INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 31, číslo 3, září 2020

NEÚPLNÉ VZORKY Z POISSONOVA ROZDĚLENÍ
INCOMPLETE POISSON SAMPLES

Ondřej Zeman, Jiří Dvořák

Adresa: Matematicko-fyzikální fakulta Univerzity Karlovy, Sokolovská 83, 186 75 Praha 8

E-mail: zeman.ond@seznam.cz, dvorak@karlin.mff.cuni.cz

Abstrakt: V tomto příspěvku se budeme zabývat odhadem parametrů z neúplného vzorku z Poissonova rozdělení, konkrétně z té části náhodného výběru, ve které chybí nulová pozorování. Cílem je odhadnout rozsah původního náhodného výběru N a parametr Poissonova rozdělení λ. Zaměříme se na odhady popsané ve dvou článcích. Novější článek uvádí, že v daném místě pouze opakuje postup staršího článku. Ve skutečnosti ale vychází z jiných věrohodnostních funkcí a dospěje k mírně odlišným odhadům. V obou článcích odvození tvaru věrohodnostních rovnic i výsledných odhadů chybí, v našem příspěvku proto podrobná odvození doplníme a vyjasníme vztah mezi oběma články. Příspěvek vychází z bakalářské práce prvního autora.

Klíčová slova: useknuté Poissonovo rozdělení, maximální věrohodnost, neúplný výběr.

Abstract: This contribution addresses the problem of parameter estimation from an incomplete Poisson sample, i.e. from the part of the random sample where zero values are missing. The aim is to estimate the size of the original sample N and the parameter λ of the Poisson distribution. We focus on the estimators from two papers. The newer one claims that, in the relevant part, it only reviews the procedure from the older paper. However, it uses different likelihood functions and arrives at slightly different estimators. Derivations are missing in both of the papers and hence we present detailed derivations here and we clarify the connections between the papers. This contribution is based on the bachelor thesis of the first author.

Keywords: truncated Poisson distribution, maximum likelihood, incomplete sample.

1. Úvod

V této práci se budeme zabývat bodovými odhady parametrů z náhodného výběru z Poissonova rozdělení, kde chybí všechna nulová pozorování. Vliv chybějících pozorování může být velmi výrazný v případě, že skutečná hodnota parametru Poissonova rozdělení je blízká nule. Jako modelovou situaci,

kdy nelze pozorovat nulové hodnoty, můžeme použít příklad z [5], který se zabývá epidemií cholery v nejmenované indické vesnici.

Cholera má jako většina infekčních nemocí určitou inkubační dobu a lidé mohou být nakaženi, přestože se u nich zatím neprojevily příznaky. Označme N počet domácností zasažených nákazou a předpokládejme, že počet lidí s příznaky v domácnosti zasažené nákazou má Poissonovo rozdělení s parametrem λ. Nahlášené jsou ale jen počty lidí s příznaky z těch domácností, kde má příznaky alespoň jeden člověk (32 domácností reportuje jednoho nemocného s příznaky, 16 domácností reportuje dva, 6 domácností reportuje tři, 1 domácnost reportuje čtyři). Zůstává však neznámý počet domácností zasažených cholerou, kde (zatím) nikdo nemá příznaky. Nulová pozorování tedy nejsou k dispozici. Cílem je nyní odhadnout parametr λ a celkový rozsah výběru N (tedy včetně neznámého počtu nulových pozorování).

Poznamenejme na okraj, že je možné k problému přistoupit z opačné strany a pozorování si doplnit nulami až do celkového počtu domácností ve vesnici. V tomto kontextu je pak možné využít tzv. „zero-inflated Poisson“ modelů.

V tomto příspěvku představíme odhady z článků [1] a [5]. Protože v původních článcích chybí odvození použitých věrohodnostních funkcí i výsledných odhadů, tato odvození zde podrobně provedeme. Půjde vpravdě o detektivní pátrání, protože novější článek [1] v dané pasáži tvrdí, že pouze opakuje postup ze staršího článku [5], ačkoli uvidíme, že použité věrohodnostní funkce i výsledné odhady se od sebe liší. Tento rozpor je velmi překvapivý s ohledem na to, že autorské kolektivy obou článků nejsou disjunktní. Naopak jeden je podmnožinou druhého...

Příspěvek vychází z bakalářské práce prvního autora [8]. V ní jsou diskutovány i další způsoby odhadu parametrů, odvozeny některé jejich asymptotické vlastnosti a pomocí simulací porovnána přesnost odhadů pro výběry s konečným rozsahem.

2. Základní pojmy

Nechť $Y_1, \ldots, Y_N$ je náhodný výběr z Poissonova rozdělení s parametrem $\lambda > 0$. Tento výběr budeme nazývat *úplným výběrem*. Jako $X_1, \ldots, X_n$ označíme ty náhodné veličiny z úplného výběru, které nemají nulovou hodnotu. Náhodné veličiny $X_1, \ldots, X_n$ budeme nazývat *neúplným výběrem*, případně *neúplným vzorkem*. V dalším výkladu budeme považovat $n \in \mathbb{N}$ za známé, naopak $N \in \mathbb{N}, N \geq n$, považujeme za neznámé a budeme ho chtít odhadnout spolu s parametrem λ. Situaci, kdy $n = 0$, mlčky přehlížíme, což však není příliš omezující.

Dále budeme značit symbolem n_x počet pozorování s hodnotou $x \in \mathbb{N}$, symbolem R nejvyšší pozorovanou hodnotu, S bude součet všech pozorovaných hodnot, $S = \sum_{x=1}^{R} x n_x$, $\lfloor a \rfloor$ bude dolní celá část kladného čísla a.

Netřeba jistě připomínat, že pravděpodobnosti v Poissonově rozdělení mají tvar $\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$, $k = 0, 1, \ldots$ Uvádíme je však pro porovnání s pravděpodobnostmi *useknutého Poissonova rozdělení*, které mají tvar $\frac{\mathrm{e}^{-\lambda}}{1-\mathrm{e}^{-\lambda}}\frac{\lambda^k}{k!}$, $k = 1, 2, \ldots$ Jde vlastně o pravděpodobnosti Poissonova rozdělení za podmínky, že náhodná veličina má kladnou hodnotu.

3. Starší článek

V této části se zaměříme na odvození postupu ze staršího článku [5]. Vychází z věrohodnostních funkcí

$$L(\lambda, N) = \mathrm{e}^{-N\lambda} \prod_{x=1}^{R} \left[\frac{\lambda^x}{x!}\right]^{n_x}, \tag{1}$$

$$L^*(\lambda, N) = \binom{N}{n} p^n q^{N-n}, \tag{2}$$

kde $p = 1 - \mathrm{e}^{-\lambda}$, $q = 1 - p = \mathrm{e}^{-\lambda}$. V rovnici (1) jsme proti článku [5] doplnili do exponentu hodnotu n_x, protože její absenci v článku považujeme za tiskovou chybu. Bez tohoto exponentu by totiž věrohodnostní funkce neobsahovala informaci o počtu pozorování pro jednotlivé hodnoty x, což by nedávalo smysl.

Odvození věrohodnostních funkcí

Funkce (1) je klasickou věrohodnostní funkcí pro úplný náhodný výběr $Y_1, \ldots, Y_N$ z Poissonova rozdělení:

$$L(\lambda, N) = \prod_{i=1}^{N} \mathrm{e}^{-\lambda} \frac{\lambda^{Y_i}}{Y_i!} = \mathrm{e}^{-N\lambda} \prod_{x=1}^{R} \left[\frac{\lambda^x}{x!}\right]^{n_x}.$$

Tato funkce je monotónní v proměnné N a pokud bychom ji maximalizovali samu o sobě, vždy bychom jako odhad N dostali nejmenší přípustnou hodnotu, tedy n. To není žádoucí a samotná funkce (1) k odhadu nestačí.

Protože platí $p = \mathbb{P}(Y_i \neq 0)$ a $q = \mathbb{P}(Y_i = 0)$, popisuje funkce (2) pravděpodobnost, že v úplném náhodném výběru o rozsahu N je n nenulových pozorování a $N - n$ nulových pozorování. Přidání této funkce má umožnit lepší odhad parametru N.

Odhady parametrů

Odhady $\widehat{\lambda}$ a $\widehat{N}$ mají podle článku maximalizovat rovnice (1) a (2) současně. Autoři pak bez odvození uvádějí, že výsledné odhady splňují vztahy

$$\frac{\widehat{\lambda}}{1-\mathrm{e}^{-\widehat{\lambda}}} = \frac{S}{n}, \tag{3}$$

$$\widehat{N} = \frac{n}{1-\mathrm{e}^{-\widehat{\lambda}}}. \tag{4}$$

Pomocné lemma

Při maximalizaci funkce (2) využijeme následujícího pomocného lemmatu [7, s. 144]. V odkazovaném článku se objevuje bez důkazu jako „známé“, pouze s odkazem do práce [3, s. 142]. Tam je však formulováno pro pravděpodobnosti v hypergeometrickém rozdělení. Upravenou verzi důkazu, odpovídající naší situaci, proto uvádíme níže.

Lemma 1. *Nechť* $q(\lambda) = \mathrm{e}^{-\lambda}$, $p(\lambda) = 1 - q(\lambda) = 1 - \mathrm{e}^{-\lambda}$. *Pro libovolné pevné* $\lambda > 0$, $\widehat{N} = \left\lfloor \frac{n}{p(\lambda)} \right\rfloor$ *maximalizuje* $L^*(\lambda, N) = \binom{N}{n} p(\lambda)^n q(\lambda)^{N-n}$. *Pokud* $p(\lambda) = \frac{n}{N'}$ *pro nějaké* $N' \in \mathbb{N}$, *pak* $\widehat{N}$ *i* $\widehat{N} - 1$ *maximalizují* $L^*(\lambda, N)$. *Jinak* $\widehat{N}$ *je jediné maximum.*

Důkaz. Zvolme nejprve pevnou hodnotu $\lambda > 0$. Pro přehlednost budeme dále psát pouze $p = p(\lambda)$, $q = q(\lambda)$. Protože N je celočíselná proměnná, budeme nejprve uvažovat podíl dvou následujících hodnot funkce L^*, tedy

$$C(N) = \frac{L^*(\lambda, N+1)}{L^*(\lambda, N)} = \frac{\binom{N+1}{n} p^n q^{N+1-n}}{\binom{N}{n} p^n q^{N-n}} = \frac{q(N+1)}{(N+1-n)}, \quad N \geq n.$$

Dále si definujeme odpovídající funkci reálné proměnné $C(V) = \frac{q(V+1)}{(V+1-n)}$ pro $V \in \mathbb{R}$, $V \geq n$. Její derivace je záporná a funkce $C(V)$ je proto klesající:

$$C'(V) = \frac{-qn}{(V+1-n)^2} < 0, \quad V \geq n.$$

Předpokládejme nyní, že $C(n) < 1$. Z tohoto předpokladu plyne, že funkce L^* nabývá maxima pro $N = n$. Platí

$$\begin{aligned} C(n) = q(n+1) &< 1, \\ -q(n+1) + (n+1) &> -1 + (n+1), \\ n + 1 &> \frac{n}{1-q}, \\ n + 1 &> \frac{n}{p}. \end{aligned}$$

Současně platí, že $p < 1$, a proto $\frac{n}{p} > n$. Platí tedy $\left\lfloor \frac{n}{p} \right\rfloor = n$. V tomto případě nemůže L^* nabývat maxima ve dvou různých bodech. Tvrzení lemmatu tedy v tomto případě platí, protože funkce L^* skutečně nabývá maxima v bodě $n = \left\lfloor \frac{n}{p} \right\rfloor$.

Dále předpokládejme $C(n) \geq 1$. Nyní hledáme N takové, že $C(N) \geq 1$, $C(N+1) < 1$. Takové N existuje, protože C je klesající funkce, $C(n) \geq 1$ a

$$\lim_{N \to \infty} C(N) = q \ < \ 1.$$

Hledaný bod maxima pak bude

$$\widehat{N} = N + 1, \tag{5}$$

jak je vidět z definice funkce $C(N)$. Řešíme tedy soustavu nerovnic

$$\begin{aligned} \frac{q(N+1)}{(N+1-n)} &\geq 1, \\ \frac{q(N+2)}{(N+2-n)} &< 1. \end{aligned}$$

Po úpravě dostáváme podmínky

$$\frac{n}{1-q} \geq \ N + 1 \ > \frac{n}{1-q} - 1.$$

Pokud $1 - q = \frac{n}{N'}$ pro nějaké $N' \in \mathbb{N}$, dostáváme

$$N' \geq N + 1 > N' - 1. \tag{6}$$

Z toho plyne $N' = N + 1$, neboť N i N' jsou přirozená čísla, a v nerovnici (6) je neostrá nerovnost rovností. Vzhledem k (5) je tedy $\widehat{N} = N + 1 = N'$.

Z podmínky $C(N) = 1$ plyne, že $L^*(\lambda, \widehat{N}) = L^*(\lambda, \widehat{N} - 1)$. V obou těchto bodech je tedy maximum funkce $L^*(\lambda, N)$. V tomto případě platí, že $N' = n/(1-q)$, a proto $\widehat{N} = N' = \frac{n}{p}$.

Pokud $1-q \neq \frac{n}{N'}$ pro žádné $N' \in \mathbb{N}$, nemůže v (6) nastat rovnost, a proto je

$$\frac{n}{p} > \ N + 1 \ > \frac{n}{p} - 1.$$

Vzhledem k (5) je pak $\widehat{N} = \left\lfloor \frac{n}{p} \right\rfloor$ jediné maximum.

□

Odvození tvaru odhadů ze staršího článku

V článku [5] chybí postup, jak přistoupit k maximalizaci funkcí L a L^*, viz (1) a (2). Pouze se zde objevuje zmínka, že odhady mají maximalizovat obě funkce současně. Nejprve se tedy pokusíme maximalizovat obě funkce jako funkce dvou proměnných současně, přičemž proměnnou N budeme brát jako celočíselnou. Ukážeme ovšem, že tento postup nevede k cíli. Následně ukážeme, že odhady z článku je možné získat maximalizací obou funkcí vzhledem k λ a volbě $\widehat{N}$ tak, aby obě funkce nabývaly maxima pro stejné $\widehat{\lambda}$.

První postup. Začneme maximalizací $L(\lambda, N)$ vzhledem k N a uvažujeme podíl

$$\frac{L(\lambda, N)}{L(\lambda, N-1)} = \frac{\mathrm{e}^{-N\lambda}}{\mathrm{e}^{-(N-1)\lambda}} = \mathrm{e}^{-\lambda} < 1, \quad \forall \lambda > 0.$$

Tento podíl jako funkce N je proto klesající a z toho plyne, že bodem maxima je $\widehat{N_1} = n$. Nyní zafixujme $N = \widehat{N_1}$ a hledejme odhad λ. Logaritmováním a dalšími úpravami dostaneme

$$l(\lambda, \widehat{N_1}) = \ln(L(\lambda, \widehat{N_1})) = -\widehat{N_1}\lambda + \sum_{x=1}^{R} \big(x n_x \ln \lambda - n_x \ln(x!) \big),$$

$$\frac{\partial l}{\partial \lambda}(\lambda, \widehat{N_1}) = -\widehat{N_1} + \sum_{x=1}^{R} \frac{x n_x}{\lambda} = 0,$$

$$\widehat{\lambda}_1 = \sum_{x=1}^{R} \frac{x n_x}{\widehat{N_1}} = \sum_{x=1}^{R} \frac{x n_x}{n} = \frac{S}{n}.$$

Tedy funkce $L(\lambda, N)$ je maximalizována pro $\lambda = \widehat{\lambda}_1 = S/n$, $N = \widehat{N}_1 = n$.

Dále použijeme lemma 1 k maximalizaci L^* podle N. Dostáváme, že pro pevně zvolené $\lambda > 0$ je maximum této funkce v bodě $\widehat{N} = \left\lfloor \frac{n}{1-\mathrm{e}^{-\lambda}} \right\rfloor$ v případě, že $\frac{n}{1-\mathrm{e}^{-\lambda}}$ není celé číslo. Jinak jsou body maxima hodnoty $\frac{n}{1-\mathrm{e}^{-\lambda}}$ a $\frac{n}{1-\mathrm{e}^{-\lambda}} - 1$.

Dále hledejme hodnotu λ, pro kterou nabývá funkce $L^*(\lambda, \widehat{N})$ maxima. Existuje posloupnost λ_i, $i \in \mathbb{Z}$, taková, že pro každé $i \in \mathbb{Z}$ je $\lambda_i > 0$ a navíc $\frac{n}{1-\mathrm{e}^{-\lambda_i}}$ je přirozené číslo (takovou posloupností je například $\lambda_i = \ln(1+1/i)$). Takových hodnot λ_i je spočetně mnoho, protože funkce $\frac{n}{1-\mathrm{e}^{-\lambda}}$ je monotónní funkcí λ.

Funkce $L^*(\lambda, \widehat{N})$ je na každém intervalu $(\lambda_i, \lambda_{i+1})$ diferencovatelná a kladná, a proto lze na každém z těchto intervalů funkci $l^*(\lambda, \widehat{N})$ derivovat. Logaritmus je rostoucí funkce, takže se zachová i monotonie na intervalech $(\lambda_i, \lambda_{i+1})$. Dostáváme

$$
\begin{aligned}
l^*(\lambda, \widehat{N}) &= \ln L^*(\lambda, \widehat{N}) = \ln \binom{\widehat{N}}{n} + n \ln(1 - \mathrm{e}^{-\lambda}) - (\widehat{N} - n)\lambda, \\
\frac{\partial l^*}{\partial \lambda}(\lambda, \widehat{N}) &= \frac{n\mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}} - \widehat{N} + n = \\
&= \frac{n}{1 - \mathrm{e}^{-\lambda}} - \frac{n}{1 - \mathrm{e}^{-\lambda_{i+1}}} > 0, \quad \forall \lambda \in (\lambda_i, \lambda_{i+1}).
\end{aligned}
$$

Z toho dostáváme, že funkce $L^*(\lambda, \widehat{N})$ je na každém takovém intervalu rostoucí, a proto jedinými kandidáty na body maxima jsou hodnoty λ_i.

Předpokládejme tedy, že nějaké konkrétní λ_i je bodem maxima. Z lemmatu 1 a vlastnosti $\left\lfloor \frac{n}{1-\mathrm{e}^{-\lambda_i}} \right\rfloor = \frac{n}{1-\mathrm{e}^{-\lambda_i}}$, $i \in \mathbb{Z}$, plyne, že máme dva odhady parametru N maximalizující rovnici $L^*(\lambda_i, N)$ vzhledem k N:

$$
\widehat{N_2} = \frac{n}{1 - \mathrm{e}^{-\lambda_i}}, \qquad \widehat{N_3} = \frac{n}{1 - \mathrm{e}^{-\lambda_i}} - 1.
$$

Jak ukážeme dále, oba tyto odhady jsou pro každé λ_i v rozporu s odhady získanými maximalizací funkce L. To je zřejmé pro odhad $\widehat{N}_2$, který se nemůže rovnat odhadu $\widehat{N}_1 = n$, protože jmenovatel zlomku je vždy menší než 1.

Předpokládejme tedy, že $\widehat{N_3} = \widehat{N_1} = n$. Dostáváme tedy

$$\begin{aligned} n &= \frac{n}{1 - \mathrm{e}^{-\lambda_i}} - 1, \\ \frac{n}{n+1} &= 1 - \mathrm{e}^{-\lambda_i}, \\ \mathrm{e}^{-\lambda_i} &= \frac{1}{n+1}, \\ \widehat{\lambda}_3 &= \ln(n+1). \end{aligned}$$

Jenže $\widehat{\lambda}_1 = S/n$ je racionální číslo, a tedy nemůže být rovno $\widehat{\lambda}_3$ pro žádné n.

Tedy funkce $L(\lambda, N)$ a $L^*(\lambda, N)$ nemohou mít maximum ve stejném bodě a nejde je proto maximalizovat současně. Poznámka v článku [5] o současné maximalizace je tedy přinejmenším zavádějící.

Druhý postup. Nyní ukážeme jiný postup, kterým se dostaneme k odhadům uvedeným v článku [5]. Nejprve maximalizujeme funkci $L(\lambda, N)$ vzhledem k λ. Logaritmováním a dalšími úpravami dostaneme

$$\begin{aligned} \ln(L(\lambda, N)) = l(\lambda, N) &= -N\lambda + \sum_{x=1}^{R} \big(x n_x \ln\lambda - n_x \ln(x!) \big), \\ \frac{\partial l}{\partial \lambda}(\lambda, N) &= -N + \sum_{x=1}^{R} \frac{x n_x}{\lambda} = 0, \\ N &= \sum_{x=1}^{R} \frac{x n_x}{\lambda} = \frac{S}{\lambda}. \end{aligned} \tag{7}$$

Protože platí

$$\frac{\partial^2 l}{\partial \lambda^2}(\lambda, N) = -\frac{S}{\lambda^2} < 0, \quad \forall \lambda > 0,$$

je pro každé pevné $N \in \mathbb{N}$ hodnota $\lambda = S/N$ bodem maxima funkce $L(\cdot, N)$.

Postupujme nyní analogicky pro funkci L^*:

$$\begin{aligned} \ln(L^*(\lambda, N)) = l^*(\lambda, N) &= \ln \binom{N}{n} + n \ln(1 - \mathrm{e}^{-\lambda}) - (N - n)\lambda, \\ \frac{\partial l^*}{\partial \lambda}(\lambda, N) &= \frac{n \mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}} - N + n = 0, \\ \frac{n}{1 - \mathrm{e}^{-\lambda}} &= N. \end{aligned} \tag{8}$$

Platí

$$\frac{\partial^2 l^*}{\partial \lambda^2}(\lambda, N) = \frac{-n\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}) - n\mathrm{e}^{-2\lambda}}{(1-\mathrm{e}^{-\lambda})^2} = \frac{-n\mathrm{e}^{-\lambda}}{(1-\mathrm{e}^{-\lambda})^2} < 0, \quad \forall \lambda > 0,$$

a tedy pro pevné $N \in \mathbb{N}$ je hodnota λ splňující (8) bodem maxima funkce $L^*(\cdot, N)$.

Dosazením vztahu (7) do (8) dostáváme

$$\frac{n}{1-\mathrm{e}^{-\lambda}} = \frac{S}{\lambda},$$
$$\frac{\widehat{\lambda}}{1-\mathrm{e}^{-\widehat{\lambda}}} = \frac{S}{n}.$$

Hodnotu $\widehat{\lambda}$ splňující předchozí rovnici pak dosadíme do (7) a dostaneme hodnotu $\widehat{N}$. Dopracovali jsme se tedy k odhadům uvedeným v článku [5]. Hodnota $\widehat{N}$ je tedy volena tak, aby se maxima funkcí $L(\cdot, \widehat{N})$ a $L^*(\cdot, \widehat{N})$ nabývala ve stejném bodě $\widehat{\lambda}$.

4. Novější článek

V této části podrobně odvodíme postup z novějšího článku [1]. Vychází z věrohodnostních funkcí

$$L_1(\lambda, N) = \binom{N}{n} p^n q^{N-n}, \tag{9}$$

$$L_2(\lambda) = \left(\frac{q}{p}\right)^n \prod_{x=1}^{R} \left[\frac{\lambda^x}{x!}\right]^{n_x}, \tag{10}$$

kde opět $p = 1 - \mathrm{e}^{-\lambda}$, $q = 1 - p = \mathrm{e}^{-\lambda}$. Je vidět, že funkce (9) odpovídá přesně funkci (2) ze staršího článku. Odlišnost je ovšem mezi funkcemi (10) a (1).

Odvození věrohodnostních funkcí

Funkce (9) se shoduje s funkcí (2) a opět udává pravděpodobnost, že v úplném náhodném výběru o rozsahu N je n nenulových pozorování a $N-n$ nulových pozorování. Funkce (10) odpovídá věrohodnostní funkci náhodného výběru $X_1, \ldots, X_n$ o rozsahu n z useknutého Poissonova rozdělení:

$$L_2(\lambda) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda}}{1-\mathrm{e}^{-\lambda}} \frac{\lambda^{X_i}}{X_i!} = \frac{q^n}{p^n} \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!} = \left(\frac{q}{p}\right)^n \prod_{x=1}^{R} \left[\frac{\lambda^x}{x!}\right]^{n_x}.$$

Všimněme si, že rovnice (10) nezávisí na N. Proto z této jedné rovnice můžeme získat odhad $\widehat{\lambda}$, dosadit jej do (9) a odtud spočítat odhad $\widehat{N}$. Odhad $\widehat{N}$ je tedy získán metodou podmíněné maximální věrohodnosti [1, s. 182].

Odhady parametrů

Odhady $\widehat{\lambda}$ a $\widehat{N}$ mají podle článku maximalizovat funkce (9) a (10) současně. Autoři pak bez odvození uvádějí, že výsledné odhady splňují

$$\frac{\widehat{\lambda}}{1-\mathrm{e}^{-\widehat{\lambda}}} = \frac{S}{n}, \tag{11}$$

$$\widehat{N}_2 = \left\lfloor \frac{n}{1-\mathrm{e}^{-\widehat{\lambda}}} \right\rfloor. \tag{12}$$

Odvození tvaru odhadů z novějšího článku

Článek [1] uvádí, že se odhad $\widehat{\lambda}$ získá maximalizací $L_2(\lambda)$ a odhad $\widehat{N}$ pak následnou maximalizací $L_1(\widehat{\lambda}, N)$. Je tedy

$$l_2(\lambda) = \ln(L_2(\lambda)) = n \ln\left(\frac{q}{p}\right) + \sum_{x=1}^{R} x n_x \ln\lambda - \sum_{x=1}^{R} n_x \ln(x!),$$

$$l_2(\lambda) = n \ln(\mathrm{e}^{-\lambda}) - n \ln(1-\mathrm{e}^{-\lambda}) + S \ln\lambda - \sum_{x=1}^{R} n_x \ln(x!),$$

$$l_2(\lambda) = -n\lambda - n \ln(1-\mathrm{e}^{-\lambda}) + S \ln\lambda - \sum_{x=1}^{R} n_x \ln(x!),$$

$$\frac{\partial l_2}{\partial \lambda}(\lambda) = -n + \frac{-n\mathrm{e}^{-\lambda}}{1-\mathrm{e}^{-\lambda}} + \frac{S}{\lambda},$$

$$\frac{\partial l_2}{\partial \lambda}(\lambda) = \frac{-n}{1-\mathrm{e}^{-\lambda}} + \frac{S}{\lambda} = 0. \tag{13}$$

Hodnota λ splňující předchozí vztah je hledaným odhadem a splňuje také, po jednoduché úpravě, vztah

$$\frac{\widehat{\lambda}}{1-\mathrm{e}^{-\widehat{\lambda}}} = \frac{S}{n}. \tag{14}$$

Dále, funkce (13) je kladná, resp. záporná, pokud je funkce $\frac{S}{n} - \frac{\lambda}{1-\mathrm{e}^{-\lambda}}$ kladná, resp. záporná. Protože funkce $\frac{\lambda}{1-\mathrm{e}^{-\lambda}}$ je rostoucí, je

$$\frac{S}{n} - \frac{\lambda}{1-\mathrm{e}^{-\lambda}} > 0 \text{ pro } \lambda < \widehat{\lambda},$$
$$\frac{S}{n} - \frac{\lambda}{1-\mathrm{e}^{-\lambda}} < 0 \text{ pro } \lambda > \widehat{\lambda}.$$

V bodě $\widehat{\lambda}$ tedy nabývá funkce $L_2(\lambda)$ maxima. Zbývá ještě ukázat, kdy vůbec existuje nějaké $\widehat{\lambda} > 0$, splňující (14). Platí

$$\lim_{\lambda \to 0+} \frac{\lambda}{1-\mathrm{e}^{-\lambda}} = 1,$$
$$\lim_{\lambda \to \infty} \frac{\lambda}{1-\mathrm{e}^{-\lambda}} = +\infty,$$

a proto pro $S > n$ hledané $\widehat{\lambda}$ vždy existuje.

Použitím lemmatu 1 na $L_1(\lambda, N)$ nakonec dostáváme, že pro $\lambda = \widehat{\lambda}$ je maximum funkce $L_1(\widehat{\lambda}, N)$ v bodě

$$\widehat{N} = \left\lfloor \frac{n}{\widehat{p}} \right\rfloor = \left\lfloor \frac{n}{1-\mathrm{e}^{-\widehat{\lambda}_c}} \right\rfloor = \left\lfloor \frac{\sum_{x=1}^{R} x n_x}{\widehat{\lambda}_c} \right\rfloor.$$

Poznamenejme ještě, že v případě, kdy $\frac{n}{\widehat{p}}$ je přirozené číslo, odhad $\widehat{N}$ není dle lemmatu 1 jednoznačný, protože $L_1(\lambda, N)$ nemá jednoznačné maximum. V takovém případě i $\widehat{N} - 1$ je maximálně věrohodným odhadem N.

Odhad $\widehat{\lambda}$ lze získat z rovnice (14) numerickou metodou, případně podle článku [6] existuje analytické vyjádření odvozené pomocí Lagrangeových řad:

$$\widehat{\lambda} = \overline{X}^* - \sum_{j=1}^{\infty} \frac{j^{j-1}}{j!} \left(\overline{X}^* \mathrm{e}^{-\overline{X}^*}\right)^j,$$

kde $\overline{X}^* = \frac{S}{n}$.

5. Ilustrační příklad

Uveďme pro zajímavost ještě příklad převzatý z [4], který jej převzal z [2].

V pruské armádě se v letech 1875–1894 sbírala data o počtu vojáků, kteří zemřeli na následky kopnutí koněm. Data se sbírala v 10 různých vojenských

x	n_x
0	109
1	65
2	22
3	3
4	1

Tabulka 1: Nahlášené četnosti úmrtí – kopnutí koněm v pruské armádě.

jednotkách po dobu 20 let. Celkem bylo nahlášeno $S = 122$ smrtí v $N = 200$ výročních zprávách. Tabulka 1 obsahuje nahlášené četnosti úmrtí v jedné jednotce za rok.

Předpokládejme, že počet vojáků zemřelých na následky kopnutí koněm má Poissonovo rozdělení. My se budeme tvářit, že neznáme celkový počet výročních zpráv N a zkusíme tuto hodnotu odhadnout pomocí postupu z novějšího článku [1]. Dostáváme pak hodnoty $\widehat{\lambda} = 0{,}618$ a $\widehat{N} = 197$. Odhad N je tedy velmi blízký skutečné hodnotě 200. Ještě doplníme, že maximálně věrohodný odhad λ z úplného výběru je pak $S/N = 0{,}610$, tedy i odhad $\widehat{\lambda}$ je uspokojivý.

Ještě můžeme doplnit odhady parametrů pro příklad zmíněný v úvodu (výskyt cholery v domácnostech v indické vesnici). Postupem podle novějšího článku získáme odhady $\widehat{\lambda} = 0{,}972$ a $\widehat{N} = 88$. Odhadujeme tedy, že cholerou bylo zasaženo dalších $\widehat{N} - n = 88 - 55 = 33$ domácností, ve kterých zatím nikdo neměl příznaky.

6. Závěr

Naše detektivní pátrání ukázalo, že mezi postupy v článcích [1] a [5] není žádný vztah, přestože autoři v textu tvrdí, že jsou postupy shodné. Články vychází z mírně odlišných věrohodnostních rovnic a dospívají k mírně odlišným odhadům parametru N. Postup ze staršího článku [5] není z našeho pohledu zcela korektní, protože nedochází k žádné maximalizaci v proměnné N, odhad $\widehat{N}$ je volen tak, aby obě věrohodnostní funkce jako funkce λ nabývaly maxima ve stejném bodě. Postup z novějšího článku [1] je přímočařejší a přirozeně vede k tomu, že odhad $\widehat{N}$ je přirozené číslo. Z těchto dvou možností tedy doporučujeme využít postup podle novějšího článku [1].

Literatura

[1] Blumenthal, S., Dahiya, R. C., Gross, A. J. (1978): Estimating the complete sample size from an incomplete Poisson sample. *J. Amer. Statist. Assoc.* **73**, 182–187. *cit. 4, 11, 12 a 14*

[2] Bortkiewicz, L. (1898): *Das Gesetz der Kleinen Zahlen.* B. G. Teubner, Leipzig, 1898. *cit. 13*

[3] Chapman, D. G. (1951): Some properties of the hypergeometric distribution with applications to zoölogical sample censuses. *Univ. California Publ. Statist.* **1**, 131–159. *cit. 6*

[4] Cohen, A. C. (1960): Estimating the parameter in a conditional Poisson distribution. *Biometrics* **16**, 203–211. *cit. 13*

[5] Dahiya, R. C., Gross, A. J. (1973): Estimating the zero class from a truncated Poisson sample. *J. Amer. Statist. Assoc.* **68**, 731–733. *cit. 4, 5, 8, 10, 11 a 14*

[6] Irwin, J. O. (1959): 138. Note: On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. *Biometrics* **15**, 324–326. *cit. 13*

[7] Sanathanan, L. (1972): Estimating the size of a multinomial population. *Ann. Math. Statist.* **43**, 142–152. *cit. 6*

[8] Zeman, O. (2018): *Neúplné vzorky z Poissonova rozdělení.* Bakalářská práce, MFF UK, Praha, 2018. *cit. 4*

EVROPSKÁ STATISTICKÁ AKREDITACE FENSTATS
FENSTATS ACCREDITATION

Gejza Dohnal

E-mail: gejza.dohnal@csc-sro.cz

1. Statistické akreditace

V roce 2019 byl Federací evropských národních statistických společností (FENStatS), jíž je Česká statistická společnost členem, odsouhlasen vznik tzv. evropských statistických akreditací. O jejich zavedení v rámci ČR ještě není rozhodnuto. Informace o těchto akreditacích naleznete v tomto článku. Budeme rádi za případné ohlasy.

Ve světě existuje několik zavedených systémů pro udělování takzvané *akreditace statistiků*. Akreditační programy mají zavedeny Statistická společnost Austrálie (SSA)[1], Statistická společnost Kanady (SSC)[2], britská Královská statistická společnost (RSS)[3] a Americká statistická asociace (ASA)[4].

Statistická akreditace je pověření založené spíše na portfoliu (na dokumentaci poskytnuté žadatelem) než na základě zkoušek a je periodicky obnovované (zpravidla každých pět let). Cílem statistické akreditace je poskytnout záruku zaměstnavatelům, dodavatelům a spolupracovníkům statistiků, že akreditovaná osoba je kvalifikovaným odborníkem. Akreditovaný statistik byl svými vrstevníky uznán jako profesionál, kombinující vzdělávání, zkušenosti, kompetence a závazek k etice na standardizované úrovni, která je garantována profesním sdružením. Akreditace je dobrovolná; žadatelé žádají o akreditaci, protože se domnívají, že pověření pro ně stojí za to, ale není to požadavek na praxi. Akreditace je uznáním profesní organizace, že akreditovaná osoba má dostatečné zkušenosti a jiné zásluhy, které všechny dohromady naznačují, že může pracovat jako statistik nezávislým a profesionálním způsobem.

[1] Od roku 1997 nabízí SSA dvě úrovně profesionální akreditace: postgraduálního statistika (G.Stat) a akreditovaného statistika (A.Stat). Viz https://www.statsoc.org.au/Professional-accreditation

[2] SSC nabízí od roku 2004 dvě úrovně akreditace: profesionálního statistika (P.Stat.) a přidruženého statistika (A.Stat.). Viz https://ssc.ca/en/accreditation/regulations-accreditation

[3] Viz webová stránka https://rss.org.uk/membership/professional-development/accreditation-scheme/

[4] Viz https://online.stat.psu.edu/statprogram/ethics/accreditation

Aby byl akreditační systém funkční, musí

- být průhledný, jasný, objektivní a nediskriminační,
- mít jasně daná kritéria pro udělení akreditace, která jsou založena na obecně uznávaných vědeckých zásadách a obecných kritériích zásluh a která vycházejí z potřeby zachování důvěryhodnosti a užitečnosti práce statistiků ve společnosti,
- nesmí být omezen na určité oblasti aplikací, metod nebo softwaru,
- nesmí se používat jako metoda k jiným politickým, náboženským nebo obchodním účelům, než k podpoře vědecky podporovaných znalostí a skutečností,
- být konstruován tak, aby bylo dosaženo vzájemného uznávání mezi jinými podobnými systémy.

Federace evropských národních statistických společností (FENStatS) se v minulých letech rozhodla poskytnout evropským statistikům možnost akreditace, srovnatelnou s již existujícími akreditačními systémy a mající širší než národní platnost.

2. Akreditační systém FENStatS

Akreditační systém FENStatS[5] je modelován podle programů v USA, Austrálii, Kanadě a Velké Británii a jeho cílem je poskytnout společný evropský standard pro definici statistického povolání.

Akreditační systém sdružení FENStatS je budován v těsné spolupráci s národními statistickými společnotmi (NS), které jsou jeho členy. FENStatS a tyto NS budou sdílet společnou odpovědnost za akreditaci, kde FENStatS stanoví a vlastní společný standard a zástupci NS mají na starosti operativní implementaci (jsou odpovědní za schvalovací proces svých členů). FENStatS je v tomto systému zastřešující organizací a nepřijímá právní odpovědnost za jednání akreditovaných Evropských statistiků.

FENStatS zřizuje mezinárodní Accreditation Committee (AC), který se bude zabývat operativními záležitostmi jménem výkonného výboru. AC je jmenován výkonným výborem na základě návrhů NS a je ve své většině složen z akreditovaných statistiků. AC bude každoročně, bude-li to považováno za nutné, přezkoumávat opatření prováděná NS, přijímat rozhodnutí o zrušení akreditace a povede dokumentaci akreditačního systému a záznamy akreditovaných členů.

[5] Viz `https://www.fenstats.eu/accreditation`

NS jmenují svého zástupce ve FENStatS pro oblast akreditace. Současně jmenují dostatečný počet auditorů z řad akreditovaných statistiků (nejméně tři na jednu NS) k přezkoumání žádostí svých členů. V případě potřeby si mohou NS zřídit příslušnou lokální institucionální organizaci pověřenou provozováním akreditačního systému.

2.1. Postup zapojení NS do akreditačního systému FENStatS

1. NS oznámí výkonnému výboru FENStatS, že zamýšlí nabídnout svým členům akreditaci (oznámení se zasílá výkonnému výboru FENStatS podle pokynů uvedených samostatně).
2. NS, které se účastní systému akreditace, jsou odpovědné za schvalovací proces svých členů.
3. NS v oznámení souhlasí s dodržováním akreditačních stanov a dalších pokynů vydaných na jejich základě a nadcházejících změn.
4. NS jako celek může svou účast kdykoli odvolat oznámením Výkonnému výboru FENStatS. Členové, kteří jsou již akreditováni odvolávající NS, budou sdružením FENStatS stále uznáváni jako akreditovaní statistici. Žádosti, které jsou v té chvíli předmětem přezkumu, budou finalizující odstupující NS.
5. NS, která úmyslně zneužije, podnikne kroky namířené proti účelu systému nebo jeho částí, může být účast odvolaná Výkonným výborem FENStatS.

2.2. Postup žadatelů o akreditaci

1. O akreditaci může požádat statistik, který je členem NS. Žádost je předložena jeho NS.
2. Žádost by měla mít sounáležitosti uvedené dále (viz Kritéria pro akreditaci).
3. Podáním žádosti žadatel souhlasí s pravidly a podmínkami stanovenými v akreditačních stanovách, etických pravidlech a jakýchkoli jiných dokumentech vydaných na jejich základě.
4. Žadateli se při podání žádosti účtuje poplatek NS.
5. Žádost přijímá nebo zamítá lokální institucionální orgán zřízený NS. Žádost by měla obsahovat jako přílohu příslušné doklady či prohlášení o vzdělání, praxi apod.
6. Rozhodnutí NS jsou předána AC FENStatS.

7. AC vydává doklad o akreditaci (elektronicky).
8. Akreditovaní členové platí NS roční poplatek.
9. Akreditace je nepřenosná a je platná po dobu 5 let.

Akreditovaný statistik bude moci používat označení

„Accredited European Statistician" (AES).

Akreditovaný evropský statistik může nechat svou akreditaci odvolat. Akreditovaný evropský statistik, který již není členem NS, nebude mít nadále status AES.

2.3. Zrušení akreditace

Akreditace může být zrušena akreditovanému evropskému statistikovi, který

- porušuje statut akreditace, jakýkoli zákon či etický standard, ke kterému je vázán,
- porušuje vědecké zásady nebo jiným způsobem poškozuje jméno statistice a statistikům obecně,
- je vyloučen ze své členské organizace.

Rozhodnutí o zrušení rozhodne výkonný výbor FENStatS na žádost AC. Rozhodnutí nebude přijato, dokud NS a členovi nebude poskytnuta přiměřená lhůta na odpověď. Rozhodnutí výkonného výboru je konečné.

Výkonný výbor FENStatS může po projednání s AC podepsat vzájemné uznávání akreditace se zahraničními organizacemi, které poskytují systémy akreditace rovnocenné systému FENStatS. Na druhou stranu statistici, kteří splňují požadavky EAS a jsou akreditováni vzájemně uznávanou zahraniční organizací, mohou požádat AC o akreditaci předložením svých zahraničních kreditů.

V případě, že NS již používá vlastní systém akreditace, mohou její členové nechat posoudit shodu tohoto systému s FENStatS. Pokud bude systém shledán srovnatelným, budou moci jeho členové převést své akreditace.

Náklady na provoz akreditačního systému ponesou primárně akreditovaní statistici na neziskovém základě. NS budou poskytovat část vybraných poplatků FENStatS na provoz systému.

2.4. Kritéria pro akreditaci

Pro posouzení žádosti budou použita následující kritéria (a doklady):

A. **Vysokoškolské vzdělání**
Vysokoškolský titul alespoň na úrovni MSc podle Boloňských kritérií (VŠ) – obor statistiky nebo ekvivalent.
Doklad: Vysokoškolský diplom, přepisy atd.

B. **Pracovní zkušenosti**
Nejméně pět let příslušné pracovní zkušenosti v oboru statistiky.
Doklad: životopis, doporučující dopisy.

C. **Odborný rozvoj**
Průběžný profesní rozvoj v příslušných oborech (nejen ve statistice) po ukončení VŠ.
Doklady: seznam (pokud není uveden v životopisu), osvědčení atd.

D. **Komunikační dovednosti**
Zdokumentované dovednosti v komunikaci statistických pojmů.
Doklady: příklady práce (3) a doporučení (mohou být stejné jako v B).

E. **Dodržování etických standardů**
Žadatel osvědčí znalosti příslušných etických norem a dodržuje je.
Doklady: pohovor, osobní prohlášení.

F. **Člen národní statistické společnosti**
Potvrzení zúčastněné národní společnosti.

AC může na doporučení auditorů udělit akreditaci žadateli, který splnil výše uvedená kriteria, ale nemá příslušnou dokumentaci. AC může na doporučení auditorů udělit akreditaci uchazeči, který má vyšší vysokoškolské vzdělání (než bakalář nebo podobně). Žádost

- se skládá z dokumentace zmíněné v bodech A–F a dopisu se žádostí shrnujícím opodstatnění žádosti,
- musí obsahovat podepsanou dohodu o tom, že žadatel přijímá pravidla akreditace a příslušné normy, zejména kritéria E (etické standardy).

Pokud není výslovně uvedeno jinak, potom

- bude s obsahem žádosti nakládáno jako s důvěrnými a bude po celou dobu akreditace uloženo,
- formulář žádosti, jméno, kontaktní informace a rozhodnutí budou zveřejněny.

FENStatS připravuje vlastní internetový portál, který bude sloužit pro akreditační proces. Žadatel si bude moci vytvořit vlastní registraci, v rámci které vytvoří elektronickou přihlášku, k níž přiloží požadované dokumenty v elektronické podobě. Na stránkách FENStatS (`https://www.fenstats.eu`) potom bude uveřejněn seznam již akreditovaných evropských statistiků.

NOVINKY ZE SVĚTA R+KORONAVIRU
NEWS FROM THE WORLD OF R+CORONAVIRUS

Pavel Stříž

E-mail: pavel@striz.cz

The R Foundation Retweeted: Peter Dalgaard. R 4.0.0 "Arbor Day" (source version) has been released.

24. 4. 2020 mi přistála na stole tato zpráva a za pár dní na to, 28. 4. 2020, došlo k aktualizaci Bioconductoru na verzi 3.11. Je Svátek práce, jdu ty nové `lazyLoad`, `lazyEval` a `lazyData` v R vyzkoušet a sdělit svůj nezávislý pohled.

1. R v4.0.0

Bez otálení jsem na svém Xubuntu 20.04 do `/etc/apt/sources.list` přidal:

```
deb https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/
```

Zakomentoval jsem starší pokusy a provedl preventivní kroky:

```
$ sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys
    E298A3A825C0D65DFD57CBB651716619E084DAB9
$ sudo apt update
$ sudo apt upgrade
```

Jádro aktualizace pak tvořil příkaz:

```
$ sudo apt install r-base r-base-dev
```

Rychlý test prokázal, že se podařilo a provedl jsem aktualizaci knihoven:

```
$ R --version
$ R
> update.packages(.libPaths())
```

2. COVID v19

The R Foundation Retweeted: R Consortium has started new GitHub repository to centralize collaboration and data sources – looking to develop COVID-19 tools and code – Come add your information and contribute to the community! `https://bddy.me/3aAXOmb`

Z toho samého dne zaujala mou pozornost ještě tato zpráva. Řekl jsem si, že bych měl nově nainstalované R hlouběji vyzkoušet.

Na úvodní stránce na mne vyskočily 4 projekty: Coronavirus Tracker, COVID-19 Propagation, COVID-19 Tracker Map a COVID-19 Projections.

Vezmu-li to od konce. U Projections na mne vyskočil GitHub. Po určitém bádání se mi podařilo otevřít rozcestník a webovou stránku pro Českou republiku, `volzinnovation.com/covid-19_SARS-CoV-2_corona/reports/latest/Czechia.html`. Autorem je Raphael Volz.

Autorem Tracker Map je Jay Ulfelder. V RStudiu Cloud mne hned upozornili, že se jedná o dočasný projekt. Pod Shiny app na mne vedle analýz vyškočily pdf v záložce WHO Situation Reports. Zajímavý nápad.

Druhý projekt v pořadí je Propagation. Autorem je Juan Francisco Venegas Gutiérrez. Bohužel repozitář na GitHubu mi nešel otevřít, tak jsem to zahlásil (Issues). Mezi modely jsem Českou republiku nenašel.

3. Coronavirus Tracker v0.1.4

První projekt v pořadí od Johna Coeneho Coronavirus Tracker vyzývá ke spuštění R. Pustil jsem se do toho. RStudio cloud občas jel bez přístupových práv, požadavek na knihovny `shinyMobile` a `echarts4r` zněl zajímavě.

```
$ R
> install.packages("coronavirus")
```

Zkusil jsem z dokumentace první ukázku a ještě si vyžádal knihovnu `dplyr`. Proč si ji nenainstaloval sám?

```
> install.packages("dplyr")
> library(coronavirus)
> require(dplyr)
> coronavirus %>% filter(type=="confirmed") %>% group_by(Country.Region) %>%
    summarise(total=sum(cases)) %>% arrange(-total) %>% head(20)
# A tibble: 20 x 2
   Country.Region       total
   <chr>                <int>
 1 Mainland China       70446
 2 Others                 355
 3 Singapore               75
```

Tady jsem zbystřil. To jsou stará data v textové formě, nikoliv hezké mapy přes web s aktuálními daty. Ve slangové řeči: rtfm! To, co jsem právě nainstaloval, je knihovna `https://github.com/RamiKrispin/coronavirus`, která má stejný název. Mezi Issues jsem autorovi Trackeru zahlásil, že jeho název je v konfliktu s existující knihovnou z února 2020.

Pokračoval jsem v experimentování, knihovnu jsem odinstaloval a podíval se na návod, `https://coronavirus.john-coene.com`.

```
> remove.packages("coronavirus")
> install.packages("remotes")
> remotes::install_github("Johncoene/coronavirus")
```

Během instalace mi naskočila tato neobvyklá zpráva:

```
- Use ‘usethis::browse_github_pat()‘ to create a Personal Access Token.
- Use ‘usethis::edit_r_environ()‘ and add the token as ‘GITHUB_PAT‘.
```

Knihovna `usethis` se teprve instalovala. V každém případě zmínka o `Rate limit reset at: [...]`, kdy došlo k restartu za několik minut pro mne znamenalo chvíli počkat a instalaci zopakovat.

Před koncem instalace si R vyžádalo systémový balík `libpq-dev` ve starší verzi 10.3-1 (aktuální je 10.12-0). Udělal jsem hrubý krok, doporučuji čtenářům najít lepší řešení přes Docker či pečlivě projít, co se bude odinstalovávat.

```
$ sudo apt install aptitude
$ sudo aptitude install libpq-dev
```

Na první dotaz jsem dal nikoliv (`n`; starší balík by se nenainstaloval), na druhou nabídku ano (`y`). Zopakoval jsem instalaci v R a skončila úspěšně.

Zkusil jsem třířádkovou ukázku dle instalačního manuálu, `coronavirus.john-coene.com` (nutno zalistovat na webové stránce níž).

```
> library(coronavirus)
> virus<-crawl_coronavirus()
i Crawling data from John Hopkins
i Crawling data from Weixin
i Crawling data from DXY
> run_app(virus)
```

Pozn. Pokud bychom se dostali do konfliktu u příkazů, užijme:

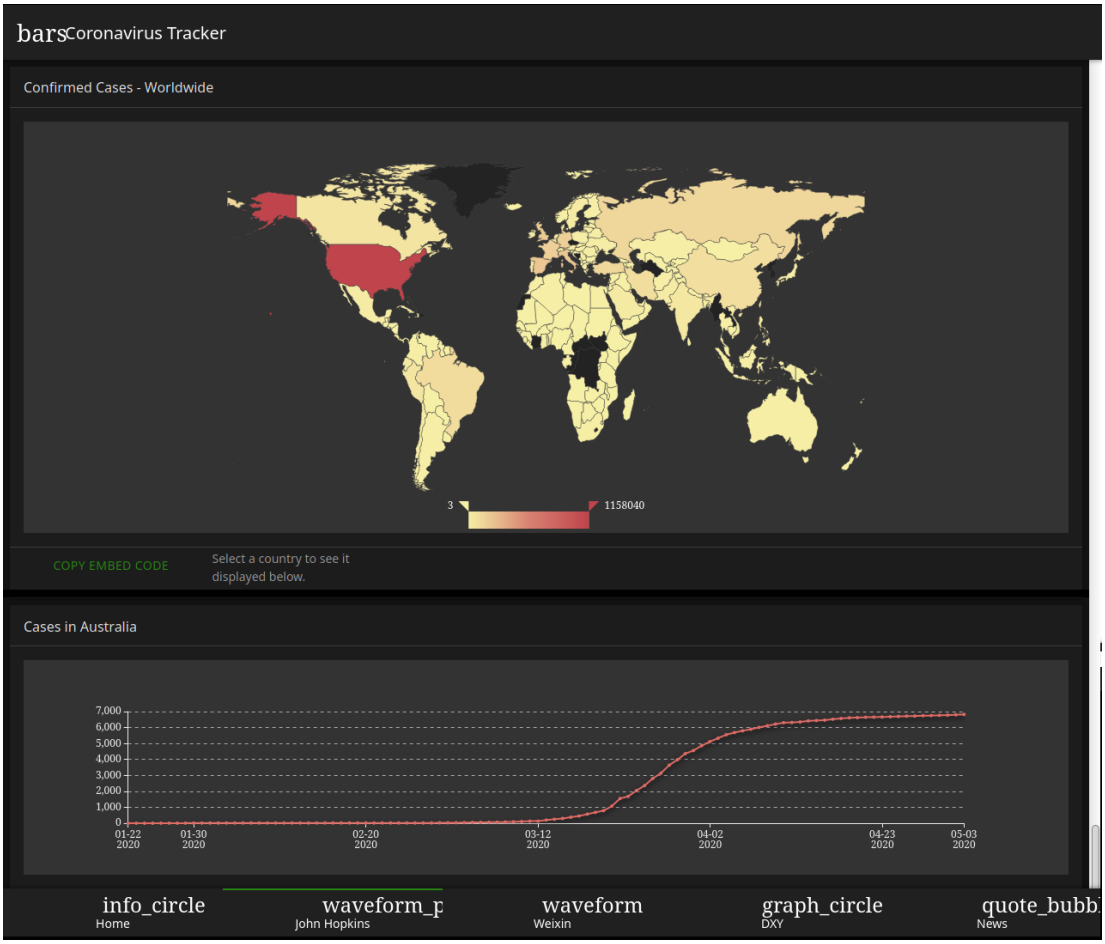
```
> virus <- coronavirus::crawl_coronavirus()
> coronavirus::run_app(virus)
```

Ve webovém prohlížeči se mi otevřela vygenerovaná stránka, pokaždé na jiném portu. Radost byla veliká! Za pozornost stojí, že má být Johns Hopkins, to již někdo zahlásil autorovi k opravení.

Ve spodní části je pět záložek. Na druhé (waveform_path) v bloku *China* a *World* a čtvrté (graph_circle) v bloku *Cities* jako kdyby něco chybělo. Po nakliknutí do bloku se otevře detailní výpis. Vrátit zpět se dá přes značku xmark_circle_fill v pravém horním rohu. Design je trochu nezvyklý, ale musíme mít na paměti, že je to zaměřené na mobilní telefony a já to zkoušel na notebooku.

Detaily kolem dat je možné nalézt v levém horním rohu pod bars nebo menu. Z R lze server zavřít přes klávesovou kombinaci Ctrl+C.

Nyní dokumentace radí si nastavit `crontab` atd. Co mne zaujalo u stažení dat z DXY je, že se občas nezadařilo připojit. V rychlosti jsem nahlédl na server `https://education.rstudio.com/`, konkrétně na `dataio`.



Jakmile se podařilo na servery připojit, mohl jsem si proměnnou `virus` uložit a opětovně užívat. Rychlá pomůcka u experimentů bez nutnosti aktualizace dat.

```
> save(virus, file="virus.RData")
> load("virus.RData")
```

4. Hašovací klíč od `newsapi.org` v2

Mou pozornost zaujala poslední záložka se zprávou `No newsapi token.` To bych rád poléčil. Autor v dokumentaci radí:

```
> library(coronavirus)
> create_config()
```

V pozadí se ze šablony vytvoří soubor `_coronavirus.yml`, blok `database` je povinný, blok `newsapi` volitelný. To byl pro mne problém. Já jsem to chtěl obráceně. Nevadí.

Přes `https://newsapi.org/register` jsem se zaregistroval a získal hašovací klíč. Zahlédl jsem jejich novou knihovnu pro R `newsanchor`, my zůstaneme u autorem užité knihovny `newsapi`.

Zkusil jsem R podsunout hašovací klíč:

```
> library(newsapi)
> newsapi::newsapi_key("41e22e9efcf64b2a9354a796b99c43b8")
```

Ale ani touto cestou ani jinou přes editaci souboru `_coronavirus.yml` se mi to nepodařilo.

Prvně jsem nahlédl na zdrojové kódy v:

```
$ cd ~/R/x86_64-pc-linux-gnu-library/4.0/coronavirus
```

Narazil jsem hlavně na binární soubory rds, rdx a rdb. Nejsem expert, abych dokázal odpovědět, jestli by se soubory daly rozluštit a editovat.

Prozkoumal jsem zdrojové kódy přímo od autora:

```
$ git clone https://github.com/JohnCoene/coronavirus
```

Došel jsem k závěru, že bych musel zdrojové kódy upravit, zkompilovat atd. To je nad rámec této sváteční zprávy.

V souboru `coronavirus/inst/app/Dockerfile` jsem si ověřil, že skutečně knihovnu `newsapi` přebírá z GitHubu od uživatele `news-r`.

5. PostgreSQL v10+190

Říkal jsem si, když už se mi podařilo nainstalovat `libpq-dev`, dokáži i zbytek. Otevřel jsem komunitní tutoriál. Nainstaloval jsem PostgreSQL:

```
$ sudo apt install postgresql postgresql-contrib
```

A nejkratší možnou cestou jsem se pustil do dalších kroků. Vytvořil jsem v databázovém systému nového uživatele `testing` a novou databázi `testing`. Vynechávám krok vytvoření uživatele pod operačním systémem.

```
$ sudo -i -u postgres createuser --interactive
Enter name of role to add: testing
Shall the new role be a superuser? (y/n) y
$ sudo -u postgres createdb testing
```

Uživatel je bez hesla, to webové rozhraní nepřijme. Nastavil jsem nové heslo přes:

```
$ sudo -i -u postgres
$ psql
postgres=# ALTER USER testing WITH PASSWORD 'testing';
postgres=# \q
$ exit
```

Rychlokurz `psql`: `help` je základní nápověda, `\l` je výpis databází, `\c testing` je připojení k naší databázi, `\dt` je výpis tabulek, `\h` je seznam

SQL příkazů, `\?` je seznam příkazů `psql` a `\q` ukončí běh programu. Verzálky u příkazů netřeba psát.

Vše zrealizované jsem zaznačil v `_coronavirus.yml`:

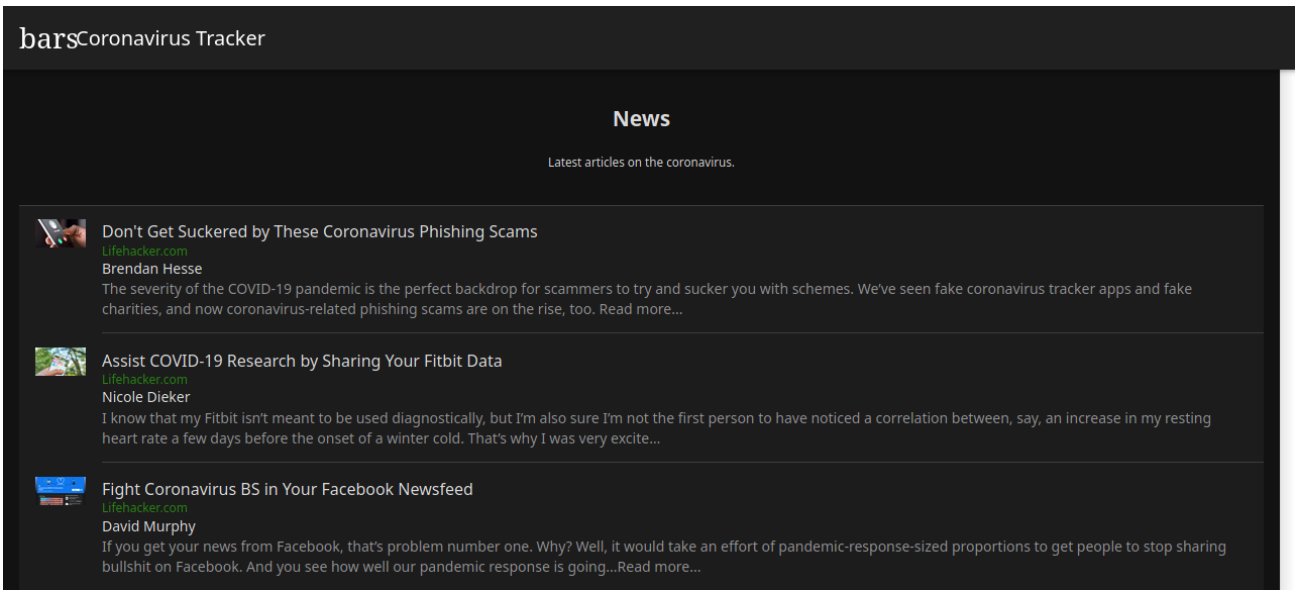
```
database:
  name: testing
  host: 127.0.0.1
  user: testing
  password: testing
newsapi:
  key: 41e22e9efcf64b2a9354a796b99c43b8
```

Když jsem opakoval tři řádky ukázkového spuštění v R, výpis se mi rozšířil o tyto dva řádky:

```
i Crawling news from newsapi.org
✔ Writing to database
```

V páté záložce mi vyběhly novinky, aktivovaný dotaz lze nalézt u autora v souboru `coronavirus/R/crawl.R`:

```
news <- newsapi::every_news("coronavirus OR covid", results = 100, language =
    "en", sort = "popularity")
```



Ověření funkčnosti můžeme zjistit i z tabulky `log`:

```
$ psql -h localhost -d testing -U testing
Password for user testing: testing
testing=# SELECT * FROM log;
```

Dostáváme přibližně takový výsledek:

```
         last_updated
------------------------------
 2020-05-01 18:19:24.547171+02
(1 row)
```

Dle chuti lze dál bádat u surových dat, např.:

```
testing=# SELECT * FROM jhu WHERE country='Czechia';
testing=# SELECT * FROM jhu WHERE country='Slovakia';
testing=# \q
```

6. Bioconductor v3.11

Před dalším krokem si obvykle nastavuji plná práva u těchto adresářů:

```
$ cd /usr/lib/R
$ sudo chmod -R 777 site-library/
$ sudo chmod -R 777 library/
$ cd /usr/share
$ sudo chmod -R 777 R/
```

Za zmínku stojí, že manažer knihoven `biocLite` ustupuje a roli nahrazuje `BiocManager`. V R lze otestovat:

```
> install.packages("BiocManager")
> library(BiocManager)
> BiocManager::install()
> BiocManager::available()
```

Můžeme ověřit instalaci knihoven:

```
> BiocManager::valid()
   [...] "coronavirus", "echarts4r", "shinyMobile" [...]
Warning message:
0 packages out-of-date; 3 packages too new
```

To souhlasí, neb `coronavirus` byl instalován z GitHubu, nikoliv z CRANu.

Pro badatele stojí za pozornost obrazy pro Docker a Amazon (Amazon Machine Image, AMI). Je zde možnost instalovat vývojářské knihovny:

```
> BiocManager::install(version="devel")
```

7. Řešení konfliktu názvu knihovny v R

Na chvíli se ještě vraťme k řešení konfliktu stejného názvu knihoven. Na Stackoverflow zmiňují v principu dvě cesty.

Stáhnout si zdrojové soubory a nic neměnit:

```
$ cd /tmp
$ wget https://cran.r-project.org/src/contrib/coronavirus_0.1.0.tar.gz
$ R CMD INSTALL -l /tmp coronavirus_0.1.0.tar.gz
```

V R si pak volat jeden z příkazů a vybrat tak chtěnou knihovnu:

```
> # library(coronavirus)
> library(coronavirus, lib.loc="/tmp")
```

Když jsem zkoušel paralelně spustit i instalovanou knihovnu z GitHubu `coronavirus` odkomentováním prvního řádku, tak to neběželo. Tuším, že se jedná o bezpečnostní pojistku.

Druhá cesta je zasáhnout do souboru `DESCRIPTION`.

```
$ tar xvf coronavirus_0.1.0.tar.gz
$ mv coronavirus coronavirusRami
$ cd coronavirusRami
$ nano DESCRIPTION
```

První řádek upravit například na `Package: coronavirusRami`.

Volitelně upravíme i `MD5`, konkrétně první řádek za výpis:

```
$ md5sum -b DESCRIPTION
324f8275940bfa7fde376934c57a28ae *DESCRIPTION
```

Chtělo by to přejmenovat i další soubory na `coronavirusRami`, u této školní ukázky vynechávám. Zabalil jsem si zpět a už podsunul R:

```
$ cd ..
$ tar cvf coronavirusRami.tar.gz coronavirusRami/
$ R CMD INSTALL coronavirusRami.tar.gz
```

Ověřit funkčnost můžeme už přímo v R a lze si spustit oba konfliktní balíčky paralelně:

```
> library(coronavirus)
> ?coronavirus::crawl_coronavirus
> library(coronavirusRami)
> ?coronavirusRami::coronavirus
```

8. Pár tipů místo Závěru

Podobně jako jsou v R knihovny setříděné podle kategorií, viz Zobrazení úloh (CRAN Task Views), lze nahlédnout u RStudia na R Views s klíčovým slovem `covid-19`. K dnešnímu dni tam jsou tři záznamy: Some Select COVID-19 Modeling Resources, Simulating COVID-19 interventions with R a COVID-19 epidemiology with R.

Kdo by dal přednost odpočinku od R a PostgreSQL, nechť nahlédne na aktuální stavy kolem koronaviru na `https://www.twitch.tv/killars`.

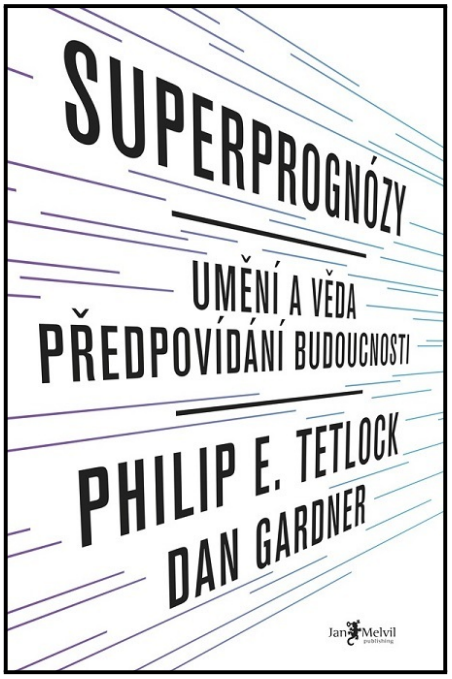
SUPERPROGNÓZY – UMĚNÍ A VĚDA PŘEDPOVÍDÁNÍ BUDOUCNOSTI

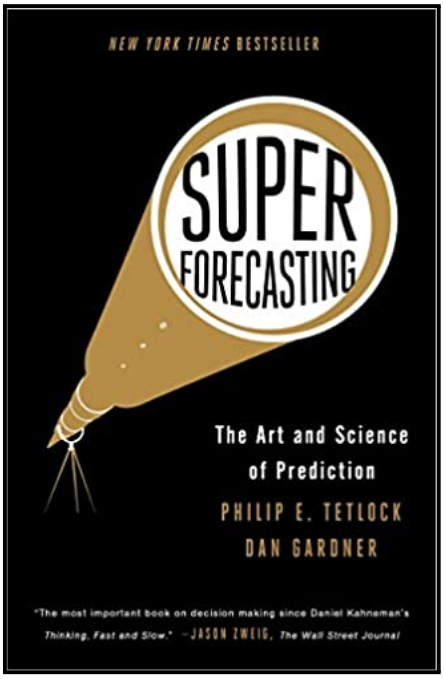
SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION

Jan Kalina

E-mail: kalina@cs.cas.cz

Z Knihovničky ČStS vybíráme recenzi. Philip E. Tetlock, Dan Gardner, vyšlo 2015, v češtině 2016 v nakladatelství Jan Melvil Publishing, 384 stran, recenzi napsal Jan Kalina (Ústav informatiky AV ČR) v srpnu 2020.





Kniha přináší velmi zajímavý vhled do prognózování politických událostí. Takové prognózy (jako např. zda v příštích 5 letech dojde k vojenskému převratu v Thajsku) jsou obtížné i spekulativní, čímž kniha od začátku získává nádech tajemna. Kritika běžně publikovaných prognóz mě rozesmála odvážnými tvrzeními: obvykle nezáleží na jejich přesnosti, představují vědomé lhaní, a i šimpanz by prognózoval stejně nepřesně jako tzv. experti (kap. 3). Kniha ale nenabízí formální definici prognostiky, rád bych zde zdůraznil (což autoři opomněli), že se kniha nevěnuje predikování na základě dostupných dat tak, jak ze statistiky známe, tj. nejde zde třeba o průzkumy volebních preferencí nebo o predikci sledovanosti televize na základě měření.

Ocenil jsem, že autoři konstatují, že dobré prognózování (a rozhodování obecně) je především otázka správné práce s pravděpodobnostmi. Zde přiznávám knize velkou zásluhu při popularizaci pravděpodobnostního uvažování široké veřejnosti, pro kterou je primárně určena; podle kap. 6 si většina (americké) veřejnosti nepřipouští vliv náhodných procesů na utváření dějin. Na druhou stranu vnímám určitý rozpor, když kniha na začátku slibuje, jak nás učiní chytřejšími a moudřejšími, a pak směřuje „jen“ k tomu, že promyšlené prognózy (superprognózy) vyžadují zjišťování **informací**, kritické uvažování a rozbor dané situace, odhad **pravděpodobností** jednotlivých variant, a vrcholnými triky pak jsou nejjednodušší verze Bayesova vzorce či regrese k průměru. Každopádně partie o rozdílech mezi analytickým a intuitivním uvažováním (kap. 2) ve mě vzbuzují zájem o myšlenky psychologa Daniela Kahnemana.

Kniha je čtivá, možná až na podrobné prognózy konkrétních amerických politických událostí, v nichž se autor projevil jako zkušený politolog. Řekl bych, že kniha by mohla směřovat k závěrům rychleji, autoři čtenáře napínají, přehled vlastností dobrých prognostiků ze str. 217 jsem čekal výrazně dřív. Z hlediska pravděpodobnosti a statistiky bych rád zmínil tyto myšlenky:

- Prognóza může být současně rozumná a špatná (kap. 4).
- Statistici nepůsobí na veřejnosti přesvědčivě, protože mluví o náhodě, neurčitosti a nejistotě, zatímco veřejnost ráda slyší sebejistou prezentaci stoprocentně platných závěrů (kap. 10).
- K lepším výsledkům vede, když se prognózy subjektivně „zextremizují“, například odhadnutou předpověď 70 % je podle knihy lepší mechanicky „natlačit“ na 85 % (tak doslovně na str. 109, viz též kap. 9).

V partiích, které kritizují žalostnou historii klinického rozhodování v medicíně a obdivují pokrok po zavedení randomizovaných experimentů, mi kniha mluví z duše. Medicínu založenou na důkazech (EBM) klade jako vzor i pro humanitní obory. Kniha ovšem mluví v kontextu EBM o důkazech, zatímco za vhodnější bych považoval překládat tyto vědecké důkazy do češtiny jako evidenci; vždyť jde o **statistickou** evidenci získanou opakovanými experimenty. O medicíně založené na důkazech se zájemci mohou víc dočíst v práci [1].

Literatura

[1] Kalina J. (2019): Mental health clinical decision support exploiting big data. In Chui K. T., Lytras M. D. (eds.): *Computational methods and algorithms for medicine and optimized clinical practice.* IGI Global, Hershey, 160 – 184. *cit. 30*

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.

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.