6. cvičení

Poznávačka náhodných veličin

Název	Pravděpodobnostní funkce	Rozsah (Im X)	Střední hodnota	Rozptyl
Bernoulliho Ber(p)	$p_X(1) = p$, $p_X(0) = 1 - p$	$\{0, 1\}$	p	$p(1 - p)$
Binomické Bin(n, p)	$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}$	$\{0, 1, \ldots, n\}$	np	$np(1 - p)$
Geometrické Geo(p)	$p_X(k) = (1 - p)^{k-1} p$	$\{1, 2, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poissonovo Poi(λ)	$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$	$\{0, 1, \ldots\}$	λ	λ
Uniformní Uni(a, b)	$p_X(k) = \frac{1}{b-a+1}$	$\{a, a+1, \ldots, b\}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$

Úloha 1 (Hacker)
Pravděpodobnost, že do našeho serveru pronikne hacker je během každého dne 0.01, nezávisle pro každý den. Označme T počet dnů do prvního průniku. Jaké je rozdělení T, $\mathbb{E}(T)$, $\mathrm{var}(T)$? Jaká je pravděpodobnost, že server zůstane bezpečný po celý rok?

Řešení
$T \sim \mathrm{Geom}(0.01), \mathbb{E}(T) = 100, \mathrm{var}(T) = \frac{1-p}{p^2} = \frac{0.99}{0.0001} = 9900$, $P[\text{365 dní bez incidentu}] = 0.99^{365} \approx 0.025518$

Úloha 2 (Test-driven probability)
Každý test programu může skončit buď nalezením chyby (úspěch) nebo ne (neúspěch). Předpokládáme, že pravděpodobnost nalezení chyby při jednom testu je 0.05 a vývojář provede 20 nezávislých testů, označíme X počet nalezených chyb. Jaké je rozdělení X, $\mathbb{E}(X)$, $\mathrm{var}(X)$? Jaká je pravděpodobnost, že nalezne právě tři chyby?

Řešení
$X \sim \mathrm{Bin}(20, 0.05), \mathbb{E}[X] = 1, \mathrm{var}(X) = np(1 - p) = 0.95$, $P[X = 3] = \binom{20}{3} 0.05^3 \cdot 0.95^{17} \approx 0.059582$

Úloha 3 (Vyřizování žádostí)
Historická data ukazují, že náš server obdrží průměrně 30 žádostí za minutu. Použijte Poissonovo rozdělení k určení pravděpodobnosti, že server obdrží přesně 40 žádostí v následující minutě.

Řešení
Použijeme tedy $Z \sim \mathrm{Poi}(30) \rightsquigarrow P(Z = 40) = \frac{30^{40}}{40!} \cdot e^{-30} = \frac{5845177845843136310577392578125 00}{392276786508598684 5929e^{30}} \approx 0.0139435$

Rozptyl

- **Definice:** rozptyl X je $\mathrm{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$
- **Věta:** $\mathrm{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- **Věta:** $\mathrm{var}(aX + b) = a^2 \,\mathrm{var}(X)$
- **Věta:** pokud $X \perp Y$, máme $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$
- **Definice:** směrodatná odchylka X je $\sigma_X = \sqrt{\mathrm{var}(X)}$
- **Definice:** variační koeficient X je $CV_X = \sigma_X / \mathbb{E}(X)$ (pokud $\mathbb{E}(X) > 0$)

Úloha 4 (Rozptyl z definice)
Spočtete přímo z definice rozptyl Unif(a, b) – stačí pro $a = -2$, $b = 2$. Srovnejte s tabulkou.

Řešení
$\mathrm{var}(X) = \frac{4+1+0+1+4}{5} = 2.$

Úloha 5 (Sliiide to the right)
Dokažte, že $\operatorname{var}(X+b) = \operatorname{var}(X)$.

Řešení
$\mathbb{E}[(X+b)^2] - \mathbb{E}[X+b]^2 = \mathbb{E}[X^2 + 2bX + b^2] - (\mathbb{E}[X] + b)^2 = \mathbb{E}[X^2] + 2b\mathbb{E}[X] + b^2 - \mathbb{E}[X]^2 + 2x\mathbb{E}[X] - b^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \operatorname{var}(X)$

Úloha 6 (Geometrické rozdělení a škálování)
Buď $X \sim \operatorname{Geo}(0.1)$ a $Y \sim \operatorname{Geo}(0.01)$. Zvolte konstantu c tak, aby veličiny X a $Z = cY$ měly stejnou střední hodnotu. Porovnejte rozptyl, směrodatnou odchylku a variační koeficient pro X, Y, Z.

Řešení
$\mathbb{E}[X] = 10, \mathbb{E}[Y] = 100$, tedy je třeba nastavit $c := 1/10$. Pak z tabulky $\operatorname{var}(X) = \frac{1-p}{p^2} = \frac{9/10}{1/100} = 90$, $\operatorname{var}(Z) = c^2 \cdot \operatorname{var}(Y) = \frac{1}{100} \cdot \frac{99/100}{1/10000} = 99$. Směrodatné odchylky jsou pak $\sigma_X = \sqrt{90}, \sigma_Y = \sqrt{99}$ a variační koeficienty jsou totéž, jen vydělené deseti.

Náhodné vektory

Úloha 7 (Opět házeme kostkami)
Označme X, Y výsledky dvou nezávislých hodů čtyřstěnnou kostkou (s čísly 1, …, 4).

a) Jaká je pravděpodobnostní funkce $Z_1 = \max(X, Y)$?

b) Jaká je pravděpodobnostní funkce $Z_2 = XY$?

Nápověda: Jakých hodnot nabývá vektor (X, Y), pokud $\max(X, Y) = k$? Resp. v druhé části, pokud $XY = k$?

Řešení
$p_{Z_1}(1) = 1/16, p_{Z_1}(2) = 3/16, p_{Z_1}(3) = 5/16, p_{Z_1}(4) = 7/16$
$p_{Z_2}(1) = 1/16, p_{Z_2}(2) = 2/16, p_{Z_2}(3) = 2/16, p_{Z_2}(4) = 3/16, p_{Z_2}(6) = 2/16, p_{Z_2}(8) = 2/16, p_{Z_2}(9) = 1/16, p_{Z_2}(12) = 2/16, p_{Z_2}(16) = 1/16$

Úloha 8 (A zase kostky)
Na kostce padne číslo i s pravděpodobností p_i pro $i = 1, \dots, k$. Hodíme n-krát a označíme X_i počet hodů, kdy padlo i.

a) Najděte sdruženou pravděpodobnostní funkci pro n.v. X_1, …, X_k. (Bylo na přednášce!)

b) Jaké je marginální rozdělení, tj. rozdělení jednotlivých n.v. X_i?

c) V téhle části je $k = 3$, $n = 10$, $p_1 = p_2 = p_3 = 1/3$. Určete $p_{X_1|X_3}(4|4)$.

Řešení
Označme $(X_1, \dots, X_k)$ počty výskytů jednotlivých hodnot v n hodech.

a) Pro nezáporná celá $x_1, \dots, x_k$ se součtem $x_1 + \dots + x_k = n$ platí $\Pr(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$. Jinak je tato pravděpodobnost rovna 0.

b) Marginální rozdělení X_i je binomické: $X_i \sim \operatorname{Bin}(n, p_i)$, $\qquad \mathbb{P}(X_i = m) = \binom{n}{m} p_i^m (1-p_i)^{n-m}$.

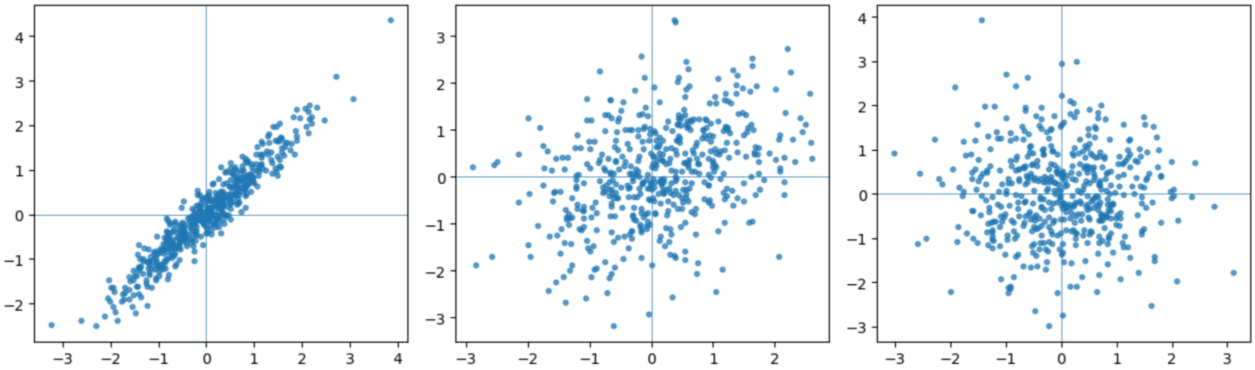
To je vidět „pohledem“, ale také jde ověřit posčítáním přes všechny hodnoty $x_2, \dots, x_k$ se součtem $n - m$ (použije se multinomická věta).

c) Podmíněně na $X_3 = 4$ zbývá 6 hodů, v nichž padá jen 1 nebo 2, a to s podmíněnými pravděpodobnostmi $1/2, 1/2$. Tedy $X_1 \mid (X_3 = 4) \sim \operatorname{Bin}(6, 1/2)$, a proto $p_{X_1|X_3}(4 \mid 4) = \binom{6}{4} \left(\frac{1}{2}\right)^6 = \frac{15}{64}$.

Kovariance a korelace

- $\mathrm{cov}(X,Y) \overset{\text{def}}{=} \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big) \overset{\text{věta}}{=} \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $\mathrm{cov}(X,X) = \mathrm{var}(X)$
- $\mathrm{cov}(X,Y)$ je lineární v obou složkách, tj. např. $\mathrm{cov}(X, a_1Y_1 + a_2Y_2) = a_1\mathrm{cov}(X,Y_1) + a_2\mathrm{cov}(X,Y_2)$
- $\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$.

Úloha 9 (Coco taxi)
Nezávislé náhodné veličiny X, Y mají střední hodnotu 0 a rozptyl 1. Položíme $Z_0 = Y$, $Z_1 = X + 0.3Y$ a $Z_2 = 0.4X + Y$. Spočtěte $\mathrm{cov}(X, Z_i)$ a také korelaci $\rho(X, Z_i)$. Který obrázek odpovídá kterému vzorci?



Řešení
Máme $\mathbb{E}X = \mathbb{E}Y = 0$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$ a X, Y jsou nezávislé, tedy $\mathrm{cov}(X,Y) = 0$.

Pro $Z_0 = Y$: $\mathrm{cov}(X, Z_0) = \mathrm{cov}(X,Y) = 0, \qquad \rho(X, Z_0) = 0.$

Pro $Z_1 = X + 0.3Y$:

$$\mathrm{cov}(X, Z_1) = \mathrm{cov}(X,X) + 0.3\,\mathrm{cov}(X,Y) = 1.$$

Dále

$$\mathrm{Var}(Z_1) = \mathrm{Var}(X) + 0.3^2\mathrm{Var}(Y) = 1 + 0.09 = 1.09,$$

takže

$$\rho(X, Z_1) = \frac{\mathrm{cov}(X, Z_1)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Z_1)}} = \frac{1}{\sqrt{1.09}}.$$

Pro $Z_2 = 0.4X + Y$:

$$\mathrm{cov}(X, Z_2) = 0.4\,\mathrm{cov}(X,X) + \mathrm{cov}(X,Y) = 0.4.$$

Dále

$$\mathrm{Var}(Z_2) = 0.4^2\mathrm{Var}(X) + \mathrm{Var}(Y) = 0.16 + 1 = 1.16,$$

tedy

$$\rho(X, Z_2) = \frac{0.4}{\sqrt{1.16}}.$$

Nejsilnější závislost má Z_1 (levý obrázek), střední Z_2 (prostřední obrázek), nulovou Z_0 (pravý obrázek).