UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Marvin Herzog

Pólyeve žare

Delo diplomskega seminarja

Mentor: prof. dr. Janez Bernik

Ljubljana, 2016/2017

Kazalo

Zahvala

Ob zaključku diplomske naloge se iskreno zahvaljujem mentorju prof. dr. Janezu Berniku za vso pomoč in za vsebinske ideje. Zahvala gre tudi asist. dr. Matiji Vidmarju za obrazložitev nekaterih konceptov iz njegovih člankov. Ravno tako se bi rad zahvalil moji družini za vso podporo ter moji mački Maci, ki me je pomirajala ob pisanju naloge.

Pólyeve žare

Povzetek

V diplomski nalogi bom predstavil več različic in posplošitev Pólyevih žar. Večkrat v nalogi bom s pomočjo programa R demonstriral obnašanje modelov in simuliral primere nekaterih žar. Natančneje bom obdelal Pólya-Eggenbergerjev model, kjer bom sprva izpeljal porazdelitev kroglic ob poljubnih začetnih parametrih in nekatere lastnosti modela. Dokazal bom, da je proces deleža črnih kroglic v taki žari martingal in da konvergira proti beta porazdelitvi. Izračunal bom verjetnost, da do poljubnega časa pride do izenačenja v številu belih in črnih kroglic ter pripadajočo limitno verjetnost. Preko definicije matrik zamenjav bom nato vpeljal večbarvne modele Pólyevih žar s splošnimi pravili za zamenjavo kroglic. Nazadnje bom s pomočjo markovskih verig in stohastičnih matrik zamenjav predstavil tri primere uporabnih aplikacij Pólyevih žar: Ehrenfestov model difuzije plinov, model epidemije ter "Play-the-Winner" shemo žar za klinične raziskave.

Pólya urn models

Abstract

In my thesis I present multiple variations and generalizations of the Pólya urn model. Regularly throughout the text I use the R software to demonstrate the behaviour of models and to simulate urn examples. I then present the Pólya-Eggenberger urn scheme in detail, where I derive the distribution of the balls in the urn and some of the model's properties. I prove that the black-to-white ball ratio process is a martingale and that it converges to a beta distribution. Next, I calculate the probability of an equalization in the number of black and white balls up to a certain time, and the associated limit probability. With the addition of replacement matrices I then introduce multicolor Pólya urn models with arbitrary replacement rules. Lastly I utilize Markov chains and stochastic replacement matrices to provide three examples of practical applications of the Pólya urn model: The Ehrenfest gas diffusion model, a model of epidemics and the "Play-the-Winner" urn scheme for clinical research trials.

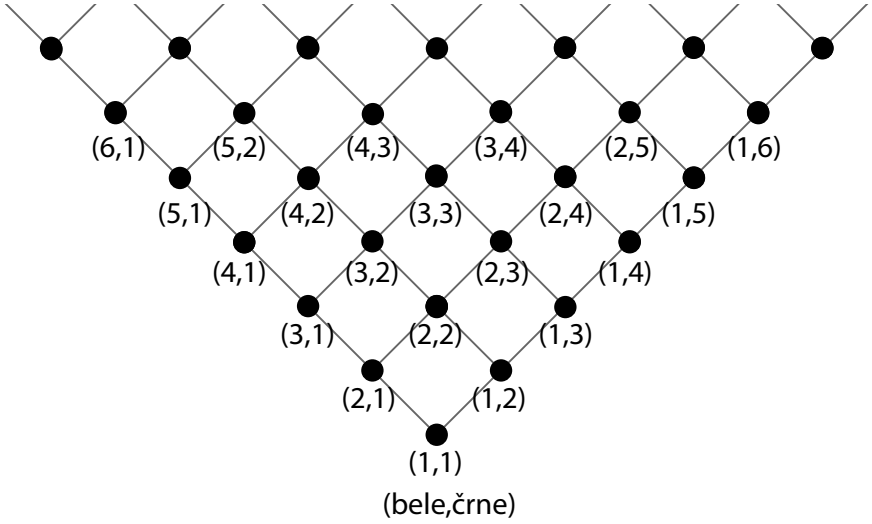
Math. Subj. Class. (2010): 60G42, 60G50, 60J10, 60J85

Ključne besede: Pólya, verjetnost, modeli žar, slučajni sprehodi, slučajni procesi

Keywords: Pólya, probability, urn models, random walks, stochastic processes

1. Uvod

Pólyeve žare so verjetnostni model, ki ga je leta 1923 predstavil madžarski matematik George (György) Pólya v sodelovanju s Florianom Eggenbergerjem v delu *Über die Statistik verketteter Vorgänge*. Gre za model žare, pri kateri imamo v osnovni različici problema eno črno in eno belo kroglico. V vsakem koraku modela naključno izvlečemo kroglico, to kroglico vrnemo v žaro, nato pa dodamo še eno kroglico iste barve kot kroglica, ki smo jo izvlekli.



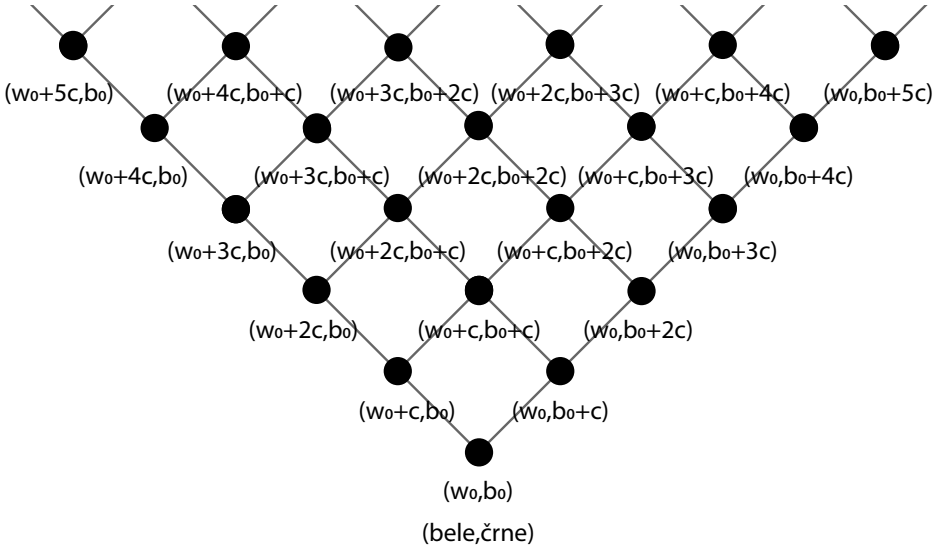
Statistična posebnost tega modela je naslednja: čim večji kot je delež kroglic neke barve (po nekaj korakih), tem bolj verjetno bo, da bomo v naslednjih korakih izvlekli kroglico prav te barve. S tovrstnim modelom lahko tako proučujemo pojave, kjer "bogati postajajo še bogatejši". Zato lahko Pólyeve žare uporabimo za modeliranje širjenja nalezljivih bolezni in epidemij, dinamike prebivalstva, evolucijskih procesov v biologiji in še kopico drugih pojavov. Zaradi svoje narave razvejanja ga lahko uporabimo kot model podatkovnih struktur v računalništvu, z njim lahko simuliramo procese odločanja, porabo internetnega paketnega prenosa, itd.

V nadaljevanju bom predstavil nekoliko posplošeno različico Pólyevih žar, kjer je na začetku v žari $B_0 \in \mathbb{N}$ črnih, $W_0 \in \mathbb{N}$ belih kroglic. Vsakič ko izvlečemo kroglico (in jo nato vrnemo), dodamo v žaro še $c \in \mathbb{Z}$ kroglic iste barve. V prvem poglavju bom vpeljal verjetnostno teorijo za procesom in se ukvarjal s porazdelitvami kroglic v žari pri različnih začetnih parametrih. Posebeno pozornost bom namenil procesu deleža črnih kroglic v žari in izpeljal njegovo limitno porazdelitev. Naslednji dve podpoglavji bom posvetil izhodni verjetnosti, tj. verjetnosti, da se po začetku procesa kadarkoli vrnemo v začetno stanje glede na razliko v številu belih in črnih kroglic. V drugem poglavju bom dodatno razširil koncept Pólyevih žar tako, da bom sprostil pravilo za vračanje kroglic in dopustil žare s kroglicami večih barv. V zadnjem poglavju bom izhajal iz posplošenih žar, spoznanih v prejšnjem poglavju. Hkrati bom vpeljal nekaj dodatnih konceptov, kot so osnove markovskih verig in stohastične žare. V tem poglavju bom tako iz zgodovinskega kot iz matematičnega vidika zajel nekaj primerov praktičnih aplikacij Pólyevih žar.

2. Pólya-Eggenbergerjeva žara

2.1. **Slučajni proces indikatorskih spremenljivk**

V naslednjih poglavjih bomo obravnavali model žare, v kateri je sprva $W_0 \in \mathbb{N}$ belih in $B_0 \in \mathbb{N}$ črnih kroglic. Na vsakem koraku žrebamo in vrnemo kroglico, nato pa dodamo še $c \in \mathbb{Z}$ kroglic iste barve. Razvoj nastalega slučajnega procesa lahko prikažemo z binomskim drevesom stanj.



Uvedimo indikatorsko slučajno spremenljivko

$$X_i := \begin{cases} 1, & \text{v času } i \text{ izvlečemo črno kroglico,} \\ 0, & \text{v času } i \text{ izvlečemo belo kroglico.} \end{cases}$$

Te lahko uredimo v (končen) slučajni proces $X = (X_1, X_2, \ldots, X_n)$.

Naj bodo $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ in naj bo $k := x_1 + x_2 + \ldots + x_n$. Zanima nas skupna porazdelitev X. Če vemo, da je v času i v žari B_i črnih ter W_i belih kroglic, potem poznamo verjetnost, da bo naslednja izvlečena črna

$$\mathbb{P}(X_{i+1} = 1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i) = \frac{B_i}{B_i + W_i}.$$

Po izreku o popolni verjetnosti sledi

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) =$$

$$= \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \cdots \mathbb{P}(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}).$$

V zgornjem produktu je v imenovalcu vsakega člena takratno število kroglic v žari. Torej dobimo produkt $(B_0 + W_0)(B_0 + W_0 + c) \ldots [B_0 + W_0 + (n-1)c]$.

V številcih k izmed faktorjev predstavlja dogodek, da smo izbrali črno kroglico. Ti tvorijo produkt $B_0(B_0+c) \ldots (B_0+(k-1)c)$. Preostalih $n-k$ faktorjev pa dogodek, da smo izvlekli belo kroglico, torej $W_0(W_0 + c) \ldots (W_0 + (n-k-1)c)$. Sledi

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) =$$

$$= \frac{[B_0(B_0 + c) \ldots (B_0 + (k-1)c)][W_0(W_0 + c) \ldots (W_0 + (n-k-1)c]}{(B_0 + W_0)(B_0 + W_0 + c) \ldots [B_0 + W_0 + (n-1)c]}.$$

Opazimo lahko, da je skupna porazdelitev odvisna od $(x_1, x_2, \ldots, x_n)$ zgolj preko $k = \sum_{i=1}^{n} x_i$. Torej je invariantna glede na permutacijo $(x_1, x_2, \ldots, x_n)$ in sledi, da so slučajne spremenljivke znotraj slučajnega procesa X zamenljive. To pa pomeni, da ima vsaka izmed permutacij $(X_1, X_2, \ldots, X_n)$ enako porazdelitev (posledica *de Finettijevega* izreka). Z drugimi besedami to pomeni, da ima v drevesu stanj vsaka izmed poti do določenega stanja enako verjetnost. Torej je X_i porazdeljena enako kot X_1 za vsak $i \in \{1, \ldots, n\}$ in velja $\mathbb{P}(X_i = 1) = \mathbb{P}(X_1 = 1) = \frac{B_0}{B_0+W_0}$. Tako lahko izračunamo

$$\mathbb{E}[X_i] = \frac{B_0}{B_0 + W_0},$$
$$\operatorname{var}(X_i) = \frac{B_0}{B_0 + W_0}\frac{W_0}{B_0 + W_0},$$
$$\operatorname{cor}(X_i, X_j) = \operatorname{cor}(X_1, X_2) = \frac{c}{B_0 + W_0 + c}.$$

Vidimo, da so X_i nekorelirane natanko tedaj, ko je $c = 0$. Še več, če primerjamo robno in skupno porazdelitev opazimo, da so pri $c = 0$ slučajne spremenljivke X_i neodvisne. Takrat gre namreč zgolj za zaporedje Bernoullijevih poskusov. Z izjemo $c = 0$ je model Pólyjevih žar eden najbolj znanih primerov slučajnih procesov, v katerem so slučajne spremenljivke zamenljive, vendar odvisne.

2.2. **Število črnih kroglic**

Za $n \in \mathbb{N}$ je število črnih kroglic, ki smo jih **izvlekli** v prvih n korakih enaka

$$Y_n := \sum_{i=1}^{n} X_i.$$

Torej je število črnih kroglic v žari po n korakih $B_0 + cY_n$. Zanima nas $\mathbb{P}(Y_n = k)$. Ker obstaja $\binom{n}{k}$ poti do tega stanja in je vsaka izmed poti enako verjetna, sledi

$$\mathbb{P}(Y_n = k) = \binom{n}{k} \frac{[B_0(B_0+c)\ldots(B_0+(k-1)c)][W_0(W_0+c)\ldots(W_0+(n-k-1)c]}{(B_0+W_0)(B_0+W_0+c)\ldots[B_0+W_0+(n-1)c]}.$$

Zgornjo verjetnost lahko (za $c \in \mathbb{N}$) zapišemo drugače s pomočjo funkcije beta

$$\mathbb{P}(Y_n = k) = \binom{n}{k} \frac{\frac{\Gamma(\frac{B_0}{c}+k)}{\Gamma(\frac{B_0}{c})}\frac{\Gamma(\frac{W_0}{c}+n-k)}{\Gamma(\frac{W_0}{c})})}{\frac{\Gamma(\frac{B_0}{c}+\frac{W_0}{c}+n)}{\Gamma(\frac{B_0}{c}+\frac{W_0}{c})}} = \binom{n}{k} \frac{\frac{\Gamma(\frac{B_0}{c}+k)\Gamma(\frac{W_0}{c}+n-k)}{\Gamma(\frac{B_0}{c}+\frac{W_0}{c}+n)}}{\frac{\Gamma(\frac{B_0}{c})\Gamma(\frac{W_0}{c})}{\Gamma(\frac{B_0}{c}+\frac{W_0}{c})}} =$$
$$= \binom{n}{k} \frac{\beta(\frac{B_0}{c}+k, \frac{W_0}{c}+n-k)}{\beta(\frac{B_0}{c}, \frac{W_0}{c})}.$$

Ta porazdelitvena funkcija je znana tudi kot **Pólyeva** oz. **Beta-binomska porazdelitev**. V primeru $c = 0$ se ta reducira na **binomsko** porazdelitev:

$$\mathbb{P}(Y_n = k) = \binom{n}{k} \left(\frac{B_0}{B_0+W_0}\right)^k \left(\frac{W_0}{B_0+W_0}\right)^{n-k}.$$

Če je $c = -1$ (ter $n \leq B_0 + W_0$), na **hipergeometrijsko** porazdelitev:

$$\mathbb{P}(Y_n = k) = \binom{n}{k} \frac{[B_0(B_0-1)\dots(B_0-k+1)][W_0(W_0-1)\dots(W_0-n+k+1)]}{(B_0+W_0)(B_0+W_0-1)\dots(B_0+W_0-n+1)} =$$
$$= \binom{n}{k} \frac{\frac{B_0!}{(B_0-k)!k!}\frac{W_0!}{(W_0-n+k)!(n-k)!}}{\frac{(B_0+W_0)!}{(B_0+W_0-n)!n!}} \frac{(n-k)!k!}{n!} = \frac{\binom{B_0}{k}\binom{W_0}{n-k}}{\binom{B_0+W_0}{n}}.$$

V primeru $B_0 = W_0 = c$ pa dobimo kar **enakomerno diskretno** porazdelitev:

$$\mathbb{P}(Y_n = k) = \frac{n!}{(n-k)!k!} \frac{[B_0^k k!][B_0^{n-k}(n-k)!]}{B_0^n (n+1)!} = \frac{1}{n+1}.$$

Ker je $Y_n = \sum_{i=1}^n X_i$ iz $\mathbb{E}[X_i] = \frac{B_0}{B_0+W_0}$ zaradi linearnosti matematičnega upanja sledi

$$E[Y_n] = n\frac{B_0}{B_0+W_0}.$$

Zanimivo je, da upanje Y_n ni odvisno od parametra c. Lahko pa fiksiramo B_0, W_0, n in opazujemo $\lim_{c\to\infty} \mathbb{P}(Y_n = k)$. Za $k = 0$ ter $k = n$ velja

$$\lim_{c\to\infty} \mathbb{P}(Y_n = 0) = \lim_{c\to\infty} \frac{W_0}{B_0+W_0}\frac{W_0+c}{B_0+W_0+c}\cdots\frac{W_0+(n-1)c}{B_0+W_0+(n-1)c} = \frac{W_0}{B_0+W_0},$$
$$\lim_{c\to\infty} \mathbb{P}(Y_n = n) = \lim_{c\to\infty} \frac{B_0}{B_0+W_0}\frac{B_0+c}{B_0+W_0+c}\cdots\frac{B_0+(n-1)c}{B_0+W_0+(n-1)c} = \frac{B_0}{B_0+W_0}.$$

Posledično kot komplement zgornjega dobimo

$\lim_{c\to\infty} \mathbb{P}(Y_n \in \{1, 2\dots, n-1\}) = 0.$

2.3. **Delež črnih kroglic**

Naj bo $c \in \mathbb{N}_0$ (da nam ne zmanjka kroglic). Za $n \in \mathbb{N}$ je delež **izvlečenih** črnih kroglic v n-tem koraku enak

$$M_n := \frac{Y_n}{n}.$$

Delež črnih kroglic **v žari** je tedaj

$$Z_n := \frac{B_0 + cY_n}{B_0+W_0+cn}.$$

Izrek 2.1. *Naj bo $c \in \mathbb{N}_0$. Ko gre $n \to \infty$ konvergira $(M_n)_{n\in\mathbb{N}}$ skoraj gotovo proti slučajni spremenljivki M natanko tedaj, ko $(Z_n)_{n\in\mathbb{N}}$ konvergira skoraj gotovo proti slučajni spremenljivki Z. Tedaj sta njuni limiti enaki.*

Dokaz.

$$Z_n = \frac{B_0}{B_0+W_0+cn} + \frac{cn}{B_0+W_0+cn} M_n.$$

Res, ko gre $n \to \infty$, gre konstantni člen proti 0, koeficient pred M_n pa proti 1. □

Izrek 2.2. *Naj bo $(\mathcal{F}_n)_{n\in\mathbb{N}} := \sigma-(X_i : 1 \le i \le n)$ naravna filtracija procesa Pólyevih žar ($\mathcal{F}_0$ trivialna). Proces $(Z_n)_{n\in\mathbb{N}}$ je martingal glede na $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*

Dokaz. Ker je $(Z_n)_{n\in\mathbb{N}}$ prilagojen in integrabilen proces, sledi

$$E[Z_{n+1}|\mathcal{F}_n] = E\left[\frac{B_0 + c\sum_{k=1}^{n+1} X_k}{B_0+W_0+c(n+1)}\Big|\mathcal{F}_n\right] =$$
$$= \frac{B_0 + c\sum_{k=1}^{n} X_k}{B_0+W_0+cn}\frac{B_0+W_0+cn}{B_0+W_0+c(n+1)} + \frac{c}{B_0+W_0+c(n+1)}E[X_{n+1}|\mathcal{F}_n] =$$
$$= Z_n\frac{B_0+W_0+cn}{B_0+W_0+c(n+1)} + \frac{c}{B_0+W_0+c(n+1)}Z_n = Z_n,$$
saj je $E[X_{n+1}|\mathcal{F}_n] = \frac{B_i}{B_i+W_i} = Z_n.$

□

Naj bo $p \geq 1$. Ker je $0 \leq Z_n \leq 1$ sledi, da je martingal $(Z_n)_{n\in\mathbb{N}}$ enakomerno integrabilen in omejen v L^p. Torej obstaja taka slučajna spremenljivka $Z \in \mathcal{F}_\infty$, da konvergira Z_n proti Z ko $n \to \infty$ skoraj gotovo in v L^p, posledično pa tudi v porazdelitvi (*za dokaz glej vir* [4], *strani* 217-223). Zanima nas porazdelitev Z.

Izrek 2.3. *Naj bo $c \in \mathbb{N}$. Proces $(Z_n)_{n\in\mathbb{N}}$ konvergira proti slučajni spremenljivki $Z \overset{(d)}{=} Beta(\frac{B_0}{c}, \frac{W_0}{c})$.*

Dokaz. Oglejmo si karakteristično funkcijo slučajne spremenljivke Z_n.

$$P(Z_n = z) = P(Y_n = \frac{z(B_0+W_0+cn) - B_0}{c}) = P(Y_n = k),$$

$$E[e^{itZ_n}] = \sum_{k=0}^{n} e^{it\frac{B_0+kc}{B_0+W_0+nc}}P(Y_n = k) = \sum_{k=0}^{n} e^{it\frac{B_0+kc}{B_0+W_0+nc}}\binom{n}{k}\frac{\beta(\frac{B_0}{c}+k, \frac{W_0}{c}+n-k)}{\beta(\frac{B_0}{c}, \frac{W_0}{c})}.$$

Uporabimo dejstvo, da je $\beta(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}dp$ in izračunajmo limito karakteristične funkcije, ko gre $n \to \infty$.

$$\lim_{n\to\infty}\sum_{k=0}^{n} e^{it\frac{B_0+kc}{B_0+W_0+nc}}\binom{n}{k}\int_0^1 p^k(1-p)^{n-k}\frac{p^{\frac{B_0}{c}-1}(1-p)^{\frac{W_0}{c}-1}}{\beta(\frac{B_0}{c}, \frac{W_0}{c})}dp =$$
$$= \lim_{n\to\infty} e^{it\frac{B_0}{B_0+W_0+nc}}\int_0^1 \sum_{k=0}^{n}\binom{n}{k}(p\cdot e^{it\frac{c}{B_0+W_0+nc}})^k(1-p)^{n-k}\frac{p^{\frac{B_0}{c}-1}(1-p)^{\frac{W_0}{c}-1}}{\beta(\frac{B_0}{c}, \frac{W_0}{c})}dp =$$
$$= \lim_{n\to\infty} e^{it\frac{B_0}{B_0+W_0+nc}}\int_0^1 (p\cdot e^{it\frac{c}{B_0+W_0+nc}} + (1-p))^n\frac{p^{\frac{B_0}{c}-1}(1-p)^{\frac{W_0}{c}-1}}{\beta(\frac{B_0}{c}, \frac{W_0}{c})}dp.$$

Pokažemo lahko, da je funkcija $|(p\cdot e^{it\frac{c}{B_0+W_0+nc}} + (1-p))^n|$ nenaraščujoča z n.

$$|p\cdot e^{it\frac{c}{B_0+W_0+nc}} + (1-p)| \leq 1 \iff (p-1)p\sin^4\left(\frac{tc}{2n(B_0+W_0+nc)}\right) \leq 0.$$

Desna zveza očitno velja za vsak $t \in \mathbb{R}$, $p \in [0,1]$ in za poljubne parametre žare. Torej ima funkcija maksimum pri $n = 1$ in lahko uporabimo Lebesgueov izrek o

dominirani konvergenci, da izračunamo naslednjo limito:

$$\lim_{n\to\infty}(p\cdot e^{it\frac{c}{B_0+W_0+nc}}+(1-p))^n=\lim_{n\to\infty}\exp\left(\frac{\log\left(p\cdot e^{it\frac{c}{B_0+W_0+nc}}+(1-p)\right)}{1/n}\right)=$$

$$\overset{\text{L'Hôpital}}{=}\lim_{n\to\infty}\exp\left(\frac{\left(p\cdot e^{it\frac{c}{B_0+W_0+nc}}+(1-p)\right)^{-1}pe^{it\frac{c}{B_0+W_0+nc}}\frac{-itc^2}{(B_0+W_0+nc)^2}}{-1/n^2}\right)=$$

$$=\lim_{n\to\infty}\exp\left(\frac{pe^{it\frac{c}{B_0+W_0+nc}}itc^2n^2}{(pe^{it\frac{c}{B_0+W_0+nc}}+1-p)(B_0+W_0+cn)^2}\right)=e^{itp}.$$

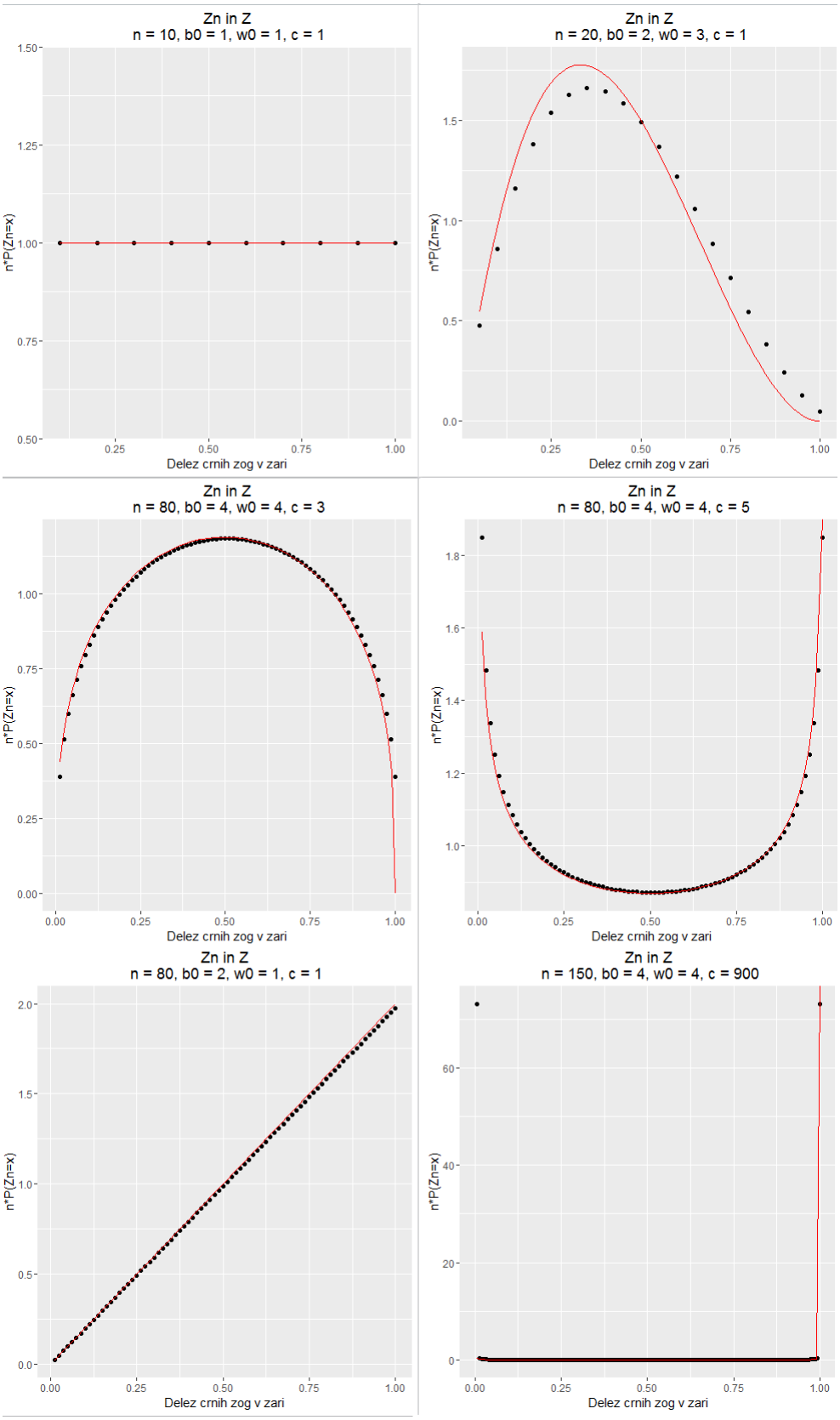
Vrnimo se k limiti karakteristične funkcije:

$$\lim_{n\to\infty}E[e^{itZ_n}]=\lim_{n\to\infty}e^{it\frac{B_0}{B_0+W_0+nc}}\int_0^1(p\cdot e^{it\frac{c}{B_0+W_0+nc}}+(1-p))^n\frac{p^{\frac{B_0}{c}-1}(1-p)^{\frac{W_0}{c}-1}}{\beta(\frac{B_0}{c},\frac{W_0}{c})}dp=$$

$$=\int_0^1 e^{itp}\cdot\frac{p^{\frac{B_0}{c}-1}(1-p)^{\frac{W_0}{c}-1}}{\beta(\frac{B_0}{c},\frac{W_0}{c})}dp,$$

kar je ravno karakteristična funkcija beta porazdelitve. Uporabimo lahko Lévyjev izrek, ki nam pove, da je zakon Z absolutno zvezen ter zaključimo, da je $Z\overset{(d)}{=}Beta(\frac{B_0}{c},\frac{W_0}{c})$.

□



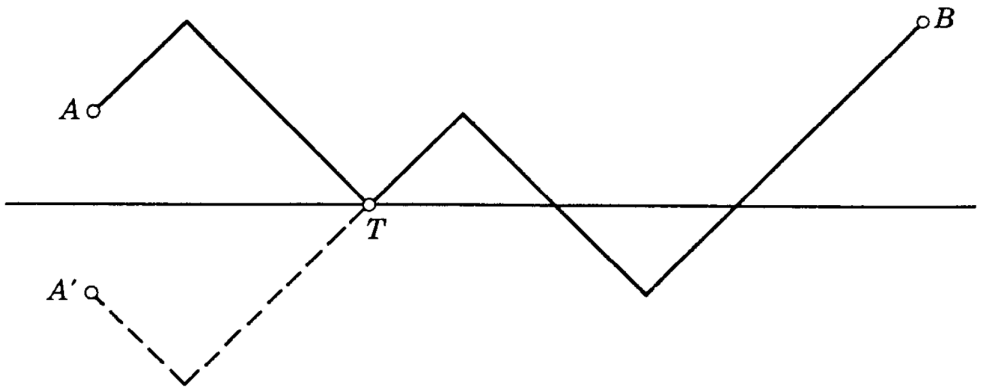
V zgornjih slikah lahko vidimo porazdelitev Z_n pri različnih začetnih parametrih skupaj z gostoto njene limitne porazdelitve. Opazimo lahko, da Z_n zelo hitro konvergira proti Z.

2.4. **Verjetnost prvega izenačenja**

V naslednjem razdelku nas bo zanimala verjetnost, da v procesu Pólyevih žar iz poljubnega neravnovesnega stanja pridemo v ravnovesje, torej da imamo enako število belih in črnih kroglic. Za lažje razumevanje bomo v tem razdelku predpostavili $c \equiv 1$. Zaradi simetrije lahko brez škode za splošnost predpostavimo, da je sprva črnih kroglic več kot belih. Splošneje si lahko ogledamo še verjetnost, da pridemo v stanje, kjer je $S_n := B_n - W_n$ presežek črnih kroglic nad belimi. Takšne množice stanj lahko v drevesu procesa predstavimo z navpičnicami. Analiza prvega izenačenja je uporabna na številnih področjih. Na tak način lahko denimo modeliramo rast bakterijskih kolonij, kjer nas zanima verjetnost, da vrsta, ki je v manjšini, sčasoma prehiti dominantno vrsto.

Lema 2.4 (Princip zrcaljenja)**.** *Naj bo $A = (0, \alpha)$ in $B = (n, \beta)$, $\alpha \geq 0$, $\beta \geq 0$ za $\alpha, \beta \in \mathbb{Z}$, $n \in \mathbb{N}$. Naj bo $A' := (0, -\alpha)$ zrcaljenje A čez x-os ter $S = (s_0, \ldots, s_n)$ diskreten sprehod s skoki velikosti 1.*

Potem je število poti od A do B, ki se bodisi dotaknejo bodisi prečkajo x-os, enako številu vseh poti od A' do B.



Princip zrcaljenja, vir: [3]

Dokaz. Naj bo $(s_0 = \alpha, s_1, \ldots, s_n = \beta)$ pot od A do B, na kateri ena ali več točk leži na x-osi. Naj bo T čas prve take točke, torej: $s_0 > 0, \ldots, s_{T-1} > 0, s_T = 0$. Potem je $(-s_0, -s_1, \ldots, -s_{T-1}, s_T = 0, s_{T+1}, s_{T+2}, \ldots, s_n)$ pot od A' do B. Opazimo, da lahko na tak način za vsako pot od A do B, ki prečka x-os dobimo ustrezno pot od A' do B (in obratno). Sledi, da je teh sprehodov enako mnogo. □

Definicija 2.5. Naj p označuje število skokov navzgor, q pa število skokov navzdol. Potem veljata zvezi $p + q = n$ ter $1 \cdot p + (-1) \cdot q = \beta - \alpha$. Število vseh poti od $(0, \alpha)$ do (n, β) je

$$N(n, \alpha, \beta) := \binom{p+q}{p} = \binom{p+q}{q} = \binom{n}{q}.$$

Izrek 2.6 (Bertrandov izrek o glasovnicah)**.** *Naj bosta n in β naravni števili. Potem obstaja natanko $\frac{\beta}{n} N(n, 0, \beta)$ poti oblike $(s_0 = 0, s_1, s_2, \ldots, s_n = \beta)$, pri čemer je $s_1 > 0, \ldots, s_n > 0$.*

Dokaz. Iz prejšnjih zvez lahko izpeljemo $p = \frac{\beta - \alpha + n}{2}$. Opazimo, da je število ustreznih poti enako, kot če bi začeli v točki $(1, 1)$. Po prejšnji lemi je število takšnih poti

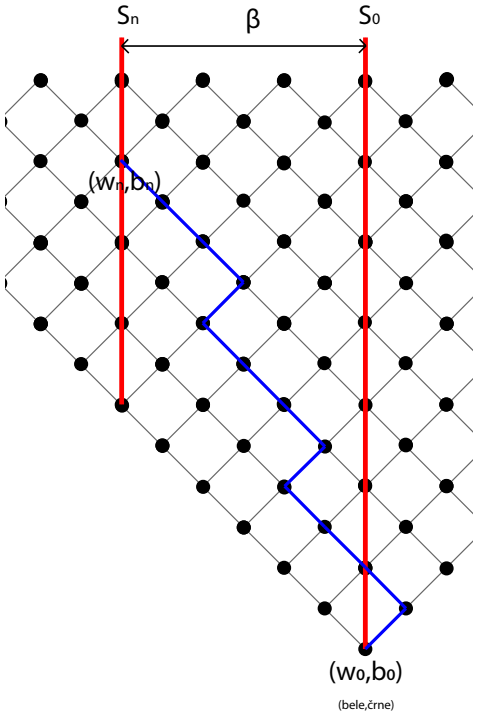
enako

$$N(n-1,1,\beta)-N(n-1,-1,\beta)=\binom{n-1}{\frac{(\beta-1)+(n-1)}{2}}-\binom{n-1}{\frac{(\beta+1)+(n-1)}{2}}=$$
$$=\binom{p+q-1}{p-1}-\binom{p+q-1}{p}=\frac{(p+q-1)!}{(p-1)!q!}-\frac{(p+q-1)!}{p!(q-1)!}=\frac{(p+q)!(p-q)}{(p+q)\cdot p!\cdot q!}=$$
$$=\frac{p-q}{p+q}\binom{p+q}{p}=\frac{\beta}{n}N(n,0,\beta).$$

□

Zgornji izrek in njegovo ime izhajata iz naslednjega praktičnega problema: "Na volitvah, kjer kandidat A dobi p glasov, kandidat B pa q glasov, tako, da je $p>q$, kakšna je verjetnost, da bo kandidat A v prednosti tekom celotnih volitev?" Odgovor je ravno $\frac{p-q}{p+q}$.

Sedaj se lahko vrnemo k Pólyevim žaram. Zanima nas verjetnost, da iz poljubnega stanja (B_0,W_0) pridemo do stanja (B_n,W_n), ne da bi pred tem prečkali navpičnico $S_n=B_n-W_n$, (BŠS $S_0>S_n$). Ker je vsaka izmed poti med dvema stanjema v procesu Pólyevih žar enako verjetna, lahko uporabimo izrek o glasovnicah, da dobimo število poti, ki ustrezajo pogoju. Količina β, ki nastopa v izreku, je v tem primeru odmik med navpičnicama S_0 in S_n.



Za $Y_n=k$ veljajo zveze: $B_n=B_0+k$, $W_n=W_0+n-k$, torej je iskana verjetnost

$$\frac{S_0-S_n}{n}P(Y_n=k)=\frac{n-2k}{n}P(Y_n=k)=\frac{B_0-W_0-(B_n-W_n)}{B_n+W_n-(B_0+W_0)}P(Y_n=B_n-B_0).$$

Bolj specifično nas zanima verjetnost, da pride v stanju (B_n, B_n) do **prvega izenačenja**, torej $S_n = 0$ in $B_1 > W_1, B_2 > W_2, \ldots, B_n = W_n$. Ob upoštevanju zvez: $W_n = B_n = B_0 + k = W_0 + n - k$, lahko izpeljemo iskano verjetnost, ki jo označimo z $G_{B_n}(B_0, W_0)$.

$$\begin{aligned} G_{B_n}(B_0, W_0) &= \frac{B_0 - W_0}{2B_n - B_0 - W_0} P(Y_n = B_n - B_0) = \\ &= \binom{2B_n - B_0 - W_0}{B_n - B_0} \frac{B_0 - W_0}{2B_n - B_0 - W_0} \frac{\beta(B_n, B_n)}{\beta(B_0, W_0)} = \\ &= \frac{B_0 - W_0}{B_0 + W_0} \binom{B_n - 1}{B_0 - 1} \binom{B_n - 1}{W_0 - 1} \binom{2B_n - 1}{B_0 + W_0}^{-1}. \end{aligned}$$

Opomba: izenačenja seveda ne moremo dobiti za poljuben $n \in \mathbb{N}$, vendar pa za vsak $m \in \mathbb{N}$ obstaja tak $n \in \mathbb{N}$, da je $B_n = m$.

Izkaže se, da se za velike B_n zgornja verjetnost asimptotsko obnaša kot

$$G_{B_n}(B_0, W_0) \simeq A(B_0, W_0) B_n^{-2},$$

kjer je

$$A(B_0, W_0) := \frac{(B_0 - W_0)(B_0 + W_0 - 1)!}{(B_0 - 1)!(W_0 - 1)!} 2^{-B_0 - W_0}.$$

2.5. **Izhodna verjetnost**

Izhodna verjetnost $\mathcal{E}_n(B_0, W_0)$ nam pove, kolikšna je verjetnost, da od začetnega stanja (B_0, W_0), do trenutka, ko je v žari $B_0 + W_0 + n$ kroglic, pride do izenačitve. Ta takoj sledi iz verjetnosti prvega izenačenja dobljene v prejšnjem razdelku.

$$\mathcal{E}_n(B_0, W_0) = \sum_{i=0}^{n} G_{B_0+i}(B_0, W_0),$$

kjer spodnja meja označuje dogodek, da žrebamo le bele kroglice, dokler ne pride do izenačenja. Posebej nas zanima *t. i.* popolna izhodna verjetnost, torej verjetnost da kadarkoli v procesu pride do izenačitve.

$$E(B_0, W_0) := \lim_{n \to \infty} \mathcal{E}_n(B_0, W_0).$$

Posebno lep je primer, ko imamo $(B_0, W_0) = (2, 1)$. Tedaj je

$$G_{B_n}(2, 1) = \frac{1}{3} \binom{B_n - 1}{1} \binom{B_n - 1}{0} \binom{2B_n - 1}{3}^{-1} = \frac{1}{(2B_n - 3)(2B_n - 1)},$$

$$E(2, 1) = \frac{1}{1 \cdot 3} + \frac{1}{3 \cdot 5} + \frac{1}{5 \cdot 7} + \cdots = \frac{1}{2}.$$

Zgornji zgled pokaže, da v procesu Pólyevih žar verjetnost izenačenja v splošnem ni enaka 1, kot bi veljalo pri enostavnem simetričnem slučajnem sprehodu. Ostane nam, da izpeljemo porazdelitev $E(B_0, W_0)$ za poljubno začetno stanje v žari. To bi lahko dobili preko limit zgornjih verjetnosti, vendar obstaja bolj eleganten način izpeljave s pomočjo limitne porazdelitve Z, izračunane v razdelku 2.3.

Izrek 2.7. *Naj bo začetno stanje v Pólyevi žari (B_0, W_0). Potem je $E(B_0, W_0) = 2F_Z(\frac{1}{2})$, kjer je $Z \overset{(d)}{=} Beta(B_0, W_0)$ in F njena porazdelitvena funkcija.*

Dokaz. Naj bo $(S_n)_{n\geq 0} := B_n - W_n = 2Y_n - n + B_0 - W_0$ presežek črnih kroglic nad belimi in $\mu_n := \frac{S_n}{n+W_0+B_0} = 2Z_n - 1$ delež tega presežka med vsemi kroglicami. Ker ima Z_n limito skoraj gotovo, jo ima tudi μ_n, torej $\mu := \lim_{n\to\infty} \mu_n = 2Z - 1$. Naj bo $T := \inf\{k \mid S_k = 0\}$. T je očitno čas ustavljanja glede na naravno filtracijo procesa Pólyevih žar $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Zanima nas verjetnost dogodka $\{T < \infty\}$.

$$P(\{T < \infty\}) = P(\{T < \infty\} \cap \{\mu > 0\}) + P(\{T < \infty\} \cap \{\mu < 0\}),$$

pri čemer je verjetnost $\mu = 0$ enaka 0, saj je gostota μ zvezna. Izkaže se, da sta zgornja dogodka enako verjetna. Res, ker je $T < \infty$, je del procesa pred prvim izenačenjem končen in nima vpliva na delež μ. V času T je v žari enako mnogo črnih in belih kroglic, kar pomeni, da je enako verjetno, da bo pot končala levo ali desno od navpičnice $S_n = 0$. Hkrati iz $\{\mu < 0\}$ sledi $\{T < \infty\}$, saj smo, ob predpostavki $B_0 > W_0$, morali na neki točki prečkati $S_n = 0$, da smo končali na drugi strani navpičnice. Sledi

$$P(T < \infty) = 2P(\{T < \infty\} \cap \{\mu < 0\}) = 2P(\mu < 0) = 2P(2Z - 1 < 0) = 2P(Z < \tfrac{1}{2}).$$

□

Če kot zgled ponovno uporabimo $B_0 = 2, W_0 = 1$, lahko vidimo, da za $Z \overset{(d)}{=} Beta(B_0, W_0)$ res velja $2P(Z < \frac{1}{2}) = 2 \cdot \frac{1}{4} = \frac{1}{2}$, kot smo izračunali prej.

3. Splošne Pólyeve žare

Dosedanji model posplošimo na dva načina. Prvič sprostimo pravilo za vračanje kroglic. Ko izvlečemo (in nato vrnemo) belo kroglico, sedaj v žaro dodamo a_{11} belih in a_{12} črnih kroglic. Podobno ob izvlečeni črni kroglici dodamo a_{21} belih in a_{22} črnih kroglic. To lahko zapišemo v matrični obliki kot

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Hkrati pa lahko namesto dveh uporabimo k-mnogo barv za kroglice. Po izvlečeni kroglici i-te barve v žaro dodamo a_{ij} kroglic j-te barve, $j = 1, 2, \ldots, k$. Tako dobimo matriko

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{bmatrix}.$$

Tako matriko imenujemo **matrika zamenjav**. Pri tem so vrstice urejene po barvah izvlečenih kroglic, stolpci pa po barvah dodanih kroglic. Tukaj je $a_{ij} \in \mathbb{Z}$ tipično deterministična. Lahko pa je tudi diskretna slučajna spremenljivka, katere vrednost se generira ob vsakem žrebu. V splošni terminologiji se izraz Pólyeve žare pogosto uporablja za žare z poljubno matriko zamenjav. Klasični model, ki smo ga obravnavali v prejšnjih poglavjih, pa je znan kot **Pólya-Eggenbergerjeva žara** in ima naslednjo matriko zamenjav:

$$\begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix}.$$

Pomemben koncept pri proučevanju limitnih lastnosti žar je **neskončna izvedljivost** modela. Za žaro rečemo da je neskončno izvedljiva, če lahko žrebamo in zamenjujemo kroglice v nedogled na vsaki stohastični poti procesa. Z drugimi besedami, na vsakem koraku je možno upoštevati pravilo matrike zamenjav, brez da bi se nam pri tem "zataknilo". Neskončna izvedljivost žare je v splošnem odvisna tako od matrike zamenjav kot od začetnega stanja kroglic v žari. Če so vsi elementi matrike nenegativni, kot v primeru

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

potem je žara neskončno izvedljiva za poljubno (neprazno) začetno stanje v žari. Prav nasprotno, žara s shemo

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

ni neskončno izvedljiva za nobeno začetno stanje, saj se žara izprazni po $W_0 + B_0$ žrebih. V resnici bi lahko razvrstili vse dvobarvne sheme po številu in razporeditvi negativnih elementov v matriki, nato pa določili potrebne in zadostne pogoje za neskončno izvedljivost v vsakem izmed primerov. Naj $\oplus$ predstavlja nenegativen element na danem mestu v matriki, $-$ pa negativen element. Tedaj naslednje sheme niso neskončno izvedljive pod nobenimi začetnimi pogoji:

$$\begin{bmatrix} - & - \\ - & - \end{bmatrix}, \begin{bmatrix} - & - \\ \oplus & - \end{bmatrix}, \begin{bmatrix} \oplus & - \\ - & - \end{bmatrix}, \begin{bmatrix} - & \oplus \\ - & - \end{bmatrix}, \begin{bmatrix} - & - \\ - & \oplus \end{bmatrix}, \begin{bmatrix} \oplus & - \\ \oplus & - \end{bmatrix}, \begin{bmatrix} - & \oplus \\ - & \oplus \end{bmatrix}, \begin{bmatrix} \oplus & - \\ - & \oplus \end{bmatrix}.$$

Pri preostalih 2x2 shemah je neskončna izvedljivost odvisna od začetnega stanja v žari ter od samih elementov matrike. Vzemimo kot primer matriko s shemo

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad A \sim \begin{bmatrix} - & - \\ \oplus & \oplus \end{bmatrix},$$

kjer zgornja notacija označuje $a < 0$, $b < 0$, $c \geq 0$ in $d \geq 0$. Možno je pokazati, da je takšna žara neskončno izvedljiva pod naslednjimi pogoji:

(1) W_0 in c sta oba večkratnika $|a|$,
(2) $\det(A) \leq 0$,
(3) $\det\begin{pmatrix} a & b \\ W_0 & B_0 \end{pmatrix} < 0$.

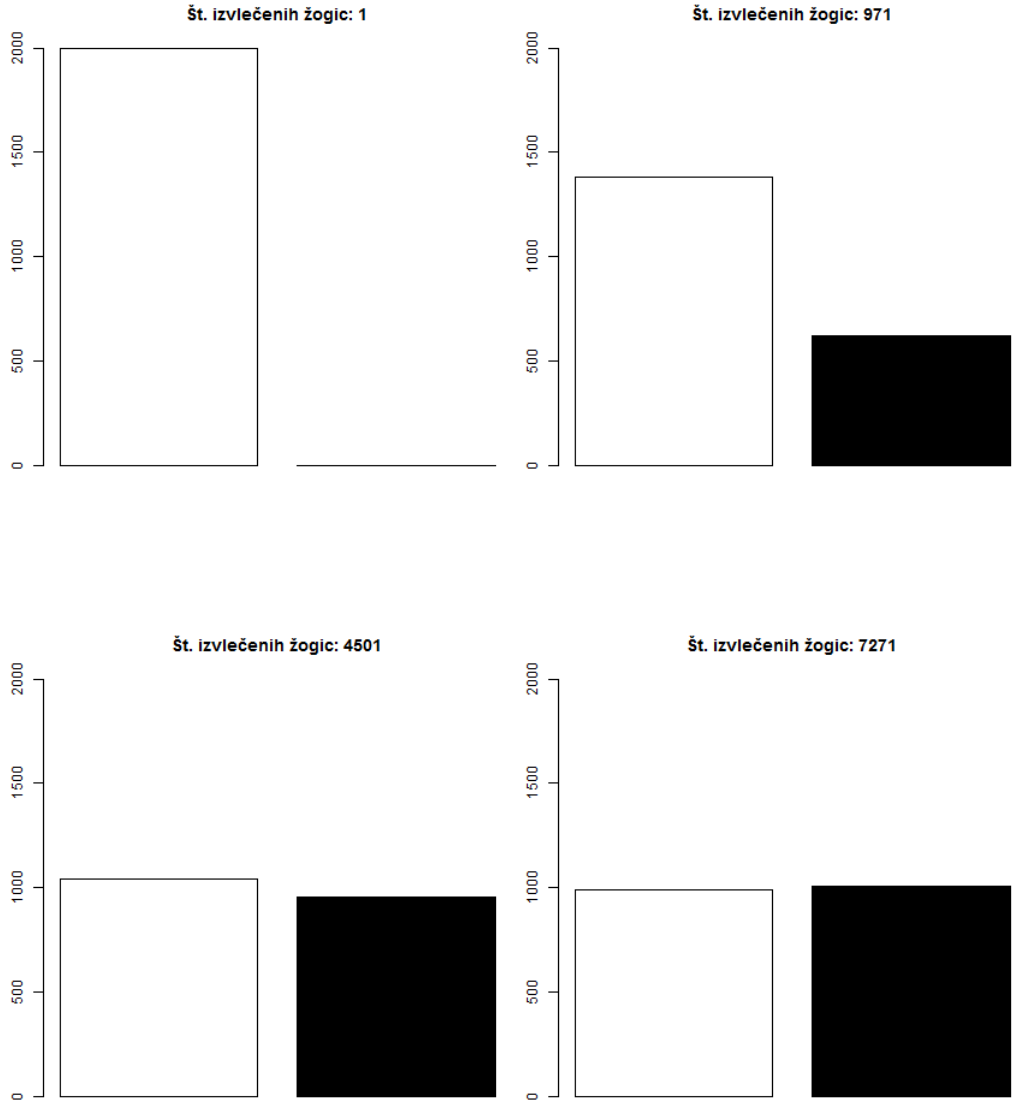
4. Praktične aplikacije Pólyevih žar

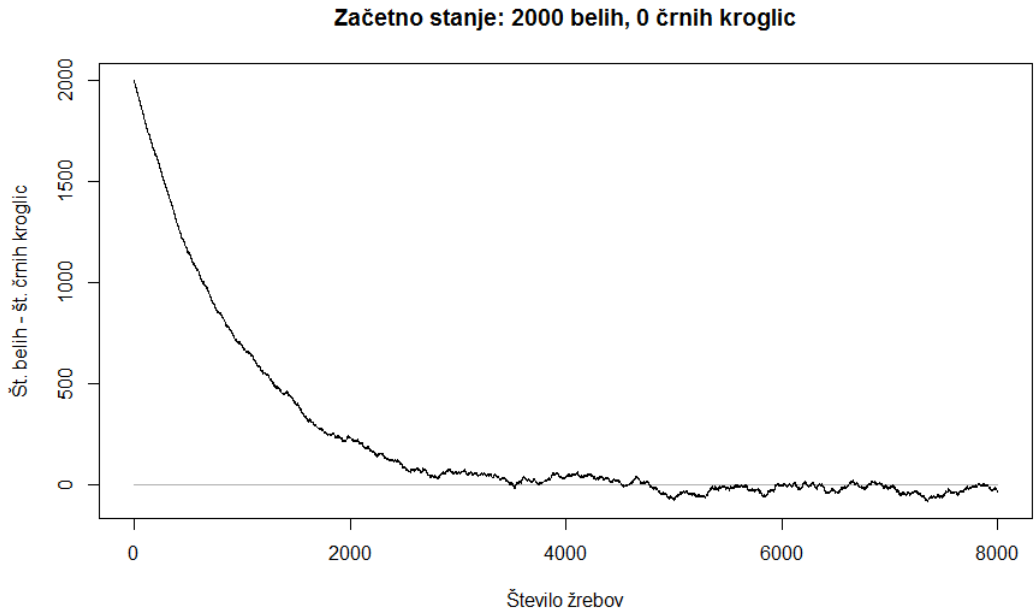
4.1. **Ehrenfestov model difuzije plinov.**

Ehrenfestova žara je fizikalni model, ki sta ga leta 1907 uvedla Paul in Tatiana Ehrenfest. Znana tudi kot model pasjih bolh, Ehrenfestova žara modelira difuzijo plinov med dvema sobama. V vsaki izmed sob je sprva neka začetna količina delcev plina. Nato na vsakem koraku iz celotne populacije delcev obeh sob naključno izberemo en delec in ga prestavimo v drugo sobo. Če sobi interpretiramo kot žari in delce plina kot kroglice, lahko zgornji model zapišemo z naslednjo matriko zamenjav:

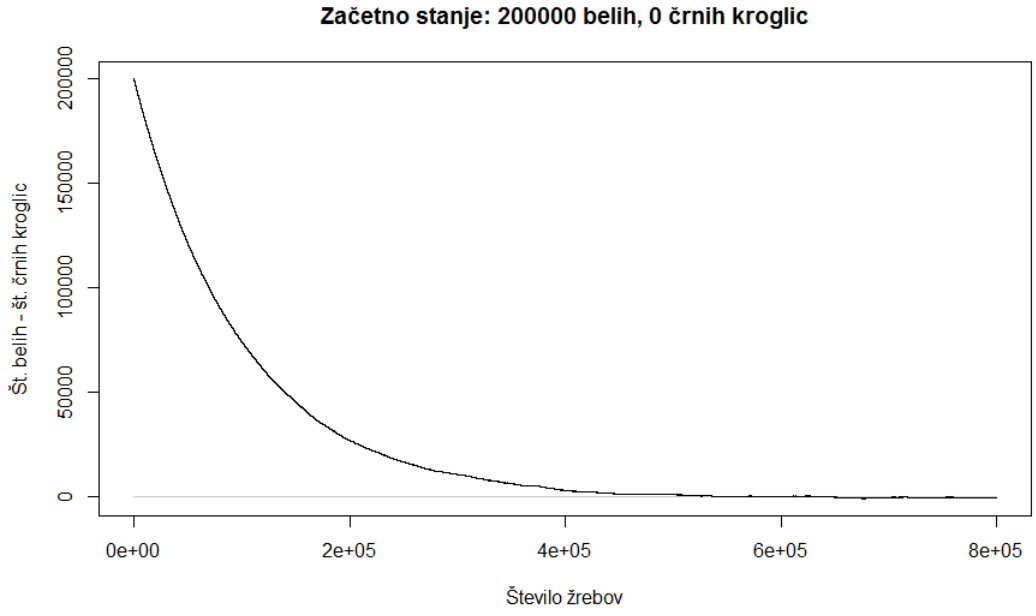
$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

V tem modelu se količina vseh kroglic v žari ne spreminja skozi čas. Spreminja se zgolj delež barv kroglic. V programu R lahko simuliramo žaro s 2000 belimi in 0 črnimi kroglicami in ilustriramo analogijo s prehodi plinov med sobama:





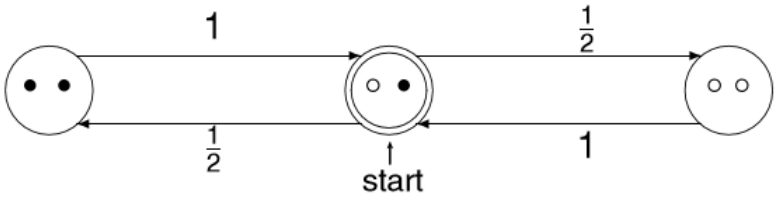
Na začetku simulacije delež belih kroglic upada hitro, a čedalje počasneje, dokler se razmerje barv ne približa vrednosti $\frac{1}{2}$, okoli katere nato naključno oscilira. Relativna velikost oscilacij se precej zmanjša, če v simulaciji drastično povečamo število kroglic, resda pa je pri tem potrebnih veliko več žrebov. Če vzamemo v obzir dejstvo, da je v kubičnem metru zraka okoli $2,53 \times 10^{25}$ molekul, se zdi model vsaj intuitivno primeren.



Definicija 4.1.
Markovska veriga je slučajni proces, ki zadošča markovski lasnosti. Naj bo na markovski verigi p_{ij} verjetnost prehoda iz i-tega v j-to stanje procesa. Zaradi markovske lastnosti je ta verjetnost odvisna zgolj od trenutnega stanja i in od izbora prihodnjega stanja j. Če so prehodne verjetnosti procesa neodvisne od časa n, rečemo, da je markovska veriga **homogena**. Na markovski verigi s K-mnogo stanji vrednosti p_{ij} za $i = 1, \ldots, K$ ter $j = 1, \ldots, K$ definirajo **prehodno matriko**

$$M = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix}.$$

Pólyeve žare, vključno z Ehrenfestovo, so homogene markovske verige. V naslednjem zgledu sta v žari bela in črna kroglica:



Tri stanja Ehrenfestove žare z dvema kroglicama. Vir: [5]

V primeru dveh kroglic lahko verjetnosti prehodov med stanji zapišemo s prehodno matriko

$$M = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}.$$

Naj bo W_n število belih kroglic v žari po n žrebih. Potem velja

$$W_n = \begin{cases} 1, & \text{če je } n \text{ sod;} \\ 2\mathrm{Ber}\left(\frac{1}{2}\right), & \text{če je } n \text{ lih.} \end{cases}$$

Definicija 4.2.
Stacionarna porazdelitev je verjetnostna porazdelitev $\boldsymbol{\pi}$ na stanjih markovske verige, za katero velja, da če začnemo v nekem naključnem stanju v skladu s $\boldsymbol{\pi}$, so prehodi na ostala stanja ponovno porazdeljeni s $\boldsymbol{\pi}$. Z drugimi besedami, če imamo markovsko verigo s K stanji in prehodno matriko M, mora $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ zadoščati naslednjim pogojem:

(1) $$\pi_i \geq 0 \quad \forall i = 1, 2, \ldots K,$$

(2) $$\sum_{i=1}^{K} \pi_i = 1,$$

(3) $$\boldsymbol{\pi} M = \boldsymbol{\pi}.$$

Ker je za poljubno začetno število kroglic (≥ 1) stanje v procesu odvisno od tega, ali je n lih ali sod, ne moremo imeti konvergence v porazdelitvi. Kljub temu pa v

procesu Ehrenfestove žare vedno obstaja stacionarna porazdelitev. V primeru žare z dvema kroglicama dobimo

$$\boldsymbol{\pi} M = \begin{bmatrix}\pi_1 & \pi_2 & \pi_3\end{bmatrix} \begin{bmatrix}0 & 1 & 0\\ \frac{1}{2} & 0 & \frac{1}{2}\\ 0 & 1 & 0\end{bmatrix} = \begin{bmatrix}\frac{1}{2}\pi_2 & \pi_1+\pi_3 & \frac{1}{2}\pi_2\end{bmatrix}.$$

Ko zgornje enačimo z $\begin{bmatrix}\pi_1 & \pi_2 & \pi_3\end{bmatrix}$, opazimo, da je stacionarna porazdelitev binomska: $\boldsymbol{\pi} = \begin{bmatrix}\frac{1}{4} & \frac{1}{2} & \frac{1}{4}\end{bmatrix}$. Slednje velja tudi za žaro s poljubno mnogo kroglicami.

Izrek 4.3. *Naj bo na začetku v Ehrenfestovi žari M kroglic, od tega $W_0 = \mathrm{Bin}(M, \frac{1}{2})$ belih in $B_0 = M - W_0$ črnih. Naj bo W_n število belih kroglic v žari po n žrebih. Potem velja*

$$W_n \xrightarrow{(\mathrm{d})} \mathrm{Bin}\Big(M, \frac{1}{2}\Big).$$

Dokaz. Binomsko stacionarno porazdelitev bi lahko dobili z nekoliko daljšo izpeljavo preko prehodne matrike markovske verige procesa. Tokrat bomo uporabili alternativen pristop. Po n žrebih je v žari W_n belih in B_n črnih kroglic. Naj bo $W_n = k$. To pomeni, da smo v prejšnjem koraku bili v enem izmed dveh stanj: ali smo imeli $k+1$ belih kroglic ter smo izvlekli belo, ali pa je bilo v žari $k-1$ belih in smo izvlekli črno kroglico.

$$P(W_n = k) = \frac{k+1}{M} P(W_{n-1} = k+1) + \frac{M-k+1}{M} P(W_{n-1} = k-1.$$

Obstoj stacionarne porazdelitve je zagotovljen, saj je markovska veriga Ehrenfestovega procesa nerazcepna (vedno je možno preiti iz vsakega stanja v poljubno drugo stanje) in ima končno množico stanj. Naj bo $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_M)$ iskana stacionarna porazdelitev. Če je začetno stanje v žari porazdeljeno s $\boldsymbol{\pi}$, mora veljati:

$$\pi_k = \frac{k+1}{M}\pi_{k+1} + \frac{M-k+1}{M}\pi_{k-1},$$
$$M(\pi_k - \pi_{k-1}) = (k+1)\pi_{k+1} - (k-1)\pi_{k-1}.$$

Zgornje lahko zapišemo kot sistem enačb za $k, k-1, \ldots, 0$ (z dodano predpostavko, da je $\pi_{-1} = 0$).

$$\begin{aligned}
M(\pi_k - \pi_{k-1}) &= (k+1)\pi_{k+1} - (k-1)\pi_{k-1}\\
M(\pi_{k-1} - \pi_{k-2}) &= k\pi_k - (k-2)\pi_{k-2}\\
&\vdots\\
M(\pi_0 - \pi_{-1}) &= \pi_1 - (-1)\pi_{-1}
\end{aligned}$$

Enačbe seštejemo in dobimo

$$\pi_{k+1} = \frac{M-k}{k+1}\pi_k.$$

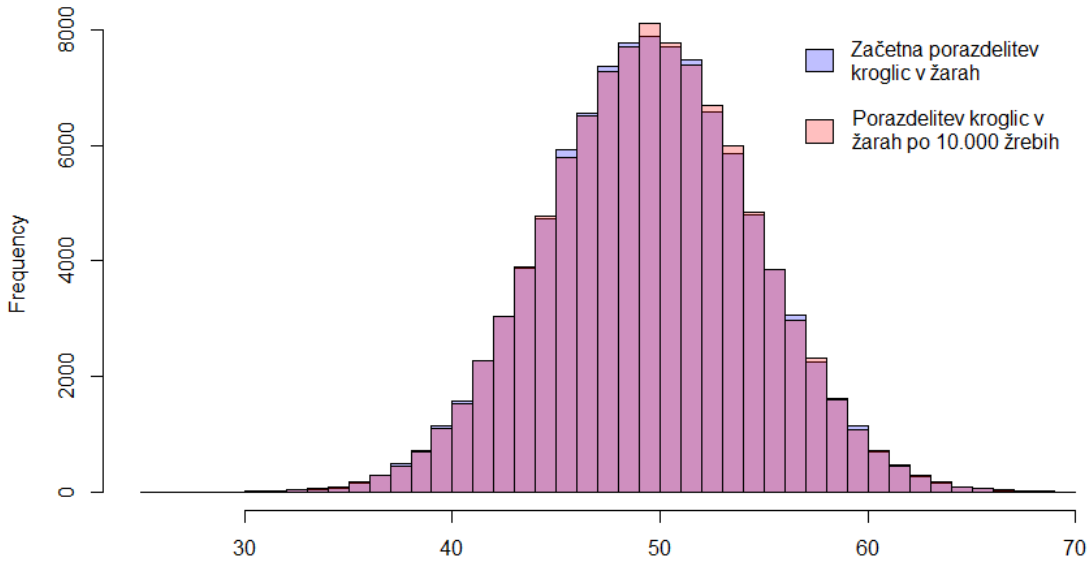
Preko iteracije lahko izračunamo

$$\begin{aligned}
P(W_n = k) = \pi_k &= \\
&= \frac{(M-(k-1))(M-(k-2))\cdots M}{k(k-1)\cdots 1}\pi_0 =\\
&= \binom{M}{k} P(W_0 = 0).
\end{aligned}$$

Če je $W_0 \overset{(d)}{=} \text{Bin}(M, \frac{1}{2})$, je $P(W_0 = 0) = (\frac{1}{2})^M$. Sledi

$$P(W_n = k) = \pi_k = \binom{M}{k}\left(\frac{1}{2}\right)^M,$$

torej je porazdelitev stacionarna. □



Simulacija v R: 100.000 Ehrenfestovih žar z naključno $\text{Bin}(100, \frac{1}{2})$ generiranim začetnim stanjem. Porazdelitev kroglic na začetku in po 10.000 žrebih.

4.2. **Model epidemije**

Zgodovinsko gledano je bil model Pólyevih žar izvorno uporabljen kot model za širjenje nalezljivih bolezni. V zgodnjem dvajsetem stoletju so številni matematiki začeli z uvajanjem matematično-statističnih modelov v epidemiologijo in s tem prispevali k njenemu razcvetu. Posebej smrtonosna je bila pandemija španske gripe med leti 1918 - 1919, ki je pobila med 50 do 100 milijonov ljudi po vsem svetu. Štiri leta kasneje sta Pólya in Eggemberger izdala svoj članek, v katerem sta postavila model epidemije s klasično matriko zamenjave

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Tak model predstavlja družbo, v katero se vključuje nova oseba. V primeru, da ima ta oseba prvi stik z okuženo osebo, postane tudi sama okužena.Nasprotno, če najprej naleti na imuno osebo, postane tudi sama imuna (dobi cepivo, ločitev preko karantene, ipd.). Iz takega modela je tipično, kot smo pokazali v prvih poglavjih, da bodo pretežno okužene skupnosti postajale čedalje bolj bolne. Nasprotno, skupnosti z večjim začetnim deležem zdravih bodo najverjetneje take tudi ostale. Z uvedbo večih barv kroglic lahko s podobno shemo modeliramo družbo, v kateri se širi več nalezljivih bolezni hkrati.

4.3. **Klinične raziskave**

Zamislimo si, da opravljamo klinično raziskavo, pri kateri imamo na voljo K med seboj izključujočih metod zdravljenja. Glavni cilj raziskave je zbrati čim več podatkov o proučevanih K metodah, da bi s tem koristili prihodnjim pacientom. Po drugi strani imamo željo čim bolje zdraviti trenutne paciente. Gre torej za moralni problem raziskav na ljudeh, saj sta naša cilja lahko nekoliko nasprotujoča. V ta namen so Robbins (1952), Zelen (1969) in kasneje Wei (1979) predlagali žare kot model za izbor alternativnih metod zdravljenja.

Denimo, da imamo na voljo dve metodi zdravljenja, a in b. Predpostavimo, da je rezultat zdravljenja znan takoj po izvedbi. Ko uporabimo metodo a, je zdravljenje uspešno z nam neznano verjetnostjo $p_a \in (0,1)$. Podobno, zdravljenje z metodo b uspe z verjetnostjo $p_b \in (0,1)$. Naj bo na začetku v žari $W_0 = B_0$ kroglic. Ko zdravnik za pacienta izbira metodo zdravljenja, iz žare naključno izvleče kroglico in jo vrne. Če je izvlekel belo kroglico, uporabi metodo a in opazuje rezultat zdravljenja. V primeru, da je zdravljenje uspešno, doda belo kroglico, kar poveča verjetnost izbire metode a v prihodnosti. Nasprotno, če je zdravljenje neuspešno, doda črno kroglico in s tem poveča verjetnost metode b. Naj bosta $X_a \overset{(\mathrm{d})}{=} \mathrm{Ber}(p_a)$ in $X_b \overset{(\mathrm{d})}{=} \mathrm{Ber}(p_b)$ neodvisni slučajni spremenljivki. Tak model žare je podan z matriko zamenjave

$$A = \begin{bmatrix} X_a & 1 - X_a \\ 1 - X_b & X_b \end{bmatrix}.$$

Posebnost te sheme je, da se "uči" iz dosedanjih informacij o poteku raziskave. Primer sheme brez te lastnosti bi bila navadna *naključna klinična raziskava*, pri kateri zdravnik vsakič, ko izbira metodo zdravljenja, vrže kovanec in z verjetnostjo $\frac{1}{2}$ izbere eno izmed metod. Katera izmed shem je torej boljša? Ali se v primeru izbora z žaro pretekle izkušnje zares prenesejo na uspešnejše zdravljenje? Začnimo z naključno raziskavo. Naj bo Y indikatorska slučajna spremenljivka, ki označuje uspeh pri posameznem pacientu.

$$Y = \begin{cases} X_a, & \text{z verjetnostjo } \frac{1}{2}; \\ X_b, & \text{z verjetnostjo } \frac{1}{2}. \end{cases}$$

$$E[Y] = \frac{1}{2}E[X_a] + \frac{1}{2}E[X_b] = \frac{p_a + p_b}{2}.$$

Naj bodo $Y_i \overset{(\mathrm{d})}{=} Y$ neodvisne slučajne spremenljivke za $i = 1, 2, \ldots, n$, kjer je n število opravljenih zdravljenj. Tedaj lahko število vseh uspehov zapišemo kot

$$S_n = Y_1 + Y_2 + \ldots + Y_n.$$

Po krepkem zakonu velikih števil dobimo delež uspešnih zdravljenj

$$\frac{S_n}{n} \overset{\text{s.g.}}{=} E[Y] = \frac{p_a + p_b}{2}.$$

Vidimo, da je uspešnost zdravljenj pri naključni klinični raziskavi kar povprečje verjetnosti uspehov metod a in b. Za primerjavo uspešnosti z žrebom s pomočjo žar potrebujemo dodaten izrek.

Definicija 4.4.
Za stohastično $k \times k$ matriko zamenjav A rečemo, da je **razširjena**, če ima naslednje lastnosti:

(1) Shema je neskončno izvedljiva.
(2) Vsi elementi matrike imajo končno varianco.
(3) $\sum_{j=1}^{k} E[A]_{ij} = c \in \mathbb{R}, \forall i = 1, 2, \ldots, n$ ($E[A]$ ima konstantno vrstično vsoto).
(4) Lastna vrednost $\lambda_1 \in \mathbb{R}$ matrike $E[A]$, za katero velja $\mathrm{Re}(\lambda_1) \geq \mathrm{Re}(\lambda_2) \geq \ldots \geq \mathrm{Re}(\lambda_k)$ (imenujmo jo vodilna lastna vrednost), je pozitivna in po vrednosti enaka vrstični vsoti matrike $E[A]$.
(5) Vse komponente lastnega vektorja, ki pripada λ_1, so pozitivne.

Izrek 4.5 (Smythe, 1996)**.** *Naj bo po n žrebih W_n število belih in B_n število črnih kroglic v dvobarvni razširjeni žari z matriko zamenjav A. Naj bo λ_1 vodilna lastna vrednost matrike $E[A]$ in (v_1, v_2) pripadajoči levi lastni vektor. Potem velja*

$$\frac{W_n}{n} \xrightarrow{p} \frac{\lambda_1 v_1}{v_1 + v_2},$$
$$\frac{B_n}{n} \xrightarrow{p} \frac{\lambda_1 v_2}{v_1 + v_2}.$$

Za dokaz Smythovega izreka je potrebno dokazati vrsto izrekov iz področja stohastičnih modelov žar. Dokazi se nahajajo v viru [5] na straneh 101-114.

Vrnimo se k shemi žare pri kliničnih raziskavah. Pričakovana vrednost matrike zamenjav je

$$E[A] = \begin{bmatrix} p_a & 1 - p_a \\ 1 - p_b & p_b \end{bmatrix}.$$

Njene lastne vrednosti so $\lambda_1 = 1$ in $\lambda_2 = p_a + p_b - 1$. Vodilna lastna vrednost matrike $E[A^T]$ je 1 s pripadajočim lastnim vektorjem

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \frac{1}{2 - p_a - p_b} \begin{bmatrix} 1 - p_b \\ 1 - p_a \end{bmatrix}.$$

Na tem koraku je očitno, da imamo opravka z razširjeno matriko zamenjav. Vsakič, ko zdravimo pacienta, dodamo kroglico v žaro. Po $n - 1$ korakih je v žari $n + W_0 + B_0$ kroglic. Podobno kot pri naključnih raziskavah, lahko uvedemo indikatorsko slučajna spremenljivko Y_n, ki označuje uspeh pri n-tem pacientu

$$Y_n = \begin{cases} X_a^{(n)}, & \text{z verjetnostjo } \dfrac{W_{n-1}}{n + W_0 + B_0 - 1}; \\ X_b^{(n)}, & \text{z verjetnostjo } \dfrac{W_{n-1}}{n + W_0 + B_0 - 1}, \end{cases}$$

pri čemer so $X_a^{(n)}$ ter $X_b^{(n)}$ neodvisne enako porazdeljene kopije X_a in X_b. Spet lahko zapišemo število uspehov po n zdravljenjih kot

$$S_n = Y_1 + \ldots + Y_n.$$

Žal so v tem primeru slučajne spremenljivke Y_i med seboj odvisne. Naj $\overline{W_n}$ in $\overline{B_n}$ predstavljata števili *izvlečenih* belih/črnih kroglic po n korakih. Tako lahko S_n zapišemo kot

$$S_n = (X_a^{(1)} + \ldots + X_a^{(\overline{W_n})}) + (X_b^{(1)} + \ldots + X_b^{(\overline{B_n})}).$$

Ker gre $\overline{W_n} \overset{\text{s.g.}}{\to} \infty$ in so $X_a^{(i)}, i = 1, \ldots, \overline{W_n}$ neodvisne, lahko uporabimo krepki zakon velikih števil.

$$\sum_{i=1}^{\overline{W_n}} X_a^{(i)} \overset{\text{s.g.}}{\to} p_a.$$

Podobno velja za $\overline{B_n} = n - \overline{W_n}$. Po Smythovem izreku dobimo

$$\frac{W_n}{n} \overset{\text{p}}{\to} \frac{\lambda_1 v_1}{v_1 + v_2} = v_1 = \frac{1 - p_b}{2 - p_a - p_b}.$$

Izkaže se, da Smythov izrek velja tudi za delež izvelečnih kroglic.

$$\begin{aligned}
W_n &= W_0 + \sum_{i=1}^{\overline{W_n}} X_a^{(i)} + \sum_{i=1}^{\overline{B_n}} (1 - X_b^{(i)}), \\
\frac{W_n}{n} &= \frac{W_0}{n} + \frac{\overline{W_n}}{n} \times \frac{\sum_{i=1}^{\overline{W_n}} X_a^{(i)}}{\overline{W_n}} + \frac{n - \overline{W_n}}{n} \times \frac{\sum_{i=1}^{\overline{B_n}} (1 - X_b^{(i)})}{\overline{B_n}}, \\
\frac{\overline{W_n}}{n} &= \Big(\frac{W_n}{n} - \frac{W_0}{n} - \frac{\sum_{i=1}^{\overline{B_n}} (1 - X_b^{(i)})}{\overline{B_n}}\Big)\Big(\frac{\sum_{i=1}^{\overline{W_n}} X_a^{(i)}}{\overline{W_n}} - \frac{\sum_{i=1}^{\overline{B_n}} (1 - X_b^{(i)})}{\overline{B_n}}\Big)^{-1}, \\
\frac{\overline{W_n}}{n} &\overset{\text{p}}{\to} \frac{v_1 - (1 - p_b)}{p_a - (1 - p_b)} = \frac{1 - p_b}{2 - p_a - p_b}.
\end{aligned}$$

□

Naposled lahko izračunamo uspešnost zdravljenj pri žrebu z žarami

$$\begin{aligned}
\frac{S_n}{n} &= \frac{X_a^{(1)} + \ldots + X_a^{(\overline{W_n})}}{n} + \frac{X_b^{(1)} + \ldots + X_b^{(\overline{B_n})}}{n} = \\
&= \frac{X_a^{(1)} + \ldots + X_a^{(\overline{W_n})}}{\overline{W_n}} \times \frac{\overline{W_n}}{n} + \frac{X_b^{(1)} + \ldots + X_b^{(\overline{B_n})}}{\overline{B_n}} \times \frac{\overline{B_n}}{n} \overset{\text{p}}{\to} \\
&\overset{\text{p}}{\to} \frac{p_a(1 - p_b) + p_b(1 - p_a)}{2 - p_a - p_b}.
\end{aligned}$$

Definirajmo sedaj funkcijo g kot razliko uspešnosti obeh načinov za izbiro zdravljenja.

$$g(p_a, p_b) = \frac{1 - p_b}{2 - p_a - p_b} - \frac{p_a + p_b}{2} = \frac{(p_a - p_b)^2}{2(2 - p_a - p_b)}.$$

Opazimo, da je funkcija g nenegativna na celotnem intervalu $p_a, p_b \in (0, 1)$. Izbira s pomočjo žar je torej boljša kot naključna izbira zdravljenja, razen v primeru, ko je $p_a = p_b$. V tem primeru sta oba načina izbire enako dobra. Naj bo kot primer $p_a = p_b = 0{,}6$. Potem je v obeh primerih uspešnost zdravljenj enaka 60%. Če

povečamo p_a na 0,9, delež uspešnih zdravljenj, izbranih z naključnimi žrebi, naraste na 75%, uspešnost teh izbranih z žaro pa na $0{,}75 + g(0,9; 0,6) = 84\%$.

Takšna shema žare za klinične raziskave je v angleščini dobila ime "Play-the-Winner" shema. V zadnjih dvajsetih letih so se pojavili številni predlogi izboljšave modela in alternativne sheme. Zanimiv primer je tako imenovana "Drop-the-Loser" shema (*Ivanova*, 2003) z naslednjo matriko zamenjave:

$$A = \begin{bmatrix} X_a - 1 & 0 & 0 \\ 0 & X_b - 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

V tej shemi imamo tri barve kroglic. Bela in črna predstavljata metodi zdravljenja a in b, kot v prejšnjem primeru. Ravno tako so enako definirane spremenljivke p_a, p_b, X_a in X_b. Novost je tako imenovana *imigracijska kroglica*, denimo rdeče barve. Na začetku je v žari $W_0 = B_0$ belih/črnih ter R_0 rdečih kroglic. Če iz žare žrebamo belo kroglico, izberemo za zdravljenje metodo a in opazujemo rezultat. V primeru, da zdravljenje uspe, kroglico vnemo v žaro, sicer jo odstranimo. Podobno s črno kroglico in metodo zdravljenja b. V primeru, da izvlečemo rdečo kroglico, v žaro dodamo po eno belo in črno kroglico, s čimer preprečimo "izumrtje" določenih barv.

Slovar strokovnih izrazov

Pólya urn Pólyeva žara

Reflection principle princip zrcaljenja

Total exit probability popolna izhodna verjetnost - verjetnost, da kadarkoli v procesu Pólyevih žar pride do izenačitve v številu kroglic

Replacement matrix matrika zamenjav - shema v skladu s katero se ob vsakem žrebu v žaro doda nove kroglice

Tenable urn neskončno izvedljiva žara - tip žare pri kateri lahko žrebamo in zamenjujemo kroglice v nedogled na vsaki stohastični poti procesa

Irreducible Markov chain nerazcepna markovska veriga - markovska veriga v kateri je iz vsakega stanja možno preiti v poljubno drugo stanje

Transition matrix / Stochastic matrix prehodna matrika - matrika verjetnosti prehodov med stanji markovske verige

Stationary distribution stacionarna porazdelitev

Principal eigenvalue vodilna lastna vrednost - lastna vrednost matrike, ki ima največjo realno komponento

Extended urn scheme razširjena shema žare

Literatura

[1] Tibor Antal, E. Ben-Naim in P. L. Krapivsky, *First Passage Properties of the Polya Urn Process*, verzija 5. 5. 2010, [ogled 13.8. 2017], dostopno na:
`https://arxiv.org/abs/1005.0867`

[2] Andrea Arfè, *Pólya Urn Models*, verzija 19. 5. 2016, [ogled 22. 6. 2017], dostopno na:
`https://andreaarfe.files.wordpress.com/2016/05/urn_models2.pdf`

[3] William Feller, *An Introduction to Probability Theory and its Applications: Volume 1*, John Wiley & Sons Inc., 1959.

[4] Achim Klenke, *Probability Theory - A Comprehensive Course*, Springer, London, 2008.

[5] Hosam Mahmoud, *Polya Urn Models*, Chapman and Hall & CRC Press, 30. 6. 2008

[6] Saad Mneimneh, *The beta density, Bayes, Laplace, and Pólya*, [ogled 3. 7. 2017], dostopno na:
`http://www.cs.hunter.cuny.edu/~saad/courses/bayes/notes/note9.pdf`

[7] Robin Pemantle, *A Time-Dependent Version of Polya's Urn*, verzija februar 1989, [ogled 23. 4. 2017], dostopno na:
`http://digitalassets.lib.berkeley.edu/sdtr/ucb/text/192.pdf`

[8] Antonija Pršlja, *Primerjava in nadgradnja metod za računanje potenc prehodnih matrik markovskih verig*, doktorska disertacija, Fakulteta za matematiko in fiziko, Univerza v Ljubljani, 2015, [ogled 20.8. 2017], dostopno na:
`http://www.matknjiz.si/doktorati/2015/Prslja-14600-3.pdf`

[9] Liu Qiang in Li Jiajin, *Polya's Urn Model And Its Application*, University of Toronto, [ogled 12. 8. 2017], dostopno na:
`http://individual.utoronto.ca/normand/Documents/MATH5501/Project-2/Polya_urn_general_distr.pdf`

[10] Matthew Rathkey, Roy Wiggins in Chelsea Yost, *Pólya Urn Models*, verzija 27. 5. 2013, [ogled 12.5. 2017], dostopno na:
`http://fac.ksu.edu.sa/sites/default/files/polyaurnmodelscompspaper.pdf`

[11] Kyle Siegrist, *The Ehrenfest Chains*, [ogled 15. 7. 2017], dostopno na:
`http://www.math.uah.edu/stat/markov/Ehrenfest.html`

[12] Kyle Siegrist, *Pólya's Urn Process*, [ogled 19.10. 2016], dostopno na:
`http://www.math.uah.edu/stat/urn/Polya.html`

[13] Matija Vidmar, *Pólya urn model*, 2012, verzija 8. 12. 2011, [ogled 12. 1. 2017], dostopno na:
`http://www2.warwick.ac.uk/study/csde/gsp/eportfolio/directory/pg/live/strlad/teaching/polya_urn.pdf`

[14] Timothy C. Wallstrom, *The equalization probability of the Polya urn*, verzija 4. 28. 2011, [ogled 18.5. 2017], dostopno na:
`https://archive.org/details/arxiv-1104.5297`

[15] L. J. Wei, *The Generalized Polya's Urn Design for Sequential Medical Trials*, v: The Annals of Statistics, Volume 7, Number 2, 1979, strani 291-296, [ogled 20.8. 2017], dostopno na:
`https://projecteuclid.org/download/pdf_1/euclid.aos/1176344614`

[16] Tong Zhu, *Nonlinear Pólya Urn Models and Self-Organizing Processes*, doktorska disertacija, University of Pennsylvania, 2009, [ogled 20.8. 2017], dostopno na:
`https://www.math.upenn.edu/grad/dissertations/tongzhudissertation.pdf`