AMathNet

síť pro transfer znalostí v aplikované matematice

evropský sociální fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Předmluva k pátému číslu časopisu Kvaternion

Vážení čtenáři, páté číslo časopisu Kvaternion je věnováno pravděpodobnosti a matematické statistice. Obsahuje tři články, jejichž autoři pocházejí z Masarykovy univerzity, Vysokého učení technického v Brně a z Univerzity Palackého v Olomouci.

První článek „Marie Budíková: Využití vícerozměrné analýzy rozptylu v psychologii“ je věnován aplikacím matematické statistiky v psychologii. Pro mě osobně je zajímavý zejména poznatek, že nadaní dyslektici vykazují vyšší potřebu poznání než průměrní studenti bez dyslexie. Druhý článek „Bohumil Maroš: Plán experimentu a statistická korektnost modelu“ zasahuje do důležité oblasti technických disciplín, konkrétně do matematického modelování a plánování experimentů. Výsledky mohou využit zejména čtenáři z technických oborů, hlavně v oblasti řízení jakosti. Třetí článek „Lucie Kreuzigerová, Ondřej Vencálek: Úloha o setkání a její zobecnění“ je věnován aplikacím tzv. geometrické pravděpodobnosti na zobecnění klasické úlohy setkávání dvou přátel.

Obsah pátého čísla časopisu Kvaternion je krásným důkazem rozmanitosti aplikací statistiky v různých oborech, které jsou většinou považovány za poměrně vzdálené – např. technické vědy a psychologie. Je rovněž důkazem toho, že matematika (v tomto konkrétním čísle pravděpodobnost a matematická statistika) je důležitým aparátem nejen pro technické disciplíny, ale i pro řadu oblastí výzkumu ve společenských vědách. Využívám této příležitosti a děkuji všem autorům za zajímavé příspěvky. Věřím, že toto číslo čtenáře zaujme a přinese jim nejen další poznání, ale především potěšení a možná také inspiraci pro vlastní práci.

V Brně dne 13. 10. 2014

Miroslav Doupovec
Ústav matematiky
Fakulta strojního inženýrství VUT v Brně

VYUŽITÍ VÍCEROZMĚRNÉ ANALÝZY ROZPTYLU V PSYCHOLOGII

MARIE BUDÍKOVÁ

ABSTRAKT. V příspěvku se zabýváme problémem, jak odhalit rozdíly v úrovni vícerozměrné normálně rozložené proměnné, jejíž hodnoty byly zjišťovány v několika skupinách objektů. Zaměřujeme se na případ, kdy variabilitu sledované vícerozměrné proměnné může ovlivňovat jeden faktor, jedná se tedy o jednofaktorovou MANOVU. Popisujeme způsob provedení MANOVY pomocí statistického programového systé-mu STATISTICA, přičemž používáme data získaná při psychologickém výzkumu středoškolských studentů.

1. ÚVOD

V celé řadě aplikací statistiky v praxi řešíme problém, zda lze určitým faktorem (tj. nominální proměnnou A) vysvětlit variabilitu pozorovaných hodnot p-rozměrné kvantitativní proměnné $\mathbf{X}$. Např. zkoumáme, zda druh kosatce (faktor A, má tři varianty neboli úrovně: Iris setosa, Iris versicolor, Iris virginica) má vliv na délku a šířku kališních lístků a na délku a šířku korunních plátků, tedy v tomto případě je kvantitativní proměnná $\mathbf{X}$ čtyřrozměrná.

Pokud bychom zkoumali závislost pouze jednoho rozměru na druhu kosatce, použili bychom analýzu rozptylu jednoduchého třídění (metoda ANOVA – Analysis of Variance). Ve vícerozměrném případě se jedná o metodu MANOVA (Multivariate Analysis of Variance). Základním principem ANOVY i MANOVY je rozklad celkové variability obsažené v datech na dvě složky: meziskupinovou variabilitu, která je způsobena sledovaným faktorem a reziduální variabilitu, která není způsobena daným faktorem a považujeme ji za náhodnou nebo nepodstatnou.

Zatímco ANOVA se přednáší v kurzech statistiky a je implementována ve všech statistických programových systémech (a také v EXCELu), MANOVA zdaleka není tak rozšířena. V tomto příspěvku popíšeme způsob provádění MANOVY na reálných datech z oblasti psychologie, a to včetně ověřování předpokladů jejího použití a interpretace výsledků.

2010 *MSC.* Primární 62H15; Sekundární 91E99.

Klíčová slova. MANOVA, systém STATISTICA, nadaní studenti s dyslexií.

Práce byla podporována projektem A-Math-Net – Síť pro transfer znalostí v aplikované matematice (CZ.1.07/2.4.00/17.0100).

2. Provedení MANOVY

2.1. Formální popis problému

Předpokládáme, že faktor A má $r \geq 3$ úrovní a přitom na h-té úrovni bylo provedeno n_h p-rozměrných pozorování, která považujeme za realizaci p-rozměrného náhodného výběru rozsahu n_h, $h = 1, \ldots, r$. Na každé úrovni faktoru musí být provedeno více pozorování než je závisle proměnných veličin, tj. $n_h > p$, $h = 1, \ldots, r$. Výsledky lze zapsat do tabulky 1:

faktor A	výsledky
úroveň 1	$x_{111}, \ldots, x_{11p}$ $x_{1n_11}, \ldots, x_{1n_1p}$
...	...
úroveň r	$x_{r11}, \ldots, x_{r1p}$ $x_{rn_r1}, \ldots, x_{rn_rp}$

Tabulka 1. Zápis výsledků MANOVY.

Zavedeme následující označení:

$n = \sum_{r=1}^{h} n_r \ldots$ celkový rozsah všech r výběrů,

$M_{hj} = \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hij} \ldots$ výběrový průměr j-té proměnné v h-té skupině, $j = 1, \ldots, p$, $h = 1, \ldots, r$,

$\mathbf{M}_h = \begin{pmatrix} M_{h1} \\ \vdots \\ M_{hp} \end{pmatrix} \ldots$ vektor výběrových průměrů v h-té skupině, $h = 1, \ldots, r$,

$\mathbf{M} = \frac{1}{n} \sum_{h=1}^{r} n_h \mathbf{M}_h \ldots$ vektor celkových průměrů,

$\mathbf{S}_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \big(\mathbf{X}_h - \mathbf{M}_h\big)\big(\mathbf{X}_h - \mathbf{M}_h\big)^T \ldots$ výběrová varianční matice v h-té skupině, $h = 1, \ldots, r$,

$\mathbf{S}_* = \frac{1}{n-r} \sum_{h=1}^{r} \big(n_h - 1\big)\mathbf{S}_h \ldots$ vážený průměr výběrových variančních matic.

Celková variabilita obsažená v datech je vyjádřena maticí $\mathbf{T}$:

$\mathbf{T} = \sum_{h=1}^{r} \sum_{i=1}^{n_h} \big(\mathbf{X}_{hi} - \mathbf{M}\big)\big(\mathbf{X}_{hi} - \mathbf{M}\big)^T$. Matici $\mathbf{T}$ lze rozložit na součet dvou matic $\mathbf{T} = \mathbf{E} + \mathbf{B}$, kde $\mathbf{E}$ je matice reziduální variability

$$\mathbf{E} = \sum_{h=1}^{r} \sum_{i=1}^{n_h} \big(\mathbf{X}_{hi} - \mathbf{M}\big)\big(\mathbf{X}_{hi} - \mathbf{M}\big)^T = \sum_{h=1}^{r} \big(n_h - 1\big)\mathbf{S}_h$$

a $\mathbf{B}$ je matice meziskupinové variability

$\mathbf{B} = \sum_{h=1}^{r} \big(\mathbf{M}_h - \mathbf{M}\big)\big(\mathbf{M}_h - \mathbf{M}\big)^T$.

Vliv faktoru, který způsobuje rozpad datové matice na skupiny, se může projevit jen v matici $\mathbf{B}$. Variabilitu projevující se v matici $\mathbf{E}$ tedy považujeme za reziduální, způsobenou buď náhodnými vlivy nebo faktory, kterou nejsou z našeho hlediska podstatné.

2.2. Test hypotézy o shodě vektorů středních hodnot

Nadále budeme předpokládat, že náhodný výběr příslušející h-té úrovni faktoru A, tedy posloupnost stochasticky nezávislých p-rozměrných náhodných vektorů $\mathbf{X}_{h1}, \ldots, \mathbf{X}_{hn_h}$, pochází z p-rozměrného normálního rozložení $N_p\big(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}\big)$, $h = 1, \ldots, r$ a jednotlivé náhodné výběry jsou stochasticky nezávislé.

Na hladině významnosti α testujeme nulovou hypotézu $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_r$ proti alternativní hypotéze H_1: aspoň jedna dvojice vektorů středních hodnot se liší. Při testování této hypotézy můžeme použít až čtyři různé testy založené na Wilksově kritériu, Lawleyově-Hotellingově kritériu, Pillaiově kritériu a na Royově kritériu [2]. Každé z těchto kritérií je určitým způsobem založeno na vlastních číslech matice $\mathbf{B^{-1}E}$. Označme λ_g g-té vlastní číslo této matice a s počet nenulových vlastních čísel, přičemž $s = \min(p, r-1)$. Uvedeme vzorce pro vyjádření jednotlivých kritérií:

Wilksovo kritérium:

$$\Lambda = \frac{\det\big(\mathbf{E}\big)}{\det\big(\mathbf{E}+\mathbf{B}\big)} = \prod_{g=1}^{s} \frac{1}{1+\lambda_g},$$

Lawleyovo-Hotellingovo kritérium:

$$T^2 = \operatorname{tr}\big(\mathbf{B^{-1}E}\big) = \sum_{g=1}^{s} \lambda_g,$$

Pillaiovo kritérium:

$$P = \operatorname{tr}\Big(\mathbf{B}\big(\mathbf{B}+\mathbf{E}\big)^{-1}\Big) = \sum_{g=1}^{s} \frac{\lambda_g}{1+\lambda_g},$$

Royovo kritérium:

$$V = \lambda_{(1)},$$

kde $\lambda_{(1)}$ je největší vlastní číslo matice $\mathbf{B^{-1}E}$.

V praxi je nejpoužívanější Wilksovo kritérium. Testová statistika F_W pro test shody vektorů středních hodnot vznikne transformací Λ:

$F_W = -\left(n - \frac{p+r}{2} - 1\right)\ln\Lambda$. V případě platnosti nulové hypotézy se statistika F_W asymptoticky řídí rozložením $\chi^2\big(p(r-1)\big)$. H_0 tedy zamítáme na asymptotické hladině významnosti α, když tato statistika nabude hodnoty větší nebo rovné $1-\alpha$ kvantilu uvedeného rozložení, tj. $F_W \geq \chi^2_{1-\alpha}\big(p(r-1)\big)$. Znamená to, že jsme s rizikem omylu nejvýše $100\alpha\%$ prokázali, že alespoň dvě skupiny nemají stejné vektory středních hodnot.

2.3. Simultánní testy o složkách vektorů středních hodnot

Prokážeme-li na zvolené hladině významnosti α rozdíl mezi vektory středních hodnot, budeme dále zjišťovat, které ze sledovaných p kvantitativních proměnných $X_1, \ldots, X_p$ způsobují rozdíl mezi skupinami. Provedeme tedy tzv. simultánní testy. Ty odhalí, které jednotlivé proměnné jsou závislé na faktoru A. Současně tedy testujeme p hypotéz
$H_{01} : \mu_{11} = \cdots = \mu_{1r}, \ldots, H_{0p} : \mu_{p1} = \cdots = \mu_{pr}$.

Použijeme testovou statistiku založenou na Wilksově kritériu:

$K_j = -\left(n - \frac{p+r}{2} - 1\right)\ln\frac{e_{jj}}{t_{jj}}$, kde e_{jj} resp. t_{jj} je j-tý diagonální prvek matice $\mathbf{E}$ resp. $\mathbf{T}$, $j = 1, \ldots, p$. V případě platnosti nulové hypotézy se statistika K_j asymptoticky řídí rozložením $\chi^2\big(p(r-1)\big)$. H_{0j} tedy zamítáme na asymptotické hladině významnosti α, když $K_j \geq \chi^2_{1-\alpha}\big(p(r-1)\big)$.

Může však nastat situace, kdy hypotéza o shodě vektorů středních hodnot byla na hladině významnosti α zamítnuta, avšak simultánní testy neprokáží žádný rozdíl mezi složkami vektorů středních hodnot. V takovém případě jsou rozdíly mezi skupinami způsobeny nějakou kombinací sledovaných p proměnných.

2.4. Vícerozměrná obdoba mnohonásobného porovnávání

Dalším krokem, který následuje po zamítnutí hypotézy o shodě vektorů středních hodnot, je provedení vícerozměrné obdoby mnohonásobného porovnávání. Chceme totiž zjistit, které dvojice vektorů středních hodnot se liší na zvolené hladině významnosti α. Budeme tedy pro všechny indexy $h, h^* = 1, \ldots, r, h \neq h^*$ testovat hypotézu $H_0 : \boldsymbol{\mu_h} = \boldsymbol{\mu_{h^*}}$ proti $H_1 : \boldsymbol{\mu_h} \neq \boldsymbol{\mu_{h^*}}$. Těchto testů je $\binom{r}{2}$. Nulovou hypotézu zamítneme na hladině významnosti α, když testová statistika (založená na Lawleyově-Hotellingově kritériu)

$$\frac{n-r-p+1}{(r-1)p} \cdot \frac{n_h n_{h^*}}{n_h + n_{h^*}} \big(\mathbf{M}_h - \mathbf{M}_{h^*}\big)^T \mathbf{E}^{-1} \big(\mathbf{M}_h - \mathbf{M}_{h^*}\big)$$

nabude hodnoty aspoň $F_{1-\alpha}\big(\nu_1, \nu_2\big)$, kde $\nu_1 = \dfrac{(r-1)p(n-r-p)}{n-2-(r-1)p}$,
$\nu_2 = n - r - p + 1$. Pak jsme s rizikem omylu nejvýše $100\alpha\%$ prokázali, že h-tá a h^*-tá skupina nemají stejné vektory středních hodnot.

2.5. Simultánní testy v mnohonásobném porovnávání

Provedení MANOVY uzavřeme tím, že odhalíme případné rozdíly mezi jednotlivými proměnnými v rámci dvojic skupin. Pro všechny indexy $h, h^*, h \neq h^*$ a všechny indexy $j = 1, \ldots, p$ testujeme na hladině významnosti α hypotézu $H_0 : \mu_{hj} = \mu_{h^*j}$ proti $H_1 : \mu_{hj} \neq \mu_{h^*j}$. Zajímá nás tedy rozdíl mezi středními hodnotami j-té proměnné v h-té a h^*-té skupině. Těchto testů je $\frac{pr(r-1)}{2}$. Testová statistika má tvar:

$$\frac{n-r-p+1}{(r-1)p(n-r)} \cdot \frac{n_h n_{h^*}}{n_h + n_{h^*}} \cdot \frac{\left(M_{hj} - M_{h^*j}\right)^2}{S^2_{*j}}$$

(S^2_{*j} je j-tý diagonální prvek matice $\mathbf{S}_*$). V případě platnosti nulové hypotézy se tato statistika asymptoticky řídí rozložením $F\big(\nu_1, \nu_2\big)$. Hypotézu o shodě j-tých složek vektorů středních hodnot v h-té a h^*-té skupině zamítneme na hladině významnosti α, když tato testová statistika nabude hodnoty větší nebo rovné kvantilu $F_{1-\alpha}\big(\nu_1, \nu_2\big)$.

Upozornění: Vícerozměrnou obdobu mnohonásobného porovnávání ani simultánní testy v mnohonásobném porovnávání systém STATISTICA neposkytuje. Pro-blém lze vyřešit tím, že na zvolenou hladinu významnosti α aplikujeme Bonferroniho korekci [2].

V prvém případě provedeme pro každou dvojici skupin vícerozměrný dvouvýběrový t-test (tzv. Hotellingův T^2 test) a jeho vypočtenou p-hodnotu porovnáme s číslem $\alpha \big/ \binom{r}{2}$. Je-li $p \leq \alpha \big/ \binom{r}{2}$, považujeme rozdíl ve vektorech středních hodnot příslušných dvojic skupin za prokázaný. Ve druhém případě provedeme pro každou proměnnou a každou dvojici skupin dvouvýběrový t-test a jeho vypočtenou p-hodnotu porovnáme s $\alpha \big/ \frac{pr(r-1)}{2}$. Je-li $p \leq \alpha \big/ \frac{pr(r-1)}{2}$, zamítáme hypotézu o shodě středních hodnot příslušné proměnné v daných dvou skupinách.

2.6. Předpoklady v MANOVĚ a jejich ověřování

Vícerozměrná normalita: V každé z r skupin bychom měli testovat hypotézu, že proměnné $X_1, \ldots, X_p$ se řídí p-rozměrným normálním rozložením. Testy na vícerozměrnou normalitu však nejsou běžnou součástí statistických programových systémů. V praxi se spokojíme s tím, že otestujeme normalitu pro každou jednotlivou proměnnou zvlášť. Výsledky těchto testů však posuzujeme jen orientačně. Menší odchylky od normality nebrání provedení MANOVY, při větším porušení používáme vhodné transformace.

Shoda variančních matic: Je-li třídění vyvážené, tj. ve všech skupinách je stejný počet pozorování, je MANOVA odolná vůči porušení předpokladu shody variančních matic. V případě nevyváženého třídění je nutné provést Boxův test shody variančních matic. Na hladině významnosti α testujeme hypotézu
$H_0 : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_r$ proti alternativní hypotéze H_1: aspoň jedna dvojice variančních matic se liší. Testová statistika má tvar:

$$T_0 = \frac{1}{C_p}\left[(n-r)\ln|\mathbf{S}_*| - \sum_{h=1}^{r}(n_h - 1)\ln|\mathbf{S}_h|\right],$$

kde

$$C_p = 1 + \frac{2p^2 + 3p - 1}{6(r-1)(p+1)} \left(\sum_{h=1}^{r} \frac{1}{n_h - 1} - \frac{1}{n-r} \right)$$

je konstanta zlepšující aproximaci. V případě platnosti nulové hypotézy se statistika T_0 asymptoticky řídí rozložením

$$\chi^2 \left(\frac{(r-1)p(p+1)}{2} \right).$$

Pokud testová statistika nabude hodnoty aspoň

$$\chi^2_{1-\alpha} \left(\frac{(r-1)p(p+1)}{2} \right),$$

hypotézu o shodě variančních matice zamítneme na asymptotické hladině významnosti α.

Linearita vztahů: Vzhledem k tomu, že MANOVA patří do skupiny obecných lineárních modelů, předpokládá se, že v každé skupině existuje mezi závisle proměn-nými veličinami přibližně lineární vztah. Tento předpoklad lze orientačně ověřit pomocí dvourozměrných tečkových diagramů. Výskyt nelineárních vztahů snižuje sílu testů v MANOVĚ.

3. Aplikace MANOVY v psychologickém výzkumu

3.1. Informace o projektu Výkonová motivace rozumově nadaných studentů s dyslexií

Institut výzkumu dětí, mládeže a rodiny je součástí Fakulty sociálních studií Masarykovy univerzity. Vědecká činnost tohoto institutu je zaměřena na sledování psychických a sociálních charakteristik dětí, adolescentů a jejich rodin. V současné době je zde mj. řešen projekt Výkonová motivace rozumově nadaných studentů s dyslexií – základní determinanty v období adolescence a časné dospělosti. Tento projekt se zaměřuje na problematiku mimořádně nadaných adolescentů a mladých dospělých se souběžnou vývojovou poruchou učení – s dyslexií. Podle současných poznatků je právě tato skupina nadaných studentů ve značně znevýhodňující vzdělávací pozici, která jí často znemožňuje dosahovat úspěchů ve škole i v životě. Hlavním cílem projektu je sledování klíčových proměnných, které mohou být zodpovědné za tento stav.

V rámci projektu byly vyšetřeny řádově stovky studentů. V tomto příspěvku se zaměříme na data o 166 studentech bez dyslexie a s diagnostikovanou dyslexií, u nichž byla změřena inteligence Ravenovým testem (maximální skóre je 60 bodů, za nadané jsou považováni studenti se skóre aspoň 56 bodů) a kteří vyplnili dotazník zaměřený na tyto aspekty:

- **vědomí vlastní účinnosti** (přesvědčení jedince, že dokáže úspěšně realizovat chování, které je potřebné k dosažení specifických cílů), výsledky jsou zaznamenány v proměnné skóre H, která může nabývat hodnot od 10 do 40;

- **osobní standardy** (tendence dávat si vysoké cíle a hodnotit se v závislosti na jejich dosažení), výsledky jsou obsaženy v proměnné skóre PS, minimální hodnota může být 7, maximální 35;
- **organizovanost** (ukazuje na schopnost udržovat pořádek a řád ve vlastních věcech), výsledky jsou shrnuty v proměnné skóre O, může nabývat hodnot mezi 6 až 30;
- **potřeba poznávat**, výsledky jsou zaznamenány v proměnné skóre G, která se může pohybovat v mezích -64 až 64.

Celý výzkumný soubor 166 studentů je rozčleněn na čtyři skupiny:

- nadaní studenti s dyslexií ($n_1 = 16$, označení ND),
- nadaní studenti bez dyslexie ($n_2 = 40$, označení NnD),
- průměrní studenti s dyslexií ($n_3 = 22$, označení PD),
- průměrní studenti bez dyslexie ($n_4 = 88$, označení PnD).

Metodami MANOVY zjistíme, zda na hladině významnosti 0,05 existují významné rozdíly mezi uvedenými čtyřmi skupinami studentů a identifikujeme proměnné, které tyto rozdíly způsobují. Ke zpracování použijeme systém STATISTICA Cz 12, a to především modul GLM, jehož součástí je rovněž ANOVA/MANOVA.

3.2. Ověřování předpokladů MANOVY

Nejprve pomocí Shapirova-Wilkova testu ověříme předpoklad o normalitě rozložení proměnných skóre H, skóre PS, skóre O, skóre G (viz 3.1) ve všech čtyřech skupinách:

proměnná	skupina	rozsah	W	p-hodnota
skóre H	ND	16	0,943706	0,396906
skóre PS	ND	16	0,920708	0,173164
skóre O	ND	16	0,974538	0,905670
skóre G	ND	16	0,984604	0,989658
skóre H	NnD	40	0,981282	0,736977
skóre PS	NnD	40	0,947461	0,062032
skóre O	NnD	40	0,950792	0,080743
skóre G	NnD	40	0,927833	0,013694
skóre H	PD	22	0,981058	0,931731
skóre PS	PD	22	0,979518	0,908287
skóre O	PD	22	0,979293	0,904593
skóre G	PD	22	0,960403	0,497479
skóre H	PnD	88	0,983965	0,350405
skóre PS	PnD	88	0,971554	0,049792
skóre O	PnD	88	0,968818	0,032215
skóre G	PnD	88	0,989775	0,728066

Tabulka 2. Výsledky S-W testů normality.

Vidíme, že S-W test zamítá na hladině významnosti 0,05 hypotézu o normalitě skóre G u nadaných nedyslektiků a dále zamítá hypotézu o normalitě skóre PS a skóre O u průměrných nedyslektiků. Normalita je však porušena jen mírně. Nedopustíme se závažné chyby, budeme-li předpokládat, že každá ze čtyř částí datové matice je realizací výběru ze čtyřrozměrného normálního rozložení.

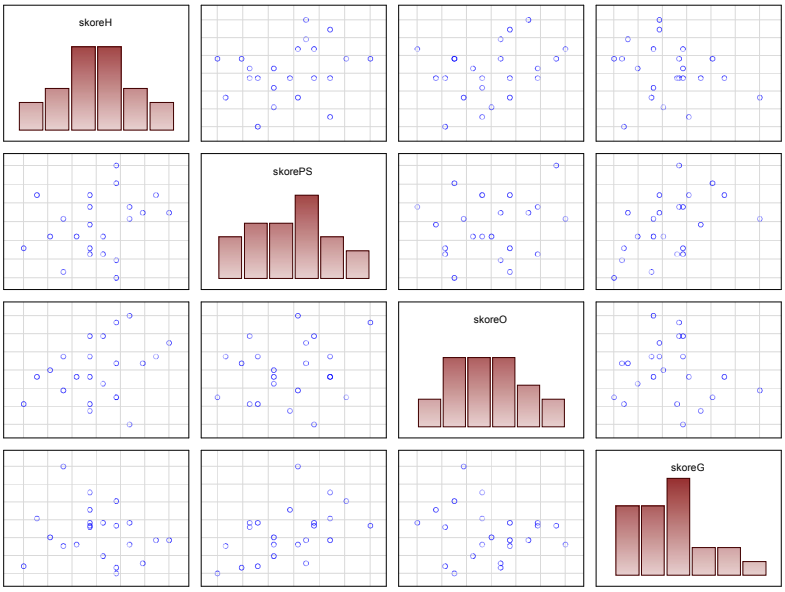
Hypotézu o shodě variančních matic otestujeme Boxovým testem:

	Boxovo M	Chí-kv.	SV	p-hodnota
Boxovo M	39,90594	37,13662	30	0,173196

Tabulka 3. Výsledky Boxova testu homogenity variančních matic.

Test shody čtyř variančních matic poskytl p-hodnotu 0,1732, tedy nadále budeme varianční matice považovat za shodné.

Linearitu vztahů mezi sledovanými proměnnými v daných čtyřech skupinách orientačně posoudíme pomocí tečkových diagramů. Uvedeme zde výsledky jen pro skupinu průměrných dyslektiků, neboť vzhled tečkových diagramů v ostatních skupinách je podobný.



Obrázek 1. Tečkové diagramy dvojic proměnných pro skupinu průměrných dyslektiků.

Výrazné nelinearity se zde neprojevují.

Můžeme konstatovat, že základní předpoklady MANOVY jsou vcelku uspokojivě splněny.

3.3. Provedení MANOVY

Před testováním hypotéz posoudíme úroveň a variabilitu sledovaných proměnných v daných čtyřech skupinách:

proměnná	skupina	rozsah	průměr	sm. odch.
skóre H	ND	16	28,6	3,6
skóre PS	ND	16	22,4	5,2
skóre O	ND	16	17,3	3,2
skóre G	ND	16	18,1	14,2
skóre H	NnD	40	27,3	4,9
skóre PS	NnD	40	20,0	4,5
skóre O	NnD	40	17,7	3,0
skóre G	NnD	40	3,1	20,4
skóre H	PD	22	27,6	2,8
skóre PS	PD	22	20,9	5,0
skóre O	PD	22	15,9	4,3
skóre G	PD	22	7,3	18,0
skóre H	PnD	88	28,3	4,2
skóre PS	PnD	88	20,9	4,4
skóre O	PnD	88	18,5	2,8
skóre G	PnD	88	0,2	19,1

Tabulka 4. Průměry a směrodatné odchylky proměnných v daných čtyřech skupinách.

Z tabulky 4 můžeme vyčíst, že průměry proměnných skóre H a skóre PS se u různých skupin příliš neliší. Průměr skóre O je poněkud nižší ve skupině průměrných dyslektiků. Největší rozdíly mezi průměry jsou pozorovatelné u skóre G, kde se velmi výrazně odlišují nadaní dyslektici a průměrní studenti bez dyslexie.

Z hlediska variability se nejvyrovnanější jeví průměrní dyslektici ve vědomí vlastní účinnosti (skóre H), naopak největší proměnlivost pozorujeme u nadaných nedyslektiků v potřebě poznání (skóre G).

Nyní podle podkapitoly 2.2 provedeme všechny čtyři testy hypotézy o shodě vektorů středních hodnot.

Všechny čtyři testy zamítají na hladině významnosti 0,05 hypotézu, že střední hodnoty proměnných skóre H, skóre PS, skóre O, skóre G jsou ve všech čtyřech skupinách shodné. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že aspoň mezi dvěma skupinami studentů existuje rozdíl z hlediska sledovaných psychologických skóre.

Test	Hodnota	F	Efekt SV	Chyby SV	p-hodnota
Wilksův	0,82122	2,711	12	420,9660	0,001535
Pillaiův	0,18498	2,645	12	483,0000	0,001932
Hotellingův	0,21022	2,762	12	473,0000	0,001213
Royův	0,16843	6,779	4	161,0000	0,000046

Tabulka 5. Výsledky testů o shodě vektorů středních hodnot.

Dále se pomocí simultánních testů pokusíme odhalit, které proměnné způsobují rozdíly mezi skupinami studentů. Simultánní testy STATISTICA neposkytuje. Můžeme však s její pomocí vypočítat matici **E** reziduální variability (viz tabulka 6) a matici **T** celkové variability (viz tabulka tabulka 7). Z těchto matic použijeme diagonální prvky pro výpočet všech čtyř testových statistik (založených na Wilksově kritériu), které jsou uvedeny v podkapitole 2.3.

	skóre H	skóre PS	skóre O	skóre G
skóre H	2788,239	1271,375	265,557	4037,19
skóre PS	1271,375	3433,392	702,182	6882,11
skóre O	265,557	702,182	1596,589	1464,28
skóre G	4037,193	6882,108	1464,277	57782,26

Tabulka 6. Matice **E** reziduální variability.

	skóre H	skóre PS	skóre O	skóre G
skóre H	2826,946	1313,458	298,530	4081,61
skóre PS	1313,458	3502,578	695,301	7257,08
skóre O	298,530	695,301	1731,928	959,94
skóre G	4081,608	7257,084	959,940	62485,28

Tabulka 7. Matice **T** celkové variability.

Diagonální prvky těchto dvou matic použijeme k výpočtu testových statistik, jejichž realizace jsou zaznamenány v tabulce 8:

K1	K2	K3	K4
2,2197	3,2120	13,0999	12,5981

Tabulka 8. Výsledky simultánních testů o složkách vektorů středních hodnot.

Kritický obor je $\langle \chi^2_{1-\alpha}(p(r-1)), \infty) = \langle \chi^2_{0,95}(12), \infty) = \langle 21,0261; \infty)$. Vidíme, že ani jedna ze čtyř statistiky se nerealizuje v kritickém oboru. Vzhledem k tomu,

že hypotéza o shodě vektorů středních hodnot byla na hladině významnosti 0,05 zamítnuta, ale simultánní testy jsou nevýznamné, musí být rozdíly mezi skupinami zapříčiněny nějakou lineární kombinací sledovaných čtyř proměnných.

Nyní provedeme vícerozměrnou analogii mnohonásobného porovnávání, tj. zjistíme, mezi kterými dvojicemi skupin existuje onen významný rozdíl, který byl odhalen při testování hypotézy o shodě vektorů středních hodnot. Jak již bylo uvedeno výše, vícenásobnou obdobu mnohonásobného porovnávání STATISTICA neposkytuje. Problém vyřešíme tak, že provedeme všech šest porovnání (1-2, 1-3, 1-4, 2-3, 2-4, 3-4) pomocí vícerozměrného dvouvýběrového t-testu založeného na Hotellingově statistice T^2 a získané p-hodnoty porovnáme s hladinou významnosti korigovanou podle Bonferroniho, tj. s číslem $\alpha \diagup \binom{r}{2} = \alpha \diagup \binom{4}{2} = \frac{0{,}05}{6} = 0,008\overline{3}$ Získané p-hodnoty jsou uvedeny v tabulce 9:

	ND	NnD	PD	PnD
ND	-	0,11150	0,28093	0,00077
NnD		-	0,29168	0,14690
PD			-	0,00236
PnD				-

Tabulka 9. p-hodnoty vícerozměrných dvouvýběrových t-testů.

Z této tabulky plyne, že s rizikem omylu nejvýše 5 % se odlišují nadaní dyslektici a průměrní studenti bez dyslexie a také průměrní studenti bez dyslexie a s dyslexií.

Poslední úkol spočívá ve zjištění, které proměnné se podílejí na významných rozdílech mezi dvojicemi skupin. Posouzení rozdílů mezi jednotlivými proměnnými v rámci skupin STATISTICA neumožňuje. Pro každou proměnnou tedy provedeme dvouvýběrový t-test, abychom ji porovnali ve dvojicích skupin 1-2, 1-3, 1-4, 2-3, 2-4, 3-4. Následně zjistíme, zda vypočtené p-hodnoty jsou menší nebo rovny korigované hladině významnosti $\alpha \diagup \frac{pr(r-1)}{2} = 0,05 \diagup 24 = 0,0021$. Vypočtené p-hodnoty obsahuje tabulka 10:

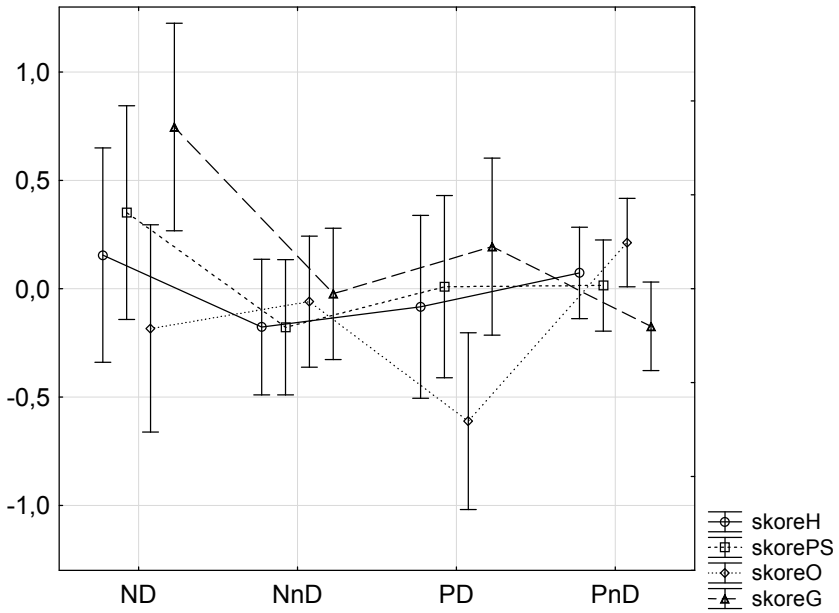
	skóre H	skóre PS	skóre O	skóre G
ND x NnD	0,3109	0,0861	0,6592	0,0096
ND x PD	0,3469	0,3508	0,2839	0,0554
ND x PnD	0,7597	0,2118	0,1058	0,0006
NnD x PD	0,7330	0,4920	0,0604	0,4176
NnD x PnD	0,2191	0,2996	0,1116	0,4347
PD x PnD	0,4914	0,9833	0,0006	0,1149

Tabulka 10. p-hodnoty dvouvýběrových t-testů pro jednotlivé proměnné.

Na základě této tabulky můžeme konstatovat, že:

- nadaní dyslektici a průměrní nedyslektici se liší ve skóre G (nadaní dyslektici vykazují vyšší potřebu poznání než průměrní studenti bez dyslexie),
- průměrní dyslektici a průměrní nedyslektici se liší ve skóre O (průměrní dyslektici mají nižší schopnost udržovat pořádek a řád ve vlastních věcech než průměrní studenti bez dyslexie).

Grafické znázornění rozdílů mezi proměnnými skóre H, skóre PS, skóre O, skóre G v rámci skupin nadaných dyslektiků, nadaných studentů bez dyslexie, průměrných dyslektiků a průměrných studentů bez dyslexie vidíme na obrázku 2, kde body značí průměry a svislé čáry meze 95% intervalů spolehlivosti. Data jsou standardizovaná, aby bylo možno lépe porovnat úroveň jednotlivých proměnných, které mají různou škálu.



Obrázek 2. Porovnání čtyř skupin studentů pro standardizované hodnoty proměnných.

Legenda:
ND – nadaní dyslektici, NnD – nadaní studenti bez dyslexie, PD – průměrní dyslektici, PnD – průměrní studenti bez dyslexie
skóre H – vědomí vlastní účinnosti, skóre PS – osobní standardy , skóre O – organizovanost, skóre G – potřeba poznání

Z obrázku 2 je zřejmé, že vědomí vlastní účinnosti a osobní standardy se v jednotlivých skupinách studentů liší jen málo, zatímco u organizovanosti jsou rozdíly poněkud větší a u potřeby poznání jsou rozdíly dosti výrazné.

4. Závěr

Cílem tohoto příspěvku bylo vysvětlit základní principy vícerozměrné jednofaktorové analýzy rozptylu a ukázat, jak se MANOVA provádí s využitím systému STATISTICA. Vzhledem k tomu, že některé postupy nejsou v tomto systému implementovány, je naznačen způsob, jak získat potřebné výsledky s využitím Bonferroniho korekce hladiny významnosti.

Aplikace MANOVY na data, která byla získána při psychologickém vyšetření jak nadaných, tak průměrných studentů s dyslexií a bez dyslexie, ukázala, že průměrní dyslektici mají nižší schopnost udržovat pořádek a řád ve vlastních věcech než průměrní studenti bez dyslexie a že nadaní dyslektici vykazují vyšší potřebu poznání než průměrní studenti bez dyslexie. Tyto poznatky by mohly být užitečné pro učitele, kteří se ve své každodenní praxi setkávají s dyslektickými studenty.

Za poskytnutí dat patří díky doc. PhDr. Šárce Portešové, Ph.D. z Institutu výzkumu dětí, mládeže a rodiny.

Reference

[1] K. Drápela: *Vícerozměrná analýza rozptylu (MANOVA) – její předpoklady a využití*, dostupné z `http://user.mendelu.cz/drapela/Statisticke_metody/Tutorialy/MANOVA.doc`.

[2] P. Hebák, J. Hustopecký, E. Jarošová, I. Pecáková: *Vícerozměrné statistické metody (1)*, INFORMATORIUM, Praha, 2004.

[3] StatSoft, Inc. (2013), STATISTICA (data analysis software system), version 12, `www.statsoft.com`.

[4] *Projekt Výkonová motivace rozumově nadaných studentů s dyslexií – základní determinanty v období adolescence a časné dospělosti*, `http://www.muni.cz/research/projects/13707?lang=cs`.

Marie Budíková, Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Kotlářská 2, 611 37 Brno, Česká republika,
e-mail: `budikova@math.muni.cz`

PLÁN EXPERIMENTU A STATISTICKÁ KOREKTNOST MODELU

BOHUMIL MAROŠ

ABSTRAKT. Kvalitní příprava, realizace a analýza vědecko-výzkumných experimentů z laboratorních či (polo)provozních pokusů předpokládá osvojení si určitého souboru poznatků a systematicky vybudované metodiky známé pod zkratkou DOE – Design of Experiments. V případě použití „klasického“ regresního přístupu k analýze dat experimentů pro původní (netransformovaná) data z experimentu sestaveného a realizovaného podle metodiky DOE může dojít k různým nesrovnalostem a numerickým, statistickým i principiálně závažným problémům s dopadem na znehodnocení celé experimentální práce. V příspěvku jsou rozebírány numerické a následně statistické aspekty zmíněného nesouladu.

1. ÚVOD

Správný návrh, provedení a následná analýza (vyhodnocení) technologicko-technických experimentů nezbytně patří (nebo rozhodně by měla patřit) k základnímu vybavení studentů i absolventů (nejen) technických univerzit, výzkumných i vědeckých pracovníků. Zanedbání určitých principů plánování a provádění experimentů může vést ke znehodnocení experimentátorské práce. Cílem příspěvku je proto stručné seznámení s pojmy, principy a metodikou *plánování experimentů*, jako i upozornění a dokumentace možných problémů na srozumitelném příkladu.

2. ZÁKLADNÍ POJMY A PRINCIPY PLÁNOVÁNÍ EXPERIMENTŮ

Častým úkolem (nejen) technologicko-technické praxe je zjištění vazeb a vztahů mezi určitými veličinami zkoumaného procesu. Jde zvláště o případy, kdy proces je velice složitý a neexistuje pro jeho popis (dostatečně) vhodný matematický fyzikálně-technický model.

2.1. Základní pojmy DOE

Základním a nejčastějším cílem experimentu je určení, zda určité *faktory* (ovlivňující, vstupní, vysvětlující veličiny) mají vliv na sledovanou (ovlivňovanou, výstupní, vysvětlovanou) veličinu, často nazývanou *odezva* (response). Další, či následující úlohou může být nalezení takové úrovně faktorů, aby bylo dosaženo *optima* (maxima, minima) sledované veličiny. Potřebná data pro sestavení modelu lze získat

2010 *MSC.* Primární 34K06, 34K25.

Klíčová slova. Regresní analýza, DOE, ortogonalita.

Práce byla podporována projektem A-Math-Net – Síť pro transfer znalostí v aplikované matematice (CZ.1.07/2.4.00/17.0100).

buď *pozorováním* veličin procesu bez cíleného zásahu do něj (tzv. *pasivní* či neplánovaný experiment) anebo uskutečněním *experimentu* s cílenými zásahy do procesu (tzv. *aktivní* či *plánovaný* experiment). Avšak ani v případě aktivních experimentů nemají vždy experimentátoři dostatek znalostí o principech efektivního provedení experimentů a vůle k jejich uskutečnění. Je třeba s politováním konstatovat, že *žádnou analýzou (žádnou metodou) experimentálních dat nelze obejít špatně či nedostatečně připravený experiment* [3].
Termínem *experiment* se označuje soustava *pokusů* (také měření, pozorování), která je v případě plánovaného experimentu *vhodným způsobem uspořádána.* V podstatě lze říci, že v plánovaném experimentu jde o vytvoření takových podmínek, aby *rozsah* experimentu byl co nejmenší, ale objem i forma *informací* co nejkvalitnější. V důsledku to také znamená, že u faktorů budeme vybírat jejich vhodné *úrovně* v rámci zvolených (technologicky vhodných) intervalů.
Požadavek na efektivitu se uplatňuje ještě mnohem výrazněji u experimentu s mnoha *faktory* (nezávislými proměnnými). Jeho důsledné respektování vedlo k vytvoření samostatného odvětví aplikované statistiky, tzv. plánovaného (řízeného) experimentu. Různé návrhy uspořádání *měření* a metody jejich *vyhodnocení* se souhrnně označují jako plánování (navrhování) experimentů. V literatuře se často označují jako *Design of Experiment* (*DOE*).

2.2. Základní principy a přístupy metody DOE

Mezi základní principy a přístupy tvorby plánů experimentů podle metody DOE patří následující:

- *replikace*: je opakování měření při stejné úrovni nebo kombinaci úrovní faktorů. Tímto způsobem lze odhadnout nepřesnost měření (náhodnou složku) a zvýšit spolehlivost závěru tím,že můžeme použít test adekvátnosti (přiměřenosti, vhodnosti, použitelnosti) modelu,
- *rozdělení do bloků*: pro poměrně stejné podmínky experimentu za účelem odlišení dalšího zdroje variability,
- *znáhodnění*: aby nevznikla systematická chyba (např. stejným pořadím úrovní faktoru v každém bloku), provedeme jednotlivé pokusy (měření) experimentu v náhodném pořadí. Celé schéma komplexního experimentu pak nazýváme *znáhodněné (replikované) bloky.* Ty dovolují rozložit celkovou variabilitu na složku odpovídající efektům úrovní zkoumaného faktoru, složku nepřesnosti měření, složku odpovídající blokům a reziduální (zbytkovou, nevysvětlenou) složku, jež zahrnuje vliv ostatních (nezkoumaných) činitelů.

Z důvodu minimalizace počtu pokusů v experimentu se používají (pouze) *dvou* a *tříúrovňové* experimenty, které mohou být *úplné* (faktorové) a *zkrácené.*
Plánovaný experiment (*DOE*) poskytuje návod jak *současně měnit všechny faktory* na několika *málo* jejich *úrovních* najednou (z toho vyplývá *minimalizace* počtu pokusů v experimentu) při získání *maxima* informací pro výstavbu vhodného a použitelného modelu zkoumaného procesu.

2.3. Možné problémy při nepoužití metodiky DOE

Pokud budeme spoléhat jen na tzv. „selský rozum“, či „expertní posouzení“ (což zvláště „láká“ u malého počtu, tj. u 2 a 3 faktorů i zkušené experimentátory), pak může (téměř určitě) dojít k několika nepříjemným skutečnostem, např.:

- špatně provedeme *výběr bodů* (hodnot pokusů, úrovní faktorů) s důsledkem *nesprávně indikované statistické nevýznamnosti* některých faktorů (přičemž tyto jsou významné a vlivné), a tím i ke stanovení nesprávného („okleštěného“) modelu,
- sestavíme nevhodný *návrh s přebytečnými údaji* pro model za „dobrým“ účelem získání maxima (ale ne všech potřebných) informací,
- *metoda „pokus-omyl“* jenom málokdy (když se „náhodou strefíme“) poskytne užitečné informace o závislostech,
- *metoda změny pouze jednoho faktoru* při dalších faktorech konstantních (tzv. jednofaktorový experiment, plán) vede jak k příliš (zbytečně) vysokému počtu pokusů (a s tím souvisejícího dlouhého času a vynaložení mnoha peněz na takovýto experiment), tak i k zásadní skutečnosti nemožnosti podchycení (často existujících) interakcí faktorů,
- použití *nevhodné metody analýzy* (při správné metodice návrhu a provedení experimentu metodou DOE) např. lineární regresní analýzy místo analýzy přístupem DOE na *původní* data může také (neočekávaně a překvapivě) vést výše uvedené *nesprávné indikaci statistické nevýznamnosti* některých (ve skutečnosti významných) faktorů s důsledkem na stanovení nesprávného (příliš zjednodušeného) modelu.

3. Plán experimentu podle DOE

S ohledem na *minimalizaci* počtu pokusů (a tím i času a peněz na jejich provedení), lze uvažovat, že každý faktor z má pouze dvě (krajní) úrovně. Ty lze pomocí lineární transformace převést na hodnoty s úrovněmi -1 a $+1$. Tyto kódované (transformované) bezrozměrné faktory budou dále označovány symboly $x_0, x_1, \ldots, x_k$ pro k faktorů.
Lineární transformace původního faktoru z_j na její kódovanou hodnotu x_j, $j = 1, \ldots, k$, se provede pomocí vztahu

$$x_j = \frac{z_j - \dfrac{z_{j,max} + z_{j,min}}{2}}{\dfrac{z_{j,max} - z_{j,min}}{2}} \tag{3.1}$$

Plán experimentu je matice $\mathbf{X}$ tvořená ze sloupců kódovaných faktorů, přičemž první sloupec je složen ze samých jedniček a odpovídá fiktivní proměnné x_0. Příslušný regresní parametr je b_0. Řádky matice plánu představují (kódované) hodnoty úrovní jednotlivých faktorů. Počet řádků matice plánu je roven počtu všech pokusů (měření), včetně opakovaných.
Z důvodu podchycení *náhodných chyb měření* a umožnění *testování adekvátnosti modelu* mohou být přidány další *opakované pokusy* (v *krajních* bodech nebo ve středu plánu čili ve *středovém, centrálním* bodu, ve kterém mají všechny kódované faktory hodnotu 0).

Ortogonální plán je takový plán experimentu $\mathbf{X}$, ve kterém jsou všechny sloupcové vektory matice na sebe *kolmé*.
Ze vztahu (3.1) je zřejmé, že kódováním dochází také k převodu z původních fyzikálních jednotek faktorů na *bezrozměrový* tvar.

3.1. Plán experimentu podle DOE v konkrétním případě

Při tváření oceli za studena nebo za ohřevu klade materiál odpor, jenž se mění jak s velikostí deformace, tak s nastavením tvářecí teploty. Úkolem je nalézt vhodný model pro ocel 14240.3, jenž bude popisovat chování tohoto *přetvárného odporu* v závislosti na *stupni deformace* a *teplotě*. Jeden faktor je tedy logaritmický stupeň deformace φ [-], druhým faktorem je teplota t [°C]. Sledovanou veličinou bude přetvárný odpor σ [MPa]. Oblast zkoumání bude v oblasti, jež je vymezena oběma faktory: $\varphi \in \langle 0,0; 1,4\rangle$, $t \in \langle 20; 700\rangle$.
Chceme nalézt co nejjednodušší vztah $\sigma = f(\varphi, t)$, ale takový, aby statistická analýza potvrdila významnost tohoto vztahu. V literatuře [3] je ukázán *postup* výstavby modelu (v součinnosti s upravovanou či doplňovanou maticí plánu) od jednoduchého *lineárního bez interakcí*, přes *lineární model s interakcemi* až po *úplný kvadratický model*. V našem případě uvažovaný hledaný model bude mít jednodušší tvar, a to:

$$\sigma = \beta_0 + \beta_1 \cdot \varphi + \beta_2 \cdot t + \beta_3 \cdot \varphi \cdot t.$$

Odhad tohoto modelu můžeme zapsat ve tvaru

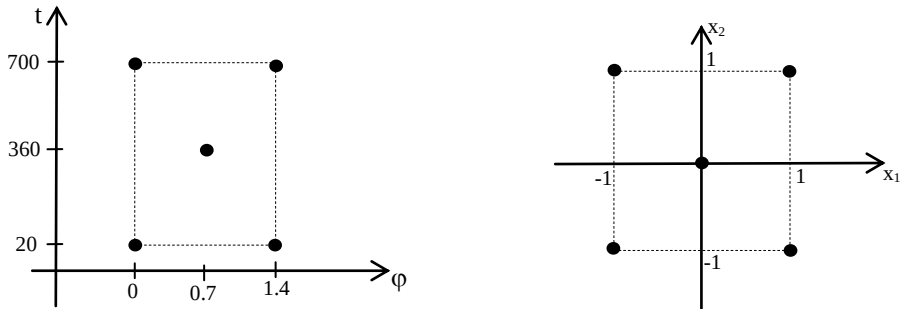
$$\hat{\sigma} = b_0 + b_1 \cdot \varphi + b_2 \cdot t + b_3 \cdot \varphi \cdot t, \tag{3.2}$$

kde b_j jsou odhady regresních parametrů β_j, $j = 0, 1, 2, 3$. Koeficient b_0 je tzv. absolutní (konstantní) člen. Obecně je vhodné s ním vždy počítat (zařazovat jej do modelu), protože podchycuje vliv měřítka (fyzikálních hodnot) odezvy, současně i vliv případné nezařazené vysvětlující proměnné, či existující nelinearitu a umožňuje korektní stanovení statistických ukazatelů (jakými jsou např. koeficient determinace modelu, Durbinova-Watsonova statistika apod.).
Pro daný rozsah uvažovaných veličin φ a t se přepočtou ze vztahu (3.1) transformované (kódované) veličiny takto:

$$x_1 = \frac{\varphi - 0,7}{0,7}, \quad x_2 = \frac{t - 360}{340}. \tag{3.3}$$

Abychom mohli *testovat adekvátnost modelu*, byla přidána ještě dvě *opakovaná* měření, a to ve *středovém* bodu $[0,7; 360]$, což po transformaci (3.3) přejde na bod $[0; 0]$. Původní a transformované hodnoty proměnných jsou názorně viditelné na obr. 1:

Měření uskutečníme v náhodném pořadí, např. v bodech $[1; -1]$, $[-1; -1]$, $[0; 0]$, $[-1; 1]$, $[1; 1]$, $[0; 0]$. Matice plánu $\mathbf{X}$ s transformovanými hodnotami úrovní faktorů v pokusech experimentu pro testovaný model tvaru (3.2) při daném pořadí měření



Obrázek 1. Původní a transformované hodnoty plánu.

pak bude:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{3.4}$$

Druhý sloupec pro x_1 odpovídá původní proměnné φ, třetí sloupec pro x_2 odpovídá původní proměnné t a poslední sloupec pro x_3 odpovídá interakci obou proměnných a získá se vynásobením hodnot z druhého a třetího sloupce.
Pomocí skalárních součinů lze jednoduše ukázat, že tento *plán* je *ortogonální*, tzn., že všechny sloupcové vektory jsou na sebe kolmé.
Po provedeném měření v pořadí, jež určuje matice plánu, jsme získali postupně tyto hodnoty přetvárného odporu: 816, 380, 510, 330, 390 a 497. To znamená, že např. hodnota přetvárného odporu $\sigma = 816\,\mathrm{MPa}$ se obdržela při logaritmickém stupni přetvoření $\varphi = 1,4$ a tvářecí teplotě $t = 20\,^\circ\mathrm{C}$, hodnota $\sigma = 510\,\mathrm{MPa}$ se obdržela při logaritmickém stupni přetvoření $\varphi = 0,7$ a tvářecí teplotě $t = 360\,^\circ\mathrm{C}$, atd.

3.2. Analýza experimentu

Pro úplný faktoriální (faktorový) dvou-úrovňový plán (ale i pro další typy experimentů) je podrobný *postup* (algoritmus, popis) návrhu i *analýzy* uskutečněného experimentu *metodou DOE* uveden např. v literatuře [5]. V této literatuře definovaným názorným postupem je možné „ručně“, „na papíře“ za pomocí kalkulačky, či elegantněji a obecněji prostřednictvím např. tabulkového procesoru Excel, uskutečnit analýzu DOE a získat tvar i koeficienty modelu. Je logické, že v dnešní době je lepší (komplexnější, spolehlivější, rychlejší) použít nějaký vhodný SW obsahující tuto metodu. Existuje speciální SW pro metodiku DOE známý pod anglickou zkratkou (akronymem) CADEX/DOE (Computer Aided Design and Analysis of Experiments / Design of Experiments), mezi něž patří např. programy *Design Expert* či *MODDE*. Avšak asi nejčastěji mají uživatelé (studenti, výzkumníci

a vědci) k dispozici obecné statistické (např. *Minitab, Statgraphics, JMP, Statistica, NCSS*), či matematické (*Matlab*) balíky programů, které lze pro dané účely použít.
Další alternativou, kterou lze použít, je klasická *lineární regrese*, jež je běžně dostupná v různých programech.
Odhad modelu (lineární 1. řádu s interakcí 2. řádu) s transformovanými proměnnými budeme potom uvažovat v analogickém tvaru jako model (3.2), tj.:

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_1 \cdot x_2,$$

kde $\hat{y}$ je vypočtená (predikovaná) hodnota σ z modelu, x_1, resp. x_2, je transformovaná proměnná φ, resp. t.
Neznámé odhadované parametry b_j, $j = 0, \ldots, 3$, získáme ze vztahu (metoda nejmenších čtverců)

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{M}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{V}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{3.5}$$

kde $\mathbf{b} = (b_0, b_1, b_2, b_3)^{\mathrm{T}}$ je vektor odhadovaných parametrů, $\mathbf{X}$ je matice plánu (3.4) a vektor $\mathbf{y} = (816, 380, 510, 330, 390, 497)^{\mathrm{T}}$ obsahuje naměřené hodnoty přetvárného odporu. Matice $\mathbf{M}$ je momentová matice a matice $\mathbf{V}$ se nazývá varianční.
Prakticky lze uvedený vektor regresních koeficientů $\mathbf{b}$ velice elegantně a názorně určit (podle vztahu (3.5) v systému *Matlab, Mathcad, Minitab* či *Statistica*, odkud dostaneme následující vektor odhadů regresních koeficientů:

$$\mathbf{b} = (487,167 \quad 124,000 \quad -119,000 \quad -94,000)^{\mathrm{T}}$$

pro transformované proměnné. Tyto hodnoty můžeme podle (3.3) zpětně přepočítat, abychom obdrželi hodnoty vektoru $\mathbf{b}$ pro původní proměnné φ a t:

$$\mathbf{b} = (389,637 \quad 319,238 \quad -0,073529 \quad -0,394958)^{\mathrm{T}}.$$

Nicméně, pro ověření statistické významnosti modelu, jako i významnosti jeho regresních koeficientů, je nutné stanovit další hodnoty různých statistik a uskutečnit potřebné testy.

4. Regresní analýza příkladu

Pokud uskutečníme statistickou vícenásobnou lineární regresní analýzu pro *kódovaná* a *původní* (netransformovaná) *data* uvedeného příkladu v systému *Matlab* (ale obecně i v libovolném jiném matematickém, statistickém, či ekonometrickém programu), pak sice dostaneme (po přepočtu, po zpětné transformaci) *numericky* stejné hodnoty regresních koeficientů, avšak jejich *statistická* významnost (statistické vlastnosti) se budou u obou modelů (pro oba druhy dat) lišit.

4.1. Statistické „pozadí“ hodnocení regrese

Před hodnocením statistických výstupů regrese je vhodné definovat vztahy vedoucí k jejich určení. Skutečnost, že vypočtený odhad regresního parametru modelu má nenulovou hodnotu ještě neznamená, že je statisticky významně odlišný od nuly (větší nebo menší než nula). To lze posoudit až po určení jeho nepřesnosti charakterizované směrodatnou odchylkou.

Nestranným odhadem rozptylu σ_2 náhodných veličin je náhodná veličina

$$s^2 = \frac{RSC}{n-k-1}, \quad RSC = (y-\hat{y})^{\mathrm{T}}(y-\hat{y}),$$

kde n je počet měření, k je počet faktorů v modelu a RSC je reziduální součet čtverců. Rozptyl jednotlivých náhodných veličin b_j je dán vztahem

$$s^2(b_j) = s^2 \cdot v_{jj}, \tag{4.1}$$

kde v_{jj} je diagonální prvek matice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Jestliže platí, že

$$T = \frac{|b_j|}{\sqrt{s^2(b_j)}} \geq t\left(1-\frac{\alpha}{2}, n-k-1\right), \tag{4.2}$$

kde $t(\dots)$ je kvantil Studentova rozdělení, pak zamítáme hypotézu o statistické nevýznamnosti („nulovosti“) regresního koeficientu na hladině α. Jinými slovy, j-tý sloupec matice $\mathbf{X}$ má významný vliv na hodnoty vektoru $\mathbf{y}$.
Další možností hodnocení statistické významnosti regresního koeficientu je určit tzv. *phodnotu* (p-value), čili dosaženou hladinu pravděpodobnosti (významnosti) zamítnutí hypotézy o nevýznamnosti tohoto regresního koeficientu a porovnat ji se zvolenou „kritickou“ hladinou významnosti testování α (většinou se volí hodnota $0,05$).

4.2. Transformovaná data

S využitím výše definovaných vztahů byla v programu *Minitab* uskutečněna lineární regresní analýza s diagnostikou statistické významnosti regresních koeficientů nejprve pro transformovaná (kódovaná) data podle (3.4) s tímto výsledkem:

Regression Analysis: sigma versus fi; t; fi*t

```
The regression equation is
sigma = 487 + 124 fi - 119 t - 94,0 fi*t


Predictor      Coef  SE Coef       T      P    VIF
Constant    487,167    8,587   56,73  0,000
fi           124,00    10,52   11,79  0,007  1,000
t           -119,00    10,52  -11,32  0,008  1,000
fi*t         -94,00    10,52   -8,94  0,012  1,000

S = 21,0337   R-Sq = 99,4%   R-Sq(adj) = 98,6%
```

(4.3)

V tomto výstupu Minitabu ve sloupci *Coef* jsou hodnoty jednotlivých koeficientů b_j, ve sloupci *SE Coef* hodnoty směrodatných odchylek b_j, ve sloupci T jsou hodnoty T ze vztahu (4.2), ve sloupci P jsou p-hodnoty a ve sloupci *VIF* jsou hodnoty Variance Inflation Factor definované jako

$$VIF_j = v_{jj}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)^2.$$

Ve výstupu Minitabu se ještě vypíše analýza rozptylu:
V řádku *Residual Error* a sloupci *SS* je hodnota RSC a sloupci *MS* je odhad rozptylu s^2. Důležitá je zde p-hodnota v řádku *Lack of Fit*, poněvadž zde se testuje

```
Analysis of Variance

Source          DF      SS     MS       F      P
Regression       3  153492  51164  115,65  0,009
Residual Error   2     885    442
  Lack of Fit    1     800    800    9,47  0,200
  Pure Error     1      85     85
Total            5  154377
```
(4.4)

adekvátnost modelu. Je-li $P < \alpha$, pak vypočtený model není adekvátní (neodpovídá naměřeným hodnotám). Tento řádek se ve výstupu neobjeví, pokud nejsou žádná opakovaná měření.

4.3. Původní data

V kapitole 4.2 byla prováděna regresní analýza z transformovaných dat na interval $\langle -1; 1\rangle$. Nyní provedeme regresní analýzu na původní data pro proměnné φ a t. To znamená, že matice plánu nebude podle (3.4), ale bude mít tvar:

$$\mathbf{X} = \begin{pmatrix} 1 & 1,4 & 20 & 28 \\ 1 & 0 & 20 & 0 \\ 1 & 0,7 & 360 & 252 \\ 1 & 0 & 700 & 0 \\ 1 & 1,4 & 700 & 980 \\ 1 & 0,7 & 360 & 252 \end{pmatrix}.$$

Opět pomocí Minitabu obdržíme:

Regression Analysis: sigma versus fi; t; fi*t

```
The regression equation is
sigma = 390 + 319 fi - 0,0735 t - 0,395 fi*t


Predictor       Coef  SE Coef      T      P    VIF
Constant     389,64    20,79  18,74  0,003
fi           319,33    21,88  14,59  0,005  2,121
t          -0,07353  0,04374  -1,68  0,235  2,000
fi*t       -0,39496  0,04419  -8,94  0,012  3,121

S = 21,0337   R-Sq = 99,4%   R-Sq(adj) = 98,6%
```
(4.5)

```
Analysis of Variance

Source          DF      SS     MS       F      P
Regression       3  153492  51164  115,65  0,009
Residual Error   2     885    442
  Lack of Fit    1     800    800    9,47  0,200
  Pure Error     1      85     85
Total            5  154377
```
(4.6)

Na první pohled je zřejmé, že analýza rozptylu v (4.4) a (4.6) je naprosto shodná. Jiné je to však ve sloupcích P pro p-hodnoty v (4.3) a (4.5). Zatímco všechny p-hodnoty v (4.3) jsou menší než $0,05$, tak p-hodnota v (4.5) pro teplotu je $0,235$, což je mnohem větší hodnota než $0,05$. Z toho plyne, že proměnná t neovlivňuje přetvárný odpor. Každý technik okamžitě ví, že něco není v pořádku, protože tvářecí teplota má významný vliv na velikost přetvárného odporu. Ovšem podle (4.6) je tento model adekvátní, takže by měl být správný.
Na druhé straně je jednoduše ověřitelné, že samotné hodnoty regresních koeficientů pro původní data i transformovaná data jsou číselně shodné po provedení zpětné transformace.

4.4. Shrnutí

Porovnáním výsledků obou přístupů (pro oba druhy dat) je jasné, že i když oba modely *stejně kvalitně* vysvětlují rozptyl odezvy (vykazují stejné hodnoty koeficientu determinace), *liší se* v hodnocení statistické významnosti vlivu (efektu, koeficientu) teploty tváření na přetvárný odpor oceli. Otázka tedy zní, čím (jakou skutečností) je tato disproporce způsobena, což je obsahem následující kapitoly.

5. Rozbor statistického problému

Pro zjištění příčin vzniklého problému byla definována některá statistická i numerická indikační kritéria.

5.1. Příčiny problému

Na základě informací z literatury, např. [2], [4], lze zjistit, že *nesprávná indikace statistické nevýznamnosti regresoru* (vysvětlující proměnné) je důsledkem tzv. *multikolinearity* (téměř souběžnosti) několika *regresorů* (jejich vzájemné korelovanosti, souvislosti), či jinak řečeno, jejich *neortogonality* (nekolmosti). Tato skutečnost způsobuje (kromě jiného) *velké rozptyly* jednotlivých odhadů regresních koeficientů, díky kterým výsledky *t-testů* indikují nesprávně statistickou významnost těchto koeficientů.
Je překvapivé, že i v tomto případě vychází *koeficient determinace* vysoký a regresní *model* může dobře *popisovat* (avšak už ne dobře a správně *vysvětlovat*) experimentální data. Nesprávně určena *statistická* významnost vychází z *numerického* hlediska tzv. *špatné podmíněnosti* tzv. momentové matice $\mathbf{M} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ (kde $\mathbf{X}$ je matice plánu).

5.2. Indikační kritéria neortogonality

Pro testování uvedeného jevu existuje a je doporučeno několik kritérií, a to jak numerických, tak i statistických, přičemž *kritéria*:

- *numerická* vycházející ze spektrálního rozkladu momentové a (její inverzní) varianční matice se stanovením tzv. *vlastních čísel, stopy* a *determinantu matic*,
- *statistická* vycházejí z *korelační* matice, používá se u nich statistických pojmů, či pro hodnocení se využívá kvantilů pravděpodobnostních rozdělení.

Další obdobné členění těchto kritérií může být podle skutečnosti, z jaké *matice* regresorů se vychází, tj. zda se vychází z matice:

- plánu $\mathbf{X}$,
- momentové $\mathbf{M} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$,
- variační $\mathbf{V}$, která je inverzní maticí k matici momentové: $\mathbf{V} = \mathbf{M}^{-1}$,
- anebo z normované verze varianční matice $\mathbf{R}$, která je formálně shodná s *korelační maticí* vysvětlujících proměnných (regresorů).

Seznam vybraných a na datech příkladu použitých kritérií vychází z literatury [1], [2], [4]. U níže uvedeného přehledu jsou u kritérií definované *kritické hodnoty*, při jejichž překročení lze uvažovat o existenci *multikolinearity* (*neortogonality*) regresorů:

A. Kritéria numerická

- *skalární součiny* dvojic vysvětlujících proměnných (z matice plánu $\mathbf{X}$): pokud nejsou všechny *nulové*, pak plán není ortogonální,
- *vlastní čísla* momentové matice $\mathbf{M}$: podíl největšího a nejmenšího vlastního čísla matice ($cond(\mathbf{M}) = \frac{\lambda_{max}}{\lambda_{min}}$) větší než 10^6,
- *vlastní čísla*: nejmenší vlastní číslo je blízké nule,
- *determinant* matice $\mathbf{M}$: hodnota větší než 10^6, nebo hodnota menší než 10^{-6},
- *stopa matice* $\mathbf{M}$: $\mathrm{tr}(\mathbf{M})$, tzn. součet diagonálních prvků, je větší než 10^6,
- *stopa matice* $\mathbf{V}$: hodnota $\mathrm{tr}(\mathbf{V})$ větší než 1.0,
- *korelační matice* $\mathbf{R}$: pokud nejsou všechny mimodiagonální prvky (Pearsonovy párové korelační koeficienty) rovné nule, pak plán není nekorelovaný (není nezávislý),
- *číslo podmíněnosti* K matice $\mathbf{R}$: $\mathrm{cond}(\mathbf{R})$ větší než 10^3,
- *determinant matice* $\mathbf{R}$: blízký nule.

B. Kritéria statistická

- *VIF-faktory* (Variance Inflation Factor - inflační faktory rozptylu regresorů [2]): rovné jedné (prediktory nejsou korelované, plán je nekorelovaný a je *ortogonální*), větší než 1, ale menší než 5 (indikace mírné korelovanosti a neortogonality plánu), větší jak 5, ale menší než 10 (významná korelovanost a neortogonalita), větší než 10 (multikolinearita, neortogonalita). V případě $VIF > 1$ jsou odhady regresních koeficientů (číselně) správné, ale nejsou správné jejich *p-hodnoty*. $VIF_j = v_{jj} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2$, $j = 1, \ldots, k$, v_{jj} je diagonální prvek variační matice $\mathbf{V}$,
- *Testy diagonality* výběrové korelační matice $\mathbf{R}$, tzn. testuje se nulová hypotéza, že tato matice je jednotková [2]: testovací veličinou je hodnota $W = -n \cdot \ln(|\mathbf{R}|)$, kde $|\mathbf{R}|$ značí hodnotu determinantu matice $\mathbf{R}$. Pokud $W > \chi^2_\alpha(\frac{k(k-1)}{2})$, pak se jedná o multikolinearitu.

- Častěji je však používaná *korigovaná* (zpřesněná) hodnota *testu diagonality* na počet faktorů (regresorů) podle *Farrara a Glaubera* [1]:

$$FG = -n \cdot \left(1 - \frac{2k+11}{6n}\right) \cdot \ln(|\mathbf{R}|) = \chi^2_\alpha \left(\frac{k(k-1)}{2}\right),$$

- *F-test* regresorů způsobujících multikolinearitu: Pokud platí

$$FD_j = \frac{n-k}{k-1}(VIF_j - 1) > F_\alpha(k-1, n-k),$$

kde $F_\alpha(k-1, n-k)$ je hodnota kvantilu F-rozdělení.

6. Způsoby řešení problému

Podle informací z literatury byly dále sestaveny návrhy způsobů řešení problému. Pro ověření navrhovaných opatření byla uskutečněna aplikace kritérií a způsobů řešení na data příkladu s hodnocením jejich detekční schopnosti a účinnosti.
V literatuře [2], [4], [5], je doporučeno v případě výskytu multikolinearity použít *lineární transformaci* dat proměnných z na proměnné x několika následujícími způsoby ($i = 1, 2, \ldots, n$, kde n je celkový počet pokusů):

- *centrování hodnot proměnných*:
  $x_i = z_i - \overline{z}$, kde $\overline{z}$ je aritmetický průměr proměnné z,
- *normování hodnot proměnných*:
  $x_i = \dfrac{z_i - \overline{z}}{s_z}$, kde s_z je směrodatná odchylka proměnné z,
- *DOE transformace* (kódování) - viz předchozí vztah (3.1).

Všechny uvedené způsoby lineární transformace dat vedou k *nulové střední hodnotě* a *symetrii* transformovaných dat s příznivým dopadem na zajištění *ortogonality* a *nekorelovanosti* plánu s takovýmito proměnnými. Nicméně, zde je třeba obecně konstatovat, že multikolinearitu regresorů (faktorů) nelze vždy uvedenými způsoby odstranit. Je to tehdy, když jsou faktory mezi sebou z principu závislé (např. délka a objem tělesa), při používání vyšších mocnin (které jsou navzájem závislé) proměnných, apod.
Pro analyzovaná data se rozmezí jejich hodnot pro původní i transformované proměnné pohybovalo v následujících intervalech:

data	φ	t	$\varphi \cdot t$
původní	$\langle 0,0; 1,4 \rangle$	$\langle 20; 700 \rangle$	$\langle 0; 980 \rangle$
centrovaná	$\langle -0,7; 0,7 \rangle$	$\langle -340; 340 \rangle$	$\langle -238; 238 \rangle$
normovaná	$\langle -1,225; 1,225 \rangle$	$\langle -1,225; 1,225 \rangle$	$\langle -1,5; 1,5 \rangle$
DOE transformovaná	$\langle -1; 1 \rangle$	$\langle -1; 1 \rangle$	$\langle -1; 1 \rangle$

Tabulka 1. Intervaly hodnot.

Prakticky to znamená, že jak *normování* tak i *DOE transformace* „stlačují“ hodnoty do intervalu $\langle -2; 2\rangle$, což je doporučené rozmezí s ohledem na zaručení numerické stability výpočtů a omezení kumulačních chyb. Uvedené rozmezí poskytuje nejmenší odchylky od správného řešení, které bychom obdrželi, kdyby počítač uměl zapsat každé číslo na nekonečně dlouhou mantisu.

6.1. Ověření kritérií a způsobů řešení

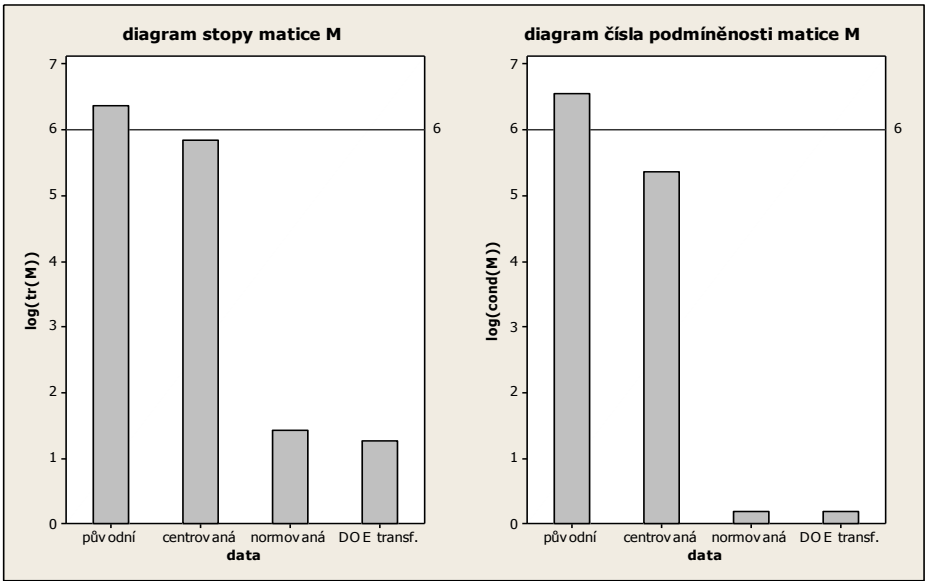
Po transformaci původních dat a výpočtu kritérií v prostředí *Minitab, Statistica* a *Matlab* obdržíme naprosto stejné výsledky, které jsou přehledně uspořádány v tab. 2 (kde jsou tučným písmem označeny hodnoty kritérií, které překročily kritické hodnoty a signalizují narušení *ortogonality* a *nekorelovanosti* plánu). Pro test Farrara a Glaubera je kritická hodnota $\chi^2_{0,05}(3) = 7,815$ a pro F-test je $F_{0,05}(2,3) = 9,552$.

kritéria		původní	centrovaná	mormovaná	DOE transf.
det(**M**)		$\mathbf{1,232 \cdot 10^{12}}$	$\mathbf{1,232 \cdot 10^{12}}$	$1,944 \cdot 10^3$	384
tr(**M**)		$\mathbf{2,328 \cdot 10^{6}}$	$6,89 \cdot 10^5$	27	18
tr(**V**)		**2,059**	0,677	0,611	0,917
cond(**M**)		$\mathbf{3,641 \cdot 10^{6}}$	$2,359 \cdot 10^5$	1,5	1,5
cond(**R**)		10,371	10,371	10,371	10,371
det(**R**)		0,321	0,321	0,321	0,321
FG		3,6	3,6	3,6	3,6
VIF	φ	**2,121**	1	1	1
	t	**2,000**	1	1	1
	$\varphi \cdot t$	**3,121**	1	1	1
FD	φ	1,682	0	0	0
	t	1,500	0	0	0
	$\varphi \cdot t$	3,182	0	0	0

Tabulka 2. Tabulka kritérií multikolinearity a korelovanosti dat.

Hodnoty a porovnání ukazatelů *stopa matice* **M**, jakož i číslo podmíněnosti matice **M** jsou pro původní a transformovaná data graficky (v logaritmickém měřítku) znázorněny na obr. 2:

Z obr. 2 je zřejmé, že *centrování* částečně (a snad i dostatečně) snižuje hodnoty stopy matice **M** a číslo podmíněnosti. Výrazné (téměř o 5 řádů) je však toto snížení u dalších způsobů transformace, tj. u *normování* a *DOE transformace*.



Obrázek 2. Sloupcový graf stopy a čísla podmíněnosti matice **M** pro analyzovaná data.

V tab. 3 můžeme sledovat, jak se výrazně mění velikost vlastních čísel, jestliže použijeme na původní data různé transformace. A nakonec je vidět to nejdůležitější - jak jednotlivé transformace ovlivňují významnost odhadů jednotlivých regresních parametrů.

		původní	centrovaná	normovaná	DOE transf.
vlastní čísla matice **M**	λ_1	$0,599$	6	6	6
	λ_2	$3,699$	$1,96$	6	4
	λ_3	$2.928 \cdot 10^5$	$4.624 \cdot 10^5$	6	4
	λ_4	$2.035 \cdot 10^6$	$2.266 \cdot 10^5$	9	4
p-hodnoty regresních koef.	b_0	$0,0028$	$0,0031$	$0,0000$	$0,0003$
	b_1	$0,0047$	$0,0071$	$0,0022$	$0,0016$
	b_2	$\mathbf{0,2348}$	$0,0077$	$0,0204$	$0,0156$
	b_3	0.0123	0.0123	0.0123	0.0123

Tabulka 3. Vlastní čísla matice **M** a p-hodnoty regresních koeficientů modelu (3.2).

6.2. Hodnocení kritérií a způsobů řešení

Z předchozích výsledků je jasné, že *neortogonalitu* (nekolmost) a *korelovanost* (*multikolinearitu* a *špatnou podmíněnost*) matice plánu s *původními* proměnnými:

- signalizovala (pouze) tři kritéria: *inflační faktory VIF* (větší než 1), *stopa matice* $\mathbf{M}$, tj. $\mathrm{tr}(\mathbf{M})$ a *číslo podmíněnosti matic* $\mathbf{M}$, které překročily kritickou mez 10^6. Je však třeba říci, že *malé* číslo podmíněnosti sice zaručuje správný numerický výsledek, avšak *velké* číslo pouze *může* (ale nemusí) způsobit numerické problémy tzn. malé číslo podmíněnosti je pouze *postačující*, ale nikoliv nutná podmínka k numerické stabilitě a statistické správnosti výsledku výpočtu,
- nejvýrazněji (pro techniky a technology) signalizovala *nesprávná* a nevěrohodná *statistická nevýznamnost* vlivu *teploty tváření* na přetvárný odpor. Jako *způsoby řešení* uvedeného problému se rýsují následující:
- použití metody *DOE*, která zajišťuje *ortogonalitu* a dobrou *numerickou podmíněnost* (malá čísla podmíněnosti díky transformovaným, normovaným, tzv. kódovaným datům) plánu experimentu s dopadem na *statistickou korektnost* a *spolehlivost* závěrů,
- při použití *regresní analýzy* (např. díky nemožnosti použít metodu DOE s příslušným SW) pro objektivní a spolehlivé rozpoznání uvedených nectností, a tím i možnosti výskytu numericko-statistické nekorektnosti indikace nevýznamnosti (některých) regresních koeficientů, je nutné a vhodné *porovnat* tuto významnost (pomocí t-statistik či p-hodnot) pro *modely* s *původními* a *transformovanými* daty. Výraznější nesouhlas (významnost versus nevýznamnost) pak signalizuje špatnou podmíněnost plánu. Hodnocení statistické významnosti koeficientů modelu je korektní pouze u modelu s *transformovanými* proměnnými, přičemž číselné hodnoty samotných regresních koeficientů jsou správné i pro model s původními proměnnými.

Obecně je zde viditelné, že doporučené způsoby *lineární transformace* původních dat vedou k odstranění multikolinearity díky ortogonalitě a nekorelovanosti plánů s *transformovanými* proměnnými a také k numericky i statisticky věrohodným výsledkům.

7. Statistický důsledek neortogonality

Při operacích se špatně podmíněnými maticemi rostou numerické chyby díky nutnému zaokrouhlení („ořezání“) čísel při omezené délce slova počítače. To se zvláště projeví při *inverzi momentové matice* $\mathbf{M}$, ze které následně vychází výpočet rozptylů i směrodatných odchylek regresních koeficientů a příslušné t-testy jejich významnosti.

Z výše uvedeného vztahu (4.1) pro výpočet rozptylu regresních koeficientů je vidět *přímý dopad nepřesně určené inverzní matice* (díky špatné podmíněnosti momentové matice vycházející z matice plánu) na následně nepřesně stanovený (většinou nadhodnocený) *rozptyl regresního koeficientu.*

Špatná podmíněnost momentové matice *může* být způsobena až o $2,5$ řádu *rozdílnými hodnotami* (úrovněmi) základních faktorů v matici plánu (průměrná hodnota

$\varphi = 0,7$, střední hodnota $t = 360\,^{\circ}\mathrm{C}$). Právě díky *nadhodnocení* odhadu směrodatné odchylky u regresního koeficientu teploty tváření t došlo k *podhodnocení* příslušné t-statistiky a nepřekročení kritické hodnoty s dopadem na nesprávnou signalizaci statistické nevýznamnosti tohoto vlivu (faktoru).

8. Závěr

Na základě matematicko-statistického rozboru jednoduchého „technologického" příkladu s lineárním modelem a interakcí faktorů, kde se pro *původní* data nečekaně vyskytl numericko-statistický problém nesprávné indikace *statistické nevýznamnosti* jednoho regresního koeficientu technologicky i fyzikálně vlivného faktoru, a to *teploty tváření*, lze vyslovit *závěry* a *doporučení*:

- Díky ortogonalitě (a malému číslu podmíněnosti) plánu je možné *v jednom kroku* (jednorázově) správně (korektně) stanovit statisticky významné a nevýznamné faktory. Užití *regresní analýzy* pro *původní* (netransformované) proměnné může (díky *neortogonalitě* plánu) poskytnout statisticky i fyzikálně, či technologicky *nesprávné* hodnocení nevýznamnosti (některých) regresních koeficientů (v analyzovaném příkladu to byl vliv, tj. regresní koeficient *teploty tváření*), přičemž však jejich hodnoty jsou numericky správné. Proto pouze regresní analýza s *transformovanými* proměnnými poskytne i statisticky *správné* výsledky,
- pokud je experiment sice *aktivní*, plánovaný, ale *ne podle metodiky DOE*, pak jeho vyhodnocení pomocí *regresní analýzy* poskytne správné hodnoty regresních koeficientů faktorů, čísla podmíněnosti jsou (většinou) malá, ale díky neortogonalitě plánu a určité multikolinearitě faktorů dochází ke statistické nekorektnosti jejich hodnocení, tj. k nesprávné indikaci statistické nevýznamnosti fyzikálně či technologicky vlivných faktorů, a to jak pro jejich *původní*, tak i pro *transformované* hodnoty. Tato skutečnost znamená, že není k dispozici žádná statisticky korektní metoda vyhodnocení experimentu, ale pouze metoda numericky korektní, kterou je *regresní analýza* (analýzu metodou DOE nelze v tomto případě použít),
- tzv. *pasivní* experiment je nejhorší variantou pro analýzu: plán je (zásadně) neortogonální, faktory jsou kolineární, podmíněnost je velká, a tak lze (asi výrazněji a častěji pro větší rozsah dat a větší počet faktorů) očekávat při analýze pomocí *regresní analýzy* jak numerickou nestabilitu (díky numerickým chybám a jejich kumulaci), tj. *numericky* nepřesné a až *nesprávné* hodnoty (některých) regresních koeficientů, tak i jejich nesprávně indikovanou *statistickou nevýznamnost*, což znamená statistickou nekorektnost výsledků,
- metoda *DOE* poskytuje *spolehlivé* a *efektivní* (optimální) přístupy k návrhu a analýze přínosných experimentů, jako i *numericky* a *statisticky* korektní výsledky. Proto je velice potřebné, aby vědci, výzkumníci, technici i technologové, ale i studenti tuto metodiku znali a zásadně ji používali,
- protože všechny tři nejpoužívanější SW *Matlab, Minitab, Statistica* poskytly shodně špatné výsledky při stanovení spolehlivosti odhadu regresních parametrů pro původní hodnoty, doporučuje se opatrnost při aplikaci výsledků regrese bez transformace.

Reference

[1] V. Garaj, I. Šujan: *Ekonometria*, ALFA/SNTL, Bratislava, 1980.
[2] P. Hebák, J. Hustopecký: *Vícerozměrné statistické metody s aplikacemi*, SNTL/ALFA, Praha, 1987.
[3] B. Maroš, T. Trávníček: *Plánování experimentu*, Sborník 5th International Conference *Aplimat* 2006, Bratislava, Ústav matematiky FSI STU Bratislava, 2006.
[4] M. Meloun, J. Militký: *Statistické zpracování experimentálních dat*, PLUS, Praha, 1994.
[5] J. Tošenovský, D. Noskievičová: *Statistické metody pro zlepšování jakosti*, Montanex, Ostrava, 2000.
[6] K. Zvára: *Regrese*, 1. vyd., Matfyzpress MFF UK Praha, Praha, 2008.

Bohumil Maroš, Ústav matematiky, Fakulta strojního inženýrství, Vysoké učení technické v Brně, Technická 2, 616 69 Brno, Česká republika,
e-mail: `maros@fme.vutbr.cz`

ÚLOHA O SETKÁNÍ A JEJÍ ZOBECNĚNÍ

LUCIE KREUZIGEROVÁ a ONDŘEJ VENCÁLEK

Abstrakt. Článek pojednává o klasické úloze o setkání dvou přátel, k jejímuž řešení se používá tzv. geometrické pravděpodobnosti. Úlohu zobecňujeme a řešíme pro $k \geq 2$ přátel. Kromě exaktního řešení je ukázáno i řešení simulační, dostupné i lidem bez širšího matematického vzdělání.

1. Úvod

S *úlohou o setkání* se dnes setká prakticky každý student základního vysokoškolského kurzu pravděpodobnosti obsahujícího pojednání o tzv. geometrické pravděpodobnosti. Ať vezmeme skripta používaná v Brně [1], v Liberci [2], v Olomouci [3], v Ostravě [4], v Praze [5] či v jiném městě, úlohu o setkání tam v kapitole o geometrické pravděpodobnosti s velkou pravděpodobností najdeme. Někdy jde o setkání dvou přátel, jindy to může být setkání dvou vlaků či lodí. Podstata úlohy je však stále stejná. My zde úlohu o setkání formulujeme podobně, jako to činí Anděl v monografii [6]:

Dva přátelé (říkejme jim pan X a pan Y) se domluvili, že přijdou na určité místo v době mezi polednem a jednou hodinou odpoledne. Na místo přijde v tomto časovém intervalu každý z nich zcela náhodně a nezávisle na příchodu toho druhého. Oba jsou připraveni čekat 10 minut, ale ne déle než do jedné hodiny odpoledne. Jaká je pravděpodobnost, že se za těchto podmínek skutečně sejdou?

V tomto článku se zamyslíme nad situací, kdy se mají (či nemají) potkat více než dva přátelé. Představme si, že se na schůzce dohodli tři přátelé. Jaká je nyní pravděpodobnost, že se všichni setkají? A jaká je pravděpodobnost, že se setkají aspoň dva z nich? A jaká je naopak pravděpodobnost, že se nesetkají vůbec? Lze si přitom představit i situace, kdy je žádoucí, aby k žádnému setkání nedošlo. Jde například o situaci, kdy se tři spolubydlící probouzejí nezávisle na sobě mezi 6:00 a 7:00 ráno a poté se odeberou do koupelny, kde se každý z nich zdrží 10 minut. Zajímá nás, jaká je pravděpodobnost, že ani jeden z nich nebude muset čekat. Řešíme tedy spíše „úlohu o nesetkání“.

Úlohu o (ne)setkání pro více než dva lidi řešíme v článku nejprve exaktně, poté své výsledky srovnáme se simulační studií, kterou je schopen realizovat i student, který neumí integrovat (pokud ovšem umí simulovat).

2010 *MSC.* Primární 60A05, 60D05.

Klíčová slova. geometrická pravděpodobnost, setkání, simulace.

Práce byla podporována projektem OP Vzdělávání pro konkurenceschopnost (CZ.1.07/2.3.00/20.0170).

2. Řešení úlohy o setkání

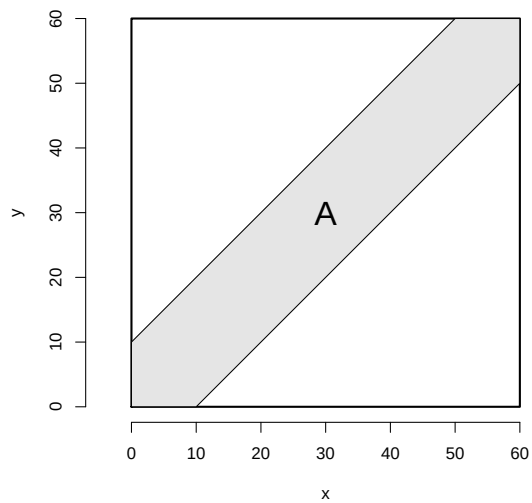
Věnujme se nejprve řešení standardní úlohy o setkání, k němuž se využívá tzv. geometrická pravděpodobnost. Symbolem x označme čas příchodu pána X vyjádřený v nějakých časových jednotkách, například v minutách, uplynulých od poledne. Je třeba si uvědomit, že čas je spojitá veličina, bez ohledu na to, v jakých časových jednotkách je měřen. Proto x může být libovolné reálné číslo z intervalu $[0, 60]$. Žádná hodnota přitom není preferována (není více pravděpodobná než jiné hodnoty). Elementární jev, že pán X přijde v čase x a pán Y v čase y budeme značit (x, y). Množinu všech elementárních jevů, tradičně označovanou symbolem Ω, tedy můžeme zapsat následovně:

$$\Omega = \{(x, y) : x \in [0, 60], y \in [0, 60]\} = [0, 60] \times [0, 60].$$

Jev, jehož pravděpodobnost chceme určit, tedy jev, že se pánové X a Y setkají, zapíšeme následovně:

$$A = \{(x, y) \in \Omega : |x - y| < 10\}.$$

Graficky je celá situace znázorněna na Obrázku 1.



Obrázek 1. Schéma řešení úlohy o setkání. Doba příchodu pána X na x-ové ose, doba příchodu pána Y na y-ové ose. Jev setkání A vyznačen šedě..

Pravděpodobnost jevu A můžeme spočítat jako podíl

$$P(A) = \frac{\lambda(A)}{\lambda(\Omega)},$$

kde symbol λ značí (v tomto případě dvourozměrnou) Lebesgueovu míru. Pravděpodobnost jevu A tedy vypočteme jako podíl šedé plochy při diagonále ku ploše

celého čtverce:

$$P(A) = \frac{60^2 - 50^2}{60^2} = \frac{11}{36} \doteq 0,306.$$

Pravděpodobnost, že se dva zmínění pánové setkají, je tedy přibližně 30 %. Naopak k „nesetkání“ dojde s pravděpodobností téměř 70 %.

3. Do hry vstupuje pán Z

Představme si nyní, že se s pány X a Y chce setkat ještě pán Z. Zde může nastat vícero situací: buďto se sejdou všichni, nebo se sejdou jen dva z nich a třetího přitom minou, nebo se nesetkají vůbec, tedy žádný z pánů nepotká žádného ze svých přátel. Položme si otázku, jaká je pravděpodobnost, že dojde právě k této situaci, tedy že dojde k „nesetkání“.

Princip řešení úlohy je stejný. Prostorem elementárních jevů je nyní množina

$$\Omega = [0, 60]^3,$$

jeho prvky značíme (x, y, z). Aspoň dva ze tří pánů se setkají, pokud platí aspoň jedna z následujících nerovností: $|x - y| < 10$, $|x - z| < 10$, $|y - z| < 10$. V opačném případě dojde k nesetkání. Tento jev tedy můžeme popsat následovně:

$$N = \{(x, y, z) \in \Omega : \min(|x - y|, |x - z|, |y - z|) > 10\}.$$

K určení pravděpodobnosti nesetkání tedy stačí vypočíst

$$P(N) = \frac{\lambda(N)}{\lambda(\Omega)}.$$

Nakreslit obrázek popisující tuto situaci již bude trochu komplikovanější (obrázek je trojrozměrný) a výpočet vyžaduje buďto velmi dobrou geometrickou představivost nebo schopnost integrovat.

Pro větší obecnost budeme uvažovat, že délka časového úseku, v němž se přátelé mají (nemají) setkat je d (v našem případě $d = 60$ minut) a délka, po kterou jsou ochotni čekat je c (v našem případě $c = 10$ minut). Budeme předpokládat, že $d > 2c$, neboť v opačném případě by k nesetkání tří osob vůbec nemohlo dojít.

Situaci si značně zjednodušíme, pokud budeme předpokládat nějaké uspořádání hodnot x, y a z, například $x < y < z$. Toto uspořádání je stejně pravděpodobné, jako kterékoliv jiné uspořádání, tedy jeho pravděpodobnost je $1/(3!)$. Lebesgueovu míru části množiny N, pro níž je splněna podmínka $x < y < z$ pak vypočteme jako integrál

$$\int_0^{d-2c} \int_{x+c}^{d-c} \int_{y+c}^{d} \mathrm{d}z\, \mathrm{d}y\, \mathrm{d}x = (d - 2c)^3/6,$$

tedy Lebesgueova míra celé množiny N bude rovna $(d - 2c)^3$ a pravděpodobnost jevu N můžeme vypočíst jako podíl

$$P(N) = \frac{(d - 2c)^3}{d^3} = \left(1 - 2\frac{c}{d}\right)^3.$$

Povšimněme si, že výsledná pravděpodobnost záleží jen na poměru c/d, tedy na tom, jakou část z celkové doby určené na setkání jsou jednotliví lidé ochotni čekat.

Při větším počtu osob, které se mají (nemají) setkat, lze postupovat obdobně. Pro $2 \leq k \leq d/c$ osob platí, že pravděpodobnost jejich nesetkání je rovna

$$P(N) = \left(1 - (k-1)\frac{c}{d}\right)^k.$$

Vrátíme-li se k situaci, kdy $d = 60$ a $c = 10$, můžeme postupně uvažovat o 2 až 6 osobách. Pravděpodobnosti jejich nesetkání jsou shrnuty v tabulce 1.

k	2	3	4	5	6
P(N)	25/36	8/27	1/16	1/243	1/46656

Tabulka 1. Pravděpodobnosti nesetkání pro různé počty osob.

Zatímco při poměru $c/d = 1/6$ se dvě osoby spíše nepotkají (s pravděp. 70 %), u tří osob už nesetkání nastává s pravděpodobností jen asi 30 %, u čtyř s pravděpodobností asi 6 % a u více než čtyř už je pravděpodobnost velmi malá. Pro více než 6 osob už je pravděpodobnost nesetkání nulová.

4. Setkání a nesetkání

V předešlé sekci jsme se zabývali „nesetkáním“, tedy situací, kdy ani jedna z osob, které se mají setkat, nepotká žádnou další osobu. Opakem tohoto jevu je jev, kdy dojde k sektání alespoň dvou osob. Pravděpodobnost tohoto jevu je tedy doplňkem do jedničky z výše uvedené pravděpodobnosti nesetkání.

Mohla by nás však zajímat i pravděpodobnost jevu, že se sejdou všichni přátelé. Zde je dobré si uvědomit, jak ve skutečnosti vypadá situace, kdy se dva z přátel opravdu sejdou. Můžeme uvažovat o různých scénářích v chování těchto osob. Pokud je bodem setkání například místo, kde se mohou občerstvit, budou zřejmě ochotni se společně občerstvovat a čekat na další účastníky schůzky delší dobu než jen stanovených 10 minut (třeba až do konce dohodnuté či otevírací doby).

Představme si však, že každý z účastníků má na setkání vyhrazeno právě jen 10 minut. Jaká je pravděpodobnost, že se setkají všichni? Jev setkání všech účastníků můžeme pro tři osoby zapsat následovně:

$$S = \{(x, y, z) \in \Omega : \max(|x-y|, |x-z|, |y-z|) < 10\}.$$

Pravděpodobnost tohoto jevu opět vypočteme pomocí podílu

$$P(S) = \frac{\lambda(S)}{\lambda(\Omega)}.$$

Znovu si pomůžeme předpokladem $x < y < z$. Lebesgueovu míru části množiny S, pro níž je splněna podmínka $x < y < z$, pak vypočteme jako součet integrálů

$$\int_0^{d-c} \int_x^{x+c} \int_y^{x+c} \mathrm{d}z\,\mathrm{d}y\,\mathrm{d}x + \int_{d-c}^{d} \int_x^{d} \int_y^{d} \mathrm{d}z\,\mathrm{d}y\,\mathrm{d}x = c^2(3d-2c)/6.$$

Pravděpodobnost jevu S tedy vypočteme následovně:

$$P(S) = \frac{c^2(3d-2c)}{d^3} = 3\,(c/d)^2 - 2\,(c/d)^3.$$

Výsledná pravděpodobnost opět záleží jen na poměru c/d, tedy na tom, jakou část z celkové doby určené na setkání jsou jednotliví lidé ochotni čekat.

Při větším počtu osob, které se mají setkat, lze postupovat obdobně. Za zmínku snad stojí skutečnost, že nyní může k setkání všech osob dojít pro libovolně velký počet osob (bez omezení), i když pochopitelně pravděpodobnost bude s rostoucím počtem účastníků setkání klesat. Pro $k \geq 2$ osob platí, že pravděpodobnost setkání všech je rovna

$$P(S) = k\,(c/d)^{k-1} - (k-1)\,(c/d)^{k}\,.$$

V situaci, kdy $d = 60$ a $c = 10$, jsou pravděpodobnosti setkání všech osob shrnuty v tabulce 2.

k	2	3	4	5	...	10	...
P(S)	11/36	2/27	7/432	13/3 888	...	17/20 155 392	...

Tabulka 2. Pravděpodobnosti setkání pro různé počty osob.

Při poměru $c/d = 1/6$ se dvě osoby setkají s pravděpodobností přibližně 30 % (11/36), tři osoby s pravděpodobností jen asi 7 % (2/27). Čím více je osob, které se mají setkat, tím je pravděpodobnost, že se potkají všechny, menší. Je však nenulová pro libovolně velký (konečný) počet osob.

5. Simulace nesetkání

Úloha o (ne)setkání je atraktivní také v tom, že se dá velmi snadno řešit simulačně. Pomocí generátoru (pseudo)náhodných čísel opakovaně generujeme z rovnoměrného rozdělení časy příchodů jednotlivých pánů (nezávisle na sobě) a poté vzhodnocujeme, zda došlo k (ne)setkání. Relativní četnost situací, kdy došlo k „nesektání“, je pak odhadem pravděpodobnosti tohoto jevu.

Ukažme si kód takovéto simulace realizovaný v prostředí statistického softwaru R. Nejprve se věnujme situaci, kdy se (ne)mají setkat jen dvě osoby:

```
opak = 100000  # pocet opakovani simulace
d    = 60      # delka casoveho useku
c    = 10      # doba cekani

x1 = runif(opak,0,d)  # casy prichodu pana X
x2 = runif(opak,0,d)  # casy prichodu pana Y

pravdep = sum(abs(x1-x2)>c)/opak
print(pravdep)
```

Naše simulace při 100 000 opakováních (výsledek máme prakticky ihned) dospěla k odhadu pravděpodobnosti nesetkání 0,69414, což je dosti blízko skutečné hodnotě 0,69444. Svou nejistotu v odhadu přitom můžeme snadno vyjádřit třeba pomocí 95%-ho (přibližného) konfidenčního intervalu. K jeho sestrojení použijeme centrální limitní větu. V našem případě má tento konfidenční interval podobu $(0,691284; 0,696996)$. Vidíme tedy, že již třetím desetinným místem si nemůžeme

být moc jisti. Zpřesnění o řád si přitom vyžádá stokrát více opakování simulace. Zde je také vidět, že simulací budeme jen velmi těžko odhadovat nízké pravděpodobnosti, třeba pravděpodobnost nesetkání pro 5 nebo 6 osob (když $c/d = 1/6$).

Věnujme se nyní úloze (ne)setkání více osob, přičemž chceme simulaci naprogramovat tak, aby fungovala pro libovolný počet $k \leq d/c$ lidí, kteří se mají (nemají) setkat. Je třeba si uvědomit, že k vyhodnocení, zda došlo k nesetkání, obecně potřebujeme porovnat všechny dvojice časů příchodů. Těch je $k(k-1)/2$. Můžeme sice nejprve časy uspořádat a poté provést jen $k-1$ porovnání, uvědomme si však, že opakované seřazení hodnot od nejmenší k největší už je poměrně náročné na čas.

Přímočará simulace má sice přehledný zápis, ale její výpočet je poněkud zdlouhavý:

```
opak = 100000  # pocet opakovani simulace
d    = 60      # delka casoveho useku
c    = 10      # doba cekani
k    = 4       # pocet osob

nesetkajise = 0
for (i in 1:opak) {
  prichody = runif(k,0,d)
  prichody = sort(prichody)
  rozdily  = diff(prichody)
  if (min(rozdily) > c) nesetkajise = nesetkajise + 1 }

pravdep = nesetkajise/opak
print(pravdep)
```

Tato simulace odhadne pravděpodobnost nesetkání v případě 4 osob, které se mají setkat během jedné hodiny, když každá z osob je ochotna čekat 10 minut. Na počítači s procesorem 2.66GHz Intel Core 2 trvá při 100 000 opakováních asi 13 sekund. Oproti tomu následující verze využívá dobrých vlastností softwaru R při práci s vektory. Při stejných parametrech trvá jen asi 1 sekundu. Úloha o nesetkání tedy nemusí být napříště spojována jen s geometrickou pravděpodobností, ale stejně dobře může posloužit i k zdokonalování programátorských schopností studentů.

```
# matice pro ulozeni casu prichodu:
x = matrix(ncol=k,nrow=opak)

# generovani casu prichodu:
for (i in 1:k)   x[,i] = runif(opak,0,d)

# porovnani casu prichodu:
rozdil = matrix(ncol=choose(k,2),nrow=opak)
index = 1
for (i in 1:(k-1))
  for (j in (i+1):k)
    {rozdil[,index] = abs(x[,i]-x[,j])
```

```
    index = index+1   }

m = apply(rozdil,1,min)
pravdep = sum(m>c)/opak
print(pravdep)
```

6. Závěr

Úloha o sektání dvou přátel se v různých obměnách objevuje snad až příliš často jako ukázkový příklad využití tzv. geometrické pravděpodobnosti. V článku jsme ukázali možnost, jak úlohu zobecnit pro setkání vícero osob. Nalezli jsme exaktní tvar řešení této úlohy a poukázali na možnost využití simulace k nalezení přibližného řešení. Úloha je zajímavá i z pohledu programátorského.

Reference

[1] M. Budíková, Š. Mikoláš, P. Osecký: *Teorie Pravděpodobnosti a matematická statistika – sbírka příkladů*, 3. vydání, Masarykova univerzita, Brno, 2007.

[2] L. Moc: *Cvičení z inženýrské statistiky (řešené příklady)*, Technická univerzita v Liberci, 2006.

[3] K. Hron, P. Kunderová: *Základy počtu pravděpodobnosti a metod matematické statistiky*, Univerzita Palackého, Olomouc, 2013.

[4] P. Otipka, V. Šmajstrla: *Pravděpodobnost a statistika [elektronický zdroj]*, Technická univerzita Ostrava, 2008.

[5] V. Dupač, M. Hušková: *Pravděpodobnost a matematická statistika*, Karolinum, Praha, 2003.

[6] J. Anděl: *Matematika náhody*, Matfyzpress, Praha, 2003.

Lucie Kreuzigerová, Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc, Česká republika, *e-mail*: `lucie.kreuzigerova01@upol.cz`

Ondřej Vencálek, Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc, Česká republika, *e-mail*: `ondrej.vencalek@upol.cz`