JAK JÁ BYCH VYKLÁDAL METODU „BARONA PRÁŠILA“ ZNÁMOU V STATISTICKÉ LITERATUŘE JAKO BOOTSTRAP

NOTES ON THE BOOTSTRAP METHOD

Josef Machek

Milé dámy a pánové[1], vážené čtenářky a vážení čtenáři našeho Bulletinu,

dovolte mi předem, abych převzal na svá bedra zodpovědnost za to, co je Vám zde předkládáno. Autor za nic nemůže, bohužel již s námi delší dobu není. A jak jsem jej znal, jistě by říkal, že to za plýtvání papírem nestojí. Ale já si myslím, že stojí. Dále se omlouvám za to, že nosím dříví do lesa. Dlouhé diskuse s profesorem Lvem Klebanovem, které jsme v poslední době vedli, mne ale přiměly opět přemýšlet o tak zvaných „resampling metodách“ všeho druhu. Při listování starými poznámkami na mne přitom vypadl útlý balíček dopisů, který jsem si založil do Efronovy knihy [4]. Když jsem si je opakovaně přečetl a připomněl si diskuse, které jsme s Josefem na toto téma vedli, přišlo mi, že by kus této korespondence mohl být zajímavý i pro některé z Vás.

Jistě se ptáte, jak a proč tyto dopisy vznikly. Bylo to takto: Začátkem devadesátých let přišla Marie Hušková s nápadem použít techniku bootstrap v oblasti detekce změn statistických modelů (*change-pointu*). Článek byl napsán, metodika byla vyzkoušena na datech reálných i simulovaných, a posléze i publikován, viz [2][2].

Přes zajímavé výsledky jsem pojal vůči bootstrapu řadu podezření, s nimiž jsem otravoval nejenom spoluautory, ale také lidi z katedry. Jak už to bývá, každý má řadu svých úkolů a zájmů, asymptoticky to krásně fungovalo, dalo se o tom publikovat, takže se moje stížnosti moc „neuchytily“. Kupodivu jedním z mála, kdo si na diskuse se mnou našel nezvykle mnoho času, byl Josef Machek. Jeho první závěry po návratu z prázdnin, jež jako každoročně rozdělil mezi Kubu a Bílou Třemešnou, shrnuje začátek dopisu datovaném 11. září 1996.

[1] Jak autor rád říkával. Dnes je to bohužel nemoderní, a soudě podle médií, se bojím, že se to pokusí zakázat i u nás, srovnejte [9], [10], [11] či [12].

[2] Přestože se náš článek neobjevil v žádném „hvězdném“ časopise, kam samozřejmě patřil, ale my jsme nepatřili ke správnému klanu, a nedostal se ani do WOSu, kde časopis sice je, ale náš článek nikoliv, čtenáři si jej všimli a patří mezi tři nejcitovanější (sobecky myslím, že zaslouženě) práce, které jsme díky Marii Huškové v oblasti change-pointu publikovali.

Buď zdráv, Járo,

listoval jsem si trochu v literatuře, kterou jsi mi o „bootstrapu“ půjčil, a dospěl jsem k následujícímu – možná ukvapenému – závěru:
Závěr: Nestojí to za ty řeči, které se o tom vedou. Je to metoda z jedné strany jasná, z druhé strany pochybná.

Důvody: Je potřeba začít trochou historie… Howgh! *Tvůj Josef*

V Josefovi ale problém doutnal, a tak následovaly další diskuse, simulační pokusy, a také další dopisy, jimiž Vás nechci obtěžovat. Místo nich jsem vybral až dopis poslední, který mi Josef doručil na konci zimního semestru 1996/1997.

Ahoj Járo,

hned úvodem se omlouvám kolegům:

- B. Efronovi a M. Huškové za to, že jsem považoval *Bootstrap* za zbytečný, nevhodný, přímo za kovopreclíkářství[3] (jak říkáme my, co spolu mluvíme[4]).

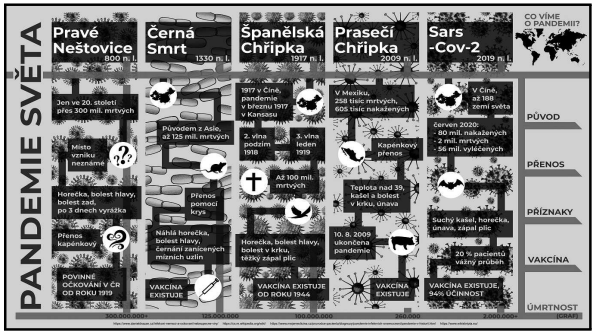
- J. Antochovi i jiným na katedře za to, že jsem jim odmítavý názor na bootstrap našeptával nebo, pokud ho sami dříve zastávali, je v něm utvrzoval.

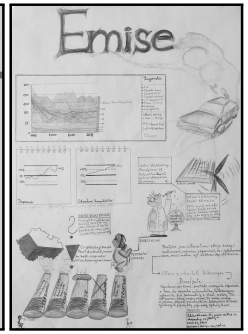
Uznávám po promyšlení celé věci, že metoda Barona Prášila[5] má jakési „zdravé jádro“. Mám však obavu, že jsou vážná omezení její praktické použitelnosti, a že je často velmi pochybným způsobem chápána a aplikována. *Proč mi pan Efron hned na začátku nevyložil svou metodu tak, jak se ji pokusím naznačit v následujících řádkách? Bude to možná jenom fikce, hypotetická povídka o tom, jak se mohlo k metodě bootstrap dospět. Ale nedovedu si to jinak představit a myslím, že ta povídka pomůže pochopit, kdy je bootstrap na místě a kdy je k ničemu.*

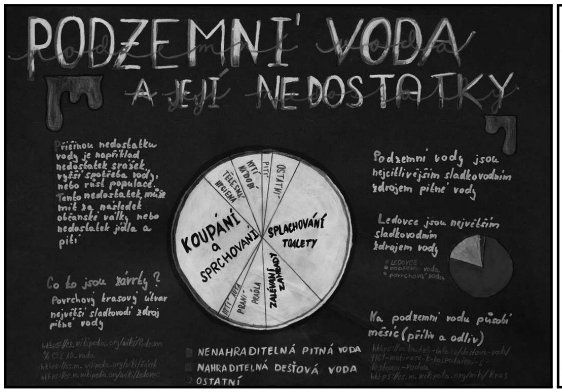
[3] O historii kovopreclíkářství se lze poučit v Bulletinu číslo 2 z roku 1999, viz [1].

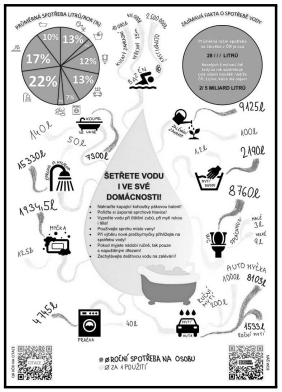
[4] Karel Poláček, Bylo nás pět. František Borový, Praha, 1946 (první knižní vydání).

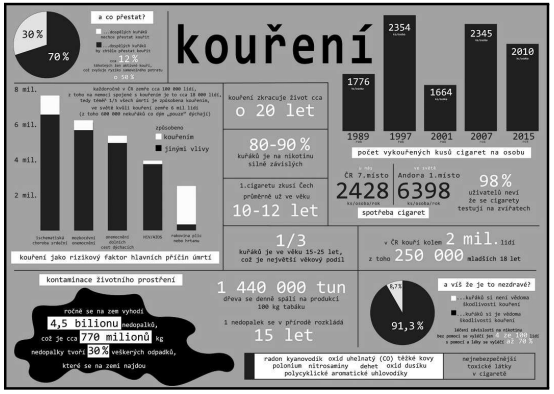
[5] Podivuhodné cesty po vodě i souši, polní tažení a veselá dobrodružství Barona Prášila, jak je vypravuje při víně v kruhu přátel (1786, Wunderbare Reisen zu Wasser und zu Lande, Feldzüge und lustige Abenteuer des Freiherrn von Münchhausen: wie er dieselben bei der Flasche im Zirkel seiner Freunde selbst zu erzählen pflegt), česky zkráceně Baron Prášil, je kniha německého preromantického spisovatele Gottfrieda Augusta Bürgera, ve které její hrdina vypráví o svých neuvěřitelných až fantastických příhodách. První české vydání Barona Prášila: Přijhody pana Žamputáře, Martin Neureuter, Praha 1824, přeložil Jan Josef Charvát. Mladší bych rád upozornil i na pozoruhodný Zemanův film z roku 1961.

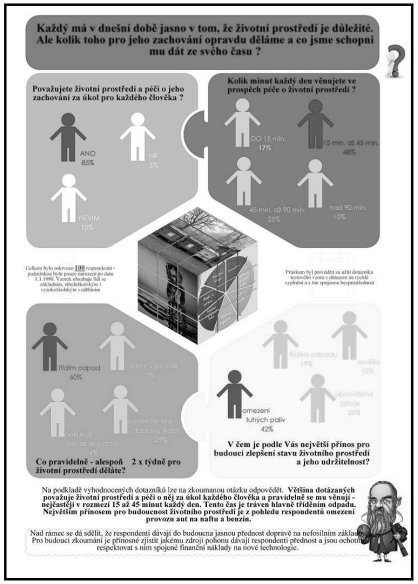


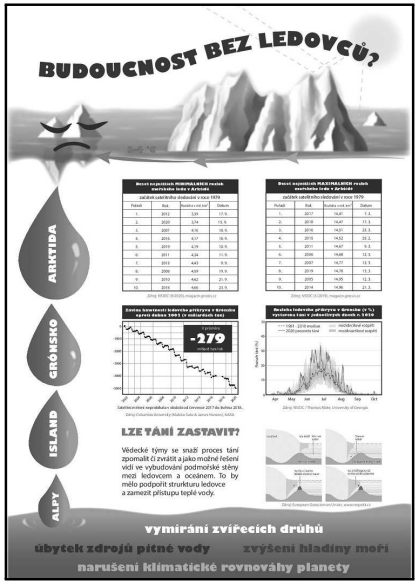


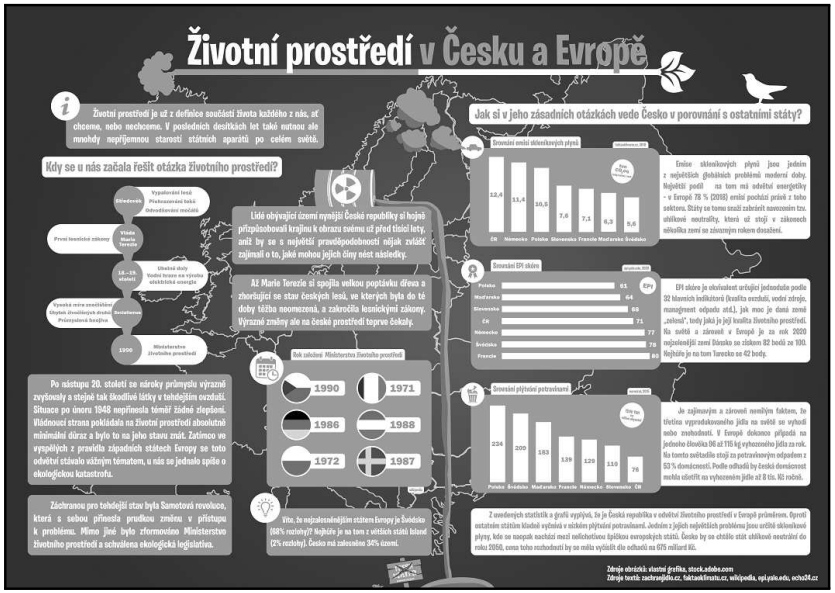












1) Všichni víme, že ke konstrukci intervalů spolehlivosti pro parametry nějakého statistického modelu a pro testování hypotéz o těchto parametrech používáme distribuční funkce statistik, jejichž pomocí parametry odhadujeme. Všichni víme, že aritmetický průměr n nezávislých pozorování z $N(\mu, \sigma^2)$ má rozdělení $N(\mu, \sigma^2/n)$, že statistika $\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ má rozdělení jako $\sigma^2 \chi^2$, kde χ^2 má rozdělení chí-kvadrát s $n-1$ stupni volnosti, atd. Podobně je známo, že pro vzájemně nezávislá pozorování $X_1, \ldots, X_n$ veličin s exponenciální hustotou $f(x;\theta) = \theta \mathrm{e}^{-\theta x},\ x > 0,\ \theta > 0$ má statistika $T = \sum_i X_i$ Erlangovo rozdělení s hustotou $g(t) = \theta^n t^{n-1} \mathrm{e}^{-\theta t}/(n-1)!,\ t > 0,\ \theta > 0$, atd.

Všimněme si, že ve všech těchto případech, a mnoha jim podobných, je důležitá následující skutečnost. *Dovedeme odvodit explicitní výraz pro distribuční funkci příslušné statistiky a můžeme danou distribuční funkci, i když je někdy velmi složitá, dána součtem řady nebo integrálem, který nelze vyjádřit pomocí elementárních funkcí, při vhodném normování numericky vypočítat.*

A všichni máme, nebo bychom alespoň měli mít, někde v záloze či zásuvce tabulky kvantilů těchto rozdělení. Mladá generace je možná již střelila do antikvariátu nebo dokonce dala do sběru, a má je místo toho někde na disketě nebo v počítači.

2) Jak čas plynul, lidé se zajímali o různé problémy. V druhé polovině tohoto století pánové Shapiro a Wilk navrhli hezký a velice účinný test normality. Tento test je založen na podílu, v jehož čitateli je odhad σ založený na součtu čtverců odchylek pořádkových statistik $X_{(i)}$ od přímky $X_{(i)} = \mu + \sigma U_{(i)}$, kde $U_{(i)}$ je pořádková statistika výběru z $N(0,1)$, a ve jmenovateli je obvyklý odhad směrodatné odchylky $N(\mu, \sigma^2)$. O odvození analytického tvaru rozdělení takovéto statistiky nemůže být ani řeč. A i kdyby se analytické vyjádření našlo, bylo by prakticky k ničemu pro svou obrovskou složitost. Ale to už jsme žili v době počítačové (a počítače asi neměly dost jiné práce, takže byly propůjčovány statistikům pro různé teoretické studie) a pánové Shapiro a Wilk získali tabulky kritických hodnot (kvantilů tohoto rozdělení) metodou Monte Carlo. A uveřejnili potřebné tabulky, které se používají dodnes.

Podobných příkladů, kdy výběrové rozdělení je nalezeno metodou Monte Carlo, je mnoho. Maně mne napadá rozdělení rozdělení statistiky průměr/rozptyl pro rozdělení exponenciality, rozdělení odhadu entropie pro test normality, apod. Ty si jistě vzpomeneš na řadu dalších.

3) V šedesátých a sedmdesátých letech minulého století se hodně pracovalo v teorii spolehlivosti. Do módy přišlo jako model pro dobu do poruchy Weibullovo rozdělení. Maximálně věrohodné odhady tohoto rozdělení vůbec nelze vyjádřit vzorcem, a je nutno numericky řešit soustavu věrohodnostních rov-

nic. Ale k testování hypotéz o parametrech nebo ke stanovení intervalů spolehlivosti či přesnosti odhadů potřebujeme pokud možno výběrové rozdělení těchto odhadů, které zde bohužel neznáme.

Autoři [8] ale ukázali, že dvě funkce maximálně věrohodného odhadu a skutečné hodnoty mají rozdělení nezávislé na skutečných hodnotách. A lze jich použít jako tak zvaných pivotálních veličin jak ke konstrukci intervalových odhadů, tak k testování hypotéz.

Jako příklad uveďme situaci, kdy $X_1, \dots, X_n$ jsou vzájemně nezávislé náhodné veličiny s hustotou

$$f(x; m) = m x^{m-1} \mathrm{e}^{-x^m}, \qquad x > 0,\ m > 0, \tag{1}$$

kde m je neznámý kladný parametr. Každý jistě poznal, že jde o náhodný výběr z Weibullova rozdělení, ve kterém je pro jednoduchost příkladu zvolen pevný „škálový“ parametr roven jedné[6]. Maximálně věrohodný odhad $\widehat{m}$ parametru m je řešením rovnice

$$\frac{1}{n}\sum_{i=1}^{n} \log x_i = \frac{1}{n}\sum_{i=1}^{n} x_i^{\widehat{m}} \log x_i - \frac{1}{\widehat{m}}. \tag{2}$$

Tuto rovnici je nutno řešit numericky a je zřejmé, že jestliže není znám žádný explicitní výraz pro $\widehat{m}$, tím méně bude možné odvodit distribuční funkci statistiky $\widehat{m}$.

Jen při vysokých hodnotách n, to jest velkých rozsazích výběru, bude možno použít vět o asymptotickém rozdělení maximálně věrohodných odhadů. *Co to je ale „vysoká hodnota“ n?* Je to deset, sto nebo tisíc? A co dělat při „nízkých hodnotách“ n? Připomeňme, že v aplikacích tohoto modelu v oblasti spolehlivosti seženeme málokdy více než jednu či dvě desítky experimentů.

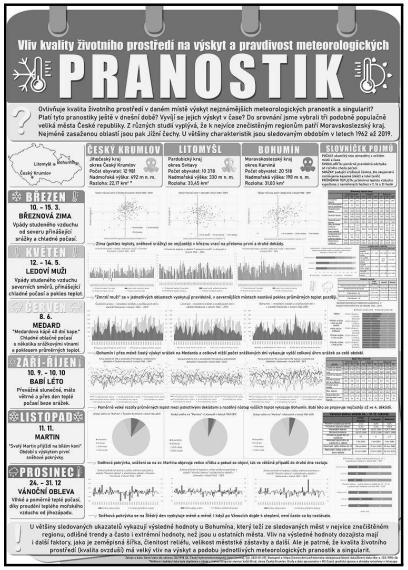
Na tuto otázku se našla překvapivě jednoduchá odpověď. Jednoduchými algebraickými úpravami zjistíme toto: Je-li $\widehat{m}$ řešením rovnice (2), pak $\widehat{l} = \widehat{m}/m$ je řešením rovnice

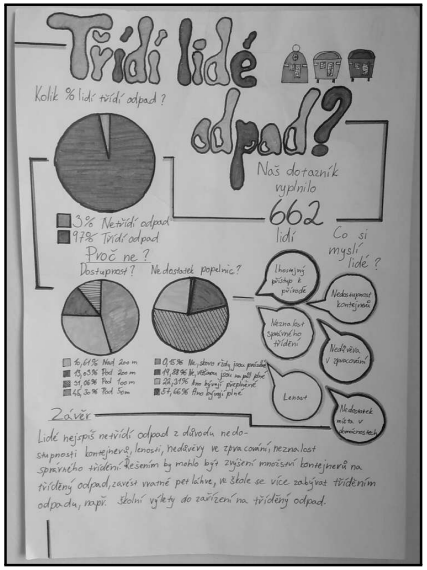
$$\frac{1}{n}\sum_{i=1}^{n} \log x_i^m = \frac{1}{n}\sum_{i=1}^{n} \left(x_i^m\right)^{\widehat{l}} \log x_i^m - \frac{1}{\widehat{l}}. \tag{3}$$

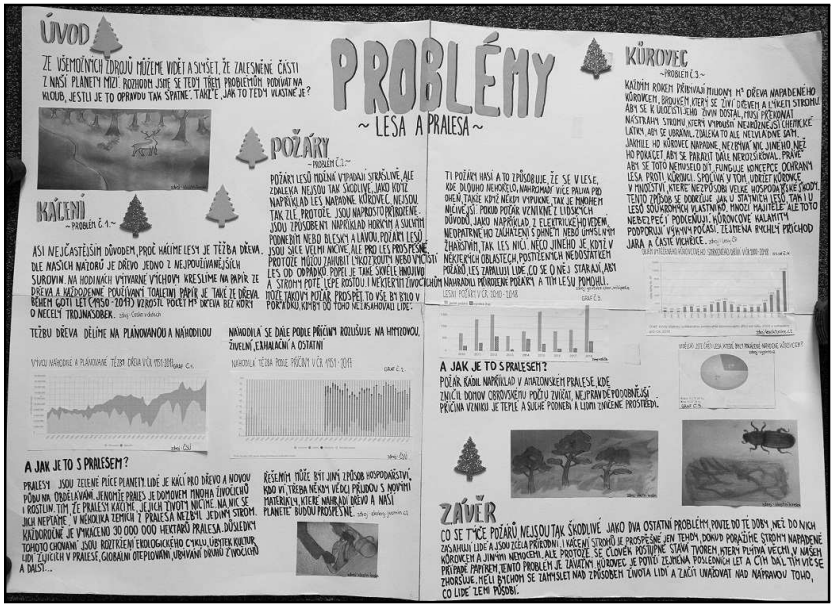
A teď to hlavní: Jestliže X má hustotu (1) s jakýmkoliv $m > 0$, pak $Y = X^m$ má hustotu

$$f(y) = \mathrm{e}^{-y}, \quad y > 0, \tag{4}$$

[6] V [8] čtenář, kterého myšlenka zaujala a nechce se mu ji znovu vymýšlet, najde řešení pro obecné Weibullovo rozdělení.







učiliště Na Veselí 51 v Praze 4. Druhé místo získal plakát Budoucnost bez ledovců? vypracovaný žáky SŠ Brno, Charlubova a třetí příčku obsadil plakát Životní prostředí v Česku a Evropě zaslaný opět Gymnáziem Dr. Emila Holuba. V kategorii vysokých škol porota vítěze nevybrala.

Plakáty posuzovala odborná komise ve složení:

- prof. Ing. Hana Řezanková, CSc., předsedkyně poroty (VŠE v Praze, Fakulta informatiky a statistiky, katedra statistiky a pravděpodobnosti),
- Ing. Petra Kuncová (Český statistický úřad, ředitelka odboru informačních služeb) a
- Ing. Michal Novotný (Český statistický úřad, ředitel odboru komunikace).

Výsledky soutěže včetně všech soutěžních plakátů najdete na webu Českého statistického úřadu:

`https://bit.ly/3qXP7jB`

2. Výsledky v kategorii základních škol, strana 23

- 1. místo, **Pranostiky**, Anna Samková, Matyáš Cafourek, Vítek Trávníček, Daniel Klement. ZŠ Litomyšl, U Školek 1117.
- 2. místo, **Třídění odpadu**, Pavlína Zemanová, Alžběta Svobodová. ZŠ Glowackého, Praha 8.
- 3. místo, **Problémy lesa a pralesa**, Barbora Dolanová, Terezie Fefková, Viktorie Kaplanová. Gymnázium Dr. Emila Holuba, Holice.

3. Výsledky v kategorii středních škol, strana 24

- 1. místo, **Životní prostředí**, Michal Lalik, Michal Sarközi. SOU Na Veselí 51, Praha 4.
- 2. místo, **Budoucnost bez ledovců?** Kristýna Marčíková, Veronika Voráčová, Nikol Nohelová. SŠ Brno, Charlubova.
- 3. místo, **Životní prostředí v Česku a Evropě** Patrik Černý, Matěj Nádeník. Gymnázium Dr. Emila Holuba, Holice.

4. Několik ukázek z ostatních plakátů, strana 25

- Pandemie světa — Emise.
- Podzemní voda a její nedostatky — Spotřeba vody.
- Kouření — Úbytek Včelstva.

to jest standardní exponenciální rozdělení. Dále odtud plyne, že řešení $\widehat{l}$ je vlastně řešením rovnice

$$\frac{1}{n}\sum_{i=1}^{n}\log y_i = \frac{1}{n}\sum_{i=1}^{n} y_i^{\widehat{l}}\log y_i - \frac{1}{\widehat{l}}, \tag{5}$$

kde y_i mají hustotu (4). To znamená, že $\widehat{l} = \widehat{m}/m$ má rozdělení nezávislé na m. Toto rozdělení ovšem nedovedeme vypočítat, protože ani $\widehat{l}$, tj. řešení (5) není vyjádřeno explicitně, ani nejsou známy jeho momenty. Bylo však vypočítáno metodou Monte Carlo, tj. simulacemi. To znamená, že pro každé $n = 2, 3, 4, \ldots$ byl N-krát generován vektor $Y_1, \ldots, Y_n$ nezávislých náhodných veličin Y_i s hustotou (4), vyřešena rovnice (5) a zaznamenán výsledek $\widehat{l}$. Tak byla získána empirická distribuční funkce veličiny $\widehat{l} = \widehat{m}/m$; chceme-li použít Fisherovy terminologie, pivotální veličiny $\widehat{m}/m$. A při dost vysokém N je tato empirická distribuční funkce dost blízká skutečné. Po vyrovnání empirické distribuční funkce veličiny $\widehat{l}$ vhodnou „hladkou“ formulí[7] pak byly získány kvantily $l_{p,n}$ statistiky $\widehat{l}$, tj. hodnoty, pro které platí

$$P\big(\widehat{l} \le l_{p,n}\big) = p.$$

Tyto byly tabelovány a publikovány, a inženýři v oboru spolehlivosti jich vesele využívají dodnes. V oné době autoři použili $N = 4\,500$. Tabelace přitom jde až k takovým hodnotám n, pro něž již lze užít asymptotickou teorii, podle níž $\widehat{l}$ má přibližně normální rozdělení.

Jsem možná hloupý, ale považuji tento krok za svého druhu průlom do teorie matematické statistiky. Možná také do způsobu vyučování. Za mých mladých let jedna z hlavních věcí při zkouškách byla obratnost při odvozování výběrových rozdělení. Na výše uvedeném příkladu je vidět, že můžeme mít výběrová rozdělení, která vůbec odvodit neumíme, a přesto jsou nám stejně užitečná. Tak například 90 % interval spolehlivosti pro m dostaneme snadno ze vztahu

$$\begin{aligned} P\big(l_{0.05,n} \le \widehat{l} \le l_{0.95,n}\big) &= P\Big(l_{0.05,n} \le \tfrac{\widehat{m}}{m} \le l_{0.95,n}\Big) = \\ &= P\Big(\tfrac{\widehat{m}}{l_{0.95,n}} \le m \le \tfrac{\widehat{m}}{l_{0.05,n}}\Big) = 0.90 \end{aligned}$$

jako $\big(\frac{\widehat{m}}{l_{0.95,n}}, \frac{\widehat{m}}{l_{0.05,n}}\big)$. Podobně, hypotézu $m = 1$, to jest vlastně hypotézu o exponenciálním rozdělení X, zamítneme na desetiprocentní hladině významnosti když $\widehat{m} < l_{0.05,n}$ nebo $\widehat{m} > l_{0.95,n}$, atd.

[7] Například Bernsteinovými polynomy, pozn. JA.

4) A kde je konečně ten Baron Prášil? ptá se netrpělivý čtenář. Tady:

V předchozím odstavci jsme ukázali, jak lze použít metodu Monte Carlo k získání výběrového rozdělení. Šlo ale o případ, kdy byl známý model, a výběr tak pocházel z rozdělení daného typu. V neparametrických úlohách se zabýváme výběry z rozdělení jen neurčitě popsaných, zpravidla jen „spojitý typ na dané polopřímce“, nebo „spojitý typ na celé reálné ose, symetrické rozdělení“, apod.

A tady je idea bootstrap, idea pana Efrona [3]: *Když neznáme typ rozdělení pravděpodobnosti, použijeme stejné metody jako v bodě 2) tohoto vyprávění, jenže na místě dané distribuční funkce $F(x)$, ze které se budou vytvářet výběry, použijeme empirickou distribuční funkci $F_n(x)$ získanou z dat samotných.* To, spolu s ukázkou z předcházejícího odstavce, pro mne znamená následující pokyny, varování a omezení.

a. Metoda bootstrap bude na místě jen tenkrát, když půjde o úlohu „ryze neparametrickou“, to jest úlohu, ve které se skutečně neodvážím formulovat žádný předpoklad o typu rozdělení výsledků pozorování. Tudíž odhadované veličiny nesmějí být vyjádřeny jako funkce parametrů v nějakém parametrickém systému distribučních funkcí, musí jít o nějaké funkcionály $\theta(F)$ na systému všech distribučních funkcí, například spojitého typu, jako jsou momenty, kvantily, atd.

b. Metoda je v podstatě asymptotická, platná jen při velmi vysokých hodnotách n. Jinak může být empirická distribuční funkce, kterou budu simulovat, i dost odlišná od skutečné. A vzniká otázka, jaké je ono „dost vysoké n“.

c. Abych mohl použít techniku simulací, měl bych asi mít možnost vhodného normování, vhodné standardizace své statistiky podobně jako v odstavci 3. Například bych asi měl vědět, že jde o rozdělení závislé na parametru polohy a měřítka, tj. typu $F(x) = G\left(\frac{x-\mu}{\sigma}\right)$, i když G není specifikováno.

Když se případným uživatelům[8] dost nezdůrazní asymptotická povaha metody, může se stát, že pět tisíckrát budou opakovat simulaci náhodného výběru z rozdělení diskrétního typu (z empirické distribuční funkce), která bude třeba dost odlišná od „skutečné“ distribuční funkce pozorované veličiny. Při malém rozsahu původního výběru pak jakýmkoliv „resampling postupem“ dostanu zase jen nějaké hodnoty podmíněné výsledky prvního výběru.

[8] Některým to můžete říci stokrát, a kde nic tu nic.

VYHODNOCENÍ NÁRODNÍHO KOLA SOUTĚŽE O NEJLEPŠÍ STATISTICKÝ PLAKÁT 2021

ISLP POSTER COMPETITION IN THE CZECH REPUBLIC IN 2021

Michal Novotný, Jakub Fischer

E-mail: michal.novotny@czso.cz, fischerj@vse.cz

1. Známe nejlepší statistický plakát roku!

Soutěž je určena dvou- až pětičlenným týmům žáků druhého stupně základních škol, studentů středních škol a bakalářských programů vysokých škol. Koná se v řadě zemí po celém světě a jejím hlavním smyslem je podporovat a rozvíjet statistickou gramotnost. V České republice ji organizují Vysoká škola ekonomická v Praze a Český statistický úřad. Nejlepší plakáty budou Českou republiku reprezentovat v mezinárodním kole soutěže, které proběhne v rámci červencového 63. světového statistického kongresu Mezinárodního statistického institutu v nizozemském Haagu.

„Soutěž o nejlepší statistický plakát se koná po celém světě a jejím hlavním smyslem je podporovat a rozvíjet statistickou gramotnost. Právě znalost, jak s daty správně pracovat a jak je chápat a vyhodnocovat, je v dnešní rychle rozvíjející se době jednou z klíčových dovedností,“ upozorňuje Marek Rojíček, předseda Českého statistického úřadu.

Pro letošní ročník tradiční celosvětové soutěže byla vybrána tři témata – životní prostředí, biologie a udržitelný rozvoj. Českého národního kola se zúčastnilo 482 žáků ve 171 týmech z 20 škol. Ty nejlepší plakáty pak školy zasílaly k hodnocení odborné porotě, která letos posuzovala 23 plakátů, z toho 10 v kategorii základních škol, 12 v kategorii středních škol a 1 v kategorii vysokých škol.

„Přestože byl letošní ročník poznamenán nepříznivou pandemickou situací, měla porota z čeho vybírat. Je velmi potěšující, že účastníci soutěže zvládli plakáty kvalitně připravit i v době omezeného osobního kontaktu a dokázali se s touto skutečností vypořádat,“ říká národní koordinátor soutěže a děkan Fakulty informatiky a statistiky VŠE v Praze Jakub Fischer.

V kategorii základních škol zvítězil plakát Pranostiky vytvořený týmem Základní školy Litomyšl. Druhé a třetí místo patří plakátům Třídění odpadu z dílny ZŠ Glowackého v Praze 8 a Problémy lesa a pralesa od autorů z Gymnázia Dr. Emila Holuba v Holicích. V kategorii středních škol vyhodnotila porota jako nejlepší plakát Životní prostředí ze Středního odborného

Úlohy soutěžících:

- Připravit si jeden PowerPoint slide, kde bude shrnutý obsah a cíl projektu
  Vzor prezentace najdete zde:
  https://www.humusoft.cz/event/technical-camp-2021/contest/vzor-tcc-2021.pptx
- Zaslat připravený PowerPoint slide o projektu na adresu: studnicka@humusoft.cz
  (do pondělí 6.9.2021)
- Připravit si cca. 5 minut prezentaci / popis svého projektu
- **Během soutěže ukázat výsledky své práce návštěvníkům u svého pracoviště**
  (s použitím vlastního hardware - notebook a pod.)

Soutěžící mají k dispozici:

- Prezentační stolek s možností elektrického napájení
- 5 minut na prezentaci svých projektů před publikem na začátku soutěže
- Experty v podobě aplikačních inženýrů společnosti Humusoft v případě potřeby

Průběh soutěže:

1. Začátek v 9:00
2. Přivítání a představení soutěžících
3. Prezentace projektů před publikem – pořadí soutěžících se určí na místě
4. Veletrh – diskuze a představení projektů ostatním účastníkům na pracovištích
5. Hlasování všemi účastníky s použitím hlasovacích lístků
6. Vyhodnocení a odevzdání hodnotných cen

Co Vám účast přináší?

- Možnost prezentovat výsledky své práce na fóru před dalšími uživateli
- Příležitost porovnat se s ostatními soutěžícími v oboru
- Účastníci–studenti mají šanci představit se případným budoucím zaměstnavatelům

Co můžete vyhrát?

- Bezplatné 1-denní školení ve školicím centru v Praze / Bratislavě dle výběru
- Studentská verze programu MATLAB
- Další hodnotné ceny (pevné i tekuté)

Může tak dojít ke „směšným“ aplikacím tohoto druhu. Z poměrně „chudých“ dat získal autor (ne já, ale ten jehož publikaci jsi mi zapůjčil a kterou zrovna nemám po ruce) odhad určitého ukazatele θ. Chtěl ale dostat odhad pravděpodobnosti $p_k(\theta)$ jevu $Y = k$ (kde Y je náhodná veličina diskrétního typu), které jsou závislé na θ. Simuloval tedy N krát N velké hodnotu Y při θ rovném odhadnuté hodnotě. Tvrdil přitom, že to je aplikace metody bootstrap. Podle mého to nebyl žádný bootstrap, nýbrž použití metody Monte Carlo k získání odhadů, které neuměl jinak vypočítat. Při generování stejně používal jistého modelu, který pravděpodobnosti $p_k(\theta)$ určoval. Podrobnosti snad dodám jindy.

5) Snad by se to mohlo ještě převyprávět v krátkosti takto. Když jsem byl student, říkal nám pan profesor Janko na jedné přednášce, že statistická práce má tři etapy:

1. problém specifikace;
2. problém odhadu;
3. problém distribuční.

Já těmto etapám rozumím takto:

1. To je to, co se dnes nazývá „volba modelu“, tj. formulace úlohy ve statistických termínech, v pojmech teorie pravděpodobnosti, atd. Tak například, jde o n nezávislých náhodných veličin s rozdělením gamma, je třeba odhadnout 90 % kvantil, apod. Ten problém má někdy řešení jasné, jindy je to hlavolam, při kterém vyjde najevo rozdíl mezi statistickým řemeslem a statistikou jako jistým druhem umění.
2. Rozhodnutí jaké statistiky použiji k odhadu neznámých parametrů modelu, volba „optimálního“ odhadu, apod. Tento problém je z různých hledisek pro mnoho standardních úloh vyřešen v teorii odhadu. Fantazii je ponecháno volné pole působnosti v tom, jakých kritérií při volbě odhadu použijeme, a jak přizpůsobíme odhad dostupným prostředkům.
3. Toto je také pro mnoho běžných situací vyřešeno. Všimni si, že po mnoho let byl značný podíl statistické literatury věnován právě úloze hledání jak výběrových, tak asymptotických rozdělení. Byly tomu věnovány metry knih, kilogramy tabulek, a dnes spousta programů. *Metodu boostrap vidím jako návrh řešení tohoto problému v případě, kdy řešení problému jedna je jen velmi obecné a neobsahuje předpoklad o tvaru rozdělení pozorovaných veličin.*

Na závěr otázka. Jestliže mám pravdu v otázce podstaty metody bootstrap, vnucuje se myšlenka: Při velkých rozsazích výběru by snad bylo možné určit příslušnost rozdělení X k některé třídě parametrických modelů, například k některému z rozdělení Johnsonova systému, a použít parametrické metody vhodné pro tento systém. Je něco známo o srovnání metody bootstrap a tohoto? Stejně by to ale asi nakonec musel rozhodnout počítačový experiment.

V Praze, neděle 16. února 1997 Tvůj J. M.

Od té doby uteklo mnoho let a Vltavou proteklo mnoho vody. Během ní vše nakynulo jako dobře zadělaná buchtička. Tak například:

- Google mi sdělil, že má k dispozici více než 98 500 000 odkazů na slovo bootstrap;
- Google Scholar byl skromnější a nabídl „jenom“ 1 350 000 odkazů;
- WOS k dnešnímu dni nabídl 44 374 prací se slovem boostrap v položce *topic*, a „nicotných“ 8 126 článků, kde se toto slovo vyskytuje přímo v názvu. Nic moc, čekal jsem jich víc. Atd.

O bootstrapu a jeho variantách [5] pravidelně slýcháme také na Robustu a čteme o nich v našem Bulletinu. Čtenářům bych rád připomněl především moc hezké články a vystoupení kolegyně Zuzanky Práškové, především [6] a [7]. Bohužel, pod názvem *bootstrap and other (re/sub) sampling methods* se dnes prodává leccos. Vidím to nejenom jako redaktor časopisů a sborníků, ale i jako autor, čtenář či oponent.

Přisuzuji to tomu, že i mnozí statistici si rádi popotáhnou za nazouvací poutko na patě boty, když se dostanou do potíží. A to mnohdy i tam, kde by nemuseli, a především tam, kde by neměli. Osobně jsem k bootstrapu nezahořkl, ale studentům, kteří zabrousí na přednášky o simulacích, se snažím našeptávat jistou „zdravou nedůvěru a opatrnost“ vůči některým postupům, jež si přinášejí z jiných přednášek a z četby „populárních statistických rodokapsů“.

Čtyřicátý den nouzového stavu léta páně 2020 v Praze sepsal a poznámkami opatřil J. Antoch

Soutěž o nejlepší uživatelský projekt

v rámci akce **Technical Computing Camp 2021**

s využitím programových prostředků MATLAB nebo COMSOL Multiphysics

Propozice soutěže

Termín konání soutěže: 10.9.2021 od 9:00 do 12:00

Místo konání: Hotel Rakovec, Brněnská přehrada, Brno

Soutěž o nejlepší uživatelský projekt proběhne ve druhém dni akce Technical Computing Camp 2021. Cílem soutěže je prezentování projektů, které účastníci realizovali v uplynulém období s použitím nástrojů MATLAB nebo COMSOL Multiphysics. Zvítězí ten, který svým projektem zaujme nejvíce ostatních účastníků Technical Computing Campu.

https://www.humusoft.cz/tcc

Soutěžní příspěvek:

- **Projekt nebo aplikace vytvořená v programu MATLAB či COMSOL Multiphysics**
- Zaměření projektu může být libovolné
  - spolupráce s hardware, zpracování obrazu / signálu, matematické modelování, ...
- Autorem může být jednotlivec či skupina, alespoň jeden člen skupiny musí být osobně přítomen na akci Technical Computing Camp
- Studenti se mohou zúčastnit se svou diplomovou či bakalářskou prací

Doporučené čtení

[1] *Informační Bulletin České statistické společnosti* **2**, 1999, `https://www.statspol.cz/oldstat/bulletiny/ib-99-2.pdf`. *cit. 4*

[2] Antoch, J., Hušková, M., Veraverbeke, N.: Change-point problem and bootstrap. *Journal of Nonparametric Statistics* **5**, 123 – 144. *cit. 3*

[3] Efron, B.: Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1 – 26, 1979. *cit. 8*

[4] Efron, B.: *The Jaknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia, 1982. *cit. 3*

[5] Mammen, E.: *When Does Bootstrap Work? Asymptotic Results and Simulations.* Springer-Verlag, New York, 1992. *cit. 10*

[6] Prášková, Z.: *Metoda bootstrap.* ROBUST 2004, 299 – 314, JČMF Praha, 2004. *cit. 10*

[7] Prášková, Z.: *Metoda subsampling a její aplikace v časových řadách.* ROBUST 2000, 227 – 236, JČMF Praha, 2001. *cit. 10*

[8] Thoman, D. R., Bain, L. J., Antler, Ch. E.: Maximum likelihood estimation, exact confidence intervals for reliability, and tolerance limits in the Weibull distribution. *Technometrics* **12**, 363 – 371, 1970. *cit. 6*

[9] Dámy a pánové v Londýně, `https://www.idnes.cz/zpravy/zahranicni/britanie-londyn-lgbt-bezpohlavni-metro-gender-neutralni-pozdrav.A170714_104108_zahranicni_lre`. *cit. 3*

[10] Dámy a pánové v New Yorku, `https://www.reflex.cz/clanek/zpravy/83130/dalsi-zakaz-osloveni-damy-a-panove-v-mhd-po-londynu-se-pridava-new-york.html`. *cit. 3*

[11] Dámy a pánové v Air Canada, `https://refresher.cz/70298-Aerolinka-bude-pasazery-vitat-genderove-neutralnim-pozdravem-Osloveni-damy-a-panove-je-minulosti`. *cit. 3*

[12] Dámy a pánové v Easy Jet, `https://zdopravy.cz/zadne-damy-a-panove-easyjet-zacne-cestujici-oslovovat-genderove-neutralne-39526/`. *cit. 3*

VZPOMÍNKA NA PROFESORA JIŘÍHO ANDĚLA
THE MEMORY OF THE PROFESSOR JIŘÍ ANDĚL

Tomáš Cipra
E-mail: cipra@karlin.mff.cuni.cz

Dlouholetý vedoucí katedry pravděpodobnosti a matematické statistiky MFF UK v Praze, dlouholetý proděkan téže fakulty a jedna z nejvýraznějších postav české matematické statistiky prof. RNDr. Jiří Anděl, DrSc., zemřel dne 29. dubna 2021 ve věku 82 let.[1]

Narodil se dne 7. března 1939 v Jenišovicích v okrese Jablonec nad Nisou. Základní školu navštěvoval ve svém rodišti. Vzhledem k zájmu o matematiku se přihlásil ke studiu na MFF UK, kde studoval v letech 1956–1961. Již během studia si ho prof. Janko vybral jako asistenta na katedře statistiky. Vědeckou přípravu absolvoval na katedře pravděpodobnosti a matematické statistiky pod vedením prof. Hájka a stal se důstojným pokračovatelem Hájkova díla jak v oblasti vědecké, tak pedagogické. Kandidátskou disertační práci *Lokální asymptotická mohutnost testů typu Kolmogorova-Smirnova* obhájil v roce 1965 (výsledky této práce byly v roce 1967 publikovány v prestižních Annals of Mathematical Statistics). Na docenta MFF UK se habilitoval v roce 1977 na základě habilitační práce *Mnohorozměrné autoregresní posloupnosti*. Od téhož roku byl pověřeným vedoucím a od roku 1981 pak řádným vedoucím katedry pravděpodobnosti a matematické statistiky. Po vypracování a obhájení doktorské disertace na téma *Některé míry závislosti v časových řadách* mu byla udělena v roce 1981 vědecká hodnost DrSc. V roce 1986 byl jmenován vysokoškolským profesorem. V letech 1993–1996 působil jako proděkan pro matematiku a od roku 1996 jako pedagogický proděkan na MFF UK. Kontakt s fakultou neztratil ani ve vysokém důchodovém věku jako emeritní profesor.

Prof. Anděl odborně pracoval především v oblasti matematické statistiky a časových řad. O rozsahu a úspěšnosti jeho odborné činnosti svědčí mimo jiné

[1] Tato vzpomínka na pana profesora paralelně vychází v časopisu Pokroky matematiky, fyziky a astronomie (zkr. PMFA).

PLÁNOVANÉ AKCE V ROCE 2021
SELECTED CONFERENCES IN 2021

Redakce časopisu

Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava
www.ssds.sk

Predbežný plán akcií SŠDS v roku 2021

Variabilný symbol	Termín konania	Názov podujatia
2101	22.4.2021	Pohľady na ekonomiku Slovenska 2021 – **online** *Téma*: (bude upresnená neskôr)
2102	**31.5.2021**	**EKOMSTAT 2021 – online** (tematické okruhy konferencie sú uvedené v pozvánke na EKOMSTAT)
	preložené na začiatok leta 2022	**21. letná škola ROBUST, Bardejov (ČStS + SŠDS) – preložená akcia z roku 2020 sa uskutoční v novom termíne, niekedy začiatkom leta roku 2022. Presný termín a ďalšie podrobnosti budú zverejnené neskôr.** *Téma*: Súčasné trendy štatistiky, pravdepodobnosti a analýzy údajov
2104	9. – 10.9.2021	20. Slovenská štatistická konferencia (SŠK) + 18. Slovenská demografická konferencia (SDK), Banská Bystrica *Téma SŠK*: (bude upresnená neskôr) *Téma SDK*: (bude upresnená neskôr)
2105	2. – 3.12.2021	29. medzinárodný seminár „Výpočtová štatistika“, Bratislava, (prípadne online)
2106	2. – 3.12.2021	Prehliadka prác mladých štatistikov a demografov, Bratislava, (prípadne online)
2107	priebežne	regionálne akcie
2108	priebežne	diskusné popoludnia, prednášky

odhadnout pravděpodobnosti některých jevů kolem nás. Tyto příklady, v nichž jde třeba o to bez počítání odhadnout aposteriorní pravděpodobnosti z daných apriorních pravděpodobností, pochopitelně působí uměle, protože při reálném rozhodování nebýváme nuceni se rozhodovat naprosto zbrkle. Na druhou stranu by málokdo v reálném životě aplikoval v dané situaci Bayesův vzorec, tím spíš vzhledem k tomu, že naše logické uvažování (slovy autora) ani nebývá ochotno zpochybňovat názor, ke kterému předtím došla (třeba i jen bleskovou rychlostí) naše intuice.

Přestože lidský úsudek je v knize vykreslen jako iracionální a interpretován doslova jako absurdní, směšný či trapný, kniha vyznívá překvapivě optimisticky a naše vlastní nedokonalost je popisována s vtipem. I to přispívá ke čtivosti knihy, zrovna tak jako (často udivující) výsledky zajímavých psychologických experimentů.

K celé knize mám snad jen tyto drobné připomínky:

- Při popisu dvou systémů lidského myšlení (rychlého intuitivního a pomalého logického) je zarážející, že kniha totálně ignoruje individuální odlišnosti (snad až na drobnou zmínku na str. 148) a nezmiňuje dostupné studie, podle nichž mají některé osobnostní typy větší sklony k intuitivnímu uvažování než jiné, viz Miková Š. (2018): *Nejsou stejné. Jak díky Teorii typů porozumět dětem i sami sobě.* Mea Gnosis, Praha.
- Na str. 220 uvádí autor patrně ve snaze maximálně přiblížit základní pravděpodobnostní pojmy laikům, že korelační koeficient r vždy nabývá hodnot z intervalu [0,1]. Tím spíš může být laik zaskočen, když v příkladu na str. 281 vychází r záporné.
- Na str. 196 nemusí být jasné, co znamená, že korelace je méně než dokonalá. Nedokonalý je zde spíš překlad; zřejmě se rozumí, že nemusí platit $r = 1$ ani $r = -1$, tj. nemusí nastat situace označovaná jako *perfect correlation*.
- V obecném kontextu autor zpochybňuje experty a jejich úsudky, nicméně myslím, že se zde vyjadřuje zejména k oborům, v nichž sám pracoval (psychologie, ekonomie). Např. v medicíně se i přes pokroky v oblasti umělé inteligence uznává, že lékař má oproti algoritmům možnost využívat své tacitní znalosti. Zatímco Kahneman je ve své knize ani nezmiňuje, jde o oblíbený pojem v klinickém rozhodování popisující neformalizované znalosti expertů získané dlouhodobými zkušenostmi.

92 vědeckých prací (často ve velmi prestižních odborných časopisech), 6 knih, 4 skripta, 54 popularizačních prací, 26 výzkumných zpráv a velké množství zahraničních citací. Z jeho citovaných výsledků lze uvést práce týkající se různých typů časových řad: autoregresních, mnohorozměrných, nelineárních, nezáporných, inverzních, s náhodnými či periodickými parametry, s dlouhou pamětí aj. V oblasti časových řad jsou dále citovány práce prof. Anděla věnované interpolování a extrapolování (predikcím), závislosti mezi časovými řadami, řadám s daným marginálním rozdělením či danými momenty, speciálním (např. bayesovským) odhadovým procedurám, spektrálním vlastnostem a další problematice.

Renomovaný časopis Journal of Time Series Analysis mu za publikační činnost udělil cenu Distinguished Author Award. Za soubor prací *Statistické modely časových řad a jejich simulace* mu byla v roce 1990 udělena Národní cena ČR. V této souvislosti je také nutné zmínit členství prof. Anděla v mezinárodní vědecké společnosti The Biometric Society, ve Vědeckém kolegiu matematiky ČSAV a předsednictví v komisi pro udělování hodnosti DrSc. v oboru Pravděpodobnost a matematická statistika. Své výsledky prezentoval (často jako zvaný řečník) na řadě zahraničních univerzit, mezinárodních konferencí a kongresů. Některé práce také vznikly ve spolupráci s renomovanými zahraničními autory. Byl vedoucím řešitelem úspěšných grantových projektů (např. GAČR *Časové řady a příbuzné modely*).

Vedle teoretického výzkumu se prof. Anděl také významně věnoval činnosti aplikační (27 aplikačních prací), která byla mimo jiné motivována spoluprací s některými praktickými institucemi z oblasti zdravotnictví (např. Institut hygieny a epidemiologie v Praze), průmyslu (např. Škoda Plzeň) nebo hydrologie (Vodohospodářský ústav). Podílel se tak na řešení praktických problémů typu periodicity v průtocích vodních toků či analýzy biosignálů EEG a dalších. Praktické výsledky, které pracovníkům z praxe při řešení konkrétních problémů předkládal, jsou velmi úspěšnými a přesvědčivými argumenty o užitečnosti matematické statistiky pro praxi. Byl členem kolektivu, který v roce 1982 získal cenu Purkyňovy společnosti a v roce 1983 cenu ministra zdravotnictví.

Zvláštní pozornost si zasluhují knihy, které prof. Anděl napsal. Monografie *Statistická analýza časových řad* (SNTL 1976) je dodnes používána jako základní referenční materiál v pracích věnovaných teorii či aplikacím časových řad (to samé platí pro její německý překlad z roku 1984 v německy mluvících zemích). Jeho nejznámější knihou je ovšem *Matematická statistika*: pokud měl u nás kdokoli co do činění s (matematickou) statistikou, určitě se setkal s touto vynikající publikací nebo dostal radu, aby se podíval do „modré knihy“. Na ni později navázala její pozměněná verze *Základy ma-*

tematické statistiky označovaná pro odlišení jako „žlutá kniha“. Velký zájem byl ovšem také o jeho *Statistické metody* (zatím poslední 5. vydání v roce 2019). Zůstával aktivní i v pozdním věku: v roce 2018 mu vyšla kniha *Statistické úlohy, historky a paradoxy* a pro studenty zveřejňoval na síti materiály k přednáškám a seminářům.

Velkou zásluhu má Prof. Anděl na propagaci statistiky u nás. Je s podivem, že při intenzivní vědecké práci, pedagogické činnosti a náročných funkcích našel čas na práce propagující statistiku a publikované např. v časopisech PMFA, Rozhledy matematicko-fyzikální, Věda a technika mládeži, Statistika, Informační bulletin České statistické společnosti apod. (přinejmenším 20 prací). Řada z těchto problémů byla použita v jeho knize *Matematika náhody* (zatím poslední čtvrté vydání v roce 2020), jejíž anglický překlad *Mathematics of Chance* vyšel v roce 2009 v nakladatelství Wiley. V této souvislosti je také nutné zdůraznit, že prof. Anděl byl v roce 1990 zakládajícím předsedou České statistické společnosti, v jejímž čele stál do roku 1993.

Nelze ovšem zapomenout ještě na jednu důležitou skutečnost. Prof. Anděl vždy spatřoval smysl své práce v činnosti pedagogické. Nejenže se na své přednášky pečlivě připravoval, ale vlastnil dar vyložit i velmi složité partie názornou a snadno pochopitelnou formou. I v nejstresovějších situacích měla u něj výuka vždy přednost. V anketním hodnocení studentů získával tradičně nejvyšší počet bodů a platilo, že právě díky jeho úvodním přednáškám se relativně velký počet posluchačů hlásil na statistický obor. Velmi důkladně a poctivě se také věnoval svým diplomantům a doktorandům (a to i zahraničním, např. ze Španělska), kteří se pod jeho vedením učili preciznosti a odpovědnému přístupu k vědecké práci, přičemž mnozí z nich působí v oboru dodnes, a tím rozvíjejí a předávají jeho odkaz.

Odchod Jiřího Anděla znamená velkou ztrátu nejen pro obor matematické statistiky, ale pro všechny, kteří ho blíže osobně potkali.

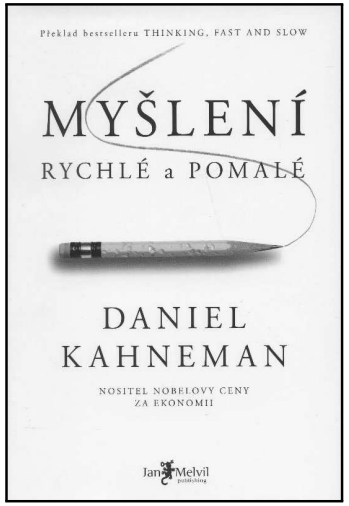
Tomáš Cipra

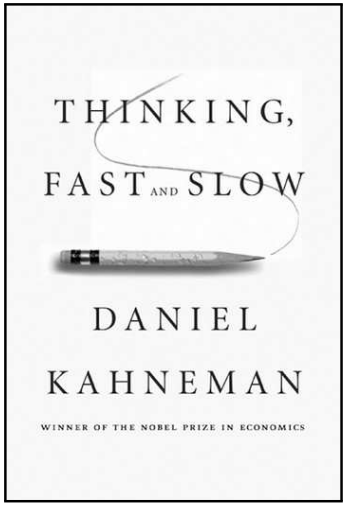
RECENZE: MYŠLENÍ RYCHLÉ A POMALÉ
BOOK REVIEW: THINKING, FAST AND SLOW

Jan Kalina

E-mail: kalina@cs.cas.cz

Kniha vyšla v angličtině v roce 2011, v češtině v roce 2012 v nakladatelství Jan Melvil Publishing. Kniha má 542 stran, recenze je z března 2021.





Knihu hodnotím jako mimořádně zdařilé dílo o rozhodování za nejistoty a o nedokonalosti lidské intuice. Nejprve popisuje dva základní systémy lidského myšlení (kap. 1). Ty tvoří psychologické pozadí pro kapitoly 2 a 3 věnované rozhodování za neurčitosti; statistikům budou jistě připadat tyto kapitoly za nejhodnotnější. Dále se kniha věnuje ekonomickému rozhodování (kap. 4) a teorii štěstí (kap. 5); právě za ekonomické aplikace poznatků o rozhodování získal Daniel Kahneman Nobelovu cenu za ekonomii v roce 2002. Kniha je hned od první stránky čtivá a přináší řadu převratných myšlenek.

Čím kniha zaujme *statistickou* komunitu? Mě především zaujaly některé úlohy, v nichž se má čtenář rychle zorientovat a intuitivně (tedy co nejrychleji) si tipnout výsledek. V některých takových hádankách jsem se nechal také nachytat. Autor zde argumentuje, že není dobré se spoléhat na svou statistickou intuici; tím ovšem má na mysli pravděpodobnostní intuici nebo schopnosti