ROBUST'2002, 1 – 15

O SPOLEHLIVOSTI VÝVOJOVÝCH STROMŮ

MARTIN BETINEC

Abstract. Evolutionary tree is a dendrogram that is used to asses genetical similarity of biological objects. Contribution concerns two methods that use bootstrap to verify the confidence of the built tree. First one – suggested by Felsenstein – is relatively simple and is based on counting of occurences of agreeing parts of trees in a bootstrap sample to the original ones. Despite that it is reasonable by the arguments that will be presented. The second method can be considered as a refinement of the previous one reflecting more precisely the shape of the feature space of the model. Results are illustrated on a real data (malaria gene sequences).

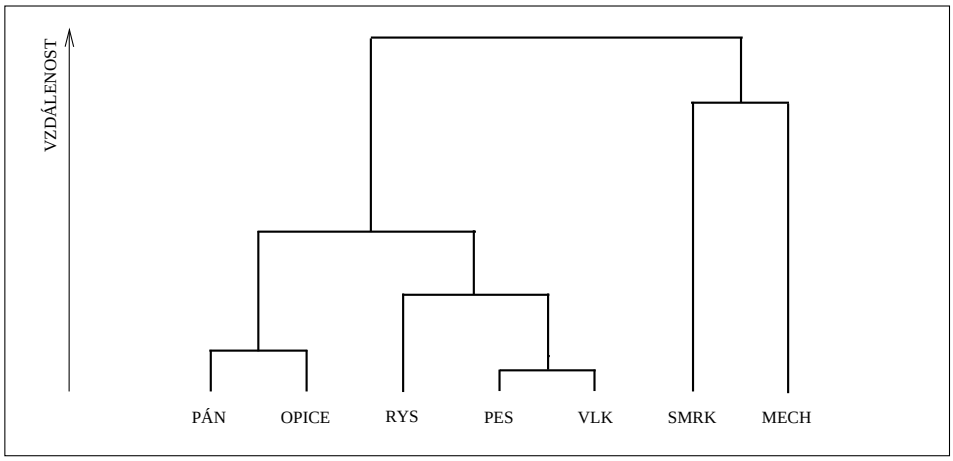
Резюме. Эволуционное дерево представляет дендрограмму, которая используется для достижений генетического совпадения биологических обектов. Результаты моей работы касаются двух методов использующих "bootstrap" для построения доверительной вероятности совпадения построенных деревь. Первый, предложенный Фелзенштейном, относительно прост и основан на подставе частей деревь, согласованных с оригинальным деревом в выборке генерированной при помощи "bootstrapa" . Приводятся аргументы в пользу этого метода. Второй метод может рассматриватся как уточнение предыдущего, более аккуратно учитывая форму элементов модели. Результаты иллустрированы используя реальные данные, точнее, последовательности генов маларии.

1. Úvod

Představme si na chvíli, že jsme zvídaví biologové, kteří si na prvního máje vyšli na procházku do Stromovky. Při kochání se pozorováním rozličných organismů vůkol nás pojednou zachvátí sžíravá potřeba zjistit, jak probíhala jejich evoluce. Co si v tuto chvíli počít? Někoho napadne, že organismy navzájem si vývojově příbuznější by si měly být podobnější i po stránce genetické. Počneme tedy odebírat vzorky tkání a extrahovat DNA. Jak však z těchto nukleotidových shluků něco smysluplného vyčíst? Zde již naše statistické já vstupuje do myšlenkového experimentu: „Nyní nastal čas pro shlukovou analýzu".
Upřesnění: *Nyní nastal čas pro vývojové stromy.*

Vývojové stromy, v anglické literatuře označované jako *Phylogenetic (Evolutionary) Trees*, představují způsob, jímž se v biologii popisuje evoluce sledovaných organismů. Z matematického hlediska se jedná o klasifikační stromy (dendrogramy). Tato metoda umožňuje nejen názorně vyjádřit „co s čím" je si v evoluci blízké, ale zároveň také kvantifikovat, jak je ta která skupina vzdálena od ostatních, a to pomocí výšky větve spojující dotyčné skupiny.

2000 *Mathematics Subject Classification.* Primary 62H30; Secondary 62P10.
Klíčová slova. Vývojové stromy, shlukování, klasifikace, bootstrap, DNA, malárie.
Tato práce vznikla za podpory grantů MSM 113200008 a GAČR č. 201/00/0769.



Obr. 1 *Evoluce ve Stromovce.*

Vraťme se k našemu příkladu; vypěstujeme-li vývojový strom, mohl by vypadat (dle mých přírodovědných znalostí) asi jako na obrázku 1. Otázkou zůstává, nakolik se na takto zvolený popis situace mohu spolehnout, tj. do jaké míry výsledný strom (tedy to, co spočtu z naměřených dat) popisuje skutečnou příbuznost sledovaných organismů.

V tomto článku představíme dvě metody využívající bootstrapu pro odhad výše zmíněné spolehlivosti a pokusíme se ukázat, proč lze právě ze vztahu mezi datovým souborem a „boot-populací“ tvrdit něco o souvisloti dat a reality.

Podotkněme jen, že metodikou pěstování evolučního stromu se podrobněji zabývat nebudeme, nýbrž k výpočtům použijeme některého z osvědčených a již dostatečně známých postupů shlukové analýzy.

Pro snadnější orientaci ve značení ještě poznamenejme, že metoda bootstrapu se dotýká tří oblastí reality. Jsou jimi **roviny:**

teoretická: týkající se odhadovaných vlastností populace, tj. neznámých parametrů;

praktická: odvozená od naměřených dat – odpovídající statistiky budeme označovat stříškou (odhad neznámého parametru μ budeme značit pomocí $\widehat{\mu}$);

virtuální: vztahující se k populacím vzniklým z bootstrapu – příslušné veličiny označíme hvězdičkami ($\widehat{\mu}^*$).

2. Metoda relativních četností

2.1. **Data.** Nejprve se budeme věnovat asi nejjednoduššímu postupu, jímž bychom mohli spolehlivost stromu odhadnout, který bývá někdy nazýván jako Felsensteinova metoda. Data, jichž využijeme pro názornější vysvětlení, nashromáždila skupina vědců (viz [4]) zkoumajících část genomu živočicha *Plasmodium falciparum*, parazita žijícího v krvi člověka, který způsobuje malárii. Data, jež jsou přístupná na adrese `www.ncbi.nlm.nih.gov`, tvoří kompletní sekvence genu kódujícího antigen AMA-1 o délce 1866 nukleotidů a pocházejí z pěti různých oblastí (Keňa, Indie, Thajsko a Venezuela – provincie Bolívar a Amazonas). K ilustraci použijeme 13 vzorků zmíněného genu, které pro zkracení označme počátečním písmenem oblasti, z níž pocházejí, a v ramci jedné oblasti je rozlišme čísly, tedy např. $B6$ označuje vzorek číslo 6 z provincie Bolívar, $T3$ třetí vzorek z Thajska apod., viz obrázek 2.

Pro $i = 1, \ldots, 13$ označme jednotlivé vzorky:

$$\boldsymbol{x}_i = \underbrace{\big(\mathrm{A,T,G,}\ \ldots,\ \mathrm{C,T,A,T}\big)}_{1866\ \text{nukleotidů}},$$

kde jsou jednotlivým nukleotidům přiřazena písmena: *A – adenin, T – thymin, C – cytosin, G – guanin.*

Důležitá vlastnost těchto vzorků spočívá v tom, že každý z nich začína přesně na začátku genu a končí na jeho konci, přičemž žádné nukleotidy mezi nimi nechybí, jinými slovy – složky vektorů $\boldsymbol{x}_i$ si navzájem odpovídají.

2.2. **Pěstování.** Z řádkových vektorů $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{13}$ sestavíme datovou matici

$$\boldsymbol{X}_{(13\times 1866)} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{13} \end{pmatrix}$$

Klíčovou otázkou pro určení vzdáleností mezi vektory matice $\boldsymbol{X}$ zůstává, krom volby metriky, kvantifikace nukleotidů. Zde je nutno uvážit jejich chemickou podobnost, pravděpodobnost bodových mutací, apod. Toto zakódování může zásadně ovlivnit podobu vývojového stromu. Já jsem nukleotidy překódoval následovně:

$$\big(\mathrm{A,G,C,T}\big) = \big(1, 2, 5, 6\big)\,. \tag{1}$$

Toto přiřazení zachovává příbuznost jak u purinů (A, G), tak i u pyrimidinů (C,G), zároveň odráží i různost obou skupin. Zde by jistě bylo možno vzdálit obě skupiny od sebe mnohem více, čímž bychom vypěstovali odlišný strom. Stejně tak lze u zvoleného kódování diskutovat o nesymetrii mezi dvojcemi komplementárních bazí (A,T), (G,C). Zde se otevírá možnost, pro konzultace s odborníky v oblasti genetiky a případné pokračování výzkumu. Faktem nicméně zůstává, že pro opačně zvolenou nesymetrii $(\mathrm{A,G,C,T}) = (2, 1, 6, 5)$, a stejně tak i pro symetrii $(\mathrm{A,G,C,T}) = (1, 2, 6, 5)$ vyrostly stromy velice podobné původnímu. Proto jsem v souladu s literaturou ([3]) zvolil výše zmíněný způsob.

Metriku jsem volil Eukleidovskou:

$$d_E(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{k=1}^{1866} \big(x_i^k - x_j^k\big)^2} \qquad i,\, j = 1, \ldots, 13\,, \tag{2}$$

kde x_i^k značí k -tou složku i -tého vzorku. Zřejmě lze vynechat konstatní sloupcové vektory matice $\boldsymbol{X}$, které nepřinášejí žádnou informaci o vzdálenosti řádků, neboť jim odpovídající sčítance jsou nulové. Tímto se sníží dimenze matice $\boldsymbol{X}$ z původních (13×1866) na (13×73).

Strom $\widehat{\Psi}$ vypěstovaný z $\boldsymbol{X}$ na základě principu nejbližšího souseda ukazuje obrázek 2.

2.3. **Spolehlivost.** Zřejmě nejjednodušší odhad hladiny spolehlivosti jednotlivych větví stromu $\widehat{\Psi}$ bychom získali jako relativní četnosti výskytu jeho větví ve stromech vzniklých prostřednictvím bootstrapu. Pro pevně zvolenou větev v bychom mohli postupovat např. dle následujícího **algoritmu:**

```
count := 0
FOR b = 1 TO B
```

(1) náhodným výběrem s vracením ze sloupců matice $\boldsymbol{X}$ vytvoříme boot-matici $\boldsymbol{X}_b^*$ stejné dimenze (zde předpokládáme, že každý ze sloupců matice $\boldsymbol{X}$ může být vybrán v každém kroku se stejnou pravděpodobností)

(2) z matice $\boldsymbol{X}_b^*$ vypěstuj strom $\widehat{\Psi}_b^*$

(3) `IF` $\widehat{\Psi}_b^* \ni v$ `THEN count:=count+1`

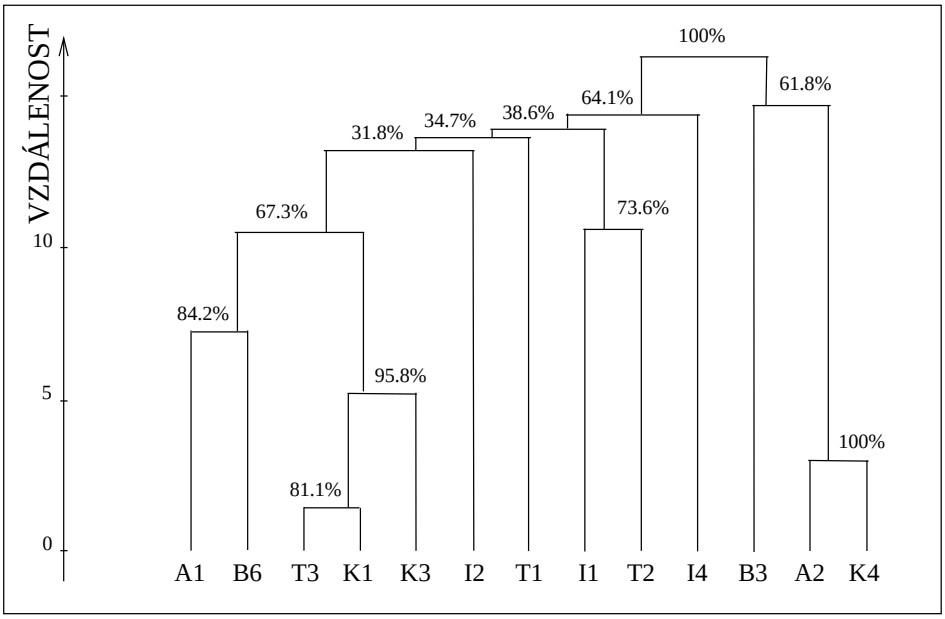
`ENDFOR`

Spolehlivostní hladinu větve v pak definujme jako:

$$\alpha_{_F}(v) \stackrel{df}{=} \frac{\texttt{count}}{B} = \frac{\#\{b : \widehat{\Psi}_b^* \ni v\}}{B}, \tag{3}$$

kde # značí „počet prvků“ a B je počet opakování při bootstrapu.

V praxi nás zpravidla zajímají spolehlivostní hladiny pro více než jen jednu větev. Přizpůsobení algoritmu je v tomto případě nasnadě.



Obr. 2 *polehlivostní hladiny vývojového stromu bakterie Plasmodium falciparum. Jsou nadepsány příslušným větvím.*

V našem příkladu (kde jsem zvolil $B = 5000$) se větev skládající se právě jen ze vzorků označených jako $T3$ a $K1$ (viz obrázek 2) vyskytla ve 4055 případech. Odhad hladiny spolehlivosti pro zmíněnou větev tedy bude $\alpha_{_F} = 4055/5000 = 0.811$. U původního stromu $\widehat{\Psi}$ k této větvi následně přirostl vzorek $K3$, což se opakovalo jen u části z 4055 případů. Větev obsahující právě jen vzorky $T3$, $K1$ a $K3$ se mezi 5000 bootstrapovanými stromy objevila 4790 krát, tzn. její $\alpha_{_F} = 4790/5000 = 0.958$. Ovšem pouze některé z těchto 4790 větví vznikly tak, že se nejdříve spojily vzorky $T3$ a $K1$, k nimž se poté přidal vzorek $K3$. Zbylé srůstaly v jiném pořadí. Větev složenou právě jen ze vzorků $A2$ a $K4$ mělo všech 5000 stromů, proto je její odhadnutá spolehlivost 100%. Její nadvětev, tj. větev, v níž se nacházejí právě jen prvky $A2$, $K4$ a $B3$, bylo možno nalézt už pouze u necelých 62% bootstrapovaných stromů.

V ostatních případech k původní větvi (tj. $A2$ a $K4$) buď přirostlo dříve něco jiného než $B3$, anebo se $B3$ už dříve spojilo s nejakým jiným vzorkem, a pak teprve takto vzniklý shluk přirostl k $A2$ a $K4$.

Podívejme se nyní podrobněji na vztah takto spočteného odhadu spolehlivosti ke skutečnosti.

2.4. **Teoretické pozadí.** O sloupcích $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{73}$ matice

$$\boldsymbol{X} = \big(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{73}\big)\,, \qquad \text{kde} \quad x_l^k \in \big\{\text{A,G,T,C}\big\} \qquad \begin{matrix} k = 1, \ldots, 73, \\ l = 1, \ldots, 13, \end{matrix} \tag{4}$$

předpokládáme, že tvoří náhodný výběr o rozsahu $n = 73$ z nějakého pravděpodobnostního rozdělení na prostoru **všech možných nekonstantních** 13-ti rozměrných vektorů $\boldsymbol{\xi}$ nad abecedou $\big\{\text{A,G,T,C}\big\}$, tento prostor označme $\mathcal{X}$.

$$\mathcal{X} = \big\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K\big\} \qquad\qquad K = 4^{13} - 4 = 67\,108\,860\,. \tag{5}$$

Schématem

$$\boldsymbol{X} \Longrightarrow \text{Shluková analýza} \Longrightarrow \widehat{\Psi} \quad \text{resp.} \quad \boldsymbol{X}^* \Longrightarrow \text{Shluková analýza} \Longrightarrow \widehat{\Psi}^*, \tag{6}$$

lze popsat jak oblast praktickou, tak oblast virtuální, nicméně těžko rovinu teoretickou, jejíž vztah k oběm předešlým nás zajímá. Proto se přesuneme do prostoru $\mathcal{P}$, tedy prostoru **pravděpodobností zahrnutí** π_i vektorů $\boldsymbol{\xi}_i \in \mathcal{X}$ do výběru:

$$\mathcal{P} \stackrel{df}{=} \Big\{\boldsymbol{\pi} = \big(\pi_1, \ldots, \pi_K\big)' \colon \quad \pi_i = \mathbf{P}\big(\text{zahrnutí } \boldsymbol{\xi}_i \text{ do výběru}\big), \quad i = 1, \ldots, K\Big\} \tag{7}$$

Teoretickým pravděpodobnostem $\boldsymbol{\pi}$ by na rovině praktické odpovídaly **pozorované četnosti** výskytu vektorů $\boldsymbol{\xi}_i$ ve sloupcích datové matice $\boldsymbol{X}$, tj.

$$\boldsymbol{\Upsilon} = \big(\Upsilon_1, \ldots, \Upsilon_K\big)', \qquad \text{kde} \quad \Upsilon_i \stackrel{df}{=} \#\big\{j \colon \boldsymbol{x}^j = \boldsymbol{\xi}_i\big\} \qquad i = 1, \ldots, K\,, \tag{8}$$

resp. **relativní četnosti**

$$\widehat{\boldsymbol{\pi}} = \big(\widehat{\pi}_1, \ldots, \widehat{\pi}_K\big)', \quad \text{kde} \quad \widehat{\pi}_i \stackrel{df}{=} \frac{1}{n}\#\big\{j \colon \boldsymbol{x}^j = \boldsymbol{\xi}_i\big\} = \frac{\Upsilon_i}{n}\,, \qquad i = 1, \ldots, K\,. \tag{9}$$

Analogicky bychom z bootstrapovaných matic $\boldsymbol{X}_b^* = \big(\boldsymbol{x}_b^{*1}, \ldots, \boldsymbol{x}_b^{*73}\big)$, $b = 1, \ldots, B$, získali:

$$\boldsymbol{\Upsilon}_b^* = \big(\Upsilon_{b,1}^*, \ldots, \Upsilon_{b,K}^*\big)', \quad \text{kde} \quad \Upsilon_{b,i}^* \stackrel{df}{=} \#\big\{j \colon \boldsymbol{x}_b^{*j} = \boldsymbol{\xi}_i\big\}\,, \quad i = 1, \ldots, K\,, \tag{10}$$

$$\widehat{\boldsymbol{\pi}}_b^* = \big(\widehat{\pi}_{b,1}^*, \ldots, \widehat{\pi}_{b,K}^*\big)', \qquad \text{kde} \quad \widehat{\pi}_{b,i}^* \stackrel{df}{=} \frac{1}{n}\,\Upsilon_{b,i}^*\,. \tag{11}$$

Zřejmě platí:

$$\boldsymbol{\Upsilon} \sim \text{Multi}_K(n, \boldsymbol{\pi}) \qquad\qquad \text{a} \qquad\qquad \boldsymbol{\Upsilon}^* \sim \text{Multi}_K(n, \widehat{\boldsymbol{\pi}})\,, \tag{12}$$

kde $K = 67\,108\,860$ a $n = 73$.

Lze namítnout, že právě zmíněný přístup nás nutí pracovat s velmi dlouhými vektory obsahujícími navíc téměř výhradně samé nuly. Nicméně, pro teoretická odvození se tato reprezentace ukazuje být přínosnou. Krom toho, z důvodů, jež budou vysvětleny později, vyplyne, že při výpočtech se lze omezit jen na nenulové složky.

Pěstování stromů, popsané schématem (6), lze nyní vyjádřit jako:

$$\widehat{\boldsymbol{\pi}} \Longrightarrow \widehat{\Psi}, \qquad\qquad \text{resp.} \quad \widehat{\boldsymbol{\pi}}_b^* \Longrightarrow \widehat{\Psi}_b^*\,. \tag{13}$$

Podobně si lze představit, že i teoretickým pravděpodobnostem $\boldsymbol{\pi}$ odpovídá nějaký teoretický strom Ψ vystihující skutečné vztahy mezi sledovanými organismy.

Stojí za zmínku, že ze dvou různých pravděpodobnostních vektorů $\widehat{\boldsymbol{\pi}}^{(1)}$ a $\widehat{\boldsymbol{\pi}}^{(2)}$ může vzniknout tentýž strom. Prostor $\mathcal{P}$ je tak rozložen na disjunktní třídy ekvivalence $\mathcal{P} = \dot{\bigcup}_i \mathcal{P}_i$, pro něž platí:

$$\left\{\mathcal{P}_i \ni \widehat{\boldsymbol{\pi}}^{(1)} \quad \& \quad \mathcal{P}_j \ni \widehat{\boldsymbol{\pi}}^{(2)}\right\} \qquad \Longrightarrow \qquad \left\{\,\widehat{\Psi}^{(1)} \equiv \widehat{\Psi}^{(2)} \;\Leftrightarrow\; i = j\,\right\}, \tag{14}$$

kde $\widehat{\Psi}^{(j)}$ odpovídá stromu, který vznikl z $\widehat{\boldsymbol{\pi}}^{(j)}$, $j = 1, 2$.

Vraťme se nyní k původní otázce po spolehlivosti námi vypěstovaného stromu $\widehat{\Psi}$. Tážeme se po pravděpodobnosti, s níž se $\widehat{\Psi}$ bude shodovat s teoretickým vzorem Ψ. Zajímá nás tedy:

$$\alpha = \mathbf{P}\big(\widehat{\Psi} \equiv \Psi\big) = \mathbf{P}\big(\boldsymbol{\pi} \in \mathcal{P}_i \,|\widehat{\boldsymbol{\pi}} \in \mathcal{P}_i\big) \tag{15}$$

Tuto pravděpodobnost jsme odhadli pomocí (3) jako empirickou pravděpodobnost (relativní četnost) toho, že $\widehat{\Psi}^*$ vypěstovaný nad $\boldsymbol{X}^*$ má danou větev shodnou s původním stromem $\widehat{\Psi}$, tedy:

$$\alpha_{_F} = \widehat{\mathbf{P}}_{\widehat{\boldsymbol{\pi}}^* \,|\widehat{\boldsymbol{\pi}}}\big(\widehat{\boldsymbol{\pi}}^* \in \mathcal{P}_i \,|\widehat{\boldsymbol{\pi}} \in \mathcal{P}_i\big)\,. \tag{16}$$

Abychom si mohli udělat představu o chybě, jíž jsme se tímto dopustili, potřebovali bychom znat rozdělení $\widehat{\boldsymbol{\pi}}^*$ (resp. aposteriorní rozdělení $\boldsymbol{\pi}$) při daném $\widehat{\boldsymbol{\pi}}$, tedy $\mathcal{L}\big(\widehat{\boldsymbol{\pi}}^* \,|\widehat{\boldsymbol{\pi}}\big)$ (resp. $\mathcal{L}\big(\boldsymbol{\pi} \,|\widehat{\boldsymbol{\pi}}\big)$).

První z nich známe, neboť ze vztahů (11) a (12) plyne, že:

$$\mathcal{L}\big(n\widehat{\boldsymbol{\pi}}^* \,|\widehat{\boldsymbol{\pi}}\big) = \mathrm{Multi}_K(n, \widehat{\boldsymbol{\pi}})\,. \tag{17}$$

Co se týče aposteriorního rozdělení $\mathcal{L}\big(\boldsymbol{\pi} \,|\widehat{\boldsymbol{\pi}}\big)$, lze je vyjádřit přes výběrové rozdělení $\widehat{\boldsymbol{\pi}}$ při daném $\boldsymbol{\pi}$, pro které díky uvažovanému modelu platí:

$$\mathcal{L}\big(n\widehat{\boldsymbol{\pi}} \,|\boldsymbol{\pi}\big) = \mathrm{Multi}_K(n, \boldsymbol{\pi})\,. \tag{18}$$

Multinomické rozdělení tvoří konjugovaný systém spolu s rozdělením Dirichletovým, a tudíž lze snadno odvodit (viz též např. [5]), že platí:

Tvrzení 2.1. *Řídí-li se apriorní rozdělení $\boldsymbol{\pi}$ principem neurčitosti, pak*

$$\mathcal{L}\big(\boldsymbol{\pi} \,|\widehat{\boldsymbol{\pi}}\big) = \mathit{Dir}_K(n\widehat{\boldsymbol{\pi}})\,. \tag{19}$$

Tvrzení 2.2. *$\mathcal{L}\big(\widehat{\boldsymbol{\pi}}^* \,|\widehat{\boldsymbol{\pi}}\big)$ a $\mathcal{L}\big(\boldsymbol{\pi} \,|\widehat{\boldsymbol{\pi}}\big)$ mají následující vlastnosti:*

1. *Obě jsou soustředěna jen na takových $\boldsymbol{\xi}_i$, pro něž $\pi_i > 0$, tj. na sloupcích matice $\boldsymbol{X}$.*
2. *Mají stejnou střední hodnotu.*
3. *Pro kovarianční matice platí:*

$$\Big(\mathsf{Cov}\left(\widehat{\boldsymbol{\pi}}^*_i,\, \widehat{\boldsymbol{\pi}}^*_j \,|\widehat{\boldsymbol{\pi}}\right)\Big)_{i,j=1}^{K} = \frac{n+1}{n} \cdot \Big(\mathsf{Cov}\left(\boldsymbol{\pi}_i,\, \boldsymbol{\pi}_j \,|\widehat{\boldsymbol{\pi}}\right)\Big)_{i,j=1}^{K}\,. \tag{20}$$

Z tvrzení je patrné, že odhad hladin spolehlivosti větví vývojového stromu (pomocí relativních četností výskytu těchto větví v bootstrapových populacích) definovaný vztahem (3) má krom předností, jež spočívají především v jednoduchosti, snadné implementovatelnosti a v rychlosti výpočtu i své omezení. Problém nastává v okamžiku, kdy nelze splnit předpoklady tvrzení 2.1 o apriorním rozdělení. Potom nemusí platit závěry tvrzení 2.2 a rozdělení $\mathcal{L}\big(\widehat{\boldsymbol{\pi}}^* \,|\widehat{\boldsymbol{\pi}}\big)$ a $\mathcal{L}\big(\boldsymbol{\pi} \,|\widehat{\boldsymbol{\pi}}\big)$ mohou být natolik odlišná, že odhadnutá hladina spolehlivosti α_F může být značně vzdálena od skutečné hladiny α. Vzhledem ke složitosti prostoru $\mathcal{P}$ není příliš zřejmé, jak tyto situace rozeznat. Pokusme se tedy nalézt lepší přístup.

3. Kterak postupovat lépe

3.1. **Teorie.** Podobně jako v odstavci 2 vyjdeme z modelu (7). Index $i = 1$ vyhraďme pro tu část prostoru $\mathcal{P} = \dot{\bigcup}_i \mathcal{P}_i$ při dělení dle pravidla (14), která obsahuje $\widehat{\boldsymbol{\pi}}$, zatímco $\mathcal{P}_1^C = \dot{\bigcup}_{i\neq 1} \mathcal{P}_i$ budiž její doplněk.

Jedna z cest, která se nyní nabízí, je vést v patrnosti nejen skutečnost, že $\widehat{\boldsymbol{\pi}} \in \mathcal{P}_1$, jako jsme činili v předchozím, ale zohlednit i vzdálenost vektoru $\widehat{\boldsymbol{\pi}}$ od hranice oblasti $\mathcal{P}_1$. Tato snaha je motivována úvahou, že vektor $\boldsymbol{\pi}$, z něhož je vektor pozorovaných relativních četností $\widehat{\boldsymbol{\pi}}$ dle předpokládaného modelu (9) nagenerován, by se měl nacházet „někde poblíž“ vektoru $\widehat{\boldsymbol{\pi}}$, a proto by hladina spolehlivosti

$$\alpha = \mathbf{P}\big(\boldsymbol{\pi} \in \mathcal{P}_1 \,|\widehat{\boldsymbol{\pi}} \in \mathcal{P}_1\big)\,, \tag{21}$$

měla být vyšší, pokud by $\widehat{\boldsymbol{\pi}}$ leželo „někde uprostřed“ $\mathcal{P}_1$, než by tomu bylo v případě $\widehat{\boldsymbol{\pi}}$ umístěného v blízkosti hranice oblasti $\mathcal{P}_1$.

Označme $\boldsymbol{\pi}_0$ jako nejbližší hraniční bod k $\widehat{\boldsymbol{\pi}}$, tj.

$$\boldsymbol{\pi}_0 \stackrel{df}{=} \arg\min_{\breve{\boldsymbol{\pi}} \in \partial\mathcal{P}_1} d(\widehat{\boldsymbol{\pi}}, \breve{\boldsymbol{\pi}})\,, \tag{22}$$

kde $\partial\mathcal{P}_1$ značí hranici $\mathcal{P}_1$ a $d(.\,,.)$ je vhodná metrika na $\mathcal{P}$, tj. nezáporná funkce splňující navíc pro libovolné $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{P}$, že $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$, dále že $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ právě tehdy, když $\boldsymbol{x} = \boldsymbol{y}$, a $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$.

V předchozím případě jsme boot-populaci obdrželi prostřednictvím četností výskytu vygenerovaných z $\widehat{\boldsymbol{\pi}}$, viz (12), nyní je nagenerujeme z $\boldsymbol{\pi}_0$, tedy pro $b = 1, \ldots, B_2$

$$\boldsymbol{\Upsilon}_b^{**} = \big(\Upsilon_{1b}^{**}, \ldots, \Upsilon_{Kb}^{**}\big)'\,, \quad \text{kde} \quad \Upsilon_{ib}^{**} \stackrel{df}{=} \#\big\{j:\ \boldsymbol{x}_b^{**j} = \boldsymbol{\xi}_i\big\}\,, \quad i = 1, \ldots, K\,, \tag{23}$$

$$\boldsymbol{\Upsilon}_b^{**} \sim \mathrm{Multi}_K(n, \boldsymbol{\pi}_0) \qquad \text{a} \qquad \widehat{\boldsymbol{\pi}}_b^{**} = \frac{1}{n}\,\boldsymbol{\Upsilon}^{**}\,. \tag{24}$$

Chceme-li odhadnout hladinu spolehlivosti (21), můžeme uvažovat takto: čím blíže budou prvky boot-populace $\widehat{\boldsymbol{\pi}}_b^{**}$ vygenerované z hraničního bodu $\boldsymbol{\pi}_0$ ležet k $\mathcal{P}_1^C$ ve srovnání s polohou $\widehat{\boldsymbol{\pi}}$ (tj. čím menší bude jejich odhadnutý rozptyl), tím menší je pravděpodobnost, že by vektor skutečných pravděpodobností $\boldsymbol{\pi}$ ležel až v $\mathcal{P}_1^C$, neboť předpokládáme, že právě z $\boldsymbol{\pi}$ byl nagenerován i vektor pozorovaných relativních četností $\widehat{\boldsymbol{\pi}}$, a to stejným způsobem jako byly nagenerovany vektory $\widehat{\boldsymbol{\pi}}_b^{**}$ z $\boldsymbol{\pi}_0$, srov. (12) a (24) . Tím větší je tedy α. Hladinu spolehlivosti zavedeme následovně:

$$\alpha_{_E} \stackrel{df}{=} \widehat{\mathbf{P}}_{\widehat{\boldsymbol{\pi}}^{**}\,|\boldsymbol{\pi}_0}\big(d(\widehat{\boldsymbol{\pi}}^{**}, \mathcal{P}_1^C) \leq d(\widehat{\boldsymbol{\pi}}, \mathcal{P}_1^C)\,|\,\widehat{\boldsymbol{\pi}} \in \mathcal{P}_1\big)\,. \tag{25}$$

Takto definovaná hladina nám zaručuje, že $1 - \alpha_{_E}$ je obdobou dosažené hladiny testu (tzv. p-hodnoty) v klasickém testování hypotéz, což lze nahlédnout např. z následujícího příkladu.

Příklad 3.1. Nechť $Z_1, \ldots, Z_n \overset{iid}{\sim} N(\mu, 1)$.

Testujme hypotézu:

$$\text{H: } \underbrace{\mu \leq \mu_0}_{\widehat{\pi} \in \mathcal{P}_1^C} \qquad \text{vs.} \qquad \text{K: } \underbrace{\mu > \mu_0}_{\widehat{\pi} \in \mathcal{P}_1} \tag{26}$$

Zde $\boldsymbol{\pi}_0 \in \partial\mathcal{P}_1$ odpovídá hraničnímu bodu hypotézy a alternativy, tj. hodnotě $\mu = \mu_0$. Provedením experimentu spočteme odhad $\widehat{\mu} = \widehat{\mu}(Z_1, \ldots, Z_n)$ – analogii $\widehat{\pi}$. Při známé hodnotě $\widehat{\mu}$, a tedy i známé $\mathcal{P}_1 \ni \widehat{\mu}$, chceme potvrdit, že také $\mu \in \mathcal{P}_1$, tedy zamítnout

opak. Proto $H \equiv \mathcal{P}_1^C$. Označme dále $Z_1^{**}, \ldots, Z_n^{**} \overset{iid}{\sim} N(\mu_0, 1)$ bootstrapovaný výběr z hypotézy H a příslušný odhad $\widehat{\mu}^{**} = \widehat{\mu}^{**}(Z_1^{**}, \ldots, Z_n^{**})$ (analogie $\widehat{\boldsymbol{\pi}}^{**}$), potom

$$p = \mathbf{P}_{\mu_0}\left(\widehat{\mu}^{**} > \widehat{\mu}\right) = \mathbf{P}_{\mu_0}\left(\widehat{\mu}^{**} - \mu_0 > \widehat{\mu} - \mu_0\right). \tag{27}$$

Zároveň platí:

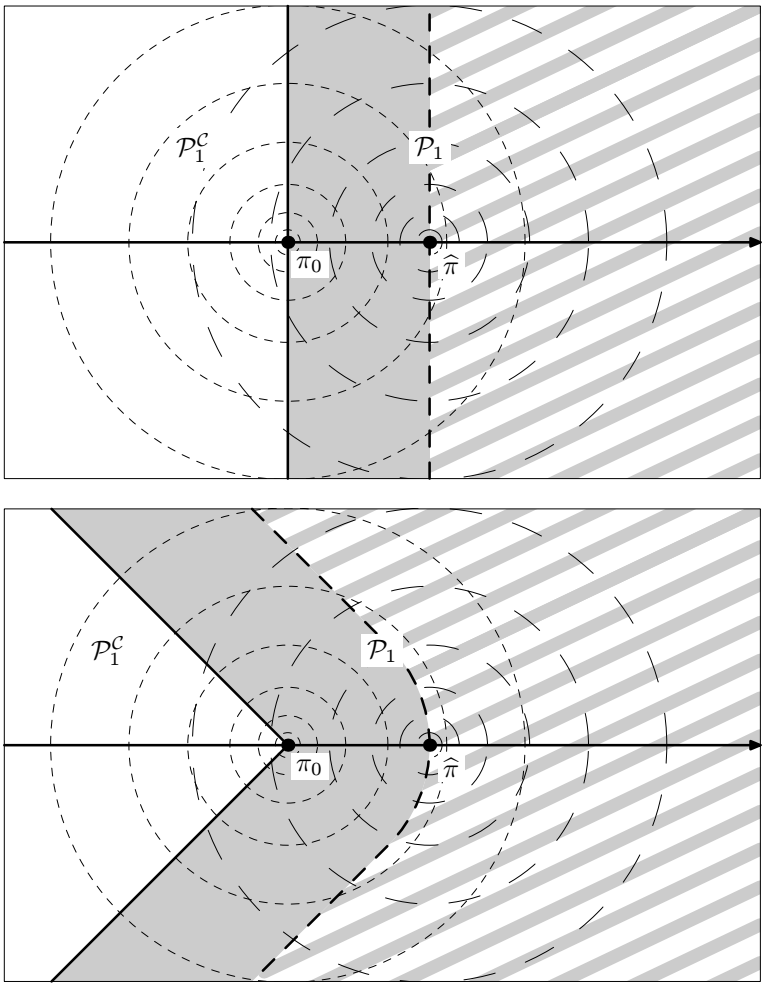
$$\mathrm{d}\left(\widehat{\boldsymbol{\mu}}^{**}, \mathcal{P}_1^C\right) = \begin{cases} \widehat{\mu}^{**} - \mu_0 & \text{pro} \quad \widehat{\mu}^{**} > \mu_0 \\ 0 & \text{jinak}, \end{cases} \tag{28}$$

odkud

$$\mathbf{P}_{\mu_0}\left(\mathrm{d}\left(\widehat{\boldsymbol{\mu}}^{**}, \mathcal{P}_1^C\right) > \widehat{\mu} - \mu_0\right) = \mathbf{P}_{\mu_0}\left(\widehat{\mu}^{**} - \mu_0 > \widehat{\mu} - \mu_0 > 0\right) + 0\,, \tag{29}$$

a tedy

$$p = 1 - \alpha_{_E} \tag{30}$$



Obr. 3, 4 *Závislost hladin $\alpha_{_F}$ a $\alpha_{_E}$ na tvaru oblastí: nezakřivené hranice (nahoře), zakřivené hranice (dole).*

Nově zavedená hladina $\alpha_{_E}$ bere více v potaz tvar oblasti $\mathcal{P}_1$ než $\alpha_{_F}$. Jakým způsobem je zřejmé z obrázků 3 a 4, kde je chování obou hladin demonstrováno opět na

příkladu (tentokrát dvourozměrném) normálních rozdělení se stejným rozptylem a se střední hodnotou v $(\boldsymbol{\pi}_0, 0)$, resp. v $(\widehat{\boldsymbol{\pi}}, 0)$. Nevybarvená plocha odpovídá oblasti $\mathcal{P}_1^C$ (tj. odpovídá $1-\alpha_F$). Z vybarvené oblasti $\mathcal{P}_1$ je pak šrafováním vyznačena ta část, kde $\mathrm{d}\left(\widehat{\boldsymbol{\pi}}^{**}, \mathcal{P}_1^C\right) > \mathrm{d}\left(\widehat{\boldsymbol{\pi}}, \mathcal{P}_1^C\right)$, tedy ta, jež opovídá $1-\alpha_E$. Zatímco pro nezakřivené hranice se obě hladiny shodují, v případě ryze konvexního tvaru oblasti $\mathcal{P}_1^C$ platí:

$$1-\alpha_{_F} = \widehat{\mathbf{P}}_{\widehat{\boldsymbol{\pi}}^* \,|\, \widehat{\boldsymbol{\pi}}}\left(\widehat{\boldsymbol{\pi}}^* \in \mathcal{P}_1^C \,|\, \widehat{\boldsymbol{\pi}} \in \mathcal{P}_1\right) \leq \tag{31}$$

$$\leq \mathbf{P}_{\widehat{\boldsymbol{\pi}}^{**} \,|\, \boldsymbol{\pi}_0}\Big(\mathrm{d}\left(\widehat{\boldsymbol{\pi}}^{**}, \mathcal{P}_1^C\right) > \mathrm{d}\left(\widehat{\boldsymbol{\pi}}, \mathcal{P}_1^C\right) \,|\boldsymbol{\pi}_0 \in \partial\mathcal{P}_1\Big) = 1-\alpha_{_E}. \tag{32}$$

Tedy

$$\alpha_{_E} \leq \alpha_{_F}\,. \tag{33}$$

Nyní zbývá vyřešit otázku, jak spočíst $\alpha_{_E}$.

3.2. **Intervaly spolehlivosti.** Vzhledem ke složitosti prostoru $\mathcal{P}$ není snadné vypočítat α_E přímo. Odložíme-li však na chvíli tento záměr, můžeme se přesunout do zdánlivě nesouvisejících oblastí matematické statistiky – do problematiky intervalových odhadů, které nám v konečném důsledku pomohou nalézt vhodnou aproximaci α_E.

Představme si na okamžik, že stojíme před opačným problémem: kterak při zadané hladině určit interval spolehlivosti, tedy oblast, jež s danou spolehlivostí pokryje skutečnou hodnotu neznámého parametru.

Pro názornost pracujme s jednoduchým modelem, kde data pocházejí z rozdělení určeného jednoparametrickým systémem hustot. Parametr θ odhadujeme reálnou statistikou $\widehat{\theta}$ s rozdělením:

$$\widehat{\theta} \sim f_\theta \tag{34}$$

3.2.1. *Standardní asymptotický interval spolehlivosti.* Využijeme-li normální aproximace

$$\frac{\left(\widehat{\theta}-\theta\right)}{\widehat{\sigma}} \overset{as}{\sim} N(0,1)\,, \tag{35}$$

kde $\overset{as}{\sim}$ označuje, že rozdělení statistiky $\sigma^{-1}\left(\widehat{\theta}-\theta\right)$ je asymptoticky normální a kde $\widehat{\sigma}$ je konzistentní odhad rozptylu, obdržíme často užívaný $100(1-\alpha)$ procentní asymptotický interval spolehlivosti pro θ ve tvaru

$$\left(\widehat{\theta} + \widehat{\sigma} z^{(\frac{\alpha}{2})};\ \widehat{\theta} + \widehat{\sigma} z^{(1-\frac{\alpha}{2})}\right)\,, \tag{36}$$

kde

$$z^{(\alpha)} = \Phi^{-1}(\alpha). \tag{37}$$

Tento interval ovšem může být v případě konečných výběrů značně nepřesný.

3.2.2. *Percentilová metoda.* Problému s normalitou $\left(\widehat{\theta}-\theta\right)$ bychom se mohli pokusit vyhnout například odhadnutím percentilů pomocí bootstrapu, kdy odhadneme θ prostřednictvím statistiky $\widehat{\theta}^*$ spočtené z B-tice bootstrapovaných populací. Z empirické distribuční funkce bootstrapovaného odhadu $\widehat{\theta}^*$

$$\widehat{F}_{\widehat{\theta}}(t) = \widehat{\mathbf{P}}_{\widehat{\theta}}\left(\widehat{\theta}^* \leq t\right) = B^{-1}\#\{b : \widehat{\theta}_b^* \leq t\}\,, \tag{38}$$

získáme pro pevné $\alpha \in (0;\, 0.5)$ meze $t_d^{(\frac{\alpha}{2})}$, $t_h^{(\frac{\alpha}{2})}$ intervalu spolehlivosti jakožto příslušné percentily. Přibližný $100(1-\alpha)\%$-ní interval spolehlivosti pro θ pak má tvar

$$\left(t_d^{(\frac{\alpha}{2})};\, t_h^{(\frac{\alpha}{2})}\right) = \left(\widehat{F}_{\widehat{\theta}}^{-1}\Big(\frac{\alpha}{2}\Big); \widehat{F}_{\widehat{\theta}}^{-1}\Big(1-\frac{\alpha}{2}\Big)\right). \tag{39}$$

Tento odhad – ač mnohdy lepší než standardní interval – není příliš důvěryhodný, jak co se týče vychýlení, tak i co do své délky (blíže viz např. [2]). Lze jej nicméně vhodně modifikovat do té míry, že se stane použitelným.

3.2.3. *Oprava vychýlení percentilové metody.* Předpokládejme, že hustota rozdělení $\widehat{\theta}^*$ náleží do stejného systému jako hustota statistiky $\widehat{\theta}$, s tím rozdílem, že namísto hodnotou θ je určena jejím bodovým odhadem, tj. $\widehat{\theta}$. Tedy:

$$\widehat{\theta}^* \sim f_{\widehat{\theta}}\,. \tag{40}$$

Distribuční funkci $\widehat{\theta}^*$ lze tedy vyjádřit jako:

$$F_{\widehat{\theta}}(t) = \int_{-\infty}^{t} f_{\widehat{\theta}}\big(\widehat{\theta}^*\big) d\widehat{\theta}^* = \mathbf{P}_{\widehat{\theta}}\big(\widehat{\theta}^* \le t\big) \tag{41}$$

Dále předpokládejme, že normality $\big(\widehat{\theta}-\theta\big)$ i stability rozptylu lze dosáhnout vhodnou transformací; tedy, že existuje taková monotónně rostoucí funkce $g : \Theta \to \Gamma$ a konstanty a a $\sigma > 0$, že pro veličiny

$$\gamma = g(\theta)\,, \qquad \widehat{\gamma} = g\big(\widehat{\theta}\big) \qquad \text{a} \qquad \widehat{\gamma}^* = g\big(\widehat{\theta}^*\big) \tag{42}$$

platí:

$$\big(\widehat{\gamma}-\gamma\big) \sim N(-z_0\sigma; \sigma^2) \qquad \text{a} \qquad \big(\widehat{\gamma}^*-\widehat{\gamma}\big) \sim N(-z_0\sigma; \sigma^2)\,. \tag{43}$$

Rádi bychom stanovili interval spolehlivosti pro θ, tedy $(t_d^{(\frac{\alpha}{2})}, t_h^{(\frac{\alpha}{2})})$ tak, aby platilo

$$\mathbf{P}_\theta\big(t_d^{(\frac{\alpha}{2})} < \theta < t_h^{(\frac{\alpha}{2})}\big) = 1-\alpha\,. \tag{44}$$

Pro zjednodušení značení nebudeme nadále zdůrazňovat zřejmou závislost mezí intervalu spolehlivosti $t_d^{(\frac{\alpha}{2})}$ a $t_h^{(\frac{\alpha}{2})}$ na hladině α. Napříště budeme tedy psát pouze t_d a t_h, nebude-li nutno jinak.
Díky monotonnii transformace g lze pracovat v prostoru Γ:

$$\begin{aligned} 1-\alpha &= \mathbf{P}_\theta\big(g(t_d) < \gamma < g(t_h)\big) = \\ &= \mathbf{P}_\theta\Big(\frac{\widehat{\gamma}-g(t_d)}{\sigma} + z_0 > \frac{\widehat{\gamma}-\gamma}{\sigma} + z_0 > \frac{\widehat{\gamma}-g(t_h)}{\sigma} + z_0\Big)\,. \end{aligned} \tag{45}$$

Z normovaného normálního rozdělení prostředního výrazu plyne:

$$g(t_d) = \widehat{\gamma} + z_0\sigma - z^{(1-\frac{\alpha}{2})}\sigma \qquad g(t_h) = \widehat{\gamma} + z_0\sigma + z^{(1-\frac{\alpha}{2})}\sigma\,. \tag{46}$$

Pro distribuční funkci veličiny $\widehat{\gamma}^*$

$$G(s) = \mathbf{P}_{\widehat{\gamma}}\big(\widehat{\gamma}^* \le s\big) \tag{47}$$

díky monotonii funkce g platí:

$$G(g(t)) = \mathbf{P}_{\widehat{\gamma}}\big(\widehat{\gamma}^* \le g(t)\big) = \mathbf{P}_{\widehat{\gamma}}\big(g(\widehat{\theta}^*) \le g(t)\big) = \mathbf{P}_{\widehat{\theta}}\big(\widehat{\theta}^* \le t\big) = F_{\widehat{\theta}}(t)\,. \tag{48}$$

Tedy

$$F_{\widehat{\theta}} = G \circ g\,, \qquad \text{odkud} \qquad F_{\widehat{\theta}}^{-1} = g^{-1} \circ G^{-1}\,. \tag{49}$$

Odtud z normality (43) obdržíme jednak opravu vychýlení z_0:

$$F_{\widehat{\theta}}(\widehat{\theta}) = G(g(\widehat{\theta})) = \mathbf{P}_{\widehat{\gamma}}\big(\widehat{\gamma}^* \le \widehat{\gamma}\big) = \widehat{\mathbf{P}}_{\widehat{\gamma}}\left(\frac{\widehat{\gamma}^* - \widehat{\gamma} + z_0\sigma}{\sigma} \le z_0\right) = \Phi(z_0)\,, \tag{50}$$

a tedy

$$z_0 = \Phi^{-1}\big(F_{\widehat{\theta}}(\widehat{\theta})\big)\,, \tag{51}$$

navíc díky (43), tj. stejnému rozdělení rozdílů $\big(\widehat{\gamma} - \gamma\big)$ a $\big(\widehat{\gamma}^* - \widehat{\gamma}\big)$, vzhledem k (46) platí:

$$\begin{aligned} F_{\widehat{\theta}}(t_h) = G(g(t_h)) &= \mathbf{P}_{\widehat{\gamma}}\big(\widehat{\gamma}^* \le \widehat{\gamma} + z_0\sigma + z^{(1-\frac{\alpha}{2})}\sigma\big) = \\ &= \mathbf{P}_{\gamma}\big(\widehat{\gamma} \le \gamma + z_0\sigma + z^{(1-\frac{\alpha}{2})}\sigma\big) = \mathbf{P}_{\gamma}\Big(\frac{\widehat{\gamma} - \gamma + z_0\sigma}{\sigma} \le \frac{2z_0\sigma + z^{(1-\frac{\alpha}{2})}\sigma}{\sigma}\Big) \\ &= \Phi\big(2z_0 + z^{(1-\frac{\alpha}{2})}\big)\,. \end{aligned} \tag{52}$$

Dosadíme-li rovnici (52) do argumentu $F^{-1}_{\widehat{\theta}}$, získáme horní a analogicky i dolní mez intervalu spolehlivosti:

$$t_d^{(\frac{\alpha}{2})} = F^{-1}_{\widehat{\theta}}\big(\Phi(2z_0 - z^{(1-\frac{\alpha}{2})})\big) \qquad t_h^{(\frac{\alpha}{2})} = F^{-1}_{\widehat{\theta}}\big(\Phi(2z_0 + z^{(1-\frac{\alpha}{2})})\big) \tag{53}$$

Poznámka 3.1. Povšimněme si, že o zobrazení g stačí předpokládat pouze jeho existenci; nemusíme tedy znát jeho explicitní tvar.

Poznámka 3.2. Od rozdělení statistik $\big(\widehat{\gamma} - \gamma\big)$ a $\big(\widehat{\gamma}^* - \widehat{\gamma}\big)$ navíc nemusíme nutně žádat, aby bylo normální, postačilo by libovolné symetrické rozdělení, jehož kvantily jsme schopni spočítat.

Poznámka 3.3. $F_{\widehat{\theta}}$ lze odhadnout např. pomocí empirické distribuční funkce $\widehat{F}_{\widehat{\theta}}$, viz (38)

Poznámka 3.4. Pro rozdělení $F_{\widehat{\theta}}$, pro něž je $\widehat{\theta}$ rovno mediánu, vyjde korekční člen $z_0 = 0$.

Poznámka 3.5. Požadavky na stabilitu rozptylu rozdílů $\big(\widehat{\gamma} - \gamma\big)$ a $\big(\widehat{\gamma}^* - \widehat{\gamma}\big)$ a jejich normalitu si mohou protiřečit.

3.2.4. *Opravená percentilová metoda bez konstantního rozptylu.* Upustíme-li od předpokladu stability rozptylu rozdílů (43) a požadujeme-li, aby pro vhodné konstanty a a z_0 (tj. $a > -1/\gamma$ pro $\gamma > 0$ a $a < -1/\gamma$ pro $\gamma < 0$) platilo:

$$\begin{aligned} &\big(\widehat{\gamma} - \gamma\big) \sim N(-z_0\sigma_\gamma; \sigma_\gamma^2)\,, && \text{kde} \quad \sigma_\gamma = 1 + a\gamma\,, \\ \text{a} \quad &\big(\widehat{\gamma}^* - \widehat{\gamma}\big) \sim N(-z_0\sigma_{\widehat{\gamma}}; \sigma_{\widehat{\gamma}}^2)\,, && \text{kde} \quad \sigma_{\widehat{\gamma}} = 1 + a\widehat{\gamma}\,, \end{aligned} \tag{54}$$

dospějeme pomocí stejných úvah jako v předchozím případě, tj. (44) – (46), k nerovnostem udávajícím interval spolehlivosti pro γ:

$$\begin{aligned} \widehat{\gamma} + z_0\sigma_\gamma - z^{(1-\frac{\alpha}{2})}\sigma_\gamma < \gamma < \widehat{\gamma} + z_0\sigma_\gamma + z^{(1-\frac{\alpha}{2})}\sigma_\gamma \\ \widehat{\gamma} + (z_0 - z^{(1-\frac{\alpha}{2})})(1 + a\gamma) < \gamma < \widehat{\gamma} + (z_0 + z^{(1-\frac{\alpha}{2})})(1 + a\gamma) \\ \widehat{\gamma} + z_0 - z^{(1-\frac{\alpha}{2})} + \gamma a(z_0 - z^{(1-\frac{\alpha}{2})}) < \gamma < \widehat{\gamma} + z_0 + z^{(1-\frac{\alpha}{2})} + \gamma a(z_0 + z^{(1-\frac{\alpha}{2})}) \end{aligned} \tag{55}$$

$$\gamma_d \overset{df}{=} \frac{\widehat{\gamma} + z_0 - z^{(1-\frac{\alpha}{2})}}{1 - a(z_0 - z^{(1-\frac{\alpha}{2})})} < \gamma < \frac{\widehat{\gamma} + z_0 + z^{(1-\frac{\alpha}{2})}}{1 - a(z_0 + z^{(1-\frac{\alpha}{2})})} \overset{df}{=} \gamma_h\,. \tag{56}$$

Obě meze lze ještě upravit:

$$\gamma_d = \frac{\widehat{\gamma} + z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})} = \widehat{\gamma}\Big(1 + a\frac{z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})}\Big) + \frac{z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})}$$

$$= \widehat{\gamma} + \frac{z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})}\,(1 + a\widehat{\gamma}) = \widehat{\gamma} + \frac{z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})}\,\sigma_{\widehat{\gamma}}\,,$$

$$\gamma_h = \widehat{\gamma} + \frac{z_0 + z^{(1-\frac{\alpha}{2})}}{1 - a(z_0 + z^{(1-\frac{\alpha}{2})})}\,\sigma_{\widehat{\gamma}} \tag{57}$$

Je patrné, že obě meze lze vyjádřit jedním předpisem jako funkci hladiny α. Nechť

$$\gamma^{(\alpha)} \stackrel{df}{=} \widehat{\gamma} + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\,\sigma_{\widehat{\gamma}}\,, \quad \text{pak} \quad \gamma_d = \gamma^{(\frac{\alpha}{2})} \quad \text{a} \quad \gamma_h = \gamma^{(1-\frac{\alpha}{2})}\,. \tag{58}$$

Vezmeme-li v úvahu vztahy mezi distribučními funkcemi $F_{\widehat{\theta}}$ a G, tj. (49), dospějeme analogicky jako v předchozím případě k vyjádření intervalu spolehlivosti pro θ, viz (52).

$$F_{\widehat{\theta}}(t_d) = G(g(t_d)) = G(\gamma_d) = \mathbf{P}_{\widehat{\gamma}}\big(\widehat{\gamma}^* \le \gamma_d\big) = \mathbf{P}_{\widehat{\gamma}}\Big(\frac{\widehat{\gamma}^* - \widehat{\gamma}}{\sigma_{\widehat{\gamma}}} + z_0 \le \frac{\gamma_d - \widehat{\gamma}}{\sigma_{\widehat{\gamma}}} + z_0\Big)$$

$$= \Phi\bigg(\frac{\gamma_d - \widehat{\gamma}}{\sigma_{\widehat{\gamma}}} + z_0\bigg) = \Phi\bigg(\frac{z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})} + z_0\bigg)\,. \tag{59}$$

Odtud

$$t_d^{(\frac{\alpha}{2})} = F_{\widehat{\theta}}^{-1}\bigg(\Phi\Big(\frac{z_0 + z^{(\frac{\alpha}{2})}}{1 - a(z_0 + z^{(\frac{\alpha}{2})})} + z_0\Big)\bigg) \tag{60}$$

$$t_h^{(\frac{\alpha}{2})} = F_{\widehat{\theta}}^{-1}\bigg(\Phi\Big(\frac{z_0 + z^{(1-\frac{\alpha}{2})}}{1 - a(z_0 + z^{(1-\frac{\alpha}{2})})} + z_0\Big)\bigg)\,. \tag{61}$$

3.2.5. *Parametry z_0 a a.* Opravu vychýlení z_0 získáme úplně stejným způsobem jako v části 3.2.3, tj. viz (51). Tzv. „urychlovací konstantu“ a je možné (viz [1]) aproximovat výrazem:

$$a \doteq \frac{\mathrm{SKEW}_{\theta=\widehat{\theta}}(\dot{l}_\theta)}{6} = \frac{\mathsf{E}\,(\dot{l}_\theta - \mathsf{E}\,\dot{l}_\theta)^3}{6\Big(\mathsf{E}\,(\dot{l}_\theta - \mathsf{E}\,\dot{l}_\theta)^2\Big)^{\frac{3}{2}}}\bigg|_{\theta=\widehat{\theta}}\,, \text{ kde } \quad \dot{l}_\theta(\widehat{\theta}) = \frac{\partial}{\partial\theta}\log f_\theta(\widehat{\theta})\,. \tag{62}$$

Poznámka 3.6. Analogicky bychom obdrželi i jednostranné $100(1-\alpha)$ procentní intervaly spolehlivosti:

$$\bigg(-\infty\,;\; F_{\widehat{\theta}}^{-1}\bigg(\Phi\Big(\frac{z_0 + z^{(1-\alpha)}}{1 - a(z_0 + z^{(1-\alpha)})} + z_0\Big)\bigg)\bigg) \tag{63}$$

$$\bigg(F_{\widehat{\theta}}^{-1}\bigg(\Phi\Big(\frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})} + z_0\Big)\bigg)\,;\; +\infty\bigg) \tag{64}$$

Poznámka 3.7. Stanovení opraveného intervalu spolehlivosti, lze interpretovat jako opravu argumentu kvantilové funkce $F_{\widehat{\theta}}^{-1}(.)$, což je zřejmé ze srovnání výchozího inervalu spolehlivosti u neopravené percetilové metody, s výsledným. Např. jednostrannému intervalu

$$\Big(\widehat{F}_{\widehat{\theta}}^{-1}(\alpha)\,;\,+\infty\Big) \tag{65}$$

odpovídá interval (64). Argumentem je u prvního z nich α, a ta se změní na

$$\Phi\Big(\frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})} + z_0\Big)\,, \qquad \text{což označme jako} \quad \Phi\big(z[\alpha]\big)\,. \tag{66}$$

3.3. **Aproximace.** Poznámku 3.7 můžeme dále rozvést, uvědomíme-li si, že α v intervalu (65) je zároveň doplňkem jeho požadované spolehlivosti do jedné. Odtud $1 - \Phi\big(z[\alpha]\big)$ odpovídá spolehlivostní hladině intervalu (64). Jak jsme již zmínili na počátku části 3.2 je náš původní problém opačný: k dané oblasti hledáme její spolehlivost. Vrátíme-li se k analogii s příkladem 3.1, má tato oblast (ve zmíněném príkladu alternativa) tvar $(\pi_0; +\infty)$ a odpovídá tedy situaci (64). Proto:

$$\alpha_E \doteq 1 - \Phi\big(z[\alpha]\big) = \Phi\big(-z[\alpha]\big)\,. \tag{67}$$

Zde narazíme na problém, jak odhadnout α. Nabízí se využít předchozí odhad hladiny spolehlivosti α_F, který umíme snadno spočíst. Ve shodě s výše zmíněným pozorováním o α by mělo platit:

$$\alpha \doteq 1 - \alpha_F\,. \tag{68}$$

Upravme dále:

$$\begin{aligned} z[\alpha] \doteq z[1-\alpha_F] &= \frac{z_0 + z^{(1-\alpha_F)}}{1 - a(z_0 + z^{(1-\alpha_F)})} + z_0 \frac{z_0 - z^{(\alpha_F)}}{1 - a(z_0 - z^{(\alpha_F)})} + z_0 = \\ &= -\frac{z^{(\alpha_F)} - z_0}{1 - a(z^{(\alpha_F)} - z_0)} + z_0\,. \end{aligned} \tag{69}$$

Odtud:

$$\alpha_E \doteq \Phi\Big(\frac{z^{(\alpha_F)} - z_0}{1 - a(z^{(\alpha_F)} - z_0)} - z_0\Big)\,, \tag{70}$$

kde Φ je distribuční funkce N(0,1),

$$z^{(\alpha_F)} = \Phi^{-1}(\alpha_{_F}) \tag{71}$$

$$a \quad z_0 = \Phi^{-1}\Big(\mathbf{P}_{\boldsymbol{\pi}_0}\big(\widehat{\pi}^{**} \in \mathcal{P}_1\big)\Big)\,, \tag{72}$$

„Urychlovací člen“ a spočteme v následující části vhodnou úpravou výrazu (62) (viz 76).

3.4. **Výpočet.** (1) Pro rozsah B_1 spočteme algoritmem popsaným v 2.3 hladinu $\alpha_{_F}$ pro danou větev. Odtud získáme $z^{(\alpha_F)}$ dle (71). Generování boot-populace ze sloupců matice $\boldsymbol{X}$ by zároveň šlo popsat jako vygenerování

$$\boldsymbol{Y}^*_1, \ldots, \boldsymbol{Y}^*_{B_1} \overset{iid}{\sim} \mathrm{Multi}_n(n, \boldsymbol{p})\,, \qquad \text{kde} \quad \boldsymbol{p} = \tfrac{1}{n}\big(1, \ldots, 1\big)'\,. \tag{73}$$

Vektory $\boldsymbol{Y}^*(b)$ jsou analogií vektorů pozorovaných četností $\boldsymbol{\Upsilon}$ (viz 8), v nichž se neuvažují nulové složky odpovídající „nepozorovaným“ vektorům $\boldsymbol{\xi}$ z prostoru $\mathcal{X}$. Mají tedy $n = 73$ složek. Stejným způsobem je svázán i vektor $\boldsymbol{p}$ se svým protějškem $\boldsymbol{\pi}$. Opět platí, že z vektoru pravděpodobností $\boldsymbol{p}^*_b = \boldsymbol{Y}^*_b/n$, kde $b = 1, \ldots, B_1$ lze vypěstovat stromy $\widehat{\Psi}^*_b$, podobně jako v (13).

(2) Zapamatujme si ty stromy, které se v dané větvi neshodují s původním stromem $\widehat{\Psi}$. Označme příslušné pravděpodobnostní vektory, které je reprezentují, jako $\boldsymbol{p}^j$, $j = 1, \ldots, J$. Víme o nich, že nemohou náležet do $\mathcal{P}_1$, proto je můžeme využít k nalezeni hraničních vektorů $\boldsymbol{p}^j_0$, analogických k $\boldsymbol{\pi}_0$. Nejprve zkusíme, zda bod ležící v polovině úsečeky mezi body $\boldsymbol{p}$ a $\boldsymbol{p}^j$ patří do $\mathcal{P}_1$. Pokud ano, nahradíme tímto bodem bod $\boldsymbol{p}$ a pokračujeme s půlením této nové úsečky. Pokud ne, budeme

novým bodem nahrazovat bod $\boldsymbol{p}^j$. Pokračujeme-li takto u každého $j = 1, \ldots, J$ až do požadované přesnosti (počet kroků označuji ve výsledcích parametrem L), obdržíme J-tici hraničních bodů $\boldsymbol{p}_0^j$.

(3) Nyní už nic nebrání nagenerovat pro $j = 1, \ldots, J$:

$$\boldsymbol{Y}_{j1}^{**}, \ldots, \boldsymbol{Y}_{jB_2}^{**} \overset{iid}{\sim} \mathrm{Multi}_n(n, \boldsymbol{p}_0^j) \tag{74}$$

a zjistit počty stromů vypěstovaných z $\boldsymbol{Y}_{j1}^{**}, \ldots, \boldsymbol{Y}_{jB_2}^{**}$ souhlasících v dané větvi s $\widehat{\Psi}$, označme je W_j. Průměrem z pozorovaných relativních četností W_j/B_2 odhadneme $\mathbf{P}_{\boldsymbol{\pi}_0}\big(\widehat{\pi}^{**} \in \mathcal{P}_1\big)$, odtud dle (72)

$$z_0 = \Phi^{-1}\left(\frac{\sum_j W_j}{J \cdot B_2}\right). \tag{75}$$

(4) Poslední veličinou, která nám zbývá je urychlovací člen a. Závisí na směru vektoru spojujícího vektor $\boldsymbol{p}$ a $\partial\mathcal{P}_1$. Položme tedy $\boldsymbol{u}_j = \boldsymbol{p}^j - \boldsymbol{p}$ pro $j = 1, \ldots, J$

$$a(\boldsymbol{u}) = \frac{\sum_{k=1}^n u_k^3}{6\left(\sum_{k=1}^n u_k^2\right)^{\frac{3}{2}}} \tag{76}$$

Odvození tohoto vztahu lze nalézt v [1], jedná se vlastně o rozpis vztahu (62) pro případ exponenciálního systému hustot. Výrazy $\boldsymbol{u}_j$ odpovídají empirickým influenčním funkcím pro odhad vektoru $\boldsymbol{\pi}$, tedy střední hodnoty rozdělení $\mathcal{L}\big(\widehat{\boldsymbol{\pi}}\,|\boldsymbol{\pi}\big)$ – viz (18). V našem případě vezměme $a = J^{-1}\sum_j a(\boldsymbol{u}_j)$

(5) Dle vztahu (70) můžeme nyní dopočíst α_{E}.

3.5. **Výsledky.** V tabulce jsou větve číslovány vzestupně podle pořadí, v němž se slučovaly. Obě metody se výrazně liší v odhadu hladiny spolehlivosti především u větví č. 7, 8 a 9, tj. u větví, které, soudě dle nízké hodnoty α_{F}, mají při bootstrapu větší tendenci se přeskupovat.

	α_{F}				α_{E}		
	B			B_1	**2000**	**2000**	**10000**
č. větve	**2000**	**5000**	**10000**	B_2	**40**	**200**	**50**
1	0. 796	0. 811	0. 891		0. 727	0. 754	0. 751
2	1. 000	1. 000	1. 000		1. 000	1. 000	1. 000
3	0. 958	0. 958	0. 954		0. 912	0. 904	0. 878
4	0. 848	0. 842	0. 841		0. 787	0. 765	0. 768
5	0. 666	0. 673	0. 670		0. 737	0. 724	0. 723
6	0. 752	0. 736	0. 739		0. 745	0. 739	0. 716
7	0. 297	0. 318	0. 308		0. 734	0. 745	0. 752
8	0. 357	0. 347	0. 357		0. 740	0. 750	0. 742
9	0. 387	0. 386	0. 386		0. 725	0. 713	0. 726
10	0. 672	0. 641	0. 653		0. 718	0. 696	0. 705
11	0. 630	0. 618	0. 620		0. 712	0. 714	0. 706

Tab. 1 *Srovnání hladin,* $L = 20$

Literatura

[1] Efron, B.: Better bootstrap confidence intervals. JASA, 82, 171–185, 1987.
[2] Efron, B.: The Jacknife, the Bootstrap, and Other Resampling Plans. CBMS, 38, SIAM-NSF, 1982.
[3] Efron, B., Halloran, E., Holmes, S.: Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA, 93, 13429–13434, 1996.
[4] Escalante, A. A., Grebert, H. M., Chaiyaroj, S. C., Magris, M., Biswas, S., Nahlen, B. L., Lal A. A.: Polymorphism in the gene encodindg the apical membrane antigen-1 (AMA-1) of Plasmodium falciparum. X. Asembo Bay Cohort Project. Molecular & Biochemical Parasitology, 113, 279–287, 2001.
[5] Robert, C. P.: The Bayesian Choice. Springer-Verlag, New York, 1996.

KPMS MFF UK Praha, Sokolovská 83, 186 75 Praha 8 - Karlín
E-mail: betinec@karlin.mff.cuni.cz