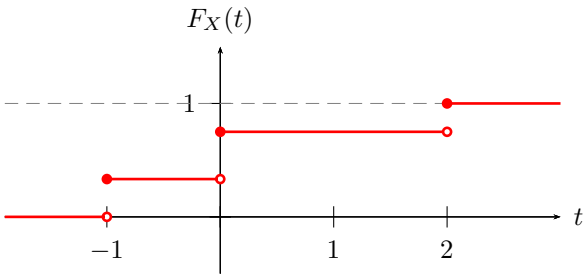
14. cvičení z STP

20. - 24. května 2019

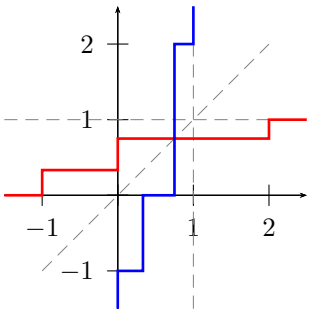
Poznámky ke kvantilům:

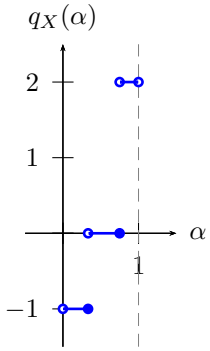
Pro náhodnou veličinu X a pravděpodobnost $\alpha \in (0,1)$ často potřebujeme najít $t \in \mathbb{R}$, že $P(X \leq t) = \alpha$, tj. $F_X(t) = \alpha$. Takové t obecně nemusí existovat (např. kdy F_X má skoky) nebo nemusí být určeno jednoznačně (když F_X je místy konstantní). Například si vezměme tuto distribuční funkci:



Chtěli bychom tedy ideálně mít inverzní funkci k F_X, která ale obecně neexistuje. Přesto můžeme něco podobného, tzv. *kvantilovou funkci* $q_X : (0,1) \to \mathbb{R}$, definovat (díky tomu, že F_X je neklesající) a to následujícím způsobem:

- graf F_X doplníme na "souvislou čáru", tj. případné skoky funkce F_X nahradíme spojitou svislou úsečkou,
- tento útvar převrátíme podle osy 1. a 3. kvadrantu (tj. podle přímky "$x = y$"),
- tam, kde převrácený útvar není funkcí (tj. obsahuje svislé čáry) tyto úseky odstraníme a nahradíme jedinou hodnotou, a sice limitou zleva (a případné krajní úseky v bodech 0 a 1 odstraníme úplně, protože tam se kvantil q_X nedefinuje)
- výsledným útvarem si pak definujeme graf funkce q_X.





Jak je tedy vidět, grafy funkcí F_X a q_X (po doplnění na souvislé čáry) si budou navzájem zrcadlovými obrazy (vhledem k ose $x = y$). Takováto definice kvantilu je sice názorná, ale chtělo by to i explicitní popis. Platí:

- $q_X(\alpha) = \min\{t \in \mathbb{R} \mid F_X(t) \geq \alpha\}$ pro všechna $\alpha \in (0,1)$

- $P\Big(X \le q_X(\alpha)\Big) = F_X\Big(q_X(\alpha)\Big) \ge \alpha$ pro všechna $\alpha \in (0,1)$
- $P\Big(X < q_X(\alpha)\Big) \le \alpha$ pro všechna $\alpha \in (0,1)$
- q_X je neklesající a zleva spojitá funkce
- Jestliže je F_X spojitá a ostře rostoucí, pak q_X je inverzní funkcí k F_X.
  V tom případě pak pro všechna $\alpha \in (0,1)$ platí:
  - $q_X(\alpha) = (F_X)^{-1}(\alpha)$
  - $P\Big(X \le q_X(\alpha)\Big) = \alpha$

Poznámky k empirickému rozdělení:
Nechť $x_1 \le \cdots \le x_n$ jsou naměřené hodnoty (veličiny X). Pro ně si můžeme přirozeně definovat *empirickou* náhodnou veličinu Emp s diskrétním rozdělením, oborem hodnot

$$A = \{a \in \mathbb{R} \mid a = x_i \ \text{ pro nějaké } \ i\}$$

a jejich pravděpodobnostmi

$$P(\mathrm{Emp} = a) = \frac{\text{“ počet výskytů } a \text{ mezi hodnotami } x_1, \ldots, x_n \text{”}}{n} \ .$$

Když si k této veličině zjistíme distribuční funkci, dostaneme známou *empirickou distribuční funkci*:

$$F_{\mathrm{Emp}}(t) = P(\mathrm{Emp} \le t) = \frac{\#\{i \mid x_i \le t\}}{n}$$

Od ní si pak vytvoříme kvantilovou funkci q_{Emp}, která má tvar

$$q_{\mathrm{Emp}}(\alpha) = \min\{t \in \mathbb{R} \mid F_{\mathrm{Emp}}(t) \ge \alpha\} = \min\{x_j \mid F_{\mathrm{Emp}}(x_j) \ge \alpha\} \ .$$

a nakonec se dá přepsat jako

$$q_{\mathrm{Emp}}(\alpha) = x_{\lceil n\alpha \rceil} \quad \text{pro} \quad \alpha \in (0,1)$$

kde $\lceil u \rceil$ je horní celá část z $u \in \mathbb{R}$, tj. zaokrouhlení desetinných čísel nahoru. Speciální hodnoty se pak jmenují

- 1. kvartil $= q_{\mathrm{Emp}}(\frac{1}{4}) = x_{\lceil \frac{n}{4} \rceil}$
- 2. kvartil $= q_{\mathrm{Emp}}(\frac{2}{4}) = x_{\lceil \frac{n}{2} \rceil}$ (tzv. medián)
- 3. kvartil $= q_{\mathrm{Emp}}(\frac{3}{4}) = x_{\lceil \frac{3n}{4} \rceil}$

Podobným způsobem se kvartily definují pro libovolnou veličinu X (jako $q_X(\frac{1}{4})$, $q_X(\frac{1}{2})$ a $q_X(\frac{3}{4})$).
Pro libovolnou veličinu X (a speciálně pro $X = \mathrm{Emp}$) platí:

$$P\Big(\ X \ \le \ \textit{1. kvartil}\ \Big) \ge \frac{1}{4}$$

$$P\Big(\ \textit{1. kvartil} \ \le \ X \ \le \ \textit{3. kvartil}\ \Big) \ge \frac{1}{2} \ .$$

$$P\Big(\ X \ \ge \ \textit{3. kvartil}\ \Big) \ge \frac{1}{4}$$

- -

Příklad 14.1 *Uvažujme následující data:*

(1) počty výskytů jistého druhu rostliny na ploše $1\,\mathrm{m}^2$*:*

$$0,\quad 2,\quad 1,\quad 4,\quad 4,\quad 5,\quad 2,\quad 3,\quad 7$$

(2) časy (v sekundách) mezi impulzy v mozku:

$$4.25,\quad 0.65,\quad 1.35,\quad 0.20,\quad 0.55,\quad 6.63,\quad 1.38,\quad 0.22,\quad 0.27$$

(3) venkovní teploty naměřené v různých letech při pravidelné podzimní akci:

$$8.07,\quad 19.23,\quad 9.27,\quad 5.71,\quad 12.62,\quad 11.24,\quad 11.92,\quad 17.30,\quad 14.87$$

Nakreslete pro tato data

(a) histogramy

(b) boxploty

(c) empirickou distribuční funkci

a odhadněte, z jakého rozdělení mohou tato data pocházet.

Řešení:
Histogram (pro četnosti): Naměřená data si rozdělíme do disjunktních intervalů I_i (stejné délky) pro $i = 1, \ldots, k$, které na sebe budou navazovat. Nad I_i nakreslíme sloupec výšky m_i, která znamená četnost dat, jež spadnou do I_i. Abychom z histogramu něco mohli vyčíst a uměli ho (ručně) nakreslit, volíme "rozumný" počet sloupců (např. něco mezi 5 a 15).

Boxplot (neboli krabicový graf): Na rozdíl od histogramu je vždy definován stejně. Krajní vousy ("whiskers") jsou dány krajními naměřenými hodnotami a krabice ("box") uprostřed je pak určena hodnotami jednotlivých kvartilů.

Počet měření je zde ve všech případech stejný: $n = 9$. Při uspořádaných datech $x_1 \leq \cdots \leq x_9$ tak budou hodnoty kvartilů tyto:

- 1. kvartil $= x_{\lceil \frac{9}{4} \rceil} = x_3$
- medián $= x_{\lceil \frac{9}{2} \rceil} = x_5$
- 3. kvartil $= x_{\lceil \frac{3 \cdot 9}{4} \rceil} = x_7$

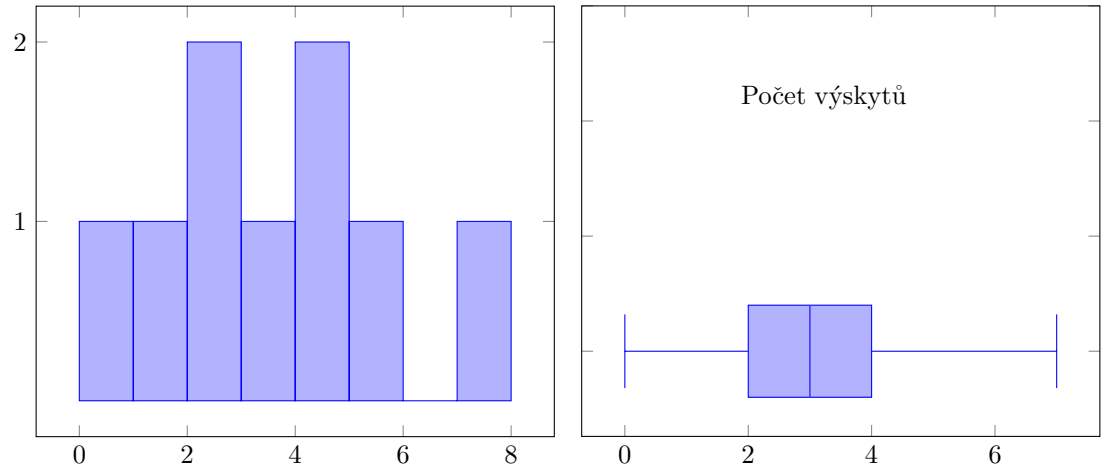
Medián je (v rámci uspořádání podle indexu) tedy přibližně uprostřed naměřených hodnot a podobně je to s okolními kvartily. Data si tudíž před výpočtem vždy uspořádáme.

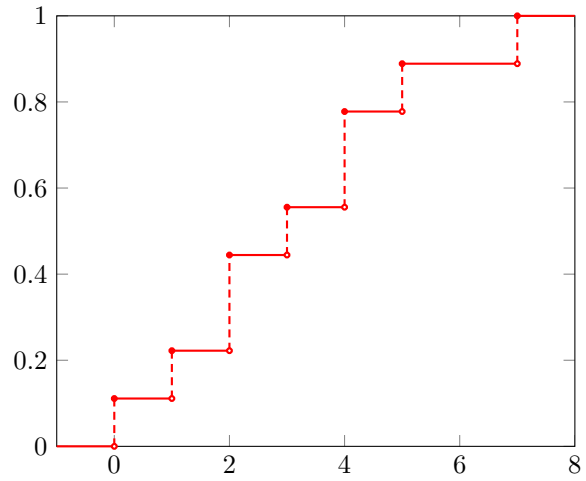
(1) Uspořádaná data:

$0,$	$1,$	$2,$	$2,$	$3,$	$4,$	$4,$	$5,$	7
x_1		$1.kvar.$		$med.$		$3.kvar.$		x_n

Rozdíl mezi největší a nejmenší hodnotou je $x_n - x_1 = 7 - 0 = 7$. Tuto délku tedy budeme potřebovat pokrýt několika disjunktními intervaly a protože se zde jedná o diskrétní veličinu (počty

výskytů), bude vhodné si zvolit šířku sloupce rovnou 1. Intervaly pak budou $\langle 0,1), \langle 1,2), \ldots, \langle 7,8)$.



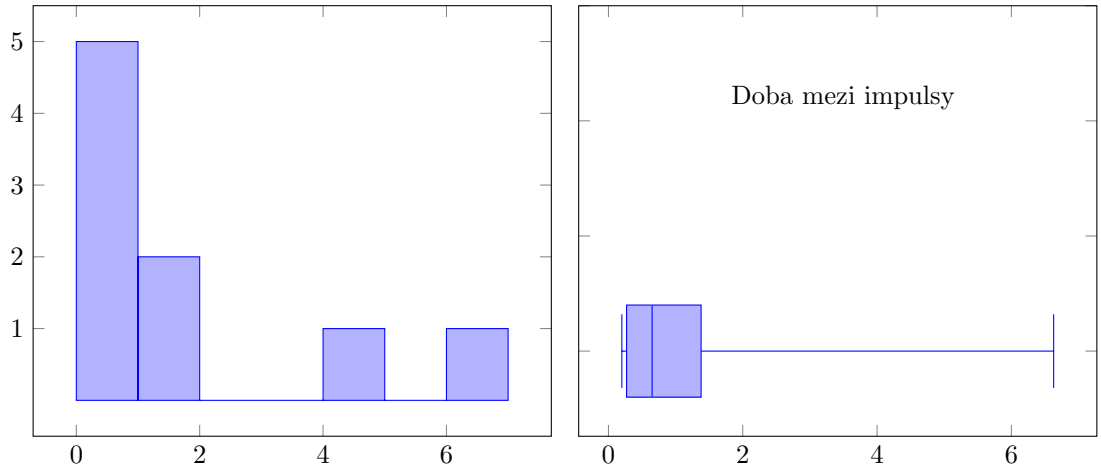


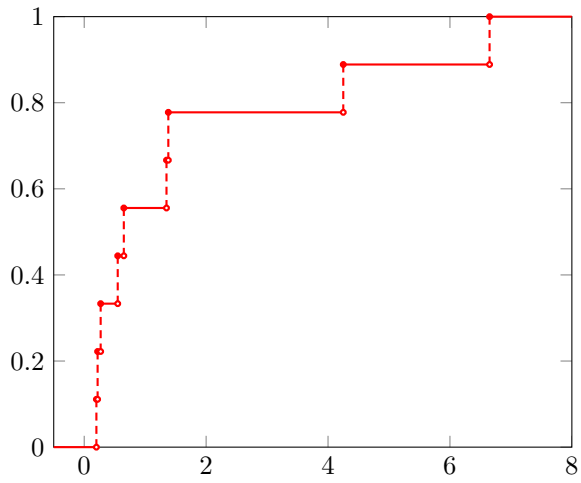
Vzhledem k popisu dat (počty výskytů na dané ploše) to vypadá na Poissonovo rozdělení. Tomu také zhruba odpovídají i grafická znázornění (histogram, boxplot, emp. distr. funkce).

(2) Uspořádaná data:

0.20,	0.22,	0.27,	0.55,	0.65,	1.35,	1.38,	4.25,	6.63
x_1		1.*kvar.*		*med.*		3.*kvar.*		x_n

Rozdíl mezi největší a nejmenší hodnotou je $x_n - x_1 = 6.63 - 0.2 = 6.42$. Tuto délku budeme zase potřebovat pokrýt několika disjunktními intervaly. Zkusíme si opět vzít šířku sloupce rovnou 1. Intervaly si pro změnu zvolíme jako $(0,1\rangle, (1,2\rangle, \ldots, (6,7\rangle$. Výběr toho, do kterého z intervalů přiřadíme dělicí body, není podstatný. Zde jsme si to takto zvolili čistě jen proto, že hodnoty čekací doby jsou vždy nenulové (tj. první interval by ideálně neměl začínat nulou).



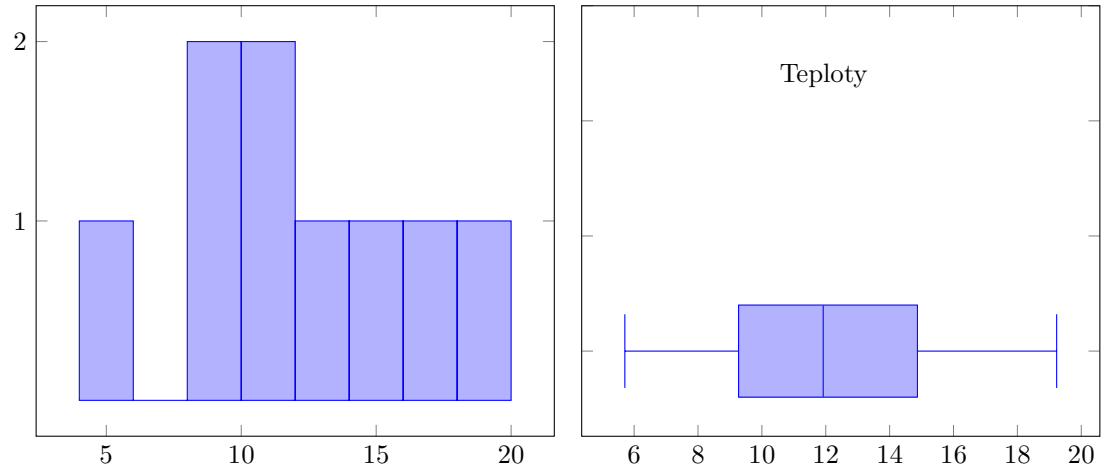


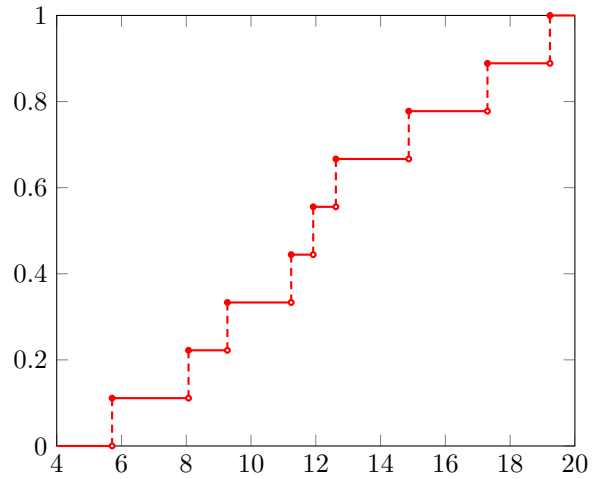
Vzhledem k popisu dat (doba čekání na další událost) to vypadá na exponenciální rozdělení. Tomu také zhruba odpovídají i grafická znázornění, kde boxplot je hodně posunutý doleva a empirická distribuční funkce připomíná exponenciálu.

(3) Uspořádaná data:

5.71,	8.07,	9.27,	11.24,	11.92,	12.62,	14.87,	17.30,	19.23
x_1		$1.kvar.$		$med.$		$3.kvar.$		x_n

Rozdíl mezi největší a nejmenší hodnotou je $x_n - x_1 = 19.23 - 5.71 = 13.52$ a tuto délku potřebujeme pokrýt několika disjunktními intervaly. Tady se nabízí vzít si větší (ideálně celočíselnou šířku), takže zkusíme šířku sloupce rovnou 2. Intervaly si zvolíme např. $\langle 4, 6), \langle 6, 8), \ldots, \langle 18, 20)$.





Vzhledem k popisu dat (hodnota ovlivněná mnoha malými výkyvy) to vypadá na normální rozdělení. Tomu také zhruba odpovídají i grafická znázornění (celkem symetrický boxplot i emp. distribuční funkce).

Příklad 14.2 *Délka hrany krychle je náhodná veličina* $X \sim \mathrm{Ro}(1,2)$. *Určete distribuční funkci náhodné veličiny* Y *popisující plochu povrchu této krychle.*

Řešení:
Máme veličiny

$X =$ *"délka hrany krychle"*

$Y =$ *"plocha povrchu krychle"*

takže $Y = 6 \cdot X^2$ a pro distribuční funkci náhodné veličiny Y dostáváme

$$F_Y(y) = P(Y \le y) = P(6X^2 \le y) = P(X^2 \le \tfrac{y}{6}) = \begin{cases} P\left(|X| \le \sqrt{\tfrac{y}{6}}\right) = F_X\left(\sqrt{\tfrac{y}{6}}\right) & , \ y \ge 0 \\ P(\emptyset) = 0 & , \ y < 0 \ . \end{cases}$$

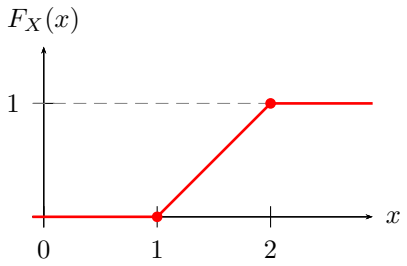
kde jsme využili toho, že obor hodnot pro X je $\langle 1,2\rangle$, tedy $X \geq 0$ a speciálně tak platí, že $|X| = X$.

Teď si už si jen vyjádříme F_X a dosadíme:

Pro veličinu $X \sim \mathrm{Ro}(1,2)$ je její hustota $f_X(x) = \begin{cases} 1, & 1 \leq x \leq 2 \\ 0, & \text{jinak} \end{cases}$ a distribuční funkce

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

s grafem



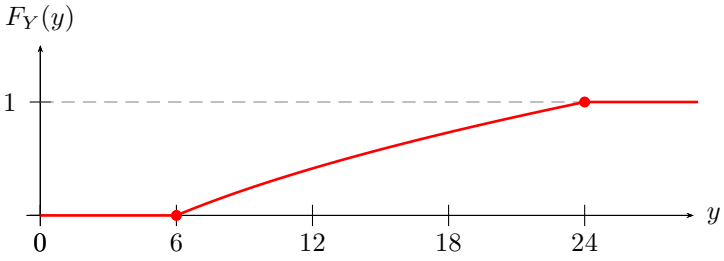
Do F_X (správně!) dosadíme $x = \sqrt{\frac{y}{6}}$ (pro $y \geq 0$) a přepíšeme podmínky pro y:

$$1 \leq \sqrt{\tfrac{y}{6}} \leq 2 \quad \Leftrightarrow \quad 1 \leq \tfrac{y}{6} \leq 4 \quad \Leftrightarrow \quad 6 \leq y \leq 24$$

(zbylé podmínky jsou podobné) a dostaneme tak

$$F_Y(y) = \begin{cases} 0, & y < 6 \\ \sqrt{\frac{y}{6}} - 1, & 6 \leq y \leq 24 \\ 1, & y > 24 \end{cases}$$

s grafem



Příklad 14.3 *Nechť* $X \sim \mathrm{Ro}(0,2)$ *a* $Y = X^2 + 1$.

(a) Sestrojte distribuční funkci náhodné veličiny Y.

(b) Spočtěte $\mathrm{cov}(X,Y)$.

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

Řešení:

(a) Distribuční funkce F_Y se dá získat pomocí distribuční funkce F_X. Před výpočtem si ještě uvědomme, že $X \geq 0$ (protože její obor hodnot je $\langle 0, 2\rangle$). Speciálně tedy $|X| = X$. Pro distribuční funkci náhodné veličiny Y máme

$$F_Y(y) = P(Y \leq y) = P(X^2 + 1 \leq y) =$$

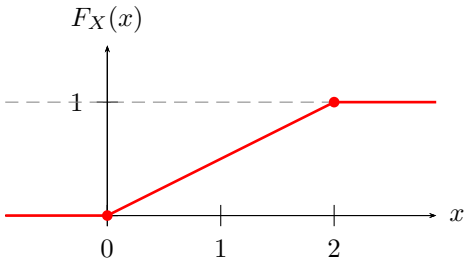
$$= P(X^2 \leq y - 1) = \begin{cases} P\left(|X| \leq \sqrt{y-1}\,\right) = F_X\left(\sqrt{y-1}\,\right) & , \; y - 1 \geq 0 \\ P(\emptyset) = 0 & , \; y - 1 < 0 \; . \end{cases}$$

Teď už si stačí jen vyjádřit F_X a dosadit.

Pro veličinu $X \sim \mathrm{Ro}(0, 2)$ je její hustota $f_X(x) = \begin{cases} \frac{1}{2}, & 0 \leq x \leq 2 \\ 0, & \text{jinak} \end{cases}$ a distribuční funkce

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

s grafem



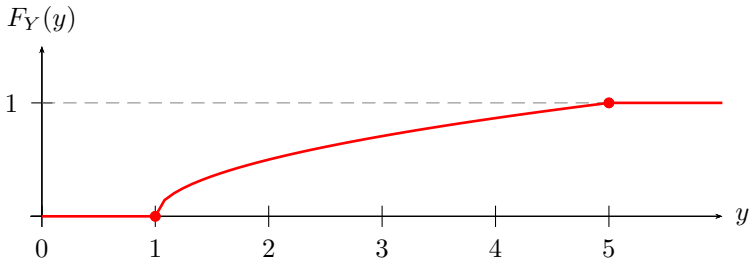
Do F_X (správně!) dosadíme $x = \sqrt{y-1}$ (pro $y - 1 \geq 0$) a přepíšeme podmínky pro y:

$$0 \leq \sqrt{y-1} \leq 2 \quad \Leftrightarrow \quad 0 \leq y - 1 \leq 4 \quad \Leftrightarrow \quad 1 \leq y \leq 5$$

(zbylé podmínky jsou podobné) a dostaneme tak

$$F_Y(y) = \begin{cases} 0, & y < 1 \\ \frac{\sqrt{y-1}}{2}, & 1 \leq y \leq 5 \\ 1, & y > 5 \end{cases}$$

s grafem



(b) Kovarianci vypočteme ze vztahu

$$\operatorname{cov}(X,Y) = E(XY) - E(X)\cdot E(Y)$$

Máme

$$E(XY) = E\big(X(X^2+1)\big) = E(X^3+X) = E(X^3) + E(X)$$

a

$$E(Y) = E(X^2+1) = E(X^2) + 1$$

Stačí si tedy pro $n \geq 1$ zjistit

$$E(X^n) = \int_{-\infty}^{\infty} x^n \cdot f_X(x)\ dx = \int_0^2 \frac{x^n}{2}\ dx = \left[\frac{x^{n+1}}{2(n+1)}\right]_0^2 = \frac{2^n}{n+1}$$

a dosazením dostaneme

$$\operatorname{cov}(X,Y) = \big(2+1\big) - 1\cdot\left(\frac{4}{3}+1\right) = \frac{2}{3}\ .$$

(c) Protože $\operatorname{cov}(X,Y) \neq 0$, jsou veličiny X a Y závislé. Toto zjištění ovšem můžeme udělat i bez výpočtu kovariance:

Veličiny X a Y jsou funkčně propojené, takže stačí najít podmínky, které naráz nemůžou splnit, ale jednotlivě, s nenulovými pravděpodobnostmi, ano. Z předchozích úprav už víme, že pro $1 < y < 5$ platí

$$Y \leq y \quad\Leftrightarrow\quad X^2+1 \leq y \quad\Leftrightarrow\quad X \leq \sqrt{y-1}$$

a pravděpodobnosti těchto jevů jsou (z tvaru F_X a F_Y) ostře mezi 0 a 1. Takže např. z volby $y=2$ dostaneme, že

$$Y \leq 2 \quad\Leftrightarrow\quad X \leq 1$$

takže

$$P(\underbrace{Y\leq 2,\ X>1}_{\emptyset}) = 0 \neq \underbrace{P\,(Y\leq 2)}_{F_Y(2)=0.5}\cdot\underbrace{P\,(X>1)}_{1-F_X(1)=0.5}$$

z čehož plyne, že veličiny X a Y jsou závislé.

Poznámka: 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 $P\big(h(X)=c\big)=1$).

Příklad 14.4 *Diskrétní náhodný vektor (X,Y) má sdružené pravděpodobnosti dány tabulkou:*

	$Y=0$	$Y=1$	$Y=2$
$X=1$	1/8	0	1/8
$X=3$	0	1/4	1/4
$X=4$	1/8	1/8	0

(a) Určete pravděpodobnost $P(X\cdot Y \geq 2.5)$.

*(b) Pokud X **a** Y **jsou závislé**, popište rozdělení náhodného vektoru (X', Y') se stejnými marginálními rozděleními, jehož složky **jsou nezávislé**.*
*Pokud X **a** Y **jsou nezávislé**, popište rozdělení náhodného vektoru (X', Y') se stejnými marginálními rozděleními, jehož složky **jsou závislé**.*

Řešení:
(a) Máme

$$X \cdot Y \geq 2.5 \quad \Leftrightarrow \quad (X,Y) \in \{\ (3,1),\ (3,2),\ (4,1),\ (4,2)\ \}$$

a tedy

$$P\left(X \cdot Y \geq 2.5\right) = P\big(X=3,\ Y=1\big) + P\big(X=3,\ Y=2\big) + P\big(X=4,\ Y=1\big) + P\big(X=4,\ Y=2\big) =$$

$$= \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + 0 = \frac{5}{8}\ .$$

(b) Marginální (diskrétní) rozdělení (tj. rozdělení jednotlivých složek vektoru) získáme pro jednotlivé hodnoty sečtením pravděpodobností v řádcích (pro X) a sloupcích (pro Y) naší tabulky:

	$Y=0$	$Y=1$	$Y=2$	$P(X=i)$
$X=1$	1/8	0	1/8	1/4
$X=3$	0	1/4	1/4	1/2
$X=4$	1/8	1/8	0	1/4
$P(Y=j)$	1/4	3/8	3/8	

Protože např.

$$\underbrace{P(X=4,\ Y=2)}_{=0} \neq \underbrace{P(X=4)\cdot P(Y=2)}_{=\frac{1}{4}\cdot\frac{3}{8}}$$

jsou X a Y **závislé**. (Nejjednodušší k tomu účelu je najít si v tabulce právě nulovou hodnotu.)

Nechť (X', Y') je nyní náhodný vektor s **nezávislými** složkami, které mají stejné marginální rozdělení jako má vektor (X, Y) tedy

$$P(X'=i) = P(X=i) \quad \text{a} \quad P(Y'=j) = P(Y=j) \quad \text{pro všechna} \quad i,j \in \mathbb{R}\ .$$

Pak sdružené pravděpodobnosti vektoru (X', Y') tedy platí, že

$$P(X'=i,\ Y'=j) = P(X'=i)\cdot P(Y'=j) = P(X=i)\cdot P(Y=j)$$

a můžeme je tak popsat následující tabulkou:

	$Y'=0$	$Y'=1$	$Y'=2$	$P(X'=i)$
$X'=1$	1/16	3/32	3/32	1/4
$X'=3$	1/8	3/16	3/16	1/2
$X'=4$	1/16	3/32	3/32	1/4
$P(Y'=j)$	1/4	3/8	3/8	

Abychom ukázali, jak řešit druhou možnost, předpokládejme nyní naopak, že takovouto tabulku pro nezávislé složky (X', Y') dostaneme. Speciálně vidíme, že všechny sdružené pravděpodobnosti v tabulce budou nenulové. Jak teď vyrobit nějaké jiné rozdělení (X'', Y''), se závislými složkami, které budou mít

stejné součty v řádcích a sloupcích? Stačí si vybrat dva řádky a dva sloupce (pro přehlednost např. první a druhý)

	$Y' = j_1$	$Y' = j_2$	
$X' = i_1$	a	b	
$X' = i_2$	c	d	

a tam, kde se protínají (celkově tedy jen ve 4 buňkách), udělat úpravu o hodnotu ε a tím vytvořit tabulku pro (X'', Y'') (zbylé hodnoty v tabulce necháme stejné):

	$Y'' = j_1$	$Y'' = j_2$	
$X'' = i_1$	$a - \varepsilon$	$b + \varepsilon$	
$X'' = i_2$	$c + \varepsilon$	$d - \varepsilon$	

Přitom pochopitelně musíme dodržet, aby upravené hodnoty byly nezáporné (protože to musí být pravděpodobnosti), takže máme toto omezení:

$$-\min(b, c) \leq \varepsilon \leq \min(a, d) \ .$$

V našem případě tedy

$$-3/32 = -\min(3/32,\ 1/8) \leq \varepsilon \leq \min(1/16,\ 3/16) = 1/16 \ .$$

a my si můžeme zvolit např. $\varepsilon = 1/16$, čímž si v první buňce vyrobíme nulu.

Protože jsme změnou hodnot v buňkách alespoň na jednom místě (dokonce však na čtyřech místech) porušili původní rovnosti pro nezávislé složky X' a Y', tedy konkrétně

$$P(X'' = i_1, Y'' = j_1) = a - \varepsilon \neq a = P(X'' = i_1) \cdot P(Y'' = j_1)$$

budou veličiny X'' a Y'' závislé.