fresh concrete mixtures. Their numerical prediction at the macro-scale level requires the non-trivial physical analysis based on the thermodynamic principles, making use of micro-structural information from both theoretical and experimental research. The paper demonstrates the derivation of a corresponding nonlinear system of macroscopic equations of evolution with certain effective material characteristics and suggests the algorithm of its numerical analysis.

**Key words:** massive concrete, early-age volume changes, computational modelling, thermo-dynamic principles, more-scale analysis.

## 1. Physical and technical background

Concrete is the most commonly used material of construction throughout the world and there is a growing need to understand its behaviour from the properties of its components, particularly the cracking and fracture behaviour. The design and realization of large building actions in last years needs still more modern building materials and technologies with the expected results of relatively inexpensive constructions with optimal user properties. However, the preparation of such actions needs much deeper study of material properties, coming not only from the traditional phenomenological description, but from the deeper physical information about their microstructure, with the perspectives of their optimal usage in a construction, whose engineering application is a priori known. Because of the lack of practical experience and of the high cost of experiments with real structures the significance of the following activities is increasing: both of the development of numerical methods and computational tools for modelling and simulation of behaviour of constructions (including such situations where the formal existence of a mathematical problem has not been verified yet) and of the modern laboratory approaches and observations on real objects, whose aim is the reliable and sufficiently exact identification of material properties.

One of the still unclosed problems of modern civil engineering, important for the whole society, is the optimization of preparation and application of silicate mixtures for concrete structures with the intention to constitute, as a result of physical and chemical processes in fresh mixtures, such stress distributions that could help to decrease the stresses caused by operational loads or loads from stochastic climatic influences, respectively. The model defined here relies upon numerical analysis performed on a relevant representative volume element (RVE), derived from an experimental analysis. In the design of constructions the influence of time-dependent strains and support changes, caused by shape modification and shrinkage, and the consequences of stiffness modification in fractured zones cannot be neglected. The static quantities, related to external loads, are usually evaluated without taking the presence of fractures into account (although some stiffness decrease due to presence of fractures could be considered).

The simple design of such constructions is impossible because of the complicated composite material structure, its creep deformability, accompanied with the formation of microstructural fractured zones, eventually later of visible macroscopic cracks, of the need of appropriate reinforcement to eliminate tension stresses and of the sensibility of the whole process to outer climatic conditions. The model defined here relies upon numerical analyses

performed on a relevant representative volume element (RVE), derived from an experimental analysis. The dominant physical processes (and hidden chemical ones) are:

a) the reversible elastic deformation,
b) the viscous material flow,
c) the volume changes, unlike a) and b) independent of external loads.

For the above mentioned stresses the decisive component is c), influenced by:

c1) autogenous volume changes, driven by chemical shrinkage of cement particles,
c2) subsequent thermal expansion,
c3) drying, connected with water loss in the environment,
c4) later carbonation.

The methods of volume changes measurement during the stiffness increase in silicate mixtures, namely the method of flexible rubber membrane, the weight (reduced buoyancy) method, etc., are described in details in [12]. The observation if volume changes in case of real constructions needs moreover the development of non-contact methods, based on the image processing from photographs at sub-pixel level; the overview of possible approaches is contained in [27].

The prediction of shrinkage and creep of concrete structures forms a rather complicated class of problems with various aspects accented in the literature: e.g. [3] refers to:

a) the formulation of activation energy for the corruption of bindings between particular components of a composite,
b) the diffusion theory in the process of drying and shrinkage,
c) the modelling of generation of fractured zones on base of the shrinkage mechanism,
d) the analysis of the stiffness growth as a special phase transformation,
e) the microstructural mechanism leading to the pre-stress status that cannot be explained from the increasing volume of hydration products only,
f) the development of macroscopic cracks in time.

The micromechanical substance of the viscous concrete flow and of the potential creation of fractured zones (e. g. [2] and [22]) should be respected. However, the simplest theories evaluate the deformation in concrete using the relation, presented e. g. in [16]:

$$\varepsilon = \varepsilon_p (1 - V)^m ,$$

where $\varepsilon$ denotes the relative strain, $\varepsilon_p$ the relative strain in mortar and V the volume ratio of aggregated particles, the exponent $m$ (constant for a given system) being usually some number between 0 and 2, but in general as a complicated function of elastic modules and aggregated particles, influenced also by their mutual position, size and other factors. The effort to verify the model with effective material characteristics leads in [16] to the description of periodic material structures and to the derivation of constitutive relations using the homogenization approach; moreover the approaches with the probabilistic description of particle positions are available (cf. [15]) as well as the correct mathematical homogenization techniques (as the applications of the least squares method but also the two-scale convergence in Lebesgue and Sobolev spaces and their generalizations, etc.).

The complex approach to the study of physical processes, active in stiffening material, declares to come out from the classical conservation laws, namely of mass, inertia and energy, as usual in the computational fluid dynamics. Since such complex models are very complicated (physically, mathematically and computationally), involving (among others) heat transfer driven by air flow, modified by (only partially reversible) moisture propagation in pores, all real computational tools implement substantial simplifications. The formulations of [24] evaluate the entropy production in the system: the constitutive equations for the heat and

moisture transfer in the stiffening silicate mixture are derived from the Onsager reciprocity relations and from the Gibbs-Duhem conditions where the phenomenological coefficients have to be set from experiments (discussed e. g. in [25]). From the mathematical point of view, the resulting relations are certain systems of partial differential equations of evolution (or their integral equivalents, respectively) with prescribed initial and boundary conditions, containing particular fields of unknown quantities, namely temperature, moisture content, strains and stresses. Only one equation of heat conduction, with some inner time-dependent heat sources q from binder hydration, is being obtained in the most drastic simplification:

$$c\rho\,\dot{T} + \operatorname{div}(\lambda \operatorname{grad} T) + q = 0\ ; \tag{1}$$

here $T$ is the unknown absolute temperature, the dot symbol refers to the partial derivative with respect to time and the above mentioned phenomenological coefficients are reduced to the heat conduction factor $\lambda$ and to the coefficient $c\rho$ where $\rho$ is the material density and $c$ the heat capacity. Analytical or simple semi-analytical solutions (e.g. those applying the Fourier analysis) of such evolution equations and systems are available only in very special cases: so the algorithm in [19] refers to the linearized models of Luikov type, and therefore offers no possibility to include strongly nonlinear dependencies as sorption isotherms, decisive for the temperature redistribution (and consequently for the slow stress and strain evolution) caused by the propagation of various phases of moisture, including their phase changes (see [17]).

The computational model of strain and stress evolution in a reinforced silicate composite should respect the following internal and external influences (cf. [11] and [28]):

1) internal hydration heat, generated by the hydration hydraulic processes,
2) ambient temperature variation, connected with ambient humidity variation (natural or artificial ones),
3) external mechanical loads.

In the period of intense hydration the thermal deformation (a), accompanied by the autogenous shrinkage (b), is dominant. In the later period of slow hydration the role of (a) decreases, but the effect of the carbonation (c) has to be taken into account. The external mechanical loads cause the elastic and creep deformation (d) (creep especially in the early age), the external temperature changes force e) additional thermal deformation, modified by the drying shrinkage and swelling (f). The presence of reinforcement causes the strain and stress redistribution, conditioned by the cohesion of these components – cf. [9], [14]. The experimental method of structural monitoring of such hybrid specimens (from contactless photogrammetric analysis, fibre optic sensors, etc.) gives typically total deformation, not particular contributions (a)-(f). However, it is possible to analyze namely the evolution of hydration heat and temperature in time and also the corresponding local stiffness increase; this enables the calibration of macroscopic (effective) concrete parameters as Young modulus, strength, creep ratio, etc., related to the viscoelastic constitutive relations, implemented to a coupled system of partial differential equations of evolution (and to its integral equivalent), including initial and boundary conditions, for unknown strains (and stresses), temperature and humidity, coming from the conservation principles of classical thermodynamics (see [13]).

A phenomenological approach is traditionally used in modelling of simultaneous hygro-thermal and mechanical behaviour of silicate composites – cf. [3]: if no distinction between different phases of moisture is made then phase changes cannot be considered in a proper way. However, phase transition and chemical reactions are of importance when performance of early-age massive concrete structures is studied; more important changes of material properties as density, porosity, permeability, compressive strength, etc. during concrete hardening should be taken into account. In a phenomenological description their effect on material behaviour is lumped together to some model parameters which must be identified by long-lasting tests in the whole range of model applicability. On the contrary, the mechanistic

approach of [18], [5], [6] makes it possible to consider such effects explicitly because they appear directly in the model equations. Nevertheless, in is nearly impossible to apply a purely mechanistic approach for such complex problem as maturing of concrete: some elements of phenomenological description are necessary to avoid the detailed analysis. Namely [18] distinguishes between for length scales, characterized as:

i) anhydrous-cement scale (typical length of a representative volume element from $10^{-8}$ to $10^{-6}$ m),
ii) cement-paste scale (from $10^{-6}$ to $10^{-4}$ m),
iii) mortar scale (about from $10^{-2}$ m),
iv) macroscale (about $10^{-1}$);

in more details i) is decomposed into 3 detailed scales where the qualitative estimate of activity of four main clinker phases, water and air requires the detailed micromechanical evaluation of all corresponding chemical reactions, consequently also the implementation of appropriate heterogeneous multiscale methods by [26], bridging models of very different nature from molecular dynamics to continuum mechanics (cf. Chap. 5, 6 and 7 in [21]), containing a lot of problems also in mathematical solvability, convergence of numerical algorithms, etc., forcing the development of non-classical multiscale computational techniques.

The approach presented in [10] applies certain mechanistic-type method to obtain the governing equations only, using the averaging hybrid mixture theory: the developments starts at the micro-scale and balance equations for phases and interfaces are introduced at this level and then averaged for obtaining macroscopic balance equations. Four phases are distinguished: solid skeleton, liquid water, vapour and dry air, whose densities are considered (under the passive air assumption) as constants; the whole hygro-thermo-chemo-mechanical process is then studied as the time evolution of capillary pressure, gas pressure, temperature and displacement of points related to the reference (initial) configuration, driven by balance equations of classical thermodynamics and conditioned by corresponding constitutive laws. The detailed geometrical analysis in [20] offers the possibility to extend such considerations beyond the assumption of small deformations and involve some elements of fracture mechanics. The thermo-mechanical analysis of balance of mass, (linear and angular) momentum and energy [23] for computational HAM ("heat, air and moisture") models in civil engineering is able to be extended to a complex model, including the mass source or solid skeleton related to the cement hydration process (and corresponding sink of liquid water mass), as well as the vapour mass source caused by the liquid water evaporation or desorption (using micromechanical arguments from the theory of porous media by [7]), similarly to Chap. 18 in [4].

## 2. Model description

The basic idea of the simulation of early-age strain and stress time redistributions in a massive concrete structure, presented in this paper, refers at most to [10], generalized in several directions, using the thermodynamic laws in the form [23]. The original research, motivated by some requirements from the building practice (namely from the large building corporation OHL-ŽS Brno), is performed at the Faculty of Civil Engineering of the Brno University of Technology; because of the complexity of physical formulations the collaboration with the Institute of Physics of Materials of the Academy of Sciences of the Czech Republic is needed, as well as the discussions with specialists from TU Vienna and CTU Prague.

We shall assume that a deformable body occupies certain domain in the 3-dimensional Euclidean space, supplied by the Cartesian coordinate system $x = (x_1, x_2, x_3)$, in the time $t$, increasing from $0$. We shall work with partial derivatives of scalar variables $\psi$ with respect

to $t$, expressed briefly as $\psi^{\cdot} = \partial\psi/\partial t$, and with respect to $x_i$ for $i \in \{1, 2, 3\}$, expressed briefly as $\psi_{,i} = \partial\psi/\partial x_i$. We shall consider four material phases $\varepsilon \in \{s, w, v, a\}$: solid material ($s$), liquid water ($w$), water vapour ($v$) and dry air ($a$) and make use of the physical balance laws, namely of the balance of mass, (linear and angular) momentum and energy; formally the same can be done both for one RVE and for the whole massive structure. If $\omega^{\varepsilon}$ is a source corresponding to a scalar quantity $\psi^{\varepsilon}$, the conservation of a scalar quantity $\psi^{\varepsilon}$ in [23], p. 33, reads

$$\psi^{\cdot\varepsilon} + (\psi^{\varepsilon} v_i^{\varepsilon})_{,i} = \omega^{\varepsilon}. \tag{2}$$

The velocities $v_i^{\varepsilon}$ are time derivatives of corresponding displacements $u_i^{\varepsilon}$, i. e.

$$v_i^{\varepsilon}(x,t) = \dot{u}_i^{\varepsilon}(x,t) + u_{i,j}^{\varepsilon}(x,t) u_j^{\varepsilon}(x,t), \tag{3}$$

from the initial reference configuration of a deformable body in the following sense: if $x_i^{\varepsilon 0}$ refers to the position of such point for $t = 0$ and $x_i^{\varepsilon}$ for general $t$ then $u_i^{\varepsilon} = x_i^{\varepsilon} - \overline{x}_i^{\varepsilon}$; the dependence of geometrical and physical quantities on $x$ and $t$ is usually not emphasized explicitly. Moreover, the strain characteristics, needed in strain-stress constitutive relations, can be derived from Jacobi matrices $J^{\varepsilon}$ for particular phases, compound from elements

$$J_{ij}^{\varepsilon} = \delta_{ij} + \frac{\partial u_i^{\varepsilon}}{\partial x_j^{\varepsilon 0}}$$

where $\delta$ is reserved for Kronecker symbols (everywhere in this paper) and $j$ is an arbitrary index from $\{1, 2, 3\}$, similarly to $i$; later $i$ and $j$ will be also used as sum indices in sense of the Einstein summation rule. Consequently all corresponding stresses $\sigma_{ij}^{\varepsilon}$ can be expressed (typically, if no linearization procedure is applied, in a rather complicated form) using $u_i$, $v_i$, etc.; for more details see [20], p. 140. For the analysis of the accelerations $a_i$ as time derivatives of velocities, $u$ is allowed to be replaced by $v$ and $v$ by $a$ in (3). In addition to $u$, the unknown fields in our model will be the (absolute) temperature $T$, the total gas pressure $p^g$ and the capillary pressure $p^c$; in terms of corresponding pressures for particular phases we have $p^g = p^v + p^a$ and $p^c = p^g - p^w$ (cf. the Dalton law in [4], p. 111).

At the RVE level, for some physical quantities $\psi$ related to a phase $\varepsilon$, we must distinguish between their resulting averaged values $\psi_{\varepsilon}$ and their intrinsic averaged values $\psi^{\varepsilon}$. The needed relation $\psi_{\varepsilon} = \eta^{\varepsilon}\psi^{\varepsilon}$ contains the volume fraction $\eta^{\varepsilon}$, a function of the material porosity $n$ and the saturation $S$; both $n$ and $S$ are varying substantially during the hydration process. We have $\eta^{s} = 1 - n$, $\eta^{w} = nS$ and $\eta^{v} = \eta^{a} = n(1 - S)$ for the particular phases.

The *mass balance* works with $\psi^{\varepsilon} = \rho_{\varepsilon}$ (for any $\varepsilon$) and with $\omega^{s} = -m_w$, $\omega^{w} = m_w - m_v$, $\omega^{w} = m_v$ and $\omega_a = 0$ where $m_w$ means the rate of (usually increasing) mass of skeleton (and corresponding sink of liquid water mass) and $m_v$ means the analogous rate for vapour mass caused by (liquid) water evaporation or desorption. Both $m_w$ and $m_v$ depend evidently (in a complicated way) on $T$, $p^g$, $p^c$, $n$, $S$, water / cement ratio, etc. The direct reliable determination of $m_w$ and $m_v$ from the analysis of chemical reaction occurring in the irreversible process of cement hydration is impossible in practice; thus the hydration kinetics has to be described using some empirical parameter(s). The so-called hydration degree $\Gamma$, a normalized time-dependent variable with values between $0$ and $1$, is introduced in [10], p.

308; its evaluation requires always to solve an auxiliary evolution problem for certain nonlinear ordinary differential equation, whose motivation comes from the micro- and macro-scale thermodynamics of chemical transformations (applying the molar Gibbs energy of unhydrated and hydrated cement and water).

The *momentum balance* equations should contain both linear and angular one. Since all phases are considered microscopically non-polar (as usual in the theory of Boltzmann continuum), this angular momentum balance forces only the symmetry for the partial Cauchy stress tensor $\tau$, i. e. $\tau_{ij} = \tau_{ji}$ for each real matrix $\tau$ generated by such stress components. The formulation of the linear momentum balance in 3 directions is much more difficult. The first step is to introduce $w_i^\varepsilon = \rho_\varepsilon v_i^\varepsilon$ and to choose, step by step, $\psi^\varepsilon = w_i^\varepsilon$. The second step, based on the setting

$$\sigma_{ij}^\varepsilon = \tau_{ij}^\varepsilon \delta^{s\varepsilon} - \delta_{ij} p_\varepsilon \ , \qquad \omega^\varepsilon = \sigma_{ij,j}^\varepsilon + \rho_\varepsilon (g_i - a_i + t_i^\varepsilon) \ , \tag{4}$$

with the gravity accelerations $g_i$ and the additional (time variable) accelerations $t_i^\varepsilon$ due to interactions with other phases needs more detailed explanation. The first relation (4) defines the total Cauchy stress $\sigma^\varepsilon$; it is expressed as a symmetrical matrix consisting of components $\sigma_{ij}^\varepsilon$ again. Let us remark that in the (hypothetical) case $v_i = 0$ and $t_i^\varepsilon = 0$ the whole formulation (4) degenerates to the well-known Cauchy (static) equilibrium condition $0 = \sigma_{ij,j}^\varepsilon + \rho_\varepsilon g_i$ only. In practical applications the matrices $\sigma^\varepsilon$ (for particular $\varepsilon$) may be not separable; thus the effective total Cauchy stress

$$\sigma_{ij} = \tau_{ij} - \delta_{ij} \sum_{\varepsilon \in \{w,v,a\}} p_\varepsilon$$

is useful, too. The constitutive relationship $\sigma(u, v, ...)$ for the (linearized) case of small deformation and viscoelastic (creep) material behaviour is studied in [10], p. 343; its generalization to finite deformation needs the (rather formally complicated) so-called kinematic equations (which means the proper description of evolution of geometrical characteristics) by [20], p. 135. For the pressure components, the dependencies $p^a(\rho^a, T)$ and $p^v(\rho^v, T)$ by [10], p. 314 (compatible with the Clapeyron state equation), are available. The setting of $t_i^\varepsilon$ is possible from the Darcy law (cf. [7], p. 25): for all phase identifiers $\varepsilon \neq s$ (thus $\varepsilon \in \{v, w, a\}$) there holds

$$\mu^\varepsilon \rho_\varepsilon (v_i^\varepsilon - v_i^s) = K_{ij}^\varepsilon (\rho_\varepsilon (g_j - a_i + t_j^\varepsilon) - p_{\varepsilon,j}) \ , \tag{5}$$

where the new material characteristics occur: $\mu^\varepsilon$ is the dynamical viscosity and $K_{ij}^\varepsilon$ refer to the elements of the permeability matrix, depending (in general) on $\rho_\varepsilon$ again. Let us notice that various derivatives of $n$ and $S$ are hidden in (5), namely in the term $p_{\varepsilon,j}$; thus the proper (both theoretical and experimental) study of dependencies of the types $n(J, \Gamma)$ and $S(p^c, T)$ cannot be avoided.

Let $q_i^\varepsilon$ denote the internal heat fluxes. The Fourier law

$$q_i^\varepsilon = -\lambda_{ij}^\varepsilon T_{,j}$$

with the heat thermal conductivity components $\lambda_{ij}^\varepsilon$ enables us (formally) to express them using $T$ only; nevertheless, in general $\lambda_{ij}^\varepsilon$ are functions of $T$ and $p^c$, as discussed in [4], p. 42. The *energy balance* comes similarly to the momentum balance from the choices

$$\psi^{\varepsilon} = \frac{1}{2} w_i v_i + \rho_{\varepsilon} \kappa^{\varepsilon}, \qquad \omega^{\varepsilon} = (\sigma^{\varepsilon}_{ij,j} + q_i^{\varepsilon})_{,i} + \rho_{\varepsilon}(g_i - a_i + t_i^{\varepsilon}) + \tilde{\omega}^{\varepsilon},$$

where $\kappa^{\varepsilon}$ is usually defined as $\kappa = c^{\varepsilon} T$ for the thermal capacity $c^{\varepsilon}$, in general a function of $T$ and $p^c$ again (cf. (1) for one very simple homogeneous isotropic case); applying the same arguments as in the mass balance considerations, $\tilde{\omega}^{s} = -m_w h_w$, $\tilde{\omega}^{w} = m_w h_w - m_v h_v$, $\tilde{\omega}^{v} = m_v h_v$ and $\tilde{\omega}^{a} = 0$ and two additional characteristics are the specific enthalpy of hydration $h_w$ and the specific enthalpy of evaporation $h_v$. To determine all $v_i^{\varepsilon}$, following [4], p. 138, in addition to the Fourier heat flow we cannot neglect also the Fick diffusion: for the diffusion fluxes $r_i^{\varepsilon}$ with $\varepsilon \in \{v, w\}$ (moreover $r_i^a + r_i^v = 0$) we have

$$r_i^{\varepsilon} = -D_{ij} \eta^{\varepsilon}_{,j}, \qquad r_i^{\varepsilon} = \rho_{\varepsilon}(v_i^{\varepsilon} - v_i^{s});$$

the identification of the diffusive characteristics $D_{ij}$ (analogous to $\lambda_{ij}$), is studied in [7], p. 64, applying the microstructurally motivated arguments on (quasi)periodic homogenization.

## 3. Computational techniques

For the simultaneous heat and moisture transfer (eventually with contaminants, too) in the porous environment (although the mathematical theory is not closed) there exists a lot of computational algorithms, including software implementations. In the Central Europe the most used software packages seem to be WUFI (Fraunhofer Institut Holzkirchen) and DELPHIN (TU Dresden), with the insufficient support of the proper analysis of volume changes; the detailed overview of software of this type can be found in [22]. The large commercial software systems as ANSYS, ABAQUS, etc., offer usually only weak support of specific properties of silicate composites (although e.g. ANSYS has been applied in [16] to the homogenization of their model periodic structures). For the development of original software, the most user-friendly is the environment MATLAB / FEMLAB (COMSOL). In the terminology of numerical mathematics the above sketched software applies the standard modern variational methods (especially the finite element method and the finite volume method for a fixed time step, or the method of lines, the Rothe method of discretization in time, and the method of characteristics, respectively) for the construction of certain sequences of approximate solutions of problems, making use (because of the presence of nonlinearities) various iterative algorithms. For the calculations of strain and stress distributions in concrete structures at various stages of their existence (including the concrete / reinforcement cooperation, the prediction of fracture and the behaviour under extreme loads) the software ATHENA is available, open (see [8]) to the enrichment by further physical processes and corresponding constitutive relations. Nevertheless, no software for the analysis of such complex more-scale problems, as discussed above, is known to the authors; thus is development seems to be reasonable. In this paper we shall only sketch the main ideas, omitting all technical details, the examples of computational results and their discussion, etc.

To simulate the redistribution of strains and stresses, we have to solve the initial-value and boundary-value problem, from the mathematical point of view represented by a system of partial differential equations of evolution of $u$ (and corresponding time derivatives), $T$, $p^c$ and $p^g$. Clearly the natural initial condition is the prescription of all values of these quantities for $t = 0$. However, the quite general formulation includes (as its special cases) some open physical and mathematical problems, e. g. in the solvability (i. e. in the verification of existence and uniquness of a solution in some reasonable function space) of the non-stationary equations of Navier-Stokes type. Therefore, similarly to all above referenced software packages, some further simplifications are necessary. Following [20], p. 140, we can

suppose that at the micro-level the porous medium is constituted of incompressible solid and water constituents, while gas in considered compressible, but the passive air assumption (on constant $p^g$) enables us to eliminate all spatial and time variations of $\rho^s$, $\rho^w$, $\rho^v$ and $\rho^a$ (not of $\rho_s$, $\rho_w$, $\rho_v$ or $\rho_a$). From other assumptions, presented in [20], p. 140, the assumption on quasi-static processes, neglecting the derivatives of $v_i$ (including $a_i$) may be acceptable, unlike the assumption on purely isothermal processes; in [10], p. 303, thanks to linearized strains, some other (in some situations unpleasant) simplifications are involved a priori.

From the practical reason, some boundary conditions are useful to be formulated for phased-independent quantities, as the total Cauchy stress tensor $\sigma$. Let $\nu = (\nu_1, \nu_2, \nu_3)$ be the unit normal vector (preserving an appropriate orientation) for some computational (sub)domain. Then, following [10], p. 316, we can have prescribed values of unknown fields on the boundary (or its part), i. e. the Dirichlet boundary conditions, or more general boundary conditions of Cauchy, Neumann, Robin, etc. types. The most frequent choices are

$$\sigma_{ij}\nu_j = \overline{t}_i$$

with imposed tractions $\overline{t}_i$,

$$\left(\rho_a(v_i^a - v_i^s) + r_i^a\right)\nu_i = \overline{r}^a$$

with imposed air fluxes $\overline{r}^a$,

$$\left(\rho_w(v_i^w - v_i^s) + r_i^w + \rho_v(v_i^v - v_i^s) + r_i^v\right)\nu_i = \overline{r}^w + \overline{r}^v + \beta(\rho_w)$$

with imposed liquid water and vapour fluxes $\overline{r}^w$, $\overline{r}^v$ and some mass exchange function $\beta$ or

$$\left(\rho_w(v_i^w - v_i^s)h_v - \lambda_{ij}T_{,j}\right)\nu_i = \overline{q} + \alpha(T)$$

with imposed heat fluxes $\overline{q}$ and some heat exchange function $\alpha$.

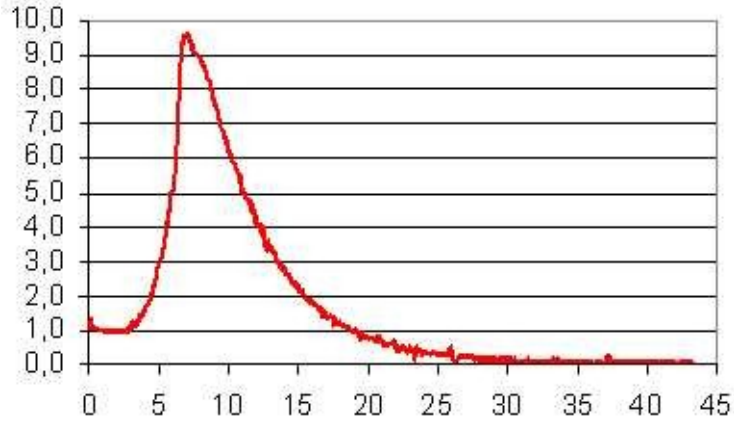



***Figure 1: Experimental intensity of hydration heat Q (W/kg) from 0 to 45 h***

We can incorporate the reinforcement as one additional phase into the model in the case of reinforced concrete; to formulate the transfer condition on the concrete / reinforcement interface, we need more quantified information about the cohesion between both phases. The same is true for the mechanism of evolution of potential damage, especially if some non-negligible tensions, not eliminated by the reinforcement system, occur in the massive concrete structure. The development of the complete original software simulation tools covering all scales is not realistic because of a small number of researchers and limited financial support. Under such conditions the optimal all simulation experiments are designed as combinations of some original functions, created e. g. in the MATLAB environment, with the software tools

mentioned above. The nonlinearity of all equations, at least in the material characteristics, dependent on the a priori unknown fields, forces both the design of iterative procedures and weak formulations, involving some types of boundary conditions in a natural way and friendly both to the finite or volume element techniques and to the discrete element methods (and similar ones) in the sense of [21], p. 344. The software scale bridging can be done properly, as described in [21], p. 386; however, more questions than complete answers are still in the macro- and microscopic identification problems (even in such seemingly simple ones, as the simultaneous determination of some effective heat capacity and thermal conductivity in a wet concrete mixture), as well as in the convergence of both theoretical and approximate numerical solutions of microscale formulations to macroscale ones, applying advanced homogenization techniques.

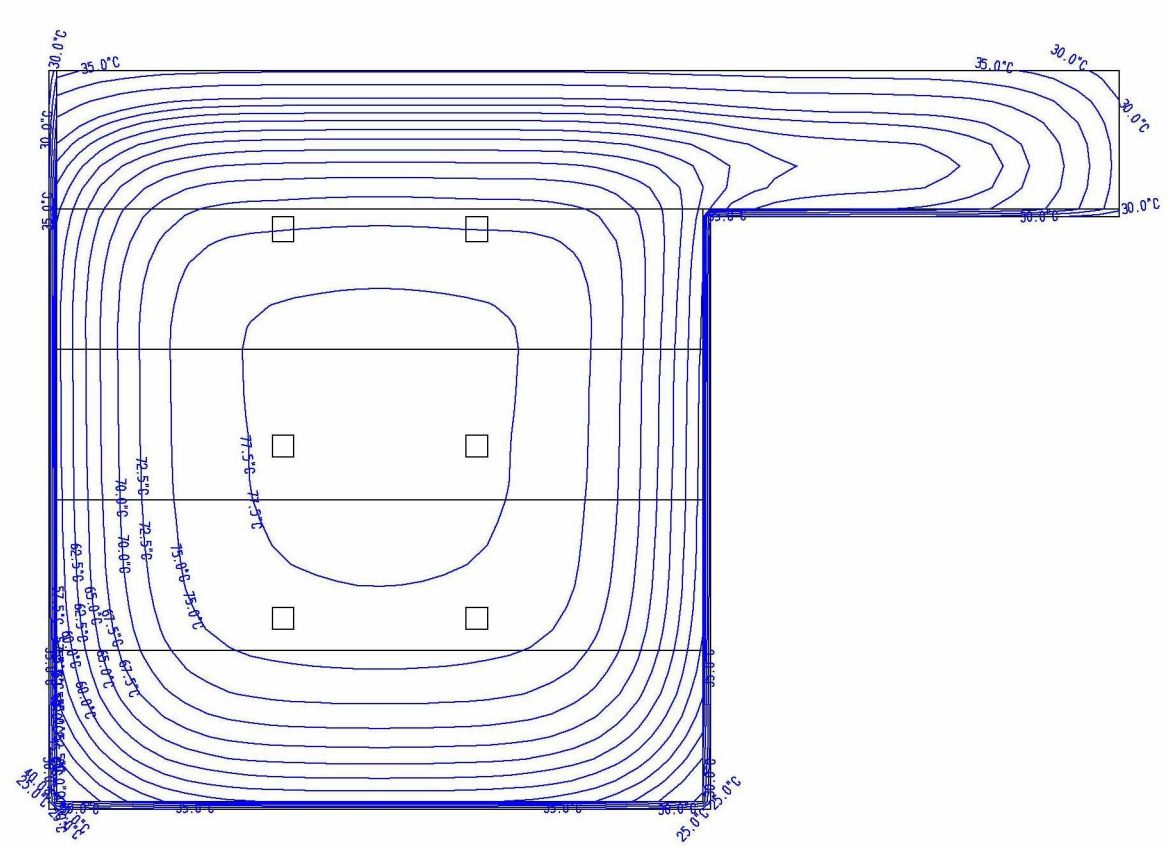


***Figure 2: Distribution of the temperature T (°C) for t = 36 h ,***
***left half of the concrete arch, size 2.1 (+1.4) × 2.4 m , no cooling***

The illustrative example documents the numerical simulation of the temperature development in certain massive concrete bridge structure, namely in one 2,5 m thick slab, formed in 5 layers in 5 time steps. The basic thermal technical characteristics by (1) were

$$\lambda = 1.8 \text{ W/(m.K)} , \qquad \rho = 2440 \text{ kg/m}^3 , \qquad c = 840 \text{ J/(kg.K)} .$$

The thermo-chemical evaluation of hydration heats corresponding to 8 dominant minerals included in the applied Portland cement was compared with experimental macroscopic results, as the time evolution of hydration heat $Q$ ($t$) at Figure 1: following [10], p. 309, the hydration degree can be evaluated from the formula

$$\Gamma(t) = \frac{\int_0^t Q(\zeta)\,d\zeta}{\int_0^{\infty} Q(\zeta)\,d\zeta}.$$

The environment temperature was 15 °C. Figures 2 and 3 demonstrate the output from the computational simulation of the distribution of $T$ after 36 hours (when the hydration heat by Figure 1 seems to be negligible) in the cases without and with active artificial water cooling. The more detailed discussion of such cooling effects (including the complete set of material characteristics, boundary conditions, etc.) is being prepared, namely for the *Modelling* conference, held in June 2009 in Rožnov pod Radhoštěm (Czech Republic).

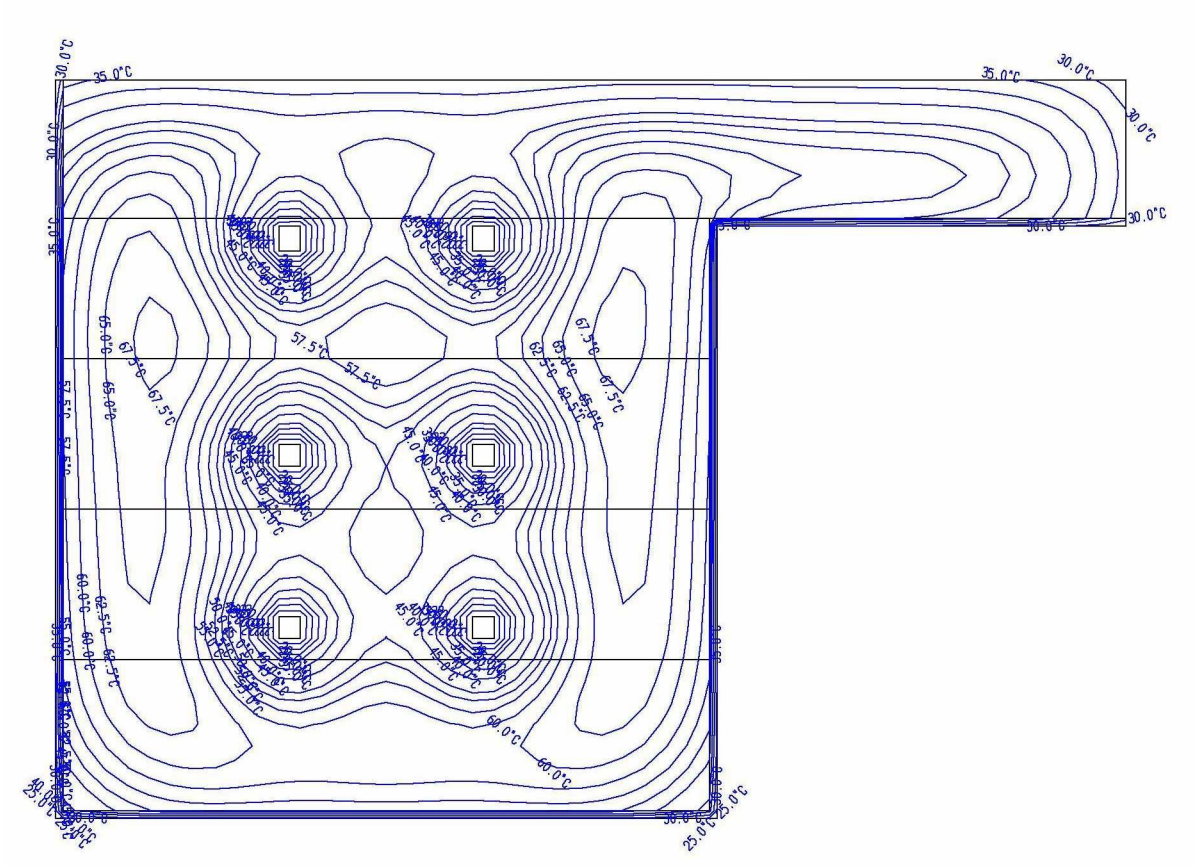



***Figure 3: Distribution of the temperature T (°C) for t = 36 h , left half of the concrete arch, size 2.1 (+1.4) × 2.4 m , artificial water cooling***

**Acknowledgement.** This work was partially supported by the Ministry of Education, Youth and Sports of the Czech Republic, research & development project No. 0021630511.

## 4. References

[1]Azenha, M. – Faria, R. 2008. Temperatures and stresses due to cement hydration on the R/C foundation of a wind tower - A case study. In: Engineering Structures, Vol. 30, 2008, 2392-2400.

[2]Bažant, Z.P. 2002. Concrete fracture models: testing and practice. In: Engineering Fracture Mechanics, Vol. 69, 2002, p. 165-205.

[3]Bažant, Z.P. 2001. Prediction of concrete creep and shrinkage: past, present and future. In: Nuclear Engineering and Design, Vol. 203, 2001, p. 27-38.

[4]Bermúdez de Castro, A. 2005. Continuum Thermomechanics. Birkhäuser, 2005.

[5]Bernard, F. – Kamali-Bernard, S. – Prince, W. 2008. 3D multi-scale modelling of mechanical behaviour of sound and leached mortar. In: Cement and Concrete Research, Vol. 38, 2008, p. 449-458.

[6]Bernard, O. – Ulm, F.-J. – Lemarchand, E. 2003. A multiscale micromechanics-hydration model for the early-age elastic properties of cement-based materials. In: Cement and Concrete Research, Vol. 33, 2003, p. 1293-1309.

[7]Dormieux, L. – Kondo, D. – Ulm, F.-J. 2006. Microporomechanics. John Wiley & Sons, 2006.

[8]Červenka, V. – Jendele, L. – Červenka, J. 2005. ATENA Program Documentation, Part 1: Theory. Prague: Červenka Consulting, 2005.

[9]Elices, M. – Guinea, G.V. – Gomez, J. – Planas, J. 2002. The cohesive zone model: advantages, limitations and challenges. In: Engineering Fracture Mechanics, Vol. 69, 2002, p. 137-163.

[10] Gawin, D. – Pesavento, F. – Schrefler, B.A. 2006. Hygro-thermo-chemo-mechanical modelling of concrete at early ages and beyond. In: International Journal for Numerical Methods in Engineering, Vol. 67, 2006, p. 299-331 (Part I) and 332-363 (Part II).

[11] Grondin, F. – Dumontet, H., et al. 2007. Multi-scales modelling for the behaviour of damaged concrete. In: Cement and Concrete Research, Vol. 37, 2007, p. 1453-1462.

[12] Holt, E.E. 2001. Early age autogenous shrinkage of concrete. Espoo: Technical Research Centre of Finland, VTT Publications 446, 2001.

[13] Hillerborg, A. – Modeer, M. – Peterson, P.E. 1976. Analysis of crack formation and crack growth in concrete by means of fracture mechanics and finite elements. In: Cement and Concrete Research, Vol. 6, 1976, p. 773-782.

[14] Klein, P.A., et al. 2001. Physics-based modeling of brittle fracture: cohesive formulations and application of meshfree methods. In: Theoretical and Applied Fracture Mechanics, Vol. 37, 2001, p. 99-166.

[15] Kulish, V.V. – Lage, J.L. 2000. Diffusion within a porous medium with randomly distributed heat sinks. In: International Journal of Heat and Mass Transfer, Vol. 43, 2000, p. 3481-3496.

[16] Moon, J.-H. – Rajabipour, F. – Pease, B. – Weiss, J. 2005. Autogenous shrinkage, residual stress, and cracking in cementitious composites: the influence of internal and external restraint. In: Self-Desiccation and Its Importance in Concrete Technology – Proceedings of 4th International Research Seminar in Gaithersburg (Maryland, USA, editors: Persson, B. – Bentz, D. – Nilsson, L.-O.), Lund Institute of Technology, 2005, p. 1-20.

[17] Nilsson, L.-O. – Mjörnell, K. 2005. A macro-model for self-desiccation in high performance concrete. In: Self-Desiccation and Its Importance in Concrete Technology - Proceedings of 4th International Research Seminar in Gaithersburg (Maryland, USA, editors: Persson, B., Bentz, D., Nilsson, L.-O.), Lund Institute of Technology 2005, p. 49-66.

[18] Pichler, Ch. – Lackner, R. – Mang, H.A. 2007. A multiscale micromechanics model for the autogenous-shrinkage deformation of early-age cement-based materials. In: Engineering Fracture Mechanics, Vol. 74, 2007, p. 34-58.

[19] Qin, M. – Belarbi, R. – Aït-Mokhtar, A. – Seigneurin, A. 2006. An analytical method to calculate the coupled heat and moisture transfer in building materials. In: International Communications in Heat and Mass Transfer, Vol. 33, 2006, p. 39-48.

[20] Sanavia, L. – Schrefler, B.A. – Steinmann, P. 2002. A formulation for an unsaturated porous medium undergoing large inelastic strains. Computational Mechanics, Vol. 28, 2002, p. 137-151.

[21] STEINHAUSER, M.O. 2008. Computational Multiscale Modelling of Fluids and Solids. Springer, 2008.

[22] VALA, J. 2001. Modelling of creep in composites. In: Building Research Journal, Vol. 49, 2001, p. 147-166.

[23] VALA, J. – ŠŤASTNÍK, S. 2004. On the thermal stability in dwelling structures. In: Building Research Journal, Vol. 52, 2004, p. 31-56.

[24] WATANABE, K. – TADA, S. 2005. From simultaneous heat and moisture transfer models to the modeling of deterioration of concrete. Tsukuba (Japan): Building Research Institute, 2005, available at www.texte.co.jp/heat_moisture/Heat_mas_transf.pdf.

[25] WATANABE, K. 2005. Rapid measurement of moisture diffusive material properties. In: International Conference on Durability of Building Materials and Components in Lyon, 2005, Part TT2-248.

[26] WEINAN, E. – ENGQUIST, B. – LI, X. – REN, W. – VANDEN-EIJNDEN, E. 2007. Heterogeneous multiscale methods: a review. In: Communications in Computational Physics, Vol. 2, 2007, p. 367-450.

[27] YILMAZTÜRK, F. – KULUR, S. – PEKMEZCI, B.Y. 2003. Measurement of shrinkage of samples by using digital photogrammetric methods. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 34, 2003, p. 258-261.

[28] ZHANG, J. – LI, V.C. 2004. Simulation of crack propagation in fiber-reinforced concrete by fracture mechanics. In: Cement and Concrete Research, Vol. 34, 2004, p. 333-339.

## Authors' address:

Vala Jiří doc. Ing., CSc.
Šťastník Stanislav doc. RNDr. Ing., CSc.
Faculty of Civil Engineering
Brno University of Technology
Veveří 95
602 00 Brno, Czech Republic
vala.j@fce.vutbr.cz, stastnik.s@fce.vutbr.cz

Kozák Vladislav Ing., CSc.
Institute of Physics of Materials
Academy of Sciences of the Czech Republic
Žižkova 22
616 62 Brno, Czech Republic
kozak@ipm.cz

# Another view on the fuzzy regression*

Štefan Varga

**Abstract:** Usual assumptions in fuzzy regression models are that observations and unknown regression parameters are fuzzy numbers. It is natural that estimators of unknown parameters and predictions of the observed variable are fuzzy numbers too. In this paper we introduce fuzzy estimators of unknown regression parameters and fuzzy predictions of the observed variable in the classical regression model (unknown parameters and observations are crisp).

**Key words:** Fuzzy estimators, Fuzzy regression models, Predictions.

## 1. Introduction

The fuzzy regression model is usually studied in the form [5]

$$Y = A_1 f_1(x) + A_2 f_2(x) + \ldots + A_m f_m(x)$$

where the input variable $x$ (predictor) is a crisp (real) variable, $f_i(x)$ are known real functions ($i = 1, 2, \ldots, m$) of the variable $x$, $Y$ is an output fuzzy variable (response) and $A = (A_1, A_2, \ldots, A_m)^T$ is the vector of unknown fuzzy parameters. The most commonly, the observation (value of the variable $Y$) is considered as a symmetric, triangular fuzzy number

$$Y_i = \langle y_i, z_i \rangle$$

($y_i \in R$, $z_i \in R^+$) where $y_i$ is a center and $z_i$ is a spread of this fuzzy number ($i = 1, 2, \ldots, m$). Similarly unknown fuzzy parameter

$$A_i = \langle a_i, s_i \rangle$$

is a symmetric, triangular fuzzy number ($a_i \in R$, $s_i \in R^+$, $a_i$ is a center and $s_i$ is a spread, $i = 1, 2, \ldots, m$). The least square estimators of the vector of the centers $a = (a_1, a_2, \ldots, a_m)^T$ and the vector of the spreads $s = (s_1, s_2, \ldots, s_m)^T$ of the unknown fuzzy parameters in the fuzzy regression model is [5]

$$est_{LS}(a, s) = est_{LS}(a_1, \ldots, a_m, s_1, \ldots s_m) = \underset{a_j \in R,\, s_j \in R^+}{\arg\min} \sum_{i=1}^{n} \left[ \left( y_i - a^T f_i \right)^2 + \frac{2}{3} \left( z_i - \sqrt[w]{s^w |f_i|} \right)^2 \right]$$

* This paper was supported by the grant VEGA 1/0374/08, by the APVV 0375-06 and by the project Kniha.sk

where $f_i = ( f_1(x_i), ..., f_m(x_i) )^T$, $|f_i| = ( |f_1(x_i)|, ..., |f_m(x_i)| )^T$, $s^w = (s_1^w, ..., s_m^w)$ and the arithmetic parameter $w \in [0, \infty]$. It is evident that the estimator of the parameter $A_i = \langle a_i, s_i \rangle$ is the fuzzy number

$$est\ A_i = \langle est\ a_i, est\ s_i \rangle$$

The fuzzy number is also the prediction of the variable *Y* in the point *x* [5]

$$est\ Y(x) = \left\langle est\ a\ f,\quad \sqrt[w]{est\ s^w\ |f|} \right\rangle$$

where $est\ a = (est\ a_1, ..., est\ a_m)$, $est\ s^w = \left((est\ s_1)^w, ..., (est\ s_m)^w\right)$, $f = \left(f_1(x), f_2(x), ..., f_m(x)\right)^T$ and $|f| = \left(|f_1(x)|, |f_2(x)|, ..., |f_m(x)|\right)^T$.

In this paper we are interesting in the classical regression model

$$Y = a_1 f_1(x) + a_2 f_2(x) + ... + a_m f_m(x)$$

where observations $y_i$ of the variable $Y$ in the point $x_i$ ($i$ = 1, 2,…, $n$) and the unknown parameters $a_i$ ($i$ = 1, 2,…, $n$) are not fuzzy but real (crisp) numbers. The question is what is a fuzzy estimator of the unknown crisp parameter $a_i$ ($i$ = 1, 2,…, $n$) and what is a fuzzy prediction of the function value $Y(x)$ in the classical regression model? Answers of these questions are in the part 2 and 3.

## 2. Fuzzy estimators of real unknown parameters

Consider a random variable *X* (population) with density $f(x,\theta)$, $\theta$ is an unknown parameter and a random sample $(x_1, x_2, ..., x_n)$ from the variable *X* (observations of *X*). The point estimator of the unknown parameter $\theta$ is a function of the observations

$$est\ \theta = g(x_1, x_2, ..., x_n)$$

with good statistical properties (good point estimator is for example unbiased estimator with minimum variance [1], [3]). The interval estimator (confidence interval) of the unknown parameter $\theta$ is an interval

$$[a(\alpha),\ b(\alpha)]$$

for which the probability that the unknown parameter $\theta \in [a(\alpha),\ b(\alpha)]$ is equal to (1-$\alpha$). Both, left and right point of the interval, depend not only on the random sample $(x_1, x_2, ..., x_n)$ but also on the probability level $\alpha$.

Now the question is **what is a fuzzy estimator of the unknown parameter** $\theta$ ? Simple said, the fuzzy estimator of the parameter $\theta$ is a fuzzy number *T* [2] whose $\alpha$-cat (for $\alpha = 1$) is the point estimator $T(1) = est\ \theta = g(x_1, x_2, ..., x_n)$ and $\alpha$-cats (for $\alpha \in (0, 1)$) are the interval estimators $T(\alpha) = [a(\alpha),\ b(\alpha)]$. More on the fuzzy numbers and more on the $\alpha$-cats of the fuzzy numbers (fuzzy sets) you can find for example in [4].

Very simple example of the fuzzy estimator of a crisp unknown parameter could be the fuzzy estimator of the unknown mean $\mu$ of the normal distributed population (random variable) $N(\mu, \sigma)$ with known standard deviation $\sigma$. If $(x_1, x_2, ..., x_n)$ is a random sample from this population, the point estimator of $\mu$ is [1], [3]

$$est\ \mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the interval estimator of the mean $\mu$ (depend on the probability $\alpha$) is [1], [3]

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2},\ \bar{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right]$$

where $u_{1-\alpha/2}$ is the quantil of the normal distribution N(0, 1). We use these two formulas for the fuzzy estimator of the unknown mean $\mu$ of the normal distributed population $N(\mu, \sigma = 0.5)$ in the case when we have $n = 20$ observations with average $\bar{x} = 5.4$. The fuzzy estimator $(est_f\ \mu)$ is the fuzzy number $T$ (Figure 1)

$$est_f\ \mu = T$$

whose $\alpha$-cat (for $\alpha = 1$) is $T(1) = 5.4$ and $\alpha$-cats (for example for $\alpha = 0.5, \alpha = 0.2, \alpha = 0.1$, $\alpha = 0.05$) are the intervals $T(0.5) = [5.10, 5.70]$, $T(0.2) = [4.83, 5.97]$, $T(0.1) = [4.66, 6.14]$, $T(0.05) = [4.52, 6.28]$ respectively.

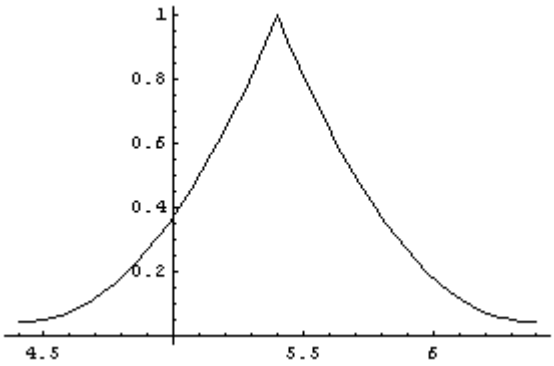



***Figure 1: Fuzzy estimator of the unknown mean μ***

## 3 Fuzzy estimators in classical regression models

The classical regression model (linear in parameters) is studied in the form

$$y = a_1 f_1(x) + a_2 f_2(x) + ... + a_m f_m(x) \qquad (1)$$

where $f_i(x)$ $(i = 1, 2, ..., m)$ are known functions of the input variable $x$ (predictor), $y$ is an output variable (response) and $a = (a_1, a_2, ..., a_m)^T$ is the vector of unknown parameters. An observed value

$$y_i = a_1 f_1(x_i) + a_2 f_2(x_i) + ... + a_m f_m(x_i) + e_i$$

measured in the point $x_i$ with the error $e_i$ $(i = 1, 2, ..., n)$ is a random variable with normal probability distribution. The expectation of the value $y_i$ is

$$E(y_i) = a_1 f_1(x_i) + a_2 f_2(x_i) + ... + a_m f_m(x_i)$$

$(E(e_i) = 0)$ and the variance

$$D(y_i) = D(e_i) = \sigma^2$$

It is known [1] that the best linear unbiased estimator (point estimator) of the unknown parameter $a_i$ is the $i$-th element ($i = 1, 2,\ldots, m$) of the vector

$$est\ a = est\ (a_1,...,a_m)^T = (X^T X)^{-1} X^T Y \tag{2}$$

where the $i$-th raw ($i = 1, 2,\ldots, n$) of the matrix $X$ is $X_i = \left(f_1(x_i), f_2(x_i),..., f_m(x_i)\right)$ and the vector $Y = (y_1, y_2,..., y_n)^T$ contains the observed values $y_i$ ($i = 1, 2,\ldots, n$). On the other hand, the interval estimator of the unknown parameter $a_i$ ($i = 1, 2,\ldots, m$) is

$$\left[est\ a_i - SD(est\ a_i)\, t_{1-\alpha/2,\, n-m}\ ,\quad est\ a_i + SD(est\ a_i)\, t_{1-\alpha/2,\, n-m}\right] \tag{3}$$

where $SD(est\ a_i)$ is the standard deviation of the estimated parameter $a_i$ $(i = 1, 2, ..., m)$ and $t_{1-\alpha/2,\, n-m}$ is a quantil of the Student distribution with ($n$-$m$) degrees of freedom, see [1]. Now, we define a fuzzy estimator of the unknown parameter $a_i$ in the model (1).

**Definition 3.1** The fuzzy estimator of the unknown regression parameter $a_i$ ($i = 1,\ldots, m$) in the classical regression model (1) is the fuzzy number $A_i$ whose $\alpha$-cat $A_i(1)$ (for $\alpha = 1$) is the point estimator $est\ a_i$ (2) and $\alpha$-cats $A_i(\alpha)$ (for $\alpha \in (0, 1)$) are the interval estimators (3)

$$A_i(1) = est\ a_i$$

$$A_i(\alpha) = \left[est\ a_i - SD(est\ a_i)\, t_{1-\alpha/2,\, n-m}\ ,\quad est\ a_i + SD(est\ a_i)\, t_{1-\alpha/2,\, n-m}\right] \tag{4}$$

Furthermore, we introduce prediction ($pred\ y(x)$) of the observed variable $y$ in the point $x$ in the classical regression model (1). The point prediction and the interval prediction you can find, for example in [1] and the fuzzy prediction is defined in Definition 3.2.

The point prediction of the value $y(x)$ in the model (1) is

$$pred\ y(x) = est\ a_1 f_1(x) + est\ a_2 f_2(x) + ... + est\ a_m f_m(x) \tag{5}$$

and the interval prediction is

$$\left[pred\ y(x) - SD(pred\ y(x))\, t_{1-\alpha/2,\, n-m}\ ,\quad pred\ y(x) + SD(pred\ y(x))\, t_{1-\alpha/2,\, n-m}\right] \tag{6}$$

Similarly as in the formula (3), $SD(est\ a_i)$ is the standard deviation of the point prediction of the value $y(x)$, see [1].

**Definition 3.2** The fuzzy prediction of the value $y(x)$ in the classical regression model (1) is the fuzzy number $Y_x$ whose $\alpha$-cat $Y_x(1)$ (for $\alpha = 1$) is the point prediction $pred\ y(x)$ (5) and $\alpha$-cats $Y_x(\alpha)$ (for $\alpha \in (0,\ 1)$) are the interval predictors (6)

$$Y_x(1) = pred\ y(x)$$

$$Y_x(\alpha) = \left[pred\ y(x) - SD(pred\ y(x))\, t_{1-\alpha/2,\,n-m}\ ,\quad pred\ y(x) + SD(pred\ y(x))\, t_{1-\alpha/2,\,n-m}\right] \tag{7}$$

**Example.** Consider the regression model $y = a_1 + a_2\, e^{-x}$ linear in parameters and the normal distributed observations $y_i$ in the points $x_i$

| $x_i$ | -2.0 | -1.4 | -0.6 | 1.3 | 2.1 | 2.9 | 3.5 | 4.1 |
|---|---|---|---|---|---|---|---|---|
| $y_i$ | 24.05 | 13.64 | 7.46 | 2.99 | 2.38 | 2.42 | 2.27 | 2.15 |

Find the fuzzy estimators of the unknown parameters $a_1, a_2$ and the fuzzy prediction of the function value $y(0)$.

Using the expression (2), we obtain the point estimators of the parameters $a_1, a_2$ and using the expression (3) their interval estimators (the probability that $a_1, a_2$ belong to these intervals is (1-α))

$$est\ a_1 = 2.11251,\quad est\ a_2 = 2.93973$$

$$a_1 \in \left[2{,}11251 - 0.08599\, t_{1-\alpha/2,\,6}\ ,\ 2{,}11251 + 0.08599\, t_{1-\alpha/2,\,6}\right]$$
$$a_2 \in \left[2{,}93973 - 0.02819\, t_{1-\alpha/2,\,6}\ ,\ 2{,}93973 + 0.02819\, t_{1-\alpha/2,\,6}\right]$$

The fuzzy estimator of the unknown parameter $a_1$ (Figure 2) is the fuzzy number $A_1$ whose $\alpha$-cats $A_1(\alpha)$ (for $\alpha \in (0,\ 1]$) are

$$A_1(1) = 2.11251$$

$$A_1(\alpha) = \left[2{,}11251 - 0.08599\, t_{1-\alpha/2,\,6}\ ,\ 2{,}11251 + 0.08599\, t_{1-\alpha/2,\,6}\right]$$

and the fuzzy estimator of the unknown parameter $a_2$ (Figure 2) is the fuzzy number $A_2$ whose $\alpha$-cats $A_2(\alpha)$ (for $\alpha \in (0,\ 1]$) are

$$A_2(1) = 2.93973$$

$$A_2(\alpha) = \left[2{,}93973 - 0.02819 * t_{1-\alpha/2,\,6}\ ,\ 2{,}93973 + 0.02819 * t_{1-\alpha/2,\,6}\right]$$

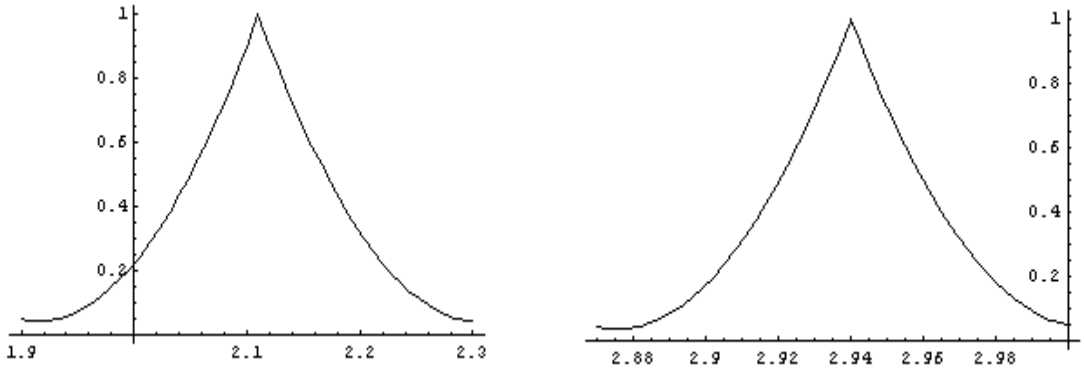



***Figure 2: Fuzzy estimators of the unknown parameters $a_1$, $a_2$***

Finally, using the expressions (5), (6), we obtain the point and the interval prediction of the function value $y(0)$ (the probability that $y(0)$ belong to the interval is (1-α))

$$pred\ y(x) = 2.11251 + 2.93973\, e^{0} = 5.05224$$

$$y(0) \in \left[5.05224 - 0.07386\, t_{1-\alpha/2,6},\ \ 5.05224 + 0.07386\, t_{1-\alpha/2,6}\right]$$

The fuzzy predictor (Figure 3) of the unknown value $y(0)$ is the fuzzy number $Y_0$ whose $\alpha$-cats $Y_0(\alpha)$ (for $\alpha \in (0, 1]$) are

$$Y_0(1) = 5.05224$$

$$Y_0(\alpha) = \left[5{,}05224 - 0.07386\, t_{1-\alpha/2,6},\ 5{,}05224 + 0.07386\, t_{1-\alpha/2,6}\right]$$

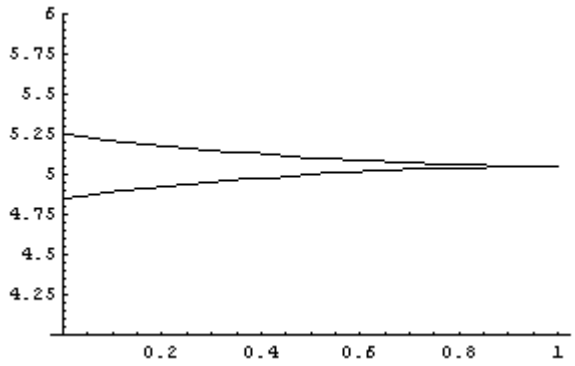



***Figure 3: Fuzzy prediction of the unknown value y(0)***

## 4. Conclusion

We have defined the fuzzy estimators of the unknown regression parameters and the fuzzy predictions of the function values of the observed variable in the classical regression model.

**References**


[1] ANDĚL, J. 1985. Matematická statistika. Praha: SNTL / Alfa 1985.

[2] BUCKLEY, J.J. 2006. Fuzzy Probability and Statistics. Berlin, Heidelberg: Springer – Verlag 2006.

[3] HOOG, R.V. - TANIS, E.A. 2004. Probability and Statistical Inference. Heidelberg: Springer 2006.

[4] KLIR, G.J. – YUAN, B. 1995. Fuzzy Sets and Fuzzy Logic – Theory and Applications. New Jork: Prentice - Hall PTR 1995.

[5] VARGA, Š. 2005. Classical regression models versus fuzzy regression models. Journal of Applied Mathematics, Statistics and Informatics JAMSI 1/2005, No. 2, 95 – 102.

[6] WENDLOVÁ, J. 2006. Which statistical tests for estimating osteoporotic fracture risk? Bratisl. Lek. Listy 2006, 107 (11-12), 453 – 458.



**Author's address:**
Štefan Varga, Doc., RNDr., CSc.
Department of Mathematics
Faculty of Chemical and Food Technology STU
Radlinského 9
812 37 Bratislava
stefan.varga@stuba.sk

# Aplikácia plánovania experimentu –DOE metóda
# Applying of DOE method - design of experiment

Božena Viktorínová

**Abstract:** The article details with reduction of entered variables into experiment and its evaluation, i.e. their influences on the lowering absenteeism of students in schools.

**Key words:** Design, experiment, effect, interraction

**Kľúčové slová:** Návrh, experiment, efekt, interakcia

## 1. Úvod

Analýza rozptylu a regresné metódy slúžia na to, aby sme mohli posúdiť, či sú štatisticky významné rozdiely medzi rôznymi ošetreniami a úrovňami premenných, pričom regresné metódy (techniky) popisujú efekty parametrov na vstupe a ich pôsobenie na výstup procesu. Tieto metódy sú statické, lebo nepopisujú zmeny v rámci prebiehajúceho procesu (napr. ak prebieha pokus v rámci ktorého sa neustále mení teplota).Presnejšou metódou na zachytenie zmien prebiehajúcich v rámci takéhoto skúmaného procesu by mala byť DOE metóda (design of experiments – plánovanie experimentu).Pomocou DOE metódy by sme mohli vstupovať do procesu a zlepšovať ho (podrobnejšie [1], str.409). Touto metódou by sa mala dať posúdiť úroveň faktorov (premenných) v tom zmysle, ktorá odpoveď(+alebo-) zhoršuje, alebo zlepšuje výsledok experimentu, a tak predvídať, aká bude reakcia nášho modelu, teda aký bude výsledok experimentu.

## 2. Popis uvedenej problematiky

Celú problematiku a jej riešenie si vysvetlíme na nasledujúcom príklade. Predstavme si, že chceme znížiť absenciu vysokoškolákov na vyučovacom procese (prednášky, cvičenia...). Za tým účelom im rozdáme dotazník, v ktorom majú odpovedať na niektoré otázky. Počet faktorov (premenných), na ktoré majú študenti odpovedať si dopredu vytypujeme. Tak isto aj počet možností pre odpovede. Ak uznáme za vhodné, že počet možných kombinácií odpovedí je veľký, môžeme ich zredukovať. Napríklad, ak vychádzame z týchto faktorov (označíme ich veľkým písmenom):

A: Deň v týždni, kedy študent chýbal v škole ( pondelok, piatok)
B: Ospravedlnenie sa študenta, ak chýbal ( áno, nie)
C: Umiestnenie školy od bydliska v km ( 1, 2, 3, 4 ), pričom 1 = do 10 km, 2 = do 20 km, 3 = do 30 km, 4 = nad 50 km.
D : Ročník ( 1, 2, 3 )
E : Rada od vyučujúceho, ak veľa chýbal ( áno, nie )
F : Dohoda so školou, ak veľa chýbal ( áno, nie )
G : Pohlavie študenta ( muž, žena ).

Počet možných kombinácií všetkých odpovedí je 2x2x4x3x2x2x2= 384, čo je príliš veľa. Ak znížime počet možných alternatív odpovedí na dve u každého zo siedmich faktorov, zníži sa počet kombinácií odpovedí na $2^7 = 128$. Ak vyselektujeme najdôležitejšie faktory, dostaneme napríklad tri dvojúrovňové faktory ( t. zn. u každého faktora sú možné iba 2 odpovede ). Napríklad:

A : Deň v týždni kedy študent chýbal ( pondelok, piatok )
B : Ospravedlnenie sa študenta ak chýbal ( áno, nie )

C : Umiestnenie školy od bydliska v km ( 1, 2 ), ak 1 = do 50 km, 2 = nad 50 km.

Potom $2^3 = 8$, čiže všetkých možných kombinácií odpovedí je osem.

Označme si faktory a ich úrovne nasledovne:

***Tabuľka 1: Faktory a ich úrovne***

| **Faktor** | **Úroveň** | |
|---|---|---|
| | - | + |
| A : deň v týždni | piatok | pondelok |
| B : ospravedlnenie | áno | nie |
| C : vzdialenosť školy | 1 | 2 |

Aby sme mohli tento experiment vyhodnotiť, náhodne sme vybrali 800 študentov z dvoch vysokých škôl a každých náhodne vybratých 100 študentov odpovedalo na jednu z ôsmich kombinácií faktorov ( t. zn., že bolo 100 študentov v každom „ pokuse“). Celkový počet dní absencií študentov v každej kombinácii je uvedený v stĺpci „odpovede“ a je podrobený analýze. Dostali sme nasledujúce výsledky:

***Tabuľka 2: Počet absencií v rámci kategórií***

| **Číslo pokusu** | **Označenie faktora** | | | **Odpovede** |
|---|---|---|---|---|
| | **A** | **B** | **C** | |
| 1 | + | + | + | 190 |
| 2 | + | + | - | 200 |
| 3 | + | - | + | 160 |
| 4 | + | - | - | 170 |
| 5 | - | + | + | 180 |
| 6 | - | + | - | 179 |
| 7 | - | - | + | 90 |
| 8 | - | - | - | 95 |

Tak napríklad pokus číslo 2 by sme čítali nasledovne: 100 študentov absentovalo 200 krát v pondelok, neospravedlnilo sa a má bydlisko vzdialené od školy menej ako 50 km. Efekt faktorov A, B, a C vypočítame ako priemer odpovedí so znamienkom +, mínus priemer odpovedí so znamienkom - . Napríklad efekt faktora A by sme vypočítali nasledovne:

$$\left[\left(\bar{x}_{A^+1}\right) - \left(\bar{x}_{A^-1}\right)\right] = \frac{190 + 200 + 160 + 170 +}{4} - \frac{180 + 179 + 90 + 95}{4} = 180 - 136 = 44 \qquad (1)$$

Podobne sme vypočítali aj efekty faktorov B a C, teda:

A : + 44

B : + 58,5

C : - 6

Ak si všimneme veľkosť faktora C, jeho efekt je malý. Zato efekt faktorov A a B je veľký. Znamienko faktorov A a B indikuje, ktorá úroveň faktora je najlepšia. V tomto prípade odpovede so znamienkom – sú najlepšie, preto piatok ( -) a ospravedlnenie sa študenta ( - ) vykazuje najlepšie hodnoty. Pozri Tab.1. a aj číslo pokusu 7 a 8 pre A a B v Tab.2. V stĺpci „odpovede“ sú najnižšie absencie.

Rozšírme teraz náš model o interakcie medzi faktormi, ktoré dostaneme násobením kombinácií stĺpcov faktorov A, B, C, podľa matematického pravidla:(+ . + = +, + . - = -, - . - = = + ).

*Tabuľka 3: Interakcia faktorov A, B, C.*

| Číslo | Označenie faktora | | | | | | | Odpovede |
|---|---|---|---|---|---|---|---|---|
| pokusu | A | B | C | AB | BC | AC | ABC | |
| 1 | + | + | + | + | + | + | + | 190 |
| 2 | + | + | - | + | - | - | - | 200 |
| 3 | + | - | + | - | - | + | - | 160 |
| 4 | + | - | - | - | + | - | + | 170 |
| 5 | - | + | + | - | + | - | - | 180 |
| 6 | - | + | - | - | - | + | + | 179 |
| 7 | - | - | + | + | - | - | + | 90 |
| 8 | - | - | - | + | + | + | - | 95 |

Z tejto tabuľky si vypočítame efekty pre interakcie faktorov AB, BC, AC a ABC podobne, ako vo vzorci (1). Ich výsledky uvádzame nižšie:

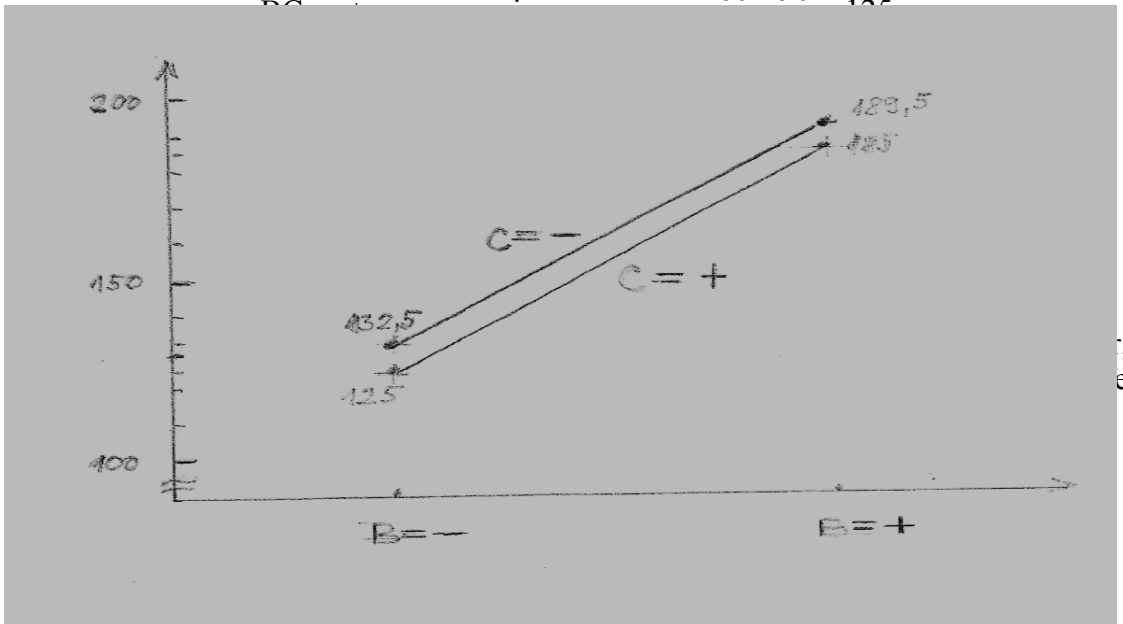
AB: - 28,5
BC: 1,5
AC: - 4
ABC: - 1,5

Tieto výsledky ukazujú, že ani jeden efekt kombinácie faktorov nie je signifikantný a najmenej je signifikantný efekt interakcie faktorov A a B, teda AB, čo sme neočakávali. „Najvyšší“ efekt sa ukazuje byť pri kombinácií faktorov BC. Pokúsme sa teda graficky znázorniť túto dvojfaktorovú interakciu. Za tým účelom si vypočítajme štyri priemery štyroch kombinácií odpovedí, faktorov B a C ( teda BC= ++, BC=+-, BC=-+, a BC=++) a znázornime si naše výsledky graficky.

BC=++ priemer $\frac{190+180}{2} = 185$

BC=+- priemer $\frac{200+179}{2} = 189,5$

$\frac{160+90}{}$ 125




, že ani
ekt a C

nesignifikantný efekt ,nemá signifikantný efekt. V opačnom prípade by mali byť priamky neparalelné ( rôznobežné).

## 3. Záver

Pri tvorbe nášho článku sme sa opierali o metódu prezentovanú v [1], konkrétne str. 418 – 424, zistili sme, že autori [1] na str. 423 urobili chybný výpočet, ktorý ďalej spracovávali a nesprávne vyhodnotili. Postupujúc podľa tam uvedenej metodiky autorka tohto článku nezistila efektívne interakcie faktorov, čomu nasvedčuje aj obrázok 1.

## 4. Literatúra

[1] FORREST,W., - BREIFOGLE,III.: Implementing six sigma, John Wiley& Sons, inc., Toronto 2003, ISBN 978 – 0- 471 – 265 – 72 - 6

TEREK,M., - HRNČIAROVÁ,Ľ.,: Štatistické riadenie kvality, Bratislava, IURA Edition, 2004, ISBN 80 – 89047 – 97 – 1.

Tento príspevok vznikol s prispením grantovej agentúry VEGA, v rámci projektu číslo 1/0437/08: Kvantitatívne metódy v stratégii šesť sigma.

## Adresa autora :

Viktorínová Božena, Mgr., CSc.
Dolnozemská 1
852 35 Bratislava
Ekonomická univerzita, FHI, KŠ
viktorin@dec.euba.sk

# On Positive Dependence of Random Variables
# O pozitívnej závislosti náhodných premenných

Volauf Peter

**Abstract:** The aim of this paper is to discuss various qualitative dependence concepts that describe a dependence structure of a random vector and express them using the concept of copula. We focus on concepts of positive dependence as comonotonicity, positive regression dependence, positive quadrant (orthant) dependence and conditional increasing in sequence. The goal is to express them through analytical properties of associated copulas.

**Key words:** Positive dependence, Comonotonicity, Positive regression dependence, Positive quadrant dependence, Positive orthant dependence, Copula, Survival function.

## 1. Introduction and notation

The association between components of random vectors can be studied in quantitative and in qualitative ways. When $(X, Y)$ is not normal vector, Pearson's correlation coefficient is far from the best measure of dependence. Instead, as quantitative measures of dependence Spearman's rho and Kendall's tau are widely used. They are scale invariant measures based on a form of dependence known as *concordance*. But the interrelation between components of random vectors can be studied also in a qualitative way. The purpose of this paper is to discuss various *qualitative* dependence concepts that describe a dependence structure of a random vector.

We are concerned about concepts of positive dependence – comonotonicity, positive regression dependence, positive quadrant (orthant) dependence and their trivariate extensions. Alghouth they were examined mainly in the second half of the last century, recently the concept of copula has been successfully involved in this study (see [2], [3], [5]). Our goal is to characterize them using the concept of copula. It is not surprising that an extension from a bivariate case to a trivariate one is not always evident and is not always unique.

In two dimensional case we can say that there is an absolute lack of dependence between $X$ and $Y$ when variables are stochastically independent. On the other hand, as a 'perfect' positive dependence we can consider the case when $Y = \alpha(X)$, where $\alpha$ is an increasing function. These two extremes form a space for the concepts of positive dependence that lie between them. In trivariate case the dependence structure is much more complex. However, it is not too much complex in the case when there is a *positive* dependence between all components of $(X, Y, Z)$. As it was mentioned above, we apply the concept of copula to study these dependences so for the convenience of the reader we refresh some basic facts about copulas.

First, let us introduce some notations, abbreviations and conventions. *rv* stands for a random variable (or a random vector), a distribution function $F$ of rv $X$ is defined by $F(x) = \mathrm{P}(X \le x)$ and is abbreviated as *cdf*. We say that $X$ is continuous when $F$ is a continuous function. When $\overline{R}$ stands for the extended real line $[-\infty; \infty]$ and $X$ is continuous then cdf $F$ maps $\overline{R}$ on $[0, 1]$, i.e. the range of $F$, Ran$F$, is equal to $[0, 1]$.

A joint cdf of $(X, Y, Z)$ is denoted by $F$, or $F_{X,Y,Z}$, while its one dimensional margins by $F_1$, $F_2$, and $F_3$. Two dimensional margins of $F$ are denoted by $F_{12}$, $F_{23}$, and $F_{13}$.

Instead of 'nondecreasing' and 'increasing' we use 'increasing' and 'strictly increasing'. For a univariate cdf $F$, the quantile function $F^{-1}$ is defined by $F^{-1}(y) = inf\{x: F(x) \ge y\}$, i.e. it is increasing and always left-continuous. In the next we use the facts that

$$F^{-1} \text{ is continuous} \iff F \text{ is strictly increasing}$$
$$F^{-1} \text{ is strictly increasing} \iff F \text{ is continuous}$$

We note that in the case when $F$ is continuous, it can happen $F^{-1}(F(x)) < x$ while continuity of $F$ guarantees that $F(F^{-1}(y)) = y$.

Symbol $X \sim U(0, 1)$ means that $X$ has the uniform distribution on the interval (0, 1). Symbol $=^{d}$ is used for 'equality in distribution'.

Using probability concepts, the copula can be introduced as follows. Let $F$ be the joint cdf of $(X, Y, Z)$ whose components are *continuous* random variables with cdf $F_1, F_2, F_3$. Note that due to continuity of $F_1$, $F_2$, and $F_3$, their ranges are equal to the interval [0, 1]. The copula $C$ is a mapping from $[0, 1]^3$ to [0, 1] that maps points $(F_1(x), F_2(y), F_3(z))$ to values $F(x, y, z)$:

$$C: (F_1(x), F_2(y), F_3(z)) \rightarrow F(x, y, z)$$

In this way, copula $C$ couples margin values $F_1(x), F_2(y), F_3(z)$ with the joint value $F(x, y, z)$. The formal definition can avoid probabilistic concepts. Here is its an $n$-dimensional version.

An $n$-dimensional *copula* ($n$-copula) is a function $C: [0, 1]^n \rightarrow [0, 1]$ which has the following properties:

(c1) $C(x_1, x_2, \ldots, x_n)$ is increasing in each component $x_i$.

(c2) $C(1, \ldots, 1, x_i, 1, \ldots, 1) = x_i$ for all $i \in \{1, \ldots, n\}$, $x_i \in [0, 1]$.

(c3) For all $a, b \in [0, 1]^n$, $a \le b$, it holds

$$\sum_{j_1=1}^{2} \ldots \sum_{j_n=1}^{2} (-1)^{j_1 + \ldots + j_n} C(x_{1 j_1}, \ldots, x_{n j_n}) \ge 0.$$

where $x_{i1} = a_i$ and $x_{i2} = b_i$ for all $i \in \{1, \ldots, n\}$.

As this contribution concerns with $n$-copulas for $n = 2$ and 3, let us write explicitely what (c3) means. For $n = 2$ (c3) gives: If $a = (a_1, a_2)$, $b = (b_1, b_2)$, $a \le b$ then

$$C(b_1, b_2) - C(a_1, b_2) - C(b_1, a_2) + C(a_1, a_2) \ge 0.$$

For $n = 3$ (c3) states: If $a = (a_1, a_2, a_3)$, $b = (b_1, b_2, b_3)$, $a \le b$ then

$$C(b_1, b_2, b_3) - C(a_1, b_2, b_3) - C(b_1, a_2, b_3) - C(b_1, b_2, a_3) + \\ + C(a_1, a_2, b_3) + C(a_1, b_2, a_3) + C(b_1, a_2, a_3) - C(a_1, a_2, a_3) \ge 0.$$

Note that if $C$ is a 3-dim copula then $C_2: [0, 1]^2 \rightarrow [0, 1]$ defined by $C_2(u, v) = C(u, v, 1)$ is a 2-dim copula. It is easy to verify that (c1), (c2), (c3) are fulfilled, e.g. $C(u, v, 0) = 0$, due to $C(u, v, 0) \le C(1, 1, 0) = 0$, according to (c2). It can be proved that copula satisfies the *Lipschitz* condition, i.e. for every $u$ and $v$ in $[0, 1]^n$ it holds

$$| C(v) - C(u) | \le \sum_{k=1}^{n} | v_k - u_k |.$$

This means that $C$ is uniformly continuous on $[0, 1]^n$. The most important result about copulas is Sklar's theorem.

**Theorem** (Sklar 1959). Let $F$ be an $n$-dimensional cdf with margins $F_1, F_2, \ldots, F_n$. Then there exists an $n$-copula $C$ such that for all $x$ in $\overline{R}^n$ it holds

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)).$$

If $F_1, F_2, \ldots, F_n$ are continuous then $C$ is unique; otherwise $C$ is uniquely determined on $\mathrm{Ran}F_1 \times \ldots \times \mathrm{Ran}F_n$. Conversely, if $C$ is an $n$-copula and $F_1, F_2, \ldots, F_n$ are cdf, then function $H$ defined by

$$H(x_1, x_2, \ldots, x_n) = C(F(x_1), F(x_2), \ldots, F(x_n))$$

is an $n$-dimensional cdf whose margins are $F_1, F_2, \ldots, F_n$.

For the next we restrict ourselves to the cases $n = 2$ and 3. If $F_1, F_2, F_3$ are continuous Sklar's theorem states that a copula is unique so we can derive a copula from the joint cdf $F$:

$$F(x, y, z) = C(F_1(x), F_2(y), F_3(z)),\ u = F_1(x), v = F_2(y), w = F_3(z) \Rightarrow$$
$$C(u, v, w) = F(F_1^{-1}(u), F_2^{-1}(v), F_3^{-1}(w))$$

Moreover, it is known that if rv $X$ has continuous cdf $F$ then rv $F(X) \sim U(0, 1)$. If $F_1, F_2, F_3$ are continuous and we set $U = F_1(X), V = F_2(Y), W = F_3(Z)$ then $U, V, W \sim U(0, 1)$. For the joint cdf $F_{U,V,W}$ and for $0 \le u, v, w \le 1$ we can write

$$F_{U,V,W}(u, v, w) = P(F_1(X) \le u, F_2(Y) \le v, F_3(Z) \le w) = P(X \le F_1^{-1}(u), Y \le F_2^{-1}(v), Z \le F_3^{-1}(w))$$
$$= F_{X,Y,Z}(F_1^{-1}(u), F_2^{-1}(v), F_3^{-1}(w)) = C(F_1(F_1^{-1}(u)), F_2(F_2^{-1}(v)), F_3(F_3^{-1}(w))) = C(u, v, w)$$

so that the restriction of $F_{U,V,W}$ on $[0, 1]^3$ is equal to $C$. It allows us to interpret the copula $C$ as the the joint cdf with uniform margins.

In the next discussion about orthant dependence we use the joint survival function $\overline{C}$. If $C$ is a 3-copula we know that $C$ can be considered as the joint cdf of uniformly distributed variables $U, V, W$. Survival function $\overline{C}$ (associated with $C$) is defined by

$$\overline{C}(u, v, w) = P(U > u, V > v, W > w).$$

In the end of this short presentation we mention very important result independently derived by Hoeffding and Fréchet.

**Theorem** (*Fréchet-Hoeffding bounds*). For every $n$-copula it holds

$$max\{u_1 + ... + u_n + 1 - n,\ 0\} \le C(u_1, \ldots, u_n) \le min\{u_1, \ldots, u_n\}.$$

Let us denote $W^n(\mathbf{u}) = max\{u_1 + ... + u_n + 1 - n, 0\}$ and $M^n(\mathbf{u}) = min\{u_1, \ldots, u_n\}$. These functions are known as Fréchet-Hoeffding lower and upper bound, respectively. Note that $M^n$ is a copula for all $n \ge 2$ whereas $W^n$ is a copula only for $n = 2$.

## 2. Positive dependencies and copulas

Let us consider rv $(X, Y, Z)$ with a joint cdf $F$ and margins $F_1, F_2, F_3$. Obviously, there is an absolute lack of dependence when components are stochastically independent, i.e. when for all real $x, y, z$ it holds

$$F(x, y, z) = F_1(x).F_2(y).F_3(z)$$

This is the case when the *product* copula $\Pi(u, v, w) = u.v.w$ is a possible copula for $(X, Y, Z)$. On the other hand, it is natural to state that there is a 'perfect' positive dependence among the components when $X, Y, Z$ are *common monotone*, i.e. if there exist rv $S$ and increasing functions $\alpha, \beta, \gamma$ such that $X = \alpha(S)$, $Y = \beta(S)$, $Z = \gamma(S)$ hold. The following theorem gives equivalent formulations.

**Theorem** ([7]). Let $(X, Y, Z)$ have a joint cdf $F$ and marginals $F_1, F_2, F_3$. The following conditions are equivalent:

1. $F(x, y, z) = min\{F_1(x), F_2(y), F_3(z)\}$
2. $(X, Y, Z) =^d (F_1^{-1}(U), F_2^{-1}(U), F_3^{-1}(U))$ where $U \sim U(0, 1)$
3. There exist rv $S$ and increasing functions $\alpha, \beta, \gamma$ such that $(X, Y, Z) =^d (\alpha(S), \beta(S), \gamma(S))$.

If this is the case, we say that rv $(X, Y, Z)$ is *comonotonic*. It is evident that $M^3(\mathbf{u}) = min\{u_1, u_2, u_3\}$ is a possible copula for $(X, Y, Z)$. Moreover, if $F_1, F_2, F_3$ are continuous, $M^3$ is a unique copula. Now the question is arising: Suppose that $(X, Y)$ and $(Y, Z)$ are comonotonic. What can be said about $(X, Z)$? The following proposition gives a sufficient condition.

**Proposition.** Let $(X, Y, Z)$ have a joint cdf $F$ and *continuous* margins $F_1, F_2, F_3$. If $(X, Y)$ and $(Y, Z)$ are comonotonic then also $(X, Z)$ is comonotonic and $F(x, y, z) = min\{F_1(x), F_2(y), F_3(z)\}$.

Instead of a complete proof we give at least its important point. If $(X, Y)$ is comonotonic and $F_1, F_2$ are continuous then $Y = F_2^{-1}(F_1(X))$ so that $Y = \alpha(X)$, a.s., where $\alpha$ is increasing.

The next relation of positive dependence between $X$ and $Y$ is a relation of *positive regression dependence* (PRD). For the bivariate case it was introduced by Lehman in [4]:

$Y$ is *positively regression dependent* on $X$ iff $P(Y > y \mid X = x)$ is increasing in $x$, for all $y$.

PRD states that it is more likely that $Y$ takes larger values as $X$ increases. Joe [3] points out that this concept can be extended to the trivariate case in two ways. The first one is positive dependence through the stochastic ordering:

$(X, Y, Z)$ is positive dependent through the stochastic ordering (PDS) if $P(Y > y, Z > z \mid X = x)$ is increasing with respect to $x$ and $P(X > x, Z > z \mid Y = y)$ and $P(X > x, Y > y \mid Z = z)$ are increasing with respect to $y$ and $z$, respectively.

The second extension of PRD is the concept of CIS:

$(X, Y, Z)$ is *conditional increasing in sequence* (CIS) if $P(Y > y \mid X = x)$ and $P(Z > z \mid X = x, Y = y)$ are increasing in $x$ and in $(x, y)$, respectively.

Let us describe CIS in term of a copula. Let us start with the characterization of PRD of $(X, Y)$ through its copula $C_{XY}$ assuming that $X, Y$ are continuous. The conditional probability $P(Y \le y \mid X = x)$ can be derived from the copula $C_{XY}$

$$P(Y \le y \mid X = x) = \lim_{\delta \to 0} \frac{P(x - \delta < X \le x,\ Y \le y)}{P(x - \delta < X \le x)} = \lim_{\delta \to 0} \frac{F_{12}(x, y) - F_{12}(x - \delta, y)}{F_1(x) - F_1(x - \delta)} =$$

$$= \lim_{\delta \to 0} \frac{C_{XY}(F_1(x), F_2(y)) - C_{XY}(F_1(x - \delta), F_2(y))}{F_1(x) - F_1(x - \delta)} =$$

$$= \lim_{\delta \to 0} \frac{C_{XY}(u, v) - C_{XY}(u_\delta, v)}{u - u_\delta} = \frac{\partial C_{XY}}{\partial u}(F_1(x), F_2(y))$$

Conditional probability $P(Y > y \mid X = x)$ is increasing in $x$ iff $P(Y \le y \mid X = x)$ is decreasing in $x$ and assuming the sufficient regularity condition it happens iff $\partial C_{XY} / \partial u$ is decreasing, i.e., $C_{XY}$ is concave in its first variable.

Let us now discuss CIS for $(X, Y, Z)$ with a joint cdf $F$, a copula $C$ and continuous margins $F_1, F_2, F_3$. Suppose that $C_{XY}$ is concave in its first variable so that $(X, Y)$ is PRD. Obviously, $C_{XY}(u, v) = C(u, v, 1)$ and if the function $C(u, v, 1)$ is concave in $u$, $P(Y > y \mid X = x)$ is increasing in $x$.

Now we are interested in a condition which guarantees that $P(Z > z \mid X = x, Y = y)$ is increasing in $x, y$. Equivalently, $P(Z \le z \mid X = x, Y = y)$ should be decreasing in $x, y$.

The conditional probability can be derived from a copula $C$ of $(X, Y, Z)$

$$P(Z \le z \mid X = x, Y = y) = \lim_{\substack{\delta \to 0 \\ \eta \to 0}} \frac{P(x - \delta < X \le x,\ y - \eta < Y \le y,\ Z \le z)}{P(x - \delta < X \le x, y - \eta < Y \le y)} =$$

$$= \lim_{\substack{\delta \to 0 \\ \eta \to 0}} \frac{F(x, y, z) - F(x - \delta, y, z) - F(x, y - \eta, z) + F(x - \delta, y - \eta, z)}{F_{12}(x, y) - F_{12}(x - \delta, y) - F_{12}(x, y - \eta) + F_{12}(x - \delta, y - \eta)}$$

Now we apply Sklar's theorem and we get

$$P(Z \le z \mid X = x, Y = y) = \lim_{\substack{\delta \to 0 \\ \eta \to 0}} \frac{C(u, v, w) - C(u_\delta, v, w) - C(u, v_\eta, w) + C(u_\delta, v_\eta, w)}{C_{12}(u, v) - C_{12}(u_\delta, v) - C_{12}(u, v_\eta) + C_{12}(u_\delta, v_\eta)}$$

where $u = F_1(x)$, $v = F_2(y)$, $w = F_3(z)$, $u_\delta = F_1(x - \delta)$, $v_\eta = F_2(y - \eta)$. Let us divide the numerator and denominator by the product $\delta\eta$ and take limits. Under the natural regularity conditions we get

$$P(Z \le z \mid X = x, Y = y) = \frac{\dfrac{\partial^2 C(u, v, w)}{\partial u \, \partial v}}{\dfrac{\partial^2 C(u, v, 1)}{\partial u \, \partial v}}$$

should be decreasing in $u$, $v$. Denoting the right side of the above equation by $\varphi(u, v, w)$, we have just proved the next proposition.

**Proposition.** Let $C$ be a copula of $(X, Y, Z)$. If $C(u, v, 1)$ is concave in $u$ and $\varphi(u, v, w)$ is decreasing in $u, v$ then $(X, Y, Z)$ is CIS.

The next relation is a relation of *positive quadrant dependence* (PQD). For the bivariate case it was introduced by Lehman in [4] and studied also, for example, in [1]:

($X$, $Y$) is *positively quadrant dependent* (PQD) iff

$$P(X \le x, Y \le y) \ge P(X \le x) P(Y \le y), \text{ for all } x, y.$$

It can be proved that ($X$, $Y$) is PQD iff $P(X > x, Y > y) \ge P(X > x) P(Y > y)$, for all $x, y$. But this is not the case for a trivariate distribution. According to ([3]) the extension of PQD can be considered in two directions :

$(X, Y, Z)$ is *positive lower orthant dependent* (PLOD) iff

$$P(X \le x, Y \le y, Z \le z) \ge P(X \le x) P(Y \le y) P(Z \le z), \text{ for all real } x, y, z.$$

$(X, Y, Z)$ is *positive upper orthant dependent* (PUOD) iff

$$P(X > x, Y > y, Z > z) \ge P(X > x) P(Y > y) P(Z > z), \text{ for all real } x, y, z.$$

It is known that PLOD and PUOD are not equivalent ([5], Example 5.22). That is the reason why POD is defined as follows: $(X, Y, Z)$ is POD if it is PLOD and PUOD simultaneously.

Now let us discuss POD and characterize it through a copula. It is clear that if $(X, Y, Z)$ is POD then subvectors are PQD (a probability is continuous). From this we can conclude that if $C$ is a copula of $(X, Y, Z)$ and $X, Y, Z$ are continuous then for 2-copulas $C_{12}$, $C_{23}$, $C_{13}$ it must hold ([5], Chap. 5] )

$$C_{12}(u, v) \ge u v, \quad C_{13}(u, w) \ge u w, \quad C_{23}(v, w) \ge v w \quad \text{for all } u, v, w \in [0, 1].$$

It is obvious that the condition PLOD is equivalent with the requirement $C(u, v, w) \ge u v w$, for all $u, v, w \in [0, 1]$. Now we need to guarantee

$$P(X > x, Y > y, Z > z) \ge P(X > x) P(Y > y) P(Z > z), \text{ for all real } x, y, z.$$

Nelsen ([5], p.180) uses the joint survival function $\overline{C}$ (associated with $C$) and the condition PUOD is equivalent with

$$\overline{C}(u, v, w) \ge (1 - u)(1 - v)(1 - w)$$

Thus POD can be formulated by two requirements:

$$C(u, v, w) \ge u v w \quad \text{and} \quad \overline{C}(u, v, w) \ge (1 - u)(1 - v)(1 - w) \quad \text{for all } u, v, w \in [0, 1].$$

But we wish to formulate the condition PUOD explicitely for $C$ itself. The left side of the condition PUOD, the probability $P(X > x, Y > y, Z > z)$ is equal to

$$1 - F_1(x) - F_2(x) - F_3(x) + F_{12}(x, y) + F_{13}(x, z) + F_{23}(y, z) - F_{123}(x, y, z)$$

and the right side is the product

$$(1 - F_1(x))(1 - F_2(y))(1 - F_3(z))$$

after cancelling and applying Sklar' theorem we obtain the condition

$$C(u,v,1) + C(u,1,w) + C(1,v,w) - C(u,v,w) \geq uv + uw + vw - uvw$$

Thus we have proved the next proposition.

**Proposition.** Let $C$ be a copula of rv $(X, Y, Z)$ whose marginal cdf are continuous. Then $(X, Y, Z)$ is POD iff $C$ satisfies the conditions

$$C(u, v, w) \geq uvw$$

$$C(u,v,1) + C(u,1,w) + C(1,v,w) - C(u,v,w) \geq uv + uw + vw - uvw$$

for all $u, v, w \in [0, 1]$.

As an example let us consider $(X, Y, Z)$ with the copula $C$ given by $C(u, v, w) = u \min(v, w)$. It is easy to verify that $C$ fulfils the above conditions so that $(X, Y, Z)$ is POD.

## 3. References


[1] ESSARY, J., PROSCHAN, F., WALKUP, D.: Association of random variables with applications. Ann. Math. Statist. 38, 1967, pp. 1466-1474.

[2] GENEST, Ch., MacKAY, J.: The joy of copulas: bivariate distributions with uniform marginals. American Statistician, 40, No4, 1986, pp. 280-283.

[3] JOE, H.: Multivariate models and dependence concepts. Chapman and Hall, London 1997.

[4] LEHMAN, E. L.: Some concepts of dependence. Ann. Math. Statist. 37, 1966, pp. 1137-1153.

[5] NELSEN, R. B.: An Introduction to Copulas. Lecture Notes in Statistics. Springer, New York, 1999.

[6] SKLAR, A.: Fonctions de répartition a n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, (1959), pp. 229-231.

[7] VANDUFFEL, S.: Comonotonicity: From Risk Measurement to Risk Management. Universiteit van Amsterdam, Faculteit der Economische Wetenschappen en Econometrie, Leuven - Prom.: Dhaene J. & Goovaerts M., 2005, pp. 1 – 174.

[8] VOLAUF, P.: A note on dependence concepts for bivariate distributions. Proceedings of *APLIMAT 2007*, 6th International Conference, Bratislava, 6-9. February, 2007, pp. 479-485.

[9] VOLAUF, P.: On dependence concepts and copulas for trivariate distributions. Proceedings of *APLIMAT 2008*, 7th International Conference, Bratislava, 5-8. February, 2008, pp. 1217-1223.



**Acknowledgements**

This contribution was supported by grant VEGA 1/0198/09.



**Current address**

Volauf, Peter , doc., RNDr., PhD.,
Katedra matematiky, FEI STU, Ilkovičova 13,
812 19 Bratislava,
e-mail: peter.volauf@stuba.sk

## Pokyny pre autorov

Jednotlivé čísla vedeckého časopisu FORUM STATISTICUM SLOVACUM sú prevažne tematicky zamerané zhodne s tematickým zameraním akcií SŠDS. Príspevky v elektronickej podobe prijíma zástupca redakčnej rady na elektronickej adrese uvedenej v pozvánke na konkrétne odborné podujatie Slovenskej štatistickej a demografickej spoločnosti. Názov word-súboru uvádzajte a posielajte v tvare: ***priezvisko_nazovakcie.doc***

**Forma:** Príspevky písané výlučne len v textovom editore MS WORD, verzia 6 a vyššia do verzie 2003, písmo Times New Roman CE 12, riadkovanie jednoduché (1), formát strany A4, všetky okraje 2,5 cm, strany nečíslovať. Tabuľky a grafy v čierno-bielom prevedení zaradiť priamo do textu článku a označiť podľa šablony. Bibliografické odkazy uvádzať v súlade s normou STN ISO 690 a v súlade s medzinárodnými štandardami. Citácie s poradovým číslom z bibliografického zoznamu uvádzať priamo v texte.

**Rozsah:** Maximálny rozsah príspevku je 6 strán.

**Príspevky sú recenzované.** Redakčná rada zabezpečí posúdenie príspevku členom redakčnej rady alebo externým oponentom.

**Štruktúra príspevku:** ***(Pri písaní príspevku využite elektronickú šablónu: http://www.ssds.sk/ v časti Vedecký časopis, Pokyny pre autorov.)***

Názov príspevku v slovenskom jazyku (štýl Názov: Time New Roman 14, Bold, centrovať)

Názov príspevku v anglickom jazyku (štýl Názov: Time New Roman 14, Bold, centrovať)

*Vynechať riadok*

Meno1 Priezvisko1, Meno2 Priezvisko2 (štýl normálny: Time New Roman 12, centrovať)

*Vynechať riadok*

**Abstract:** Text abstraktu v anglickom jazyku, max. 10 riadkov (štýl normálny: Time New Roman 12).

*Vynechať riadok*

**Key words:** Kľúčové slová v anglickom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

*Vynechať riadok*

**Kľúčové slová:** Kľúčové slová v jazyku v akom je napísaný príspevok, max. 2 riadky (štýl normálny: Time New Roman 12).

*Vynechať riadok*

*Vlastný text príspevku v členení:*

**Úvod** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

**Názov časti 1** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

**Názov časti 1. . .**

**Záver** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

Vlastný text jednotlivý častí je písaný štýlom Normal: písmo Time New Roman 12, prvý riadok odseku je odsadený vždy na 1 cm, odsek je zarovnaný s pevným okrajom. Riadky medzi časťami nevynechávajte.

**Literatúra** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

[1] Písať podľa normy STN ISO 690

[2] GRANGER, C.W. – NEWBOLD, P. 1974. Spurious Regression in Econometrics. In: Journal of Econometrics, č. 2, 1974, s. 111 – 120.

**Adresa autora (-ov)** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, adresy vpísať do tabuľky bez orámovania s potrebným počtom stĺpcov a s 1 riadkom):*


| | |
|---|---|
| Meno1 Priezvisko1, tituly1<br>Ulica1<br>970 00 Mesto1<br>meno1.priezvisko1@mail.sk | Meno2 Priezvisko2 , tituly2<br>Ulica2<br>970 00 Mesto2<br>meno2.priezvisko2@mail.sk |

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

# Naše najbližšie akcie:

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

**Aplikácie metód na podporu rozhodovania vo vedeckej, technickej a spoločenskej praxi**
30. jún 2009, STU Bratislava

**12. SLOVENSKÁ DEMOGRAFICKÁ KONFERENCIA**
tematické zameranie: Demografická budúcnosť Slovenska
23. – 25. 9. 2009, Hotel DAMONA REGIA, Bojnice

**FernStat 2009**
VI. medzinárodná konferencia aplikovanej štatistiky
(**F**inancie, **E**konomika, **R**iadenie, **N**ázory)
tematické zameranie: *Aplikovaná, demografická, matematická štatistika, štatistické riadenie kvality.*
1. - 2. október 2009, hotel Lesák, Tajov pri Banskej Bystrici

**18. Medzinárodný seminár VÝPOČTOVÁ ŠTATISTIKA,**
3. – 4. 12. 2009, Bratislava, Infostat

**Prehliadka prác mladých štatistikov a demografov**
3. 12. 2009, Bratislava, Infostat

**NITRIANSKE ŠTATISTICKÉ DNI 2010**
4. - 5. február 2010, Nitra

**Pohľady na ekonomiku Slovenska 2010**
13. 4. 2010, Bratislava, Aula EU

**EKOMSTAT 2010, 24. škola štatistiky**
tematické zameranie: *Štatistické metódy vo vedecko-výskumnej, odbornej a hospodárskej praxi.*
jún 2010, Trenčianske Teplice

**Regiónálne akcie**
priebežne

FORUM STATISTICUM SLOVACUM

vedecký recenzovaný časopis Slovenskej štatistickej a demografickej spoločnosti

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

*Redakcia*
Miletičova 3
824 67 Bratislava 24
Slovenská republika

*Fax*
02/39004009

*e-mail*
chajdiak@statis.biz
Jan.Luha@chello.sk

*Registráciu vykonalo*
Ministerstvo kultúry Slovenskej republiky

*Registračné číslo*
3416/2005

*Evidenčné číslo*
EV 3287/09

*Tematická skupina*
B1

*Dátum registrácie*
22. 7. 2005

*Objednávky*
Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava 24
Slovenská republika
*IČO*: 178764
*DIČ*: 2021504276
*Číslo účtu:* 0011469672/0900
*ISSN 1336-7420*



*Ročník*
V.

*Číslo*
3/2009

*Cena výtlačku* 20 EUR
*Ročné predplatné* 80 EUR