Obsah

Vážené kolegyně, vážení kolegové,

výbor společnosti si Vás dovoluje pozvat na výroční zasedání, jež se uskuteční ve čtvrtek 28. ledna 2010 od 13.00 na VŠE v Praze. Pozvané přednášky přednesou kolegové Jiří Militký a Jan Picek z Technické univerzity v Liberci.

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Časopis je zařazen do Seznamu Rady, více viz `http://www.vyzkum.cz/`.

Předseda společnosti: doc. RNDr. Gejza Dohnal, CSc.
ÚTM FS ČVUT v Praze, Karlovo náměstí 13, Praha 2, CZ-121 35
E-mail: `gejza.dohnal@fs.cvut.cz`

Redakční rada: prof. Ing. Václav Čermák, DrSc. (předseda), prof. RNDr. Jaromír Antoch, CSc., doc. Ing. Josef Tvrdík, CSc., RNDr. Marek Malý, CSc., doc. RNDr. Jiří Michálek, CSc., doc. RNDr. Zdeněk Karpíšek, CSc., prof. Ing. Jiří Militký, CSc.

Technický redaktor: ing. Pavel Stříž, Ph.D., `striz@fame.utb.cz`

Informace pro autory jsou na stránkách `http://www.statspol.cz/`

ISSN 1210–8022

INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 20, číslo 3, prosinec 2009

SIMULACE JEDNOFRONTOVÉHO SYSTÉMU GI/GI/N V PROGRAMU R

SIMULATION OF A SINGLE QUEUE GI/GI/N SYSTEM IN THE R ENVIRONMENT

Jiří Dvořák, Pavel Kříž, Vojtěch Skubanič

Adresa: KPMS MFF UK, Sokolovská 83, Praha 8

E-mail: jiri.dvorak.rce@gmail.com, pavel-kriz@post.cz, kotw@centrum.cz

Abstrakt

V následujícím článku se budeme zabývat simulací jednofrontového systému hromadné obsluhy GI/GI/N se zaměřením na implementaci v programu R. Představíme také skript kontrétního simulačního algoritmu v syntaxi programu R a ukážeme některé výstupy ze simulace systému M/M/1 provedené tímto algoritmem. Výsledky simulace poté srovnáme s teoretickými výsledky, které lze pro tento jednoduchý systém explicitně vyjádřit.

In the present work we study simulation of GI/GI/N system with one queue with regard to the implementation in R. We also present a script of a specific simulation algorithm. We give some results of the simulation of M/M/1 system and compare them with theoretical results which can be expressed explicitly in this simple case.

1. Motivace

Jednou z nejčastějších aplikací teorie pravděpodobnosti jsou systémy obsluhy. Teorie front, jak se nazývá tato část pravděpodobnosti a statistiky, je úzce spjata s Markovovými procesy, s procesy obnovy a obecně s náhodnými procházkami, ale také s částicovými systémy. V praxi se systémy obsluhy používají k modelování telekomunikačního provozu, počítačových sítí, výrobních linek a mnoha dalších systémů. Není divu, že jejich teoretický popis a statistická analýza jsou tak populární.

Popišme si nejprve stručně systémy obsluhy. Základem je soubor zákazníků a obslužných zařízení, ve kterých jsou zákazníci odbavováni. Obslužné zařízení zpravidla nemůže obsluhovat více zákazníků najednou. Jsou-li všechny obslužné stanice obsazeny, čekají zákazníci ve frontě. Po obsloužení zákazník systém zase opouští. Rozlišujeme přitom několik typů systémů hromadné obsluhy, a to zejména podle počtu a uspořádání obslužných stanic, počtu

front, jejich režimu a kapacity, podle rozdělení doby obsluhy či rozdělení doby mezi příchody dvou zákazníků (tyto doby jsou většinou náhodné).

V praxi se s těmito systémy setkáváme běžně. Vezměme například velký sklad s několika naskladňovacími rampami. Ke skladu přijíždí kamiony a jsou vykládány u těchto ramp. Jsou-li všechny rampy obsazeny, čekají kamiony na vyložení ve společné frontě. Nový kamion se řadí vždy na konec fronty a k jeho vyložení dochází až po odbavení všech kamionů, co čekají před ním, a po uvolnění některé rampy.

Při modelování takového problému většinou neznáme přesné časy příjezdů jednotlivých kamionů ani přesnou dobu vykládky. Doby mezi příjezdy dvou kamionů a doby vykládky proto modelujeme jako dvě posloupnosti náhodných veličin (obvykle nezávislých a nejčastěji i stejně rozdělených). Toto je jednoduchý příklad systému hromadné obsluhy s několika paralelně uspořádanými obslužnými stanicemi a společnou frontou s režimem FIFO (First In First Out).

Analogický systém můžeme pozorovat třeba na poště, kde je několik přepážek pro obsluhu zákazníků, kteří v případě obsazení všech přepážek čekají ve společné frontě a poté přistupují k té přepážce, která se první uvolní. Podobnou situaci můžeme sledovat u pokladen hypermarketů. Tyto pokladny také tvoří paralelně uspořádané obslužné systémy, narozdíl od předchozích příkladů má však v tomto případě každé obslužné zařízení svou vlastní frontu.

Mnohdy je však situace složitější, obslužná zařízení nejsou řazena pouze paralelně, nýbrž tvoří síť. Zákazník tedy při průchodu systémem může navštívit několik obslužných stanic, přičemž navštívené stanice se mohou u jednotlivých zákazníků lišit. Příkladem takového síťového modelu může být třeba nemocnice, kde pacient může navštívit i více ordinací na různých odděleních a to podle jeho choroby. Jiným příkladem může být výrobní závod, kde každý výrobek prochází několika operacemi na různých strojích. Simulací síťových modelů se však v tomto článku zabývat nebudeme.

2. Analytické řešení versus simulace

Řešením různých problémů souvisejících se systémy hromadné obsluhy pomocí matematických prostředků (zejména teorie pravděpodobnosti) se zabývá teorie front (viz např. [1]). Analyticky však dokážeme řešit jen relativně jednoduché systémy. Nejjednodušší je systém s jediným obslužným zařízením, neomezenou frontou s režimem FIFO a s nezávislými, exponenciálně rozdělenými dobami obsluhy i dobami mezi příchody zákazníků (tzv. M/M/1 v Kendallově klasifikaci). Pro tento systém dokážeme relativně snadno spočítat stacionární charakteristiky (tj. charakteristiky pro stabilizovaný systém),

jako např. stacionární rozdělení počtu zákazníků v systému (tzv. stav systému), stacionární rozdělení počtu zákazníků ve frontě, stacionární rozdělení doby čekání zákazníka na začátek obsluhy atd.

Výše zmíněné charakteristiky lze spočítat za předpokladu, že systém se po určité době stabilizuje, tedy je-li intenzita příchodů (průměrný počet zákazníků, kteří přijdou za jednotku času) menší než intenzita obsluhy (tj. průměrný počet obsloužených zákazníků za jednotku času). V opačném případě se systém přehltí a doba čekání i počet zákazníků v systému konvergují skoro jistě k nekonečnu.

O něco těžší je spočítat pro systém M/M/1 nestacionární rozdělení počtu zákazníků v systému v libovolném, pevně daném čase t (tj. v čase, kdy systém ještě není stabilizován). Jde totiž o řešení nekonečné soustavy lineárních diferenciálních rovnic.

Analyticky lze částečně řešit i obecnější modely GI/M/1 (resp. M/GI/1), kdy, na rozdíl od modelu M/M/1, doby mezi příchody zákazníků (resp. doby obsluhy) nemusí mít exponenciální rozdělení, jsou však i nadále nezávislé a stejně rozdělené. U těchto modelů jsou však výpočty velmi komplikované a dokážeme pomocí nich spočítat pouze některé charakteristiky. Rozdělení stavu systému například umíme spočítat (za předpokladu stacionarity) v okamžicích příchodu zákazníka do systému, zatímco u M/M/1 lze vypočítat toto rozdělení v libovolném okamžiku.

Některé výsledky byly odvozeny i pro obecné modely GI/GI/1, kde předpoklad exponenciálního rozdělení opouštíme jak pro doby mezi příchody zákazníků, tak pro doby obsluhy. V tomto obecném případě jsou výpočty již velmi obtížné a vedou často jen k dílčím výsledkům (např. Laplaceova transformace doby čekání na obsluhu). Používají se zde výsledky ze stochastické analýzy, procesy obnovy, Lindleyovy procesy, principy invariance atd.

Jak jsme mohli vidět v úvodu, reálné systémy jsou mnohdy dosti složité a nelze je modelovat pomocí modelu M/M/1. Jelikož charakteristiky reálných systémů obvykle neumíme spočítat analyticky, nezbývá než přikročit k jejich simulaci. Pomocí simulace dokážeme řešit i relativně dosti komplikované systémy. Další výhodou simulace je, že po snadné modifikaci můžeme prověřovat různé scénáře (jako např. různá rozdělení dob mezi příchody a dob obsluhy, různé uspořádání front atd.). To může být užitečné kupříkladu při vybírání vhodného modelu (srovnáváme výsledky ze simulací různých variant s empirickými daty) nebo při hledání optimálního nastavení systému.

Během simulace generujeme doby mezi příchody jednotlivých zákazníků a doby jejich obsluhy. Podle těchto dob měníme množství zákazníků ve frontě a obsazení jednotlivých stanic. Sledujeme přitom požadované ukazatele a po dostatečně dlouhé době simulace odhadujeme jejich rozdělení či některé jeho

charakteristiky. Tento postup však předpokládá ergodicitu sledovaného systému. To znamená, zhruba řečeno, že můžeme (po případném „odříznutí“ prvních kroků simulace) nahradit stavové průměry průměry časovými. Předpoklad ergodicity je splněn u systémů, kde celková intenzita obsluhy (tj. součet intenzit obsluhy všech stanic) převýší intenzitu příchodů.

3. Simulace v programu R

Stojíme před úkolem simulovat vývoj relativně komplikovaného systému. Budeme chtít sledovat a zaznamenávat různé údaje – hlavně počet zákazníků v systému a jeho změny v čase. Z pohledu zákazníka je podstatnou informací doba čekání na začátek obsluhy, z pohledu provozovatele je důležité znát navíc např. vytížení jednotlivých obslužných stanic, případně počet jejich pracovních cyklů (tj. počet nastartování stanice). Tyto údaje jsou užitečné například při optimalizaci počtu pracovních stanic.

Na sledování těchto dat se tedy v dalším textu zaměříme a odvodíme postup, jak systém v obecném případě simulovat tak, aby bylo možné model snadno upravit pro konkrétní aplikace.

Největším problémem, se kterým je potřeba se vypořádat, je čas. Doby mezi jednotlivými událostmi mohou být spojité náhodné veličiny (např. doba mezi příchody zákazníků má v mnoha modelech exponenciální rozdělení), z podstaty věci proto nestačí systém sledovat v nějaké předem dané posloupnosti diskrétních časů a musíme čas chápat jako spojitý.

Nemůžeme-li systém sledovat v daných diskrétních časech, zkusíme to „vyřídit celé najednou.“ Mohli bychom nagenerovat náhodnou posloupnost dob mezi příchody zákazníků, každému zákazníkovi nagenerovat náhodnou dobu obsluhy, a z těchto dat dopočítat to, co nás zajímá.

U systému GI/GI/1 bude tento přístup fungovat, v případě systému s více obslužnými stanicemi už může narazit. Co když budou mít stanice rozdílnou intenzitu obsluhy? Jak zjistit předem, která stanice bude daného zákazníka obsluhovat, a jak mu tedy nagenerovat správnou dobu obsluhy?

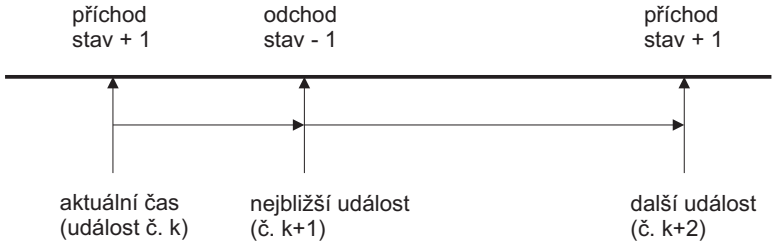
Alternativou je generovat doby obsluhy ne hromadně na začátku, ale teprve ve chvíli, kdy obsluha zákazníka skutečně začíná, a je tedy jasné, která stanice jej bude obsluhovat. Jak uvidíme později, umožní nám tento přístup ještě zobecnit proces generování dob mezi příchody a dob obsluhy zákazníků.

Uvažujme tedy situaci, kdy máme v každém okamžiku naplánováno jen několik nejbližších událostí: víme, kdy přijde do systému další zákazník (to jsme nagenerovali buď předem nebo ve chvíli, kdy přišel minulý zákazník), a víme také, kdy skončí obsluha aktuálně obsluhovaných zákazníků (tento údaj jsme nagenerovali v momentě, kdy začala jejich obsluha).

Pouze v těchto časech dochází ke změně počtu zákazníků v systému, a jde tedy přesně o ty události, které chceme zaznamenat. Mezi těmito okamžiky nedochází k ničemu, co by ovlivnilo dobu čekání zákazníků na obsluhu nebo vytíženost obslužných stanic.

Rozlišujeme tedy dva druhy událostí (příchod nového zákazníka do systému a ukončení obsluhy nějakého zákazníka), které, každá jiným způsobem, znamenají změnu systému.

Použijeme-li pomocnou proměnnou, v níž bude zaznamenán aktuální čas, můžeme se vždy po zpracování některé události s aktuálním časem posunout do minima z časů naplánovaných událostí. Potom si zaznamenáme aktuální čas (čas, kdy k události došlo), všechny další informace, které nas zajímají, a příslušným způsobem upravíme hodnoty všech ukazatelů, které se při dané události mění. Následně aktualizujeme seznam naplánovaných událostí (například po ukončení obsluhy zákazníka může okamžitě začít obsluha dalšího zákazníka ve frontě, a je potřeba vygenerovat dobu jeho obsluhy a naplánovat tak, kdy obsluha skončí), a opět se posunout do času nejbližší naplánované události. Schéma průběhu simulace je naznačeno na obrázku 3.



Obrázek 1: Schéma průběhu uvažovaného simulačního algoritmu.

Tímto postupem tedy sledujeme systém v náhodných okamžicích a zaznamenáváme informace o jeho **změnách** – jde vlastně o **vnořený řetězec v procesu se spojitým časem**.

Jaké výhody nám postupné posouvání aktuálního času do okamžiku nejbližší naplánované události přináší? Můžeme teď například zobecnit způsob generování dob mezi příchody zákazníků a dob obsluhy, a zohlednit v něm větší množství informací, které o systému máme.

Zaveďme funkce `příchod` a `obsluha`, jejichž výstupem bude doba do příchodu dalšího zákazníka a doba obsluhy zákazníka. Tyto funkce budou volány

teprve ve chvíli, kdy jeden zákazník přišel do systému a potřebujeme naplánovat příchod dalšího, resp. ve chvíli, kdy začíná obsluha zákazníka a my potřebujeme vědět, kdy jeho obsluha skončí. Proto mohou mít tyto funkce jako parametry například aktuální čas, počet zákazníků v systému nebo číslo obslužné stanice, tyto údaje jsou totiž v okamžiku volání funkce známé.

Využití může být následující: protože funkce `příchod` závisí na aktuálním čase, můžeme ji upravit tak, aby zohledňovala například ranní špičku nebo naopak mírnější víkendový provoz. Zavisí také na počtu zákazníků ve frontě, můžeme tedy nastavit omezení na délku fronty nebo pracovat s netrpělivostí přišedšího zákazníka – pokud je ve frontě moc lidí, zákazník se může s kladnou pravděpodobností rozhodnout systém opustit hned a doba do příchodu dalšího zákazníka se odpovídajícím způsobem prodlouží. Dále skutečnost, že funkce `obsluha` má jako parametr číslo obslužné stanice, nám umožní modelovat systémy, v nichž obslužné stanice mají různé intenzity obsluhy.

Použití funkcí `příchod` a `obsluha`, které závisí na aktuálním čase a dalších parametrech, nám tedy dává široké možnosti, jak obecný model jednoduše upravit pro konkrétní aplikace. Výsledný skript je navíc přehlednější – generování dob mezi příchody a dob obsluhy je možné upravovat na jednom jasně definovaném místě a ne kdesi uprostřed hlavního simulačního cyklu.

Jádro simulace potom může vypadat nějak takto (uvedený kód je v syntaxi programu R, většina technických detailů je ale pro větší přehlednost nahrazena slovním popisem):

```
for (i in 1:maximální počet zákazníků){
  if (aktcas >= maximální čas){break}
  # simulace končí, pokud je dosaženo maximálního času

  while (min(časy naplánovaných událostí) < čas příchodu
    dalšího zákazníka){
    # někdo skončí obsluhu dřív, než dorazí další zákazník
    aktcas <- min(časy naplánovaných událostí)
    stav <- stav - 1
    záznam události, odebrání zákazníka z obsluhy

    # začátek obsluhy dalšího zákazníka
    # (uvolnila se obslužná stanice)
    if (fronta není prázdná){
      délka obsluhy <- obsluha(aktcas,stav,stanice)
      začátek obsluhy zákazníka, aktualizace fronty
    }
```

```
}
# nejbližší událostí je příchod i-tého zákazníka
aktcas <- čas příchodu dalšího zákazníka
stav <- stav + 1
záznam události

# pokud je volná stanice, bude zákazník rovnou obsluhován
if (některá stanice je volná){začátek obsluhy zákazníka}
  else{zařazení zákazníka na konec fronty}

# určení času příchodu dalšího zákazníka
příchod dalšího zákazníka <- aktcas + prichod(aktcas,stav)}
```

Podmínky ukončující simulaci mohou být různé, často se však používá některá z těchto dvou: dosažení cílového času a zpracovnání daného množství zákazníků. V uvedeném skriptu jsou zkombinovány dvě podmínky – omezení počtu zákazníků, kteří mohou přijít do systému, a kontrola dosažení maximálního času. V úvahu přicházejí i další možnosti, z nichž si můžeme vybírat při řešení konkrétní úlohy, i v tomto případě je obecný model velmi flexibilní.

Použití podmínky ukončení simulace po příchodu daného počtu zákazníků má tu výhodu, že nám dává horní odhad na délku vektorů, které potřebujeme pro záznam informací o vývoji systému. Tato délka je dvojnásobkem maximálního počtu zákazníků (u každého můžeme zaznamenat příchod a odchod ze systému). Práce s vektory předem dané délky je totiž v programu R efektivnější než opakované použití příkazu `cbind(vektor,nový údaj)`, který bychom využili, kdybychom neměli žádný odhad délky vektorů.

Doby čekání zákazníků na začátek obsluhy můžeme sledovat tak, že si u každého zákazníka zaznamenáme čas jeho příchodu (ten je nagenerován v okamžiku příchodu předchozího zákazníka) a čas začátku jeho obsluhy (ten je současně buď časem jeho příchodu a doba čekání je nulová, nebo je časem ukončení obsluhy jiného zákazníka). V obou těchto časech se náš uvažovaný model zastavuje a máme tak možnost událost zaznamenat. Po dokončení simulace můžeme jednoduchým odečtením těchto údajů zjistit doby čekání na začátek obsluhy a podrobit je statistické analýze.

Podobně můžeme sledovat i to, jak dlouho jednotlivé obslužné stanice pracují. Pro každou stanici uvažujme proměnnou, do které budeme nasčítávat jednotlivé doby obsluhy, například v okamžicích jejího ukončení. To nám umožní na konci simulace vyhodnotit, jakou část doby stanice pracovaly.

Stejným způsobem budeme sledovat počet nastartování obslužných stanic. Dá se rozmyslet, že jedinou situací, kdy je potřeba vypnutou stanici

nastartovat, je tato: zákazník přichází do systému, fronta je prázdná, některá stanice nepracuje a zákazník je tedy rovnou obsluhován. V takovém momentě si do proměnné odpovídající této stanici přičteme jedno nastartování a po doběhnutí simulace můžeme tyto údaje analyzovat.

Vhodnou datovou strukturou pro zaznamenávání uvedených informací je *dataframe.* V jedné proměnné tak můžeme mít uchovány všechny údaje o zákaznících, v druhé o obslužných stanicích a v poslední pak záznam o vývoji systému.

4. Ukázka výstupů pro systém M/M/1

Systémem M/M/1 rozumíme takový model s jednou obslužnou stanicí, v němž doby mezi příchody zákazníků i doby obsluhy jsou nezávislé a stejně rozdělené s exponenciálním rozdělením.

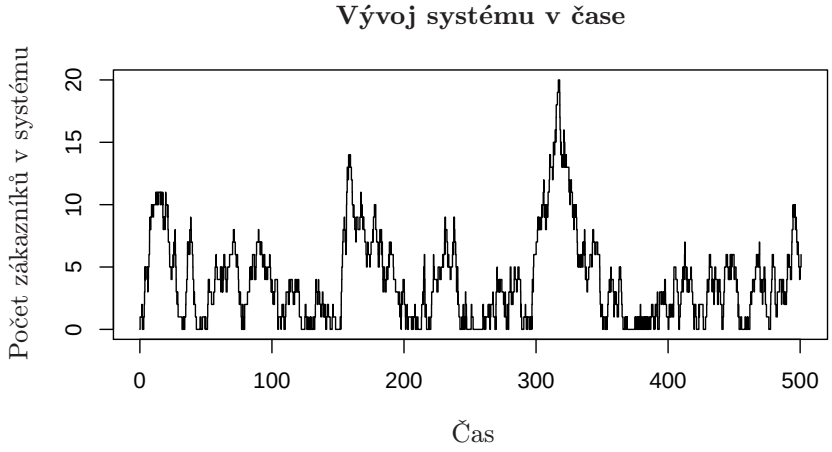
Pomocí výše uvedeného postupu jsme v programu R simulovali vývoj tohoto jednoduchého systému s následujícími parametry: intenzita obsluhy $\mu = 1.2$, intenzita příchodu zákazníků $\lambda = 1$, maximální čas $T = 500$.

Délka simulace je dost velká na to, aby došlo ke stabilizaci systému a mohli jsme odhadovat stacionární charakteristiky. Výsledky simulace jsou následující (obrázky byly vytvořeny v programu R):

- do systému přišlo celkem 518 zákazníků (intenzita příchodu je 1 a délka simulace je 500 časových jednotek),
- obslouženo bylo dohromady 512 zákazníků,
- obrázek 2 ukazuje časový vývoj počtu zákazníků v systému,
- obslužná stanice pracovala přibližně 83.33 % času a nastartována byla 88krát,
- bez čekání bylo odbaveno 88 zákazníků, naopak 424 jich na začátek své obsluhy muselo čekat; histogram popisující dobu jejich čekání na začátek obsluhy je na obrázku 3.

5. Srovnání s teoretickými výsledky – odhad stacionárního rozdělení v modelu M/M/1

Stacionární rozdělení počtu zákazníků v tomto modelu existuje, pokud je intenzita příchodu zákazníků λ menší než intenzita obsluhy μ. Dá se snadno

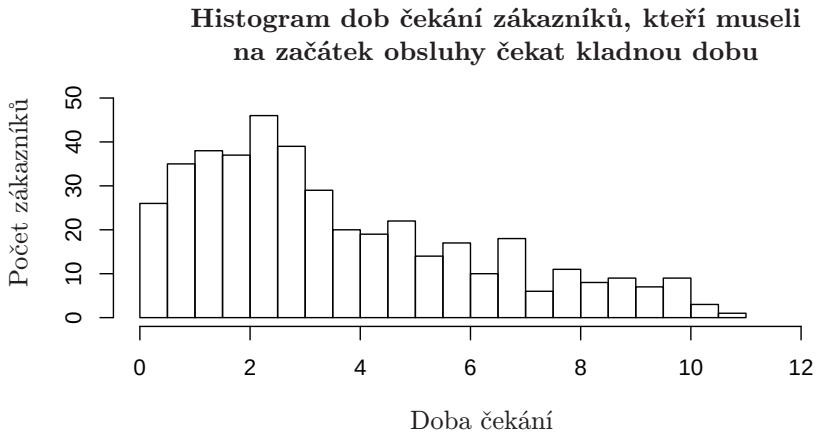


Obrázek 2: Časový vývoj počtu zákazníků v systému M/M/1 s parametry $\lambda = 1$ a $\mu = 1.2$.

ukázat, že v takovém případě je stacionární rozdělení počtu zákazníků v systému geometrické s parametrem $1 - \rho = 1 - \frac{\lambda}{\mu}$, tj. $P[X_t = k] = (1 - \rho)\rho^k$, $k = 0, 1, 2, \ldots$ Hodnotu $\rho = \frac{\lambda}{\mu}$ nazýváme intenzita provozu.

Jednou z možností, jak odhadnout stacionární rozdělení z nasimulovaných údajů, je metoda maximální věrohodnosti. Tak můžeme získat MLE odhady intenzit $\widehat{\lambda}$ a $\widehat{\mu}$ (podle vzorce $\widehat{\lambda} = \frac{1}{\overline{x}}$, kde $\overline{x}$ je aritmetický průměr dob mezi příchody zákazníků; odhad $\widehat{\mu}$ určíme analogicky). Potom už můžeme vypočítat jednotlivé pravděpodobnosti ve stacionárním (geometrickém) rozdělení, v němž za parametr vezmeme hodnotu $1 - \frac{\widehat{\lambda}}{\widehat{\mu}}$ (odhad teoretické hodnoty parametru $1 - \frac{\lambda}{\mu}$).

Alternativou, která se nám nabízí k odhadu stacionárního rozdělení počtu zákazníků v systému, je použití neparametrické metody. Nechť t_k je celková doba, během níž se proces nacházel ve stavu k (tj. v systému bylo k zákazníků) a T je celkový čas simulace. Potom lze odhad stacionárního rozdělení $\{\pi_k\}_{k=0}^{\infty}$ získat jako podíly $\frac{t_k}{T}$, tedy odhad stacionární pravděpodobnosti výskytu stavu k je podíl doby v něm strávené ku celkovému času.



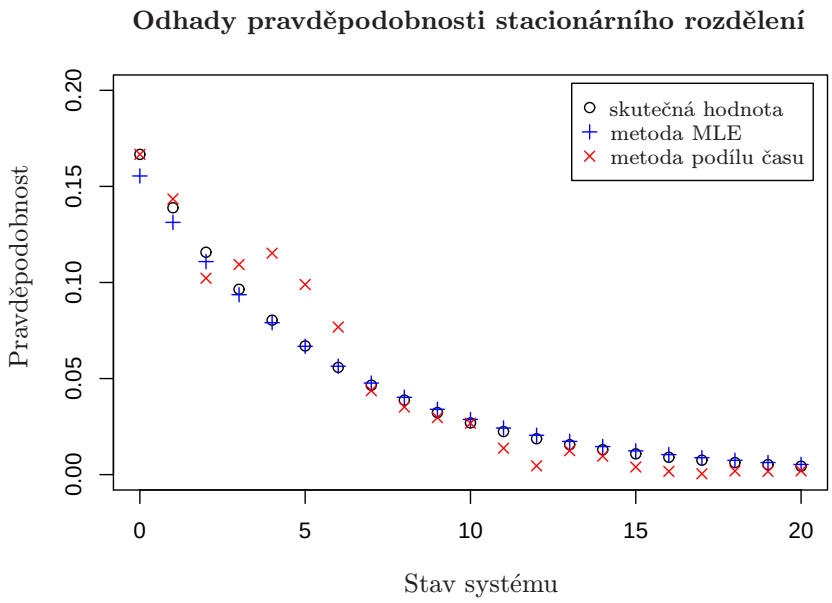
Obrázek 3: Histogram doby čekání na začátek obsluhy těch zákazníků, kteří museli čekat kladnou dobu (v systému M/M/1 s parametry $\lambda = 1$ a $\mu = 1.2$).

Obrázek 4 nabízí srovnání obou popsaných metod odhadování stacionárního rozdělení počtu zákazníků v systému (použitá data pochází ze stejného běhu simulace jako v ukázce výstupů v předchozím oddílu). Metoda vycházející z MLE odhadů poskytuje přesnější výsledky, ale je závislá na předpokladu exponenciálního rozdělení dob mezi příchody zákazníků a dob obsluhy. Naproti tomu metoda založená na podílech časů je použitelná i v obecných modelech a také v situacích, kdy analyzujeme reálná data a platností modelu si nejsme jisti.

6. Možnosti rozšíření modelu

Navrhovaný algoritmus je velmi obecný a po snadné modifikaci jej lze použít na širokou škálu úloh. Některé reálné systémy jsou však velmi komplikované a pro jejich simulaci bude nutné tento model ještě dále rozšířit a zobecnit. Pro představu uveďme alespoň několik typických příkladů.

Výše popsaný algoritmus předpokládá, že zákazníci jsou homogenní. Často se však jednotliví zákazníci liší a při simulaci je třeba jim přiřadit nějaké atributy (s diskrétním či spojitým oborem hodnot), jimiž se pak řídí obsluha zákazníků. Tento atribut se většinou generuje v okamžiku příchodu zákaz-



Obrázek 4: Srovnání metod odhadu pravděpodobností stacionárního rozdělení v systému M/M/1 s parametry $\lambda = 1$ a $\mu = 1.2$.

níka. Příkladem spojitého atributu může být třeba velikost nákladu u nákladního automobilu, která determinuje střední dobu vykládky. Diskrétní atribut bychom použili třeba při simulaci víceúčelového výrobního zařízení. Atributem by zde byl typ výrobku a mohl by určovat dobu přenastavení výrobního zařízení i střední dobu samotné výroby.

Dále můžeme vzpomenout síťové modely z úvodu. Pokud obslužná zařízení nejsou řazena čistě paralelně, ale zákazník během pobytu v systému musí navštívit více stanic, je nutné uvažovaný algoritmus rozšířit. Museli bychom zavést více front a po obsloužení zákazníka by se nejen uvolnila příslušná stanice, ale zároveň by se zvýšila fronta před následující stanicí (případně by tato stanice začala s obsluhou zákazníka). U tohoto typu systémů velmi často rozlišujeme několik typů zákazníků, což vede opět k použití atributů. Hodnota atributu pak určuje, které stanice a v jakém pořadí zákazník navštíví.

Mnohdy se setkáme se systémy, kde je k obsluze zákazníka v obslužném zařízení nutný ještě nějaký dodatečný zdroj, jehož kapacity jsou omezené. Obsluha tedy nemůže začít, dokud není tento zdroj k dispozici, což může prodloužit dobu čekání zákazníka a snížit vytížení stanice. Typickým příkladem jsou poloautomatizovaná zařízení, kde je před zahájením obsluhy nutný zásah pracovníka, samotná obsluha však již probíhá bez jeho asistence. Počet pracovníků pak bývá menší než počet obslužných stanic. Při simulování těchto modelů pak sledujeme nejen vytížení obslužných zařízení, ale také vytížení jednotlivých pracovníků. Samotnou obsluhu pak můžeme rozdělit na dvě fáze – fáze s asistencí pracovníka a fáze bez jeho asistence. Asistovaná fáze začíná při uvolnění zařízení a pracovníka. Po jejím ukončení uvolňujeme pracovníka a zahájíme automatizovanou fázi. Uvolněný pracovník se v případě potřeby přesouvá k jinému zařízení, kde zahájí asistovanou fázi obsluhy.

7. Závěr

V minulých odstavcích byly nastíněny některé problémy, s nimiž se setkáváme při simulaci jednofrontového systému GI/GI/N, a naznačen jeden z možných přístupů k jejich řešení. Čtenář se může tímto postupem inspirovat při vlastní práci se simulacemi podobných či složitějších systémů, nebo jej naopak použít k porovnání s výstupy svých simulací založených na použití jiných metod.

Zájemci, které popsaný přístup zaujal, mohou získat ukázkový skript v programu R po zaslání e-mailu na adresu `jiri.dvorak.rce@gmail.com`.

Tento skript byl vytvořen s cílem aplikovat zde odvozenou metodu simulací v co nejobecnějším případě, a je možné jej snadno upravit pro simulaci konkrétního systému podle volby uživatele. Věříme, že tento skript umožní čtenáři rychle se seznámit s technickými detaily použitého řešení a případně pak aplikovat podobné metody při provádění vlastních simulací.

Reference

[1] Asmussen S.: *Applied Probability And Queues*, 2nd Ed., Springer, 2003.

O ROZDĚLENÍCH S TĚŽKÝMI CHVOSTY
ON DISTRIBUTIONS WITH VERY HEAVY TAILS

Zdeněk Fabián

Adresa: Ústav informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8

E-mail: zdenek@cs.cas.cz

Abstrakt

V článku se zabýváme zobecněnou momentovou metodou v situaci, kdy data pocházejí z distribucí s velmi těžkými chvosty. Místo klasického průměru a rozptylu, která pro ně zpravidla neexistují, jsou tyto distribuce charakterizovány odpovídající skórovou funkcí. Na závěr je uvedena řada příkladů které, jak se autor domnívá, ukazují přednosti navržené metody.

Generalized moment method is applied to the distributions with very heavy tails. Instead of using for their characterization classical mean and variance, corresponding score function is used. A series of examples illustrates advantages of the chosen approach.

1. Úvod

S rozděleními s těžkými chvosty je těžké pořízení. Obyčejná rozdělení mají střední hodnotu, rozptyl, šikmost a špičatost, rozdělení s těžkými chvosty nic takového mít nemusí. Data vybraná z těchto rozdělení jsou podivná, vždyť co je to za data, která se nedají popsat svým průměrem ani výběrovou odchylkou. Můžeme z nich ovšem odhadovat parametry předpokládaného rozdělení, ale jak odhady pro různé modely porovnat, každé rozdělení má jiné, jedno řecké, druhé latinské, a tak statistici často zanedbávají „přízemní“ část dat a věnují se, podle rady pana Hilla, pouze chvostu, který lze, konec konců, popsat jedním γ.

V této práci se pokusím ukázat, že rozdělení s těžkými chvosty jsou rozdělení jako každá jiná a že s daty z nich pocházejícími si není třeba dělat zas až tak těžkou hlavu.

2. Skalární inferenční funkce

Buď $\mathcal{X} \subseteq \mathbb{R}$ otevřený interval a $\eta : \mathcal{X} \to \mathbb{R}$ nějaké spojité rostoucí zobrazení. V pracech publikovaných ve sbornících Robust (viz [1] a seznam prací tam uvedený) jsem ukázal, že rozdělení spojité náhodné veličiny X s nosičem $\mathcal{X} \neq \mathbb{R}$ lze popsat, kromě distribuční funkce F a hustoty f, i funkcí $T(x) =$

$T_G(\eta(x))$ kde g je hustota a $T_G(y) = -g'(y)/g(y)$ skórová funkce rozdělení G, ‚prototypu' náhodné veličiny X, kterým je $Y = \eta(X)$ s nosičem $\mathbb{R}$. Funkci $T(x)$ lze vyjádřit i bez pomoci prototypu jakožto

$$T(x) = \frac{1}{f(x)} \frac{\mathrm{d}}{\mathrm{d}x}(-\frac{1}{\eta'(x)} f(x)), \tag{1}$$

je však třeba definovat transformaci η. Pro porovnání momentů funkce T různých rozdělení je třeba, aby funkce η byla pro dané $\mathcal{X}$ a všechna rozdělení táž. Nejlépe ta, pro kterou je (1) vyjádřena jednoduchým vzorcem pro většinu prakticky užívaných rozdělení, a tou je

$$\eta(x) = \begin{cases} x & \text{if} \quad \mathcal{X} = \mathbb{R} \\ \log(x-a) & \text{if} \quad \mathcal{X} = (a, \infty) \\ \log \frac{x}{1-x} & \text{if} \quad \mathcal{X} = (0,1) \end{cases} \tag{2}$$

(vzorce pro nosič ve formě obecného intervalu jsou v [1]). Po dosazení (2) do (1) dostaneme explicitní tvar funkce T pro různé nosiče,

$$T(x) = \begin{cases} -\frac{f'(x)}{f(x)} & \mathcal{X} = \mathbb{R} \\ -1 - (x-a)\frac{f'(x)}{f(x)} & \mathcal{X} = (a, \infty) \\ -1 + 2x - x(1-x)\frac{f'(x)}{f(x)} & \mathcal{X} = (0,1). \end{cases} \tag{3}$$

Funkce (3) je skalární funkce podobná věrohodnostnímu skóru, je však sestrojená pouze pomocí derivace podle proměnné. Budeme jí namísto dřívějšího core funkce říkat *t-skór* (od *transformation-based score*).

Několik příkladů. t-skór standardního normálního rozdělení je $T(x) = x$. Pro normální rozdělení tedy nic nového. Weibullovo rozdělení s nosičem $\mathcal{X} = (0, \infty)$ má hustotu

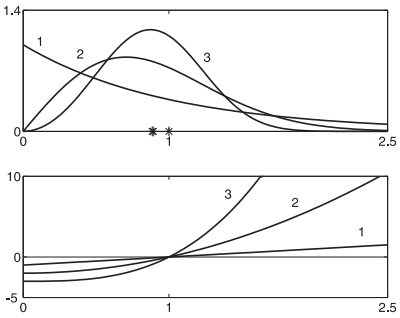
$$f(x) = \frac{c}{\tau}(x/\tau)^{c-1} e^{-(x/\tau)^c}$$

a t-skór (obr. 1)

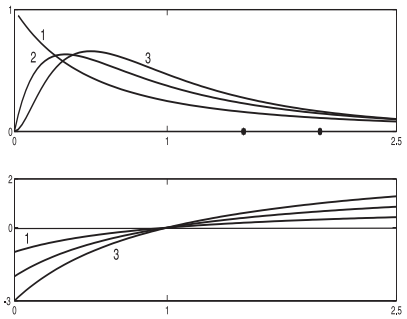
$$T(x) = -1 - x\left(\frac{c-1}{x} + c\left(\frac{x}{\tau}\right)^c\right) = c\left(\left(\frac{x}{\tau}\right)^c - 1\right).$$

Rozdělení beta-prime (beta druhého druhu) s hustotou

$$f(x) = \frac{1}{B(p,q)} \frac{x^{p-1}}{(x+1)^{p+q}}$$



Obrázek 1: Hustoty a t-skóry Weibullova rozdělení, $\tau = 1$, c =1,2,3.



Obrázek 2: Hustoty a t-skóry rozdělení beta-prime, $p = q$=1,2,3.

je rozdělení s těžkým chvostem a jeho t-skór je podle (3) omezená funkce (obr. 2)

$$T(x) = \frac{qx - p}{x + 1}.$$

3. Numerické charakteristiky rozdělení

Těžištěm x^* rozdělení F nazveme řešení rovnice

$$T(x) = 0.$$

Protože

$$0 = T(x^*) = T_G(\eta(x^*)) = -\frac{g'(y^*)}{g(y^*)},$$

je x^* ‚obrazem' módu prototypu (což je lepší než ‚obraz' střední hodnoty prototypu, protože ta nemusí existovat). Střední hodnoty Weibullových rozdělení (hvězdičky v obr. 1) jsou blízké těžišti v bodě $x = 1$. Naopak, v případě beta-prime rozdělení s těžkými chvosty střední hodnoty nevyjadřují nic zajímavého (hvězdičky v obr. 2), jedna ani neexistuje, kdežto těžiště (opět $x^* = 1$) myslím dobře charakterizuje polohu všech tří zobrazených rozdělení na reálné ose.

Uvažujme parametrická rozdělení $F_\theta, \theta \in \Theta \subseteq \mathbb{R}^m$. Může se stát, že x^* je přímo jedním z parametrů. Je-li μ parametrem polohy prototypu G_θ a $\theta = (\mu, \theta_2, \ldots, \theta_m)$, je parametrem $F_{\theta'}$ ‚obraz' parametru polohy prototypu

$$\tau = \eta^{-1}(\mu). \tag{4}$$

Pak $\theta' = (\tau, \theta_2, \ldots, \theta_m)$. Příkladem je třeba Weibullovo rozdělení. Pro třídu těchto rozdělení platí (viz sborníky ROBUST)

$$\eta'(\tau)T(x;\theta) = \frac{\partial}{\partial\tau}\log f(x;\theta). \tag{5}$$

Pomocí derivace podle proměnné jsme tak zkonstruovali věrohodnostní skór pro nejdůležitější parametr, který se obvykle považuje za parametr měřítka, ale který vidím jako parametr polohy těžiště na $(0, \infty)$.

Pro ostatní rozdělení (viz např. beta-prime) položíme

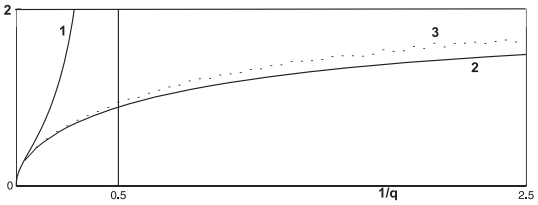
$$S(x) = \eta'(x^*)T(x;\theta) \tag{6}$$

a budeme předpokládat, že (6) má podobný význam jako (5) pro třídu rozdělení s parametrem τ , t.j. že je věrohodnostním skórem pro těžiště. Jako charakteristiku variability rozdělení F pak definujeme veličinu

$$\omega^2 = \frac{1}{\mathrm{E}S^2}, \tag{7}$$

které budeme říkat *t-variance*. Pro pro třídu rozdělení s parametrem τ je to podle (5) převrácená hodnota Fisherovy informace $I_\tau(\theta)$ pro τ, pro ostatní to je něco jako převrácená hodnota Fisherovy informace pro těžiště. Podobně jako x^*, existuje ω^2 i pro rozdělení s těžkými chvosty, její existence vyplývá podle (6) ze vztahu $0 < I_\tau(\theta) < \infty$, což je obvyklá podmínka regularity rozdělení.

t-variance Weibullova rozdělení je zřejmě $\omega^2 = \tau^2/c^2$. Těžiště beta-prime rozdělení je $x^* = p/q$ a protože $\eta'(x^*) = 1/(x^*)^2$ a $\mathrm{E}T^2 = pq/(p+q+1)$, t-variance je $\omega^2 = \frac{(x^*)^2}{\mathrm{E}T^2} = \frac{p(p+q+1)}{q^3}$. Směrodatná odchylka σ a ω rozdělení beta-prime jsou porovnány v obr. 3, σ nemá smysl ani v oboru, ve kterém existuje.



Obrázek 3: σ (1) a ω (2) rozdělení beta-prime ($p = q$) v závislosti na $1/q$, (3) je simulovaná median absolute deviation (MAD).

4. Charakteristiky datových souborů

Namísto výběrové střední hodnoty a výběrového rozptylu, za charakteristiky datového souboru $(X_1, \ldots, X_n)$ vybraného z rozdělení F_θ můžeme považovat odhady těžiště a t-variance. Pro výběry z normálního rozdělení dostaneme totéž, pro výběry z rozdělení s lehkými konci (např. Weibullovo rozdělení) přibližně totéž, pro rozdělení s těžkými konci jsou to nové charakteristiky. Položíme-li $\hat{x}^* = x^*(\hat{\theta}_{ML})$ a $\hat{\omega}^2 = \omega^2(\hat{\theta}_{ML})$, zkonstruujeme odhady těžiště a t-variance z maximálně věrohodných odhadů parametrů. Spočteme-li třeba maximálně věrohodné odhady $\hat{p} = \hat{p}_{ML}$ a $\hat{q} = \hat{q}_{ML}$ datového souboru vybraného z beta-prime rozdělení, určíme těžiště a t-varianci souboru jako $\hat{x}^* = \hat{p}/\hat{q}$ a $\hat{\omega}^2 = \hat{p}(\hat{p} + \hat{q} + 1)/\hat{q}^3$.

Rodina, do které patří F_θ, je obvykle známa jen přibližně. Rozumné je vyzkoušet několik předpokladů, výsledné odhady je však obtížné porovnat neboť různé rodiny jsou parametrizovány různými způsoby. Je však snadné porovnávat odhady těžiště a t-variance. Z tohoto hlediska nemusí být odhady parametrů konečným cílem zpracování, ale prostředkem k sestrojení výběrového těžiště a výběrové t-variance a snad i nějakých vyšších t-skórových momentů, to ale ještě neumím.

5. Zobecněné momentové odhady

Definujeme-li obecný t-skórový moment rozdělení F s t-skórem T jako

$$\mathrm{E}T^k = \int_{\mathcal{X}} T^k(x;\theta)\ \mathrm{d}F_\theta,$$

mají rovnice pro odhady parametrů zobecněnou momentovou metodou tvar

$$\frac{1}{n}\sum_{i=1}^{n} T^k(x_i;\theta) = \mathrm{E}T^k(\theta) \qquad k = 1, \ldots, m. \tag{8}$$

Podle [2] jsou odhady parametrů z (8) konzistentní a asymptoticky normální. t-skóry jsou často dány jednoduchými vzorci a ‚hodí se ke svému rozdělení', takže momenty $\mathrm{E}T^k(\theta)$ jsou často jednoduchými funkcemi parametrů.

V nejjednodušším případě lze t-skór vyjádřit ve tvaru $T(x;x^*)$. Snadno se ukáže, že v tomto případě odhad

$$\hat{x}^* : \qquad \sum_{i=1}^{n} T(x_i;x^*) = 0 \tag{9}$$

je $AN(x^*, \sigma_*^2/n)$, kde

$$\sigma_*^2 = \frac{\mathrm{E}T^2}{\mathrm{E}[\partial T(x;x^*)/\partial(x^*)]^2}. \tag{10}$$

Konfidenční intervaly pro těžiště se určí docela snadno. V [1] jsme zavedli jakousi ‚vzdálenost' (diferenci) bodů $x_1, x_2 \in \mathcal{X}$ ve výběrovém prostoru jako

$$d(x_1, x_2) = T(x_2; \theta) - T(x_1; \theta).$$

Snadno se ověří, že platí

Věta 1. $\sqrt{n}d(x^*, \hat{x}^*)$ je $AN(0, [T'(x^*)]^2\sigma_*^2)$, kde $T'(x^*) = \frac{\mathrm{d}T(x;\theta)}{\mathrm{d}x}|_{x=x^*}$ a σ_*^2 je dáno vzorcem (10).

Označíme-li $\lambda_n = u_{\alpha/2}/\sqrt{n}$ kde $u_{\alpha/2}$ je $(\alpha/2)$-tý kvantil normálního rozdělení a nahradíme-li x^* a σ_* ve jmenovateli jejich odhady, dostaneme přibližný $(100-\alpha)\%$ konfidenční interval pro x^* z podmínky

$$|T(x^*; \hat{x}^*)| \leq \lambda_n |T'(x^*)| \hat{\sigma}_*. \tag{11}$$

6. Příklady

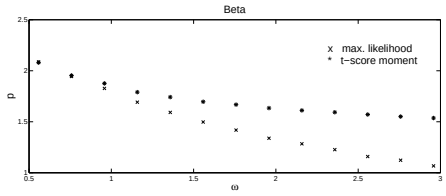
Příklad 1. t-skór rozdělení beta ($Beta(p, q)$) s nosičem $(0, 1)$ a hustotou

$$f(x) = \frac{1}{B(p, q)} x^{p-1}(1-x)^{q-1}$$

je podle (1) funkce $T(x) = (p+q)x - p$. Těžiště rozdělení, řešení rovnice $T(x) = 0$, je $x^* = p/(p+q)$. Protože $T(x)$ je lineární, jsou rovnice (8) obyčejnými momentovými rovnicemi. Je známo, že odhady parametrů p a q z nich určené nejsou vydatné, protože však je t-skór omezená funkce, lze očekávat že budou s rostoucím ω robustní. Protože $\mathrm{E}T^2 = pq/(p+q+1)$, je podle (7) $\omega^2 = \frac{pq(p+q+1)}{(p+q)^4}$. Na obrázku 4 je porovnání průměrů maximálně věrohodných odhadů $\hat{p}_{ML}$ s průměrnými momentovými odhady $\hat{p}$ z kontaminovaného rozdělení $F_{kont.} = 0.9 * Beta(2, 2) + 0.1 * Beta(\omega)$ při 5000 výběrech. Momentový odhad je méně vychýlený.

Příklad 2. Paretovo rozdělení s nosičem $(1, \infty)$ a hustotou $f(x; c) = cx^{-(c+1)}$. Maximálně věrohodný odhad c je $\hat{c}_1 = n/\sum_{i=1}^n \log x_i$. Zvolíme-li podle (2) $\eta(x) = \log(x-1)$, je t-skór $T(x) = c(1 - x^*/x)$, kde $x^* = (c+1)/c$. Z rovnice (9) plyne, že $\hat{x}^* = \bar{x}_H$ je harmonický průměr a odhad parametru c je

$$\hat{c}_2 = \frac{1}{\bar{x}_H - 1}.$$



Obrázek 4: Maximálně věrohodný a momentový odhad těžiště při rostoucí kontaminaci beta rozdělení.

Z průměrných hodnot obou odhadů z 5000 výběrů délky N pro $c = 1$ (Tabulka 1) plyne, že $\hat{c}_2$ má sice větší rozptyl, ale odhad je méně vychýlený.

odhad	N=12	N=25	N=50	N=75
$\hat{c}_1$	1.086	1.041	1.023	1.014
$\hat{c}_2$	1.052	1.028	1.010	1.009
$\mathrm{Var}(\hat{c}_1)$	0.342	0.215	0.144	0.116
$\mathrm{Var}(\hat{c}_2)$	0.371	0.241	0.166	0.135

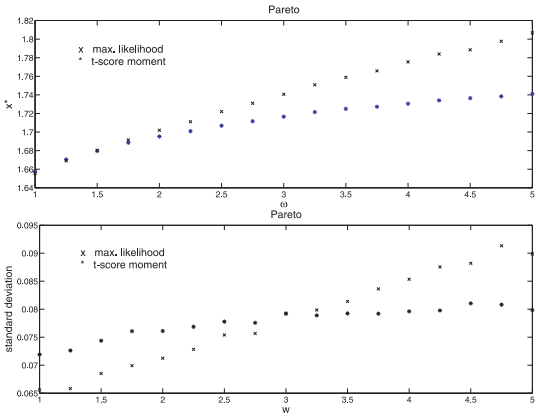
Tabulka 1: Odhady těžiště Paretova rozdělení, $c = 1$.

Nové odhady jsou ovšem výrazně lepší v případě kontaminovaného rozdělení. Protože $\mathrm{E}T^2 = c/(c+2)$, t-variance je $\omega^2 = \frac{(c+2)}{c^3}$. Hodnoty $\hat{x}^*_{ML} = x^*(\hat{c}_1)$ a $\hat{x}^*$ pro výběry z kontaminovaného rozdělení $F_k = 0.9P(1)+0.1P(\omega)$, kde $P(\omega)$ značí Paretovo rozdělení s t-variancí ω^2, jsou znázorněny na obr. 5. Momentový odhad je robustní.

Příklad 3. Výběry z pěti různých rozdělení s nosičem $(0, \infty)$ byly generovány s parametry určenými tak, že pro každé rozdělení $x^* = 1$ a $\omega = 1.118$ (první sloupec Tabulky 2). Odhady těžiště byly určovány za předpokladu, že výběry délky 50 bodů pocházejí z každého z těchto rozdělení (první řádek Tabulky 2). Poslední dvě jsou rozdělení s těžkým chvostem. Hodnoty v Tabulce 2 jsou průměry odhadů z 5000 experimentů.

Příklad 4. Odhadneme dolní mez *Paretova rozdělení* na nosiči (γ, ∞) s hustotou

$$f(x) = \frac{c\gamma^c}{x^{c+1}}.$$



Obrázek 5: Maximálně věrohodné a momentové odhady těžiště a jejich standardní odchylka při rostoucí kontaminaci Paretova rozdělení.

$\hat{x}^*$	gamma	Weibull	lognorm	beta-prime	inv.gamma
gamma	1.000	0.94	0.60	0.49	0.12
Weibull	1.06	1.005	0.64	0.53	0.15
lognormal	1.66	1.66	1.01	1.010	0.63
beta-prime	2.00	1.77	1.008	1.01	0.54
inv.gamma	84.4	4.71	1.70	2.13	1.022

Tabulka 2: Odhady těžiště z různých rozdělení za různých předpokladů o modelu.

Podobně jako v příkladu 2 určíme t-skór $T(x) = c(1 - x^*/x)$, těžiště $x^* = \gamma(c+1)/c$ a t-varianci $\omega^2 = \frac{\gamma^2(c+2)}{c^3}$. Momentové rovnice (8) mají tvar

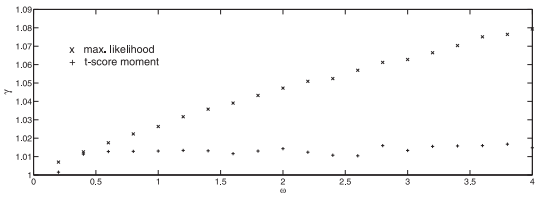
$$\begin{aligned} \sum_{i=1}^{n}(1 - x^*/x_i) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}(1 - x^*/x_i)^2 &= \frac{1}{c(c+2)}, \end{aligned}$$

takže opět $\hat{x}^* = \bar{x}_H$. Označme $\bar{x}_{2H} = n/\sum_{i=1}^{n} 1/x_i^2$ a $\rho = \frac{\bar{x}_{2H}}{\bar{x}_H^2 - \bar{x}_{2H}}$, z druhé rovnice $\hat{c} = \sqrt{(1+\rho)} - 1$ a nakonec $\hat{\gamma} = \bar{x}_H \hat{c}/(\hat{c}+1)$. Odhad bere v úvahu, že dolní mez může být nižší než pozorované hodnoty.

Na obrázku 6 je zachycena závislost maximálně věrohodného odhadu $\gamma_{ML} = x_{(1)}$ a odhadu

$$\hat{\gamma}_M = \min(\hat{\gamma}, x_{(1)}).$$

Je patrné, že upravený momentový odhad dolní meze zůstává i s ‚těžknoucím chvostem' rozdělení blízko skutečné hodnoty.



Obrázek 6: Odhad dolní meze Paretova rozdělení.

7. Závěr

Nově zavedené charakteristiky polohy a variability rozdělení umožňují porovnávat modely s různými parametry. Jejich odhadů lze použít jako charakteristiky datových souborů včetně souborů z rozdělení s těžkými chvosty.

Reference

[1] Fabián Z. (2006). *Nové charakteristiky rozdělení a výběrů z rozdělení.* Sborník Robust'2006, 459 – 466.

[2] Fabián, Z. (2001). *Induced cores and their use in robust parametric estimation.* Communication in Statistics, Theory Methods **30**, 537 – 556.

[3] Fabián, Z. (2008). *New measures of central tendency and variability of continuous distributions.* Communication in Statistics, Theory Methods **37**, 159 – 174.

[4] Fabián, Z., (2009). *Confidence intervals for a new characteristic of central tendency of distributions.* Communication in Statistics, Theory Methods **38**, 1804 – 1814.

Poděkování: Práce byla podpořena grantem AV ČR 1ET 400300513.

JE VĚTŠÍ ROZDÍL MÉNĚ VÝZNAMNÝ?
IS A BIGGER DIFFERENCE LESS SIGNIFICANT?

Josef Tvrdík

Adresa: Ostravská univerzita, PřF, KIP, 30. dubna 22, CZ-701 03 Ostrava

E-mail: josef.tvrdik@osu.cz

Klíčová slova: párové testy, předpoklady a jejich porušení, významnost.

Abstrakt

Sdělení se zabývá jednou úlohou klinického šetření, kde jedna odlehlá hodnota způsobila porušení předpokladů pro aplikaci párového t-testu a její ponechání ve zpracovávaných datech způsobilo, že zlepšení pacientů bylo statisticky nevýznamné. Po vyřazení tohoto nejvíce léčbou zlepšeného pacienta bylo paradoxně celkové zlepšení stavu pacientů léčbou významné při menším průměrném rozdílu zlepšení.

A real-world simple task of clinical data analysis is discussed. One outlaying value in pair comparison caused the violation of assumptions for application of paired t-test and non-significant result of paired comparison. If the outlying value is deleted, then the less average difference becomes significant.

1. Úvod

V jednom klinickém šetření bylo zjišťováno, zda léčení ovlivňuje hodnoty 10 veličin změřených na 16 pacientech před a po léčbě. Všechny veličiny byly měřeny v rozdílové škále a pro statistické zpracování je bylo možné považovat za spojité. Zcela přirozeně se nabízí užít párový t-test. Pro prvních devět porovnání to byl vhodný postup, který v několika případech vedl k oprávněnému zamítnutí nulové hypotézy o shodě středních hodnot ve prospěch oboustranné alternativy.

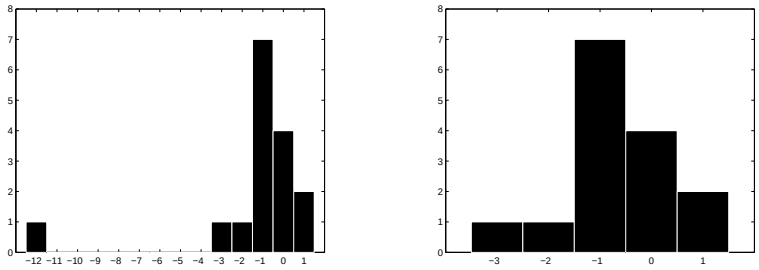
2. Méně je více

U desáté veličiny nastala zajímavě podivná situace. Hodnoty této veličiny jsou uvedeny v tabulce 1, hodnoty před léčbou jsou označeny x_i, hodnoty zjištěné po léčbě jsou označeny y_i, $d_i = y_i - x_i$, $i = 1, 2, \ldots, 16$. Vidíme, že u pacienta $i = 6$ byla hodnota x_6 extrémně velká, rovněž tak i rozdíl d_6. Ukázalo se, že výjimečné zlepšení léčbou u pacienta číslo 6 „kazí“ hodnocení léčby, což vidíme na první výsledkovém řádku v tabulce 2. Tato odlehlá hodnota ovlivní

Tabulka 1: Naměřené hodnoty před a po léčbě a jejich rozdíl.

i	x_i	y_i	$d_i = y_i - x_i$
1	3	2	-1
2	3	4	1
3	4	3	-1
4	3	2	-1
5	1	0	-1
6	**12**	0	$\mathbf{-12}$
7	0	0	0
8	4	1	-3
9	3	3	0
10	2	1	-1
11	2	2	0
12	3	2	-1
13	2	1	-1
14	2	0	-2
15	0	0	0
16	1	2	1

empirické rozdělení rozdílů tak výrazně, že musíme zamítnout normalitu, viz sloupce *Skew*, *Curt* a *Omni*, kde je počtem hvězdiček vyznačena dosažená úroveň významnosti ve třech obvykle užívaných testech normality [2], viz také histogram na obrázku 1a. Pak ovšem bychom párový t-test neměli použít a jeho výsledek je v tabulce 2 uveden jen proto, abychom ukázali, že vyjde nevýznamný, tzn. neopravňuje k zamítnutí nulové hypotézy.



a) všichni pacienti **b)** bez $i = 6$

Obrázek 1: Histogramy empirických rozdělení rozdílů d.

			Párový t-test					Wilcoxon	
d_6	$\overline{d}$	s_d	t	p	*Skew*	*Curt*	*Omni*	z	p
-12	$-1{,}38$	3,01	$-1{,}828$	0,0875	$\star\star\star$	$\star\star\star$	$\star\star\star$	2,374	0,0176
mis	$-0{,}67$	1,05	$-2{,}467$	0,0271				2,153	0,0313
-4	$-0{,}88$	1,31	$-2{,}671$	0,0174				2,374	0,0176

Tabulka 2: Výsledky párových testů pro různé hodnoty d_6.

Statistickému hodnocení léčby párovým t-testem však můžeme paradoxně pomoci, když vyloučíme ze zpracování právě nejvíce zlepšeného pacienta $i = 6$, jak vidíme na druhém výsledkovém řádku v tabulce 2. Pak normalita narušena není (viz také obrázek 1b) a dosažená úroveň významnosti u párového t-testu vede ke kýženému zamítnutí nulové hypotézy. Pacient 6 by hodnocení také nekazil, kdyby se zlepšil sice nejvíce, ale „přiměřeně“, např. tak, aby hodnota $d_6 = -4$. Pak vidíme na posledním řádku v tabulce 2, že dosažená úroveň významnosti párového t-testu je 0,0174, a přitom opět předpoklad normality porušen není.

Pokud se nechceme trápit s tím, že nejvíce zlepšený pacient zhoršuje hodnocení léčby, pak zapomeňme na párový t-test a použijme neparametrický Wilcoxonův test, který zachází s extrémní hodnotou stejně, ať je odlehlá hodně či nepatrně, viz poslední dva sloupce v tabulce 2, kde jsou uvedeny hodnoty asymptoticky normálně rozdělené z statistiky s korekcí nespojitosti a dosažená úroveň významnosti při oboustranné alternativě. Pro pochybovače, i užití přesných kritických hodnot pro tento test vede k zamítnutí nulové hypotézy, kritická hodnota pro $\alpha = 0,05$ je 29 [1], hodnota Wilcoxonovy statistiky je 18. K zamítnutí nulové hypotézy by dokonce v této úloze postačil i znaménkový test, jak se čtenář může snadno a rychle přesvědčit vlastním výpočtem.

3. Závěr

Původní verse tohoto článku se vešla na jednu a půl stránky. Pohled na nová čísla Informačního Bulletinu ale ukázal, že je potřeba dodat klíčová slova, abstrakt, členění na sekce a další náležitosti. Jsem tedy nucen vyslovit i nějaký závěr, který byl původně ponechán na čtenáři.

Příčiny zdánlivého paradoxu, že menší rozdíl je významnější, jsou v této jednoduché úloze okamžitě viditelné, stačí se podívat na hodnoty průměrného rozdílu a jeho směrodatné odchylky v tabulce 2. Obávám se, že u komplikovanějších úloh příčiny podobně paradoxních výsledků tak průhledné nejsou.

Nezbývá než připomenout úsloví, zmiňované v jedné přednášce na CompStatu 1984 a jehož autora neznám: *„There is no routine statistical question, there is questionable statistical routine“*. Pokud znáte autora tohoto stále platného úsloví, sdělte mi jeho jméno, ať vím, koho si mám hodně vážit.

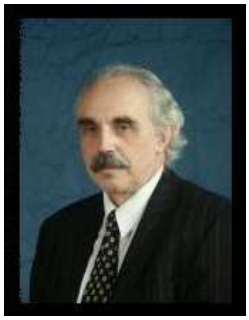
Reference

[1] Anděl J. (1993) *Statistické metody*, první vydání, MATFYZPRESS, Praha.

[2] Hintze J. (2001) *NCSS, Number Cruncher Statistical System*, Kaysville, Utah, `http://www.ncss.com/`.

Doc. RNDr. Felix Koschin, CSc. (1946–2009)

Jitka Langhamrová



Doc. RNDr. Felix Koschin, CSc., se narodil 15. února 1946 v Olomouci. Vystudoval obor matematická statistika na Univerzitě Karlově v Praze. Zde také v roce 1969 promoval. V roce 1981 získal titul RNDr. v oboru pravděpodobnost a matematická statistika. Titul CSc. v oboru ekonomická statistika obhájil v roce 1984 na VŠE v Praze a zde se také v roce 1994 habilitoval docentem v oboru statistika.

Po studiích pracoval na Vysoké škole ekonomické v Praze. Od roku 1969 do roku 1979 zde působil jako odborný asistent na katedře statistiky. V následujících letech až do roku 1983 pracoval také ve Výzkumném pracovišti VŠE Praha, v letech 1983–1990 opět na katedře statistiky. Součástí tehdejší katedry statistiky byla i Laboratoř demografie. Od jejího vzniku se významně podílel na všech analytických výstupech tohoto pracoviště. V roce 1990 vznikla na VŠE samostatná katedra demografie, tehdy ještě pod vedením prof. Vladimíra Roubíčka.

Doc. Felix Koschin, CSc., byl jedním z nejaktivnějších členů této nově vzniklé katedry. Na Vysoké škole ekonomické pracoval doc. Koschin celých čtyřicet let. Působil také v řadě akademických funkcí. V roce 1990 byl proděkanem pro pedagogiku Národohospodářské fakulty VŠE v Praze. V letech 1990–1995 byl zástupcem vedoucího katedry demografie VŠE v Praze, od roku 1995 až do konce svého života byl jejím vedoucím. Od roku 2006 pracoval na VŠE také jako prorektor pro studijní a pedagogickou činnost. Ve funkci prorektora pro studijní a pedagogickou činnost se výrazně zasloužil

o to, že VŠE v Praze získala jako první vysoká škola v České republice pro roky 2009–2013 certifikáty Evropské komise ECTS Label a DS Label.

Doc. Koschin byl znám především jako demograf. Byl místopředsedou České demografické společnosti, mezi aktivní členy patřil již od doby ukončení studia v roce 1969. Byl členem redakční rady časopisu Demografie i členem dalších organizací jako například: Jednota Českých matematiků a fyziků, Česká statistická společnost, Česká společnost aktuárů. Dále byl členem SŠDS (Slovenská štatistická a demografická spoločnosť) a EAPS (European Association for Population Studies) a řady dalších institucí. Byl též dlouholetým zástupcem České republiky v projektu OECD INES.

Také jako pedagog odvedl obrovský kus práce. Byl přednášejícím i cvičícím kurzu statistiky a přednášel demografii na všech typech studia VŠE. Prakticky se podílel na tvorbě nových demografických kurzů a byl garantem řady z nich.

Doc. Koschin patřil k uznávaným osobnostem v oboru demografie nejenom v České republice, ale i v zahraničí. Prakticky po celý svůj život úzce spolupracoval s Českou demografickou společností, byl nejenom jejím členem, ale aktivně pracoval v Hlavním výboru ČDS. Jako odborník a nadšený propagátor demografie vystupoval často v médiích. Rozsáhlá byla také jeho publikační činnost (byl autorem a spoluautorem tří monografií, členem autorského kolektivu čtyř encyklopedií, autorem více než 100 článků a příspěvků na národních i mezinárodních konferencích, desítek vysokoškolských učebních textů, řady výzkumných studií, odborných překladů, recenzí a odborných posudků atp.).

Mezi oblasti jeho odborného zájmu patřila problematika spojená s úmrtností, plodností, zabýval se demografickou metodologií, demografickými prognózami, zajímal se o aktuárskou demografii. Byl zakladatelem kurzu aplikované demografie, demografických modelů, vícestavové demografie. V demografii spojoval svoje matematické schopnosti a využíval znalosti statistických metod. Pod jeho vedením vznikaly na katedře demografie VŠE projekce obyvatelstva nejenom České republiky, ale také za menší územní celky. Vypracoval jako autor či vedoucí kolektivu řadu demografických studií využívaných v praxi, byl řešitelem a spoluřešitelem řady grantů. V poslední době se intenzivně věnoval problematice reprodukce lidského kapitálu a lidským zdrojům z pohledu národního hospodářství. Velmi dobře si uvědomoval jaký je význam demografických studií a prognóz pro řízení jednotlivých regionů nejenom v České republice.

V docentu Felixi Koschinovi ztrácíme výraznou osobnost, která prakticky celý svůj profesní život spojila s demografií, a především skvělého člověka a zaníceného pedagoga. Demografie v něm ztratila velmi výraznou osobnost,

která je nezastupitelná. Připomeňme si jeho živé diskuze s ostatními kolegy, nejenom demografy. Schopnost přesvědčivé argumentace byla jeho výrazným rysem. Často přišel s něčím, s čím nikdo nepočítal nebo to nikoho ani nenapadlo. Doc. Koschin se také významně zasloužil o zviditelnění demografie jako oboru důležitého pro praxi. Patřil k jedněm z mála demografů, kteří dokázali zpracovat kvalitní prognózy obyvatelstva. Byl řešitelem řady demografických studií, využívaných v praxi.

Docent Koschin nás opustil nečekaně. Zemřel 18. 8. 2009 v plném pracovním nasazení. Prakticky do poslední chvíle života dělal to, co měl rád. Byl zapáleným turistou, miloval hory. Především slovenské hory. V létě také rád jezdil na kole, v zimě na běžkách. Ti z nás, co měli možnost se výletů s ním účastnit, mně dají za pravdu, že z každé společné akce jsme měli nezapomenutelné zážitky.

Osobnost docenta Koschina je stále mezi námi. Česká demografie a statistika ztratila ve Felixovi Koschinovi svého čelného představitele. Chybí jistě všem, kteří ho blíže znali.

SPOLEK MLADÝCH STATISTIKŮ VŠE, O. S.

Jana Langhamrová, Kristýna Vltavská

Adresa: VŠE, nám. W. Churchilla 4, 130 67 Praha 3

Dne 2. června 2009 bylo Ministerstvem vnitra ČR zaregistrováno nově vzniklé občanské sdružení *Spolek mladých statistiků VŠE*. Sdružení založila čtveřice studentů Vysoké školy ekonomické Ing. Petra Coufalová, Jana Langhamrová, Ing. Tomáš Löster a Ing. Kristýna Vltavská. Prezidentkou spolku byla zvolena Kristýna Vltavská. Zakladatelé spolku si dali za cíl popularizovat statistiku a s ní související obory a to nejen mezi studenty Vysoké školy ekonomické v Praze. Sdružení se nebude úzce zaměřovat pouze na obor statistika, ale rádo by podpořilo rozvoj takových oborů jako například ekonomická statistika, demografie a dalších. Spolek bude pořádat a spolupořádat vzdělávací semináře, konference a také by rád zapojil studenty statistických oborů do vědecké činnosti Fakulty informatiky a statistiky VŠE. Spolek mladých statistiků bude podporovat své členy v aktivní účasti na mezinárodních a národních konferencích, soutěžích a seminářích.

Spolek očekává budoucí spolupráci i s dalšími sdruženími podobného typu. Kupříkladu se *Studentským demografickým klubem*, působícím pod záštitou České demografické společnosti. V budoucnu zakladatelé spolku předpokládají i možnost mezinárodní spolupráce s mladými statistiky a demografy na

zahraničních vysokých školách a institucích. Spolek mladých statistiků je určen všem zájemcům o statistiku a demografii ve věku 15 až 35 let. Ostatní zájemci se mohou stát sympatizanty sdružení a spolupodílet se tak na jeho chodu. Cílem spolku není separace začínajících statistiků, ale naopak propojení různých generací českých i zahraničních statistiků, demografů a nebo případných zájemců o tyto obory. Podle zakladatelů Spolku je výměna zkušeností a názorů se zkušenými pedagogy a odborníky z praxe pro nastupující generaci mladých statistiků neocenitelná. Další informace a přihlášku do Spolku mladých statistiků VŠE naleznete na `http://sms.vse.cz/`.

V úterý 1. 12. 2009 se od 17:00 na VŠE konala první akce Spolku mladých statistiků. Přednášku na téma *Consumer's Preferences for Italien Wine* přednesla Dr. Maria Bonaria Lai z univerzity v Cagliari, Itálie. Po přednášce následovala ochutnávka vybraných moravských vín a setkání s předsedou vlády ČR Ing. Janem Fischerem, CSc., místopředsedou naší společnosti, nazvané *Česko očima nepolitika v politice*.

20. ROČNÍK MEZINÁRODNÍ MATEMATICKÉ SOUTĚŽE VOJTĚCHA JARNÍKA

Jaroslav Hančl, Jan Šustek, Jan Štěpnička a T. Sochor

Adresa: Ostravská univerzita, PřF, KM, 30. dubna 22, CZ-701 03 Ostrava

Ve čtvrtek dne 25. 3. 2010 se na Katedře matematiky Přírodovědecké fakulty Ostravské univerzity uskuteční už 20. ročník Mezinárodní matematické soutěže Vojtěcha Jarníka. Tato soutěž vznikla před devatenácti lety a je určena pro studenty matematiky studující na vysoké škole. Pořádá se v Ostravě každým rokem a dělí se do dvou kategorií. První kategorie je pro studenty prvního a druhého ročníku vysoké školy a druhá kategorie je pro studenty třetího, čtvrtého a pátého ročníku. Za první místo v každé kategorii je 15 000 Kč za druhé 10 000 Kč a za třetí 5 000 Kč.

Minulého ročníku se zúčastnilo 164 studentů matematiky ze 40 nejdůležitějších univerzit a 15 států nejen v Evropě. Ostravská soutěž Vojtěcha Jarníka je nejstarší a nejprestižnější matematickou soutěží v Evropské unii. Její organizátoři jsou student studenta a studenti studenta studenta jednoho z nejlepších českých matematiků všech dob, pražského matematika Vojtěcha Jarníka.

Více informací o této každoroční soutěži naleznete na internetových stránkách `http://vjimc.osu.cz/`.

Vážené kolegyně, vážení kolegové,

výbor společnosti si Vás dovoluje pozvat na výroční zasedání, jež se uskuteční ve čtvrtek 28. ledna 2010 od 13.00 na VŠE v Praze. Pozvané přednášky přednesou kolegové Jiří Militký a Jan Picek z Technické univerzity v Liberci.

ISSN 1210 – 8022. Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.
Časopis je zařazen do Seznamu Rady, více viz `http://www.vyzkum.cz/`.

Předseda společnosti: doc. RNDr. Gejza Dohnal, CSc., ÚTM FS ČVUT v Praze, Karlovo náměstí 13, 121 35 Praha 2, e-mail: `gejza.dohnal@fs.cvut.cz`

Ediční rada: prof. Ing. Václav Čermák, DrSc. (předseda), prof. RNDr. Jaromír Antoch, CSc., doc. Ing. Josef Tvrdík, CSc., RNDr. Marek Malý, CSc., doc. RNDr. Jiří Michálek, CSc., doc. RNDr. Zdeněk Karpíšek, CSc. a prof. Ing. Jiří Militký, CSc.

Techničtí redaktoři: doc. RNDr. Gejza Dohnal, CSc., `gejza.dohnal@fs.cvut.cz`
a ing. Pavel Stříž, Ph.D., `striz@fame.utb.cz`

Pokyny autorům: `<http://www.statspol.cz/bulletiny/sablony.htm>`

WEB server: `<http://www.statspol.cz/>`