9. cvičení z PST

11. - 15. dubna 2022

Poznámky ke kovarianci a korelaci: Náhodné veličiny (jako funkce na pravděpodobnostním prostoru Ω) tvoří přirozeně (reálný) vektorový prostor (kde ještě navíc dvě veličiny budeme pokládat za totožné, pokud se rovnají s pravděpodobností 1). Na vektorovém *pod*prostoru veličin s konečnou střední hodnotou a konečným rozptylem pak můžeme přirozeným způsobem zavést skalární součin jako

$$\langle X|Y\rangle := E(X \cdot Y)$$

Díky němu můžeme přirozeně zavést *normu* $\|X\|$ (neboli "délku" vektoru X) jako

$$\|X\| := \sqrt{\langle X|X\rangle} = \sqrt{E(X^2)}\ .$$

Mimo jiné si všimněme, že pro X je $\mathrm{var}(X) = \|X - E(X)\|^2$, takže platí

$$\|norm(X)\| = \left\| \frac{X - EX}{\sqrt{\mathrm{var}(X)}} \right\| = \frac{\|X - EX\|}{\sqrt{\mathrm{var}(X)}} = 1$$

neboli $norm(X)$ má délku skutečně znormovanou na hodnotu 1.

Skalární součin nám dále umožňuje měřit také úhel mezi dvěma vektory. Pro veličiny X a Y je užitečné znát, jestli jejich výchylky vůči středním hodnotám (tj. veličiny $X - EX$ a $Y - EY$) mají podobné chování (tj. jestli korelují). Zavádíme proto korelaci mezi veličinami X a Y jako kosinus úhlu α mezi vektory $X - EX$ a $Y - EY$, tedy

$$\mathrm{corr}(X,Y) := \frac{\langle X - EX \mid Y - EY\rangle}{\|X - EX\| \cdot \|Y - EY\|} = \cdots = \frac{E(XY) - E(X) \cdot E(Y)}{\sqrt{\mathrm{var}(X) \cdot \mathrm{var}(Y)}}\ .$$

(Veličiny $X - EX$ a $Y - EY$ mají nulovou střední hodnotu).

A kromě toho máme:

$$\mathrm{cov}(X,Y) := \langle X - EX \mid Y - EY\rangle = \cdots = E(XY) - E(X) \cdot E(Y)\ .$$

Praktický důsledek korelace:

Pokud máme dvě veličiny X a Y takové, že

$$X - EX \geq 0 \ \Leftrightarrow\ Y - EY \geq 0 \quad \Big(\text{což implikuje, že} \quad (X - EX)(Y - EY) \geq 0\Big)$$

pak je $\mathrm{corr}(X,Y) \geq 0$.

Obdobně platí: Jestliže

$$X - E(X) \geq 0 \ \Leftrightarrow\ Y - E(Y) \leq 0$$

pak je $\mathrm{corr}(X,Y) \leq 0$.

Ačkoliv zpětné implikace v obou případech neplatí, přesto nám korelace umožňuje nějakým způsobem zachytit jistou míru kauzální závislosti dvou veličin.

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

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

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

Definice: Náhodný vektor (X, Y) má *diskrétní rozdělení* $\Leftrightarrow$ existuje $A \subseteq \mathbb{R}^2$, která je konečná nebo spočetná a taková, že $P((X, Y) \in A) = 1$. (Tedy vektor má nejvýše spočetně mnoho "zajímavých" hodnot.)
V tomto případě pak pro *sdruženou distribuční funkci* máme

$$F_{(X,Y)}(a, b) = P(X \leq a, Y \leq b) = \sum_{\substack{u \leq a \\ t \leq b}} P(X = u, Y = t) \ .$$

Věta: Nechť náhodný vektor (X, Y) má *diskrétní* rozdělení. Pak:

X a Y jsou nezávislé $\Leftrightarrow P(X = i, Y = j) = P(X = i) \cdot P(Y = j)$ pro všechna $i, j \in \mathbb{R}$.

(Srovnejte to s obecně NEdiskrétním případem, kdy je nezávislost popsána jen nerovnostmi:

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

Příklad 9.1 *Sdružené pravděpodobnosti náhodných veličin X a Y jsou dány následující tabulkou:*

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

(a) Jaká jsou jejich marginální rozdělení? Určete pravděpodobnost $P(X + Y > 1)$.

*(b) Jsou veličiny X a Y nezávislé? Zdůvodněte. Pokud X **a** Y **jsou závislé**, popište (jednoznačně určené) rozdělení náhodného vektoru (X', Y') se stejnými marginálními rozděleními jako (X, Y), jehož složky **jsou nezávislé**.*

(c) Určete střední hodnotu $E(X \cdot Y)$. Určete varianční a korelační matici.

Řešení:
Na začátku bychom si měli pro pořádek ještě ověřit, že součet všech pravděpodobností v tabulce je $= 1$ (pokud by byl např. < 1, pak nemáme úplnou informaci o rozdělení a nemůžeme dál pokračovat).

(a) U veličin X a Y předpokládáme obory hodnot určené tabulkou. Pak máme

$$X + Y > 1 \quad \Leftrightarrow \quad (X, Y) \in \{(0, 2), (1, 1), (1, 2)\}$$

a tedy

$$P\left(X + Y > 1\right) = P\Big((X, Y) \in \{(0, 2), (1, 1), (1, 2)\}\Big) = \frac{1}{4} + \frac{1}{8} + 0 = \frac{3}{8} \ .$$

Marginální (česky: okrajová) rozdělení náhodného vektoru (X, Y) jsou rozdělení jeho jednotlivých složek, tedy veličin X a Y. Vektor (X, Y) má diskrétní rozdělení a obě veličiny X a Y budou proto mít také diskrétní rozdělení a pro jejich rozdělení platí:

$$P(X = i) = P(X = i, \ Y \in \mathbb{R}) = \sum_{j \in \mathbb{R}} P(X = i, \ Y = j)$$

$$P(Y = j) = P(X \in \mathbb{R}, \ Y = j) = \sum_{i \in \mathbb{R}} P(X = i, \ Y = j)$$

Hodnoty pravděpodobností získáme tedy sečtením v řádcích (pro X) a sloupcích (pro Y) naší tabulky:

	$X=0$	$X=1$	$X=2$	$P(Y=j)$
$Y=0$	1/4	1/8	0	3/8
$Y=1$	1/4	1/4	1/8	5/8
$P(X=i)$	1/2	3/8	1/8	

Tedy

$$P(X=i)=\begin{cases}1/2, & i=0\\ 3/8, & i=1\\ 1/8, & i=2\\ 0, & \text{jinak}\end{cases} \qquad \text{a} \qquad P(Y=j)=\begin{cases}3/8, & j=0\\ 5/8, & j=1\\ 0, & \text{jinak.}\end{cases}$$

(b) Protože nyní je např.

$$P(X=2,Y=0)=0\neq\tfrac{1}{8}\cdot\tfrac{3}{8}=P(X=2)\cdot P(Y=0).$$

jsou X a Y **závislé**.

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

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

Pro sdružené pravděpodobnosti vektoru (X',Y') pak 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:

	$Y'=0$	$Y'=1$	$Y'=2$	$P(X'=i)$
$X'=0$	$\frac{1}{2}\cdot\frac{3}{8}=\frac{3}{16}$	$\frac{3}{8}\cdot\frac{3}{8}=\frac{9}{64}$	$\frac{1}{8}\cdot\frac{3}{8}=\frac{3}{64}$	3/8
$X'=1$	$\frac{1}{2}\cdot\frac{5}{8}=\frac{5}{16}$	$\frac{3}{8}\cdot\frac{5}{8}=\frac{15}{64}$	$\frac{1}{8}\cdot\frac{5}{8}=\frac{5}{64}$	5/8
$P(Y'=j)$	1/2	3/8	1/8	

(c) Pro diskrétní vektor (X,Y) a borelovskou (tj. téměř každou, např. spojitou) funkci $h:\mathbb{R}^2\to\mathbb{R}$ platí podobná věc jako jsme měli už u náhodné veličiny a sice:

$$E\big(h(X,Y)\big)=\sum_{(i,j)\in\mathbb{R}^2} h(i,j)\cdot P(X=i,\ Y=j)\ .$$

Odsud tedy snadno spočítáme střední hodnotu $E(XY)$:

$$E\big(X\cdot Y\big)=\sum_{(i,j)\in\mathbb{R}^2} i\cdot j\cdot P(X=i,\ Y=j)=$$
$$=0\cdot 0\cdot\tfrac{1}{4}+0\cdot 1\cdot\tfrac{1}{8}+0\cdot 2\cdot 0+1\cdot 0\cdot\tfrac{1}{4}+1\cdot 1\cdot\tfrac{1}{4}+1\cdot 2\cdot\tfrac{1}{8}=\tfrac{1}{2}$$

Kovarianci pak vypočteme ze vztahu

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

K tomu potřebujeme znát také

$$E(X) = 0\cdot\tfrac{1}{2} + 1\cdot\tfrac{3}{8} + 2\cdot\tfrac{1}{8} = \tfrac{5}{8}$$
$$E(Y) = 0\cdot\tfrac{3}{8} + 1\cdot\tfrac{5}{8} = \tfrac{5}{8}$$

Takže

$$\mathrm{cov}(X,Y) = E(XY) - E(X)\cdot E(Y) = \tfrac{1}{2} - \tfrac{5}{8}\cdot\tfrac{5}{8} = \tfrac{7}{64}.$$

Pro korelaci

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

potřebujeme ještě znát rozptyly, takže si je dopočítáme:

$$E(X^2) = 0^2\cdot\tfrac{1}{2} + 1^2\cdot\tfrac{3}{8} + 2^2\cdot\tfrac{1}{8} = \tfrac{7}{8},$$
$$E(Y^2) = 0^2\cdot\tfrac{3}{8} + 1^2\cdot\tfrac{5}{8} = \tfrac{5}{8},$$
$$\mathrm{var}(X) = E(X^2) - \Big(E(X)\Big)^2 = \tfrac{7}{8} - (\tfrac{5}{8})^2 = \tfrac{31}{64},$$
$$\mathrm{var}(Y) = E(Y^2) - \Big(E(Y)\Big)^2 = \tfrac{5}{8} - (\tfrac{5}{8})^2 = \tfrac{15}{64}.$$

Všimněme si ještě, že $Y \sim \mathrm{Alt}(\frac{5}{8})$, takže rozptyl jsme mohli spočítat jako $\mathrm{var}(Y) = \frac{5}{8}\cdot(1-\frac{5}{8}) = \frac{15}{64}$.
Korelace tedy je

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\cdot\sqrt{\mathrm{var}(Y)}} = \frac{\frac{7}{64}}{\sqrt{\frac{31}{64}}\cdot\sqrt{\frac{15}{64}}} = \frac{7}{\sqrt{465}} \doteq 0.32462,$$

Varianční matice je tudíž

$$\mathrm{Var}(X,Y) = \begin{pmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{pmatrix} = \begin{pmatrix} \frac{31}{64} & \frac{7}{64} \\ \frac{7}{64} & \frac{15}{64} \end{pmatrix}$$

a korelační matice je

$$\mathrm{Corr}(X,Y) = \begin{pmatrix} \mathrm{corr}(X,X) & \mathrm{corr}(X,Y) \\ \mathrm{corr}(Y,X) & \mathrm{corr}(Y,Y) \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{corr}(X,Y) \\ \mathrm{corr}(X,Y) & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{7}{\sqrt{465}} \\ \frac{7}{\sqrt{465}} & 1 \end{pmatrix}.$$

Pro zajímavost si ještě můžeme zjistit úhel $\alpha \in \langle 0, \pi\rangle$ mezi našimi náhodnými veličinami $X - EX$ a $Y - EY$:

$$\alpha = \arccos\big(\mathrm{corr}(X,Y)\big) \doteq \arccos\big(0.32462\big) \doteq 71.06^\circ\ .$$

(Důvodem pro závislost X a Y je také fakt, že $\mathrm{cov}(X,Y) \neq 0$. Zdůrazněme ale, že pokud je kovariance nulová, nemůžeme (pouze na základě její znalosti) o nezávislosti obecně nic říct!)

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 $\iint_A 1\ \mathrm{d}x\mathrm{d}y = 0$, tj. pokud A má nulový plošný obsah).

Příklad 9.2 *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?

(d) Určete pravděpodobnost $P(X > Y)$.

Řešení:

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

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

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

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.

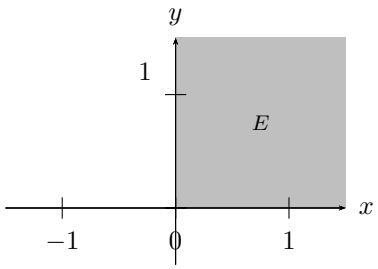
(Podmínce "skoro všude" se nelze vyhnout z toho důvodu, že hustoty nejsou jednoznačně definovány svými hodnotami, ale svými integrály.)

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

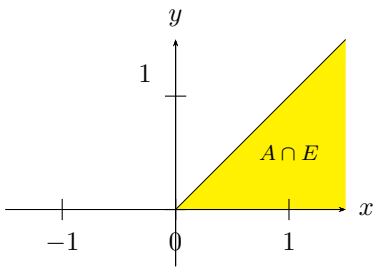
$$\mathrm{Corr}(X,Y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

(d) Sdružená hustota $f_{(X,Y)}$ je nenulová na množině $E := (0,+\infty)^2$ (vyznačena šedě):



Jev "$X > Y$" je popsán jako $(X,Y) \in \underbrace{\{(x,y) \in \mathbb{R}^2 \mid x > y\}}_{A}$. Pravděpodobnost tohoto jevu tak dostaneme zintegrováním hustoty přes množinu A, neboli

$$P(X > Y) = P\Big((X,Y) \in A\Big) = \iint\limits_A f_{X,Y}(x,y)\ dx\ dy = \iint\limits_{A \cap E} f_{X,Y}(x,y)\ dx\ dy =$$



$$= \Big\{A \cap E :\ 0 < y < x\Big\} = \int_0^\infty \int_0^x \tfrac{1}{2} e^{-x-\frac{y}{2}}\ dy\ dx = \int_0^\infty e^{-x} \Big[-e^{-\frac{y}{2}}\Big]_{y=0}^{y=x}\ dx =$$

$$= \int_0^\infty e^{-x}(1 - e^{-\frac{x}{2}})\ dy = \int_0^\infty e^{-x} - e^{-\frac{3}{2}x}\ dx = \Big[-e^{-x} + \tfrac{2}{3} e^{-\frac{3}{2}x}\Big]_{x=0}^{x=\infty} = 1 - \tfrac{2}{3} = \tfrac{1}{3} \ .$$

Poznámky ke kovarianci:
Kovariance $\mathrm{cov}(\cdot,\cdot)$ 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:

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

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

Pro rozptyl $\mathrm{var}(\cdot)$ díky tomu máme:

- $\mathrm{var}(aX + c) = \mathrm{var}(aX) = a^2 \cdot \mathrm{var}(X)$,
- $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2 \cdot \mathrm{cov}(X, Y)$.

Příklad 9.3 *Náhodný vektor (X, Y) má kovarianční matici.*

$$\begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}$$

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

(b) Určete $\mathrm{corr}(X, Y)$*, var$(X + Y)$ a cov$(X + Y, -2Y + 1)$.*

Řešení:

(a) Kovarianční matice pro (X, Y) představuje tyto parametry:

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

Máme tedy $\mathrm{cov}(X, Y) = -2 \neq 0$ a proto veličiny X a Y musí být závislé. (Jestliže by kovariance byla nulová, nemohli bychom udělat o nezávislosti žádný závěr.)

(b) Korelaci určíme jako

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X) \cdot \mathrm{var}(Y)}} = \frac{-2}{\sqrt{3 \cdot 4}} = -\frac{1}{\sqrt{3}} \; .$$

a dále máme

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

a

$$\mathrm{cov}(X + Y, -2Y + 1) = \mathrm{cov}(X + Y, -2Y) = \mathrm{cov}(X, -2Y) + \mathrm{cov}(Y, -2Y) =$$
$$= -2\mathrm{cov}(X, Y) - 2\mathrm{cov}(Y, Y) = (-2) \cdot (-2) + (-2) \cdot 4 = -4 \; .$$

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 $\Omega = \{\mathit{rub},\ \mathit{líc}\}$ a veličina třeba $X(\mathit{líc}) = 1$ a $X(\mathit{rub}) = 0$ 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 $n \in \mathbb{N}$):

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

- -

Odhad chyby v CLV pro alternativní rozdělení: Pokud mají veličiny X_i alternativní rozdělení s parametrem p, tj. $P(X_i = 1) = p$, pak

$$\mu = E(X_i) = p, \qquad \sigma = \sqrt{D(X_i)} = \sqrt{p(1-p)}, \qquad \varrho = E\left(|X_i - \mu|^3\right) = p(1-p)\big(p^2 + (1-p)^2\big) = \sigma^2\big(p^2 + (1-p)^2\big)$$

čímž pro binomické rozdělení $Z_n = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n,p)$ dostáváme odhad

$$\Big|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\Big| \quad < \quad 0.4748 \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}} = 0.4748 \cdot \frac{p^2 + (1-p)^2}{\sqrt{D(Z_n)}} \ .$$

Aproximace CLV se obvykle používá pro $D(Z_n) \ge 9$. Pak je odhad chyby nejvýše: $\Big|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\Big| < \frac{0.4748}{\sqrt{9}} = 0.159$.

Příklad 9.4 *Pravděpodobnost toho, že se za dobu T porouchá přístroj je $p = 0.2$. S jakou pravděpodobností se za dobu T ze 100 (nezávisle pracujících) přístrojů porouchá*

(a) alespoň 20,

(b) méně než 28,

(c) 14 *až* 26 *přístrojů?*

Řešení:
Pro $i = 1, \ldots, n$ (kde $n = 100$) si zavedeme veličiny

$$X_i = \begin{cases} 1 & , \ i\text{-tý přístroj se porouchá,} \\ 0 & , \ i\text{-tý přístroj bude v pořádku.} \end{cases}$$

Veličiny X_i budou nezávislé s alternativním rozdělením $\mathrm{Alt}(p) = \mathrm{Alt}(0.2)$, protože $P(X_i = 1) = 0.2$. Počet porouchaných přístrojů je tedy veličina

$$Z = \sum_{i=1}^{100} X_i$$

která má tudíž binomické rozdělení $\mathrm{Bi}(n, p) = \mathrm{Bi}(100, 0.2)$. To sice umíme přesně popsat, ale vyčíslování součtu mnoha velmi malých členu by vedlo ke značným numerickým chybám (a bez softwaru by ani nebylo možné). Proto použijeme CLV, která velmi dobře aproximuje hledané pravděpodobnosti.

Pro $Z = \sum_{i=1}^{n} X_i$ tedy máme

$$\begin{aligned} E(Z) &= n \cdot E(X_1) = n \cdot p = 100 \cdot 0.2 = 20 \\ \mathrm{var}(Z) &= n \cdot \mathrm{var}(X_1) = n \cdot p \cdot (1-p) = 100 \cdot 0.2 \cdot 0.8 = 16 \\ &\Rightarrow \quad \sqrt{\mathrm{var}(Z)} = \sqrt{16} = 4 \end{aligned}$$

Podle CLV můžeme předpokládat, že veličina $norm(Z) = \frac{Z - E(Z)}{\sqrt{\mathrm{var}(Z)}} = \frac{Z-20}{\sqrt{16}}$ má přibližně normované normální rozdělení $N(0, 1)$.

To také můžeme chápat tak, že veličina Z má *přibližně* normální rozdělení

$$N\Big(E(Z), \mathrm{var}(Z)\Big) = N(20, 4^2)$$

tedy že

$$F_Z(t) \doteq \Phi\left(\frac{t - 20}{\sqrt{16}}\right) \quad \text{pro} \quad t \in \mathbb{R} \ .$$

.

Pak tedy máme:

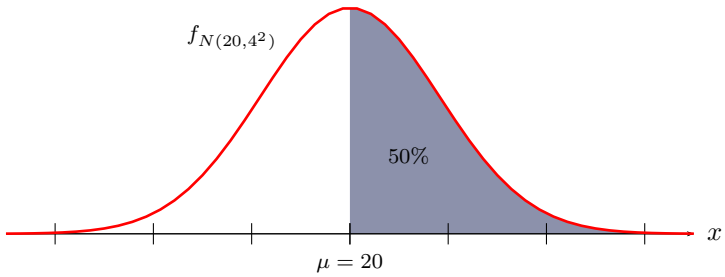
(a)

$$P(Z \geq 20) = P\left(\frac{Z - 20}{\sqrt{16}} \geq \frac{20 - 20}{\sqrt{16}}\right) = P\big(norm(Z) \geq 0\big) =$$

$$= 1 - P\big(norm(Z) < 0\big) \overset{(CLV)}{\doteq} 1 - \Phi\left(0\right) = 1 - 0.5 = \mathbf{0.5} \ .$$

(Pro srovnání: skutečná hodnota pro binomické rozdělení je **0.5398**.)

Když budeme uvažovat (spojité) rozdělení $N(20, 4^2)$, které ALE POUZE aproximuje *původní diskrétní* binomické rozdělení veličiny Z, můžeme si představit hledanou pravděpodobnost (z HLEDISKA VÝPOČTU) takto:



(b)

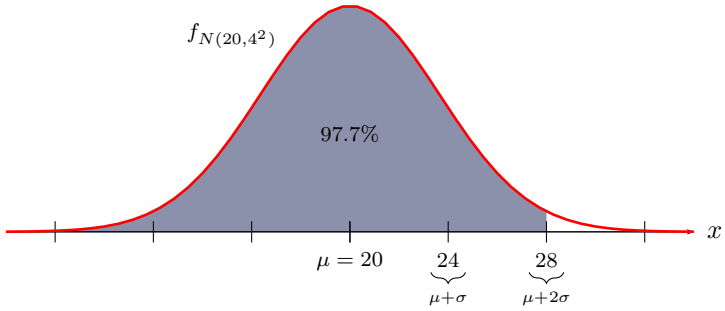
$$P(Z < 28) = P\left(\frac{Z-20}{\sqrt{16}} < \frac{28-20}{\sqrt{16}}\right) = P\Big(norm(Z) < 2\Big) \overset{(CLV)}{\doteq}$$

$$\overset{(CLV)}{\doteq} \Phi(2) \doteq \mathbf{0.977}\ .$$

(Pro srovnání: skutečná hodnota pro binomické rozdělení je **0.9658**.)

Když opět budeme uvažovat (spojité) rozdělení $N(20, 4^2)$, které POUZE aproximuje *původní diskrétní* binomické rozdělení veličiny Z a vezmeme do úvahy pravidlo tří sigma, můžeme uvažovat (z HLEDISKA VÝPOČTU) takto:

Protože $28 = \mu + 2\sigma$, tak hodnota $P(Z < 28) \doteq P\Big(N(20, 4^2) < 28\Big)$ nám musí vyjít větší než 95% (viz náčrt).



(c)

$$P(14 \le Z \le 26) = P\left(\frac{14-20}{\sqrt{16}} \le \frac{Z-20}{\sqrt{16}} \le \frac{26-20}{\sqrt{16}}\right) = P\Big(-1.5 \le norm(Z) \le 1.5\Big) =$$

$$= P\Big(norm(Z) \le 1.5\Big) - P\Big(norm(Z) < -1.5\Big) \overset{(CLV)}{\doteq} \Phi(1.5) - \Phi(-1.5) =$$

$$= 2 \cdot \Phi(1.5) - 1 \doteq 2 \cdot 0.933 - 1 = \mathbf{0.866}\ .$$

(Pro srovnání: skutečná hodnota pro binomické rozdělení je **0.8973**.)

Když opět budeme uvažovat (spojité) rozdělení $N(20, 4^2)$, které POUZE aproximuje *původní diskrétní* binomické rozdělení veličiny Z a vezmeme do úvahy pravidlo tří sigma, můžeme uvažovat (z HLEDISKA VÝPOČTU) takto:

Protože $\mu - 2\sigma = 12 < 14 < 16 = \mu - \sigma$ a $\mu + \sigma = 24 < 16 < 28 = \mu + \sigma$, tak hodnota $P(14 \leq Z \leq 26) \doteq P\Big(14 \leq N(20, 4^2) \leq 26\Big)$ nám musí vyjít mezi 68% a 95% (viz náčrt).