9. cvičení z PRA

15. - 19. dubna 2024

Definice: Náhodný vektor (X, Y) má spojité rozdělení se *sdruženou hustotou pravděpodobnosti* $f_{X,Y} : \mathbb{R}^2 \to \langle 0, +\infty)$ $\Leftrightarrow$ $f_{X,Y}$ je integrabilní funkce a pro každou "rozumnou" množinu $A \subseteq \mathbb{R}^2$ (tj. takovou, která se dá získat z intervalu v $\mathbb{R}^2$ pomocí sjednocování, průniku a doplňku) platí, že

$$P\big((X,Y) \in A\big) = \iint\limits_A f_{X,Y}(x,y)\ \mathrm{d}x\mathrm{d}y\ .$$

To nastává právě když

$$F_{X,Y}(a,b) = \int\limits_{-\infty}^{a}\int\limits_{-\infty}^{b} f_{X,Y}(x,y)\ \mathrm{d}x\mathrm{d}y$$

pro každé $a, b \in \mathbb{R}$.

Sdružená hustota $f_{X,Y}$ opět (jako u veličin) NENÍ zdaleka určena jednoznačně, co se týče její funkční hodnoty, ale pouze hodnotami integrálů z této funkce (např. její změnou v konečně mnoha bodech nebo na nějaké hladké křivce se nezmění příslušné integrály, takže i změněná funkce bude také hustotou). Přesněji, dvě nezáporné funkce $f_{X,Y}$ a $g_{X,Y}$ (s integrálem rovným jedné) jsou hustotami pro tutéž sdruženou distribuční funkci $F_{X,Y}$ právě když se rovnají *skoro všude* a zapisuje se to jako

$$f_{X,Y} = g_{X,Y} \quad \text{(s.v.)}\ .$$

(tj. mohou se lišit jen na takové množině $A \subseteq \mathbb{R}^2$, že $\iint\limits_A 1\ \mathrm{d}x\mathrm{d}y = 0$, tj. pokud A má nulový plošný obsah).

Příklad 9.1 *Sdružená hustota náhodných veličin X a Y je*

$$f_{(X,Y)}(x,y) = \begin{cases} \frac{1}{2}e^{-x-\frac{y}{2}}\,, & x > 0, y > 0\,, \\ 0\,, & \textit{jinak}\,. \end{cases}$$

(a) Jaká jsou jejich marginální rozdělení?

(b) Jsou veličiny X a Y nezávislé? Zdůvodněte.

(c) Jak vypadá jejich korelační matice?

Řešení:

(a) Marginální hustoty (tj. hustoty jednotlivých veličin X a Y) jsou

$$f_X(x) = \int\limits_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy = \begin{cases} \int\limits_0^{\infty} \frac{1}{2}e^{-x-\frac{y}{2}}dy = e^{-x}\cdot\big[-e^{-\frac{y}{2}}\big]_0^{\infty} = e^{-x} & \text{pro } x > 0, \\ \int\limits_{-\infty}^{\infty} 0\ dy = 0 & \text{pro } x \leq 0\ . \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx = \begin{cases} \int_0^{\infty} \frac{1}{2}e^{-x-\frac{y}{2}}dx = \frac{1}{2}e^{-\frac{y}{2}} \cdot \big[-e^{-x}\big]_0^{\infty} = \frac{1}{2}e^{-\frac{y}{2}} & \text{pro } y > 0, \\ \int_{-\infty}^{\infty} 0\ dx = 0 & \text{pro } y \le 0\ . \end{cases}$$

Vidíme tedy, že obě rozdělení jsou exponenciální, konkrétně $X \sim \mathrm{Exp}(1)$ a $Y \sim \mathrm{Exp}(\frac{1}{2})$.

(b) Složky X a Y jsou nezávislé právě tehdy, když

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \quad \text{pro skoro všechna} \quad (x,y) \in \mathbb{R}^2\ ,$$

což znamená, že množina bodů, kde uvedená rovnost neplatí má nulový plošný obsah.

(Podmínce "skoro všude" se nelze vyhnout z toho důvodu, že hustoty nejsou jednoznačně definovány svými hodnotami, ale svými integrály.)

Jak je hned vidět, v našem případě je rovnost splněna dokonce všude, takže X a Y JSOU nezávislé.

(c) Z nezávislosti X, Y plyne okamžitě $\mathrm{cov}(X,Y) = 0$, tedy také $\mathrm{corr}(X,Y) = 0$ a korelační matice je tak

$$\mathrm{Corr}(X,Y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

A konečně, protože pro $Z \sim \mathrm{Exp}(\lambda)$ je $E(Z) = \frac{1}{\lambda}$ a $\mathrm{var}(Z) = \frac{1}{\lambda^2}$, tak díky bodu (1) je $\mathrm{var}(X) = 1$ a $\mathrm{var}(Y) = 4$. Tedy dostáváme varianční matici:

$$\mathrm{Var}(X,Y) = \begin{pmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\ .$$

Připomeňme si, co říká **Centrální limitní věta (CLV):**

Nechť X_i, pro $i = 1, 2, \ldots$ je posloupnost nezávislých náhodných veličin, které mají stejná rozdělení se střední hodnotou μ a (konečným) rozptylem σ^2. Pak pro veličiny

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

platí, že

$$\lim_{n\to\infty} P\Big(\mathrm{norm}(Z_n) \le t\Big) = \Phi(t) \quad \text{pro každé} \quad t \in \mathbb{R}\ .$$

Neboli: pro velká n má veličina $\mathrm{norm}(Z_n)$ přibližně normální rozdělení $N(0,1)$.

Centrální limitní větu můžeme formulovat (namísto pro Z_n) také pro tzv. výběrový průměr, tj. veličiny

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} \cdot Z_n\ .$$

protože pro ně platí $\mathrm{norm}(\overline{X}_n) = \mathrm{norm}(Z_n)$.

Poznámka: V rámci Centrální limitní věty (níže) se vyskytuje posloupnost nezávislých náhodných veličin, která nejčastěji vzniká následujícím způsobem:

Mějme náhodnou veličinu $X : \Omega \to \mathbb{R}$ na pravděpodobnostním prostoru Ω (např. pro házení mincí je $\Omega = \{rub,\ líc\}$ a veličina třeba $X(líc) = 1$ a $X(rub) = 0$ s rozdělením $\mathrm{Alt}(p)$). Jestliže nyní budeme opakovat (nekonečně) nezávislých pokusů, pak jejich výsledky tvoří posloupnost $\widetilde{\omega} = (\omega_1, \omega_2, \dots)$, kde $\omega_i \in \Omega$ pro $i \in \mathbb{N}$. Množina všech takovýchto možných posloupností je tedy $\Omega^{\mathbb{N}}$ (tj. spočetná kartézská mocnina množiny Ω).

Na této množině $\Omega^{\mathbb{N}}$ lze opět vybudovat pravděpodobnostní prostor tj. σ-algebru $\widetilde{\mathcal{A}}$ na $\Omega^{\mathbb{N}}$ (která se bude skládat ze spočetných sjednocení množin typu $\bigtimes_{i=1}^{\infty} A_i = A_1 \times A_2 \times \cdots$, kde $A_i \subseteq \Omega$ je jev pro každé i) a pravděpodobnost bude dána jako $\widetilde{P}\left(\bigtimes_{i=1}^{\infty} A_i\right) = \prod_{i=1}^{\infty} P(A_i)$.

Výsledek při i-tém pokusu nyní bude veličina $X_i : \Omega^{\mathbb{N}} \to \mathbb{R}$, definovaná prostě jako $X_i(\widetilde{\omega}) = \omega_i$ pro $\widetilde{\omega} = (\omega_1, \omega_2, \dots)$. Takovéto veličiny pak budou nezávislé a budou mít rozdělení stejné jako veličina X.

Rychlost konvergence v CLV: Pokud pro veličiny X_i v CLV navíc ještě je $\varrho := E(|X_i - \mu|^3) < \infty$, pak platí Berry–Esseenův odhad chyby (pro všechna $t \in \mathbb{R}$ a $n \in \mathbb{N}$):

$$\left|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\right| < 0.4748 \cdot \frac{\varrho}{\sigma^3\sqrt{n}}$$

Odhad chyby v CLV pro Poissonovo rozdělení: Pro veličinu $Z \sim \mathrm{Poiss}(\lambda)$ platí

$$\left|F_{\mathrm{norm}(Z)}(t) - \Phi(t)\right| \ \leq \ \frac{0.4748}{\sqrt{\lambda}} \text{ pro všechna } t \in \mathbb{R}.$$

V praxi se obvykle CLV používá už pokud $\lambda \geq 10$ jako dobrá aproximace (v tomto případě je odhad chyby $\leq \frac{0.4748}{\sqrt{10}} = 0.1502$, ale ve skutečnosti je tento odhad příliš nadsazený a skutečná chyba je menší.)

Důkaz: Veličinu Z můžeme rozepsat jako $Z = \sum_{i=1}^{n} X_i$, kde $X_i \sim \mathrm{Poiss}\left(\frac{\lambda}{n}\right)$ jsou nezávislé veličiny. Položme $\omega = \frac{\lambda}{n}$. Pro použití odhadu Berry-Esseen máme $\sigma = \sqrt{\mathrm{var}(X_i)} = \sqrt{\omega} = \sqrt{\frac{\lambda}{n}}$ a potřebujeme ještě odhadnout tuto hodnotu:

$$\varrho := E\left(|X_i - E(X_i)|^3\right) = E\left(|X_i - \omega|^3\right) = \sum_{k=0}^{\infty} |k-\omega|^3 \frac{\omega^k}{k!} e^{-\omega} =$$

$$= \sum_{k=0}^{[\omega]} (\omega - k)^3 \frac{\omega^k}{k!} e^{-\omega} + \sum_{k=[\omega+1]}^{\infty} (k-\omega)^3 \frac{\omega^k}{k!} e^{-\omega} = 2\sum_{k=0}^{[\omega]} (\omega - k)^3 \frac{\omega^k}{k!} e^{-\omega} + \underbrace{\sum_{k=0}^{\infty} (k-\omega)^3 \frac{\omega^k}{k!} e^{-\omega}}_{=E\left(\left(X_i - E(X_i)\right)^3\right) = \omega} =$$

$$= \omega + 2\sum_{k=0}^{[\omega]} \underbrace{(\omega - k)^3}_{\leq \omega^3} \frac{\omega^k}{k!} e^{-\omega} \leq \omega + 2\omega^3 \underbrace{\sum_{k=0}^{[\omega]} \frac{\omega^k}{k!} e^{-\omega}}_{\leq 1} \leq \omega + 2\omega^3 = \frac{\lambda}{n} + 2\left(\frac{\lambda}{n}\right)^3$$

V Berry-Esseen odhadu tedy pro všechna $n \in \mathbb{N}$ a všechna $t \in \mathbb{R}$ máme

$$\left|F_{\mathrm{norm}(Z)}(t) - \Phi(t)\right| \ < \ 0.4748 \cdot \frac{\varrho}{\sigma^3\sqrt{n}} = 0.4748 \cdot \frac{\frac{\lambda}{n} + 2\left(\frac{\lambda}{n}\right)^3}{\left(\sqrt{\frac{\lambda}{n}}\right)^3 \sqrt{n}} = 0.4748 \cdot \left(\frac{1}{\sqrt{\lambda}} + 2\left(\frac{\lambda^{3/2}}{n^2}\right)\right)$$

a protože λ zůstává pevné, zatímco s n můžeme jít libovolně vysoko, dostaneme v limitě odhad $\left|F_{\mathrm{norm}(Z)}(t) - \Phi(t)\right| \ \leq \ \frac{0.4748}{\sqrt{\lambda}}$.

Příklad 9.2 *V lese se narodí průměrně 4 zajíci denně. Předpokládejme, že počet narozených zajíců se řídí Poissonovým rozdělením. Jaká je pravděpodobnost, že v následujících 7 týdnech se v lese narodí alespoň 175 zajíců?*

Řešení:
Pro veličinu

$$Z = \text{“počet narozených zajíců za } 49 \text{ dnů”}$$

nás zajímá $P(Z \geq 175)$. U této veličiny sice snadno zjistíme její rozdělení (bude to $Z \sim \text{Poiss}(4 \cdot 49)$), ale k přesnějšímu vyčíslení by bylo při tomto přístupu potřeba sečíst kolem 175 velmi malých čísel, což by bylo jednak náročné a také by vznikalo hodně chyb.

K řešení proto použijeme centrální limitní větu a tudíž budeme chtít veličinu Z “rozsekat” na více stejně rozdělených nezávislých veličin. Označme si tedy pro $i = 1, 2, ..., n$, kde $n = 7 \cdot 7 = 49$, veličiny

$$X_i = \text{“počet narozených zajíců v } i\text{-tý den”} .$$

Veličiny pokládáme za nezávislé s rozdělením $X_i \sim \text{Poiss}(4)$, tedy $E(X_i) = 4 = \text{var}(X_i)$. Protože platí $Z = \sum_{i=1}^{n} X_i$, dostaneme

$$\begin{aligned} E(Z) &= n \cdot E(X_1) = 49 \cdot 4 = 196 \\ \text{var}(Z) &= n \cdot \text{var}(X_1) = 49 \cdot 4 = 196 \quad \Rightarrow \quad \sqrt{\text{var}(Z)} = \sqrt{196} = 14 \end{aligned}$$

což v případě rozptylu platí díky nezávislosti veličin.

Podle CLV (a kritéria použitelnosti CLV pro Poissonovo rozdělení, tj. $196 = E(Z) \geq 10$) bude mít veličina $norm(Z) = \frac{Z - E(Z)}{\sqrt{var(Z)}} = \frac{Z-196}{14}$ přibližně rozdělení $N(0,1)$. Můžeme proto psát

$$\begin{aligned} P(Z \geq 175) &= P\left(\frac{Z - 196}{14} \geq \frac{175 - 196}{14}\right) = P\Big(norm(Z) \geq -1.5\Big) = \\ &= 1 - P\Big(norm(Z) < -1.5\Big) \overset{(CLV)}{\doteq} 1 - \Phi(-1.5) = 1 - \Big(1 - \Phi(1.5)\Big) = \\ &= \Phi(1.5) \doteq \mathbf{0.9332} . \end{aligned}$$

(Pro srovnání: skutečná hodnota pro Poissonovo rozdělení je **0.9398**. Ovšem uvědomme si, že ”rozsekání” veličiny na součet dalších je v tomto případě spíš kvůli procvičení než jako argument pro použití CLV. Základním důvodem použití CLV je dostatečně velká hodnota $\lambda = E(Z)$, viz výše)

Odhad chyby v CLV pro rovnoměrné rozdělení: Pokud mají veličiny X_i rovnoměrné rozdělení na intervalu $\langle a, b \rangle$, pak

$$\mu = E(X_i) = \tfrac{a+b}{2}, \qquad \sigma = \sqrt{D(X_i)} = \sqrt{\tfrac{(b-a)^2}{12}} = \tfrac{b-a}{2\sqrt{3}}, \qquad \varrho = E\left(|X_i - \mu|^3\right) = \tfrac{(b-a)^3}{32}$$

čímž pro $Z_n = \sum_{i=1}^{n} X_i$ dostáváme odhad

$$\Big|F_{\text{norm}(Z_n)}(t) - \Phi(t)\Big| \quad < \quad 0.4748 \cdot \frac{3\sqrt{3}}{4\sqrt{n}} \quad < \quad \frac{0.62}{\sqrt{n}} .$$

Příklad 9.3 *Tramvaj má intervaly mezi příjezdy* 10 *minut. Jaká je pravděpodobnost, že během* 24 *pracovních dnů stráví člověk při cestách do práce a zpět čekáním na tramvaj nejvýše* 3 *hodiny?*

Řešení:
Pro veličinu

$$Z = \text{“celková doba čekání během 24 dnů při cestách tam a zpět”} \text{ [v hodinách]}$$

nás zajímá $P(Z \leq 3)$.

K řešení opět použijeme centrální limitní větu. Označme si tedy pro $i = 1, 2, ..., n$, kde $n = 24 \cdot 2 = 48$, veličiny

$$X_i = \text{“doba strávená čekáním při } i\text{-tém příchodu na zastávku”} \text{ [v hodinách]}$$

které pokládáme za nezávislé. Tramvaj jezdí přesně po 10 minutách, zatímco naše příchody na zastávku budeme pokládat za náhodné s rovnoměrným rozdělením v rámci 10 minutového intervalu. Proto i doba čekání X_i bude mít rovnoměrné rozdělení (v jednotkách hodin) tvaru $\text{Ro}\,(a, b) = \text{Ro}\left(0, \frac{1}{6}\right)$.

Protože opět platí $Z = \sum_{i=1}^{n} X_i$, dostaneme

$$E(X_i) = \frac{a+b}{2} = \frac{0+\frac{1}{6}}{2} = \frac{1}{12} \quad \Rightarrow \quad E(Z) = n \cdot E(X_1) = 48 \cdot \frac{1}{12} = 4$$

$$\text{var}(X_i) = \frac{(b-a)^2}{12} = \frac{(\frac{1}{6}-0)^2}{12} = \frac{1}{12 \cdot 36} \quad \Rightarrow \quad \text{var}(Z) = n \cdot \text{var}(X_1) = 48 \cdot \frac{1}{12 \cdot 36} = \frac{1}{9}$$

$$\Rightarrow \quad \sqrt{\text{var}(Z)} = \sqrt{\frac{1}{9}} = \frac{1}{3}\,.$$

Podle CLV bude mít veličina $norm(Z) = \frac{Z - E(Z)}{\sqrt{var(Z)}} = 3(Z-4)$ přibližně rozdělení $N(0,1)$. Můžeme proto psát

$$P(Z \leq 3) = P\Big(\underbrace{3 \cdot (Z-4)}_{norm(Z)} \leq 3 \cdot (3-4)\Big) = P\Big(norm(Z) \leq -3\Big) \overset{(CLV)}{\doteq}$$

$$\overset{(CLV)}{\doteq} \Phi(-3) = 1 - \Phi(3) \doteq 1 - 0.9987 = \mathbf{0.0013}\,.$$

Odhad chyby je maximálně $\left|F_{\text{norm}(Z_n)}(t) - \Phi(t)\right| < \frac{0.62}{\sqrt{n}} = \frac{0.62}{\sqrt{48}} \doteq 0.0895$. Ale pro $t = -3$, kde nás hodnota pravděpodobnosti zajímá, je tento odhad zbytečně hrubý (protože pravděpodobnost už bude blízká k 0).

Poznámka: Proč čekací doba tentokrát nemá exponenciální rozdělení:

Veličina, která má exponenciální rozdělení měří čekací dobu buď od nějakého pevného okamžiku k nejbližší události (která přichází náhodně), anebo je to přímo čekací doba, co uplyne od poslední události než přijde (náhodná) událost další.

V tomto případě je ale náhodnou událostí příchod na zastávku a tato událost už jednoznačně určí čekací dobu, která je jen čas do dalšího pevně stanoveného příjezdu tramvaje.

Kromě toho, u Poissonova rozdělení je také předpoklad, že události jsou v časovém intervalu $\langle a, b \rangle$ rozděleny rovnoměrně. Protože Poisson je propojen s exponenciálním rozdělením, tak i v exponenciálním rozdělení se objevuje předpoklad rovnoměrnosti rozdělení událostí.

K čemu se ale toto rovnoměrné rozdělení vlastně vztahuje? Zde už nejde o dobu po sobě jdoucích událostí, ale dobu výskytu události samotné (bez ohledu na to, kdy nastaly ostatní události, tj. předchozí a následující). Tj. kdyby se dlouhodobě zaznamenávaly časy výskytu události (v nějakém časovém intervalu), tak by tyto události měly přibližně vyplnit rovnoměrně tento interval. Např. si vybereme určitou hodinu během dne a místo na silnici a opakovaně po mnoho dní měříme, v jakém čase během této hodiny projede tímto místem auto.