MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁRSKA PRÁCA

Matej Rada

Pólyov-Aeppliho proces

Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. RNDr. Zbyněk Pawlas, Ph.D.
Študijný program: Finanční matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V .... dne ....

Podpis autora

Poďakovanie by som rád venoval hlavne svojim rodičom a starým rodičom, ktorí mi umožnili štúdium na vysokej škole. Nezabúdam pritom na ochotu a obetavosť môjho školiteľa, s ktorým sme sa intenzívne venovali tejto práci. Preto moje poďakovanie patrí samozrejme aj jemu. Okrem toho by som sa rád poďakoval Matematicko-Fyzikální fakulte Karlovej univerzity za vysokú kvalitu štúdia a za organizáciu mnohých mimoškolských aktivít.

Názov práce: Pólyov-Aeppliho proces

Autor: Matej Rada

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúcí bakalárskej práce: doc. RNDr. Zbyněk Pawlas, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Táto práca je venovaná skúmaniu Pólyovho-Aeppliho procesu a zároveň Pólyovho-Aeppliho rozdelenia, ktoré sa v tomto procese využíva. Pri Pólyovom-Aeppliho rozdelení sú uvedené dva tvary pravdepodobnostnej funkcie – rekurzívny a explicitný. Popísané sú aj vlastnosti tohto rozdelenia. Pólyov-Aeppliho proces je zadefinovaný rôznymi spôsobmi a odvodené sú vzťahy medzi týmito definíciami. Takisto sú popísané vlastnosti tohto procesu. Praktická časť je venovaná rôznym spôsobom, ako odhadnúť parametre Pólyovho-Aeppliho rozdelenia pre počty odohraných zápasov účastníkov grandslamových turnajoch. Nakoniec je uvedené porovnanie týchto spôsobov.

Kľúčové slová: čítací proces, geometrické rozdelenie, Pólyov-Aeppliho proces, zložené Poissonovo rozdelenie, zložený Poissonov proces

Title: Pólya-Aeppli process

Author: Matej Rada

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zbyněk Pawlas, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis examines the Pólya-Aeppli process along with Pólya-Aeppli distribution, which is utilized in this process. Two forms of probability function of Pólya-Aeppli distribution are listed – recursive form and explicite form. The characteristics of this distribution are also derived. Pólya-Aeppli process is defined in different ways and relations between those definitions are shown. Also the characteristics of this process are derived. In practical part of this thesis, different methods of parameter estimation of Pólya-Aeppli distribution are shown. The data, that are modeled by this distribution are number of played matches by participants of grandslam tournaments. Finally, the comparison of these methods is shown.

Keywords: counting process, geometric distribution, Pólya-Aeppli process, compound Poisson distribution, compound Poisson process

Obsah

Úvod

Pri pozorovaní nejakého sledu udalostí v čase nás v praxi často zaujíma, ako sa vyvíjal ich celkový počet. Tieto udalosti prichádzajú v čase v náhodných časových intervaloch, preto musíme použiť stochastické metódy. Nástroj na takéto modelovanie ponúkajú čítacie procesy. Štandardný, široko používaný model je Poissonov proces. Jedna z najdôležitejších vlastností tohoto procesu je jeho ekvidisperzia. To znamená že jeho rozptyl a stredná hodnota počtu udalostí sú si rovné. V praxi je ale táto vlastnosť nereálnym predpokladom pri pozorovaných udalostiach. Preto je vhodné použiť na reálne dáta zložený Poissonov proces. Príkladom takého procesu je Pólyov-Aeppliho proces. V tejto práci si odvodíme, akými spôsobmi vieme Pólyov-Aeppliho proces zadefinovať, ďalej si ukážeme vzťahy medzi týmito definíciami a nejaké z jeho vlastností.

V prvej kapitole zavedieme všetku teóriu, z ktorej budeme v práci vychádzať. Jedná sa hlavne o teóriu k pravdepodobnostným rozdeleniam a náhodným procesom.

V druhej kapitole zadefinujeme geometrické Poissonovo rozdelenie, nazývané niekedy aj Pólyovo-Aeppliho rozdelenie, ktoré využíva Pólyov-Aeppliho proces. Geometrické Poissonovo rozdelenie je špeciálny prípad zloženého Poissonovho rozdelenia, v ktorom sa jednotlivé sčítance riadia geometrickým rozdelením. Ukážeme si rekurzívny, ale aj implicitný vzorec na výpočet jeho pravdepodobností a porovnáme, ako vyzerá graf pravdepodobnostného rozdelenia pri rôznych parametroch.

V tretej kapitole sa budeme venovať samotnému procesu. Ukážeme si štyri uhly pohľadu, ako sa vieme pozerať na tento proces a to ako na proces s nezávislými prírastkami, zložený Poissonov proces, proces obnovy a Markovov reťazec. Všetky tieto definície si budú navzájom ekvivalentné, to znamená že budú definovať ten istý proces. Keď už budeme mať Pólyov-Aeppliho proces zavedený, ukážeme si niektoré z jeho jednoduchších vlastností, ako napríklad strednú hodnotu a rozptyl, ale aj niektoré komplexnejšie vlastnosti, napríklad distribučnú funkciu doby čakania na príchod udalosti.

Posledná kapitola bude venovaná praktickej časti, v ktorej si porovnáme dva spôsoby odhadovania parametrov pre geometrické Poissonovo rozdelenie, ktoré ako sme si už spomínali je kľúčové pre Pólyov-Aeppliho proces. Toto rozdelenie vystihuje dáta s takou povahou, že z nejakého počtu udalostí, ktorých je náhodný počet, nastane s nejakou pravdepodobnosťou úspechu ďalší náhodný počet pozorovaných udalostí. V článku Özel a Inal (2010) autori používajú toto rozdelenie na modelovanie počtu fatalít pri autonehodách, v článku Nuel (2008) zase na modelovanie DNA substitúcií baktérie E. coli. My budeme čerpať zo sveta športu, presnejšie sa budeme snažiť modelovať pravdepodobnosti počtu odohraných kôl v play-off grandslamov, ktorý je tenista schopný dosiahnuť a ktorý sa aspoň raz v kariére dostal na post svetovej jednotky.

1. Potrebná teória

V tejto kapitole si pripomenieme pár základných pojmov, definícií, lemmat a tvrdení, ktoré budú kľúčové pre zavedenie geometrického Poissonovho (alebo aj Pólyovho-Aeppliho) rozdelenia a následne Pólyovho-Aeppliho náhodného procesu.

1.1 Pravdepodobnostné rozdelenia

Pri Pólyovom-Aeppliho rozdelení (a teda aj procese) sa využívajú niektoré diskrétne rozdelenia. Uvedieme si preto ich definície a zopár užitočných vlastností.

Začneme pripomenutím definície geometrického rozdelenia.

Definícia 1. *Nech $0 < \theta \leq 1$ a majme diskrétnu náhodnú veličinu X. Hovoríme, že X má geometrické rozdelenie s parametrom θ, pokiaľ*

$$\mathbb{P}(X = k) = (1-\theta)^{k-1}\theta, \quad k = 1, 2, 3, \ldots \tag{1.1}$$

Prvý moment a druhý centrálny moment vyzerajú nasledovne:

$$\mathsf{E}\,[X] = \frac{1}{\theta} \quad a \quad \mathrm{var}[X] = \frac{1-\theta}{\theta^2}. \tag{1.2}$$

Značíme $X \sim Geo(\theta)$.

Pokiaľ považujeme θ za pravdepodobnosť úspechu, udáva geometrické rozdelenie pravdepodobnosť, že prvý úspech nastane po k nezávislých pokusoch, každý s pravdepodobnosťou úspechu θ. Takýto predpis geometrického rozdelenia sa používa pre modelovanie počtu pokusov do a vrátane prvého úspechu. Definíciu 1 vieme modifikovať tak, aby k reprezentovalo iba počet neúspešných pokusov, kým nastane prvý úspech a to nasledovne:

$$\mathbb{P}(Y = k) = \mathbb{P}(X = k+1) = (1-\theta)^k\theta, \quad k = 0, 1, 2, \ldots$$

V nasledujúcej lemme sa pozrieme na ďalšiu charakteristiku geometrického rozdelenia a to na jeho pravdepodobnostnú vytvárajúcu funkciu.

Lemma 1. *Majme náhodnú veličinu X z geometrického rozdelenia s parametrom $\theta \in (0,1]$. Potom pravdepodobnostná vytvárajúca funkcia X je daná vzťahom*

$$\Pi_X(s) = \frac{\theta s}{1-(1-\theta)s}. \tag{1.3}$$

Dôkaz. Pretože $X \sim Geo(\theta)$, platí podľa (1.1) $p_X(k) = \mathbb{P}(X = k) = (1-\theta)^{k-1}\theta$. Počítame:

$$\Pi_X(s) = \mathsf{E}\; s^X = \sum_{i=0}^{\infty} p_X(i)s^i = \sum_{i=1}^{\infty}(1-\theta)^{i-1}\theta s^i = \theta s\sum_{i=1}^{\infty}((1-\theta)s)^{i-1} =$$
$$= \theta s \sum_{i=0}^{\infty}((1-\theta)s)^i = \frac{\theta s}{1-(1-\theta)s}.$$

□

Pre zavedenie Pólyovho-Aeppliho rozdelenia budeme ďalej potrebovať definíciu Poissonovho rozdelenia.

Definícia 2. *Diskrétna náhodná veličina N má Poissonove rozdelenie s parametrom $\lambda > 0$, pokiaľ*

$$\mathbb{P}(N = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots \tag{1.4}$$

Pre Poissonovo rozdelenie platí

$$\mathsf{E}[N] = \mathrm{var}[N] = \lambda. \tag{1.5}$$

Táto vlastnosť sa nazýva ekvidisperzia. Značíme $N \sim \mathcal{P}(\lambda)$.

Posledné z diskrétnych rozdelení, ktoré dokopy vyskladajú Pólyovo-Aeppliho rozdelenie, je zložené Poissonovo rozdelenie.

Definícia 3. *Majme množinu indexov $I \subset \mathbb{N}_0$. Náhodná veličina S má zložené Poissonove rozdelenie $\mathcal{CP}((\lambda_k)_{k \in I})$ s parametrami $(\lambda_k)_{k \in I}$ takými, že $\lambda_k > 0$ a $\sum_{k \in I} \lambda_k = \lambda < \infty$, pokiaľ*

$$S = \sum_{m=1}^{N} X_m, \tag{1.6}$$

kde $N \sim \mathcal{P}(\lambda)$ je nezávislé na X_m, ktoré sú medzi sebou nezávislé rovnako rozdelené podľa

$$\mathbb{P}(X = k) = \frac{\lambda_k}{\lambda}, \quad k \in I. \tag{1.7}$$

V prípade, že $N = 0$, definujeme

$$S = \sum_{m=1}^{0} X_m = 0.$$

Značíme $S \sim \mathcal{CP}((\lambda_k)_{k \in I})$ a v prípade, že máme iba jeden parameter λ, môžeme značenie zjednodušiť na $S \sim \mathcal{CP}(\lambda)$.

Tieto tri diskrétne rozdelenia budú dokopy tvoriť Póylovo-Aeppliho rozdelenie, o ktorom bude pojednané v ďalšej kapitole. Teraz si ešte ukážeme zopár užitočných vlastností ako napríklad strednú hodnota a rozptyl zloženého Poissonovho rozdelenia.

Tvrdenie 2. *Nech $S \sim \mathcal{CP}(\lambda)$, kde $\lambda > 0$, potom*

1. $\mathsf{E}[S] = \lambda \, \mathsf{E}[X]$,
2. $\mathrm{var}[S] = \lambda \, \mathsf{E}[X^2]$.

Dôkaz. Pretože $S \sim \mathcal{CP}(\lambda)$, platí $N \sim \mathcal{P}(\lambda)$ a teda $\mathsf{E}\,[N] = \mathrm{var}[N] = \lambda$. Ďalej počítame:

1. $\mathsf{E}\,[S] = \mathsf{E}\,[\mathsf{E}\,(S|N)] \ = E[\,N\,\mathsf{E}\,(X)] \ = \mathsf{E}\,[N]\,\mathsf{E}\,[X] = \lambda\,\mathsf{E}\,[X]$,
2. $\mathrm{var}[S] \ = \ \mathsf{E}\,[\mathrm{var}(S|N)] + \mathrm{var}[\mathsf{E}\,(S|N)] \ = \ \mathsf{E}\,[N\,\mathrm{var}(X)] + \mathrm{var}[N\,\mathsf{E}\,(X)] \ = \mathsf{E}\,[N]\,\mathrm{var}[X] + (\mathsf{E}\,[X])^2\,\mathrm{var}[N] = \lambda(\mathrm{var}[X] + (\mathsf{E}\,[X])^2) = \lambda\,\mathsf{E}\,[X^2]$.

□

Pravdepodobnostnú funkciu nepodmieneného rozdelenia náhodnej veličiny S vieme vyjadriť nasledovne.

Lemma 3. *Ak $S \sim \mathcal{CP}((\lambda_k)_{k\in I})$, kde $\sum_{k\in I} \lambda_k = \lambda$ a $I \subset \mathbb{N}$ je nejaká indexová množina, potom pre všetky $n \in \mathbb{N}$ je*

$$\mathbb{P}(S = n) = e^{-\lambda} \sum_{m=1}^{n} \frac{1}{m!} \sum_{k_1,\ldots,k_m \in I} \mathbb{I}_{(k_1+\cdots+k_m=n)} \lambda_{k_1} \cdots \lambda_{k_m} \tag{1.8}$$

a $\mathbb{P}(S = 0) = e^{-\lambda}$.

Dôkaz. V prípade, že $n = 0$ je $\mathbb{P}(S = 0) = \mathbb{P}(N = 0) = e^{-\lambda}$. Využijeme vety o úplnej pravdepodobnosti pre rozpísanie $\mathbb{P}(S = n)$:

$$\mathbb{P}(S = n) = \sum_{m=0}^{\infty} \underbrace{\mathbb{P}(X_1 + X_2 + \cdots + X_m = n|N = m)}_{A} \underbrace{\mathbb{P}(N = m)}_{B}, \quad n = 1, 2, \ldots$$

Uvedomíme si, že výraz A hovorí o pravdepodobnosti, že súčet náhodných veličín X_i sa bude rovnať n za podmienky, že počet sčítancov je m, teda $i = 1, 2, \ldots, m$. To isté reprezentuje výraz

$$\sum_{k_1,\ldots,k_m \in I} \mathbb{I}_{(k_1+\cdots+k_m=n)} \frac{\lambda_{k_1} \cdots \lambda_{k_m}}{\lambda^m},$$

pretože náhodné veličiny X_i, $i = 1, 2, \ldots, m$, sa riadia rozdelením (1.7). Člen B hovorí o pravdepodobnosti, že počet sčítancov N sa rovná nejakému pevnému m. Keďže má N Poissonovo rozdelenie, platí (1.4) a tým pádom máme dôkaz hotový.

□

V Özel a Inal (2010) sú pravdepodobnosti najprv odvodené pre $I = \{1, 2\}$, čo znamená, že číslo n rozkladáme na dva sčítance. Potom autori tento postup zovšeobecňujú pre $I = \mathbb{N}$. V tom prípade ich môžeme rozpísať pomocou rovnice (1.8) nasledovne:

$$\begin{aligned}
\mathbb{P}(S = 0) &= e^{-\lambda}, \\
\mathbb{P}(S = 1) &= e^{-\lambda} \frac{\lambda_1}{1!}, \\
\mathbb{P}(S = 2) &= e^{-\lambda} \left[\frac{\lambda_1^2}{2!} + \frac{\lambda_2}{1!} \right],
\end{aligned}$$

$$
\begin{aligned}
\mathbb{P}(S=3) &= e^{-\lambda}\left[\frac{\lambda_1^3}{3!}+\frac{\lambda_1\lambda_2}{1!1!}+\frac{\lambda_3}{1!}\right],\\
\mathbb{P}(S=4) &= e^{-\lambda}\left[\frac{\lambda_1^4}{4!}+\frac{\lambda_1^2\lambda_2}{2!1!}+\frac{\lambda_1\lambda_3}{1!1!}+\frac{\lambda_2^2\lambda_4}{2!1!}\right],\\
\mathbb{P}(S=5) &= e^{-\lambda}\left[\frac{\lambda_1^5}{5!}+\frac{\lambda_1^3\lambda_2}{3!1!}+\frac{\lambda_1^2\lambda_3}{2!1!}+\frac{\lambda_1\lambda_2^2}{2!1!}+\frac{\lambda_1\lambda_4}{1!1!}+\frac{\lambda_2\lambda_3}{1!1!}+\frac{\lambda_5}{1!}\right],\\
&\ \vdots
\end{aligned}
\tag{1.9}
$$

Pravdepodobnosti $\mathbb{P}(S=k)$, $k=1,2,3,\ldots$ na ľavej strane vyjadrujeme výrazmi na pravej strane, ktoré sa na prvý pohľad vzdajú dosť zložité. Keď sa nad tým zamyslíme, ide v podstate iba o to, koľkými možnosťami vieme poskladať prirodzené číslo k používajúc iba čísla $1,2,\ldots,k$, pričom na poradí záleží. Napríklad ak $k=5$, potom máme sedem možností ako rozložiť toto číslo: $\{1,1,1,1,1\}$, $\{1,1,1,2\}$, $\{1,2,2\}$, $\{1,1,3\}$, $\{2,3\}$, $\{1,4\}$, $\{5\}$.

Poslednou z vlastností zloženého Poissonovho rozdelenia, ktorú si uvedieme, je jeho pravdepodobnostná vytvárajúca funkcia.

Lemma 4. *Ak má náhodná veličina S zložené Poissonovo rozdelenie, potom pravdepodobnostná vytvárajúca funkcia S má tvar*

$$\Pi_S(s)=e^{\lambda[\Pi_X(s)-1]}, \tag{1.10}$$

kde $\Pi_X(s)$ je pravdepodobnostná vytvárajúca funkcia sčítancov.

Dôkaz. Z predpokladov $S\sim\mathcal{CP}(\lambda)$, teda platí pre S vzťah (1.6). Počítame pravdepodobnostnú vytvárajúcu funkciu

$$
\begin{aligned}
\Pi_S(s) &= \mathsf{E}\left[\mathsf{E}\left[s^{\sum_{m=1}^n X_m}|N=n\right]\right]=\mathsf{E}\left[\mathsf{E}\left[\Pi_X(s)^n|N=n\right]\right]=\\
&=\sum_{n=0}^{\infty}e^{-\lambda}\frac{\lambda^n}{n!}[\Pi_X(s)]^n=e^{\lambda[\Pi_X(s)-1]}.
\end{aligned}
$$

□

Z teórie k pravdepodobnostným rozdeleniam sme si spomenuli všetko dôležité, čo budeme neskôr potrebovať. V nasledujúcej podkapitole si uvedieme zase potrebnú teóriu z náhodných procesov, na ktorej budeme v tejto práci stavať.

1.2 Náhodné procesy

Čítacie procesy sa používajú na modelovanie počtu udalostí v čase. Keďže našou témou je čítací proces so spojitým časom a diskrétnymi stavmi, uvedieme si niekoľko definícií, s ktorými budeme pracovať, keď si budeme definovať Pólyov-Aeppliho proces. Začneme definíciou čítacieho procesu.

Definícia 4. *Čítací proces $\{N(t), t \geq 0\}$ je stochastický proces, pre ktorý platia nasledujúce vlastnosti:*

1. *$N(t) \geq 0$,*
2. *$N(t)$ je celočíselné,*
3. *ak $s \leq t$, potom $N(s) \leq N(t)$.*

Navyše ak $s < t$, potom rozdiel $N(t)-N(s)$ určuje počet prírastkov v časovom intervale $(s, t]$.

Čítacie procesy vieme rozlišovať podľa typu prírastkov do mnohých kategórií. Ako si ukážeme neskôr, bude náš Pólyov-Aeppliho proces prípad procesu s nezávislými a stacionárnymi prírastkami. Definujeme si teraz, kedy sa jedná o proces s nezávislými prírastkami.

Definícia 5. *Majme čítací proces $\{N(t), t \geq 0\}$. Hovoríme, že $N(t)$ má nezávislé prírastky, ak pre všetky $n \in \mathbb{N}$ a pre ľubovolnú postupnosť $t_0, t_1, t_2, \ldots, t_n$, pre ktorú platí*

$$0 \leq t_0 < t_1 < \cdots < t_n,$$

sú náhodné veličiny $N(t_1)-N(t_0), N(t_2)-N(t_1), \ldots, N(t_n)-N(t_{n-1})$ nezávislé.

V ďalšej definícii si uvedieme, kedy sa jedná o proces so stacionárnymi prírastkami.

Definícia 6. *Majme čítací proces $\{N(t), t \geq 0\}$. Hovoríme, že $N(t)$ má stacionárne prírastky, ak pre všetky $n \in \mathbb{N}$, $0 \leq t_0 < t_1 < \cdots < t_n$ a $h > 0$ nezávisí rozdelenie náhodného vektora $(N(t_1+h)-N(t_0+h), \ldots, N(t_n+h)-N(t_{n-1}+h))$ na h.*

Proces s nezávislými prírastkami má stacionárne prírastky, keď rozdelenie

$$R_{t,h} = N(t+h) - N(t)$$

závisí iba na h a nie na t. Inými slovami povedané, zmena $N(t)$ závisí na veľkosti časového intervalu medzi pozorovaniami a nie na tom, v akom časovom okamihu sme prírastky pozorovali.

Pólyov-Aeppliho proces je čítací proces založený na zloženom Poissonovom procesom, čo je jeden zo špecifických prípadov všeobecného náhodného procesu. Uvedieme si jeho definíciu.

Definícia 7. *Nech $X_1, X_2, \ldots$ sú nezávislé, rovnako rozdelené, celočíselné, nezáporné náhodné veličiny a nech $N_1(t)$ je homogénny Poissonov proces s intenzitou $\lambda > 0$ nezávislý na X_i, $i = 1, 2, \ldots$ Čítací proces $\{N(t), t \geq 0\}$ taký, že*

$$N(t) = \sum_{i=1}^{N_1(t)} X_i, \tag{1.11}$$

sa nazýva zložený Poissonov proces. V prípade, že $N_1(t) = 0$, je aj $N(t) = 0$.

Mohli sme si všimnúť, že v definícii zloženého Poissonovho procesu sa využíva obyčajný Poissonov proces. Jedná sa o ďalší zo špecifických prípadov všeobecného čítacieho procesu. Je to jeden z najčastejšie používaných procesov pre modelovanie počtov nejakých udalostí v čase.

Definícia 8. *Špecifický prípad všeobecného čítacieho procesu* $\{N_1(t), t \geq 0\}$, *pre ktorý platí*

1. *začína v nule skoro isto, t.j.* $N_1(0) = 0$,
2. *je to proces s nezávislými prírastkami,*
3. *pre každé* $t \geq 0$ *a* $h > 0$ *má počet prírastkov* $R_{t,h} = N_1(t+h) - N_1(t)$ *v ľubovolnom intervale dĺžky* h *Poissonovo rozdelenie s parametrom* λh,

sa nazýva homogénny Poissonov proces.

Poissonov proces má mnoho vlastností a charakteristík, z ktorých budeme aj my nejaké využívať. Uvedieme si preto niektoré z nich.

Prvá vlastnosť sa bude týkať distribučnej funkcie doby čakania na udalosť v Poissonovom procese. Doba prvého príchodu Poissonovho procesu je také t, kedy sa nám počítadlo $N_1(t)$ zväčší aspoň o 1. Matematicky zapísané

$$U_1 = \inf\{t \geq 0, N_1(t) \geq 1\}. \tag{1.12}$$

Takéto U_1 reprezentuje čas príchodu prvej udalosti. Dobu čakania U_2 na druhú udalosť, ktorá nastala po prvej, vieme matematicky vyjadriť ako

$$U_2 = \inf\{t \geq 0, N_1(t) \geq 2\} - U_1.$$

Podobne zavedieme aj ostatné doby čakania

$$U_n = \inf\{t \geq 0, N_1(t) \geq n\} - U_{n-1} \quad n = 2, 3, \ldots \tag{1.13}$$

V nasledujúcej vete si uvedieme, že tieto doby čakania majú pri Poissonovom procese exponenciálne rozdelenie.

Veta 5. *Doby medzi príchodmi* $U_1, U_2, \ldots$ *definované ako v* (1.12) *a* (1.13) *jednotlivých udalostí Poissonovho procesu sú nezávislé a rovnako rozdelené náhodné veličiny z exponenciálneho rozdelenia s parametrom* λ, *kde* $\lambda > 0$ *je intenzita Poissonovho procesu.*

Dôkaz. Viď Rolski a kol. (1999), Theoreom 5.2.1.

□

Jeden zo spôsobov, ako zadefinovať Pólyov-Aeppliho proces, je cez proces obnovy. V knihe Rolski a kol. (1999), podkapitola 6.1, alebo v skriptách Prášková a Lachout (2012), str. 133, je definovaný proces obnovy nasledovne.

Definícia 9. *Nech* $\{T_n, n \in \mathbb{N}\}$ *je postupnosť nezávislých náhodných veličín, ktoré nadobúdajú iba nezáporné hodnoty. Ďalej predpokladajme, že* $T_2, T_3, \ldots$ *majú rovnaké rozdelenie s distribučnou funkciou* F*, pre ktorú platí* $F(0) < 1$ *a so strednou hodnotou* μ*. Položme* $S_0 = 0$ *a*

$$S_n = \sum_{k=1}^{n} T_k, \quad n = 1, 2, 3, \ldots$$

Potom proces náhodných veličín $\{N(t), t \geq 0\}$ *takých, že*

$$N(t) = \sup\{n \in \mathbb{N}_0 : S_n \leq t\},$$

sa nazýva proces obnovy. Pokiaľ $\mathbb{P}(T_1 > 0) > 0$*, hovoríme, že* $\{N(t), t \geq 0\}$ *je proces obnovy s oneskorením.*

Vzhľadom k tomu, že predpokladáme $T_2 \geq 0$ s pravdepodobnosťou jedna a $F(0) = \mathbb{P}(T_2 = 0) < 1$, je $\mu > 0$. Náhodná veličina S_n značí čas, kedy dôjde k výskytu n-tej udalosti (obnovy), $N(t)$ je počet obnov v intervale $[0, t]$. Ak je $T_1 = S_1 = 0$, považujeme počiatok za čas obnovy. Niekedy v tomto prípade hovoríme o čistom procese obnovy. Náhodné veličiny $T_2, T_3, \ldots$ sú doby medzi obnovami. Typickým príkladom procesu obnovy je proces výmeny súčiastok zariadenia v nepretržitej prevádzke.

Existuje viacero špecifických prípadov procesu obnovy. Budeme z nich potrebovať prípad procesu obnovy s oneskorením a prípad procesu so stacionárnymi prírastkami. O proces obnovy s oneskorením sa jedná vtedy, ak $F_{T_1}(0) = \mathbb{P}(T_1 = 0) < 1$. V nasledujúcom tvrdení si ukážeme, kedy sa jedná o proces obnovy so stacionárnymi prírastkami.

Tvrdenie 6. *Nech* $\{N(t), t \geq 0\}$ *je proces obnovy s oneskorením a distribučná funkcia* T_1 *spĺňa vzťah* $F_{T_1}(t) = \frac{1}{\mu}\int_0^t (1 - F(u))\,\mathrm{d}u$*, potom hovoríme, že* $N(t)$ *má stacionárne prírastky.*

Dôkaz. Viď Rolski a kol. (1999), Theorem 6.1.8. □

V tejto práci sa používa výraz martingála. Uvedieme si preto význam tohto pojmu v matematike.

Definícia 10. *Nech* $N(t)$ *je náhodný proces so spojitým časom prispôsobený sprava spojitej filtrácii* H_t*,* $t \in [0, \infty)$*. Hovoríme, že* $N(t)$ *je martingála, pokiaľ spĺňa nasledujúce vlastnosti:*

1. $\mathsf{E}\left[|N(t)|\right] < \infty$ *pre všetky* t *a*
2. $\mathsf{E}\left[N(t)|H_s\right] = N(s)$ *pre všetky* $s < t$*.*

Proces nazývame submartingálou (resp. supermaringálou), ak je rovnosť nahradená $\geq$ *(resp.* $\leq$*).*

V predošlej definícii je spomenutý výraz sprava spojitá filtrácia. Filtrácia je postupnosť podmnožín H_t (ktoré sú zároveň σ-algebry), $t \in [0,\infty)$, σ-algebry $\mathcal{A}$, takých že pre každé $s \leq t$ je $H_s \subset H_t$. Hovoríme, že filtrácia je sprava spojitá, ak

$$\bigcap_{\epsilon>0} H_{t+\epsilon} = H_t.$$

V tejto práci využívame okrem iného aj Markovove reťazce. Rovnako ako pri definícii procese obnovy budeme čerpať z Prášková a Lachout (2012), str. 77.

Definícia 11. *Systém celočíselných náhodných veličín $\{N_t, t \geq 0\}$ definovaných na pravdepodobnostnom priestore $(\Omega, \mathcal{A}, \mathbb{P})$ sa nazýva Markovov reťazec so spojitým časom a spočetnou množinou stavov S, pokiaľ*

$$\mathbb{P}(N_t = j | N_s = i, N_{t_n} = i_n, \dots, N_{t_1} = i_1) = \mathbb{P}(N_t = j | N_s = i)$$

pre všetky $i, j, i_1, \dots, i_n \in S$ a pre všetky $0 \leq t_1 < t_2 < \dots < t_n < s < t$, pre ktoré $\mathbb{P}(N_s = i, N_{t_n} = i_n, \dots, N_{t_1} = i_1) > 0$. Pravdepodobnosť $\mathbb{P}(N_t = j | N_s = i)$ označíme ako $p_{ij}(s,t)$ a budeme ju nazývať pravdepodobnosť prechodu zo stavu i v čase s do stavu j v čase t.

My sa budeme zaoberať iba s homogénnymi Markovovými reťazcami so spojitým časom. Sú to také Markovove reťazce, ktoré spĺňajú nasledujúcu vlastnosť:

$$p_{ij}(s, s+t) = p_{ij}(t), \quad s \geq 0, t \geq 0.$$

Pokiaľ pre každé $i, j \in S$ budeme uvažovať celý systém pravdepodobností $\{p_{ij}(t), t > 0\}$, pre ktorý platí $\sum_{j \in S} p_{ij}(t) = 1$, môžeme tento systém zapísať maticovým spôsobom $\{\mathbf{P}(t), t > 0\}$. Zvyčajne sa definuje $p_{ij}(0) = \delta_{ij}$, kde δ_{ij} je Kroneckerov symbol, t.j. $\mathbf{P}(0) = \mathbf{I}$. Kroneckerove delta je také δ_{ij}, pre ktoré platí

$$\delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Príslušné intenzity prechodu homogénneho reťazca sú definované nasledovne.

Definícia 12. *Majme množinu stavov S a nezáporné čísla*

$$q_i := \lim_{h \to 0_+} \frac{1 - p_{ii}(h)}{h},$$
$$q_{ij} := \lim_{h \to 0_+} \frac{p_{ij}(h)}{h}.$$

Tieto budeme nazývať intenzity prechodu zo stavu i do stavu j, $i, j \in S$. Číslo q_i sa nazýva celková intenzita. Matica $\mathbf{Q} = \{q_{ij}, i, j \in S\}$, kde $q_{ii} = -q_i$, sa nazýva matica intenzít prechodu.

V tejto práci použijeme Kolmogorovove diferenciálne rovnice. Keďže sme si pripravili všetky nástroje na ich zavedenie, môžeme si ich sformulovať v ďalšom tvrdení.

Tvrdenie 7. *Predpokladajme, že $q_i < \infty$ pre všetky $i \in S$ a platí*

$$q_i = \sum_{j \neq i} q_{ij} \quad \textit{pre všetky } i \in S.$$

Potom pravdepodobnosti prechodu $p_{ij}(t)$ sú diferencovateľné pre všetky $i, j \in S$ a $t > 0$ a sústavu prospektívnych Kolmogorovových diferenciálnych rovníc vieme zapísať maticovo ako

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}. \tag{1.14}$$

Dôkaz. Viď Prášková a Lachout (2012), veta 3.10.

□

Posledný pojem, ktorý si spomenieme, je index disperzie.

Definícia 13. *Majme náhodný proces $N(t)$ so strednou hodnotou μ a rozptylom σ^2. Index disperzie je definovaný ako*

$$D(N(t)) = \frac{\sigma^2}{\mu}. \tag{1.15}$$

Tento pomer nám hovorí o tom, či je pozorovaný počet výskytov zhluknutý alebo rozptýlený oproti nejakému porovnávanému modelu. Ak má náhodná veličina Poissonovo rozdelenie, je tento index rovný jednej. Pri geometrickom rozdelení začínajúce v nule je zase tento index väčší ako 1. Ak máme dostatočne veľa dát o počte výskytov z nejakého časového intervalu a index disperzie je blízko 1, hovorí nám to, že počet výskytov bude mať pravdepodobne Poissonove rozdelenie a bolo by vhodné na modelovanie použiť Poissonov proces. Ak je tento index väčší ako 1, znamená to, že sa častejšie vyskytujú intervaly s vyšším a nižším počtom výskytov oproti prípadu, kedy by boli tieto počty výskytov nejakých udalostí z Poissonovho rozdelenia. Vtedy hovoríme, že dáta majú nadmerný rozptyl. Ako si ukážeme neskôr v práci, bude Pólyov-Aeppliho proces presne tento prípad.

2. Geometrické Poissonovo rozdelenie

V tejto kapitole si zadefinujeme rozdelenie, ktoré bude kľúčové pre Pólyov-Aeppliho proces. Bude sa jednať o zložené Poissonovo rozdelenie, ktorého sčítance sa riadia geometrickým rozdelením. Preto sa aj nazýva geometrické Poissonovo rozdelenie (alebo aj Pólyovo-Aeppliho rozdelenie). Ukážeme si aj zopár vlastností a aj to, ako vyzerá graf pravdepodobnostnej funkcie rozdelenia pri rozličných parametroch.

2.1 Zavedenie

V predošlej kapitole sme si pripomenuli niekoľko základných a užitočných definícií a tvrdení. Máme teda všetko potrebné pre zadefinovanie nášho rozdelenia.

Definícia 14 (Geometrické Poissonove rozdelenie)**.** *Náhodná veličina S má geometrické Poissonove rozdelenie $\mathcal{GP}(\lambda, \theta)$, kde parameter $\theta \in (0, 1]$ náleží geometrickej časti rozdelenia, teda sčítancom a parameter $\lambda > 0$ zase Poissonovej časti rozdelenia, teda počtu sčítancov, práve keď $S \sim \mathcal{CP}((\lambda_k)_{k\in\mathbb{N}})$ a*

$$\lambda_k = \lambda(1-\theta)^{k-1}\theta, \quad k = 1, 2, \dots \tag{2.1}$$

V prípade kedy $\theta = 1$, sa jedná o degenerovaný prípad a toto rozdelenie sa stáva Poissonovým.

V tvrdení 2 sme si odvodili vzorce pre strednú hodnotu a rozptyl zloženého Poissonovho rozdelenia. Ako jeho dôsledok vyplynú stredná hodnota a rozptyl geometrického Poissonovho rozdelenia.

Dôsledok 8. *Pokiaľ $S \sim \mathcal{GP}(\lambda, \theta)$, potom stredná hodnotu a rozptyl geometrického Poissonovho rozdelenia majú tvar*

$$\mathsf{E}[S] = \frac{\lambda}{\theta} \quad a \quad \mathrm{var}[S] = \frac{\lambda(2-\theta)}{\theta^2}. \tag{2.2}$$

.

Dôkaz. Triviálne. Spočítame druhý necentrálny moment geometrického rozdelenia s parametrom θ a spolu s prvým momentom v (1.2) dosadíme do vzorcov v tvrdení 2.

□

V nasledujúcej vete si uvedieme rekurzívny vzťah pre rozdelenie pravdepodobností geometrického Poissonovho rozdelenia.

Veta 9. *Majme náhodnú veličinu* $S \sim \mathcal{GP}(\lambda, \theta)$. *Potom platí*

$$\mathbb{P}(S = 0) = e^{-\lambda} \quad a \quad \mathbb{P}(S = 1) = e^{-\lambda}(1 - \theta)z \tag{2.3}$$

a pre $n \geq 2$, $n \in \mathbb{N}$,

$$\mathbb{P}(S = n) = \frac{(2n - 2 + z)}{n}(1-\theta)\mathbb{P}(S = n-1) + \frac{(2 - n)}{n}(1-\theta)^2\mathbb{P}(S = n-2), \tag{2.4}$$

kde $z = \frac{\lambda\theta}{1-\theta}$.

Dôkaz. Dôkaz je uvedený v Nuel (2008), Proposition 7.

□

Rekurzívna formula v rovnici (2.4) vyžaduje výpočet predošlých pravdepodobností $\mathbb{P}(S = k)$, $k = 0, 1, 2, \ldots, n - 1$ na to, aby sme určili $\mathbb{P}(S = n)$. Takýto algoritmus na počítanie pravdepodobností je časovo aj pamäťovo náročný pre veľké hodnoty n. Tento problém vieme ale odstrániť tak, že vzorec prevedieme do explicitného tvaru, o čom budeme pojednávať v nasledujúcej podkapitole.

2.2 Rozdelenie pravdepodobností

Pre dôvody spomenuté na konci minulej podkapitoly zavedieme iné spôsoby ako počítať pravdepodobnosti geometrického Poissonovho rozdelenia. Každý je založený na niečom inom. Jeden zo spôsobov ako spočítať pravdepodobnosti geometrického Poissonovho rozdelenia explicitným vzorcom, ktorý si uvedieme neskôr v tejto podkapitole, bude využívať pravdepodobnostnú vytvárajúcu funkciu tohto rozdelenia. Teraz si ukážeme prvý spôsob, ako spočítať pravdepodobnosti geometrického Possonovho rozdelenia. V nasledujúcom tvrdení si zhrnieme všetky naše doterajšie poznatky a zavedieme explicitný vzorec pre výpočet pravdepodobnosti $\mathbb{P}(S = k)$ geometrického Poissonovho rozdelenia.

Tvrdenie 10. *Ak* $S \sim \mathcal{GP}(\lambda, \theta)$, *potom pre všetky* $n \in \mathbb{N}$ *dostávame*

$$\mathbb{P}(S = n) = e^{-\lambda} \sum_{k=1}^{n} \frac{\lambda^k}{k!}(1 - \theta)^{n-k}\theta^k \binom{n-1}{k-1} \tag{2.5}$$

a $\mathbb{P}(S = 0) = e^{-\lambda}$.

Dôkaz. K dôkazu využijeme rovnice (1.8) a (2.1). Pre všetky $n \in \mathbb{N}$ dostávame

$$\mathbb{P}(S = n) = e^{-\lambda} \sum_{m=1}^{n} \frac{\lambda^m}{m!}(1 - \theta)^{n-m}\theta^m \underbrace{\sum_{k_1,\ldots,k_m \in \mathbb{N}} \mathbb{I}_{(k_1+\cdots+k_m=n)}}_{A(n,m)},$$

čiže stačí nám ukázať, že $A(n,m) = \binom{n-1}{m-1}$. Uvedomíme si, že člen $A(n,m)$ reprezentuje počet možností (na poradí záleží), ako vieme sčítať m prirodzených čísel tak, aby dávali súčet n. Majme množinu $M = \{1, \ldots, n\}$. Množina M obsahuje práve $n-1$ čiarok (alebo prepážok) a výber $m-1$ týchto čiarok zrejme dáva rozklad čísla n. Všetkých týchto možností je dohromady $\binom{n-1}{m-1}$. Keďže sa jedná o počet možností, ako sčítať m prirodzených čísel tak, aby dávali súčet n a člen $A(n,m)$ hovorí to isté, ukázali sme, že $A(n,m) = \binom{n-1}{m-1}$ a tvrdenie je dokázané.

$\square$

V Özel a Inal (2010) autori ukazujú ďalší spôsob, ako sa dopracovať k týmto pravdepodobnostiam, ktorý je založený na derivovaní pravdepodobnostnej vytvárajúcej funkcie. Potrebujeme ale vedieť, ako vyzerá pravdepodobnostná vytvárajúca funkcia geometrického Poissonovho rozdelenia. Ukážeme si to v nasledujúcej lemme.

Lemma 11. *Nech $S \sim \mathcal{GP}(\lambda, \theta)$, potom pravdepodobnostná vytvárajúca funkcia S má tvar*

$$\Pi_S(s) = e^{-\lambda[(1-s)/(1-(1-\theta)s)]}. \tag{2.6}$$

Dôkaz. Dosadením (1.3) do (1.10) dostávame vzťah

$$\Pi_S(s) = e^{-\lambda} e^{\lambda[\theta s/(1-(1-\theta)s)]}.$$

Postupnými úpravami sa dostaneme až na tvar

$$\Pi_S(s) = e^{-\lambda[(1-s)/(1-(1-\theta)s)]},$$

čo je rovné (2.6). Tým pádom je lemma dokázané.

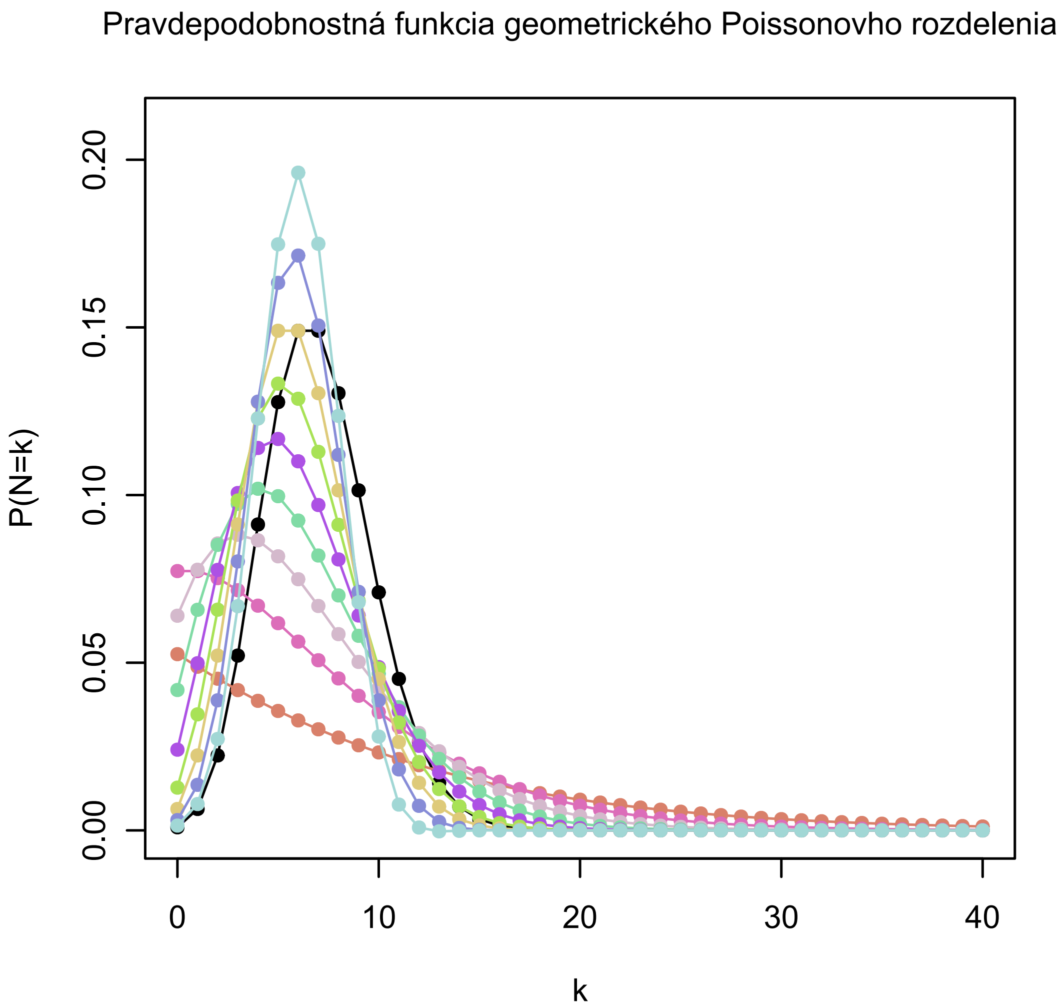
$\square$

Teraz keď máme pravdepodobnostnú vytvárajúcu funkciu geometrického Poissonovho rozdelenia, spočítame k-tu deriváciu v bode $s = 0$:

$$\mathbb{P}(S=0) = \Pi_S(0), \quad \mathbb{P}(S=k) = \frac{\partial^k/\partial s^k(\Pi_S(s))|_{s=0}}{k!}, \quad k = 1, 2, \ldots \tag{2.7}$$

Z rovnice (2.7) nám vyjdú rovnaké pravdepodobnosti ako v (2.5). Máme teda k dispozícií dva spôsoby pre výpočet pravdepodobností $\mathbb{P}(S=k)$, ktorý nie je rekurzívny. Tým pádom môžeme rovno určiť tieto pravdepodobnosti $\mathbb{P}(S=k)$ napríklad podľa (2.5) bez toho, aby sme predtým museli počítať všetky predošlé $\mathbb{P}(S=l)$, $l = 0, 1, 2, \ldots, k-1$, čo výrazne ušetrí čas výpočtu.

Na obrázku 2.1 si ukážeme graf pravdepodobnostnej funkcie geometrického Poissonovho rozdelenia s rôznou voľbou parametrov θ a λ tak, že $\mathsf{E}[S] = \lambda/\theta = 7$. Všimnime si, že so zmenšujúcim sa parametrom θ dostávame ťažšie konce rozdelenia.



Obr. 2.1: Rozdelenie pravdepodobností náhodnej veličiny $S \sim \mathcal{GP}(\lambda, \theta)$ s $\lambda/\theta = 7$ a θ nabývajúcim rôznych hodnôt. Od horného až po spodný graf distribučnej funkcie v bode $k \approx 7$, θ je rovné $0.9, 0.8, \ldots, 0.2, 0.1$. Čiernou farbou je znázornený Poissonov prípad, t.j. $\theta = 1.0$.

3. Pólyov-Aeppliho proces

3.1 Definície Pólyovho-Aeppliho procesu

V tejto kapitole zavedieme Pólyov-Aeppliho proces rôznymi spôsobmi. Budeme vychádzať z článku Chukova a Minkova (2013). Pri každej definícií sa budeme na tento proces pozerať z iného uhla pohľadu a ukážeme si vzťahy medzi nimi. Nakoniec si spomenieme zopár vlastností tohto procesu. Pre Pólyov-Aeppliho proces budeme používať skratku PAP.

3.1.1 PAP ako proces s nezávislými prírastkami

V minulej kapitole, tvrdenie 10, sme si odvodili vzorec pre výpočet pravdepodobností Pólyovho-Aeppliho rozdelenia. Budeme chcieť Pólyov-Aeppliho proces zaviesť tak, aby prírastky mali príslušné Pólyovo-Aeppliho rozdelenie, teda $N(t) \sim \mathcal{GP}(\lambda t, \theta)$. Potom si budeme vedieť odvodiť pravdepodobnosti pre celkový počet prírastkov v čase $(N(t))$ podľa tvrdenia 10 nasledovne:

$$\mathbb{P}(N(t) = n) = \begin{cases} e^{-\lambda t}, & n = 0 \\ e^{-\lambda t} \sum_{k=1}^{n} \frac{(\lambda t)^k}{k!} (1-\theta)^{n-k} \theta^k \binom{n-1}{k-1}, & n = 1, 2, \ldots \end{cases} \tag{3.1}$$

Ukážeme si prvý spôsob, ako sa dá pozerať na Pólyov-Aeppliho proces. Jedná sa o zovšeobecnenie definície 8 Poissonovho procesu.

Definícia 15. *Čítací proces $\{N(t), t \geq 0\}$ sa nazýva Pólyov-Aeppliho proces s parametrami λ a θ ak platí*

1. *začína v nule skoro isto, t.j. $N(0) = 0$,*
2. *je to proces s nezávislými prírastkami,*
3. *pre každé $t \geq 0$ má počet prírastkov procesu v ľubovolnom intervale dĺžky t geometrické Poissonovo rozdelenie s parametrami λt a θ.*

Značíme $\{N(t), t \geq 0\} \sim \mathcal{PAP}(\lambda, \theta)$. Ak $\theta = 1$, proces sa zjednoduší na prípad $\mathcal{PAP}(\lambda, 1)$, čo je homogénny Poissonov proces s intenzitou λ.

3.1.2 PAP ako zložený Poissonov proces

V tejto sekcii si ukážeme, že Pólyov-Aeppliho proces sa dá zaviesť ako zložený Poissonov proces.

Definícia 16. *Uvažujme zložený Poissonov proces $\{N(t), t \geq 0\}$ ako v definícii 7. Vezmeme $X_i \sim Geo(\theta)$ z definície 1 a dosadíme do (1.11). Potom*

$$N(t) = X_1 + X_2 + \cdots + X_{N_1(t)}, \quad t \geq 0,$$

je Pólyov-Aeppliho proces.

3.1.3 PAP ako oneskorený proces obnovy

Ďalší spôsob ako definovať Pólyov-Aeppliho proces je cez oneskorený proces obnovy tak, aký ho máme uvedený v definícii 9. Budeme používať exponenciálne rozdelenie s parametrom λ, ktoré je s pravdepodobnosťou $1-\theta$ sústredené v nule a ktoré značíme ako $Exp(\lambda, \theta)$. Preto predtým, ako si uvedieme ďalšiu definíciu Pólyovho-Aeppliho procesu, si ukážeme, aký predpis má distribučná funkcia tohto rozdelenia.

Definícia 17. *Náhodná veličina $T \sim Exp(\lambda, \theta)$ má tvar distribučnej funkcie*

$$F_T(t) = 1 - \theta e^{-\lambda t}, \quad t \geq 0.$$

Predtým ešte, ako si uvedieme ďalšiu definíciu Poissonovho procesu z formálnych dôvodov vysvetlíme značenie $T \sim Exp(\lambda)$. Týmto značením je myslené, že náhodná veličina T má exponenciálne rozdelenie s parametrom λ. Sme teda pripravený uviesť si ďalšiu definíciu Pólyovho-Aeppliho procesu.

Definícia 18. *Oneskorený proces obnovy $\{T_1, T_2, \dots\}$ sa nazýva Pólyov-Aeppliho proces s parametrami λ a θ, ak $T_1 \sim Exp(\lambda)$ a $T_i \sim Exp(\lambda,\, \theta)$, $i = 2, 3, \dots$*

Z definícií 17 a 18 vyplýva, že všetky T_i, $i = 2, 3, \dots$ majú distribučnú funkciu v tvare

$$F_{T_i}(t) = 1 - \theta e^{-\lambda t}, \quad t \geq 0. \tag{3.2}$$

Existuje vzťah medzi distribučnou funkciou prvého času príchodu T_1 a ostatnými $T_2, T_3, \dots$, ktorý si uvedieme v ďalšej lemme.

Lemma 12. *Pre $T_1 \sim Exp(\lambda)$ a $T_i \sim Exp(\lambda,\, \theta)$ platí, že*

$$F_{T_1}(t) = \frac{1}{\mathsf{E}\, T_i} \int_0^t (1 - F_{T_i}(u))\, \mathrm{d}u, \quad i = 2, 3, \dots \tag{3.3}$$

Dôkaz. Pretože T_1 je z exponenciálneho rozdelenia s parametrom λ, dostávame $F_{T_1}(t) = 1 - e^{-\lambda t}$. Spočítame najprv strednú hodnotu

$$\mathsf{E}\, T_i = \int_0^\infty (1 - F_{T_i}(t))\, \mathrm{d}t = \theta \int_0^\infty e^{-\lambda t}\, \mathrm{d}t = \frac{\theta}{\lambda},$$

$i = 2, 3, \dots$ Ďalej počítame

$$\frac{1}{\mathsf{E}\, T_i} \int_0^t (1 - F_{T_i}(u))\, \mathrm{d}u = \frac{\lambda}{\theta} \int_0^t \theta e^{-\lambda u}\, \mathrm{d}u = \int_0^t \lambda e^{-\lambda u}\, \mathrm{d}u = 1 - e^{-\lambda t},$$

čo je presne $F_{T_1}(t)$ a tým pádom je lemma dokázané.

□

3.1.4 PAP ako Markovov reťazec

Keď si uvedomíme, že pri konštrukcii $\mathcal{PAP}(\lambda, \theta)$ sa dá postupovať cez pravdepodobnosti prechodu, môžeme Pólyov-Aeppliho proces definovať ako Markovov reťazec.

Definícia 19. *Majme Markovov reťazec $\{N(t), t \geq 0\}$ s množinou stavov $S = \mathbb{N}_0$ a počiatočným rozdelením $N(0) = 0$. Nech pre pravdepodobnosti prechodu plaita vzťahy*

$$\mathbb{P}(N(t+h) = n | N(t) = m) = \begin{cases} 1 - \lambda h + o(h), & n = m \\ \theta(1-\theta)^{i-1}\lambda h + o(h), & n = m + i, i \in \mathbb{N}, \\ 0, & n < m, \end{cases} \tag{3.4}$$

pre každé $m \in \mathbb{N}_0$, $t, h \geq 0$, kde $o(h) \xrightarrow{h \to 0} 0$. Potom sa jedná o Pólyov-Aeppliho proces s parametrami λ a θ. Príslušná matica intenzít prechodu má tvar

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda\theta & \lambda(1-\theta)\theta & \lambda(1-\theta)^2\theta & \dots \\ 0 & -\lambda & \lambda\theta & \lambda(1-\theta)\theta & \dots \\ 0 & 0 & -\lambda & \lambda\theta & \dots \\ 0 & 0 & 0 & -\lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

3.2 Ekvivalencia definícií PAP

V predošlej podkapitole sme si ukázali niekoľko spôsobov, ako sa dá nahliadať na Pólyov-Aeppliho proces. Aby tieto štyri spomenuté spôsoby definície PAP definovali ten istý proces, treba ukázať, že sú si navzájom ekvivalentné.

Tvrdenie 13. *Proces z definície 16 je presne taký PAP, ako je definovaný v definícii 15.*

Dôkaz.

- Najprv si dokážeme implikáciu 16 $\implies$ 15. Majme proces ako v definícii 16 a overíme, či sú splnené všetky tri body definície 15.

    1. Pretože PAP je zložený Poissonov proces, máme $N_1(t)$ ako v definícii 8, pre ktorý je $N_1(0) = 0$ skoro iste. Tým pádom je aj $N(0) = 0$ skoro iste.
    2. Označme
    $$R_{1,2} = N(t_2) - N(t_1) = \sum_{i=N_1(t_1)+1}^{N_1(t_2)} X_i,$$
    ktoré reprezentuje počet prírastkov v časovom intervale $(t_1, t_2]$ a
    $$R_{3,4} = N(t_4) - N(t_3) = \sum_{i=N_1(t_3)+1}^{N_1(t_4)} X_i,$$

ktoré zase reprezentuje počet prírastkov v časovom intervale $(t_3, t_4]$. Pretože N_1 má nezávislé prírastky a náhodné veličiny X_i sú nezávislé na $\{N_1(t), t \geq 0\}$ dostávame, že $R_{1,2}$ a $R_{3,4}$ sú nezávislé. Podobne sa ukáže nezávislosť prírastkov v n disjunktných intervaloch.

3. Veličiny $X_1, X_2, \ldots, X_{N_1(t)}$ v definícii 16 majú geometrické rozdelenie a $N_1(t)$ je Poissonov proces ako v definícii 8, teda má Poissonovo rozdelenie s parametrom λt. Potom pre sumu

$$\sum_{i=1}^{N_1(t)} X_i$$

platí, že má geometrické Poissonovo rozdelenie s parametrami λt a θ a tým pádom pravdepodobnosti $\mathbb{P}(N(t) = n)$ majú predpis ako v (3.1).

- Obrátená implikácia 15 $\implies$ 16 sa dokáže nasledovne. Majme proces z definície 15, potom podľa Daley a kol. (2003), Theorem 2.2.II, sa jedná o zložený Poissonov proces, pričom X_i majú geometrické rozdelenie.

□

Tvrdenie 14. *Definície 16 a 18 sú si navzájom ekvivalentné.*

Dôkaz.

- Začneme implikáciou 16 $\implies$ 18.
Majme proces $N(t)$ ako v definícii 16. Označme doby čakania na udalosť U_i Poissonovho procesu $N_1(t)$, pre ktoré platia vzťahy (1.12) a (1.13). Ďalej označme $T_1 = U_1$ ako dobu čakania na prvú udalosť procesu $N(t)$. Keďže udalosti $1, 2, \ldots, X_1$ prišli v čase spoločne, sú $T_2 = \cdots = T_{X_1} = 0$. Nasledujúce nenulové T_i bude $T_{X_1+1} = U_2$. Takto pokračujem až po U_n. Z vety 5 vieme, že doby medzi príchodmi udalostí $U_1, U_2, \ldots, U_n$ v Poissonovom procese majú exponenciálne rozdelenie s parametrom λ. To platí pre $T_1 = U_1$. Pre T_2 platí $\mathbb{P}(T_2 = 0) = \mathbb{P}(X_1 > 1) = 1 - \theta$ a $T_2 = U_2$, pokiaľ $X_1 = 1$. To znamená, že $T_2 \sim Exp(\lambda, \theta)$. Nezávislosť T_1 a T_2 môžeme ukázať nasledovne:

$$\begin{aligned}
\mathbb{P}(T_1 \leq t_1, T_2 \leq t_2) &= \mathbb{P}(T_1 \leq t_1, T_2 \leq t_2, X_1 = 1) + \\
&\quad + \mathbb{P}(T_1 \leq t_1, T_2 \leq t_2, X_1 > 1) = \\
&= \mathbb{P}(U_1 \leq t_1, U_2 \leq t_2, X_1 = 1) + \mathbb{P}(U_1 \leq t_1, X_1 > 1) = \\
&= \mathbb{P}(U_1 \leq t_1)\mathbb{P}(U_2 \leq t_2)\mathbb{P}(X_1 = 1) + \\
&\quad + \mathbb{P}(U_1 \leq t_1)\mathbb{P}(X_1 > 1) = \\
&= (1 - e^{-\lambda t_1})(1 - e^{-\lambda t_2})\theta + (1 - e^{-\lambda t_1})(1 - \theta) = \\
&= (1 - e^{-\lambda t_1})(1 - \theta e^{-\lambda t_2}) = \mathbb{P}(T_1 \leq t_1)\mathbb{P}(T_2 \leq t_2),
\end{aligned}$$

kde sme využili vetu 5 a nezávislosť U_1, U_2 a X_1. Analogicky by sme ukázali, že $T_3, \cdots \sim Exp(\lambda, \theta)$ a že $T_1, T_2, \ldots$ sú nezávislé.

- Teraz si dokážeme 18 $\implies$ 16.
  Pre proces obnovy z definície 18 označme $S_n = T_1 + \cdots + T_n$ a $N(t) = \sup\{n : S_n \le t\}$. Budeme chcieť ukázať, že $N(t)$ je zložený Poissonov proces. Položme $U_1 = S_1$ a $I_1 = 1$. Ďalej definujme indukciou $I_n = \inf\{k : S_k > S_{I_{n-1}}, U_n = S_{I_n} - S_{I_{n-1}}$ a $X_{n-1} = I_n - I_{n-1}$ pre $n = 2, 3, \ldots$ Vieme, že $U_1 \sim Exp(\lambda)$ a ďalej z konštrukcie plyne, že $U_2, U_3, \ldots$ majú rovnaké rozdelenie ako T_2 za podmienky $T_2 > 0$. Toto je rozdelenie je $Exp(\lambda)$. Z nezávislosti T_i plyne nezávislosť U_i. Preto $N_1(t) = \sup\{n : U_1 + \cdots + U_n \le t\}$ je homogénny Poissonov proces s intenzitou λ. Náhodné veličiny X_i sú nezávislé a majú rozdelenie $\mathbb{P}(X_i = k) = \mathbb{P}(T_{j+1} = 0, T_{j+2} = 0, \ldots, T_{j+i-1} = 0, T_{j+i} > 0) = (1-\theta)^{i-1}\theta$ pre vhodné j. Z toho dostaneme, že $N(t) = \sum_{i=1}^{N_1(t)} X_i$ má rozdelenie $\mathcal{GP}(\lambda, \theta)$. Iné odvodenie toho, že $N(t) \sim \mathcal{GP}(\lambda, \theta)$ založené na Laplaceovo-Stieltjesovej transformácii môžeme nájsť v Minkova (2004).

□

Tvrdenie 15. *Definície 15 a 19 sú si navzájom ekvivalentné.*

Dôkaz.

- Najprv si ukážeme implikáciu 19 $\implies$ 15.
  Rovnica (3.4) sa dá interpretovať tak, že geometricky rozdelené prírastky prichádzajú náhodne s intenzitou λ. Označme $\mathbb{P}_m(t) = \mathbb{P}(N(t) = m)$, $m = 0, 1, 2, \ldots$ Potom z (3.4) získame Kolmogorovove diferenciálne rovnice (tvrdenie 7)

$$\begin{aligned} \mathbb{P}_0'(t) &= -\lambda \mathbb{P}_0(t) \\ \mathbb{P}_m'(t) &= -\lambda \mathbb{P}_m(t) + \theta\lambda \sum_{j=1}^{m} (1-\theta)^{j-1} \mathbb{P}_{m-j}(t), \quad m = 1, 2, \ldots, \end{aligned} \tag{3.5}$$

  s počiatočnými podmienkami

$$\mathbb{P}_0(0) = 1 \quad a \quad \mathbb{P}_m(0) = 0, \quad m = 1, 2, \ldots \tag{3.6}$$

  Ukážeme, že riešenie rovnice (3.5) s počiatočnými podmienkami (3.6) je dané (3.1).

  Majme pravdepodobnostnú vytvárajúcu funkciu

$$h(s,t) = \sum_{m=0}^{\infty} s^m \mathbb{P}_m(t)$$

  procesu $N(t)$. Vynásobením m-tej rovnice v (3.5) faktorom s^m a sčítaním cez všetky $m = 0, 1, 2, \ldots$ dostaneme nasledujúce diferenciálne rovnice

$$\frac{\partial h(s,t)}{\partial t} = -\lambda\,[1 - \Pi_X(s)\,]h(s,t). \tag{3.7}$$

  Riešením tejto rovnice berúc v ohľad počiatočnú podmienku $h(1,t) = 1$ je

$$h(s,t) = e^{-\lambda t(1-\Pi_X(s))},$$

  čo je pravdepodobnostná vytvárajúca funkcia Pólyovho-Aeppliho procesu s parametrami λ a θ, presne sme si ju uviedli v (1.10). Nezávislosť prírastkov plynie z markovskej vlastnosti. Týmto je implikácia dokázaná.

- Zostáva implikácia 15 $\implies$ 19.
  Z nezávislosti prírastkov máme markovskú vlastnosť a $\mathbb{P}(N(t+h) = n|N(t) = m) = \mathbb{P}(N(t+h)-N(t) = n-m)$. Zoberieme rovnice (3.1) a spravíme Taylorov rozvoj v bode 0. Odtiaľ nám vyjdú rovnaké pravdepodobnosti prechodu ako v (3.4) a tým pádom je tvrdenie dokázané.

□

Ukázali sme teda, že všetky definície PAP procesu uvedené v predošlej podkapitole sú si ekvivalentné. Každá z nich nám podáva iný pohľad na PAP proces, čo sa zíde v rozličných situáciach a pri rôznych aplikáciach procesu.

3.3 Vlastnosti PAP

Pri definovaní Pólyovho-Aeppliho procesu sme si už uviedli zopár základných triviálnych vlastností ako napríklad, že proces začína od nuly. V tejto podkapitole sa pozrieme na zopár triviálnych aj netriviálnych vlastností týkajúcich sa nášho procesu. Začneme jeho strednou hodnotou a rozptylom

Tvrdenie 16. *Stredná hodnota a rozptyl $N(t) \sim \mathcal{PAP}(\lambda, \theta)$ sú rovné*

$$\mathsf{E}\left[N(t)\right] = \frac{\lambda t}{\theta} \quad a \quad \mathrm{var}[N(t)] = \frac{\lambda t(2-\theta)}{\theta^2}.$$

Dôkaz. Z definície 15 vieme, že $N(t)$ má geometrické Poissonovo rozdelenie s parametrami λt a θ. Ďalej stačí aplikovať dôsledok 8.

□

Príslušný index disperzie PAP je podľa (1.15) rovný

$$D(N(t)) = \frac{\mathrm{var}[N(t)]}{\mathsf{E}\left[N(t)\right]} = \frac{2-\theta}{\theta} \geq 1.$$

V prípade ak $\theta \neq 1$, má PAP nadmerný rozptyl ($D(N(t)) > 1$), čo poskytuje väčšiu flexibilitu pri modelovaní diskrétnych náhodných veličín než bežný Poissonov proces, ktorého index disperzie je rovný 1.

Jedna z netriviálnych vlastností PAP je tvar distribučnej funkcie doby čakania. Dobu čakania na n-tý príchod udalostí značíme $S_n = T_1 + T_2 + \cdots + T_n$.

Tvrdenie 17. *Distribučná funkcia doby čakania S_n je určená vzťahom*

$$F_{S_n}(t) = 1 - e^{-\lambda t} \sum_{i=0}^{n-1} \binom{n-1}{i} \theta^i (1-\theta)^{n-1-i} \sum_{j=0}^{i} \frac{(\lambda t)^j}{j!}, \quad t > 0, n = 1, 2, \ldots \tag{3.8}$$

Dôkaz. Na dokázanie tohto tvrdenia využijeme matematickú indukciu.

- Pre $n = 1$ dostávame distribučnú funkciu $S_1 = T_1$:

$$F_{S_1}(t) = 1 - \mathbb{P}(N(t) = 0).$$

Z (3.1) pre $n = 0$ dostávame

$$F_{S_1}(t) = 1 - e^{-\lambda t}, \quad t > 0.$$

Zo všeobecných vlastností čítacích procesov vyplýva vzťah

$$F_{S_n}(t) = F_{S_{n-1}}(t) - \mathbb{P}(N(t) = n-1), \quad n = 2, 3, \dots \tag{3.9}$$

- Pre $n = 2$ vyzerá distribučná funkcia $S_2 = T_1 + T_2$ nasledovne:

$$F_{S_2}(t) = F_{S_1}(t) - \mathbb{P}(N(t) = 1) = 1 - (1 + \theta\lambda t)e^{-\lambda t}.$$

- Predpokladajme teraz, že pre $n \geq 2$ je distribučná funkcia doby čakania

$$F_{S_{n-1}}(t) = 1 - e^{-\lambda t} \sum_{i=0}^{n-2} \binom{n-2}{i} \theta^i (1-\theta)^{n-2-i} \sum_{j=0}^{i} \frac{(\lambda t)^j}{j!}. \tag{3.10}$$

Potom aplikovaním (3.1) pre $\mathbb{P}(N(t) = n-1)$ a po dosadení (3.10) do (3.9) a následnom upravení dostaneme (3.8).

□

Môžeme si všimnúť, že distribúcia (3.8) je zovšeobecnením Erlangovho rozdelenia. V prípade, že $\theta = 1$, zjednoduší sa táto distribúcia na Erlangovo rozdelenie s parametrami λ a n. Poslednou vlastnosťou, ktorú si uvedieme a ktorá na prvý pohľad nie je zrejmá je ekvivalencia PAP s martingálou.

Tvrdenie 18. *Majme náhodný proces $N(t) \sim \mathcal{PAP}(\lambda t, \theta)$, potom proces $M(t) = N(t) - \frac{\lambda}{\theta} t$ je martingála.*

Dôkaz. Pretože $\frac{\lambda}{\theta} t$ je nenáhodné, platí $\mathsf{E}\left(N(t) - \frac{\lambda}{\theta} t\right) = 0$ a $M(t)$ má nezávislé prírastky. Teda pre $s \leq t$ a $\mathcal{F}_t = \sigma\{N(s), s \leq t\}$ dostávame

$$\mathsf{E}\left[N(t) - \frac{\lambda}{\theta} t | \mathcal{F}_s\right] = \mathsf{E}\left[N(t) - N(s) - \frac{\lambda}{\theta}(t-s) | \mathcal{F}_s\right] + N(s) - \frac{\lambda}{\theta} s = N(s) - \frac{\lambda}{\theta} s.$$

□

4. Praktická časť

V praktickej časti sa budeme venovať odhadom parametrov geometrického Poissonovho rozdelenia. Ukážeme si dva spôsoby, ako ich odhadnúť a porovnáme výsledky. Dáta, ktoré na to využijeme budú z oblasti športu, presnejšie z tenisu. Zoberieme si tenistov a tenistky, ktorý už ukončili svoju profesionálnu kariéru a v nejakom momente počas nej dosiahli post svetovej jednotky. Týmto kritériom si oddelíme lepších tenistov od tých horších. Zo skupiny lepších tenistov potom u každého budeme sledovať dva údaje a to koľkých grandslamov sa zúčastnil a koľko zápasov v turnaji odohral. Označme Z ako počet výhier hráča v jednom grandslame. Jedná sa o náhodnú veličinu, ktorá môže nadobúdať hodnôt z množiny $\{0, 1, 2, \ldots, 7\}$, pretože systém vo vyraďovacej časti pri grandslamoch je nastavený tak, že hráč môže odohrať nanajvýš sedem zápasov. V prípade, že $Z = 7$, znamená to, že hráč vyhral daný grandslam. Označme vektor

$$\mathbf{z} = (Z_7, Z_6, \ldots, Z_0)^T$$

ako vektor počtu víťazných sérií jedného hráča, kde Z_7 značí koľkokrát sa mu podarilo vyhrať sedem po sebe idúcich zápasov, Z_6 šesť po sebe idúcich zápasov atď. Vezmime si teraz vektor váh

$$\mathbf{w} = (7, 7, 6, 5, 4, 3, 2, 1)^T.$$

Každá zložka tohto vektoru odpovedá počtu zápasov, koľko hráč odohral pri tom, ako ďaleko sa mu v turnaji podarilo dostať. Teda ak sa hráč dostal do semifinále grandslamu (kde aj skončil), máme $\mathbf{z} = (0, 0, 1, 0, 0, 0, 0, 0)^T$ a jeho celkový počet odohraných zápasov je

$$\mathbf{w}^T\mathbf{z} = (7, 7, 6, 5, 4, 3, 2, 1)\begin{pmatrix}0\\0\\1\\0\\0\\0\\0\\0\end{pmatrix} = 6. \tag{4.1}$$

Označíme tento počet odohraných zápasov ako X.

Dáta budeme čerpať z Wikipédie, kde nájdeme počty víťazných sérií každého hráča. Dáta pre mužov siahajú do roku 1973, pre ženy do roku 1975. Budeme chcieť pre tieto pozorovania nájsť parametre geometrického Poissonovho rozdelenia. V týchto dátach vieme nájsť interpretačnú analógiu s týmto rozdelením. Počet, koľkých grandslamov sa hráč zúčastnil, bude Poissonova časť rozdelenia a geometrická časť zase to, koľko zápasov odohral. Zhrnieme si teraz všetky dáta do jednej tabuľky. V stĺpci štarty nájdeme údaj o tom, na koľkých grandslamoch sa daný hráč zúčastnil. Nasleduje údaj o počte zápasov, ktorý hráč odohral. Tento počet dostávame zo vzťahu (4.1). Vektor $\mathbf{z}$ pritom dostaneme z nasledujúcich stĺpcov 7S, 6S, ..., 1S, 0S, ktoré reprezentujú to, koľkokrát sa hráčovi podarilo dosiahnuť 7 zápasovú, 6 zápasovú, ..., 1 zápasovú, 0 zápasovú šnúru víťazstiev za sebou. Príprava dát je tým pádom hotová.

Tabuľka 4.1: Svetové tenisové jednotky

Meno	Štarty	Zápasy	7S	6S	5S	4S	3S	2S	1S	0S
R. Federer	81	434	20	11	15	12	11	5	1	6
S. Williamsová	81	425	23	10	7	14	10	11	3	3
M. Navrátilová	67	366	18	14	12	9	5	2	2	5
Ch. Evertová	56	362	18	16	18	2	0	2	0	0
S. Grafová	54	312	22	9	6	5	5	3	1	3
J. Connors	57	284	8	7	16	10	2	3	5	6
A. Agassi	61	275	8	7	11	10	5	3	8	9
I. Lendl	55	263	7	11	9	5	8	3	5	7
A. Sánchézová Vicariová	57	260	4	8	9	13	8	2	6	7
M. Sharapovová	58	251	5	5	10	5	15	7	4	7
L. Davenportová	55	250	3	4	11	13	10	5	6	3
P. Sampras	52	241	14	4	5	6	7	2	7	7
S. Edberg	54	230	6	5	8	7	7	4	13	4
J. Newcombe	51	223	7	3	6	12	5	8	3	7
L. Hewitt	66	212	2	2	4	7	15	10	9	17
M. Selesová	40	212	9	4	5	13	3	3	2	1
J. McEnroe	45	208	7	4	8	7	7	2	4	6
E. Goolang Cawleyová	35	192	7	11	4	4	0	4	5	0
M. Hingisová	37	185	5	7	7	5	3	5	2	3
M. Wilander	44	184	7	4	3	6	6	5	7	6
J. Capriatová	42	179	3	0	10	10	8	1	3	7
A. Roddick	46	176	1	4	5	9	5	9	6	7
A. Mauresmová	46	175	2	1	5	9	9	8	7	5
K. Clijstersová	36	165	4	4	8	3	6	3	5	3
J. Jankovičová	57	170	0	1	5	2	14	10	12	13
K. Wozniacka	51	170	1	2	4	3	11	13	10	7
J. Heninová	35	169	7	5	5	2	8	1	3	4
B. Borg	27	158	11	5	1	4	3	2	1	0
J. Courier	42	157	4	3	4	4	7	4	8	8
A. Ivanovičová	48	155	1	2	2	3	11	13	8	8
I. Nastase	43	139	2	3	1	6	5	7	8	11
Y. Kafelnikov	38	135	2	1	3	7	3	8	11	3
M. Safin	41	134	2	2	3	2	7	7	11	7
C. Moyá	47	124	1	1	1	5	5	5	15	14
T. Muster	39	116	1	0	3	5	5	9	3	13
P. Rafter	35	109	2	2	3	0	6	6	5	11
D. Safinová	33	98	0	3	2	2	4	6	5	11
G. Kuerten	33	95	3	0	0	5	3	3	9	10
T. Austinová	17	82	2	0	3	8	1	1	1	1
A. Bartyová	27	80	3	0	1	2	3	5	4	9
M. Rios	26	77	0	1	0	5	4	4	5	7

4.1 Odhadovanie z úplných dát

Ukážeme prvú metódu odhadovania parametrov. Označme $S^{(k)} = \sum_{i=1}^{N^{(k)}} X_i^{(k)}$ ako celkový počet odohratých zápasov na grandslamových turnajoch k-teho hráča, kde sčítací index i značí i-ty turnaj a náhodná veličina $N^{(k)}$ počet turnajov, ktorých sa k-ty hráč zúčastnil. Náhodnú veličinu $X_i^{(k)}$ sme si už vysvetlili v úvode kapitoly. Chceme z dát $S^{(k)}$ odhadnúť pravdepodobnostné rozdelenie náhodnej veličiny S, pre ktorú tým pádom dostaneme odhady pravdepodobnosti $\mathbb{P}(S = m)$. Tieto pravdepodobnosti znamenajú to, že náhodne vybraný hráč zo zoznamu svetových jednotiek odohrá na grandslamoch počas svojej práve m zápasov. Aby sme mohli na naše dáta použiť geometrické Poissonovo rozdelenie, budeme uvažovať tieto predpoklady:

1. Počet absolvovaných grandslamov $N^{(k)}$ má Poissonovo rozdelenie,
2. počty postúpených kôl $X_i^{(k)}$ majú geometrické rozdelenie a
3. náhodné veličiny $N^{(k)}$ a $X_i^{(k)}$ sú navzájom nezávislé.

Budeme predpokladať, že sú splnené. Pretože máme k dispozícii úplné dáta ako o počte účastí, tak aj o počte zápasov a jednotlivých víťazných šnúr, odhadneme parametre pre každú časť rozdelenia (Poissonovu a geometrickú) samostatne a potom ich spojíme dokopy. Najprv odhadneme parameter pre Poissonovu časť $\widehat{\lambda_1}$. Odhad parametra pri Poissonovom rozdelení je priemer cez všetky pozorovania $\frac{1}{n}\sum_{i=1}^{n=41} N_i^{(k)}$. Priemer je nestranný a konzistentný odhad parametra $\widehat{\lambda_1} = 46.634$. Podobne budeme postupovať pri odhade parametra geometrického rozdelenia $\widehat{\theta_1}$. Využijeme znovu nestrannosti a konzistencie priemeru $\overline{X}_n$ a parameter θ_1 odhadneme ako $\widehat{\theta_1} = \frac{1}{\overline{X}_n/N_n} = 0.232$. Túto metódu pre určenie parametrov používajú autori v článku Özel a Inal (2010).

4.2 Odhadovanie z kumulatívnych súčtov

Druhý spôsob, ako odhadnúť parametre geometrického Poissonovho rozdelenia bude podobne ako v predošlého spôsobu pomocou momentovej metódy s tým rozdielom, že sa obmedzíme iba na dáta o počte odohratých zápasov. Tým pádom nemáme k dispozícii dáta o počte štartov na grandslamoch a nemôžeme odhadovať zvlášť Poissonovu a zvlášť geometrickú časť rozdelenia. Vyjdeme teda z rovníc (2.2), kde nahradíme strednú hodnotu výberovým priemerom $\overline{X}_n$ a rozptyl zase výberovým rozptylom S_n^2. Dostávame nasledujúcu sústavu rovníc:

$$\overline{X}_n = \frac{\widehat{\lambda_2}}{\widehat{\theta_2}} \quad a \quad S_n^2 = \frac{\widehat{\lambda_2}(2 - \widehat{\theta_2})}{\widehat{\theta_2^2}}.$$

Riešime sústavu dvoch rovníc o dvoch neznámych. Z prvej rovnice si vyjadríme $\widehat{\lambda_2} = \widehat{\theta_2}\overline{X}_n$, dosadíme do druhej a dostávame:

$$S_n^2 = \frac{\overline{X}_n(2 - \widehat{\theta_2})}{\widehat{\theta_2}}$$

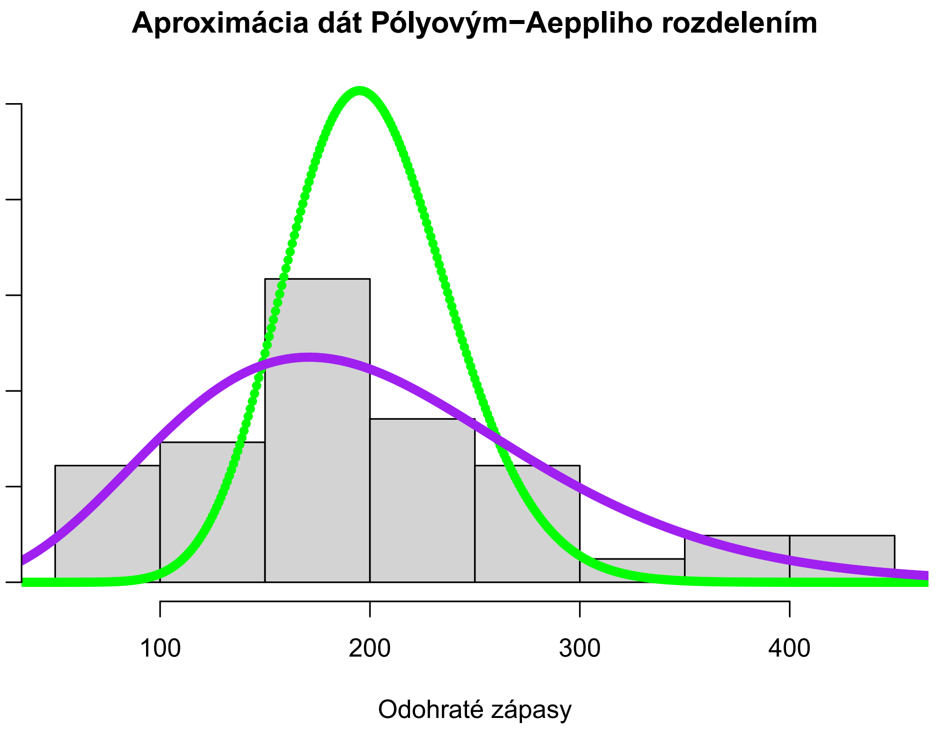
$$\widehat{\theta_2}\frac{S_n^2}{\overline{X}_n} + \widehat{\theta_2} = 2$$

$$\widehat{\theta_2} = \frac{2\overline{X}_n}{\overline{X}_n + S_n^2}.$$

Po dosadení dostávame odhady $\widehat{\lambda_2} = 10.121$ a $\widehat{\theta_2} = 0.05$.

4.3 Porovnanie

Dostali sme dve dvojice odhadov parametrov $(\widehat{\lambda_1}, \widehat{\theta_1})$ a $(\widehat{\lambda_2}, \widehat{\theta_2})$. V tejto sekcií sa teraz pozrieme na to, ktorý z uvedených spôsobov lepšie aproximuje naše dáta. Najlepšie to uvidíme na grafe, do ktorého si spolu s histogramom dát vykreslíme aj obe rozdelenia s odhadnutými parametrami.



Obr. 4.1: Histogram počtu zápasov. Rozdelenie s parametrami odhadnuté z úplných dát je znázornené zelenou farbou, fialová zase vykresľuje rozdelenie s parametrami získanými z kumulatívnych súčtov.

Z grafu je jasno vidieť, že rozdelenie s parametrami získanými z kumulatívnych súčtov lepšie aproximuje naše dáta. Dôvodom, prečo odhady parametrov, ktoré boli získané z úplných dát, ich nevystihujú až tak dobre je to, že pri tomto spôsobe máme reštriktívnejšie požiadavky. Pri kumulatívnych súčtoch nemáme informácie o tom, či dáta naozaj spĺňajú predpoklady geometrického Poissonovho rozdelenia, ako napríklad či počty účastí na grandslamoch majú naozaj Poissonovo rozdelenie. Pri odhadovaní z úplných dát ich k dispozícii máme a keďže sme ich predpokladali, môžeme si ich teraz overiť. Spočítame si napríklad index disperzie D pre štarty na grandslamoch. Po dosadení nám vyjde $D = 4.027$, čo naznačuje, že počty štartov asi nebudú pochádzať z Poissonovho rozdelenia, pri ktorom by sme tento index očakávali blízko 1.

Pozrime sa ešte na dáta o sériach víťazstiev, z ktorých potom vzniknú dáta o počte zápasov. Pre tieto sme predpokladali, že budú mať geometrické rozdelenie, čiže so zvyšujúcou sa hodnotou k by mala počet udalostí klesať. Je to dané vlastnosťami geometrického rozdelenia, ktoré nadobúda menšie hodnoty s väčšou pravdepodobnosťou. Môžeme si všimnúť, že tieto dáta klesajúcu tendenciu nemajú. Tým pádom zrejme nebudú pochádzať z geometrického rozdelenia. Čiastočne to môže byť spôsobené tým, že berieme iba tých najlepších. Ak by sme sa sústredili na priemerných tenistov, počty krátkych sérií víťazstiev za sebou by sa isto vyskytovali častejšie. Okrem toho máme pozorovania od Rogera Federera, Sereny Wiliamsovej, Martiny Navrátilovej a Chris Evertovej, ktorý sú na čele počtu odohraných zápasov s celkom veľkým rozdielom od ostatných tenistov. Tieto pozorovania nám spôsobujú ťažké konce rozdelenia, ktoré sa pri pravdepodobnostnej mase sústredenej okolo 170 ťažko odhadujú.

Čo sa nezávislosti veličín $N^{(k)}$ a $X_i^{(k)}$ týka, môžeme si previesť formálny test na jej overenie. Podobne ako to robia autori v článku Özel a Inal (2010), použijeme Spearmanov ρ-test. Na 5% α-hladine testu dostávame Spearmanovo $\rho = 0.827$ a p-hodnota je rovná $2.627e^{-11}$. Zamietame teda, že by počty štartov a počty odohraných zápasov na grandslamových turnajoch neboli nezávislé.

4.4 Zhrnutie

Prišli sme na to, že oboma spôsobmi sa dostaneme k zmysluplným odhadom parametrov (parameter θ vyšiel v oboch prípadoch v intervale $(0, 1]$). Pri modelovaní odhadovaní pomocou kumulatívnych súčtov nemáme až také prísne reštrikcie, preto sa krivka rozdelenia snaží čo najviac priblížiť pozorovaným dátam. Pri odhadovaní parametrov z úplných dát máme k dispozícii všetky údaje a preto je na mieste overiť, či spĺňajú predpoklady pravdepodobnostného modelu. Ak sú porušené, nebude výsledný model až tak presne vystihovať realitu.

Záver

V tejto práci sme si zaviedli najprv Pólyovo-Aeppliho rozdelenie ako zložené Poissonovo rozdelenie, v ktorom sa sčítance riadia geometrickým rozdelením. Uviedli sme si jeho pravdepodobnostnú funkciu v rekurzívnom tvare. Keďže je ale takýto spôsob počítania pravdepodobností $\mathbb{P}(S = k)$ časovo aj pamäťovo náročný, je nepraktické ho používať pre veľké čísla k. Z toho dôvodu sme si odvodili implicitný vzťah na ich výpočet. Vykreslili si potom pravdepodobnostné rozdelenie pre rôzne kombinácie parametrov a všímali sme si, ako sa chová.

Potom sme si zadefinovali Pólyov-Aeppliho proces ako nadstavbu zloženého Poissonovho procesu, ktoré využíva toto rozdelenie. Ako sme si už spomínali, tento proces ponúka riešenie pre problém s ekvidisperziou dát, ktorú predpokladá obyčajný Poissonov proces a s ktorou sa v praxi často nestretávame. Ukázali sme, že Pólyov-Aeppliho proces je špecifickým prípadom zloženého Poissonovho procesu, ktorý má index disperzie väčší ako 1. Znamená to, že modelované dáta neprichádzajú v čase v rovnako veľkých zhlukoch, ale sú viac rozptýlené. Uviedli sme si štyri spôsoby, ako vieme zadefinovať Pólyov-Aeppliho proces a ukázali sme, že sú si ekvivalentné. Poskytuje to výhodu v tom, že môžeme na proces pozerať z viacerých uhľov pohľadu a nemusíme problém, ktorý práve riešime, prevádzať na jeden konkrétny prípad čítacieho procesu. Odvodili sme si aj niektoré z vlastností Pólyovho-Aeppliho procesu, ako napríklad jeho strednú hodnotu a rozptyl, z ktorých sme potom zistili, že má tento proces naozaj nadmerný rozptyl (index disperzie > 1).

V praktickej časti sme si ukázali dva spôsoby odhadovania parametrov a porovnali ich. Zistili sme, že ak odhadujeme parametre rozdelenia zo všetkých dostupných dát, kladieme aj dôraz na to, či dáta splňujú predpoklady modelu. Ak tieto splnené nebudú, nebude ani výsledný model dobre vystihovať realitu.

Čo sa vlastného prínosu týka, spočíval hlavne vo rozpísaní jednotlivých krokov, ktoré boli v článkoch, z ktorých som čerpal, iba opomenuté. Bolo treba úplne doplniť niektoré dôkazy. Takisto, keďže som nečerpal iba z jedného článku, bolo potrebné zjednotiť značenie. Okrem toho nedávali niektoré formulácie zmysel, takže ich bolo nutné opraviť. Vo vzorcoch sa taktiež vyskytovali buď chybné indexy alebo iné preklepy, čiže aj tie som opravil. Okrem toho som v praktickej časti porovnal dva spôsoby odhadovania parametrov Pólyovho-Aeppliho rozdelenia.

Zoznam použitej literatúry

CHUKOVA, S. a MINKOVA, L. D. (2013). Characterization of the pólya-aeppli process. *Stochastic Analysis and Applications*, **31**(4), 590–599.

DALEY, D. J., VERE-JONES, D. A KOL. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

MINKOVA, L. D. (2004). The pólya-aeppli process and ruin problems. *Journal of Applied Mathematics and Stochastic Analysis*, **2004**(3), 221–234.

NUEL, G. (2008). Cumulative distribution function of a geometric poisson distribution. *Journal of Statistical Computation and Simulation*, **78**(3), 385–394.

ÖZEL, G. a INAL, C. (2010). The probability function of a geometric poisson distribution. *Journal of Statistical Computation and Simulation*, **80**(5), 479–487.

PRÁŠKOVÁ, Z. a LACHOUT, P. (2012). Základy náhodných procesů I. matfyzpress.

ROLSKI, T., SCHMIDLI, H., SCHMIDT, V. a TEUGELS, J. L. (1999). *Stochastic processes for insurance and finance*. John Wiley & Sons.