11. Galtonův – Watsonův proces větvení

11.1. Definice: Nechť jedinec tvořící nultou generaci může dát vznik 0, 1, 2, ... jedincům (potomkům) první generace. Analogicky každý jedinec z první generace může dát vznik 0, 1, 2, ... jedincům druhé generace atd. Přitom předpokládáme, že

a) počet potomků X náhodně zvoleného jedince má pravděpodobnostní funkci

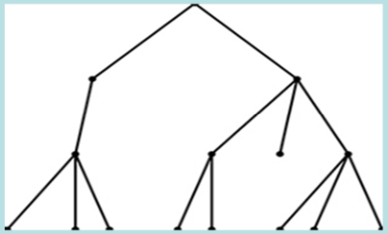
$$P(X = k) = \begin{cases} p_k & \text{pro } k = 0,1,2,\ldots \\ 0 & \text{jinak} \end{cases}$$

, která nezávisí na zvoleném jedinci ani na generaci, do níž přísluší;

b) jedinci z dané generace dávají vzniknout svým potomkům vzájemně nezávisle.

Označme X_n počet jedinců n-té generace (speciálně je $X_0 = 1$). Za uvedených předpokladů posloupnost náhodných veličin $\{X_n ; n \in N_0\}$ tvoří homogenní markovský řetězec s množinou stavů J = {0, 1, 2, ...}. Tento řetězec se nazývá Galtonův – Watsonův proces větvení.

11.2. Označení: Zavedeme náhodné veličiny U_{nk}, k = 1, 2, ... které mají stejné rozložení jako náhodná veličina X_1 a jsou stochasticky nezávislé jak mezi sebou, tak na veličinách X_0, X_1, ... Veličina U_{nk} udává počet potomků k-tého jedince v n-té generaci. Je zřejmé, že
$X_{n+1} = U_{n1} + U_{n2} + ... + U_{nX_n}$.
Ilustrace:



11.3. Věta: Pravděpodobnosti přechodu $p_{ij} = P(X_n = j / X_{n-1} = i)$ splňují vztah: $\forall i, j \in J : p_{ij} = \{p_j\}^{i*}$

s počáteční podmínkou $p_{0j} = \begin{cases} 1 \text{ pro } j = 0 \\ 0 \text{ jinak} \end{cases}$, tedy matice přechodu má tvar

$$P = \begin{pmatrix} 1 & 0 & 0 & \dots \\ p_0 & p_1 & p_2 & \dots \\ p_0^{\,2} & 2p_0p_1 & p_1^{\,2} + 2p_0p_2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}.$$

Důkaz: $p_{ij} = P(X_n = j / X_{n-1} = i) = P\left(\sum_{k=1}^{X_{n-1}} U_{n-1,k} = j / X_{n-1} = i\right) = P\left(\sum_{k=1}^{i} U_{n-1,k} = j\right) = \{p_j\}^{i*}$ podle věty 10.10.

11.4. Příklad: Nechť $\{X_n; n \in N_0\}$ je Galtonův – Watsonův proces větvení s množinou stavů J = {0, 1, 2, ...} a vektorem počátečních pravděpodobností **p**(0) = (¼, ¼ , ½ , 0, ...). Najděte matici přechodu **P**.

Řešení:

$$P = \begin{pmatrix} 1 & 0 & 0 & \dots \\ p_0 & p_1 & p_2 & \dots \\ p_0^{\ 2} & 2p_0p_1 & 2p_0p_2 + p_1^{\ 2} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots \\ 1/4 & 1/4 & 1/2 & \dots \\ 1/16 & 2/16 & 5/16 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

11.5. Věta: Pro pravděpodobnostní vytvořující funkci náhodné veličiny X_{n+1} platí:

$$g_{X_{n+1}}(z) = \begin{cases} g_{X_n}(g_X(z)) \text{pro} n = 1, 2, \dots \\ 0 \text{pro} n = 0 \end{cases}$$

, kde $g_X(z)$ je pravděpodobnostní vytvořující funkce náhodné veličiny X_1.

Důkaz: Protože $X_{n+1} = \sum_{k=1}^{X_n} U_{nk}$ je součet náhodného počtu náhodných veličin, tvrzení plyne z věty 10.13.

11.6. Příklad: Nechť $\{X_n; n \in N_0\}$ je Galtonův – Watsonův proces větvení, přičemž náhodná veličina X_1 se řídí rozložením A(ϑ), tedy její pravděpodobnostní vytvořující funkce má tvar $g_X(z) = 1 - \vartheta + z\vartheta$. Najděte pravděpodobnostní vytvořující funkci náhodné veličiny X_n.

Řešení:

$$g_X(z) = 1 - \vartheta + z\vartheta = 1 - \vartheta(1 - z)$$

$$g_{X_2}(z) = g_X(g_X(z)) = 1 - \vartheta + \vartheta(1 - \vartheta + z\vartheta) = 1 - \vartheta^2(1 - z)$$

$$g_{X_3}(z) = g_{X_2}(g_X(z)) = 1 - \vartheta^2\left[1 - (1 - \vartheta + z\vartheta)\right] = 1 - \vartheta^3(1 - z)$$

Obecně: $g_{X_n}(z) = 1 - \vartheta^n(1 - z)$ pro n = 1, 2, ...

11.7. Věta: Nechť μ je střední hodnota a σ^2 je rozptyl náhodné veličiny X_1. Pak pro střední hodnotu a rozptyl náhodné veličiny X_n platí:

$$E(X_n) = \mu^n, \quad D(X_n) = \begin{cases} \dfrac{\sigma^2\mu^{n-1}\left(\mu^n - 1\right)}{\mu - 1} \text{pro} \mu \neq 1 \\ n\sigma^2 \text{pro} \mu = 1 \end{cases}.$$

Důkaz: Provedeme matematickou indukcí. Z věty 10.16. plyne:

$$E(X_{n+1}) = E(U_{n+1,1} + \ldots + U_{n+1,X_n}) = E(X_n)\mu = \mu^n\mu = \mu^{n+1}.$$

Nechť $\mu \neq 1$. Pak

$$D(X_{n+1}) = D(U_{n+1,1} + \ldots + U_{n+1,X_n}) = D(X_n)\mu^2 + E(X_n)\sigma^2 = \frac{\sigma^2\mu^{n-1}\left(\mu^n - 1\right)}{\mu - 1}\mu^2 + \mu^n\sigma^2 =$$

$$= \frac{\sigma^2\mu^{2n+1} - \sigma^2\mu^{n+1} + \sigma^2\mu^{n+1} - \sigma^2\mu^n}{\mu - 1} = \frac{\sigma^2\mu^n\left(\mu^{n+1} - 1\right)}{\mu - 1}$$

Nechť $\mu = 1$.

Pak $D(X_{n+1}) = D(X_n) + \sigma^2 = n\,\sigma^2 + \sigma^2 = (n + 1)\,\sigma^2$.

11.8. Příklad: Pro zadání příkladu 11.4. vypočtěte střední hodnotu a rozptyl počtu potomků v n-té generaci.
(Připomeňme, že vektor počátečních pravděpodobností **p**(0) = (¼, ¼ , ½ , 0, ...).)
Řešení:

$$E(X_1) = \mu = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} = \frac{5}{4},$$

$$D(X_1) = \sigma^2 = E\left(X_1^{\,2}\right) - [E(X_1)]^2 = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{2} - \left(\frac{5}{4}\right)^2 = \frac{11}{16}$$

$$E(X_n) = \left(\frac{5}{4}\right)^n$$

$$D(X_n) = \frac{\sigma^2 \mu^{n-1}\left(\mu^n - 1\right)}{\mu - 1} = 4 \cdot \frac{11}{16} \cdot \left(\frac{5}{4}\right)^{n-1} \cdot \left[\left(\frac{5}{4}\right)^n - 1\right] = \frac{11}{4} \cdot \left(\frac{5}{4}\right)^{n-1} \cdot \left[\left(\frac{5}{4}\right)^n - 1\right]$$

11.9. Věta: Označme $q_n = P(X_n = 0)$ pravděpodobnost vyhynutí v n-té generaci. Pak platí:
$q_n = g_{X_n}(0)$.

Důkaz: $g_{X_n}(z) = \sum_{k=0}^{\infty} P(X_n = k) z^k \Rightarrow g_{X_n}(0) = P(X_n = 0) = q_n$.

11.10. Věta: Nechť $0 < p_0 < 1$. (Krajní případy $p_0 = 0$ a $p_0 = 1$ vylučujeme, protože pro $p_0 = 0$ je $q_n = 0$ pro všechna n = 1, 2, ... a pro $p_0 = 1$ je $q_n = 1$ pro všechna n = 1, 2, ...)

a) Je-li $\mu \le 1$, pak $\lim\limits_{n\to\infty} q_n = 1$.

b) Je-li $\mu > 1$, pak $\lim\limits_{n\to\infty} q_n = \xi$, kde $\xi \in (0,1)$ je nejmenší kladný kořen rovnice $z = g_X(z)$.

Interpretace: Je-li $\mu \le 1$, pak s pravděpodobností 1 proces dosáhne jen konečně mnoha generací. Je-li $\mu > 1$, pak s pravděpodobností ξ proces dosáhne konečně mnoha generací a s pravděpodobností $1 - \xi$ dosáhne nekonečně mnoha generací.

Důkaz: nebudeme provádět.

11.11. Příklad: Pro Galtonův – Watsonův proces z příkladu 11.4. najděte limitní hodnotu pravděpodobnosti vyhynutí.

Řešení:

V příkladu 11.8. bylo vypočteno, že $E(X_1) = \frac{5}{4}$. Protože $\frac{5}{4} > 1$, podle tvrzení b) věty 11.10.

$\lim\limits_{n\to\infty} q_n = \xi$, kde $\xi \in (0,1)$ je nejmenší kladný kořen rovnice $z = g_X(z) = \sum_{k=0}^{2} p_k z^k =$

$= \frac{1}{4} + \frac{1}{4}z + \frac{1}{2}z^2 \Rightarrow 4z = 1 + z + 2z^2, 2z^2 - 3z + 1 = 0 \Rightarrow z_{12} = \frac{3 \pm \sqrt{9-8}}{4} = \left\langle \begin{matrix} 1 \\ \frac{1}{2} \end{matrix} \right.$. Podmínku splňuje kořen ½,

tedy limitní hodnota pravděpodobnost i vyhynutí je 0,5.

11.12. Poznámka: Předchozí výsledky lze snadno zobecnit na případ, kdy nultá generace je tvořena $k_0 \geq 1$ jedinci. Pak pro n = 1, 2, ... platí.

a) $E(X_n) = k_0 \mu^n$

b) $$D(X_n) = \begin{cases} k_0 \dfrac{\sigma^2 \mu^{n-1}(\mu^n - 1)}{\mu - 1} \text{ pro } \mu \neq 1 \\ k_0 n \sigma^2 \text{ pro } \mu = 1 \end{cases}$$

c) $$\lim_{n\to\infty} q_n = \begin{cases} 1 \text{ pro } \mu \leq 1 \\ \xi^{k_0} \text{ pro } \mu > 1 \end{cases}$$

Lze si totiž představit, že vedle sebe se navzájem nezávisle větví k_0 populací, z nichž každá vznikla z právě jednoho jedince.

Proces zániku příjmení

V r. 1845 francouzský statistik **Irenée Jules Bienaymé** uveřejnil článek „O zákonu násobení a trvání rodin“. V něm si položil otázku, jaká je pravděpodobnost, že muž bude mít po n generacích mužské potomky nesoucí jeho příjmení. Uvádí zde, že je-li průměrný počet synů jednoho muže nejvýše 1, je pravděpodobnost zániku příjmení 1. Je-li však průměrný počet synů větší než 1, není zánik příjmení jistý jev a pravděpodobnost zániku příjmení lze vypočítat jako řešení jisté algebraické rovnice.

Na Bienaymého práci navázal v roce 1847 jeho přítel, matematik a ekonom **Antoine-Augustin Cournot**. Uvedeme zde speciální případ, kterým se Cournot zabýval.

Předpokládejme, že každý muž má nejvýše dva syny.
p_0 … pravděpodobnost, že nemá žádného syna
p_1 … pravděpodobnost, že má jednoho syna
p_2 … pravděpodobnost, že má dva syny
$p_0 + p_1 + p_2 = 1$
Pravděpodobnost vymření v 1. generaci: $q_1 = p_0$
Pravděpodobnost vymření ve 2. generaci: mohou nastat tři různé situace.

1. Rod vymře již v 1. generaci s pravděpodobností $q_1 = p_0$.
2. Muž má právě jednoho syna, který nemá žádné mužské potomky. To nastane s pravděpodobností $p_1 q_1$.
3. Muž má právě dva syny, žádný z nich nemá mužské potomky. K tomu dojde s pravděpodobností $p_2 q_1^2$.

Celkem: $q_2 = p_0 + p_1 q_1 + p_2 q_1^2$.
Pravděpodobnost vymření ve 3. generaci: opět mohou nastat tři situace, stejně jako ve 2. generaci. Jejich pravděpodobnosti jsou p_0, $p_1 q_2$, $p_2 q_2^2$. Tedy
$q_3 = p_0 + p_1 q_2 + p_2 q_2^2$.
Obecně: $q_n = p_0 + p_1 q_{n-1} + p_2 q_{n-1}^2$.

Platí, že posloupnost $\{q_n\}_{n=1}^{\infty}$ pravděpodobností vymření je rostoucí posloupnost, $q_n \le 1$ pro $\forall n \in N$. Tato posloupnost má limitu $x \le 1$, která je řešením kvadratické rovnice $x = p_0 + p_1 x + p_2 x^2$. Protože $p_0 + p_1 + p_2 = 1$, vyjádříme p_1 jako $p_1 = 1 - p_0 - p_2$. Pak

$$x = p_0 + (1 - p_0 - p_2)x + p_2 x^2,$$

$$0 = p_0(1-x) - p_2 x(1-x) = (1-x)(p_0 - p_2 x) / : p_2$$

$$0 = (1-x)\left(\frac{p_0}{p_2} - x\right)$$

Odtud je vidět, že rovnice má dva kořeny $x = 1$ a $x = \frac{p_0}{p_2}$.

Dáme tyto kořeny do souvislosti se střední hodnotou počtu synů libovolného muže: $\mu = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 = p_1 + 2p_2 = 1 - p_0 - p_2 + 2p_2 = 1 - p_0 + p_2$

Mohou nastat tři různé situace.

Je-li $\mu < 1$, tj. $1 - p_0 + p_2 < 1$, pak $\frac{p_0}{p_2} > 1$. Avšak x je pravděpodobnost, tedy jedinou možnou limitou je 1.

Je-li $\mu = 1$, tj. $1 - p_0 + p_2 = 1$, pak $\frac{p_0}{p_2} = 1$ a limitou je 1.

Je-li $\mu > 1$, tj. $1 - p_0 + p_2 > 1$, pak $\frac{p_0}{p_2} < 1$ a limita je $\frac{p_0}{p_2}$.

Zamená to, že při střední hodnotě počtu synů $\mu \leq 1$ příjmení určitě zanikne, zatímco při $\mu > 1$ příjmení zanikne jenom s pravděpodobností $\frac{p_0}{p_2}$.

Praktická aplikace Galtonova – Watsonova procesu větvení v demografii

Výchozí data

Budeme vyšetřovat sled generací ženské populace Československa. Máme k dispozici údaje z roku 1961, které popisují rozdělení žen ve věkovém intervalu (45 let, 50 let) (tedy na konci reprodukčního období) podle počtu živě narozených dětí. Žen v tomto věkovém intervalu bylo 450259.

počet dětí	označení	počet žen
0	c0	65387
1	c1	78901
2	c2	136150
3	c3	79878
4	c4	39387
5	c5	19856
6	c6	15365
7	c7	7683
8	c8	3841
9	c9	1921
10	c10	960
11	c11	480
12	c12	240
13	c13	120
14	c14	60
15	c15	30

Tabulka 1 - výchozí data

Stanovení pravděpodobnosti narození dcery, která se dožije reprodukčního věku 25 let

Zajímají nás pouze potomci ženského pohlaví. Je známo, že v r. 1961 byl poměr počtu živě narozených chlapců k počtu živě narozených dívek 1,055 (tzv. ukazatel maskulinity). Tedy pravděpodobnost narození dívky je $\frac{1}{1+1{,}055}$.

Dále je známo z úmrtnostních tabulek ČSR pro rok 1961, že pravděpodobnost dožití 25 let pro ženu je 0,96788. Tedy hodnotu h = 0,96788 . $\frac{1}{1+1{,}055}$ = 0,470988 lze považovat za pravděpodobnost, že živě narozený potomek je dcera, která se dožije věku 25 let.

Rozložení počtu žen podle počtu dcer, které se dožijí reprodukčního věku 25 let

Nechť c_i je počet žen s i potomky. Z vlastností binomického rozložení s parametry i a h plyne, že z počtu c_i připadá $(1-h)^i c_i$ na ženy s žádnou 25 letou dcerou, $ih(1-h)^{i-1} c_i$ na ženy s právě jednou 25 letou dcerou atd.

počet potomků	počet žen	rozdělení žen podle počtu 25 letých dcer			
		0	1	2	…
0	c_0	c_0	0	0	…
1	c_1	$(1-h)c_1$	hc_1	0	…
2	c_2	$(1-h)^2c_2$	$2h(1-h)c_2$	h^2c_2	…
3	c_3	$(1-h)^3c_3$	$3h(1-h)^2c_3$	$3h^2(1-h)c_3$	…
4	c_4	$(1-h)^4c_4$	$4h(1-h)^3c_4$	$6h^2(1-h)^2c_4$	…
.	.	.	.	.	…
.	.	.	.	.	…
.	.	.	.	.	…
suma	c	p_0c	p_1c	p_2c	

Tabulka 2 – rozdělení p_0, p_1, ... žen podle počtu 0, 1, ... 25 letých dcer konstruované na základě počtů c_0, c_1, ... žen s 0, 1, ... potomky.

Numerické vyhodnocení tabulky 2:

počet potomků	počet žen	rozdělení žen podle počtu 25 letých dcer					
		0	1	2	3	4	…
0	65387	65387	0	0	0	0	…
1	78901	41740	37161	0	0	0	…
2	136150	38102	67846	30202	0	0	…
3	79878	11826	31586	28121	8645	0	…
4	39387	3045	10985	14671	8708	1938	0
.	.	.	.		.		…
.	.	.	.		.		…
.	.	.	.		.		…
suma	450259	161420	153833	85789	31330	11549	

Na základě této tabulky lze stanovit pravděpodobnosti p_0, p_1, p_2, …

$$p_0 = \frac{161420}{450259} = 0{,}358504,$$

$p_1 = 0{,}3416544$, $p_2 = 0{,}1905325$, $p_3 = 0{,}0695821$, …

Střední hodnota $\mu = p_1 + 2p_2 + 3p_3 + \ldots = 0{,}341655 + 2.0{,}190533 + 3.0{,}069583 + \ldots = 1{,}111166 > 1$

Matice pravděpodobností přechodu

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ p_0 & p_1 & p_2 & \cdots \\ {p_0}^2 & 2p_0p_1 & {p_1}^2 + 2p_0p_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

V našem případě tedy dostáváme:

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ 0{,}358504 & 0{,}341655 & 0{,}190533 & 0{,}069583 & \cdots \\ 0{,}128525 & 0{,}244969 & 0{,}253342 & 0{,}180085 & \cdots \\ 0{,}046077 & 0{,}131733 & 0{,}199067 & 0{,}206734 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Pravděpodobnostní rozložení počtu 25 letých dcer v n-té generaci

Pro vektor absolutních pravděpodobností platí zákon evoluce: $\mathbf{p}(n) = \mathbf{p}(0)\mathbf{P}^n$.
V následující tabulce jsou uvedeny složky vektoru $\mathbf{p}(n)$ pro n = 0, 1, ..., 10

n \ i	0	1	2	3	4	5	6	7	...
0	1,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	...
1	0,358504	0,341655	0,190533	0,065830	0,025651	0,009428	0,003191	0,001013	...
2	0,509170	0,174495	0,131214	0,078222	0,046294	0,026754	0,015134	0,008447	...
3	0,593159	0,105974	0,090446	0,064523	0,045799	0,031962	0,022014	0,015040	...
4	0,646756	0,071021	0,065064	0,051054	0,039920	0,030799	0,023537	0,017881	...
5	0,683843	0,050729	0,048627	0,040485	0,033619	0,027604	0,022493	0,018246	...
6	0,710930	0,037884	0,037477	0,032484	0,028103	0,024071	0,020485	0,017367	...
7	0,731494	0,029234	0,029590	0,026409	0,023533	0,020781	0,018246	0,015970	...
8	0,747563	0,023131	0,023824	0,021736	0,019805	0,017894	0,016085	0,014416	.
9	0,760402	0,018665	0,019489	0,180087	0,016768	0,015422	0,014117	0,012887	.
10	0,770845	0,015300	0,016151	0,015194	0,014281	0,013322	0,012371	0,011460	
.	.	.	.	.	.	.	.	.	.

Tabulka 3 – pravděpodobnostní rozložení $p_i(n)$ počtu i 25 letých dcer v n-té generaci

V prvním sloupci jsou pravděpodobnosti vyhynutí (v ženské linii). Je vidět, že v 10. generaci je pravděpodobnost vyhynutí $p_0(10) = 0{,}770845$ hodně vysoká. Naproti tomu pravděpodobnosti libovolného nenulového počtu dcer jsou velice malé (např. pro jednu dceru $p_1(10) = 0{,}015300$ a s rostoucím počtem generací se stále snižují.

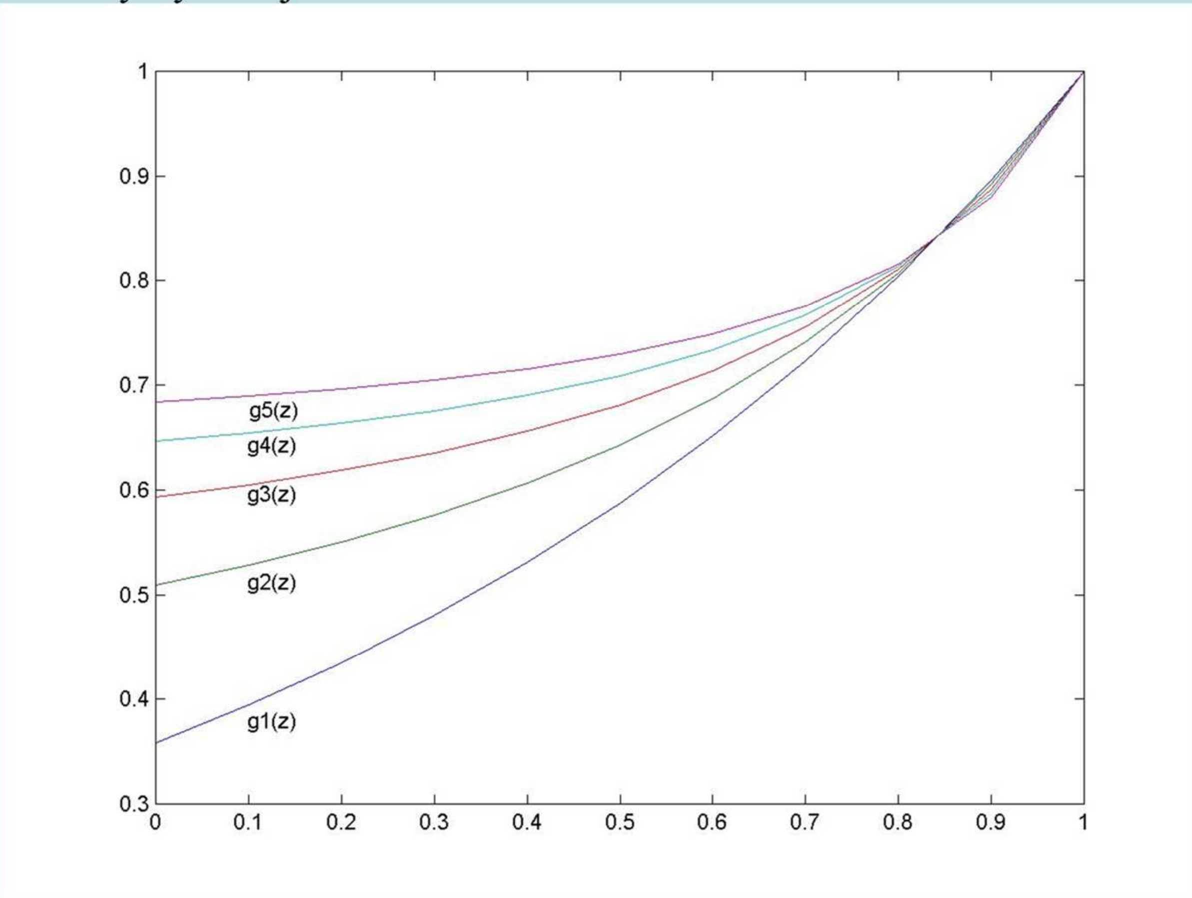
Vytvořující funkce počtu 25 letých dcer v n-té generaci

$$g_{X_n}(z) = \sum_{k=0}^{\infty} p_{nk} z^k$$

z	$g_1(z)$	$g_2(z)$	$g_3(z)$	$g_4(z)$	$g_5(z)$
0,0	0,358504	0,509170	0,593159	0,646756	0,683843
0,1	0,394647	0,528015	0,604730	0,654564	0,689446
0,2	0,435057	0,550027	0,618573	0,664047	0,696322
0,3	0,480260	0,575893	0,635302	0,675717	0,704892
0,4	0,530872	0,606507	0,655773	0,690318	0,715785
0,5	0,587619	0,643056	0,681205	0,708964	0,729979
0,6	0,651360	0,687140	0,713398	0,733401	0,749073
0,7	0,723115	0,740967	0,755101	0,766513	0,775872
0,8	0,804109	0,807566	0,810729	0,813400	0,815726
0,9	0,895811	0,891733	0,887770	0,883881	0,880022
1,0	1,000000	1,000000	1,000000	1,000000	1,000000

Tabulka 4 – hodnoty vytvořující funkce $g_n(z)$ počtu 25 letých dcer v n-té generaci

Průběhy vytvořujících funkcí



Pravděpodobnost vyhynutí

Limitní hodnotu pravděpodobnosti vyhynutí lze získat jako nejmenší kladný kořen rovnice $z = g(z)$, kde $g(z) = g_1(z) = p_0 + p_1 z + p_2 z^2 + \ldots$

V našem případě řešíme rovnici $z = 0{,}358504 + 0{,}341655z + 0{,}190533z^2 + 0{,}069583z^3 + \ldots$

Výsledek: $\xi = 0{,}834043$.

Je zajímavé posoudit rychlost konvergence posloupnosti q_n pravděpodobnosti vyhynutí potomků v ženské linii v jednotlivých generacích k příslušné limitní hodnotě $\xi = 0{,}834043$ - viz 1. sloupec v tabulce 3.

Pro zajímavost ještě uvedeme limitní hodnoty pravděpodobnosti vyhynutí potomků v ženské linii pro různé země (údaje jsou z roku 1960).

země	ξ
Peru	0,2620
Japonsko	0,3242
Mexiko	0,4066
Maďarsko	0,7130
USA	0,8209