UNIVERZITET U NOVOM SADU
PRIRODNO - MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I INFORMATIKU

Milana Radmanović

Stohastički procesi u genetici

-Master rad-

Mentor: Prof dr. Danijela Rajter - Ćirić

Novi Sad, 2023

Sadržaj

Predgovor

Genetika nam pruža ključne uvide u nasleđivanje, evoluciju i funkcionisanje živih organizama, pa je neiscrpan izvor interesovanja za konstantna istraživanja. Jasno da genetska nasleđivanja nisu uvek deterministički određena, stoga stohastički procesi igraju značajnu ulogu u oblikovanju genetske raznolikosti i evolucije.

Ovaj rad je posvećen razumevanju i proučavanju stohastičkih procesa u genetici. Rad će obuhvatiti pet ključnih modela koji ilustruju primene pojmova i teorija iz stohastičkih procesa kao što su Markovski lanci, martingali itd.

Videćemo da kod prva tri modela (Hardy - Weinberg, Wright - Fisher, Moran) posmatrana populacija teži da dostigne stabilno stanje i izgubi genetsku raznolikost. Kako je očuvanje genetske raznolikosti u populaciji veoma važno zbog toga što veća raznolikost može pružiti populaciji veću sposobnost da se prilagodi promenljivim uslovima sredine ili da se nosi sa štetnim efektima bolesti ili drugih stresora, bitno nam je da znamo koliko brzo neka populacija teži gubljenju genetske raznolikosti. Pomoću te informacije možemo "upravljati" populacijama posebno u očuvanju ugroženih ili retkih vrsta. Gubitak genetske raznolikosti, može dovesti do povećanja rizika od bolesti i smanjenja adaptabilnosti populacije. Stoga je proučavanje kako se genetska raznolikost održava ili gubi u populacijama igra ključnu ulogu u razumevanju evolucije i dugoročnog opstanka različitih organizama.

Pored pomenuta tri modela, od ključnog značaja za razumevanje evolucije i genetske promene u populacijama su koalescentna teorija i Galton - Warsonov proces grananja. Koalescentna teorija se fokusira na retrospektivno praćenje genetskih događaja unazad u vremenu, dok se proces grananja bavi trenutnim procesima reprodukcije i nasleđivanja u populacijama. Koalescentna teorija je posebno korisna za analizu genetskog stabla, koje je grafički prikaz kako su geni nasleđeni i kako se razvijaju kroz generacije. Ovaj pristup omogućava istraživanje genetske raznolikosti, efekata genetskog drifta i mnoge druge aspekte evolucije u populacijama. Galton - Watsonov proces grananja se često koristi za razumevanje širenja bolesti, populacija organizama, i drugih procesa.

Na samom početku rada u poglavlju *Osnovni pojmovi* će biti opisani svi relavantni pojmovi iz genetike i biologije koji će se spominjati u toku rada. Takođe, na slikovit način će biti opisani tipovi nasleđivanja. Dat je pregled osnovnih pojmova i teorema iz verovatnoće i stohastičke analize, ali je najviše obraćena pažnja na pojmove koji se koriste u toku rada.

U poglavlju *Hardy - Weinberg-ov model* pravimo razliku između genetske i genotipske učestalosti i na slikovit način se objašnjava pojam genetičkog drifta. Dokazana su dva osnovna tvrđenja za nasleđivanje telesnih gena i time dolazimo do zaključka da ovaj model opisuje stabilno stanje pop-

ulacije u kojem genetska raznolikost ostaje nepromenjena ako se ispunjavaju određeni uslovi. Ovaj model se često koristi za analizu genetske ravnoteže u populacijama.

U poglavlju pod nazivom *Moranov model* je obrađen teorijski model u populacijskoj genetici koji se koristi za proučavanje dinamike alelnih frekvencija u populacijama. Iz same postavke modela se vidi da je on relativno jednostavan model, ali služi kao početna tačka za razmatranje složenijih modela koji uključuju selekciju, mutacije i druge faktore evolucije. Biće pokazano da u slučaju posmatranja jednog lokusa sa dva alela koja se nasleđuju, populacija teži da postane homozigotna, međutim vreme koje je potrebno za to je veoma veliko.

Sledeće poglavlje je *Wright - Fisher model* koji je idejno veoma blizak Moranovom modelu, pa će slične stvari biti i posmatrane. U priču uvodimo i Markovske lance i povezujemo ih sa brojem jednog alela u posmatranoj generaciji. Takođe, biće reči o brzini dostizanja apsorbcionog stanja, pa to može biti jedan način za poređenje ovog modela sa Moranovim. Pored toga, videćemo neke zaključke do kojih se dođe daljom analizom ukoliko u priču uvedemo različite veličine populacije koju možemo posmatrati kao determinističku veličinu ili kao slučajnu promenljivu.

Nakon toga, prelazimo na poglavlje *Koalescentna teorija* čija je osnovna ideja da se može pratiti kako se aleli iz različitih jedinki u populaciji "kolektivno vraćaju" na zajedničkog pretka. To znači da se može odrediti koliko generacija unazad je potrebno da se aleli iz dve ili više jedinki "susretnu" i postanu zajednički. Naravno, u svakom trenutku, postoji slučajnost u tome koji aleli će biti preneti na sledeću generaciju. Pomoću koalescentne teorije, moguće je konstruisati genetsko stablo koje prikazuje evoluciju alela i identifikuje zajedničke pretke. Razdvojićemo diskretni i kontinuirano vremenski koalascent gde će biti spomenute veoma bitne slučajne promenljive kao što su: promenljiva koja meri vreme koje prođe do pronalaska pomenutog zajedničkog pretka (mereno u generacijama), visina koalescentnog stabla, ukupno grananje koalescentnog stabla.

Poslednji model o kome će biti reči je *Proces grananja*. Razdvojićemo dva veoma slična pristupa, prvi je Galton - Watson model koji sve čvorove grananja posmatra jednako i drugi, proces grananja sa višestrukim čvorovima. Kod Galton - Watson procesa grananja posmatraće se broj jedinki u jednoj generaciji kao slučajna promenljiva, za koju će biti dokazano da ima približno geometrijsku raspodelu. Biće pokazano koji su to trigeri za izumiranje ili ekspanzije populacije. U programskom jeziku "R" su urađene dve simulacije Galton – Watsonovog procesa grananja, jedna u kojoj će čitava populacija izumreti, i druga u kojoj će doći do ekpanzija određenih podpopulacija. Grafički je prikazano kako izgleda raspodela broja jedinki u generacijama što, očekivano, približno liči na geometrijsku raspodelu kao što je teorijski i bilo pokazano. Detaljno je opisano šta koja linija koda predstavlja i zašto je bitna i šta bi se desilo ako bi se neki uslov izostavio ili napisao drugačije. Kod procesa grananja sa višestrukim čvorovima, naglasak je na samoj primeni modela koji na primeru raka jajnika kod žena utvrđuje optimalni interval za sprovođenje preventivnih pregleda kako bi se izbegle smrtonosne posledice.

Osnovni pojmovi

1.1 Osnovni pojmovi u genetici

Ćelija je osnovna jedinica građe i funkcije svih živih bića. Ćelija se sastoji od:

- ćelijske membrane
- citoplazme
- organela
- jedra (nukleusa).

Ćelijska membrana obavija ćeliju, daje joj oblik, a kako je selektivno propustljiva uspostavlja kontakte sa drugim ćelijama i vanćelijskom sredinom i razmenjuje materije sa njima. Citoplazma predstavlja unutrašnji sadržaj ćelije u kome se nalaze organele. Ćelijske organele su odeljci za obavljanje različitih funkcija unutar ćelija kao što su sinteza proteina, ćelijsko disanje, lučenje hormona, kretanje ćelije itd. Jedro predstavlja najznačajniji deo ćelije i ima dve glavne funkcije: nosi genetske informacije i koordinira aktivnost ćelije. Jedro, pored ostalih stvari sadrži hromatin (u periodu ćelijskog ciklusa između dve deobe, a u toku koga se ćelija priprema za deobu, poznatiji kao *interfaza*). Tokom ćelijske deobe dolazi do kondezovanja hromatinskih vlakana tako da ona postaju samostalna telašca – hromozomi.

Hromozomi su končaste strukture u jedru, na kojima se nalaze geni. Svaki gen ima određeno mesto na hromozomu. Broj hromozoma je stalan i karakterističan za svaku biološku vrstu i naziva se kariotip. U određenim vrstama, hromozomi se javljaju u parovima i takve vrste se nazivaju diploidne. Pored diploidnih postoje i haploidne (hromozomi se pojavljuju singularno), triploidne i u opštem slučaju poliploidne vrste. Uglavnom ćemo razmatrati diploidne vrste.

Telesne (somatske) ćelije imaju diploidan (grč. diploos = dvostruk) broj hromozoma. Normalna telesna ćelija čoveka ima 46 hromozoma ili dve garniture po 23 hromozoma, pri čemu jedna garnitura potiče od majke, a druga od oca pa se tako obrazuje 23 para homologih hromozoma. Kariotip žene sadrži 23 homologa para hromozoma, od čega su 22 para autozomni (telesni) hromozomi a jedan par su polni XX hromozomi. Muški kariotip takođe ima 23 para hromozoma, ali je homologih 22 para autozomnih, dok su polni hromozomi heterologi (različiti) X i Y.

Polne ćelije ili gameti (kod čoveka su to spermatozoidi i jajna ćelija) sadrže upola manji broj hromozoma u odnosu na telesne ćelije, nazvan haploidan (grč. haploos = jednostruk). Broj hromozoma u polnim ćelijama čoveka je 23.

Geni se nalaze na hromozomima i to na tačno određenom mestu nazvanom genski lokus. Geni koji zauzimaju ista mesta na homologim hromozomima nazivaju se aleli. Aleli su različiti oblici jednog istog gena. Aleli jednog gena mogu međusobno da deluju jedan na drugi na tri osnovna načina, koji istovremeno predstavljaju i tipove nasleđivanja:

- dominantno-recesivno nasleđivanje
- nepotpuno dominantno nasleđivanje
- kodominantno nasleđivanje.

Navešćemo 3 primera koja ilustruju gore navedena tri tipa nasleđivanja:

Primer 1.

Razmatraćemo nasleđivanje gena za visinu kod graška. Gen koji određuje ovu karakteristiku ima dva alela, obeležićemo ih sa T i t, gde T utiče da biljka bude visoka, dok t utiče na to da biljka bude "patuljak". Sada biljka graška može da ima jednu od tri kombinacije na svom homologom paru hromozoma: TT, Tt i tt. Kombinacije TT i Tt rezultuju fizički visokim biljkama, dok jedino tt kombinacija daje "patuljke". Zaključujemo da iako imamo tri kombinacije, u stvarnosti imamo samo dva rezultata: visoku biljku i ona koja je "patuljak". Kako je T alel onaj koji određuje osobinu visine, njega zovemo dominantnim alelom, a t nazivamo recesivnim alelom.
Napomena: $Tt = tT$

Primer 2.

Gen koji određuje boju biljke zevalice ima dva alela, R za crvenu boju i r za belu boju. Kao i u primeru 1, imamo tri moguće kombinacije genotipa: RR, Rr, i rr. Primećeno je da kombinacija RR daje crvenu boju cvetova, rr daje belu boju, dok Rr daje rozu boju. Vidimo da 3 genotipa[1] rezultiraju sa 3 fenotipa[2], pa nemamo dominantan alel. U ovom slučaju kažemo da su aleli R i r kodominantni.

Primer 3.

Sada ćemo razmatrati način na koji se nasleđuju krvne grupe. Gen koji određuje krvnu grupu ima tri alela: O, A i B. Imamo šest mogućih genotipa: OO, OA, OB, AA, AB i BB. Međutim, postoje samo četiri kombinacije fenotipa: L^A (odgovara genotipovima OA, AA), L^B (odgovara genotipovima OB, BB), L^O (genotip OO) i L^{AB} (genotip AB). Ovde su A i B kodominantni aleli, dok je O recesivan u odnosu na A i B. Odatle proizilazi činjenica da genotip OA se manifestuje isto kao i genotip AA; OB kao i BB dok se genotip AB manifestuje u drugačiji fenotip, u L^{AB} grupu. Ideja iza ovakvog nasleđivanje krvnih grupa je sledeća. Svaka osoba pored antigena (na eritrocitima) ima i antitela u krvnoj plazmi. Antitela su proteini koji se bore protiv antigena, dok su antigeni hemijska jedinjenja čija prisutnost prisiljava telo da proizvodi antitela kako bi se borilo protiv ovih antigena. Postoje dve različite vrste antigena nazvane A i B koji mogu ili ne moraju biti prisutni u krvi. Osnovno pravilo je da organizam neće proizvesti antitela koja bi se borila protiv svojih sopstvenih antigena. Ova četiri tipa fenotipova opisana gore odgovaraju krvi koja nema nikakve antigene, ili ima samo antigen A, ili ima samo antigen B, ili ima i A i B antigene.

[1] **genotip** - genska konstitucija nekog organizma koja može da se odnosi na jedan par alela (uži smisao genotipa) ili celovitu naslednu osnovu (sve gene koje taj organizam poseduje), što predstavlja širi smisao genotipa.

[2] **fenotip** - u širem smislu: skup svih morfoloških i fizioloških svojstava po kojima se prepoznaje neki organizam i po čemu se razlikuje od drugih organizama. Uži smisao: kako se ispoljava samo jedna osobina.

U realnom svetu postoje mutacije, koje su glavni izvor genetičke raznolikosti. Na primer, razmotrimo autozomni gen sa dva alela A i a. Pretpostavimo da je genotip oca AA, a majke aa. Sledi da svaka jedinka koja je njihov potomak mora dobiti alel A od oca i alel a od majke, što znači da će biti genotip Aa. Međutim, u praksi postoji mala verovatnoća da se tokom formiranja gameta ili zigota, alel A može pretvoriti u a. Takva promena rezultiraće u potomstvu sa genotipom aa. Tako, alel A se može mutirati u alel a, a alel a se može mutirati u A.
Ponekad, alel mutira u nešto novo, neki drugi alel, koji trenutno ne postoji u populaciji. Ovo navodi na pomisao kako je mogući broj alela za gen teorijski veoma beskonačan, što zapravo nije slučaj jer su verovatnoće za nastanak ovakvih alela veoma, veoma male. Ako je ovaj novi alel dobar, potomci će širiti ovaj novi alel u populaciji. Ako je novi alel štetan, on će pre ili kasnije nestati iz populacije. Ponekad, ovaj novi alel može biti ne samo dobar, već i povoljan za preživljavanje, stoga će ga populacija dalje prenositi što ide u prilog evoluciji.

Modele sa mutacijama i uticajem prirodne selekcije (evolucije) nećemo razmatrati u ovom radu jer su veoma komplikovane za analizu, a osnovne zakonitosti svakog pojedinačnog modela se mogu pokazati i bez njihovog uključivanja.

1.2 Osnovni pojmovi stohastičkih procesa

Sa Ω ćemo označavati skup svih mogućih ishoda nekog eksperimenta, a sa $w_i \quad i = 1, 2, ...$ ćemo označavati elementarne događaje odnosno sve moguće ishode eksperimenta. Pod pojmom eksperiment podrazumevamo događaj čiji ishodi nisu deterministički određeni.

Definicija 1.2.1. *Proizvoljan podskup A skupa Ω naziva se slučajni događaj i on se sastoji od onih elementarnih događaja koji imaju svojstvo kojim se događaj A definiše.*

Definicija 1.2.2. *(aksiom σ-polja) Podskup $\mathcal{F}$ skupa $\mathcal{P}(\Omega)$ se zove σ-polje odnosno σ-algebra događaja iz Ω ako važi:*

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \;\Rightarrow\; \overline{A} \in \mathcal{F}$
3. *ako* $A_1, A_2, ... \in \mathcal{F} \;\Rightarrow\; \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.2.3. *(aksiom verovatnoće) Preslikavanje $P : \mathcal{F} \to [0, 1]$ takvo da važi:*

- $P(\Omega) = 1$
- *za $A_1, A_2, ... \in \mathcal{F}$ takve da je $A_i \cap A_j = \emptyset \;\; , i \neq j, \;\; i, j = 1, 2, ...$ važi:*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

zove se funkcija verovatnće (ili samo verovatnoća) na prostoru $(\Omega, \mathcal{F})$

Uređena trojka $(\Omega, \mathcal{F}, P)$ zove se prostor verovatnoća.

Borelova σ-algebra na $\mathbb{R}^n$ je najmanja σ-algebra koja sadrži sve otvorene podskupove od $\mathbb{R}^n$.

Definicija 1.2.4. *Preslikavanje $X : \Omega \to \mathbb{R}$ je slučajna promenljiva na prostoru verovatnoća $(\Omega, \mathcal{F}, P)$ ako za svaki Borelov skup $S \in \mathcal{B}(\mathbb{R})$ imamo da $X^{-1}(S) = \{w : X(w) \in S\} \in \mathcal{F}$.*

Definicija 1.2.5. *Slučajna promenljiva X je diskretna ako je njen skup slika (skup vrednosti) prebrojiv.*

Definicija 1.2.6. *Neka je X proizvoljna slučajna promenljiva na $(\Omega, \mathcal{F}, P)$. Preslikavanje $F_X : \mathbb{R} \to [0, 1]$ takvo da važi:*

$$F_X(\mathrm{x}) = P\{X < \mathrm{x}\} \qquad \mathrm{x}, \in \mathbb{R}$$

zove se funkcija raspodele slučajne promenljive X.

Definicija 1.2.7. *Slučajna promenljiva X je apsolutno-neprekidna ako postoji nenagativna, integrabilna funkcija $\varphi_X(\mathrm{x}), \quad -\infty < \mathrm{x} < \infty$, takvo da:*

$$P\{X \in S\} = \int_S \varphi_X(\mathrm{x}) d\mathrm{x}, \qquad \forall S \in \mathcal{B}(\mathbb{R})$$

Funkcija $\varphi_X(\mathrm{x})$ se zove funkcija gustine raspodele slučajne promenljive X.

Definicija 1.2.8. *Očekivanje slučajne promenljive X:*

- *kada je X diskretnog tipa je:*
$$E(X) = \sum \mathrm{x}_i P(X = \mathrm{x}_i)$$
i ono postoji akko suma apsolutno konvergira.

- *kada je X apsolutno-neprekidnog tipa:*
$$E(X) = \int_{-\infty}^{\infty} \mathrm{x}\varphi_X(\mathrm{x})dx$$
i ono postoji akko integral apsolutno konvergira.

Definicija 1.2.9. *Disperzija slučajne promenljive X je njen centralni momenat reda 2:*

$$D(X) = \sigma^2(X) = E\Big(\Big(X - E(X)\Big)^2\Big)$$

Definicija 1.2.10. *Neka je $(\Omega, \mathcal{F}, P)$ prostor verovatnoća i neka slučajne promenljive $A, B \in \mathcal{F}$. Neka je $P(B) > 0$. Verovatnoća događaja A pod uslovom da se realizovao događaj B se naziva uslovna verovatnoća, obeležava se sa $P(A \mid B)$ i jednaka je:*

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

Definicija 1.2.11. *Slučajne promenljive $X_1, X_2, ...$ su nezavisne akko važi:*

$$F_{(X_1,...,X_n)}(\mathrm{x}_1, ..., \mathrm{x}_n) = F_{X_1}(\mathrm{x}_1) \cdot ... \cdot F_{X_n}(\mathrm{x}_n)$$

Teorema 1.2.1. *Neke bitne osobine očekivanja:*

- $E(c) = c$ *gde je c konstanta*
- $E(cX) = cE(X)$
- *Ako slučajne promenljive $X_1, ..., X_n$ imaju očekivanja tada važi:*
$$E\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} E(X_i)$$
ako su pritom ove promenljive i nezavisne tada važi:
$$E\Big(\prod_{i=1}^{n} X_i\Big) = \prod_{i=1}^{n} E(X_i)$$

Teorema 1.2.2. *Neke bitne osobine disperzije:*

- $D(X) = E(X^2) - E^2(X)$
- $D(X) = 0 \Leftrightarrow X = c$ *gde je c konstanta*
- $D(cX) = c^2 D(X)$

- $D(X+c) = D(X)$
- *Ako su* $X_1, ..., X_n$ *nezavisne slučajne promenljive i ako imaju disperzije, tada važi:*

$$D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i)$$

Pomenućemo neke osnovne raspodele diskretnog i apsolutno-neprekidnog tipa sa svojim očekivanjima i disperzijama.

Diskretne raspodele:

1. Binomna $X:\ \mathcal{B}(n,p)$ raspodela

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

$$E(X) = np$$

$$D(X) = np(1-p)$$

2. Poasonova $X:\ \mathcal{P}(\lambda)$ raspodela

$$P\{X=k\} = \frac{\lambda^k}{k!} e^{-k}$$

$$E(X) = D(X) = \lambda$$

3. Geometrijska $X:\ \mathcal{G}(p)$ raspodela

$$P(X=k) = (1-p)^{k-1} p$$

$$E(X) = \frac{1}{p}$$

$$D(X) = \frac{1-p}{p^2}$$

Apsolutno-neprekidne raspodele:

1. Uniformna $X : \mathcal{U}(a,b)$ raspodela

$$\varphi_X(\mathrm{x}) = \begin{cases} \frac{1}{b-a} & \text{, ako } \mathrm{x} \in (a,b) \\ 0 & \text{, inače} \end{cases}$$

$$F_X(\mathrm{x}) = \begin{cases} 0 & \text{, ako } \mathrm{x} \leqslant a \\ \frac{x-a}{b-a} & \text{, ako } a < \mathrm{x} \leqslant b \\ 1 & \text{, ako } \mathrm{x} > b \end{cases}$$

$$E(X) = \frac{a+b}{2}$$

$$D(X) = \frac{(b-a)^2}{12}$$

2. Eksponencijalna $X : \mathcal{E}(\lambda)$ raspodela

$$\varphi_X(x) = \begin{cases} \lambda e^{-\lambda x} & , \text{ako } x \geqslant 0 \\ 0 & , \text{ako } x < 0 \end{cases}$$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & , \text{ako } x \geqslant 0 \\ 0 & , \text{ako } x < 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}$$

$$D(X) = \frac{1}{\lambda^3}$$

3. Normalna $X : \mathcal{N}(m, \sigma^2)$ raspodela

$$\varphi_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$$

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2t^2}} dt$$

$$E(X) = m$$

$$D(X) = \sigma^2$$

Definicija 1.2.12. *Slučajni (stohastički) proces $\{X(t), t \in I\}$ (označava se još i kao X_t) je familija realnih slučajnih promenljivih definisanih na istom prostoru verovatnoća $(\Omega, \mathcal{F}, P)$. Skup I ćemo zvati parametarski skup, a realni prostor $\mathbb{R}^d$ skup stanja procesa.*

Ukoliko je parametarski skup I prebrojiv, reč je o diskretnom slučajnom procesu (zovemo ga još nizom ili lancem slučajnih promenljivih), u suprotnom proces je neprekidan. Stohastički proces ima dve promenljive (t, w), gde $t \in [t_0, T]$ i $w \in \Omega$, ali je praksa da se w izostavlja iz zapisa.

Ako je $X_t, t \in [t_0, T]$ stohastički proces, za svako fiksirano t - X_t je slučajna promenljiva koja ima svoj zakon raspodele koji je određen odgovarajućom funkcijom raspodele. Ovi jednodimenzionalni zakoni raspodele nisu dovoljni za karakterizaciju stohastičkog procesa, stoga je neophodno znati i višedimenzionalne zakone raspodela, odnosno konačno-dimenzionalne raspodele procesa.

Definicija 1.2.13. *Konačno-dimenzionalne raspodele stohastičkog procesa $\{X_t, t \in [t_0, T]\}$ su date sa:*

$$F_t(x) = P\{X_t < x\}$$

$$F_{t_1 t_2}(x_1, x_2) = P\{X_{t_1} < x_1, X_{t_2} < x_2\}$$

$$\ldots$$

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = P\{X_{t_1} < x_1, \ldots, X_{t_n} < x_n\}$$

gde $t, t_1, \ldots, t_n \in [t_0, T]$; $x, x_1, \ldots, x_n \in \mathbb{R}^d$ i $n \geqslant 1$

Konačno dimenzionalne raspodele zadovoljavaju sledeće uslove:

1. uslov simetrije:
   Ako je $\{i_1, ..., i_n\}$ permutacija brojeva $1, ..., n$ tada je:

$$F_{t_{i_1},...,t_{i_n}}(x_{i_1}, ..., x_{i_n}) = F_{t_1,...,t_n}(x_1, ..., x_n).$$

2. uslov saglasnoti:
   Za $m < n$ i za proizvoljne $t_{m+1}, ..., t_n \in [t_0, T]$ važi:

$$F_{t_1,...,t_m,t_{m+1},...,t_n}(x_1, ..., x_m, \infty, ..., \infty) = F_{t_1,...,t_m}(x_1, ..., x_m).$$

Teorema 1.2.3. *(Kolmogorova fundamentalna teorema)*
Za svaku familiju funkcija raspodela koje zadovoljavaju uslove simetrije i saglasnosti, postoji prostor verovatnoća $(\Omega, \mathcal{F}, P)$ i na njemu definisan stohastički proces $\{X_t, t \in [t_0, T]\}$ koji ima date funkcije raspodela kao svoje konačno-dimenzionalne raspodele.

Definicija 1.2.14. *Srednja vrednost procesa $X_t, t \in [t_0, T]$ odnosno očekivanje procesa je preslikavanje $m_X : [t_0, T] \to \mathbb{R}$:*

$$m_X(t) = m(t) = E(X_t).$$

Definicija 1.2.15. *Autokovarijansna funkcija ili korelaciona funkcija stohastičkog procesa je:*

$$K_X(t,s) = K(t,s) = E\Big(\Big(X_t - m(t)\Big)\Big(X_s - m(s)\Big)\Big) = \\ = E(X_t X_s) - m(t)m(s).$$

Definicija 1.2.16. *Uzajamna korelaciona funkcija procesa X_t i Y_t je:*

$$K_{X,Y}(t,s) = E\Big(\Big(X_t - m_X(t)\Big)\Big(Y_s - m_Y(s)\Big)\Big).$$

Definicija 1.2.17. *Disperzija stohastičkog procesa X_t je:*

$$D_X(t) = D(t) = K_X(t,t) = K(t,t) = E(X_t^2) - (m(t))^2.$$

Definicija 1.2.18. *Koeficijent korelacije stohastičkog procesa X_t je:*

$$\varphi_X(t,s) = \varphi(t,s) = \frac{K_X(t,s)}{\sqrt{D_X(t)D_X(s)}} = \frac{K(t,s)}{\sqrt{D(t)D(s)}}.$$

Definicija 1.2.19. *Posmatrajmo stohastički proces $\{X_n, n = 0, 1, 2, ...\}$ sa konačnim ili prebrojivim skupom vrednosti. Niz slučajnih promenljivih sa istim skupom stanja $\{x_1, x_2, ...\}$ zove se lanac Markova ako za proizvoljno $r \in \mathbb{N}$, $n \geqslant k_1 > k_2 > ... > k_r$ važi takozvano Markovsko svojstvo:*

$$P\{X_n = x_n \mid X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, ..., X_{k_r} = x_{k_r}\} = P\{X_n = x_n \mid X_{k_1} = x_{k_1}\}.$$

tj. verovatnoća da se proces (sistem) nađe u stanju x_n u trenutku n zavisi samo od stanja u sadašnjem trenutku k_1, a ne od stanja u prošlim trenucima $k_2, ..., k_r$

Definicija 1.2.20. *Verovatnoća prelaza iz i-tog u j-to stanje u jednom koraku je:*

$$p_{i,j}^{n,n+1} = P\{X_{n+1} = \mathrm{x}_j \mid X_n = \mathrm{x}_i\}.$$

Ako gore navedene verovatnoće ne zavise od n, kažemo da je lanac (vremenski) homogen. Matrica prelaska za jedan korak je:

$$P = \left[p_{i,j}\right]_{i,j}.$$

Definicija 1.2.21. *Verovatnoća prelaska iz i-tog u j-to stanje u n koraka je:*

$$p_{i,j}(n) = P\{X_{m+n} = \mathrm{x}_j \mid X_m = \mathrm{x}_i\}.$$

Matrica prelaza za n koraka je:

$$P_n = \left[p_{i,j}(n)\right]_{i,j}.$$

Teorema 1.2.4. *(Jednačine Chapmen Kolmogorov-a)*
Za $i, j, m, n \geqslant 0$ važi:

$$p_{i,j}(n+m) = \sum_{k=0}^{\infty} p_{i,k}(n) p_{k,j}(m).$$

Gornja jednakost se u matričnom obliku može zapisati kao:

$$P_{n+m} = P_n \cdot P_m \quad \vee \quad P_m = P_n \cdot P_{m-n}, \;\; m > n$$

Primetimo da važi:

$m = 1 \quad \Rightarrow \quad P_1 = P$

$m = 2 \quad \Rightarrow \quad P_2 = P_1 \cdot P_1 = P \cdot P = P^2$

$m = 3 \quad \Rightarrow \quad P_3 = P_2 \cdot P_1 = P^2 \cdot P = P^3$

...

$P_n = P^n$

Sa $p_i(n)$ označavamo verovatnoću i-tog stanja u trenutku n. Početna verovatnoća i-tog stanja se označava sa $p_i(0)$. Pomoću jednakosti Chapmen Kolmogorov-a dobijamo da važi:

$$\mathbf{p}(k) = \mathbf{p}(0) \cdot P^k$$

Gde je $\mathbf{p}(k) = [p_1(k), ..., p_n(k)]$

Definicija 1.2.22.

- *Stanje x_j je dostižno iz stanja x_i ako postoje $n_0 \in \mathbb{N}$ tako da je $p_{i,j}(n_0) > 0$*
- *Stanje x_j je apsorbujuće ako je $p_{j,j} = 1$*
- *Stanje x_j je povratno ako $\exists n_0 \in \mathbb{N}$ takvo da $p_{j,j}(n_0) > 0$*

Definicija 1.2.23. *Nad prostorom verovatnoća $(\Omega, \mathcal{F}, P)$ niz sigma algebri $\mathcal{F}_1, \mathcal{F}_2, ...$ takvih da*

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_n \subset \mathcal{F}_{n+1} \subset ... \subset \mathcal{F}$$

zove se filtracija.

Definicija 1.2.24. *Kažemo da je niz slučajnih promenljivih $\xi_1, \xi_2, ...$ adaptiran (prilagođen) filtraciji $\mathcal{F}_1, \mathcal{F}_2, ...$ ako je ξ_n $\mathcal{F}_n$-merljivo za svako $n = 1, 2, ...$*

Definicija 1.2.25. *Niz $\xi_1, \xi_2, ...$ slučajnih promenljivih zove se diskretni martingal u odnosu na filtraciju $\mathcal{F}_1, \mathcal{F}_2, ...$ ako:*

1. *ξ_n je integrabilno za sve $n = 1, 2, ...$*
2. *Niz $\xi_1, \xi_2, ...$ je adaptiran filtraciji $\mathcal{F}_1, \mathcal{F}_2, ...$*
3. *$E(\xi_{n+1} \mid \mathcal{F}_n) = \xi_n$ skoro sigurno $n = 1, 2, ...$ - martingalsko svojstvo*

Teorema 1.2.5. *(Doob-ova teorema za konvergenciju martingala).*
Ako je $(X_n)_{n \geqslant 0}$ martingal sa $sup_n E|X_n| < \infty$, onda X_n konvergira sa verovatnoćom 1 ka slučajnom promenljivom X koja ima konačno očekivanje.

Hardy - Weinberg-ov model

Posmatrajmo populaciju od N diploidnih jedinki i lokus sa genom sa dva alela A i a. Znači imamo ukupno 2N ovih gena. Obeležimo sada sa N_1 broj jedinki sa AA genotipom, N_2 broj jedinki sa Aa genotipom, a sa N_3 broj sa aa. Učestalost genotipova je u stvari udeo jedinki sa određenim genotipom u čitavoj posmatranoj populaciji. Tako dobijamo:

$$f = \frac{N_1}{N} \quad g = \frac{N_2}{N} \quad h = \frac{N_3}{N}$$
$$\Rightarrow f + g + h = 1$$

Broj gena (alela) od iste vrste iznosi:

$$p = 2N_1 + N_2 \text{ za alel } A$$
$$q = N_2 + 2N_3 \text{ za alel } a$$

Vidimo da je učestalost gena, odnosno alela, A i a redom:

$$x = \frac{p}{2N} = \frac{2N_1 + N_2}{2N} = f + \frac{1}{2}\,g$$
$$y = \frac{q}{2N} = \frac{N_2 + 2N_3}{2N} = \frac{1}{2}\,g + h$$
$$\Rightarrow x + y = 1$$

Na primeru ćemo pokazati da populacije sa različitim genotipskim učestalostima mogu imati jednaku učestalost gena. Ako imamo populaciju od 20 jedinki od kojih je 10 tipa AA, a 10 sa aa genotipom. Vidimo da je genotipska učestalost:

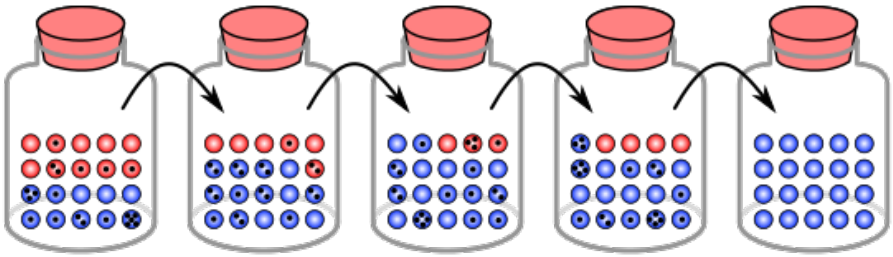
$$f = \tfrac{1}{2} \ , g = 0 \ , h = \tfrac{1}{2}$$

dok je učestalost alela : $x = y = \frac{1}{2}$.
Ako sada, u drugom slučaju, imamo opet populaciju od 20 jedinki gde njih 5 ima genotip AA, 10 Aa i 5 aa. Sledi da je učestalost alela jednaka kao i u prvom primeru, $x = y = \frac{1}{2}$ dok je genotipska učestalost

$$f = \tfrac{1}{4} \ , g = \tfrac{1}{2} \ , h = \tfrac{1}{4}$$

Videli smo da je ovo pojednostavljen prikaz činjenice da u populaciji beskonačne veličine, gde se jedinke razmnožavaju slučajno, genotipska učestalost ostaje konstantna. U konačnim populacijama ćemo videti da postoji genetički drift koji dovodi do eliminacije heterozigotnih genotipova (Aa). Osobine Markovih lanaca će se promeniti ukoliko uključimo postojanje mutacije gena. Napomena, ovaj model u svoju analizu ne uključuje postojanje mutacija i prirodne selekcije.

Genetički drift je promena u učestalosti postojeće varijante gena (alela) u populaciji usled slučajnog uzorkovanja organizama. Ako se od jedne velike populacije izdvoji manja grupa jedinki i oformi novu populaciju, ona ne mora biti ista već se čak može veoma razlikovati od matične populacije. Proces genetskog drifta može se ilustrovati korišćenjem 20 klikera koji predstavljaju jedinke u početnoj populaciji. Polovina klikera u tegli je crvena, a polovina plava, pri čemu svaka boja odgovara različitom alelu jednog gena u populaciji. U svakoj novoj generaciji, organizmi se slučajno razmnožavaju. Slučajnom reprodukcijom ćemo smatrati biranje jednog klikera iz originalne tegle i stavljanje novog kliker iste boje u novu teglu. Ovo je potomak originalnog klikera, što znači da originalni kliker ostaje u svojoj tegli. Ovaj proces se ponavlja sve dok se u drugoj tegli ne nađe 20 novih klikera. Druga tegla će sada sadržati 20 potomaka ili klikera raznih boja. Osim ako druga tegla ne sadrži tačno 10 crvenih klikera i 10 plavih klikera, došlo je do promene u učestalosti alela, odnosno do genetičkog drifta.
Slikovni prikaz ovog primera kroz par iteracija možemo videti na slici, gde crne tačke na klikeru predstavljaju onaj kliker, jedinku, koji je odabran da se preslika u drugu teglu, odnosno jedinku koja će se razmnožavati i dati potomak u narednoj generaciji.



Kroz generacije $n = 0, 1, 2...$ posmatramo učestalost genotipova, f_n, g_n, h_n, kao i učestalost alela, x_n, y_n.

2.1 Autozomni (telesni) geni

Teorema 2.1.1 (Hardi - Vajnbergov princip za autozomne hromozome)**.** *Posmatrajmo populaciju sa dva pola gde se jedinke razmnožavaju polno i slučajno. Pretpostavimo dalje, da su jedinke ove populacije diploidne i posmatrajmo autozomni (telesni) hromozom sa dva alela* A *i* a*. Tada važe sledeća tvrđenja:*

- *učestalost alela jednaka je iz generacije u generaciju*
- *bez obzira kako je izgledala inicijalna genotipska učestalost, učestalost genotipova* AA, Aa, aa *u prvoj generaciji (n=1) pa na dalje postaje stacionarna i određena samo početnom učestalošću alela A i a, odnosno genetskom učestalošću.*

Dokaz. Treba da dokažemo:

1. $x_n = x_0$ i $y_n = y_0$ $n = 1, 2, 3...$
2. $f_n = f_1$, $g_n = g_1$, $h_n = h_1$ $n = 2, 3...$
3. f_1, g_1 i h_1 zavise samo od x_0

Interesuje nas koja će biti učestalost genotipova u narednoj generaciji. Postavljamo uslov slučajnog razmnožavanja koje podrazumeva sledeće: verovatnoća da se pojavi $\alpha \times \beta$ genotip u narednoj generaciji gde je α genotip potekao od oca, a β od majke, jednak je proizvodu udela genotipova α u populaciji muškaraca i β u populaciji žena. Kako pričamo o autozomima gde oba roditelja imaju jednaku ulogu, nema potrebe da razdvajamo slučajeve $\alpha M \times \beta F$ i $\beta M \times \alpha F$. Za oba slučaja ćemo koristiti zapis $\alpha \times \beta$.

	genotip roditelja	v-ća ovakvog parenja	v-ća genotipa deteta
M_1:	$AA \times AA$	f_0^2	AA
M_2 :	$aa \times aa$	h_0^2	aa
M_3 :	$AA \times Aa$	$2f_0g_0$	$\frac{1}{2}AA + \frac{1}{2}Aa$
M_4 :	$aa \times Aa$	$2g_0h_0$	$\frac{1}{2}Aa + \frac{1}{2}aa$
M_5 :	$AA \times aa$	f_0h_0	Aa
M_6 :	$Aa \times Aa$	g_0^2	$\frac{1}{4}AA + \frac{1}{2}Aa + \frac{1}{4}aa$

Vidimo da su razmnožavanja tipa M_i $i = 1, ..., 6$ međusobno disjunktna, a njihova unija predstavlja čitav skup svih mogućih vrsta parenja, pa možemo da koristimo formulu totalne verovatnoće:

$$
\begin{aligned}
f_1 &= P(\textit{jedinka prve generacije ima } AA \textit{ genotip}) \\
&= \sum_{i=1}^{6} P(\textit{jedinka prve generacije ima } AA \textit{ genotip} \mid M_i) \cdot P(M_i) \\
&= f_0^2 \ + f_0 g_0 \ + \frac{1}{4} g_0^2 \\
&= \left(f_0 \ + \frac{1}{2} g_0 \right)^2 \\
&= x_0^2
\end{aligned}
$$

$$
\begin{aligned}
g_1 &= P(\textit{jedinka prve generacije ima } Aa \textit{ genotip}) \\
&= f_0 g_0 \ + g_0 h_0 \ + f_0 h_0 \ + \frac{1}{2} g_0^2 \\
&= 2 \left(f_0 \ + \frac{1}{2} g_0 \right) \left(\frac{1}{2} g_0 + h_0 \right) \\
&= 2 x_0 y_0
\end{aligned}
$$

$$
\begin{aligned}
h_1 &= P(\textit{jedinka prve generacije ima } AA \textit{ genotip}) \\
&= h_0^2 \ + h_0 g_0 \ + \frac{1}{4} g_0^2 \\
&= \left(h_0 \ + \frac{1}{2} g_0 \right)^2 \\
&= y_0^2
\end{aligned}
$$

Zaključujemo da je učestalost genotipova u bilo kojoj generaciji potpuno određena genetskom učestalošću prethodne generacije, kakva god bila genotipska učestalost prethodne generacije. Time je dokazan deo pod rednim brojem 3.

Dalje, učestalost alela A u prvoj generaciji je:

$$x_1 = f_1 + \frac{1}{2} g_1 = x_0^2 + x_0 y_0 = x_0(x_0 + y_0) = x_0$$

$$y_1 = h_1 + \frac{1}{2} g_1 = y_0^2 + x_0 y_0 = y_0(x_0 + y_0) = y_0$$

Kako su $x_1 = x_0$ i $y_1 = y_0$, znači da učestalost alela u novoj generaciji mora biti jednaka kao ona u prethodnoj, čime smo dokazali deo pod rednim brojem 1.
2. sledi iz $f_2 = x_1^2 = x_0^2$, nastavimo i za f_3, f_4,...,$f_n = x_0^2$. Analogno i za g_n, h_n □

U ovako formulisanom modelu sem što smo pretpostavili slučajno parenje među jedinkama, pretpostavili smo i sledeće stvari:

1. isključili smo mogućnost mutacija
2. isključena je prirodna selekcija (svim genotipovima je data jednaka verovatnoća preživljavanja i parenja; mogućnost stvaranja novog genotipa zavisi samo od zastupljenosti genotipova roditelja u čitavoj populaciji)
3. populacija je zatvorena, nema migracija
4. nema razmnožavanja između dve generacije. Imamo nultu generaciju koja proizvodi prvu itd.

Da bi naš sistem slučajnog razmnožavanja imao smisla, moramo pretpostaviti da je odnos muških u odnosu na ženske jedinke 1:1 i da je populacija beskonačno velika.

2.2 Nasleđivanje gena na X hromozomu

Neka se sada gen čije nasleđivanje posmatramo nalazi na X hromozomu (polni hromozom) i neka on ima dva alela A i a. Kako ženska jedinka nasleđuje i od majke i od oca X hromozom, ona će imati tri moguća genotipa AA, Aa i aa, dok će muška jedinka imati samo dva AY i aY (kraće A i a).

Teorema 2.2.1 (Hardi - Vajnbergov zakon za gene na X hromozomima)**.** *Neka je data diploidna, dvopolna, zatvorena populacija koja se razmnožava polno i slučajno. Pretpostavimo da nemamo mutacija i da su svi genotipovi imaju jednaku verovatnoću razmnožavanja i preživljavanja. Posmatrajmo prenos gena koji se nalazi na X polnom hromozomu, koji ima dva alela* A *i* a. *Ako je inicijalna učestalost muških genotipova* $p_0\,A + q_0\,a = 1$ *, a ženskih* $u\,AA + 2v\,Aa + w\,aa = 1$ *, tada sledi da će populacija dostići učestalost muških, odnosno ženskih jedinki redom:*

$$\alpha\,A + (1-\alpha)\,a$$

$$\alpha^2\,AA + 2\alpha(1-\alpha)\,Aa + (1-\alpha)^2\,aa + 1$$

gde je $\alpha = \frac{2}{3}(u+v) + \frac{1}{3}p_0$.

Dokaz. Analogno kao i u Teoremi 2.1.1 interesuje nas koja će učestalost genotipova biti u narednoj generaciji.

genotip roditelja	v-ća ovakvog parenja	v-ća genotipa deteta
$AA \times A$	up_0	$\frac{1}{2}AA + \frac{1}{2}A$
$AA \times a$	up_0	$\frac{1}{2}Aa + \frac{1}{2}A$
$Aa \times A$	$2vp_0$	$\frac{1}{4}AA + \frac{1}{4}Aa + \frac{1}{4}A + \frac{1}{4}a$
$Aa \times a$	$2vq_0$	$\frac{1}{4}Aa + \frac{1}{4}aa + \frac{1}{4}A + \frac{1}{4}a$
$aa \times A$	wp_0	$\frac{1}{2}Aa + \frac{1}{2}a$
$aa \times a$	wq_0	$\frac{1}{2}aa + \frac{1}{2}a$

Dakle, prva generacija ženskih jedinki će imati tri genotipa sa verovatnoćama:

$$p_0p_1\,AA + (p_0q_1 + q_0p_1)\,Aa + q_0q_1\,aa = 1$$

dok će muške jedinke imati dva genotipa sa verovatnoćama:

$$p_1\,A + q_1\,a = 1$$

gde su $p_1 = u + v$ i $q_1 = 1 - p_1$
Nastavimo ovaj postupak dalje, i dobijamo da će n-ta generacija ženskih odnosno muških jedinki, redom, izgledati:

$$p_np_{n-1}\,AA + (p_nq_{n-1} + q_np_{n-1})\,Aa + q_nq_{n-1}\,aa = 1$$

$$p_n\,A + q_n\,a = 1$$

gde su $p_n = p_{n-1}p_{n-2} + \frac{1}{2}(p_{n-1}q_{n-2} + q_{n-1} + p_{n-2})$ i $q_n = 1 - p_n$
Kada uvrstimo formulu za q_n u formulu za p_n dobijamo:

$$p_n = \frac{1}{2}(p_{n-1} + p_{n-2})$$

Šta će se desiti kada $n \to \infty$?

$$p_n = \frac{1}{2}(p_{n-1} + p_{n-2})$$

Neka je $\alpha_n = p_n - p_{n-1} \quad \Rightarrow \; p_n = \alpha_n + p_{n-1}$

$$\Rightarrow \alpha_n + p_{n-1} = \frac{1}{2}\, p_{n-1} + \frac{1}{2}\, p_{n-2}$$

$$\alpha_n = \frac{1}{2}\, p_{n-2} - \frac{1}{2}\, p_{n-1} = -\frac{1}{2}\, \alpha_{n-1}$$

$$\alpha_n = -\frac{1}{2}\alpha_{n-1} = -\frac{1}{2}\left(-\frac{1}{2}\right)\, \alpha_{n-2} = ... = \left(-\frac{1}{2}\right)^n\, \alpha_0$$

$$\Rightarrow \; p_n = \alpha_n + p_{n-1} = \alpha_n + \alpha_{n-1} + p_{n-2} = ... = \alpha_n + \alpha_{n-1} + ... + \alpha_0 + p_0 =$$

$$\left(-\frac{1}{2}\right)^n\, \alpha_0 + \left(-\frac{1}{2}\right)^{n-1}\, \alpha_0 + ... + \left(-\frac{1}{2}\right)^1\, \alpha_0 + \alpha_0 + p_0 = p_0 + \alpha_0 \sum_{k=0}^{n} \left(-\frac{1}{2}\right)^k$$

Pustimo da $n \to \infty$

$$\lim_{n\to\infty} p_n = p_0 + \alpha_0 \sum_{k=0}^{\infty} \left(-\frac{1}{2}\right)^k = p_0 + \frac{2}{3}\, \alpha_0 = \frac{2}{3}\, p_1 + \frac{1}{3}\, p_0 = \frac{1}{3}\, (u + v) + \frac{1}{3}\, p_0$$

□

Moranov model

Posmatrajmo haploidnu populaciju koja ima dva genotipa, A i a. Naš model ima sledeću postavku: iz trenutne generacije na slučajan način biramo jedinku koja se razmnožava bespolno, odnosno odabrana jedinka stvara jedinku nove generacije koja ima identičan genotip. U isto vreme, druga slučajno izabrana jedinka iz iste populacije umire. Vidimo da je veličina populacije konstantna. Uvodimo oznake:

X_n - broj A alela u n-toj generaciji

N - veličina populacije

Ako trenutna generacija ima i jedinki sa A genotipom, onda ćemo imati $N - i$ jedinki sa a genotipom, gde $i \in 1, ..., N$. Verovatnoće prelaza su:

$$p_{i,i-1} = p_{i,i+1} = \frac{i}{N}\left(1 - \frac{i}{N}\right)$$
$$p_{i,i} = \left(\frac{i}{N}\right)^2 + \left(1 - \frac{i}{N}\right)^2$$

Primetimo, $p_{i,j} = 0$ ukoliko $j \notin i - 1, i, i + 1$. Takođe, jasno je da je $p_{0,0} = p_{N,N} = 1$ zbog toga što jednom kada alel A ili a nestane iz populacije, ne može više biti vraćen. Stoga, stanja 0 i N su apsorbujuća.

$$\begin{aligned}
E(X_{n+1} \mid X_n = i) &= (i-1)p_{i,i-1} + ip_{i,i} + (i+1)p_{i,i+1} = \\
&= (i-1)\frac{i}{N}\left(1 - \frac{i}{N}\right) + i \cdot \left(\left(\frac{i}{N}\right)^2 + \left(1 - \frac{i}{N}\right)^2\right) + (i+1)\frac{i}{N}\left(1 - \frac{i}{N}\right) = \\
&= i \cdot \left(\left(\frac{i}{N}\right)^2 + 2 \cdot \frac{i}{N}\left(1 - \frac{i}{N}\right) + \left(1 - \frac{i}{N}\right)^2\right) = \\
&= i \cdot \left(\frac{i}{N} - 1 - \frac{i}{N}\right)^2 = \\
&= i
\end{aligned}$$

Kako je $E(X_{n+1} \mid X_n = i) = i$ sledi da je $(X_n)_{n \geqslant 0}$ martingal.

$$
\begin{aligned}
D(X_{n+1} \mid X_n = i) &= E(X_{n+1}^2 \mid X_n = i) - E^2(X_{n+1} \mid X_n = i) = \\
&= (i-1)^2 p_{i,i-1} + i^2 p_{i,i} + (i+1)^2 p_{i,i+1} - i^2 = \\
&= (i-1)^2 \frac{i}{N}\left(1 - \frac{i}{N}\right) + i^2 \cdot \left(\left(\frac{i}{N}\right)^2 + \left(1 - \frac{i}{N}\right)^2\right) + (i+1)^2 \frac{i}{N}\left(1 - \frac{i}{N}\right) - i^2 = \\
&= (i-1)^2 \frac{i}{N}\left(1 - \frac{i}{N}\right) + i^2 \cdot \left(\left(\frac{i}{N}\right)^2 + 1 - 2\frac{i}{N} + \left(\frac{i}{N}\right)^2 - 1\right) + (i+1)^2 \frac{i}{N}\left(1 - \frac{i}{N}\right) = \\
&= (i-1)^2 \frac{i}{N}\left(1 - \frac{i}{N}\right) - 2i^2 \cdot \frac{i}{N}\left(1 - \frac{i}{N}\right) + (i+1)^2 \frac{i}{N}\left(1 - \frac{i}{N}\right) = \\
&= \frac{i}{N}\left(1 - \frac{i}{N}\right)(i^2 - 2i + 1 - 2i^2 + i^2 + 2i + 1) = \\
&= 2 \cdot \frac{i}{N}\left(1 - \frac{i}{N}\right)
\end{aligned}
$$

Neka je $t = min(X_n = 0 \vee X_n = N)$ vreme fiksacije, odnosno vreme nakon kojeg sve jedinke imaju jednake genotipove. Kako je moguće dostići apsorbciono stanje iz bilo kog početnog, imamo da važi:

$$P(t < \infty \mid X_0 = i) = 1 \qquad \forall\, 0 < i < N$$

Neka je matrica $P = [p_{i,j}]_{i,j}$. Zanimaju nas njeni karakteristični koreni uz pomoć kojih ćemo videti kojom brzinom se populacija približava apsorbcionom stanju u kojem ostaje zauvek, odnoso u genetici, videćemo kojom brzinom će populacija postati homozigotna.

Lema 3.0.1. *Neka je data matrica $R = [r_{i,j}]_{i,j}$ gde su $r_{i,j} = \binom{i}{j}$, $0 \leqslant i, j \leqslant N$. Tada sledi da je $R^{-1} = S = [s_{i,j}]_{i,j}$ gde su vrednosti $s_{i,j} = (-1)^{i+j}\binom{i}{j}$.*

Dokaz.

$$[R \cdot S]_{i,j} = \sum_{k=0}^{N} \binom{i}{k} (-1)^{k+j} \binom{k}{j}$$

Treba da dokažemo: $R \cdot S = E$.

1. $i < j$
   Posmatrajmo proizvod $\binom{i}{k}\binom{k}{j}$:
   Ukoliko je $k < i$ sledi da je $k < j$ pa je $\binom{k}{j} = 0$.
   Ukoliko je $k \leqslant i$ sledi da je $\binom{i}{k} = 0$.

   $$\Rightarrow \quad \binom{i}{k}\binom{k}{j} = 0$$

2. $i = j$
   Proizvod koji posmatramo je sada: $\binom{i}{k}\binom{k}{i}$.
   Ukoliko je $k < i$ sledi da je $\binom{k}{i} = 0$.
   Ukoliko je $k > i$ sledi da je $\binom{i}{k} = 0$.
   Ukoliko je $k = i$ sledi da je $\binom{i}{k}\binom{k}{i} = 1$.

3. $i > j$
Posmatramo sumu: $\sum_{k=0}^{N} \binom{i}{k}(-1)^{k+j}\binom{k}{j}$.
Ukoliko ke $k < j$ sledi da je $\binom{k}{j} = 0$.
Ukoliko je $k > i$ sledi da je $\binom{i}{k} = 0$.

$$\Rightarrow \quad \sum_{k=0}^{N} \binom{i}{k}(-1)^{k+j}\binom{k}{j} = \sum_{k=j}^{i} \binom{i}{k}(-1)^{k+j}\binom{k}{j} =$$

$$= \sum_{k=j}^{i} \frac{i!}{(i-k)!\cdot k!} \cdot \frac{k!}{(k-j)!\cdot j!} \cdot \frac{(i-j)!}{(i-j)!} \cdot (-1)^{k+j} =$$

$$= \sum_{k=j}^{i} \frac{(i-j)!}{(k-j)!\cdot (i-k)!} \cdot \frac{i!}{(i-j)!\cdot j!} \cdot (-1)^{k+j} =$$

$$= \binom{i}{j} \cdot \sum_{k=j}^{i} \binom{i-j}{i-k} \cdot (-1)^{k+j}$$

Kako je $i > j$ i $i \geqslant k \geqslant j \quad \Rightarrow \quad i-k > i-j$

$$\Rightarrow \quad \binom{i-j}{i-k} = 0 \qquad \forall k \quad i \geqslant k \geqslant j$$

$$\Rightarrow \quad \sum_{k=0}^{N} \binom{i}{k}(-1)^{k+j}\binom{k}{j} = 0$$

Zaključujemo da matrica $R \cdot S$ ima sve nule sem na dijagonali gde su jedinice, odnosno $R \cdot S = E$. $\square$

Lema 3.0.2. *Matrica $R^{-1}PR = [a_{i,j}]$ ima sledeće vrednosti:*

$$a_{i,i} = 1 - \frac{i(i-1)}{N^2} \ , \ a_{i,i+1} = \frac{i(N-i)}{N^2} \ , \ a_{i,j} = 0 \ za \ j \neq i, i+1$$

Dokaz. Obeležićemo elemente matrica R^{-1}, P, R sa $s_{i,j}$, $p_{i,j}$, $r_{i,j}$ redom, gde $i, j = 0, ..., N$.
Kako je $P = [p_{k,l}] = 0$ za $l \notin k-1, k, k+1$, element (i,j) matrice $R^{-1} \cdot P \cdot R$, odnosno koristeći prethodnu lemu, matrice $S \cdot P \cdot R$ jednak je:

$$\sum_{k=0}^{N} s_{i,k} \cdot (p_{k,k-1}r_{k-1,j} + p_{k,k}r_{k,j} + p_{k,k+1}r_{k+1,j}) \qquad (\star)$$

Kako je $s_{i,k} = (-1)^{i+k}\binom{i}{k}$ vidimo da će $s_{i,k} = 0$ ukoliko je $k > i$.

Kako u jednačini $\star$ imamo vrednosti $r_{k-1,j}, r_{k,j}, r_{k+1,j}$ koje iznose $\binom{k-1}{j}, \binom{k}{j}, \binom{k+1}{j}$ redom, vidimo da ukoliko je $j > k+1$ vrednosti ovih faktorijela će biti 0. Sledi da $j \leqslant k+1$, odnosno $k \geqslant j-1$.
Imamo da važi $j-1 \leqslant i \ \Rightarrow \ j \leqslant i+1$
Dakle, naša suma iz $\star$ izgleda ovako:

$$\sum_{k=0}^{N} s_{i,k}\cdot(p_{k,k-1}r_{k-1,j}+p_{k,k}r_{k,j}+p_{k,k+1}r_{k+1,j}) =$$

$$= \sum_{k=j-1}^{i} s_{i,k}\cdot(p_{k,k-1}r_{k-1,j}+p_{k,k}r_{k,j}+p_{k,k+1}r_{k+1,j}) =$$

$$= \sum_{k=j-1}^{i}(-1)^{i+k}\binom{i}{k}\left[\frac{k}{N}\left(1-\frac{k}{N}\right)\binom{k-1}{j}+\binom{k}{j}\left(\left(\frac{k}{N}\right)^2+\left(1-\frac{k}{N}\right)^2\right)+\frac{k}{N}\left(1-\frac{k}{N}\right)\binom{k+1}{j}\right] =$$

$$= \sum_{k=j-1}^{i}(-1)^{i+k}\binom{i}{k}\left[\frac{k}{N}\left(1-\frac{k}{N}\right)\binom{k-1}{j}+\binom{k}{j}\left(\left(\frac{k}{N}+1-\frac{k}{N}\right)^2-2\cdot\frac{k}{N}\left(1-\frac{k}{N}\right)\right)+\frac{k}{N}\left(1-\frac{k}{N}\right)\binom{k+1}{j}\right]$$

$$= \sum_{k=j-1}^{i}(-1)^{i+k}\binom{i}{k}\frac{k}{N}\left(1-\frac{k}{N}\right)\left[\binom{k-1}{j}-2\cdot\binom{k}{j}+\binom{k+1}{j}\right]+\sum_{k=j-1}^{i}(-1)^{i+k}\binom{i}{k}\binom{k}{j}$$

Primetimo da važi:

$$\binom{k}{j}=\binom{k-1}{j-1}+\binom{k-1}{j} \quad \Rightarrow \quad \binom{k-1}{j}=\binom{k}{j}-\binom{k-1}{j-1}$$

$$\binom{k+1}{j}=\binom{k}{j}+\binom{k}{j-1}$$

$$\binom{k-1}{j}-2\cdot\binom{k}{j}+\binom{k+1}{j}=\binom{k}{j}-\binom{k-1}{j}-2\cdot\binom{k}{j}+\binom{k}{j}+\binom{k}{j-1}=\binom{k-1}{j-2}$$

Sada izraz iz ✯ ima oblik:

$$\sum_{k=j-1}^{i}(-1)^{i+k}\binom{i}{k}\frac{k}{N}\left(1-\frac{k}{N}\right)\binom{k-1}{j-2}+\sum_{k=j-1}^{i}(-1)^{i+k}\binom{i}{k}\binom{k}{j}$$

Može se pokazati da je taj izraz jednak nuli ukoliko je $j<i$. Znači da je $j\geqslant i$ u obe sume. Ranije smo pokazali da je $j\leqslant i+1$, što nam daje da su jedine dve mogućnosti za j, $j=i,i+1$.

Naša matrica A ima vrednosti:

$$a_{i,i+1}=(-1)^{2i}\binom{i}{i}\frac{i}{N}\left(1-\frac{i}{N}\right)\binom{i-1}{i-1}+(-1)^{2i}\binom{i}{i}\binom{i}{i+1}=\frac{i}{N}\left(1-\frac{i}{N}\right)$$

$$\begin{aligned}a_{i,i} &= (-1)^{2i-1}\binom{i}{i-1}\frac{i-1}{N}\left(1-\frac{i-1}{N}\right)\binom{i-2}{i-2}+(-1)^{2i-1}\binom{i}{i-1}\binom{i-1}{i}+\\ &+(-1)^{2i}\binom{i}{i}\frac{i}{N}\left(1-\frac{i}{N}\right)\binom{i-1}{i-2}+(-1)^{2i}\binom{i}{i}\binom{i}{i}=\\ &=\frac{i(i-1)}{N}\cdot\left(-1+\frac{i-1}{N}+1-\frac{i}{N}\right)+1=\\ &=1-\frac{i(i-1)}{N^2}\end{aligned}$$

□

Teorema 3.0.3. *Karakteristični koreni matrice prelaza P Moranovog modela su*

$$\lambda_i = 1 - \frac{i(i-1)}{N^2} \qquad i = 0, ..., N$$

Dokaz. Dokaz sledi iz prethodne leme, kako je matrica $R^{-1} \cdot P \cdot R$ donja trougaona matrica, njeni karakteristični koreni su vrednosti sa dijagonale. Takođe, karakteristični koreni matrice P i $R^{-1} \cdot P \cdot R$ su jednaki, čime je tvrđenje dokazano. □

Vidimo da su 0 i N apsorbciona stanja, zanima nas koliko vremena treba da se ova stanja dostignu? Ovde ćemo koristiti slučajnu promenljivu koja predstavlja vreme potrebno da se dođe u apsorbciono stanje.

$$H^A = \inf \{n \geqslant 0 : X_n \in A\}$$

Vidimo da je ovo najkraće vreme nakon koga je čitava populacija homozigotna, odnosno nakon koga je alel a zauvek izgubljen. Ukoliko nemamo takvu populaciju pišemo $\inf \{\emptyset\} = \infty$.

Sa $k_i^A = E(H^A \mid X_0 = i)$ ćemo označiti očekivano vreme potrebno da dođemo u apsorbciono stanje ukoliko je inicijalna populacija imala i jedinki sa alelom A.

Kako se pojedinaci iz jedne generacije nasumično odaberu za reprodukciju, a u isto vreme se jedinka iz iste generacije zamenjuje potomkom druge, a ovaj proces se nastavlja dok sve jedinke ne budu u potpunosti zamenjene, zaključujemo da je vreme između tranzicija je slučajna promenljiva sa eksponencijalnom rapodelom. Sledi pojašnjenje:

1. Svojstvo nepamćenja: Verovatnoća tranzicije u svakom trenutku je konstantna i ne zavisi od prethodne tranzicije. Ovo svojstvo je poznato kao svojstvo nepamćenja, što je karakteristično za eksponencijalnu raspodelu.

2. Poasonov proces: Moranov model se može shvatiti kao Poasonov proces, gde je verovatnoća da se tranzicija dogodi u malom vremenskom intervalu proporcionalna dužini tog intervala. Ovo je još jedna karakteristična osobina eksponencijalne raspodele.

3. Markovsko svojsto: Moranov model zadovoljava Markovsko svojstvo, što znači da verovatnoća tranzicije zavisi samo od trenutnog stanja, a ne od prethodnih.

Ove osobine zajedno impliciraju da je vreme između tranzicija u Moranovom modelu eksponencijalno raspoređeno.

Populacija od N jedinki prema Moranovom modelu evoluira, kao što smo rekli, eksponencijalnom brzinom sa parametrom $\binom{N}{2}$. Ovaj model se zasniva na pretpostavci da se samo jedan pojedinac bira za reprodukciju i da se jedan pojedinac bira za eliminaciju u svakom diskretnom vremenskom koraku. Kako je populacija veličine N, tada postoji $N(N-1)$ različitih parova jedinki koje mogu učestvovati u interakciji reprodukcije i eliminacije u svakom koraku. Ovaj izraz se deli sa 2, pošto svaki par dolazi u obzir tačno dva puta (npr. par (1,2) je isti kao i par (2,1)), pa dobijamo vrednost od $\frac{N(N-1)}{2}$ parova jedinki koje učestvuju u ovoj razmeni.

Teorema 3.0.4. *Vektor očekivanog vremena dostizanja apsorbcionog stanja* $k^A = (k_i^A \quad : \quad i \in \Omega)$ *je minimalno nenegativno rešenje sistema linearnih jednačina:*

$$k_i^A = 0 \quad i \in A$$
$$k_i^A = 1 + \sum_{j \in A} p_{ij} \cdot k_j^A \quad i \notin A$$

Teorema 3.0.5. *Ukoliko je učestalost alela A u inicijalnoj populaciji p, onda je očekivano vreme za apsorpciju približno jednako:*

$$t(p) = -2(p \log(p) + (1-p) \log(1-p)).$$

Dokaz. Da bismo izračunali očekivano vreme za apsorbciju Moranovog modela, prvo je potrebno da izračunamo očekivani broj tranzicija fiksiranog diskretnog lanca do apsorbcije. Taj broj ćemo pomnožiti sa očekivanim vremenom između tranzicija kako bismo dobili traženu vrednost.

Koristimo prethodnu teoremu i imamo:

$$\begin{aligned} k_i &= 1 + \sum_p p_{ij} \cdot k_j = \\ &= 1 + \frac{i(N-i)}{N^2} k_{i+1} + \frac{i(N-i)}{N^2} k_{i-1} + \left(\frac{i^2}{N^2} + \frac{(N-i)^2}{N^2} \right) k_i = \\ &= 1 + \frac{i(N-i)}{N^2} k_{i+1} + \frac{i(N-i)}{N^2} k_{i-1} + \left(1 - \frac{2i(N-i)}{n^2} \right) k_i \end{aligned}$$

$$\Rightarrow \quad k_{i+1} - 2k_i + k_{i-1} = -\frac{N^2}{i(N-i)} \quad i = 1, ..., N-1.$$

Sada nam još treba očekivano vreme između tranzicija. Znamo da se tranzicije dešavaju sa stopom $\frac{N(N-1)}{2}$ po jedinici vremena, što znači da je očekivano vreme da se jedna tranzicija desi u jedinici vremena zapravo recipročna vrednost, odnosno jednako je: $\frac{2}{N(N-1)}$.

Dakle, očekivano vreme za apsorpciju je jednako

$$\frac{2}{N(N-1)} \cdot \left(\sum_{j=1}^{i} \frac{N-i}{N-j} + \sum_{j=i+1}^{N-1} \frac{i}{j} \right).$$

Zapišimo i kao $i = pN$ i dobićemo da je:

$$\frac{2}{N(N-1)} \cdot \left(\sum_{j=1}^{i} \frac{N-i}{N-j} + \sum_{j=i+1}^{N-1} \frac{i}{j} \right) = \frac{2}{N(N-1)} \cdot \left(\sum_{j=1}^{pN} \frac{N-pN}{N-j} + \sum_{j=pN+1}^{N-1} \frac{pN}{j} \right) =$$

$$= \frac{2N}{N(N-1)} \cdot \left(\sum_{j=1}^{pN} \frac{1-p}{N-j} + \sum_{j=pN+1}^{N-1} \frac{p}{j} \right) =$$

$$= \frac{2}{N-1} \cdot \left((1-p) \sum_{j=1}^{pN} \frac{1}{N-j} + p \cdot \sum_{j=pN+1}^{N-1} \frac{1}{j} \right) \approx$$

$$\approx -2((1-p)\log(1-p) + p\log(p))$$

Poslednju jednakost (približnu jednakost), ukoliko je N veliko, ćemo dobiti iz aproksimacije Rimanovim sumama na podintervalima:

$$\sum_{j=pN+1}^{N-1} \frac{1}{j} = \int_{pN+1}^{N} \frac{1}{x}\, dx = \ln\left(\frac{N}{pN+1}\right) \approx \ln\left(\frac{N}{pN}\right) = -\ln(p)$$

$$\sum_{j=1}^{pN} \frac{1}{N-j} = \sum_{j=pN+1}^{N} \frac{1}{j} \approx \int_{N(1-p)}^{N} \frac{1}{x}\, dx = \ln\left(\frac{N}{N(1-p)}\right) = -\ln(1-p)$$

Napomena: Granice integrala su određene vremenskim intervalom u kojem se traži očekivano vreme do apsorpcije. Ukoliko početno imamo učestalost alela A u populaciji jednaka p, onda imamo pN jedinki sa alelom A, a $N(1-p)$ jedinki sa alelom a. □

Wright - Fisher model

4.1 Osnovni model

Hardi - Vajnbergov zakon iz prethodnog poglavlja ima dosta ograničenja - slučajnost proizilazi samo iz slučajnosti razmnožavanja određene generacije; smatra se da je populacija na kojoj primenjujemo ovaj model beskonačno velika. Očito je da ovako konstruisan model ne može da pokrije slučaj genetičkog drifta u populacijama konačne veličine.

Postavka modela:

- Populacija je diploidna, razmnožavanje je polno i slučajno
- Posmatramo alele A i a i genotipove AA, Aa i aa
- Učestalost genotipova je jednaka kod oba pola. Ako govorimo o *hermafroditima*[1], npr. biljkama, jasno je da ovo ne predstavlja nikakvo uprošćavanje sistema.
- Veličina populacije je u svakoj generaciji u toku vremena konstantna i ima N jedinki. Ovo očito odstupa od realnosti pa možemo interpretirati ovu postavku na sledeći način: od čitave populacije kojoj varira velličina mi izaberemo N jedinki koje uzimamo u analizu modela. Vidimo da slučaj u kome se veličina populacije smanjuje nije pokriven ovim modelom.
- posmatramo učestalost gena a ne genotipova

Uvodimo sledeće oznake:

X_n - broj A alela u n-toj generaciji $n = 0, 1, ..., 2N$

$\Rightarrow$ broj a alela u t-toj generaciji isnosi $2N - X_n$

Vidimo da je X_n diskretna slučajna pomenljiva sa binomnom raspodelom $n = 0, 1, ..., 2N$

$$X_n = i \;\Rightarrow\; X_{n+1} : B\left(2N, \frac{i}{2N}\right)$$

$\Rightarrow$ $X = X_0, X_1, ..., X_n$ je niz slučajnih promenljivih, a kako raspodela alela $n+1$-e generacije zavisi isključivo od raspodele alela u n-toj, X je takođe i Markov lanac:

[1] **hermafrodit** - jedinka koja stvara i muške i ženske polne ćelije

$$P(X_{n+1} = j|X_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$
$$i, j = 0, 1, ...2N$$

Stoga, matrica prelaza za jedan korak je $\mathbf{P} = [p_{ij}] = P(X_{n+1} = j|X_n = i)$. Dimenzije matrice $\mathbf{P}$ su $2N + 1 \times 2N + 1$.

Iako ovaj model nema tendenciju za usmerenu promenu u učestalosti gena, slučajni odabir jedinki za razmnožavanje će nakon nekog vremena dovesti do toga da sistem dostigne apsorbujuća stanja. Kada se apsorbujuća stanja dostignu, genetska varijacija će zauvek biti izgubljena. Vidimo da su stanja 0 i 2N apsorbujuća, važi:

$$\lim_{t\to\infty} X_n = 0 \ \vee \ \lim_{t\to\infty} X_n = 2N$$

Ukoliko su karakteristični koreni matrice prelaza P $\lambda_i \quad i = 0, ..., 2N$, kako imamo dva apsorpciona stanja znamo: $\lambda_0 = \lambda_1 = 1$, dok su ostali karakteriistični koreni različiti.

Ako je broj alela A u inicijalnoj populaciji i $(X_0 = i)$ koje su verovatnoće da Markov lanac bude apsorbovan u stanju 0, odnosno 2N? Obeležićemo ove verovatnoće sa $b_0(i)$ i $b_{2N}(i)$ redom.

$$b_0(i) = P(\lim_{n\to\infty} X_n = 0 \mid X_0 = i)$$
$$b_{2N}(i) = P(\lim_{n\to\infty} X_n = 2N \mid X_0 = i)$$

Vidimo da slučajni proces $(X_n)_{n\geqslant 0}$ ima sledeću osobinu:

$$E(X_{n+1} \mid X_n) = X_n \ \ \forall n$$
$$E(X_{n+1} \mid X_n = j) = 2N \cdot \frac{j}{2N} = j$$

A kako govorimo o lancu Markova, imamo:

$$E(X_{n+1} \mid X_0, X_1, ..., X_n) = X_n \ \ \forall n$$

Dakle, $(X_n)_{n\geqslant 0}$ je martingal za koji važi $E(X_n \mid X_0 = i) = i$. Kako je martingal uniformno ograničen imamo:

$$E\left(\lim_{n\to\infty} X_n \mid X_0 = i\right) = i$$
$$E\left(\lim_{n\to\infty} X_n \mid X_0 = i\right) = 0 \cdot b_0(i) + 2N \cdot b_{2N}(i)$$
$$\Rightarrow \ b_{2N}(i) = \frac{i}{2N} \ \wedge \ b_0(i) = 1 - \frac{i}{2N}$$

Sada nas zanima kojim će brzinom naši lanci Markova dostići apsorbciono stanje. Ovaj vremenski interval predstavlja slučajnu promenljivu koju ćemo zvati vreme za apsorbciju. Kada govorimo o genetici, ovo vreme zapravo predstalja vreme koje je potrebno kako bi heterozigoti, odnosno aleli Aa u našem slučaju, u potpunosti nestali iz populacije.

Lema 4.1.1. *Neka je $2 \leqslant r \leqslant 2N$ i neka je $a_0, a_1, ..., a_r$ niz brojeva različitih od nule. Tada sledi da vektor $v = (x_0, x_1, ..., x_{2N})$ gde je $x_k = \sum_{i=0}^{r} a_i \cdot k^i$ je nenula vektor.*

Dokaz. $P(k) = \sum_{i=0}^{r} a_i k^i$ je polinom reda $r \leqslant 2N$ i ima najviše r rešenja. Kako je $x_k = P(k)$ za $k = 0, 1, ..., 2N$ sledi da ne mogu svi x_k biti istovremeno nule, tako da je v nenula vektor. □

Lema 4.1.2. *Ako su $2 \leqslant r \leqslant 2N$ i $a_0, ..., a_r$ brojevi različiti od 0 tada važi da je vektor $v = (x_0, ..., x_{2N})$ nenula vektor gde je $x_k = \sum_{i=0}^{r} a_i \cdot \frac{k!}{(k-i)!}$*

Lema 4.1.3. *Za svaki broj $k \geqslant 0$ polinom $x(x-1) \cdot ... \cdot (x-i+1)$ gde je $0 \leqslant i \leqslant k$ predstavlja bazu vektorskog prostora svih polinoma stepena manjeg ili jednakog sa k.*

Teorema 4.1.4. *Brojevi λ_r $1 \leqslant r \leqslant 2N$ su karakteristični koreni Wright-Fisherova matrice prelaza $\boldsymbol{P}$, gde su λ_r definisani na sledeći način:*

$$\lambda_0 = 1 \;\; i \;\; \lambda_r = 1\left(1 - \frac{1}{2N}\right) \cdot ... \cdot \left(1 - \frac{r-1}{2N}\right) \; za \; 1 \leqslant r \leqslant 2N$$

Dokaz. Posmatrajmo dva vektora: $v_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ i $v_1 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 2N \end{bmatrix}$

Primetimo da su to da linearno nezavisna vektora dimenzija $1 \times (2N+1)$. Važe jednakosti:

$$Pv_0 = v_0 \quad \wedge \quad Pv_1 = v_1$$

Kako je P matrica prelaza, zbir elemenata u svakoj vrsti (ili koloni) je 1, tako da je prva jednakost jasna. Takođe, znamo da j-ta vrsta matrice predstavlja binomnu raspodelu slučajne promenljive $B : (2N, \frac{j}{2N})$ iz čega sledi da važi i druga jednakost.

U Markovljevim lancima, karakteristični koren 1 uvek odgovara apsorburajućem stanju. Apsorbirajuća stanja su stanja u kojima se lanac završava i više ne može preći u bilo koje drugo stanje. Kada je lanac u apsorbirajućem stanju, verovatnoća ostanka u tom stanju je 1, što znači da će karakteristični koren za to stanje biti 1. Kako mi imamo dva apsorbujuća stanja sledi da će 1 biti dvostruki karakteristični koren matrice prelza P.

$$\Rightarrow \quad \lambda_0 = \lambda_1 = 1$$

Treba dokazati da $\forall r = 2, ..., 2N$ i v_r nenula vektor važi:

$$Pv_r = \lambda_r v_r$$

Tada će slediti da su λ_r karakteristični koreni matrice P
Formiraćemo v kao u lemi 4.1.2:

$$v = (x_0, ..., x_{2N}) \text{ gde su } x_k = \sum_{l=0}^{r} a_l \cdot \frac{k!}{(k-l)!}$$

v zadovoljava uslov $Pv = \lambda_r v$ akko $\forall j = 0, 1, ...2N$

$$\begin{aligned}
\lambda_r x_j &= \sum_{k=0}^{2N} p_{jk} x_k \\
&= \sum_{k=0}^{2N} p_{jk} \sum_{l=0}^{r} a_l \cdot \frac{k!}{(k-l)!} \\
&= \sum_{k=0}^{2N} \binom{2N}{k} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k} \sum_{l=0}^{r} a_l \cdot \frac{k!}{(k-l)!} \\
&= \sum_{k=0}^{2N} \sum_{l=0}^{r} \binom{2N}{k} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k} \frac{k!}{(k-l)!} \cdot a_l \\
&= \sum_{l=0}^{r} \sum_{k=0}^{l} \binom{2N}{k} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k} \frac{k!}{(k-l)!} \cdot a_l + \sum_{l=0}^{r} \sum_{k=l}^{2N} \binom{2N}{k} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k} \frac{k!}{(k-l)!} \cdot a_l \\
&= \sum_{l=0}^{r} \sum_{k=l}^{2N} \frac{(2N)!}{(2N-l)!} \cdot \frac{(2N-l)!}{(2N-k)!(k-l)!} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k} \cdot a_l \\
&= \sum_{l=0}^{r} \frac{(2N)!}{(2N-l)!} \left(\frac{j}{2N}\right)^l \cdot a_l \cdot \sum_{k=l}^{2N} \binom{2N-l}{k-l} \left(\frac{j}{2N}\right)^{k-l} \left(1 - \frac{j}{2N}\right)^{2N-k} \\
&= \sum_{l=0}^{r} \frac{(2N)!}{(2N-l)!} \left(\frac{j}{2N}\right)^l \cdot a_l
\end{aligned}$$

$$\Rightarrow \quad \lambda_r x_j = \sum_{l-0}^{r} \frac{(2N)!}{(2N-l)!} \left(\frac{j}{2N}\right)^l \cdot a_l \qquad (*)$$

Napomena, važe jednakosti:

$$\forall k < l \;\; \frac{k!}{(k-l)!} = 0 \quad \wedge \quad \sum_{k=l}^{2N} \binom{2N-l}{k-l} \left(\frac{j}{2N}\right)^{k-l} \left(1 - \frac{j}{2N}\right)^{2N-k} = 1$$

Postavlja se pitanje kako da izaberemo vrednosti $a_0, ..., a_r$ da bismo dobili baš ono što nam treba.

Koristićemo lemu 4.1.3 iz koje imamo da $\forall r \;\; 0 \leqslant r \leqslant 2N$ postoje koeficijenti $c_{r,0}, c_{r,1}, ..., c_{r,r}$ tako da važi u našem slučaju:

$$\left(\frac{j}{2N}\right)^r = \sum_{m=0}^{r} c_{r,m} \cdot j(j-1) \cdot ... \cdot (j-m+1)$$

$$\Rightarrow \quad c_{r,r} = \frac{1}{(2N)^r}$$

Zapišimo λ_r na malo drugačiji način:

$$\lambda_r = \left(1 - \frac{1}{2N}\right) \cdot ... \cdot \left(1 - \frac{r-1}{2N}\right) = \frac{2N-1}{2N} \cdot ... \cdot \frac{2N-(r-1)}{2N} = \frac{(2N)!}{(2N-r)!} \cdot \frac{1}{(2N)^r}$$

$$\Rightarrow \quad \lambda_r = \frac{(2N)!}{(2N-r)!} \cdot c_{r,r} \qquad (**)$$

$$(*)+(**) \quad \Rightarrow \quad \lambda_r x_j = \lambda_r \sum_{l=0}^{r} a_l \cdot \frac{j!}{(j-l)!} = \sum_{l=0}^{r} \frac{(2N)!}{(2N-l)!} \left(\frac{j}{2N}\right)^l \cdot a_l = \sum_{l=0}^{r} \frac{(2N)!}{(2N-l)!} \cdot a_l \cdot \sum_{m=0}^{l} c_{l,m} \cdot \frac{j!}{(j-m)!}$$

$$\sum_{m=0}^{r} \lambda_r \cdot a_m \cdot \frac{j!}{(j-m)!} = \sum_{m=0}^{r} \left[\sum_{l=m}^{l} a_l \cdot \frac{(2N)!}{(2N-l)!} \cdot c_{l,m} \right] \frac{j!}{(j-m)!}$$

$$\Rightarrow \quad \lambda_r a_m = \sum_{l=m}^{r} a_l c_{l,m} \cdot \frac{(2N)!}{(2N-l)!} \quad \forall m = 0, 1, ..., r \qquad (1)$$

Vidimo da ukoliko je $m = r$ ova jednakost važi za svaki izbor broja a_r zbog toga što je $a_r \neq 0$ i zvog same definicije λ_r. Pošto važi za svako a_r uzimamo $a_r = 1$. Zapišimo (1) na malo drugačiji način:

$$\lambda_r a_m = a_m c_{m,m} \cdot \frac{(2N)!}{(2N-m)!} + \sum_{l=m+1}^{r} a_l c_{l,m} \cdot \frac{(2N)!}{(2N-l)!}$$

$$(\lambda_r + \lambda_m) \cdot a_m = \sum_{l=m+1}^{r} a_l c_{l,m} \cdot \frac{(2N)!}{(2N-l)!}$$

$$\Rightarrow \quad a_m = \frac{1}{\lambda_r + \lambda_m} \cdot \sum_{l=m+1}^{r} a_l c_{l,m} \cdot \frac{(2N)!}{(2N-l)!} \qquad (2)$$

Kako je $m \geqslant 2 \ \wedge \ m < r$ sledi da je $\lambda_r \neq \lambda_m$. Kako smo uzeli da je $a_r = 1$ iz jednačine (2) možemo odrediti sve $a_{m+1}, ...a_r$ tako da su vrednosti za λ_r, $1 \leqslant r \leqslant 2N$ karakteristični koreni matrice prelaza P.

□

Najveći karakteristični koren matrice prelaza P koji nije jednak jedinici je $\lambda_2 = 1 - \frac{1}{2N}$ koji je u slučaju velikih populacija veoma blizu jedinici. Iz toga zaključujemo da, iako u ovakvom modelu genetička varijacija mora vremenom biti izgubljena usled gubitka jednog alela pa će samim tim biti dostignuto apsorpciono stanje, potrebno je mnogo vremena da se to desi. Sada nas zanima koje je očekivano vreme koje je potrebno da se dostigne apsorbciono stanje?

Veoma je teško doći do eksplicitnog izraza za čekivano vreme do apsorpcije, ali možemo da dobijemo dosta dobru aproksimaciju. Obeležimo učestalost alela A u populaciji sa $p = \frac{i}{2N}$ i pretpostavimo da očekivano vreme do apsorpcije može biti aproksimiramo funkcijom klase $C^2([0,1])$, označićemo tu funkciju sa $t(p) \in C^2([0,1])$. Znamo da važi:

$$X_n = i \quad \Rightarrow \quad B\left(2N, \frac{i}{2N}\right)$$

$$\Rightarrow \quad E(X_{n+1} \mid X_n = i) = 2N \cdot \frac{i}{2N} = i$$

$$Var(X_{n+1} \mid X_n = i) = 2N \cdot \frac{i}{2N} \cdot \left(1 - \frac{i}{2N}\right) = \frac{i(2N-i)}{2N}$$

Sada nam treba promena udela alela A u populaciji. To će biti slučajna promenljiva, koju ćemo obeležiti sa δp i koja je jednaka:

$$\delta_p = \frac{X_n}{2N} - p$$

$$\Rightarrow \quad E(\delta_p) = E\left(\frac{X_n}{2N} - p\right) = \frac{1}{2N} \cdot E(X_n) - \frac{i}{2N} = \frac{i}{2N} - \frac{i}{2N} = 0$$

$$Var(\delta_p) = Var\left(\frac{X_n}{2N} - p\right) = \frac{1}{(2N)^2} \cdot Var(X_n) = \frac{1}{(2N)^2}\frac{i(2N-i)}{2N} = \frac{1}{2N}p(1-p)$$

Vidimo da je očekivanje δ_p jednako 0, i varijansu koja je veoma mali broj sledi da je promena veoma mala vrednost.

Koristićemo teoremu kao i u Moranovom modelu, gde znamo da je k_i^A vektor očekivanog vrema potrebnog da dođemo u apsorpciono stanje ukoliko je inicijalna populacija imala i jedinki sa alelom A. Takođe, važi:

$$k_i^A = 1 + \sum_p p_{ij} \cdot k_j^A$$

gde je p_{ij} verovatnoća prelaza iz i-tog u j-to stanje. Prema tome, izraz $p_{ij}k_j^A$ predstavlja očekivano vreme potrebno da dođemo u apsorpciono stanje ukoliko smo iz i-tog stanja prešli u j-to.

Kako je funkcija $t(p) \in C^2([0,1])$ možemo je razviti u Tejlorov red u okolini p:

$$\begin{aligned}
\Rightarrow \quad t(p) &= \sum_{\delta_p} P(p \to p + \delta_p) \cdot (t(p + \delta_p) + 1) \approx \\
&\approx \sum_{\delta_p} P(p \to p + \delta_p) \cdot \left(t(p) + \delta_p t'(p) + \frac{\delta_p^2}{2} t"(p) + 1\right) = \\
&= t(p) + t'(p)E(\delta_p) + \frac{1}{2}t"(p)E(\delta_p^2) + 1
\end{aligned}$$

Kako je

$$E(\delta_p) = 0 \qquad \wedge \qquad Var(\delta_p) = E(\delta_p^2) - E^2(\delta_p) = E(\delta_p^2) = \frac{1}{2N}p(1-p)$$

gornji izraz je jednak:

$$t(p) \approx t(p) + \frac{1}{4N}p(1-p)t"(p) + 1$$

$$p(1-p)t"(p) \approx -4N$$

Očigledno, ukoliko je $p = 0$ ili $p = 1$ to znači da je već dostignuto apsorpciono stanje pa je očekivano vreme do apsorpcije u tom slučaju jednako 0. Zbog toga imamo početne uslove:

$$t(0) = t(1) = 0$$

Dakle, rešavamo diferencijalnu jednačinu drugog reda sa početnim uslovom:

$$t"(p) = -\frac{4N}{p(1-p)}$$
$$t(0) = t(1) = 0$$

$$t'(p) = \int -\frac{4N}{p(1-p)}dp = -4N \cdot \int \left(\frac{1}{p} + \frac{1}{1-p}\right) dp = -4N \cdot \Big(ln|p| + ln|1-p|\Big) + C_1$$

$$\begin{aligned} \Rightarrow \qquad t(p) &= \int \Big(-4N \cdot \Big(ln|p| + ln|1-p|\Big) + C_1 \Big) dp = \\ &= -4N\Big(\int ln|p|dp + \int ln|1-p|dp \Big) + C_1 \cdot p + C_2 = \\ &= -4N\Big(\int ln(p)dp - \int ln(1-p)dp \Big) + C_1 \cdot p + C_2 \end{aligned}$$

Rešićemo $\int ln(p)$ parcijalnom integracijom:

$$u = ln(p) \quad \Rightarrow \quad du = \frac{1}{p}dp$$
$$dv = dp \quad \Rightarrow \quad v = p$$
$$\int ln(p) = uv - \int vdu = p \cdot ln(p) - \int \frac{1}{p}pdp = p \cdot ln(p) - p$$

Analogno, rešavamo $\int ln(1-p)$ parcijalnom integracijom, ali prvo uvedemo smenu:

$$1 - p = m \quad \Rightarrow \quad -dp = dm$$
$$\int ln(1-p) = -\int ln(m)dm = -m \cdot ln(m) + m = -(1-p) \cdot ln(1-p) + (1-p)$$

Vraćamo se na naš izraz za $t(p)$:

$$\begin{aligned} t(p) &= -4N\Big(p \cdot ln(p) - p + (1-p) \cdot ln(1-p) + 1 - p\Big) + C_1 p + C_2 = \\ &= -4N\Big(p \cdot ln(p) + (1-p) \cdot ln(1-p)\Big) + C_1 p + C_2 \end{aligned}$$

gde su C_1 i C_2 neke konstante koje ćemo odrediti iz početnih uslova.

$$t(0) = 0 \quad \Rightarrow \quad C_2 = 0$$
$$t(1) = 0 \quad \Rightarrow \quad C_1 = 0$$

Na kraju dobijamo da je očekivano vreme do apsorpcije približno jednako:

$$-4N\left(\frac{i}{2N} \cdot ln\Big(\frac{i}{2N}\Big) + \Big(1 - \frac{i}{2N}\Big) \cdot ln\Big(1 - \frac{i}{2N}\Big)\right)$$

4.2 W-F model sa različitim veličinama populacije

Uvodimo mogućnost promene veličine populacije iz generacije u generaciju. Naš model će u n-toj generaciji imati N_n jedinki, odnosno $2N_n$ alela.

Razmatraćemo dva slučaja: onaj u kome je veličina populacije deterministički određena i onaj gde je veličina populacije slučajna promenljiva.

4.2.1 Deterministički određena veličina populacije

Neka je $(N_n)_{n\geqslant 0}$ niz pozitivnih brojeva gde sa N_n obeležavamo veličinu n-te generacije. Kao i ranije, X_n je broj alela A u n-toj generaciji. Ako je $X_n = j$, uslovna raspodela slučajne promenljive X_{n+1} je binomna $B : \left(2N_{n+1}, \frac{j}{2N_n}\right)$. Vidimo da i ovde ne utiču prethodne generacije $0, 1, ..., n-1$ na raspodelu slučajne promenljive X_{n+1}, tako da pretpostavljamo da će $(X_n)_{n\geqslant 0}$ imati Markove osobine.

Posmatrajmo slučajnu promenljivu $Y_n = \frac{X_n}{2N_n}$

Teorema 4.2.1.

1. *$(Y_n)_{n\geqslant 0}$ je martingal i $Y = \lim_{n\to\infty} Y_n$*
2. *Pretpostavimo da je $P(0 < Y_0 < 1) > 0$. Tada, $P(Y = 0 \vee Y = 1) = 1 =$ akko $\sum_{n=0}^{\infty} \frac{1}{N_n} = \infty$*

Dokaz.

1. Kako je za date $X_0, X_1, ..., X_n$ uslovna raspodela slučajne promenljive $X_{n+1} : B\left(2N_{n+1}, \frac{X_n}{2N_n}\right)$ sledi da je Y_n martingal. Vidimo iz same formulacije Y_n da $Y_n \in [0, 1]$.
   Koristićemo *Doob-ovu teoremu za konvergenciju martingala* koja kaže da ako je $(X_n)_{n\geqslant 0}$ martingal sa $sup_{n\to\infty} E|X_n| < \infty$ tada X_n konvergira sa verovatnoćom 1 ka slučajnoj promenljivoj X koja ima konačno očekivanje.
   Imamo sada da je Y_n martingal i da $\forall n \;\; 0 \leqslant Y_n \leqslant 1$ pa Doob-ova teorema za konvergenciju martingala implicira da Y postoji sa verovatnoćom 1:
   $$Y = \lim_{n\to\infty} Y_n$$
2. Prvo, primetimo sledeće, ukoliko imamo slučajnu promenljivu Z sa binomnom raspodelom, $Z : B(m, p)$ tada važi:
   $$\begin{aligned} E\left(\frac{Z}{m}\left(1 - \frac{Z}{m}\right)\right) &= \frac{1}{m}E(Z) - \frac{1}{m^2}E(Z^2) \\ &= \frac{1}{m}E(Z) - \frac{1}{m^2}(D(Z) + E^2(Z)) \\ &= \frac{1}{m}mp - \frac{1}{m^2}\left(mp(1-p) + m^2p^2\right) \\ &= p - \frac{1}{m}p(1-p) - p^2 \\ &= p\left(1 - \frac{1}{m} + \frac{p}{m} - p\right) \\ &= p\left(1 - \frac{1}{m}\right)(1-p) \end{aligned} \qquad (\bigstar)$$

Kako $Y_n = \frac{X_n}{2N_n}$,gde $X_{n+1} : B\left(2N_{n+1}, \frac{X_n}{2N_n}\right)$ uz pomoć gore navedene jednakosti imamo:

$$E\left(\frac{X_{n+1}}{2N_{n+1}}\left(1-\frac{X_{n+1}}{2N_{n+1}}\right) \mid X_0, ..., X_n\right) = \frac{X_n}{2N_n}\left(1-\frac{X_n}{2N_n}\right)\left(1-\frac{1}{2N_{n+1}}\right)$$

$$E(Y_{n+1}(1-Y_{n+1} \mid X_0, ...X_n) = Y_n(1-Y_n)\left(1-\frac{1}{2N_{n+1}}\right)$$

Tražimo očekivanje i od leve i od desne strane jednakosti i iskoristimo osobinu uslovne verovatnoće $E(E(X \mid Y)) = E(X)$:

$$E(Y_{n+1}(1-Y_{n+1})) = \left(1-\frac{1}{2N_{n+1}}\right)E(Y_n(1-Y_n))$$

Dobili smo rekurentnu vezu iz koje sledi:

$$E(Y_n(1-Y_n)) = E(Y_0(1-Y_0))\prod_{k=1}^{n}\left(1-\frac{1}{2N_k}\right) \qquad (\bigstar)$$

Koristićemo *Teoremu dominantne konvergencije* koja kaže: Neka je $(X_n)_{n\geqslant 0}$ niz slučajnih promenljivih koje skoro sigurno konvergiraju ka slučajnoj promenljivoj X. Pretpostavimo da postoji slučajna promenljiva Y takva da $|X_n| < Y$ skoro sigurno za svako n i $E(Y) < \infty$ tada sledi $\lim_{n\to\infty} E(X_n) = E(X)$

Sada u našem slučaju, ako uradimo $\lim_{n\to\infty}$ leve i desne strane ($\bigstar$) imamo

$$E(Y(1-Y)) = E(Y_0(1-Y_0))\prod_{n=1}^{\infty}\left(1-\frac{1}{2N_n}\right)$$

$$P(0 < Y_0 < 1) > 0 \quad \Rightarrow \quad E(Y(1-Y)) = 0 \;\Leftrightarrow\; \prod_{n=1}^{\infty}\left(1-\frac{1}{2N_n}\right) = 0$$

$$\Leftrightarrow \sum_{n=1}^{\infty}\frac{1}{N_n}$$

$$0 \leqslant Y \leqslant 1 \quad \Rightarrow E(Y(1-Y)) = 0 \;\Leftrightarrow\; P(Y = 0 \vee Y = 1) = 1$$

□

4.2.2 Proizvoljna veličina populacije

Sada ćemo da posmatramo veličinu populaije u n-toj generaciji N_n kao slučajnu promenljivu. Pretpostavićemo da za date $(X_0, N_0), ..., (X_n, N_n)$ i N_{n+1} uslovna raspodela slučajne promenljive X_{n+1} je $B\left(2N_{n+1}, \frac{X_n}{2N_n}\right)$. Kao i ranije, posmatraćemo slučajnu promenljivu $y_n = \frac{X_n}{2N_n}$. Primećujemo:

$$\begin{aligned}
E(Y_{n+1} \mid X_0, ...X_n, N_0, ...N_{n+1}) &= E\left(\frac{X_{n+1}}{2N_{n+1}} \mid X_0, ..., X_n, N_0, ..., N_{n+1}\right) = \\
&= \frac{1}{2N_{n+1}} \cdot E(X_{n+1} \mid X_0, ..., X_n, N_0, ...N_{n+1}) = \\
&= \frac{1}{2N_{n+1}} \cdot 2N_{n+1} \cdot \frac{X_n}{2N_n} = \\
&= \frac{X_n}{2N_n} = \\
&= Y_n
\end{aligned}$$

Uz pomoć (☆) dobijamo:

$$E(Y_{n+1}(1 - Y_{n+1}) \mid X_0, ...X_n, N_0, ...N_{n+1}) = Y_n(1 - Y_n)\left(1 - \frac{1}{2N_{n+1}}\right)$$

Dakle, $(Y_n)_{n\geqslant 0}$ je martingal i po *Doob-ovoj teoremi za konvergenciju martingala* $Y = \lim_{n\to\infty} Y_n$ postoji sa verovatnoćom 1.

Teorema 4.2.2.

1. *Neka je* $U_n = Y_n(1 - Y_n) + \sum_{k=0}^{n-1} \frac{1}{2N_{k+1}} Y_k(1 - Y_k) \quad \Rightarrow \quad (U_n)_{n\geqslant 0}$ *je martingal.*

2. *Red* $\sum_{n=0}^{\infty} \frac{1}{N_{n+1}} Y_n(1 - Y_n)$ *konvergira sa verovatnoćom 1.*

Dokaz.

1.

$$\begin{aligned}
E(U_n \mid X_0, ..., X_n, N_0, ..., N_{n+1}) &= E\left(Y_n(1 - Y_n) + \sum_{k=0}^{n-1} \frac{1}{2N_{k+1}} \cdot Y_k(1 - Y_k) \mid X_0, ..., X_n, N_0, ..., N_{n+1}\right) = \\
&= E(Y_n(1 - Y_n) \mid X_0, ..., X_n, N_0, ..., N_{n+1}) + \sum_{k=0}^{n-1} \frac{1}{2N_{k+1}} \cdot Y_k(1 - Y_k) = \\
&= Y_{n-1}(1 - Y_{n-1})\left(1 - \frac{1}{2N_n}\right) + \sum_{k=0}^{n-1} \frac{1}{2N_{k+1}} \cdot Y_k(1 - Y_k) = \\
&= Y_{n-1}(1 - Y_{n-1}) + \sum_{k=0}^{n-2} \frac{1}{2N_{k+1}} \cdot Y_k(1 - Y_k) = \\
&= U_{n-1}
\end{aligned}$$

$\Rightarrow$ $(U_n)_{n\geqslant 0}$ je martingal.

2. Kako su svi $(U_n)_{n\geqslant 0}$ nenegativne slučajne promenljive, po *Doob-ovoj teoremi za konvergenciju martingala* U_n konvergira sa verovatnoćom 1.
   $Y_n(1-Y_n)$ takođe konvergira sa verovatnoćom 1.
   Sledi da $\sum_{n=0}^{\infty}\frac{1}{N_{n+1}}Y_n(1-Y_n)$ konvergira sa verovatnoćom 1.

□

Teorema 4.2.3. *Ako je* $P\left(\sum_{n=1}^{\infty}\frac{1}{N_n}=\infty\right)=1$ *, tada sledi* $P(Y=0\vee Y=1)=1$

Dokaz. Iz Teoreme 4.2.2 imamo da red $\sum_{n=0}^{\infty}\frac{1}{N_{n+1}}Y_n(1-Y_n)$ konvergira.

Kako je u našem slučaju $P\left(\sum_{n=1}^{\infty}\frac{1}{N_n}=\infty\right)=1$ moraće $lim_{n\to\infty}Y_n(1-Y_n)=0$ odakle sledi da ili je $Y=0$ ili je $Y=1$. □

Teorema 4.2.4. *Neka je* $P(0<Y_0<1)>0$

1. *Ako* $\forall n \quad \frac{1}{N_n}<\alpha_n$ *gde je* $(\alpha_n)_n$ *konvergentan niz realnih brojeva,* $\sum_{n=0}^{\infty}\alpha_n<\infty$ *sledi* $P(0<Y<1)>0$.
2. *Ako su* $\forall n \quad N_{n+1}$ *i* (X_n,N_n) *je nezavisne slučajne promenljive i suma očekivanja slučajnih promenljivih* $\frac{1}{N_n}$ *konačna,* $\sum_{n=0}^{\infty}E\left(\frac{1}{N_n}\right)<\infty$ *tada sledi* $P(0<Y<1)>0$.

Dokaz. Pretpostavka $P(0<Y_0<1)>0$ implicira $E(Y_n(1-Y_n))>0 \quad ,\forall n$.
Imamo da za svako n važi:

$$E(Y_{n+1}(1-Y_{n+1}) \mid X_0,...X_n,N_0,...N_{n+1})=Y_n(1-Y_n)\left(1-\frac{1}{2N_{n+1}}\right)$$

Tražimo očekivanje leve i desne strane jednakosti i koristimo osobinu uslovne verovatnoće $E(E(X|Y))=E(X)$, kao i to da je $1-\frac{1}{2N_{n+1}}\leqslant 1$ i dobijamo:

$$E(Y_{n+1}(1-Y_{n+1}))\leqslant E(Y_n(1-Y_n)). \qquad (\bullet)$$

Iz Teoreme 4.2.2 imamo:

$$U_n=Y_n(1-Y_n)+\sum_{k=0}^{n-1}\frac{1}{2N_{k+1}}Y_k(1-Y_k),$$

pa za bilo koje $n<m$ imamo:

$$U_m-U_n-(Y_m(1-Y_m)-Y_n(1-Y_n))=\sum_{k=n}^{m-1}\frac{1}{2N_{k+1}}\cdot Y_k(1-Y_k).$$

Sada ćemo da tražimo očekivanje od leve i desne strane jednakosti i iskoristićemo činjenicu da je $(U_n)_{n\geqslant 0}$ martingal:

$$E((Y_n(1-Y_n))-E(Y_m(1-Y_m))=\sum_{k=n}^{m-1}E\left(\frac{1}{2N_{k+1}}\cdot Y_k(1-Y_k)\right)$$

Puštamo $m \to \infty$:

$$E((Y_n(1-Y_n)) - \lim_{m\to\infty} E(Y_m(1-Y_m)) = \sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}} \cdot Y_k(1-Y_k)\right)$$

Koristimo *Teoremu dominantne konvergencije* i dobijamo:

$$E((Y_n(1-Y_n)) - E(Y(1-Y)) = \sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}} \cdot Y_k(1-Y_k)\right)$$

1. Iskoristimo pretpostavku teoreme da je $\forall n \quad \frac{1}{N_n} < \alpha_n$ gde je $(\alpha_n)_n$ konvergentan niz realnih brojeva i (●):

$$\begin{aligned} E((Y_n(1-Y_n)) - E(Y(1-Y)) &= \sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}} \cdot Y_k(1-Y_k) =\right) \\ &\leqslant \sum_{k=n}^{\infty} \frac{1}{2}\alpha_{k+1} E(Y_k(1-Y_k)) = \\ &\leqslant E(Y_n(1-Y_n)) \sum_{k=n}^{\infty} \frac{1}{2}\alpha_{k+1} \end{aligned}$$

   Sada kako je $E(Y_n(1-Y_n)) > 0 \quad \forall n$, a $\sum_{n=0}^{\infty} \alpha_n < \infty$. Biramo n takvo da $\sum_{k=n}^{\infty} \alpha_{k+1} < 2$ i dobijamo:

$$\begin{gathered} E((Y_n(1-Y_n)) - E(Y(1-Y)) < E(Y_n(1-Y_n)) \\ E(Y(1-Y)) > 0 \\ \Rightarrow \quad P(0 < Y < 1) > 0 \end{gathered}$$

2. Koristimo (●) i dobijamo:

$$\begin{aligned} E((Y_n(1-Y_n)) - E(Y(1-Y)) &= \sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}} \cdot Y_k(1-Y_k)\right) = \\ &= \sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}}\right) \cdot E(Y_k(1-Y_k)) = \\ &\leqslant E(Y_n(1-Y_n)) \cdot \sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}}\right) \end{aligned}$$

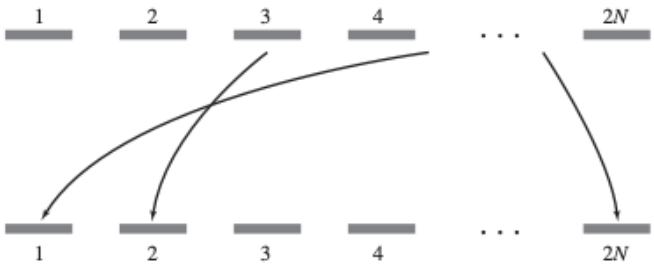
   Kako kao pretpostavku imamo da je $\sum_{n=0}^{\infty} E\left(\frac{1}{N_n}\right) < \infty$, biramo takvo n da $\sum_{k=n}^{\infty} E\left(\frac{1}{2N_{k+1}}\right) < 1$ i dobijamo:

$$\begin{gathered} E(Y_n(1-Y_n)) - E(Y(1-Y)) < E(Y_n(1-Y_n)) \\ E(Y(1-Y)) > 0 \\ \Rightarrow \quad P(0 < Y < 1) > 0 \end{gathered}$$

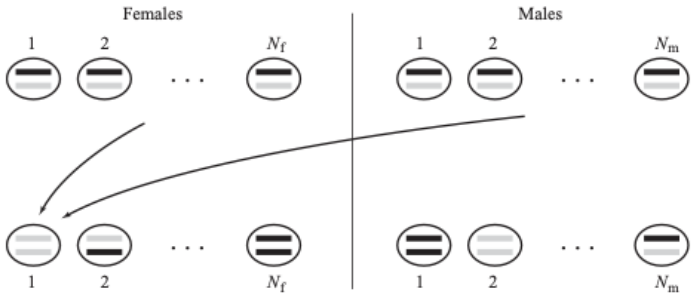
□

Koalescentna teorija

Posmatraćemo ovu teoriju u Wright-Fisher modelu. Ovaj osnovni model reprodukcije daje opis evolucije idealne populacije i prenosa gena iz jedne generacije u drugu. Mi ćemo, kao i uvek posmatrati jedan lokus sa dva alela i njihovo ponašanje kroz generacije. Ilustrovan prikaz modela u haploidnoj i diploidnoj verziji, redom, prikazan je na slikama ispod:



Slika 5.1. Haploidni model reprodukcije. Geni koji čine trenutnu generaciju (donja linija) se nasumično biraju sa zamenom iz prethodne generacije.



Slika 5.2. Diploidni model reprodukcije. Jedinka u trenutnoj generaciji (donja linija) nasumično bira, s zamenom, jedan od svojih alela iz ženske populacije i drugi alel iz muške populacije.

Radi lakšeg upoređivanja haploidnog i diploidnog modela, možemo pretpostaviti veličinu populacije od 2N alela, što odgovara N diploidnih ili 2N haploidnih jedinki. Dakle, haploidni model

reprodukcije se modelira pretpostavljajući 2N jedinki. U haploidnom modelu, svaki gen generacije $n+1$ se dobija kopiranjem gena (alela) nasumične jedinke iz generacije n. Ovaj proces se nezavisno ponavlja dok se ne izabere 2N gena (Slika 5.1). Svaki gen u generaciji $n+1$ ima jednog roditeljskog gena u generaciji n. Gen u generaciji n ne mora imati potomke u generaciji $n+1$ i ukoliko nema, njegova linija je izumrla.

Diploidna reprodukcija kod vrsta sa odvojenim polovima pretpostavlja postojanje dve podpopulacije - ženske i muške - veličine N_f i N_m, pri čemu je $N = N_f + N_m$, što ponovo predstavlja 2N gena. Svaka jedinka bira mužjaka (otac) i ženku (majku) iz muške odnosno ženske populacije iz generacije n. Unutar genotipa oca i majke, jedan od dva gena, odnosno alela, se bira sa verovatnoćom $0,5$. Ova reproduktivna šena je prikazana na Slici 5.2 Kao i u haploidnom modelu, svaki gen ima jednog roditeljskog gena koji potiče od mužjaka ili od ženke, ali ovde svaka jedinka ima dva roditelja.

Mi ćemo pratiti poreklo dva alela jedne jedinke unazad kroz vreme u haploidnom i diploidnom modelu. Postoje određena ograničenja u vezi sa tim koje roditelje gen (alel) može izabrati u diploidnom u odnosu na haploidni model. U haploidnom modelu, svi geni biraju nezavisno jedni od drugih, dok u diploidnom modelu drugi gen mora izabrati drugog roditelja u odnosu na prvi gen (ako je prvi alel izabrao roditelja iz muške podpopulacije, drugi alel mora izabrati roditelja iz ženske podpopulacije). Genealoški stabla u diploidnom modelu i haploidnom modelu su probabilistički slična za velike vrednosti N, N_f i N_m, zbog toga ćemo radi praktičnosti razmatrati haploidni model.

Broj potomaka određenog gena, i, u generaciji n predstavlja slučajnu promenljivu. Njegova raspodela se lako može izračunati, jer svaki put kada se stvara novi gen u generaciji $n+1$, ima verovatnoćom $\frac{1}{2N}$ da odabere roditelja i u generaciji n, i ovo uzorkovanje se ponavlja 2N puta s ponavljanjem. Neka v_i predstavlja broj potomaka gena i u generaciji n, $i = 1, 2, ..., 2N$, tada:

$$P(v_i = k) = \binom{2N}{k} \cdot \left(\frac{1}{2N}\right)^k \cdot \left(1 - \frac{1}{2N}\right)^{2N-k}$$

Ovo je primer binomne raspodele, $B_i(m, p)$, sa parametrima $m = 2N$ i $p = \frac{1}{2N}$. Dakle, broj gena koji potiču od određenog gena ima binomnu raspodelu. Znamo i očekivanu vrednost, kao i varijansu:

$$E(v_i) = mp = 2N \cdot \frac{1}{2N} = 1$$
$$D(v_i) = mp(1-p) = 2N \cdot \frac{1}{2N} \cdot \left(1 - \frac{1}{2N}\right) = 1 - \frac{1}{2N}$$

Da je srednja vrednost jednaka 1 je posledica konstantne veličine populacije: Ako bi prosečan broj potomaka gena bio veći/manji od 1, populacija bi se povećavala/smanjivala.

Kovarijanca broja potomaka za dva gena i i j je:

$$Cov(V_i, V_j) = E(V_i \cdot V_j) - E(V_i) \cdot E(V_j) = -\frac{1}{2N}$$

Koeficijent korelacije je:

$$Cor(V_i, V_j) = \frac{Cov(V_i, V_j)}{\sqrt{D(V_i)D(V_j)}} = \frac{-\frac{1}{2N}}{1 - \frac{1}{2N}} = -\frac{1}{2N-1}$$

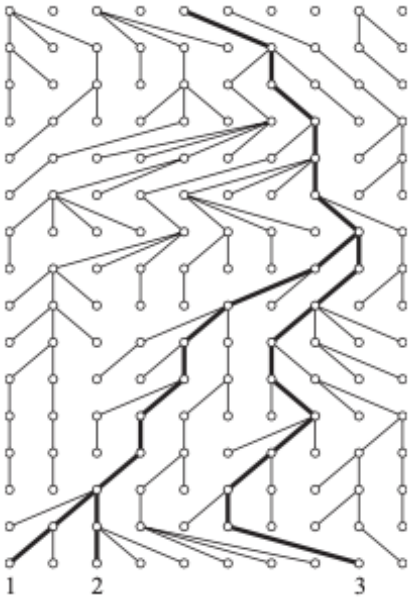
Dakle, V_i i V_j su gotovo nezavisni jedno od drugog u slučaju velikog 2N. Intuitivno, očekuje se negativna kovarijansa (ili korelacija) jer ako gen i ostavi mnogo potomaka u narednoj generaciji, tada se očekuje da će gen j ostaviti malo potomaka. To je zato što je ukupan broj potomaka svih gena u jednoj generaciji jednak 2N. Prirodno, ovaj efekat je izraženiji u malim nego u velikim populacijama.

Ako je 2N veliko, tada slučajna promenljiva v_i ima približno Poasonovu raspodelu sa parametrom 1:

$$P(v_i = k) \approx \frac{1}{k} \cdot e^{-1}$$

Verovatnoća da gen ne ostavi potomke je $P(v_i = 0) = e^{-1} \approx 0,37$, a otprilike $1 - e^{-1} \approx 0,63$ svih gena ima potomke. Dakle, u velikoj populaciji sa slučajnim razmnožavanjem, današnja (posmatrana) populacija potiče od relativno malog dela gena nekoliko generacija unazad, otprilike $0,63^n$ ako se posmatra n generacija unazad. Na primer, populacija veličine 10 000 jedinki (odnosno 10 000 posmatranih gena, alela) potiče od otprilike deset predaka (0.1% ukupne populacije) pre otprilike petnaest generacija ($10000 \cdot 0,6315^{15} \approx 10$). Poreklo preostalih gena (otprilike $10000 - 10 = 9990$) u precima populacije pre petnaest generacija nije preživelo do današnje generacije.

Slika 5.3 prikazuje Wright-Fisher model reprodukcije u populaciji veličine 10 jedinki tokom petnaest reprodukcionih ciklusa, što odgovara šesnaest generacija. Svaki gen je povezan sa svojim prethodnim genetskim pretkom. Naravno, moguće je da je potrebno više od šesnaest generacija da bi se pronašao prvi zajednički predak svih gena (eng. **MRCA** - Most Recent Common Ancestor), iako to nije slučaj u ovom primeru. U određenoj generaciji sve jedinke koje potiču od prvobitnih N će izumreti osim jedne (u slučaju diploidnih jedinki, ovom šemom možemo posmatrati rodoslov). Da bismo to videli, uzorkujemo celu populaciju i pratimo njihovo poreklo unazad dok se ne pronađe njihov MRCA. Svi ostali geni (jedinke) iz početne generacije nemaju potomke u najnovijoj generaciji. Kako je veoma teško pratiti poreklo cele populacije unazad, samo se uzorak n (obično je n mnogo manje od 2N, tj. $n << 2N$) gena uzima iz trenutne populacije i interesuje nas genetsko poreklo tog uzorka. Na slici 5.3 su slučajno odabrana tri gena (1, 2 i 3) u trenutnoj populaciji i prikazane su veze sa genima njihovih predaka.



Slika 5.3. Genealogija tri slučajno izabrane jedinke označene brojevima 1, 2, 3. Pretci ovih jedinki su označeni podebljanim linijama šesnaest generacija unazad.

Kao i u prethodnim modelima, i u koalescentnoj teoriji je prisutna Markovljeva osobina modela. U genetici je prirodno pretpostaviti da verovatnoća da se nešto desi (na primer, mutacija ili pronalaženje zajedničkog pretka) zavisi samo od trenutne situacije. U procesima gde se vreme diskretno meri (na primer, u generacijama), Markovljeva osobina je blisko povezana sa geometrijskom raspodelom. Ako je vreme kontinuirano (na primer, mereno štopericom), analogna raspodela je eksponencijalna raspodela. Posmatrajmo povratak unazad u vreme gde u svakoj generaciji nas zanima da li su se dva gena spojila sa zajedničkim pretkom. Imamo verovatnoću p za uspeh i $1-p$ za neuspeh. Jasno je da ovo predstavlja geometrijsku raspodelu. Neka je sa T obeleženo vreme koje prođe do prvog pronalaska zajedničkog pretka dva gena. Važi:

$$P(T = j) = (1-p)^{j-1}p$$

Kod geometrijske raspodele važi:

$$P(T > t_2 \quad | \quad T > t_1) = P(T > t_2 - t_1)$$

Kada se gleda verovatnoća da ćemo naći zajedničkog pretka dve jedinke (dva gena) posle vremena t_2, pri uslovu da se prvi uspeh desio nakon t_1, gledamo situaciju u kojoj smo već imali najmanje t_1 neuspeha. Verovatnoća da će se prvi uspeh desiti posle t_2, ukoliko je $T > t_1$ će biti ista kao da smo krenuli od početka i imali $t_2 - t_1$ eksperimenata bez uspeha, odnosno, tek da smo nakon tog vremena pronašli zajedničkog pretka dve jedinke. Ova osobina geometrijske raspodele predstavlja manjak memorije.

Ukoliko bismo sada vreme posmatrali na sve finijoj lestvici vremenskih tačaka, naša do sada geometrijska raspodela će postati ekpsponencijalna. Eksponencijalna raspodela ima takođe Markovsko svojstvo: ako je T slučajna promenljiva sa eksponencijalnom raspodelom, a $t_2 > t_1$ važi:

$$P(T > t_2 \mid T > t_1) = \frac{P(T > t_2,\, T > t_1}{P(T > t_1)} = \frac{P(T > t_2)}{P(T > t_1)} = \frac{e^{-\lambda \cdot t_2}}{e^{-\lambda \cdot t_1}} = e^{-\lambda \cdot (t_2 - t_1)} = P(T > t_2 - t_1)$$

$$\Rightarrow \quad P(T > t_2 \mid T > t_1) = P(T > t_2 - t_1)$$

Sada želimo da pređemo geometrijske raspodele, na eksponencijalnu. To ćemo uraditi na sledeći način. T je kao i do sada slučajna promenljiva koja predstavlja vreme koje prođe do prvog pronalaska zajedničkog pretka sa geometrijskom raspodelom sa parametrom p što znači:

$$P(T \geqslant j) = (1-p)^j$$

Ako je p malo, T je obično veliko i može se meriti na manjoj vremenskoj skali. Pretpostavimo da je M neki veliki broj takav da je $\lambda = pM$ i $t = \frac{j}{M}$. U kontekstu koelescentne teorije, M će biti reda veličine 2N, p će biti reda veličine 1/(2N), a j će biti reda veličine 2N. Zamenjujući $(1-p)^j$ kao:

$$(1-p)^j = \left(1 - \frac{pM}{M}\right)^{\frac{Mj}{M}} = \left(1 - \frac{\lambda}{M}\right)^{t \cdot M} \approx e^{-\lambda \cdot j}$$

dobijamo

$$P(T \geqslant j) = P\left(\frac{T}{M} \geqslant t\right) \approx e^{-\lambda \cdot t}$$

5.1 Diskretno vremenski koalescent

5.1.1 Uzorak od dva gena

Zanima nas koja je raspodela vremena koje prođe do pronalaska MRCA (poslednji zajednički predak) dva gena iz uzorka u haploidnom modelu sa 2N gena. Verovatnoća da ova dva gena pronađu zajedničkog pretka u prvoj generaciji je $\frac{1}{2N}$ - prvi gen može slobodno izabrati roditelja, ali drugi gen mora izabrati istog roditelja kao i prvi gen. Verovatnoća da ova dva gena imaju različitog roditelja je $1 - \frac{1}{2N}$.

Koristićemo oznaku T_k za vreme koje prođe do pronalaska MRCA k posmatranih gena.
Kako je uzorkovanje u različitim generacijama nezavisno, verovatnoća da ova dva gena pronađu zajedničkog pretka j generacija unazad, koristeći istu logiku je:

$$P(T_2 \geqslant j) = \left(1 - \frac{1}{2N}\right)^{j-1} \cdot \frac{1}{2N}, \qquad j = 1, 2, ...$$

Vidimo da slučajna promenljiva T_2 ima geometrijsku raspodelu sa parametrom $\frac{1}{2N}$.

Očekivano vreme do MRCA je:

$$E\left(\frac{1}{1/2N}\right) = 2N.$$

Dakle, isto kao i broj gena u posmatranoj populaciji.

5.1.2 Uzorak od n gena

Zanima nas koja je raspodela slučajne promenljive T_n, odnosno koja je raspodela vremena koje prođe do pronalaska poslednjeg zajedničkog pretka uzorka od n gena.

Posmatrajmo prvo sledeće, koja je verovatnoća da n gena ima n različitih predaka u prvoj generaciji? Analogno kao i u slučaju kod dva gena, prvi gen može slobodno birati roditelja od $2N$ gena (jedinki); drugi gen mora odabrati različitog roditelja u odnosu na prvi, tako da on bira od $2N-1$ jedinki; treći gen bira od $2N-2$ moguća gena i tako dalje. Zaključujemo da je ta verovatnoća jednaka:

$$\frac{2N-1}{2N}\cdot\frac{2N-2}{2N}\cdot\ldots\cdot\frac{2N-n+1}{2N}=\prod_{i=1}^{n-1}\left(1-\frac{i}{2N}\right)=e^{\ln\left(\prod_{i=1}^{n-1}\left(1-\frac{i}{2N}\right)\right)}=e^{\sum_{i=1}^{n-1}\ln\left(1-\frac{i}{2N}\right)}.$$

Koristićemo sledeće jednakosti koje predstvljaju razvoje funkcija u Tejlorov red:

$$e^x=\sum_{j=0}^{\infty}\frac{x^j}{j!},\qquad\forall x.$$

$$\ln(1-x)=-\sum_{j=1}^{\infty}\frac{x^j}{j},\qquad|x|\leqslant 1\wedge x\neq 1.$$

Dakle, uz pomoć gore navedenih jednakosti, naš izraz sada postaje:

$$e^{\sum_{i=1}^{n-1}\ln\left(1-\frac{i}{2N}\right)}=e^{-\sum_{i=1}^{n-1}\frac{i}{2N}+\left(\frac{1}{N^2}\right)}=1-\sum_{i=1}^{n-1}\frac{i}{2N}+\left(\frac{1}{N^2}\right)=1-\binom{n}{2}\frac{1}{2N}+\left(\frac{1}{N^2}\right)$$

Kako je $\left(\frac{1}{N^2}\right)$ veoma malo, možemo ga zanemariti. Ova aproksimacija zapravo znači da isključujemo mogućnost da više od jednog para gena nađe zajedničkog pretka u istoj generaciji. Ako pretpostavimo da je n mnogo manji broj od N, ovo deluje veoma smisleno.
Zaključujemo da u posmatranoj generaciji, verovatnoća da se desi koalescencija, odnosno da pronađemo poslednjeg zajedničkog pretka od n posmatranih gena (jedinki) je:

$$\binom{n}{2}\frac{1}{2N}$$

Odnosno, verovatnoća da ne dođe do koalescencije je:

$$1-\binom{n}{2}\frac{1}{2N}$$

Ovo nam služi kako bismo konačno došli do raspodele T_n koja je približno geometrijska sa parametrom $\binom{n}{2}\frac{1}{2N}$.

Vremena $T_2,\ldots,T_n$ su nezavisne slučajne promenljive.

5.2 Kontinuirano vremenski koalescent

Do sada smo vreme do koalescencije posmatrali u diskretnim jedinicama, sada želimo da pređemo na kontinuirano. Da bismo dobili kontinuirani koalescentni proces, koristimo $t = \frac{i}{2N}$, gde je i vreme izraženo u generacijama.

$$T_n \sim \text{Exp}\binom{n}{2}$$

$$\Rightarrow \quad P(T_n \leqslant t) = 1 - e^{-\binom{n}{2} \cdot t}$$

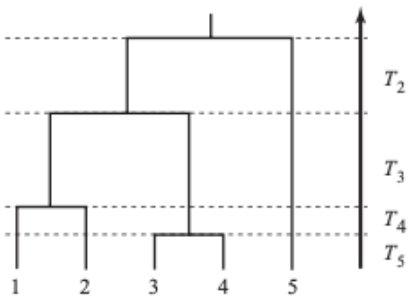
Kontinuirani vremenski koalascent sa svojom raspodelom nam omogućava lako računanje nekih važnih veličina u genealogiji, kao što su: visina koalescentnog stabla, veličina grananja stabla, efektivna veličina populacije...

Posmatrajmo stablo prikazano na slici 5.4. Visina koalescentnog stabla je slučajna promenljiva koju ćemo obeležiti sa H_n je n veličina posmatranog uzorka gena (jedinki). Visina stabla predstavlja sumu vremena koje je proteklo dok su se pretci u stablu razdvajali i koalescirali sve do poslednjeg zajedničkog pretka svih uzoraka, odnosno ova slučajna promenljiva opisuje vreme potrebno da se svi geni u uzorku vrate do njihovog poslednjeg zajedničkog pretka.

Dakle, visina stabla H_n u koalescentnom stablu može se modelirati kao suma nezavisnih eksponencijalnih slučajnih promenljivih. Preciznije, ako imamo n gena, visina stabla se može izračunati kao:

$$H_n = T_2 + T_3 + ... + T_n$$

gde su $T_2, T_3, ..., T_n$ nezavisne eksponencijalne slučajne promenljive koje predstavljaju vreme koalescencije dva, tri, ..., n gena.



Slika 5.4. Različiti vremenski periodi u koalescentnom stablu.

Primetimo da važi:

$$E(H_n) = E(\sum_{i=2}^{n} T_i) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2 \cdot \sum_{i=2}^{n} \left(\frac{1}{i-1} - \frac{1}{i} \right) = 2 \left(1 - \frac{1}{n} \right)$$

$$Var(H_n) = Var(\sum_{i=2}^{n} T_i) = \sum_{i=2}^{n} Var(T_i) = \sum_{i=2}^{n} \frac{4}{i^2(i-1)^2}$$

Informacija o očekivanoj vrednosti visine koalescentnog stabla pruža nam uvid u evolutivni proces i vreme povratka gena do njihovog zajedničkog pretka, što može biti korisno za različite analize i istraživanja. Na primer, ova informacija može biti korisna za proučavanje migracija, istorije populacija, za procenu brzine evolucije, za procenu starosti uzorkovane populacije itd.

Izrazi koje smo izveli za očekivanu vrednost i varijansu visine koalescentnog stabla pokazuju da u kontinuiranom koalescentu možemo bar formalno razmatrati beskonačno velik uzorak, jer beskonačan uzorak pronalazi zajedničkog pretka u konačnom očekivanom vremenu:

$$E(H_\infty) = 2\left(1 - \frac{1}{\infty}\right) = 2$$

Druga bitna stohastička promenljiva je slučajna promenljiva ukupnog grananja koalescentnog stabla, označena kao L_n, koja predstavlja ukupnu dužinu svih grana koje se javljaju u stablu koalescentnog procesa. Ova slučajna promenljiva meri koliko vremena je prošlo od poslednjeg zajedničkog pretka svih n gena do današnjeg trenutka. Ova promenljiva je suma svih vremenskih intervala (grana) u kojima dolazi do spajanja genetskih linija.

Kao što znamo, u koalescentnom modelu pretpostavlja se da se populacija razmnožava putem slučajnog odabira parova jedinki koji se međusobno ukrštaju, a potom se javlja koalescencija - spajanje linija potomaka nazad na zajedničkog pretka. Vreme do koalescencije između dva gena ima eksponencijalnu raspodelu sa stopom 1/2. Kako imamo uzorak od n gena, a vreme između svakog koalescentnog događaja je raspoređeno na opisan način, sledi:

$$P(L_n \leqslant t) = (1 - e^{-\frac{t}{2}})^{n-1}$$

Ova promenljiva daje informaciju o trajanju i intenzitetu procesa koalescencije, odnosno o evolutivnoj istoriji uzorka. Služi kao mera koliko genetskog porekla dele geni u uzorku. Veća vrednost L_n ukazuje na to da geni u uzorku imaju više zajedničkog porekla i da su bliže zajedničkom pretku. S druge strane, manja vrednost L_n ukazuje na to da geni u uzorku imaju manje zajedničkog porekla i da su se razvijali nezavisno jedni od drugih. Najčešće se za ovu meru koristi očekivana vrednost:

$$E(L_n) = \sum_{i=2}^{n} i \cdot E(T_i) = \sum_{i=2}^{n} i \cdot \frac{2}{i(i-1)} = 2 \cdot \sum_{i=1}^{n-1} \frac{1}{i} \approx 2\log(n)$$

$$Var(L_n) = \sum_{i=2}^{n} i^2 \cdot Var(T_i) = \sum_{i=2}^{n} i^2 \cdot \frac{4}{i^2(i-1)^2} = 4 \cdot \sum_{i=1}^{n-1} \frac{1}{i^2}$$

$$\Rightarrow \qquad n \to \infty \Rightarrow Var(L_n) \to \frac{2}{3}\pi^2$$

Treća bitna slučajna promenljiva za koalescentnu teoriju je efektivna veličina populacije. Za stvarnu populaciju ili neku model populaciju (na primer, diploidni Wright-Fisher model), veličina populacije koja najbolje aproksimira stvarnu populaciju (ili model) naziva se efektivna veličina populacije i obeležava se sa N_e. Dakle, N_e je broj genetski efektivnih jedinki koji bi imali isti uticaj na genetsku strukturu populacije kao i stvarni broj jedinki. Ova mera uzima u obzir faktore poput veličine populacije, broja potomaka svake jedinke i nivoa mešanja među jedinkama... Efektivna veličina populacije može biti manja od stvarnog broja jedinki u populaciji, posebno u populacijama koje su podložne genetičkom driftu ili imaju neproporcionalno razmnožavanje među jedinkama. Takođe, može varirati tokom vremena u zavisnosti od promena u populacijskoj strukturi. Postoje različite definicije ove mere, mi ćemo koristiti onu koja se fokusira na genetsku strukturu populacije u prethodnoj generaciji.

$$N_e = \frac{1}{2P(T_2 = 1)}$$

$P(T_2 = 1)$ je verovatnoća da su dve slučajno odabrane jedinke u prethodnoj generaciji direktni potomci istog pretka. Drugim rečima, to je verovatnoća da se dve jedinke nisu razdvojile ili se nisu granale od zajedničkog pretka u prethodnoj generaciji. Kada se bira jedna slučajna jedinka iz populacije, postoji jednaka verovatnoća da je ta jedinka bila muški ili ženski predstavnik. Kada se traži verovatnoća da su dve slučajno odabrane jedinke direktni potomci istog pretka, potrebno je uzeti u obzir oba moguća slučaja: da su obe jedinke potomci muške ili ženske jedinke. Zato se u formuli za N_e deli sa 2.

Ako je verovatnoća $P(T_2 = 1)$ veća, to znači da je veća verovatnoća da su dve jedinke direktno potomci istog pretka i da se populacija genetski manje razgranava, pa je samim tim i efektivna veličina populacije manja jer manje jedinki doprinosi genetskoj raznolikosti populacije.

Druga mera koja se često koristi je:

$$N_e = \frac{E(T_2)}{2}$$

Glavna razlika između ove dve mere je ta da je u prvoj definiciji N_e povezano sa prethodnom generacijom, dok je u drugoj povezano sa brojem generacija do pronalaska najskorijeg zajedničkog pretka (MRCA).Za haploidni Wright-Fisher model, ove dve definicije se podudaraju:

$$N_e = \frac{1}{2P(T_2 = 1)} = \frac{1}{2 \cdot 1/2N} = N$$

$$N_e = \frac{E(T_2)}{2} = \frac{2N}{2} = N$$

Naravno, ako veličina populacije varira tokom vremena, tada se ove dve definicije ne slažu jer prva formula za N_e zavisi od konkretne generacije na koju se fokusiramo.

Proces grananja

6.1 Galton - Watson proces grananja

6.1.1 Teorijski deo

Proces grananja je fundamentalni koncept u populacionoj genetici koji opisuje kako se genetska raznolikost širi i menja tokom generacija. Ovaj proces se odvija kroz formiranje novih linija potomaka iz postojećih jedinki. On omogućava proučavanje evolucije, genetske raznolikosti i populacionih struktura.

Zamislimo jedinke koje stvaraju potomke iste vrste: to mogu, kao i do sada, biti ljudi ili bakterije, ali možemo posmatrati proces grananja i kod neurona. Početni skup jedinki nazivamo nultom generacijom. Svaki član nulte generacije rađa određeni broj potomaka, a skup svih tih potomaka čini prvu generaciju. Članovi prve generacije stvaraju drugu itd. Možemo zamisliti ovo kao stablo u kojem se svaka generacija grana u sledeću generaciju, pa otuda i ovaj naziv.

Brojevi potomaka za posmatranu generaciju obeleženi sa $\xi_1, \xi_2, \xi_3, ...$ za različite pojedince koji su indeksirani, su međusobno nezavisne slučajne promenljive. Takođe, nezavisne su i od broja potomaka pojedinaca iz ranijih generacija. Pretpostavka modela je da su identično su distribuirani, sa raspodelom $\tilde{p} = (p_0, p_1, p_2, ...)$, koja ostaje nepromenjena kroz generacije. Naravno, važi:

$$\sum_{i=0}^{\infty} p_i = 1$$

Obeležimo, kao i ranije, broj jedinki u n-toj generaciji sa $(X_n)_{n\geqslant 0}$. Formalno, Galton-Watsonov proces $(X_n)_{n\geqslant 0}$ sa raspodelom potomaka $\tilde{p} = (p_0, p_1, p_2, ...)$ je diskretni Markovljev lanac čije vrednosti pripadaju skupu $\mathbb{Z}^+$, a čije verovatnoće prelaza u jednom koraku su date kao:

$$p_{ij} = P(X_{n+1} = j \mid X_n = i) = P(\xi_1 + \xi_2 + ... + \xi_i = j) = p_j^{*i}$$

gde je p_j^{*i} i-ta konvolucijaska potencija raspodele $\tilde{p}$.

Fiksirajmo $X_0 = 1$, i zaključimo uz pomoć gore navedenih pojmova da važi:

$$X_{n+1} = \sum_{i=1}^{X_n} \xi_i^{n+1}$$

Za razumevanje samog procesa, biće nam interesantno da saznamo koja je očekivana veličina n-te generacije ($E(X_n)$), zatim varijansa. Drugo zanimljivo pitanje je: koja je verovatnoća da $X_n \to 0$ kada $n \to \infty$, kako $X_n \in \mathbb{Z}^+$ ova verovatnoća predstavlja verovatnoću izumiranja.

Standardan pristup Markovim lancima bi podrazumevao analize matrice prelaza, što je u ovom slučaju gotovo nemoguće, tako da su Galton i Watson pribegli drugačijoj analizi, koristeći funkciju generisanja verovatnoće.

Definisaćemo funkciju generisanja verovatnoće:

$$\phi(s) = \sum_{k=0}^{\infty} p_k \cdot s^k, \qquad 0 \leqslant s \leqslant 1$$

$$\phi_n(s) = E(s^{X_n}) = \sum_{k=0}^{\infty} P(X_n = k) \cdot s^k = \sum_{k=0}^{\infty} p_k \cdot s^k, \qquad 0 \leqslant s \leqslant 1$$

Sada možemo izvesti:

$$\begin{aligned}
\phi_{n+1}(s) = E(s^{X_{n+1}}) &= \sum_{j=0}^{\infty} E(s^{X_{n+1}} \mid X_n = j) P(X_n = j) = \\
&= \sum_{j=0}^{\infty} E(s^{\xi_1+\xi_2+\ldots+\xi_j}) P(X_n = j) = \\
&= \sum_{j=0}^{\infty} E(s^{\xi_1} E(s^{\xi_2} \ldots E(s^{\xi_j}) P(X_n = j) = \\
&= \sum_{j=0}^{\infty} E(s^{\xi})^j P(X_n = j) = \\
&= \sum_{j=0}^{\infty} (\phi(s))^j P(X_n = j) = \\
&= \phi_n(\phi(s))
\end{aligned}$$

$$\Rightarrow \quad \phi_{n+1}(s) = \phi_n(\phi(s))$$

Ova rekurentna relacija će nam kasnije biti od značaja.

Sada pretpostavimo da ukoliko je $X_0 = 1$ da su očekivana vrednost i varijansa broja potomaka konačne, odnosno:

$$E(X_1) = m < \infty \qquad \wedge \qquad Var(X_1) = E(X_1)^2 - E(X_1^2) = \sigma^2 < \infty$$

Znamo,

$$E(X_n) = \phi_n^{'}(1) = \left.\frac{d\phi_n(s)}{ds}\right|_{s=1}$$

Iz naše rekurentne veze imamo:

$$\phi_n^{'}(1) = \phi_{n-1}^{'}(\phi(1))\phi^{'}(1) = \phi_{n-1}^{'}(1)\phi^{'}(1) = (\phi_{n-2}(\phi(1)))^{'} \cdot \phi^{'}(1) = \phi_{n-2}^{'}(1)(\phi^{'}(1))^2 = ... = \left(\phi^{'}(1)\right)^n$$

Kako je: $\phi^{'}(1) = \phi_1^{'}(1) = E(X_1) = m$

$$\Rightarrow \quad E(X_n) = \phi_n^{'}(1) = m^n$$

Ovo je intuitivno, kako je $X_0 = 1$ sledi da je $E(X_0) = 1$. Imamo zadato da je $E(X_1) = m$, a kako je X_1, kao i X_2 slučajna promenljiva sa raspodelom $\tilde{p}$, svaki od ovih m potomaka iz prve generacije daje potomke prema raspodeli $\tilde{p}$, tako da svako od njih daje m potomaka opet. Dakle, očekivana veličina druge generacije je m^2. Ovom logikom, očekivana veličina n-te generacije je m^n.

$$\lim_{n\to\infty} E(X_n) = m^n = \begin{cases} \infty & \text{, ako } m > 1 \\ 1 & \text{, ako } m = 1 \\ 0 & \text{, ako } m < 1 \end{cases}$$

Sada nas zanima varijansa.

$$\phi_n(s) = \sum_{k=0}^{\infty} P(X_n = k) \cdot s^k \quad \Rightarrow \quad \phi_n^{"}(s) = \sum_{k=0}^{\infty} k(k-1)P(X_n = k) \cdot s^k$$

Za $s = 1$:

$$\phi_n^{"}(1) = \sum_{k=0}^{\infty} k(k-1)P(X_n = k) = E(X_n^2) - E(X_n) = E(X_n^2) - m^n$$

$$\Rightarrow \quad E(X_n^2) = \phi_n^{"}(1) + m^n$$

$$Var(X_n) = E(X_n^2) - E^2(X_n) = \phi_n^{"}(1) + m^n - m^{2n}$$

$$\Rightarrow \quad \phi^{"}(1) = \sigma^2 - m - m^2$$

Kako bismo izračunali $Var(X_n)$, treba nam $\phi_n^{"}(1)$ koje nam je nepoznato. Pokušaćemo dobiti ovu vrednost iz ranije izračunate rekurentne relacije.

$$\phi_n(s) = \phi_{n-1}(\phi(s)) \quad \Big/ \frac{d}{ds}$$

$$\phi_n^{'}(s) = \phi_{n-1}^{'}(\phi(s)) \cdot \phi^{'}(s) \quad \Big/ \frac{d}{ds}$$

$$\phi_n^{"}(s) = \phi_{n-1}^{"}(\phi(s)) \cdot \left(\phi^{'}(s)\right)^2 + \phi_{n-1}^{'}(\phi(s)) \cdot \phi^{"}(s)$$

Za $s = 1$, koristeći $\phi(1) = 1$, $\phi^{'}(1) = m$, $\phi_{n-1}^{'}(1) = m^{n-1}$ i $\phi^{"}(1) = \sigma^2 - m - m^2$ gore navedena rekurentna relacija izgleda ovako:

$$\begin{aligned}
\phi_n^{"}(1) &= \phi_{n-1}^{"}(1)m^2 + m^{n-1}\left(\sigma^2 + m^2 - m\right) = \\
&= \left[\phi_{n-2}^{"}(1)m^2 + m^{n-2}\left(\sigma^2 + m^2 - m\right)\right] m^2 + m^{n-1}\left(\sigma^2 + m^2 - m\right) = \\
&= \phi_{n-2}^{"}(1)m^4 + \left(\sigma^2 + m^2 - m\right)\left(m^n + m^{n-1}\right) = \\
&= ... = \phi_1^{"}(1)m^{2n-2} + \left(\sigma^2 + m^2 - m\right)\left(m^{2n-3} + ... + m^n + m^{n-1}\right) = \\
&= \left(\sigma^2 + m^2 - m\right)\left(m^{2n-2} + ... + m^{n-1}\right)
\end{aligned}$$

Vratimo se na varijansu:

$$\begin{aligned}
Var(X_n) &= \phi_n^{"}(1) + m^n - m^{2n} = \\
&= \left(\sigma^2 + m^2 - m\right)\left(m^{2n-2} + ... + m^{n-1}\right) + m^n - m^{2n} = \\
&= \sigma^2 m^{n-1}\left(m^{n-1} + ... + m + 1\right) + m^n(m-1)(m^{n-1} + ... + m + 1) + m^n(1 - m^n) = \\
&= \sigma^2 m^{n-1}(m^{n-1} + ... + m + 1)
\end{aligned}$$

Konačno, dobijamo:

$$Var(X_n) = \begin{cases} n\sigma^2 & \text{, ako } m = 1 \\ \sigma^2 m^{n-1}\dfrac{m^n - 1}{m - 1} & \text{, ako } m \neq 1 \end{cases}$$

Ono što možemo zapaziti je:
Ako je $m = 1$, očekivana vrednost populacije se ne menja, ali varijansa raste linearno.
Ako je $m > 1$, očekivana vrednost populacije i varijansa rastu geometrijski.
Ako je $m < 1$, očekivana vrednost populacije i varijansa opadaju geometrijski.

Jedno od najbitnijih pitanja kod Galton - Watsonovog procesa grananja koja je verovatnoća izumiranja familije (posmatrane populacije). Izumiranje predstavlja događaj da određena generacija nema potomaka.

Teorema 6.1.1. *Za Galton-Watsonov proces grananja koji ima funkciju generisanja verovatnoće ϕ, verovatnoća izumiranja q je najmanje rešenje jednačine $\phi(s) = s$ na intervalu $[0, 1]$.*

Dokaz. Neka je:

$$q_n = P(X_n = 0) = P(X_n = X_{n+1} = X_{n+2}... = 0)$$
$$q_n = \phi_n(0) \qquad \forall n$$

Primetimo da je $\phi(s) = \sum_{k=0}^{\infty} p_k s^k$ neopadajuća funkcija po s na intervalu $[0, 1]$. Tu činjenicu možemo iskoristiti i uz pomoć indukcije pokazati sledeće:

$$q_n \geqslant q_{n+1}$$

- *baza indukcije:* $n = 1$

$$X_0 = 1 \quad \Rightarrow \quad q_0 = 0$$
$$q_1 = \phi(0) = P(X_1 = 0) \geqslant 0 = q_0$$
$$\Rightarrow \quad q_1 \geqslant q_0$$

- *indukcijska hipoteza:* Pretpostavimo da za n važi: $q_n \geqslant q_{n-1}$, odnosno $\phi(q_n) \geqslant \phi(q_{n-1})$.

- *indukcijski korak:* Dokažimo da važi za $n+1$

$$q_{n+1} = \phi_{n+1}(0) = \phi(\phi_n(0)) = \phi(q_n) \geqslant \phi(q_n - 1) = \phi(\phi_{n-1}(0)) = \phi_n(0) = q_n$$

$$\Rightarrow \quad q_{n+1} \geqslant q_n$$

Nas zanima verovatnoća q koja je jednaka:

$$q = \lim_{n\to\infty} P(X_n = 0) = \lim_{n\to\infty} q_n = \lim_{n\to\infty} \phi(q_{n-1}) = \phi\left(\lim_{n\to\infty} q_{n-1}\right) = q$$

$$\Rightarrow \quad q = \phi(q) \quad , kada \;\; n \to \infty$$

Dakle, moramo rešiti sledeću jednačinu da bismo našli traženu verovatnoću:

$$\phi(s) = s \quad 0 \leqslant s \leqslant 1 \tag{*}$$

Sada smo pokazali da je verovatnoća q rešenje jednačine *. Ostaje nam još da pokažemo da je najmanje rešenje.

Neka je neko π takođe rešenje jednačine (*) tako da je $\pi > q_n \qquad \forall n$. S druge strane, ako iskoristimo činjenicu da je ϕ neopadajuća funkcija važi:

$$q_n = \phi(q_{n-1}) \leqslant \phi(\pi) = \pi$$

$$\Rightarrow \quad q_n \leqslant \pi$$

Što predstavlja kontradikciju sa prvobitnom pretpostavkom.

□

Teorema 6.1.2. *Za Galton-Watson-ov proces grananja sa sa raspodelom potomaka $\tilde{p} = (p_0, p_1, p_2, ...)$ važi da ako je $p_0 > 0$ onda $s = 0$ nije nikada rešenje jednačine (*), dok je $s = 1$ uvek jedno rešenje. Takođe u zavisnosti od $E(X_1) = m$ važi sledeće:*

$$m \leqslant 1 \quad \Rightarrow \textit{verovatnoća izumiranja populacije je 1}$$
$$m > 1 \quad \Rightarrow \textit{verovatnoća izumiranja populacije je u otvorenom intervalu } (0, 1)$$

Teorema 6.1.3. *Neka je $p_1 < 1$. Tada važi:*

$$\forall\, k \geqslant 1 \qquad P(X_n = k) \to 0 \;\; kada \;\; n \to \infty$$

$$P(X_n \to \infty) = 1 - q = 1 - P(X_n \to 0)$$

Iz prethodne teoreme smo izuzeli slučaj kada je $p_1 = 1$ zato što tada važi da je $P(X_n = 1) = 1$ za sve n. Gore navedene teoreme tvrde sledeće: ukoliko jedna populacija koju posmatramo počinje od jednog pretka i $m > 1$, ona ima šanse za preživljavanje, odnosno nije za sigurno izumiranje te populacije. Ako, dalje, ta populacija zaista preživi, tada ona beskonačno raste. Pokazaćemo da populacija raste geometrijskom progresijom. (Ovo je tvrdio i kontroverzni demograf i ekonomista engl. *Thomas Robert Malthus* - stanovništvo ima tendenciju da raste eksponencijalno, dok su resursi, npr. hrana, ograničeni i rastu linearno. Zbog toga rast stanovništva premašuje raspoložive resurse, hrane i životnog prostora, što dalje dovodi do prirodne kontrole stanovništva kroz faktore kao što su glad, bolesti, ratovi i nesigurnost.).

Lema 6.1.4. *Ako je* $0 < m < \infty$ *i* $r \in \mathbb{Z}^+$ *važi* $E(X_{n+r} \mid X_n) = m^r X_n$.

Dokaz. Dokazaćemo lemu indukcijom po r:

- *baza indukcije:* $r = 1$
$$E\left(\sum_{i=1}^{X_n} \xi_i^{n+1} \mid X_n\right) = X_n E(\xi_i) = mX_n$$
- *indukcijska hipoteza:* Pretpostavimo da za r važi:
$$E\left(X_{n+r} \mid X_n\right) = m^r X_n$$
- *indukcijski korak:* Dokažimo da važi za $r+1$.
Koristićemo jednakost $E(X) = E(E(X \mid Y))$ i činjenicu da je X_n Markov lanac:
$$\begin{aligned} E(X_{n+r+1} \mid X_n) &= E(E(X_{n+r+1} \mid X_{n+r}, ..., X_n) \mid X_n) = \\ &= E(E(X_{n+r+1} \mid X_{n+r}) \mid X_n) \end{aligned}$$
Koristićemo deo iz baze indukcije za $n = n + r$.
$$\begin{aligned} E(X_{n+r+1} \mid X_n) &= E(E(X_{n+r+1} \mid X_{n+r}) \mid X_n) = \\ &= E(mX_{n+r} \mid X_n) = \\ &= mE(X_{n+r} \mid X_n) = \\ &= m \cdot m^r X_n = m^{r+1} X_n \end{aligned}$$
Poslednja jednakost je dobijena iz indukcijske hipoteze.

□

Pretpostavimo da je $0 < m < \infty$ i definišimo slučajnu promenljivu
$$W_n = \frac{X_n}{m^r} \qquad n = 0, 1, 2, ...$$
Sada posmatrajmo uslovno očekivanje W_n i iskoristimo prethodnu lemu i dobijamo:
$$E(W_{n+r} \mid W_n) = E\left(\frac{X_{n+r}}{m^{n+r}} \mid W_n\right) = \frac{1}{m^{n+r}} \cdot E(X_{n+r} \mid X_n) = \frac{1}{m^{n+r}} \cdot m^r X_n = \frac{X_n}{m^n} = W_n$$
Koristeći činjenicu da X_n ima Markovsko svojstvo, možemo zaključiti:
$$E(W_{n+r} \mid W_n, ..., W_0) = E(W_{n+r} \mid W_n) = W_n$$
Iz čega zaključujemo da je W_n martingal.

Teorema 6.1.5. *Neka je* $m \neq 0$ *tada važi da niz slučajnih promenljivih* $(W_n)_{n \geqslant 0}$ *konvergira ka slučajnoj promenljivoj* W *skoro sigurno. Ako je* $m > 1$ *i* $E(X_1^2) < \infty$ *onda* $(W_n)_{n \geqslant 0}$ *konvergira u prostoru* L_2 *i važi:*
$$E(W) = 1 \qquad \wedge \qquad Var(W) = \frac{Var(X_1)}{m^2 - m} = \frac{\sigma^2}{m^2 - m}$$

Dokaz. Pokazali smo da je W_n martingal. Lako je zaključiti da je $W_n \geqslant 0$, pa važi Doob-ova teorema za konvergenciju martingala koja nam daje da $(W_n)_{n\geqslant 0}$ konvergira sa verovatnoćom 1 ka nekoj slučajnoj promenljivoj W koja ima konačno očekivanje, $E(W) < \infty$.

U slučaju da je $m > 1$ i $E(X_1^2) < \infty$ želimo pokazati da $(W_n)_{n\geqslant 0}$ konvergira u prostoru L_2 odnosno da je $\sup\left\{E(W_n^2)\right\} < \infty$

$$Var(W_n) = E(W_n^2) - E^2(W_n) \qquad \Rightarrow \qquad E(W_n^2) = Var(W_n) + E^2(W_n)$$

$$E(W_n^2) = Var(W_n) - E^2(W_n) = Var\left(\frac{X_n}{m^n}\right) + E^2\left(\frac{X_n}{m^n}\right) =$$
$$= \frac{1}{m^{2n}} \cdot Var(X_n) + \frac{1}{m^n} m^n = 1 + Var(X_1)(1 - m^{-n})\frac{1}{m^2 - m} < \infty$$

Ovo implicira da $\lim\limits_{n\to\infty} W_n \to W$ u L_2.
Dalje, važi:

$$E(W) = \lim_{n\to\infty} E(W_n) = \lim_{n\to\infty} E\left(\frac{X_n}{m^n}\right) = 1$$

$$Var(W) = \lim_{n\to\infty} Var(W_n) = \lim_{n\to\infty} Var(X_1)(1 - m^{-n})\frac{1}{m^2 - m} = \frac{\sigma^2}{m^2 - m}$$

□

Dakle, $\frac{X_n}{m^n}$ konvergira ka slučajnoj promenljivoj W sa verovatnoćom 1, odnosno

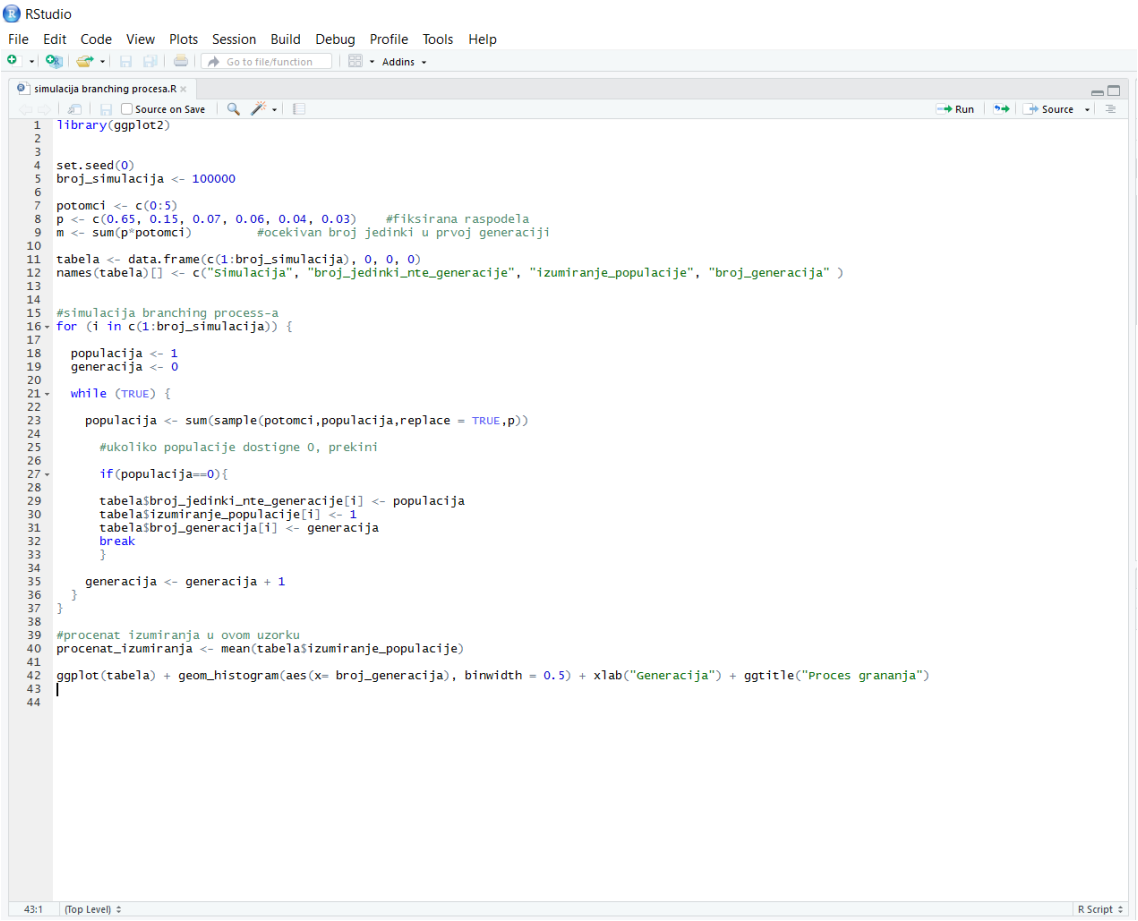
$$X_n \sim W \cdot m^n$$

6.1.2 Simulacija procesa grananja u R-u

U programu R ćemo simulirati proces grananja za zadatu fiksnu raspodelu i videćemo grafički kako je broj jedinki u n-toj generaciji, X_n raspoređen. Videćemo dva različita slučaja, jednan kada je očekivani broj jedinki u prvoj generaciji manji od jedan i drugi kada je veći od jedan.

$E(X_1) < 1$

Prvi slučaj koji ćemo posmatrati je onaj u kome je očekivani broj jedinki prve generacije manji od jedan. To ćemo osigurati zadavanjem fiksne raspodele. Ispod se nalazi slika koda simulacije kroz koju ćemo detaljno proći.



Funkcija *set.seed(0)* se koristi za postavljanje početnog stanja generatora slučajnih brojeva, a broj u zagradi označava to početno stanje. Kada koristimo generatore slučajnih brojeva, u našem slučaju to će biti *sample()* funkcija, svaki put kada iznova pokrenemo kod, bitno je da imamo isto početno stanje kako bismo osigurali doslednost rezultata novih simulacija.

Promenljiva *broj simulacija*, je logično, broj simulacija koji ćemo raditi. Ovaj broj bi trebao da je što veći kako bi rezultati bili što tačniji.

Potomci predstavljaju broj potomaka koji jedna jedinka može imati. Ja sam stavila da je to vektor $0, 1, 2, 3, 4, 5$.

Vektor p predstavlja sa kojom verovatnoćom će jedinka imati određeni broj potomaka. U našem primeru će bilo koja jedinka imati redom: 0, 1, 2, 3, 4 ili 5 potomaka sa verovatnoćom 65%, 15%, 7%, 6%, 4% i 3%. Sledi da je očekivani broj jedinki u prvoj generaciji:

$$E(X_1) = 0 \cdot 65\% + 1 \cdot 15\% + 2 \cdot 7\% + 3 \cdot 6\% + 4 \cdot 4\% + 5 \cdot 3\% = 0.78$$

Vidimo da je $E(X_1) < 1$ što znači da očekujemo da će u svakoj simulaciji populacija izumreti zauvek sa verovatnoćom 1, što ćemo videti i kroz slučajne simulacije. U kodu smo $E(X_1)$ obeležili sa m.

Formiramo tabelu pod nazivom *tabela* sa 4 kolone:

1. kolona - broj simulacije
2. kolona - broj jedinki n-te simulacije
3. kolona - izumiranje populacije, uzima vrednost 0 ukoliko je populacija preživela i 1 ukoliko je populacije u posmatranoj simulaciji izumrla
4. kolona - broj generacija koji je bio potreban da populacija izumre, ako izumre

For petljom prolazimo kroz broj simulacija i u svakoj simulaciji radimo sledeće: kada počne nova simulacija, populaciju stavljamo na 1, a izumiranje populacije na 0 (populacija je živa jer ima jednu jedinku). Koristimo *while* uslov jer ne znamo koliko ćemo jedinki u simulaciji imati, a za svaku hoćemo da radimo istu stvar. Naravno, moraćemo imati uslov koji će prekinuti ponavljanje kako ne bismo ušli u beskonačnu petlju. Intuitivno, ponavljanje bi prekinulo izumiranje populacije.
U *while*-u radimo sledeće:

sample(potomci, populacije, replace = TRUE, p)

Uzimamo uzorak od promenljive *potomci*, odnosno biramo broj od 0-5 (koliko smo zadali da jedinka može imati potomaka). To radimo u skladu sa fiksnom raspodelom p i to za svaku jedinku iz promenljive *populacija*. Uslov *replace = TRUE* nam daje da se broj potomaka u uzorku može ponavljati. Bez ovog uslova bi bilo nemoguće da dve generacije u populaciji (u jednoj simulaciji) u posmatranoj generaciji imaju isti broj potomaka. Na primer, u prvoj generaciji imamo 1 jedinku, koja da 3 potomka koji čine drugu generaciju, bilo bi nemoguće da oni daju $1, 0, 1$ potomaka bez ovog uslova. Znamo da se jedinke razmnožavaju nezavisno jedne od drugih, tako da ovaj uslov ne smemo da izostavimo.
Dalje, sumiramo generisani broj jedinki kako bismo dobili broj jedinki koji predstavlja novu populaciju za uzorkovanje.

Ukoliko čitava populacija u jednom trenutku postane 0, što znači da iz prethodne generacije ni jedna od jedinki nije imala potomke, nemamo više šta uzorkovati, populacija je izumrla pa je logično da postavimo *break* uslov kako bismo izašli iz petlje.
Dakle, postavljamo

if(populacija==0)

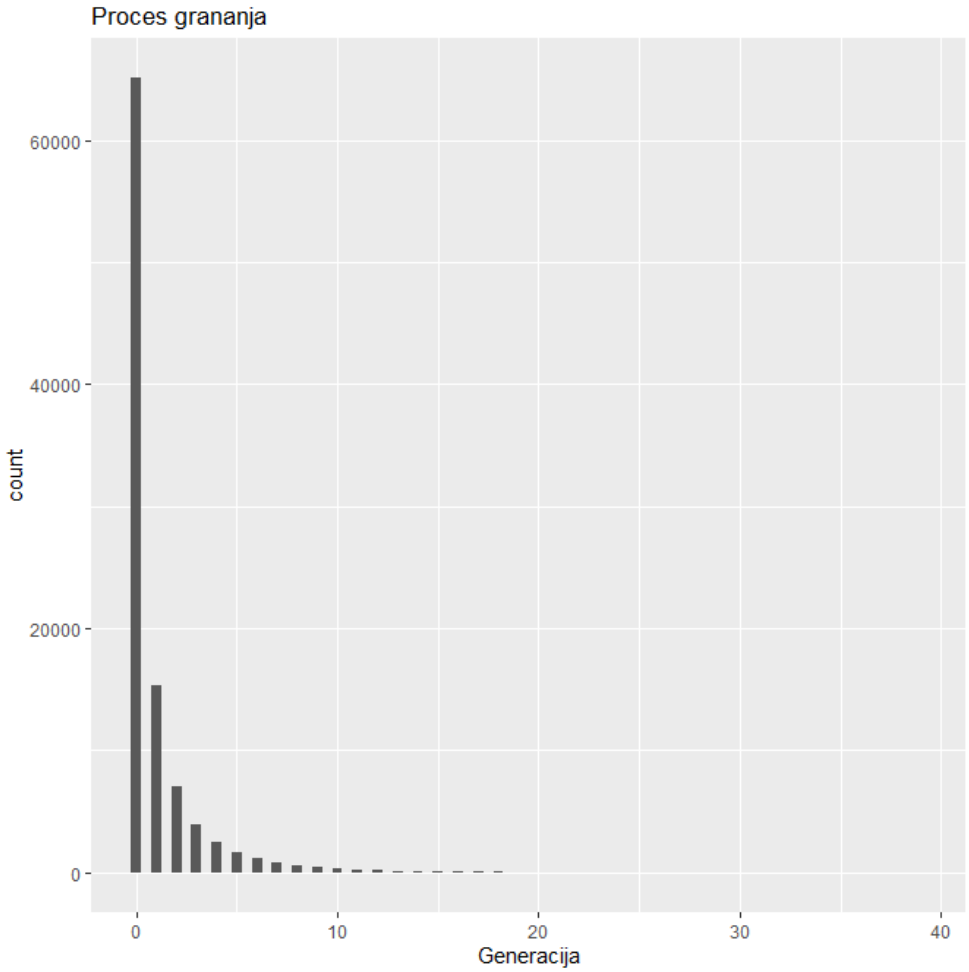
uslov u kome kažemo da se u i-ti red naše tabele upiše:

- broj jedinki i-te generacije je naša sumirana populacija (u ovom primeru će biti 0 jer su svi izumrli)
- u *izumiranje populacije* kolonu upiši 1
- u *broj generacija* upiši generaciju nakon koje je prekinuta *while* petlja

Ovaj uslov ćese proveravati nakon što završimo uzorkovanje u određenoj generaciji. Ako populacija nije izumrla, prelazimo na uzorkovanj u sledećoj generaciji i zbog toga brojač za generaciju uvećavamo za 1. Nakon uvećavanja brojača, ponovo ulazimo u *while* petlju i ponavljamo postupak.

Važno je napomenuti da u ovom slučaju, kada je $E(X_1) < 1$ nismo postavili ni jedan dodatan uslov za izlazak iz *while* petlje zato što će svaka simulacija završiti izumiranjem populacije nakon nekoliko generacija. Ovo neće biti slučaj ukoliko je $E(X_1) > 1$, ali o tome malo kasnije.

Sada ćemo da iscrtamo šta smo dobili pomoću funkcije *ggplot* kojoj zadamo podatke koje će iscrtati, u našem slučaju će iscrtavati našu tabelu; *aes* funkcija definiše estetiku grafičkih elemenata. Mi ćemo zadati da podatke prikaže kao histogram. Histogram je u R-u grafički prikaz raspodele numeričkih podataka u diskrete intervale gde svaki interval broji koliko podataka spada unutar tog intervala i iscrtava grafik.



$E(X_1) > 1$

Videli smo u prvom slučaju da će u svakoj simulaciji svaka populacija izumreti. Sada ćemo simulaciju da uradimo za $E(X_1) > 1$. To ćemo dobiti promenom raspodele. Kod je sličan kao prethodni, uz par izmena. Prva izmena je raspodela, sada ćemo opet imati broj potomaka 0, 1, 2, 3, 4, 5 sa verovatnoćama 40%, 25%, 20%, 7%, 5%, 3% redom.

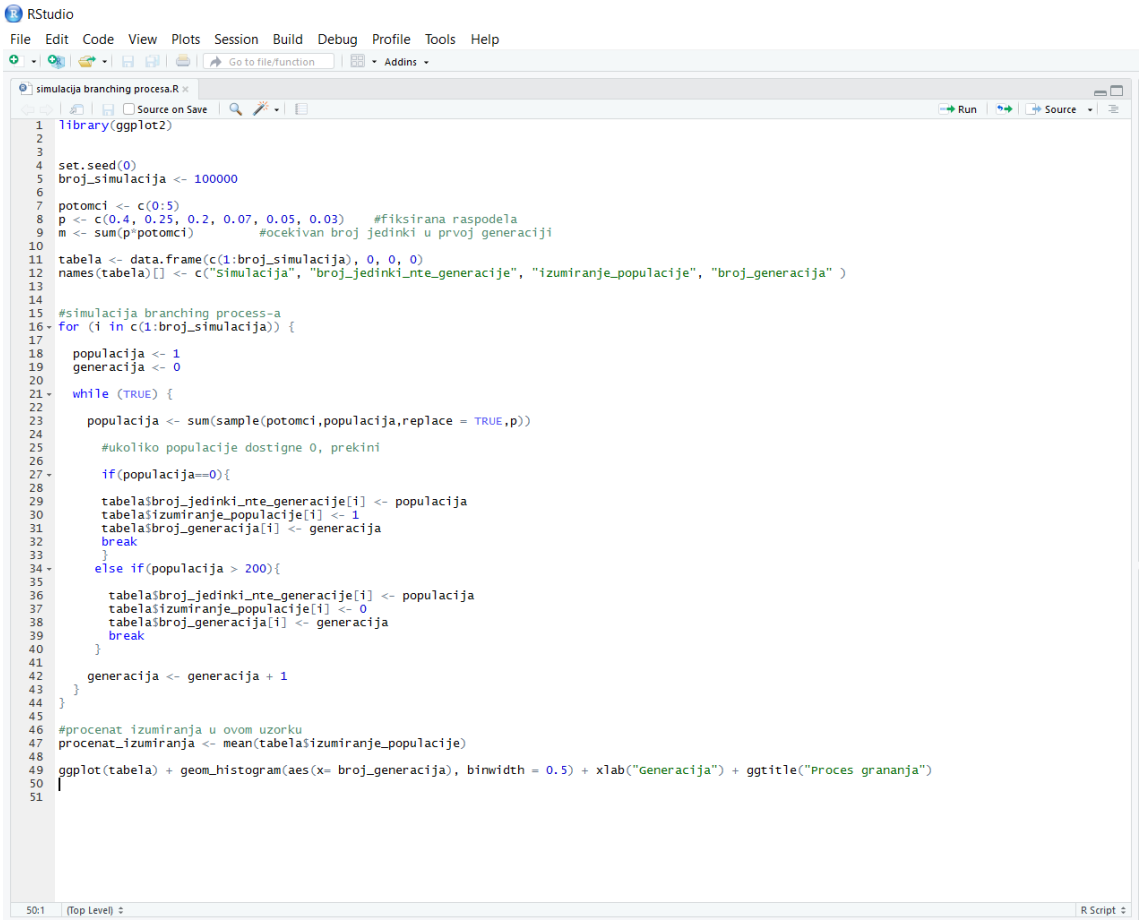
Međutim, ova izmena koda nije dovoljna. Teorijski smo pokazali kako će postojati populacije koje neće izumreti, već će se njihova veličina beskonačno povećavati. To znači da ćemo u kodu morati imati dodatan *break* uslov jer bismo u suprotnom ušli u beskonačnu *while* petlju. Moj dodatan uslov će biti: ukoliko populacija pređe 200 jedinki ja pretpostavljam da će populacija preživeti i izlazim iz petlje. Dodajem dodatan uslov:

$$else\ if(populacija > 200)$$

Ako je zadovoljen taj uslov u i-ti red naše tabele pišem:

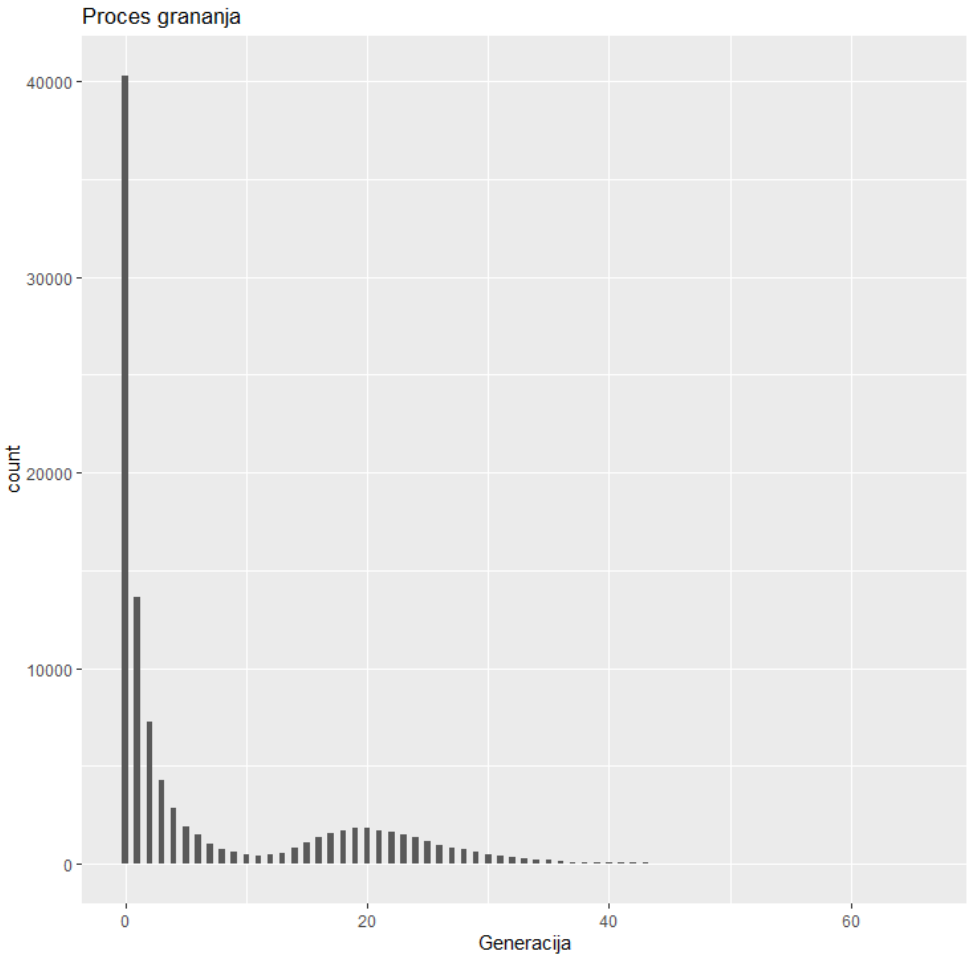
- broj jedinki i-te generacije je naša sumirana populacija

- u *izumiranje* kolonu upisujem 0
- u *broj generacija* kolonu upisuje se broj generacije u kome je broj jedinki prešao broj 200, odnosno ona generacija gde smo prekinuli uzorkovanje



Ovde sam još izračunala promenljivu *procenat izumiranja* koja predstavlja procenat izumrlih populacija u ovakvoj simulaciji i dobila sam vrenost 0.7654, odnosno procenat populacija koje su izumrle je 76.54%. Kako sam simulaciju pokrenula više puta mogla sam da vidim da li će se ovaj broj menjati i uvek je bio sličan procenat. Ovu promenljivu sam takođe izračunala i u prethodnom kodu, ali kako je već više puta ponovljeno, taj procenat je 100% jer će svaka populacija u jednom trenutku izumreti.

Slika koju ćemo na kraju uzorkovanja dobiti je sledeća:



Ono što vidimo kod oba grafika je to da grafički prikazan broj jedinki u n-toj generaciji liči na grafik geometrijske raspodele, kako smo ranije i teorijski pokazali.

6.2 Proces grananja sa višestrukim tipovima čvorova

6.2.1 Teorijski deo

Proces grananja sa višestrukim tipovima čvorova i Galton-Watson proces grananja su dva različita tipa procesa grananja, a razlikuju se po nekoliko ključnih karakteristika.

- Tipovi čvorova:
  U Galton - Watson procesu, svi čvorovi u procesu grananja su istog tipa. To znači da svaki čvor može imati istu stopu reprodukcije i istu raspodelu potomaka, dok u procesu grananja sa višestrukim tipovima čvorova, čvorovi se mogu podeliti u različite tipove ili kategorije. Svaka kategorija čvorova može imati svoju jedinstvenu stopu reprodukcije i raspodelu potomaka.

- Raspodela potomaka:
  U Galton - Watson procesu, obično se koristi Poasonova raspodela za broj potomaka svakog čvora, a ta raspodela je ista za sve čvorove. Kod drugog pomenutog procesa, raspodela broja potomaka može biti različita za svaku kategoriju čvorova. Na primer, jedna kategorija čvorova može koristiti Poasonovu raspodelu, dok druga može koristiti negativnu binomnu raspodelu.

- Cilj:
  Galton - Watson se obično koristi za modeliranje procesa u kojima su svi čvorovi ekvivalentni, kao što je širenje epidemije. Proces sa višestrukim tipovima čvorova se koristi za modeliranje složenijih procesa u kojima različite kategorije čvorova igraju različite uloge, kao što su genetski procesi ili evolucija različitih vrsta.

- Primeri primene:
  Galton - Watson proces se može primeniti na modele širenja epidemije, gde svaka zaražena osoba može zaraziti druge s određenom verovatnoćom. Proces sa višestrukim tipovima čvorova se može primeniti u biološkim modelima evolucije gde različite genetske vrste imaju različite stope reprodukcije i interakcije.

U suštini, razlika između ova dva procesa leži u složenosti i mogućnosti različitih tipova čvorova. Proces sa višestrukim tipovima čvorova je fleksibilniji i omogućava modeliranje raznolikih populacija čvorova, dok je Galton - Watson proces jednostavniji i koristi se za modeliranje homogenih populacija.

Od sada, pa na dalje ćemo pod *proces grananja* podrazumevati *proces grananja sa višestrukim tipovima čvorova*, osim ukoliko nije drugačije naglašeno.

Ovaj proces grananja podrazumeva proces grananja sa mutacijom u kojem je $X_i(t)$ broj ćelija tipa i u trenutku t. Ćelije tipa i rađaju se brzinom (verovatnoćom) a_i i umiru brzinom b_i. Uvek pretpostavljamo da je stopa rasta $\lambda_i = a_i - b_i > 0$. Kako bismo uzeli u obzir mutacije, pretpostavljamo da jedinke tipa i rađaju jedinke tipa $i+1$ brzinom u_{i+1}.

Počnimo proučavanje broja ćelija tipa 0, $X_0(t)$, koji je proces grananja u kojem se svaka ćelija rađa stopom a_0 i umire stopom b_0. Što se tiče teorije Markovljevih lanaca u kontinuiranom vremenu, matrica prelaza iz i-tog u j-to stanje ima oblik:

$$p_{ij} = \begin{cases} a_0 \cdot i & \text{, ako } j = i+1 \\ b_0 \cdot i & \text{, ako } j = i-1 \\ 0 & \text{, inače} \end{cases}$$

Pretpostavlja se da je $X_0(0) = 1$. Kako svaka jedinka rađa sa verovatnoćom a_0, a umire sa verovatnoćom b_0 imamo da važi:

$$\frac{d}{dt}E(X_0(t)) = \lambda_0 \cdot E(X_0(t))$$

gde je $\lambda_0 = a_0 - b_0$. Kako je $E(X_0(0)) = 1$ imamo:

$$E(X_0(t)) = e^{\lambda_0 \cdot t}$$

Kao i kod Galton - Watsonovog procesa, interesovaće nas verovatnoća izumiranja, gde ćemo opet koristiti funkciju generisanja verovatnoće.

Kako bismo izračunali funkciju generisanja verovatnoće $\phi(s,t) = E(s^{Z_0(t)})$ koristićemo sledeću lemu:

Lema 6.2.1.

$$\frac{d}{dt}\phi(s,t) = -(a_0+b_0)\phi + a_0\phi^2 + b_0 = (1-\phi)(b_0 - a_0\phi)$$

Proof. Ako je h malo, tada je verovatnoća više od jednog događaja u intervalu $[0,h]$ veoma mala, odnosno $\mathcal{O}(h^2)$. Verovatnoća rađanja je približno a_0h, a verovatnoća umiranja je približno b_0h. U slučaju kada nema čestica, funkcija generisanja verovatnoće od $X_0(t+h)$ će biti jednaka 1. U slučaju kada imamo dve čestice u vremenu h koje daju po dve nezavisne kopije procesa grananja, funkcija generisanja verovatnoće od $X_0(t+h)$ će biti $\phi(s,t)^2$. Dakle, imamo:

$$\phi(s,t+h) = a_0h\phi^2(s,t) + b_0h + \Big(1 - (a_0+b_0)\Big)\phi(s,t) + \mathcal{O}(h^2)$$

$$\frac{\phi(s,t+h) - \phi(s,t)}{h} = a_0\phi^2(s,t) + b_0 - (a_0+b_0)\phi(s,t) + \mathcal{O}(h)$$

Kada pustimo da $h \to 0$ tvrđenje je dokazano. □

Može se pokazati da je rešavanjem gore navedene jednačine dobijamo da je:

$$\phi(s,t) = \frac{b_0(s-1) - e^{-\lambda_0 \cdot t}(a_0s - b_0)}{a_0(s-1) - e^{\lambda_0 \cdot t}(a_0s - b_0)}$$

Pomoću funkcije generisanja verovatnoće možemo pronaći raspodelu koja joj odgovara:

$$p_0 = \alpha \qquad \wedge \qquad p_n = (1-\alpha)(1-\beta)\beta^{n-1} \;\; n \geqslant 1$$

gde su:

$$\alpha = \frac{b_0e^{\lambda_0 \cdot t} - b_0}{a_0e^{\lambda \cdot t} - b_0} \qquad \beta = \frac{a_0e^{\lambda_0 \cdot t} - a_0}{a_0e^{\lambda \cdot t} - b_0}$$

$$1 - \phi(s,t) = 1 - \frac{b_0(s-1) - e^{-\lambda_0 \cdot t}(a_0s - b_0)}{a_0(s-1) - e^{\lambda_0 \cdot t}(a_0s - b_0)} =$$
$$= \frac{\lambda_0(s-1)}{a_0(s-1) - e^{\lambda_0 \cdot t}(a_0s - b_0)}$$

Ukoliko stavimo da je $s = 0$ dobijamo sledeće verovatnoće:

$$P\Big(X_0(t) = 0\Big) = \frac{b_0 - b_0 e^{-\lambda_0 \cdot t}}{a_0 - b_0 e^{-\lambda_0 \cdot t}}$$

$$P\Big(X_0(t) > 0\Big) = 1 - \phi(0,t) = \frac{\lambda_0}{a_0 - b_0 e^{-\lambda_0 \cdot t}}$$

Primetimo da ova druga verovatnoća konvergira eksponencijalno ka $\frac{\lambda_0}{a_0}$.

Teorema 6.2.2. *Neka je $a_0 > b_0$. Kada $t \to \infty$ tada $e^{\lambda_0 \cdot t} X_0(t) \to W_0$, gde je*

$$W_0 = \frac{b_0}{a_0}\delta_0 + \frac{\lambda_0}{a_0}A_0,$$

gde je δ_0 "tačkasta masa u 0", a $A_0 : \mathcal{E}\left(\frac{\lambda_0}{a_0}\right)$.
To znači:

$$P(W_0 = 0) = \frac{b_0}{a_0} \qquad P(W_0 > x \mid W_0 > 0) = e^{-\frac{x\lambda_0}{a_0}}$$

Obeležimo sa $\Omega_0^0 = \{X_0(t) = 0,\ za\ neko\ t \geqslant 0\}$. Logično, $P(\Omega_0^0) = \frac{b_0}{a_0}$.
Kako $W_0 = 0 \in \Omega_0^0$, prethodna teorema implicira da je $W_0 > 0$ kada proces ne izumire.

$$\Omega_\infty^0 \{X_0(t) > 0 \ \ \forall t \geqslant 0\},$$

pa imamo:

$$lim_{t\to\infty} e^{-\lambda_0 \cdot t} X_0(t)\Big|_{\Omega_\infty^0} = V_0 : \mathcal{E}\left(\frac{\lambda_0}{a_0}\right).$$

6.2.2 Primena: Rak jajnika kod žena

Počinjemo opisivanjem četiri opšta stadijuma koji se koriste za klasifikaciju bolesti u kliničkom kontekstu:

1. rak ograničen na jedan jajnik
2. rak zahvata oba jajnika ili se širi na druga tkiva unutar karlice
3. rak se širi na trbuh
4. rak se širi na udaljene organe

Prema *SEER*[1] bazi podataka, raspodela stadijuma pri dijagnozi je (približno) 1.: 20%, 2.: 10%, 3.: 40% i 4.: 30%. Statistike preživljavanja u periodu od pet godina, zasnovane na stadijumu pri dijagnozi, iznose: 1.: 90%, 2.: 65%, 3.: 25%, 4.: 10%. Obzirom na ove statistike, sposobnost tačnog otkrivanja bolesti u ranim stadijumima mogla bi značajno poboljšati verovatnoću za preživljavanje ove bolesti. Naš cilj je da procenimo optimalni period vremena ili interval za sprovođenje preventivnih pregleda, kako bi se izbegle katastrofalne posledice. Konkretnije, to je vreme tokom kojeg je primarni tumor dovoljno velik da ga je moguće otkriti ga transvaginalnim ultrazvukom, dok količina metastaza nije značajno povećala šansu za smrtnost.

Karcinom jajnika započinje kao tumor na površini jajnika ili jajovoda, koje nazivamo primarni tumor. Metastaza se dešava ili direktnim širenjem iz primarnog tumora na susedne organe, kao što su mokraćna bešika ili debelo crevo, ili kada se ćelije raka odvoje sa površine primarnog tumora putem prelaska iz epitelnih u mezenhimske (EMT) ćelije. EMT podrazumeva promene u svojstvima ćelija, uključujući njihovu sposobnost da se odvoje od matičnog tumora, postanu pokretne, prodiru u krvne sudove, i naseljavaju se na novim mestima u telu. Kada se ćelije odvoje, plivaju u peritonealnoj tečnosti kao pojedinačne ćelije ili višećelijski sferoidi. Zatim se ponovo pričvršćuju za omentum i peritoneum i počinju agresivniji metastatski rast. Možemo stoga razmišljati o karcinomu jajnika kao o tri opšta podsistema tumorskih ćelija:

1. Primarni (ćelije u jajniku ili jajovodu), tip 0.
2. Peritonealni (žive ćelije u peritonealnoj tečnosti), tip 1.
3. Metastatski (ćelije implantirane na drugim unutartrbušnim površinama), tip 2.

Da bismo parametrizovali model grananja, koristićemo podatke iz studije[2] koja je ispitivala pojavu neotkrivenih karcinoma jajnika kod zdravih žena koje su podvrgnute hirurškom postupku u kojem se uklanjaju oba jajnika i oba jajovoda kako bi se sprečila mogućnost razvoja raka. Ova procedura se obično izvodi kod žena koje imaju povećan rizik od razvoja ovih vrsta karcinoma, na primer, ako postoji nasledna predispozicija za ovu bolest ili ako su prethodno imale rak dojke. Procenili su da karcinomi jajnika imaju dvofazni eksponencijalni rast sa $\lambda_0 = \frac{1}{4}ln2$ i $\lambda_2 = \frac{2}{5}ln2$ mesečno, tj. u ranom stadijumu vreme udvostručenja veličine raka je 4 meseca, dok je kasnije 2,5

[1] **SEER (Surveillance, Epidemiology, and End Results)** - baza podataka koje je program američke nacionalne agencije za rak (National Cancer Institute) koji prikuplja podatke o obolelima od raka i njihovom ishodu u Sjedinjenim Američkim Državama. Ova baza podataka osnovana je 1973. godine i obuhvata informacije o dijagnozama raka, lečenju i ishodima pacijenata, uključujući preživljavanje i smrtnost.

[2] Brown, Patrick O., and Chana Palmer. *The preclinical natural history of serous ovarian cancer: defining the target for early detection.* PLoS medicine 6.7 (2009): e1000114.

meseca. Stopa rasta λ_1 se ne može direktno proceniti iz ovih podataka, pa ćemo je uzeti kao 0 kako bismo dobili gornju granicu vremenskog okvira za efikasne preventivne preglede.
Nije moguće iz dostupnih podataka proceniti stope migracija u_1 i u_2. Izabrana je vrednost $u_1 \cdot u_2 = 10^{-4}$ kako bi se postigla saglasnost sa veličinama kao što je veličina primarnog tumora u trenutku kada se dostigne stadijum 3.

Ćelije tipa 0 predstavljaju proces grananja koji raste eksponencijalno brzinom $\lambda_0 > 0$. Ne zanima nas slučaj kada on izumire, tako da gledamo $\Omega^0_\infty = \{X_0(t) > 0 \ \ \forall t \geqslant 0\}$.
Od ranije znamo da:

$$lim_{t\to\infty} e^{-\lambda_0 \cdot t} X_0(t)\Big|_{\Omega^0_\infty} = V_0 : \mathcal{E}\left(\frac{\lambda_0}{a_0}\right)$$

Vreme t predstavlja količinu vremena od trenutka kada je počela početna mutacija koja je započela tumor. Taj događaj je nemoguće posmatrati, pa tako pomeranjem početka vremena možemo pretpostaviti da je $X_0(t) = e^{\lambda_0 \cdot t}$ kako bismo se oslobodili V_0.
Tip 1 ćelije napuštaju površinu primarnog tumora brzinom u_1 puta površine tumora. Ako je $\gamma_1 = 2\lambda_0/3$ imamo:

$$\begin{aligned} E\Big(X_1(t)\Big) &= \int_0^t u_1 E^{\gamma_1 \cdot s} e^{\lambda_1 (t-s)} ds = \\ &= \frac{u_1}{\gamma_1 - \lambda_1}(e^{\gamma_1 \cdot t} - e^{\lambda_1 \cdot t}) \approx \\ &\approx \left(\frac{u_1}{\gamma_1 - \lambda_1}\right) e^{\gamma_1 \cdot t} \end{aligned}$$

Pošto su tip 1 ćelije ćelije koje plivaju u peritonealnoj tečnosti i imaju manje pristupa hranljivim materijama, prirodno je pretpostaviti da je $\lambda_1 < \gamma_1$.

Teorema 6.2.3. *Ako je* $\gamma_1 > \lambda_1 \geqslant 0$ *sledi da je* $P\left(\frac{X_1(t)}{E(X_1(t))}\right) \to 1$ *kada* $t \to \infty$

U trenutku s mutacije u ćelije tipa 2 se javljaju sa verovatnoćom $u_2 \frac{u_1}{\gamma_1} e^{\gamma_1 \cdot s}$.
s_2 je trenutak u kome je verovatnoća mutacije na ćelije tipa 2 jednaka 1.

$$s_2 = \frac{1}{\gamma_1} ln\left(\frac{\gamma_1}{u_1 u_2}\right)$$

Teorema 6.2.4. *Ako je* $\lambda_2 > \gamma_1 > 0 \quad \Rightarrow \quad e^{-\lambda_2(t-s_2)} X_2(t) \to V_2$ *gde je* V_2 *suma tačaka Poasonovog procesa sa očekivanom vrednošću:* $\mu(x, \infty) = c_2 \cdot x^{-\frac{\gamma_1}{\lambda_2}}$.

$$c_2 = \frac{1}{a_2}\left(\frac{a_2}{\lambda_2}\right)^{\frac{\gamma_1}{\lambda_2}} \Gamma\left(\frac{\gamma_1}{\lambda_2}\right)$$

Kako bismo mogli da izračunamo trajanje vremenskog intervala tokom kojeg preventivni pregled može biti efikasan, neophodno je da precizno definišemo njegova dva kraja. Što se tiče gornje granice, definišemo vreme kada pacijent ulazi u treći stadijum kao T_2, gde je $T_2 = min\{t \ : \ X_2(t) = 10^9\}$.

Ovde je korićeno veoma često pravilo da 10^9 ćelija $= 1cm^3 = 1$ gram. Što se tiče donje granice, fokusiramo se na detekciju putem transvaginalnog ultrazvuka, pa definišemo $T_0 = min\{t \ : \ X_0(t) = 6.5 \cdot 10^7\}$, što odgovara sferičnom tumoru prečnika 0.5 cm. Ove definicije se baziraju na približnim procenama detektabilnosti i "značajne" metastaze.

Da bismo dobili gore pomenuta vremena, koristićemo stope rasta koje imamo:

$$\lambda_0 = \frac{ln2}{4} = 0,1733 \qquad \lambda_2 = \frac{2}{5}\ln(2) = 0,2772$$

$$\Rightarrow \quad e^{0,1733 \cdot T_0} = 6,5 \cdot 10^7 \qquad \Rightarrow \qquad T_0 = \frac{1}{0,1733} ln(6,5 \cdot 10^7) = 103,8 \ \ meseci$$

Da bismo "ugrubo" izračunali T_2, ignorišući slučajnost X_2, koristićemo gore navedenu formulu za trenutak u kome je verovatnoća mutacije do ćelije tipa 2 jednaka 1 i parametre $u_1 u_2 = 10^{-4}$ i $\gamma_1 = 0,1155$

$$\Rightarrow \quad s_2 = \frac{1}{0,1155} ln(1155) = 61,05$$

$$\frac{1}{\lambda_2} ln(10^9) = 74,76$$

Što znači da će od tog trenutka s_2, trebati približno 74,76 meseci da ćelije tipa 2 porastu do veličine 10^9.

Dakle, $T_2 = 61,05 + 74,76 = 135,81 \ \ meseci$

$$\Rightarrow \quad T_2 - T_0 = 135,81 - 103,8 = 32,01 \ \ meseci \approx 2,67 \ \ godina$$

Uključujemo i slučajnost X_2 i iz teoreme 6.2.4 imamo da važi

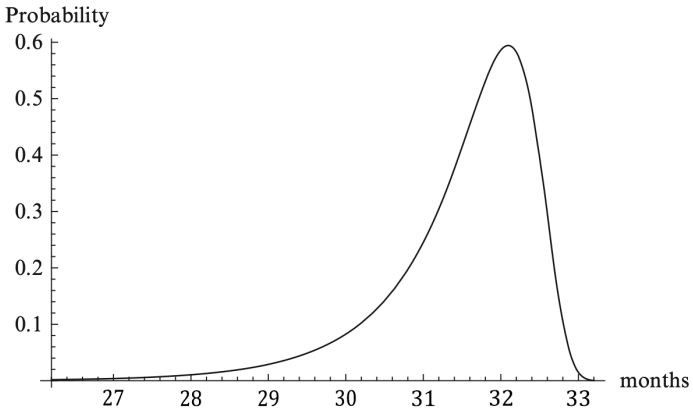
$$X_2(t) \approx e^{\lambda_2 (t - s_2)} \cdot V_2$$

Naravno, umesto t, nas interesuje veličina T_2 pa gledamo kada je gornja jednaČina jednaka sa 10^9.

$$\Rightarrow \quad T_2 \approx s_2 + \frac{1}{\lambda_2} ln\left(\frac{10^9}{V_2}\right)$$

Sada imamo da je naše $T_2 = 135,81 - ln\left(\frac{1}{V_2}\right)$.

Naša tražena vrednost intervala $T_2 - T_0$ ima distribuciju prikazanu na grafiku ispod.



Zaključak

U ovom radu su istraženi različite modeli koji se koriste u populacijskoj genetici i evoluciji kako bismo bolje razumeli procese koji oblikuju genetsku strukturu populacija. Ovi modeli su alati koji omogućavaju genetičarima, biolozima, epidemiolozima i drugim istraživačima da bolje razumeju i analiziraju genetske i evolucijske procese u prirodi i razvijaju strategije za očuvanje raznolikosti i kontrolu širenja bolesti.

Hardy - Weinberg-ov model nam pomaže da razumemo ravnotežu alelnih frekvencija u populaciji pod pretpostavkom odsustva mutacija, selekcije i drugih faktora.
Moranov model i Wright - Fisher-ov model pružaju uvid u efekte genetskog drifta i nasleđivanja u diskretnim i nepreklapajućim generacijama.
Koalescentna teorija nam omogućava da pratimo povratke unazad u vremenu i identifikujemo zajedničke pretke gena u populaciji.
Galton - Watson-ov proces grananja se koristi za modeliranje slučajnih procesa reprodukcije i nasleđivanja.

Svi ovi modeli imaju svoje prednosti i ograničenja, ali zajedno nam omogućavaju dublje razumevanje genetske evolucije i dinamike populacija.

Jedna stvar koja se mogla zaključiti i kroz sam rad je da se nismo bavili efektima mutacije, prirodne selekcije, migracija, uključivanje više lokusa... Uključujući pomenute faktore u modele populacijske genetike dobila bi se preciznija slika o tome kako se genetske osobine razvijaju tokom vremena i vernije bi prikazivala realnost. Naravno, ova analiza je veoma složena, evo nekih razloga i zašto.
Mutacije dovode do stvaranja novih alela, što povećava broj mogućih alelnih varijanti u populaciji. To znači da se treba pratiti veći broj alela i njihovih frekvencija, što povećava složenost analize. Analiza sa mutacijama može uključivati više genetskih lokusa koji su podložni mutacijama. Svaki lokus može imati svoju vlastitu stopu mutacije i dinamiku alelnih frekvencija. Mutacije se dešavaju tokom vremena, pa je potrebno pratiti kako se nove mutacije uvode u populaciju, kako se šire i kako utiču na genetsku raznolikost tokom vremena.
Prirodna selekcija često zavisi od okoline. Promene u okolini mogu značiti da će se selekcija različito odvijati tokom vremena ili na različitim mestima. Različite osobine mogu biti podložne različitim vrstama selekcije, uključujući pozitivnu, negativnu i neutralnu selekciju. Selekcija se dešava tokom vremena, pa je potrebno pratiti kako se genetske karakteristike menjaju tokom generacija. Postoji pojam koji se naziva epistaza i on se odnosi na interakciju između različitih gena, gde efekti jednog gena zavise od prisustva ili odsustva drugog gena. Na primer, jedan lokus može da ima alele koji određuju boju kose kod čoveka. takođe može postojati gen za ćelavost čoveka i za njega kažemo da

je epistatski u odnosu na gene za boju kose. Logično, ukoliko ljudi oćelave, ovaj gen će se jednako ispoljiti i kod osobe koja je prvobitno imala plavu i kod osobe koja je imala smeđu kosu.

Iako su modeli koji su uključeni u ovaj rad naizgled jednostavni i zbog gore pomenutih razloga ne oslikavaju baš najvernije realnu sliku u prirodi, laki su za razumevanje i pružaju dovoljno dobru osnovu za dalju nadogradnju. Takođe, predstavljaju lepu primenu stohastičkih procesa u svetu oko nas, što je i bio cilj.

Literatura

1. (Texts and Readings in Mathematics 40) A. Goswami & B. V. Rao, *A Course in Applied Stochastic Processes* - Hindustan Book Agency, 2006
2. Hein, Jotun, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory.* Oxford University Press, USA, 2004.
3. Ewens Warren John. *Mathematical population genetics: theoretical introduction.* Vol. 27. New York: Springer, 2004
4. Skancke, Jørgen. *Two models of population genetics.* MS thesis. Universitetet i Tromsø, 2008.
5. Etheridge, Alison. *Diffusion Process Models in Mathematical Genetics.*
6. Tuckwell, Henry C. *Elementary applications of probability theory.* Chapman and Hall/CRC, 2018.
7. Kimura, Motoo. *The neutral theory of molecular evolution.* Cambridge University Press, 1983.
8. Ewens, Warren John. *Population genetics.* Springer Science & Business Media, 2013.
9. Rajter-Ćirić, Danijela. *Verovatnoća.* Univerzitet u Novom Sadu, Prirodno-matematički fakultet (2009).
10. Durrett, Richard, and Richard Durrett. *Branching process models of cancer.* Springer International Publishing, 2015.
11. Brown, Patrick O., and Chana Palmer. *The preclinical natural history of serous ovarian cancer: defining the target for early detection.* PLoS medicine 6.7 (2009): e1000114.

Biografija

Zovem se Milana Radmanović, rođena sam 14. septembra 1997. godine u Somboru. Završila sam osnovnu školu "Ivo Lola Ribar", a potom i prirodno - matematički smer Gimnazije "Veljko Petrović".
2016. godine upisujem osnovne akademske studije na Prirodno - matematičkom fakultetu u Novom Sadu, smer teorijska matematika. Diplomirala sam 14, marta 2020. godine sa prosečnom ocenom 8,44. Iste godine upisujem master studije na smeru primenjene matematike. Sve ispite položila sam u oktobru 2022. godine čime sam stekla pravo na odbranu master rada.
Od septembra 2022. godine zaposlena sam u RBA Banci, a od maja 2023. godine u Raiffeisen Banci u Integrated Risk Unitu.

UNIVERZITET U NOVOM SADU
PRIRODNO - MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: Monografska dokumentacija
TD

Tip zapisa: Tekstualni štampani materijal
TZ

Vrsta rada: Master rad
VR

Autor: Milana Radmanović
AU

Mentor: dr Danijela Rajter - Ćirić
ME

Naslov rada: Stohastički procesi u genetici
NR

Jezik publikacije: Srpski (latinica)
JP

Jezik izvoda: s / en
JI

Zemlja publikovanja: Republika Srbija
ZP

Uže geografsko područje: Vojvodina
UGP

Godina: 2023.
GO

Izdavač: Autorski reprint
IZ

Mesto i adresa: Novi Sad, Trg D. Obradovića 4
MA

Fizički opis rada: (6/84/11/0/10/0/0)(broj poglavlja/broj strana/broj literarnih citata/broj tabela/broj slika/broj grafika/broj priloga)
FO:

Naučna oblast: Matematika
NO

Naučna disciplina: Stohastička analiza
ND

Ključne reči: Hardy - Weinbergov model, Moranov model, Wright - Fisher model, Koalescentna teorija, Galton - Watson proces grananja, Proces grananja sa višestrukim tipovima čvorova, aleli, lokus, genotip, učestalost, izumiranje, populacija, jedinka, potomak
PO, UDK

Čuva se: U biblioteci Departmana za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu
ČU

Važna napomena:
VN

Izvod: U ovom master radu opisano je pet ključnih modela u populacionoj genetici. Za svaki model su navedene pretpostavke i najbitniji zaključci do kojih se dubljom analizom dolazi, kao i mane jednog u odnosu na drugi (uporedivi) model. Kod prva tri modela (Hardy - Weinberg, Wright - Fisher, Moran) posmatrana populacija teži da dostigne stabilno stanje i izgubi genetsku raznolikost, iako je za to potrebno mnogo vremena. Koalescentna teorija nam omogućava da pratimo povratke unazad u vremenu i identifikujemo zajedničke pretke gena u populaciji. Proces grananja se koristi za modeliranje slučajnih procesa reprodukcije i nasleđivanja. Data je simulacija Galton - Watson procesa grananja u programskom jeziku "R", kao i primena procesa grananja sa višestrukim tipovima čvorova za optimalno vremensko otkrivanje raka jajnika kod žena.
IZ

Datum prihvatanja teme od strane NN veća: 22.09.2023.
DP

Datum odbrane:
DO

Članovi komisije:
ČK

Predsednik: dr Dora Seleši, redovni profesor Prirodno – matematičkog fakulteta u Novom Sadu
Mentor: dr Danijela Rajter - Ćirić, redovni profesor Prirodno – matematičkog fakulteta u Novom Sadu
Član: dr Sanja Rapajić, Prirodno – matematičkog fakulteta u Novom Sadu

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
KEY WORDS DOCUMENTATION

Accession number:
ANO

Identification number:
INO

Document type: Monograph type
DT

Type of record: Printed text
TR

Contents Code: Master's thesis
CC

Author: Milana Radmanović
AU

Mentor: dr Danijela Rajter - Ćirić
MN

Title: Stochastic Processes in Genetics
TI

Language of text: Serbian (Latin)
LT

Language of abstract: s / en
LA

Country of publication: Republic of Serbia
CP

Locality of publication: Vojvodina
LP

Publication year: 2023.
PY

Publisher: Author's reprint

PU

Publication place: Novi Sad, Trg D. Obradovića 4
PP

Physical description: (6/84/11/0/10/0/0)(chapters/ pages/ quotations/ tables/ pictures/ graphics/ enclosures)
PD

Scientific field: Mathematics
SF

Scientific discipline: Stochastic Analysis
SD

Subject/Key words: Hardy - Weinberg model, Moran model, Wright - Fisher model, Coalescent theory, Galton - Watson branching process, Multi-type branching process, alleles, locus, genotype, frequency, extinction, population, organism, offspring
SKW

Holding data: The Library of the Department of Mathematics and Informatics, Faculty of Science and Mathematics, University of Novi Sad
HD

Note:
N

Abstract: This master's thesis describes five key models in population genetics. For each model, assumptions and the most important conclusions reached through deeper analysis are provided, as well as the drawbacks of one compared to the other (comparable) model. In the first three models (Hardy - Weinberg, Wright - Fisher, Moran), the observed population tends to reach a stable state and lose genetic diversity, although it takes a very long time. Coalescent theory enables us to trace backward returns in time and identify common gene ancestors in the population. Branching process is used to model random processes of reproduction and inheritance. A simulation of the Galton-Watson branching process in the "R" programming language is given, as well as the application of the multi-type branching process for optimal time detection of ovarian cancer among women.
AB

Accepted by the Scientific Board on: 22.09.2023.
ASB

Defended:
DE

Thesis defend board:
DB

President: dr Dora Seleši, Full professor, Faculty of Science, University of Novi Sad

Mentor: dr Danijela Rajter – Ćirić, Full professor, Faculty of Science, University of Novi Sad
Member: dr Sanja Rapajić, Full professor, Faculty of Science, University of Novi Sad