14. cvičení z PSI

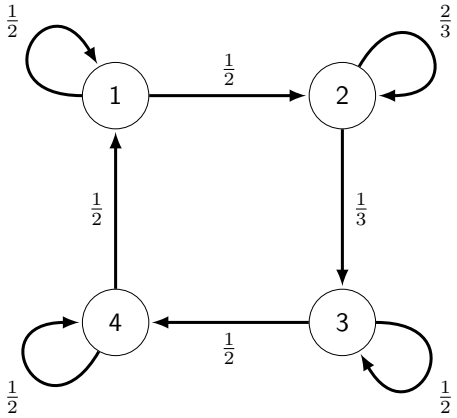
9. - 13. ledna 2023

14.1 (maximálně věrohodné odhady)
Odhadněte stav i a k Markovova řetězce s maticí přechodu

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 2/3 & 1/3 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}$$

z pozorované posloupnosti stavů $(2, i, k, 3)$.

Řešení:
Pro větší názornost si nakreslíme diagram:



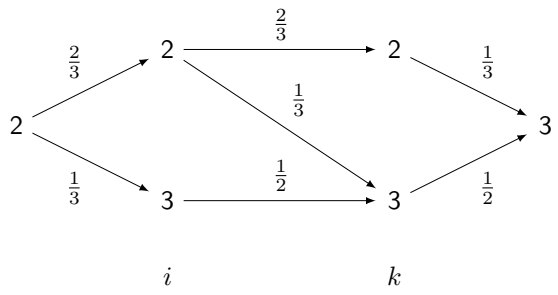
Stav odhadneme pomocí maximální věrohodnosti

$$L(i,k) = P\Big(X_0 = 2,\ X_1 = i,\ X_2 = k,\ X_3 = 3\Big).$$

V našem případě tak máme

$$L(i,k) = P(X_0 = 2) \cdot p_{2,i} \cdot p_{i,k} \cdot p_{k,3}\ .$$

Hodnotu počáteční pravděpodobnosti $c := P(X_0 = 2)$ sice neznáme, ale ani jí nepotřebujeme k výpočtu (za předpokladu, že byla nenulová). Abychom zjistili, které stavy i a k vůbec přicházejí (pro nenulovou věrohodnost) v úvahu, nakreslíme následující obrázek:



Vypsali jsme všechny stavy, na které přejde v jednom kroku počáteční stav 2 (druhý sloupec) a pak všechny stavy, které přejdou na koncový stav 3 (třetí sloupec). Mezi těmito dvěma sloupci nakreslíme všechny možné způsoby přechodu. Celkem máme tři možné cesty z počátečního 2 do koncového 3. Pak snadno dostáváme:

$$\begin{aligned} L(2,2) &= c \cdot \tfrac{2}{3} \cdot \tfrac{2}{3} \cdot \tfrac{1}{3} &&= \tfrac{4}{27} \cdot c \\ L(2,3) &= c \cdot \tfrac{2}{3} \cdot \tfrac{1}{3} \cdot \tfrac{1}{2} &&= \tfrac{1}{9} \cdot c = \tfrac{3}{27} \cdot c \\ L(3,3) &= c \cdot \tfrac{1}{3} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} &&= \tfrac{1}{12} \cdot c \\ L(i,k) &= 0\,, && \text{jinak.} \end{aligned}$$

Případ, pro který je hodnota věrohodnosti nejvyšší, je tedy $i = 2$ a $k = 2$ (za předpokladu, že $P(X_0 = 2) > 0$, jinak jsou všechny čtyři stavy stejně věrohodné).

14.2 (maximálně věrohodné odhady)

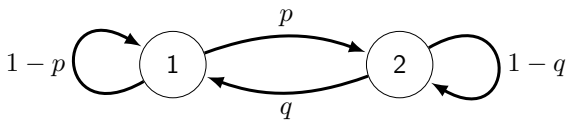
Markovův řetězec má dva stavy 1 a 2. Pravděpodobnost přechodu ze stavu 1 do stavu 2 je p, pravděpodobnost přechodu ze stavu 2 do stavu 1 je q. Z pozorované posloupnosti stavů

$$(1, 2, 2, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1)$$

odhadněte parametry p, q.

Řešení:

Markovovův řetězec má graf



a matici přechodu

$$\mathbb{P} = (p_{ij})_{i,j\in\{1,2\}} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Nechť n_{ij} jsou naměřené četnosti přechodu ze stavu i do j v naměřené posloupnosti stavů $(i_0, i_1, \ldots, i_k)$. Věrohodnostní funkce je pak tvaru

$$L(p,q) = P(X_0 = i_0,\ X_1 = i_1,\ \ldots, X_k = i_k) = P(X_0 = i_0) \cdot \prod_{\ell=0}^{k-1} p_{i_\ell, i_{\ell+1}} =$$

$$= P(X_0 = i_0) \cdot \underbrace{(1-p)^{n_{11}} \cdot p^{n_{12}}}_{=L_1(p)} \cdot \underbrace{q^{n_{21}} \cdot (1-q)^{n_{22}}}_{=L_2(q)}$$

Jestliže budeme předpokládat, že $P(X_0 = i_0) \neq 0$ (jinak bychom neměli co zjišťovat, protože všechny možnosti by byly stejně věrohodné), pak je jasné, že maxima bude dosaženo, pokud budou maximalizovány obě funkce L_i každá v dané proměnné.

Stačí tedy zjistit, kdy obecně nabývá maxima funkce

$$L(\alpha) := (1-\alpha)^n \cdot \alpha^m$$

pro proměnnou $\alpha \in \langle 0, 1\rangle$. Po zlogaritmovaní a zderivovaní máme

$$\lambda(\alpha) = \ln(L(\alpha)) = n \ln(1-\alpha) + m \ln \alpha$$

$$0 = \lambda'(\alpha) = -\frac{n}{1-\alpha} + \frac{m}{\alpha} = \frac{m - \alpha(n+m)}{(1-\alpha)\alpha} \quad \Leftrightarrow \quad \alpha = \frac{m}{n+m}$$

Pro původní $L(p,q)$ pak dostaneme maximum pro $p = \frac{n_{11}}{n_{11}+n_{12}}$ a $q = \frac{n_{22}}{n_{21}+n_{22}}$.

Pro naše zadání

$$(1,2,2,1,1,1,2,1,1,1,2,2,1,1,1,1)$$

a odpovídající tabulku n_{ij}:

i \ j	1	2
1	7	3
2	3	2

dostaneme tedy nejvěrohodnější matici přechodu tvaru

$$\mathbb{P} = \begin{pmatrix} \frac{7}{7+3} & \frac{3}{7+3} \\ \frac{3}{3+2} & \frac{2}{3+2} \end{pmatrix} = \begin{pmatrix} \frac{7}{10} & \frac{3}{10} \\ \frac{3}{5} & \frac{2}{5} \end{pmatrix}.$$

Tento výsledek je i heuristicky očekávatelný, protože pokud jsme např. ze stavu 1 šli 7-krát do stavu 1 a 3-krát do stavu 2, pak přirozeně nejlepší poměr mezi $1-p$ a p je $(1-p) : p = 7 : 3$, neboli $p = \frac{3}{7+3}$.

14.3 (aplikace Markovových řetězců, asymptotické pravděpodobnosti stavů)

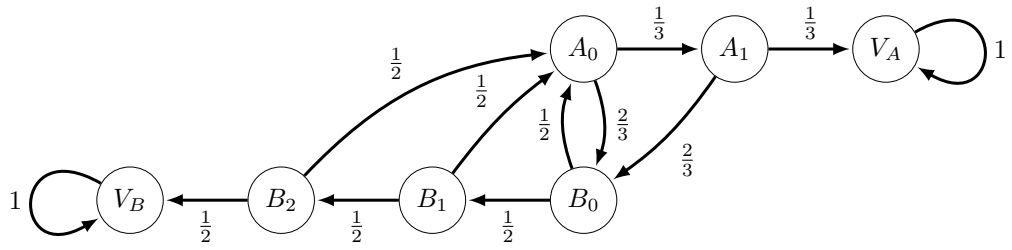
Alice trefí terč s pravděpodobností 1/3, Bob s pravděpodobností 1/2. Pokud hráč zasáhne terč, střílí dále, pokud mine, je na řadě druhý hráč. Začíná Alice. Alice vyhrává, pokud trefí terč 2× za sebou, Bob vyhrává, pokud trefí terč 3× za sebou. Pro oba hráče stanovte pravděpodobnosti výhry.

Řešení:

Pokud bychom rozlišovali nejen to, který hráč je na řadě, ale i kolik již má úspěšných pokusů, potřebovali bychom 7 stavů:

- V_A - vyhrála Alice,
- V_B - vyhrál Bob,
- A_i - Alice má právě za sebou i úspěšných pokusů $i \in \{0,1\}$,
- B_i - Bob má právě za sebou i úspěšných pokusů $i \in \{0,1,2\}$.

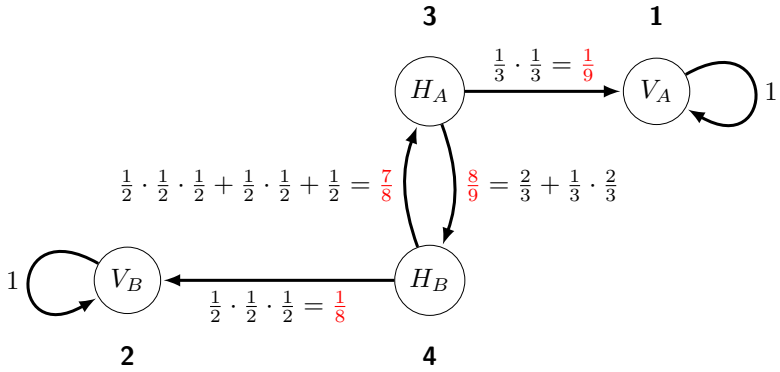
Odpovídající diagram by byl tento:



Protože nás zajímají pouze pravděpodobnosti výhry obou hráčů, můžeme si situaci popsat jednodušším způsobem a to tak, že rozlišíme pouze stavy:

- V_A - vyhrála Alice,
- V_B - vyhrál Bob,
- H_A - na řadě je Alice,
- H_B - na řadě je Bob,

kde celou sérii úspěšných pokusů daného hráče považujeme za jeden krok. Tento krok pak končí výhrou hráče s pravděpodobností $\left(\frac{1}{3}\right)^2$ pro Alici, $\left(\frac{1}{2}\right)^3$ pro Boba, nebo se na řadu dostává druhý hráč:



Stavy si očíslujeme tak, aby první byly ty absorpční a teprve po nich následovali ty přechodné:

$1 := V_A$ (vyhrála Alice), $2 := V_B$ (vyhrál Bob), $3 := H_A$ (na řadě je Alice), $4 := H_B$ (na řadě je Bob).

pravděpodobnosti výher Alice a Boba zjistíme z asymptotického rozdělení pravděpodobnosti $\boldsymbol{p}(\infty)$ s počátečním rozdělením

$$\boldsymbol{p}(0) = (0, 0, 1, 0)\ .$$

Je tedy potřeba spočítat $\mathbf{P}^\infty$ pro matici přechodu

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/9 & 0 & 0 & 8/9 \\ 0 & 1/8 & 7/8 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix},$$

kde

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1/9 & 0 \\ 0 & 1/8 \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} 0 & 8/9 \\ 7/8 & 0 \end{pmatrix}.$$

Fundamentální matice tohoto řetězce je

$$\mathbf{F} = \big(\mathbf{I}_2 - \mathbf{Q}\big)^{-1} = \begin{pmatrix} 1 & -8/9 \\ -7/8 & 1 \end{pmatrix}^{-1} = (9/2) \cdot \begin{pmatrix} 1 & 8/9 \\ 7/8 & 1 \end{pmatrix} = \begin{pmatrix} 9/2 & 4 \\ 63/16 & 9/2 \end{pmatrix}$$

a tedy

$$\big(\mathbf{I}_2 - \mathbf{Q}\big)^{-1}\mathbf{R} = \begin{pmatrix} 9/2 & 4 \\ 63/16 & 9/2 \end{pmatrix} \cdot \begin{pmatrix} 1/9 & 0 \\ 0 & 1/8 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 7/16 & 9/16 \end{pmatrix}.$$

Celkem tedy dostaneme

$$\mathbf{P}^{\infty} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ \big(\mathbf{I}_2 - \mathbf{Q}\big)^{-1}\mathbf{R} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 7/16 & 9/16 & 0 & 0 \end{pmatrix}$$

a asymptotické rozdělení tak je

$$\boldsymbol{p}(\infty) = \boldsymbol{p}(0) \cdot \mathbf{P}^{\infty} = (0,0,1,0) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 7/16 & 9/16 & 0 & 0 \end{pmatrix} = (1/2, 1/2, 0, 0)\ .$$

Zjistili jsme tak (vcelku překvapivě), že pravděpodobnosti výhry Alice i Boba jsou stejné (a sice $\frac{1}{2}$), pokud bude začínat Alice jako první.

Alternativní postup: K nalezení hodnoty výhry Alice (a tím i výhry Boba) stačí znát první sloupec matice $\mathbf{P}^{\infty}$, který je tvaru $\mathbf{u} = (1,0,a,c)^T$. Místo počítání fundamentální matice můžeme zase využít vztahu

$$\mathbf{P} \cdot \mathbf{P}^{\infty} = \mathbf{P}^{\infty}$$

ze kterého speciálně máme

$$\mathbf{P} \cdot \mathbf{u}^T = \mathbf{u}^T$$

neboli

$$\begin{pmatrix} 1 \\ 0 \\ a \\ c \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/9 & 0 & 0 & 8/9 \\ 0 & 1/8 & 7/8 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ a \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1/9 + 8/9 \cdot c \\ 7/8 \cdot a \end{pmatrix}$$

Vyřešením rovnic

$$a = \tfrac{1}{9} + \tfrac{8}{9}c$$

$$c = \tfrac{7}{8}a$$

tak máme $P(V_A) = a = \frac{1}{2}$ a $c = \frac{7}{16}$.

Poznámka: Uvažujme následující obecnější případ. Pro $n, m, a, b \in \mathbb{N}$ předpokládejme, že

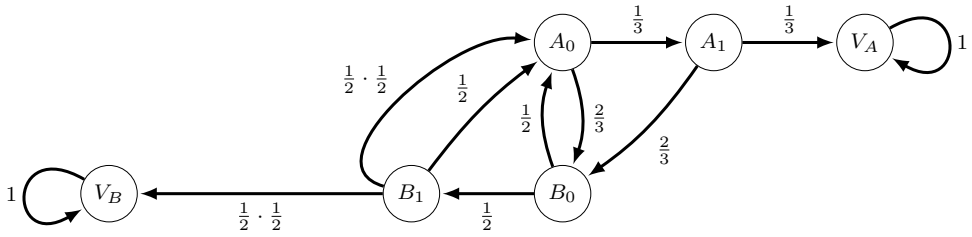
- Alice má pravděpodobnost zásahu $\frac{1}{n}$ a k výhře musí mít sérii a úspěšných pokusů a podobně
- Bob má pravděpodobnost zásahu $\frac{1}{m}$ a k výhře musí mít sérii b úspěšných pokusů a dále, že
- $\big(\frac{1}{n}\big)^a < \big(\frac{1}{m}\big)^b$ a proto opět necháme začít Alici.

Kdybychom opět chtěli, aby Alice a Bob měli stejné šance na výhru, zjistíme, že to nastane právě když bude platit

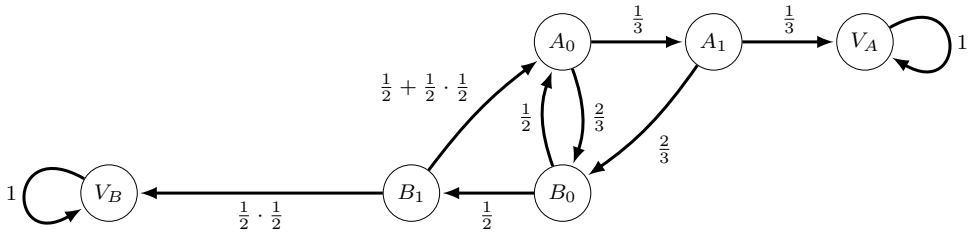
$$n^a - m^b = 1 \ .$$

V rámci teorie čísel se řešeními této rovnice zabýval Eugène Charles Catalan a v roce 1844 vyslovil hypotézu (tzv. Catalan's conjecture), že jediné řešení této rovnice v kladných přirozených číslech je právě jen $3^2 - 2^3 = 1$. Hypotézu potvrdil Preda Mihăilescu v roce 2002.

Ještě jedna poznámka: Můžeme si ještě podrobněji zdůvodnit, proč pravděpodobnosti výhry vycházejí u obou Markovových řetězců (složitějšího i jednoduššího) stejně. Pravděpodobnost výhry Boba (v obou grafech) představuje součet pravděpodobností všech cest v daném grafu vycházejících z A_0 (případně H_A) a končících v V_B. Nechť takováto cesta v prvním grafu obsahuje vrchol B_2. Protože tento vrchol je pouze průchozí a neobsahuje smyčku, můžeme např. cestu $A_0 \to \cdots \to B_1 \overset{\frac{1}{2}}{\to} B_2 \overset{\frac{1}{2}}{\to} V_B$ nahradit cestou $A_0 \to \cdots \to B_1 \overset{\frac{1}{2}\cdot\frac{1}{2}}{\to} V_B$, která bude mít stejnou hodnotu pravděpodobnosti a bude cestou v novém grafu, kde jsme vrchol B_2 vynechali a příslušně změnili hodnoty šipek:



Podobně jsme v cestách, které v původním grafu obsahovaly úsek $\cdots \to B_1 \overset{\frac{1}{2}}{\to} B_2 \overset{\frac{1}{2}}{\to} A_0 \to \ldots$ nahradili tento úsek takto: $\cdots \to B_1 \overset{\frac{1}{2}\cdot\frac{1}{2}}{\to} A_0 \to \ldots$ a tato nová cesta má opět stejnou pravděpodobnost jako ta, ze které vznikla. Celkově tedy zachováváme pravděpodobnost cest. Nyní máme nový graf, který má jen jednu nedokonalost a sice, že dva vrcholy jsou propojeny více než jednou šipkou. Tyto dvě šipky nyní spojíme do jedné a přiřadíme jí hodnotu součtu obou původních šipek. Tím opět zachováme pravděpodobnost cest - zde je jednodušší to zdůvodnit přes interpretaci s kapalinou: při jednom kroku proudí kapalina z B_1 do A_0 dvěma cestami, ale celkově ji přeteče stejně jako kdyby šla jen jednou cestou s odpovídajícím vyšším průtokem. Tímto tedy získáme nový graf, který už je skutečně Markovovým řetězcem:

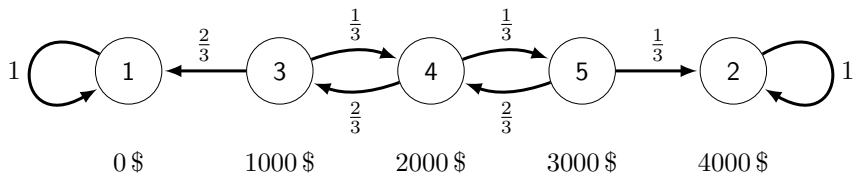


Opakováním tohoto postupu dospějeme k jednoduššímu řetězci se čtyřmi vrcholy (viz výše), přičemž pravděpodobnosti výher zůstanou zachovány.

14.4 (aplikace Markovových řetězců, asymptotické pravděpodobnosti stavů)

Alice hraje v kasinu hru, kde s pravděpodobností 1/3 vyhraje. V každém kole vsadí 1000 dolarů. V případě výhry získá 1000 dolarů, v případě prohry o 1000 dolarů přijde. Alice odejde z kasina, jestliže prohraje všechny své peníze nebo bude mít 4000 dolarů. Jaká je pravděpodobnost, že Alice odejde s prázdnou, měla-li na začátku 3000 dolarů?

Řešení:
Nakreslíme si příslušný orientovaný graf:



Pro Alici uvažujme stavy

1 - odchází s prázdnou, 2 - má 4000 dolarů (a tedy odchází), 3 - má 1000 dolarů, 4 - má 2000 dolarů a 5 - má 3000 dolarů.

Stavy jsme si očíslovali tak, aby první byly ty absorpční a teprve po nich následovali ty přechodné. Na začátku má Alice 3000 dolarů, tedy je ve stavu číslo 5 a počáteční rozdělení pravděpodobnosti tak je

$$\boldsymbol{p}(0) = (0, 0, 0, 0, 1) \ .$$

pravděpodobnost, že Alice odejde s prázdnou odpovídá složce pro stav 1 v asymptotickém rozdělení pravděpodobnosti $\boldsymbol{p}(\infty)$.

Pro výpočet asymptotického rozdělení pravděpodobnosti si opět zapíšeme matici přechodu

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 2/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 2/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 2/3 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix},$$

kde

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 2/3 & 0 \\ 0 & 0 \\ 0 & 1/3 \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} 0 & 1/3 & 0 \\ 2/3 & 0 & 1/3 \\ 0 & 2/3 & 0 \end{pmatrix}.$$

Opět si určíme matici

$$\mathbf{P}^\infty := \lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ (\mathbf{I}_3 - \mathbf{Q})^{-1}\mathbf{R} & \mathbf{0} \end{pmatrix} .$$

Spočítáme fundamentální matici $\mathbf{F} = \big(\mathbf{I}_3 - \mathbf{Q}\big)^{-1}$:

$$\Big(\mathbf{I}_3 - \mathbf{Q} \ \Big|\ \mathbf{I}_3\Big) = \left(\begin{array}{ccc|ccc} 1 & -1/3 & 0 & 1 & 0 & 0 \\ -2/3 & 1 & -1/3 & 0 & 1 & 0 \\ 0 & -2/3 & 1 & 0 & 0 & 1 \end{array}\right) \sim \cdots \sim$$

$$\sim \cdots \sim \left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{1}{5}\begin{pmatrix} 7 & 3 & 1 \\ 6 & 9 & 3 \\ 4 & 6 & 7 \end{pmatrix} \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \end{array}\right) = \Big(\mathbf{I}_3 \ \Big|\ \big(\mathbf{I}_3 - \mathbf{Q}\big)^{-1}\Big) \ .$$

Tedy

$$\mathbf{F} = \big(\mathbf{I}_3 - \mathbf{Q}\big)^{-1} = \begin{pmatrix} 1 & -1/3 & 0 \\ -2/3 & 1 & -1/3 \\ 0 & -2/3 & 1 \end{pmatrix}^{-1} = \frac{1}{5}\begin{pmatrix} 7 & 3 & 1 \\ 6 & 9 & 3 \\ 4 & 6 & 7 \end{pmatrix}$$

a

$$\big(\mathbf{I}_3 - \mathbf{Q}\big)^{-1}\mathbf{R} = \frac{1}{5}\begin{pmatrix} 7 & 3 & 1 \\ 6 & 9 & 3 \\ 4 & 6 & 7 \end{pmatrix} \cdot \frac{1}{3}\begin{pmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \frac{1}{15}\begin{pmatrix} 14 & 1 \\ 12 & 3 \\ 8 & 7 \end{pmatrix} \ .$$

Celkem tedy dostaneme

$$\mathbf{P}^\infty = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 14/15 & 1/15 & 0 & 0 & 0 \\ 12/15 & 3/15 & 0 & 0 & 0 \\ 8/15 & 7/15 & 0 & 0 & 0 \end{pmatrix}$$

Pravděpodobnost, že Alice vše prohraje, pokud na začátku měla 3000 dolarů, nyní odpovídá hodnotě

$$\left(\mathbf{P}^\infty\right)_{5,1} = \frac{8}{15} .$$

Alternativní postup: Hodnota pravděpodobnosti toho, že Alice odejde s prázdnou je $\left(\mathbf{P}^\infty\right)_{5,1}$. Stačí zase znát první sloupec matice $\mathbf{P}^\infty$, který je tvaru $\mathbf{u} = (1, 0, a, b, c)^T$. Místo počítání fundamentální matice můžeme zase využít vztahu

$$\mathbf{P} \cdot \mathbf{P}^\infty = \mathbf{P}^\infty$$

ze kterého speciálně máme

$$\mathbf{P} \cdot \mathbf{u}^T = \mathbf{u}^T$$

neboli

$$\begin{pmatrix} 1 \\ 0 \\ a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 2/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 2/3 & 0 & 1/3 \\ 0 & 1/3 & 0 & 2/3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2/3 + 1/3 \cdot b \\ 2/3 \cdot a + 1/3 \cdot c \\ 2/3 \cdot b \end{pmatrix}$$

Vyřešením rovnic

$$a = \tfrac{2}{3} + \tfrac{1}{3}b$$
$$b = \tfrac{2}{3}a + \tfrac{1}{3}c$$
$$c = \tfrac{2}{3}b$$

tak máme $\left(\mathbf{P}^\infty\right)_{5,1} = c = \frac{8}{15}$.

Poznámky k empirickému rozdělení:

Nechť $x_1 \leq \cdots \leq x_n$ jsou naměřené hodnoty (veličiny X). Pro ně si můžeme přirozeně definovat *empirickou* náhodnou veličinu Emp s diskrétním rozdělením, oborem hodnot

$$A = \{a \in \mathbb{R} \mid a = x_i \text{ pro nějaké } i\}$$

a jejich pravděpodobnostmi

$$P(\mathrm{Emp} = a) = \frac{\text{“ počet výskytů } a \text{ mezi hodnotami } x_1, \ldots, x_n \text{”}}{n} .$$

Když si k této veličině zjistíme distribuční funkci, dostaneme známou *empirickou distribuční funkci*:

$$F_{\mathrm{Emp}}(t) = P(\mathrm{Emp} \leq t) = \frac{\#\{i \mid x_i \leq t\}}{n}$$

Od ní si pak vytvoříme kvantilovou funkci q_{Emp}, která má tvar

$$q_{\mathrm{Emp}}(\alpha) = x_{\lceil n\alpha \rceil} \quad \text{pro} \quad \alpha \in (0, 1)$$

kde $\lceil u \rceil$ je horní celá část z $u \in \mathbb{R}$, tj. zaokrouhlení desetinných čísel nahoru. Speciální hodnoty se pak jmenují

- 1. kvartil $= q_{\mathrm{Emp}}(\frac{1}{4}) = x_{\lceil \frac{n}{4} \rceil}$
- 2. kvartil $= q_{\mathrm{Emp}}(\frac{2}{4}) = x_{\lceil \frac{n}{2} \rceil}$ (tzv. medián)
- 3. kvartil $= q_{\mathrm{Emp}}(\frac{3}{4}) = x_{\lceil \frac{3n}{4} \rceil}$

14.5 *Uvažujme následující data:*

(1) počty výskytů jistého druhu rostliny na ploše $1\,\mathrm{m}^2$*:*

$$0, \quad 2, \quad 1, \quad 4, \quad 4, \quad 5, \quad 2, \quad 3, \quad 7$$

(2) časy (v sekundách) mezi impulzy v mozku:

$$4.25, \quad 0.65, \quad 1.35, \quad 0.20, \quad 0.55, \quad 6.63, \quad 1.38, \quad 0.22, \quad 0.27$$

(3) venkovní teploty naměřené v různých letech při pravidelné podzimní akci:

$$8.07, \quad 19.23, \quad 9.27, \quad 5.71, \quad 12.62, \quad 11.24, \quad 11.92, \quad 17.30, \quad 14.87$$

Nakreslete pro tato data

(a) histogramy

(b) boxploty

(c) empirickou distribuční funkci

a odhadněte, z jakého rozdělení mohou tato data pocházet.

Řešení:
Histogram (pro četnosti): Naměřená data si rozdělíme do disjunktních intervalů I_i (stejné délky) pro $i = 1, \ldots, k$, které na sebe budou navazovat. Nad I_i nakreslíme sloupec výšky m_i, která znamená četnost dat, jež spadnou do I_i. Abychom z histogramu něco mohli vyčíst a uměli ho (ručně) nakreslit, volíme "rozumný" počet sloupců (např. něco mezi 5 a 15).

Boxplot (neboli krabicový graf): Na rozdíl od histogramu je vždy definován stejně. Krajní vousy ("whiskers") jsou dány krajními naměřenými hodnotami a krabice ("box") uprostřed je pak určena hodnotami jednotlivých kvartilů.

Počet měření je zde ve všech případech stejný: $n = 9$. Při uspořádaných datech $x_1 \leq \cdots \leq x_9$ tak budou hodnoty kvartilů tyto:

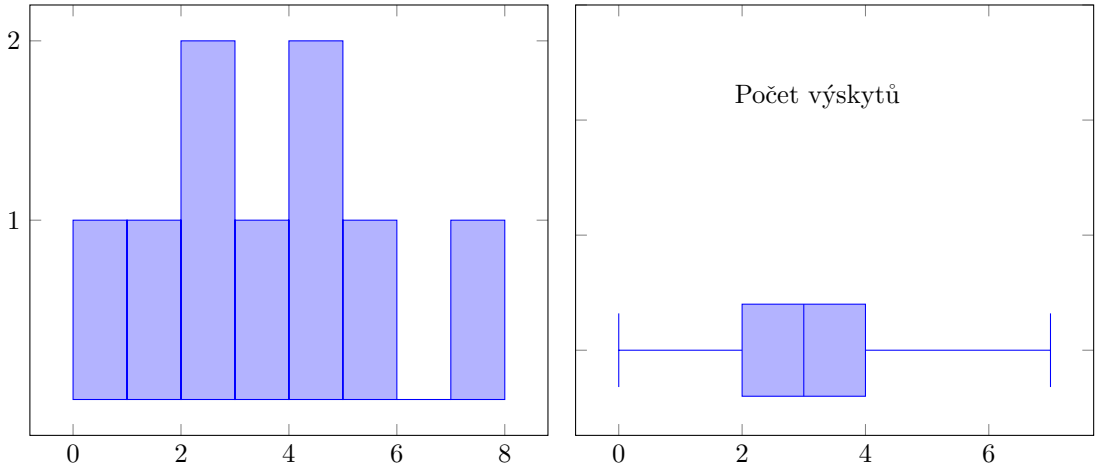
- 1. kvartil $= x_{\lceil \frac{9}{4} \rceil} = x_3$
- medián $= x_{\lceil \frac{9}{2} \rceil} = x_5$
- 3. kvartil $= x_{\lceil \frac{3 \cdot 9}{4} \rceil} = x_7$

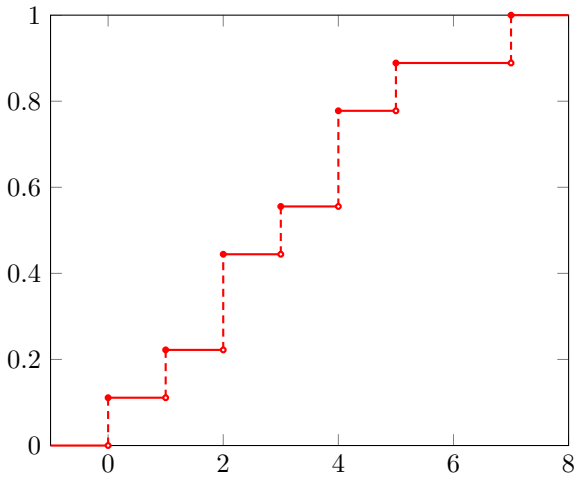
Medián je (v rámci uspořádání podle indexu) tedy přibližně uprostřed naměřených hodnot a podobně je to s okolními kvartily. Data si tudíž před výpočtem vždy uspořádáme.

(1) Uspořádaná data:

$$\begin{array}{cccccccccc} 0, & 1, & & 2, & 2, & 3, & 4, & 4, & 5, & 7 \\ x_1 & & 1.kvar. & & med. & & 3.kvar. & & & x_n \end{array}$$

Rozdíl mezi největší a nejmenší hodnotou je $x_n - x_1 = 7 - 0 = 7$. Tuto délku tedy budeme potřebovat pokrýt několika disjunktními intervaly a protože se zde jedná o diskrétní veličinu (počty výskytů), bude vhodné si zvolit šířku sloupce rovnou 1. Intervaly pak budou $\langle 0, 1), \langle 1, 2), \ldots, \langle 7, 8)$.



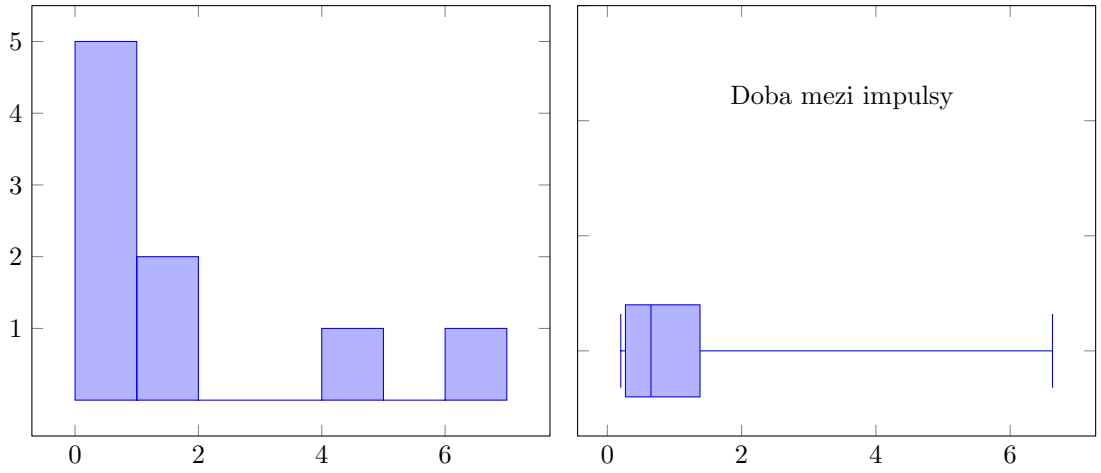


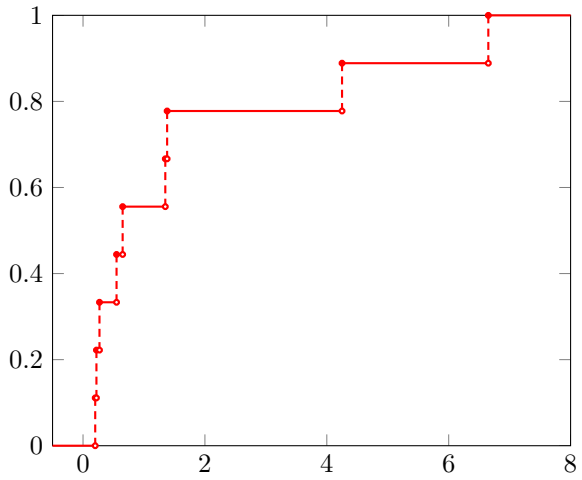
Vzhledem k popisu dat (počty výskytů na dané ploše) to vypadá na Poissonovo rozdělení. Tomu také zhruba odpovídají i grafická znázornění (histogram, boxplot, emp. distr. funkce).

(2) Uspořádaná data:

$$\begin{array}{cccccccccc} 0.20, & 0.22, & & 0.27, & 0.55, & 0.65, & 1.35, & 1.38, & 4.25, & 6.63 \\ x_1 & & 1.kvar. & & & med. & & 3.kvar. & & x_n \end{array}$$

Rozdíl mezi největší a nejmenší hodnotou je $x_n - x_1 = 6.63 - 0.2 = 6.42$. Tuto délku budeme zase potřebovat pokrýt několika disjunktními intervaly. Zkusíme si opět vzít šířku sloupce rovnou 1. Intervaly si pro změnu zvolíme jako $\langle 0, 1\rangle, \langle 1, 2\rangle, \ldots, \langle 6, 7\rangle$. Výběr toho, do kterého z intervalů přiřadíme dělicí body, není podstatný. Zde jsme si to takto zvolili čistě jen proto, že hodnoty čekací doby jsou vždy nenulové (tj. první interval by ideálně neměl začínat nulou).



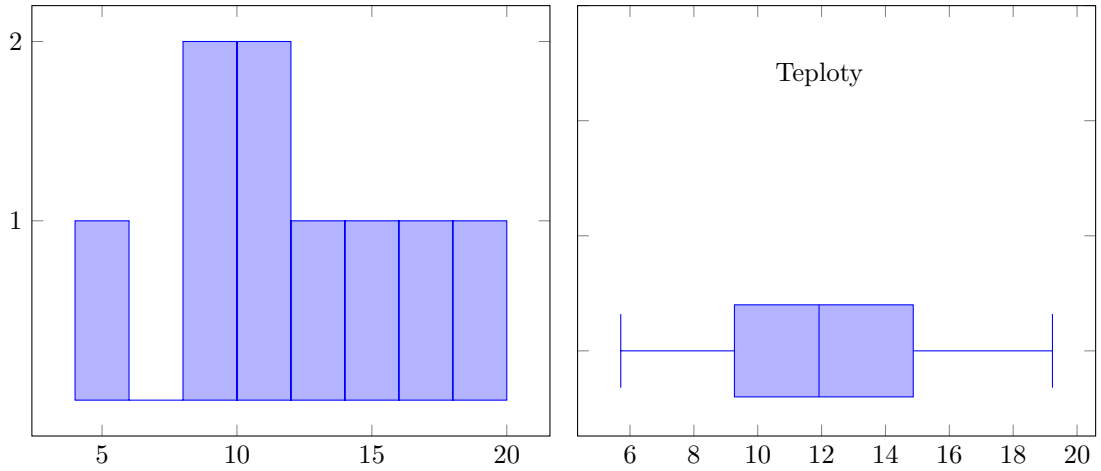


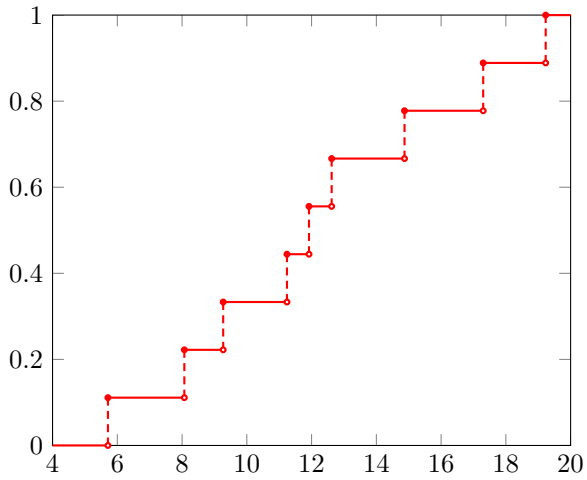
Vzhledem k popisu dat (doba čekání na další událost) to vypadá na exponenciální rozdělení. Tomu také zhruba odpovídají i grafická znázornění, kde boxplot je hodně posunutý doleva a empirická distribuční funkce připomíná exponenciálu.

(3) Uspořádaná data:

5.71,	8.07,	9.27,	11.24,	11.92,	12.62,	14.87,	17.30,	19.23
x_1		$1.kvar.$		$med.$		$3.kvar.$		x_n

Rozdíl mezi největší a nejmenší hodnotou je $x_n - x_1 = 19.23 - 5.71 = 13.52$ a tuto délku potřebujeme pokrýt několika disjunktními intervaly. Tady se nabízí vzít si větší (ideálně celočíselnou šířku), takže zkusíme šířku sloupce rovnou 2. Intervaly si zvolíme např. $\langle 4, 6), \langle 6, 8), \ldots, \langle 18, 20)$.





Vzhledem k popisu dat (hodnota ovlivněná mnoha malými výkyvy) to vypadá na normální rozdělení. Tomu také zhruba odpovídají i grafická znázornění (celkem symetrický boxplot i emp. distribuční funkce).