11. cvičení z STP

25. - 29. dubna 2022

Příklad 11.1 *Počet bodů ze zkouškové písemky je náhodná veličina pohybující se v rozmezí* $0-100$*, s průměrem* 53 *a rozptylem* 839*. Celkem* 200 *studentů psalo zkouškový test. Určete pravděpodobnost, že průměrný počet bodů u těchto studentů byl menší než* 50*, a uveďte za jakých předpokladů.*

Řešení:
Označme si X_i počet bodů i-tého studenta, $i = 1, 2, \ldots, n$, pro $n = 200$. Ze zadání je $E(X_i) = 53$ a $\mathrm{var}(X_i) = 839$. Veličina $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ (tzv. výběrový průměr) představuje průměrný počet bodů skupiny n studentů. Zajímá nás tedy pravděpodobnost

$$P\left(\overline{X}_n < 50\right).$$

Předpokládáme, že veličiny X_i jsou nezávislé (tj. např. že studenti opisují), a odsud dostaneneme

$$E\left(\overline{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}E(X_1) = 53$$

$$\mathrm{var}\left(\overline{X}_n\right) = \frac{1}{n^2}\mathrm{var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{\mathrm{var}(X_1)}{n} = \frac{839}{200}$$

Podle CLV tedy máme

$$P\left(\overline{X}_n < 50\right) = P\left(\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\mathrm{var}(\overline{X}_n)}} < \frac{50-53}{\sqrt{\frac{839}{200}}}\right) = P\left(\mathrm{norm}(X) < \underbrace{-\frac{3\sqrt{200}}{\sqrt{839}}}_{\doteq -1.46}\right) \doteq$$

$$\doteq \Phi(-1.46) = 1 - \Phi(1.46) \doteq 1 - 0.93 = 0.07.$$

Příklad 11.2 *Náhodná veličina* X *nabývá hodnot s pravděpodobnostmi dle tabulky, kde* c, q *jsou reálné parametry rozdělení. Z četností hodnot v náhodném výběru, uvedených v tabulce, odhadněte parametry* c *a* q*.*

hodnota i	1	2	3
pravděpodobnost $P_{(c,q)}(X = i)$	$c - q$	c	$c + q$
četnost n_i	8	10	5

Řešení:
Protože součet pravděpodobností všech hodnot je 1, musí být

$$1 = (c - q) + c + (c + q) = 3c$$

tedy $c = \frac{1}{3}$. Současně musí být pravděpodobnosti nezáporné, tj. $0 \le c - q = \frac{1}{3} - q$ a $0 \le c + q = \frac{1}{3} + q$, takže $|q| \le \frac{1}{3}$. Zbývá tedy odhadnout parametr q.

Metoda maximální věrohodnosti:

Hledáme hodnotu q, která maximalizuje funkci věrohodnosti

$$L(q) = P_{(\frac{1}{3},q)}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{j=1}^{n} \underbrace{P_{(\frac{1}{3},q)}(X_j = x_j)}_{P_{(\frac{1}{3},q)}(X = x_j)} =$$

$$= \left(\frac{1}{3} - q\right)^{n_1} \cdot \left(\frac{1}{3}\right)^{n_2} \cdot \left(\frac{1}{3} + q\right)^{n_3} = \left(\frac{1}{3} - q\right)^{8} \cdot \left(\frac{1}{3}\right)^{10} \cdot \left(\frac{1}{3} + q\right)^{5}$$

kde X_j jsou jednotlivé nezávislé veličiny (v pokusech) a x_j naměřené hodnoty. Funkce L je nezáporná a spojitá na uzavřené množině $\langle -\frac{1}{3}, \frac{1}{3} \rangle$, takže zde nabývá maxima. To nemůže být v krajních bodech (tam je nulová) a proto je nabyto uvnitř dané množiny. To odpovídá hledání maxima funkce

$$\ell(q) = \ln\big(L(q)\big) = 8 \cdot \ln\left(\frac{1}{3} - q\right) + 5 \cdot \ln\left(\frac{1}{3} + q\right) + \mathit{konst.}$$

na intervalu $(-\frac{1}{3}, \frac{1}{3})$. Protože maximum existuje, musí pro něj platit

$$0 = \ell'(q) = \frac{-8}{\frac{1}{3} - q} + \frac{5}{\frac{1}{3} + q}$$

Odhad parametru q je

$$q = -\frac{1}{13} \doteq -0.077 \ .$$

Odhady pravděpodobností hodnot $1, 2, 3$ jsou tedy

$$p_X(1) = \frac{16}{39} \doteq 0.41 \qquad p_X(2) = \frac{1}{3} \doteq 0.333 \qquad p_X(3) = \frac{10}{39} \doteq 0.256$$

což vyhovuje zadání.

Metoda momentů:

Střední hodnota je

$$E(X) = \left(\frac{1}{3} - q\right) + 2 \cdot \frac{1}{3} + 3 \cdot \left(\frac{1}{3} + q\right) = 2 + 2\,q$$

její odhad z realizace je

$$\overline{x} = \frac{1}{n} \sum_i i \cdot n_i = \frac{1}{8 + 10 + 5} \cdot (1 \cdot 8 + 2 \cdot 10 + 3 \cdot 5) = \frac{43}{23} \ .$$

Porovnáním dostaneme

$$2 + 2q = E(X) = \overline{x} = \frac{43}{23}$$

což odpovídá hodnotě

$$q = -\frac{3}{46} \doteq -0.065 \ .$$

Odhady pravděpodobností hodnot $1, 2, 3$ jsou tedy

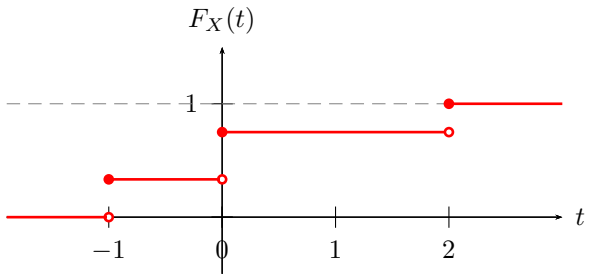
$$p_X(1) = \frac{55}{138} \doteq 0.399 \qquad p_X(2) = \frac{1}{3} \doteq 0.333 \qquad p_X(3) = \frac{37}{138} \doteq 0.268$$

což opět vyhovuje zadání.

Jak je vidět, metoda max. věrohodnosti by vyšla stejně, ať by veličina X měla jakékoliv hodnoty (dokonce i nečíselné), zatímco metoda momentů velmi podstatně závisí právě na tom, jaké hodnoty veličina X má (např. kdybychom místo hodnot $\{1, 2, 3\}$ zvolili třeba $\{-3, 10, 4\}$, byl by výsledek úplně jiný a dokonce bychom ani žádný přijatelný odhad nemuseli takto získat.)

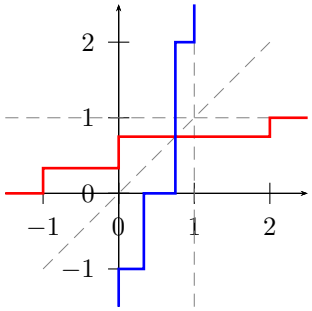
Poznámky ke kvantilům:

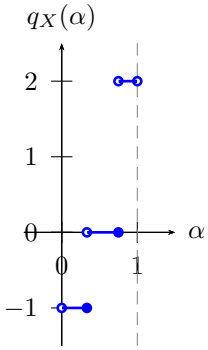
Pro náhodnou veličinu X a pravděpodobnost $\alpha \in (0,1)$ často potřebujeme najít $t \in \mathbb{R}$, že $P(X \leq t) = \alpha$, tj. $F_X(t) = \alpha$. Takové t obecně nemusí existovat (např. kdy F_X má skoky) nebo nemusí být určeno jednoznačně (když F_X je místy konstantní). Například si vezměme 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 $= x_{\lceil \frac{3 \cdot 9}{4} \rceil} = x_7$

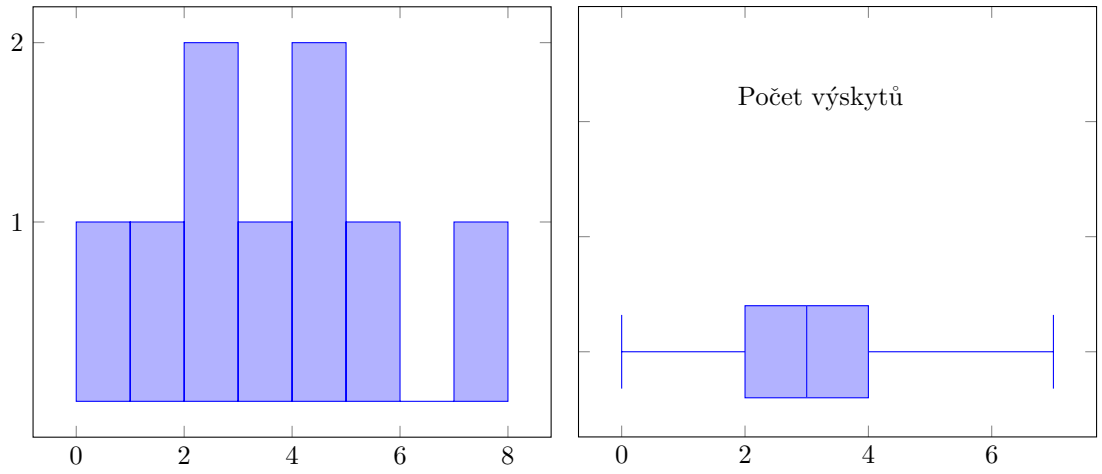
Medián je (v rámci uspořádání podle indexu) tedy přibližně uprostřed naměřených hodnot a podobně je to s okolními kvartily. Data si tudíž před výpočtem vždy uspořádáme.

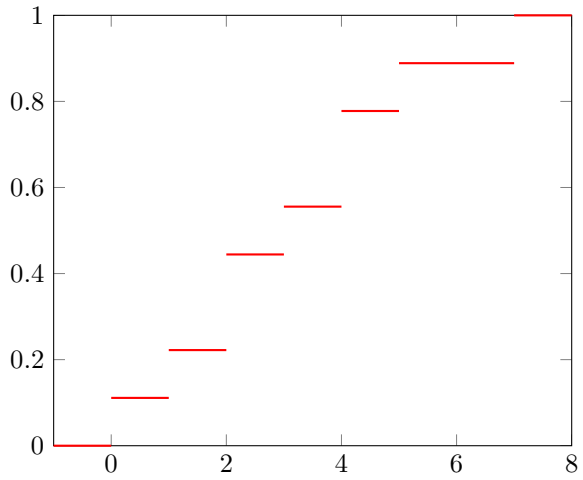
(1) Uspořádaná data:

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

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



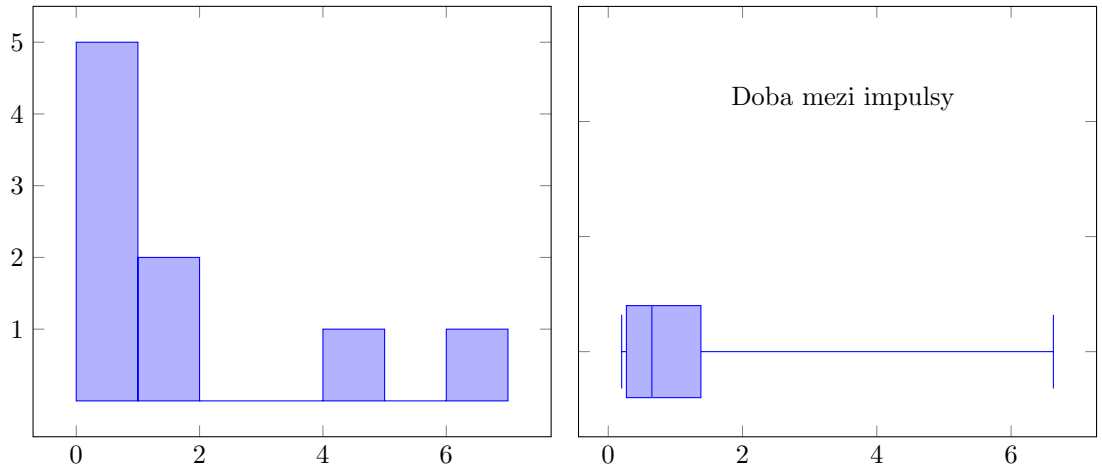


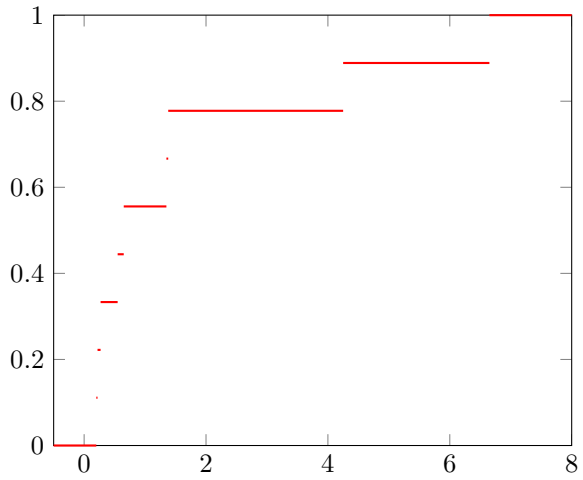
Vzhledem k popisu dat (počty výskytů na dané ploše) to vypadá na Poissonovo rozdělení. Tomu také zhruba odpovídají i grafická znázornění (histogram, boxplot, emp. distr. funkce).

(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).



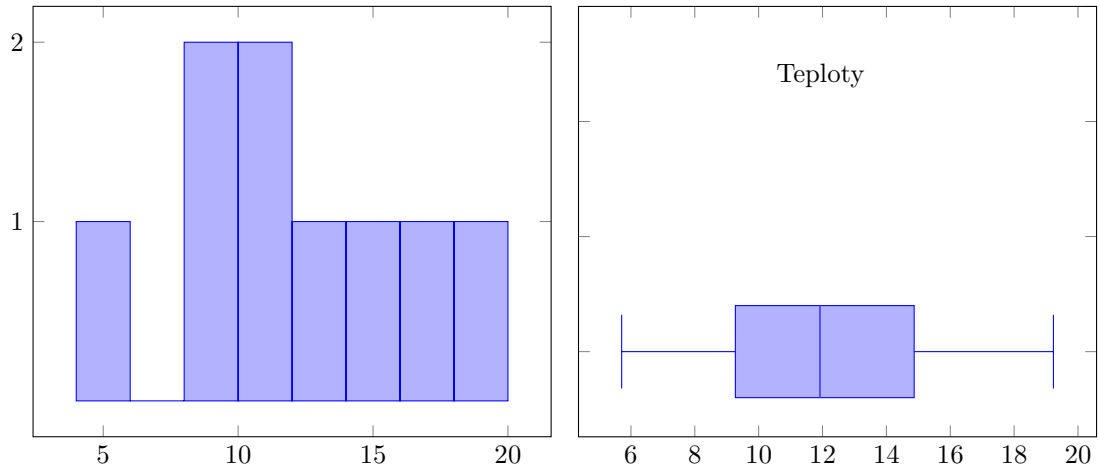


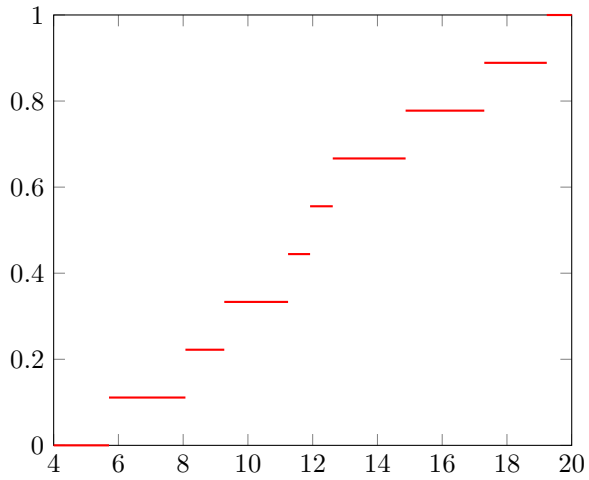
Vzhledem k popisu dat (doba čekání na další událost) to vypadá na exponenciální rozdělení. Tomu také zhruba odpovídají i grafická znázornění, kde boxplot je hodně posunutý doleva a empirická distribuční funkce připomíná exponenciálu.

(3) Uspořádaná data:

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

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).