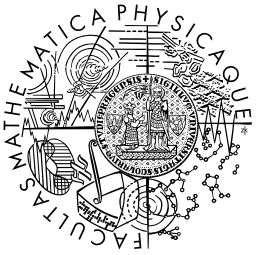


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Natálie Šuléřová

Markovský binomický model

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	RNDr. Šárka Hudecová, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2019

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 má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

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

Na tomto místě bych ráda poděkovala RNDr. Šárce Hudecové, Ph.D. za cenné rady a připomínky, za čas, který mi věnovala, a také za odbornou pomoc a ochotu při zpracování mé bakalářské práce.

Název práce: Markovský binomický model

Autor: Natálie Šuléřová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá Markovským binomickým modelem. Jedná se o zobecnění standardního binomického rozdělení, kde místo součtu nezávislých náhodných veličin uvažujeme součet veličin, které tvoří stacionární Markovův řetězec. Cílem práce je popsat tento model a odvodit jeho vlastnosti jako jsou střední hodnota, rozptyl nebo vytvořující funkce. Část práce je věnována také odhadům neznámých parametrů tohoto modelu pomocí momentové metody a metody maximální věrohodnosti. Přesnost odhadů jednotlivých metod je porovnána na simulovaných datech. Na závěr je představený model aplikován na reálná data o nehodovosti pod vlivem alkoholu.

Klíčová slova: Markovský binomický model, Markovův řetězec, binomické rozdělení

Title: Markov binomial model

Author: Natálie Šuléřová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis we study the Markov chain binomial model, which generalizes the standard binomial distribution. Instead of the sum of independent random variables, we consider the sum of random variables that form a stationary Markov chain. The goal of this thesis is to describe this model along with its properties, such as the expected value, variance and probability generating function. A part of this thesis is dedicated to estimating parameters of this model using the method of moments and the maximum likelihood estimation. The accuracy of the methods is compared in a simulation study and obtained results are discussed. The presented model is then applied on a real dataset based on rate of alcohol-impaired car accidents.

Keywords: Markov binomial model, Markov chain, binomial distribution

Obsah

Úvod **2**

1 Základní pojmy **3**
1.1 Binomické rozdělení . . . 3
1.1.1 Vlastnosti binomického rozdělení . . . 3
1.1.2 Odhady parametrů binomického rozdělení . . . 4
1.1.3 Zobecnění binomického rozdělení . . . 4
1.2 Markovův řetězec . . . 5
1.2.1 Základní definice . . . 5
1.2.2 Markovův řetězec se dvěma stavy . . . 7

2 Markovský binomický model **9**
2.1 Definice . . . 9
2.2 Vytvořující funkce . . . 11
2.3 Střední hodnota a rozptyl . . . 14
2.4 Odhady parametrů . . . 16
2.4.1 Odhadování parametrů založené pouze na X . . . 16
2.4.2 Odhadovaní parametrů založené na průběhu řetězců . . . 18

3 Ilustrace na datech **21**
3.1 Simulovaná data . . . 21
3.2 Reálná data . . . 23

Závěr **25**

Seznam použité literatury **26**

A Přílohy **27**
A.1 Data o nehodovosti pod vlivem alkoholu v Praze . . . 27
A.2 Data o nehodovosti pod vlivem alkoholu ve Zlínském kraji . . . 28

Úvod

Tématem této bakalářské práce je Markovský binomický model. Jedná se o zobecnění klasického binomického modelu, kde místo součtu nezávislých alternativních veličin budeme uvažovat součet veličin, které nebudou nezávislé, ale budou tvořit Markovův řetězec. Cílem práce je tento model představit, studovat jeho základní vlastnosti a možné metody odhadů jeho parametrů a v neposlední řadě také ilustrovat tento model na datech.

V první kapitole si nejdříve připomeneme binomické rozdělení a dále se seznámíme se základními pojmy, které se týkají Markovových řetězců. Uvedeme si důležité definice a terminologii, dále popíšeme vlastnosti Markovova řetězce se dvěma stavy.

Ve druhé kapitole si představíme samotný model a dále se budeme zabývat jeho základními vlastnostmi jako je střední hodnota, rozptyl a vytvořující funkce. Další část této kapitoly bude věnována odhadům parametrů Markovského binomického modelu. Popíšeme celkem tři různé metody, které lze použít k odhadu neznámých parametrů.

V poslední kapitole potom budeme ilustrovat Markovský binomický model na datech. Nejprve si nasimulujeme data z tohoto modelu a budeme porovnávat, jak dobře odhadují jednotlivé metody. Bude nás zajímat, jak moc se liší metody, které využívají jen souhrnnou informaci o datech, od metody, která využívá detailnější informaci. Ve druhé části této kapitoly pak budeme model ilustrovat na reálných datech, konkrétně na datech o nehodovosti pod vlivem alkoholu. Opět pomocí všech tří metod vypočítáme odhady parametrů a budeme diskutovat výsledky.

1. Základní pojmy

V této práci se budeme zabývat zobecněným binomickým rozdělením, kde místo nezávislých alternativních veličin budeme uvažovat tzv. Markovův řetězec. Proto by bylo dobré si na začátku připomenout i základní vlastnosti binomického rozdělení a stejně tak i odhady jeho parametrů. Můžeme pak porovnávat, jak se budou naše nové poznatky o zobecněném binomickém rozdělení od toho standardního lišit.

1.1 Binomické rozdělení

Binomické rozdělení udává počet úspěchů v n bernoulliovských pokusech, tj. nezávislých pokusech, které končí buď úspěchem, nebo neúspěchem. Potom tedy můžeme binomické rozdělení definovat jako rozdělení součtu nezávislých stejně rozdělených náhodných veličin s alternativním rozdělením.

Definice 1. *Nechť $n \in \mathbb{N}$ a $Y_1, Y_2, \ldots, Y_n$ jsou nezávislé stejně rozdělené náhodné veličiny s alternativním rozdělením s parametrem $p \in (0,1)$. Potom náhodná veličina $X = \sum_{i=1}^{n} Y_i$ má* binomické rozdělení s parametry n a p. *Značíme $X \sim Bi(n,p)$.*

1.1.1 Vlastnosti binomického rozdělení

Nechť X je náhodná veličina s binomickým rozdělením s parametry n a p. Potom je známo, že můžeme pro $k = 0, 1, \ldots, n$ spočítat příslušné pravděpodobnosti následovně:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Pro střední hodnotu a rozptyl dále platí:

$$\begin{aligned} \mathsf{E}\, X &= np\,, \\ \operatorname{var} X &= np(1-p). \end{aligned}$$

To plyne ihned z definice binomického rozdělení jakožto součtu n nezávislých alternativních veličin.

Také nás bude zajímat vytvořující funkce náhodné veličiny X. K tomu si nejprve vypočítáme vytvořující funkci P_Y náhodné veličiny Y_1.

$$P_Y(s) = \mathsf{E}\,(s^Y) = (1-p)s^0 + ps^1 = 1 - p + ps.$$

Potom platí

$$P_X(s) = [P_Y(s)]^n = (1 - p + ps)^n.$$

Dostáváme tedy, že vytvořující funkce náhodné veličiny X je $(1 - p + ps)^n$.

1.1.2 Odhady parametrů binomického rozdělení

V této části si odvodíme bodový odhad parametru p. Nejprve momentovou metodou a poté i metodou maximální věrohodnosti.

Nechť $m \in \mathbb{N}$ a $X_1, X_2, \ldots, X_m$ je náhodný výběr z binomického rozdělení s parametry n a p.

Momentová metoda spočívá v tom, že pokládá výběrový průměr $\overline{X}_m$ rovný střední hodnotě. Dostaneme tedy:

$$n\widehat{p} = \overline{X}_m$$
$$\widehat{p} = \frac{\overline{X}_m}{n}.$$

Nyní najdeme bodový odhad metodou maximální věrohodnosti. Věrohodnostní funkce je rovna

$$L(n,p) = \prod_{i=1}^{m} \binom{n}{X_i} p^{X_i}(1-p)^{n-X_i}$$

a logaritmická věrohodnostní funkce je

$$l(n,p) = \log \prod_{i=1}^{m} \binom{n}{X_i} + \log p \sum_{i=1}^{m} X_i + \log(1-p)(mn - \sum_{i=1}^{m} X_i).$$

Zderivujeme $l(n,p)$ podle p a výraz položíme rovný nule.

$$\frac{\sum_{i=1}^{m} X_i}{\widetilde{p}} - \frac{mn - \sum_{i=1}^{m} X_i}{1-\widetilde{p}} = 0$$
$$\sum_{i=1}^{m} X_i - \widetilde{p}\sum_{i=1}^{m} X_i = \widetilde{p}mn - \widetilde{p}\sum_{i=1}^{m} X_i$$
$$\sum_{i=1}^{m} X_i = \widetilde{p}mn$$
$$\widetilde{p} = \frac{\overline{X}_m}{n}$$

Tedy vidíme, že momentovou metodou i metodou maximální věrohodnosti jsme získali stejný bodový odhad $\overline{X}_m/n$ parametru p. Je známo, že tento odhad je konzistentní, což plyne ze silného zákona velkých čísel. Z centrální limitní věty pak dostáváme, že je i asymptoticky normální.

1.1.3 Zobecnění binomického rozdělení

Situace se stává zajímavější, pokud bychom uvažovali součet stejně rozdělených náhodných veličin, které by však nebyly nezávislé. Předpokládejme tedy, že $Y_1, Y_2, \ldots, Y_n$ jsou stejně rozdělené náhodné veličiny s alternativním rozdělením s parametrem p, které jsou závislé, a uvažujme náhodnou veličinu $X = \sum_{i=1}^{n} Y_i$.

Potom střední hodnota náhodné veličiny X je díky linearitě stejná, jako kdyby byly veličiny $Y_1, Y_2, \ldots, Y_n$ nezávislé, tedy $\mathsf{E}\ X = np$.

V případě rozptylu to už samozřejmě bude rozdílné. Obecně víme, že pro rozptyl náhodné veličiny X platí

$$\operatorname{var} X = \sum_{i=1}^{n} \operatorname{var} Y_i + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \operatorname{cov}(Y_i, Y_j) = np(1-p) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \operatorname{cov}(Y_i, Y_j).$$

Pokud by byly veličiny $Y_1, Y_2, \ldots, Y_n$ nezávislé, potom je druhý sčítanec v rozptylu nulový, neboť kovariance dvou nezávislých veličin je nulová. Pokud jsou ale veličiny $Y_1, Y_2, \ldots, Y_n$ závislé, tato závislost se v rozptylu projeví právě skrze tyto kovariance. Tedy rozptyl a případné další vlastnosti veličiny X se oproti standardnímu binomickému rozdělení budou lišit podle toho, jak na sobě budou veličiny Y_i pro $i = 1, \ldots, n$ závislé.

Samozřejmě možností, jak by mohly být veličiny závislé, je spousta. V této práci se zaměříme zejména na situaci, kdy veličiny Y_i tvoří tzv. Markovův řetězec, neboť se s ním v praxi setkáváme poměrně často. Budeme vycházet ze článku Islam a O'shaughnessy (2013), který se právě touto problematikou zabývá. Dá se říci, že autoři v něm navazují na článek Rudolfer (1990) a zároveň opravují některé jeho chyby.

1.2 Markovův řetězec

Jak už bylo zmíněno, budeme pracovat se součtem náhodných veličin, které tvoří Markovův řetězec, konkrétně Markovův řetězec se dvěma stavy 0 a 1. V první části této podkapitoly si tedy nejdříve zavedeme obecně pojem Markovova řetězce a další definice s ním spjaté, v druhé části se potom budeme blíže věnovat právě Markovovu řetězci se stavy 0 a 1.

1.2.1 Základní definice

Jelikož je Markovův řetězec náhodný proces, nejprve uvedeme pár základních definic z teorie náhodných procesů.

Definice 2. *Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor a nechť $T \subset \mathbb{R}$. Rodina náhodných veličin $\{X_t, t \in T\}$ definovaných na $(\Omega, \mathcal{A}, P)$ se nazývá* náhodný (stochastický) proces. *Pokud $T = \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ nebo $T = \mathbb{N}_0 = \{0, 1, \ldots\}$, mluvíme o* procesu s diskrétním časem.

Definice 3. *Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor a uvažujme na něm posloupnost náhodných veličin $\{X_n, n \in \mathbb{N}_0\}$, které nabývají pouze celočíselných hodnot. Nechť S je množina celých čísel i takových, že $i \in S$ právě tehdy, když existuje $n \in \mathbb{N}_0$ tak, že $P(X_n = i) > 0$. Množina S může být buď konečná, nebo spočetná a budeme ji nazývat* množina stavů *náhodného procesu $\{X_n, n \in \mathbb{N}_0\}$. Prvky stavové množiny budeme nazývat* stavy. *V tomto případě také říkáme, že jde o* proces s diskrétními stavy.

Stěžejní definicí bude samozřejmě definice Markovova řetězce.

Definice 4. *Posloupnost celočíselných náhodných veličin* $\{X_n, n \in \mathbb{N}_0\}$ *se nazývá* Markovův řetězec s diskrétním časem *a množinou stavů S, jestliže*

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

pro všechna $n = 0, 1, \ldots$ *a všechna* $i, j, i_{n-1}, \ldots, i_0 \in S$ *taková, že*

$$P(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) > 0.$$

Vztah (1.1) se nazývá *markovská vlastnost*. Znamená, že pravděpodobnost výsledku v budoucím čase $n + 1$, pokud známe výsledek v přítomném čase n a výsledky z minulých časů $n - 1, n - 2, \ldots, 0$, je stejná, jako kdybychom znali jen výsledek v přítomném čase n.

Zavedeme si ještě základní terminologii spjatou s Markovovými řetězci.
Podmíněné pravděpodobnosti $P(X_{n+1} = j | X_n = i) = p_{ij}(n, n + 1)$ (pokud jsou definovány) budeme nazývat *pravděpodobnosti přechodu* ze stavu i v čase n do stavu j v čase $n + 1$.
Podmíněné pravděpodobnosti $P(X_{n+m} = j | X_n = i) = p_{ij}(n, n + m)$ pro $m \in \mathbb{N}$, $m \geq 1$ (pokud jsou definovány) potom budeme nazývat *pravděpodobnosti přechodu* ze stavu i v čase n do stavu j v čase $n + m$, nebo také *pravděpodobnosti přechodu m-tého řádu.*

V celé práci budeme pracovat s homogenním Markovovým řetězcem, který má stacionární rozdělení, tudíž zadefinujeme i tyto pojmy.

Definice 5. *Řekneme, že Markovův řetězec* $\{X_n, n \in \mathbb{N}_0\}$ *s množinou stavů S je* homogenní*, jestliže pro každé* $i,j \in S, m \in \mathbb{N}, n_1, n_2 \in \mathbb{N}_0$ *splňující*

$$P(X_{n_1} = i) > 0,\ P(X_{n_2} = i) > 0$$

platí

$$p_{ij}(n_1, n_1 + m) = p_{ij}(n_2, n_2 + m).$$

Pro homogenní Markovův řetězec můžeme dále zavést i matici pravděpodobností přechodu a počáteční rozdělení.
Všechny pravděpodobnosti přechodu lze seskládat do čtvercové matice

$$\mathbf{P} = \{p_{ij}, i,j \in S\},$$

kterou budeme nazývat *matice pravděpodobností přechodu.*
Dále pro $i \in S$ označme $p_i = P(X_0 = i)$. Potom pravděpodobnostní rozdělení $\mathbf{p} = \{p_i, i \in S\}$ nazveme *počáteční rozdělení.*

Definice 6. *Nechť* $\{X_n, n \in \mathbb{N}_0\}$ *je homogenní řetězec s množinou stavů S a maticí pravděpodobností přechodu* $\boldsymbol{P}$*. Nechť* $\boldsymbol{\pi} = \{\pi_j, j \in S\}$ *je nějaké pravděpodobnostní rozdělení na množině S, tj.* $\pi_j \geq 0, j \in S, \sum_{j \in S} \pi_j = 1$*. Potom* $\boldsymbol{\pi}$ *se nazývá* stacionární rozdělení *daného řetězce, jestliže platí*

$$\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \boldsymbol{P},$$

neboli

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

Řekneme, že homogenní řetězec je stacionární*, pokud jeho počáteční rozdělení je stacionární.*

Nepodmíněné pravděpodobnosti $p_j(n) = P(X_n = j)$ se nazývají *absolutní pravděpodobnosti* v čase n. Označíme-li $\mathbf{p}(n) = \{p_j(n), j \in S\}$, potom platí

$$\mathbf{p}(n)^\top = \mathbf{p}^\top \mathbf{P}^n, \quad n \in \mathbb{N}_0.$$

Zavedeme ještě značení $p_{ij}^{(n)}$.
Položme $p_{ij}^{(0)} = \delta_{ij}$, kde

$$\delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Dále položme $p_{ij}^{(1)} = p_{ij}$. Potom pro přirozené $n \geq 1$ definujme rekurentně

$$p_{ij}^{(n+1)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}.$$

Prvky $p_{ij}^{(n)}$ pak tvoří matici $\mathbf{P}^{(n)}$ a platí $\mathbf{P}^{(n)} = \mathbf{P} \cdot \mathbf{P}^{(n-1)} = \mathbf{P}^n$.

Na závěr ještě uvedeme dvě důležité věty, které shrnují vlastnosti Markovova řetězce, jež budeme využívat, viz Prášková a Lachout (2012) podkapitola 2.1.

Věta 1. *Nechť $\{X_n\}$ je homogenní Markovův řetězec s množinou stavů S. Potom platí*

$$P(X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) = p_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n}$$

pro všechna $n \in \mathbb{N}_0$ a $i_0, i_1, \ldots, i_n \in S$.

Věta 2. *Nechť $\{X_n\}$ je homogenní Markovův řetězec s maticí pravděpodobností přechodu $\boldsymbol{P}$ a množinou stavů S. Potom pro pravděpodobnosti přechodu n-tého řádu platí*

$$P(X_{m+n} = j | X_m = i) = p_{ij}^{(n)}, \quad i, j \in S$$

pro všechna $m, n \in \mathbb{Z}, m \geq 0, n \geq 0$ a $P(X_m = i) > 0$.

1.2.2 Markovův řetězec se dvěma stavy

Předpokládejme, že $\{Z_i, i = 1, 2, \ldots\}$ je dvoustavový Markovův řetězec se stavy 0 a 1, kde stav 0 představuje neúspěch a stav 1 úspěch. Označme jeho počáteční rozdělení $(q, p)^\top$, kde $p = P(Z_1 = 1)$ a $q = P(Z_1 = 0) = 1 - p$, a pravděpodobnosti přechodu $p_{ij} = P(Z_{m+1} = j | Z_m = i)$ pro $i, j = 0, 1, \ldots$ a $m = 1, 2, \ldots$ Tyto pravděpodobnosti můžeme napsat do matice pravděpodobností přechodu

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

Označme si pravděpodobnosti v matici $\mathbf{P}$ jako:

$$\mathbf{P} = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix},$$

kde $0 < \alpha, \beta < 1$.
Pro $k \geq 2$ pravděpodobnost úspěchu v k-tém kroku, $p_k = P(Z_k = 1)$, závisí

jak na k, tak na počátečním rozdělení $(q,p)^\top$. To můžeme nahlédnout, pokud si vyjádříme pravděpodobnost p_k podle toho, jestli jsme v $(k-1)$-ním kroku byli ve stavu 1, nebo 0. To nám stačí, neboť $\{Z_i, i = 1,2,\ldots\}$ je Markovův řetězec, tudíž stav v k-tém kroku závisí pouze na stavu v $(k-1)$-ním kroku. Dále využijeme matice pravděpodobností přechodu a dostaneme rekurentní vztah:

$$\begin{aligned} p_k &= P(Z_k = 1) \\ &= P(Z_{k-1} = 1, Z_k = 1) + P(Z_{k-1} = 0, Z_k = 1) \\ &= (1-\beta)p_{k-1} + \alpha(1-p_{k-1}) \\ &= \alpha + (1-\alpha-\beta)p_{k-1}. \end{aligned}$$

Pokud bychom si opět vyjádřili pravděpodobnost p_{k-1} a dosadili do vztahu pro p_k, dostali bychom:

$$\begin{aligned} p_k &= \alpha + (1-\alpha-\beta)[\alpha + (1-\alpha-\beta)p_{k-2}] \\ &= \alpha + (1-\alpha-\beta)\alpha + (1-\alpha-\beta)^2 p_{k-2}. \end{aligned}$$

A takto dále rekurzivně získáme výraz:

$$\begin{aligned} p_k &= \alpha\left[1 + (1-\alpha-\beta) + (1-\alpha-\beta)^2 + \ldots + (1-\alpha-\beta)^{k-1}\right] \\ &\quad + (1-\alpha-\beta)^{k-1}p \\ &= \alpha \sum_{n=0}^{k-1} (1-\alpha-\beta)^n + (1-\alpha-\beta)^{k-1}p. \end{aligned}$$

Řadu $\alpha\sum_{n=0}^{k-1}(1-\alpha-\beta)^n$ sečteme podle vzorce pro součet prvních $k-1$ členů geometrické řady.

$$\alpha \sum_{n=0}^{k-1} (1-\alpha-\beta)^n = \alpha \cdot \frac{(1-\alpha-\beta)^{k-1} - 1}{-(\alpha+\beta)} \quad = \frac{-\alpha(1-\alpha-\beta)^{k-1}}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}$$

Tedy celkem dostáváme pravděpodobnost p_k vyjádřenou pomocí pravděpodobnosti p:

$$p_k = \frac{\alpha}{\alpha+\beta} + (1-\alpha-\beta)^{k-1}\left(p - \frac{\alpha}{\alpha+\beta}\right).$$

Zřejmě $|1-\alpha-\beta| < 1$, neboť $0 < \alpha, \beta < 1$. Pro $k \to \infty$ pak máme, že

$$\lim_{k\to\infty} p_k = \frac{\alpha}{\alpha+\beta}$$

a tato limita už nezávisí na počátečním stavu (0 nebo 1) ani na počátečním rozdělení $(q,p)^\top$. Tedy stacionární rozdělení Markovova řetězce je $(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})^\top$. Navíc, pokud pro počáteční rozdělení platí, že $p = \frac{\alpha}{\alpha+\beta}$, potom je pravděpodobnost úspěchu p_k ve všech pokusech konstantně rovna $\frac{\alpha}{\alpha+\beta}$.

2. Markovský binomický model

V této kapitole si nejprve odvodíme Markovský binomický model se všemi parametry a dále se budeme zabývat jeho základními vlastnostmi.

2.1 Definice

Nechť $\{Z_i, i = 1, 2, \ldots\}$ je dvoustavový stacionární Markovův řetězec se stavy 0 a 1, počátečním rozdělením $(q, p)^\top$, kde $p = P(Z_1 = 1)$ a $q = P(Z_1 = 0) = 1-p$, a maticí pravděpodobností přechodu

$$\mathbf{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

Protože řetězec je stacionární, počáteční rozdělení je stacionární, tudíž

$$p = \frac{\alpha}{\alpha+\beta}, \;\; q = 1 - p = \frac{\beta}{\alpha+\beta}.$$

Označme $\delta = 1 - \alpha - \beta$. Pak platí následující rovnosti:

$$\begin{aligned}
q + p\delta &= \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \cdot (1-\alpha-\beta) = \frac{\beta(1-\alpha)+\alpha(1-\alpha)}{\alpha+\beta} = 1-\alpha, \\
p - p\delta &= \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta} \cdot (1-\alpha-\beta) = \frac{\alpha(\alpha+\beta)}{\alpha+\beta} = \alpha, \\
q - q\delta &= \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta} \cdot (1-\alpha-\beta) = \frac{\beta(\alpha+\beta)}{\alpha+\beta} = \beta, \\
p + q\delta &= \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta} \cdot (1-\alpha-\beta) = \frac{\alpha(1-\beta)+\beta(1-\beta)}{\alpha+\beta} = 1-\beta.
\end{aligned}$$

Tudíž můžeme matici pravděpodobností přechodu napsat ve tvaru

$$\mathbf{P} = \begin{pmatrix} q+p\delta & p-p\delta \\ q-q\delta & p+q\delta \end{pmatrix}.$$

Nyní nás bude zajímat význam parametru δ. K tomu si napočítáme kovariance a korelace náhodných veličin Z_i a Z_{i+k}. Abychom ale mohli spočítat tyto kovariance, budeme potřebovat pravděpodobnosti přechodu k-tého řádu, tedy podle věty 2 budeme muset vyjádřit matici pravděpodobností přechodu po k krocích $\mathbf{P}^k$. Takže si nejdříve vypočítáme k-tou mocninu matice $\mathbf{P}$. Chceme si matici $\mathbf{P}$ diagonalizovat, abychom ji mohli snadno umocnit. Vypočítáme tedy charakteristický polynom matice $\mathbf{P}$, neboli determinant matice $(\mathbf{P} - \lambda\mathbf{I_2})$, a budeme hledat jeho kořeny.

$$\mathbf{P} = \begin{pmatrix} q+p\delta & p-p\delta \\ q-q\delta & p+q\delta \end{pmatrix}$$

$$\mathbf{P} - \lambda\mathbf{I_2} = \begin{pmatrix} q+p\delta-\lambda & p-p\delta \\ q-q\delta & p+q\delta-\lambda \end{pmatrix}$$

$$\begin{aligned}0 = \det(\mathbf{P} - \lambda\mathbf{I_2}) &= (q + p\delta - \lambda)(p + q\delta - \lambda) - (1-\delta)^2 pq \\ &= (q + p\delta)(p + q\delta) - \lambda(q + p\delta) - \lambda(p + q\delta) + \lambda^2 - (1-\delta)^2 pq \\ &= \lambda^2 - \lambda[p + q + \delta(p + q)] - pq(1-\delta)^2 + (q + p\delta)(p + q\delta) \\ &= \lambda^2 - \lambda(1 + \delta) - pq(1-\delta)^2 + (q + p\delta)(p + q\delta)\end{aligned}$$

Vypočítáme vlastní čísla matice $\mathbf{P}$ jakožto kořeny charakteristického polynomu.

$$\begin{aligned}\lambda_{1,2} &= \frac{1 + \delta \pm \sqrt{(1+\delta)^2 - 4(p^2\delta + 2\delta pq + q^2\delta)}}{2} \\ &= \frac{1 + \delta \pm \sqrt{(1+\delta)^2 - 4\delta(p+q)^2}}{2} \\ &= \frac{1 + \delta \pm \sqrt{(1-\delta)^2}}{2}\end{aligned}$$

Tedy dostáváme vlastní čísla $\lambda_1 = 1$ a $\lambda_2 = \delta$. K nim dopočítáme i příslušné vlastní vektory.

Pro $\lambda_1 = 1$ matice $\mathbf{P} - \lambda\mathbf{I_2}$ vypadá následovně:

$$\begin{pmatrix} q + p\delta - 1 & p - p\delta \\ q - q\delta & p + q\delta - 1 \end{pmatrix}.$$

Vidíme, že tuto soustavu rovnic řeší například vektor $\binom{1}{1}$, tedy vlastní vektor příslušný vlastnímu číslu $\lambda_1 = 1$ je $v_1 = \binom{1}{1}$.

Pro $\lambda_1 = \delta$ matice $\mathbf{P} - \lambda\mathbf{I_2}$ vypadá potom takto:

$$\begin{aligned}\begin{pmatrix} q + p\delta - \delta & p - p\delta \\ q - q\delta & p + q\delta - \delta \end{pmatrix} &= \begin{pmatrix} q + p\delta - \delta(p+q) & p - p\delta \\ q - q\delta & p + q\delta - \delta(p+q) \end{pmatrix} \\ &= \begin{pmatrix} q(1-\delta) & p(1-\delta) \\ q(1-\delta) & p(1-\delta) \end{pmatrix}.\end{aligned}$$

Tuto soustavu zase řeší například vektor $\binom{-p}{q}$, tedy vlastní vektor příslušný vlastnímu číslu $\lambda_2 = \delta$ je $v_2 = \binom{-p}{q}$.

Potom platí:

$$\mathbf{P}^k = \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix} \begin{pmatrix} 1^k & 0 \\ 0 & \delta^k \end{pmatrix} \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix}^{-1},$$

kde

$$\begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix}^{-1} = \begin{pmatrix} q & p \\ -1 & 1 \end{pmatrix}.$$

Tedy dopočítáme k-tou mocninu matice $\mathbf{P}$:

$$\begin{aligned}\mathbf{P}^k &= \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \delta^k \end{pmatrix} \begin{pmatrix} q & p \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -p\delta^k \\ 1 & q\delta^k \end{pmatrix} \begin{pmatrix} q & p \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} q + p\delta^k & p - p\delta^k \\ q - q\delta^k & p + q\delta^k \end{pmatrix}.\end{aligned}$$

Nyní již můžeme napočítat kovariance

$$\begin{aligned}\operatorname{cov}(Z_i,Z_{i+k}) &= \mathsf{E}\,(Z_iZ_{i+k}) - \mathsf{E}\,(Z_i)\,\mathsf{E}\,(Z_{i+k}) \\ &= P(Z_i = 1, Z_{i+k} = 1) - p^2 \\ &= p(p + q\delta^k) - p^2 \\ &= pq\delta^k.\end{aligned}$$

Dostali jsme, že

$$\operatorname{cov}(Z_i,Z_{i+k}) = pq\delta^k. \tag{2.1}$$

Korelace pak dopočítáme snadno jako:

$$\operatorname{corr}(Z_i,Z_{i+k}) = \frac{\operatorname{cov}(Z_i,Z_{i+k})}{\sqrt{\operatorname{var} Z_i \operatorname{var} Z_{i+k}}} = \frac{pq\delta^k}{pq} = \delta^k.$$

Tedy vidíme, že parametr δ vyjadřuje pro $i \geq 1$ korelaci mezi veličinami Z_i a Z_{i+1}. Zároveň z toho, že prvky v matici pravděpodobností přechodu $\mathbf{P}$ jsou mezi 0 a 1, dostáváme omezení na δ v závislosti na p. Z nerovnic

$$\begin{aligned}0 < q + p\delta < 1 \\ 0 < p + q\delta < 1\end{aligned}$$

dostáváme, že $\delta > \max\{-(1-p)/p, -p/(1-p)\}$ a zároveň $\delta < 1$.

Předpokládejme tedy, že máme posloupnost $\{Z_i, i = 1, 2, \ldots\}$ z homogenního stacionárního Markovova řetězce se stavy 0 a 1 a parametry p a δ. Nyní uvažujme náhodnou veličinu X, která bude představovat počet úspěchů v n markovských pokusech, tj. $X = \sum_{i=1}^n Z_i$. Potom řekneme, že náhodná veličina X má rozdělení z *Markovského binomického modelu s parametry* $n, p, \delta, \boldsymbol{P}$.

2.2 Vytvořující funkce

První, čím se budeme zabývat, bude vytvořující funkce. Předpokládejme, že $X = \sum_{i=1}^n Z_i$ je náhodná veličina s rozdělením z Markovského binomického modelu s parametry $n, p, \delta, \mathbf{P}$. Potom vytvořující funkce náhodné veličiny X je

$$P_X = \mathsf{E}\; s^X = \mathsf{E}\; s^{\sum_{i=1}^n Z_i}.$$

Z vlastností podmíněné střední hodnoty pak platí

$$\mathsf{E}\; s^{\sum_{i=1}^n Z_i} = \mathsf{E}\left[\mathsf{E}\left[s^{\sum_{i=1}^n Z_i} | Z_1\right]\right].$$

Počítejme tedy

$$\begin{aligned}\mathsf{E}\left[\mathsf{E}\left[s^{\sum_{i=1}^n Z_i} | Z_1\right]\right] &= \mathsf{E}\left[s^{\sum_{i=1}^n Z_i} | Z_1 = 1\right] \cdot P(Z_1 = 1) \\ &\quad + \mathsf{E}\left[s^{\sum_{i=1}^n Z_i} | Z_1 = 0\right] \cdot P(Z_1 = 0) \\ &= ps\,\mathsf{E}\left[s^{\sum_{i=2}^n Z_i} | Z_1 = 1\right] + q\,\mathsf{E}\left[s^{\sum_{i=2}^n Z_i} | Z_1 = 0\right] \\ &= (q\ ps)\begin{pmatrix} \mathsf{E}\left[s^{\sum_{i=2}^n Z_i} | Z_1 = 0\right] \\ \mathsf{E}\left[s^{\sum_{i=2}^n Z_i} | Z_1 = 1\right] \end{pmatrix}.\end{aligned}$$

Vyjádříme si podmíněné střední hodnoty.

$$\begin{aligned}
\mathsf{E}\,[s^{\sum_{i=2}^{n} Z_i}|Z_1 = 0] &= \mathsf{E}\,[s^{\sum_{i=2}^{n} Z_i}|Z_1 = 0, Z_2 = 0] \cdot P(Z_2 = 0|Z_1 = 0) \\
&\quad + \mathsf{E}\,[s^{\sum_{i=2}^{n} Z_i}|Z_1 = 0, Z_2 = 1] \cdot P(Z_2 = 1|Z_1 = 0) \\
&= p_{00}\,\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_1 = 0, Z_2 = 0] \\
&\quad + p_{01}s\,\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_1 = 0, Z_2 = 1] \\
&= p_{00}\,\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 0] + p_{01}s\,\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 1] \\
&= (p_{00}\ p_{01}s) \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 0] \\ \mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 1] \end{pmatrix}
\end{aligned}$$

Ve třetí rovnosti jsme využili markovské vlastnosti (1.1). Obdobně si můžeme vyjádřit $\mathsf{E}\,[s^{\sum_{i=2}^{n} Z_i}|Z_1 = 1]$ jako

$$(p_{10}\ p_{11}s) \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 0] \\ \mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 1] \end{pmatrix}.$$

Dostáváme tedy, že

$$\begin{aligned}
\mathsf{E}\,\left[\mathsf{E}\,[s^{\sum_{i=1}^{n} Z_i}|Z_1]\right] &= (q\ ps) \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=2}^{n} Z_i}|Z_1 = 0] \\ \mathsf{E}\,[s^{\sum_{i=2}^{n} Z_i}|Z_1 = 1] \end{pmatrix} \\
&= (q\ ps) \begin{pmatrix} p_{00} & p_{01}s \\ p_{10} & p_{11}s \end{pmatrix} \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 0] \\ \mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 1] \end{pmatrix}.
\end{aligned}$$

Stejným způsobem bychom si opět mohli odvodit

$$\begin{aligned}
\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 0] &= (p_{00}\ p_{01}s) \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 0] \\ \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 1] \end{pmatrix} \\
\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2 = 1] &= (p_{10}\ p_{11}s) \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 0] \\ \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 1] \end{pmatrix}.
\end{aligned}$$

Tedy

$$\mathsf{E}\,[s^{\sum_{i=3}^{n} Z_i}|Z_2] = \begin{pmatrix} p_{00} & p_{01}s \\ p_{10} & p_{11}s \end{pmatrix} \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 0] \\ \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 1] \end{pmatrix}.$$

Tudíž bychom dostali, že

$$\mathsf{E}\ s^{\sum_{i=1}^{n} Z_i} = (q\ ps) \begin{pmatrix} p_{00} & p_{01}s \\ p_{10} & p_{11}s \end{pmatrix}^2 \begin{pmatrix} \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 0] \\ \mathsf{E}\,[s^{\sum_{i=4}^{n} Z_i}|Z_3 = 1] \end{pmatrix}.$$

A pokud bychom takto rekurzivně pokračovali dál, dospěli bychom k vytvořující funkci

$$\mathsf{E}\ s^X = \mathsf{E}\ s^{\sum_{i=1}^{n} Z_i} = (q\ ps) \begin{pmatrix} p_{00} & p_{01}s \\ p_{10} & p_{11}s \end{pmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Tedy vytvořující funkce náhodné veličiny X z Markovského binomického modelu s parametry $n, p, \delta, \mathbf{P}$ je

$$P_X(s) = (q\ ps) \begin{pmatrix} q+p\delta & (p-p\delta)s \\ q-q\delta & (p+q\delta)s \end{pmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

V článku Islam a O'shaughnessy (2013) při odvozování vytvořující funkce vycházeli ze článku Edwards (1960), kde byla vytvořující funkce uvedena správně, avšak autoři Islam a O'shaughnessy (2013) ji do svého článku uvedli ve špatném tvaru

$$(qps)\ \mathbf{P}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

namísto

$$(q\ ps) \begin{pmatrix} q+p\delta & (p-p\delta)s \\ q-q\delta & (p+q\delta)s \end{pmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Kdybychom chtěli pomocí vytvořující funkce explicitně vyjádřit pravděpodobnosti $P(X = k)$ pro $k = 1,2,\ldots,n$, bylo by to poměrně obtížné, neboť bychom potřebovali $(n-1)$-ní mocninu uvedené matice. To lze například pomocí Perronova vzorce.

Pro malá n můžeme tyto pravděpodobnosti napočítat i přímým výpočtem, kde využíváme jen počáteční rozdělení $(q, p)^\top$ a matici pravděpodobností přechodu $\mathbf{P}$. Pro $n = 1$ máme

$$P(X = k) = P(Z_1 = k) = \begin{cases} q, & k = 0, \\ p, & k = 1. \end{cases}$$

Pro $n = 2$

$$P(X = k) = P(Z_1 + Z_2 = k) = \begin{cases} q(q+p\delta), & k = 0, \\ q(p-p\delta) + p(q-q\delta), & k = 1, \\ p(p+q\delta), & k = 2. \end{cases}$$

Pro $n = 3$

$$P(X = k) = P(Z_1 + Z_2 + Z_3 = k) = \begin{cases} q(q+p\delta)^2, & k = 0, \\ pq(1-\delta)[2(q+p\delta) + q(1-\delta)], & k = 1, \\ pq(1-\delta)[2(p+q\delta) + p(1-\delta)], & k = 2, \\ p(p+q\delta)^2, & k = 3. \end{cases}$$

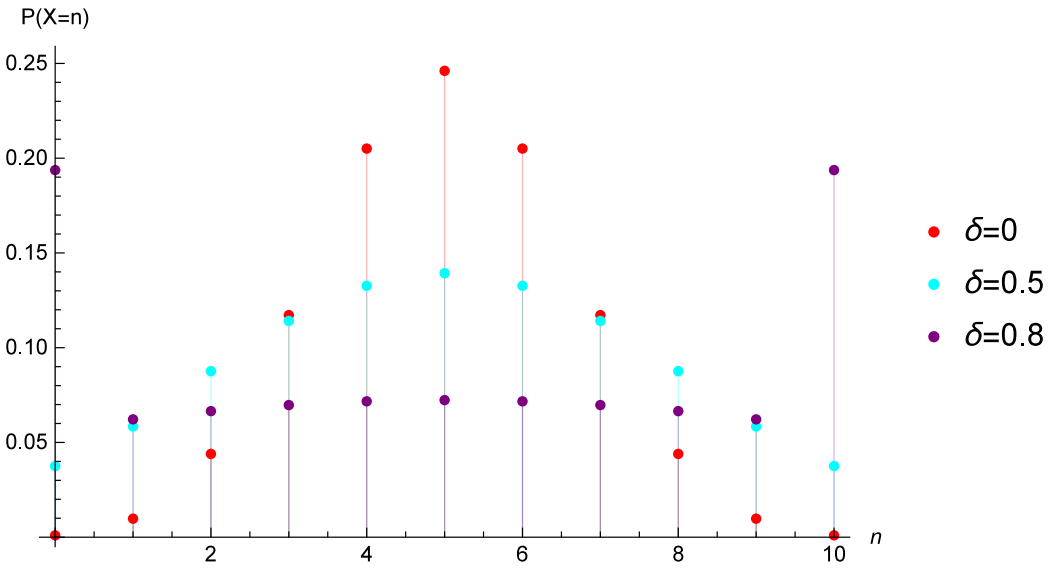
Pro $n = 4$ pak

$$P(X = k) = \begin{cases} q(q+p\delta)^3, & k = 0, \\ pq(1-\delta)(q+p\delta)[4q + 2\delta(p-q)], & k = 1, \\ pq(1-\delta)[2(p+q\delta)(q+p\delta) + 2pq(1-\delta)^2 & \\ \quad +p(1-\delta)(q+p\delta) + q(1-\delta)(p+q\delta)], & k = 2, \\ pq(1-\delta)(p+q\delta)[4p + 2\delta(q-p)], & k = 3, \\ p(p+q\delta)^3, & k = 4. \end{cases}$$

Nicméně pro zadané p, δ a n lze tyto pravděpodobnosti pomocí vytvořující funkce

napočítat například pomocí softwaru Wolfram Research Inc. (2017). Ten pro zadané n vypočítá vytvořující funkci, čímž vznikne polynom stupně n proměnné s. Potom pravděpodobnosti $P(X = k)$ budou koeficienty u s^k. Na následujícím obrázku 2.1 můžeme porovnat pravděpodobnosti pro $n = 10$ a $p = 1/2$ jak pro binomické rozdělení, tak pro rozdělení z Markovského binomického modelu. Pravděpodobnosti jsou napočítané celkem pro tři různé volby δ, kde $\delta = 0$ odpovídá binomickému rozdělení. Z obrázku vidíme, že i přesto, že jde o rozdělení se stejnou střední hodnotou, pravděpodobnosti se s δ jdoucím k jedné už poměrně liší.

Obrázek 2.1: Porovnání pravděpodobností $P(X = n)$ pro různé volby δ



2.3 Střední hodnota a rozptyl

Samozřejmě nás budou zajímat také základní vlastnosti Markovského binomického modelu jako jsou střední hodnota a rozptyl. O nich hovoří následující lemma.

Lemma 3. *Nechť $X = \sum_{i=1}^{n} Z_i$ je náhodná veličina s rozdělením z Markovského binomického modelu s parametry $n, p, \delta, \boldsymbol{P}$. Dále nechť platí $\delta = 1 - \alpha - \beta$ a $p = \frac{\alpha}{\alpha+\beta}$. Potom $\mathsf{E}\, X = np$ a $\mathrm{var}\, X = npq + 2pq\frac{\delta}{(1-\delta)^2}[n(1-\delta) - 1 + \delta^n]$.*

Důkaz. Zřejmě platí

$$\mathsf{E}\, X = \mathsf{E}\left(\sum_{i=1}^{n} Z_i\right) = \sum_{i=1}^{n} \mathsf{E}\,(Z_i) = \sum_{i=1}^{n} (0 \cdot q + 1 \cdot p) = np.$$

Pro rozptyl dále platí

$$\mathrm{var}\, X = \mathrm{var}\left(\sum_{i=1}^{n} Z_i\right) = \sum_{i=1}^{n} \mathrm{var}(Z_i) + 2\sum_{i=1}^{n-1}\sum_{k=1}^{n-i} \mathrm{cov}(Z_i, Z_{i+k}).$$

Použijeme napočítané kovariance z (2.1) a dostáváme:

$$\begin{aligned}
\operatorname{var} X &= \sum_{i=1}^{n}[\mathsf{E}\,(Z_i^2) - (\mathsf{E}\ Z_i)^2] + 2\sum_{i=1}^{n-1}\sum_{k=1}^{n-i} pq\delta^k \\
&= n(p-p^2) + 2pq\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\delta^k \\
&= npq + 2pq\sum_{i=1}^{n-1}\frac{\delta}{1-\delta}(1-\delta^{n-i}) \\
&= npq + 2pq\cdot\frac{\delta}{1-\delta}\sum_{j=1}^{n-1} 1-\delta^j \\
&= npq + 2pq\cdot\frac{\delta}{1-\delta}[n-1-\frac{\delta}{1-\delta}(1-\delta^{n-1})] \\
&= npq + 2pq\cdot\frac{\delta}{(1-\delta)^2}[n(1-\delta)-1+\delta^n].
\end{aligned}$$

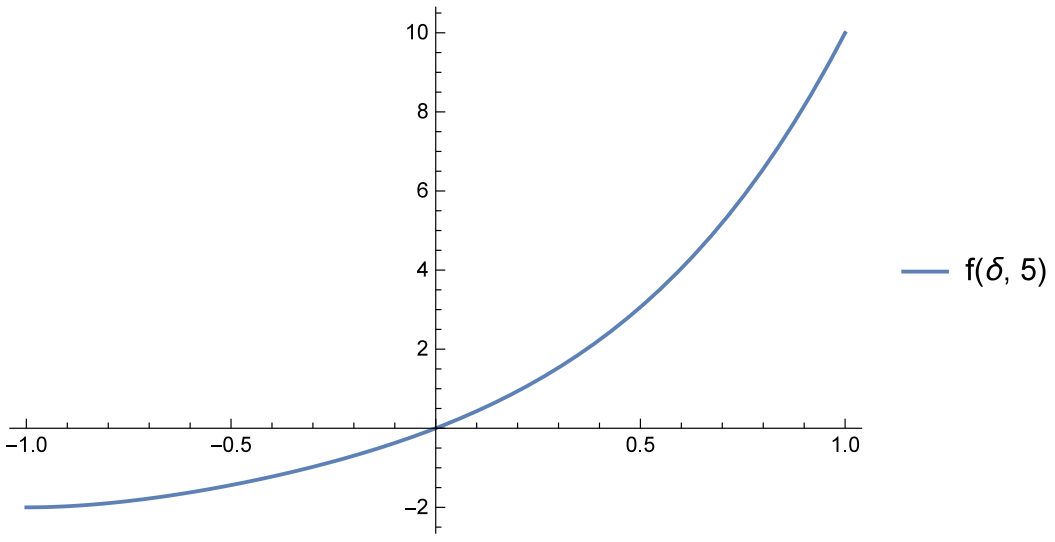
□

Podíváme se, jak rozptyl náhodné veličiny X závisí na parametru δ. Označme

$$f(\delta, n) = \frac{\delta}{(1-\delta)^2}[n(1-\delta)-1+\delta^n].$$

Tato funkce je rostoucí v δ, tedy rozptyl bude maximální, pokud δ bude maximální, a naopak rozptyl bude minimální, pokud δ bude minimální. Monotonii

Obrázek 2.2: Závislost rozptylu na δ



funkce f můžeme nahlédnout z obrázku 2.2, který vykresluje funkci f pro $n = 5$. Rozptyl tedy bude maximální, pokud se bude δ blížit k 1. To by znamenalo, že vlastně sčítáme n stejných náhodných veličin, tj. $X = \sum_{i=1}^{n} Z_1 = nZ_1$. Pokud bychom odtud počítali rozptyl, dostali bychom

$$\operatorname{var} X = n^2 \operatorname{var} Z_1 = n^2 pq.$$

Stejný výsledek bychom získali, pokud bychom v poslední rovnosti v důkazu lemmatu 3 poslali δ k 1:

$$\operatorname{var} X = npq + 2pq\left(-\frac{n}{2} + \frac{n^2}{2}\right) = npq - npq + n^2pq = n^2pq.$$

Rozptyl bude naopak minimální, pokud bude $\delta > \max\{-(1-p)/p, -p/(1-p)\}$ minimální možné.

Pokud by nás zajímalo limitní chování X pro $n \to \infty$, využijeme toho, že X představuje počet návratů do stavu 1. Tudíž můžeme pomocí analogie centrální limitní věty (viz Prášková a Lachout (2012) podkapitola 2.7) odvodit limitní rozdělení X. Pro n velké pak máme, že X má přibližné rozdělení

$$N\left(np, \frac{npq(1+\delta)}{1-\delta}\right).$$

2.4 Odhady parametrů

Nyní nás budou zajímat odhady parametrů náhodné veličiny z Markovského binomického modelu, a to konkrétně parametry p a δ. Použijeme dvě metody k nalezení bodového odhadu, momentovou metodu a metodu maximální věrohodnosti. Zároveň odlišíme dva různé přístupy pro zadaná data, ze kterých odhady počítáme. Můžeme mít totiž k dispozici buď jen realizace náhodné veličiny X, nebo můžeme znát i průběh celých řetězců.

2.4.1 Odhadování parametrů založené pouze na X

Uvažujeme N nezávislých Markovových řetězců $\{Z_j, j = 1, \ldots, n\}$ s parametry $p, \delta, \mathbf{P}$. Pro každé $i = 1, \ldots, N$ označme $X_i = \sum_{j=1}^{n} Z_{ij}$. Potom tedy $X_1, X_2, \ldots, X_N$ tvoří náhodný výběr o rozsahu N z Markovského binomického modelu s parametry $n, p, \delta, \mathbf{P}$.

Momentová metoda

Označme $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ výběrový průměr a $M_2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2$ výběrový druhý necentrální moment. Potom momentovou metodou získáme následující rovnosti.

$$np = \overline{X}_N \tag{2.2}$$

$$npq + 2pq \cdot \frac{\delta}{(1-\delta)^2}[n(1-\delta) - 1 + \delta^n] + n^2p^2 = M_2 \tag{2.3}$$

Z rovnice (2.2) získáme bodový odhad $\widehat{p} = \overline{X}_N/n$ parametru p. Ten už tedy můžeme dosadit do rovnice (2.3) a rovnici upravíme:

$$n\widehat{p}\widehat{q} + 2\widehat{p}\widehat{q} \cdot \frac{\delta}{(1-\delta)^2}[n(1-\delta) - 1 + \delta^n] + n^2\widehat{p}^2 = M_2$$

$$n\widehat{p}\widehat{q} - 2\delta n\widehat{p}\widehat{q} + \delta^2 n\widehat{p}\widehat{q} + 2\delta n\widehat{p}\widehat{q} - 2\delta^2 n\widehat{p}\widehat{q} - 2\delta\widehat{p}\widehat{q} + 2\delta^{n+1}\widehat{p}\widehat{q} + n^2\widehat{p}^2 - 2\delta n^2\widehat{p}^2 + \delta^2 n^2\widehat{p}^2 - M_2 + 2\delta M_2 - \delta^2 M_2 = 0,$$

kde $\widehat{q} = 1 - \widehat{p}$. Využijeme této rovnosti a dostaneme výraz závislý pouze na $\widehat{p}$:

$$\delta^{n+1}2\widehat{p}(1-\widehat{p}) - \delta^2[n\widehat{p}(1-\widehat{p}) - n^2\widehat{p}^2 + M_2] + 2\delta[M_2 - n^2\widehat{p}^2 - \widehat{p}(1-\widehat{p})] + n\widehat{p}(1-\widehat{p}) + n^2\widehat{p}^2 - M_2 = 0. \tag{2.4}$$

Tuto polynomiální rovnici můžeme pro δ vyřešit pomocí softwaru. Musíme avšak uvážit omezení na δ. Odhady parametru δ budou tedy ta $\widehat{\delta}$, která řeší (2.4) a splňují $\widehat{\delta} > \max\{-(1-\widehat{p})/\widehat{p}, -\widehat{p}/(1-\widehat{p})\}$ a $\widehat{\delta} < 1$. Vzhledem k tomuto omezení tedy může nastat i situace, kdy neexistuje žádné takové řešení. Uvedeme příklady pro obě možné situace:

- Rovnice (2.4) má právě jedno řešení.
  Například pro $X = (0,4,0,1,0,4,0,0,4,0,0,0,5,4,0,5,0,0,0,0)^\top$, kde $N = 20$ a $n = 5$, dostáváme odhady $\widehat{p} = 0{,}27$ a $\widehat{\delta} = 0{,}3248$.

- Rovnice (2.4) nemá žádné řešení.
  Například pro

  $$X = (1,2,1,1,2,2,2,2,2,2,1,1,2,1,2,2,2,3,2,2,2,3,3,1,2, 2,1,1,1,2,1,1,2,0,1,2,2,2,2,2,1,2,2,1,1,2,2,2,2,2)^\top,$$

  kde $N = 50$ a $n = 5$, dostáváme odhad $\widehat{p} = 0{,}34$ a odhad pro δ nedostaneme, protože neexistuje žádné reálné řešení, které by splňovalo omezení dané parametrem p.

Pokud nastane první situace, metoda nám dá odhad a jsme spokojeni. Avšak pokud nastane druhá situace, odhad požadovaného parametru bohužel nezískáme.

Metoda maximální věrohodnosti

Nyní budeme chtít z náhodného výběru $X_1, \ldots, X_N$ odhadnout parametry p a δ pomocí metody maximální věrohodnosti. Již víme, že pro malá n lze pravděpodobnosti $P(X = k)$ napočítat přímým výpočtem. Uvažujme tedy například situaci, kdy $n = 3$. Na konci podkapitoly 2.2 jsme si už vyjádřili pravděpodobnosti

$$P(X=k) = \begin{cases} q(q+p\delta)^2, & k=0, \\ pq(1-\delta)[2(q+p\delta)+q(1-\delta)], & k=1, \\ pq(1-\delta)[2(p+q\delta)+p(1-\delta)], & k=2, \\ p(p+q\delta)^2, & k=3. \end{cases}$$

Označme si $p_k = P(X = k)$ pro $k = 0{,}1{,}2{,}3$. A ještě označme N_k počet výskytů k v náhodném výběru $X_1, \ldots, X_N$ opět pro $k = 0{,}1{,}2{,}3$. Potom věrohodnostní funkci můžeme zapsat ve tvaru

$$L(p,\delta) = \prod_{k=0}^{3} p_k^{N_k} = p_0^{N_0} p_1^{N_1} p_2^{N_2} p_3^{N_3}.$$

Logaritmická věrohodnostní funkce pak bude

$$l(p,\delta) = N_0 \log p_0 + N_1 \log p_1 + N_2 \log p_2 + N_3 \log p_3.$$

Maximálně věrohodné odhady můžeme hledat buď rovnou numericky jako maximum funkce $l(p,\delta)$, nebo bychom funkci $l(p,\delta)$ parciálně zderivovali podle p a δ a tyto derivace bychom položily rovné nule. Odhady by pak byly řešeními této soustavy rovnic. Z teorie odhadu pomocí metody maximální věrohodnosti pak vyplývá, že dané odhady budou konzistentní a asymptoticky normální, viz Anděl (2011) podkapitola 7.6.

Pro $n = 3$ uvažme ještě jednou náhodný výběr $X_1, \ldots, X_M$ z Markovského binomického modelu. Označme opět N_k četnosti X_j pro $k = 0, 1, 2, 3$. Dále označme $\boldsymbol{\theta} = (p, \delta)^\top \in \Theta$ a $\mathbf{p}(\boldsymbol{\theta}) = (p_0(\boldsymbol{\theta}), p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), p_3(\boldsymbol{\theta}))^\top$, kde $p_k(\boldsymbol{\theta})$ pro $k = 0, 1, 2, 3$ jsou jako výše. Pak vektor $\mathbf{N} = (N_0, N_1, N_2, N_3)^\top$ má multinomické rozdělení $\mathrm{Mult}_4(M, \mathbf{p}(\boldsymbol{\theta}))$. Za předpokladu, že máme dostatečně velký počet pozorování v každé složce $\mathbf{N}$, pak můžeme testovat pomocí χ^2 testu dobré shody, zda vektor $\mathbf{N}$ lze skutečně popsat tímto modelem, tj. testujeme

$$H_0 : \exists\, \boldsymbol{\theta} \in \Theta \quad \mathbf{p} = \mathbf{p}(\boldsymbol{\theta})$$

$$H_1 : \forall\, \boldsymbol{\theta} \in \Theta \quad \mathbf{p} \neq \mathbf{p}(\boldsymbol{\theta}).$$

K tomu budeme potřebovat odhad $\widehat{\boldsymbol{\theta}}$, který získáme již popsanou metodou maximální věrohodnosti. Použijeme testovou statistiku

$$\chi^2 = \sum_{k=1}^{4} \frac{[N_k - M p_k(\widehat{\boldsymbol{\theta}})]^2}{M p_k(\widehat{\boldsymbol{\theta}})},$$

která má za H_0 asymptotické rozdělení χ^2_{4-d-1}, kde d je počet testovaných parametrů. V našem případě je tedy $d = 2$. A protože proti nulové hypotéze svědčí velké hodnoty testové statistiky, H_0 zamítáme právě tehdy, když $\chi^2 \geq \chi^2_1(1-\alpha)$, kde $\chi^2_1(1-\alpha)$ značí $(1-\alpha)$-kvantil rozdělení χ^2_1 pro $\alpha \in (0,1)$.

Výše uvedený postup metody maximální věrohodnosti i χ^2 test dobré shody můžeme samozřejmě provést pro jakékoliv n, avšak pro větší n už je nutné napočítat pravděpodobnosti $P(X = k)$ pomocí softwaru. Na konci podkapitoly 2.2 jsou tyto pravděpodobnosti uvedeny pro $n = 1, \ldots, 4$.

2.4.2 Odhadovaní parametrů založené na průběhu řetězců

Předpokládejme, že $X_1, X_2, \ldots, X_N$ je náhodný výběr o rozsahu N s rozdělením z Markovského binomického modelu s parametry $n, p, \delta, \mathbf{P}$ stejně jako výše v minulé podkapitole. Navíc ale budeme znát i informaci o průběhu všech N řetězců $\{Z_j, j = 1, \ldots, n\}$. Opět budeme chtít odhadnout parametry p a δ tentokrát pouze pomocí metody maximální věrohodnosti.

Díky větě 1 si můžeme pravděpodobnost jednotlivých stavů Markovova řetězce rozepsat do součinu pravděpodobností, tudíž věrohodnostní funkci můžeme

napsat v následujícím tvaru:

$$L(p,\delta) = p^{N_1}(1-p)^{N_0} \prod_{i,j=0}^{1} p_{ij}^{N_{ij}},$$

kde N_1 je počet Markovových řetězců začínajících ve stavu 1 (začínajících úspěchem), N_0 je počet Markovových řetězců začínajících ve stavu 0 (začínajících neúspěchem), tedy $N_0 = N - N_1$, a N_{ij} je celkový počet přechodů ze stavu i do stavu j napozorovaných v N Markovových řetězcích. Protože máme N Markovových řetězců a každý řetězec obsahuje n pokusů, celkový počet napozorovaných přechodů bude $(n-1)N$.

Pokud tedy uvažujeme počáteční rozdělení $(q,p)^\top$ a matici pravděpodobností přechodu

$$\mathbf{P} = \begin{pmatrix} q+p\delta & p-p\delta \\ q-q\delta & p+q\delta \end{pmatrix},$$

můžeme věrohodnostní funkci upravit následovně

$$\begin{aligned} L(p,\delta) &= p^{N_1} q^{N_0} (q+p\delta)^{N_{00}} (p-p\delta)^{N_{01}} (q-q\delta)^{N_{10}} (p+q\delta)^{N_{11}} \\ &= p^{N_1} q^{N_0} (q+p\delta)^{N_{00}} p^{N_{01}} (1-\delta)^{N_{01}} q^{N_{10}} (1-\delta)^{N_{10}} (p+q\delta)^{N_{11}} \\ &= p^{N_1+N_{01}} q^{N_0+N_{10}} (q+p\delta)^{N_{00}} (p+q\delta)^{N_{11}} (1-\delta)^{N_{01}+N_{10}}. \end{aligned}$$

Kdybychom chtěli hledat maximálně věrohodné odhady parametrů p a δ pomocí funkce $L(p,\delta)$, řešení by bylo poměrně komplikované. Proto budeme uvažovat trochu modifikovanou věrohodnostní funkci, která zanedbává pravděpodobnosti počátečního stavu Z_1

$$L^*(p,\delta) = p^{N_{01}} q^{N_{10}} (q+p\delta)^{N_{00}} (p+q\delta)^{N_{11}} (1-\delta)^{N_{01}+N_{10}}.$$

Potom je řešení jednodušší než v případě, kdy zohledňujeme i pravděpodobnosti počátečního stavu. Pro výpočet odhadů budeme uvažovat věrohodnostní funkci ve tvaru

$$L^*(p) = \prod_{i,j=0}^{1} p_{ij}^{N_{ij}},$$

kde p_{ij} jsou pravděpodobnosti přechodu z matice $\mathbf{P}$. Logaritmická věrohodnostní funkce pak bude

$$l^*(p) = \sum_{i,j=0}^{1} N_{ij} \log p_{ij} = N_{00} \log p_{00} + N_{01} \log p_{01} + N_{10} \log p_{10} + N_{11} \log p_{11}.$$

Před tím, než budeme parciálně derivovat, musíme vzít ještě v potaz platnost následujících vztahů:

$$\begin{aligned} p_{00} &= 1 - p_{01}, \\ p_{11} &= 1 - p_{10}. \end{aligned}$$

Potom můžeme $l^*(p)$ přepsat tak, aby závisela pouze na p_{00} a p_{11}:

$$l^*(p) = N_{00} \log p_{00} + N_{01} \log(1-p_{00}) + N_{10} \log(1-p_{11}) + N_{11} \log p_{11}.$$

Nyní už můžeme parciálně derivovat.

$$\frac{\partial l^*(p)}{\partial p_{00}} = \frac{N_{00}}{p_{00}} - \frac{N_{01}}{1-p_{00}}$$
$$\frac{\partial l^*(p)}{\partial p_{11}} = \frac{N_{11}}{p_{11}} - \frac{N_{10}}{1-p_{11}}$$

Tyto derivace položíme rovné nule a vyjádříme p_{00} a p_{11}. Dostaneme maximálně věrohodné odhady

$$\widehat{p}_{00} = \frac{N_{00}}{N_{00}+N_{01}} \quad , \quad \widehat{p}_{11} = \frac{N_{11}}{N_{10}+N_{11}}$$

parametrů p_{00} a p_{11}. K nalezení maximálně věrohodného odhadu parametrů p a δ využijeme následující věty, viz Anděl (2011) podkapitola 7.6.

Věta 4 (Zehnova)**.** *Je-li $\widehat{\theta}$ maximálně věrohodný odhad parametru θ, pak $u(\widehat{\theta})$ je maximálně věrohodný odhad parametrické funkce $u(\theta)$.*

Díky vztahům mezi parametry p_{00}, p_{11} a parametry p, δ získáme dvě rovnice o dvou neznámých p a δ:

$$\widehat{p}_{00} = 1 - p + p\delta,$$
$$\widehat{p}_{11} = p + (1-p)\delta.$$

Řešením této soustavy dostaneme bodové odhady

$$\widehat{p} = \frac{1-\widehat{p}_{00}}{2-\widehat{p}_{00}-\widehat{p}_{11}} \quad , \quad \widehat{\delta} = \widehat{p}_{00} + \widehat{p}_{11} - 1.$$

Z věty 4 pak tedy máme, že tyto odhady jsou maximálně věrohodné.

3. Ilustrace na datech

V této kapitole budeme nejprve na simulovaných datech porovnávat jednotlivé metody výpočtu odhadů parametrů p a δ. V další části pomocí těchto metod vypočítáme parametry p a δ z napozorovaných reálných dat.

3.1 Simulovaná data

Pomocí softwaru R Core Team (2018) si nasimulujeme pro $n = 5$ a zadané p, δ, N celkem N Markovových řetězců. Sečtením úspěchů v každém řetězci pak získáme náhodný výběr $X_1, \ldots, X_N$. Pro výpočet parametrů p a δ použijeme celkem tři metody – momentovou metodu (MM), metodu maximální věrohodnosti založenou pouze na náhodném výběru $X_1, \ldots, X_N$ (MLEX) a metodu maximální věrohodnosti založenou i na informaci o průběhu jednotlivých řetězců (MLE).

Nasimulovali jsme si data nejprve pro $N = 50, p = 1/3$ a tři volby δ. V tabulce 3.1 jsou uvedeny četnosti X_j pro $j = 1, \ldots, 50$.

Tabulka 3.1: Tabulka četností $X_j = k$ pro různé volby δ

k	0	1	2	3	4	5
$\delta = -0{,}4$	5	14	28	3	0	0
$\delta = 0{,}1$	11	12	17	9	1	0
$\delta = 0{,}8$	27	5	2	2	6	8

Výsledné odhady pro každou ze tří metod shrnuje tabulka 3.2.

Tabulka 3.2: Odhady parametrů pro $N = 50$ a $p = 1/3$

		$\delta = -0{,}4$	$\delta = 0{,}1$	$\delta = 0{,}8$
MLE	$\widehat{p}$	0,3235	0,3166	0,3434
	$\widehat{\delta}$	-0,4320	0,0976	0,7859
MLEX	$\widehat{p}$	0,3155	0,3079	0,3092
	$\widehat{\delta}$	-0,4610	0,0686	0,8008
MM	$\widehat{p}$	0,3160	0,3080	0,3160
	$\widehat{\delta}$	-0,3887	0,0577	0,8201

Abychom mohli porovnat jednotlivé metody, provedeme 1000 opakování, tedy získáme 1000 odhadů parametru p a δ pomocí každé metody. Tyto odhady zprůměrujeme a vypočítáme směrodatnou odchylku a tzv. bias, kde skutečnou hodnotu parametru odečteme od průměru. Toto uděláme pro stejné volby parametrů jako výše, tedy pro $n = 5$, $p = 1/3$ a tři volby $\delta = -0{,}4;\ 0{,}1;\ 0{,}8$. Tentokrát ale výpočty provedeme pro různé rozsahy výběrů, pro N rovno 50, 200 a 1000.

Jak už bylo zmíněno dříve, při výpočtu odhadu parametru δ pomocí momentové metody může dojít k tomu, že výsledný odhad nezískáme. To nastane, pokud neexistuje reálné řešení, které by splňovalo dané omezení. V tomto případě tedy metoda selže. Tabulka níže ukazuje v kolika procentech metoda selže pro jednotlivé případy.

Tabulka 3.3: Procenta selhání MM při výpočtu odhadu δ

	$\delta = -0{,}4$	$\delta = 0{,}1$	$\delta = 0{,}8$
N=50	24,5%	0,0%	0,0%
N=200	3,8%	0,0%	0,0%
N=1000	0,0%	0,0%	0,0%

Výsledky simulací jsou k nahlédnutí v následujících tabulkách, kde jsou pro lepší přehlednost všechny hodnoty přenásobeny 100.

Tabulka 3.4: Simulace pro $\delta = -0{,}4$; hodnoty jsou přenásobeny 100

$\widehat{p}$						
	MLE		MLEX		MM	
	BIAS	SD	BIAS	SD	BIAS	SD
N=50	0,022	2,177	-0,035	2,127	-0,051	2,129
N=200	0,004	1,100	0,003	1,081	0,006	1,083
N=1000	-0,005	0,497	-0,005	0,481	-0,004	0,005
$\widehat{\delta}$						
	MLE		MLEX		MM	
	BIAS	SD	BIAS	SD	BIAS	SD
N=50	-0,304	5,258	-0,997	8,959	2,305	8,784
N=200	-0,052	2,672	-0,530	5,108	-0,080	5,232
N=1000	0,017	1,193	-0,057	2,190	-0,009	2,414

Tabulka 3.5: Simulace pro $\delta = 0{,}1$; hodnoty jsou přenásobeny 100

$\widehat{p}$						
	MLE		MLEX		MM	
	BIAS	SD	BIAS	SD	BIAS	SD
N=50	-0,037	3,819	-0,124	3,313	-0,125	3,314
N=200	-0,003	1,835	0,016	1,579	0,016	1,579
N=1000	-0,034	0,820	-0,021	0,710	-0,021	0,709
$\widehat{\delta}$						
	MLE		MLEX		MM	
	BIAS	SD	BIAS	SD	BIAS	SD
N=50	-0,352	7,322	-1,591	11,021	-1,565	11,088
N=200	-0,103	3,683	-0,382	5,430	-0,367	5,448
N=1000	0,058	1,589	0,025	2,354	0,028	2,358

Tabulka 3.6: Simulace pro $\delta = 0{,}8$; hodnoty jsou přenásobeny 100

$\widehat{p}$						
	MLE		MLEX		MM	
	BIAS	SD	BIAS	SD	BIAS	SD
N=50	0,959	10,277	-0,203	5,572	-0,175	5,616
N=200	0,034	4,924	-0,004	2,798	0,007	2,848
N=1000	-0,008	2,248	-0,011	1,260	-0,011	1,269
$\widehat{\delta}$						
	MLE		MLEX		MM	
	BIAS	SD	BIAS	SD	BIAS	SD
N=50	-0,111	4,885	-0,602	5,328	-0,611	5,485
N=200	-0,048	2,372	-0,160	2,572	-0,192	2,654
N=1000	0,004	1,047	-0,039	1,103	-0,055	1,127

Z tabulek vidíme, že odchylky odhadů pomocí momentové metody a metody maximální věrohodnosti založené pouze na X jsou podobné. Při odhadování parametru p dávají obě tyto metody většinou lepší odhady než metoda maximální věrohodnosti založená na průběhu řetězců, například pro $\delta = 0{,}8$. Naopak u parametru δ jsou odchylky menší právě u odhadů pomocí metody MLE. To se dalo očekávat, neboť tato metoda využívá informaci o průběhu řetězců a může tedy lépe odhadovat závislost mezi veličinami Z_i a Z_{i+1}.

Jako nejhorší z metod se jeví momentová metoda, přestože například pro $\delta = -0{,}4$ a $N = 1000$ nám dává nejmenší odchylku odhadu parametru p i δ. Na druhou stranu, momentová metoda není schopna vždy dát odhad parametru δ, v čemž spočívá její největší nevýhoda.

3.2 Reálná data

V této části budeme odhadovat parametry z reálných dat. Budeme se zabývat nehodovostí pod vlivem alkoholu v roce 2018 v Praze a ve Zlínském kraji. Data o počtu nehod pod vlivem alkoholu pro každý den roku 2018 pro oba kraje jsme získali z webových stránek Českého rozhlasu https://www.irozhlas.cz/nehody.

Nejdříve každému pracovnímu dni v roce 2018 přiřadíme 0, pokud se v daný den nestala žádná dopravní nehoda pod vlivem alkoholu. Pokud se v daný den stala aspoň jedna dopravní nehoda pod vlivem alkoholu, přiřadíme mu 1. Budeme předpokládat, že pracovní dny v jednom týdnu budou tvořit jeden Markovův řetězec délky 5. Záznamy o nehodách v sobotu a v neděli vynecháme a protože má aplikace spíše ilustrativní charakter, zanedbáme zbytkovou závislost mezi řetězci z každých dvou po sobě jdoucích týdnů. Budeme tudíž předpokládat, že jednotlivé řetězce budou na sobě nezávislé. Dále budeme pro jednoduchost předpokládat, že všechny řetězce budou stacionární a budou mít stejné parametry. Zanedbáváme tedy případné sezonní chování.

Pokud si to tedy přepíšeme matematicky, budeme mít pro každé $j = 1, \ldots, 52$ (pro každý týden v roce 2018) Markovův řetězec $\{Z_{ji}, i = 1, \ldots, 5\}$ délky 5. Pro každé $j = 1, \ldots, 52$ dále označíme $X_j = \sum_{i=1}^{5} Z_{ji}$ počet nehodových dnů v j-tém

týdnu. Potom náhodné veličiny $X_1, \ldots, X_{52}$ budou představovat počet nehodových dnů v každém týdnu roku 2018 a budou tvořit náhodný výběr z Markovského binomického modelu.

Budou nás zajímat bodové odhady parametrů p a δ, tedy odhad pravděpodobnosti, že se v daný den stane alespoň jedna nehoda pod vlivem alkoholu, a odhad korelace mezi nehodovostí dvou následujících dnů. Opět použijeme tři metody nalezení odhadu jako výše pro simulovaná data.

Data, ze kterých jsme odhady počítali, jsou k nahlédnutí v přílohách A.1 a A.2. Zde uvádíme alespoň tabulku četností X_j pro $j = 1, \ldots, 52$ pro oba kraje.

Tabulka 3.7: Tabulka četností $X_j = k$

k	0	1	2	3	4	5
četnost Praha	0	6	10	22	12	2
četnost Zlín	1	13	19	12	6	1

Následující tabulka shrnuje výsledné odhady obou parametrů pro každou metodu jak pro kraj Praha, tak pro Zlínský kraj.

Tabulka 3.8: Odhady parametrů z reálných dat

Praha	$\widehat{p}$	$\widehat{\delta}$
MM	0,57692	-0,10818
MLE	0,56295	0,04497
MLEX	0,57690	-0,10337
Zlínský kraj	$\widehat{p}$	$\widehat{\delta}$
MM	0,44615	-0,05065
MLE	0,42741	0,03186
MLEX	0,44617	-0,04787

Vidíme, že odhady pomocí metod založených pouze na náhodném výběru X (MM, MLEX) jsou podobné odhadům pomocí metody založené i na informaci o průběhu řetězců (MLE), jen odhady pomocí metody MLE vychází o něco menší.

Dále můžeme nahlédnout, že parametr p pro Prahu vychází mírně nad jednu polovinu. Ve Zlínském kraji je pravděpodobnost, že se v daný den stane alespoň jedna nehoda pod vlivem alkoholu menší, naopak mírně pod jednou polovinou.

Zajímavé je, že odhad parametru δ se v případě obou krajů pohybuje kolem 0. Pokud by to 0 opravdu byla, znamenalo by to, že veličiny Z_i jsou nekorelované, což v našem případě znamená i nezávislé, tudíž by netvořily Markovův řetězec. Mohli bychom tedy na ně nahlížet jako na nezávislé náhodné veličiny z binomického rozdělení. Pak by se dalo říci, že pravděpodobnost, jestli se v daný den stane alespoň jedna nehoda pod vlivem alkoholu nezávisí na tom, jestli se stala alespoň jedna nehoda pod vlivem alkoholu předchozí den.

Závěr

Cílem této práce bylo zobecnit standardní binomické rozdělení na situaci, kdy pracujeme se součtem náhodných veličin, které jsou závislé. Uvažovali jsme posloupnost náhodných veličin, která tvořila Markovův řetězec. Pomocí tohoto Markovova řetězce jsme si potom odvodili Markovský binomický model se všemi parametry a vysvětlili význam těchto parametrů. Dále jsme se zabývali vlastnostmi tohoto modelu jako je střední hodnota, rozptyl nebo vytvořující funkce. Uvedli jsme, jak je možné využít vytvořující funkci například k výpočtu pravděpodobnostního rozdělení.

Část práce byla také věnována odhadům neznámých parametrů Markovského binomického modelu. Popsali jsme si tři metody – momentovou metodu, metodu maximální věrohodnosti založenou pouze na náhodném výběru X a metodu maximální věrohodnosti, která využívá i informaci o průběhu jednotlivých řetězců. K porovnání těchto metod jsme si data nasimulovali. Ze simulací nám poté vyplynulo, že obě metody založené pouze na X, tedy metody MM a MLEX, dávají lepší odhady parametru p. Naopak u odhadů parametru δ se lépe jeví metoda založená na průběhu řetězců, což odpovídá tomu, že má k dispozici více informace a může lépe odhadovat závislost. Jako nejhorší se ukázala momentová metoda, neboť nám v některých situacích ani nebyla schopna dát odhad požadovaného parametru.

V závěru práce jsme ilustrovali Markovský binomický model na reálných datech o nehodovosti pod vlivem alkoholu a pomocí všech tří metod jsme odhadli jeho parametry. V případné budoucí práci by se dalo pomocí χ^2 testu dobré shody testovat, zda tato data opravdu pochází z Markovského binomického modelu, či nikoliv. Pokud by se ukázalo, že data neodpovídají tomuto modelu, mohla by se uvažovat nějaká jiná varianta závisloti mezi jednotlivými veličinami než ta, že tvoří Markovův řetězec. V článku Rudolfer (1990) autor uvádí i několik dalších možných typů závislostí. Jednou z možností je například posloupnost náhodných veličin s předepsaným korelačním koeficientem ρ, tedy $\mathrm{corr}(Z_i,Z_j) = \rho$ pro každé $i \neq j$. Takový model by popisoval veličiny, které se nevyvíjí v čase, neboť každé dvě veličiny by měly stejnou korelaci, narozdíl od Markovova řetězce, kde mají stejnou korelaci jen každé dvě po sobě jdoucí veličiny.

Práce vycházela převážně ze článku Islam a O'shaughnessy (2013), avšak v něm se vyskytlo několik tvrzení, která nebyla dostatečně okomentována či odvozena. Navíc článek obsahoval několik obsahových chyb, například vytvořující funkce byla bez jakéhokoliv odvození uvedena ve špatném tvaru. Vlastní přínos práce tedy spočívá v opravení těchto chyb a důkladném odvození jednotlivých tvrzení a vlastností. To se týče například již zmíněné vytvořující funkce, dále pak třeba odvození rozptylu nebo postupu odhadování parametrů pomocí metody maximální věrohodnosti založené na průběhu řetězců. Dalším přínosem je i poslední kapitola o ilustraci modelu na reálných datech o nehodovosti v České republice v roce 2018.

Seznam použité literatury

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

Edwards, A. W. F. (1960). The Meaning of Binomial Distribution. *Nature*, **186**, 1074.

Islam, M. N. a O'shaughnessy, C. D. (2013). On the Markov Chain Binomial Model. *Applied Mathematics*, **5**, 1726–1730.

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rudolfer, S. M. (1990). A Markov Chain Model of Extrabinomial Variation. *Biometrika*, **77**(2), 255–264.

Wolfram Research Inc. (2017). Mathematica, Version 11.2. Champaign, IL, 2017.

A. Přílohy

A.1 Data o nehodovosti pod vlivem alkoholu v Praze

Dny 2018	Z_i	X	Dny 2018	Z_i	X
1.1. – 5.1.	1 0 1 0 0	2	2.7. – 6.7.	1 0 0 0 0	4
8.1. – 12.1.	1 1 0 1 0	3	9.7. – 13.7.	1 1 1 0 1	4
15.1. – 19.1.	0 0 0 0 1	1	16.7. – 20.7.	0 0 1 1 0	2
22.1. – 26.1.	1 1 1 1 1	5	23.7. – 27.7.	0 0 1 0 0	1
29.1. – 2.2.	1 0 0 1 1	3	30.7. – 3.8.	1 1 0 0 0	2
5.2. – 9.2.	1 1 0 1 1	4	6.8. – 10.8.	0 0 1 1 1	3
12.2. – 16.2.	1 0 0 0 1	2	13.8. – 17.8.	1 0 0 1 1	3
19.2. – 23.2.	1 1 0 0 0	2	20.8. – 24.8.	1 0 0 1 1	3
26.2. – 2.3.	1 1 0 1 1	4	27.8 – 31.8.	1 0 1 0 1	3
5.3. – 9.3.	0 0 1 1 1	3	3.9. – 7.9.	0 1 1 1 1	4
12.3. – 16.3.	1 0 1 1 0	3	10.9. – 14.9.	1 1 1 0 0	3
19.3. – 23.3.	1 0 1 1 1	4	17.9. – 21.9.	0 0 1 0 1	2
26.3. – 30.3.	0 0 1 1 1	3	24.9. – 28.9.	0 0 0 1 0	1
2.4. – 6.4.	1 0 0 1 0	2	1.10. – 5.10.	1 1 0 1 1	4
9.4. – 13.4.	1 1 1 0 1	4	8.10. – 12.10.	0 0 1 1 1	3
16.4. – 20.4.	1 1 1 1 1	5	15.10. – 19.10.	1 1 0 1 0	3
23.4. – 27.4.	0 0 1 1 1	3	22.10. – 26.10.	1 0 1 1 0	3
30.4. – 4.5.	0 1 1 1 1	4	29.10. – 2.11.	0 0 1 0 1	2
7.5. – 11.5.	1 1 0 0 1	3	5.11. – 9.11.	0 1 1 0 1	3
14.5. – 18.5.	1 1 1 0 1	4	12.11. – 16.11.	0 0 1 0 1	2
21.5. – 25.5.	1 0 1 1 1	4	19.11. – 23.11.	0 0 0 1 1	2
28.5. – 1.6.	0 1 1 1 0	3	26.11. – 30.11.	1 0 0 0 0	1
4.6. – 8.6.	1 1 0 0 1	3	3.12. – 7.12.	1 0 1 1 0	3
11.6. – 15.6.	0 0 1 1 1	3	10.12. – 14.12.	1 1 0 0 1	3
18.6. – 22.6.	1 1 0 0 1	3	17.12. – 21.12.	1 1 0 1 1	4
25.6. – 29.6.	1 1 1 0 1	4	24.12. – 28.12.	0 0 1 0 0	1

A.2 Data o nehodovosti pod vlivem alkoholu ve Zlínském kraji

Dny 2018	Z_i	X	Dny 2018	Z_i	X
1.1. – 5.1.	1 0 0 1 0	2	2.7. – 6.7.	1 0 0 1 0	2
8.1. – 12.1.	1 0 0 0 0	1	9.7. – 13.7.	1 1 1 0 1	4
15.1. – 19.1.	1 1 0 0 0	2	16.7. – 20.7.	0 0 0 1 0	1
22.1. – 26.1.	0 1 0 1 0	2	23.7. – 27.7.	1 1 1 1 1	5
29.1. – 2.2.	1 0 1 1 0	3	30.7. – 3.8.	0 0 1 1 1	3
5.2. – 9.2.	0 1 0 1 0	2	6.8. – 10.8.	0 0 1 1 0	2
12.2. – 16.2.	1 0 0 0 1	2	13.8. – 17.8.	0 0 1 1 1	3
19.2. – 23.2.	1 0 1 0 0	2	20.8. – 24.8.	1 1 1 0 0	3
26.2. – 2.3.	0 0 0 1 1	2	27.8 – 31.8.	0 0 0 1 1	2
5.3. – 9.3.	0 0 0 0 1	1	3.9. – 7.9.	1 1 0 1 1	4
12.3. – 16.3.	0 0 0 0 0	0	10.9. – 14.9.	0 1 0 1 0	2
19.3. – 23.3.	1 0 0 0 0	1	17.9. – 21.9.	1 0 0 0 0	1
26.3. – 30.3.	0 0 1 1 0	2	24.9. – 28.9.	0 0 1 1 1	3
2.4. – 6.4.	0 0 1 1 0	2	1.10. – 5.10.	1 0 0 1 1	3
9.4. – 13.4.	1 0 1 0 1	3	8.10. – 12.10.	1 0 1 1 1	4
16.4. – 20.4.	0 0 1 0 0	1	15.10. – 19.10.	0 0 1 1 1	3
23.4. – 27.4.	0 1 0 1 1	3	22.10. – 26.10.	1 0 0 0 1	2
30.4. – 4.5.	1 1 0 0 0	2	29.10. – 2.11.	1 0 1 0 1	3
7.5. – 11.5.	1 0 0 1 1	3	5.11. – 9.11.	1 1 0 1 1	4
14.5. – 18.5.	1 0 0 0 0	1	12.11. – 16.11.	0 1 1 1 1	4
21.5. – 25.5.	0 1 0 1 0	2	2 19.11. – 23.11.	0 0 0 0 1	1
28.5. – 1.6.	0 0 1 1 1	3	26.11. – 30.11.	0 0 0 0 1	1
4.6. – 8.6.	1 0 1 1 1	4	3.12. – 7.12.	0 0 1 0 0	1
11.6. – 15.6.	1 0 0 1 0	2	10.12. – 14.12.	1 0 0 0 0	1
18.6. – 22.6.	0 0 1 1 0	2	17.12. – 21.12.	0 1 0 0 0	1
25.6. – 29.6.	1 0 0 0 0	1	24.12. – 28.12.	1 0 0 1 0	2