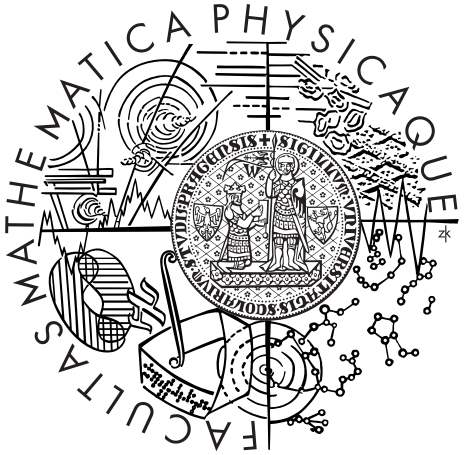
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Dominik Marko

Metoda bootstrap v Markovových řetězcích

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Zuzana Prášková, CSc.
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2014

Rád by som poďakoval doc. RNDr. Zuzane Práškovej, CSc., vedúcej bakalárskej práce, za pomoc, cenné rady, poskytnuté materiály, trpezlivosť a ochotu a množstvo venovaného času pri vypracovaní bakalárskej práce.
Taktiež by som chcel poďakovať svojim rodičom za podporu počas celého štúdia.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Metoda bootstrap v Markovových řetězcích

Autor: Dominik Marko

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zuzana Prášková, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zabýváme odhadováním pravděpodobností přechodu v Markovových řetězcích s konečnou množinou stavů a diskrétním časem. Použijeme dvě metody, konkrétně metodu maximální věrohodnosti a metodu bootstrap, pro získaní odhadů těchto pravdepodobností přechodu a odvodíme asymptotické rozdělení takto získaných odhadů. Popíšeme základní charakteristiky metody bootstrap a ukážeme aplikaci dvou bootstrapových metod pro odhadování pravděpodobností přechodů, konkrétně podmíněný a standardní bootstrap. Na numerické studii ukážeme výsledky aplikace jednotlivých metod pro odhadování pravděpodobností přechodu a výpočet intervalů spolehlivosti a porovnáme s výsledky založenými na asymptotické normalitě.

Klíčová slova: Markovovy řetězce, metoda bootstrap, matice pravděpodobností přechodu, intervaly spolehlivosti

Title: Boostrapping Markov Chains

Author: Dominik Marko

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zuzana Prášková, CSc., Department of Probability and Mathematical Statistics

Abstract: In this thesis we deal with estimating the transition matrix probabilities of discrete time Markov chains with finite state space. We will use two methods, specifically the maximum likelihood method and the bootstrap method, for obtaining estimators of these matrix probabilities and then we will develop the asymptotic distribution of these estimators. We will describe the basic characteristics of the bootstrap method and show the application of two bootstrap methods used for estimating transition probabilites, specifically conditional bootstrap and standard bootstrap. The results of the application of every method used for obtaining transition probabilities and computing confidence intervals will be presented in a numerical study and compared with the results based on asymptotic normality.

Keywords: Markov chains, bootstrap method, transition probability matrix, confidence intervals

Názov práce: Metóda bootstrap v Markovových reťazcoch

Autor: Dominik Marko

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. RNDr. Zuzana Prášková, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V tejto práci sa zaoberáme odhadovaním pravdepodobností prechodu v Markovových reťazcoch s konečnou množinou stavov a diskrétnym časom. Použijeme dve metódy, konkrétne metódu maximálnej vierohodnosti a metódu bootstrap, pre získanie odhadov týchto pravdepodobností prechodu a odvodíme asymptotické rozdelenie takto získaných odhadov. Popíšeme základné charakteristiky metódy bootstrap a ukážeme aplikáciu dvoch boostrapových metód pre odhadovanie pravdepodobností prechodu, konkrétne podmienený a štandardný bootstrap. Na numerickej štúdii ukážeme výsledky aplikácie jednotlivých metód pre odhadovanie pravdepodobností prechodu a výpočet intervalov spoľahlivosti a porovnáme s výsledkami založenými na asymptotickej normalite.

Kľúčové slová: Markovove reťazce, metóda bootstrap, matica pravdepodobností prechodu, intervaly spoľahlivosti

Obsah

Úvod

Markovove reťazce obvykle slúžia k opisu náhodných procesov s diskrétnymi stavmi a diskrétnym časom, pre ktoré platí, že pravdepodobnosť prechodu do nasledujúceho stavu $n + 1$ závisí iba na súčasnom stave n, v ktorom sa reťazec nachádza a nezávisí na minulých stavoch. Používajú sa napríklad vo finančnej sfére pri modelovaní zmien ratingov, v genetike, pri predpovedaní počasia a v mnohých ďalších oblastiach reálneho života. V praxi sa stáva, že pravdepodobnosti prechodu nepoznáme, ale máme k dispozícii dáta z minulosti, na základe ktorých by sme radi tieto pravdepodobnosti odhadli a zistili ich intervaly spoľahlivosti.

Cieľom tejto práce bude ukázať metódy vhodné na získanie neznámych pravdepodobností prechodu a ich intervalov spoľahlivosti.

V kapitole 1 zavedieme základné pojmy, definície a tvrdenia týkajúce sa Markovových reťazcov.

V druhej kapitole si ukážeme výpočet odhadov pravdepodobností prechodu metódou maximálnej vierohodnosti a popíšeme asymptotické rozdelenie takto získaných odhadov.

V tretej kapitole uvedieme základné charakteristiky metódy bootstrap. Táto metóda je vhodná na odhadovanie parametrov a nájdenie konfidenčných intervalov pre parametre, ktoré su funkciou neznámeho rozdelenia v situácii, kedy máme náhodný výber z tohto rozdelenia. Preto ju môžeme použiť aj na odhadnutie pravdepodobností prechodu, ak máme k dispozícii realizáciu Markovovho reťazca.

Vo štvrtej kapitole popíšeme dve bootstrapové metódy vhodné pre odhadovanie pravdepodobností prechodu, konkrétne sa jedná o podmienený a štandardný bootstrap. Ukážeme asymptotické rozdelenie takto získaných odhadov a porovnáme s asymptotickým rozdelením maximálne vierohodných odhadov.

Na numerickej štúdii ukážeme aplikáciu spomenutých metód na odhady pravdepodobností prechodu a výpočet konfidenčných intervalov pre tieto odhady. Dáta získame vygenerovaním Markovovho reťazca v programe Wolfram Mathematica 9.0 Student Edition. Porovnáme jednotlivé metódy a určíme záver našej štúdie.

Použité značenie

$\boldsymbol{\Pi}$	stacionárne rozdelenie Markovovho reťazca
S	množina stavov Markovovho reťazca
$\boldsymbol{P}$	matica pravdepodobností prechodu
$\widehat{\boldsymbol{P}}$	maximálne vierohodný odhad matice $\boldsymbol{P}$
n_i	počet prechodov zo stavu i v realizácii Markovovho reťazca
n_{ij}	počet prechodov zo stavu i do stavu j v realizácii Markovovho reťazca
$Mult_K(n,\boldsymbol{p})$	multinomické rozdelenie s K priehradkami a parametrami n a $\boldsymbol{p}$
δ_{ij}	Kroneckerovo delta
$\boldsymbol{I_n}$	matica $n \times n$, ktorej prvky na diagonále sú 1 a prvky mimo diagonály sú 0
$Diag\{a_1,\ldots,a_n\}$	diagonálna matica, ktorej prvky na diagonále sú $a_1,\ldots,a_n$ a prvky mimo diagonály sú 0
$N_K(\boldsymbol{\mu},\boldsymbol{\Sigma})$	K-rozmerné normálne rozdelenie so strednou hodnotou $\boldsymbol{\mu}$ a kovariančnou maticou $\boldsymbol{\Sigma}$
$\stackrel{D}{=}$	rovnosť distribučných funkcií
$\stackrel{D}{\rightarrow}$	konvergencia v distribúcii
$\stackrel{P}{\rightarrow}$	konvergencia v pravdepodobnosti

Kapitola 1

Základné definície a tvrdenia o Markovových reťazcoch

V tejto kapitole zavedieme základné pojmy, definície a vety týkajúce sa Markovových reťazcov podľa Prášková a Lachout (2012).

Definícia 1 (Markovov reťazec)**.** *Postupnosť celočíselných náhodných veličín* $\{X_n, n \in N_0\}$ *sa nazýva Markovov reťazec s diskrétnym časom a množinou stavov S, ak*

$$P(X_{n+1} = j|X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j|X_n = i) \quad (1.1)$$

pre všetky $n = 0,1\ldots$ *a všetky* $i,j,i_{n-1},\ldots,i_0 \in S$ *také,že* $P(X_n = i,X_{n-1} = i_{n-1},\ldots,X_0 = i_0) > 0$. *O množine stavov S predpokladáme, že* $i \in S$ *práve vtedy, ak existuje* $n \in \mathbb{N}_0$ *také, že* $P(X_n = i) > 0$.

Vzťah (1.1) vyjadruje markovskú vlastnosť; znamená, že pravdepodobnosť výsledku v budúcom čase n+1, ak poznáme výsledok v prítomnom čase n a výsledky z minulých časov $n-1,n-2,\ldots,0$ je rovnaká, ako keď poznáme výsledok len v prítomnom čase.

Podmienené pravdepodobnosti

$$P(X_{n+1} = j|X_n = i) = p_{ij}(n,n+1)$$

nazývame *pravdepodobnosti prechodu* zo stavu i v čase n do stavu j v čase $n+1$, niekedy pravdepodobnosti prechodu 1. rádu. Podmienené pravdepodobnosti

$$P(X_{n+m} = j|X_n = i) = p_{ij}(n,n+m)$$

pre prirodzené $m \geq 1$ sa nazývajú pravdepodobnosti prechodu m-tého rádu. Markovov reťazec sa nazýva *homogénny*, ak pravdepodobnosti prechodu $p_{ij}(n,n+m)$ nezávisia na časových okamihoch n a $n+m$, ale len na ich rozdiele m. V takomto prípade ich budeme značiť $p_{ij}^{(m)}$.

Uvažujme homogénny Markovov reťazec $\{X_n\}$. Pravdepodobnosti prechodu $P(X_{n+1} = j|X_n = i)$ sú v tomto prípade nezávislé na čase n, označíme ich p_{ij} a vynecháme prívlastok 1. rádu. Všetky tieto pravdepodobnosti môžeme zostaviť do štvorcovej matice $\boldsymbol{P} = \{p_{ij},i,j \in S\}$, keďže p_{ij} sú definované pre všetky $i,j \in S$. Zrejme pre každé $n \in N_0$ platí

$$p_{ij} \geq 0, i,j \in S; \sum_{j \in S} p_{ij} = 1, i \in S. \quad (1.2)$$

Štvorcová matica majúca vlastnosť (1.2) sa nazýva stochastická matica, matica $\boldsymbol{P}$ je teda stochastická a nazýva sa matica pravdepodobností prechodu.

Označme

$$p_i = P(X_0 = i), i \in S.$$

Platí:

$$p_i \geq 0, i \in S; \sum_{i \in S} p_i = 1. \tag{1.3}$$

Pravdepodobnostné rozdelenie $\boldsymbol{p} = \{p_i, i \in S\}$ sa nazýva *počiatočné rozdelenie* Markovovho reťazca.

Veta 1. *Nech* $\{X_n,\ n \in \mathbb{N}_0\}$ *je náhodný proces s množinou stavov* $S = \{0,1,\dots\}$, *nech* $\boldsymbol{p}$ *je vektor spĺňajúci (1.3) a* $\boldsymbol{P} = \{p_{ij}, i,j \in S\}$ *je matica, ktorá spĺňa (1.2). Potom* $\{X_n,\ n \in \mathbb{N}_0\}$ *je homogénny Markovov reťazec s počiatočným rozdelením* $\boldsymbol{p}$ *a maticou pravdepodobností prechodu* $\boldsymbol{P}$ *vtedy a len vtedy, keď všetky konečnerozmerné rozdelenia tohto procesu sú v tvare*

$$P(X_0 = i_0, X_1 = i_1, \dots, X_k = i_k) = p_{i_0} p_{i_0 i_1} \dots p_{i_{k-1} i_k}$$

pre všetky $i_0, i_1, \dots, i_k \in S$ *a všetky* $k \in \mathbb{N}_0$.

Dôkaz vety 1 je uvedený v (Prášková a Lachout (2012), veta 2.1).

Definícia 2. *Stav Markovovho reťazca j je dosiahnuteľný zo stavu i, ak existuje* $n \in \mathbb{N}_0$ *také, že* $p_{ij}^{(n)} > 0$. *Reťazec, ktorého všetky stavy sú vzájomne dosiahnuteľné, sa nazýva* nerozložiteľný.

Definícia 3 (Trvalý nenulový stav)**.** *Položme*

$$\tau_j(1) = \inf\{n > 0 : X_n = j\}$$

s konvenciou $\inf\{\emptyset\} = \infty$. *Označme* $P(.|X_0 = j) = P_j(.)$. *Stav j Markovovho reťazca sa nazýva trvalý, ak reťazec, ktorý vychádza z j, sa do j vráti s pravdepodobnosťou 1 po konečne mnoho krokoch, tj.*

$$P_j(\tau_j(1) < \infty) = 1.$$

Trvalý stav definujeme ako nenulový, ak $E_j(\tau_j) < \infty$, *kde* $E_j(.) = E(.|X_0 = j)$.

Definícia 4. *Nech* d_j *je najväčší spoločný deliteľ čísel* $n \geq 1$, *pre ktoré* $p_{jj}^{(n)} > 0$. *Ak je* $d_j > 1$, *hovoríme, že stav j je* periodický *s periódou* d_j, *ak* $d_j = 1$, *hovoríme, že stav j je* neperiodický.

Definícia 5 (Stacionárne rozdelenie)**.** *Nech* $\{X_n, n \in N_0\}$ *je homogénny reťazec s množinou stavov S a maticou pravdepodobností prechodu* $\boldsymbol{P}$. *Nech* $\boldsymbol{\Pi} = \{\pi_j, j \in S\}$ *je nejaké pravdepodobnostné rozdelenie na množine S, tj.* $\pi_j \geq 0, j \in S, \sum_{j \in S} \pi_j = 1$. *Potom* $\boldsymbol{\Pi}$ *sa nazýva* stacionárne rozdelenie *daného reťazca, ak platí*

$$\boldsymbol{\Pi}^\intercal = \boldsymbol{\Pi}^\intercal \boldsymbol{P} \tag{1.4}$$

alebo

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, j \in S,$$

keď uvažujeme stĺpcové vektory.

Veta 2. *V nerozložiteľnom reťazci s konečne mnoho stavmi sú všetky stavy trvalé nenulové.*

Veta 3. *Ak sú všetky stavy nerozložiteľného Markovovho reťazca trvalé nenulové, potom stacionárne rozdelenie existuje a je jediné.*

Dôkazy viet 2 a 3 sú uvedené v (Prášková a Lachout, 2012).

Veta 4. *Nech* $\{X_n\}$ *je Markovov reťazec s konečne mnoho stavmi a maticou pravdepodobností prechodu P, kde* $p_{ij} > 0 \ \forall \ i,j \in S$. *Potom sú všetky stavy trvalé, nenulové a neperiodické.*

Dôkaz. Z $p_{ij} > 0$ pre všetky i,j vyplýva, že všetky stavy sú vzájomne dosiahnuteľné, teda reťazec je nerozložiteľný. Podľa vety 2 sú potom všetky stavy trvalé nenulové. Keďže $p_{ij} > 0 \ \forall \ i,j \in S$, tak aj $p_{jj} > 0$ pre všetky $j \in S$, z čoho vyplýva, že všetky stavy sú neperiodické.

Kapitola 2

Odhady pravdepodobností prechodov

2.1 Maximálne vierohodný odhad pravdepodobností prechodu

Nech $\boldsymbol{X} = \{X_1,X_2,\ldots,X_n; n \in \mathbb{N}\}$ je realizáciou homogénneho nerozložiteľného neperiodického Markovovho reťazca pozorovaného do času n s konečnou množinou stavov $S = \{1,2,\ldots,K\}$ a nami neznámou maticou pravdepodobností prechodu $\boldsymbol{P} = (p_{ij}, i,j = 1,\ldots,K)$, ktorú chceme odhadnúť na základe dát (realizácie Markovovho reťazca). Taktiež chceme zistiť asymptotické rozdelenie tohto odhadu. V tejto práci budeme vždy uvažovať homogénny nerozložiteľný neperiodický Markovov reťazec. Budeme predpokladať, že $p_{ij} > 0 \ \forall \ i,j \in S$, čo nám podľa vety 4 zaručí všetky tieto požadované vlastnosti Markovovho reťazca.

Pravdepodobnosť realizácie $(X_1,\ldots,X_n)$ je podľa vety 1

$$P(X_1 = a_1,X_2 = a_2,\ldots,X_n = a_n) = p_{a_1}p_{a_1a_2}\ldots p_{a_{n-1}a_n} = p_{a_1}\prod_{i=1}^{K}\prod_{j=1}^{K}p_{ij}^{n_{ij}}, \quad (2.1)$$

kde $p_{a_1} = P(X_1 = a_1)$ a n_{ij} je pozorovaný počet prechodov zo stavu i do stavu j v reťazci $\{X_1,\ldots,X_n\}$, tj. $n_{ij} = \sum_{k=1}^{n-1} I[X_k = i,X_{k+1} = j]$. Pozorovaný počet prechodov zo stavu i v reťazci $\{X_1,\ldots,X_n\}$ označíme ako n_i, tj. $n_i = \sum_{j=1}^{n-1} I[X_j = i] = \sum_{j=1}^{K} n_{ij}$.

Odhad $\widehat{\boldsymbol{P}}$ matice $\boldsymbol{P}$ spočítame metódou maximálnej vierohodnosti (anglicky MLE-maximum likelihood estimation). Vierohodnostná funkcia vyzerá nasledovne:

$$L(p_{ij},\ i = 1,\ldots,K,\ j = 1,\ldots,K) = p_{a_1}\prod_{i=1}^{K}\prod_{j=1}^{K}p_{ij}^{n_{ij}}.$$

Jej zlogaritmovaním získame logaritmickú vierohodnostnú funkciu

$$\log L(p_{ij},\ i = 1,\ldots,K,\ j = 1,\ldots,K) = \log p_{a_1} + \sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij}\log p_{ij}.$$

Keďže maximálne vierohodný odhad je taký odhad, ktorý maximalizuje hodnotu vierohodnostnej funkcie a keďže $\log L$ nadobúda maximum v rovnakom bode ako L (pretože logaritmus je rýdzo rastúca funkcia), budeme chcieť maximalizovať logaritmickú vierohodnostnú funkciu vzhľadom k $p_{ij},\ i,j = 1,\dots,K$ za podmienok $\sum_{j=1}^{K} p_{ij} = 1$ pre všetky $i = 1,\dots,K$. Použijeme metódu Lagrangeových multiplikátorov, chceme teda nájsť maximum funkcie

$$h(p_{ij},\ i,j = 1,\dots,K; \lambda_1,\dots,\lambda_K) = \log p_{a_1} + \sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij} \log p_{ij} - \sum_{i=1}^{K} \lambda_i (\sum_{j=1}^{K} p_{ij} - 1). \tag{2.2}$$

Deriváciou funkcie (2.2) podľa jednotlivých $p_{ij},\ i,j = 1,\dots K$ a λ_i, $i = 1,\dots K$, a položením každej derivácie rovnej nule dostávame sústavu rovníc

$$\frac{\partial h}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} - \lambda_i = 0,\ \ i,j = 1,\dots,K,$$
$$\frac{\partial h}{\partial \lambda_i} = \sum_{j=1}^{K} p_{ij} - 1 = 0,\ \ i = 1,\dots,K.$$

Úpravou získame

$$p_{ij} = \frac{n_{ij}}{\lambda_i} \text{ pre } i,j = 1,\dots K.$$

Sčítaním tejto sústavy rovníc cez všetky $j = 1,\dots,K$ nám vychádza

$$\sum_{j=1}^{K} p_{ij} = \frac{n_i}{\lambda_i}.$$

Využitím faktu, že $\sum_{j=1}^{K} p_{ij} = 1,\ \ i = 1,\dots,K$, dostaneme $\lambda_i = n_i$ pre $i = 1,\dots,K$. Maximálne vierohodný odhad $\widehat{p}_{ij}$ pravdepodobností prechodu p_{ij} teda vychádza

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i},\ \ i,j = 1,\dots,K. \tag{2.3}$$

Aby sme ošetrili prípady, kedy $n_i = 0$, definujeme odhad p_{ij} ako

$$\widehat{p}_{ij} = \begin{cases} \frac{n_{ij}}{n_i}, & \text{ak je } n_i > 0,\ \ i,j = 1,\dots,K, \\ \delta_{ij}, & \text{ak } n_i = 0,\ i,j = 1,\dots,K. \end{cases}$$

2.2 Asymptotické rozdelenie odhadu matice pravdepodobností prechodu

Radi by sme zistili, aké je asymptotické rozdelenie $\sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P})$.
Definujme $\xi_{ij} = \frac{n_{ij} - n_i p_{ij}}{\sqrt{n_i}}$, $i,j = 1,\dots,K$ a vektor $\boldsymbol{\xi_i} = (\xi_{i1},\dots,\xi_{iK})^\intercal$, $i = 1,\dots,K$.

Veta 5. *Rozdelenie K^2-rozmerného náhodného vektora $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\intercal, \ldots, \boldsymbol{\xi}_K^\intercal)^\intercal$ konverguje pre $n \to \infty$ ku K^2-rozmernému normálnemu rozdeleniu s kovariančnou maticou $\boldsymbol{\Sigma}$, tj. $\boldsymbol{\xi} \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0}, \boldsymbol{\Sigma})$, kde*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Sigma_K} \end{pmatrix}, \tag{2.4}$$

pričom

$$\boldsymbol{\Sigma_i} = \begin{pmatrix} (1-p_{i1})\,p_{i1} & -p_{i1}p_{i2} & \cdots & -p_{i1}p_{iK} \\ -p_{i2}p_{i1} & (1-p_{i2})\,p_{i2} & \cdots & -p_{i2}p_{iK} \\ \vdots & \ddots & \ddots & \vdots \\ -p_{iK}p_{i1} & \cdots & -p_{iK}p_{iK-1} & (1-p_{iK})\,p_{iK} \end{pmatrix}.$$

V nasledujúcich riadkoch tejto práce bude $\boldsymbol{\Sigma}$ vždy označovať variančnú maticu definovanú ako (2.4).

Predtým, ako dokážeme vetu 5, si uvedieme dve vety, ktoré pri jej dôkaze využijeme.

Veta 6. *V nerozložiteľnom reťazci s trvalými nenulovými stavmi platí*

$$\frac{n_i}{n} \to \pi_i, \text{ pri } n \to \infty \text{ s pravdepodobnosťou 1 pre každé } i \in S, \tag{2.5}$$

kde π_j značí j-tú zložku stacionárneho rozdelenia.

Dôkaz vety 6 sa nachádza v (Prášková a Lachout (2012), veta 2.28).

Veta 7. *Nech $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\intercal \sim Mult_K(n, \boldsymbol{p})$, kde $\boldsymbol{p} = (p_1, \ldots, p_K)$. Potom*

$$\frac{1}{\sqrt{n}}(\boldsymbol{Z} - n\boldsymbol{p}) \xrightarrow[n\to\infty]{D} N_K(\mathbf{0}, \overline{\boldsymbol{\Sigma}}),$$

kde $\overline{\boldsymbol{\Sigma}}$ je matica, ktorá má na diagonále prvky $\sigma_{jj} = p_j(1-p_j)$, $j = 1, \ldots, K$ a mimo diagonály prvky $\sigma_{ij} = -p_i p_j$, $1 \le i, j \le K$, $i \ne j$.

Dôkaz. Nech $\boldsymbol{X} = (X_1, \ldots, X_K)^\intercal \sim M_K(n; p_1, \ldots, p_K)$. Položme

$$Y_i = \frac{X_i - np_i}{\sqrt{np_i}}, \quad i = 1, \ldots, K, \quad \boldsymbol{Y} = (Y_1, \ldots, Y_K)^\intercal.$$

Potom podľa vety 12.4 v (Anděl, 2011) pre $n \to \infty$ platí $\boldsymbol{Y} \xrightarrow{D} N_K(\mathbf{0}, \boldsymbol{Q})$, kde matica $\boldsymbol{Q} = \boldsymbol{I_K} - \boldsymbol{u}\boldsymbol{u}^\intercal$ a vektor $\boldsymbol{u} = (\sqrt{p_1}, \ldots, \sqrt{p_K})^\intercal$. Položme $Z_i = \frac{X_i - np_i}{\sqrt{n}} = \sqrt{p_i} Y_i$ a $\tilde{\boldsymbol{D}} = Diag\{\sqrt{p_1}, \ldots, \sqrt{p_K}\}$. Potom $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\intercal = \tilde{\boldsymbol{D}}\boldsymbol{Y}$. Zo vzťahu $\boldsymbol{Y} \xrightarrow{D} N_K(\mathbf{0}, \boldsymbol{Q})$ máme $\boldsymbol{Z} \xrightarrow{D} N_K(0, \tilde{\boldsymbol{D}}\boldsymbol{Q}\tilde{\boldsymbol{D}}^\intercal)$, pričom $\tilde{\boldsymbol{D}}\boldsymbol{Q}\tilde{\boldsymbol{D}}^\intercal = \overline{\boldsymbol{\Sigma}}$.

Dôkaz vety 5 (spracovaný podľa (Billingsley, 1961)). Dôkaz je založený na reprezentácii Markovovho reťazca pomocou nezávislých náhodných veličín.

O Markovovom reťazci $\{X_t,\ t = 1,\ldots,n\}$ môžeme predpokladať, že bol generovaný nasledujúcim spôsobom. Majme nezávislý výber náhodných veličín X_1 a $w_{in},\ i = 1,2,\ldots,K,\ n = 1,2,\ldots$ takých, že

$$P[X_1 = i] = \pi_i \text{ a } P[w_{in} = j] = p_{ij}.$$

Predstavme si veličiny w_{in} usporiadané do poľa nasledovne:

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{\mathrm{K1}} & w_{\mathrm{K2}} & \cdots & w_{\mathrm{Kn}} \end{pmatrix}.$$

Najprv sa vytvorí zložka X_1. X_2 sa následne definuje ako $w_{x_1 1}$ a ďalšie zložky vytvárame rekurentne, teda ak máme definované $X_1,\ldots,X_t$, potom X_{t+1} je definované ako $w_{x_t m}$, kde $m-1$ je počet takých l, $1 \le l < t$, pre ktoré $X_l = X_t$. Potom platí

$$\begin{aligned} P[X_k = a_k,\ 1 \le k \le n,\ a_k \in S] &= P[X_1 = a_1, w_{a_{k-1} m_k} = a_k,\ 2 \le k \le n,\ a_k \in S] \\ &= \pi_{a1} p_{a_1 a_2} \ldots p_{a_{n-1} a_n}, \end{aligned} \tag{2.6}$$

kde prvá rovnosť vyplýva z definície procesu na základe náhodných veličín w_{in} a druhá z nezávislosti náhodných veličín X_1 a w_{in}.

Keďže proces vygenerovaný týmto spôsobom má združené rozdelenie dané (2.6), ktoré zodpovedá združenému rozdeleniu veličín Markovovho reťazca podľa vety 1, môžeme ho použiť na výpočet rozdelení n_{ij}. Z definície jednotlivých w_{in} je zrejmé, že z náhodných veličín$(w_{i1},\ldots,w_{in_i})$ môžeme získať náhodné veličiny $(n_{i1},\ldots,n_{iK})$ nasledovne

$$n_{ij} = \sum_{k=1}^{n_i} I[w_{ik} = j] \text{ pre } i,j = 1,\ldots,K.$$

Pretože sa zaoberáme asymptotickým rozdelením, na základe (2.5) môžeme náhodné veličiny $(n_{i1},\ldots,n_{iK})$ porovnať s náhodnými veličinami $(g_{i1},\ldots,g_{iK})$, ktoré získame totožným spôsobom z $(w_{i1},\ldots,w_{i[n\pi_i]})$. Definujme náhodný vektor $\boldsymbol{\nu} = (\boldsymbol{\nu_1}^\intercal,\ldots,\boldsymbol{\nu_K}^\intercal)^\intercal$, kde $\boldsymbol{\nu_i} = (\nu_{i1},\ldots,\nu_{iK})^\intercal$ pre $i = 1,\ldots,K$, obsahujúci K^2 náhodných veličín

$$\nu_{ij} = \frac{(g_{ij} - [n\pi_i]p_{ij})}{\sqrt{n\pi_i}},\ \ i,j = 1,\ldots,K.$$

Z nezávislosti poľa náhodných veličín $\{w_{in},\ i = 1,\ldots,K,\ n \ge 1\}$ a centrálnej limitnej vety pre multinomické výbery (viď veta 7) potom vyplýva

$$\boldsymbol{\nu} \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0},\boldsymbol{\Sigma}).$$

Podľa sekcie 20.6 v (Cramér, 1946) má K^2-rozmerný náhodný vektor $\boldsymbol{\eta} = (\eta_1^\intercal,\ldots,\eta_K^\intercal)^\intercal$, kde $\boldsymbol{\eta_i} = (\eta_{i1},\ldots,\eta_{iK})^\intercal$ pre $i = 1,\ldots,K$ a

$$\eta_{ij} = \frac{(n_{ij} - n_i p_{ij})}{\sqrt{n\pi_i}},$$

rovnaké asymptotické rozdelenie ako $\boldsymbol{\nu}$, ak ukážeme, že pre fixné i a j platí

$$\frac{g_{ij} - [n\pi_i]p_{ij}}{\sqrt{n}} - \frac{n_{ij} - n_i p_{ij}}{\sqrt{n}} \xrightarrow{P} 0.$$

Dôkaz tejto konvergencie v pravdepodobnosti je uvedený na stranách 20 a 21 v (Billingsley, 1961) a teda

$$\boldsymbol{\eta} \xrightarrow{D} N_{K^2}(\mathbf{0},\boldsymbol{\Sigma}).$$

Keďže platí $\xi_{ij} \xrightarrow{P} \eta_{ij}$ na základe (2.5), má (podľa sekcie 20.6 v (Cramér, 1946)) náhodný vektor $\boldsymbol{\xi}$ rovnaké asymptotické rozdelenie ako $\boldsymbol{\eta}$, tj. $\boldsymbol{\xi} \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0},\boldsymbol{\Sigma})$.

Z asymptotického rozdelenia $\boldsymbol{\xi}$ si ľahko odvodíme asymptotické rozdelenie pre $\sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P})$. Upravme si ξ_{ij} nasledovným spôsobom

$$\begin{aligned}\xi_{i,j} &= \frac{n_{ij} - n_i p_{ij}}{\sqrt{n_i}} = \frac{n_i(\frac{n_{ij}}{n_i} - p_{ij})}{\sqrt{n_i}} \\ &= \sqrt{n}\frac{\sqrt{n_i}}{\sqrt{n}}(\widehat{p}_{ij} - p_{ij}) = \frac{\sqrt{n_i}}{\sqrt{n}}(\sqrt{n}(\widehat{p}_{ij} - p_{ij})).\end{aligned}$$

Definujme $\vartheta_{ij} = \sqrt{n}(\widehat{p}_{ij} - p_{ij})$, tj. $\vartheta_{ij} = \sqrt{\frac{n}{n_i}}\xi_{i,j}$. Položme vektor

$$\boldsymbol{\vartheta} = (\vartheta_1^\intercal, \ldots, \vartheta_K^\intercal)^\intercal, \text{ kde } \boldsymbol{\vartheta_i} = (\vartheta_{i1}, \ldots, \vartheta_{iK})^\intercal,$$

a $K^2 \times K^2$-rozmernú maticu

$$\boldsymbol{L} = \begin{pmatrix} \sqrt{\frac{n}{n_1}}\boldsymbol{I_K} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \sqrt{\frac{n}{n_K}}\boldsymbol{I_K} \end{pmatrix}. \tag{2.7}$$

Platí $\boldsymbol{\vartheta} = \boldsymbol{L\xi}$. Z vety (2.5) a z vety o spojitej transformácii (viď Anděl (2011), veta B.9) vieme, že $\frac{n}{n_i} \xrightarrow{P} \frac{1}{\pi_i}$. Z toho vyplýva, že

$$\boldsymbol{L} \xrightarrow{P} \begin{pmatrix} \frac{1}{\sqrt{\pi_1}}\boldsymbol{I_K} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{\sqrt{\pi_K}}\boldsymbol{I_K} \end{pmatrix}.$$

Vieme, že $\boldsymbol{\xi} \xrightarrow{D} \bar{\boldsymbol{\xi}}$, kde $\bar{\boldsymbol{\xi}} \sim N_{K^2}(\mathbf{0},\boldsymbol{\Sigma})$.

Z Cramérovej-Sluckého (viď Anděl (2011), veta B.10) vety potom môžeme odvodiť

$$\boldsymbol{\vartheta} = \boldsymbol{L\xi} \xrightarrow[n\to\infty]{D} \begin{pmatrix} \frac{1}{\sqrt{\pi_1}}\boldsymbol{I_K} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{\sqrt{\pi_1}}\boldsymbol{I_K} \end{pmatrix} \bar{\boldsymbol{\xi}}.$$

To vedie na vzťah

$$\boldsymbol{\vartheta} = \sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P}) \xrightarrow[n\to\infty]{D} N_{K^2}(0,\tilde{\boldsymbol{\Sigma}}), \tag{2.8}$$

$$\text{kde } \tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{\boldsymbol{\Sigma_1}}{\pi_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \frac{\boldsymbol{\Sigma_2}}{\pi_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{\boldsymbol{\Sigma_K}}{\pi_K} \end{pmatrix}, \tag{2.9}$$

čím dostávame požadované asymptotické rozdelenie pre $\sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P})$.

Poznámka: V literatúre (Athreya a Fuh, 1992) je matica $\tilde{\boldsymbol{\Sigma}}$ uvedená chybne.

Kapitola 3

Základy metódy bootstrap

3.1 Popis metódy bootstrap

Metóda bootstrap patrí medzi intenzívne počítačové metódy pre štatistickú analýzu dát. Často sa používa na nájdenie štandardných chýb pre odhady, konfidenčných intervalov pre neznáme parametre alebo nájdenie p hodnôt pre testové štatistiky za platnosti nulovej hypotézy. V tejto kapitole vysvetlíme základné vlastnosti tejto metódy(podľa Prášková (2004)).

Majme nezávislé rovnako rozdelené náhodné veličiny $X_1,\ldots,X_n$, ktorých distribučná funkcia F nie je bližšie špecifikovaná. Nech $\theta = \theta(F)$ značí nejakú charakteristiku rozdelenia F. Tento parameter nepoznáme a chceme ho odhadnúť na základe realizácie náhodného výberu.

Nech $\theta_n = \theta(X_1,\ldots,X_n)$ je štatistika pre odhad parametru θ a štatistika $R_n = R_n(X_1,\ldots,X_n)$ je jej vhodne štandardizovaná verzia, napríklad $R_n = \sqrt{n}(\theta_n - \theta)$. Distribučnú funkciu R_n si označme ako:

$$H_n(x) = P[R_n(X_1,\ldots,X_n,F) \leq x].$$

V prípade, že nevieme explicitne vyjadriť rozdelenie H_n alebo je vyjadrenie príliž obtiažne, riešením môže byť použitie metódy bootstrap. Táto metóda kombinuje tzv. substitučný princíp a metódu Monte Carlo. Vysvetlíme si najprv substitučný princíp. Neznámu distribučnú funkciu F nahradíme nejakým jej odhadom F_n. Najčastejšie sa ako odhad používa empirická distribučná funkcia založená na náhodnom výbere $X_1,\ldots,X_n$, tj.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x].$$

Pri daných hodnotách $X_1,\ldots,X_n$ je F_n známa funkcia.

Nech $X_1^*,\ldots,X_n^*$ je nezávislý náhodný výber z F_n. Pri daných pozorovaniach $X_1,\ldots,X_n$ sú $X_1^*,\ldots,X_n^*$ (podmienene) nezávislé, rovnako rozdelené náhodné veličiny, z ktorých každá nadobúda hodnôt $X_1,\ldots,X_n$ s pravdepodobnosťou $\frac{1}{n}$. Súbor $X_1^*,\ldots,X_n^*$ sa nazýva *bootstrapový výber*. Nahradením pôvodného výberu bootstrapovým výberom a neznámej distribučnej funkcie F empirickou distribučnou funkciou F_n dostávame parameter $\theta^* = \theta(F_n)$ a štatistiku $R_n^* = R_n(X_1^*,\ldots,X_n^*,F_n)$.

Zadefinujeme tzv. teoretické charakteristiky

$$E^*R_n^* = \int R_n(x_1,\ldots,x_n)d(F_n(x_1)\ldots F_n(x_n)),$$

$$var^*R_n^* = \int [R_n(x_1,\ldots,x_n) - E^*R_n^*]^2 d(F_n(x_1),\ldots,F_n(x_n))$$

získané metódou bootstrap a teoretickú distribučnú funkciu

$$\begin{aligned} H_n^*(x) &= P^*(R_n(X_1^*,\ldots,X_n^*,F_n) \leq x) \\ &= P(R_n(X_1^*,\ldots,X_n^*,F_n) \leq x \mid X_1,\ldots,X_n) \end{aligned}$$

získanú metódou bootstrap.

Najčastejšie sa metóda používa tak, že na bootstrapový výber $X_1^*,\ldots,X_n^*$ a známu distribučnú funkciu F_n aplikujeme metódu Monte Carlo. Tá spočíva v tom, že sa mnohokrát (B-krát) generuje nezávislý nezávislý náhodný výber z rozdelenia F_n, zakaždým sa spočítajú hodnoty štatistík θ_n^*,R_n^*, stanoví sa z nich aritmetický priemer a dostaneme tak bootstrapové odhady pôvodného rozdelenia a pôvodných charakteristík.

Takto môžeme odhadnúť napríklad H_n, tj. distribučnú funkciu štatistiky R_n, ako:

$$\widehat{H}_n^*(x) = \frac{1}{B}\sum_{b=1}^{B} I[R_n(X_{1,b}^*,\ldots,X_{n,b}^*) \leq x],$$

kde $\{X_{1,b}^*,\ldots,X_{n,b}^*\}, b = 1,\ldots,B$, sú nezávislé výbery z F_n.

3.2 Konzistencia metódy bootstrap

Povieme, že $H_n^*(x)$ je konzistentný odhad $H_n(x)$, ak

$$\rho(H_n^*,H_n) \to 0 \text{ keď } n \to \infty$$

v pravdepodobnosti (slabá konzistencia) alebo skoro iste (silná konzistencia), kde ρ je nejaká metrika na priestore distribučných funkcií. Najčastejšie sa používa supremálna metrika:

$$\rho_\infty(G,H) = \sup_{x\in\mathbb{R}} |G(x) - H(x)|.$$

Obvykle sa konzistencia bootstrapového rozdelenia dokazuje nasledovným spôsobom. Dokážeme,že

$$\sup_{x\in\mathbb{R}} |H_n(x) - \phi(x)| \to 0 \text{ pre } n \to \infty$$

a

$$\sup_{x\in\mathbb{R}} |H_n^*(x) - \phi(x)| \to 0 \text{ pre } n \to \infty \text{ v pravdepodobnosti alebo skoro iste.}$$

Ak je toto splnené, bootstrapový odhad je potom konzistentný, pretože platí

$$\begin{aligned}\rho_\infty(H_n,H_n^*) &= \sup_{x\in\mathbb{R}} |H_n(x) - H_n^*(x)| \\ &\leq \sup_{x\in\mathbb{R}} \big(|H_n(x) - \phi(x)| + |H_n^*(x) - \phi(x)|\big) \\ &\leq \sup_{x\in\mathbb{R}} |H_n(x) - \phi(x)| + \sup_{x\in\mathbb{R}} |H_n^*(x) - \phi(x)| \to 0\end{aligned}$$

v pravdepodobnosti alebo skoro iste. Funkcia ϕ obvykle býva distribučná funkcia normálneho rozdelenia.

3.3 Bootstrapové intervaly spoľahlivosti

3.3.1 Studentizované intervaly spoľahlivosti

Označme $\widehat{\theta}_n$ odhad parametru θ a uvažujeme studentizovanú štatistiku

$$R_n = \frac{\widehat{\theta}_n - \theta}{S_n}$$

a jej bootstrapovú verziu

$$R_n^* = \frac{\widehat{\theta}_n^* - \theta^*}{S_n^*}.$$

Označme β_p p-kvantil distribučnej funkcie štatistiky R_n, β_p^* p-kvantil distribučnej funkcie štatistiky R_n^* spočítaný ako $\beta_p^* = R_{[B_p]}$, tj. výberový kvantil spočítaný z usporiadaného výberu $R_{(1)}^*,\ldots,R_{(B)}^*$. Potom interval spoľahlivosti pre θ s koeficientom $1-\alpha$ je

$$(\widehat{\theta}_n - \beta_{1-\frac{\alpha}{2}} S_n, \widehat{\theta}_n - \beta_{\frac{\alpha}{2}} S_n).$$

Bootstrapový interval spoľahlivosti s koeficientom $1-\alpha$ vyzerá

$$(\widehat{\theta}_n - \beta^*_{1-\frac{\alpha}{2}} S_n, \widehat{\theta}_n - \beta^*_{\frac{\alpha}{2}} S_n).$$

3.3.2 Percentilové intervaly

Táto metóda počíta intervalový odhad parametru θ pomocou kvantilov distribučnej funkcie neštandardizovanej štatistiky $\widehat{\theta}_n^*$, tj. z distribučnej funkcie $G_n^*(x) = P^*(\widehat{\theta}_n^* \leq x)$. Dostaneme intervalový odhad s koeficientom $1-\alpha$

$$\big(G_n^{*^{-1}}(\frac{\alpha}{2}), G_n^{*^{-1}}(1-\frac{\alpha}{2})\big).$$

3.3.3 Hybridné intervaly

Ak je $H_n(x) = P(\sqrt{n}(\widehat{\theta}_n - \theta) \leq x)$, potom interval spoľahlivosti pre θ s koeficientom $1-\alpha$ je

$$(\widehat{\theta}_n - c_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}}, \widehat{\theta}_n - c_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}}),$$

kde c_p je kvantil distribučnej funkcie H_n.

Ak je $H_n(x) \approx H_n^*(x) = P^*(\sqrt{n}(\theta_n^* - \widehat{\theta}_n) \leq x)$, môžeme neznáme kvantily c_p nahradiť odpovedajúcimi kvantilmi c_p^* distribučnej funkcie H_n^* a dostaneme tak interval spoľahlivosti s koeficientom $1 - \alpha$

$$\big(\widehat{\theta}_n - c^*_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n}}, \widehat{\theta}_n - c^*_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}}\big).$$

Kapitola 4

Aplikácia metódy bootstrap na odhady pravdepodobností prechodu

V tejto kapitole si ukážeme dve bootstrapové metódy pre odhadovanie pravdepodobností prechodu v Markovovom reťazci s konečným počtom stavov. Prvou metódou je podmienený bootstrap, ktorý používa multinomické rozdelenie, pričom berie marginálne početnosti každého stavu v pôvodnom výbere ako fixné. Druhou metódou je štandardný parametrický bootstrap, pri ktorom sú pravdepodobnosti prechodu brané ako parametre modelu. Ukážeme, že podmienené združené asymptotické rozdelenie bootstrapových odhadov pravdepodobností prechodu za danej realizácie $X_1,\ldots,X_n$ má rovnaké rozdelenie ako združené asymptotické rozdelenie maximálne vierohodných odhadov (viď 2.8).

Budeme podobne ako v kapitole 2 predpokladať, že máme realizáciu homogénneho nerozložiteľného neperiodického Markovovho reťazca $\boldsymbol{X} = \{X_1,X_2, \ldots, X_n; n \in \mathbb{N}\}$ pozorovaného do času n s konečnou množinou stavov $S = \{1,2,\ldots,K\}$ a maticou pravdepodobností prechodu $\boldsymbol{P} = (p_{ij}, i,j = 1,\ldots,K)$, kde $p_{ij} > 0$ pre všetky $i,j = 1,\ldots,K$.

4.1 Podmienený bootstrap

Prvou bootstrapovou metódou, ktorú si ukážeme, je podmienený bootstrap. Predtým, ako si ju ukážeme, si zavedieme pre nás vhodnú reprezentáciu Markovovho reťazca $\{X_t; t = 1,\ldots,n\}$ podobne ako v dôkaze vety 5 v odstavci 2.2. Nech U, W_{it}, $i = ,1,\ldots,K,\ t \geq 1$ sú nezávislé náhodné veličiny s pravdepodobnostnými funkciami

$$P[U = j] = \pi_j,\ j = 1,\ldots,K,\ \text{kde } \pi_j \text{ je j-tá zložka stacionárneho rozdelenia}$$

a

$$P[W_{it} = j] = p_{ij},\ 1 \leq i,j \leq K,\ t \geq 1.$$

Definujme proces $\tilde{X} = (\tilde{X}_1,\ldots,\tilde{X}_n)$ ako

$$\tilde{X}_1 = U, \tilde{X}_2 = W_{\tilde{X}_1 1},\ldots,\tilde{X}_k = W_{\tilde{X}_{k-1} m},\ k > 2, \tag{4.1}$$

kde $m-1$ je počet $\tilde{X}_l$ v $\{\tilde{X}_1,\ldots,\tilde{X}_{k-2}\}$ takých, že $\tilde{X}_l = \tilde{X}_{k-1}$. Potom

$$(\tilde{X}_1,\ldots,\tilde{X}_n) \stackrel{D}{=} (X_1,\ldots,X_n), \tag{4.2}$$

ak je počiatočné rozdelenie počiatočného stavu X_1 definované stacionárnym rozdelením. Symbolom $\stackrel{D}{=}$ označujeme rovnosť distribučných funkcií.

Z (4.2) vyplýva, že

$$n_{ij} = \sum_{k=1}^{n-1} I[X_k = i]I[X_{k+1} = j] \stackrel{D}{=} \sum_{k=1}^{n-1} I[\tilde{X}_k = i]I[\tilde{X}_{k+1} = j],\ i,j = 1,\ldots,K,$$

čo sa rovná počtu takých W_{it} medzi $W_{i1},\ldots,W_{i\tilde{n}_i}$, že $W_{it} = j$, pričom $\tilde{n}_i$ je počet takých W_{kl} v realizácii (4.1), že $k = i$, tj. $\tilde{n}_i = \sum_{k=1}^{n-1} I[\tilde{X}_t = i]$.

Potom platí

$$n_{ij} \stackrel{D}{=} \sum_{k=1}^{\tilde{n}_i} I[W_{ik} = j] \text{ a } n_i \stackrel{D}{=} \tilde{n}_i \tag{4.3}$$

Z (4.3) môžeme vidieť, že pozorované počty prechodov n_{ij}, $i,j = 1,\ldots,K$, v realizácii Markovovho reťazca majú rozdelenie súčtu náhodného počtu náhodných veličín, ktoré majú alternatívne rozdelenie s parametrom p_{ij}.

Nahradenie pôvodného výberu bootstrapovým výberom spočíva v tom, že pôvodné nezávislé náhodné veličiny W_{it} nahradíme náhodnými veličinami W^*_{it}, ktoré sú za daných $X_1,\ldots,X_n$ podmienene nezávislé a majú nasledovnú pravdepodobnostnú funkciu

$$P[W^*_{it} = j|X_1,\ldots,X_n] = \widehat{p}_{ij},\ \ 1 \le i,j \le K,\ \ t \ge 1.$$

kde $\widehat{p}_{ij}$ je odhad p_{ij} vypočítaný metódou maximálnej vierohodnosti (viď 2.3).

Teraz môžme zadefinovať bootstrapové odhady n_{ij} ako

$$n^*_{ij} = \sum_{k=1}^{n_i} I[W^*_{ik} = j],\ \ 1 \le i,j \le K. \tag{4.4}$$

Z (4.4) si môžeme všimnúť, že za podmienky $X_1,\ldots,X_n$ majú náhodné veličiny n^*_{ij} binomické rozdelenie $B(n_i,\widehat{p}_{ij})$.

Pre $i \neq j$ sú vektory

$$(n^*_{i1},\ldots,n^*_{iK}) \text{ a } (n^*_{j1},\ldots,n^*_{jK})$$

podmienene nezávislé a pre všetky $i \in S$ platí, že za podmienky $X_1,\ldots,X_n$ majú náhodné vektory $(n^*_{i1},\ldots,n^*_{iK})$ multinomické rozdelenie $Mult_K(n_i,\widehat{p}_{i1},\ldots,\widehat{p}_{iK})$.

Táto metóda generovania n^*_{ij} sa nazýva podmienený bootstrap, pretože n^*_{ij} sú podmienené pozorovanými n_i.

Definujme

$$p^*_{ij} = \frac{n^*_{ij}}{n_i},\ \ 1 \le i,j \le K. \tag{4.5}$$

Veta 8. *Definujme náhodné veličiny* ξ_{ij}^* *ako*

$$\xi_{ij}^* = \frac{n_{ij}^* - n_i\widehat{p}_{ij}}{\sqrt{n_i}} = \frac{n_{ij}^* - n_{ij}}{\sqrt{n_i}}, \;\; i,j = 1,\ldots,K$$

a náhodný vektor $\boldsymbol{\xi_i^*}$ *ako*

$$\boldsymbol{\xi_i^*} = (\xi_{i1}^*,\ldots,\xi_{iK}^*)^\intercal, \;\; i = 1,\ldots,K.$$

Potom pre náhodný K^2*-rozmerný vektor* $\boldsymbol{\xi}^* = (\boldsymbol{\xi_1^{*\intercal}},\ldots,\boldsymbol{\xi_K^{*\intercal}})^\intercal$ *platí*

$$\boldsymbol{\xi}^* \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0},\boldsymbol{\Sigma}) \text{ pre takmer všetky výbery } X_1,\ldots,X_n.$$

Dôkaz vety 8 je uvedený v (Basawa a kol., 1990).

ξ_{ij}^* si môžme prepísať nasledovne

$$\xi_{ij}^* = \frac{n_i(\frac{n_{ij}^*}{n_i} - \widehat{p}_{ij})}{\sqrt{n_i}} = \sqrt{n_i}(p_{ij}^* - \widehat{p}_{ij}) = \sqrt{\frac{n_i}{n}}(\sqrt{n}(p_{ij}^* - \widehat{p}_{ij})).$$

Definujme

$$\vartheta_{ij}^* = \sqrt{n}(p_{ij}^* - \widehat{p}_{ij}), \;\; i,j = 1,\ldots,K.$$

Položme vektor $\boldsymbol{\vartheta^*} = (\vartheta_1^{*\intercal},\ldots,\vartheta_K^{*\intercal})^\intercal$, kde $\boldsymbol{\vartheta_i^*} = (\vartheta_{i1}^*,\ldots,\vartheta_{iK}^*)^\intercal$ pre $i = 1,\ldots,K$. $K^2 \times K^2$-rozmernú maticu L definujme rovnako ako v kapitole 2 (viď 2.7). Potom $\boldsymbol{\vartheta} = \boldsymbol{L\xi^*}$. Z (2.5) vieme, že $\frac{n}{n_i} \to \pi_i$ s pravdepodobnosťou 1. Rovnakým postupom ako v podkapitole 2.2 dospejeme ku nasledovnému asymptotickému rozdeleniu

$$\sqrt{n}(\boldsymbol{P}^* - \widehat{\boldsymbol{P}}) \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0},\tilde{\boldsymbol{\Sigma}}) \text{ s pravdepodobnosťou 1,}$$

kde $\tilde{\boldsymbol{\Sigma}}$ je matica daná (2.9). Z toho vyplýva, že asymptotické rozdelenie $\sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P})$ môžeme aproximovať asymptotickým rozdelením $\sqrt{n}(\boldsymbol{P}^* - \widehat{\boldsymbol{P}})$.

4.2 Metóda štandardného boootstrapu

Druhá bootstrapová metóda, ktorú si ukážeme, je štandardná bootstrapová metóda. Pri podmienenom bootstrape boli n_{ij}^* generované za podmienky n_i. Avšak pri tradičnom parametrickom bootstrape berieme odhady parametrov, ako keby boli skutočnými hodnotami parametrov. Pri tejto metóde sú n_i nahradené $n_i^* = \sum_{j=1}^{K} n_{ij}^*$.

Aby sme sa vyhli tomu, že odhadovaná matica pravdepodobností prechodu môže byť periodická alebo rozložiteľná, definujeme

$$\tilde{n}_{ij} = n_{ij} + \frac{1}{K}, \;\; i,j = 1,\ldots,K \text{ a } \tilde{n}_i = n_i + 1, \;\; i = 1,\ldots,K.$$

Definujme

$$\tilde{p}_{ij} = \frac{\tilde{n}_{ij}}{\tilde{n}_i} \;\; i,j = 1,\ldots,K.$$

Označme stacionárne rozdelenie matice $\tilde{\boldsymbol{P}} = \{\tilde{p}_{ij},\ i,j = 1,\ldots,K\}$ ako $\tilde{\boldsymbol{\Pi}} = (\tilde{\pi}_1, \ldots,\tilde{\pi}_K)$. Použijeme podobnú reprezentáciu generovania Markovského reťazca ako pri podmienenom bootstrape.

Nech U, W_{it} sú za daných $X_1,\ldots,X_n$ podmienene nezávislé náhodné veličiny a majú nasledovné pravdepodobnostné funkcie

$$P[U = j|X_1,\ldots,X_n] = \tilde{\pi}_j,\ \ j = 1,\ldots,K$$

a

$$P[W_{it} = j|X_1,\ldots,X_n] = \tilde{p}_{ij},\ \ i,j = 1,\ldots,K, t \geq 1.$$

Potom definujeme bootstrapovú realizáciu Markovského reťazca ako

$$X_1^* = U, X_2^* = W_{X_1^* 1},\ldots,X_n^* = W_{X_{n-1}^* m},$$

pričom princíp generovania je rovnaký ako pri podmienenom bootstrape (viď 4.1).

Teraz môžeme definovať bootstrapové odhady n_{ij} ako

$$n_{ij}^* = \sum_{k=1}^{n-1} I[X_k^* = i] I[X_{k+1}^* = j],\ \ i,j = 1,\ldots K$$

a bootstrapové odhady n_i ako

$$n_i^* = \sum_{k=1}^{n-1} I[X_k^* = i],\ \ i = 1,\ldots,K.$$

Bootstrapové odhady pravdepodobností prechodu p_{ij} získame ako

$$p_{ij}^* = \frac{n_{ij}^*}{n_i^*},\ \ i,j = 1,\ldots,K.$$

Ak $n_i^* = 0$ definujeme $p_{ij}^* = \delta_{ij}$ pre všetky $j \in S$. Početnosti n_{ij}^* môžeme zapísať aj ako $n_{ij}^* = \sum_{t=1}^{n_i^*} I[W_{it} = j]$.

Veta 9. *Definujme náhodné veličiny ξ_{ij}^* ako*

$$\xi_{ij}^* = \frac{n_{ij}^* - n_i^* \tilde{p}_{ij}}{\sqrt{n_i^*}} \ \textit{pre}\ i,j = 1,\ldots,K$$

a náhodný vektor $\boldsymbol{\xi_i}$ ako

$$\boldsymbol{\xi_i^*} = (\xi_{i1}^*,\ldots,\xi_{iK}^*)^\intercal,\ \ i = 1,\ldots,K.$$

Potom pre náhodný vektor $\boldsymbol{\xi}^* = (\xi_1^{*\intercal},\ldots,\xi_K^{*\intercal})^\intercal$ platí

$$\boldsymbol{\xi}^* \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0},\boldsymbol{\Sigma})\ \textit{s pravdepodobnosťou 1.}$$

Dôkaz vety 9 nájdeme v (Basawa a kol., 1990).
Opäť si môžme odvodiť

$$\xi_{ij}^* = \frac{n_i^*(\frac{n_{ij}^*}{n_i^*} - \tilde{p}_{ij})}{\sqrt{n_i^*}} = \frac{\sqrt{n_i^*}}{\sqrt{n}}(\sqrt{n}(p_{ij}^* - \tilde{p}_{ij})).$$

Definujme

$$\vartheta_{ij}^* = \sqrt{n}(p_{ij}^* - \tilde{p}_{ij}) = \sqrt{\frac{n}{n_i^*}}\xi_{ij}^*, \;\; i,j = 1,\ldots,K.$$

Položme vektor $\boldsymbol{\vartheta^*} = (\vartheta_1^{*\intercal},\ldots,\vartheta_K^{*\intercal})^\intercal$, kde $\boldsymbol{\vartheta_i^*} = (\vartheta_{i1}^*,\ldots,\vartheta_{iK}^*)^\intercal$ pre $i = 1,\ldots,K$.
Podľa lemmy 3.1 v (Basawa a kol., 1990) platí

$$\frac{n_i^*}{n} \xrightarrow{P^*} \pi_i \text{ s pravdepodobnosťou 1 pre } i = 1,\ldots,K, \tag{4.6}$$

čo znamená, že pre všetky $\varepsilon > 0$ a pre skoro všetky výbery $X_1,X_2,\ldots$ platí

$$P\big[|\frac{n_i^*}{n\pi_i} - 1| > \varepsilon | X_1,\ldots,X_n\big] \to 0 \text{ pre } n \to \infty.$$

Z vety o spojitej transformácii máme $\sqrt{\frac{n}{n_i^*}} \xrightarrow{P^*} \frac{1}{\sqrt{\pi_i}}$ s pravdepodobnosťou 1 pre $i = 1,\ldots,K$.
Definujme $K^2 \times K^2$-rozmernú maticu

$$\boldsymbol{M} = \begin{pmatrix} \sqrt{\frac{n}{n_1^*}}\boldsymbol{I_K} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \sqrt{\frac{n}{n_K^*}}\boldsymbol{I_K} \end{pmatrix}.$$

Môžeme sa presvedčiť, že platí $\boldsymbol{\vartheta^*} = \boldsymbol{M\xi^*}$. Rovnakým postupom ako v podkapitole 2.2 si potom môžeme odvodiť asymptotické rozdelenie
pre $\sqrt{n}(\boldsymbol{P}^* - \tilde{\boldsymbol{P}})$. Konkrétne

$$\sqrt{n}(\boldsymbol{P}^* - \tilde{\boldsymbol{P}}) \xrightarrow[n\to\infty]{D} N_{K^2}(\mathbf{0},\tilde{\boldsymbol{\Sigma}}) \text{ s pravdepodobnosťou 1.}$$

Rozdelenie $\sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P})$ môžeme teda aproximovať rozdelením $\sqrt{n}(\boldsymbol{P}^* - \tilde{\boldsymbol{P}})$.

Kapitola 5

Numerická štúdia

V tejto kapitole uvedieme odhady matíc pravdepodobností prechodu spočítané metódou maximálnej vierohodnosti a metódou bootstrap a intervaly spoľahlivosti spočítané na základe týchto odhadov. Porovnáme výsledky získané jednotlivými metódami.

Dáta, tj. realizáciu Markovského reťazca budeme generovať v programe Wolfram Mathematica 9.0 Student Edition. Príslušný kód sa nachádza na priloženom CD.

5.1 Príklad

Generujeme realizáciu homogénneho nerozložiteľného neperiodického Markovovho reťazca $\boldsymbol{X} = (X_1, \ldots, X_{100})$ dĺžky 100 s maticou pravdepodobností prechodu

$$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.5 & 0.35 & 0.15 \\ 0.3 & 0.43 & 0.27 \end{pmatrix}.$$

Počiatočné rozdelenie $\boldsymbol{p}$ si určíme ako $\boldsymbol{p} = \boldsymbol{\Pi}$, kde $\boldsymbol{\Pi}$ je stacionárne rozdelenie reťazca $\boldsymbol{X}$.

Stacionárne a teda zároveň aj počiatočné rozdelenie reťazca vychádza

$$\boldsymbol{\Pi} = (\pi_1, \pi_2, \pi_3) = (0.4505, 0.3022, 0.2473).$$

Na základe realizácie $\boldsymbol{X}$ vypočítame odhad $\widehat{\boldsymbol{P}}$ matice $\boldsymbol{P}$, ktorý sme si uviedli, že vychádza (viď 2.3)

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i} \quad i,j = 1, \ldots, K.$$

5.1.1 Asymptotický interval spoľahlivosti

Z odvodeného asymptotického rozdelenia pre $\sqrt{n}(\widehat{\boldsymbol{P}} - \boldsymbol{P})$ (viď 2.8) vieme, že pre jednotlivé p_{ij} platí

$$\sqrt{n}(\widehat{p}_{ij} - p_{ij}) \xrightarrow[n\to\infty]{D} N\big(0, \frac{p_{ij}(1-p_{ij})}{\pi_i}\big), \quad i,j = 1, \ldots, K.$$

To si môžme s využitím (2.5) a Cramérovej-Sluckého vety prepísať na

$$\sqrt{n_i}(\widehat{p}_{ij} - p_{ij}) \xrightarrow[n\to\infty]{D} N(0,p_{ij}(1-p_{ij})).$$

Teraz ukážeme, že $\frac{1}{\sqrt{\widehat{p}_{ij}(1-\widehat{p}_{ij})}} \xrightarrow{P} \frac{1}{\sqrt{p_{ij}(1-p_{ij})}}$. Podľa vety o spojitej transformácii stačí dokázať, že $\widehat{p}_{ij} \xrightarrow{P} p_{ij}$. To ukážeme nasledovne. Označme $Z_n = \sqrt{n}(\widehat{p}_{ij} - p_{ij})$ a $c_n = \frac{1}{\sqrt{n}}$. Vieme, že $Z_n \xrightarrow{D} Z$, kde $Z \sim N(0,\frac{p_{ij}(1-p_{ij})}{\pi_i})$. Potom z Cramérovej-Sluckého vety máme $c_n Z_n \xrightarrow{D} 0$, čo implikuje, že $c_n Z_n \xrightarrow{P} 0$. Keďže $c_n Z_n = \widehat{p}_{ij} - p_{ij}$, dostávame $\widehat{p}_{ij} - p_{ij} \xrightarrow{P} 0$.

S využitím tejto konvergencie v pravdepodobnosti dostaneme

$$P\left[u_{0.025} \le \sqrt{n_i}\frac{\widehat{p}_{ij} - p_{ij}}{\sqrt{\widehat{p}_{ij}(1-\widehat{p}_{ij})}} \le u_{0.975}\right] \xrightarrow[n\to\infty]{} 0.95,$$

kde u_p značí p-kvantil rozdelenia $N(0,1)$. Po úprave

$$P\left[\widehat{p}_{ij} - \frac{\sqrt{\widehat{p_{ij}}(1-\widehat{p}_{ij})}}{\sqrt{n_i}}u_{0.975} \le p_{ij} \le \widehat{p}_{ij} + \frac{\sqrt{\widehat{p_{ij}}(1-\widehat{p}_{ij})}}{\sqrt{n_i}}u_{0.975}\right] \xrightarrow[n\to\infty]{} 0.95.$$

95%-né intervaly spoľahlivosti pre jednotlivé pravdepodobnosti prechodu p_{ij} sú potom v tvare

$$\left(\widehat{p}_{ij} - \frac{\sqrt{\widehat{p_{ij}}(1-\widehat{p}_{ij})}}{\sqrt{n_i}}u_{0.975} \ , \ \widehat{p}_{ij} + \frac{\sqrt{\widehat{p_{ij}}(1-\widehat{p}_{ij})}}{\sqrt{n_i}}u_{0.975}\right).$$

Ukážeme si typy intervalových odhadov, ktoré budeme počítať.

5.1.2 Percentilový interval spoľahlivosti

95%-ný percentilový interval je podľa (3.3.2) v tvare

$$\big(G_n^{*^{-1}}(0.025),G_n^{*^{-1}}(0.975)\big),$$

kde v našom prípade $G_n^{*^{-1}}(\alpha)$ je empirický α-kvantil, ktorý definujeme ako k_α -tú poriadkovú štatistiku náhodného výberu $p^*_{ij1},\dots,p^*_{ijB}$, kde

$$k_\alpha = \begin{cases} \alpha B, & \text{ak } \alpha B \text{ je celé číslo,} \\ \lfloor \alpha B \rfloor + 1, & \text{ak } \alpha B \text{ nie je celé číslo} \end{cases}$$

a p^*_{ijb} značí bootstrapový odhad pravdepodobnosti prechodu p_{ij} získaný z b-tej realizácie bootstrapového výberu. B označuje celkový počet bootstrapových výberov.

5.1.3 Hybridný interval spoľahlivosti

Keďže sme si ukázali, že rozdelenie $\sqrt{n}(\widehat{p}_{ij} - p_{ij})$ môžeme aproximovať rozdelením $\sqrt{n}(p^*_{ij} - \widehat{p}_{ij})$, získame podľa (3.3.3) 95%-ný bootstrapový interval spoľahlivosti pre p_{ij} ako

$$\big(\widehat{p}_{ij} - c^*_{0.975}\frac{1}{\sqrt{n}}, \widehat{p}_{ij} - c^*_{0.025}\frac{1}{\sqrt{n}}\big),$$

kde c_α^* je empirický α-kvantil, ktorý počítame rovnako ako u percentilového intervalu s výnimkou, že tentokrát vyberáme k_α-tú poriadkovú štatistiku náhodného výberu $\sqrt{n}(p_{ij1}^* - \widehat{p}_{ij}), \ldots, \sqrt{n}(p_{ijB}^* - \widehat{p}_{ij})$.

Z tohto si môžeme odvodiť nasledovný 95%-ný hybridný interval pre p_{ij}

$$\big(2\widehat{p}_{ij} - \tilde{c}_{0.975} \ , \ 2\widehat{p}_{ij} - \tilde{c}_{0.025}\big),$$

kde $\tilde{c}_\alpha$ je empirický α-kvantil spočítaný z náhodného výberu $p_{ij1}^*, \ldots, p_{ijB}^*$.

Najprv si ukážeme porovnania matíc spočítaných jednotlivými metódami. Bootstrapový odhad matice pravdepodobností prechodu $\boldsymbol{P}$ získame tak, že budeme generovať 1000 realizácií Markovovho reťazca s maticou pravdepodobností prechodu $\widehat{\boldsymbol{P}}$, pre každú realizáciu spočítame bootstrapový odhad matice $\boldsymbol{P}$ metódami uvedenými v kapitole 4 a konečný bootstrapový odhad matice dostaneme ako aritmetický priemer zo všetkých bootstrapových odhadov pre jednotlivé realizácie. Teda ak označíme bootstrapový odhad matice $\boldsymbol{P}$ ako $\boldsymbol{P^*} = \{p_{ij}^*, \ i = 1, \ldots, 3, \ \ j = 1, \ldots, 3\}$, potom $p_{ij}^* = \frac{1}{1000} \sum_{b=1}^{1000} p_{ijb}^*$. Bootstrapové odhady $\boldsymbol{P}^*$ matice $\boldsymbol{P}$ spočítané podmienenou metódou, resp. štandardnou metódou označíme ako $\boldsymbol{P}_C^*$, resp. $\boldsymbol{P}_S^*$.

Pripomeňme si pôvodnú maticu pravdepodobností prechodu

$$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.5 & 0.35 & 0.15 \\ 0.3 & 0.43 & 0.27 \end{pmatrix}.$$

Maximálne vierohodný odhad matice pravdepodobností prechodu nám vyšiel

$$\widehat{\boldsymbol{P}} = \begin{pmatrix} 0.4792 & 0.2292 & 0.2917 \\ 0.6429 & 0.2143 & 0.1429 \\ 0.3043 & 0.4783 & 0.2174 \end{pmatrix}.$$

Odhad matice pravdepodobností prechodu získaný metódou podmieneného bootstrapu vyšiel

$$\boldsymbol{P}_C^* = \begin{pmatrix} 0.4757 & 0.2311 & 0.2932 \\ 0.6411 & 0.2127 & 0.1462 \\ 0.302 & 0.48 & 0.2179 \end{pmatrix}.$$

Nakoniec sme spočítali odhad matice pravdepodobností prechodu získaný metódou štandardného bootstrapu

$$\boldsymbol{P}_S^* = \begin{pmatrix} 0.4687 & 0.2286 & 0.2929 \\ 0.6385 & 0.2056 & 0.146 \\ 0.3051 & 0.4748 & 0.209 \end{pmatrix}.$$

Vidíme, že odhad $\widehat{\boldsymbol{P}}$ matice $\boldsymbol{P}$ spočítaný metódou maximálnej vierohodnosti nevyšiel z realizácie dĺžky 100 úplne presne. Preto sa aj bootstrapové odhady matice $\widehat{\boldsymbol{P}}$ odlišujú od pôvodnej matice, keďže ich počítame na základe realizácií Markovových reťazcov s maticou pravdepodobností prechodu $\boldsymbol{P}$.

Z tabuliek, kde máme udané intervaly spoľahivosti pre jednotlivé pravdepodobnosti prechodu (viď 5.1, 5.2, 5.3, 5.4, 5.5), vidíme, že výsledné intervaly nevyšli najpresnejšie a dĺžky jednotlivých intervalov sú dosť veľké. Pri pohľade

na graf (5.2) vidíme, že jednotlivé intervaly spoľahlivosti sú si dosť podobné a z našej numerickej štúdie sa nedá povedať, že by bola niektorá metóda vhodnejšia na odhadnutie intervalov spoľahlivosti pre p_{ij}. Z grafu (5.2) taktiež vidíme, že jednotlivé intervaly spoľahlivosti pokrývajú skutočné hodnoty pravdepodobností prechodu, avšak v prípade pravdepodobnosti prechodu p_{22} sa skutočná hodnota nachádza až pri hornej hranici intervalov spoľahlivosti. Môžeme teda usúdiť, že realizácie dĺžky 100 nie sú dostatočné pre presný odhad pravdepodobností prechodu a intervalov spoľahlivosti a na ich presnejšie určenie by sme potrebovali dlhšie realizácie. Problémom bootstrapových výberov je to, že pri takejto malej dĺžke Markovovho reťazca bude odhad pravdepodobností prechodu spočítaný metódou maximálnej vierohodnosti často nie úplne presný, čo ovplyvní výsledky bootstrapových odhadov, pretože maximálne vierohodné odhady pravdepodobnosti prechodu využívame pri generovaní bootstrapových realizácií Markovovho reťazca.

Z grafu (5.1), v ktorom porovnávame empirické distribučné funkcie pre p_{11} a normálne rozdelenie $N\big(0,\frac{p_{ij}(1-p_{ij})}{\pi_i}\big)$ je vidno, že bootstrapové rozdelenia $\sqrt{n}(p^*_{ij}-\widehat{p}_{ij})$ oboch bootstrapových metód aproximujú rozdelenie $\sqrt{n}(\widehat{p}_{ij}-p_{ij})$ lepšie ako normálne rozdelenie. Pri prehliadnutí ostatných grafov vidíme, že bootstrapové empirické distribučné funkcie aproximujú empirickú distribučnú funkciu pôvodho rozdelenia pre tisíc bootstrapových realizácii Markovovho reťazca dĺžky 100 veľmi dobre.

Z grafu (5.4)je zrejmé, že zväčšením dĺžky Markovovho reťazca na 350 sa intervaly spoľahlivosti pre p_{ij} zreteľne skrátia.

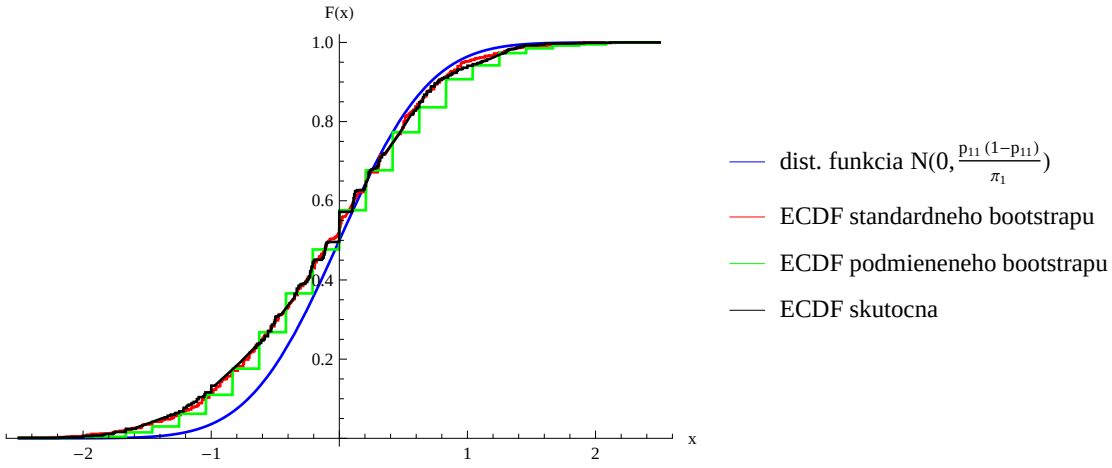
Podľa grafu (5.3), ako aj podľa ostatných grafov vygenerovaných v programe Wolfram Mathematica, vidíme, že bootstrapové rozdelenia aproximujú pôvodné rozdelenie s pribúdajúcou dĺžkou Markovovho reťazca ešte lepšie.

Zväčšením dĺžky Markovovho reťazca na 1000 už dostávame výrazne lepšie výsledky. Grafy bootstrapových rozdelení takmer splývajú so skutočným rozdelením, čo môžme vidieť z obrázku (5.5).

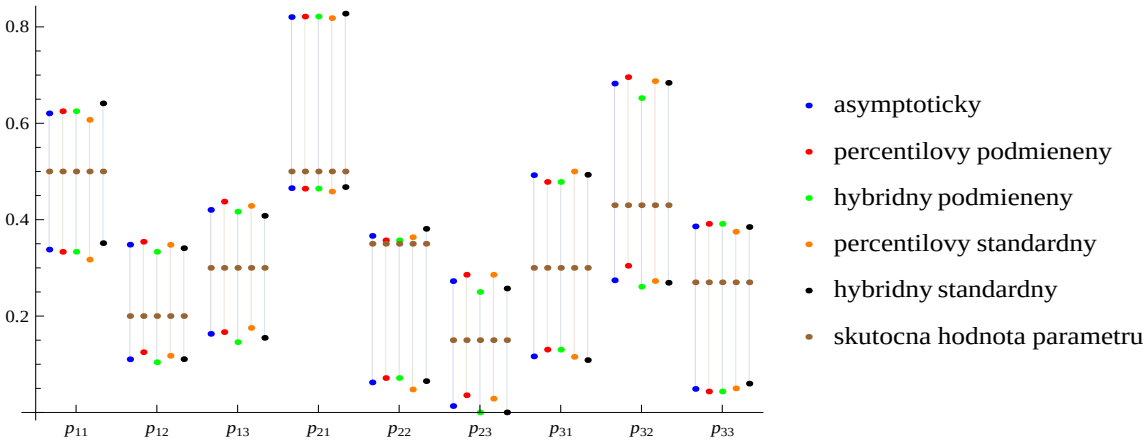
Jednotlivé intervaly spoľahlivosti pre p_{ij} vychádzajú taktiež lepšie, dĺžky jednotlivých intervalov sa skrátili približne na tretinu oproti intervalom spoľahlivosti počítaným pre Markovove reťazce dĺžky 100.

Parameter	Interval spoľahlivosti	Skutočná hodnota parametru
p_{11}	(0.3379 , 0.6205)	0.5
p_{12}	(0.1103 , 0.3481)	0.2
p_{13}	(0.1631 , 0.4203)	0.3
p_{21}	(0.4654 , 0.8204)	0.5
p_{22}	(0.0623 , 0.3663)	0.35
p_{23}	(0.0133 , 0.2725)	0.15
p_{31}	(0.1163 , 0.4923)	0.3
p_{32}	(0.2742 , 0.6824)	0.43
p_{33}	(0.0488 , 0.3860)	0.27

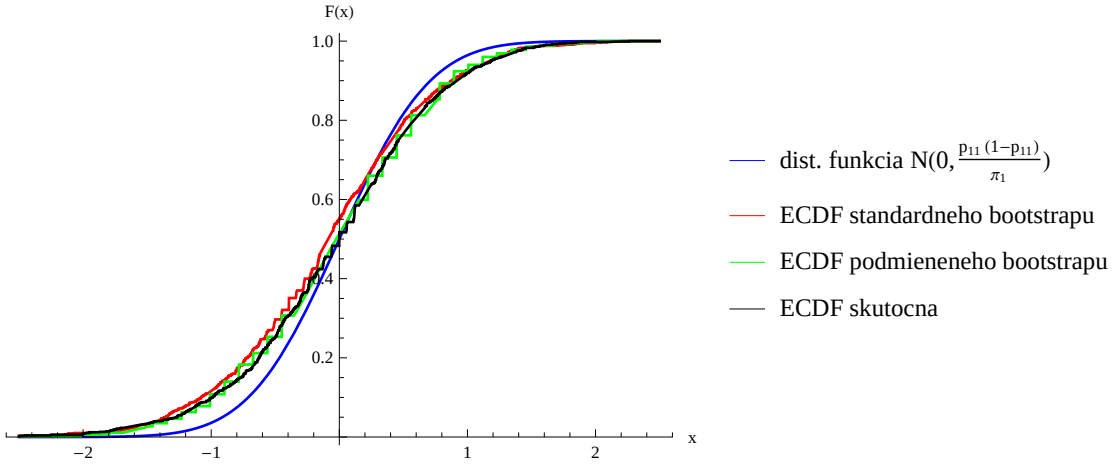
Tabuľka 5.1: Asymptotické intervaly spoľahlivosti



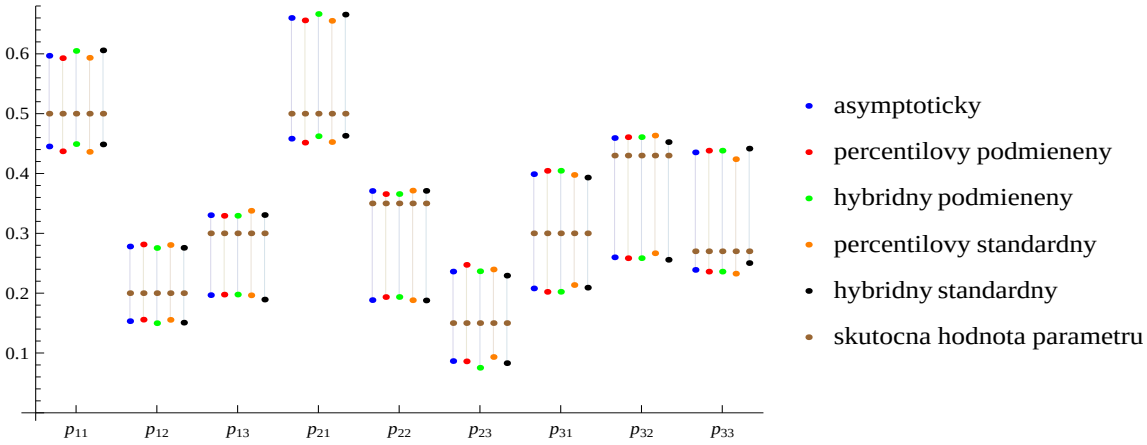
Obr. 5.1: Porovnania distribučných funkcií pre p_{11} pre Markovov reťazec dĺžky 100



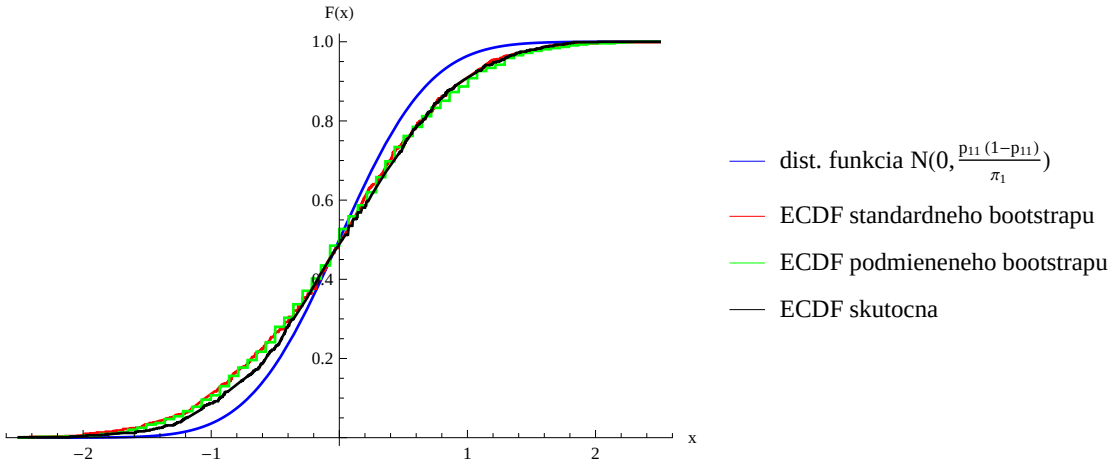
Obr. 5.2: Intervaly spoľahlivosti pre jednotlivé $p_{11}, \ldots, p_{33}$ pre Markovov reťazec dĺžky 100



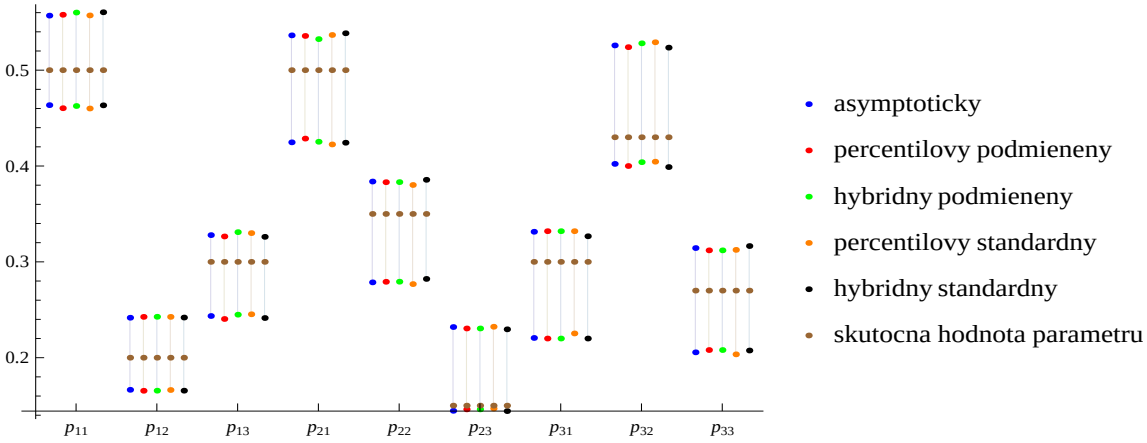
Obr. 5.3: Porovnania distribučných funkcií pre p_{11} pre Markovov reťazec dĺžky 350



Obr. 5.4: Intervaly spoľahlivosti pre jednotlivé $p_{11}, \ldots, p_{33}$ pre Markovov reťazec dĺžky 350



Obr. 5.5: Porovnania distribučných funkcií pre p_{11} pre Markovov reťazec dĺžky 1000



Obr. 5.6: Intervaly spoľahlivosti pre jednotlivé $p_{11}, \ldots, p_{33}$ pre Markovov reťazec dĺžky 1000

Parameter	Interval spoľahlivosti	Skutočná hodnota parametru
p_{11}	(0.3333 , 0.6250)	0.5
p_{12}	(0.1250 , 0.3542)	0.2
p_{13}	(0.1667 , 0.4375)	0.3
p_{21}	(0.4643 , 0.8214)	0.5
p_{22}	(0.0714 , 0.3571)	0.35
p_{23}	(0.0357 , 0.2857)	0.15
p_{31}	(0.1304 , 0.4783)	0.3
p_{32}	(0.3043 , 0.6957)	0.43
p_{33}	(0.0435 , 0.3913)	0.27

Tabuľka 5.2: Percentilové intervaly spočítané metódou podmieneného bootstrapu

Parameter	Interval spoľahlivosti	Skutočná hodnota parametru
p_{11}	(0.3171 , 0.6071)	0.5
p_{12}	(0.1176 , 0.3478)	0.2
p_{13}	(0.1754 , 0.4286)	0.3
p_{21}	(0.4583 , 0.8182)	0.5
p_{22}	(0.0476 , 0.3636)	0.35
p_{23}	(0.0286 , 0.2857)	0.15
p_{31}	(0.1154 , 0.5000)	0.3
p_{32}	(0.2727 , 0.6875)	0.43
p_{33}	(0.0500 , 0.3750)	0.27

Tabuľka 5.3: Percentilové intervaly spočítané metódou štandardného bootstrapu

Parameter	Interval spoľahlivosti	Skutočná hodnota parametru
p_{11}	(0.3333 , 0.6250)	0.5
p_{12}	(0.1250 , 0.3542)	0.2
p_{13}	(0.1667 , 0.4375)	0.3
p_{21}	(0.4643 , 0.8214)	0.5
p_{22}	(0.0714 , 0.3571)	0.35
p_{23}	(0.0357 , 0.2857)	0.15
p_{31}	(0.1304 , 0.4783)	0.3
p_{32}	(0.3043 , 0.6957)	0.43
p_{33}	(0.0435 , 0.3913)	0.27

Tabuľka 5.4: Hybridné intervaly spočítané metódou podmieneného bootstrapu

Parameter	Interval spoľahlivosti	Skutočná hodnota parametru
p_{11}	(0.3513 , 0.6413)	0.5
p_{12}	(0.1106 , 0.3408)	0.2
p_{13}	(0.1548 , 0.4080)	0.3
p_{21}	(0.4676 , 0.8275)	0.5
p_{22}	(0.0650 , 0.3810)	0.35
p_{23}	(0.0001 , 0.2572)	0.15
p_{31}	(0.1086 , 0.4932)	0.3
p_{32}	(0.2691 , 0.6839)	0.43
p_{33}	(0.0598 , 0.3848)	0.27

Tabuľka 5.5: Hybridné intervaly spočítané metódou štandardného bootstrapu

Záver

V teoretickej časti práce sme uviedli postupy, ako odhadnúť pravdepodobnosti prechodu, ak máme k dispozícii realizáciu homogénneho nerozložiteľného neperiodického Markovovho reťazca pozorovaného do času n s konečnou množinou stavov. Odvodili sme asymptotické rozdelenie odhadov pravdepodobností prechodu a zistili, že odhady získané metódou maximálnej vierohodnosti majú rovnaké asymptotické normálne rozdelenie ako bootstrapové odhady. Preto môžme ich rozdelenie aproximovať bootstrapovým rozdelením.

V poslednej kapitole sme si odvodili intervaly spoľahlivosti získané jednotlivými metódami a spočítali ich na simulovaných dátach. Zistili sme, že intervaly spoľahlivosti spočítané z asymptotického rozdelenia sú približne rovnaké ako bootstrapové intervaly spoľahlivosti pri pravdepodobnosti pokrytia 0.95. Rovnako aj odhady matíc pravdepodobností prechodu sú si podobné pre jednotlivé metódy. Podľa grafov distribučných funkcií v numerickej štúdii sme ale zistili, že aproximácia bootstrapovým rozdelením je dosť presná. Pri výberoch väčšej dĺžky sme zistili, že bootstrapové odhady pravdepodobností prechodu sú dostatočne presné.

Na základe výsledkov našej práce môžeme skonštatovať, že oproti odhadom spočítaným metódou maximálnej vierohodnosti a oproti intervalom spoľahlivosti spočítaným z asymptotického rozdelenia týchto odhadov nepriniesol bootstrap pozorovateľne lepšie výsledky, avšak ukázal svoje opodstatnenie pri odhade nami neznámeho rozdelenia $\sqrt{n}(\widehat{p}_{ij} - p_{ij})$, keď sa ukázalo, že ho aproximuje presnejšie ako normálne rozdelenie.

Literatúra

Anděl, J. (2011). *Základy matematické statistiky*. Tretie vydanie. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Athreya, K. a Fuh, C. (1992). Bootstrapping markov chains. *Exploring the Limits of Bootstrap*, pages 49–64.

Basawa, I., Green, T., McCormick, W. a Taylor, R. (1990). Asymptotic bootstrap validity for finite markov chains. *Communications in Statistics-Theory and Methods*, **19**(4), 1493–1510.

Billingsley, P. (1961). Statistical methods in markov chains. *The Annals of Mathematical Statistics*, **32**(1), 12–40.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Prášková, Z. (2004). Metoda bootstrap. *ROBUST 04*, pages 299–314.

Prášková, Z. a Lachout, P. (2012). *Základy náhodných procesů I.* Druhé vydanie. Matfyzpress, Praha. ISBN 978-80-7378-210-8.

Zoznam obrázkov

Zoznam tabuliek