Informační Bulletin
České Statistické Společnosti
č. 2. listopad 2000, ročník 11

Využití statistických metod ve výrobě polovodičů

Milan Hutyra Tesla Sezam, a. s., Rožnov p. R.,
VŠB - TU Ostrava

1. Úvod

Tesla Sezam, a.s. je výrobce bipolárních integrovaných obvodů a dalších součástek působící v ČR. Tesla Sezam, a.s. byla založena 1. 5. 1992 jako jedna z nástupnických organizací podniku Tesla Rožnov v oblasti výroby polovodičových prvků. V průběhu několika let se společnost vypracovala z bankrotujícího podniku, navrženého státem k likvidaci, na prosperující společnost s dlouhodobým výrobním programem, který je úspěšně exportován na světové trhy.

Ve výrobním programu dnes již neexistujícího podniku Tesla Rožnov byl v roce 1989 kromě celé řady výrobků (jako např. černobílé a barevné obrazovky, základní materiály pro výrobu vakuových a polovodičových součástek, nástroje a jednoúčelová technologická zařízení) i rozsáhlý sortiment polovodičových součástek vyráběných bipolární technologií.

V důsledku zániku trhu RVHP, se kterým byla Tesla Rožnov existenčně svázána a postupného rozpadu českého elektronického průmyslu po roce 1990 nebyl zajištěn odbyt vyráběné produkce. Její uplatnění na světovém trhu bylo nemožné, neboť díky extenzivnímu vývoji a isolaci od světového trhu se jednalo o výrobky konkurence neschopné zejména z důvodu vysokých výrobních nákladů a stagnaci úrovně jakosti.

Vedení společnosti Tesla Sezam, a.s. si uvědomilo, že bez nalezení solidního strategického partnera, který je schopen rozpoznat naše možnosti a poskytne příležitost pro využití existujícího potenciálu dojde k zániku výroby polovodičových součástek v ČR. Po téměř třech letech pokusů a složitých vyjednávání byla v roce 1993 úspěšně ukončena kvalifikace čipu

integrovaného stabilizátoru napětí z produkce Tesla Sezam, a.s. firmou Motorola, a tím se otevřela možnost spolupráce Tesla Sezam, a.s. s předním světovým výrobcem polovodičových součástek. Podmínky spolupráce na prvním místě uváděly požadavek zabezpečení minimálně stejné jakosti výroby Tesla Sezam, a.s. jako je dosahováno v závodech Motorola.

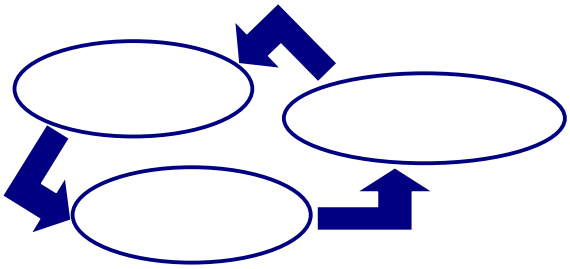
2. Jakost SIX SIGMA

Za základní cíl společnosti Motorola je považována „Úplná spokojenost zákazníka“. Pro dosažení tohoto cíle definovala Motorola klíčové iniciativy – jak to budeme dělat. Na prvním místě mezi těmito iniciativami je jakost SIX SIGMA. Požadavek zabezpečení stejné jakosti výroby Tesla Sezam, a.s. jaké je dosahováno v závodech Motorola si vynutil přijmout iniciativu SIX SIGMA za základní strategii společnosti Tesla Sezam, a.s.

Jakost SIX SIGMA je filosofie zabezpečování jakosti, která spočívá v tom, že jakost výrobku dodávaného zákazníkovi je zajišťována procesy u dodavatele, které prakticky vylučují vznik neshodného výrobku. Strategie SIX SIGMA vyžaduje dosažení takové míry způsobilosti procesu, kdy střední hodnota charakteristiky výrobního procesu je vzdálena 6σ (σ - směrodatná odchylka charakteristiky procesu) od obou specifikačních mezí reprezentujících požadavky zákazníka. Plocha křivky hustoty pravděpodobnosti mezi oběma specifikačními mezemi představuje podíl výrobků vyhovujícím požadavkům specifikace. Plocha mimo tyto meze představuje podíl neshodných výrobků. Pro způsobilost na úrovni 6σ je podíl neshodných výrobků jen 0,002 ppm. Toto je prakticky proces s nulovou úrovní vad (zero defect). Dlouhodobě je ale téměř nemožné udržet střední hodnotu charakteristiky procesu přesně ve středu tolerančního pole. Běžný je její posun o $\pm 1{,}5\sigma$ od ideální hodnoty. I při tomto posunu lze očekávat při úrovni způsobilosti 6σ pouze 3,4 neshodných výrobků na milión vyrobených.

3. Výroba polovodičových součástek a její řízení

Výroba polovodičových součástek je hromadná výroba. Denní produkce se pohybuje v řádu miliónů vyráběných součástek. Svým charakterem patří mezi náročné a složité procesy. Např. výrobní postup výroby systémů integrovaného obvodu střední hustoty integrace na křemíkových substrátech se skládá z více než 250 na sebe navazujících operací. Jestliže celková výtěžnost celého procesu musí být nad 95%, což je současná úroveň dosahovaná předními světovými výrobci polovodičových součástek, pak podíl neshodných výrobků u jednotlivých dílčích operací musí být menší než 200 ppm. Dosažení takové úrovně výroby nelze zabezpečit bez komplexního řízení všech procesů. Komplexně řídit proces znamená proces poznat, proces řídit a proces zdokonalovat v uzavřené smyčce.



Hromadnost a složitost výroby polovodičových součástek vyžaduje využívat při komplexním řízení výroby statistických metod. Odpověď na otázku proč je nutné využívat statistické metody je jednoduchá. Protože využívání statistických metod šetří peníze a tím umožňuje dosažení nízkých výrobních nákladů.

3.1. Poznání procesu

Poznat proces znamená zjistit jeho způsobilost. Jako metodu poznání jsme proto v Tesla Sezam, a.s. zavedli vyhodnocování způsobilosti jednotlivých procesů. Pod pojmem způsobilost chápeme měřenou, procesu vlastní reprodukovatelnost výrobku prošlého výrobním procesem.

Měřená – to znamená, že způsobilost musí být kvantifikována pomocí údajů (dat) o výrobku prošlého výrobním procesem.

Procesu vlastní – to znamená, že reprodukovatelnost existuje trvale v procesu, je dána procesem a není od něj oddělitelná.

Způsobilost procesu vyhodnocujeme prostřednictvím indexů C_P (potenciál způsobilosti) C_{PK} (index způsobilosti). Oba tyto indexy kvantifikují schopnost procesu vyrábět výrobky, které vyhovují technické nebo zákaznické specifikaci. Princip spočívá v hodnocení poměru požadavků (tedy specifikace) a skutečného stavu procesu.

$$Cp, Cpk = \frac{\text{Co požaduje zákazník}}{\text{Co nabízí proces}}$$

Potenciál způsobilosti C_P udává, jestli je proces vůbec schopen plnit požadavky, tedy být způsobilý. Ale ani vysoká hodnota C_P nemusí ještě zaručovat, že tomu tak skutečně je. Potenciál způsobilosti totiž nebere v úvahu „polohu“ procesu vzhledem k jeho mezním hodnotám. Tuto informaci poskytuje index způsobilosti C_{PK}.

Vlastnímu vyhodnocování způsobilosti procesu předchází studie způsobilosti, která se skládá z etap:

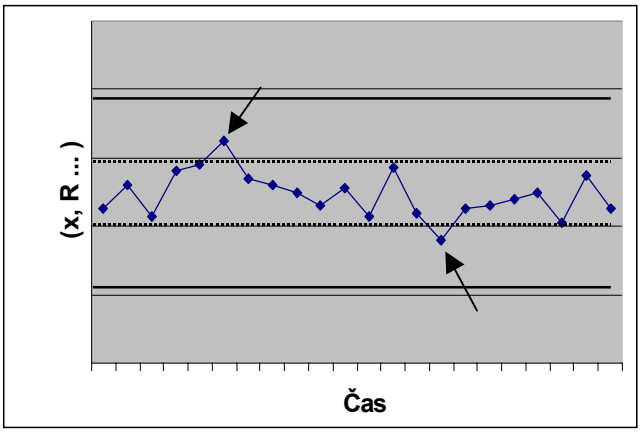
- charakterizace operace
- charakterizace měřících metod
- určení způsobilosti měřících systémů
- vlastní určení způsobilosti procesu (určení normality, test statistické stability, výpočet indexů způsobilosti).

Při provádění studie způsobilosti i vlastním vyhodnocování způsobilosti je využíváno celé řady statistických technik (výpočet výběrových charakteristik, testování statistických hypotéz apod.).
Indexy způsobilosti jsou pravidelně vyhodnocovány. Dosažené výsledky jsou porovnávány s požadovanou způsobilostí (např. pro rok 1999 byla požadována způsobilost procesů na úrovni C_{PK}=1.67). V případě nedosažení jsou definována nápravná opatření.

3.2. Řízení procesu

Jako metodu řízení procesu využíváme v Tesla Sezam, a.s. statistickou regulaci procesu. Účelem statistické regulace výrobního procesu je nastavit a udržet výrobní proces na přípustné a stabilní úrovni tak, aby byla dlouhodobě zajištěna požadovaná jakost vyráběné produkce. Statistická regulace umožňuje včasné rozpoznání poruchy výrobního procesu (působení zvláštních příčin) a zavedení nápravných opatření ještě dříve, než dojde k výrobě neshodného výrobku. Statistická regulace *"nehledá žádné zmetky, ale hlídá proměnlivost"*.

Statistická regulace procesu (SPC) je založena na porovnání výsledků kontrol náhodných výběrů s předem stanovenými optimálními hodnotami polohy a proměnlivosti procesu. Zásadně se tedy statistickou regulací procesu nekontroluje jakost právě vyráběných výrobků, ale zjišťuje se, jak proces probíhá (dobře nebo špatně), a to srovnáním hodnot výběrové charakteristiky s ideálními mezními hodnotami. Tyto ideální mezní hodnoty se nazývají regulační meze. V případě překročení regulačních mezí se realizují předem definovaná nápravná opatření.



Statistickou regulaci máme v Tesla Sezam, a.s. zavedenou nejen pro řízení jednotlivých procesů, ale využíváme ji i pro řízení vstupů do těchto procesů. Pro vedení a vyhodnocování regulačních diagramů využíváme jak klasický způsob s využitím papírových formulářů, tak i vedení a vyhodnocování zabezpečované prostřednictvím softwarové podpory na PC. Účinnost statistické regulace je pravidelně vyhodnocována prostřednictvím ″instability indexu″.

3.3. Zdokonalování procesu

Jako metodu zdokonalování procesu využíváme v Tesla Sezam, a.s. metodu plánovaných experimentů (DOE – Design of Experiment). Plánovaný experiment je založen na vědomých změnách hodnot vstupů procesu (faktorů) za účelem zjištění odpovídajících hodnot výstupů. Dobře připravený a důsledně vykonaný plánovaný experiment poskytne informace o skutečném stavu sledovaného procesu/operace. Tyto informace ve formě dat je potřeba analyzovat tak, aby na jejich základě bylo možné uvedení procesu do stavu, aby pokud možno splňoval požadavky na něj kladené. První etapa analýzy dat získaných při provádění plánovaného experimentu bývá zpravidla zaměřena na analýzu procesu (z relativně velkého počtu faktorů ovlivňujících proces zjistit ty, které „stojí za řeč“). Následuje pak etapa modelování procesu s následnou optimalizací. Cílem DOE je:

- neustále zvyšování způsobilosti procesů
- snižování nákladů.

Pochopení a zavedení metodiky plánovaných experimentů vyžaduje potřebné znalosti. V minulosti tato metodika nebyla zařazena v náplni inženýrského vzdělání. Pro naprostou většinu inženýrů v Tesla Sezam, a.s. byl v roce 1995 plánovaný experiment pojem naprosto neznámý. Proto bylo v roce 1996 zahájeno rozsáhlé školení pro tým vybraných techniků. Tito pracovníci absolvovali za podpory našeho strategického partnera sérii sedmi školení:

- Podstata plánování experimentů
- Plánovaný experiment a jeho příprava
- Typy plánů experimentů

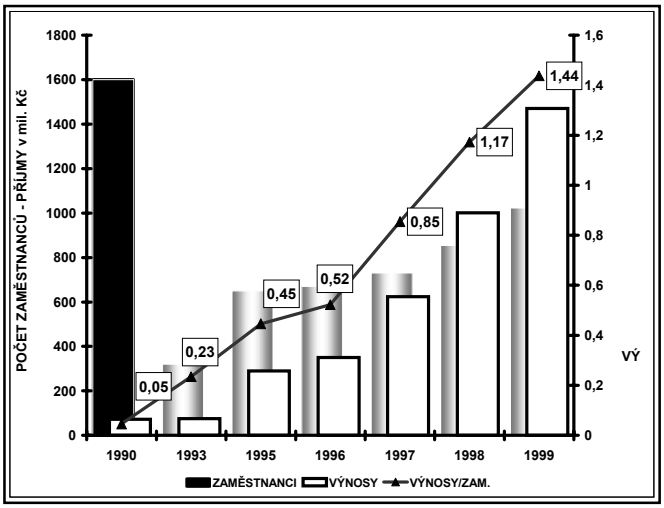
- Statistické techniky
- Odolný návrh
- Taguchiho přístup
- Metoda plošné rezervy

Po teoretické průpravě pak následoval praktický výcvik, kdy pod vedením lektorů vyškolených v Motorola University byly provedeny vzorové experimenty přímo v podmínkách běžících reálných procesů. Tento postup se ukázal být účinným a přispěl k postupnému osvojování metodiky DOE v naší společnosti. V roce 1997 jsme absolvovali zdokonalovací kurz DOE, jehož praktickou náplní bylo vyhodnocení do té doby provedených DOE s ukázkou možnosti případného zlepšení.

Metodika DOE je postupně rozšiřována na stále větší část procesů, průběžně jsou vyškolováni další pracovníci. DOE se tak stává účinným nástrojem využívaným nejen při zdokonalování již běžících procesů, ale i fungujícím nástrojem při charakterizaci nově zaváděných procesů v rámci expanzních projektů Tesla Sezam, a.s. Pojmy jako lineární regrese, mnohonásobná regrese, ANOVA, metoda plošné odezvy a další se pro stále větší počet techniků stávají důvěrně známými názvy statistických metod a technik, které využívají při své každodenní činnosti při zajišťování výroby polovodičových součástek.

4. Závěr

Charakter výroby polovodičových součástek vyžaduje poměrně značný rozsah využívání statistických metod. Praktické využívání statistických metod a technik je jedním z faktorů, které výrazně přispívají k stále se zlepšujícím výsledkům Tesla Sezam, a.s. Můžeme potvrdit, že statistika skutečně šetří peníze.



Nicméně statistické techniky jsou jenom pomůckou, nástrojem pro rozhodování na základě více nebo méně důvěryhodných dat. Statistika rozhodně nemůže nahradit technické poznatky a zkušenosti a již vůbec ne „zdravý rozum", kterým bychom se měli neustále řídit.

Statistika a řízení jakosti podle norem ISO 9000+

Jiří Mrověc

(JM) Jakost - Motivace, Frýdek-Místek, e-mail: jmrovec@iol.cz

Klasický model řízení a zabezpečování jakosti (odpovídá modelu podle ISO 9003:94)

Základním principem je *následná kontrola výrobku*. V konečném důsledku jde o *řízení distribuce dobrých a špatných výrobků*, nikoliv o řízení jakosti. Ta je dána přecházejícími procesy.

Použitá statistická metoda - výběrová kontrola

- Umožňuje kontrolu skupiny výrobků, i když je zkouška destruktivní nebo drahá
- Statistické přejímky - umožňují snížit náklady, vynakládané na získání informací pro rozhodnutí přijmout, nebo zamítnout

Modernější modely řízení a zabezpečování jakosti (odpovídají modelům podle ISO 9002:94 resp. ISO 9001:94)

Základním principem je *preventivní regulace procesu* (je uplatněn procesní přístup, prevence a systémový přístup).

Procesní přístup. Respektuje skutečnost, že parametry výrobku řídíme pouze do té míry, do jaké jsme schopni regulovat parametry procesu a do jaké hloubky rozumíme vztahům mezi parametry procesu a parametry výrobku. Kontrolu výrobku nahrazujeme kontrolou parametrů procesu. Metoda překračuje omezení klasické metody; je možno poskytnout záruky i v těch případech, kdy není proveditelná kontrola parametru výrobku (např. pevnost při creepu).

Prevence je schopnost předpovědět budoucí stav, a pokud se nám předpověď nelíbí, učinit opatření, které umožní dosáhnout příznivějšího stavu.

*Systémový přístup j*e vymezení kompetencí, zajištění a přidělení zdrojů, stanovení postupů, vytvoření zpětné vazby. Formalizovanými nástroji jsou: a) Normalizace a certifikace, b) Komplexní přístup, c) Statistické metody z nichž převládají regulační diagramy (číselná data a znakové proměnné), analýza rozptylu – ANOVA, data mining, a DOE (navrhování experimentů).
+++--

V současné době platné normy ISO 9001, 9002 resp. 9003:94 obsahují článek 20 Statistické metody. Jak je tedy možné, že v řadě případů se v Příručce jakosti českých organizací setkáváme s formulací typu:

"V naší organizaci žádné statistické metody nepoužíváme, ale pokud by si to zákazník přál, ve spolupráci s ním vhodné metody vybereme a zavedeme"?

Domnívám se, že existují nejméně dvě hlavní příčiny tohoto stavu.

- Manažeři, kteří svými postoji blokují uplatnění statistických metod v organizaci, kterou řídí. Obvykle jde o manažery, kteří se s aplikovanou statistikou nesetkali ani v době studií, ani v předcházející praxi. Výsledkem je, že nejenže nemají zkušenosti s jejím využitím, ale navíc se stydí tento stav přiznat, protože by přiznávali nekompetentnost. Důsledkem je, že případná aplikace statistických metod v organizaci, kterou řídí nemá "zákazníka", nikdo ji nechce. A lidé v našich podnicích jsou dost chytří na to, aby nevěnovali pozornost tomu, co zřejmě nepovažují nadřízení za důležité.
- Způsob školení, resp. výcviku v rámci podnikového zvyšování kvalifikace zaměstnanců. Výuka je zaměřena na *zvyšování znalosti* (jako zdroje pro zvyšování výkonnosti), ale bez *změny postoje pracovníků* je snaha o změnu znalostí neúčinná. Významnější než *znalost* "*statistických metod*" (zahrnují *variabilitu, data, stat. nástroje)* je "*statistické myšlení*" (zahrnuje proces, variabilitu, data). Vytvoření této *dovednosti* je prakticky nezávislé na předcházejícím vzdělání, ale v konkrétním případě silně podmíněné znalostí procesu (kterou obvykle nemá lektor).

Proces ➔ Variabilita ➔ Data ➔ Stat. nástroje, Stat. myšlení ➔Stat. metody

Základním úkolem každého výkladu statistického myšlení je vysvětlit, že variabilita procesu je sice v jisté směru nežádoucí, ale jinak zcela přirozeným projevem objektivně existující skutečnosti. Začínat s procesem je pro lidi z praxe přirozené. Znají proces a postupují od známého a konkrétního k méně známému a obecnému. Pokud si osvojí statistické myšlení, změní svůj postoj a následně začnou aktivně uplatňovat statistické metody.

Nelze očekávat rozsáhlou a bezproblémovou aplikaci metod, které jsou postaveny na faktech, jestliže nositelům nepříjemných zpráv jsou stínány hlavy. Na druhé straně aplikace statistických metod může usnadnit seznámení manažerů se souborem tzv. "soft skills", protože výcvik je možno provádět tak, aby přímo podporoval týmovou práci.

Postupy, které se osvědčily při vnitropodnikových seminářích:

(Vysvětlení pojmů intuitivní statistika, popisná statistika a matematická statistika, resp. navrhování experimentů DOE - Taguchiho metodou formou hry)

- ***Intuitivní statistika***

  Sběr a hodnocení dat provádíme bez záznamů, které by byly přístupné jiným osobám; máme je v mysli, v paměti. To však neznamená, že neměříme, nehodnotíme a nečiníme kvalifikovaná rozhodnutí.

  Otázka pro úvodní diskusi:

  Jak se rozhodujeme ráno, při vstávání, kolik času dnes potřebujeme pro přípravu snídaně, na cestu do práce, ...?

  Umožní účastníky semináře vtáhnout do diskuse a na základě známých pojmů vysvětlit pojmy nové.

- ***Popisná statistika***

  Nepoužívá matematické modely. Používá záznamy a charakteristiky: medián (polohy), rozpětí, (rozptýlení), kvantily, grafy (bodové diagramy), run chart.

  Otázka pro úvodní diskusi:

  Kolik máte doma elektromotorů?

  Umožní po záznamu „známého faktu“ a jeho konfrontaci s připraveným seznamem vysvětlit význam záznamu pozorování, resp. vhodného grafického vyjádření.

- ***Matematická statistika***

  Používá matematické modely a celkově se vyznačuje systematickým postupem. Systematickým přístupem, který využívá netriviálních znalostí, je např. organizování pokusu (DOE). Taguchiho přístup k organizování pokusu je "inženýrským přístupem", který je srozumitelný lidem provozů.

  Příkladem pro cvičení je stavba „papírových letadel“ a následné testování jejich letové výkonnosti.

Tento workshop umožňuje s minimálními finančními náklady vysvětlit podstatu rozdělení faktorů na faktory, které jsou řízeny projektantem - operátorem procesu a faktory, které v praxi nejsou pod kontrolou toho, kdo proces řídí. Numerické zpracování získaných experimentálních dat má podpůrný charakter. Workshop usnadňuje pochopení principů statistického myšlení.

Připravovaná norma (ISO/DIS 9001:2000) již nemá článek "Statistické metody", nicméně v odstavci 8.1 Plánování článku 8 Měření, analýza a zlepšování je výslovná zmínka o statistických metodách. V návrhu normy existuje řada dalších požadavků, které mohou vést k aplikaci statistických metod, např. hodnocení spokojenosti zákazníků, hodnocení spokojenosti zaměstnanců. Jinou oblastí aplikace statistických metod představuje metrologie, zvláště oblast "*Analýzy systému měření*", podrobněji viz následující schéma (převzaté z J. Raffaldi and S. Ramsier: 5 Ways to Verify Your Gages, Quality Magazine, March 2000):

Chyba! Chybné propojení.

Závěrem je možno konstatovat, že systém norem pro řízení jakosti poskytuje dobrou příležitost uplatnění pro statistiků, zvláště respektují-li potřeby praxe.

Hodnocení způsobilosti technologického procesu u poloviční tolerance

Josef Tošenovský
VŠB-TU Ostrava, Katedra řízení jakosti

V posledních letech se hodnocení způsobilosti technologického procesu stává pozvolna součástí standardní dokumentace o výrobci. Také řada větších odběratelů žádá konkrétní hodnoty indexů způsobilosti, které jsou v drtivé většině reprezentovány indexem C_{pk} a to i tehdy, když to je naprosto nevhodné. Dodavatel však musí vyhovět svému odběrateli a jiný index k dispozici nemá. O jedné takové situaci pojednává tento příspěvek.

Ve výrobní dokumentaci se dosti časti vyskytují vedle nesymetrické tolerance také poloviční tolerance, které jsou navíc často spojené s nenormálním rozdělením ukazatele kvality. U poloviční tolerance (kdy $USL = T$ resp. $LSL = T$, T je cílová = ideální hodnota) není index C_{pk} schopen

rozlišit různou úroveň procesů, jak ilustruje následující příklad. Proto jsou nutné speciální metody. Je-li například stanovena cílová hodnota a příslušné tolerance $T = 3.5^{+0.010}_{-0.000}$ (takže *LSL* = *T*), vychází pro různé rozptyly a průměry tyto indexy C_{pk}:

a) $\mu = 3.502, 3\sigma = 0.002$:

$$C_{pk} = \min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right) =$$

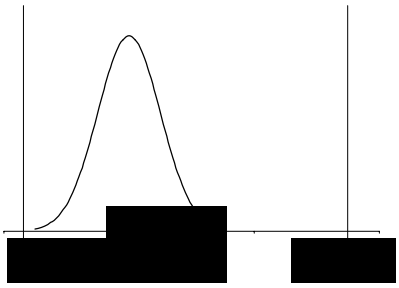
$$= \min\left(\frac{3{,}51-3{,}502}{0{,}002}, \frac{3{,}502-3{,}5}{0{,}002}\right) = \min\left(\frac{0{,}008}{0{,}002}, \frac{0{,}002}{0{,}002}\right) = 1$$

b) $\mu = 3.505, 3\sigma = 0.005$:

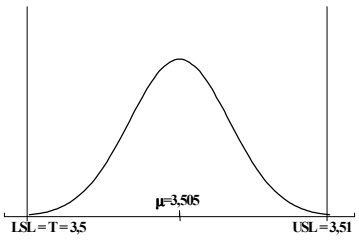
$$C_{pk} = \min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right) =$$

$$= \min\left(\frac{3{,}51-3{,}505}{0{,}005}, \frac{3{,}505-3{,}5}{0{,}005}\right) = \min\left(\frac{0{,}005}{0{,}005}, \frac{0{,}005}{0{,}005}\right) = 1$$

V případě b) je větší rozptyl a také větší vzdálenost μ od *T*, ale $C_{pk} = 1$ je stejné jako v případě a). Oba případy jsou na obr. 1 a 2.



Obr. 1
$C_{pk} = 1$ pro $\mu = 3.502, 3\sigma = 0.002$



Obr. 2
$C_{pk} = 1$ *pro* $\mu = 3.505, 3\sigma = 0.005$

Uvedenou situaci řeší například indexy C_{pp} a C_{pT}. Jsou definovány takto:

a) Pro poloviční toleranci, je-li $USL = T$, je

$$C_{pp} = \frac{\mu - LSL}{\mu - x_{0.00135}}, \qquad C_{pT} = \frac{T - LSL}{T - x_{0.00135}} \tag{1}$$

b) Pro poloviční toleranci, je-li $LSL = T$, je

$$C_{pp} = \frac{USL - \mu}{x_{0.99865} - \mu}, \qquad C_{pT} = \frac{USL - T}{x_{0.99865} - T} \tag{2}$$

Problém s použitím C_{pp} a C_{pT} je v tom, že potřebné kvantily $x_{0,00135}$ a $x_{0,99865}$ lze určit z dat pouze v případě, že rozsah souboru je alespoň 740. Tak velký soubor však většinou nebývá k dispozici. V takovém případě doporučují autoři [2] nahradit kvantily $x_{0,00135}$ resp. $x_{0\cdot 9865}$ hodnotami $x_{\min}$ resp. $x_{\max}$. Použití indexů ukážeme na příkladech.

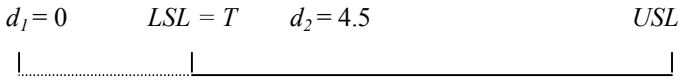
1) Specifikace procesu: $125^{+0.0}_{-2.5}$ tj. $LSL = 122.5, USL = T = 125$,
charakteristiky: $\bar{x} = 123.3$, $x_{\min} = 122.8$, $x_{max} = 123{,}8$.
Má se vypočítat C_{pp} a C_{pT}.

Ze zadání je zřejmé, že tolerance jsou: dolní $d_1 = 2{,}5$ a horní $d_2 = 0$.

$$C_{pp} = \frac{\bar{x} - LSL}{\bar{x} - x_{\min}} = \frac{123.3 - 122.5}{123.3 - 122.8} = 1.6$$

$$C_{pT} = \frac{T - LSL}{T - x_{\min}} = \frac{125 - 122.5}{125 - 122.8} = 1.14$$

2) Specifikace procesu: $209.5^{+4.5}_{-0.0}$ tj. $LSL{=}T = 209.5$, $USL = 214$, $x_{max} = 213{,}5$. Má se vypočítat C_{pT}.



$$C_{pT} = \frac{USL - T}{x_{\max} - T} = \frac{214 - 209.5}{213.2 - 209.5} = 1.22$$

Existují i další možnosti hodnocení způsobilosti u polovičních tolerancí, například tzv. ratios (reciproké indexy) a nebo modifikace nejrozšířenějšího indexu C_{pk}. Podrobnosti včetně specializovaného programu lze nalézt například v učebnici [1].

Literatura

[1] Tošenovský,J., Noskievičová,D., *Statistické metody pro zlepšování jakosti* (včetně software). Montanex, a.s., Ostrava 2000, 362 s.

[2] Schneider, H., Pruett, J., Lagrange, C.*, Uses of process capability indices in the supplier certification process*. Quality Engineering 8(2), 1996.

Projekt CESAR – Vnímání rizika v šesti zemích aneb jak se staví mosty mezi statistikou a prostým lidem

Hana Šlachtová, KHS Ostrava

Na letošních Ostravských statistických dnech jsem přednesla výše uvedený referát s cílem jednak podat informaci o studii CESAR, jejich částech, použitých metodách zpracování dat – MLM (multilevel modelling) a připravovaných publikacích; a na základě dat z RPQ (dotazníkové šetření vnímání rizika) upozornit na důležitost přípravy dat, vážení statistických postupů, komunikace mezi výzkumníkem a statistikem, nutnosti volby statistických výstupů, které je možno interpretovat, ujistit se, že interpretace opravdu vychází ze sesbíraných dat a interpretuje to, co je získáno statistickou analýzou.

Výzkumný projekt (nazvaný CESAR) probíhal ve dvou etapách: sběr dat v rámci programu EU Phare v letech 1994-97 a MLM z prostředků EU Inco-

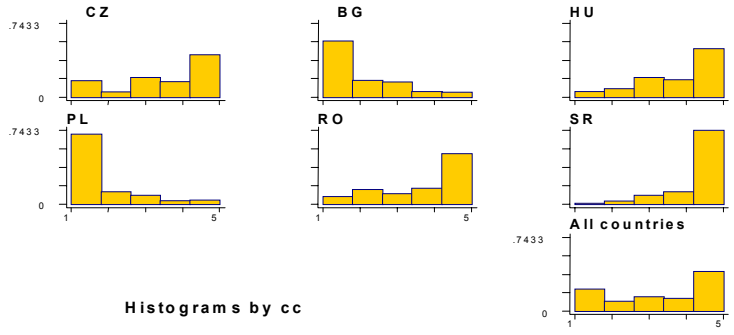
Copernicus v období 1998-2000. Projekt CESAR byl rozsáhlou epidemiologickou studií zaměřenou na zjištění vztahu znečištění vnějšího ovzduší a chronických respiračních onemocnění u dětí ve věku 7-11 let, na zajištění kvality dat a použitých metod a na zjištění vnímání rizika. Realizován byl ve 24 (resp. 25 – v Maďarsku 5 oblastí) oblastech Bulharska, České republiky, Maďarska, Polska, Rumunska a Slovenské republiky, koordinaci projektu zajišťovaly výzkumné ústavy a univerzity z Nizozemí a Velké Británie. V každé zemi byly vybrány čtyři oblasti s rozdílnou úrovní znečištění ovzduší. Pro představu o velikosti projektu, který je jednou z největších realizovaných studií na světě: jenom v ČR bylo analyzováno 3 672 zdravotních dotazníků, 1 789 vzorků ovzduší, 1 753 vyšetření funkce plic, 528 vzorků krve a 716 dotazníků vnímání rizika.

Návazný projekt CESAR II byl zaměřen na ověření možnosti použití víceúrovňové statistické techniky pro hierarchickou strukturu dat (MLM) při analýze pečlivě očištěných databází získaných v rámci studie CESAR I.

Tato v epidemiologii nová metoda byla vhodnější než tradiční regresní metody, protože umožňuje v modelu lépe zachytit hierarchii dat. MLM umožňuje modelování vztahu mezi subjekty v rámci stejné oblasti. Výsledky získané užitím MLM byly porovnány s výsledky získanými tradičními metodami. Tomuto srovnání je z metodologického pohledu věnován článek S. Pattendena a kol. Methodological issues in the analysis of hierarchical studies of PM and children's respiratory health v Journal of Exposure Assessment and Environmental Epidemiology, který bude připraven k publikaci do konce r. 2000. Při vyhledávání příčin některých nekonzistentností ve výsledcích vyšetření funkce plic byla využita i metoda GUHA. Celkovým výstupem projektu má být série cca 30 publikací v prestižních světových odborných časopisech.

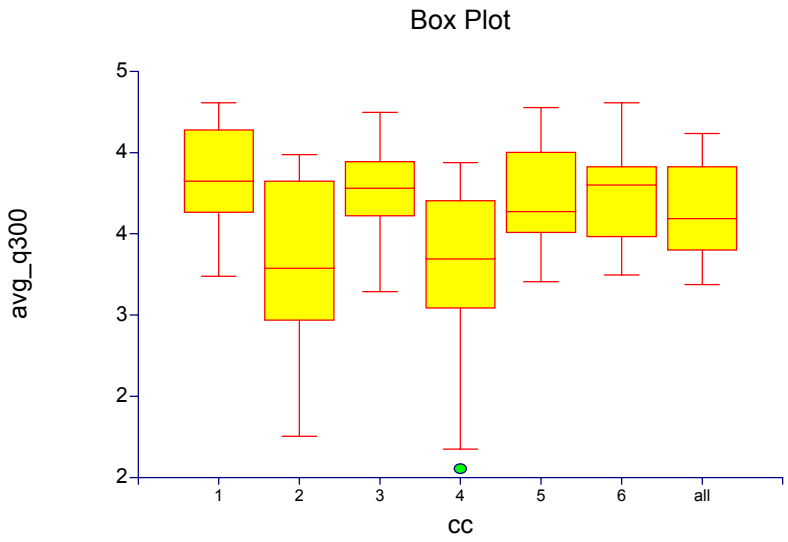
Jedním z cílů studie vnímání a komunikace rizika (risk perception -dále jen RP), která byla součástí projektu CESAR, bylo zjistit základní popisná data o vnímání rizika souvisejícího se životním prostředím, představy o riziku všeobecně a konkrétně, jak je znečištění ovzduší vnímáno v relaci k ostatním problémům životního prostředí a ostatním sociálním problémům vůbec. Krajská hygienická stanice v Ostravě se stala garantem mezinárodní analýzy RP dotazníků – celkem 6 043. První připravovaná publikace se týká zjištění vlivu geografických, demografických, ekonomických a sociálních faktorů na určení priorit mezi vyjmenovanými problémy. Tyto problémy byly různé úrovně (globální, celostátní, lokální). různých oborů (sociální, politika, zdravotnictví, školství, životní prostředí, doprava) a různé míry osobního dopadu (zdraví vlastních dětí, stav životního prostředí ve světě).

Priority byly analyzovány z otázky dotazníku, která nabízela výčet 20 problémů na pětistupňové škále, jak se daný problém respondenta týká. Rozdělení odpovědí na některé z 20 sledovaných problémů se v jednotlivých zemích dosti liší, viz např. závažnost AIDS na obr. 1



Obr. 1: Empirická rozdělení v jednotlivých zemích a celkové rozdělení - vnímání závažnosti AIDS

Na základě literatury a pilotní studie položených rozhovorů v jednotlivých zemích existovala hypotéza, že největší rozdíly budou podmíněny kulturně, tzn. rozdíly mezi zeměmi. To ukazuje i grafické srovnání průměrů všech 20 problémů v jednotlivých zemích, z něhož je patrné, že na pětistupňové škále je vnímání závažnosti problémů odlišné mezi jednotlivými zeměmi – obr. 2 (zleva doprava – Bulharsko, Česká republika, Maďarsko, Polsko, Rumunsko, Slovenská republika, všechny země); respondenti ve sledovaných oblastech ČR a Polska nepovažují dopad problémů za tak významný, jako v jiných zemích.



Obr. 2: Porovnání průměrných závažností v jednotlivých zemích

Pro analýzu vnímání rizika podle jednotlivých problémů v různých zemích byla užita ordinální logistická regrese (ordered logistic regression), což je lineární model pro odds ratio (OR) ordinální vysvětlované proměnné. Výsledky odds ratio (jako referenční země bylo zvoleno Bulharsko) jsou v tab. 1. Šedě jsou vyznačena políčka, ve kterých OR adjustované o vliv faktorů jsou nevýznamně odlišná od jedné, tzn. vnímání daného problému je shodné s referenční zemí.

Tabulka 1:Adjustované OR

	CZ	HU	PL	RO	SR
AIDS	0.08	1.16	0.04	1.19	3.81
Drogová závislost /+ alkohol/	0.06	1.09	0.05	0.56	4.60
Nedostatečné vytápění v zimě	0.05	0.57	0.13	0.64	0.32
Kouření tabáku	0.24	1.05	0.27	0.79	1.85
Politická nestabilita	0.64	1.48	0.65	1.52	1.45
Mé zdraví	0.27	0.29	0.29	1.35	0.39
Kvalita školství v oblasti	0.29	0.63	0.71	2.67	0.49
Kvalita mého bydlení	0.29	0.29	0.26	0.97	0.14
Ekonomická situace v zemi	0.29	0.51	0.25	0.46	0.29
Stav živ. prostř. v bydlišti	0.58	0.45	0.53	0.55	0.26
Zdraví mých dětí	0.25	0.37	0.19	0.36	0.17
Stav živ. prostředí ve světě	0.69	1.67	0.77	0.83	1.03
Nezaměstnanost	0.17	1.19	0.24	0.60	0.54
Riziko havárie jader. reaktoru	0.31	1.00	0.37	0.66	1.02
Hrozba zbrojení	0.29	0.63	0.34	0.63	1.01
Kvalita zdravotní péče	0.31	0.49	0.34	0.78	0.17
Kval. veř. dopravy v bydlišti	0.46	0.27	0.41	0.74	0.39
Kriminalita	0.34	0.48	0.21	0.11	0.46
Dopravní nehody	0.47	0.73	0.56	0.60	0.57
Korupce ve vládě	0.32	0.63	0.27	0.45	0.89

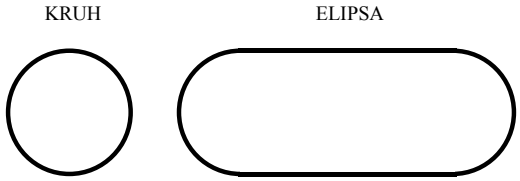
Mezi zkoumanými faktory (země, oblast, pohlaví, věk, vzdělání, povolání, hustota zalidnění bytu, příjem nebo dávka od státu, ekonomická situace rodiny) byly nejčastěji významné (kromě země), pohlaví, věk a ekonomická situace rodiny. Výsledky této části projektu CESAR budou připraveny k publikaci v Journal of Risk Analysis do konce roku 2000.

Za spolupráci na analýze dat v projektu CESAR i na přípravě této prezentace děkuji Josefu Tvrdíkovi z katedry informatiky Přírodovědecké fakulty Ostravské univerzity.

Je libo normální rozdělení?

Karel Kupka (kupka@trilobyte.cz)

Toto úsměvné zamyšlení je věnováno oslavě pečlivosti v technické a vědecké literatuře, jakožto indikátoru určité úcty autora či učitele ke svému oboru a tím i ke čtenářům, respektive k žákům. Žák, který se poctivě snaží studovat a potřebuje pochopit předloženou tématiku, a je navíc schopen technického, ne-li přímo matematického uvažování, je mnohdy citlivý k prvnímu setkání s novou informací. Je-li podána pokřiveně nebo nepoctivě, může mít žák později potíže s chápáním navazujících souvislostí. Lze si představit studenta prvního semestru, kterého se na přednášce o kuželosečkách a křivkách druhého stupně zmocní podivný neklid a nevysvětlitelná vnitřní dissonance, která má svůj původ v dávno zapomenutém obrázku, jímž jej znásilnil necitlivý kantor v páté třídě (Obr. 1).

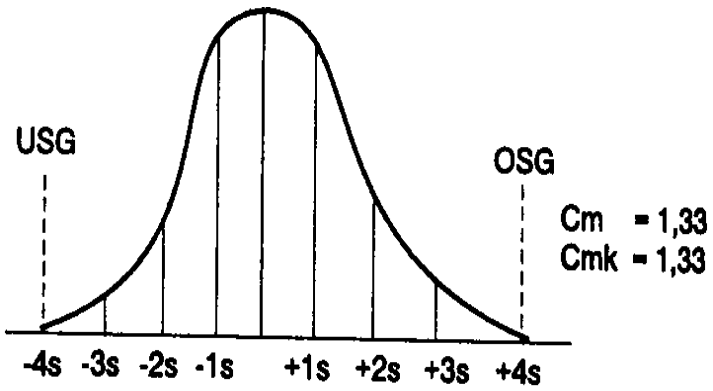


Obr. 1 Znásilnění ubohého páťáka.

Stále více je vyžadováno široké nasazení statistických metod v praktickém světě. Technicky nebo matematicky vzdělaný pracovník bez úzkého statistického zázemí si musí v krátkém čase osvojit základy aplikované statistické analýzy, v níž je normalita většinou základním předpokladem. Vysvětlení a pochopení normálního rozdělení se tedy pochopitelně věnuje ústřední pozornost v každé knize, která hovoří o statistice. To se děje celosvětově v intenzivních kurzech konzultačních firem podle renomovaných učebnic a autorských učebních textů. Z jedné strany tlačí čas, z druhé naléhavost (konkurence, akreditace, atd.), a tak se takové kurzy zkracují a zdražují do závratných výšek. Vzhledem k tomu, že grafická informace je vnímána účinněji než text, není kvalita grafu zcela nedůležitá. Co jiného je pak polovina níže uvedených obrázků, než varianta Obr. 1? A tak čtenář učebnice za 200 dolarů na vlastní oči vidí, že normální rozdělení

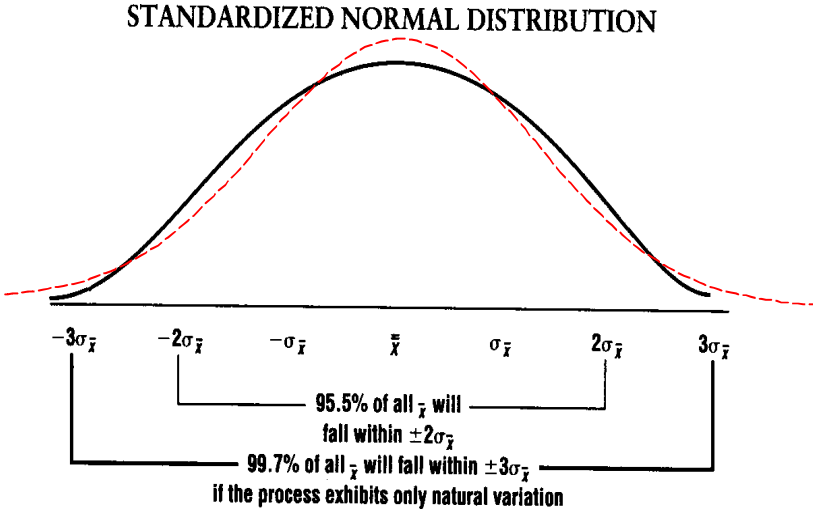
(vždy jen jeho hustota, ta se jaksi lépe kreslí) je mírně asymetrické, podobá se kružnici, protíná osu X, má nevlastní integrál, je konkávní, má nespojité derivace, omezený definiční interval, atd. Nechceme v žádném případě kritizovat autory odborné literatury za jeden odbytý obrázek. V citlivém čtenáři se však (doufejme, že neprávem) může vynořit znepokojivá otázka: jestliže autorovi nezáleží na nejdůležitějších grafech, záleží mu na ostatním obsahu? A tak i odborná literatura někdy nenápadně přispívá k obecné inflaci elektronické i tištěné informace, jejímž původcem je snadná dostupnost publikačních nástrojů, jejichž výkonnost byla ještě nedávno nepředstavitelná. Jen z čiré zvědavosti jsme namátkou vybrali pár publikací, oskenovali grafy normálního rozdělení a ověřili, zda scanner nezkresluje. Pak jsme k tyto grafy porovnali s Gaussovou křivkou vygenerovanou na počítači. Porovnáván byl pouze tvar, ne absolutní měřítka. Dokonalá shoda byla nalezena pouze u W. Shewharta, ČSN ISO 3951, R. A. Fishera a na staré desetimarkové bankovce. Zajímavá je jedna převažující nelogická skutečnost: Obrázek je tím horší, čím je snazší ho nakreslit. Křivku $y=1/\exp(x^2)$ nakreslí a vytiskne každý středoškolák v Excelu. Proč to neumí téměř žádný ze současných uznávaných autorů? A co je nejzáhadnější: Jak to dělal sir Fisher a Dr. Shewhart bez počítače a laserové tiskárny? Dotkli jsme se snad podstaty velikosti osobnosti, která spočívá v malých dokonalostech, poctivosti a úctě k sobě i druhým? Jsou čas a objem tak důležité, že je třeba jim obětovat tyto přednosti? Chtělo by se citovat na závěr klasika: Před několika málo sty lety málo lidí umělo psát, ale ti, co uměli psát - ti uměli *psát*! To se dá číst i dnes, to byste nevěřili!

Obrázková příloha.
(Tam, kde není pro srovnání čárkovaná skutečná Gaussova křivka, jde buď o evidentní paskvil, nebo naopak o shodu zcela dokonalou, kterou nechceme rušit.)

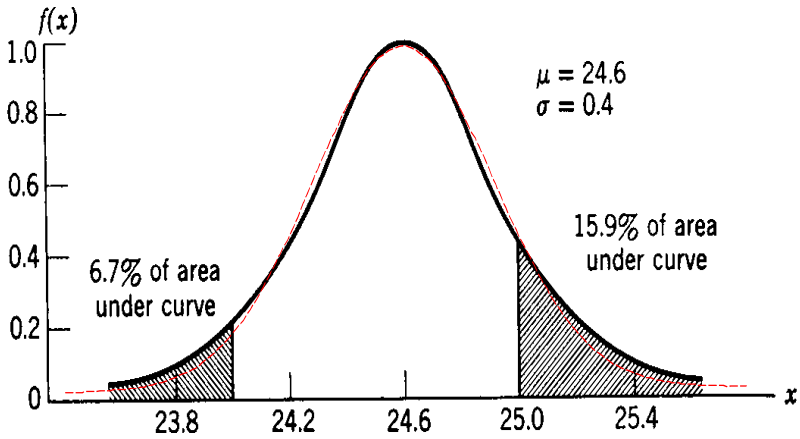


Obr. 62

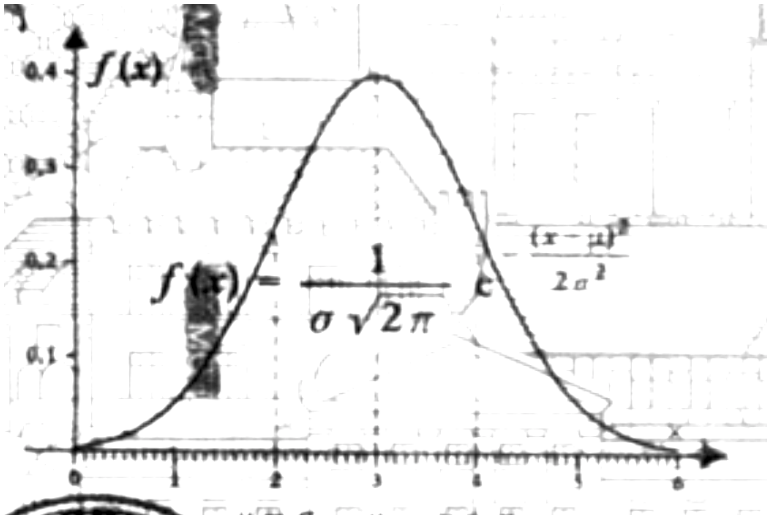
Hans-Ulrich Frehr: Total Quality Management, Unis Brno 1995



Hamid Noori, Russel Radford: Production and operation management, McGraw-Hill 1955



Hahn, G. J., Shapiro, S. S.: Statistical models in Engineering, J. Wiley 1994
R.A. Fisher: The Design of Experiments, 1935



Deutches Bundesbank, Frankfurt am Main 1993

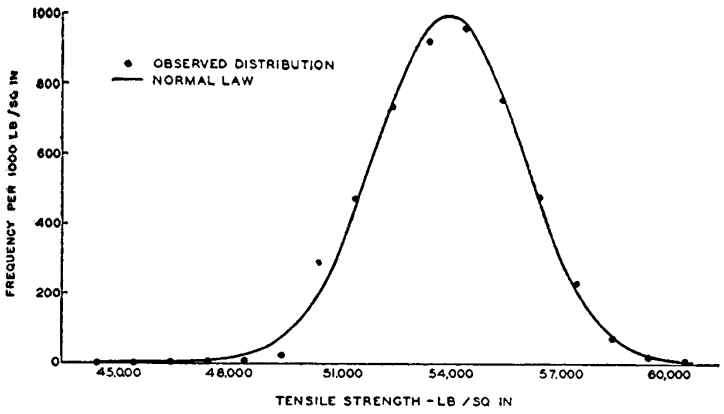
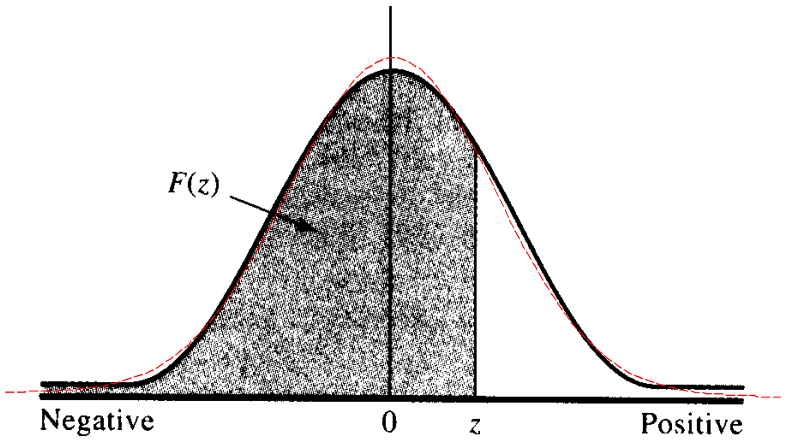


FIG. 12

W. Shewhart: Statistical Methods from the Wiewpoint of Quality Control, GSDA 1939



Grant, E.L., Leavenworth, R.S.: Statistical Quality Control, McGraw-Hill 1996

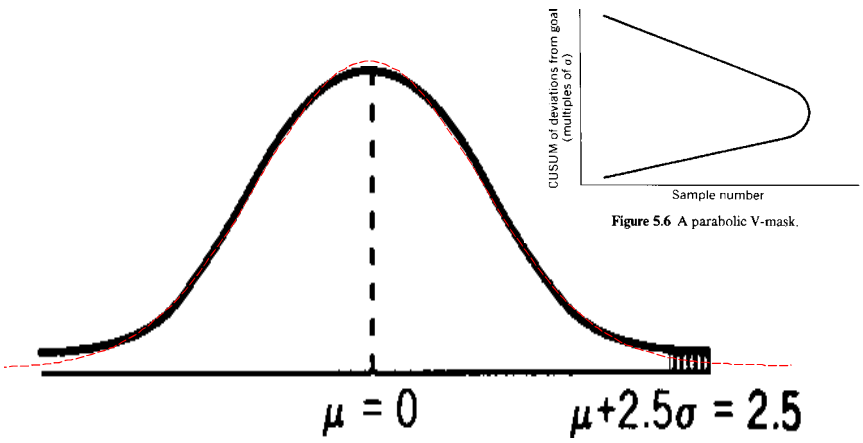
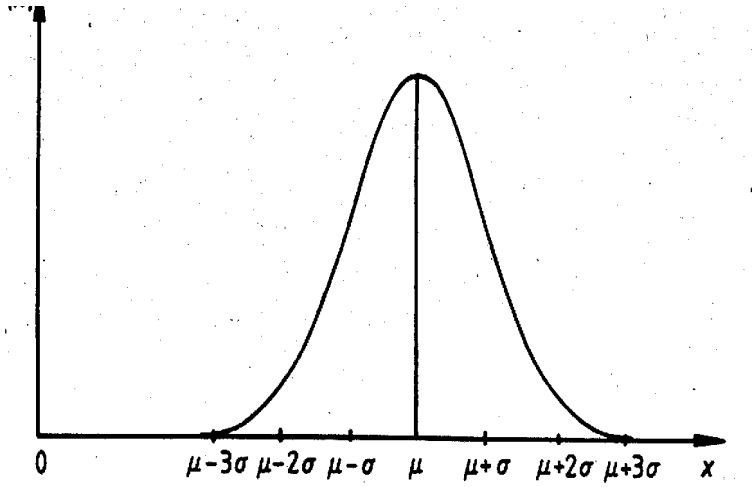


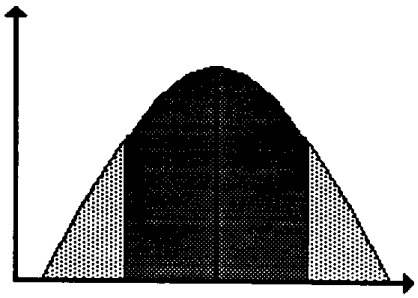
Figure 5.6 A parabolic V-mask.

P. Ryan: Statistical Methods for Quality Control, J. Wiley 1989



Obrázek 5 – Normální rozdělení

ČSN ISO 3951, 1993



Normal Distribution

Eurachem: Quantifying Uncertainty in Analytical Measurement, Draft, May 1994

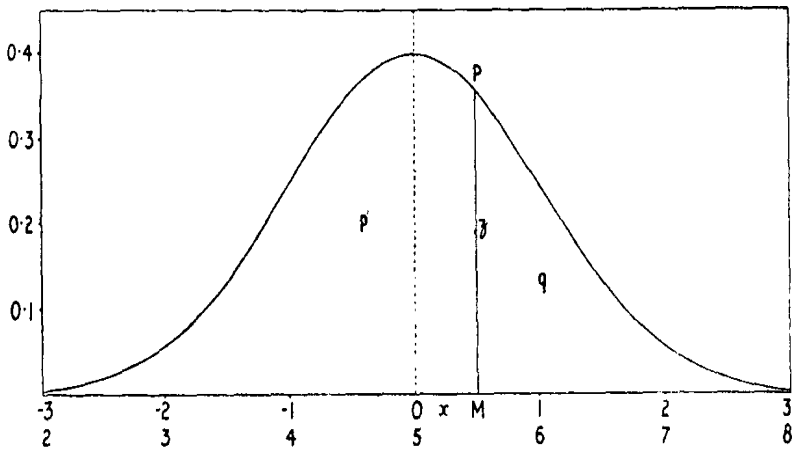


FIG. 4.—Normal distribution curve

R.A. Fisher: The Design of Experiments, 1935

Hledání optimálního počtu shluků metodami fuzzy shlukové analýzy

Martin Malčík,Centrum informačních technologií OU

1. Úvod

V době exploze informací je potřeba zpracování dat, která jsou zpravidla uchovávána ve formě tabulek v nichž řádky představují sledované objekty a sloupce jejich atributy, větší než kdy předtím. Shluková analýza je metoda pomáhající řešit problémy klasifikace dat. Může být použita tehdy, pokud známe málo nebo nic o struktuře skupiny dat. Snaží se najít vnitřní strukturu dat jejich organizací do skupin neboli shluků. Slouží především jako prostředek generování hypotéz o klasifikaci objektů nebo atributů.

Podle (Andenberg, 1973) je předmětem shlukové analýzy třídění objektů nebo atributů do skupin tak, že stupeň nepodobnosti je menší mezi členy stejné třídy a větší mezi členy různých skupin.

2. Fuzzy shlukovací metody

Fuzzy shlukování je možno částečně srovnat se známějším nehierarchickým shlukováním množiny objektů. Nechť $X=\{x_1, \ldots, x_n\}$ je množina n vektorů $\mathbf{x_i} \in R^p$ reprezentující data. Pro $c \geq 2$ ($c \in$ N) sestává nehierarchické shlukování množiny X do c shluků z c disjunktních podmnožin množiny X, $S_1, \ldots, S_c$, jejichž sjednocením je množina X. Tedy pro každé $i = 1, \ldots, c$ definujeme funkce $u_i: X \rightarrow \{0,1\}$ tak, že $u_i(x)=1$, jestliže $x \in S_i$ a $u_i(x)=0$, jestliže $x \notin S_i$. Tyto funkce se nazývají funkce příslušnosti, protože určují, do kterého shluku každý objekt patří.

Fuzzy shlukování množiny X do c shluků je tvořeno funkcemi $u_1, \ldots, u_c$ kde $u_i: X \rightarrow [0,1]$ a $\Sigma_i u_i(x)=1$, pro všechna $x \in X$. Tyto funkce se také nazývají funkce příslušnosti, avšak nedefinují podmnožiny v obvyklém smyslu, ale jsou příklady fuzzy množin (Novák 1990). Hodnota fuzzy funkce příslušnosti může být libovolné číslo z intervalu [0,1] a je míněna jako matematické vyjádření „množiny“, která není přesně definována. Funkce příslušnosti indikují vnitřní strukturu dat takto: Jestliže dva body mají hodnotu příslušnosti blízkou jedné pro stejnou funkci příslušnosti, jsou považovány za navzájem podobné. Podmínka $\Sigma_i u_i(\mathbf{x})=1$ odpovídá příslušnosti každého prvku do X. Pro $i=1,\ldots,c$ můžeme psát $S_i=\{\mathbf{x} \in X: u_i(\mathbf{x}) \geq u_j(\mathbf{x})$ pro $j=1, \ldots c\}$. Tyto množiny nemusí být disjunktní.

3. Fuzzy c-Means algoritmus

Fuzzy c-Means algoritmus provádí fuzzy shlukování množiny dat. Pro $c \geq 2$ a reálné $m' \in [1, \infty)$ algoritmus vybírá u_i: $X \rightarrow [0,1]$ tak, že pro $\Sigma_i u_i(\mathbf{x})=1$ a $\mathbf{v}_i \in \mathrm{R}^p$, kde $i= 1,..,c$ minimalizuje objektovou funkci

$$J_p(\underset{\sim}{U}, \mathbf{v}) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^{m'} (d_{ik})^2 ,$$

kde $d_{ik} = d(\mathbf{x}_k - \mathbf{v}_i) = \left[\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right]^{1/2}$ a u_{ik} je hodnota i-té funkce příslušnosti v bodě x_k. Vznikne matice funkcí příslušnosti $\mathbf{U}^{\mathrm{T}} = [\mathbf{u}_1, \mathbf{u}_{2,...,} \mathbf{u}_n]$, kde $\mathbf{u}_i^{\mathbf{T}} = [u_1, u_2, ..u_c]$ množiny dat X. Předpokládáme, že nepodobnost objektů je vyjádřena vzdáleností mezi dvěma vektory.

Fuzzy c-Means (FCM) (Bezdek, 1981)

Zvolíme c ($2 \leq c < n$) a hodnotu parametru m'. Parametr m', který nazveme váhový parametr, určuje "fuzičnost" klasifikačního procesu, $m' \in [1, \infty)$. Inicializujeme matici $\underset{\sim}{U}^{(0)}$ tak, že n bodů náhodně rozmístíme do c shluků. Každý průchod algoritmu bude označen r, kde $r = 0, 1, 2,\ldots$

Vypočteme těžiště každého z c shluků $\{ \mathbf{v}_i^{(r)} \}$ pro každý r-tý průchod

$$v_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \cdot x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m'}} .$$

Aktualizujeme matici shluků pro r-tý průchod $\underset{\sim}{U}^{(r)}$ takto:

$$u_{ik}^{(r+1)} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^{2/(m'-1)} \right]^{-1} \quad \text{pro } I_k = \varnothing,\ I_k = \left\{ i \middle| 2 \leq c < n; d_{ik}^{(r)} = 0 \right\}$$

nebo

$u_{ik}^{(r+1)} = 0$ pro všechny třídy i, kde $i \in \tilde{I}_k$, $\tilde{I}_k = \{1,2,...,c\} - I_k$

a hodnoty $u_{ik}^{(r+1)}$ jsou normovány tak, aby $\sum_{i \in I_k} u_{ik}^{(r+1)} = 1$

Jestliže $\left\| \underset{\sim}{\mathrm{U}}^{(r+1)} - \underset{\sim}{\mathrm{U}}^{(r)} \right\| \leq \varepsilon_L$, stop; jinak nastavíme $r = r + 1$ a pokračujeme na krok 2.

4. Hledání optimálního počtu shluků

Kvalitu shlukování indikuje těsnost asociace bodů ke středům shlukování a je to funkce příslušnosti, která měří míru asociace. Pro komplexní zpracování neznámého souboru dat navrhuji následující postup:

1) Data analyzujeme hierarchickou aglomerativní metodou až po výsledný jeden shluk obsahující všechny body. Podle typu dat použijeme některou míru a koeficient asociace. Na příklady bylo použito pro výpočet Wishart Wardovo rekurzivní schéma (Lukasová 1985):

$\mathrm{D}(\{\mathbf{x}_i\},\{\mathbf{x}_j\}) = d_{\mathrm{E}}^2(\mathbf{x}_i, \mathbf{x}_j)$, kde $\mathbf{x}_\mathrm{i}$, $\mathbf{x}_\mathrm{j}$ jsou jednotlivé objekty, i,j = 1, …, n

$$\mathrm{D}(\mathrm{U},\mathrm{R}) = \frac{1}{|\mathrm{R}|+|\mathrm{U}|}\left[(|\mathrm{U}|+|\mathrm{P}|)\mathrm{D}(\mathrm{U},\mathrm{P}) + (|\mathrm{U}|+|\mathrm{L}|)\mathrm{D}(\mathrm{U},\mathrm{L}) - |\mathrm{U}|\mathrm{D}(\mathrm{P},\mathrm{L})\right],$$

kde $\mathrm{R} = \mathrm{P} \cup \mathrm{L}$ a $|\mathrm{U}|,|\mathrm{P}|,|\mathrm{R}|,|\mathrm{L}|$ jsou počty objektů ve shluku U, P, R, L. Jako koeficient nepodobnosti objektů počítáme čtverec Euklidovské metriky

$d_{\mathrm{E}}^2 = \sum_{i=1}^{p} (a_i - b_i)^2$ pro body A=(a_1, a_2,…a_p), B=(b_1,b_2,…b_p).

1) V každém kroku hierarchické aglomerativní metody vypočteme koeficient optimality rozkladu K.

$K = \frac{1}{n}\sum_{v} u_{iv}(x_v)$, kde $u_{iv}(x_v) \geq u_{jv}(x_v)$ a n je počet bodů, c je počet shluků v rozkladu, i, j=1,…$c, i \neq j$, v=1..n.

ii) Nalezneme lokální maxima K.

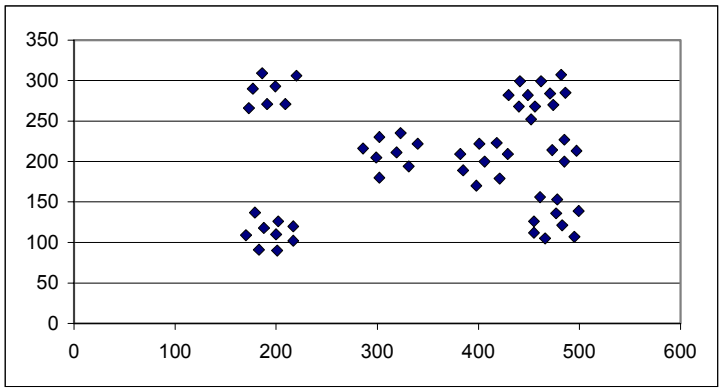
iii) Podle pořadí lokálního maxima, maxim, vybereme nejvhodnější rozklad, rozklady, které můžeme ještě dále zkoumat nehierarchickou klasickou nebo fuzzy shlukovací metodou.

5. Příklady

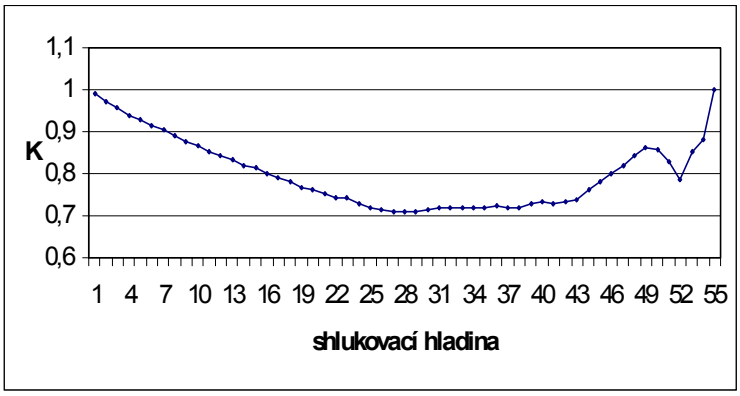
Použití metody je ilustrováno na čtyřech jednoduchých příkladech, kde p=2. První obrázek znázorňuje vždy rozložení bodů v souřadnicové rovině. Na druhém obrázku je ke každé shlukovací hladině zobrazena velikost koeficientu K.

5.1 Několik shluků

V souboru je 56 bodů a sedm víceméně zřetelných shluků. Byly nalezeny tři lokální maxima koeficientu *K*, největší z nich indikuje 7 shluků na 49. shlukovací hladině.

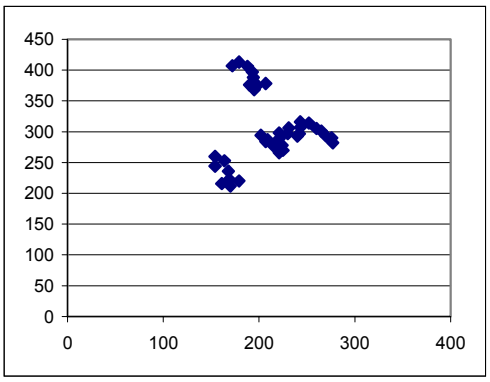


Obrázek 1: Zobrazení dat př. 5.1Obrázek 2: Průběh koeficientu

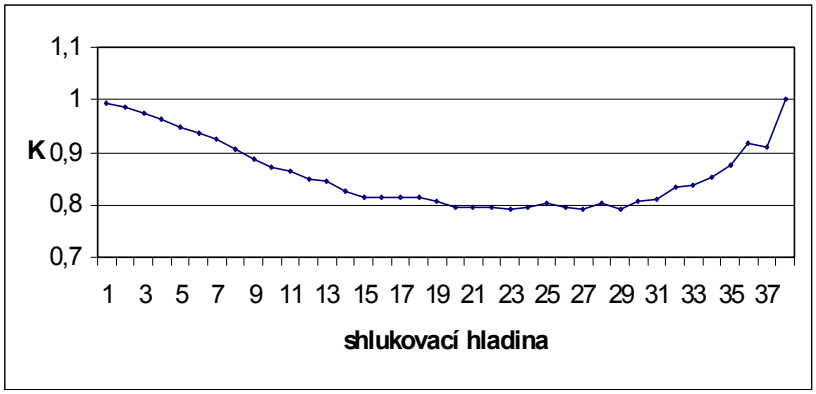


5.2 Tři shluky

Soubor obsahuje 39 bodů, bylo nalezeno pět lokálních maxim koeficientu *K*, největší z nich indikuje tři shluky na 36. shlukovací hladině.



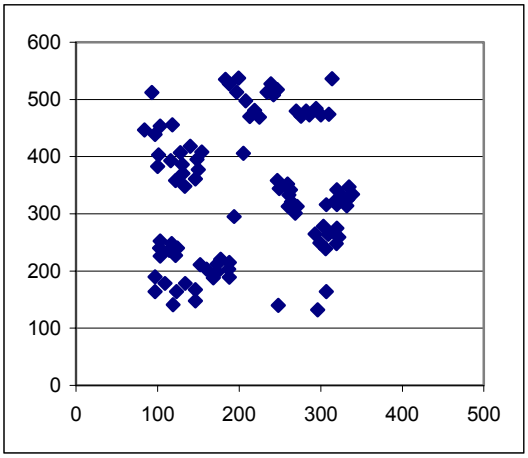
Obrázek 3: Zobrazení dat př. 5.2



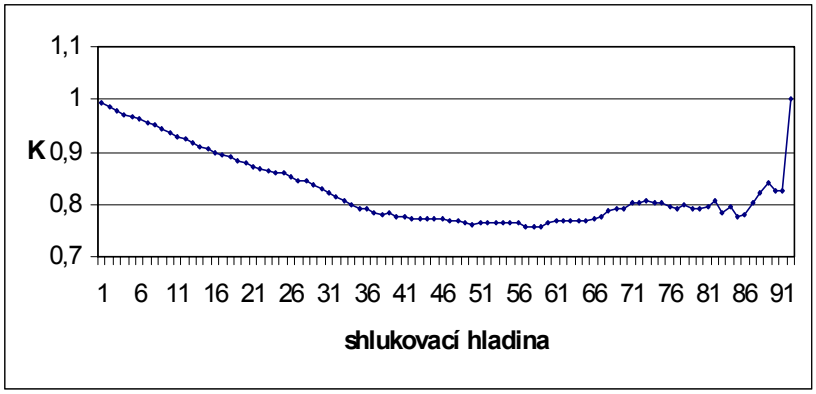
Obrázek 4: Průběh koeficientu *K*

5.3 Zřetelné shluky na dvou hladinách

Soubor obsahuje 93 bodů, bylo nalezeno 10 lokálních maxim koeficientu *K*. Na křivce koeficientů *K* lze pozorovat dva výrazné vrcholy, které odpovídají dvěma největším hodnotám lokálních maxim, tedy 11 a 4 shlukům. To odpovídá i vizuálnímu rozdělení bodů na dvě výraznější shlukovací hladiny.



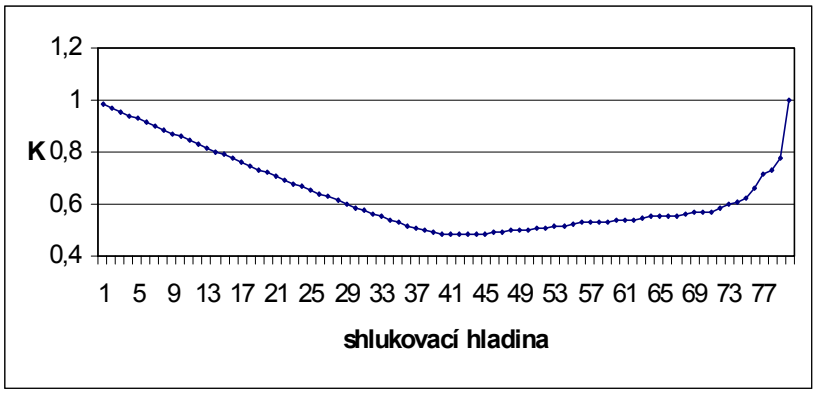
Obrázek 5: Zobrazení dat př. 5.3



Obrázek 6: Průběh koeficientu *K*

5.4 Data bez shluku

Soubor obsahuje 81 bodů rovnoměrně rozmístěných v rovině. Nebylo nalezeno žádné



Obrázek 8: Průběh koeficientu *K*

6 Závěr

Nehierarchické metody, zvláště s fuzzy přístupem, které můžeme použít na následnou analýzu zvoleného rozkladu, mohou ukázat míru příslušnosti jednotlivých bodů ke shlukům. Kombinací tradičního iteračního přístupu s fuzzy metodou – prověřováním velikosti funkce příslušnosti jednotlivých bodů ke shlukům bychom měli najít optimální rozklad. To může pomoci řešit stop problém v hierarchických metodách.

Literatura

Andenberg, M.R.: Cluster Analysis for Applications. Academic Press, New York, 1973

Backer, E.: Computer-assisted Reasoning in Cluster Analysis.Prentice Hall, 1995

Bezdek, J.: Pattern recognition with fuzzy objective function algoritmus. Plenum Press, New York, 1981

Bezdek, J.,C., Pal, S., K.: Fuzzy Models for Pattern Recognition: Methods that search for structures in data. IEEE, New York, 1992

Jain, A.K., Dubes, R.C.: Algorithmus for Clustering Data. Prentice Hall, Englewood Cliffs, 1988

Kahounová, J.: Měření podobnosti struktur. VŠE Praha, 1994.

Lukasová, A., Šarmanová, J.: Metody shlukové analýzy. SNTL Praha, 1985.

Novák, V.: Fuzzy množiny a jejich aplikace. SNTL Praha, 1990.

Novák, V., Perfilieva, I., Močkoř, J.: Mathematical Principles of Fuzzy Logic., Kluwer Academic Publishers, Boston, 2000

Ross, T.: Fuzzy Logic with Engineering Applications.McGraw-Hill,. New York, 1995

Věřit statistickému software?

Josef Tvrdík

Ostravská universita [1]

Abstrakt. Příspěvek se zabývá selháními softwarových statistických procedur, zjištěnými v průběhu jejich dlouhodobého užívání. Jsou diskutovány některé chyby v jednoduchých popisných metodách nalezené v Excelu, nesprávné odhady parametrů nelineárních regresních modelů vyskytující se až příliš často v běžně prodávaných statistických paketech (NCSS, SYSTAT, SPSS, S-PLUS) a také numerické nesrovnalosti i triviální implementační chyby ve statistikách pro test shody funkcí přežití v NCSS.

Klíčová slova: statistický software, Excel 97, algoritmy, chyby, nelineární regrese.

Úvod

Text je rozšířenou versí sdělení předneseného na Statistických dnech České statistické společnosti v Ostravě v červnu 2000. Pod názvem „Opravdu jen drobné vady na kráse statistického software?“ byl přednesen také na letní škole ROBUST 2000 v Nečtinách v září téhož roku. Obě tato upozornění na některé zjevné i méně zjevné numerické nesrovnalosti či chyby zjištěné při užívání známých statistických paketů a tabulkového procesoru Excel 97 vyvolala dosti živý ohlas u účastníků zmíněných akcí. Po dohodě s editory sborníku ROBUST je tento text zveřejněn v Bulletinu České statistické společnosti, kde snad má příležitost se dostat k většímu okruhu zainteresovaných čtenářů.

Excel je přidán ke kritizovaným statistickým programům, neboť je pro statistické výpočty velmi často užíván zejména lidmi potřebujícími statistiku pouze občas a u nich je rozpoznání chybného výsledku ještě méně pravděpodobné než u zkušeného statistika. Chyby v Excelu mohou tedy působit škodu velice často. Navíc škody působí také česká lokalizace Excelu, viz Tvrdík (1998).

Po těchto zkušenostech s nespolehlivostí statistického software se stává naléhavou otázka, zda úsilí, které statistici věnují hledání rigorosních řešení různých statistických problémů (občas i dosti vyumělkova-

[1] Tato práce byla podporována z grantu 402/00/1165 GA ČR a z projektu institucionálního výzkumu CEZ: J09/98:179000002.

ných) není z hlediska aplikací statistiky zbytečné a zda by podobné úsilí nemělo být orientováno na výběr vhodnějších, numericky spolehlivějších algoritmů a důkladnějšímu testování jejich implementace. Naprostá většina aplikací statistiky je opřena o výpočty provedené s využitím statistického software a pokud jsou jejich výsledky numericky chybné, jsou sofistikované statistické metody na nic.

Testy numerické spolehlivosti

Testováním spolehlivosti statistických programů se zabývají všichni výrobci software, ale patrně některé výsledky si nechávají pro sebe. Objektivní pohled na spolehlivost statistických procedur je asi dost problematický. O jednu z možných cest se pokouší americký Státní institut standardů a technologie (National Institute of Standards and Technology, NIST, viz citaci na jeho web-stránku). Tam je shromážděna sada testovacích úloh z několika oblastí statistiky, u kterých jsou známé jejich výsledky na jistý (tzv. certifikovaný) počet platných cifer. Přehled je uveden v tab. 1.

Tabulka 1: Standardní referenční úlohy NIST – přehled

Druh úloh	počet úloh	certifikovaný počet platných míst
jednorozměrné statistiky	9	15
lineární regrese	11	15
analýza rozptylu	11	15
nelineární regrese	27	11

Numerickou správností výsledků některých statistických procedur v Excelu 97 se nedávno zabývali McCullough a Wilson (1999). Zjišťovali, jak se shodují výsledky získané Excelem s certifikovanými výsledky NIST. Poněkud paradoxní je, že v jejich článku je chyba v definici veličiny, kterou sledovali. Podle slovního popisu má veličina λ vyjadřovat míru shody výsledné hodnoty x spočítané Excelem s certifikovanou hodnotou c a znamená vlastně počet platných číslic shodných s certifikovaným výsledkem. V recenzovaném článku v poměrně renomovaném časopisu jim prošla následující definiční rovnice

$$\lambda = \log_{10}\left(|x - c|\right)/|c| \tag{1}$$

Správná definice λ má mít zřejmě tvar

$$\lambda = \begin{cases} 0 & \text{když } \frac{|x-c|}{|c|} \geq 1 \\ 15 & \text{když } \frac{|x-c|}{|c|} < 1 \times 10^{-15} \\ -\log_{10}\left(\frac{|x-c|}{|c|}\right) & \text{jinak} \end{cases} \tag{2}$$

V úlohách, kde se počítá více než jeden parametr, je výsledná míra shody λ pro úlohu chápána jako

$$\lambda = \min(\lambda_1, \lambda_2, \ldots, \lambda_k), \tag{3}$$

kde k je počet vypočítávaných parametrů.

Jednorozměrné statistiky

Přes výše uvedené výhrady však McCullough a Wilson (1999) důvěryhodně zjistili, že Excel selhává i v jednorozměrných statistikách, kdy dokonce mezi úlohami v sadě NIST byly nalezeny takové. pro které $\lambda = 0$. Bylo to způsobeno většinou užitím nevhodného algoritmu pro výpočet výběrového rozptylu. V mnoha učebnicích základních statistických metod se tradičně uvádí, že výběrový rozptyl je

$${s_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}{x_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right], \tag{4}$$

přičemž výraz za druhým rovnítkem se doporučuje jako výpočetně vhodnější. Na ošidnost toho doporučení upozorňuje Ekblom (1994). Jak je rovněž v učebnicích elementárních statistických metod uváděno, rozptyl je invariantní vůči posunu, tj. pro $y_i = a + x_i$, a je konstantní, je pak ${s_y}^2 = {s_x}^2$. Tento vztah má však při numerických výpočtech omezenou platnost, neboť musíme uvažovat chyby ze zaokrouhlování. Pokud je průměr $\overline{x}$ velký a rozptyl malý, pak druhá rovnost v rov. 4 platí jen přibližně, za jistých okolností může být počítačová hodnota výrazu

$$S = \sum_{i=1}^{n}{x_i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \tag{5}$$

dokonce záporná. Podle experimentálních výsledků několika testovacích příkladů lze usoudit, že v Excelu je tento problém vyřešen vskutku velice úsporně. Byl přímo ukázkově uplatněn i odjinud známý racionální přístup Microsoftu: Pokud nastane situace, že hodnota výrazu S z rov. 5 je menší než 0, pak výsledný výběrový rozptyl v Excelu je roven nule. Ve standardním statistickém software podobná hrubá implementační chyba zjištěna nebyla, např. NCSS dává správné výsledky pro všechny úlohy z testovací sady NIST.

Nelineární regrese

Na sadě příkladů NIST McCullough a Wilson (1999) ověřovali rovněž numerickou správnost odhadů parametrů nelineárních regresních modelů získaných Excelem. Za selhání programu se považuje, když program skončí výpočet v lokálním minimu součtu čtverců residuálních odchylek, tzn. $\lambda = 0$. Na stejných úlohách testovala Valchařová (2000) statistický paket NCSS 6.0, pro každou úlohu 10 opakování s náhodně volenými počátečními hodnotami odhadů z jejich přijatelného oboru. Za selhání se považuje, když ve více než třetině opakování výpočtu pro danou úlohu skončí odhad parametrů v lokálním minimu. Výsledky uvedeny v tab. 2. Pohled na tabulku důvěru ve spolehlivost statistických programů nepovzbudí.

Tabulka 2: Nelineární regrese – počty selhání

obtížnost úloh	počet úloh	Excel McCullough, Wilson	NCSS Valchařová
nízká	8	4	2
střední	11	10	6
vysoká	8	7	3
celkem	27	21	11

Na 8 z 11 v NCSS selhávajících úloh zkusila Valchařová užít pro odhad parametrů stochastické algoritmy, popis algoritmů viz Tvrdík a Křivý (1999). Ač byly výpočty prováděny v časové tísni před termínem odevzdání diplomové práce bez hlubšího rozmyslu (ale to je dosti častý přístup uživatelů statistického software), v polovině z těchto úloh stochastické algoritmy neselhaly ani jednou, takže z testu vyšly přece jen o trochu lépe než softwarová klasika.

Zajímavé srovnání úspěšnosti statistického software (Excel, S-Plus, SPSS, SAS, STATA, Mathematica) na sadě úloh NIST publikoval nedávno McCoullogh (2000). Jediným plně úpěšným softwarovým produktem na všech 59 úlohách NIST byla Mathematica.

Tabulka 3: Procento selhání standardních statistických paketů

Model	NCSS	SYSTAT G-N	SYSTAT Simplex	S-PLUS	SPSS	Průměr
1	0	3	0	6	97	21.2
2	90	35	37	73	77	62.2
3	89	69	67	100	100	85.0
4	4	0	76	0	0	16.0
5	100	0	16	57	3	35.2
6	45	8	0	100	0	30.6
7	100	100	100	81	69	90.0
8	78	11	30	18	0	27.4
9	0	0	3	0	0	0.6
10	81	2	75	79	76	62.6
11	59	7	49	34	20	33.8
12	33	100	1	51	9	38.8
13	68	100	100	64	62	78.8
14	0	2	36	37	8	16.8
Průměr	53.4	31.2	42.1	49.9	37.2	42.8

Pro testování stochastických algoritmů globální optimalizace byla v 90. létech sestavena sada 14 obtížných úloh nelineární regrese. Některé z nich jsou léta v literatuře zmiňované - Jennrich a Sampson (1968), Meyer a Roth (1972), Militký a Meloun (1994), část z těchto úloh pochází z nepublikovaných příkladů Militkého. Přehled modelů je uveden v Křivý et al (2000), úplná data byla uvedena v Tvrdík a Křivý (1995) a v Tvrdík a Křivý (1998), v elektronické formě je můžete obdržet na adrese tvrdik@osu.cz. Na této sadě úloh byly testovány komerčně dostupné statistické pakety NCSS 2000, SYSTAT 8.0, S-PLUS 4.5 a SPSS 8.0. Pro odhad parametrů nelineárních modelů je v S-PLUS a SYSTATu užíván Gauss-Newtonův algoritmus, v SYSTATu je i možnost užití simplexové metody, NCSS užívá Levenberg-Marquartův algoritmus a SPSS modifikovaný Levenberg-Marquartův algoritmus. Pro každou úlohu byla

vygenerována náhodně stovka k-tic počatečních odhadů parametrů (k je počet parametrů modelu, $2-7$ pro úlohy z této sady). Za selhání se považuje, když program skončí v lokálním minimu (hodnota kriteriální funkce, tj. residuální součet čtverců je o více jak 5 % větší než hodnota v globálním minimu). Výsledky pro statistický software dosti smutné jsou uvedeny v tabulce 3 (Krpec, 1999, Křivý et al, 2000).

Globální optimalizace multimodálních funkcí je těžký problém, není znám algoritmus, který by tento problém obecně řešil v polynomiálním čase. Jak však ukazuje tabulka 4, lze užít spolehlivější algoritmy než ty, které jsou běžně implementovány ve statistickém software. V tabulce 4 jsou uvedena procenta selhání dvou stochastických algoritmů (MCRS, ES2, Křivý et al, 2000). Kromě metody nejmenších čtverců (sloupec RSS) byly jako kriteriální funkce užity i nejmenší uřezáváné čtverce (sloupec LTS) a součet absolutních odchylek (SAD). V tabulce jsou uvedeny jen modely, u kterých procento selhání při testování bylo nenulové.

Tabulka 4: Procento selhání stochastických algoritmů

	MCRS			ES2		
Model	RSS	LTS	SAD	RSS	LTS	SAD
2	1	88	0	0	97	0
5	0	3	0	0	0	0
8	0	100	0	0	85	0
11	24	19	20	0	0	5
13	0	20	0	0	0	0

Testy shody funkcí přežití v NCSS

Pozoruhodně podivné výsledky NCSS 2000 při testech shody funkcí přežití byly nedávno objeveny shodou náhod. Pan primář Vodvářka z radioterapie FNsP v Ostravě potřeboval jen „takovou drobnost na počkání“, takže jsme několik hodin u počítače přeskupovali data a chrlili funkce přežití a výsledky testů jejich shody. On s neutuchající pozorností nahlížel na výsledky na obrazovce a neunikla mu následující nesrovnalost ve výstupu z programu:

. . .
Gehans-Wilcoxon Section: . . .
Chi Square = 0.63 DF = 2 Prob>CS = 0.730450

Peto-Wilcoxon Section: ...
Chi Square = 0.63 DF = 2 Prob>CS = 0.000000
Log-Rank Section: ...
Chi Square = 1.76 DF = 2 Prob>CS = 0.415603

Na první pohled je zřejmé, že hodnota $P = 0.000000$ u Peto-Wilcoxonovy statistiky $\chi^2_2 = 0.63$ je nesprávná. Otázkou je, zda je dobře spočítaná hodnota statistiky. Porovnání s výsledky získanými jinými pakety však přineslo další pochybnosti, viz tab. 5. Hodnoty statistik shodné u S-Plus a STATA se liší od NCSS. Má se snad při statistické analýze dat užívat vždy více programů a o správném výsledku má rozhodovat většinová shoda?

Tabulka 5: Hodnoty statistik χ^2

		soubor1	soubor2	soubor3
Wilcoxon	NCSS 2000	6.53	8.26	0.63
	STATA 6.0	7.80	8.02	0.62
	S-Plus 4.5	7.8	8.0	0.6
Log-Rank	NCSS 2000	6.08	11.32	1.76
	STATA 6.0	10.15	8.47	1.35
	S-Plus 4.5	10.1	8.5	1.3

Reklamoval jsem zjištěné nesrovnalosti v NCSS přes dodavatele tohoto software (Statistical Solutions, Cork) u výrobce. Netušil jsem, že se tím stávám podezřelým a budu se muset po tři měsíce obhajovat. První reakce J. Hintze byla, že rozdílné výsledky byly získány na různých datech. Reklamované výsledky byly totiž spočítány jako podskupiny jednoho souboru pomocí funkce FILTER a zřejmě ani autor NCSS nevěří ve spolehlivost její implementace. Další reklamace tentokrát už s daty rozdělenými do více souborů (výsledky byly shodné s předchozími) přinesla jediný pozitivní výsledek celého dlouhého reklamačního procesu. J. Hintze připustil, že hodnota $P = 0.000000$ u Peto-Wilcoxonovy statistiky $\chi^2_2 = 0.63$ je chyba NCSS a pyšně sdělil, že ji „fixoval“: Je způsobena nulovým počtem cenzorovaných pozorování. K neshodám v hodnotách statistik oznámil, že rozdíly ve statistikách Wilcoxonova typu jsou způsobeny odlišnými variantami těchto statistik v různých programech (což jsem akceptoval, i když manuály zmíněných statistických paketů jednoznačnou odpověď nedávají), kromě toho oznámil, že NCSS byl znovu

prověřen na příkladech z knihy Lee (1992) a bylo shledáno vše v pořádku. O rozdílech v log-rank testu pomlčel a tak to zůstalo. Asi jsme se dostali do nekonečného cyklu, na otázku, proč se liší NCSS v log-rank testu, vždy přišla předchozí odpověď. Po dvou měsících jsem rezignoval. Usoudil jsem, že softwarové firmy více zajímá, zda zákazník platí, než to, zda mají chybu v programu. Časy se mění, pamatuji se, že rychlé opravení reklamované chyby bylo považováno za nutnou ohajobu řemeslnické cti programátora a také tomu bylo věnováno patřičné osobní úsilí.

Závěr

Užívání statistického software nepřináší jen pohodlí, ale občas také jistou frustraci ze zbořených pocitů jistoty a důvěry. Upozornění na problémy a chyby v implementacích generování náhody a jejich možné důsledky (Antoch, 1998) byly zaručeně silnou ranou. Podobné rány nás však mohou potkávat i v situacích deterministických, kde bychom je očekávali ještě méně.

Pokud laskavý čtenář dočetl text až sem, je patrně zvědav, zda mu bude nabídnuta nějaká odpověď na otázku vyslovenou v názvu článku. Vyčerpávající odpověď asi nečeká, ale snad dvě možnosti se nabízejí. Pro optimisty: *Doveraj, no proveraj!* A doufejme, že důsledná aplikace tohoto přístupu pomůže zvýšit spolehlivost statistického software, když v 80. létech dokázala změnit svět. Pro zdravě skeptické realisty: *Nevěřte ničemu!*

Literatura:

Antoch, J., Jak pomocí simulací dokázat nemožné, *Informační Bulletin České statistické společnosti*, **9**(1), 1–14, 1998

Ekblom, H., What can numerical analysis do for statistics, *COMPSTAT 1994, Proceedings in Computational Statistics* (eds R.Dutter and W. Grossmann), 31–45, Physica Verlag, 1994

Jennrich, R. I. and Sampson, P. F., Application of stepwise regression to non-linear estimation, *Technometrics*, **10**(1), 63–72, 1968

Krpec, R., Optimalizace nelineárních regresních modelů, In: Sborník semináře *Moderní matematické metody v inženýrství*, VŠB-TUO, 66–69, 1999

Křivý, I., Tvrdík, J., Krpec, R., Stochastic algorithms in nonlinear regression, *Comput. Statist. Data Anal.* **33**, 278–290, 2000

Lee, Elisa T., *Statistical Methods for Survival Data Analysis*, Second Edition, Wiley-Interscience, 1992

McCullough, B.D., The Accuracy of Mathematica 4 as a Statistical Package, *Computational Statistics*, 2000 (September), viz http://www.wolfram.com/news/statistics.html

McCullough, B.D., Wilson, B., On the accuracy of statistical procedures in Microsoft Excel 97, *Comput. Statist. Data Anal.* **31**, 27–37, 1999

Meyer, R. R. and Roth, P. M., Modified damped least squares: An algorithm for non-linear estimation, *J. Inst. Math. Applics.*, **9**, 218–233, 1972

Militký, J., Meloun, M.: Modus operandi of the least squares algorithm MINOPT. *Talanta*, **40**(2), 269–277, 1994

NCSS 97, Statistical System for Windows, Number Cruncher Statistical Systems, Dr. Jerry Hintze, Kaysville, Utah, 1997

NIST, Statistical Reference Datasets, http://www.itl.nist.gov/div898/strd

S-PLUS 4.5, Data Analysis Products Division, MathSoft, Seattle, 1998

SPSS ver. 8.0, SPSS Inc., Michigan, 1998

STATA 6.0, StataCorp. College Station, TX, 1999

SYSTAT 8.0, SYSTAT, Chicago, 1997

Tvrdík, J., Excel, statistika, lokalizace a zmatek, *Informační Bulletin České statistické společnosti*, **9**(2), 13–20, 1998

Tvrdík, J., Křivý, I., Stochastic algorithms in estimating regression parameters, in: J. Hančlová (Ed.), *Proceedings of the MME'95 Symposium*, AIMES Press, Ostrava, 217–228, 1995

Tvrdík, J., Křivý, I., Evoluční algoritmy a odhad parametrů nelineárních regresních modelů, In: Sborník konference *Analýza dat'98*, 56–69, Trilobyte, Pardubice, 1998

Tvrdík, J. and Křivý, I., Simple Evolutionary Heuristics for Global Optimization, *Comput. Statist. Data Anal. (in SSN)*, **30**, 345-352, 1999

Valchařová, A., Aplikace evolučních algoritmů v odhadech parametrů nelineárních regresních modelů, diplomová práce, Ostravská universita, Přírodovědecká fakulta, 2000