10. cvičení z PST

22. - 26. listopadu 2021

10.1 (normální rozdělení)
Oštěpařky Anna a Barbora mají střední hodnoty hodů po řadě 67 m a 75 m a směrodatné odchylky 6 m a 3 m. Předpokládejme nezávislá normální rozdělení. Odhadněte pravděpodobnost, že při jednom hodu hodí Anna dál než Barbora.

Řešení:
Náhodná veličina

$A =$"*délka hodu Anny*"

má rozdělení $N\big(67\text{ m}, (6\text{ m})^2\big)$ a veličina

$B =$"*délka hodu Barbory*"

má rozdělení $N\big(75\text{ m}, (3\text{ m})^2\big)$.

Zajímá nás $P(A > B) = P(A - B > 0)$. Protože veličiny A a B jsou nezávislé, tak veličina $Z := A - B$ má také normální rozdělení, a sice

$$Z \sim N(67 - 75, 6^2 + 3^2) = N(-8, 45)$$

(jednotky už pro přehlednost nepíšeme).
Takže

$$P(A > B) = P(Z > 0) = P\Bigg(\underbrace{\frac{Z - (-8)}{\sqrt{45}}}_{norm(Z)} > \frac{0 - (-8)}{\sqrt{45}}\Bigg) = 1 - P\left(norm(Z) \leq \frac{8}{\sqrt{45}}\right) =$$

$$= 1 - \Phi\left(\tfrac{8}{\sqrt{45}}\right) \doteq 1 - \Phi\left(1.1926\right) \doteq 1 - 0.883 = 0.117\,.$$

POZOR! Zatímco střední hodnota je lineární zobrazení, tak rozptyl se chová jinak! Konkrétně je to takto:
Nechť X a Y jsou veličiny se střední hodnotou a konečným rozptylem. Pak

- $E(X \pm Y) = E(X) \pm E(Y)$
- $D(X \pm Y) = D(X) + D(Y) \pm 2 \cdot \mathrm{cov}(X, Y)$ (≥ 0)

Speciálně, pokud X a Y jsou nezávislé, je $\mathrm{cov}(X, Y) = 0$. Máme tedy:

- X a Y nezávislé $\Rightarrow$ $D(X \pm Y) = D(X) + D(Y)$

Tedy v tomto případě se rozptyly VŽDY sčítají!

10.2 (alternativní rozdělení, odhad pravděpodobnosti - CLV a Čebyševova nerovnost)
Házíme $n = 1000$-krát symetrickou mincí a počítáme počet rubů. Nechť X je náhodnou veličinu X, která je počtem rubů během těchto pokusů.

(a) Určete pravděpodobnost, že počet rubů bude mezi 455 a 545, pomocí centrální limitní věty a odhadem pomocí Čebyševovy nerovnosti.

(b) Určete $k \in \mathbb{R}$ tak, abychom mohli říct, že s 90% pravděpodobností bude počet rubů menší nebo roven k.

Řešení:

Veličina X počítá počet úspěchů v n pokusech a má tedy binomické rozdělení $\mathrm{Bi}(n,p)$.

(a) Kvůli velkých hodnotám, se kterými pracujeme je ale výhodnější použít CLV.
Pro $i = 1, \ldots, n$ (kde $n = 1000$) si označme diskrétní veličiny s alternativním rozdělením

$$X_i = \begin{cases} 1 & , \text{ při } i\text{-tém hodu padne rub,} \\ 0 & , \text{ při } i\text{-tém hodu padne líc} \end{cases}$$

s alternativním rozdělením $\mathrm{Alt}(p)$, $p = \frac{1}{2}$ (tj. $P(X_i = 1) = p$). Veličiny X_i považujeme za nezávislé, se středními hodnotami $E(X_i) = p = \frac{1}{2}$ a rozptyly $D(X_i) = p(1-p) = \frac{1}{4}$. Máme

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

se střední hodnotou

$$E(X) = \sum_{i=1}^{n} E(X_i) = np = 1000 \cdot \frac{1}{2} = 500$$

a rozptylem

$$D(X) = \sum_{i=1}^{n} D(X_i) = np(1-p) = 1000 \cdot \frac{1}{4} = 250 \ .$$

Podle CLV má veličina $\mathrm{norm}(X) = \frac{X-500}{\sqrt{250}}$ přibližně normální rozdělení $N(0,1)$. Máme teď určit

$$P(455 < X < 545) = P\left(\frac{455-500}{\sqrt{250}} < \frac{X-500}{\sqrt{250}} < \frac{545-500}{\sqrt{250}}\right) = P\left(-\frac{9}{\sqrt{10}} < \mathrm{norm}(X) < \frac{9}{\sqrt{10}}\right) =$$

$$\doteq \Phi\Big(\frac{9}{\sqrt{10}}\Big) - \Phi\Big(-\frac{9}{\sqrt{10}}\Big) = 2 \cdot \Phi\Big(\frac{9}{\sqrt{10}}\Big) - 1 \doteq$$

$$\doteq 2 \cdot \Phi(2.846) - 1 \doteq 2 \cdot 0.99779 - 1 = 0.99558 \ .$$

Poznámka: Postup se dá ještě trochu zpřesnit protože veličina X má binomické rozdělení a její hodnoty jsou pouze přirozená čísla od 0 do n:

Nechť Y je veličina s normálním rozdělením aproximujícím X, tj. $Y \sim N\Big(E(X), D(X)\Big)$.

Pro $a = 0, 1, \ldots, n-1$ je F_X na intervalu $\langle a, a+1)$ konstantní (s hodnotou $F_X(a)$) a funkce F_Y pro odpovídající normální rozdělení, je zde ostře rostoucí. Z tohoto důvodu je lepší vzít jako aproximaci pro $F_X(a)$ hodnotu $F_Y(a + \frac{1}{2})$.

Tedy pro $a, b \in \{0, 1, \ldots, n-1\}$ se níže uvedená pravděpodobnost aproximuje jako:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) \doteq$$
$$\doteq F_Y(b + \tfrac{1}{2}) - F_Y(a + \tfrac{1}{2}) = P(a + \tfrac{1}{2} < Y \leq b + \tfrac{1}{2})$$

Abychom se tedy při výpočtu dostali k tomuto tvaru je dobré uvědomit si, že (díky diskrétnosti X) přesně platí

$$P(a < X \leq b) = P(a + \tfrac{1}{2} < X \leq b + \tfrac{1}{2})$$

a toto "posunutí" o $\frac{1}{2}$ se pak používá při přesnější aproximaci (přitom u nerovnosti na pravé straně rovnosti už je jedno, jestli jsou ostré nebo ne):

$$P(455 < X < 545) = P(455 < X \le 544) = P(455.5 < X \le 544.5) = P\Big(\frac{455.5-500}{\sqrt{250}} < \frac{X-500}{\sqrt{250}} \le \frac{544.5-500}{\sqrt{250}}\Big) =$$

$$= P\Big(-\frac{8.9}{\sqrt{10}} < \mathrm{norm}(X) \le \frac{8.9}{\sqrt{10}}\Big) \doteq$$

$$\doteq \Phi\Big(\frac{8.9}{\sqrt{10}}\Big) - \Phi\Big(-\frac{8.9}{\sqrt{10}}\Big) = 2\cdot\Phi\Big(\frac{8.9}{\sqrt{10}}\Big) - 1 \doteq$$

$$\doteq 2\cdot\Phi(2.814) - 1 \doteq 2\cdot 0.99755 - 1 = 0.9951\ .$$

Předchozí postup dal 99.558% a tento postup zase 99.51%, což je rozdíl 0.048%. Větší rozdíly bychom zaznamenali, kdybychom byli v intervalech blíže ke střední hodnotě.

Odhad pomocí Čebyševovy nerovnosti:

$$P(455 < X < 545) = P\Big(\big|X - \underbrace{500}_{E(X)}\big| < 45\Big) \ge 1 - \frac{D(X)}{45^2} = 1 - \frac{250}{45^2} = 1 - \frac{10}{81} \doteq 0.8765$$

Dostáváme tedy spodní odhad 87.65%.

(b) Zde použijeme opět CLV. Hledáme $k \in \mathbb{R}$, že

$$0.9 = P(X \le k) = P\Big(\frac{X-500}{\sqrt{250}} \le \frac{k-500}{\sqrt{250}}\Big) = P\Big(\mathrm{norm}(X) \le \frac{k-500}{\sqrt{250}}\Big) \doteq \Phi\Big(\frac{k-500}{\sqrt{250}}\Big)\ .$$

Řešíme tedy přibližnou rovnost

$$0.9 \doteq \Phi\Big(\frac{k-500}{\sqrt{250}}\Big)$$

$$\frac{k-500}{\sqrt{250}} \doteq \Phi^{-1}(0.9) \doteq 1.282$$

$$k = 500 + \sqrt{250}\cdot 1.282 \doteq 520.27\ .$$

10.3 (alternativní rozdělení, odhad počtu - CLV)

Při táborové hře je potřeba zasáhnout šiškou vzdálený cíl. Ze zkušenosti se ví, že pravděpodobnost zásahů je 0.2. Kolikrát nejméně musí družstvo dětí provést hod, aby zasáhly cíl alespoň 15-krát s pravděpodobností alespoň 80%?

Řešení:

Nechť $n \in \mathbb{N}$ je hledaný nejmenší počet hodů. Pro $i = 1, \dots, n$ si zavedeme veličiny

$$X_i = \begin{cases} 1 & , \text{ při } i\text{-tém zásahu jsme se trefili,} \\ \\ 0 & , \text{ při } i\text{-tém zásahu jsme se netrefili.} \end{cases}$$

Veličiny X_i považujeme za nezávislé, s alternativním rozdělením, kde $P(X_i = 1) = 0.2$.

Celkový počet úspěšných zásahu bude veličina

$$Z_n = \sum_{i=1}^{n} X_i\ ,$$

která má binomické rozdělení $\mathrm{Bi}\Big(n, 0.2\Big)$ se střední hodnotou $E(Z_n) = 0.2\cdot n$ a směrodatnou odchylkou

$\sqrt{D(Z_n)} = \sqrt{0.2 \cdot 0.8 \cdot n} = 0.4 \cdot \sqrt{n}$.

Zajímá nás teď největší n takové, že

$$P\Big(Z_n \geq 15\Big) \geq 0.8 \ .$$

Použijeme centrální limitní větu a pomocí úprav nerovností můžeme psát :

$$0.8 \leq P\Big(Z_n \geq 15\Big) = P\left(\frac{Z_n - E(Z_n)}{\sqrt{D(Z_n)}} \geq \frac{15 - 0.2 \cdot n}{0.4\sqrt{n}}\right) =$$

$$= P\left(\mathrm{norm}(Z_n) \geq \frac{15 - 0.2 \cdot n}{0.4\sqrt{n}}\right) = 1 - P\left(\mathrm{norm}(Z_n) < \frac{15 - 0.2 \cdot n}{0.4\sqrt{n}}\right) \doteq 1 - \Phi\left(\frac{15 - 0.2 \cdot n}{0.4\sqrt{n}}\right)$$

tedy dostáváme přibližnou nerovnost (kterou budeme řešit, jako přesnou nerovnost)

$$0.8 \leq 1 - \Phi\left(\frac{15 - 0.2 \cdot n}{0.4\sqrt{n}}\right)$$

$$\Phi\left(\frac{15 - 0.2 \cdot n}{0.4\sqrt{n}}\right) \leq 0.2$$

$$\frac{15 - 0.2 \cdot n}{0.4\sqrt{n}} \leq \Phi^{-1}(0.2) = -\Phi^{-1}(0.8) \doteq -0.842$$

a

$$0.2\left(\sqrt{n}\right)^2 - \underbrace{0.4 \cdot 0.842}_{0.3368} \cdot \sqrt{n} - 15 \geq 0 \ .$$

Dostali jsme kvadratickou nerovnost a hledáme nejmenší $n \in \mathbb{N}$, které ji splňuje. Z kořenů kvadratické rovnice si tedy vezmeme ten, co je více vpravo (druhý totiž bude záporný). Dostaneme tak

$$\sqrt{n} \geq \frac{0.3368 + \sqrt{0.3368^2 + 4 \cdot 0.2 \cdot 15}}{0.4} \doteq 9.54$$

a

$$n \geq (9.54)^2 \doteq 91.01 \ .$$

Je tedy potřeba udělat alespoň 92 hodů.

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

V praxi máme hodnotu parametru danou (označme si ji ϑ_0), ale bohužel ji neznáme. Snažíme se ji proto určit (jako hodnotu $\widehat{\vartheta}$) z naměřených hodnot $(x_1, \ldots, x_n) \in \mathbb{R}^n$ a to co "nejlépe" (tím, že si stanovíme nějaké vhodné podmínky, které chceme splnit). Hodnotě $\widehat{\vartheta}$ pak říkáme *bodový odhad* (té skutečné hodnoty parametru ϑ_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

10.4 (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 neplyne.

Metoda maximální věrohodnosti:

Hledáme hodnotu w, která maximalizuje funkci věrohodnosti

$$L(w) = P(X_1 = x_1, \dots, X_n = x_n;\ w) = \prod_{j=1}^{n} \underbrace{P(X_j = x_j;\ w)}_{p_X(x_j;\ w)} = \prod_{i=0}^{3} p_X(i;\ w)^{n_i} =$$

$$= \Big(1-w\Big)^{20} \Big(w(1-w)\Big)^{10} \Big(w^2(1-w)\Big)^{7} \Big(w^3(1-w)\Big)^{3} = w^{33}(1-w)^{40}$$

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

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

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

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

$$= w + \sum_{i=2}^{\infty}(i-i+1)w^i = \sum_{i=1}^{\infty} w^i = w\sum_{i=1}^{\infty} w^{i-1} = \frac{w}{1-w}$$

a její odhad z realizace je

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

Porovnáním dostaneme

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

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.

10.5 (metoda momentů a max. věrohodnosti - diskrétní 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. $0 \le c-q = \frac{1}{3} - q$ 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

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

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

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

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

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

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

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

což vyhovuje zadání.

Metoda momentů:

Střední hodnota je

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

její odhad z realizace je

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

Porovnáním dostaneme

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

což odpovídá hodnotě

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

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

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

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.)

10.6 (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 jí 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 $X = \mathrm{Mix}_c(X_1, X_2)$ složenou z náhodných veličin

$X_1 =$"*hodnota, která padne na falešné kostce*"
X_1 : "množina falešných kostek" $\to \mathbb{R}$

$X_2 =$"*hodnota, která padne na správné kostce*"
X_2 : "množina správných kostek" $\to \mathbb{R}$.

Metoda momentů: Máme

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

a

$$E(X_2) = \tfrac{1}{6}(1 + \cdots + 6) = \tfrac{7}{2}$$

a z definice směsi tak dostaneme

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

Realizace výběrového průměru je

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

Srovnáním dostaneme

$$\widehat{p} + 3.5 = E(X) = \overline{\boldsymbol{x}} = 3.54 \ ,$$

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

Metoda maximální věrohodnosti:
Z definice směsi máme pro její pravděpodobnostní funkci, že

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

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

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

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

$$0 = \ell'(\widehat{c}) = \frac{-150}{5 - 2\,\widehat{c}} + \frac{50}{1 + 2\,\widehat{c}} = 50 \cdot \frac{2 - 8\widehat{c}}{(5 - 2\,\widehat{c})(1 + 2\,\widehat{c})} \, ,$$
$$\widehat{c} = \tfrac{1}{4} \in \langle 0, 1 \rangle \, .$$

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

Poznámka:
Kdybychom měli trochu jiné hodnoty, např.

hodnota i	$\cdots$	5	6
četnost n_i	$\cdots$	15	20

dostali bychom

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

Protože ale $E(X) = c \cdot 4.5 + (1 - c) \cdot 3.5 \in \langle 3.5,\ 4.5 \rangle$, tak rovnice $c + 3.5 = E(X) = \overline{\boldsymbol{x}} = 3.49$ nemá řešení pro $c \in \langle 0, 1 \rangle$.

V tomto případě metoda momentů prostě nedává žádnou odpověď.

10.7 (metoda momentů a max. věrohodnosti - směs)

Dvě diskrétní náhodné veličiny X, Y mají pravděpodobnostní funkce dané tabulkou. Odhadněte koeficient c směsi $Z = \mathrm{Mix}_c(X, Y)$ 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 $Z = \mathrm{Mix}_c(X, Y)$ dostaneme

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

Pro střední hodnoty X a Y máme

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

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

Takže dostaneme

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

Hodnota realizace výběrového průměru je

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

Jejich srovnáním dostáváme

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

takže výsledek je

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

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

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

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

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

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

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

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

je nulová ve stacionárním bodě

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

V intervalu $\big(-\frac{1}{4}, \widehat{c}\big)$ je ℓ' evidentně kladná (o znaménku rozhoduje jen výraz v čitateli, výraz ve jmenovateli je kladný) a v intervalu $\big(\widehat{c}, \frac{5}{4}\big)$ je ℓ' zase záporná. Takže v bodě $\widehat{c} = \frac{29}{52} \doteq 0.5577$ 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 $P\Big(X \in (x_0 - \varepsilon, x_0 + \varepsilon)\Big)$ pro $\varepsilon \to 0_+$. Dá se ukázat, že pokud je hustota f_X spojitá v x_0, pak platí

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

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

$$\lim_{\varepsilon \to 0_+} \frac{P\Big(X \in (x_0 - \varepsilon, x_0 + \varepsilon) \cap H\Big)}{\text{“délka intervalu } (x_0 - \varepsilon, x_0 + \varepsilon) \cap H\text{“}} = f_X(x_0)\ .$$

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. V některých případech (viz Příklad **10.9**) se ovšem u exponenciálního rozdělení uvažuje i obor hodnot $\langle 0, +\infty)$, kde jsme přidali i hodnotu 0. Je to do jisté míry umělý předpoklad a jeho jediným důvodem je to, aby daný postup dal nějaký výsledek.

10.8 (metoda momentů a max. věrohodnosti - spojité rozdělení)

Datový soubor $\boldsymbol{x} = (-4, -3, -2, -1.5, 0.5, 1, 2.5, 3)$ je realizací náhodné veličiny X, která má spojité 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 $|x_i| \le h$ pro všechna i, neboli musí platit, že

$$h \ge \max\{|x_1|, \ldots, |x_n|\} = 4\ .$$

Metoda maximální věrohodnosti:

V metodě max. věrohodnosti pro spojité 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

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

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

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

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

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 $k = 1, 2, \ldots$.

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

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

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

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

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

$$\frac{\widehat{h}^2}{3} = E(X^2) = m_2 = \frac{47.75}{8} \quad \Longrightarrow \quad \widehat{h} = \sqrt{17.90625} \doteq 4.2316\,.$$

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

10.9 (metoda momentů a max. věrohodnosti - spojité rozdělení)
Náhodná veličina X s oborem hodnot $\langle a, +\infty)$ má hustotu

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

kde $a \in \mathbb{R}$ je parametr. Pomocí metody maximální věrohodnosti i metody momentů odhadněte parametr a.
Úlohu vyřešte obecně pro realizaci $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ a také pro konkrétní realizaci

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

rozsahu $n = 7$.

Řešení:
Realizované výsledky musí spadat do oboru hodnot, tj.

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

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

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

Dále si všimněme, že funkce f_X je posunutá hustota exponenciálního rozdělení s parametrem $\tau = 1$ (neboli veličina $Y = X - a$ má exponenciální rozdělení 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

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

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

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

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

$$\widehat{a} = \min\{x_1, \ldots x_n\} \ .$$

Pro konkrétní zadání je to pak

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

Metoda momentů:

Porovnáme teoretickou střední hodnotu

$$E(X) = \int_{-\infty}^{\infty} t \cdot f_X(t)\,\mathrm{d}t = \int_a^{\infty} t\, e^{a-t}\,\mathrm{d}t = [-t\, e^{a-t}]_{t=a}^{\infty} + \int_a^{\infty} e^{a-t}\,\mathrm{d}t =$$

$$= a + [-e^{a-t}]_{t=a}^{\infty} = a + 1$$

a výběrový průměr $\overline{\boldsymbol{x}}$. Odtud tak pro parametr $\widehat{a}$ dostaneme

$$\widehat{a} = \overline{\boldsymbol{x}} - 1 \ ,$$

POKUD je ovšem splněno, že $\widehat{a} \leq \min\{x_1, \ldots, x_n\}$!

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

10.10 (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 λ tohoto exponenciálního rozdělení.

Řešení:

Máme tedy veličinu

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

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

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

$$\Lambda(\tau) = \prod_{i=1}^{n} f(x_i;\ \tau) = \prod_{i=1}^{n} \tfrac{1}{\tau}\,\mathrm{e}^{-\frac{x_i}{\tau}} = \tfrac{1}{\tau}\,\mathrm{e}^{-\frac{4}{\tau}} \cdot \tfrac{1}{\tau}\,\mathrm{e}^{-\frac{7}{\tau}} \cdot \tfrac{1}{\tau}\,\mathrm{e}^{-\frac{12}{\tau}} \cdot \tfrac{1}{\tau}\,\mathrm{e}^{-\frac{2.5}{\tau}} \cdot \tfrac{1}{\tau}\,\mathrm{e}^{-\frac{24.5}{\tau}} =$$

$$= \tfrac{1}{\tau^5}\,\mathrm{e}^{-\frac{1}{\tau}\cdot(4+7+12+2.5+24.5)} = \tfrac{1}{\tau^5}\,\mathrm{e}^{-\frac{50}{\tau}} \ .$$

Logaritmicko-věrohodnostní funkce je

$$\lambda(\tau) = \ln \Lambda(\tau) = -5 \ln \tau - \frac{50}{\tau}.$$

Z její derivace

$$\lambda'(\tau) = -\frac{5}{\tau} + \frac{50}{\tau^2} = \frac{-5\tau + 50}{\tau^2}.$$

získáme řešení

$$-5\hat{\tau} + 50 = 0 \qquad \Longrightarrow \qquad \widehat{\tau} = 10 \text{ [dnů]}$$

(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 $k = 1, 2, \ldots$.

Máme

- střední hodnotu $E(X) = \tau$
- 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 $\hat{\tau} = E(X) = \overline{\boldsymbol{x}} = 10$ dostáváme opět odhad $\widehat{\tau} = 10$. Tato shoda je opět způsobena tím, že parametr τ 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 $\frac{1-\hat{p}}{\hat{p}} = E(X) = \overline{x} \;\Rightarrow\; \hat{p} = \frac{1}{\overline{x}+1}$
- λ pro Poissonovo $Poiss(\lambda)$, odhad je $\hat{\lambda} = E(X) = \overline{x}$
- τ pro exponenciální $Exp(\tau)$, 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.