INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 34, číslo 1, březen 2023

Obsah

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je sázen v programu TEX, ve formátu LuaHBTEX s písmy balíku CSfonts.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

Předseda společnosti: Mgr. Ondřej Vencálek, Ph.D., Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta Univerzity Palackého, 17. listopadu 12, 771 46 Olomouc, e-mail: ondrej.vencalek@upol.cz.

Redakce: prof. RNDr. Gejza Dohnal, CSc. (šéfredaktor), prof. RNDr. Jaromír Antoch, CSc., doc. RNDr. Zdeněk Karpíšek, CSc., RNDr. Marek Malý, CSc., doc. RNDr. Jiří Michálek, CSc., prof. Ing. Jiří Militký, CSc., doc. Ing. Iveta Stankovičová, PhD., Mgr. Ondřej Vencálek, Ph.D.

Redaktor časopisu: Mgr. Ondřej Vencálek, Ph.D., ondrej.vencalek@upol.cz.
Informace pro autory jsou na stránkách společnosti, http://www.statspol.cz/.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.

K OSMDESÁTINÁM ZDEŇKA FABIÁNA
ON THE 80TH BIRTHDAY OF ZDENĚK FABIÁN

Jan Klaschka

Adresa: Ústav informatiky AV ČR, Pod Vodárenskou věží 271/2, 182 07 Praha 8

E-mail: klaschka@cs.cas.cz

Redakce IB mě požádala o několik slov u příležitosti životního jubilea Zdeňka Fabiána. Rád vyhovím, jen jaksi po svém.

Zdeněk se projevuje jako velmi kreativní osobnost na poli vědeckém a uměleckém a musím se přiznat, že jsem odjakživa věnoval větší pozornost jeho písničkám i jiné literární tvorbě než jeho vědeckým výsledkům. V tom si ostatně s časopisem ČStS můžeme podat ruce. Nejen že v posledních letech je na stránkách IB o Zdeňkovi slyšet hlavně jako o autorovi beletrie, viz anonci tří jeho knih v IB 4/2020 a čtvrté i s ukázkami veršovaných aktualit z covidových časů v mimořádném čísle téhož ročníku, ale v letech ještě dřívějších Zdeněk svá dílka z oblasti krásného písemnictví publikoval přímo v IB. Jeho beletrizované zápisky ze zahraničních pracovních cest zaplnily polovinu mimořádného čísla v r. 1994 a celé číslo 4/1997, jakož i několik stran v č. 4/2017, a monotematické mimořádné číslo ročníku 2000 tvořil pro změnu zpěvník jeho písní. Aniž bych musel podnikat zevrubnou rešerši, jsem si jist, že vědecké tvorbě Zdeňka Fabiána se v IB tak velkého prostoru nikdy nedostalo.

Tento dluh, svůj vlastní i našeho časopisu, bych chtěl ne snad zapravit, to by tak snadno nešlo, ale aspoň trochu umenšit textem věnovaným Zdeňkově vědecké kariéře. Bude to ovšem text písňový. S texty odbornými má společné to, že vyžaduje poznámkový aparát. To je ve Zdeňkových očích, jak dobře vím, textařský smrtelný hřích; čemu není rozumět rovnou, nemá v písni co dělat. Tak honem do toho, ať mám tu ostudu za sebou.

Ve třetí sloce narazíme na jméno Kovanic, které zvláště mladším z nás dost možná už nic neříká. Ing. Pavel Kovanic, DrSc. se v 80. letech minulého století postaral o pořádný rozruch, když přišel s „úplně jinou statistikou“, tzv. gnostickou teorií dat (viz např. https://www.statspol.cz/robust/1984_kovani84.pdf). Statistická komunita na ni hleděla s velkou nedůvěrou, ale jen málokdo do této záhadné nauky hlouběji pronikl. Zdeněk Fabián patřil k nečetným výjimkám, nebyl-li vůbec výjimkou jedinou. Neznamená to, že by se stal vyznavačem gnostické teorie, nicméně časem mi svou zarputilou sna-

hou o originální „statistickou teorii všeho“ začal Ing. Kovanice připomínat. Jestli to nebude tím, že oba jsou původním zaměřením fyzici...

A pak je tu ještě hned v úvodní sloce ona blíže nedefinovaná „ta z Petřína“. Tady bych, pokud někomu zvědavost nedá a nechce se nechat jen unášet vlastní fantazií, doporučil pobídnout někdy Zdeňka, aby povyprávěl, s kým a jak se seznámil cestou z obhajoby disertace.

Rád bych ještě poznamenal, že co do faktografické přesnosti následujícího „průletu vědeckým životem Zdeňka Fabiána“ se hlásím ke škole Václava Hájka z Libočan. Ostatně, když Zdeněk věnoval mé maličkosti dva verše ve své písni, viz `http://uivty.cs.cas.cz/zdenek/ČEPLYN/COMPSTAT.wma`, pravda také neslavila kdovíjaký triumf: Že bych kdy na ranní přednášce na kongresu Compstat pojídal perník na konferenční tašce, mohu kategoricky popřít. Takže o co jde... Dost ale řečí. Kdo chce, ať spolu se mnou na počest našeho oslavence pozvedne číši (vím, patrně virtuálně, jak máme natrénováno) a dá se do zpěvu.

Píseň strašlivá k jubileu Zdeňka Fabiána[1]

Když dostal Fábo CSc., to bylo slávy,
divili se všichni chlapi, šílely baby.
Ta z Petřína vypadala, že mu hnedle dá,
ale lístek s jejím telefonem ještě dnes hledá.

Takže Fábo je CSc.,
bádat se mu ale nechce.
Když jsem zdolal tento milník,
budu pijan už jen a smilník.

Komise, jíž účty z vědy skládat se musí,
tázala se Tak, jaké máš, Fábo, opusy?
Vypadali, jako když je bolí osmička,
když děl Mrtvotechna, Pašerácká, Osička.

Zda je z toho vzala depka,
nebo rovnou klepla pepka,
to už dávno není známý,
jen ty písně tu jsou furt s námi.

[1] Na melodii Jackovy písně (Můj strýček Jack když narodil se, to bylo slávy, ...) Jiřího Voskovce, Jana Wericha a Jaroslava Ježka, viz `https://www.youtube.com/watch?v=_rIq2W6e6aE`.

Jak šel čas, na radovánky ubývalo sil,
tak že by se o bádání přec jen pokusil.
Ale Kolmogorov, R. A. Fisher, toť všechno na nic,
mám-li mít vzor, tak to bude jenom Kovanic.

Budu novým Kovanicem,
každý slovo do pranice.
Cesta, jíž budu sledovat,
ať je šílená, jen když nová.

Jak tak dumal, tak ho napad jedinečnej fór,
že na všechno se napasuje Fabiánův skór.
Ten skór všechno změří, zváží, přesně zhodnotí,
ale proč to, nač to, po tom, milý brachu, h[illegible]o ti.

Nacpu vás tím dolem, horem,
vše se bude řešit skórem
a kdo na svět jinak brejlí,
ten je vůl a trapně se mejlí.

Ve věku, kdy každý myslí hlavně na penzi,
on se začal ohlížet po nový dimenzi.
Vlít k šéfovi, You have to upgrade my salary!
Jsem světovej – objevil jsem právě skaláry (v akvárku).

Takže od nynějška sorry,
zapomeňte na vektory,
co bych se s tím, medle, páral,
každej skór je od dneška skalár.

Když završil osmdesátý života rok,
říkali mu Už toho nech, nebuď, Fábo, cvok!
Rád bych, jenže kdo by hlídal nový obzory?
Co když skóry zítřka budou fuzzy tenzory?

Tak takovej argument k věci
nevyvrátěj žádný kecy.
Tak, barde ty neratovický,
žij a bádej s námi navždycky!

DVĚ POZNÁMKY K HONZOVĚ PÍSNIČCE
TWO NOTES ON HONZA'S SONG

Zdeněk Fabián

Adresa: Ústav informatiky AV ČR, Pod Vodárenskou věží 271/2, 182 07 Praha 8

E-mail: zdenek@cs.cas.cz

Abstrakt: V článku popisuji cestu k nové metodě inference založené na zjištění, že ke každé spojité náhodné veličině lze přiřadit náhodnou veličinu s významem relativního vlivu pozorování na konstrukci typické hodnoty rozdělení. Odtud plyne nový popis standardních rozdělení, snadno zobecnitelný pro parametrická rozdělení a použitelný pro řešení statistických úloh.
Klíčová slova: Kovanic, skaláry, skalární skórová funkce.

Abstract: In the paper the way to a new paradigm in probabilistic and statistical reasoning is described. The new paradigma is based on the finding that a scalar-valued score random variable expressing relative influence of items generated from the distribution of X with respect to its typical value can be assigned to any continuous univariate random variable X. The approach leads to a new description of standard distributions. The methodology is generalized for parametric families and used for solutions of some estimation problems.
Keywords: Kovanic, scalars, scalar-valued score function.

.

1. Úvod

„Co teď čtete, pane Werichu?" „Teď, pane Horníčku, Švejka." „Poprvé?" „To ne, ale teď poprvé s vysvětlivkama ... Jsou delší než celá kniha."

Tolik nepřesná citace. Já mám k Honzově dílku pouze dvě připomínky: k jednomu pojmu mladším asi neznámému a k jednomu taky nepřesnému. Písnička je to ovšem pěkná.

2. Kovanic

Když jsem před mnoha lety zapustil kořeny v Ústavu Informatiky AV (jmenovalo se to tehdy Centrální výpočetní středisko ČSAV) se zadáním založit knihovnu programů pro analýzu signálů, řekla mi má nová šéfová OK (Olga

Kufudaki): „OK. Když umíš ty seismický vlny, vemu tě do Ústavu fyziologických regulací, tam mají časových řad to budeš koukat.“ A jednou později: „Deme se podívat na Kovanice. Skládá data podle teorie relativity.“

Kovanicova gnostická teorie [1], deklarovaná jako alternativa ke statistice, pracovala s kladnými daty pomocí podivuhodných fyzikálních postupů. Data se prezentovala svou „fidelity“ a „irrelevancí“. Tou dobou vrcholil boom robustních odhadů. Nové metody se testovaly na známých Stiglerových datech [2] z historických měření fyzikálních veličin. Kovanicovy „gnostické“ odhady se ukázaly být nejblíže v současnosti známým hodnotám.

Nevím, co si o tom mysleli statistici, ale na Kovanicovy přednášky a následné diskuse už nechodili. Já se zrovna zabýval tzv. „mrkací“ časovou řadou, registrovanou při relaxaci a při duševní činnosti. Programy si s ní neporadily, protože každý proband občas narušil svou mrkací frekvenci a chvilku zíral bez mrknutí. Naprogramoval jsem si adaptaci Kovanicových odhadů na časové řady a ejhle, spektra dostala smysl. Už ani nevím, zda v nich fyziologové něco objevili; já jsem objevil, že Kovanic nepodvádí a že na jeho odhadech asi něco je.

Tak jsem si četl Kovanicovy práce pořád dokola a pátral po něčem podobném v učebnicích statistiky. Časem jsem pochopil: Kovanic vkládá data do své irrelevance, funkce, kterou jsem ani v teorii pravděpodobnosti, ani v matematické statistice nenašel.

Po delším čase jsem publikoval článek [3] o tom, že Kovanicova „fidelity“ je pravděpodobnostní hustota loglogistického rozdělení a „irrelevance“ něco jako věrohodnostní funkce, příslušná ne parametru, ale přímo tomu rozdělení. Zhruba řečeno, Kovanicova teorie je v zásadě teorie jednoho, v tehdejší době asi nepříliš používaného, pravděpodobnostního rozdělení. Rozdělení, které je pro kladná data velmi šikovné. Zmíněná funkce je omezená (proto jsou odhady robustní) a velice nesymetrická (a proto jsou odpovídající odhady lepší než většina robustních odhadů založených na symetrických omezených funkcích). Těžko uvěřit, že tato funkce v teorii pravděpodobnosti i matematické statistice chyběla. Přesněji, v teorii pravděpodobnosti chyběla úplně, ve statistice je v některých případech identická s Fisherovou skórovou funkcí pro centrální parametr. Ta je však definována pouze pro parametrická rozdělení a rozpadá se při vektorovém parametru do několika větví (důsledkem čehož je třeba to, že ve výběrovém prostoru je obtížné definovat vzdálenost bez pomoci matic). Irrelevance fungovala i bez parametrů.

Ukázalo se, že Kovanicova volba výběrového prostoru na kladné části reálné osy je podstatná. Funkce, jejíž smysl je shodný s Kovanicovou irrelevancí, je pro rozdělení na celé reálné ose dávno známá jako skórová funkce. Známá, ale nepoužívaná ze zřejmého důvodu: dosadíme-li do jejího vzorce hustotu

rozdělení definovaného jen na části reálné osy, dostaneme cosi nesmyslného. Třeba pro rovnoměrné rozdělení nulu.

3. Skaláry

3.1. Skórová funkce

Teď už o statistice. Buď Y náhodná veličina se spojitým rozdělením G s nosičem $\mathcal{R}$ („na $\mathcal{R}$“), pro jednoduchost s unimodální spojitě diferencovatelnou hustotou $g(y)$. Její *skórová funkce* je

$$S_G(y) = -\frac{g'(y)}{g(y)}. \tag{1}$$

Pro rodinu rozdělení s parametrem polohy $G(y-\mu)$, $\mu \in \mathcal{R}$, je skórová funkce fundamentální funkcí klasické statistiky a Fisherovým skórem pro μ,

$$S_G(y-\mu) = \frac{\partial}{\partial \mu} \log g(y-\mu). \tag{2}$$

Odhad μ z dat $y_1, \ldots, y_n$, realizací náhodných veličin $(Y_1, \ldots, Y_n)$ s rozdělením $G(y-\mu)$ s neznámým μ z rovnice

$$\frac{1}{n} \sum_{i=1}^{n} S_G(y_i - \mu) = 0 \tag{3}$$

je eficientní, a tedy asymptoticky nejlepší. Pro $\mu = 0$ je (1) nepochybně funkcí charakterizující samotné rozdělení.

Co ale když má rozdělení těch parametrů víc?

R. A. Fisher před sto lety zobecnil rovnici (3) pro vektorový parametr $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ na soustavu m rovnic

$$\frac{1}{n} \sum_{i=1}^{n} U_j(y_i; \boldsymbol{\theta}) = 0, \qquad j = 1, \ldots, m, \tag{4}$$

kde funkce $U_j(y; \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} \log f(x; \boldsymbol{\theta})$ jsou dnes známé jako Fisherovy skóry. Řešením soustavy rovnic (4) získáme maximálně věrohodné (maximum likelihood: ML) odhady $(\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$, asymptoticky normální, eficientní, takže nejlepší, ovšem jen když data nejsou kontaminována.

Pro tento častý případ dosadil do (3) P. J. Huber, model nemodel, omezenou funkci (které se bohužel taky říká score function) a nějakých těch „ulítlých“ pozorovaní si robustní odhady ani nevšimnou.

Ale co když na svém modelu z nějakých důvodů trváme?

Napadlo mne se blíže podívat na skórovou funkci (řekněme v užším slova smyslu) (1). Ukázalo se, že má úžasné vlastnosti:

a/ Podle chování $S_G(y)$ v plus mínus nekonečnu je na $\mathcal{R}$ pouhých 6 typů rozdělení (viz Tabulka 1).

b/ Momenty

$$\mathrm{E}S_G^k(Y) = \int_{\mathcal{X}} S_G^k(y) f(y)\, \mathrm{d}y \tag{5}$$

jsou pro $k \in \mathcal{N}$ konečné (je-li S_G omezená, je to zřejmé, hustoty rozdělení s neomezenou skórovou funkcí rychle klesají k nule).

c/ Za typickou hodnotu lze vzít modus $y^* : S_G(y^*) = 0$.

d/ $\mathrm{E}S_G^2$ se v [4] považuje za Fisherovu informaci rozdělení G.

Tabulka 1: S_G: nejjednodušší skórové funkce typu N neomezená, O omezená, E exponenciální, P polynomiální, R redescentní. g je příslušná hustota z (1).

$S_G(y)$	Typ	$g(y)$	Rozdělení
$\frac{\mathrm{e}^y - \mathrm{e}^{-y}}{2}$	NE	$\frac{1}{2K_0(1)}\mathrm{e}^{-\cosh y}$	hyperbolic
y	NP	$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}y^2}$	normal
$\mathrm{e}^y - 1$	ON	$\mathrm{e}^y \mathrm{e}^{-\mathrm{e}^y}$	Gumbel
$1 - \mathrm{e}^{-y}$	NO	$\mathrm{e}^{-y}\mathrm{e}^{-\mathrm{e}^{-y}}$	extreme value
$\frac{\mathrm{e}^y - 1}{\mathrm{e}^y + 1}$	OO	$\frac{\mathrm{e}^y}{(1+\mathrm{e}^y)^2}$	logistic
$\frac{2y}{1+y^2}$	OR	$\frac{1}{\pi(1+y^2)}$	Cauchy

Kupodivu, není o ní v její čisté podobě (1) v teorii pravděpodobnosti a ve statistice ani vidu ani slechu. Důvod je jasný. Jak už bylo řečeno, pro rozdělení definované na otevřeném intervalu $\mathcal{X} \neq \mathcal{R}$ vzorec (1) generuje nesmysly.

Ale Kovanic na polopřímce nějakou skórovou funkci má. Jenže jeho postup pro jiná rozdělení nefunguje.

3.2. Skórová funkce na intervalu a polopřímce

Nakonec mi to došlo. Jak se dostaneme z prostoru A do prostoru B? V životě všelijak, v matematice transformací.

Buď $\mathcal{X} \subseteq \mathcal{R}$ otevřený interval a $\eta : \mathcal{X} \to \mathcal{R}$ rostoucí spojitá a diferencovatelná. Uvažujme náhodnou veličinu Y na $\mathcal{R}$ a říkejme jejímu rozdělení G s hustotou $g(y)$ prototyp. Transformovaná náhodná veličina

$$X = \eta^{-1}(Y) \tag{6}$$

má rozdělení $F(x) = G(\eta(x))$ s hustotu

$$f(x) = g(\eta(x))\eta'(x). \tag{7}$$

A *skórová funkce* na $\mathcal{R}^+$ bude, logicky, transformovanou skórovou funkcí prototypu [4]:

$$S_F(x) = S_G(\eta(x)). \tag{8}$$

Transformované logistické rozdělení pro $\eta(x) = \log x$ je loglogistické a jeho $S_F(x)$ je ta Kovanicova irrelevance [5].

Místo (8) lze psát

$$S_F(x) = -\frac{1}{f(x)} \frac{\mathrm{d}}{\mathrm{d}x} \left[\frac{1}{\eta'(x)} f(x) \right], \tag{9}$$

ježto podle (1) a (7)

$$S_F(x) = -\frac{1}{g(\eta(x))\eta'(x)} \frac{\mathrm{d}}{\mathrm{d}x} g(\eta(x)) = -\frac{g'(\eta(x))}{g(\eta(x)} = S_G(\eta(x))$$

a skórová funkce rozdělení na $\mathcal{X}$ nezávisí na prototypu.

Závisí však na té transformaci. Ale tu si nevolíme. Většinou je z (7) „vidět“ g nebo ten Jakobián. Pro $\eta : \mathcal{R} \to \mathcal{R}$ vždycky. Pro $\mathcal{X} \neq \mathcal{R}$ to může být vidět: rozdělení na $\mathcal{X} = (\pi/2, \pi/2)$ s hustotou

$$f(x) = \frac{1}{\sqrt{2\pi}\cos^2 x} \mathrm{e}^{-\frac{1}{2}\tan^2 x}$$

má jistě za prototyp Gaussovo rozdělení a $\eta(x) = \tan x$. „Vidět“ to ale být nemusí, a tehdy je nejlépe předpokládat, že

$$\eta(x) = \begin{cases} \log(x-a) & \text{pro} \quad \mathcal{X} = (a, \infty) \\ \log\left(\frac{x-a}{b-x}\right) & \text{pro} \quad \mathcal{X} = (a, b), \end{cases}$$

neb ji má většina běžně užívaných rozdělení. S_F je pak jednoznačná [6].

Typickou hodnotou x^* rozdělení F s unimodálním prototypem je x^* : $S_F(x^*) = 0$, což je podle (6) projekce modu prototypu: $x^* = \eta^{-1}(y^*)$. Není to obecně ani střední hodnota, ani modus, ani medián. Se střední hodnotou je identická pro F s lineární skórovou funkcí, což je rozdělení Gaussovo ($S_F(x) = x$), exponenciální ($S_F(x) = x - 1$) a rovnoměrné ($S_F(x) = 2x - 1$). Vlastnosti x^* (kterému říkám *score mean*) viz [7].

Skórové momenty F jsou podle (5) a (8)

$$\mathrm{E}S_F^k = \int_{\mathcal{X}} S_G^k(\eta(x)) g(\eta(x)\, \mathrm{d}x = \mathrm{E}S_G^k. \tag{10}$$

Speciálně, $\mathrm{E}S_F = 0$. Za míru variability rozdělení F lze považovat *skórový rozptyl*

$$\omega_F^2 \equiv \mathrm{Var}_S X = \frac{\mathrm{E}S_F^2}{[S_F'(x^*)]^2}. \tag{11}$$

Důvody pro tuto volbu viz [8].

Funkci

$$w_F(x) = S_F'(x) = \frac{\mathrm{d}S_F(x)}{\mathrm{d}x} \tag{12}$$

lze interpretovat [7] jako váhovou funkci rozdělení F. Vzdálenost bodů $x_1, x_2 \in \mathcal{X}$ lze rozumně definovat jako $d_F(x_1, x_2) = |\delta_F(x_1, x_2)|$ kde

$$\delta_F(x_1, x_2) = \frac{S_F(x_2) - S_F(x_1)}{S_F'(x^*)} = \frac{1}{w_F(x^*)} \int_{x_1}^{x_2} w_F(x) f(x)\, \mathrm{d}x. \tag{13}$$

Tím je jednoznačně definována metrika (či pseudometrika) ve výběrovém prostoru, indukovaná pravděpodobnostní mírou reprezentovanou distribuční funkcí F, viz [8].

V Tabulce 2 uvádím hustoty, skórové a váhové funkce na $\mathcal{R}^+$ transformovaných prototypů z Tabulky 1 při $\eta(x) = \log x$ (bez log-Cauchy, to má příliš těžký chvost). Jsou stejného typu jako jejich prototypy.

3.3. Centrální limitní věta pro skórové náhodné veličiny

Jsou-li $X_1, \ldots, X_n$ nezávislé náhodné veličiny na $\mathcal{X}$ s rozdělením F a skórovou funkcí $S_F(x)$, jsou nezávislé i náhodné veličiny $S_F(X_1), \ldots, S_F(X_n)$. Buď

$$\overline{S}_F = \frac{1}{n} \sum_{i=1}^{n} S_F(X_i).$$

Věta. Pro $n \to \infty$ je $\sqrt{n}\, \overline{S}_F \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{E}S_F^2)$.

Tabulka 2: Charakteristiky transformovaných standardních rozdělení (GIG je Generalized Inverse Gaussian).

Typ	Rozdělení	$f(x)$	$T_F(x)$	$w_F(x)$
NE	GIG	$\frac{1}{2K_0(1)x}\mathrm{e}^{-\frac{1}{2}(x+1/x)}$	$\frac{1}{2}(x-1/x)$	$\frac{1}{2}(1+1/x^2)$
NP	lognormal	$\frac{1}{\sqrt{2\pi}x}\mathrm{e}^{-\frac{1}{2}\log^2 x}$	$\log x$	$1/x$
ON	exponential	e^{-x}	$x-1$	1
ON	Fréchet	$\frac{1}{x^2}\mathrm{e}^{-1/x}$	$1-1/x$	$1/x^2$
OO	loglogistic	$\frac{1}{(x+1)^2}$	$\frac{x-1}{x+1}$	$\frac{2}{(x+1)^2}$

Důkaz. Ježto $\mathrm{E}S_F = 0$ a $\mathrm{E}S_F^2$ je konečná, věta platí podle Lindebergovy-Lévyho centrální limitní věty.

Když se tedy „pracuje“ se skórovými náhodnými veličinami, není nutno (teoreticky) rozlišovat rozdělení „obyčejná“ a s těžkými chvosty (heavy-tailed).

3.4. Skalární skórová funkce parametrických rozdělení

Vraťme se k funkci (9). Je asi už patrné, jak je možné ji jednoduše zobecnit pro F s vektorovým parametrem. Takto:

$$S_F(x;\boldsymbol{\theta}) = -\frac{1}{f(x;\boldsymbol{\theta})}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1}{\eta'(x)}f(x;\boldsymbol{\theta})\right]. \tag{14}$$

(Pro parametrická rozdělení je ve vzorci přídavný konstantní člen, aby $\mathrm{E}S_F^2$ byla Fisherova informace, viz [6, 7]).

Ač je $\boldsymbol{\theta}$ vektor, skórová funkce (funkce, Honzo) je skalární („scalar-valued“). A nejen to: v (14) se derivuje pouze podle proměnné, takže všechny výše uvedené vzorce platí i pro parametrická rozdělení. Několik takových s těžkými chvosty uvádím v Tabulce 3.

Příklad 1. Pro $c \leq 2$ neexistuje konečný rozptyl Fréchetova rozdělení, skórový rozptyl je $\omega_F^2 = \tau^2/c^2$. Na obrázku 1 nahoře MATLAB[1] nakreslil tři hustoty evidentně blízkých rozdělení s velice odlišnými rozptyly, skórové rozptyly jsou postupně 0,17, 0,21 a 0,25. Dole máme závislost obou rozptylů na c: ω_F^2 se blíží k nekonečnu až u hranice definičního intervalu.

[1] Poznámka redakce. Grafy byly překresleny s pomocí nástrojů Lua, ULua (Universal Lua Distribution, viz `https://ulua.io`) a TEXového balíčku `pgfplots`.

Tabulka 3: Hustoty, skórové funkce, typické hodnoty a skórové rozptyly několika rozdělení s těžkými chvosty na $\mathcal{R}^+$.

F	$f(x)$	$S_F(x)$	x^*	ω_F^2
Fréchet	$\frac{c}{x}(\tau/x)^c \mathrm{e}^{-(\tau/x)^c}$	$\frac{c}{\tau}[1-(\tau/x)^c]$	τ	τ^2/c^2
loglogistic	$\frac{c}{x}\frac{(x/\tau)^c}{[(x/\tau)^c+1]^2}$	$\frac{c}{\tau}\frac{(x/\tau)^c-1}{(x/\tau)^c+1}$	τ	$\frac{4}{3}\tau^2/c^2$
beta-prime	$\frac{1}{B(p,q)}\frac{x^{p-1}}{(x+1)^{p+q}}$	$\frac{q}{p}\frac{qx-p}{x+1}$	$\frac{p}{q}$	$\frac{p(p+q)^2}{q^3(p+q+1)}$

Příklad 2. Jak už bylo (mírně nepřesně) řečeno, Kovanicova „fidelity“ je úměrná hustotě a „irrelevance“ skórové funkci loglogistického rozdělení. Typická hodnota (score mean) je Kovanicova ideální hodnota, míra variability, zde $\omega_F^2 = \frac{4}{3}\tau^2/c^2$, je v gnostické teorii jiná.

Poznamenejme, že Fréchetovo i loglogistické rozdělení mají strukturní parametr $\tau = \exp(\mu)$, který je projekcí modu μ prototypu. Parametr c se transformuje podle

$$\frac{y-\mu}{\sigma} \to \frac{\log x - \log \tau}{\sigma} = \log\left(\frac{x}{\tau}\right)^c.$$

Je patrné, že $1/c$ rozdělení na $\mathcal{R}^+$ lze považovat za parametr měřítka.

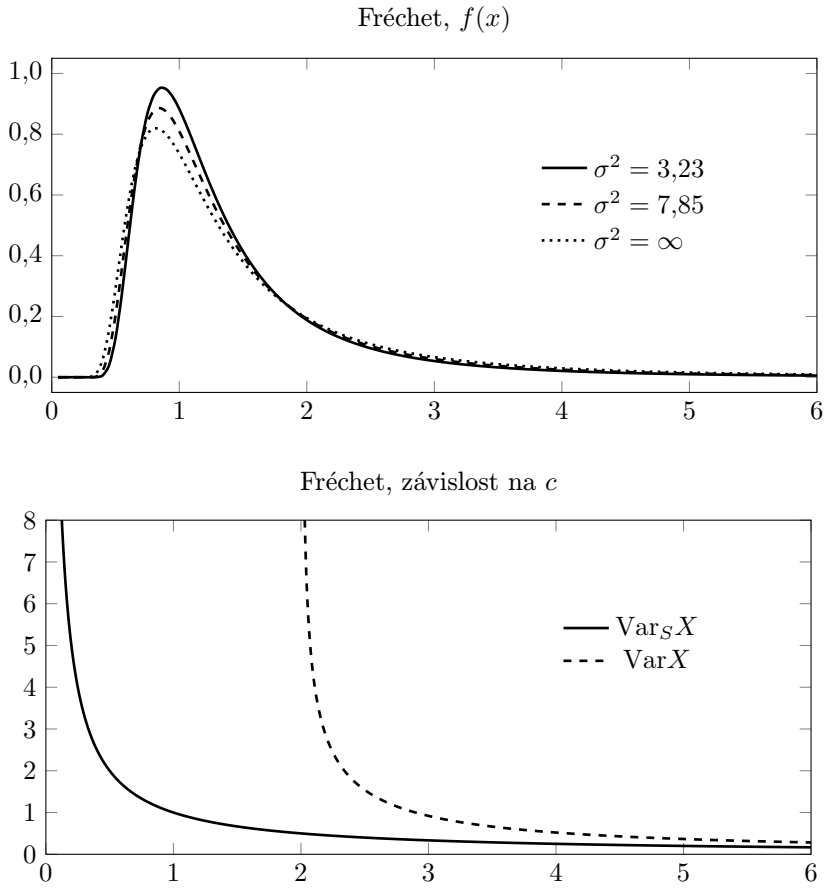
Příklad 3. Beta-prime $BP(p,q)$ na $\mathcal{R}^+$ je příkladem rozdělení druhého typu – těch, která mají typickou hodnotu vyjádřenou funkcí (často poměrem) parametrů. Hustoty a skórové funkce $BP(2p,p)$ viz obr. 2. Skórové funkce jsou omezené tak jako i u jiných rozdělení s těžkými chvosty, což je z hlediska odhadů slibné.

3.5. Odhady

Jsou-li data $(x_1, \dots x_n)$ z rozdělení $F(x;\boldsymbol{\theta_0})$, je možné $\boldsymbol{\theta}_0 \in \Theta^m$ odhadovat pomocí $S_F(x;\boldsymbol{\theta})$, i když je tato scalar-valued, a to zobecněnou momentovou (SM) metodou [6] z rovnic

$$\frac{1}{n}\sum_{i=1}^{n} S_F^k(X_i;\boldsymbol{\theta}) = \mathrm{E}S_F^k(\boldsymbol{\theta}), \qquad k = 1, \dots, m. \tag{15}$$

Momenty $\mathrm{E}S_F^k$ jsou podle (5) a (10) konečné. SM odhady sice nejsou většinou eficientní, zato však jsou pro rozdělení s těžkými chvosty robustní pro

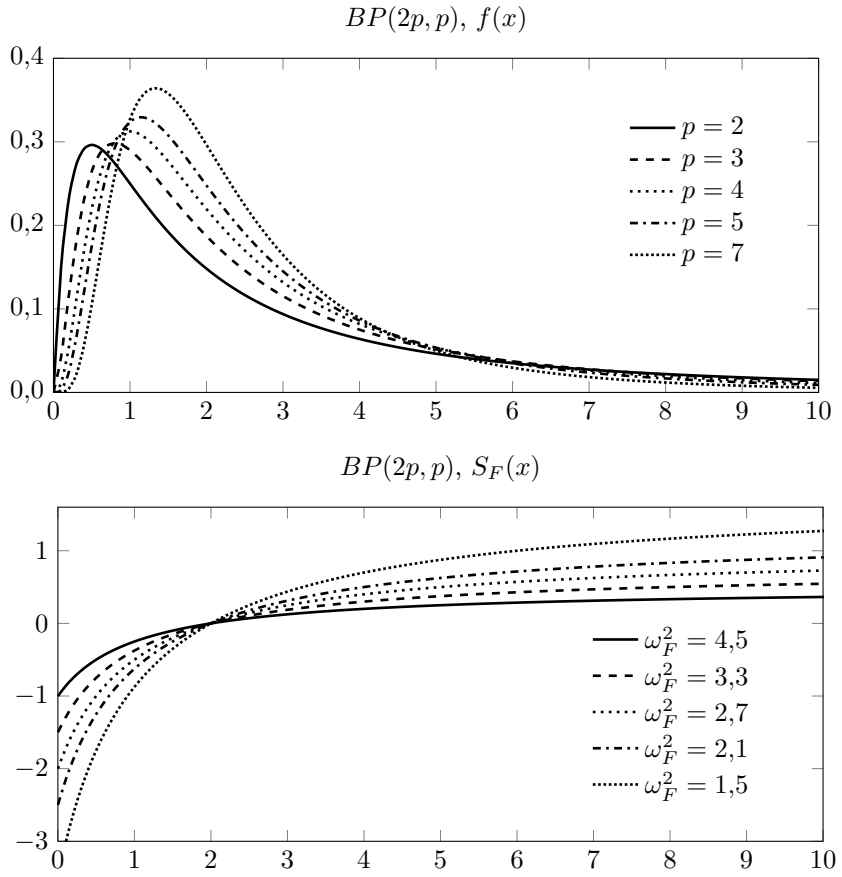


Obrázek 1: Podobné hustoty Fréchetových rozdělení s velmi různými rozptyly. Spodní obrázek závislost $\mathrm{Var}X$ a $\mathrm{Var}_S X$ na rostoucím c.

všechny komponenty vektoru $\boldsymbol{\theta}$. A neomezenou skórovou funkci lze v (15) poměrně snadno huberizovat, což je v ML metodě obtížné až nemožné (spočítat momenty useknuté funkce ale taky není úplně triviální).

Pro rozdělení $F(x; x^*)$ s explicitně vyjádřeným x^* najdeme jeho odhad z první z rovnic (15)

$$\frac{1}{n}\sum_{i=1}^{n} S_F(x_i; \widehat{x}^*) = 0. \tag{16}$$



Obrázek 2: Hustoty a skórové funkce rozdělení $BP(2p, p)$.

Je-li x^* strukturním parametrem, je odhad ML odhadem [6]. Pro beta-prime a loglogistické rozdělení má rovnice (16) tvar

$$\sum_{i=1}^{n} \frac{x_i - x^*_{BP}}{x_i + 1} = 0, \qquad \sum_{i=1}^{n} \frac{x_i - x^*_{LL}}{x_i + x^*_{LL}} = 0.$$

Pro beta-prime je ten odhad de facto zobecněným ML odhadem typické hodnoty $x^* = p/q$, „Kovanicovo ML“ je třeba spočíst iteračně.

Výběrová typická hodnota rozdělení s lineární scalar-valued skórovou funkcí (normální, gamma, beta) je aritmetický průměr, lognormálního rozdělení geometrický průměr a Fréchetova či Paretova rozdělení harmonický průměr [7]. Naopak, zvolený typ průměru vlastně implikuje předpokládané rozdělení datového souboru.

4. Závěrem

Každá spojitá náhodná veličina X má kromě hustoty $f(x)$, popisující relativní pravděpodobnost x vzhledem k ostatním hodnotám, také skórovou (vlivovou) funkci popisující relativní vliv pozorované hodnoty x na konstrukci typické hodnoty F, váhovou funkci vyjadřující relativní váhu x, vzdálenost bodů ve výběrovém prostoru generovanou danou pravděpodobnostní mírou, a nové numerické charakteristiky: místo střední hodnoty a rozptylu, které nemusí v případech rozdělení s těžkými chvosty existovat, máme konečnou typickou hodnotu a konečný skórový rozptyl.

Nabízí se samozřejmě využití skalární skórové funkce v řadě dalších úloh, viz třeba [9]. Skórová korelace a lineární regrese [10, 11] nabízejí alternativu k metodám robustní statistiky. Zdá se, že jsou dobře použitelné jak pro data z rozdělení s těžkými chvosty, tak (zejména) pro data z rozdělení velmi nesymetrických a špičatých. Zatímco skórová vzdálenost typické hodnoty a jejího odhadu je pro rozdělení s jedním strukturním parametrem Raova vzdálenost [8], vzájemná vzdálenost rozdělení na základě čtverce rozdílu skórových funkcí [12] je v rodině divergencí novinkou. Praktické využití má zatím jen robustní verze Hillova estimátoru extremal value indexu [13, 14].

Teorie založená na skalar-valued funkcích je, Honzo, možný alternativní postup opomenutý klasickou statistikou, patrně z důvodu ohromujícího úspěchu metody maximum-likelihood. Zda bude užitečná v praxi, se teprve ukáže.

A dík za písničku, která mě mimo jiné i donutila k sepsání článku v rodném jazyce.

Reference

[1] Kovanic, P. (1986). A new theoretical and algorithmical tool for estimation, identification and control. *Automatica* **22**, 657–674. *cit. 7*

[2] Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, **5**, 1055–1098. *cit. 7*

[3] Fabián, Z. (1987). On the relation between gnostical and probability theory. *Kybernetika*, **33**, 259–270. *cit. 7*

[4] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, second edition, Wiley, New York. *cit. 9, 10*

[5] Fabián, Z. (2001). Induced cores and their use in robust parametric estimation. *Communs. Stat. Theory Methods*, **30**, 537–556. *cit. 10*

[6] Fabián, Z. (2016). Score function of distribution and revival of the moment method. *Communs. Stat. Theory Methods*, **45**, 1118–1136. *cit. 10, 12, 13 a 15*

[7] Fabián, Z. (2021). Mean, mode or median? The score mean. *Communs. Stat. Theory Methods*, **50**, 2360–2370. *cit. 11, 12 a 16*

[8] Fabián, Z. (2022). A measure of variability within parametric families of continuous distributions. *Communs. Stat. Theory Methods*, `https://doi.org/10.1080/03610926.2022.2155792`. *cit. 11, 16*

[9] Fabián, Z. (2011). A New Statistical Tool: Scalar Score Function. *Computer Technology and Application*, **2**, 109–116. ISSN 1934-7332. *cit. 16*

[10] Fabián, Z. (2010). Score correlation. *Neural Network World*, **20**, 793–798. *cit. 16*

[11] Stehlík, M. et al. (2019). On ecological aspects of dynamics for zero slope regression for water pollution in Chile. *Stochastic Analysis and Applications,* **37**, 574–601. *cit. 16*

[12] Fabián, Z., Vajda, I. (2003). Core functions and core divergences of regular distributions. *Kybernetika*, **39**, 29–42. *cit. 16*

[13] Fabián, Z., Stehlík, M. (2009). On robust and distribution sensitive Hill-like method. *Tech. Rep. of IFAS Research Paper Series*, **43**, 4–14. *cit. 16*

[14] Jordanova, P. et al. (2016). Weak properties and robustness of t-Hill estimators. *Extremes*, **19**, 591–626. *cit. 16*

STATISTICKÉ DNY V ROCE 2023
STATISTICAL DAYS IN 2023

Redakce časopisu

Česká statistická společnost pokračuje i v roce 2023 v tradici pořádání Statistických dnů, jejichž hlavním cílem je potkat se a ukázat si, co je v naší práci nového.

Kdy: 19. – 21. května 2023.

Kde: Na břehu Tiché Orlice, Penzion Mítkov, `www.penzionmitkov.cz`.

Zaměření konference

Chystáme čtyři tematické bloky, dva úzce specifikované věnované aktuálním tématům a dva pojaté velmi široce, aby prostor dostala i další zajímavá témata.

1/ Covidové ozvěny. Epidemiologie, imunologie a statistika – právě tyto disciplíny se dostaly v roce 2020 do popředí zájmu široké veřejnosti. Jak jsme my statistici v době covidové obstáli a jaká poučení si z ní odnášíme? Se svými příspěvky vystoupí zajímaví hosté, následovat bude moderovaná diskuse.

2/ Socioekonomické ozvěny covidu. Jak se protipandemická opatření promítla do vývoje statistických ukazatelů zachycujících klíčové ekonomické a sociální jevy? Jak se změnila struktura socioekonomických jevů a procesů, co se stalo s dlouhodobými vazbami mezi ukazateli? Jaké jsou možnosti a limity ekonomického modelování v podobných situacích?

3/ Statistické metody. Toto téma nabízí účastníkům se zájmem o rozvoj metodologie statistiky a analýzy dat, včetně statistiky matematické, prostor k prezentaci výsledků svého výzkumu.

4/ Aplikace statistiky. V roce 2022 byly na Statistických dnech prezentovány výsledky analýz z oblasti letectví, finančnictví (vývoj hodnoty kryptoměn), textilních materiálů či zpracování komunálního odpadu. Organizátoři jsou přesvědčeni, že i letos budou účastníci prezentovat neméně zajímavé problémy.

Organizační výbor: Jaromír Antoch, Jakub Fischer, Martina Litschmannová, Tomáš Löster, Ondřej Vencálek a Adéla Kondé.

Těšíme se na viděnou!