5. cvičení z PST

19. října 2022

Co je to spojité rozdělení:

Veličina $X : \Omega \to \mathbb{R}$ má tzv. *spojité rozdělení* $\Leftrightarrow$ její distribuční funkce F_X je spojitá.

Často je v tomto případě F_X tzv. *absolutně spojitá* $\Leftrightarrow$ ex. $f_X : \mathbb{R} \to \langle 0, +\infty)$, která je integrabilní a platí

$$F_X(t) = \int_{-\infty}^{t} f_X(u)\ \mathrm{d}u$$

pro každé $t \in \mathbb{R}$.

Funkce f_X se nazývá *hustotou pravděpodobnosti* veličiny X.

Poznámka: Stojí za to ještě poznamenat, že vlastnost absolutní spojitosti není samozřejmá pro spojité distribuční funkce - příkladem je tzv. *Cantorova funkce* c (viz např. `https://en.wikipedia.org/wiki/Cantor_function`). Tato funkce (rozšířená z intervalu $\langle 0, 1\rangle$ přirozeně na celé $\mathbb{R}$) má nulovou derivaci c' až na (z hlediska integrálu) nepodstatnou množinu bodů. Přesto však Cantorova funkce není konstantní (a tudíž nelze získat integrováním své derivace - příslušné integrály $\int_{-\infty}^{t} c'(t)\ \mathrm{d}t = 0$ jsou všechny nulové). Pro veličinu s touto distribuční funkcí neexistuje hustota pravděpodobnosti.

5.1 *Určete konstantu $c \in \mathbb{R}$ tak, aby funkce*

$$f(x) = \begin{cases} c \cdot xe^{-2x} & , x \in (0,1) \\ 0 & , \textit{jinak.} \end{cases}$$

byla hustotou nějaké náhodné veličiny.

Řešení:

Máme charakterizační větu:

Nezáporná integrabilní funkce $f : \mathbb{R} \to \mathbb{R}$ je hustotou pravděpodobnosti nějaké náhodné veličiny X právě když $\int_{\mathbb{R}} f(x)\ dx = 1$.

Použitím integrace per partes dostaneme

$$1 = \int_{-\infty}^{\infty} f(x)\ dx = \int_{-\infty}^{0} 0\ dx + \int_{0}^{1} c \cdot xe^{-2x}\ dx + \int_{1}^{\infty} 0\ dx =$$

$$= c\left[x \cdot \frac{e^{-2x}}{-2}\right]_0^1 + c\int_0^1 \frac{e^{-2x}}{2}\ dx = -\frac{c}{2}e^{-2} + c\left[\frac{e^{-2x}}{-4}\right]_0^1 = c \cdot \frac{1-3e^{-2}}{4},$$

a tudíž $c = \frac{4}{1-3e^{-2}}$. Protože $c > 0$, je splněna i nezápornost funkce f.

K významu hustoty pravděpodobnosti:

Nechť náhodná veličina X má hustotu pravděpodobnosti veličiny f_X, tj. $F_X(t) = \int_{-\infty}^{t} f_X(u)\ du$ pro každé $t \in \mathbb{R}$. Pak pro interval $I \subseteq \mathbb{R}$ (nebo i nějakou množinu poskládanou z intervalů) platí, že

$$P(X \in I) = \int_I f_X(u)\ \mathrm{d}u$$

tedy pravděpodobnost, že hodnoty veličiny X padnou do I, zjistíme prostě zintegrováním hustoty přes I (podobně zjišťujeme např. hmotnost nějaké křivky, když zintegrujeme hustotu (hmotnosti) přes danou křivku).

Poznamenejme ještě důležitou věc a sice, že hustota f_X NENÍ zdaleka určena jednoznačně jako funkce (např. její změnou v konečně mnoha bodech se nezmění příslušné integrály, takže i změněná funkce bude také hustotou).

Střední hodnotu pro náhodnou veličinu X se *spojitým* rozdělením a s hustotou pravděpodobnosti f_X si definujeme analogicky jako pro diskrétní případ

$$E(X) = \int_{-\infty}^{\infty} t \cdot f_X(t) \, \mathrm{d}t$$

(pokud tento integrál existuje.)

Názornou interpretací střední hodnoty $E(X)$ pak je, že hodnota $E(X)$ je vodorovná souřadnice těžiště plochy, která je určena grafem hustoty f_X (a vodorovnou osou).

Poznámka: Ve všech případech (diskrétním i spojitém) platí pro střední hodnotu následující vztah

$$E(X) = -\int_{-\infty}^{0} F_X(t) \, dt + \int_{0}^{\infty} \big(1 - F_X(t)\big) \, dt \ .$$

Co je rozptyl náhodné veličiny:

Rozptyl $\mathrm{var}(X)$ (libovolné) náhodné veličiny X určuje, jak moc se hodnoty odchylují od střední hodnoty. Rozptyl je proto definován jako střední hodnota veličiny $\big(X - E(X)\big)^2$, která vyjadřuje kvadrát odchylky X od své střední hodnoty, tedy předpisem:

$$\mathrm{var}(X) = E\Big(\big(X - E(X)\big)^2\Big) = \cdots = E\big(X^2\big) - \Big(E\big(X\big)\Big)^2$$

(ovšem opět jen pokud uvedené střední hodnoty existují.)

Věta: Pro (borelovsky) měřitelnou funkci $h : \mathbb{R} \to \mathbb{R}$ (v praxi to je téměř libovolná funkce, např. po částech spojitá) a diskrétní veličinu X platí, že

$$E\big(h(X)\big) = \sum_{k \in \mathbb{R}} h(k) \cdot P(X = k) \ .$$

a analogický vzorec platí pro veličinu X s spojitým rozdělením a hustotou pravděpodobnosti f_X:

$$E\big(h(X)\big) = \int_{-\infty}^{\infty} h(t) \cdot f_X(t) \, \mathrm{d}t \ .$$

5.2 *Mějme funkci*

$$f(x) = \begin{cases} 3e^{-3x} & , \ x \in \langle 0, \infty) \\ 0 & , \ \textit{jinak}. \end{cases}$$

(a) Ověřte, že f je hustota pravděpodobnosti nějaké náhodné veličiny X.

(b) Určete distribuční funkci F_X veličiny X příslušnou této hustotě.

(c) Spočtěte pravděpodobnost $P(-1 \leq X \leq 1)$.

(d) Spočtěte střední hodnotu $E(X)$ a rozptyl $\mathrm{var}(X)$.

Řešení:

(a) Funkce f je nezáporná a platí, že

$$\int_{-\infty}^{\infty} f(x)\ dx = \int_{-\infty}^{0} 0\ dx + \int_{0}^{\infty} 3e^{-3x}\ dx = \left[e^{-3x}\right]_0^{\infty} = 1,$$

tudíž vlastnost $\int_{\mathbb{R}} f(x)\ dx = 1$ je splněna a f tedy je hustotou nějaké veličiny X. Graf hustoty f viz níže.

(b) Příslušná distribuční funkce F_X pro veličinu X je

pro $t \in (0, \infty)$: $F_X(t) = \int\limits_{-\infty}^{t} f(x)\ dx = \int\limits_{-\infty}^{0} 0\ dx + \int\limits_{0}^{t} 3e^{-3x}\ dx = \left[e^{-3x}\right]_0^{t} = 1 - e^{-3t}$.

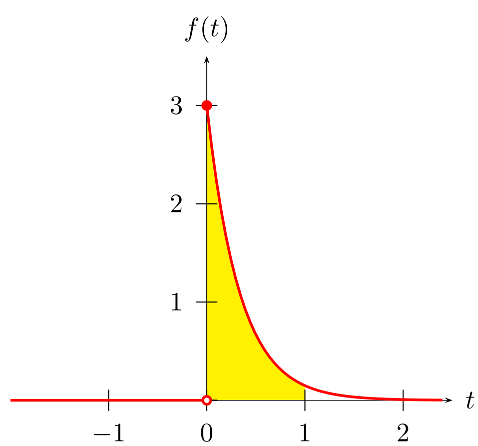
pro $t \in (-\infty, 0\rangle$: $F_X(t) = \int\limits_{-\infty}^{t} 0\ dx = 0$.

Graf distribuční funkce F_X viz níže.

(c) Hledaná pravděpodobnost je

$$P(-1 \leq X \leq 1) = \int_{-1}^{1} f(x)dx = \int_{0}^{1} 3e^{-3x}dx = 1 - e^{-3}.$$

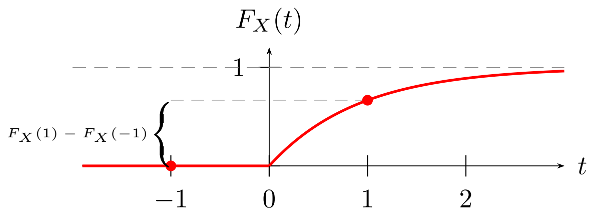
Geometrická interpretace této hodnoty je plocha pod grafem hustoty v intervalu $\langle -1, 1\rangle$:



Lze využít také distribuční funkce:

$$P(-1 \leq X \leq 1) = \underbrace{P(X \leq 1)}_{F_X(1)} - \underbrace{P(X < -1)}_{\substack{\text{limita zleva } F_X(t) \\ \text{v bodě } -1}} \overset{(\text{spojitost } F_X)}{=} F_X(1) - F_X(-1) = 1 - e^{-3\cdot 1} - 0 = 1 - e^{-3}.$$

Geometrická interpretace v tomto případě je rozdíl funkčních hodnot distribuční funkce:



(d) Použitím integrace per partes dostaneme

$$E(X) = \int_{-\infty}^{\infty} x f(x)\ dx = \int_{0}^{\infty} 3xe^{-3x} dx = \frac{1}{3},$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\ dx = \int_{0}^{\infty} 3x^2 e^{-3x} dx = \frac{2}{9}.$$

Střední hodnota je tudíž $E(X) = 1/3$ a rozptyl

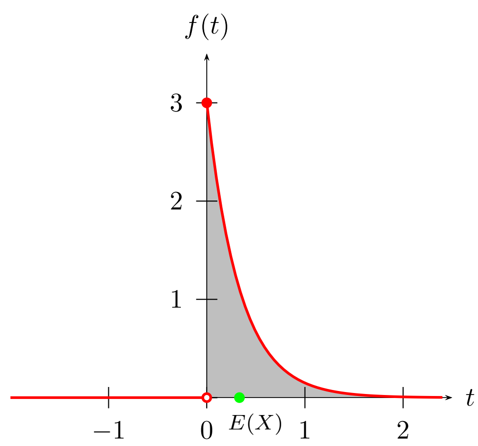
$$\operatorname{var}(X) = E\Big(\big(X - E(X)\big)^2\Big) = E\big(X^2\big) - \big(E(X)\big)^2 = \frac{2}{9} - \frac{1}{9} = \frac{1}{9}.$$

Poznamenejme ještě, že veličina X má tzv. *exponenciální* rozdělení, tj.

$$F_X(t) = \begin{cases} 0 & ,\ t < 0 \\ 1 - e^{-\frac{t}{\tau}} & ,\ t \geq 0 \end{cases}$$

kde $\tau > 0$ je parametr, jehož význam je, že $E(X) = \tau$. Dále ještě platí, že $\operatorname{var}(X) = \tau^2$.

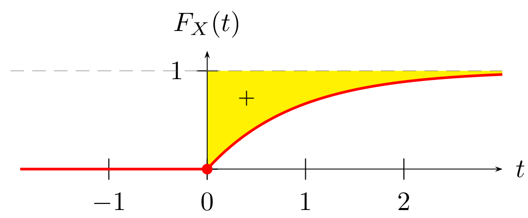
Geometrická interpretace hodnoty $E(X)$ - je to vodorovná souřadnice (zelený bod) těžiště plochy pod grafem hustoty (šedá plocha):



Ještě jedna geometrická interpretace hodnoty $E(X)$ - je to rozdíl velikosti plochy *nad* grafem distr. funkce F_X *v intervalu* $(0, +\infty)$ a velikosti plochy *pod* grafem distr. funkce F_X *v intervalu* $(-\infty, 0)$. Konkrétně:

$$E(X) = -\int_{-\infty}^{0} F_X(t)\ dt + \int_{0}^{\infty} \big(1 - F_X(t)\big)\ dt$$

Tento vztah platí dokonce pro jakoukoliv náhodnou veličinu X (diskrétní, spojitou i smíšenou).



V tomto případě je velikost plochy pod grafem F_X na intervalu $(-\infty, 0)$ nulová, takže v obrázku nejde zvýraznit jako plocha se záporným znaménkem.

Co je směs: Mějme Kolmogorův model $(\Omega, \mathcal{A}, P)$, kde množina výsledků Ω se skládá ze dvou *disjunktních* jevů Ω_1 a Ω_2. Na každé z částí Ω_i vznikne Kolmogorův model s pravděpodobností $P_i(\,\cdot\,) := P(\,\cdot\,|\Omega_i)$ pro $i = 1, 2$.

Jestliže nyní máme náhodné veličiny $X_i : \Omega_i \to \mathbb{R}$ na každém z odvozených Kolmogorových modelů, pak veličina $X : \Omega \to \mathbb{R}$ definovaná sloučením obou veličin, tedy jako

$$X(\omega) = \begin{cases} X_1(\omega) & ,\omega \in \Omega_1 \\ X_2(\omega) & ,\omega \in \Omega_2 \end{cases}$$

se nazývá *směsí* veličin X_1 a X_2 a označuje se jako $X = \mathrm{Mix}_c(X_1, X_2)$, kde $c = P(\Omega_1)$.

V této chvíli se může zdát zbytečné vypisovat ještě konstantu c, která označuje pravděpodobnost výsledku z Ω_1. Způsob, který jsme teď popsali, je vlastně rozdělením původního modelu na dva odvozené.

Můžeme však také začít opačně - tedy vzít dva "nesouvisející" modely, tj. disjunktní množiny výsledků Ω_1 a Ω_2 s příslušnými pravděpodobnostmi P_1 a P_2 a z nich složit nový Kolmogorův model. Tento nový model bude přirozeně mít množinu výsledků $\Omega := \Omega_1 \cup \Omega_2$. Pravděpodobnost P na novém modelu ovšem nebude určena, dokud si nestanovíme, jakou chceme hodnotu $c = P(\Omega_1)$ (a tím i doplňkovou hodnotu $1 - c = P(\Omega_2)$). Protože opět chceme, abychom měli $P(\,\cdot\,|\Omega_i) = P_i(\,\cdot\,)$ pro $i = 1, 2$, tak nyní bude už pravděpodobnost P určená (z věty o úplné pravděpodobnosti) pro jevy $A_1 \subseteq \Omega_1$ a $A_2 \subseteq \Omega_2$ jako

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = P(\Omega_1) \cdot P(A_1|\Omega_1) + P(\Omega_2) \cdot P(A_2|\Omega_2) =$$

$$= c \cdot P_1(A_1) + (1 - c) \cdot P_2(A_2)$$

tedy zde nutně *musíme* používat konstantu c, která pro různé hodnoty vytvoří různé Kolmogorovy modely a tím i různá rozdělení veličiny

$$X = \mathrm{Mix}_c(X_1, X_2)$$

jejíž předpis ovšem zůstane stále stejný (bez ohledu na c)!

(Rozdělení veličiny X se bude ale pochopitelně měnit podle toho, jaké pravděpodobnosti vzniknou na základě volby c.)

Příklad smíšeného rozdělení: Mějme veličinu $X =$ *"kolik naprší milimetrů srážek v daný den"*. Označme si $\Omega_1 =$ *"dny, kdy neprší"*, $\Omega_2 =$ *"dny, kdy prší"* a $\Omega = \Omega_1 \cup \Omega_2 =$ *"všechny dny"*. Pak konstantní veličina $D = X_{|\Omega_1} = 0$ má diskrétní rozdělení v rámci Ω_1 a u veličiny $S = X_{|\Omega_2}$ předpokládáme spojité rozdělení. Pak $X = \mathrm{Mix}_c(D, S)$, kde $c = P(\Omega_1) \in (0, 1)$, má smíšené rozdělení.

Připomenutí: Nechť $a \in \mathbb{R}$ je bod *spojitosti distribuční funkce* F_X náhodné veličiny X. Pak máme

$$P(X = a) = P(X \le a) - P(X < a) = F_X(a) - \lim_{x \to a_-} F_X(t) = 0$$

a tedy

$$P(X \le a) = P(X < a)$$

neboli v bodech spojitosti nezáleží na typu nerovnosti (neostré vs. ostré).

Speciálně pro veličinu X se spojitým rozdělením je $P(X = a) = 0$ pro každé $a \in \mathbb{R}$.

Poznámka: Pro veličinu X a její distribuční funkci F_X se definuje funkce

$$p_X(t) := P(X = t) = F_X(t) - F_X(t_-) \quad \text{pro všechna } t \in \mathbb{R}$$

Jestliže X má spojité rozdělení, pak je $p_X(t) = 0$ pro všechna $t \in \mathbb{R}$, tj. tento případ není zajímavý.

Jestliže X má diskrétní rozdělení, pak se p_X nazývá *pravděpodobnostní funkce* a $\sum_{t\in\mathbb{R}} p_X(t) = 1$.

5.3 (rozdělení veličiny vytvořené jako směs)

Stroj se pohybuje po úsečce $\langle -1, 2\rangle$ rovnoměrně přímočaře sem a tam. V bodech 0 a 1 se zastaví, aby tu vykonal konkrétní činnost. Doba, kterou při zastavení stráví je 25% celkového času, přičemž poměr časů strávených v bodech 0 a 1 je $2 : 3$. Náhodná veličina Z představuje polohu stroje. Určete:

(a) distribuční funkci F_Z.

(b) pravděpodobnosti $P(0 \leq Z \leq 1)$ a $P(Z \geq 0.5)$.

(c) střední hodnotu $E(Z)$.

(d) $t \in \mathbb{R}$ takové, že $P(Z \leq t) = 0.9$.

Řešení:

Pravděpodobnost rozdělení veličiny

$$Z = \text{“\textit{poloha stroje}”}$$

budeme přirozeně určovat podle času, který popisuje pohyb stroje. Konkrétně si za Ω můžeme zvolit časový interval, který odpovídá jednomu projetí úsečky $\langle -1, 2\rangle$. Pak budeme mít $Z : \Omega \to \mathbb{R}$ a $Z(\omega) =$ “*poloha stroje v čase ω*”.

Veličinu Z rozdělíme na ty polohy, kdy se stroj hýbe a ty, kdy stojí:

$$X = \text{“\textit{polohy, kdy stroj stojí}”}$$

$$Y = \text{“\textit{polohy, kdy se stroj hýbe}”}$$

tj. Z je pak směs $Z = \mathrm{Mix}_{(c_1,c_2)}(X, Y)$, kde $c_1 = 0.25$ je podíl času, který odpovídá stojícím polohám stroje a $c_2 = 1 - c_1 = 0.75$ je podíl času, které odpovídají polohám, kdy se stroj hýbe.

Pokud bychom to chtěli vyjádřit konkrétněji, máme:

$$\Omega_1 = \text{“\textit{časové okamžiky, kdy stroj stojí}”}, \qquad P(\Omega_1) = c_1 = 0.25$$

a

$$\Omega_2 = \text{“\textit{časové okamžiky, kdy se stroj hýbe}”}, \qquad P(\Omega_2) = c_2 = 0.75$$

a tedy $X = Z|_{\Omega_1}$ a $X : \Omega_1 \to \mathbb{R}$ a podobně $Y = Z|_{\Omega_2}$ a $Y : \Omega_2 \to \mathbb{R}$.

(a) Ze vztahu $Z = \mathrm{Mix}_{(c_1,c_2)}(X, Y)$ plyne

$$F_Z(t) = c_1 F_X(t) + c_2 F_Y(t)$$

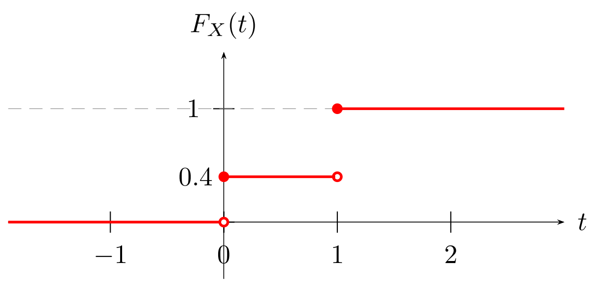
takže je potřeba určit distribuční funkce F_X a F_Y.

Veličina X má hodnoty $\{0, 1\}$, tedy má alternativní rozdělení $\mathrm{Alt}(p)$, kde p je pravděpodobnost hodnoty 1. Protože pravděpodobnosti hodnot 0 a 1 veličiny X jsou rozděleny v poměru $2 : 3$, máme $p = p_X(1) = \frac{3}{5} = 0.6$ a $p_X(0) = \frac{2}{5} = 0.4$.

Připomeňme, že pravděpodobnost hodnot veličiny X se počítá jako $p_X(u) = P_1(X = u) = P(Z = u \mid \Omega_1)$ tj. jako podmíněná pravděpodobnost vzhledem k Ω_1.

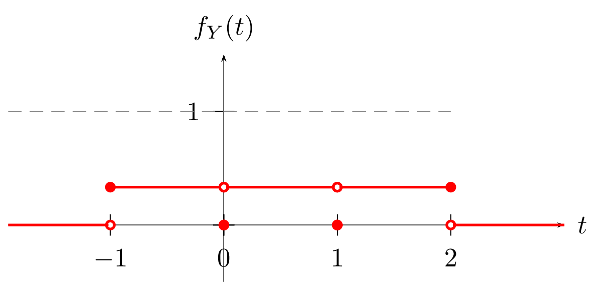
Tudíž X má distribuční funkci

$$F_X(t) = \sum_{u \leq t} p_X(u) = \begin{cases} 0\,, & t < 0\,, \\ 0.4\,, & 0 \leq t < 1\,, \\ 1\,, & t \geq 1\,. \end{cases}$$



Veličina Y bude mít (spojité) rovnoměrné rozdělení na množině svých hodnot $I = \langle -1, 2 \rangle \setminus \{0, 1\}$. To znamená, že existuje hustota pravděpodobnosti f_Y veličiny Y tvaru

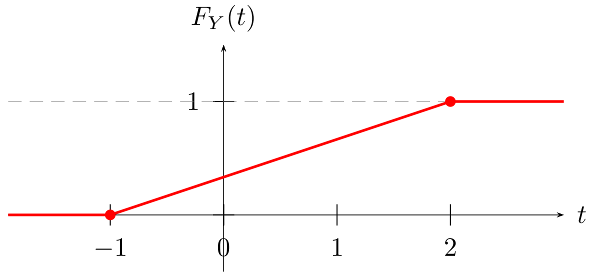
$$f_Y(t) = \begin{cases} c\,, & t \in I\,, \\ 0\,, & \text{jinak}\,, \end{cases}$$



Z podmínky, že plocha pod grafem hustoty je 1, snadno určíme konstantu $c = \frac{1}{3}$ (současně si můžeme hustotu pozměnit v konečně mnoha bodech, tj. např. v $t = 0$ a $t = 1$ můžeme f_Y udělat spojitou).

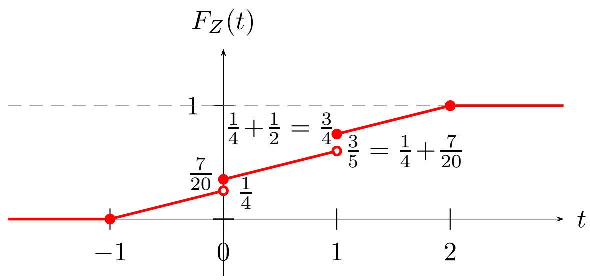
Distribuční funkce F_Y vznikne integrací hustoty f_Y a protože integrujeme konstantní funkci na intervalu $\langle -1, 2 \rangle$, bude grafem F_Y na tomto intervalu úsečka. Konkrétně:

$$F_Y(t) = \int_{-\infty}^{t} f_Y(u)\,\mathrm{d}u = \begin{cases} 0, & t < -1\,, \\ \int_{-1}^{t} \frac{1}{3}\,\mathrm{d}u = \frac{t+1}{3}, & -1 \leq t < 2\,, \\ 1, & t \geq 2\,. \end{cases}$$



Pro distribuční funkci veličiny $Z = \mathrm{Mix}_{0.25,\ 0.75}(X, Y)$ pak platí

$$F_Z(t) = \frac{1}{4}F_X(t) + \frac{3}{4}F_Y(t) = \begin{cases} & 0, & t < -1\,, \\ \frac{1}{4}\cdot 0 + \frac{3}{4}\cdot\frac{t+1}{3} = & \frac{1}{4}t + \frac{1}{4}, & -1 \le t < 0\,, \\ \frac{1}{4}\cdot 0.4 + \frac{3}{4}\cdot\frac{t+1}{3} = & \frac{1}{4}t + \frac{7}{20}, & 0 \le t < 1\,, \\ \frac{1}{4}\cdot 1 + \frac{3}{4}\cdot\frac{t+1}{3} = & \frac{1}{4}t + \frac{1}{2}, & 1 \le t < 2\,, \\ & 1, & t \ge 2\,. \end{cases}$$



(b) $P(0 \le Z \le 1) = P(Z \le 1) - P(Z < 0) = F_Z(1) - \lim\limits_{t\to 0_-} F_Z(t) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$
$P(Z \ge 0.5) = 1 - P(Z < 0.5) = 1 - \lim\limits_{t\to 0.5_-} F_Z(t) = 1 - \left(\frac{1}{4}\cdot 0.5 + \frac{7}{20}\right) = \frac{21}{40}$.

(c) Pro alternativní rozdělení veličiny X je $E(X) = 0\cdot p_X(0) + 1\cdot p_X(1) = 0.6 = \frac{3}{5}$ a pro rovnoměrné rozdělení veličiny Y na intervalu $(a, b) = (-1, 2)$ je $E(Y) = \frac{a+b}{2} = \frac{1}{2}$. Takže pro $Z = \mathrm{Mix}_{1/4}(X, Y)$ je

$$E(Z) = \frac{1}{4}E(X) + \frac{3}{4}E(Y) = \frac{1}{4}\cdot\frac{3}{5} + \frac{3}{4}\cdot\frac{1}{2} = \frac{21}{40}\ .$$

(d) Hledáme $t \in \mathbb{R}$ tak, že $0.9 = P(Z \le t) = F_Z(t)$. K tomu potřebujeme vědět, kterou část předpisu pro F_Z máme použít. Protože $\frac{3}{4} \le 0.9 \le 1$, což je rozmezí hodnot na poslední rostoucí části předpisu, tak musíme použít právě tuto část:

$$0.9 = F_Z(t) = \frac{1}{4}t + \frac{1}{2} \qquad \Rightarrow \qquad t = 3.6 - 2 = 1.6 \in \langle 1, 2\rangle\ .$$