8. cvičení z PST

13. listopadu 2019

Připomenutí: Kovariance $\text{cov}(\cdot,\cdot)$ má tyto vlastnosti (X, Y, Z jsou veličiny, $a, b, c, d \in \mathbb{R}$):

- je lineární v každé složce zvlášť (tj. je bilineární), tedy:

$$\text{cov}(aX + bY, Z) = a \cdot \text{cov}(X, Z) + b \cdot \text{cov}(Y, Z)$$
$$\text{cov}(Z, aX + bY) = a \cdot \text{cov}(Z, X) + b \cdot \text{cov}(Z, Y)$$

- symetrická, tj. $\text{cov}(X, Y) = \text{cov}(Y, X)$
- pozitivně semi-definitní, tj. $\text{cov}(X, X) \geq 0$, kde navíc platí, že:
  $\text{cov}(X, X) = 0 \ \Rightarrow\ \exists \alpha \in \mathbb{R}$, že $P(X = \alpha) = 1$ (neboli: X odpovídá konstantní veličině)
- $\text{cov}(X + c, Y + d) = \text{cov}(X, Y)$,
- $\text{cov}(X, X) = D(X) = (\sigma_X)^2$.

Dále je $\varrho(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{D(X)\cdot D(Y)}} = \frac{\text{cov}(X,Y)}{\sigma_X\cdot\sigma_Y}$ a proto platí

$$\varrho(aX + c, bY + d) = \text{sgn}(ab) \cdot \varrho(X, Y)$$

pro $a \neq 0$ a $b \neq 0$ a c, d libovolné konstanty (kde sgn je znaménková funkce).

Pro rozptyl $D(\cdot)$ díky tomu máme:

- $D(aX + c) = D(aX) = a^2 \cdot D(X)$,
- $D(X + Y) = D(X) + D(Y) + 2 \cdot \text{cov}(X, Y)$ (viz níže).

Poznámka: Uvědomme si, že existuje několik stupňů "nezávislosti" veličin:

X a Y jsou nezávislé $\xRightarrow{\text{pokud cov ex.}}$ $\text{cov}(X, Y) = 0$ (tj. $X - E(X)$ a $Y - E(Y)$ jsou kolmé) $\xRightarrow{\text{pokud } X, Y \text{ nejsou konst.}}$ X a Y jsou lineár. nezáv.

Konstantní veličina X spolu s jakoukoliv jinou veličinou Y vždy tvoří vzájemně nezávislé veličiny X a Y (tento případ je ale celkem nezajímavý).

8.1 (kovariance, korelace)

Náhodný vektor (X, Y) má následující parametry:

$$E(X) = 10,\quad \sigma_X = 5,\quad E(Y) = 150,\quad \sigma_Y = 20,\quad \varrho(X, Y) = 0.5 \text{ (korelace)}.$$

(a) Stanovte střední hodnotu a rozptyl náhodných veličin

$$T = 2X + 3,\quad U = 200 - Y,\quad V = X + Y\ .$$

(b) Určete kovarianci $\text{cov}(T, U)$ a rozhodněte, zda jsou náhodné veličiny T a U závislé či nezávislé.

Řešení:

(a) Střední hodnota je lineární zobrazeni ("veličina"$\mapsto$ "její střední hodnota") :

$$E(T) = E(2X + 3) = 2E(X) + 3 = 2 \cdot 10 + 3 = 23$$

$$E(U) = E(200 - Y) = 200 - E(Y) = 200 - 150 = 50$$

$$E(V) = E(X + Y) = E(X) + E(Y) = 10 + 150 = 160$$

Rozptyl na druhou stranu NENÍ lineární zobrazení. (Je to tzv. kvadratická forma, tedy vzniká z bilineárního zobrazení - viz výše.):

$$D(T) = D(2X+3) = D(2X) = 2^2 \cdot D(X) = 2^2 \cdot (\sigma_X)^2 = 4 \cdot 25 = 100$$

$$D(U) = D(200 - Y) = (-1)^2 \cdot D(Y) = (\sigma_Y)^2 = 400$$

$$\begin{aligned} D(V) &= D(X+Y) = \operatorname{cov}(X+Y, X+Y) = \\ &= \operatorname{cov}(X,X) + \operatorname{cov}(X,Y) + \operatorname{cov}(Y,X) + \operatorname{cov}(Y,Y) = \\ &= D(X) + 2 \cdot \operatorname{cov}(X,Y) + D(Y) = \\ &= (\sigma_X)^2 + 2 \cdot \sigma_X \cdot \sigma_Y \cdot \varrho(X,Y) + (\sigma_Y)^2 = \\ &= 25 + 2 \cdot 5 \cdot 20 \cdot 0.5 + 400 = 525 \ . \end{aligned}$$

(b) $\operatorname{cov}(T,U) = \operatorname{cov}(2X+3, 200-Y) = \operatorname{cov}(2X, -Y) =$

$$= 2 \cdot (-1) \cdot \operatorname{cov}(X,Y) = (-2) \cdot \sigma_X \cdot \sigma_Y \cdot \varrho(X,Y) = (-2) \cdot 5 \cdot 20 \cdot 0.5 = -100$$

Protože kovariance vyšla nenulová, můžeme hned říct, že U z T musí být *závislé*. V případě, že by kovariance byla nulová, bychom o nezávislosti nic usoudit nemohli!

8.2 (kovariance, kovarianční a korelační matice)

Pro náhodné veličiny X a Y platí, že

$$E(X) = 2, \quad E(Y) = -1, \quad D(X) = 3, \quad D(Y) = 4, \quad \operatorname{cov}(X,Y) = -2.$$

Pro náhodné veličiny

$$U = 3X + 4Y - 1 \quad \text{a} \quad V = -2X + 2Y + 3$$

určete

(a) koeficient kovariance $\operatorname{cov}(U,V)$ a střední hodnotu $E(U)$.

(b) rozptyl $D(X+Y)$.

(c) kovarianční a korelační matice náhodných vektoru (X,Y) a $(X,-2Y)$.

Řešení:

(a) Díky bilinearitě kovariance můžeme jednotlivé složky "roznásobit":

$$\operatorname{cov}(U,V) = \operatorname{cov}(3X+4Y-1,\ -2X+2Y+3) = \operatorname{cov}(3X+4Y,\ -2X+2Y) =$$

$$= 3 \cdot (-2) \cdot \operatorname{cov}(X,X) + 3 \cdot 2 \cdot \operatorname{cov}(X,Y) + 4 \cdot (-2) \cdot \operatorname{cov}(Y,X) + 4 \cdot 2 \cdot \operatorname{cov}(Y,Y) =$$

$$= (-6) \cdot D(X) + (-2) \cdot \operatorname{cov}(X,Y) + 8 \cdot D(Y) = (-6) \cdot 3 + (-2) \cdot (-2) + 8 \cdot 4 = 18 \ .$$

Jak je vidět, znalosti středních hodnot jsme zatím vůbec nepotřebovali!

Poznámka: Je dobré si všimnout, ze díky bilinearitě můžeme také používat přehlednější maticový zápis:

$$\operatorname{cov}(aX+bY, cX+dY) = (a,b) \begin{pmatrix} \operatorname{cov}(X,X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y) \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix}$$

V našem případě tedy

$$\operatorname{cov}(U,V) = (3,4) \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} -2 \\ 2 \end{pmatrix} = (3,4) \begin{pmatrix} -10 \\ 12 \end{pmatrix} = 18 \ .$$

Střední hodnota:

$$E(U) = E(3X + 4Y - 1) = 3E(X) + 4E(Y) - 1 = 3 \cdot 2 + 4 \cdot (-1) - 1 = 1$$

(b) Využijeme vlastnosti kovariance:

$$D(X+Y) = \operatorname{cov}(X+Y, X+Y) = \operatorname{cov}(X,X) + \operatorname{cov}(X,Y) + \operatorname{cov}(Y,X) + \operatorname{cov}(Y,Y) =$$

$$= D(X) + 2 \cdot \operatorname{cov}(X,Y) + D(Y) =$$

$$= 3 + 2 \cdot (-2) + 4 = 3\ .$$

(c) Kovarianční matice pro (X, Y):

$$\begin{pmatrix} \operatorname{cov}(X,X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} D(X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & D(Y) \end{pmatrix} = \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}$$

Korelační matice pro (X, Y):

$$\begin{pmatrix} \varrho(X,X) & \varrho(X,Y) \\ \varrho(Y,X) & \varrho(Y,Y) \end{pmatrix} = \begin{pmatrix} 1 & \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X) \cdot D(Y)}} \\ \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X) \cdot D(Y)}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} & 1 \end{pmatrix}$$

protože máme $\varrho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X) \cdot D(Y)}} = \frac{-2}{\sqrt{3 \cdot 4}} = -\frac{1}{\sqrt{3}}$.

Podobně pro vektor $(X, -2Y)$ máme korelační matici:

$$\begin{pmatrix} D(X) & \operatorname{cov}(X,-2Y) \\ \operatorname{cov}(X,-2Y) & D(-2Y) \end{pmatrix} = \begin{pmatrix} D(X) & -2 \cdot \operatorname{cov}(X,Y) \\ -2 \cdot \operatorname{cov}(X,Y) & (-2)^2 \cdot D(Y) \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 4 & 16 \end{pmatrix}$$

a korelační matici:

$$\begin{pmatrix} 1 & \frac{4}{\sqrt{3 \cdot 16}} \\ \frac{4}{\sqrt{3 \cdot 16}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & 1 \end{pmatrix}$$

což není překvapení, protože díky vlastnostem korelace (viz výše) máme, že:

$$\varrho(X, -2Y) = \operatorname{sgn}(-2) \cdot \varrho(X,Y) = -\varrho(X,Y)\ .$$

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 integrabilni 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).

Je dobré si uvědomit, že náhodný vektor (X, Y) můžeme snadno sestavit z libovolných dvou náhodných veličin X a Y. Zatímco ale k počítání s veličinou X nám stačí znát jen její distribuční funkci F_X, k práci s vektorem nám NESTAČÍ znalost distribučních funkcí jeho složek! Potřebujeme totiž znát, jaký je vztah mezi veličinami X a Y, a ten je schovaný právě ve sdružené distribuční funkci.

8.3 (spojitý náhodný vektor)

Náhodný vektor (X, Y) má spojité rovnoměrné rozdělení v množině

$$E = \{(x, y) \in \mathbb{R}^2 \mid y \geq 0,\ x + y \leq 1,\ y - x \leq 1\} .$$

Určete:

(a) sdruženou hustotu $f_{X,Y}$.

(b) marginální hustoty f_X, f_Y.

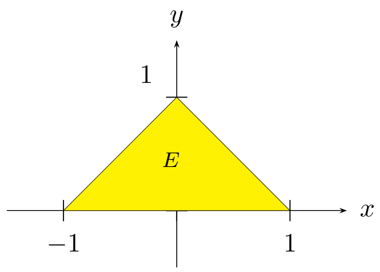
(c) zda jsou náhodné veličiny X a Y závislé či nezávislé.

(d) hodnotu $P(X + Y \geq \frac{1}{2})$.

(e) koeficient korelace $\varrho(X, Y)$.

Řešení:

Množina E je trojúhelník



(a) Spojité rovnoměrné rozdělení na E odpovídá modelu geometrické pravděpodobnosti na E (můžeme si pro větší názornost představovat E jako terč, do kterého se trefujeme všude se stejnou "intenzitou"). Sdružená hustota je tedy dána jako

$$f_{X,Y}(x, y) = \begin{cases} c & , (x, y) \in E \\ 0 & , \text{jinak}, \end{cases}$$

kde c je konstanta taková, aby integrál z hustoty byl roven 1, tedy

$$1 = \iint\limits_{\mathbb{R}^2} f_{X,Y}(x, y)\ dx\ dy = \iint\limits_{E} c\ dx\ dy = c \iint\limits_{E} 1\ dx\ dy = c \cdot \text{obsah}(E) = c \cdot 1$$

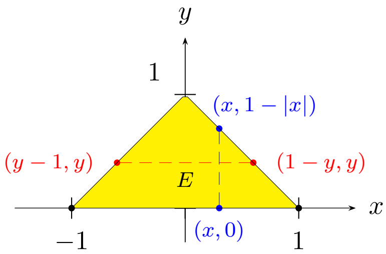
neboli

$$c = 1 .$$

Integrál $\iint\limits_{E} 1\ dx\ dy$ jsme mohli buď skutečně spočítat nebo prostě využít jeho geometrickou interpretaci, tj. že je to obsah trojúhelníka.

(b) Marginální rozdělení vektoru (X, Y) jsou rozdělení jeho jednotlivých složek, tj. veličin X a Y. Marginální hustoty f_X a f_Y jsou tedy hustoty náhodných veličin X a Y. V tomto případě je nejsnadněji získáme částečným zintegrováním sdružené hustoty, tedy integrací podél vhodného řezu množiny E, kterou si ještě vyjádříme jako

$$E: \quad -1 \leq x \leq 1 \;\;\&\;\; 0 \leq y \leq 1 - |x|$$



Pro $x \in \langle -1, 1\rangle$ tak máme

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\; dy = \int_0^{1-|x|} 1\; dy = 1 - |x|$$

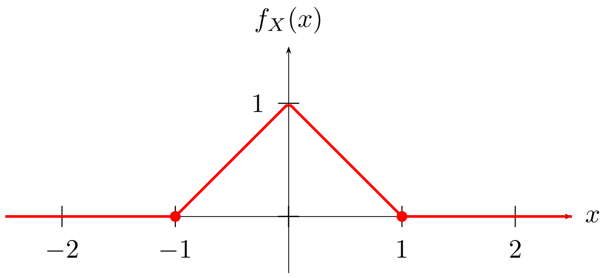
a pro $y \in \langle 0, 1\rangle$ máme

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\; dx = \int_{y-1}^{1-y} 1\; dx = 2(1 - y)\;.$$

Celkově tedy

$$f_X(x) = \begin{cases} 1 - |x| & ,\; x \in \langle -1, 1\rangle \\ 0 & ,\; \text{jinak,} \end{cases}$$

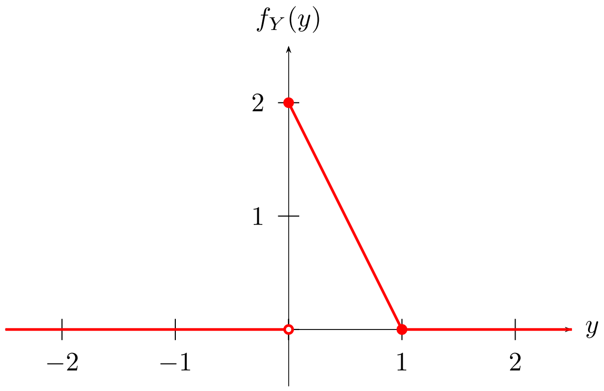
s grafem



a podobně

$$f_Y(y) = \begin{cases} 2(1 - y) & ,\; y \in \langle 0, 1\rangle \\ 0 & ,\; \text{jinak,} \end{cases}$$

s grafem



(c) Veličiny X a Y jsou nezávislé právě když

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y) \quad \text{pro všechna} \quad (x,y) \in \mathbb{R}^2 \ .$$

V případě existence sdružené hustoty je to ekvivalentní tomu, že

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

(tj. množina bodů, kde rovnost NENASTÁVÁ, má nulový plošný obsah). Obě strany se mohou lišit např. v konečně mnoha bodech, nebo na nějaké křivce atd. Rovnost hustot *skoro všude* se velmi často opomíjí a nevyznačuje se. Nicméně to nic nemění na tom, že je potřeba tuhle věc mít na paměti.

Speciálně, pro nezávislé veličiny musí platit následující (pouze nutná podmínka!):

Nechť

$$S = \{(x,y) \in \mathbb{R}^2 \mid f_{X,Y}(x,y) \neq 0\}$$

a $\pi_1, \pi_2 : \mathbb{R}^2 \to \mathbb{R}$ jsou projekce na jednotlivé souřadné osy, tj. $\pi_1(x,y) = x$ a $\pi_2(x,y) = y$. Pokud jsou veličiny X a Y *nezávislé*, má množina

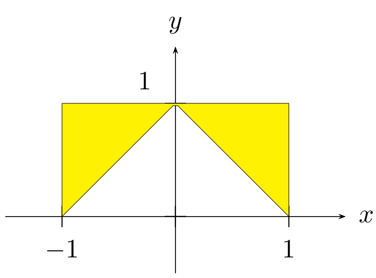
$$\Big(\pi_1(S) \times \pi_2(S)\Big) \setminus S$$

nulový obsah.

V našem konkretním případě $S = E$ a $\pi_1(E) = \langle -1, 1\rangle$ a $\pi_2(E) = \langle 0, 1\rangle$. Ovšem množina

$$\Big(\pi_1(E) \times \pi_2(E)\Big) \setminus E \quad = \quad \text{“obdélník”} \ \ \setminus \ \ \text{“trojúhelník”}$$

tj.



zřejmě nulový obsah NEMÁ. Veličiny X a Y tudíž NEJSOU nezávislé.

Naopak, pokud bychom hledali sdruženou hustotu vektoru (X', Y'), který by měl mít stejná marginální rozdělení jako vektor (X, Y), ale složky X' a Y' by byly nezávislé, stačilo by (podobně jako u diskrétních vektorů a jejich tabulek) položit

$$f_{X',Y'}(x,y) = f_X(x) \cdot f_Y(y) = \begin{cases} 2(1-|x|)(1-y), & (x,y) \in \langle -1,1\rangle \times \langle 0,1\rangle \\ 0, & \text{jinak.} \end{cases}$$

(d) Jev "$X + Y \geq \frac{1}{2}$" je množina $\Phi^{-1}(A)$, kde $\Phi = (X,Y) : \Omega \to \mathbb{R}^2$ je náš náhodný vektor a

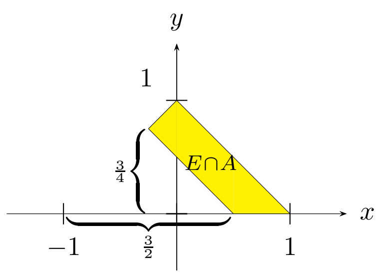
$$A = \{(x,y) \in \mathbb{R}^2 \mid x + y \geq \frac{1}{2}\}$$

tj. $\Phi^{-1}(A)$ je vzor množiny $A \subseteq \mathbb{R}^2$ při zobrazení Φ. Pravděpodobnost tohoto jevu tak dostaneme zintegrováním hustoty přes množinu A, neboli

$$P(X + Y \geq \tfrac{1}{2}) = \iint\limits_{A=\{(a,b)\in\mathbb{R}^2 \mid a+b\geq\frac{1}{2}\}} f_{X,Y}(x,y)\ dx\ dy = \iint\limits_{E\cap A} 1\ dx\ dy = \text{obsah}(E \cap A)\ .$$

Množina $E \cap A$ je tvaru

$$E \cap A : \quad \tfrac{1}{2} \leq x + y \leq 1 \quad \wedge \quad y - x \leq 1 \quad \wedge \quad y \geq 0$$



Velikost plochy $E \cap A$ spočítáme snadněji pomocí doplňkové plochy (tj. trojúhelníka s výškou $\frac{3}{4}$ a základnou $\frac{3}{2}$) do původního trojúhelníka E (s plochou 1):

$$P(X + Y \geq \tfrac{1}{2}) = 1 - \tfrac{1}{2} \cdot \tfrac{3}{4} \cdot \tfrac{3}{2} = \tfrac{7}{16}\ .$$

(e) Pro kovarianci máme

$$\varrho(X,Y) = \frac{E(XY) - E(X) \cdot E(Y)}{\sqrt{D(X) \cdot D(Y)}}\ .$$

Přitom je

$$E(XY) = \iint\limits_{\mathbb{R}^2} xy \cdot f_{X,Y}(x,y)\ dx\ dy = \iint\limits_{A} 2xy\ dx\ dy =$$

$$= \int\limits_0^1 \int\limits_{y-1}^{1-y} 2xy\ dx\ dy = \int\limits_0^1 y \underbrace{\left[x^2\right]_{x=y-1}^{x=1-y}}_{=0} dy = 0.$$

To, že tento integrál vyjde nula bylo vidět už na začátku z lichosti integrované funkce (tj. $xy \cdot f_{X,Y}(x,y)$) vzhledem k proměnné x (lichost této funkce je pochopitelně určena i tím, že množina A je symetrická podle osy y). Ze stejných důvodů bude nulový i následující integrál:

$$E(X) = \iint_{\mathbb{R}^2} x \cdot f_{X,Y}(x,y)\; dx\; dy = 0$$

což bylo vidět už z toho, že hustota f_X pro X byla sudá.

Takže $\varrho(X,Y) = 0$, ačkoliv veličiny X a Y nezávislé nejsou (viz (c)).

Poznamenejme ještě, že $E(Y) = \frac{1}{3}$, což snadno zjistíme integrováním pomocí hustoty f_Y a nebo z toho, že plocha pod hustotou je (pravoúhlý) trojúhelník, jehož x-ova složka těžiště je právě ve vzdálenosti $\frac{1}{3}$ od svislé odvěsny.

8.4 (spojitý náhodný vektor)

Náhodný vektor (X,Y) má sdruženou hustotu $f_{X,Y}$. Jeho sdružená distribuční funkce je

$$F_{X,Y}(x,y) = \frac{1}{250}(20xy - x^2y - xy^2) \quad \text{pro} \quad (x,y) \in \langle 0,5\rangle \times \langle 0,5\rangle.$$

Určete:

(a) sdruženou hustotu $f_{X,Y}$.

(b) $F_{X,Y}(x,y)$ v ostatních bodech.

(c) marginální distribuční funkce F_X, F_Y.

(d) zda jsou náhodné veličiny X a Y závislé či nezávislé.

(e) hodnotu $P(X > 2)$.

(f) střední hodnotu vektoru $E(X,Y)$.

Řešení:

Podobně jako u diskrétního vektoru zadaného tabulkou potřebujeme i zde ověřit, že máme všechny informace pro popis pravděpodobnosti na $\mathbb{R}^2$. Znamená to, že musí platit, že $P\big((X,Y) \in \langle 0,5\rangle^2\big) = 1$.

Proč to potřebujeme vědět: V opačném případě by totiž pro $D = \mathbb{R}^2 \setminus \langle 0,5\rangle^2$ bylo $P\big((X,Y) \in D\big) > 0$. Pak bychom ovšem měli $\iint_D f_{X,Y} \mathrm{d}S = P\big((X,Y) \in D\big) > 0$ a přitom bychom nevěděli, jaký tvar hustota $f_{X,Y}$ na D má!

K tomuto ověření využijeme existence hustoty a vlastnosti sdružené distribuční funkce:

$$P\big((X,Y) \in \langle 0,5\rangle^2\big) = \iint_{\langle 0,5\rangle^2} f_{X,Y} \mathrm{d}S = \iint_{(0,5\rangle^2} f_{X,Y} \mathrm{d}S = P\big((X,Y) \in (0,5\rangle^2\big) =$$

$$= F_{X,Y}(5,5) - F_{X,Y}(0,5) - F_{X,Y}(5,0) + F_{X,Y}(0,0) =$$

$$= \frac{1}{250}(20 \cdot 5 \cdot 5 - 5^2 \cdot 5 - 5 \cdot 5^2) - 0 - 0 + 0 = 1\ .$$

Připomenutí odvození vlastnosti pro $F_{X,Y}$: Označme si pro jednoduchost $U_{a,b} := (-\infty, a\rangle \times (-\infty, b\rangle$ Pro pravděpodobnost $P_{X,Y}$ na $\mathbb{R}^2$ máme:

$$(0,5\rangle^2 = U_{5,5} \setminus (U_{0,5} \cup U_{5,0})$$

a tedy

$$P\Big((X,Y)\in(0,5\rangle^2\Big) = P_{X,Y}\big((0,5\rangle^2\big) = P_{X,Y}\big(U_{5,5}\big) - P_{X,Y}\big(U_{0,5}\cup U_{5,0}\big) =$$
$$= P_{X,Y}\big(U_{5,5}\big) - P_{X,Y}\big(U_{0,5}\big) - P_{X,Y}\big(U_{5,0}\big) + P_{X,Y}\big(\underbrace{U_{0,5}\cap U_{5,0}}_{=U_{0,0}}\big) =$$
$$= F_{X,Y}(5,5) - F_{X,Y}(0,5) - F_{X,Y}(5,0) + F_{X,Y}(0,0)\ .$$

(a) Vzhledem k tomu, že už víme, že $P\Big((X,Y)\in(0,5)^2\Big)=1$ (okraje intervalu díky hustotě nehrají roli), musí být $P\Big((X,Y)\in\mathbb{R}^2\setminus(0,5)^2\Big)=0$. Na množině $\mathbb{R}^2\setminus(0,5)^2$ se proto hustota nemůže významně lišit od nulové funkce. Konkrétněji nám to popisuje následující věta, která zachytí i situaci pro body $(x,y)\in(0,5)^2$:

Podmínka pro sdruženou hustotu: Nechť $U\subseteq\mathbb{R}^2$ je otevřená množina a (X,Y) je náhodný vektor. Jestliže

- $P\Big((X,Y)\in\mathbb{R}^2\setminus U\Big)=0$ a
- sdružená distribuční funkce $F_{X,Y}$ je třídy C^2 na U (tj. má v U spojité druhé parciální derivace)

pak funkce

$$f_{X,Y}(x,y)=\begin{cases}\frac{\partial^2 F_{X,Y}}{\partial x\partial y}(x,y) & ,\ (x,y)\in U\\ 0 & ,\ (x,y)\in\mathbb{R}^2\setminus U\end{cases}$$

je hustotou pravděpodobnosti pro (X,Y).

Pro $(x,y)\in(0,5)^2$ tedy máme

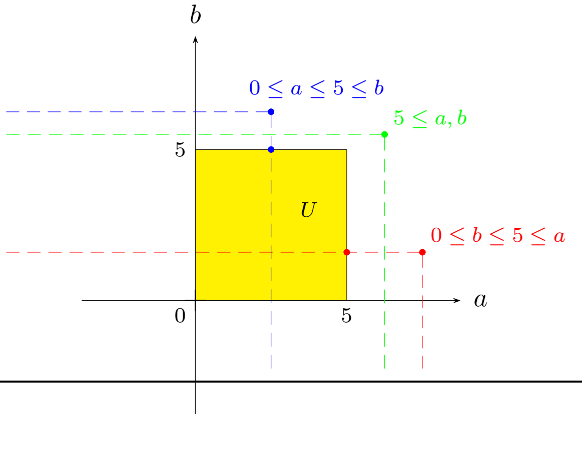
$$\frac{\partial^2 F_{X,Y}}{\partial x\partial y}(x,y)=\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}\Big(\frac{20xy-x^2y-xy^2}{250}\Big)\right)=$$

$$=\frac{\partial}{\partial x}\left(\frac{20x-x^2-2xy}{250}\right)=\frac{10-x-y}{125}$$

Celkem jsme tedy zjistili, že hustota pravděpodobnosti je

$$f_{X,Y}(x,y)=\begin{cases}\frac{10-x-y}{125} & ,\ (x,y)\in(0,5)^2\\ 0 & ,\ \text{jinak.}\end{cases}$$

(b) Ostatní hodnoty $F_{X,Y}(a,b)$ už snadno určíme z obrázku (funkce $F_{X,Y}(a,b)$ je spojitá a hustota $f_{X,Y}$ je nulová mimo žlutou oblast $U=(0,5)^2$):



Např. pro $0 \le b \le 5 \le a$ máme, že

$$F_{X,Y}(a,b) = \iint\limits_{U \cap U_{a,b}} f_{X,Y}\ dS = \iint\limits_{U \cap U_{5,b}} f_{X,Y}\ dS = F_{X,Y}(5,b)$$

Celkem máme

$$F_{X,Y}(a,b) = \begin{cases} 0 & , \ a \le 0 \text{ nebo } b \le 0 \\ \frac{1}{250}(20ab - a^2b - ab^2) & , \ 0 \le a, b \le 5 \\ F_{X,Y}(5,b) = \frac{1}{50}(15b - b^2) & , \ 0 \le b \le 5 \le a \\ F_{X,Y}(a,5) = \frac{1}{50}(15a - a^2) & , \ 0 \le a \le 5 \le b \\ 1 & , \ 5 \le a, b. \end{cases}$$

Z cvičných důvodů si ještě zpětně odvodíme hodnoty $F_{X,Y}$ z hustoty $f_{X,Y}$ pro $(a,b) \in (0,5)^2$:

$$F_{X,Y}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x,y)\ \mathrm{d}y \mathrm{d}x = \int_0^a \int_0^b \frac{10 - x - y}{125}\ \mathrm{d}y \mathrm{d}x =$$

$$= \frac{1}{125} \int_0^a (10 - x)b - \left[\frac{y^2}{2}\right]_0^b \mathrm{d}x = \frac{1}{125} \int_0^a 10b - xb - \frac{b^2}{2}\ \mathrm{d}x =$$

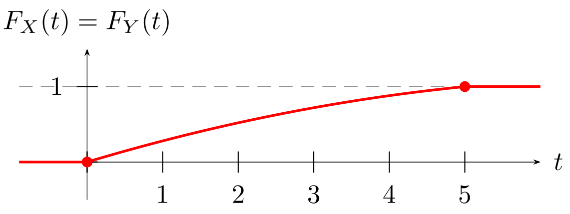
$$= \frac{1}{125}\left(10ba - \frac{a^2}{2}b - \frac{b^2}{2}a\right) = \frac{1}{250}(20ab - a^2b - ab^2)$$

(c) Marginální distribuční funkce F_X a F_Y představují distribuční funkce jednotlivých složek vektoru, tj. (samostatných) náhodných veličin X a Y. Získáme je jednoduše jako limity sdružené distribuční funkce (také s využitím obrázků):

$$F_X(a) = P(X \le a) = \lim_{b \to \infty} P(X \le a, Y \le b) = \lim_{b \to \infty} F_{X,Y}(a,b) = \begin{cases} 0 & , \ a \le 0 \\ \frac{1}{50}(15a - a^2) & , \ 0 \le a \le 5 \\ 1 & , \ 5 \le a \end{cases}$$

$$F_Y(b) = P(Y \le b) = \lim_{a \to \infty} P(X \le a, Y \le b) = \lim_{a \to \infty} F_{X,Y}(a,b) = \begin{cases} 0 & , \ b \le 0 \\ \frac{1}{50}(15b - b^2) & , \ 0 \le b \le 5 \\ 1 & , \ 5 \le b \end{cases}$$

s grafy



(d) Veličiny X a Y jsou nezávislé právě když

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y) \quad \text{pro všechna} \quad (x,y) \in \mathbb{R}^2 \ .$$

V našem případě pro $(x,y) \in \langle 0,5 \rangle^2$ máme

$$F_X(x) \cdot F_Y(y) = \frac{1}{2500}(15x - x^2)(15y - y^2)$$

což se zjevně liší od $F_{X,Y}(x,y) = \frac{1}{250}(20xy - x^2y - xy^2)$ např. v bodě $(x,y) = (1,1)$.

Veličiny X a Y tedy jsou *závislé*.

(e)

$$P(X > 2) = 1 - P(X \leq 2) = 1 - \lim_{b \to \infty} P(X \leq 2, Y \leq b) =$$

$$= 1 - \lim_{b \to \infty} F_{X,Y}(2,b) = 1 - F_{X,Y}(2,5) = 1 - \frac{130}{250} = \frac{12}{25} = 0.48$$

(f) Střední hodnota náhodného vektoru (X,Y) popisuje jakou průměrnou hodnotu dvojic (x,y) můžeme při opakovaných měřeních očekávat. Střední hodnota je přirozeně definována jako

$$E(X,Y) := (E(X), E(Y)) \ .$$

Protože X a Y mají stejná rozdělení, tak máme

$$E(Y) = E(X) = -\int_{-\infty}^{0} F_X(t)\ \mathrm{d}t + \int_{0}^{\infty} \big(1 - F_X(t)\big)\ \mathrm{d}t = \int_{0}^{5} 1 - \frac{1}{50}(15t - t^2)\ \mathrm{d}t =$$

$$= \left[t - \frac{3t^2}{20} + \frac{t^3}{150} \right]_0^5 = 5 - \frac{15}{4} + \frac{5}{6} = \frac{25}{12} \doteq 2.08 \ .$$

8.5 (transformace veličiny)

Nechť veličina X má rovnoměrné rozdělení na intervalu $(-1,2)$ a $Y = 2X^2 + 1$.

(a) Sestrojte distribuční funkci F_Y náhodné veličiny Y.

(b) Rozhodněte, zda jsou X a Y nezávislé a proč.

(c) Spočtěte $\mathrm{cov}(X,Y)$.

Řešení:

(a) Pro distribuční funkci veličiny Y máme

$$F_Y(y) = P(Y \leq y) = P(2X^2 + 1 \leq y) = P\Big(X^2 \leq \underbrace{\tfrac{y-1}{2}}_{=:u}\Big) =$$

$$= \begin{cases} P(\emptyset) = 0 & , \ u < 0 \\ P\Big(|X| \leq \sqrt{u}\Big) = P\Big(-\sqrt{u} \leq X \leq \sqrt{u}\Big) = F_X\left(\sqrt{u}\right) - F_X\left(-\sqrt{u}\right) & , \ u \geq 0 \ . \end{cases}$$

kde jsme využili toho, že F_X je spojitá, tedy nemusíme rozlišovat mezi ostrými a neostrými nerovnostmi.

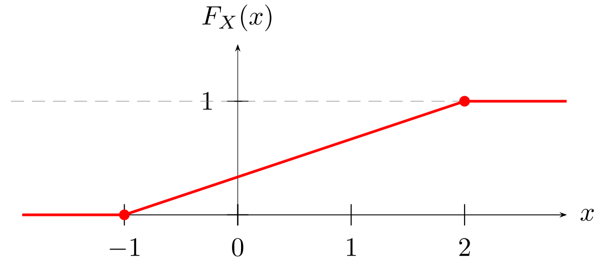
Teď si už si jen vyjádříme F_X a dosadíme. Pro veličinu $X \sim \mathrm{Ro}(-1, 2)$ je její hustota

$$f_X(x) = \begin{cases} \frac{1}{2-(-1)} = \frac{1}{3} & , \ -1 \leq x \leq 2 \\ 0 & , \ \text{jinak} \end{cases}$$

a distribuční funkce

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt = \begin{cases} 0, & x < -1 \\ \frac{x+1}{3}, & -1 \leq x \leq 2 \\ 1, & x > 2 \end{cases}$$

s grafem



Do F_X teď dosadíme argument $u = \frac{y-1}{2} \geq 0$ (tedy $y \geq 1$). Musíme však pro $\sqrt{u} \geq 0$ rozlišit jednotlivé případy hodnoty $-\sqrt{u}$:

- $0 \leq \sqrt{u} \leq 1$: Pak je $-1 \leq -\sqrt{u} \leq 0$ a tedy

$$F_Y(y) = \cdots = F_X\left(\sqrt{u}\right) - F_X\left(-\sqrt{u}\right) = \frac{\sqrt{u}+1}{3} - \frac{-\sqrt{u}+1}{3} = \frac{2}{3}\sqrt{u} = \frac{2}{3}\sqrt{\tfrac{y-1}{2}}$$

- $1 \leq \sqrt{u} \leq 2$: Pak je $-2 \leq -\sqrt{u} \leq -1$ a tedy

$$F_Y(y) = \cdots = F_X\left(\sqrt{u}\right) - F_X\left(-\sqrt{u}\right) = \frac{\sqrt{u}+1}{3} - 0 = \frac{1}{3}\sqrt{\tfrac{y-1}{2}} + \frac{1}{3}$$

- $2 \leq \sqrt{u}$: Pak je $-\sqrt{u} \leq -2$ a tedy

$$F_Y(y) = \cdots = F_X\left(\sqrt{u}\right) - F_X\left(-\sqrt{u}\right) = 1 - 0 = 1$$

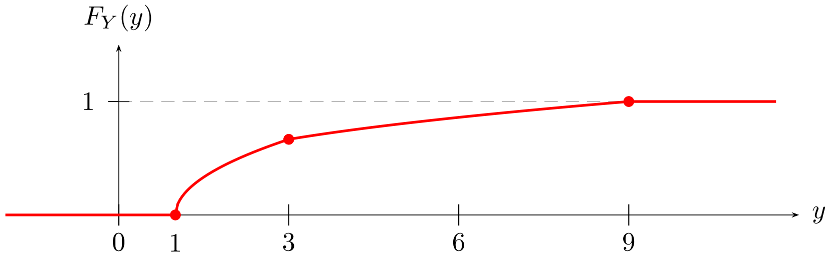
ještě přepíšeme podmínky pro y:

$$1 \leq \sqrt{u} = \sqrt{\tfrac{y-1}{2}} \leq 2 \quad \Leftrightarrow \quad 1 \leq \tfrac{y-1}{2} \leq 4 \quad \Leftrightarrow \quad 3 \leq y \leq 9$$

(zbylé podmínky jsou podobné) a celkově tak dostaneme

$$F_Y(y) = \begin{cases} 0 & , \ y < 1 \\ \frac{2}{3}\sqrt{\frac{y-1}{2}} & , \ 1 \leq y \leq 3 \\ \frac{1}{3}\sqrt{\frac{y-1}{2}} + \frac{1}{3} & , \ 3 \leq y \leq 9 \\ 1 & , \ y > 9 \end{cases}$$

s grafem



(b) Jak se dá očekávat, pokud jedna veličina závisí svými hodnotami na druhé, nejspíš nezávislé nebudou. Výjimkou je jen jeden případ a celá situaci se dá popsat takto:

Věta: Nechť X a $h(X)$ jsou obě náhodné veličiny, kde $h : \mathbb{R} \to \mathbb{R}$ je "rozumná" funkce (např. spojitá). Pak X a $h(X)$ jsou nezávislé veličiny právě jen pokud

- $h(X)$ je konstantní veličina (přesněji: ex. $c \in \mathbb{R}$, že $P\big(h(X) = c\big) = 1$).

V našem případě už víme, že $Y = h(X) = 2X^2 + 1$ konstantní není, takže veličiny X a Y jsou *závislé.*

(c) Kovarianci vypočteme ze vztahu

$$\mathrm{cov}(X,Y) = \mathrm{cov}(X, 2X^2+1) = 2 \cdot \mathrm{cov}(X, X^2) = 2 \cdot \Big[E(X \cdot X^2) - E(X) \cdot E(X^2)\Big]$$

Stačí si tedy pro $n \geq 1$ zjistit

$$E(X^n) = \int_{-\infty}^{\infty} x^n \cdot f_X(x)\ dx = \int_{-1}^{2} \frac{x^n}{3}\ dx = \left[\frac{x^{n+1}}{3(n+1)}\right]_{-1}^{2} = \frac{2^{n+1} - (-1)^{n+1}}{3(n+1)}$$

takže

$$E(X) = \frac{1}{2}, \quad E(X^2) = 1, \quad E(X^3) = \frac{5}{4}$$

a dosazením dostaneme

$$\mathrm{cov}(X,Y) = 2 \cdot \Big[E(X^3) - E(X) \cdot E(X^2)\Big] = 2 \cdot \left(\frac{5}{4} - 1 \cdot \frac{1}{2}\right) = \frac{3}{2}\ .$$

Nenulová hodnota kovariance nám nyní jen potvrzuje, že veličiny X a Y jsou závislé.

8.6 (rozdělení součtu nezávislých veličin)

Diskrétní náhodná veličina X má rovnoměrné rozdělení na množině $\{0,1\}$ a spojitá náhodná veličina Y má rovnoměrné rozdělení na intervalu $[0,1]$. Veličiny X a Y jsou nezávislé. Určete rozdělení veličiny $W = X + Y$.

Řešení:

Protože X nabývá pouze konečně mnoha hodnot, bude výhodné použít větu o úplné pravděpodobnosti:

$$F_W(t) = P(X+Y \le t) = P(X+Y \le t|X=0) \cdot P(X=0) + P(X+Y \le t|X=1) \cdot P(X=1)$$

Díky nezávislosti veličin X a Y máme

$$P(X+Y \le t|X=1) = P(1+Y \le t|X=1) = P(Y \le t-1) = F_Y(t-1)$$

a podobně

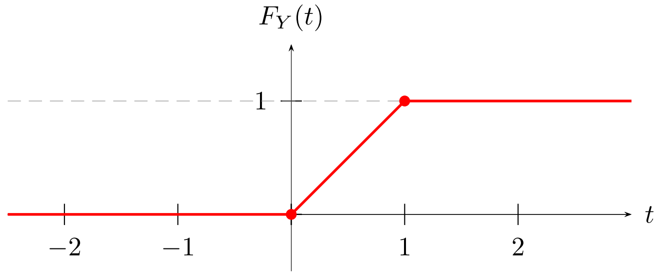
$$P(X+Y \le t|X=0) = P(Y \le t) = F_Y(t).$$

Protože ještě víme, že $P(X=0) = \frac{1}{2} = P(X=1)$, tak můžeme psát

$$F_W(t) = \frac{1}{2}\Big(F_Y(t) + F_Y(t-1)\Big)\ .$$

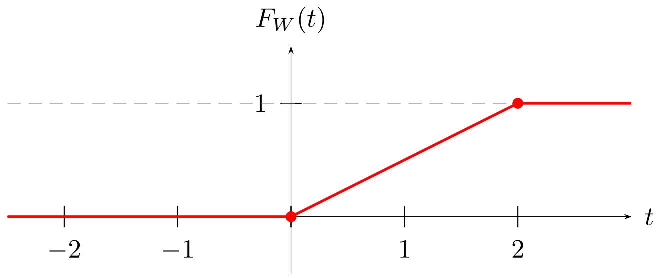
Nyní jen vyjádříme distribuční funkci veličiny Y

$$F_Y(t) = \begin{cases} 0 & ,\ t<0 \\ t & ,\ t \in \langle 0,1\rangle \\ 1 & ,\ t>1\ . \end{cases}$$



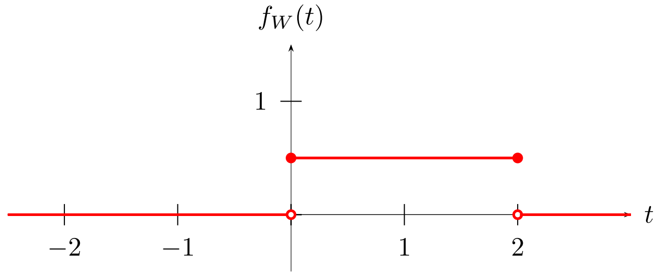
a dosadíme spolu s jejím posunutím do vzorce pro F_W, čímž získáme:

$$F_W(t) = \begin{cases} 0 & ,\ t<0 \\ \frac{t}{2} & ,\ t \in \langle 0,2\rangle \\ 1 & ,\ t>2\ . \end{cases}$$



Veličina W je tedy spojitá s rovnoměrným rozdělením na intervalu $\langle 0,2\rangle$ a hustotou:

$$f_W(t)=\begin{cases}\frac{1}{2} & ,\ t\in\langle 0,2\rangle\\ 0 & ,\ \text{jinak}\ .\end{cases}$$



Jiný způsob řešení: Na začátku jsme také mohli veličinu vyjádřit jako směs, kde by jedna ze složek odpovídala případu $X=0$ a druhá případu $X=1$ (tj. šlo by o zúžení veličiny W za daných podmínek). Pro $A=X^{-1}(\{0\})$ a $c=P(A)=P(X=0)=\frac{1}{2}$ tak máme

$$W|_A=Y|_A,\quad W|_{\overline{A}}=Y+1|_{\overline{A}}$$

a pro vyjádření směsi tak dostáváme

$$W=\mathrm{Mix}_c(W|_A,W|_{\overline{A}})=\mathrm{Mix}_{\frac{1}{2}}(Y|_A,Y+1|_{\overline{A}})$$

Zbytek postupu by pak už ale byl stejný.

Poznámky k normálnímu rozdělení:
Veličina X má normální rozdělení $N(\mu,\sigma^2)$ (kde $\mu\in\mathbb{R}$ a $\sigma>0$), jestliže má hustotu

$$f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad \text{pro}\quad x\in\mathbb{R}\ .$$

Je to tedy spojité rozdělení, $E(X)=\mu$, $D(X)=\sigma^2$ a oborem hodnot veličiny X je celá reálná osa. Všimněme si ještě, že hustota f_X je symetrická vzhledem ke středu μ a proto platí $F_X(\mu)=\frac{1}{2}$.

Toto rozdělení je limitním rozdělením, které aproximuje součty nezávislých stejně (nebo podobně) rozdělených veličin. Typicky se tedy objevuje u veličin, jejichž hodnoty jsou ovlivněny mnoha drobnými odchylkami (např. u chyb měření, výšky člověka apod.)

U zmíněné výšky člověka (která může být samozřejmě jen kladná) nebo u veličin s hodnotami omezenými na nějaký interval, je přesto použití normálního rozdělení (které může nabývat libovolných hodnot) přiměřené. Je to tím, že u dané veličiny Y předpokládáme aproximaci pomocí normálního rozdělení obvykle jen ve vhodném okolí kolem střední hodnoty $\mu:=E(Y)$. Je to podobná situace, jako když aproximujeme funkci pomocí jejího Taylorova polynomu v okolí daného bodu.

Přesněji to vystihuje toto tvrzení:

Věta: Nechť Y je veličina s hustotou f_Y, střední hodnotou μ a rozptylem $\sigma^2\neq 0$. Nechť $X\sim N(\mu,\sigma^2)$. Jestliže se hustoty f_X a f_Y rovnají na nějakém intervalu $(a,b)\subseteq\mathbb{R}$ takovém, že $\mu\in(a,b)$ a pokud $F_Y(\mu)=\frac{1}{2}$, pak

$$F_Y(t)=F_X(t)\quad\text{pro všechna}\quad t\in(a,b)\ .$$

Důležité vlastnosti normálního rozdělení:

- $X\sim N(\mu,\sigma^2)\ \Rightarrow\ Y=\frac{X-\mu}{\sigma}\sim N(0,1)$ (je to tzv. normované normální rozdělení s hodnotami v tabulkách) dist. funkce pro $N(0,1)$ se značí Φ.

  V tomto případě pak máme $F_X(t)=P(X\leq t)=P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=Y}\leq\frac{t-\mu}{\sigma}\Big)=\Phi\Big(\frac{t-\mu}{\sigma}\Big)$ pro všechna $t\in\mathbb{R}$.
- hustota $f_{N(0,1)}$ je symetrická $\Rightarrow$ $\Phi(t)+\Phi(-t)=1$ pro všechna $t\in\mathbb{R}$.
- Nechť $X_i\sim N(\mu_i,\sigma_i^2)$, pro $i=1,2$, jsou nezávislé. Pak $X_1+X_2\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ (tj. speciálně součet nezávislých normálních rozdělení je zase normální.)

8.7 (normální rozdělení)

Rychlost aut v úseku, kde je omezení na maximální povolenou rychlost 50 km/hod, je náhodná veličina X, která má normální rozdělení $N(\mu, \sigma^2)$. Z naměřených hodnot vyplývá, že

$$P(X > 60) = 0.45 \quad \text{a} \quad P(X > 70) = 0.2\ .$$

Určete

(a) parametry μ a σ^2;

(b) pravděpodobnost $P(X < 50)$.

Řešení:

(a) Máme

$$\begin{aligned} 0.45 = P(X > 60) = 1 - P(X \leq 60) = 1 - F_X(60) &= 1 - \Phi\Big(\frac{60-\mu}{\sigma}\Big) \\ \Phi\Big(\frac{60-\mu}{\sigma}\Big) &= 1 - 0.45 = 0.55 \\ \frac{60-\mu}{\sigma} &= \Phi^{-1}(0.55) \\ 60 &= \Phi^{-1}(0.55)\cdot\sigma + \mu \end{aligned}$$

a podobně

$$\begin{aligned} 0.2 = P(X > 70) = 1 - P(X \leq 70) = 1 - F_X(70) &= 1 - \Phi\Big(\frac{70-\mu}{\sigma}\Big) \\ \Phi\Big(\frac{70-\mu}{\sigma}\Big) &= 1 - 0.2 = 0.8 \\ \frac{70-\mu}{\sigma} &= \Phi^{-1}(0.8) \\ 70 &= \Phi^{-1}(0.8)\cdot\sigma + \mu \end{aligned}$$

Hodnoty funkce Φ^{-1} (tj. kvantilové funkce) zjistíme opět pomocí tabulek (nebo softwaru). Zde máme $\Phi^{-1}(0.55) \doteq 0.126$ a $\Phi^{-1}(0.8) \doteq 0.842$. Máme tedy (lineární) soustavu rovnic

$$\begin{aligned} 60 &= 0.126\cdot\sigma + \mu \\ 70 &= 0.842\cdot\sigma + \mu \end{aligned}$$

Odsud odečtením získáme $70 - 60 = (0.842 - 0.126)\cdot\sigma$ tedy

$$\begin{aligned} \sigma &= \frac{10}{0.716} \doteq 13.97 \\ \mu &= 70 - 0.842\cdot\sigma \doteq 70 - 0.842\cdot 13.97 \doteq 58.24 \end{aligned}$$

(b) Z vypočítaných μ a σ dostaneme

$$P(X < 50) = F_X(50) = \Phi\Big(\frac{50-\mu}{\sigma}\Big) = \Phi\Big(\frac{50-58.24}{13.97}\Big) \doteq$$

$$\doteq \Phi\Big(-0.59\Big) = 1 - \Phi(0.59) \doteq 1 - 0.722 = 0.278\ .$$