

# 10. cvičení z PST

25. listopadu 2020

**Připomenutí:** Mějme náhodný výběr  $(X_1, \dots, X_n)$  závislý na parametru  $\vartheta$  (tj. máme vektor z nezávislých *stejně rozdělených* náhodných veličin  $X_i$  s distribuční funkcí  $F_{\vartheta}$  závislou na parametru  $\vartheta$ ). Můžeme uvažovat i závislost na více parametrech, ale většinou budeme pracovat jen s jedním.

V praxi máme hodnotu parametru danou (označme si ji  $\vartheta_0$ ), ale bohužel ji neznáme. Snažíme se ji proto určit (jako hodnotu  $\hat{\vartheta}$ ) z naměřených hodnot  $(x_1, \dots, x_n) \in \mathbb{R}^n$  a to co “nejlépe” (tím, že si stanovíme nějaké vhodné podmínky, které chceme splnit). Hodnotě  $\hat{\vartheta}$  pak říkáme *bodový odhad* (té skutečné hodnoty parametru  $\vartheta_0$ ).

Možných metod odhadu je více. Obvykle se používají

- metoda maximální věrohodnosti

- + *výhody*: dává (v podstatě) vždy výsledek; je možné ji použít i pro veličiny, co nemají číselné hodnoty (což znamená, že nezáleží na hodnotách, ale na jejich pravděpodobnostech)
- *nevýhody*: není vytvořena pro veličiny se smíšeným rozdělením (tj. jiným než buď diskrétním nebo spojitým)

- metoda momentů

- + *výhody*: dá se použít na jakýkoliv typ veličiny  $X$  (která má konečné hodnoty  $E(X^k)$  pro prvních několik  $k = 1, 2, 3, \dots$ )
- *nevýhody*: obecně nemáme zaručeno, že dostaneme nějaký výsledek

## 10.1 (metoda momentů a max. věrohodnosti - diskrétní rozdělení)

Odhadněte parametr  $w \in (0, 1)$  veličiny  $X$  s geometrickým rozdělením

$$p_X(i; w) = w^i(1 - w), \quad i \in \mathbb{N}_0$$

na základě realizace s následujícími četnostmi výsledků:

hodnota $i$	0	1	2	3
pozorovaná četnost $n_i$	20	10	7	3

Použijte metodu momentů i metodu maximální věrohodnosti.

### Řešení:

Součet pravděpodobností všech hodnot je 1 (tj.  $\sum_{i=0}^{\infty} p_X(i; w) = 1$ ), takže žádná speciální podmínka pro  $w$  odsud neplatí.

#### Metoda maximální věrohodnosti:

Hledáme hodnotu  $w$ , která maximalizuje funkci věrohodnosti

$$\begin{aligned} L(w) &= P(X_1 = x_1, \dots, X_n = x_n; w) = \prod_{j=1}^n \underbrace{P(X_j = x_j; w)}_{p_X(x_j; w)} = \prod_{i=0}^3 p_X(i; w)^{n_i} = \\ &= (1 - w)^{20} (w(1 - w))^{10} (w^2(1 - w))^7 (w^3(1 - w))^3 = w^{33}(1 - w)^{40} \end{aligned}$$

kde  $X_j$  jsou jednotlivé nezávislé veličiny (odpovídající jednotlivým pokusům) a  $x_j$  naměřené hodnoty. Funkce  $L$  je nezáporná a spojitá na uzavřené množině  $\langle 0, 1 \rangle$ , takže zde nabývá maxima. To nemůže být v krajních bodech (tam je funkce nulová) a proto je nabyto uvnitř dané množiny. To tedy odpovídá hledání maxima funkce

$$\ell(w) = \ln(L(w)) = 33 \cdot \ln(w) + 40 \cdot \ln(1-w)$$

na otevřeném intervalu  $(0, 1)$ . Protože maximum existuje, musí pro něj platit

$$0 = \ell'(q) = \frac{33}{w} - \frac{40}{1-w}$$

neboli

$$w = \frac{33}{73} \doteq 0.45205$$

což vyhovuje zadání.

### Metoda momentů:

Porovnáváme teoretické  $k$ -té momenty  $E(X^k)$  s jejich odhady  $m_k = \frac{1}{n} \sum_{i=1}^n x_i^k$  pro prvních několik  $k = 1, 2, \dots$

Střední hodnota je

$$\begin{aligned} E(X) &= \sum_{i=0}^{\infty} iw^i(1-w) = \sum_{i=1}^{\infty} iw^i - \sum_{i=1}^{\infty} iw^{i+1} = \sum_{i=1}^{\infty} iw^i - \sum_{i=2}^{\infty} (i-1)w^i = \\ &= w + \sum_{i=2}^{\infty} (i-i+1)w^i = \sum_{i=1}^{\infty} w^i = w \sum_{i=1}^{\infty} w^{i-1} = \frac{w}{1-w} \end{aligned}$$

a její odhad z realizace je

$$\bar{x} = \frac{1}{n} \sum_i i n_i = \frac{1}{20+10+7+3} \cdot (0 \cdot 20 + 1 \cdot 10 + 2 \cdot 7 + 3 \cdot 3) = \frac{33}{40}.$$

Porovnáním dostaneme

$$\frac{w}{1-w} = E(X) = \bar{x} = \frac{33}{40}$$

což dává opět řešení

$$w = \frac{33}{73} \doteq 0.45205$$

jako v předchozí metodě.

Jak je snadno vidět, v případě geometrického rozdělení dostáváme pro jeho parametr  $w$  vždy stejné výsledky pro obě metody.

## 10.2 (metoda momentů a max. věrohodnosti - diskretní rozdělení)

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 $p_X(i; c, q)$	$c - q$	$c$	$c + q$
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.  $0 \leq c - q = \frac{1}{3} - q$  a  $0 \leq c + q = \frac{1}{3} + q$ , takže  $|q| \leq \frac{1}{3}$ . Zbývá tedy odhadnout parametr  $q$ .

**Metoda maximální věrohodnosti:**

Hledáme hodnotu  $q$ , která maximalizuje funkci věrohodnosti

$$L(q) = P(X_1 = x_1, \dots, X_n = x_n; \frac{1}{3}, q) = \prod_{j=1}^n \underbrace{P(X_j = x_j; \frac{1}{3}, q)}_{p_X(x_j; \frac{1}{3}, q)} = \left(\frac{1}{3} - q\right)^8 \cdot \left(\frac{1}{3}\right)^{10} \cdot \left(\frac{1}{3} + q\right)^5$$

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

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

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

$$q = -\frac{1}{13} \doteq -0.07692.$$

Odhady pravděpodobností hodnot 1, 2, 3 jsou tedy

$$p_X(1) = \frac{16}{39} \doteq 0.4103 \quad p_X(2) = \frac{1}{3} \doteq 0.3333 \quad p_X(3) = \frac{10}{39} \doteq 0.2564$$

což vyhovuje zadání.

**Metoda momentů:**

Střední hodnota je

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

její odhad z realizace je

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

Porovnáním dostaneme

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

což odpovídá hodnotě

$$q = -\frac{3}{46} \doteq -0.06522.$$

Odhady pravděpodobností hodnot 1, 2, 3 jsou tedy

$$p_X(1) = \frac{55}{138} \doteq 0.3986 \quad p_X(2) = \frac{1}{3} \doteq 0.3333 \quad p_X(3) = \frac{37}{138} \doteq 0.2681$$

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.)

### 10.3 (metoda momentů a max. věrohodnosti - směs)

V urně je mnoho hracích kostek, z nichž některé jsou správné, některé falešné. Na falešných padá šestka s pravděpodobností  $1/2$ , zbývající čísla mají stejnou pravděpodobnost. Opakovaně jsme vytáhli kostku, hodili ji a vrátili ji zpět. Četnost výsledků udává tabulka:

hodnota $i$	1	2	3	4	5	6
četnost $n_i$	18	20	12	15	10	25

Odhadněte, kolik procent kostek je falešných.

#### Řešení:

Podíl falešných kostek označme  $c \in \langle 0, 1 \rangle$ . Naše náhodná veličina je

$X =$  "hodnota, která padne na dané kostce"

a můžeme ji vyjádřit jako směs  $X = \text{Mix}_c(X_1, X_2)$  složenou z náhodných veličin

$X_1 =$  "hodnota, která padne na falešné kostce"

$X_1 :$  "množina falešných kostek"  $\rightarrow \mathbb{R}$

$X_2 =$  "hodnota, která padne na správné kostce"

$X_2 :$  "množina správných kostek"  $\rightarrow \mathbb{R}$ .

**Metoda momentů:** Máme

$$E(X_1) = \frac{1}{10}(1 + \dots + 5) + \frac{1}{2} \cdot 6 = \frac{9}{2}$$

a

$$E(X_2) = \frac{1}{6}(1 + \dots + 6) = \frac{7}{2}$$

a z definice směsi tak dostaneme

$$E(X) = c \cdot E(X_1) + (1 - c) \cdot E(X_2) = c \cdot \frac{9}{2} + (1 - c) \cdot \frac{7}{2} = c + \frac{7}{2}.$$

Realizace výběrového průměru je

$$\bar{x} = \frac{\sum_i n_i \cdot i}{\sum_i n_i} = \frac{18 \cdot 1 + 20 \cdot 2 + 12 \cdot 3 + 15 \cdot 4 + 10 \cdot 5 + 25 \cdot 6}{18 + 20 + 12 + 15 + 10 + 25} = \frac{354}{100} = 3.54.$$

Srovnáním dostaneme

$$\hat{p} + 3.5 = E(X) = \bar{x} = 3.54,$$

což dává  $\hat{p} = 0.04 \in \langle 0, 1 \rangle$ , a to vyhovuje zadání.

**Metoda maximální věrohodnosti:**

Z definice směsi máme pro její pravděpodobnostní funkci, že

$$p_X(i) = c \cdot p_{X_1}(i) + (1 - c) \cdot p_{X_2}(i) = \begin{cases} c \frac{1}{10} + (1 - c) \frac{1}{6} = \frac{5-2c}{30} & , i = 1, \dots, 5 , \\ c \frac{1}{2} + (1 - c) \frac{1}{6} = \frac{1+2c}{6} & , i = 6 . \end{cases}$$

Ve směsi rozdělení šestka padla  $25 \times$ , ostatní čísla padla  $75 \times$  (není třeba mezi nimi rozlišovat, protože mají stejnou pravděpodobnost). Tedy věrohodnostní funkce je

$$L(c) = \left( \frac{5-2c}{30} \right)^{75} \cdot \left( \frac{1+2c}{6} \right)^{25} ,$$

$$\ell(c) = \ln(L(c)) = 75 \ln(5-2c) + 25 \ln(1+2c) + konst.$$

Maximum nastává pro  $\hat{c}$  takové, že

$$0 = \ell'(\hat{c}) = \frac{-150}{5-2\hat{c}} + \frac{50}{1+2\hat{c}} = 50 \cdot \frac{2-8\hat{c}}{(5-2\hat{c})(1+2\hat{c})} ,$$

$$\hat{c} = \frac{1}{4} \in \langle 0, 1 \rangle .$$

protože na intervalu  $\langle 0, \frac{1}{4} \rangle$  je  $\ell' > 0$  a na  $(\frac{1}{4}, 1)$  je  $\ell' < 0$ . Tato hodnota je i v souladu s počátečními omezujícími podmínkami.

**Poznámka:**

Kdybychom měli trochu jiné hodnoty, např.

hodnota $i$	...	5	6
četnost $n_i$	...	15	20

dostali bychom

$$\bar{x} = \frac{18 \cdot 1 + 20 \cdot 2 + 12 \cdot 3 + 15 \cdot 4 + 15 \cdot 5 + 20 \cdot 6}{18 + 20 + 12 + 15 + 15 + 20} = \frac{349}{100} = 3.49 .$$

Protože ale  $E(X) = c \cdot 4.5 + (1 - c) \cdot 3.5 \in \langle 3.5, 4.5 \rangle$ , tak rovnice  $c + 3.5 = E(X) = \bar{x} = 3.49$  nemá řešení pro  $c \in \langle 0, 1 \rangle$ .  
V tomto případě metoda momentů prostě nedává žádnou odpověď.

**10.4 (metoda momentů a max. věrohodnosti - směs)**

Dvě diskrétní náhodné veličiny  $X, Y$  mají pravděpodobnostní funkce dané tabulkou. Odhadněte koeficient  $c$  směsi  $Z = \text{Mix}_c(X, Y)$  z četností jejích realizací uvedených v tabulce.

hodnota	1	2	3	4
$p_X$	0.1	0.2	0.2	0.5
$p_Y$	0.5	0.2	0.2	0.1
četnost	30	20	15	35

**Řešení:**

Z definice směsi máme pro parametr  $c$  nutnou podmínku  $0 \leq c \leq 1$ .

**Metoda momentů:** Z definice směsi  $Z = \text{Mix}_c(X, Y)$  dostaneme

$$E(Z) = c \cdot E(X) + (1 - c) \cdot E(Y) .$$

Pro střední hodnoty  $X$  a  $Y$  máme

$$E(X) = 0.1 + 2 \cdot 0.2 + 3 \cdot 0.2 + 4 \cdot 0.5 = 3.1$$

$$E(Y) = 0.5 + 2 \cdot 0.2 + 3 \cdot 0.2 + 4 \cdot 0.1 = 1.9 .$$

Takže dostaneme

$$E(Z) = c \cdot E(X) + (1 - c) \cdot E(Y) = 1.9 + 1.2 \cdot c .$$

Hodnota realizace výběrového průměru je

$$\bar{z} = \frac{30 + 2 \cdot 20 + 3 \cdot 15 + 4 \cdot 35}{100} = \frac{51}{20} = 2.55 .$$

Jejich srovnáním dostáváme

$$1.9 + 1.2 \cdot c = E(Z) = \bar{z} = 2.55$$

takže výsledek je

$$c = \frac{13}{24} \doteq 0.5417 .$$

**Metoda maximální věrohodnosti:** Z definice  $Z = \text{Mix}_c(X, Y)$  pro pravděpodobnostní funkci dostaneme

$$p_Z = c \cdot p_X + (1 - c) \cdot p_Y$$

hodnota	1	2	3	4
$p_Z$	$0.5 - 0.4c$	0.2	0.2	$0.1 + 0.4c$

Pro funkci věrohodnosti pak máme

$$L(c) = (0.5 - 0.4 \cdot c)^{30} \cdot 0.2^{20+15} \cdot (0.1 + 0.4 \cdot c)^{35}$$

Funkce  $L$  je nezáporná a spojitá na uzavřené množině  $\langle 0, 1 \rangle$ , takže zde nabývá maxima. To odpovídá hledání maxima funkce

$$\ell(c) = \ln L(c) = 30 \cdot \ln(0.5 - 0.4c) + 35 \cdot \ln(0.1 + 0.4c) + \text{konst.}$$

na stejném intervalu  $\langle 0, 1 \rangle$ . Poznamenejme, že tento interval je uvnitř většího definičního oboru daného podmínkami  $0.5 - 0.4 \cdot c > 0$  a  $0.1 + 0.4 \cdot c > 0$ , tj. jde o otevřený interval  $(-\frac{1}{4}, \frac{5}{4})$ .

Derivace

$$0 = \ell'(\hat{c}) = -\frac{30 \cdot 0.4}{0.5 - 0.4\hat{c}} + \frac{35 \cdot 0.4}{0.1 + 0.4\hat{c}} = \frac{5.8 - 10.4\hat{c}}{(0.5 - 0.4\hat{c})(0.1 + 0.4\hat{c})}$$

je nulová ve stacionárním bodě

$$\hat{c} = \frac{29}{52} \doteq 0.5577 .$$

V intervalu  $(-\frac{1}{4}, \hat{c})$  je  $\ell'$  evidentně kladná (o znaménku rozhoduje jen výraz v čitateli, výraz ve jmenovateli je kladný) a v intervalu  $(\hat{c}, \frac{5}{4})$  je  $\ell'$  zase záporná. Takže v bodě  $\hat{c} = \frac{29}{52} \doteq 0.5577$  je skutečně věrohodnost maximální.

**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  $P(X \in (x_0 - \varepsilon, x_0 + \varepsilon))$  pro  $\varepsilon \rightarrow 0_+$ . Dá se ukázat, že pokud je hustota

$f_X$  spojitá v  $x_0$ , pak platí

$$\lim_{\varepsilon \rightarrow 0^+} \frac{P(X \in (x_0 - \varepsilon, x_0 + \varepsilon))}{2\varepsilon} = f_X(x_0).$$

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

Toto můžeme ještě zobecnit v případě, že obor hodnot veličiny  $X$  bude interval  $H$ , kde funkce  $f_X$  bude spojitá vzhledem k  $H$  (tj. např. v krajních bodech intervalu  $H$  bude jednostranně spojitá). Pak pro každé  $x_0 \in H$  podobně dostaneme, že

$$\lim_{\varepsilon \rightarrow 0^+} \frac{P(X \in (x_0 - \varepsilon, x_0 + \varepsilon) \cap H)}{\text{"délka intervalu } (x_0 - \varepsilon, x_0 + \varepsilon) \cap H"} = f_X(x_0).$$

Proto se ve věrohodnostní funkci nakonec opravdu hustota používá, ale pouze za předpokladu, že je spojitá v oboru hodnot dané veličiny. Např. pro exponenciální rozdělení (které modeluje dobu čekání) je obor hodnot  $(0, +\infty)$  a tam už hustotu spojitou máme.

### 10.5 (metoda momentů a max. věrohodnosti - spojité rozdělení)

Datový soubor  $\mathbf{x} = (-4, -3, -2, -1.5, 0.5, 1, 2.5, 3)$  je realizací náhodné veličiny  $X$ , která má spojitě rovnoměrné rozdělení v intervalu  $\langle -h, h \rangle$ . Metodou momentů a metodou maximální věrohodnosti určete odhad parametru  $h$  (a ověřte, zda odhad odpovídá zadání).

#### Řešení:

Rozsah souboru je  $n = 8$ . Realizované výsledky musí spadat do oboru hodnot, což je interval  $\langle -h, h \rangle$ . Tedy musí být  $|x_i| \leq h$  pro všechna  $i$ , neboli musí platit, že

$$h \geq \max\{|x_1|, \dots, |x_n|\} = 4.$$

#### Metoda maximální věrohodnosti:

V metodě max. věrohodnosti pro spojitě rozdělení nahrazujeme pravděpodobnostní funkci  $p_X$  (která by zde byla vždy nulová) hustotou  $f_X$ , u které požadujeme, aby byla spojitá na oboru hodnot veličiny  $X$  (taková hustota už je pak jen jedna). Naše zadání toto splňuje, protože

$$f_X(x; h) = \begin{cases} \frac{1}{2h} & , |x| \leq h, \\ 0 & , |x| > h. \end{cases}$$

Naším cílem je maximalizovat funkci

$$\Lambda(h) = \prod_{i=1}^n \underbrace{f_{X_i}(x_i; h)}_{f_X(x_i; h)} = \left(\frac{1}{2h}\right)^n.$$

pro  $h \in \langle 4, +\infty \rangle$ . Tato funkce je klesající v proměnné  $h$ , takže nabývá maxima pro největší přípustnou hodnotu parametru

$$\hat{h} = 4.$$

Hledaný interval  $\langle -h, h \rangle$  je tedy nejmenší takový, který obsahuje všechna  $x_i$ , pro  $i = 1, \dots, n$ .

#### Metoda momentů:

Opět porovnáváme teoretické  $k$ -té momenty  $E(X^k)$  s jejich odhady  $m_k = \frac{1}{n} \sum_{i=1}^n x_i^k$  pro prvních několik  $k = 1, 2, \dots$

Protože hustota  $f_X$  je sudá, bude  $E(X^k) = 0$  pro  $k$  liché. Speciálně, střední hodnota je  $E(X) = 0$  a tedy požadavek  $0 = E(X) = \bar{x}$  nám žádnou podmínku pro  $h$  nedává. Dokonce tuto rovnost ani není možno pro naše zadání splnit, protože:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = -\frac{3.5}{8} = -0.4375 \neq 0.$$

To nám ale nemusí vadit, protože jen těžko můžeme očekávat, že se aritmetickým průměrem při konečném počtu měření trefíme právě do hodnoty nula.

Proto musíme použít další momenty

$$E(X^2) = \int_{-h}^h x^2 \cdot \frac{1}{2h} dx = \left[ \frac{x^3}{6h} \right]_{-h}^h = \frac{h^2}{3}.$$

Odhad druhého momentu je

$$m_2 = \frac{1}{n} \sum_{i=1}^n x_i^2 = \frac{47.75}{8} = 5.96875.$$

Odhad parametru získáme jako řešení rovnice

$$\frac{\hat{h}^2}{3} = E(X^2) = m_2 = \frac{47.75}{8} \implies \hat{h} = \sqrt{17.90625} \doteq 4.2316.$$

Protože všechny hodnoty ze souboru leží v intervalu  $\langle -\hat{h}, \hat{h} \rangle = \langle -4.2316, 4.2316 \rangle$ , můžeme nalezenou hodnotu  $\hat{h}$  tudíž považovat za hledaný odhad parametru rozdělení.

### 10.6 (metoda momentů a max. věrohodnosti - spojité rozdělení)

Náhodná veličina  $X$  s oborem hodnot  $\langle a, +\infty \rangle$  má hustotu

$$f_X(t) = \begin{cases} 0 & , t \in (-\infty, a), \\ e^{a-t} & , t \in \langle a, \infty \rangle, \end{cases}$$

kde  $a \in \mathbb{R}$  je parametr. Pomocí metody maximální věrohodnosti i metody momentů odhadněte parametr  $a$ .

Úlohu vyřešte obecně pro realizaci  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  a také pro konkrétní realizaci

$$\mathbf{x} = (1, 2, 2, 2, 3, 3, 4)$$

rozsahu  $n = 7$ .

#### Řešení:

Realizované výsledky musí spadat do oboru hodnot, tj.

$$x_i \in \langle a, +\infty \rangle$$

pro všechna  $i = 1, \dots, n$  neboli musí platit, že

$$a \leq \min\{x_1, \dots, x_n\}.$$

Dále si všimněme, že funkce  $f_X$  je posunutá hustota exponenciálního rozdělení s parametrem  $\tau = 1$  (neboli veličina  $Y = X - a$  má exponenciální rozdělení  $\text{Exp}(1)$ , což se snadno odvodí). Tedy je to opět hustota. Ale to, že  $f_X$  je hustota můžeme ukázat i přímo: funkce  $f_X$  je nezáporná a platí, že

$$\int_{-\infty}^{\infty} f_X(t) dt = \int_a^{\infty} e^{a-t} dt = e^a [-e^{-t}]_{t=a}^{t=\infty} = e^a \cdot e^{-a} = 1.$$

#### Metoda maximální věrohodnosti:



Metoda max. věrohodnosti pro spojité rozdělení je podobná jako pro diskrétní rozdělení. Pravděpodobnostní funkci zde nahradíme hustotou, která ale (jak víme) není jednoznačně definována. Aby tedy metoda měla vůbec smysl, uvažuje se zde jen případ, kdy hustota  $f_X$  je spojitá na oboru hodnot veličiny  $X$  (taková hustota už je pak jen jedna). Naše zadání toto splňuje.

Naším cílem je maximalizovat funkci

$$\Lambda(a) = \prod_{i=1}^n \underbrace{f_{X_i}(x_i; a)}_{f_X(x_i)} = \prod_{i=1}^n e^{a-x_i} = e^{na} \cdot e^{-\sum_i x_i}.$$

pro  $a \in (-\infty, \min\{x_1, \dots, x_n\})$ . Tato funkce je rostoucí v proměnné  $a$ , takže nabývá maxima pro největší přípustnou hodnotu parametru

$$\hat{a} = \min\{x_1, \dots, x_n\}.$$

Pro konkrétní zadání je to pak

$$\hat{a} = \min\{1, 2, 2, 2, 3, 3, 4\} = 1.$$

### Metoda momentů:

Porovnáme teoretickou střední hodnotu

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} t \cdot f_X(t) dt = \int_a^{\infty} t e^{a-t} dt = [-t e^{a-t}]_{t=a}^{\infty} + \int_a^{\infty} e^{a-t} dt = \\ &= a + [-e^{a-t}]_{t=a}^{\infty} = a + 1 \end{aligned}$$

a výběrový průměr  $\bar{x}$ . Odtud tak pro parametr  $\hat{a}$  dostaneme

$$\hat{a} = \bar{x} - 1,$$

**POKUD** je ovšem splněno, že  $\hat{a} \leq \min\{x_1, \dots, x_n\}$ !

Pro konkrétní zadání je  $\bar{x} = \frac{1+2+2+2+3+3+4}{7} = \frac{17}{7}$  a tedy  $\hat{a} = \frac{17}{7} - 1 \doteq 1.43$ , což ale **NENÍ** menší než  $\min\{1, 2, 2, 2, 3, 3, 4\} = 1$ . V tomto případě tedy metoda momentů **NEDÁVÁ** žádný odhad.

### 10.7 (metoda momentů a max. věrohodnosti - spojité rozdělení)

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  $\lambda$  tohoto exponenciálního rozdělení.

#### Řešení:

Máme tedy veličinu

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

s exponenciálním rozdělením  $Exp(\tau)$ , kde  $\tau > 0$ , a hustotou  $f(x; \tau) = \begin{cases} \frac{1}{\tau} e^{-\frac{x}{\tau}} & \text{pro } x > 0 \\ 0 & \text{pro } x \leq 0. \end{cases}$

Počet měření je  $n = 5$  a jejich hodnoty jsou  $x_1 = 4$  dny,  $\dots$ ,  $x_5 = 24.5$  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  $\tau > 0$ , které maximalizuje věrohodnostní funkci

$$\begin{aligned} L(\tau) &= \prod_{i=1}^n f(x_i; \tau) = \prod_{i=1}^n \frac{1}{\tau} e^{-\frac{x_i}{\tau}} = \frac{1}{\tau} e^{-\frac{4}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{7}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{12}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{2.5}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{24.5}{\tau}} = \\ &= \frac{1}{\tau^5} e^{-\frac{1}{\tau} \cdot (4+7+12+2.5+24.5)} = \frac{1}{\tau^5} e^{-\frac{50}{\tau}}. \end{aligned}$$

Logaritmicko-věrohodnostní funkce je

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

Z její derivace

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

získáme řešení

$$-\frac{5}{\hat{\tau}} + 50 = 0 \quad \implies \quad \hat{\tau} = 10 \text{ [dnů]}$$

(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  $k = 1, 2, \dots$

Máme

- střední hodnotu  $E(X) = \tau$

- výběrový průměr  $\bar{x} = m_1 = \frac{\sum_{j=1}^n x_j}{n} = \frac{4+7+12+2.5+24.5}{5} = \frac{50}{5} = 10$

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

**Poznámka:** Na následujícím příkladu si ukážeme, v čem spočívá hledání intervalu spolehlivosti pro nějaký parametr. Odvodíme si oboustranný symetrický interval o spolehlivosti  $1 - \alpha$  pro střední hodnotu  $\mu$  normálního rozdělení  $N(\mu, \sigma^2)$  při *neznámém* rozptylu  $\sigma^2$  (při známém rozptylu by výsledek vypadal jinak a jednodušeji):

Mějme realizaci náhodného výběru  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$  o rozsahu  $n$  (pro nezávislé náhodné veličiny  $X_i \sim N(\mu, \sigma^2)$ ). Hledáme teď nějaké funkce  $h_1, h_2 : \mathbb{R}^n \rightarrow \mathbb{R}$  (nezávislé na volbě  $\mu$  i  $\sigma$ ) takové, že

$$P\left(h_1(X_1, \dots, X_n) \leq \mu \leq h_2(X_1, \dots, X_n)\right) = 1 - \alpha$$

(to je ta oboustrannost a  $1 - \alpha$  spolehlivost) a současně chceme, aby pro zbylé případy ještě platilo, že

$$P\left(\mu < h_1(X_1, \dots, X_n)\right) = \frac{\alpha}{2} = P\left(h_2(X_1, \dots, X_n) < \mu\right)$$

(to je ta symetričnost - tj. symetričnost nikoliv ve "vzdálenosti", ale v pravděpodobnosti).

Při dané realizaci  $\mathbf{x} = (x_1, \dots, x_n)$  pak jako hledaný **interval spolehlivosti  $1 - \alpha$  pro  $\mu$**  chápeme (číselný) interval tvaru:

$$\left\langle h_1(x_1, \dots, x_n), h_2(x_1, \dots, x_n) \right\rangle (\subseteq \mathbb{R})$$

Je ještě dobré poznamenat, že

- pro parametr  $\mu$  žádné rozdělení pravděpodobnosti nemáme!
- daný interval spolehlivosti pro  $\mu$  vzniká čistě na základě naměřených hodnot  $\mathbf{x} = (x_1, \dots, x_n)$  a také se společně s nimi MĚNÍ! Jeho smysl je ten, že skutečná hodnota  $\mu = E(X)$  (která se NEMĚNÍ!) bude obsažena v těchto (obecně proměnných intervalech) s pravděpodobností  $1 - \alpha$ .  
Ovšem problémem zůstává, že při neznalosti skutečné hodnoty  $\mu$  nejsme schopni zjistit, které konkrétní naměřené intervaly  $\mu$  obsahují a které naopak ne. Víme jen, že těch druhých je jen 5%. V tom je rozdíl oproti např. střelení do terče, kdy před pokusem víme, že se trefíme s pravděpodobností  $q$  a po uskutečněním pokusu umíme zjistit, který z výsledků nastal, takže si tuto pravděpodobnost můžeme i ověřit.
- Pro konkrétní pokus se pak na vyčíslený interval můžeme dívat i takto: Dejme tomu, že pro parametr  $\mu$  a  $1 - \alpha = 95\%$  nám vyjde (při daných měřeních) interval jako  $\langle 59.93, 62.07 \rangle$ . Pak 95% vyjadřuje poměr, se kterým se budeme ochotni vsadit, že skutečná hodnota  $\mu$  (kterou může mít třeba někdo někde přesně zjištěnou), bude obsažena v intervalu  $\langle 59.93, 62.07 \rangle$ .

A teď pro náš konkrétní případ: Vezmeme si vhodnou veličinu, jejíž rozdělení známe. V našem případě veličinu

$$T = \frac{\bar{X} - \mu}{S_X} \sqrt{n},$$

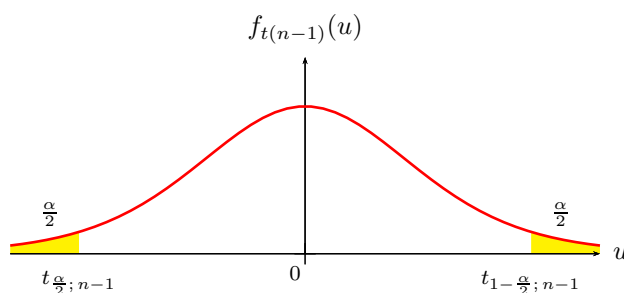
kde

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

a

$$(S_X)^2 = \frac{1}{n-1} \left( \sum_{i=1}^n (X_i - \bar{X})^2 \right) = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n \cdot (\bar{X})^2 \right).$$

Velichina  $T$  má tzv.  $t$ -rozdělení, tj. Studentovo rozdělení, s  $n - 1$  stupni volnosti. Hustota  $f_{t(n-1)}$  veličiny  $T$  je:



Pak pro kvantily platí, že

$$P\left( \underbrace{t_{\frac{\alpha}{2}; n-1}}_{-t_{1-\frac{\alpha}{2}; n-1}} \leq T \leq t_{1-\frac{\alpha}{2}; n-1} \right) = 1 - \alpha$$

a

$$P\left( T < -t_{1-\frac{\alpha}{2}; n-1} \right) = \frac{\alpha}{2} = P\left( t_{1-\frac{\alpha}{2}; n-1} < T \right).$$

Výrazy uvnitř pravděpodobnosti si teď jen přepíšeme a budeme mít hledané funkce  $h_1$  a  $h_2$ :

$$-t_{1-\frac{\alpha}{2}; n-1} \leq \frac{\bar{X} - \mu}{S_X} \sqrt{n} \leq t_{1-\frac{\alpha}{2}; n-1}$$

$$\underbrace{\bar{X} - \frac{S_X}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}; n-1}}_{h_1(X_1, \dots, X_n)} \leq \mu \leq \underbrace{\bar{X} + \frac{S_X}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}; n-1}}_{h_2(X_1, \dots, X_n)}$$

Po dosažení konkrétní realizace  $\mathbf{x}$  vektoru  $\mathbf{X}$  pak dostaneme výše uvedený interval spolehlivosti pro  $\mu$  ve tvaru

$$\langle \mu_L, \mu_U \rangle = \left\langle \bar{x} - \frac{s_x}{\sqrt{n}} t_{1-\frac{\alpha}{2}; n-1}, \bar{x} + \frac{s_x}{\sqrt{n}} t_{1-\frac{\alpha}{2}; n-1} \right\rangle$$

Pro názornost uveďme funkce  $h_1$  a  $h_2$  více rozepsané:

$$h_1(z_1, \dots, z_n) := \frac{1}{n} \left( \sum_{i=1}^n z_i \right) - \frac{t_{1-\frac{\alpha}{2}; n-1}}{\sqrt{n}} \cdot \frac{1}{n-1} \left( \sum_{i=1}^n z_i^2 - n \cdot \left( \frac{1}{n} \sum_{i=1}^n z_i \right)^2 \right)$$

$$h_2(z_1, \dots, z_n) := \frac{1}{n} \left( \sum_{i=1}^n z_i \right) + \frac{t_{1-\frac{\alpha}{2}; n-1}}{\sqrt{n}} \cdot \frac{1}{n-1} \left( \sum_{i=1}^n z_i^2 - n \cdot \left( \frac{1}{n} \sum_{i=1}^n z_i \right)^2 \right)$$

### 10.8 (intervalový odhad pro střední hodnotu a rozptyl)

Soubor dat

$$\mathbf{x} = (75, 85, 58, 72, 70, 75)$$

je náhodným výběrem z normálního rozdělení  $N(\mu, \sigma^2)$ . Stanovte

- horní a dolní 95% interval spolehlivosti pro střední hodnotu  $\mu$ ;
- oboustranný symetrický 90% interval spolehlivosti pro rozptyl  $\sigma^2$ .

#### Řešení:

- K určení intervalového odhadu použijeme statistiku

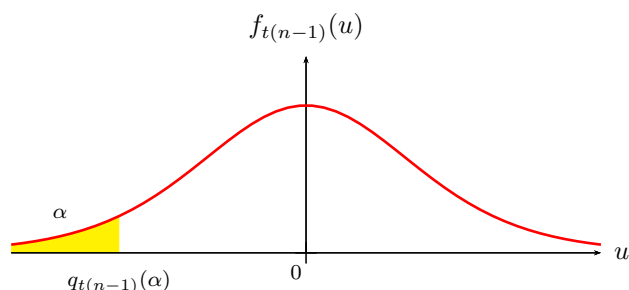
$$T = \frac{\bar{X} - \mu}{S_X} \sqrt{n}$$

která má Studentovo rozdělení  $t(n-1)$ , kde  $n = 6$  je rozsah souboru. Poznamenejme, že zatímco centrální limitní větu používáme pro velká  $n$ , protože obvykle pro  $X$  máme nějaké “obecné” rozdělení, tak v případě, kdy  $X$  má *právě* normální rozdělení, známe rozdělení veličiny  $T$  také přesně a to pro jakákoliv  $n$  (tj. i malá).

- Horní** interval spolehlivosti pro  $\mu$  (tj.  $\mu$  bude omezené *shora*) dostaneme ze vztahu

$$P\left( \underbrace{q_{t(n-1)}(\alpha)}_{-q_{t(n-1)}(1-\alpha)} \leq T \right) = 1 - \alpha$$

který vyjadřuje **dolní**  $1 - \alpha = 95\%$  intervalový odhad pro veličinu  $T$  (viz obrázek):



Pro realizaci  $t = \frac{\bar{x} - \mu}{s_x} \sqrt{n}$  tedy nastává případ

$$q_{t(n-1)}(1 - \alpha) \leq \frac{\bar{x} - \mu}{s_x} \sqrt{n}.$$

s pravděpodobností  $1 - \alpha = 95\%$ . Po úpravě máme horní interval spolehlivosti pro  $\mu$  ve tvaru:

$$\mu \leq \bar{x} + \frac{s_x}{\sqrt{n}} q_{t(5)}(0.95).$$

Pro jeho vyčíslení potřebujeme znát realizaci výběrového průměru  $\bar{x}$  a výběrového rozptylu  $s_x^2$

$$\bar{x} = \frac{1}{6} \sum_{i=1}^6 x_i = \frac{435}{6} = 72.5$$

$$s_x^2 = \frac{1}{5} \left( \sum_{i=1}^6 x_i^2 - 6 \cdot (\bar{x})^2 \right) = \frac{385.5}{5} \doteq 77.1, \quad s_x \doteq 8.781.$$

Z tabulek kvantilů Studentova rozdělení dostaneme  $q_{t(5)}(0.95) \doteq 2.02$ , a hledaný interval je tedy

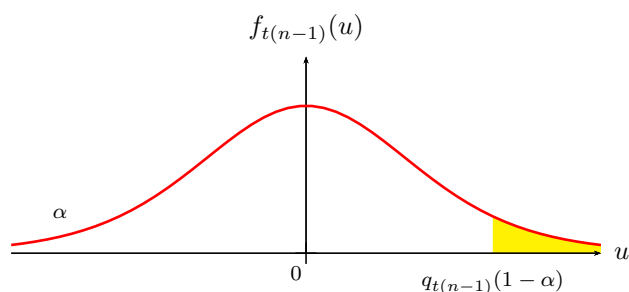
$$\mu \leq \underbrace{72.5 + \frac{8.781}{\sqrt{6}} 2.02}_{\doteq 79.74}.$$

neboli

$$\mu \in (-\infty, 79.74).$$

(a2) Podobně dostaneme **dolní** interval spolehlivosti pro  $\mu$  (tj.  $\mu$  bude omezené *zezdola*) ze vztahu pro **horní**  $1 - \alpha = 95\%$  intervalový odhad veličiny  $T$  (viz obrázek)

$$P(T \leq q_{t(n-1)}(1 - \alpha)) = 1 - \alpha$$



tedy

$$\frac{\bar{x} - \mu}{s_x} \sqrt{n} \leq q_{t(n-1)}(1 - \alpha).$$

a po úpravě

$$\bar{x} - \frac{s_x}{\sqrt{n}} q_{t(5)}(0.95) \leq \mu$$

a dosazení

$$\underbrace{72.5 - \frac{8.781}{\sqrt{6}} 2.02}_{=65.26} \leq \mu$$

máme

$$\mu \in (65.26, \infty).$$

(b) K určení intervalového odhadu použijeme statistiku

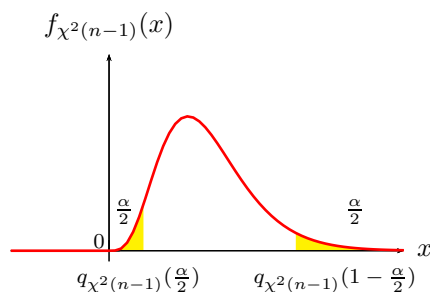
$$T = \frac{(n-1)S_{\mathbf{X}}^2}{\sigma^2}$$

která má tzv. chi-kvadrát rozdělení  $\chi^2(n-1)$  s  $n-1$  stupni volnosti, kde  $n=6$  je opět rozsah souboru (ten může být i malé číslo).

Oboustranný symetrický interval spolehlivosti  $1-\alpha=90\%$  pro  $\sigma^2$  dostaneme zase analogicky z oboustranného symetrického intervalového odhadu veličiny  $T$  (viz obrázek):

$$P\left(q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right) \leq T \leq q_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha$$

Graf zde představuje opět hustotu  $f_{\chi^2(n-1)}$  veličiny  $T$ .



Pro realizaci  $t = \frac{(n-1)s_{\mathbf{x}}^2}{\sigma^2}$  tedy nastává případ

$$q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right) \leq \frac{(n-1)s_{\mathbf{x}}^2}{\sigma^2} \leq q_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right)$$

s pravděpodobností  $1-\alpha=90\%$ . Po úpravě máme

$$\frac{(n-1)s_{\mathbf{x}}^2}{q_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-1)s_{\mathbf{x}}^2}{q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)}$$

Z tabulek kvantilů  $\chi^2$ -rozdělení pak dostaneme, že hledaný interval je pro rozptyl  $\sigma^2$  (o spolehlivosti  $0.9=1-\alpha$ ) je:

$$\begin{aligned} \sigma^2 &\in \left\langle \frac{(n-1)s_{\mathbf{x}}^2}{q_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right)}, \frac{(n-1)s_{\mathbf{x}}^2}{q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)} \right\rangle \doteq \\ &\doteq \left\langle \frac{(n-1)s_{\mathbf{x}}^2}{q_{\chi^2(5)}(0.95)}, \frac{(n-1)s_{\mathbf{x}}^2}{q_{\chi^2(5)}(0.05)} \right\rangle \doteq \\ &\doteq \left\langle \frac{5 \cdot 77.1}{11.07}, \frac{5 \cdot 77.1}{1.145} \right\rangle \doteq \\ &\doteq \langle 34.82, 336.68 \rangle. \end{aligned}$$

Všimněte si, že intervalový odhad výběrového rozptylu nemá střed ve svém bodovém odhadu  $s_{\mathbf{x}}^2 \doteq 77.1$ . Velké hodnoty rozptylu i intervalu jsou dány tím, že některé naměřené hodnoty (a sice 58 a 85) se dost odchyľují od výběrového průměru  $\bar{\mathbf{x}} = 72.5$ .