13. cvičení z STP

13. května 2021

Poznámky k testu nezávislosti: Máme veličiny

- X s (různými) hodnotami $\{a_1, \dots, a_k\}$ a
- Y s (různými) hodnotami $\{b_1, \dots, b_\ell\}$

a chceme otestovat (na hladině α), hypotézu

$\mathbf{H}_0$: rozdělení veličin X a Y jsou *nezávislá*

proti alternativní hypotéze:

$\mathbf{H}_A$: rozdělení veličin X a Y jsou *závislá*

Při n pokusech s náhodným vektorem (X, Y) si pro $i = 1, \dots, k$ označme veličiny

$$N_{i,j} = \text{“počet výskytů případu } (X,Y) = (a_i, b_j) \text{ při } n \text{ pokusech”} .$$

a opět máme náhodný vektor

$$\boldsymbol{N} = (N_{1,1}, \dots, N_{k,\ell})$$

s multinomickým rozdělením. Marginální rozdělení jednotlivých veličin $N_{i,j}$ jsou opět binomická a *za předpokladu nezávislosti* X *a* Y mají střední hodnotu

$$E(N_{i,j}) = n \cdot P(X = a_i, Y = b_j) \overset{\text{(nezáv.)}}{=} n \cdot P(X = a_i) \cdot P(Y = b_j) .$$

Podobně jako v testu dobré shody by tyto střední hodnoty představovaly teoretické četnosti až na to, že pravděpodobnosti $P(X = a_i)$ a $P(Y = b_j)$ nemáme v hypotéze uvedeny. Proto je odhadneme jako

$$P(X = a_i) \doteq \frac{n_{i,\bullet}}{n} \quad \text{a} \quad P(Y = b_j) \doteq \frac{n_{\bullet,j}}{n}$$

kde $n_{i,\bullet}$ a $n_{\bullet,j}$ jsou naměřené hodnoty veličin

$$N_{i,\bullet} = \sum_{j=1}^{\ell} N_{i,j} = \text{“počet výskytů případu } X = a_i \text{ při } n \text{ pokusech”}$$

$$N_{\bullet,j} = \sum_{i=1}^{k} N_{i,j} = \text{“počet výskytů případu } Y = b_j \text{ při } n \text{ pokusech”}$$

což jsou tzv. *marginální četnosti*.

Z tohoto důvodu jako testovací veličinu volíme:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{\left(N_{i,j} - \frac{N_{i,\bullet} \cdot N_{\bullet,j}}{n}\right)^2}{\frac{N_{i,\bullet} \cdot N_{\bullet,j}}{n}}$$

která má asymptoticky (tj. pro $n \to \infty$) opět χ^2-rozdělení, tentokrát ale s $(k-1) \cdot (\ell - 1)$ stupni volnosti. Pro praktické použití této asymptotiky se obvykle opět požaduje, aby platilo, že

$$\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} \geq 5 \quad \text{pro všechna} \quad i = 1, \dots, k \text{ a } j = 1, \dots, \ell.$$

Kritérium pro **ZAMÍTNUTÍ $\mathbf{H}_0$** (na hladině α) volíme podobně jako u testu dobré shody a sice

$$t > \chi^2_{1-\alpha;\,(k-1)\cdot(\ell-1)} \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) \,.$$

Příklad 13.1 *Na $n = 100$ osobách byla pozorována barva očí a vlasů. Naměřeny byly následující sdružené četnosti:*

Oči \ *Vlasy*	*tmavé*	*světlé*
modré	10	20
šedé	10	10
hnědé	40	10

(a) Jsou barvy očí a vlasů nezávislé? Testujte na hladině 5%.

(b) Otestujte na hladině 5%, jestli je v populaci stejně tmavovlasých jako světlovlasých.

Řešení:

Označme si veličiny

$X =$ *"barva očí daného člověka"*

$Y =$ *"barva vlasů daného člověka"*

a dál budeme pracovat s náhodným vektorem (X, Y), tj. u daného člověka budeme zjišťovat barvu očí a barvu vlasů.

(a) Budeme testovat hypotézu:

$\mathbf{H}_0$: rozdělení veličin X a Y jsou *nezávislá*

proti alternativní hypotéze:

$\mathbf{H}_1$: rozdělení veličin X a Y jsou *závislá*.

na hladině významnosti $\alpha = 5\%$.

Četnost případu $(X, Y) = (i, j)$ v tabulce označme jako $n_{i,j}$ a marginální četnosti pak budou

$n_{i,\bullet} = \sum_j n_{i,j}$ pro případ $X = i$

$n_{\bullet,j} = \sum_i n_{i,j}$ pro případ $Y = j$.

což jsou součty v řádcích a sloupcích tabulky:

$n_{i,j}$ $(Y =)\ j$ / $(X =)\ i$	tmavé	světlé	$n_{i,\bullet}$
modré	10	20	30
šedé	10	10	20
hnědé	40	10	50
$n_{\bullet,j}$	60	40	

Za předpokladu $\mathbf{H}_0$ pak jako teoretické četnosti budeme chápat hodnoty $\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}$ v této tabulce:

$\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}$ $(Y =)\ j$ $(X =)\ i$	tmavé	světlé	$n_{i,\bullet}$
modré	$\frac{30 \cdot 60}{100} = 18$	$\frac{30 \cdot 40}{100} = 12$	30
šedé	$\frac{20 \cdot 60}{100} = 12$	$\frac{20 \cdot 40}{100} = 8$	20
hnědé	$\frac{50 \cdot 60}{100} = 30$	$\frac{50 \cdot 40}{100} = 20$	50
$n_{\bullet,j}$	60	40	

Podmínka na tyto teoretické (tj. očekávané) četnosti ≥ 5 je splněna, takže test nezávislosti můžeme použít. Pro hodnotu testovací statistiky dostaneme

$$t = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}} =$$

$$= \frac{(10-18)^2}{18} + \frac{(10-12)^2}{12} + \frac{(40-30)^2}{30} + \frac{(20-12)^2}{12} + \frac{(10-8)^2}{8} + \frac{(10-20)^2}{20} = 18 + \frac{1}{18} \doteq 18.056 .$$

Tuto hodnotu dále porovnáme s hodnotou kvantilu χ^2 pro $(k-1)(\ell-1)$ stupňů volnosti, kde k je počet položek veličiny X a ℓ je počet položek veličiny Y. Tento počet je nyní jiný, než by byl u "obvyklého" testu dobré shody s $k \cdot \ell$ položkami, protože data jsme použili k odhadu marginálních pravděpodobností.

Kritérium pro **ZAMÍTNUTÍ H**$_0$ (na hladině α) bude tedy tvaru

$$t > \chi^2_{1-\alpha;\ (k-1)(\ell-1)} \quad \Leftrightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha) .$$

Hledaný kvantil je

$$\chi^2_{1-\alpha;\ (3-1)(2-1)} = \chi^2_{0.95;\ 2} \doteq 5.992 .$$

Protože

$$t \doteq 18.056 > 5.992 \doteq \chi^2_{0.95;\ 2} ,$$

hypotézu o nezávislosti **ZAMÍTÁME**.

- -

(b) V tomto případě budeme uvažovat pouze veličinu Y a testovat (na hladině $\alpha = 5\%$) hypotézu

$\widetilde{\mathbf{H}}_0$: veličina Y má rozdělení s pravděpodobnostmi $(p_1, p_2) = \left(\frac{1}{2}, \frac{1}{2}\right)$,

proti alternativní hypotéze

$\widetilde{\mathbf{H}}_A$: veličina Y má rozdělení *jiné* než $\left(\frac{1}{2}, \frac{1}{2}\right)$.

Využijeme teď opět test dobré shody. Celkový počet měření je zase $n = 100$. Naměřené četnosti odpovídají už spočítaným marginálním četnostem pro hodnoty veličiny Y, tedy $n_i = n_{\bullet,i}$. Pro přehlednost si zase vypíšeme tabulku s jednotlivými četnostmi (pozorovanými i teoretickými):

i (barvy vlasů)	tmavé	světlé
n_i (pozorované četnosti)	60	40
p_i (teoretické pravděpodobnosti)	$\frac{1}{2}$	$\frac{1}{2}$
$n \cdot p_i$ (teoretické četnosti)	$100 \cdot \frac{1}{2} = 50$	$100 \cdot \frac{1}{2} = 50$

Vidíme, že všechny teoretické četnosti jsou ≥ 5, takže můžeme použít asymptotické přiblížení pro testovací statistiku T. Teď už si jen spočítáme hodnotu této statistiky

$$t = \sum_{i=1}^{2} \frac{(n_i - np_i)^2}{np_i} = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = 2 + 2 = 4$$

a porovnáme s kvantilem χ^2-rozdělení s $k - 1 = 2 - 1 = 1$ stupněm volnosti:

$$t = 4 > 3.84 \doteq \chi^2_{0.95;\,1} = \chi^2_{1-\alpha;\,k-1}$$

Protože zamítací kritérium JE splněno, tak $\widetilde{\mathbf{H}}_0$ **ZAMÍTÁME** (na hladině α).

Příklad 13.2 (test nezávislosti veličin)
Úspěšnost u zkoušek ve vztahu k počtu přítomných studentů udává tabulka:

termín	1.	2.	3.	4.	5.
počet přítomných	20	30	40	60	50
počet úspěšných	11	8	14	43	24

Otestujte na hladině významnosti 5 % *hypotézu, že pravděpodobnost úspěchu byla u všech zkouškových termínů stejná.*

Řešení:
Označme si veličiny

$X($ *"účast daného studenta na j-tém termínu"* $) :=$ *"zda uspěl/neuspěl"*

$Y($ *"účast daného studenta na j-tém termínu"* $) :=$ j

Pravděpodobnost úspěchu v j-tém termínu je dána jako $P(X = uspěl \mid Y = j)$.

Ukážeme, že tato hodnota bude nezávislá na j, právě když veličiny X a Y budou nezávislé.

- -

Důkaz:
($\Leftarrow$): Z nezávislosti X a Y ihned máme, že $P(X = uspěl \mid Y = j) = P(X = uspěl\ ) = konst.$

($\Rightarrow$): Naopak, nechť $c := P(X = uspěl \mid Y = j) = \frac{P(X = uspěl\ \cap\ Y = j)}{P(Y = j)}$ pro všechna j. Tedy

$$P(X = uspěl\ \cap\ Y = j) = c \cdot P(Y = j)$$

a sečtením přes všechna j dostaneme, že

$$P(X = uspěl\ ) = \sum_j P(X = uspěl\ \cap\ Y = j) = \sum_j c \cdot P(Y = j) = c \cdot \underbrace{\sum_j P(Y = j)}_{=1} = c$$

neboli

$$P(X = uspěl\ \cap\ Y = j) = P(X = uspěl\ ) \cdot P(Y = j)\ .$$

Podobně z $P(X = neuspěl \mid Y = j) = 1 - c$ pro všechna j odvodíme, že

$P(X = neuspěl\ \cap\ Y = j) = P(X = neuspěl\ ) \cdot P(Y = j)$. Tedy veličiny X a Y jsou nezávislé.

- -

Úloha tedy ve skutečnosti znamená, že budeme testovat hypotézu:

$\mathbf{H}_0$: rozdělení veličin X a Y jsou *nezávislá*

proti alternativní hypotéze:

$\mathbf{H}_1$: rozdělení veličin X a Y jsou *závislá*.

na hladině významnosti $\alpha = 5\%$.

Četnosti n_{ij} jednotlivých případů pro $X = i$ a $Y = j$ přepíšeme pomocí tabulky

$n_{i,j}$ $(Y=)\ j$ / $(X=)\ i$	1	2	3	4	5	$n_{i,\bullet}$
uspěl	11	8	14	43	24	100
neuspěl	9	22	26	17	26	100
$n_{\bullet,j}$	20	30	40	60	50	200

kde

$$n_{\bullet,i} = \sum_j n_{i,j} \quad \text{a} \quad n_{j,\bullet} = \sum_i n_{i,j}$$

Za předpokladu $\mathbf{H}_0$ pak jako teoretické četnosti budeme opět chápat hodnoty $\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}$ v této tabulce:

$\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}$ $(Y=)\ j$ / $(X=)\ i$	1	2	3	4	5	$n_{i,\bullet}$
uspěl	$\frac{20\cdot 100}{200} = 10$	$\frac{30\cdot 100}{200} = 15$	20	30	25	100
neuspěl	$\frac{20\cdot 100}{200} = 10$	15	20	30	25	100
$n_{\bullet,j}$	20	30	40	60	50	200

Podmínka na teoretické (tj. očekávané) četnosti ≥ 5 je splněna, takže položky nemusíme slučovat. Pro realizaci testovací statistiky dostaneme

$$t = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}} =$$

$$= \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(8-15)^2}{15} + \frac{(22-15)^2}{15} + \frac{(14-20)^2}{20} + \frac{(26-20)^2}{20} +$$
$$+ \frac{(43-30)^2}{30} + \frac{(17-30)^2}{30} + \frac{(24-25)^2}{25} + \frac{(26-25)^2}{25} = 21.68$$

a porovnáme ji s hodnotou kvantilu χ^2 pro $(5-1)\cdot(2-1) = 4$ stupně volnosti

$$\chi^2_{1-\alpha;\,4} = \chi^2_{0.95;\,4} \doteq 9.49 \ .$$

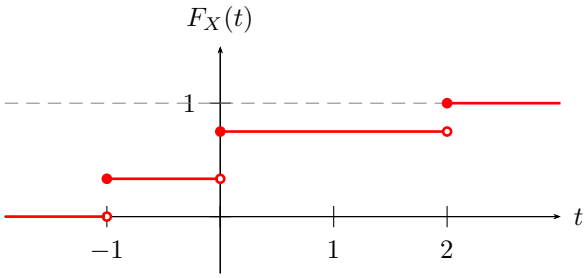
Protože

$$t \doteq 21.68 > 9.49 \doteq \chi^2_{0.95;\,4} \ ,$$

hypotézu o nezávislosti proto **ZAMÍTÁME**.

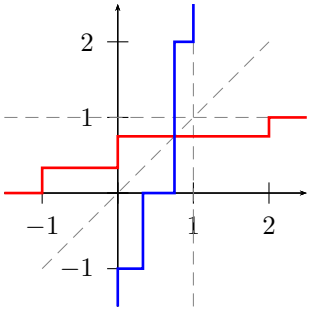
Poznámky ke kvantilům:

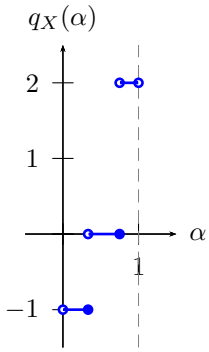
Pro náhodnou veličinu X a pravděpodobnost $\alpha \in (0,1)$ často potřebujeme najít $t \in \mathbb{R}$, že $P(X \leq t) = \alpha$, tj. $F_X(t) = \alpha$. Takové t obecně nemusí existovat (např. kdy F_X má skoky) nebo nemusí být určeno jednoznačně (když F_X je místy konstantní). Například si vezměme 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 13.3 *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 $= x_{\lceil \frac{3 \cdot 9}{4} \rceil} = x_7$

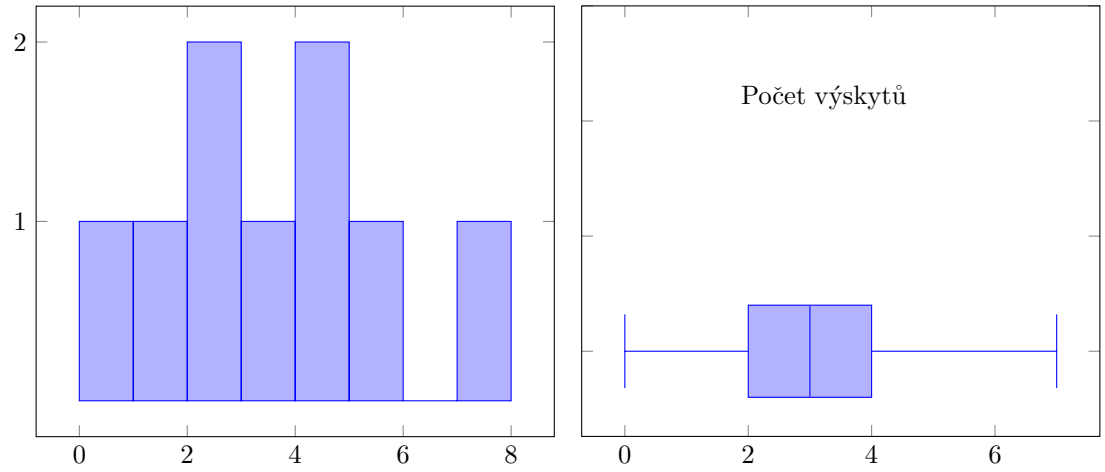
Medián je (v rámci uspořádání podle indexu) tedy přibližně uprostřed naměřených hodnot a podobně je to s okolními kvartily. Data si tudíž před výpočtem vždy uspořádáme.

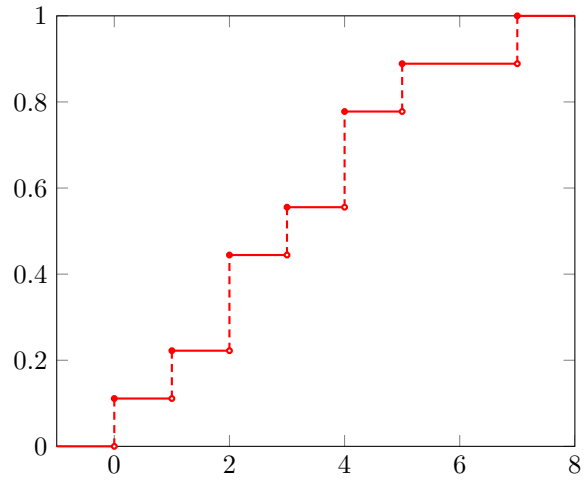
(1) Uspořádaná data:

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

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



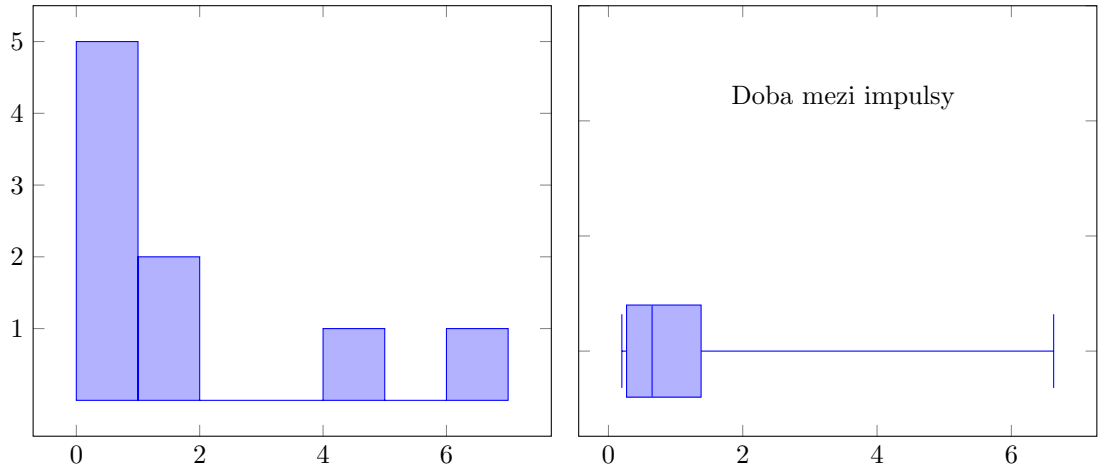


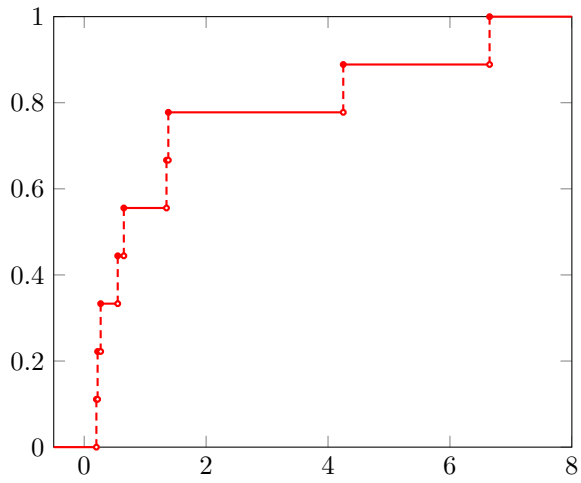
Vzhledem k popisu dat (počty výskytů na dané ploše) to vypadá na Poissonovo rozdělení. Tomu také zhruba odpovídají i grafická znázornění (histogram, boxplot, emp. distr. funkce).

(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).



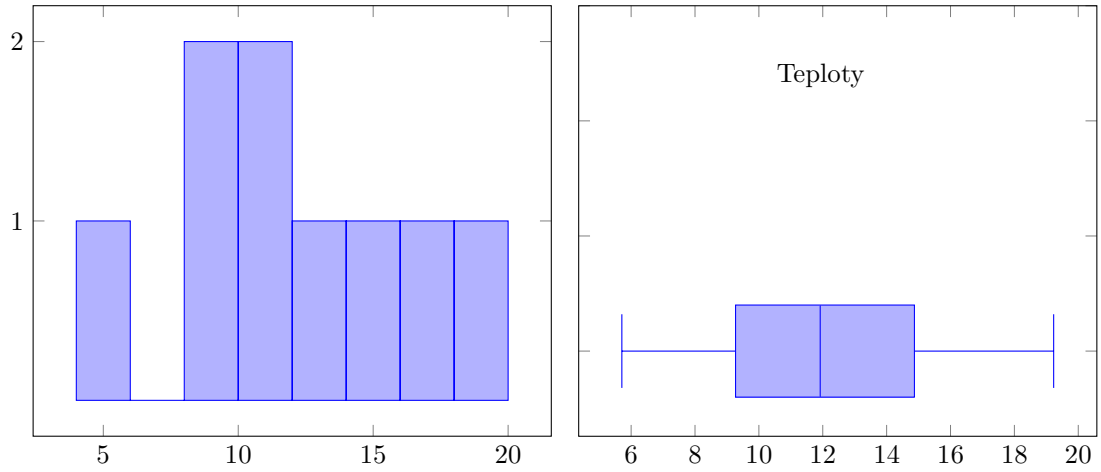


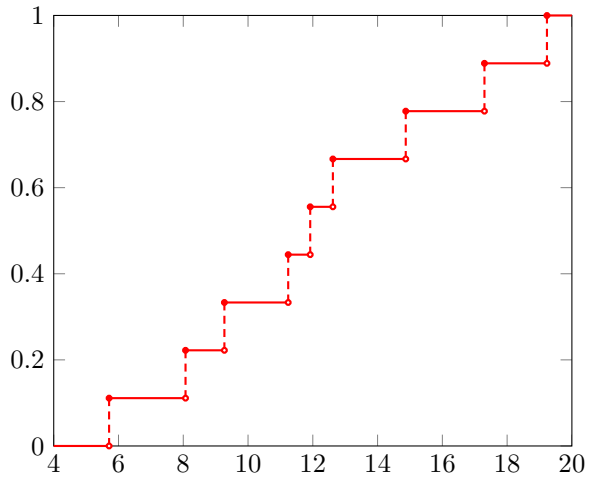
Vzhledem k popisu dat (doba čekání na další událost) to vypadá na exponenciální rozdělení. Tomu také zhruba odpovídají i grafická znázornění, kde boxplot je hodně posunutý doleva a empirická distribuční funkce připomíná exponenciálu.

(3) Uspořádaná data:

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).