Izumiranje linij

Milan Hladnik

Moderni izzivi poučevanja matematike
23. september 2011

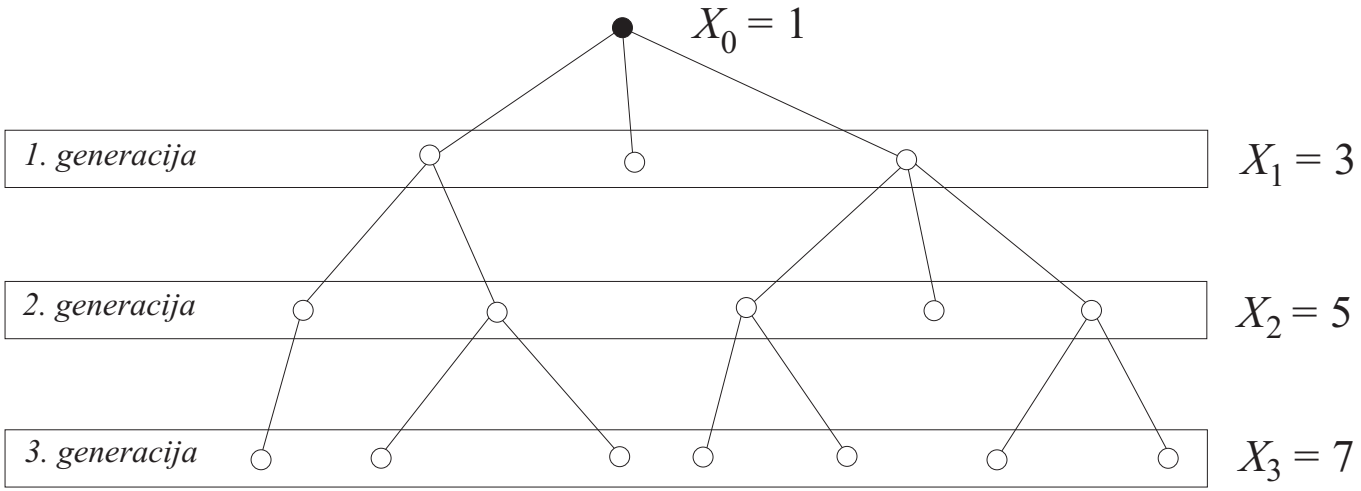
Namen seminarja

Podati preprost model o rasti števila moških potomcev iz generacije v generacijo. Moški so običajno nosilci priimkov, zato model pove, kakšne so možnosti za izginotje priimka oziroma izumrtje dane moške linije.

- Pričakovano število oseb v n-ti generaciji
- Lastnosti rodovne funkcije in verjetnost izumrtja
- Limitno vedenje rekurzivnega zaporedja in Watsonov paradoks
- Zgodovinske opombe
- Zgledi

Rodovno drevo

Predpostavimo: en oče, nekaj moških potomcev, nekaj sinov v drugi generaciji itd. Ena linija se nadaljuje iz generacije v generacijo.



Slika: Začetne generacije

Verjetnostni model oče-sin

Diskretna slučajna spremenljivka:

$$S : \begin{pmatrix} 0 & 1 & 2 & \dots & N \\ p_0 & p_1 & p_2 & \dots & p_N \end{pmatrix}$$

$p_k = P(S = k) > 0$: verjetnost, da ima oče k sinov
($k = 0, 1, 2, \dots, N$, N zelo veliko število ali $N = \infty$)
Velja $p_0 + p_1 + p_2 + \dots + p_N = 1$.

Porazdelitveni zakon: npr. *Poissonov* ali *geometrijski* ($N = \infty$), ni pomemben.

Pričakovano število sinov (**povprečje**): $r = E(S)$

n-ta generacija

X_n slučajna spremenljivka, ki šteje, koliko moških živi v n-ti generaciji ($n \geq 0$)

Predpostavke:

(a) $X_0 = 1$, linija izvira iz enega praočeta
(b) Vsak moški ima lahko $0, 1, 2, ..., N$ sinov, neodvisno od drugih
(c) Verjetnosti p_k, da ima moški k otrok, se ohranjajo ne glede na generacijo (homogenost linije)

Model za n-to generacijo

Model:

$$X_{n+1} = S_1 + S_2 + \ldots + S_{X_n} = \sum_{i=1}^{X_n} S_i, \quad n \geq 0,$$

kjer so $S_1, S_2, \ldots$, med seboj neodvisne slučajne spremenljivke, ki so porazdeljene enako kot spremenljivka S.

Npr. $X_0 = 1$, $X_1 = S$, $X_2 = S_1 + S_2 + \ldots + S_{X_1}$ itd.

Ena ali več linij

Ena linija (pogoj $X_0 = 1$):
Verjetnost, da je v n-ti generaciji j moških potomcev:

$$P(X_n = j) = P(X_n = j | X_0 = 1)$$

Več linij (pogoj $X_0 = k$):

$$P(X_n = j | X_0 = k), \text{ kjer je } k > 1.$$

Slednje pomeni verjetnost, da je v n-ti generaciji j moških potomcev, če jih je bilo na začetku k.

Pričakovano število potomcev v n-ti generaciji

Za eno linijo: $r_n = E(X_n) = E(X_n|X_0 = 1)$

Za k začetnih linij: $E(X_n|X_0 = k) = kE(X_n)$
(aditivnost matematičnega upanja)

Npr. $r_0 = 1$, $r_1 = E(X_1) = E(S) = r$, $E(X_1|X_0 = k) = kr$, ...

Posledica: $r_n = r^n$

Hitro vidimo direktno.

Pričakovano število potomcev - drugače

Po definiciji matematičnega upanja dobimo rekurzijo (za $n \geq 1$):

$$
\begin{aligned}
&E(X_n) = \sum_{j=0}^{\infty} jP(X_n = j) = \\
&\sum_{j=0}^{\infty} j \sum_{k=0}^{\infty} P(X_{n-1} = k) P(X_n = j | X_{n-1} = k) = \\
&\sum_{k=0}^{\infty} P(X_{n-1} = k) \sum_{j=0}^{\infty} jP(X_n = j | X_{n-1} = k) = \\
&\sum_{k=0}^{\infty} P(X_{n-1} = k) E(X_n | X_{n-1} = k) = \\
&\sum_{k=0}^{\infty} P(X_{n-1} = k) E(X_1 | X_0 = k) = \\
&\sum_{k=0}^{\infty} P(X_{n-1} = k) kr = r \sum_{k=0}^{\infty} kP(X_{n-1} = k) = rE(X_{n-1})
\end{aligned}
$$

oziroma rekurzijo

$$r_n = r \cdot r_{n-1}.$$

Od tod:

$$r_n = r^n$$

Kdaj linija izumre?

Dogodki:

$(X_n = 0)$ linija izumre do n-te generacije
$(X_n \neq 0)$ linija se ohrani vsaj do n-te generacije (vsota obeh je gotov dogodek)
$(X_n = 0)^k = (X_n = 0) \cdot \ldots \cdot (X_n = 0)$ (n-krat) vse linije izumrejo do n-te generacije, če imamo na začetku k neodvisnih linij

Verjetnosti izumrtja

Verjetnosti:

Verjetnost, da linija izumre do n-te generacije:

$$q_n = P(X_n = 0)$$

Npr. $q_0 = P(X_0 = 0) = 0$ in $q_1 = P(X_1 = 0) = p_0$ itd.

Verjetnost, da do n-te generacije izumre k linij:

$$P(X_n = 0 | X_0 = k) = P((X_n = 0)^k) = P(X_n = 0)^k = q_n^k$$

(neodvisnost linij!)

Rekurzija za verjetnost izumrtja linije

Za eno linijo in za $n \geq 1$ sklepamo po formuli za polno verjetnost:

$$P(X_{n+1} = 0) = \sum_{k=0}^{N} P(X_1 = k) P(X_{n+1} = 0 | X_1 = k) =$$

$$\sum_{k=0}^{N} P(X_1 = k) P(X_n = 0 | X_0 = k)$$

Torej

$$q_{n+1} = \sum_{k=0}^{N} p_k q_n^k$$

Rodovna funkcija

Definicija: $f(x) = \sum_{k=0}^{N} p_k x^k$

Lastnosti:
(i) $f(0) = p_0 > 0$, $f(1) = p_0 + p_1 + ... + p_N = 1$,
(ii) $f'(x) = \sum_{k=1}^{N} k p_k x^{k-1} > 0$ za $x \geq 0$,
(iii) $f'(0) = p_1$, $f'(1) = \sum_{k=1}^{N} k p_k = r$,
(iv) $f''(x) = \sum_{k=2}^{N} k(k-1) p_k x^{k-2} > 0$ za $x \geq 0$.

Zaradi vseh teh lastnosti funkcija f narašča in je konveksna, za vsak $x \in [0,1]$ pa velja

$$p_0 \leq f(x) \leq 1 \ , \ p_1 \leq f'(x) \leq r$$

Rekurzivno zaporedje

Zaporedje (q_n) je določeno z začetnim členom $q_0 = 0$ in rekurzivno formulo

$$q_{n+1} = f(q_n),\ n \geq 0.$$

Ker je f naraščajoča pozitivna funkcija, je zaporedje naraščajoče. Navzgor je omejeno z 1. Potem je konvergentno in konvergira k najmanjši negibni točki $q \leq 1$ funkcije f, tj. točki q, ki zadošča enačbi $q = f(q)$. Velja torej $\lim_{n\to\infty} q_n = q$.

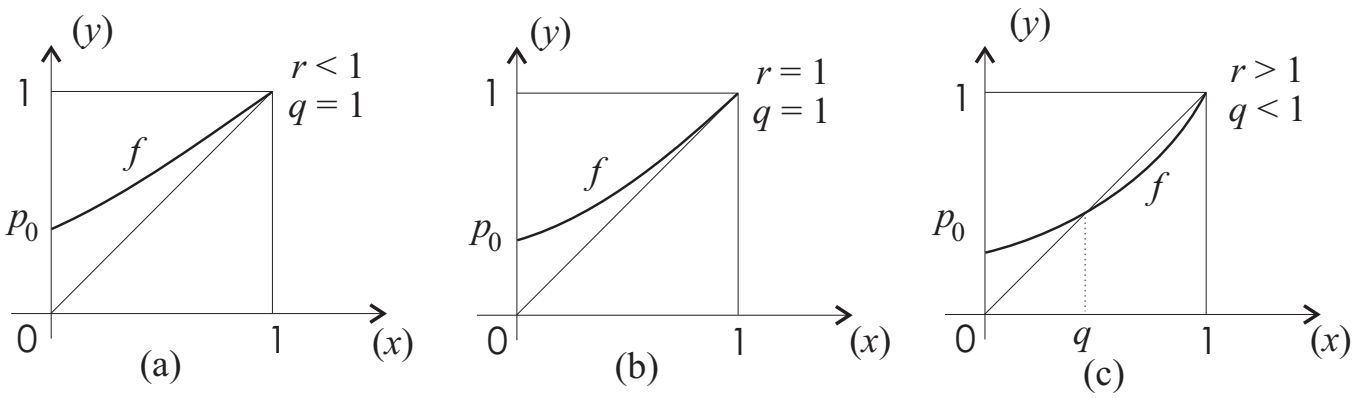
Tu pomeni:

q	verjetnost izumrtja,
$1-q$	verjetnost preživetja.

Limitno vedenje

Trije primeri (glej sliko):

(a) $r < 1$: $r^n \searrow 0$, $q = 1$.
Edina negibna točka funkcije f je enaka 1.
(b) $r = 1$: $r^n = 1$ za vsak n, $q = 1$.
To je po svoje presenetljivo, celo protislovno.
(c) $r > 1$: $r^n \nearrow \infty$, $q_n \to q < 1$.
Funkcija f ima dve negibni točki, poleg 1 še $q < 1$.



Slika: Grafi različnih primerov

Pogojna verjetnost

Opazujmo dogodke v n-ti generaciji samo *pri pogoju, da se linija do takrat ohrani*.

Pogojna verjetnost:

$$P(X_n = k | X_n \neq 0) = P((X_n = k)(X_n \neq 0))/P(X_n \neq 0) =$$

$$P(X_n = k)/P(X_n \neq 0),$$

če je $k \geq 1$, in 0, če je $k = 0$.

Torej je za $k \geq 1$

$$P(X_n = k) = P(X_n \neq 0)P(X_n = k | X_n \neq 0)$$

Pogojno matematično upanje

Matematično upanje:

$$E(X_n) = \sum_{k=0}^{\infty} kP(X_n = k) = P(X_n \neq 0) \sum_{k=1}^{\infty} kP(X_n = k | X_n \neq 0)$$

oziroma

$$E(X_n) = P(X_n \neq 0) E(X_n | X_n \neq 0)$$

—

Tu je $E(X_n | X_n \neq 0)$ pogojno matematično upanje slučajne spremenljivke X_n pri pogoju, da linija preživi do n-te generacije, torej povprečno število sinov za moške, ki sploh imajo sinove.

Pogojno povprečje po domače

X spremenljivka z vrednostmo $x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_n$, med katerimi je (prvih) m od nič različnih, ostale pa so enake nič.

Povprečje med *od nič različnimi* vrednostmi:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_m}{m}$$

Celotno povprečje:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{x_1 + x_2 + \ldots + x_m}{n} = \frac{m}{n} \cdot \frac{x_1 + x_2 + \ldots + x_m}{m}$$

Oziroma

$$E(X) = P(X \neq 0) E(X | X \neq 0)$$

Pogojno pričakovano število potomcev

Spomnimo se, da je $E(X_n) = r_n = r^n$.

Pričakovana moč n-te generacije pri pogoju, da linija preživi vsaj do nje:

$$E(X_n | X_n \neq 0) = r^n / P(X_n \neq 0)$$

V posebnem primeru pri $r = 1$ imamo torej

$$E(X_n | X_n \neq 0) = 1 / P(X_n \neq 0)$$

Watsonov paradoks

Zaradi $q_n \to q = 1$ velja

$$P(X_n \neq 0) = 1 - P(X_n = 0) = 1 - q_n \to 0$$

Če je $r = 1$, velja tudi

$$E(X_n | X_n \neq 0) = 1/P(X_n \neq 0) \to \infty$$

Watsonov paradoks:

Verjetnost, da linija traja vsaj do n, je zaradi $P(X_n \neq 0) \to 0$ *pri velikem n zelo majhna. Če pa se to vseeno zgodi, je pričakovano število članov n-te generacije zaradi* $E(X_n | X_n \neq 0) \to \infty$ *zelo veliko.*

Zgled 1: Majhno število priimkov

Dejansko lahko zlasti v manjših in izoliranih skupnostih, npr. v gorskih vaseh, pogosto opazimo, da je pri približno enakem številu pripadnikov vsake generacije veliko ljudi z enakim priimkom, različnih priimkov pa je razmeroma malo, kar je v skladu z Watsonovim paradoksom. Da se o tem prepričamo, zadošča obisk lokalnega pokopališča in bežen pregled priimkov na nagrobnikih.

Zgodovinske opombe

Henry William Watson (1827-1903) je na pobudo Francisa Galtona (1822-1911) podal prvo zadovoljivo rešitev problema, zakaj in kako izginjajo angleški aristokratski priimki.

Njun skupni članek o verjetnosti izumiranja družin *On the probability of extinction of families*, Journal of the Anthrop. Institute of Great Britain, vol. 4 (1874), 138-144, se šteje za začetek teorije ti. razvejitvenih stohastičnih procesov (Galton-Watsonovih procesov).

V poznih dvajsetih in tridesetih letih 20. stoletja so njuno delo dopolnili drugi (R. Fisher, J.B.S. Haldane, A.K. Erlang in J.F. Steffensen).

Ime razvejitveni procesi sta sicer vpeljala šele A.N. Kolmogorov in N.A. Dmitrijev leta 1947.

Galton

Francis Galton (1822-1911): angleški znanstvenik, geograf, meteorolog, izumitelj, (pred)genetik, eksperimentalni psiholog, antropolog in statistik.

Raziskoval JZ Afriko 1850-52, poimenoval pojav anticiklona, uvajal kvantitativno analizo, psihometrično testiranje, proučeval dednost različnih sposobnosti, opravil pionirsko delo pri uvedbi anket in podrobnih vprašalnikov, uvedel statistične pojme: standardna deviacija, korelacija in regresija, dosegel, da so uvedli metodo prstnih odtisov za identifikacijo, znan kot zagovornik evgenike (skoval ime), ustanovil Galtonov laboratory 1904 (prvi profesor Karl Pearson do 1933, drugi Ronald Fisher do 1943), skupaj s Pearsonom revijo *Biometrika* (1901).

Erlang

Danski matematik in inženir Agner Krarup Erlang (1878-1929) je imel zanimivo lastno izkušnjo z izginevanjem priimkov. Po materini strani je bil priimek Krarup že v njegovi mladosti zelo redek in tik pred tem, da izumre.

Erlang se je uveljavil s svojim modelom iz teorije množične strežbe. Ima tudi svojo verjetnostno porazdelitev. Ukvarjal se je predvsem s telefonskimi centralami in telefonskim prometom. Po njem se imenuje enota erlang za intenzivnost telefonskega prometa.

Zgled 2: Množična strežba

V teoriji množične strežbe (strežniki in stranke, ki slučajno prihajajo do strežnikov) pogosto nastajajo vrste čakajočih. Za zunanjega opazovalca so čakajoči nasledniki ('sinovi') stranke, ki je na vrsti. Ko zmanjka strank, blagajničarka zapre blagajno in začne šteti denar, linija 'izumre'.

O vedenju vrste odloča r, povprečno število novih strank, ki pridejo, medtem ko se eni streže. Če je $r < 1$, bo vrsta nekoč (v limiti) prazna. Če je $r > 1$, obstaja pozitivna verjetnost, da se to ne bo zgodilo. Če pa je $r = 1$, je malo verjetno, da bi se vrsta dolgo obdržala, toda če se to zgodi, je v njej veliko čakajočih.

Nauk: zaradi Watsonovega paradoksa ni dobro, da je r blizu 1. Sistem ne sme delovati blizu svoje optimalne zmogljivosti.

Zgled 3: Stare civilizacije

Znano je npr., da obstaja na **Kitajskem** okrog 200 najpogostejših priimkov, ki pokrivajo približno 96% celotne populacije. V resnici samo tri priimke nosi okrog 20% populacije (okrog 300 milijonov ljudi!). Ti trije priimki so (po rangiranju iz leta 2006) Li, Wang in Zhang. Po pogostosti sledijo priimki Zhao, Chen, Yang, Wu, Liu, Huang in Zhou. Teh deset prvih priimkov pokriva okrog 40% populacije.

Še bolj drastično se to vidi pri **Korejcih**:
Vsega skupaj imajo okrog 250 priimkov in samo trije najbolj pogosti (Kim, Lee, Park) pokrivajo kar 45% populacije.

Zgled 4: Nove civilizacije

V nasprotju s tem je na **Nizozemskem** npr. okrog 68.000 različnih priimkov z vsega skupaj več kot 100.000 variantami. Začeli so jih uporabljati šele po napoleonskih vojnah. Najpogostejši trije nizozemski priimki so De Jong, Jansen in De Vries, vendar je njihov skupni delež med vsemi samo 1.44%.

Še bolj ekstremen primer je **Tajska**, kjer je različnih priimkov skoraj toliko kot družin. Uporabljajo jih šele od leta 1920, oblasti zahtevajo, da ima vsaka družina svoj priimek, poleg tega ljudje svoje priimke, ne glede na poroke, menjavajo zelo pogosto, zato je njihov sistem priimkov zelo kompliciran. V vzorcu več kot 45.000 pregledanih ljudi so npr. odkrili 81% unikatnih priimkov.

Zgled 5: Slovenija

Konec leta 2008 je bilo v Sloveniji 90.294 različnih priimkov (85% redkih, tj. pod 5%, 68% unikatnih)

Trije najpogostejši priimki:
Novak 11.298 (12,5% = 1/8), Horvat 9.934, Kovačič 5.665

Sledijo:
Krajnc 5.661, Zupančič 5.044, Kovač 4.791, Potočnik 4.759, Mlakar 4.000, Vidmar, Kos, Golob, Turk, Božič, Kralj, Zupan, ...

Uvedba priimkov v 11. stoletju v Beneški republiki,
v 13. stoletju na Tržaškem, v 15. stoletju v osrednji Sloveniji

Zgled 6: Mitohondrijska Eva

Z modernimi raziskavami mitohondrijske DNK, ki se prenaša samo z matere na hčer, lahko danes genetiki sledijo biološki liniji daleč v preteklost. Ugotovili so, da je pred manj kot deset tisoč generacijami (tj. od 150.000 do 200.000 leti) živela ženska, pramati, iz katere izhajajo vse danes živeče osebe ženskega spola na svetu (mitohondrijska Eva). To ne pomeni, da takrat ni bilo na svetu drugih žensk, ampak da so vse druge ženske linije do danes izumrle.

V tem času je celotna človeška populacija menda obsegala nekaj tisoč posameznikov, ki so mnogo generacij živeli približno na istem območju brez velikih številčnih sprememb. Čez čas so različne linije izumrle; ostal je le en tip mitohondrijske DNK (ena linija), ki jo podedujemo še danes. Kot vemo, je človeška vrsta danes zelo številna (okrog sedem milijard prebivalcev).

Zgled 7: Y-kromosomski Adam

Podobno so z analizo DNK na moškem kromosomu Y, ki se podeduje le z očeta na sina, našli t.i. Y-kromosomskega Adama, praočeta vseh danes živečih moških. Zanimivo, da ni živel istočasno z mitohondrijsko Evo, ampak približno 100.000 let kasneje.

Najmlajši (tj. najpozneje živeči) skupni prastarši, se pravi skupni par prednikov, iz katerih izvira (po moški ali po ženski liniji) vsak izmed nas, pa naj bi živeli pred približno 5000 leti.

Zgled 8: Ostalo

Pojav izumiranja linij je dokaj splošen; lahko ga opazimo vsepovsod, npr. tudi pri prenašanju informacij, širjenju govoric, nekaterih bolezni, pri genetskih spremembah v populaciji itd.

Teoretični fizik Leo Szilard (ki je sodeloval pri projektu *Manhattan*) je konec tridestih let 20. stoletja kot model za izračun kritične mase prostih nevtronov pri kontrolirani verižni reakciji cepitve jeder uporabil ravno Galton-Watsonov model.

Körner

T. Körner, *The Pleasure of Counting*, Cambridge University Press, Cambridge 1988.

Vsebina: Snow in kolera v Londonu 1854, bitka za konvoje v Atlantiku med prvo in drugo svetovno vojno, odkritje radarja, Richardsonov model vojne, prava velikost živih bitij, teorija relativnosti (Lorentz in Einstein), številski sistemi in algoritmi, izumiranje linij in mitohondrijska Eva, kombinatorični problem maksimalnega pretoka: problem načrtovanja železnic (Ford in Fulkerson), uvedba gregorijanskega koledarja v Angliji, Alan Turing in zlom nemške kode Enigma, problem trgovskega potnika.

Wade

Za podrobnejše podatke v zvezi z mitohondrijsko Evo in Y-kromosomskim Adamom glej:

Internet (Google),

za širši pogled na sodobne genetsko podprte teorije o razvoju človeštva pa:

- Nickolas Wade, *Before the Down: Recovering the Lost History of Our Ancestors*, Penguin, 2007.

Knjiga poljudno opisuje najnovejša dognanja v zvezi s prazgodovinskim razvojem človeštva: od izhoda iz Afrike, naselitve Azije in Evrope, izginotja Neandertalcev, pojava jezika, udomačitve živali, nastanka ras, do prehoda v zgodovinsko dobo.