

Simulacija Wright-Fisherovih difuzija

Horvat, Laura

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:217:182263>

Rights / Prava: [In copyright](#)/[Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-11-10**



Repository / Repozitorij:

[Repository of the Faculty of Science - University of Zagreb](#)



SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Laura Horvat

SIMULACIJA WRIGHT-FISHEROVIH
DIFUZIJA

Diplomski rad

Voditelj rada:
dr. sc. Ivana Valentić

Suvoditelj rada:
izv. prof. dr. sc. Nikola Sandrić

Zagreb, 2024.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Mladenu. Jer si mi pokazao što znači boriti se s osmijehom.

Sadržaj

Sadržaj	iv
Uvod	1
1 Simulacije	3
1.1 Metode	3
2 Slučajni procesi	11
2.1 Markovljevi lanci	13
3 Brownovo gibanje	15
3.1 Limes slučajnih šetnji	16
3.2 Simulacija Brownovog gibanja	18
4 Stohastičke diferencijalne jednačbe	19
4.1 Numeričko rješavanje SDJ	21
5 Wright-Fisherove difuzije	27
5.1 Od genetike do difuzije	27
5.2 Simulacija Wright-Fisherovih difuzija	30
5.3 Wright-Fisherove difuzije s mutacijom	38
5.4 Egzaktna simulacija Wright-Fisherovih difuzija	42
Bibliografija	47

Uvod

Cilj većine znanosti, pogotovo prirodnih, upoznavanje je svijeta koji nas okružuje. Za postizanje tog cilja najbolje je proučavati stvarne pojave u prirodi ili društvu. Međutim, ponekad to nije izvedivo, bilo zbog moralnih implikacija, nepredvidljivosti, skupoće, kratkog ili dugog trajanja, opasnosti i slično. Primjerice, u stvarnosti je nemoguće promatrati promjenu udjela nekog gena u populaciji kroz desetak generacija počevši od trenutne ili proučavati učinkovitost novog modela zračnog jastuka u automobilskim nesrećama.

Takvi problemi u znanosti se zaobilaze koristeći simulacije, postupke kojima se oponaša događaje iz stvarnog svijeta sa svrhom njihova proučavanja. Posebnu familiju tih postupaka čine stohastičke simulacije, odnosno simulacije sustava sa slučajnim svojstvima.

Glavni su predmet ovoga rada Wright-Fisherove difuzije, slučajni procesi posebno važni za modeliranje kretanja udjela alela nekog gena u populaciji kroz generacije. Simulacije navedenih procesa problematične su iz više razloga, napose što nemaju zatvorenu formu pomoću koje bi se mogle provesti Monte Carlo simulacije. Matematičari, genetičari, biolozi i ostali znanstvenici bili su primorani smisliti alternativne načine proučavanja ovih procesa, neki od kojih su predstavljeni na nadolazećim stranicama.

Za provedbu samih simulacija korišten je programski jezik **R**.

Poglavlje 1

Simulacije

Kao što je navedeno u Uvodu, u ovome radu ograničavamo se na **stohastičke simulacije** kojima možemo simulirati bilo slučajne uzorke, bilo slučajne procese. U različitim programskim jezicima simulacije se provode pomoću različitih metoda, stoga nećemo ulaziti u detalje. Ipak, u nastavku navodimo neke metode simulacija slučajnih varijabli [13].

1.1 Metode

Metoda inverzije

Metoda 1.1.1. *Neka je F proizvoljna funkcija distribucije. Uzmemo li slučajni uzorak U iz standardne uniformne razdiobe $U(0, 1)$, tada je $Y := F^{-1}(U)$ slučajni uzorak s funkcijom distribucije F .*

Dokaz.

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = [F \text{ je rastuća}] = \mathbb{P}(U \leq F(y)) = [U \sim U(0, 1)] = F(y).$$

□

Napomena 1.1.2. *Primijetimo da proizvoljna funkcija distribucije F ne mora biti bijekcija, stoga se oznakom F^{-1} označava **generalizirani inverz funkcije**, odnosno za $p \in [0, 1]$:*

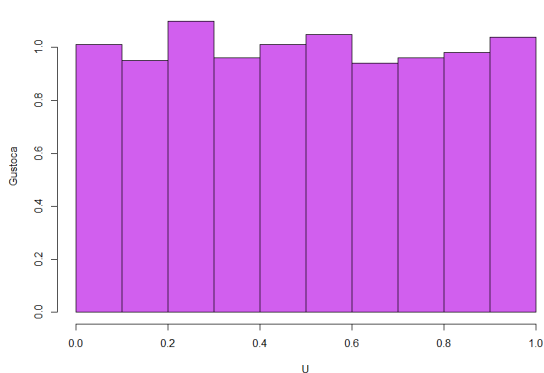
$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Primjer 1.1.3. Simulirajmo metodom inverzije slučajni uzorak duljine 1000 iz eksponencijalne razdiobe s parametrom 1. Generalizirani inverz funkcije distribucije u ovom je slučaju $F^{-1}(x) = -\ln x$. Na Slici 1.1 možemo vidjeti histogram uzorka Y iz standardne uniformne razdiobe te uzorka Z dobivenog metodom inverzije iz Y . Sam kôd naveden je u nastavku.

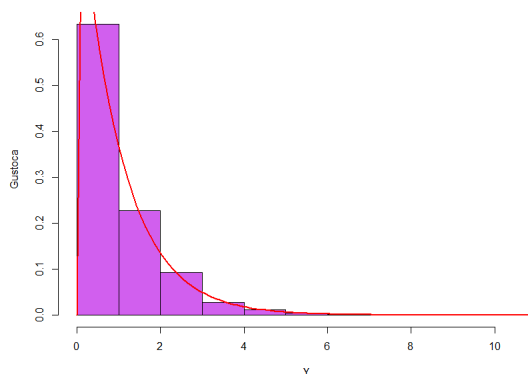
```

1 # Metoda inverzije
2
3 U <- runif(1000)
4 hist(U, main = "", probability = T, ylab = "Gustoca",
5 col='mediumorchid2')
6 Y <- qexp(U)
7 hist(Y, probability = T, ylab = "Gustoca", main = "",
8 col = 'mediumorchid2')
9 curve(dexp(x), add=TRUE, col='red', lwd=2)

```



(a) Uniformna razdioba



(b) Eksponencijalna razdioba

Slika 1.1: Metoda inverzije

Box-Mullerova metoda

Sredinom dvadesetog stoljeća, britanski statističar George E. P. Box i američki informatičar i matematičar Mervin Muller osmislili su metodu za generiranje parova nezavisnih standardnih normalnih slučajnih varijabli pomoću dvije nezavisne standardne uniformne slučajne varijable. Koristeći linearne transformacije, ovom metodom možemo simulirati par proizvoljnih nezavisnih normalnih slučajnih varijabli.

Metoda 1.1.4. Neka su U_1 i U_2 nezavisne standardne uniformne slučajne varijable. Tada su $X = \sqrt{-2 \ln(U_2)} \cos(2\pi U_1)$ i $Y = \sqrt{-2 \ln(U_2)} \sin(2\pi U_2)$ nezavisne standardne normalne slučajne varijable.

Dokaz. Neka su $X, Y \sim N(0, 1)$ nezavisne slučajne varijable. Zajednička gustoća slučajnog vektora je, zbog nezavisnosti,

$$f_{(X,Y)}(x, y) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \right) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$

Prelaskom na polarne koordinate, odnos između (X, Y) i $(R, \Theta) \in [0, +\infty) \times [0, 2\pi)$ dan je s

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} R \cos(\Theta) \\ R \sin(\Theta) \end{bmatrix},$$

dok pripadna funkcija gustoće glasi

$$f_{(R,\Theta)}(r, \theta) = \frac{1}{2\pi} r e^{-\frac{r^2}{2}} = f_{\Theta}(\theta) \cdot f_R(r).$$

Već je iz odnosa gustoća vidljivo da možemo pronaći nezavisne slučajne varijable R i Θ koje bi na opisan način generirale normalne X i Y . Pritom je $\Theta \sim U(0, 2\pi)$, odnosno $\Theta \stackrel{d}{=} 2\pi U_1$ za slučajnu varijablu $U_1 \sim U(0, 1)$.

Nadalje, funkcija distribucije slučajne varijable R sljedećeg je oblika:

$$F_R(x) = \mathbb{P}(R \leq x) = \int_0^x r e^{-\frac{r^2}{2}} dr = [\text{supstitucija}] = 1 - e^{-\frac{x^2}{2}}.$$

Pozivajući se na Metodu inverzije, znamo da je $R \stackrel{d}{=} F_R^{-1}(\tilde{U}_2)$, pri čemu je $\tilde{U}_2 \sim U(0, 1)$, a $F_R^{-1}(t) = \sqrt{-2 \ln(1-t)}$. Budući da je za $U \sim U(0, 1)$ slučajna varijabla $1-U$ također standardna uniformna, neka je $U_2 := 1 - \tilde{U}_2 \sim U(0, 1)$.

Primijetimo da su $U_1 = \frac{1}{2\pi} \Theta$ i $U_2 = 1 - F_R(R)$ nezavisne slučajne varijable te vrijedi:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \sqrt{-2 \ln(U_2)} \cos(2\pi U_1) \\ \sqrt{-2 \ln(U_2)} \sin(2\pi U_1) \end{bmatrix}.$$

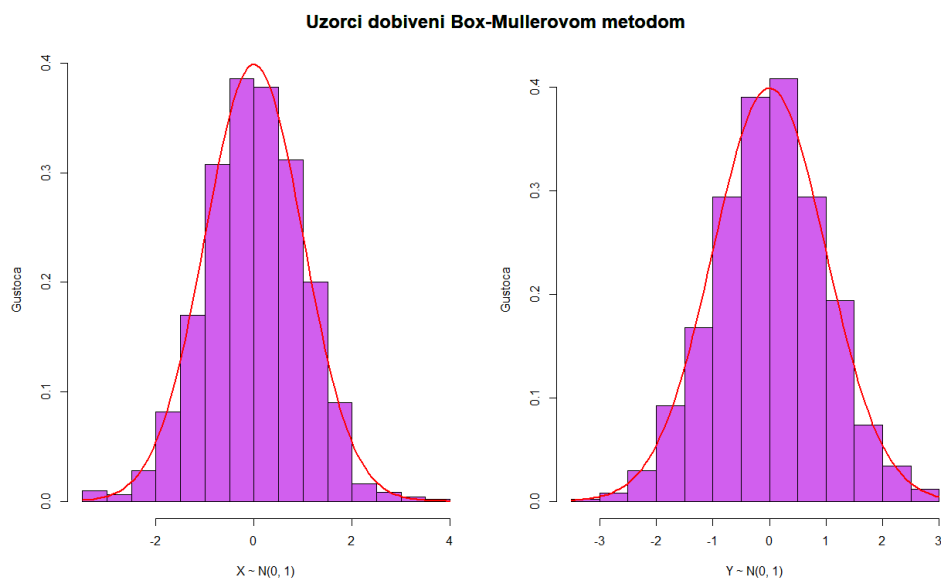
Time je dokaz završen. □

Primjer 1.1.5. U nastavku je naveden kôd kojim su Box-Mullerovom metodom simulirani uzorci duljine 1000. Na temelju Kolomogorov-Smirnovljevog testa u oba slučaja zaključak je da ne odbacujemo nultu hipotezu da uzorci dolaze iz standardne normalne razdiobe. Nadalje, Pearsonovim χ^2 -testom pokazalo se da ne možemo odbaciti početnu hipotezu o njihovoj nezavisnosti.

```

1 # Box-Mullerova metoda
2
3 U1 <- runif(1000)
4 U2 <- runif(1000)
5 X <- sqrt(-2*log(U2)) * cos(2*pi*U1)
6 Y <- sqrt(-2*log(U2)) * sin(2*pi*U1)
7
8 ks.test(X, "pnorm")
9 ks.test(Y, "pnorm")
10 chisq.test(X,Y) # test nezavisnosti uzoraka
11
12 hist(X, probability = T, xlab = "X ~ N(0, 1)", ylab = "Gustoca",
13 main = "", col = 'mediumorchid2')
14 curve(dnorm(x), add=TRUE, col='red', lwd=2)
15 hist(Y, probability = T, xlab = "Y ~ N(0, 1)", ylab = "Gustoca",
16 main = "", col = 'mediumorchid2')
17 curve(dnorm(x), add=TRUE, col='red', lwd=2)

```



Slika 1.2: Histogrami dobivenih uzoraka

```

One-sample Kolmogorov-Smirnov test
data: X
D = 0.028314, p-value = 0.3992
alternative hypothesis: two-sided

Pearson's Chi-squared test
data: X and Y
X-squared = 999000, df = 998001, p-value = 0.2397

One-sample Kolmogorov-Smirnov test
data: Y
D = 0.022222, p-value = 0.7067
alternative hypothesis: two-sided

```

(a) χ^2 -test nezavisnosti

(b) K-S testovi

Slika 1.3: Provedeni testovi

Metoda prihvatanja i odbacivanja

Metoda 1.1.6. Neka su F i G funkcije distribucije s pripadajućim funkcijama gustoće f i g , za koje postoji konstanta $c \geq 1$ takva da za svaki $x \in \mathbb{R}$ vrijedi $f(x) \leq cg(x)$.

- Neka je Y slučajna varijabla s funkcijom distribucije G .
- Neka je U njoj nezavisna standardna uniformna slučajna varijabla, $U \sim U(0, 1)$.
- Ako je $Ucg(Y) \leq f(Y)$, $X := Y$ ("prihvati"). U suprotnom, vrati se na korak a. ("odbaci").

Tada je X slučajna varijabla s funkcijom distribucije F .

Napomena 1.1.7. Prije dokaza i primjera, primijetimo sljedeće [15]:

- $f(Y)$ i $g(Y)$ su slučajne varijable, kao i omjer $\frac{f(Y)}{cg(Y)}$. Kao funkcija slučajne varijable Y , nezavisan je od U .
- Zbog svojstava funkcija f i g , za omjer vrijedi $0 < \frac{f(Y)}{cg(Y)} \leq 1$.
- Neka je N broj pokušaja do prihvatanja slučajne varijable Y . Neka je p vjerojatnost uspjeha: $p = \mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right)$. Tada je vjerojatnost da slučajnu varijablu prihvatimo u n -tom pokušaju $\mathbb{P}(N = n) = (1 - p)^{n-1}p$. Slijedi da je N slučajna varijabla s geometrijskom distribucijom, $N \sim G(p)$.

- Izračunajmo p . Neka je $h_{(U,Y)}$ funkcija gustoće slučajnog vektora (U, Y) .

$$\begin{aligned} p &= \mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \iint_{\left\{U \leq \frac{f(Y)}{cg(Y)}\right\}} h_{(U,Y)}(u, y) \, dud y = [\text{nezavisnost}] = \\ &= \int_{-\infty}^{+\infty} \int_0^{\frac{f(y)}{cg(y)}} 1 \cdot g(y) \, dud y = \int_{-\infty}^{+\infty} \frac{f(y)}{cg(y)} g(y) dy = \frac{1}{c} \int_{-\infty}^{+\infty} f(y) dy = \frac{1}{c}. \end{aligned}$$

Dakle, očekivani broj pokušaja do prihvaćanja je $\mathbb{E}[N] = \frac{1}{p} = c$. Prema tome, pomoćnu funkciju distribucije G potrebno je birati kako bi funkcije gustoće f i g bile što bliže, odnosno konstanta c približno 1.

- Za diskretne slučajne varijable također je moguće na analogan način provesti metodu. Sam opis dan je u [15].

Dokaz Metode 1.1.6.

Slučajna varijabla Y prihvaćena je pod uvjetom da je $U \leq \frac{f(Y)}{cg(Y)}$. Za $x \in \mathbb{R}$ vrijedi sljedeće:

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}\left(Y \leq x \mid U \leq \frac{f(Y)}{cg(Y)}\right) = \frac{\mathbb{P}\left(Y \leq x, U \leq \frac{f(Y)}{cg(Y)}\right)}{\mathbb{P}\left(U \leq \frac{f(Y)}{cg(Y)}\right)} = \frac{\mathbb{P}\left(Y \leq x, U \leq \frac{f(Y)}{cg(Y)}\right)}{p} = \\ &= c \cdot \iint_{\left\{Y \leq x, U \leq \frac{f(Y)}{cg(Y)}\right\}} h_{(U,Y)}(u, y) \, dud y = [\text{nezavisnost}] = c \cdot \int_{-\infty}^x \int_0^{\frac{f(y)}{cg(y)}} 1 \cdot g(y) \, dud y = \\ &= c \cdot \int_{-\infty}^x \frac{f(y)}{cg(y)} g(y) dy = \int_{-\infty}^x f(y) dy = F(x). \end{aligned}$$

□

Primjer 1.1.8. Simulirajmo metodom prihvaćanja i odbacivanja slučajni uzorak duljine 1000 iz normalne razdiobe, tj. za neke $\mu \in \mathbb{R}$, $\sigma^2 > 0$, tražimo $X \sim N(\mu, \sigma^2)$. Znamo da je u tom slučaju $X \stackrel{d}{=} \sigma Z + \mu$, za $Z \sim N(0, 1)$. Prema tome, dovoljno je simulirati uzorak iz standardne normalne razdiobe.

Kako bismo lakše pronašli pomoćnu funkciju distribucije, koristimo činjenicu da je normalna razdioba simetrična. Tako možemo generirati $Y = |Z|$ te staviti $Z = Y$ ako je nezavisna standardna uniformna slučajna varijabla $U_s \leq 0.5$ i $Z = -Y$ ako je $U_s > 0.5$.

$|Z|$ je nenegativna slučajna varijabla s funkcijom gustoće

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathbb{1}_{[0, +\infty)}.$$

Kao pomoćnu slučajnu varijablu uzimamo eksponencijalnu s parametrom 1. Pomoćna funkcija gustoće tada je $g(x) = e^{-x}$. Tražimo konstantu $c > 0$ takvu da je za svaki $x \in \mathbb{R}$ $f(x) \leq cg(x)$. Drugim riječima, dovoljno je maksimizirati funkciju $\frac{f(x)}{g(x)}$ na skupu pozitivnih brojeva. Upravo vrijednost maksimuma bit će tražena konstanta. Neka je

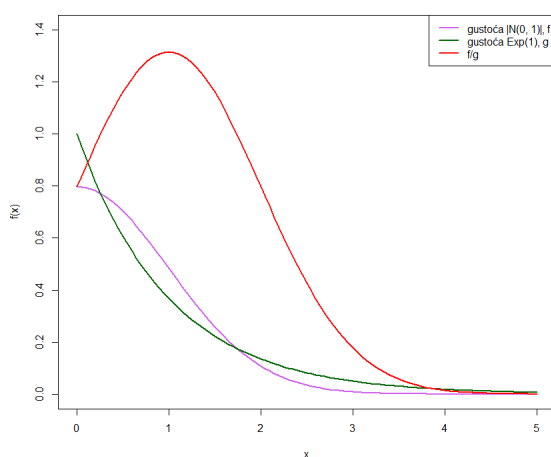
$$h(x) := \frac{f(x)}{g(x)} = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + x}.$$

Tada je $h'(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + x}(-x + 1)$, što nam daje stacionarnu točku $x^* = 1$. Budući da je h' pozitivna za $x < 1$, a negativna za $x \geq 1$, zaključujemo da se doista radi o globalnom maksimumu. Prema tome,

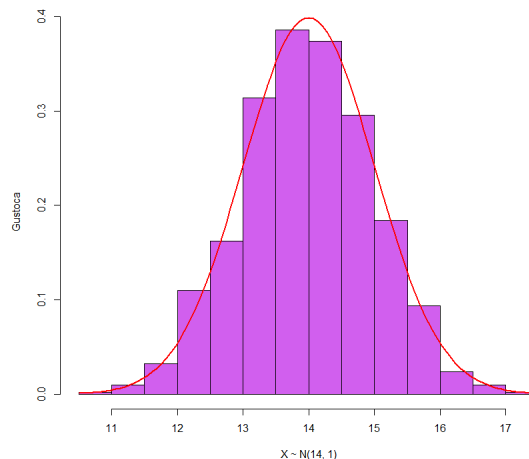
$$c := h(1) = \sqrt{\frac{2e}{\pi}} \approx 1.32.$$

Budući da je c relativno blizu 1, zaključujemo da je g dobar odabir pomoćne funkcije gustoće. Na grafu sa Slike 1.4 vidi se usporedba funkcija gustoće f i g te njihov količnik. Na istoj slici prikazan je i histogram simuliranog uzorka s pripadajućom funkcijom gustoće normalne slučajne razdiobe. Za primjer, uzeti su $\mu = 14$, $\sigma^2 = 1$.

U nastavku je naveden kôd pomoću kojeg se u \mathbf{R} -u izvršila simulacija. U ovome se primjeru petlja izvršila 1330 puta, dok je prema Napomeni 1.1.7 očekivani broj pokušaja za uzorak duljine 1000 otprilike 1320, što je relativno blizu opaženom.



(a) Usporedba funkcija gustoće



(b) Histogram dobivenog uzorka

Slika 1.4: Metoda prihvaćanja i odbacivanja

```
1 # Metoda prihvatanja i odbacivanja
2
3 U_s <- runif(1000) # za predznak
4 Z <- rep(0, 1000) # buduci normalni slucajni uzorak
5 i <- 1 # indeks
6 br <- 0 # informativno, broj pokusaja
7 while (i <= 1000) {
8   br <- br + 1
9   u1 <- runif(1) # za eksponencijalnu varijablu
10  y <- -log(u1) # metoda inverzije
11  u <- runif(1)
12  if (u <= exp(-(y-1)^2/2)){
13    if(U_s[i] <= 0.5){
14      Z[i] <- y
15    } else{
16      Z[i] <- -y
17    }
18    i <- i + 1
19  }
20 }
21
22 sigma <- 1
23 mi <- 14
24 X <- sigma * Z + mi # X ~ N(14, 1)
25
26 hist(X, probability = T, xlab = "X ~ N(14, 1)", ylab = "Gustoca",
27 main = "", col = 'mediumorchid2')
28 curve(dnorm(x, mean = 14), add=TRUE, col='red', lwd=2)
```

Poglavlje 2

Slučajni procesi

U praksi, često je korisno promatrati neku slučajnu pojavu kroz vrijeme, stoga uvodimo definiciju slučajnog procesa. [17]

Definicija 2.0.1. *Neka je (Ω, \mathcal{F}) izmjeriv prostor i $\mathcal{T} \subseteq [0, +\infty)$. **Slučajni proces** je familija $X = (X_t : t \in \mathcal{T})$ slučajnih varijabli na (Ω, \mathcal{F}) .*

*Ukoliko je $\mathcal{T} = \mathbb{N}$, radi se o slučajnom procesu s **diskretnim vremenom**. Ako pak je $\mathcal{T} = [0, +\infty)$, riječ je o slučajnom procesu s **neprekidnim vremenom**.*

Približimo definiciju primjerima.

Primjer 2.0.2. Slučajna šetnja.

Jedan od najjednostavnijih, a opet vrlo korisnih slučajnih procesa s diskretnim vremenom je slučajna šetnja. Neka je $X = (X_n : n \in \mathbb{N})$ familija nezavisnih jednakodistribuiranih slučajnih varijabli, pri čemu je

$$X_n \sim \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Definiramo slučajni proces $S = (S_n : n \in \mathbb{N}_0)$ na sljedeći način:

$$S_0 = 0, \quad S_n = X_1 + \dots + X_n, \quad n \in \mathbb{N}.$$

Proces S zovemo jednostavnom simetričnom slučajnom šetnjom. Grafički prikaz primjera slučajne šetnje vidljiv je na Slici 2.1.

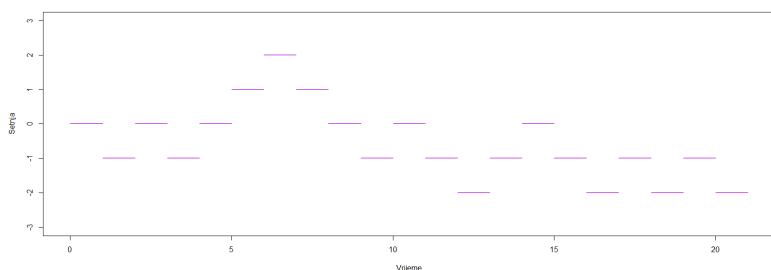
Primjer 2.0.3. Jednostavni proces grananja.

Neka je $(Z_{n,i} : n, i \in \mathbb{N})$ niz nezavisnih jednakodistribuiranih nenegativnih cjelobrojnih slučajnih varijabli. Jednostavni proces grananja, također poznat pod imenom Galton-Watsonov proces, slučajni je proces $Z = (Z_n : n \in \mathbb{N}_0)$ definiran s:

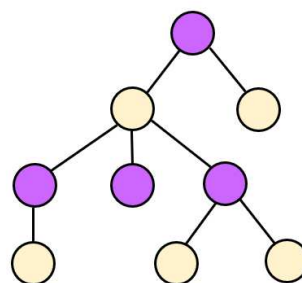
$$\begin{aligned} Z_0 &= 1 \\ Z_1 &= Z_{1,1} \\ Z_2 &= Z_{2,1} + Z_{2,2} + \cdots + Z_{2,Z_1} \\ &\dots \\ Z_n &= Z_{n,1} + Z_{n,2} + \cdots + Z_{n,Z_{n-1}}, \end{aligned}$$

pri čemu $Z_n = 0$ povlači $Z_{n+k} = 0$ za svaki $k \in \mathbb{N}$. Slučajna varijabla Z_n interpretira se kao broj jedinki u n -toj generaciji neke populacije, dok je $Z_{n,i}$ broj potomaka u n -toj generaciji poteklih od i -te jedinke $(n-1)$ -e generacije.

Kod procesa grananja jedno od zanimljivijih pitanja je kolika je vjerojatnost da populacija u nekom trenutku izumre. Ispostavlja se, a dokaz se može pronaći u [10], da će populacija sigurno izumrijeti ako je očekivani broj potomaka manji ili jednak 1, odnosno, zbog jednake distribucije, kad je $\mathbb{E}[Z_1] \leq 1$. U suprotnom, postoji pozitivna vjerojatnost da će populacija preživjeti.



(a) Primjer slučajne šetnje



(b) Galton-Watsonov proces

Slika 2.1: Primjeri slučajnih procesa

Primjer slučajnih procesa s neprekidnim vremenom su difuzije, od kojih je najjednostavnije i najpoznatije Brownovo gibanje. Njime ćemo se pozabaviti u Poglavlju 3.

2.1 Markovljevi lanci

Jedno od korisnijih svojstava kod proučavanja odnosa između slučajnih varijabli svakako je nezavisnost. Međutim, u praksi često izostaje. Stoga se u primjenama, ali i u samoj teoriji vjerojatnosti, vrlo važnima ispostavljaju Markovljevi lanci, slučajni procesi kod kojih sljedeće stanje ovisi isključivo o trenutnom stanju u kojem se nalazimo, a ne o putu kojim smo do njega došli.

Definicija 2.1.1. *Neka je S neprazan skup. Slučajni proces $X = (X_t : t \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u skupu S je **Markovljev lanac** ako vrijedi*

$$\mathbb{P}(X_{t_{n+1}} = j \mid X_{t_n} = i, X_{t_{n-1}} = i_{n-1}, \dots, X_{t_0} = i_0) = \mathbb{P}(X_{t_{n+1}} = j \mid X_{t_n} = i), \quad (2.1)$$

za svaki $n \in \mathbb{N}_0$, za sve $0 \leq t_1 < t_2 < \dots < t_{n+1}$ i za sve $i_0, \dots, i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo (2.1) naziva se **Markovljevim svojstvom**.

Valja primijetiti da Markovljevi lanci mogu imati prebrojiv i neprebrojiv skup stanja. Jednako tako, to mogu biti procesi s diskretnim ili neprekidnim vremenom.

Slučajna šetnja iz prošlog odjeljka očito je primjer Markovljevog lanca s diskretnim vremenom i skupom stanja.

Nadalje, zbog jednake distribucije broja potomaka svake jedinke, jednostavni proces grananja također je Markovljev lanac s diskretnim vremenom, s obzirom da broj potomaka sljedeće generacije ovisi isključivo o jedinkama (i potomcima jedinki) trenutne generacije.

Poglavlje 3

Brownovo gibanje

Brownovo gibanje dobilo je ime po škotskom botaničaru Robertu Brownu, koji je pod mikroskopom promatrao kretanje peluda u vodi. Primijetio je da se malene čestice neprestano gibaju u svim smjerovima. U člancima iz 1828. i 1829. godine objavio je neka zapažanja ([13]):

- ▶ gibanje je prilično nepravilno, sastavljeno od translacija i rotacija;
- ▶ čini se da se čestice gibaju jedna nezavisno od druge;
- ▶ što su čestice manje, gibanje je dinamičnije;
- ▶ gustoća i raspored čestica nemaju utjecaja na gibanje;
- ▶ gibanje je dinamičnije što je tekućina manje viskozna;
- ▶ gibanje je neprestano;
- ▶ gibanje nije uzrokovano isparavanjem ili strujanjem tekućine;
- ▶ čestice ničime nisu pokretane.

Nakon inicijalnog Brownovog zapažanja, pojavilo se nekoliko teorija. Torvald Thiele prvi je 1880. godine matematički opisao Brownovo gibanje. Nezavisno od toga, Louis Bachelier je 1900. učinio isto u svom doktoratu. Konačno, Albert Einstein je 1905. otkrio sam uzrok ovog procesa, čime je pokazao da je Brownovo gibanje način indirektno potvrde postojanja atoma i molekula.

Definicija 3.0.1. Standardno jednodimenzionalno Brownovo gibanje slučajni je proces $B = (B_t)_{t \geq 0}$ sa sljedećim svojstvima:

(BG 1) $B_0 = 0$ g.s.

(BG 2) Za svaki $n \in \mathbb{N}$, za svaki $0 \leq t_1 < \dots < t_n$, $B_{t_1}, B_{t_2-t_1}, \dots, B_{t_n-t_{n-1}}$ su nezavisne slučajne varijable. Drugim riječima, B ima nezavisne priraste.

(BG 3) Za $s, t \geq 0$ $B_{s+t} - B_s \sim N(0, t)$.

(BG 4) Za g.s. $\omega \in \Omega$ $t \mapsto B_t(\omega)$ je neprekidna funkcija.

3.1 Limes slučajnih šetnji

O Brownovom gibanju možemo razmišljati i kao o svojevrsnoj "ubrzanom" slučajnoj šetnji. [7]

Neka je $(Y_n)_n$ niz nezavisnih, jednakodistribuiranih slučajnih varijabli s očekivanjem 0 i varijancom 1. Primjerice,

$$Y_n \sim \begin{pmatrix} -1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Nadalje, neka je $S_0 = 0$ g.s. (usp. (BG 1)), $S_n = Y_1 + \dots + Y_n$ slučajna šetnja. Primijetimo da ima nezavisne priraste, odnosno, za svaki $n \in \mathbb{N}$, $k \in \mathbb{N}_0$ slučajne varijable S_n i $S_{n+k} - S_n$ su nezavisne kao sume nezavisnih slučajnih varijabli:

$$S_n = \sum_{i=1}^n Y_i, \quad S_{n+k} = \sum_{i=n+1}^{n+k} Y_i.$$

Dakako, na isti način indukcijom dolazimo do zaključka da su za $k \in \mathbb{N}$ te $0 \leq n_1 < n_2 < \dots < n_k$ nezavisne i slučajne varijable $S_{n_1}, S_{n_2} - S_{n_1}, \dots, S_{n_k} - S_{n_{k-1}}$.

Fiksirajmo $n \in \mathbb{N}$ te definirajmo "ubrzanu" slučajnu šetnju. Za $t \geq 0$ za koji je $nt \in \mathbb{N}_0$, neka je $S_t^{(n)} := S_{nt}$. Za preostale $t \geq 0$ linearno interpolirajmo na sljedeći način:

$$\begin{aligned} S_t^{(n)} &:= S_{\lfloor nt \rfloor} + (S_{\lfloor nt \rfloor + 1} - S_{\lfloor nt \rfloor})(nt - \lfloor nt \rfloor) = \\ &= S_{\lfloor nt \rfloor} + Y_{\lfloor nt \rfloor + 1}(nt - \lfloor nt \rfloor). \end{aligned}$$

Primjerice, za $t = \frac{k}{n}$ vrijedi $S_t^{(n)} = S_{k/n}^{(n)} = S_k + Y_{k+1}(k - k) = S_k$, što objašnjava naziv *ubrzanu* slučajna šetnja.

Budući da je zbog nezavisnosti $\text{Var} S_1^{(n)} = \sum_{k=1}^n \text{Var} Y_k = n$, proces $(S_1^{(n)})_n$ ne konvergira, što motivira skaliranje. Definiramo, stoga,

$$B_t^{(n)} := \frac{1}{\sqrt{n}} S_t^{(n)}.$$

Iz centralnog graničnog teorema odmah slijedi da $B_1^{(n)} = \frac{S_n}{\sqrt{n}} \xrightarrow{d} N(0, 1)$, dok dokaz analogne tvrdnje za $B_t^{(n)}$, tj. $B_t^{(n)} \xrightarrow{d} N(0, t)$ zahtijeva malo igranja. Naime,

$$B_t^{(n)} = \frac{1}{\sqrt{n}} S_{[nt]} + \frac{1}{\sqrt{n}} Y_{[nt]+1} (nt - [nt]).$$

Prvi pribrojnik je konvergentan po centralnom graničnom teoremu,

$$\frac{1}{\sqrt{n}} S_{[nt]} = \sqrt{\frac{[nt]}{n}} \cdot \frac{S_{[nt]}}{[nt]} \xrightarrow{d} \sqrt{t} \cdot N(0, 1). \quad (3.1)$$

Za konvergenciju drugog pribrojnika iskoristit ćemo teorem neprekidnosti. Označimo s ϕ karakterističnu funkciju slučajne varijable $\frac{1}{\sqrt{n}} Y_{[nt]+1} (nt - [nt])$. Zbog jednake distribucije i svojstava karakterističnih funkcija, za svaki $s \in \mathbb{R}$ vrijedi

$$\phi(s) = \phi_{Y_1} \left(\frac{nt - [nt]}{\sqrt{n}} s \right).$$

Iz uniformne neprekidnosti karakterističnih funkcija te iz teorema o sendviču, proizlazi

$$\lim_{n \rightarrow \infty} \phi(s) = \phi_{Y_1} \left(\lim_{n \rightarrow \infty} \frac{nt - [nt]}{\sqrt{n}} s \right) = \phi_{Y_1}(0) = 1 = \phi_0(s).$$

Po teoremu neprekidnosti slijedi

$$\frac{1}{\sqrt{n}} Y_{[nt]+1} (nt - [nt]) \xrightarrow{d} 0. \quad (3.2)$$

Konačno, iz Slutskyjeve leme te rezultata (3.1) i (3.2) dobivamo traženu tvrdnju, koju pak uspoređujemo sa svojstvom (BG 3) iz definicije Brownovog gibanja.

3.2 Simulacija Brownovog gibanja

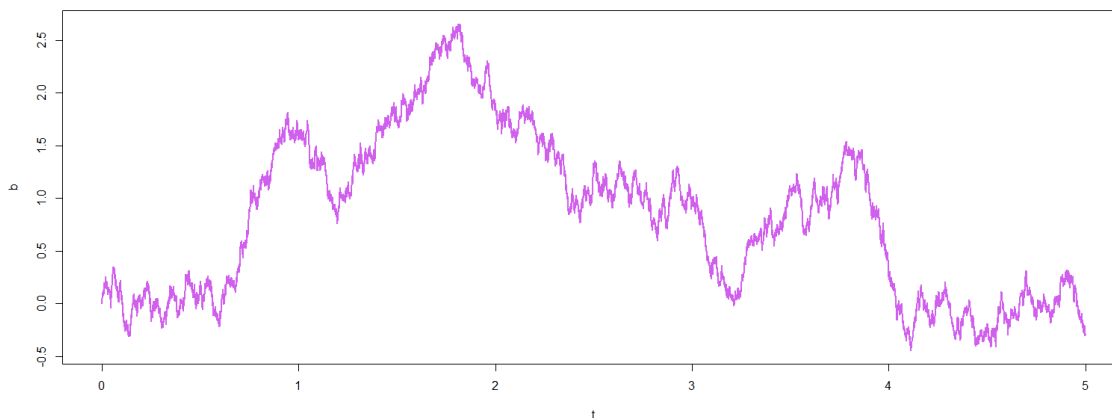
U ovom odjeljku naveden je primjer algoritma kojim se može simulirati Brownovo gibanje. Zasniva se na normalnosti i nezavisnosti prirasta.

Iako se nijednim algoritmom ne može precizno prikazati Brownovo gibanje zbog neprekidnosti vremena, dovoljno sitnim vremenskim intervalima može se postići zadovoljavajuća aproksimacija. Primjer grafa jedne takve aproksimacije prikazan je na Slici 3.1, dok su metoda i kôd navedeni u nastavku.

Metoda 3.2.1. *Neka je $n \in \mathbb{N}$ te $0 = t_0 < t_1 < \dots < t_n$ vremenska mreža. Postavimo $B_0 = 0$. U svakom od n koraka simuliramo $Y \sim N(0, t_j - t_{j-1})$ te stavimo $B_{t_j} = B_{t_{j-1}} + Y$. Tada su $(B_0, B_{t_1}, \dots, B_{t_n})$ slučajne varijable iz procesa Brownovog gibanja.*

```

1 # Brownovo gibanje
2
3 n <- 10000
4 t <- seq(0, 5, length.out = n+1)
5 b <- rep(0, n+1)
6 for (j in 1:n){
7   b[j+1] <- b[j] + rnorm(1, 0, sqrt(t[j+1] - t[j]))
8 }
9 plot(t, b, type = 'l', lwd = 2, col = 'mediumorchid2')
```



Slika 3.1: Primjer simulacije Brownovog gibanja

Poglavlje 4

Stohastičke diferencijalne jednačbe

Definicija 4.0.1. Stohastička diferencijalna jednačba (SDJ) diferencijalna je jednačba čiji su koeficijenti slučajni brojevi ili funkcije nezavisnih slučajnih varijabli. [16]

Napomena 4.0.2. [11]

- U ovome se radu ograničavamo na stohastičke diferencijalne jednačbe oblika

$$dX_t = a(X_t) dt + b(X_t) dB_t, \quad X_0 = x_0, \quad (4.1)$$

pri čemu su $a, b : \mathbb{R} \rightarrow \mathbb{R}$ funkcije, $x_0 \in \mathbb{R}$ konstanta, $a(B_t)_t$ standardno Brownovo gibanje. Funkciju a zovemo **koeficijentom drifta**, dok funkcija b predstavlja **koeficijent difuzije**.

- Gornja je SDJ zapravo integralna jednačba oblika

$$X_t = x_0 + \int_0^t a(X_s) ds + \int_0^t b(X_s) dB_s,$$

pri čemu je $\int_0^t b(X_s) dB_s$ **Itôv integral** procesa $(b(X_t))_{t \geq 0}$.

- Dok u običnom diferencijalnom računu vrijedi ulančano pravilo za derivaciju kompozicije, kod Brownovog gibanja to ne vrijedi. Konkretno, $df(B_t) \neq f'(B_t)dB_t$, već vrijedi tzv. **Itôva formula** za Brownovo gibanje:

$$df(B_t) = f'(B_t) dB_t + \frac{1}{2} f''(B_t) dt.$$

Ili, u integralnom obliku:

$$f(B_T) - f(B_0) = \int_0^T f'(B_t) dB_t + \frac{1}{2} \int_0^T f''(B_t) dt.$$

Dokaz ove tvrdnje može se pročitati u [18], kao i dodatne informacije o Itôvim procesima i integralima.

- Iz prethodne tvrdnje slijedi Itôva formula za funkciju dvije varijable koju ćemo koristiti u nadolazećem primjeru:

$$df(t, B_t) = \frac{\partial f}{\partial t}(t, B_t) dt + \frac{\partial f}{\partial x}(t, B_t) dB_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}(t, B_t) dt.$$

Primjer 4.0.3. Geometrijsko Brownovo gibanje. [8], [9]

Rješenje jednog od najjednostavnijih primjera stohastičkih diferencijalnih jednadžbi proces je koji zovemo geometrijskim Brownovim gibanjem. Vrlo je važno u financijskoj matematici jer opisuje kretanje cijena dionica u Black-Scholesovom modelu.

Neka su $\mu \in \mathbb{R}$, $\sigma > 0$ konstante te neka slučajni proces $(X_t)_{t \geq 0}$ zadovoljava SDJ

$$dX_t = \mu X_t dt + \sigma X_t dB_t, \quad X_0 = x_0 > 0.$$

Tražimo rješenje u obliku $X_t = f(t, B_t)$. Primjenom Itôve formule dobivamo

$$dX_t = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dB_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} dt = \left(\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \right) dt + \frac{\partial f}{\partial x} dB_t.$$

Izjednačavanjem koeficijenata dobivamo

$$\mu f = \frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2}, \quad \sigma f = \frac{\partial f}{\partial x}.$$

Iz drugog izraza, prebacivanjem funkcije f na drugu stranu jednakosti te integriranjem, slijedi

$$\begin{aligned} \ln f(t, x) &= \sigma x + g(t), \quad \text{odnosno,} \\ f(t, x) &= e^{\sigma x + g(t)}. \end{aligned}$$

Uvrštavanjem u prvi izraz dobije se

$$\begin{aligned} \mu f(t, x) &= g'(t) f(t, x) + \frac{1}{2} \sigma^2 f(t, x), \\ g'(t) &= \mu - \frac{1}{2} \sigma^2, \\ g(t) &= \left(\mu - \frac{1}{2} \sigma^2 \right) t + C, \quad C \in \mathbb{R}. \end{aligned}$$

Dakle,

$$f(t, x) = e^{\sigma x + (\mu - \frac{1}{2} \sigma^2) t + C}, \quad C \in \mathbb{R}.$$

Slijedi

$$\begin{aligned} X_t &= f(t, B_t) = e^{\sigma B_t + (\mu - \frac{1}{2}\sigma^2)t + C}, \quad C \in \mathbb{R} \\ x_0 &= X_0 = f(0, B_0) = e^C \end{aligned}$$

Konačno,

$$X_t = x_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma B_t}.$$

Razlika između geometrijskog i standardnog Brownovog gibanja je u koeficijentu drifta: dok je kod geometrijskog Brownovog gibanja linearna funkcija, kod standardnog on je nula.

U jednom od nadolazećih poglavlja vidjet ćemo da su i Wright-Fisherove difuzije slučajni procesi koji zadovoljavaju određene stohastičke diferencijalne jednačbe.

4.1 Numeričko rješavanje stohastičkih diferencijalnih jednačbi

Neka je $(X_t)_t$ slučajni proces koji zadovoljava SDJ (4.1). Nadalje, neka je $[0, T]$ segment na kojem taj proces promatramo.

Odaberimo $n \in \mathbb{N}$ koji određuje ekvidistantnu particiju $0 = t_0 < t_1 < \dots < t_n = T$ segmenta $[0, T]$ s korakom $\Delta t := \frac{T}{n}$.

Nadalje, neka su $\Delta_i B := B_{t_i} - B_{t_{i-1}} \sim N(0, \Delta t)$, $i \in \{1, \dots, n\}$, prirasti standardnog Brownovog gibanja.

Numeričko rješenje SDJ slučajni je proces $(X_t^n)_{t \in [0, T]}$ koji aproksimira pravo rješenje u točkama particije. U preostalim točkama proces $(X_t^n)_{t \in [0, T]}$ linearno interpoliramo. U ovom odjeljku promotrit ćemo dvije numeričke aproksimacijske metode rješavanja stohastičkih diferencijalnih jednačbi, Eulerovu shemu i Milsteinovu shemu.

Eulerova shema

Eulerova je shema najjednostavnija aproksimacijska metoda zasnovana na linearnosti integrala i pretpostavci da su funkcije a i b dovoljno glatke, odnosno da za malenu duljinu intervala Δt ne variraju previše na svakom intervalu $[t_{i-1}, t_i]$.

Konkretno, znamo da vrijedi

$$X_{t_i} = X_{t_{i-1}} + \int_{t_{i-1}}^{t_i} a(X_s) ds + \int_{t_{i-1}}^{t_i} b(X_s) dB_s,$$

na temelju čega aproksimaciju definiramo na sljedeći način:

$$\begin{aligned} X_0^n &:= x_0, \\ X_{t_i}^n &:= X_{t_{i-1}}^n + a(X_{t_{i-1}}^n) \Delta t + b(X_{t_{i-1}}^n) \Delta_t B, \quad i \in \{1, \dots, n\}. \end{aligned}$$

Milsteinova shema [4]

Vrijedi formula

$$X_t = X_0 + \int_0^t a(X_s) ds + \int_0^t b(X_s) dB_s.$$

Iz Taylorovog razvoja za funkciju b i Eulerove sheme kojom aproksimiramo X_t dobivamo

$$\begin{aligned} b(X_t) &= b(X_0) + b'(X_0)(X_t - X_0) \\ &\approx b(X_0) + b'(X_0)(a(X_0)t + b(X_0) \cdot B_t) = \\ &= b(X_0) + b'(X_0) \cdot b(X_0) \cdot B_t + O(t). \end{aligned}$$

Integriranjem slijedi

$$\begin{aligned} \int_0^t b(X_s) dB_s &= b(X_0) \int_0^t dB_s + b'(X_0) \cdot b(X_0) \int_0^t B_s dB_s + O(t)B_t = \\ &= b(X_0) \int_0^t dB_s + b'(X_0) \cdot b(X_0) \int_0^t B_s dB_s + O(t^{\frac{3}{2}}), \end{aligned}$$

pri čemu posljednja jednakost slijedi jer $B_t \sim N(0, t)$, odnosno $\frac{B_t}{\sqrt{t}} \sim N(0, 1)$, stoga vrijednosti B_t rastu proporcionalno s \sqrt{t} .

Želimo na drugačiji način izraziti posljednji Itôv integral. Za $f(B_s) = \frac{1}{2}B_s^2$ iz Itôve formule dobivamo

$$\begin{aligned} \frac{1}{2}B_t^2 - \frac{1}{2}B_0^2 &= \int_0^t B_s dB_s + \frac{1}{2} \int_0^t ds, \quad \text{odnosno} \\ \int_0^t B_s dB_s &= \frac{1}{2}(B_t^2 - t). \end{aligned}$$

Uvrštavanjem u Eulerovu shemu slijedi

$$X_t = X_0 + a(X_0)t + b(X_0) \cdot B_t + \frac{1}{2}b'(X_0) \cdot b(X_0)(B_t^2 - t).$$

Konačno, ta nas relacija motivira da uvedemo Milsteinovu metodu za numeričku aproksimaciju rješenja stohastičke diferencijalne jednačbe:

$$\begin{aligned} X_0 &= x_0, \\ X_{t_i}^n &= X_{t_{i-1}}^n + a(X_{t_{i-1}}^n)\Delta t + b(X_{t_{i-1}}^n) \cdot \Delta_i B + \\ &+ \frac{1}{2}b(X_{t_{i-1}}^n) b(X_{t_{i-1}}^n) \cdot b'(X_{t_{i-1}}^n) ((\Delta_i B)^2 - \Delta t), \quad i \in \{1, \dots, n\}. \end{aligned}$$

Primjer 4.1.1. Simulacija geometrijskog Brownovog gibanja.

U Primjeru 4.0.3 upoznali smo se s geometrijskim Brownovim gibanjem, koje ćemo sada simulirati dvjema prethodno navedenim metodama.

Neka je $x_0 = 35$, $\mu = 0.2$, $\sigma = 1$. Geometrijsko Brownovo gibanje s tim parametrima i početnim uvjetom zadovoljava stohastičku diferencijalnu jednačbu

$$dX_t = 0.2X_t dt + X_t dB_t, \quad X_0 = 35 > 0,$$

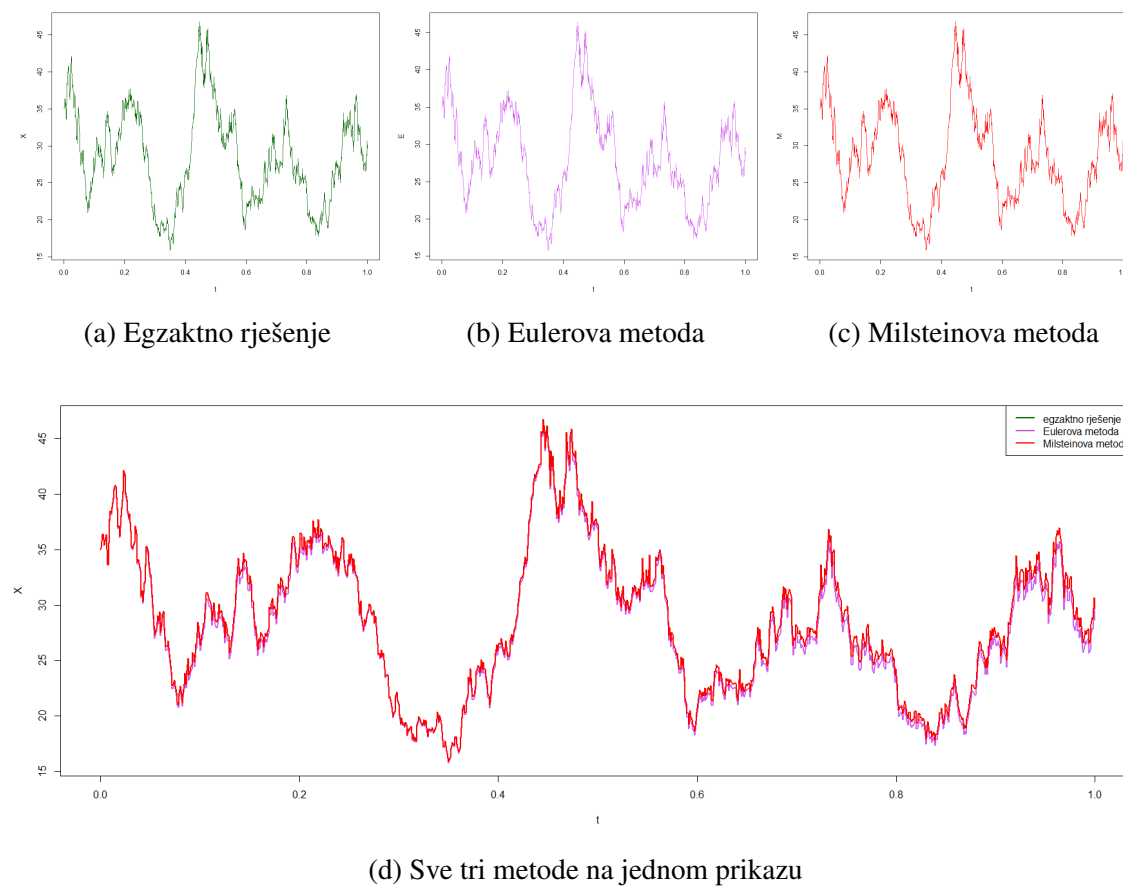
Dakle, funkcije a i b glase

$$a(X_t) = 0.2X_t, \quad b(X_t) = X_t.$$

Promatramo gibanje na segmentu $[0, 1]$, a za numeričke aproksimativne metode koristimo ekvidistantnu subdiviziju s korakom $\Delta t = 2^{-10}$. Kao što je navedeno u Primjeru 4.0.3, egzaktno rješenje ove SDJ dano je formulom

$$X_t = x_0 e^{(0.2 - \frac{1}{2})t + B_t}, \quad t \in [0, 1]. \quad (4.2)$$

Na Slici 4.1 prikazane su trajektorije dobivene egzaktno - spajanjem uređenih parova točaka subdivizije segmenta i njihovih funkcijskih vrijednosti dobivenih pomoću formule [4.2], Eulerovom metodom te Milsteinovom metodom, dok je kôd naveden u nastavku. Možemo primijetiti da na grafu uopće nije vidljivo egzaktno rješenje jer ga i numerička rješenja prilično dobro opisuju već za ovu subdiviziju.



Slika 4.1: Simulacija geometrijskog Brownovog gibanja

```

1 # Simulacija geometrijskog Brownovog gibanja
2
3 mi <- 0.2
4 sigma <- 1
5
6 x0 <- 35
7 delta <- 2^(-10)
8 seg <- seq(0, 1, by = delta)
9
10 X <- rep(0, 2^10 + 1)
11 X[1] <- x0
12 E <- rep(0, 2^10 + 1)
13 E[1] <- x0
14 M <- rep(0, 2^10 + 1)
15 M[1] <- x0
16 Bt <- 0

```

```
17 for (i in 1:2^10){
18   B <- rnorm(1, 0, sqrt(delta))
19   E[i+1] <- E[i] + mi*E[i]*delta + sigma*E[i]*B
20   M[i+1] <- M[i] + mi*M[i]*delta + sigma*M[i]*B +
21   0.5*sigma*M[i]*sigma*(B^2-delta)
22   Bt <- Bt + B
23   X[i+1] <- x0*exp((mi - sigma^2/2)*seg[i+1] + sigma * Bt)
24 }
25
26 plot(seg, X, type = 'l', lwd = 2, col = 'darkgreen', xlab = "t")
27 points(seg, E, type = 'l', lwd = 2, col = 'mediumorchid2')
28 points(seg, M, type = 'l', lwd = 2, col = 'red')
29
30 legend("topright", legend = c("egzaktno rjesenje", "Eulerova
31 metoda", "Milsteinova metoda"), lty = c(1, 1, 1), lwd = 2,
32 col = c('darkgreen', 'mediumorchid2', 'red'))
```


Poglavlje 5

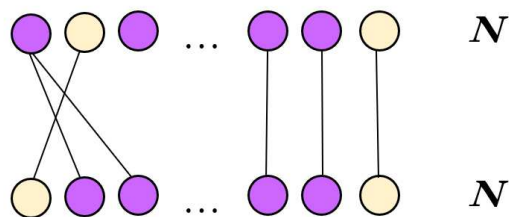
Wright-Fisherove difuzije

Familija Wright-Fisherovih difuzija skup je često korištenih evolucijskih modela kojima se opisuje promjena udjela $X_t \in [0, 1]$ alela, odnosno varijante nekog gena, u velikoj populaciji.

5.1 Od genetike do difuzije

Neutralni Wright-Fisherov model (skraćeno: W-F model) opisan je na sljedeći način. Populacija od N gena razvija se u diskretnim generacijama. $(k + 1)$ -a generacija formirana je iz k -te generacije nasumičnim odabirom N gena, pri čemu su moguća ponavljanja.

Interpretacija: Svaki roditelj (gen iz k -te generacije) može imati 0 ili više djece (geni iz $(k + 1)$ -e generacije), dok svako dijete ima točno jednog roditelja kojeg bira uniformno.



Slika 5.1: Primjer dvije generacije u W-F modelu.

Recimo da se neki gen javlja u dva alela, A i a . Promotrimo udio alela A kroz generacije.

Budući da njegov udio u sljedećoj populaciji ovisi isključivo o potomcima i roditeljima iz trenutne generacije, Wright-Fisherov model primjer je Markovljevog lanca s diskretnim vremenom. Prijelazne vjerojatnosti možemo opisati na sljedeći način.

Za $i, j \in \{0, 1, \dots, N\}$, označimo s p_{ij} vjerojatnost da iz generacije s i alela A prijeđemo u

generaciju u kojoj j jedinki ima alel A .

Za prijelazne vjerojatnosti vrijedi

$$p_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j}.$$

Naime, biramo kojih j jedinki u sljedećoj generaciji ima alel A . Sve one s vjerojatnošću $\frac{i}{N}$ "biraju" roditelje iz prethodne generacije koji imaju alel A . Ostale jedinke s vjerojatnošću $\frac{N-i}{N}$ biraju roditelje iz prethodne generacije s alelom a .

Promjena broja alela [3]

Označimo sa Z_n broj jedinki s alelom A n -te generacije. Primijetimo da je, uvjetno na Z_n , varijabla Z_{n+1} jednako distribuirana kao zbroj $I_1 + \dots + I_N$, pri čemu su I_1, \dots, I_N nezavisne, jednakodistribuirane Bernoullijeve slučajne varijable koje za svaku jedinku u generaciji označavaju ima li alel A ili ne. Konkretno,

$$\mathbb{P}(I_k = 1 | Z_n) = \frac{Z_n}{N} \quad \text{i} \quad \mathbb{P}(I_k = 0 | Z_n) = 1 - \frac{Z_n}{N}.$$

Dakle, uvjetno na Z_n , Z_{n+1} je binomna slučajna varijabla s parametrima N i $\frac{Z_n}{N}$.

Promotrimo promjenu broja alela A iz n -te generacije u $(n+1)$ -vu generaciju. Naravno, smisleno je promatrati uvjetno na trenutno stanje, odnosno Z_n . Vrijedi da je, uvjetno na Z_n , očekivana vrijednost promjene $Z_{n+1} - Z_n$ jednaka nuli, dok je varijanca promjene $N\frac{Z_n}{N}\left(1 - \frac{Z_n}{N}\right)$. Raspišimo.

Neka je $k \in \{0, \dots, N\}$.

$$\begin{aligned} \mathbb{E}[Z_{n+1} - Z_n | Z_n = k] &= \mathbb{E}[Z_{n+1} | Z_n = k] - k = \\ &= N\frac{k}{N} - k = \\ &= 0, \end{aligned}$$

$$\begin{aligned} \text{Var}[Z_{n+1} - Z_n | Z_n = k] &= \mathbb{E}[(Z_{n+1} - Z_n)^2 | Z_n = k] = \\ &= \mathbb{E}[Z_{n+1}^2 | Z_n = k] - 2k\mathbb{E}[Z_{n+1} | Z_n = k] + k^2 = \\ &= \text{Var}[Z_{n+1} | Z_n = k] + \mathbb{E}[Z_{n+1} | Z_n = k]^2 - 2k\mathbb{E}[Z_{n+1} | Z_n = k] + k^2 = \\ &= N\frac{k}{N}\left(1 - \frac{k}{N}\right) + \left(N\frac{k}{N}\right)^2 - 2k\left(N\frac{k}{N}\right) + k^2 = \\ &= N\frac{k}{N}\left(1 - \frac{k}{N}\right). \end{aligned}$$

Dakle, $\mathbb{E}[Z_{n+1} - Z_n | Z_n] = 0$ i $\text{Var}[Z_{n+1} - Z_n | Z_n] = N\frac{Z_n}{N}\left(1 - \frac{Z_n}{N}\right)$.

Primijetimo da promjena broja alela A iz generacije u generaciju ovisi samo o veličini ukupne populacije i o broju tog alela u prethodnoj generaciji, koja god ona po redu bila. Fiksirajmo, stoga, $n \in \mathbb{N}$, te promotrimo ponašanje promjene broja alela A iz n -te generacije u sljedeću s povećanjem veličine populacije N . Smisleno je promatrati uvjetno na Z_n , odnosno trenutno stanje u generaciji.

Sukladno razmatranju s početka odjeljka, promjena broja alela A uvjetno na Z_n može se zapisati kao $\Delta_N := I_1 + \dots + I_N - Z_n$, $N \in \mathbb{N}$. Tada je, uvjetno na Z_n , $(\Delta_N)_{N \in \mathbb{N}}$ niz nezavisnih modificiranih binomnih slučajnih varijabli $B(\frac{Z_n}{N})$, čije su vrijednosti pomaknute za $-Z_n$.

Prema De Moivre-Laplaceovom centralnom graničnom teoremu, vrijedi

$$\frac{\Delta_N}{\sqrt{N \frac{Z_n}{N} \left(1 - \frac{Z_n}{N}\right)}} \xrightarrow{d} N(0, 1),$$

što znači da je za velike vrijednosti N distribucija promjene broja alela A $Z_{n+1} - Z_n$ iz generacije u generaciju, uvjetno na Z_n , aproksimativno normalna s očekivanjem 0 i varijancom $N \frac{Z_n}{N} \left(1 - \frac{Z_n}{N}\right)$.

Udio alela

Problematično je kada prostor stanja može biti neograničen. Iz tog je razloga, umjesto broja alela A , zgodnije promatrati udio tog istog alela u generaciji.

Definiramo novu slučajnu varijablu $Y_n := \frac{Z_n}{N}$. Tada prostor stanja ograničavamo na segment $[0, 1]$, neovisno o vrijednosti N .

Analogno prethodnom razmatranju, promjena $Y_{n+1} - Y_n$ udjela alela A iz n -te generacije u $(n + 1)$ -vu generaciju za velike vrijednosti N uvjetno na Y_n približno je normalno distribuirana s očekivanjem 0 i varijancom $Y_n(1 - Y_n)\frac{1}{N}$.

Drugim riječima, uvjetno na Y_n , promjena $Y_{n+1} - Y_n$ za velike je N približno distribuirana kao $\sqrt{Y_n(1 - Y_n)}W$, pri čemu je $W \sim N(0, \frac{1}{N})$.

Možemo primijetiti da promjena udjela alela u populaciji kroz generacije izrazito podsjeća na definiciju Brownovog gibanja.

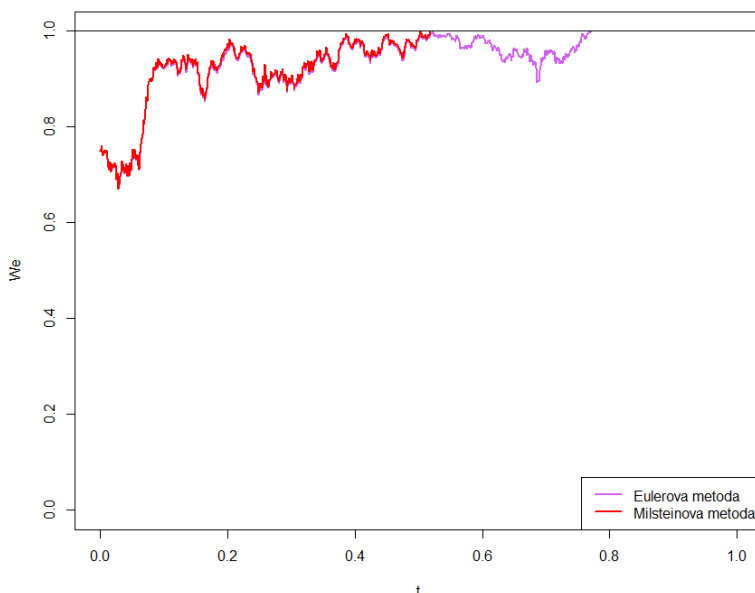
Na temelju prethodnog razmatranja te odjeljka o Brownovom gibanju kao limesu slučajnih šetnji, zaključujemo da je $Y_{\lfloor Nt \rfloor} = \frac{1}{N}Z_{\lfloor Nt \rfloor}$ distribuirano približno kao X_t , pri čemu je $X = (X_t : t \geq 0)$ slučajni proces koji zadovoljava stohastičku diferencijalnu jednadžbu

$$dX_t = \sqrt{X_t(1 - X_t)} dB_t. \quad (5.1)$$

Proces X zovemo *neutralnom Wright-Fisherovom difuzijom*.

5.2 Simulacija Wright-Fisherovih difuzija

Poput svakog neprekidnog procesa, i difuzija se može samo aproksimirati. Ipak, najveći je problem što ne postoji zatvorena forma rješenja stohastičke diferencijalne jednačbe (5.1), stoga ne možemo jednostavno provesti Monte Carlo simulaciju. Iako se ta prepreka za većinu procesa može zaobići koristeći numeričke metode aproksimacije poput Eulerove i Milsteinove iz Odjeljka 4.1, kod Wright-Fisherovih difuzija situacija je složenija. Budući da njima modeliramo *udjele*, svaka slučajna varijabla X_t mora upasti u interval $[0, 1]$. Međutim, u Eulerovoj i Milsteinovoj metodi nemamo način da spriječimo "iskakanje" varijabli izvan dozvoljenih vrijednosti.



Slika 5.2: Aproksimacije dobivene pomoću obje metode iskaču iz $[0, 1]$.

Naravno, to nije problem ukoliko se u praksi ne događa često. Za potrebe ovog rada, obje su metode provedene 10000 puta te je provjereno koliko se često dogodilo da pojedina slučajna varijabla poprimi vrijednost van segmenta $[0, 1]$.

Uzet je vremenski interval $[0, 1]$ te ekvidistantna subdivizija s korakom 2^{-10} kao u Primjeru 4.1.1. Iz SDJ (5.1) slijedi

$$a(X_t) = 0, \quad b(X_t) = \sqrt{X_t(1 - X_t)}, \quad t \in [0, 1].$$

Provedene su tri inačice eksperimenta. U prvoj je početni udio promatranog alela $\frac{1}{2}$, što dalje od krajnjih točaka segmenta. U drugoj je inačici udio $\frac{3}{4}$, dok je u trećoj $\frac{1}{4}$. Kôd simulacija, uz početnu vrijednost $\frac{1}{2}$, naveden je u nastavku.

Varijabla `br` označava u koliko se simulacija svake metode, od ukupno 10000, pojavila vrijednost izvan dozvoljenog segmenta. U Tablici 5.1 vidljivo je da su, ukoliko na početku pola populacije ima traženi alel te simuliramo jednom od navedenih numeričkih aproksimacijskih metoda, vrijednosti iskočile u više od 45% slučajeva. Iako je to prilično veliki postotak, situacija je još gora ukoliko se približimo bilo kojoj granici segmenta $[0, 1]$. Bilo da je početni udio alela u populaciji 0.25 ili 0.75, gotovo 60% simulacija u svakoj metodi iskoči iz dozvoljenog segmenta.

```

1  # Eulerova metoda
2
3  br <- 0
4  for (j in 1:10000){
5    We <- rep(0, 2^10 + 1)
6    We[1] <- 0.5
7
8    for (i in 1:2^10){
9      if(We[i] > 1 || We[i] < 0){
10       br <- br + 1
11       break
12     }
13     B <- rnorm(1, 0, sqrt(delta))
14     b <- sqrt(We[i]*(1-We[i]))
15     We[i+1] <- We[i] + b*B
16   }
17 }
18
19 # Milsteinova metoda
20
21 br <- 0
22 for (j in 1:10000){
23   Wm <- rep(0, 2^10 + 1)
24   Wm[1] <- 0.5
25
26   for (i in 1:2^10){
27     if(Wm[i] > 1 || Wm[i] < 0){
28       br <- br + 1
29       break
30     }
31     B <- rnorm(1, 0, sqrt(delta))
32     b <- sqrt(Wm[i]*(1-Wm[i]))
33     b2 <- (1-2*Wm[i])/(2*sqrt(Wm[i]*(1-Wm[i])))
34     Wm[i+1] <- Wm[i] + b*B + 0.5*b*b2*(B^2-delta)
35   }
36 }

```

početna vrijednost	Euler br	Euler postotak	Milstein br	Milstein postotak
0.5	4506	45.06	4516	45.16
0.75	5869	58.69	5855	58.55
0.25	5896	58.96	5889	58.89

Tablica 5.1: Nedozvoljene vrijednosti u Eulerovoj i Milsteinovoj metodi

BIM metoda

Za najjednostavniju Wright-Fisherovu difuziju Henry Schurz je 1996. godine u [14] razvio takozvanu BIM metodu (engleski: *Balanced Implicit Method*) u kojoj pod određenim uvjetima simulirane aproksimacije ostaju unutar segmenta $[0, 1]$.

Kao i u Odjeljku 4.1, odaberimo $n \in \mathbb{N}$ koji određuje ekvidistantnu particiju $0 = t_0 < t_1 < \dots < t_n = 1$ segmenta $[0, 1]$ s korakom $\Delta t := \frac{1}{n}$. Pretpostavimo da je $x_0 \in \langle 0, 1 \rangle$. Radi lakše notacije, neka je $Y_i = X_{t_i}^n$ za $i \in \{0, \dots, n\}$. Aproksimaciju definiramo s:

$$\begin{aligned}
 Y_0 &:= x_0, \\
 Y_{i+1} &:= \frac{Y_i + \sqrt{Y_i(1-Y_i)} \Delta_i B + c(Y_i) |\Delta_i B| Y_i}{1 + c(Y_i) |\Delta_i B|} = \\
 &= Y_i + \frac{\sqrt{Y_i(1-Y_i)} \Delta_i B}{1 + c(Y_i) |\Delta_i B|}, \quad i \in \{0, \dots, n-1\}.
 \end{aligned}$$

Pritom je $c : [0, 1] \rightarrow \mathbb{R}$ odabrana kontrolna funkcija koja ovisi o sljedećem.

Fiksirajmo $\varepsilon \in \langle 0, \frac{1}{2} \rangle$. Budući da ne želimo da nam vrijednosti padnu izvan segmenta $[0, 1]$, aproksimirat ćemo W-F difuzije na segmentu $[\varepsilon, 1 - \varepsilon]$. Naravno, želimo što manji $\varepsilon \ll \frac{1}{2}$. U stvarnosti, 0 i 1 su apsorpcijska stanja, stoga ovako definiran proces, jednom kad u njih dođe, postaje stacionaran. S $\varepsilon \ll \frac{1}{2}$ u simulaciji oponašamo stacionarnost stvarnog procesa u okolini rubova.

O tom parametru ε ovisi odabir funkcije c . Naime,

$$c_\varepsilon(y) = c(y) := \begin{cases} \sqrt{\frac{1-\varepsilon}{\varepsilon}}, & y \in [0, \varepsilon) \\ \sqrt{\frac{1-y}{y}}, & y \in [\varepsilon, \frac{1}{2}) \\ \sqrt{\frac{y}{1-y}}, & y \in [\frac{1}{2}, 1 - \varepsilon) \\ \sqrt{\frac{1-\varepsilon}{\varepsilon}}, & y \in [1 - \varepsilon, 1] \end{cases} \quad (5.2)$$

Lema 5.2.1. [2] $\mathbb{P}(Y_{i+1} \in [0, 1] \mid Y_i \in [\varepsilon, 1 - \varepsilon]) = 1$.

Dokaz. Pretpostavimo najprije da je $Y_i \in [\varepsilon, \frac{1}{2}]$. Tada je

$$Y_{i+1} = \frac{Y_i + \sqrt{Y_i(1-Y_i)} (\Delta_i B + |\Delta_i B|)}{1 + \sqrt{\frac{1-Y_i}{Y_i}} |\Delta_i B|},$$

što je svakako nenegativno.

1° $\Delta_i B \leq 0$. Tada je

$$Y_{i+1} = \frac{Y_i}{1 + \sqrt{\frac{1-Y_i}{Y_i}} |\Delta_i B|} \leq Y_i \leq \frac{1}{2}.$$

2° $\Delta_i B > 0$. U tom slučaju $Y_{i+1} \leq 1$ ako i samo ako

$$1 + \sqrt{\frac{Y_i}{1-Y_i}} \Delta_i B \geq Y_i + \sqrt{Y_i(1-Y_i)} \cdot 2\Delta_i B, \quad \text{odnosno}$$

$$1 - Y_i \geq \Delta_i B \sqrt{\frac{Y_i}{1-Y_i}} (2Y_i - 1),$$

što je zadovoljeno jer je desna strana nejednakosti negativna, dok je lijeva pozitivna.

Neka je sada $Y_i \in [\frac{1}{2}, 1 - \varepsilon]$. U tom je slučaju

$$Y_{i+1} = \frac{Y_i + \sqrt{Y_i(1-Y_i)} \Delta_i B + \sqrt{\frac{Y_i}{1-Y_i}} |\Delta_i B| \cdot Y_i}{1 + \sqrt{\frac{Y_i}{1-Y_i}} |\Delta_i B|}.$$

1° $\Delta_i B \leq 0$. Tada je $Y_{i+1} \leq 1$ ako i samo ako vrijedi

$$Y_i + \left(\sqrt{Y_i(1-Y_i)} - Y_i \sqrt{\frac{Y_i}{1-Y_i}} \right) \Delta_i B \leq 1 - \sqrt{\frac{Y_i}{1-Y_i}} \Delta_i B, \quad \text{odnosno}$$

$$1 - Y_i \geq 2 \sqrt{Y_i(1-Y_i)} \Delta_i B.$$

Budući da je desna strana nejednakosti manja ili jednaka nuli, dok je lijeva pozitivna, tvrdnja vrijedi.

S druge strane, $Y_{i+1} \geq 0$ ukoliko

$$Y_i + \left(\sqrt{Y_i(1-Y_i)} - Y_i \sqrt{\frac{Y_i}{1-Y_i}} \right) \Delta_i B \geq 0, \quad \text{odnosno, kad se podijeli s } Y_i,$$

$$1 + \left(\sqrt{\frac{1-Y_i}{Y_i}} - \sqrt{\frac{Y_i}{1-Y_i}} \right) \Delta_i B \geq 0,$$

što vrijedi jer je izraz u zagradi, kao i $\Delta_i B$, negativan s obzirom da je u ovom slučaju $1 - Y_n \leq Y_n$.

2° $\Delta_i B > 0$. U tom je slučaju svakako $Y_{i+1} \geq 0$. S druge strane, $Y_{i+1} \leq 1$ ako i samo ako

$$Y_i + \left(\sqrt{Y_i(1-Y_i)} + Y_i \sqrt{\frac{Y_i}{1-Y_i}} \right) \Delta_i B \leq 1 + \sqrt{\frac{Y_i}{1-Y_i}} \Delta_i B, \quad \text{tj.}$$

$$\left(\sqrt{Y_i(1-Y_i)} - (1-Y_i) \sqrt{\frac{Y_i}{1-Y_i}} \right) \Delta_i B \leq 1 - Y_i,$$

što, naravno, vrijedi jer je lijeva strana nejednakosti jednaka nuli.

Sumiramo li sve slučajeve, dobivamo da, pod pretpostavkom da je $Y_i \in [\varepsilon, 1-\varepsilon]$, vrijedi $Y_{i+1} \in [0, 1]$, čime je dokaz završen. \square

Postavlja se pitanje je li ova simulacija uopće "dobra" te kako to mjerimo. Naravno, želimo da aproksimacija u konačnici konvergira prema rješenju stohastičke diferencijalne jednadžbe (5.1).

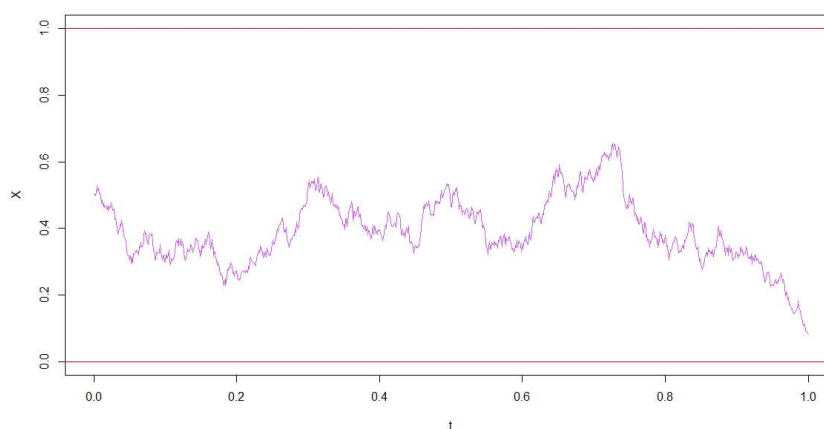
Definicija 5.2.2. *Rješenje (X_t^n) konvergira jako s redom $\beta > 0$ ako postoji konstanta $c > 0$ takva da za dovoljno mali Δt*

$$\sup_{i=0, \dots, n} \mathbb{E}|X_{t_i}^n - X_{t_i}| \leq c(\Delta t)^\beta.$$

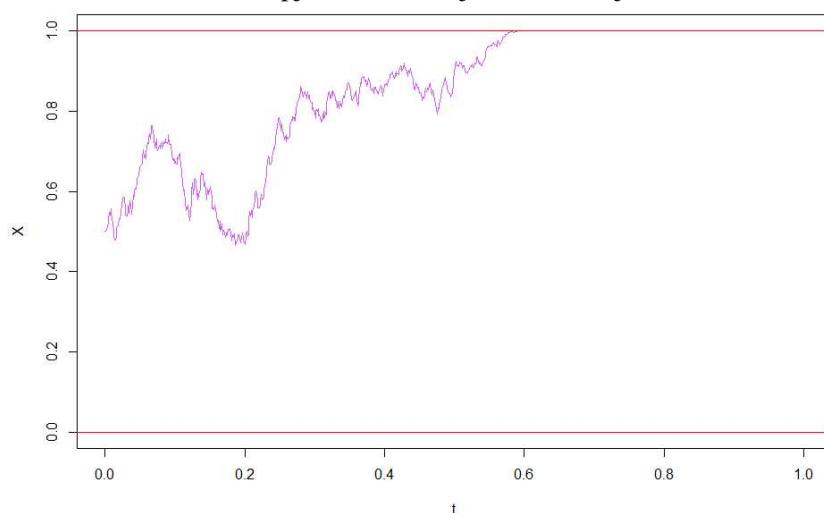
Napomena 5.2.3.

- *Primijetimo da veći red β znači bržu konvergenciju.*
- *Pokazuje se da Eulerova shema pod nekim uvjetima konvergira jako s redom $\frac{1}{2}$. S druge strane, Milsteinova shema je u tom smislu nešto bolja jer konvergira jako s redom 1. Naravno, govorimo općenito o procesima koji zadovoljavaju stohastičke diferencijalne jednadžbe bez ograničenja vrijednosti. [11]*
- *BIM metoda za rješavanje SDJ općenito konvergira jako s redom $\frac{1}{2}$ ukoliko su koeficijenti drifta i difuzije Lipschitz neprekidne funkcije. Međutim, to u slučaju Wright-Fisherove difuzije nipošto nije slučaj za koeficijent difuzije. [2]*

Primijetimo da u dokazu Leme 5.2.1 doista osiguravamo da, uz uvjet da trenutna varijabla upada u interval $[\varepsilon, 1 - \varepsilon]$, sljedeća ne ispadne iz segmenta $[0, 1]$. Ipak, ničime ne osiguravamo da ona također upadne u interval $[\varepsilon, 1 - \varepsilon]$, što nam već u idućem koraku može stvarati probleme. To je vidljivo i na Slici 5.3.



(a) Uspješna simulacija W-F difuzija



(b) Neuspješna simulacija W-F difuzija

Slika 5.3: BIM metoda simulacije neutralnih Wright-Fisherovih difuzija

Slično kao za Eulerovu i Milsteinovu shemu u prethodnom odjeljku, proveden je eksperiment u kojem se na 10000 ponavljanja promatralo koliko je puta vrijednost simulirane varijable iskočila iz segmenta $[\varepsilon, 1 - \varepsilon]$ te koliko je puta vrijednost iskočila iz dozvoljenog segmenta $[0, 1]$. Želimo utvrditi postotak slučajeva u kojima se simulacija uspije "izvući" iz rubnog intervala kada u nj uđe, a koliko često izleti i poprimi nedozvoljenu vrijednost.

Dakle, ukoliko je vrijednost varijable udaljena od rubova za manje od ε , u jednoj inačici eksperimenta povećavamo brojač `br_eps`, no ne prekidamo simulaciju. Tek ukoliko simulirana vrijednost iskoči iz dozvoljenog segmenta $[0, 1]$, povećavamo brojač `br` i prekidamo simulaciju. U drugoj inačici eksperimenta simulacija se prekida već nakon iskakanja iz segmenta $[\varepsilon, 1 - \varepsilon]$ kako bi se utvrdilo koliko se često to zapravo događa.

Sam programski kôd prve inačice eksperimenta naveden je u nastavku. Kao i za prethodne sheme, promatrane početne vrijednosti bile su $\frac{1}{2}$, $\frac{1}{4}$ i $\frac{3}{4}$. Vrijednost konstante ε je 2^{-12} .

```

1  # BIM metoda
2
3  x0 <- 0.5
4  delta <- 2^(-10)
5  seg <- seq(0, 1, by = delta)
6  X <- rep(0, 1025)
7  X[1] <- x0
8  eps <- 2^(-12)
9  br_eps <- 0
10 br <- 0
11
12 for (j in 1:10000){
13   for (i in 1:1024){
14     if (X[i] < 0 || X[i] > 1){
15       br <- br + 1
16       break
17     }
18     else if (X[i] < eps || X[i] >= 1 - eps){
19       br_eps <- br_eps + 1
20       c <- sqrt((1 - eps)/eps)
21     }
22     else if (X[i] >= eps && X[i] < 0.5){
23       c <- sqrt((1 - X[i])/X[i])
24     }
25     else {
26       c <- sqrt(X[i]/(1 - X[i]))
27     }
28     B <- rnorm(1, 0, sqrt(delta))
29     X[i+1] <- (X[i] + sqrt(X[i]*(1-X[i]))*B +
30              c*abs(B)*X[i])/(1 + c*abs(B))
31   }
32 }

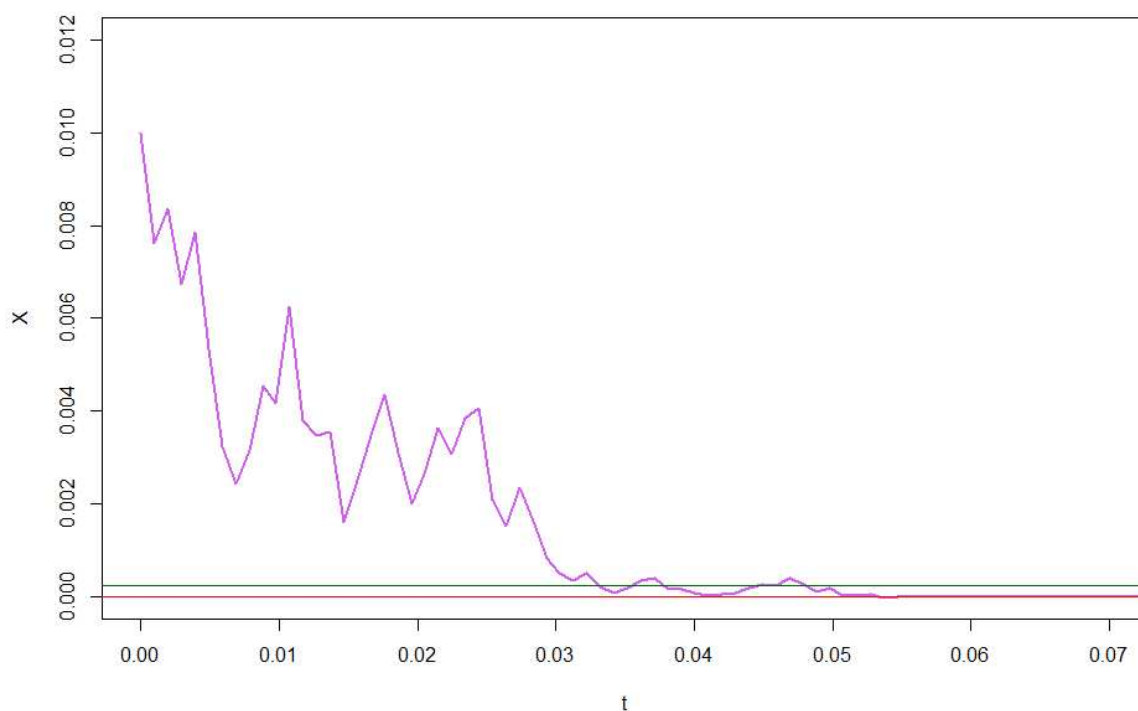
```

početna vrijednost	Izvan $[\varepsilon, 1 - \varepsilon]$ br_eps	Izvan $[0, 1]$ br	Izvan $[0, 1]$ postotak	Izvan $[\varepsilon, 1 - \varepsilon]$ prekid	Izvan $[\varepsilon, 1 - \varepsilon]$ postotak
0.5	18601	3136	31.36	3268	32.68
0.75	27953	4722	47.22	4837	48.37
0.25	27392	4630	46.30	4847	48.47

Tablica 5.2: Ispadanje iz segmenata u BIM metodi

U Tablici 5.2 navedene su brojke i postoci neuspješnih simulacija. Valja primijetiti da se otprilike u trećini simulacija s početnom vrijednosti najudaljenijoj od rubova dogodi da simulirana varijabla iskoči iz intervala $[\varepsilon, 1 - \varepsilon]$. Iz preostalih brojki u tablici može se iščitati da se tek u nešto više od 1% slučajeva simulacija uspije "izvući" i ostati unutar dozvoljenih vrijednosti ukoliko se javi varijabla čija je vrijednost u ε -okolini rubova. Naravno, sve ovisi o ε , postotak bi svakako bio veći rubni interval širi.

S druge strane, veliki brojevi u drugom stupcu tablice svjedoče kako se simulacija definitivno ne izvuče iz ε -okoline ruba u jednom koraku (ako se izvuče), već "živi na rubu" u više iteracija. To je vidljivo na Slici 5.4, koja prikazuje neuspješnu simulaciju W-F difuzije BIM metodom s malom početnom vrijednosti ($x_0 = 0.01$) kako bi se na grafu lijepo moglo vidjeti što se događa u ε -okolini točke 0.



Slika 5.4: Kretanje simulacije u ε -okolini točke 0

5.3 Wright-Fisherove difuzije s mutacijom

U stvarnosti, može se dogoditi da pojedini gen mutira pa od roditelja s alelom A nastane potomak s alelom a . Neka je $\theta_1 \in \langle 0, 1 \rangle$ vjerojatnost da alel a iz jedne generacije u drugu mutira u alel A . Analogno, neka je $\theta_2 \in \langle 0, 1 \rangle$ vjerojatnost da alel A iz jedne generacije u drugu mutira u alel a . Proces ostaje Markovljev, no mijenjaju se prijelazne vjerojatnosti. Naime, ako je u trenutnoj generaciji i jedinki s alelom A , vjerojatnost da će ih u idućoj biti j je

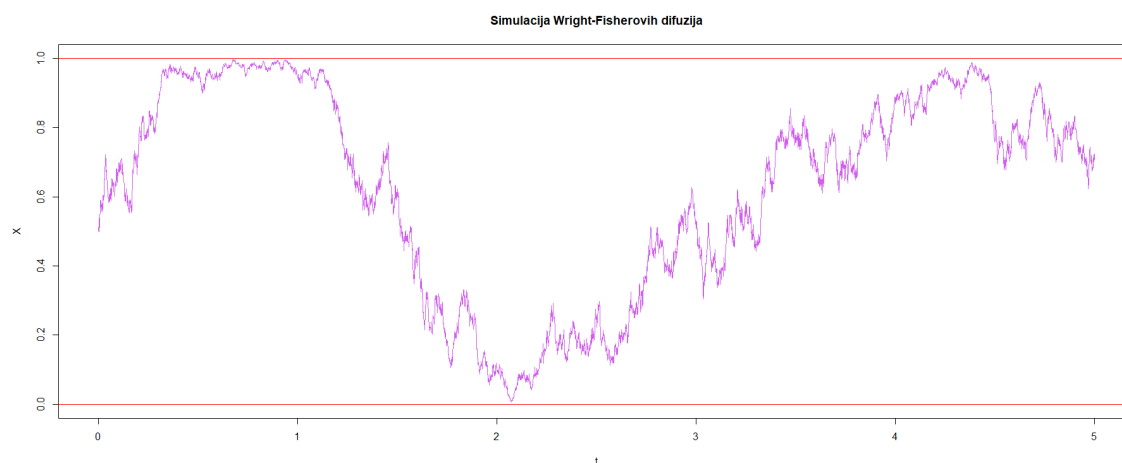
$$p_{ij} = \binom{N}{j} p_i^j (1 - p_i)^{N-j},$$

pri čemu je $p_i = \frac{i(1-\theta_1) + \theta_2(N-i)}{N}$, odnosno vjerojatnost da odaberemo roditelja s alelom A koji ne mutira u idućoj generaciji ili roditelja s alelom a koji u idućoj generaciji mutira.

Ispostavlja se da takav Wright-Fisherov model konvergira prema Wright-Fisherovoj difuziji zadanoj stohastičkom diferencijalnom jednačzbom

$$dX_t = (\theta_1(1 - X_t) - \theta_2 X_t) dt + \sqrt{X_t(1 - X_t)} dB_t, \quad X_0 = x_0 \in [0, 1]. \quad (5.3)$$

Za ovakav je proces razvijena metoda sa simulacijama koje se zadržavaju u segmentu $[0, 1]$ i koja konvergira jako s redom $\frac{1}{2}$. [2]



Slika 5.5: Simulacija W-F difuzije s mutacijom, uz $\theta_1 = \frac{1}{3}$, $\theta_2 = \frac{1}{4}$.

BISS metoda

Nadogradnjom BIM metode te razdvajanjem početne SDJ (5.3) na dvije

$$\begin{aligned}d\tilde{Y} &= \sqrt{\tilde{Y}(1 - \tilde{Y})} dB_t, \\d\hat{Y} &= (\theta_1 - (\theta_1 + \theta_2)\hat{Y}) dt,\end{aligned}$$

dobivamo takozvanu **BISS metodu** (engleski: *Balanced Implicit Split Step Method*).

Prva jednađbu rješavamo BIM metodom, dok se druga u ovom slučaju rješava Eulerovom metodom, gdje se za početni uvjet uzima rješenje prve jednađbe. Konkretno, ako je \tilde{Y}_{i+1} numerička aproksimacija rješenja prve jednađbe dobivena BIM metodom, a c kontrolna funkcija iz (5.2),

$$\begin{aligned}Y_0 &= x_0 \in \langle 0, 1 \rangle, \\ \tilde{Y}_{i+1} &= \frac{Y_i + \sqrt{Y_i(1 - Y_i)} \Delta_i B + c(Y_i) |\Delta_i B| Y_i}{1 + c(Y_i) |\Delta_i B|} = Y_i + \frac{\sqrt{Y_i(1 - Y_i)} \Delta_i B}{1 + c(Y_i) |\Delta_i B|}, \\ Y_{i+1} &= \tilde{Y}_{i+1} + (\theta_1 - (\theta_1 + \theta_2) \tilde{Y}_{i+1}) \Delta t = \\ &= Y_i + \frac{\sqrt{Y_i(1 - Y_i)} \Delta_i B}{1 + c(Y_i) |\Delta_i B|} + (\theta_1 - (\theta_1 + \theta_2) \left(Y_i + \frac{\sqrt{Y_i(1 - Y_i)} \Delta_i B}{1 + c(Y_i) |\Delta_i B|} \right)) \Delta t = \\ &= Y_i + (\theta_1 - (\theta_1 + \theta_2) Y_i) \Delta t + \frac{\sqrt{Y_i(1 - Y_i)} \Delta_i B}{1 + c(Y_i) |\Delta_i B|} (1 - (\theta_1 + \theta_2) \Delta t), \quad i \in \{0, \dots, n-1\}.\end{aligned}$$

Važno je primijetiti da kontrolna funkcija c u svojoj definiciji ovisi o odabranom ε koji je u klasičnoj BIM metodi fiksiran. U BISS metodi možemo ga mijenjati u svakom koraku, ovisno o vremenskom intervalu $t_{i+1} - t_i$. U teoriji, subdivizija segmenta $[0, T]$ na kojem promatramo W-F difuzije ne mora biti ekvidistantna, iako mi takvu promatramo radi jednostavnosti.

U svakom koraku, biramo ε kao

$$\varepsilon = \min\{\theta_1 \Delta t, \theta_2 \Delta t, 1 - \theta_1 \Delta t, 1 - \theta_2 \Delta t\} > 0.$$

Na taj način osiguravamo da je iduća numerička aproksimacija u intervalu $[0, 1]$, štoviše u $[\varepsilon, 1 - \varepsilon]$, što dokazujemo u sljedećem teoremu.

Teorem 5.3.1. *Pretpostavimo da je $x_0 \in [\varepsilon, 1 - \varepsilon] \subseteq [0, 1]$. Tada za svaki $i \in \{1, \dots, n\}$ Y_i leži unutar segmenta $[0, 1]$.*

Dokaz. Tvrđnju dokazujemo metodom matematičke indukcije.

Baza: Pretpostavimo da je $Y_0 = x_0 \in [\varepsilon, 1 - \varepsilon]$. Po Lemi 5.2.1 vrijedi da je $\bar{Y}_1 \in [0, 1]$.

$$Y_1 = \bar{Y}_1 + (\theta_1 - (\theta_1 + \theta_2)\bar{Y}_1)\Delta t = (1 - \Delta t(\theta_1 + \theta_2))\bar{Y}_1 + \theta_1\Delta t.$$

Budući da se radi o afinoj funkciji i znamo rubne vrijednosti \bar{Y}_1 , znamo u kojem se slučaju postižu ekstremi.

1° $1 - \Delta t(\theta_1 + \theta_2) > 0$. Tada je

$$(1 - \Delta t(\theta_1 + \theta_2)) \cdot 0 + \theta_1\Delta t \leq Y_1 \leq (1 - \Delta t(\theta_1 + \theta_2)) \cdot 1 + \theta_1\Delta t, \quad \text{odnosno}$$

$$\varepsilon \leq [\text{po definiciji}] \leq \theta_1\Delta t \leq Y_1 \leq 1 - \theta_2\Delta t \leq [\text{po definiciji}] \leq 1 - \varepsilon.$$

2° $1 - \Delta t(\theta_1 + \theta_2) < 0$. Sada vrijedi

$$(1 - \Delta t(\theta_1 + \theta_2)) \cdot 1 + \theta_1\Delta t \leq Y_1 \leq (1 - \Delta t(\theta_1 + \theta_2)) \cdot 0 + \theta_1\Delta t, \quad \text{odnosno}$$

$$\varepsilon \leq [\text{po definiciji}] \leq 1 - \theta_2\Delta t \leq Y_1 \leq \theta_1\Delta t \leq [\text{po definiciji}] \leq 1 - \varepsilon.$$

Dakle, $Y_1 \in [\varepsilon, 1 - \varepsilon] \subseteq [0, 1]$.

Pretpostavka: Pretpostavimo da je za $i \in \{1, \dots, n - 1\}$ $Y_i \in [\varepsilon, 1 - \varepsilon]$.

Korak: Tvrđimo da je $Y_{i+1} \in [0, 1]$.

Po Lemi 5.2.1 vrijedi da je $\bar{Y}_{i+1} \in [0, 1]$. Budući da je

$$Y_{i+1} = (1 - \Delta t(\theta_1 + \theta_2))\bar{Y}_{i+1} + \theta_1\Delta t,$$

analogno kao u bazi dobije se $Y_{i+1} \in [\varepsilon, 1 - \varepsilon] \subseteq [0, 1]$.

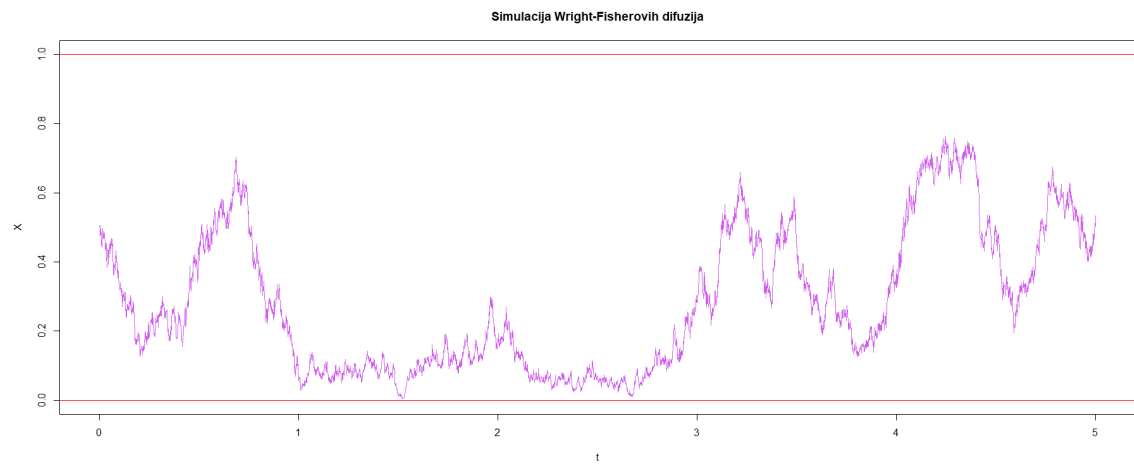
Po principu matematičke indukcije, tvrdnja je dokazana. □

Uz to što metoda osigurava da simulacije budu unutar segmenta $[0, 1]$, ona i konvergira prema Wright-Fisherovoj difuziji jako s redom $\frac{1}{2}$, što je pokazano u [2].

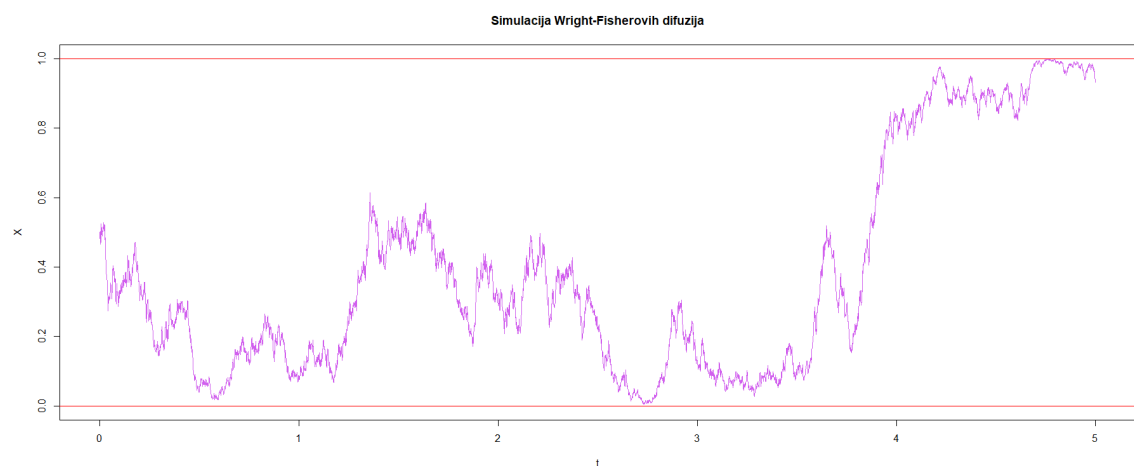
Na Slikama 5.5 i 5.6 vidljivo je kako, usprkos približavanju rubovima, simulacija BISS metodom ne iskače iz segmenta $[0, 1]$. U nastavku slijedi kôd simulacije BISS metode za rješenje stohastičke diferencijalne jednačbe

$$dX_t = \left(\frac{1}{3}(1 - X_t) - \frac{1}{4}X_t \right) dt + \sqrt{X_t(1 - X_t)} dB_t.$$

Difuziju promatramo na segmentu $[0, 5]$, a širina vremenskog intervala Δt je 2^{-10} . Uzeto je da 50% početne populacije ima alel A. Po definiciji konstante ε , ona iznosi $\frac{1}{4} \cdot 2^{-10} = 2^{-12}$.



(a) $\theta_1 = \frac{1}{4}, \theta_2 = \frac{1}{3}$



(b) $\theta_1 = \frac{3}{4}, \theta_2 = \frac{1}{3}$

Slika 5.6: BISS metoda simulacije Wright-Fisherovih difuzija s mutacijom

```

1 # BISS metoda
2
3 x0 <- 0.5
4 delta <- 2^(-10)
5 seg <- seq(0, 5, by = delta)
6 X <- rep(0, 5121)
7 X[1] <- x0
8 theta1 <- 0.75
9 theta2 <- 1/3
10 eps <- 2^(-12)

```

```

11 for (i in 1:5120){
12   B <- rnorm(1, 0, sqrt(delta))
13   if (X[i] < eps || X[i] >= 1 - eps){
14     c <- sqrt((1 - eps)/eps)
15   }
16   else if (X[i] >= eps && X[i] < 0.5){
17     c <- sqrt((1 - X[i])/X[i])
18   }
19   else {
20     c <- sqrt(X[i]/(1 - X[i]))
21   }
22   X[i+1] <- X[i] + (theta1 - (theta1+theta2)*X[i])*delta +
23   (sqrt(X[i]*(1-X[i]))*B)/(1 + c*abs(B))*(1 - (theta1+theta2)*delta)
24 }

```

5.4 Egzaktna simulacija Wright-Fisherovih difuzija

Kao što je spomenuto na samom početku poglavlja, simulacija Wright-Fisherovih difuzija sama po sebi je zahtjevna jer ne postoji zatvorena forma funkcije koja bi ju opisivala. Razvijene su, stoga, numeričke metode poput onih opisanih u prethodnom poglavlju. Međutim, navedene metode rade prilično dobro u slučaju da je simulacija udaljena od ruba. Međutim, što ako je iz nekog razloga baš potrebno promatrati ponašanje Wright-Fisherove difuzije na samom rubu intervala?

U [1] je opisana egzaktna metoda simulacije Wright-Fisherove difuzije koja rješava taj problem. Opravdanje epiteta "egzaktna" leži u tome što je metoda temeljena na razvoju funkcije prijelaza $f(x, \cdot; t)$ Wright-Fisherove difuzije pomoću funkcija prijelaza u tzv. procesu umiranja s beskonačno velikom početnom populacijom.

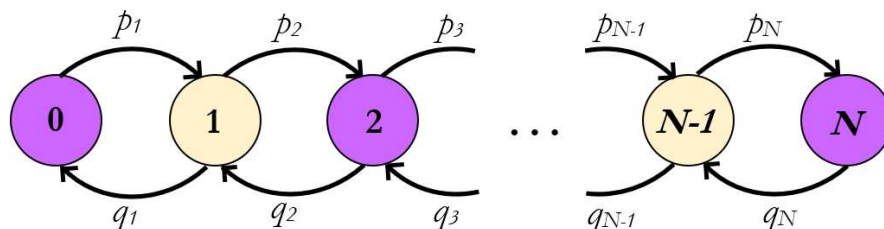
Čisti proces umiranja

Budući da se javlja u samoj simulaciji, u ovom odjeljku ukratko će biti opisan proces umiranja. Pogledajmo najprije općenitiji proces rađanja i umiranja.

Proces rađanja i umiranja je slučajni proces $(X_t)_{t \geq 0}$ u kojem promatramo koliko je jedinki u populaciji u danom trenutku. Radi se o Markovljevom procesu s neprekidnim vremenom.

Pretpostavimo da je u populaciji trenutno n jedinki. Vjerojatnost da će nakon Δt vremena u populaciji biti $n+1$ jedinka iznosi $p_n \Delta t + o(\Delta t)$, a da će biti $n-1$ jedinka $q_n \Delta t + o(\Delta t)$. Vjerojatnost da se broj neće promijeniti je $1 - (p_n + q_n) \Delta t + o(\Delta t)$.

Ukoliko je za svaki $n \in \mathbb{N}$ $p_n = 0$, odnosno u svakom se trenutku broj jedinki može samo smanjivati ili ostati isti, govorimo o **čistom procesu umiranja**.



Slika 5.7: Proces rađanja i umiranja

Egzaktna metoda

Ispostavlja se da je prijelazna funkcija gustoće Wright-Fisherove difuzije dana s:

$$f(x, y; t) = \sum_{m=0}^{\infty} q_m^{\theta}(t) \sum_{l=0}^m \mathcal{B}_{m,x}(l) \mathcal{D}_{2\theta_1+l, 2\theta_2+m-l}(y),$$

pri čemu $\mathcal{B}_{m,x}$ predstavlja funkciju gustoće binomne slučajne varijable $Y \sim B(m, x)$, dok je $\mathcal{D}_{2\theta_1+l, 2\theta_2+m-l}$ funkcija gustoće slučajne varijable s beta distribucijom uz navedene parametre. Uz to, $\theta = 2(\theta_1 + \theta_2)$, a $\{q_m^{\theta}(t) : m \in \mathbb{N}_0\}$ su funkcije prijelaza procesa umiranja $A_{\infty}^{\theta}(t)$ s beskonačnom početnom populacijom. Proces preciznije možemo opisati na sljedeći način: neka je $\{A_n^{\theta}(t) : t \geq 0\}$ čisti proces umiranja s vrijednostima u \mathbb{N} i početnom veličinom populacije n . Točnije, $A_n^{\theta}(0) = n$ g.s. i $p_m = \frac{m(m+\theta-1)}{2}$, $m = 1, \dots, n$. Tada su tražene funkcije prijelaza za proces s beskonačnom početnom populacijom dane s $q_m^{\theta}(t) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n^{\theta}(t) = m)$.

Jako neformalno približavanje dane formule glasi ovako. Čisti proces umiranja $A_{\infty}(t)$ predstavlja koliko "lozi" jedinki s alelom A (odnosno njezinih predaka) također ima alel A gledajući t vremena unatrag od sadašnjeg trenutka. Binomna razdioba vezana je uz jedinke iduće generacije, kao što je navedeno u prethodnim poglavljima ovog rada. Beta distribucija stacionarna je distribucija Wright-Fisherove difuzije s pozitivnim koeficijentima mutacije θ_1, θ_2 . Detalji u razvoju funkcije prijelaza nadilaze opseg ovoga rada te se zainteresirani čitatelj upućuje na [1], [6].

Čini se da imamo sve potrebne sastojke za simulaciju, problem je samo što ne postoji zatvorena forma funkcija prijelaza procesa umiranja $A_{\infty}^{\theta}(t)$. Međutim, sljedeći teorem iz [5] olakšava tu situaciju.

Teorem 5.4.1. *Neka je $\beta = \frac{1}{2}(\theta - 1)t$. Definiramo*

$$\mu^{(t,\theta)} := \frac{2\eta}{t},$$

$$(\sigma^{(t,\theta)})^2 := \begin{cases} \frac{2\eta}{t}(\eta + \beta)^2 \left(1 + \frac{\eta}{\eta + \beta} - 2\eta\right)\beta^{-2}, & \beta \neq 0, \\ \frac{2}{3t}, & \beta = 0, \end{cases}$$

pri čemu je

$$\eta = \begin{cases} \frac{\beta}{e^\beta - 1}, & \beta \neq 0, \\ 1, & \beta = 0. \end{cases}$$

Tada

$$\frac{A_\infty^\theta(t) - \mu^{(t,\theta)}}{\sigma^{(t,\theta)}} \xrightarrow{d} N(0, 1), \quad \text{kada } t \rightarrow 0.$$

Dakle, za dovoljno malene $t \geq 0$ proces $A_\infty^\theta(t)$ možemo aproksimirati normalnom $N(\mu^{(t,\theta)}, (\sigma^{(t,\theta)})^2)$ razdiobom.

Metoda 5.4.2. *Neka je x trenutno stanje procesa, a $t \geq 0$ vrijeme provedeno u stanju x prije prelaska u neko drugo (slučajno) stanje Y . Egzaktna metoda simulacije Wright-Fisherovih difuzija iz [1] koja prati oznake s početka ovog potpoglavlja dana je sljedećim algoritmom.*

1. Simuliraj $A_\infty^\theta(t) \sim N(\mu^{(t,\theta)}, (\sigma^{(t,\theta)})^2)$ i zaokruži na najbliži nenegativan cijeli broj.
2. Ako je $A_\infty^\theta(t) = m$, simuliraj $L \sim B(m, x)$.
3. Ako je $L = l$, simuliraj $Y \sim \text{Beta}(2\theta_1 + l, 2\theta_2 + m - l)$.

U nastavku je naveden kôd kojim je implementirana prethodna metoda. Kao i u prošlim slučajevima, dijametar subdivizije je 2^{-10} . Za početnu vrijednost x_0 te parametre θ_1 i θ_2 uzeto je nekoliko vrijednosti. Prikazi tih simulacija mogu se vidjeti na Slici 5.8.

```

1 # egzaktna metoda (Jenkins and Spano)
2
3 x0 <- 0.5
4 delta <- 2^(-10)
5 t <- seq(0, 5, by = delta)
6 X <- rep(0, 5121)
7 X[1] <- x0
8 theta1 <- 0.75
9 theta2 <- 1/3
10 theta <- 2*(theta1 + theta2)

```

```

11  for (i in 1:5120){
12      beta <- 0.5*(theta-1)*delta
13      if (beta == 0){
14          eta <- 1
15          var <- 2/(3*delta)
16      }
17      else{
18          eta <- beta/(exp(beta) -1)
19          var <- 2*eta/delta*(eta+beta)^2*(1 + eta/(eta+beta) -
20              2*eta)/(beta^2)
21      }
22      mi <- 2*eta/delta
23
24      A <- round(rnorm(1, mi, sqrt(var)))
25      L <- rbinom(1, A, X[i])
26      X[i+1] <- rbeta(1, 2*theta1+L, 2*theta2+A-L)
27  }

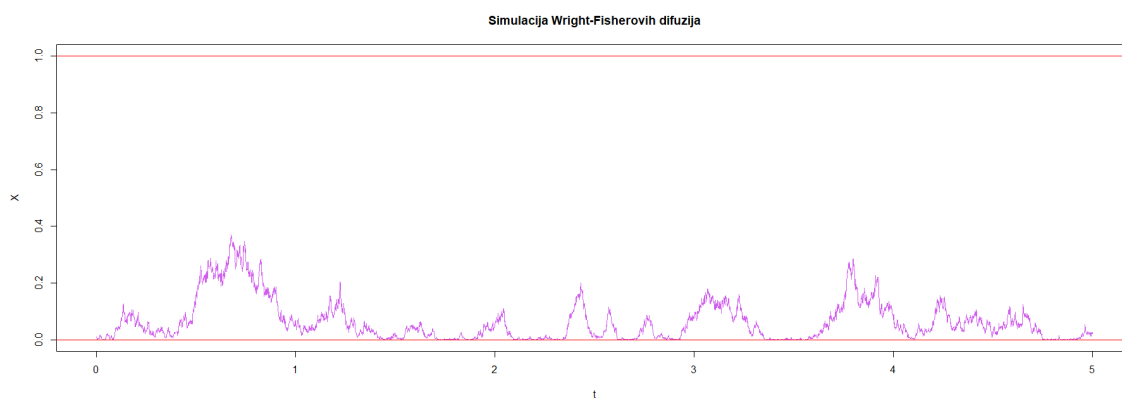
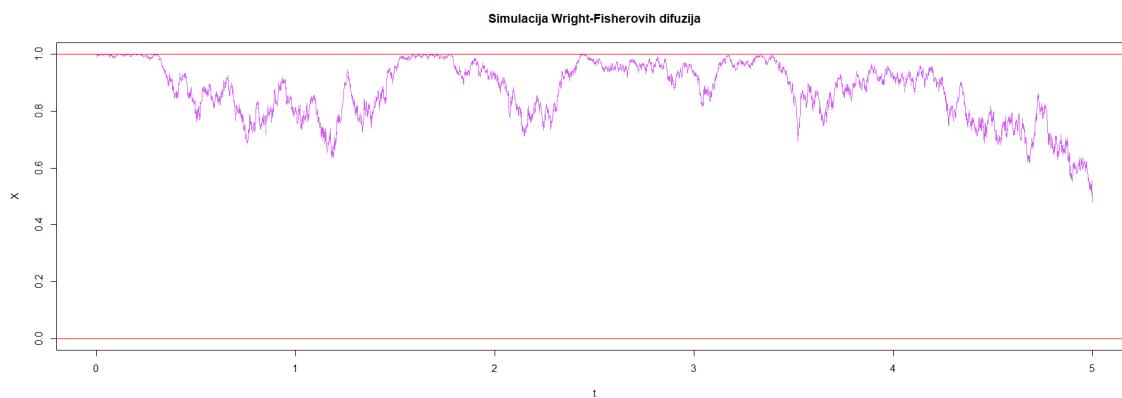
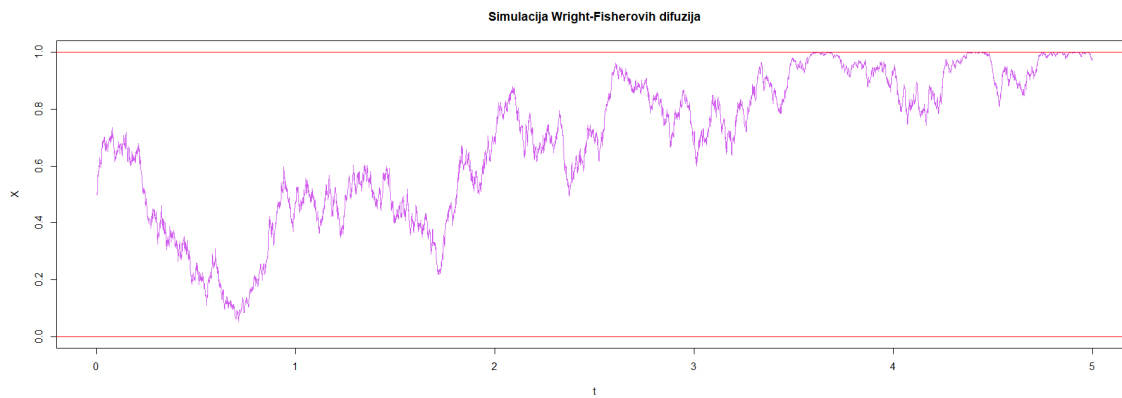
```

Valja primijetiti nekoliko stvari. Najprije, vrijednost Wright-Fisherove difuzije dobivamo simulacijom iz familije beta razdiobi, čime je osiguran ostanak difuzije unutar segmenta $[0, 1]$. Koliko se god blizu ruba nalazili (v. Slika 5.8 (b), (c)), difuzija neće poprimiti nedozvoljenu vrijednost.

Nadalje, za argument t procesa umiranja opisanog u Metodi 5.4.2 i Teoremu 5.4.1 ne uzimamo apsolutnu vrijednost varijable t koju vidimo na x -osi prikaza simulacije. Kao taj argument uzeta je vrijednost međuintervalu Δt , odnosno 2^{-10} jer se radi o prijelazu iz stanja X_t u stanje $X_{t+\Delta t}$ koje simuliramo ovom metodom.

Konačno - kao što je suptilno napomenuto u raspravi o razvoju prijelazne funkcije gustoće ovog procesa - kako bi ova metoda funkcionirala, važno je da parametri θ_1 i θ_2 u komponenti drifta oba budu pozitivni. Dakle, ova metoda nije pogodna za simulaciju neutralnih Wright-Fisherovih bez mutacije.

U ovome su radu obrađena dva osnovna oblika Wright-Fisherovih difuzija. Budući da je i sama genetika vrlo kompleksna, Wright-Fisherove difuzije koje su glavni alat za matematičko modeliranje procesa u genetici javljaju se i u složenijim formama. Neke su od metoda simulacija tih složenijih formi predstavljene u [1], no i u tim se slučajevima ograničavamo na strogo pozitivne θ_1 i θ_2 . Nedavno je, ipak, objavljen članak [12] u kojem je predstavljen EWF - C++ paket koji egzaktno simulira Wright-Fisherove difuzije s općenitim parametrima. Temeljen je na Metodi 5.4.2 te modificiran za difuzije s jednostranom mutacijom ili bez mutacije.



Slika 5.8: Egzaktna metoda simulacije Wright-Fisherovih difuzija s mutacijom

Bibliografija

- [1] Jenkins P. A. i D. Spanò, *Exact simulation of the Wright-Fisher diffusion*, The Annals of Applied Probability **27** (2017), 1478–1509.
- [2] C. E. Dangerfield, D. Kay, S. MacNamara i Burrage K., *A boundary preserving numerical algorithm for the Wright-Fisher model with mutation*, BIT Numerical Mathematics (2011), 283–304.
- [3] S. Evans, *Mathematical Models in Population Genetics III*, <https://youtu.be/f-BiaQrZ8bg?si=K8zM0sQPNbXsACj9>, snimka predavanja.
- [4] M. Giles, *Numerical Methods II, Lecture 11*, Oxford, materijali s predavanja z kolegija Monte Carlo Methods.
- [5] R. C. Griffiths, *Asymptotic line-of-descent distributions*, Journal of Mathematical Biology (1984).
- [6] R. C. Griffiths i D. Spanò, *Diffusion processes and coalescent trees*, Probability and Mathematical Genetics - Papers in Honour of Sir John Kingman (2010).
- [7] A. Gut, *Probability: A Graduate Course*, Springer, 2005.
- [8] C. Lee, *Stochastic Differential Equations*, MIT, materijali s predavanja iz kolegija Topics In Mathematics With Applications In Finance.
- [9] Geometric Brownian motion, *Geometric Brownian motion — Wikipedia, The Free Encyclopedia*, 2024., https://en.wikipedia.org/wiki/Geometric_Brownian_motion, pristupljeno: 13. lipnja 2024.
- [10] R. Mrazović i H. Planinić, *Markovljevi lanci - vježbe*, PMF-MO, nastavni materijali, 2023.
- [11] H. Planinić, *Stohastičke diferencijalne jednačbe i Black-Scholesov model*, PMF, materijali iz kolegija Financijski praktikum.

- [12] J. Sant, P. A. Jenkins, J. Koskela i D. Spanò, *EFW : simulating exact paths of the Wright–Fisher diffusion*, Bioinformatics (2023).
- [13] R. L. Schilling, *Brownian Motion*, De Gruyter, 2021.
- [14] H. Schurz, *Numerical Regularization for SDEs: Construction of Nonnegative Solutions*, Weierstrass Institute for Applied Analysis and Stochastics (1996).
- [15] K. Sigman, *Acceptance-Rejection Method*, 2007.
- [16] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry (Third Edition)*, Elsevier B. V., 2007.
- [17] Z. Vondraček, *Slučajni procesi*, PMF-MO, nastavni materijali, 2010.
- [18] V. Wagner, *Financijsko modeliranje 2*, PMF, predavanja iz kolegija Financijsko modeliranje 2.

Sažetak

Glavni predmet ovoga rada simulacije su Wright-Fisherovih difuzija.

Kao uvod u temu, definirani su slučajni procesi i navedene osnovne metode simulacija slučajnih varijabli i procesa: metoda inverzije, Box-Mullerova metoda i metoda prihvaćanja i odbacivanja.

Detaljnije je obrađeno Brownovo gibanje; dan je kratki povijesni pregled, dok je definicija približena kroz razmišljanje o "ubrzanom" slučajnoj šetnji. Naveden je primjer metode simulacije.

Budući da ne postoji zatvorena forma za Wright-Fisherove difuzije, već su one zadane stohastičkom diferencijalnom jednačbom, dan je pregled dviju metoda rješavanja istih - Eulerove i Milsteinove.

Konačno, uz motivaciju iz područja genetike, uvedene su Wright-Fisherove difuzije, s i bez mutacije, te su raspravljani problemi oko njihovih simulacija - posebno poprimanje nedozvoljenih vrijednosti u Eulerovoj i Milsteinovoj metodi te nedostatak zatvorene forme iz koje bi se jednostavno provele Monte Carlo simulacije. Obradene su dvije novije numeričke metode prikladnije za ovu vrstu procesa, BIM i BISS metoda. Najviše problema oko simulacije riješila je egzaktna metoda na samom kraju rada.

Svaka je metoda potkrijepljena primjerima, a algoritmi su implementirani u programskom jeziku **R**.

Summary

This thesis explores the simulation of Wright-Fisher diffusions.

As an introduction to the topic, random processes are defined, and the basic methods for simulating random variables and processes are listed: the inversion method, the Box-Muller method, and the acceptance-rejection method.

Brownian motion is discussed in more detail; a brief historical overview is provided, and the definition is approached through the concept of an "accelerated" random walk. An example of a simulation method is given.

Since there is no closed form for Wright-Fisher diffusions, and they are given by a stochastic differential equation, a review of two methods for solving SDEs is presented - Euler's and Milstein's method.

Finally, with motivation from the field of genetics, Wright-Fisher diffusions with and without mutation are introduced, and the issues surrounding their simulation are discussed - particularly the adoption of inadmissible values in the Euler's and Milstein's method, and the lack of a closed form from which Monte Carlo simulations could be easily conducted. Two newer numerical methods more suitable for this type of process, the BIM and BISS method, are discussed. Most of the simulation issues were resolved by the exact method provided at the end of the thesis.

Each method is supported by examples, and the algorithms are implemented in the programming language **R**.

Životopis

Rođena sam 14. siječnja 2001. godine u Čakovcu, gdje sam pohađala I. osnovnu školu Čakovec te kasnije prirodoslovno-matematički smjer Gimnazije Josipa Slavenskog Čakovec. Dobila sam nagradu za treći najbolji rezultat na Državnoj maturi 2019. godine. Kroz srednjoškolske i studentske dane oblikovala me Franjevačka mladež čija sam bila članica te vršila povjerene službe u vijećima.

U listopadu 2019. upisala sam Preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu, a tri godine kasnije nastavila obrazovanje na Diplomskom sveučilišnom studiju Matematička statistika. Od 2020. godine volontiram u udruzi *Mladi nadareni matematičari "Marin Getaldić"*, a od 2021. radim na fakultetu kao demonstrator u računalnim praktikumima. Tijekom studija bila sam studentski demonstrator iz sedam kolegija te sam dva puta dobila nagradu Matematičkog odsjeka za najuspješnije studente završnih godina studija. Uz kolegu Matea Dujića dobila sam *Rektorovu nagradu za individualni znanstveni rad u području prirodnih znanosti* za akademsku godinu 2021./2022. Autorica sam članka Faktorizacija u časopisu *Matematika i škola*.