11. cvičení z PRA

2. a 9. května 2024

Příklad 11.1 *Basketbalista hází na tréninku na koš. Skončí, až nastřílí 20 košů. Hody jsou vzájemně nezávislé. Trenér mu počítal počty neúspěšných hodů před každou úspěšnou trefou a zjistil, že se trefil*

12 *krát ihned*
6 *krát až na druhý pokus*
2 *krát až na třetí pokus.*

Odhadněte basketbalistovu úspěšnost $p \in (0,1)$

(a) metodou momentů;

(b) metodou maximální věrohodnosti.

Řešení:
Úlohu můžeme řešit dvěma způsoby - buď pomocí geometrického rozdělení nebo pomocí alternativního rozdělení.

(1) *Pomocí geometrického rozdělení:*

Veličina

$$X = \text{“počet neúspěšných zásahů, než se basketbalista trefí”}$$

má geometrické rozdělení $\mathrm{Geom}(p)$ pro $p \in (0,1)$ a

$$P_p(X = i) = (1-p)^i p, \quad i \in \mathbb{N}_0 \ .$$

Z tabulky

hodnota i veličiny X	0	1	2
pozorovaná četnost n_i	12	6	2

vidíme, že počet měření je $n = \sum_i n_i = 12 + 6 + 2 = 20$. Naměřené hodnoty $(x_1, \ldots, x_{20})$ se skládají z hodnot $i \in \{0,1,2\}$, kde každá se vyskytuje se svojí četností n_i.

Metoda momentů:
Porovnáváme teoretické k-té momenty $E(X^k)$ s jejich odhady $m_k = \frac{1}{n}\sum_{j=1}^n x_j^k$ pro prvních několik $k = 1, 2, \ldots$.

Střední hodnota geometrického rozdělení $X \sim \mathrm{Geom}(p)$ je

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

a její odhad z realizace je

$$m_1 = \overline{\boldsymbol{x}} = \frac{\sum\limits_{j=1}^{n} x_j}{n} = \frac{\sum\limits_i i \cdot n_i}{\sum\limits_i n_i} = \frac{0 \cdot 12 + 1 \cdot 6 + 2 \cdot 2}{12 + 6 + 2} = \frac{10}{20} \ .$$

Porovnáním dostaneme

$$\frac{1}{\widehat{p}} - 1 = E(X) = \overline{x} = \frac{1}{2}$$

což dává řešení

$$\widehat{p} = \frac{2}{3} \ .$$

Metoda maximální věrohodnosti:
Hledáme hodnotu $p \in (0,1)$, která maximalizuje funkci věrohodnosti

$$L(p) = P_p(X_1 = x_1, \ldots, X_{20} = x_{20}) = \prod_{j=1}^{20} P_p(X_j = x_j) = \prod_{i \in \{0,1,2\}} \big(P_p(X = i)\big)^{n_i} =$$

$$= \Big((1-p)^0 p\Big)^{12} \Big((1-p)^1 p\Big)^{6} \Big((1-p)^2 p\Big)^{2} =$$

$$= \ (1-p)^{0\cdot 12 + 1\cdot 6 + 2\cdot 2} \cdot p^{12+6+2} \ = \ (1-p)^{10} \cdot p^{20}$$

kde X_j jsou jednotlivé nezávislé veličiny (odpovídající jednotlivým pokusům) a x_j naměřené hodnoty.
Ekvivalentně budeme hledat maximum funkce

$$\ell(p) = \ln\big(L(p)\big) = 10 \cdot \ln(1-p) + 20 \cdot \ln p \ .$$

Z její derivace

$$\ell'(p) = -\frac{10}{1-p} + \frac{20}{p} = \frac{-30p + 20}{(1-p)p}$$

dostáváme opět řešení

$$\ell'(\widehat{p}) = 0 \qquad \Longrightarrow \qquad \widehat{p} = \frac{2}{3}$$

které vyhovuje zadání, tj. $\widehat{p} \in (0,1)$. Ze znamének derivace je snadno vidět, že v $\widehat{p} = \frac{2}{3}$ je skutečně maximum.

- -

(2) *Pomocí alternativního rozdělení:*

Veličina

$$Y = \begin{cases} 1, & \textit{při daném hodu se basketbalista trefí} \\ 0, & \textit{při daném hodu se basketbalista netrefí} \end{cases}$$

má alternativní rozdělení $\mathrm{Alt}(p)$ pro $p \in (0,1)$, tedy

$$P_p(Y = 1) = p \quad \text{a} \quad P_p(Y = 0) = 1 - p \ .$$

Počet pokusů s veličinou Y bude nyní odpovídat počtu všech hodů $m = 12 \cdot 1 + 6 \cdot 2 + 2 \cdot 3 = 30$. Přitom máme $12 + 6 + 2 = 20$ úspěšných pokusů a tedy $30 - 20 = 10$ pokusů je neúspěšných. Už nyní bychom z toho snadno mohli usoudit, že odhad úspěšnosti musí být $\hat{p} = \frac{\textit{počet úspěšných hodů}}{\textit{počet všech hodů}} = \frac{20}{30}$.

Budeme ale postupovat podle zadání, tj. použijeme uvedené metody.
Máme tedy tabulku:

hodnota i veličiny Y	0	1
pozorovaná četnost n_i	10	20

Naměřené hodnoty $(y_1, \ldots, y_{30})$ se skládají z hodnot $i \in \{0, 1\}$, kde každá se vyskytuje se svojí četností n_i.

Metoda momentů:
Střední hodnota alternativního rozdělení $Y \sim \mathrm{Alt}(p)$ je

$$E(Y) = p$$

a její odhad z realizace je

$$m_1 = \overline{\boldsymbol{y}} = \frac{\sum_{j=1}^{m} y_j}{m} = \frac{\sum_i i \cdot n_i}{\sum_i n_i} = \frac{0 \cdot 10 + 1 \cdot 20}{10 + 20} = \frac{2}{3} \ .$$

Porovnáním dostaneme

$$p = E(Y) = \overline{y} = \frac{2}{3}$$

tedy

$$\widehat{p} = \frac{2}{3} \ .$$

Metoda maximální věrohodnosti:
Hledáme hodnotu $p \in (0, 1)$, která maximalizuje funkci věrohodnosti

$$L(p) = P_p(Y_1 = y_1, \ldots, Y_{30} = y_{30}) = \prod_{j=1}^{30} P_p(Y_j = y_j) = \prod_{i \in \{0,1\}} \big(P_p(Y = i)\big)^{n_i} =$$

$$= (1 - p)^{10} p^{20}$$

která je stejná jako v předchozím případě geometrického rozdělení. Tedy opět je $\widehat{p} = \frac{2}{3}$.

Důležité pozorování: Všimněme si, že obě metody dávají v obou případech stejné odhady a tato hodnota je opět společná při obou rozděleních.

Binomické rozdělení zde nemůžeme použít, protože bychom museli dopředu mít dán pevný počet všech hodů, který je znám až po pokusu v závislosti na tom, jak pokus probíhal. A tento počet by pak v dalších pokusech byl opět jiný.

Poznámka k věrohodnostní funkci pro spojitá rozdělení: Pro metodu max. věrohodnosti se u diskrétního rozdělení využívá pravděpodobnosti, že daná hodnota x_0 bude *přesně* nabyta, tj. $P(X = x_0)$. Tyto pravděpodobnosti by ale byly v případě spojitého rozdělení vždy nulové. Musíme tedy použít nějakou jinou charakteristiku v daném bodě a zde se nabízí hustota f_X. Jak ale víme, hustota není určena svými hodnotami, ale jen svými integrály. My ovšem nebudeme ani tak chtít zkoumat hustotu v bodě x_0, nýbrž spíše chování výrazu $P\Big(X \in (x_0 - \varepsilon, x_0 + \varepsilon)\Big)$ pro $\varepsilon \to 0+$. Dá se ukázat, že pokud je hustota f_X spojitá v x_0, pak platí

$$\lim_{\varepsilon \to 0+} \frac{1}{2\varepsilon} \cdot P\Big(X \in (x_0 - \varepsilon, x_0 + \varepsilon)\Big) = f_X(x_0) \ .$$

Tedy v tomto případě je chování daného výrazu skutečně přibližně úměrné hodnotě $f_X(x_0)$.

Proto se ve věrohodnostní funkci nakonec opravdu hustota používá, ale za předpokladu, že je f_X je *spojitá* buď všude nebo v oboru hodnot, který je otevřenou množinou (důvodem je to, že limitu děláme z obou stran). Např. pro exponenciální rozdělení (které modeluje dobu čekání) je obor hodnot $(0, +\infty)$ a tam už hustotu spojitou máme (přestože na celém $\mathbb{R}$ spojitá není).

Příklad 11.2 *Doba do poruchy přístroje má exponenciální rozdělení. Bylo zjištěno, že se přístroj porouchal postupně za* 4 *dny,* 7 *dní,* 12 *dní,* 2.5 *dne a* 24.5 *dne. Metodou maximální věrohodnosti (příp. metodou momentů) určete parametr* λ *tohoto exponenciálního rozdělení.*

Řešení:
Máme tedy veličinu

$$X = \text{“doba do poruchy přístroje”} \text{ [ve dnech]}$$

s exponenciálním rozdělením $Exp(\lambda)$ a hustotou $f_\lambda(x) = \begin{cases} \lambda\,\mathrm{e}^{-\lambda x} & \text{pro } x > 0 \\ 0 & \text{pro } x \le 0. \end{cases}$

Počet měření je $n = 5$ a jejich hodnoty jsou $x_1 = 4$ dny, $\ldots, x_5 = 24.5$ dne.

Metoda maximální věrohodnosti:
Obor hodnot X je $(0, +\infty)$, což je otevřený interval a hustota je zde spojitá. (Hodnotu 0 neuvažujeme, protože jako čekací dobu má smysl brát jen kladné hodnoty.)

Hledáme takové $\lambda > 0$, které maximalizuje věrohodnostní funkci

$$L(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} \lambda\,\mathrm{e}^{-\lambda x_i} = \lambda\,\mathrm{e}^{-\lambda\cdot 4} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 7} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 12} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 2.5} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 24.5} =$$

$$= \lambda^5\,\mathrm{e}^{-\lambda\cdot(4+7+12+2.5+24.5)} = \lambda^5\,\mathrm{e}^{-\lambda\cdot 50}\ .$$

Logaritmicko-věrohodnostní funkce je

$$\ell(\lambda) = \ln L(\lambda) = 5\ln\lambda - 50\lambda.$$

Z její derivace

$$\ell'(\lambda) = \frac{5}{\lambda} - 50.$$

získáme řešení

$$\frac{5}{\hat\lambda} - 50 = 0 \qquad \Longrightarrow \qquad \widehat\lambda = \frac{1}{10}\ [\text{den}^{-1}] \qquad \Longrightarrow \qquad \widehat\tau = \frac{1}{\hat\lambda} = 10\ [\text{dnů}]$$

(ve kterém skutečně nastává maximum, jak je vidět ze znamének derivace.)

Metoda momentů:
Chceme, aby platily rovnosti $E(X^k) = m_k$ teoretických a výběrových momentů pro co nejvíce počátečních hodnot $k = 1, 2, \ldots$.

Máme

- střední hodnotu $E(X) = \frac{1}{\lambda}$
- výběrový průměr $\overline{\boldsymbol{x}} = m_1 = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{4+7+12+2.5+24.5}{5} = \frac{50}{5} = 10$

Z požadované rovnosti $\frac{1}{\lambda} = E(X) = \overline{\boldsymbol{x}} = 10$ dostáváme opět odhad $\widehat\lambda = \frac{1}{10}$. Tato shoda je opět způsobena tím, že parametr $\tau = \frac{1}{\lambda}$ má význam střední hodnoty X a ta se nejlépe odhaduje pomocí výběrového průměru $\overline{\boldsymbol{x}}$.

Poznámka: Pro následující rozdělení veličiny X dávají obě výše probírané metody stejné výsledky pro daný parametr:

- p pro alternativní $Alt(p)$, odhad je $\hat p = E(X) = \overline{x}$

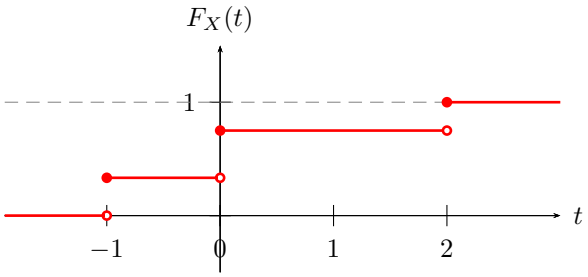
- p pro binomické $Bi(n,p)$, odhad je $n\hat{p} = E(X) = \overline{x} \;\Rightarrow\; \hat{p} = \frac{\overline{x}}{n}$
- p pro geometrické $Geom(p)$, odhad je $\frac{1}{\hat{p}} - 1 = E(X) = \overline{x} \;\Rightarrow\; \hat{p} = \frac{1}{\overline{x}+1}$
- λ pro Poissonovo $Poiss(\lambda)$, odhad je $\hat{\lambda} = E(X) = \overline{x}$
- τ pro exponenciální $Exp(\frac{1}{\tau})$, odhad je $\hat{\tau} = E(X) = \overline{x}$
- μ pro normální $N(\mu, \sigma^2)$, odhad je $\hat{\mu} = E(X) = \overline{x}$

Metoda maximální věrohodnosti je v těchto případech výpočetně složitější než metoda momentů, která je zde velmi snadná. V písemkách je ovšem smyslem zadání prověřit znalost použití zvolené metody (obvykle právě metody maximální věrohodnosti), takže znalost výsledku získaného jiným způsobem je pouze kontrola, že nám to vyšlo správně.

- -

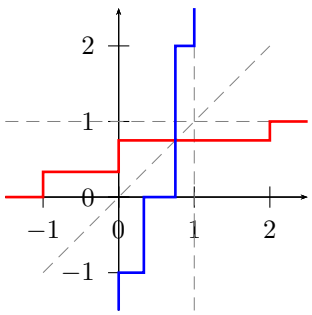
Poznámky ke kvantilům:

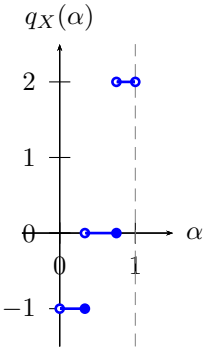
Pro náhodnou veličinu X a pravděpodobnost $\alpha \in (0,1)$ často potřebujeme najít $t \in \mathbb{R}$, že $P(X \leq t) = \alpha$, tj. $F_X(t) = \alpha$. Takové t obecně nemusí existovat (např. kdy F_X má skoky) nebo nemusí být určeno jednoznačně (když F_X je místy konstantní). Například si vezměme tuto distribuční funkci:



Chtěli bychom tedy ideálně mít inverzní funkci k F_X, která ale obecně neexistuje. Přesto můžeme něco podobného, tzv. *kvantilovou funkci* $q_X : (0,1) \to \mathbb{R}$, definovat (díky tomu, že F_X je neklesající) a to následujícím způsobem:

- graf F_X doplníme na "souvislou čáru", tj. případné skoky funkce F_X nahradíme spojitou svislou úsečkou,
- tento útvar převrátíme podle osy 1. a 3. kvadrantu (tj. podle přímky "$x = y$"),
- tam, kde převrácený útvar není funkcí (tj. obsahuje svislé čáry) tyto úseky odstraníme a nahradíme jedinou hodnotou, a sice limitou zleva (a případné krajní úseky v bodech 0 a 1 odstraníme úplně, protože tam se kvantil q_X nedefinuje)
- výsledným útvarem si pak definujeme graf funkce q_X.





Jak je tedy vidět, grafy funkcí F_X a q_X (po doplnění na souvislé čáry) si budou navzájem zrcadlovými obrazy (vhledem k ose $x = y$). Takováto definice kvantilu je sice názorná, ale chtělo by to i explicitní popis. Platí:

- $q_X(\alpha) = \min\{t \in \mathbb{R} \mid F_X(t) \geq \alpha\}$ pro všechna $\alpha \in (0,1)$
- $P\Big(X \leq q_X(\alpha)\Big) = F_X\Big(q_X(\alpha)\Big) \geq \alpha$ pro všechna $\alpha \in (0,1)$
- $P\Big(X < q_X(\alpha)\Big) \leq \alpha$ pro všechna $\alpha \in (0,1)$
- q_X je neklesající a zleva spojitá funkce
- Jestliže je F_X spojitá a ostře rostoucí, pak q_X je inverzní funkcí k F_X.
  V tom případě pak pro všechna $\alpha \in (0,1)$ platí:
  - $q_X(\alpha) = (F_X)^{-1}(\alpha)$
  - $P\Big(X \leq q_X(\alpha)\Big) = \alpha$

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

$$q_{\mathrm{Emp}}(\alpha) = \min\{t \in \mathbb{R} \mid F_{\mathrm{Emp}}(t) \geq \alpha\} = \min\{x_j \mid F_{\mathrm{Emp}}(x_j) \geq \alpha\} .$$

a nakonec se dá přepsat jako

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

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 $= q_{\mathrm{Emp}}(\frac{3}{4}) = x_{\lceil \frac{3n}{4} \rceil}$

Podobným způsobem se kvartily definují pro libovolnou veličinu X (jako $q_X(\frac{1}{4})$, $q_X(\frac{1}{2})$ a $q_X(\frac{3}{4})$).
Pro libovolnou veličinu X (a speciálně pro $X = \mathrm{Emp}$) platí:

$$P\Big(\ X \ \leq \ \textit{1. kvartil}\ \Big) \geq \frac{1}{4}$$

$$P\Big(\ \textit{1. kvartil} \ \leq \ X \ \leq \ \textit{3. kvartil}\ \Big) \geq \frac{1}{2} .$$

$$P\Big(\ X \ \geq \ \textit{3. kvartil}\ \Big) \geq \frac{1}{4}$$

Příklad 11.3 *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 $= x_{\lceil \frac{3 \cdot 9}{4} \rceil} = x_7$

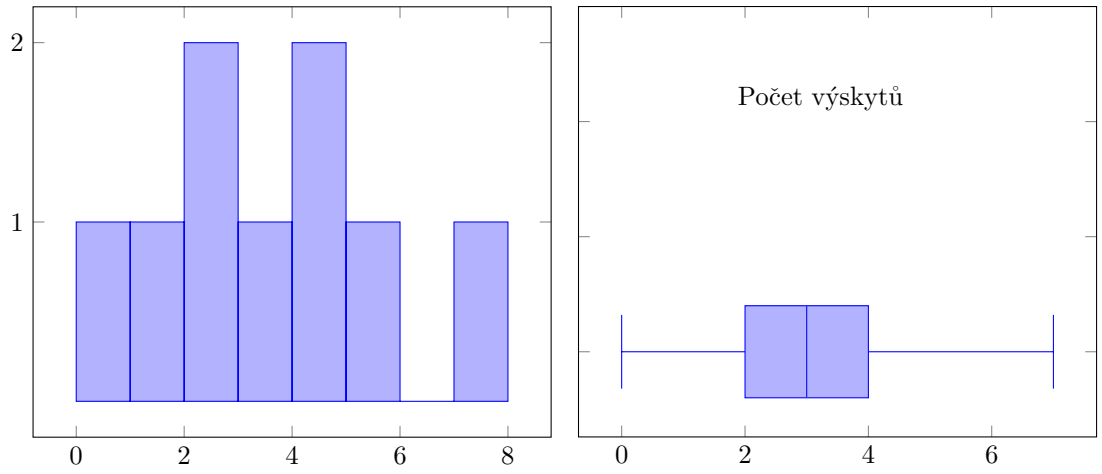
Medián je (v rámci uspořádání podle indexu) tedy přibližně uprostřed naměřených hodnot a podobně je to s okolními kvartily. Data si tudíž před výpočtem vždy uspořádáme.

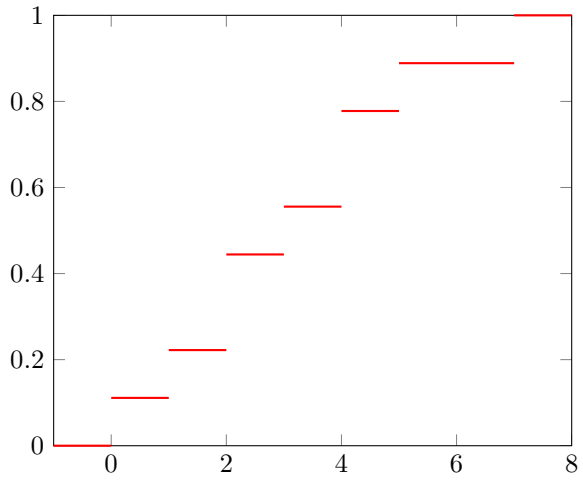
(1) Uspořádaná data:

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

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 $\langle 0,1), \langle 1,2), \ldots, \langle 7,8)$.



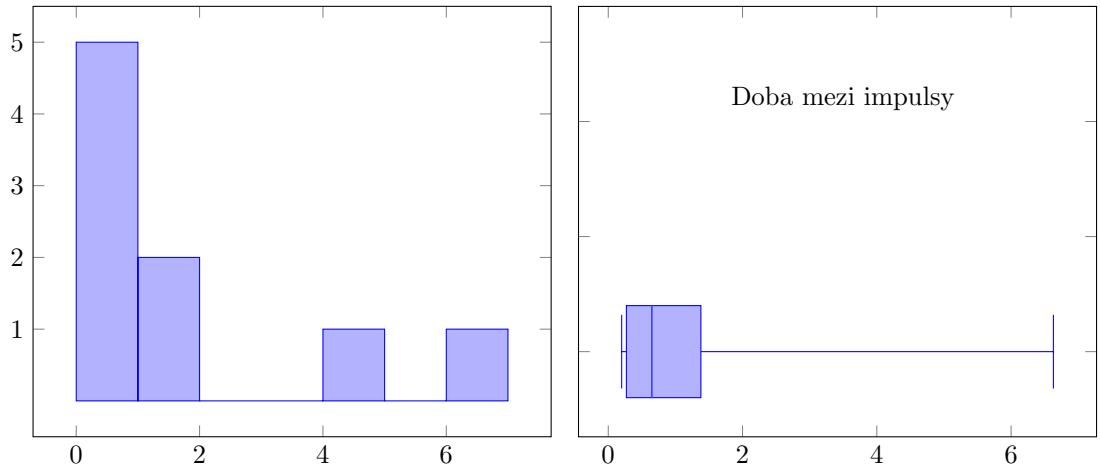


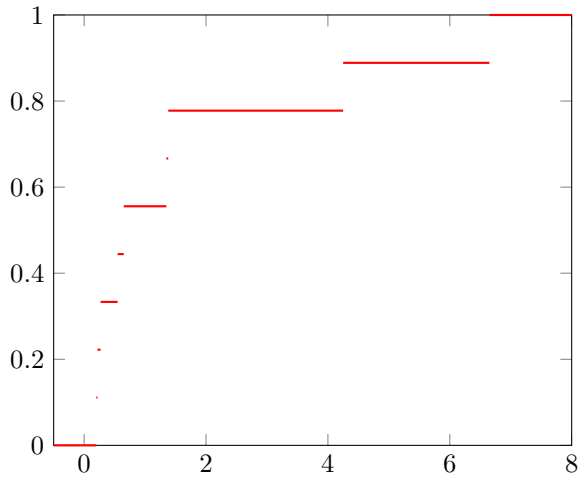
Vzhledem k popisu dat (počty výskytů na dané ploše) to vypadá na Poissonovo rozdělení. Tomu také zhruba odpovídají i grafická znázornění (histogram, boxplot, emp. distr. funkce).

(2) Uspořádaná data:

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

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 $(0,1\rangle, (1,2\rangle, \ldots, (6,7\rangle$. 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.