Poznámky k předmětu Aplikovaná statistika, 4. téma

4. Náhodné vektory

V praxi se nám může hodit postihnout více vlastností jednoho objektu najednou, např. výšku, váhu a pohlaví člověka; rychlost chemické reakce a množství katalyzátoru; či počet dnů přípravy na zkoušku, výsledek zkoušky a hodinová dotace zkoušeného předmětu. Takové situace lze modelovat ***n*–rozměrným náhodným vektorem**, což je uspořádaná n–tice náhodných veličin $(X_1, \ldots, X_n)$.

V následujícím výkladu se (pro jednoduchost) omezíme na dvourozměrný náhodný vektor (X, Y). Rozdělení pravděpodobnosti náhodného vektoru (X, Y) je jednoznačně určeno **sdruženou distribuční funkcí**

$$F(x, y) = \mathbb{P}\,(X \leq x, Y \leq y),\ \ x, y \in \mathbb{R},$$

kde

$$\{X \leq x, Y \leq y\} = \{X \leq x\} \cap \{Y \leq y\} = \{\omega \in \Omega; X(\omega) \leq x\} \cap \{\omega \in \Omega; Y(\omega) \leq y\}.$$

Vlastnosti sdružené distribuční fuknce $F(x, y)$ jsou obdobné jako vlastnosti distribuční funkce jedné náhodné veličiny:

- $0 \leq F(x, y) \leq 1$,
- F je neklesající a zprava spojitá v každé z proměnných x, y,
- $\lim_{x\to-\infty} F(x, y) = \lim_{y\to-\infty} F(x, y) = 0$ a $\lim_{x\to\infty} \lim_{y\to\infty} F(x, y) = 1$,
- $\lim_{x\to+\infty} F(x, y) = F_Y(y)$ a $\lim_{y\to+\infty} F(x, y) = F_X(x)$, kde $F_X(x)$ a $F_Y(y)$ jsou distribuční funkce náhodných veličin X a Y.

Poznámka 1. *Rozdělení náhodných veličin X, Y nazýváme* **marginální rozdělení***.*

(D) Nechť diskrétní náhodný vektor (X, Y) nabývá hodnot z množiny

$$M = \{(x_i, y_j) \in \mathbb{R}^2;\ i = 1, 2, \ldots, j = 1, 2, \ldots\}.$$

Pak jeho rozdělení lze popsat **sdruženou pravděpodobnostní funkcí**

$$p(x, y) = \mathbb{P}\,(X = x, Y = y) = \begin{cases} p(x_i, y_j) & , \quad (x, y) = (x_i, y_j) \in M, \\ 0 & , \quad (x, y) \notin M. \end{cases}$$

VLASTNOSTI SDRUŽENÉ PRAVDĚPODOBNOSTNÍ FUNKCE

- $p(x_i, y_j) \geq 0,\ (x_i, y_j) \in M$,
- $\sum\sum_{(x_i,y_j)\in M} p(x_i, y_j) = 1$,
- $\mathbb{P}\,(X \leq x, Y \leq y) = \sum\sum_{x_i\leq x, y_j\leq y, (x_i,y_j)\in M} p(x_i, y_j)$.

Často je užitečné znát rozdělení jednotlivých složek náhodného vektoru, např. náhodnou veličinu popisující rychlost chemické reakce z úvodního příkladu. Máme-li sdruženou pravděpodobnostní funkci p náhodného vektoru (X, Y), lze **marginální pravděpodobnostní funkci**

- $p_X(x_i)$ náhodné veličiny X určit jako

$$p_X(x_i) = \mathbb{P}(X = x_i) = \sum_{y_j;(x_i,y_j)\in M} p(x_i, y_j),$$

- $p_Y(y_j)$ náhodné veličiny Y spočítat jako

$$p_Y(y_j) = \mathbb{P}(Y = y_j) = \sum_{x_i;(x_i,y_j)\in M} p(x_i, y_j).$$

Příklad 1. *Náhodný vektor (M, CH) popisuje výsledky maturity z matematiky a chemie studentů přijatých na VŠCHT v Praze. Rozdělení (M, CH) je zadáno sdruženou pravděpodobnostní funkcí $p(i, j) = \mathbb{P}(M = i, CH = j), i, j = 1, \dots, 4,$ v tabulce*

CH	M			
	1	2	3	4
1	0,38	0,11	0,05	0,01
2	0,08	0,17	0,03	0,02
3	0,02	0,06	0,03	0,01
4	0	0,01	0,01	0,01

(a) Určete marginální pravděpodobnostní funkce známek z maturity z matematiky a chemie.

(b) Stanovte střední hodnotu a rozptyl M, CH a směrodatnou odchylku CH.

(c) Zjistěte pravděpodobnost, že náhodně vybraný student VŠCHT maturoval z matematiky i chemie

- *výborně,*
- *ani výborně, ani dostatečně.*

(S) Rozdělení spojitého náhodného vektoru (X, Y) lze popsat **sdruženou hustotou pravděpodobnosti** $f(x, y)$, což je taková nezáporná funkce, že

$$F(x, y) \equiv \mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s)\mathrm{d}t\,\mathrm{d}s, \ (x, y) \in \mathbb{R}^2.$$

VLASTNOSTI SDRUŽENÉ HUSTOTY PRAVDĚPODOBNOSTI

- $f(x, y) \ge 0, \ (x, y) \in \mathbb{R}^2,$
- $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)\mathrm{d}x\,\mathrm{d}y = 1,$
- $\mathbb{P}\big(X \in [a, b], Y \in [c, d]\big) = \int_a^b \int_c^d f(x, y)\mathrm{d}y\,\mathrm{d}x$. Tuto pravděpodobnost lze geometricky interpretovat jako objem tělesa pod grafem hustoty zúžené na množinu $[a, b] \times [c, d]$.

Jako u diskrétního náhodného vektoru lze i v případě spojitého náhodného vektoru (X, Y) definovat **marginální rozdělení pravděpodobnosti**

- hustotu $f_X(x)$ náhodné veličiny X jako

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\mathrm{d}y,\ x \in \mathbb{R},$$

- hustotu $f_Y(y)$ náhodné veličiny Y jako

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\mathrm{d}x,\ y \in \mathbb{R}.$$

Příklad 2. *Náhodný vektor (X,Y) má sdružené rozdělení dané hustotou*

$$f(x,y) = \begin{cases} \frac{1}{5}(3x+y) & ,\quad x \in [0,1], y \in [0,2], \\ 0 & ,\quad \text{jinde.} \end{cases}$$

(a) Stanovte marginální hustoty náhodných veličin X, Y.

(b) Určete střední hodnotu, rozptyl a směrodatnou odchylku náhodných veličin X, Y.

(c) Spočítejte $\mathbb{P}\big(X \in [0,5;1], Y \in [1,2]\big)$.

Nyní můžeme definovat nezávislost dvojice náhodných veličin.

Definice 1. Náhodné veličiny X, Y jsou (vzájemně) **nezávislé**, jestliže

(D) pro diskrétní náhodné veličiny platí

$$\forall (x_i, y_j) \in M \quad p(x_i, y_j) = p_X(x_i)p_Y(y_j),$$

(S) pro spojité náhodné veličiny platí

$$\forall (x,y) \in \mathbb{R}^2 \quad f(x,y) = f_X(x)f_Y(y).$$

Poznámka 2. *(i) V případě nezávislosti se o sdruženém rozdělení náhodného vektoru (X,Y) říká, že je součinové.*

(ii) Pojem nezávislosti dvojice diskrétních náhodných veličin lze vidět ve světle nezávislosti náhodných jevů následovně:

$$P\big(\{X = x_i\} \cap \{Y = y_j\}\big) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j),\ (x_i, y_j) \in M.$$

Příklad 3. *Jsou výsledky maturity z matematiky a chemie z Příkladu 1 nezávislé? Jsou náhodné veličiny X, Y z Příkladu 2 nezávislé?*

Nejsou-li náhodné veličiny X, Y nezávislé, zkoumáme míru jejich závislosti.

MÍRY STATISTICKÉ LINEÁRNÍ ZÁVISLOSTI NÁHODNÝCH VELIČIN (např. výsledky maturity z matematiky vs. chemie)

Buďte X, Y náhodné veličiny s vlastnostmi $\mathbb{E}X^2 < +\infty, \mathbb{E}Y^2 < +\infty$. Pak **kovariance** náhodných veličin X, Y je definována vztahem

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

(speciálně $\mathrm{cov}(X,X) = \mathrm{var}X$),
kde

- pro diskrétní náhodné veličiny je

$$\begin{aligned}\operatorname{cov}(X,Y) &= \sum_{x_i}\sum_{y_j}(x_i-\mathbb{E}X)(y_j-\mathbb{E}Y)p(x_i,y_j)\\ &= \sum_{x_i}\sum_{y_j}x_iy_jp(x_i,y_j)-\mathbb{E}X\mathbb{E}Y,\end{aligned}$$

- pro spojité náhodné veličiny je

$$\begin{aligned}\operatorname{cov}(X,Y) &= \int_{\mathbb{R}}\int_{\mathbb{R}}(x-\mathbb{E}X)(y-\mathbb{E}Y)f(x,y)\mathrm{d}x\mathrm{d}y\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}xyf(x,y)\mathrm{d}x\mathrm{d}y-\mathbb{E}X\mathbb{E}Y.\end{aligned}$$

Znormujeme-li vhodně kovarianci[1], dostaneme velmi užitečnou míru statistické lineární závislosti náhodných veličin X, Y, tzv. **korelační koeficient**

$$\rho(X,Y) := \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathrm{var}X}\sqrt{\mathrm{var}Y}}$$

za předpokladu, že $0 < \mathrm{var}X < +\infty, 0 < \mathrm{var}Y < +\infty$.

VLASTNOSTI KORELAČNÍHO KOEFICIENTU

- $-1 \leq \rho(X,Y) \leq 1$, $\rho(X,X) = 1$,
- $\rho(X,Y) = 1$ právě tehdy, když $Y = kX + q$, kde $k > 0$, $q \in \mathbb{R}$,
- $\rho(X,Y) = -1$ právě tehdy, když $Y = kX + q$, kde $k < 0$, $q \in \mathbb{R}$,
- X, Y nezávislé, pak $\rho(X,Y) = 0$,
- $\rho(X,Y) = 0$ (tedy i $\operatorname{cov}(X,Y) = 0$), pak náhodné veličiny X, Y nazýváme **nekorelované** (tzn. X, Y nejsou statisticky lineárně závislé)
- s rostoucí $|\rho(X,Y)|$ roste i statistická lineární závislost X, Y.

Na závěr si uveďme některé důležité pojmy související s dvourozměrným náhodným vektorem (X,Y):

- **střední hodnota**

$$\mathbb{E}(X,Y) = (\mathbb{E}X, \mathbb{E}Y),$$

- **kovarianční (varianční) matice**

$$\mathrm{var}(X,Y) = \begin{pmatrix} \mathrm{var}X & , & \operatorname{cov}(X,Y)\\ \operatorname{cov}(X,Y) & , & \mathrm{var}Y\end{pmatrix},$$

  což je obdoba rozptylu jedné náhodné veličiny,

[1] Laskavý čtenář si může povšimnout, že pro normované náhodné veličiny $X_N = \frac{X-\mathbb{E}X}{\sqrt{\mathrm{var}X}}$ a $Y_N = \frac{Y-\mathbb{E}Y}{\sqrt{\mathrm{var}Y}}$ platí

$$\rho(X_N, Y_N) = \operatorname{cov}(X_N, Y_N).$$

- **korelační matice**
$$\mathrm{corr}(X,Y) = \begin{pmatrix} 1 & , & \rho(X,Y) \\ \rho(X,Y) & , & 1 \end{pmatrix},$$
- pro $a, b, c \in \mathbb{R}$ platí
$$\mathbb{E}(aX + bY + c) = a\mathbb{E}X + b\mathbb{E}Y + c,$$
$$\mathrm{var}(aX + bY + c) = a^2\mathrm{var}X + b^2\mathrm{var}Y + 2ab\,\mathrm{cov}(X,Y).$$

  Druhý vztah v případě nezávislosti X, Y přejde na tvar
$$\mathrm{var}(aX + bY + c) = a^2\mathrm{var}X + b^2\mathrm{var}Y.$$

Příklad 4. *Určete kovarianci, korelaci, kovarianční a korelační matici náhodného vektoru* (X,Y) *z Příkladu 1 a 2.*

Cvičení k tomuto tématu

(i) Mějme dvě náhodné veličiny X, Y, pro které platí $\mathbb{E}X = 1$, $\mathrm{var}X = 36$, $\mathbb{E}Y = -4$, $\mathrm{var}Y = 25$ a $\rho(X,Y) = 0,8$. Definujme náhodné veličiny
$$Z := 2X - 3Y - 3,\ U := XY.$$
Určete $\mathbb{E}Z$, $\mathrm{var}Z$, $\mathbb{E}U$.

(ii) Mějme nezávislé náhodné veličiny X, Y s vlastnostmi $\mathbb{E}X = 2$, $\mathrm{var}X = 16$, $\mathbb{E}Y = -1$ a $\mathrm{var}Y = 4$. Definujme náhodné veličiny
$$U := -2X + 3,\ V := 2X - 3Y + 2.$$
Stanovte $\mathbb{E}U$, $\sqrt{\mathrm{var}U}$, $\rho(X,U)$, $\rho(X,Y)$, $\mathbb{E}V$, $\mathrm{var}V$.

(dcv1) Buď (X,Y) náhodný vektor, jehož rozdělení je dané hustotou
$$f(x,y) = \begin{cases} \frac{3}{2}xy^2 & , \quad x \in [0,2], y \in [0,1], \\ 0 & , \quad \text{jinde.} \end{cases}$$
Udělejte úkoly (a), (b) z Příkladu 2, vyšetřete nezávislost náhodných veličin X, Y, spočtěte kovarianci, korelaci, kovarianční a korelační matici náhodného vektoru (X,Y) a dále určete
$$\mathbb{P}\left(1 < X < 2, \frac{1}{2} < Y < 1\right),\ \mathbb{P}\left(0 < X < 1, 0 < Y < \frac{1}{2}\right).$$

(dcv2) Rozdělení náhodného vektoru (X,Y) je dáno sdruženou pravděpodobnostní funkcí $p(x_i, y_j)$

Y	**X**		
	-1	0	1
0	$\frac{1}{18}$	$\frac{2}{9}$	$\frac{1}{18}$
1	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

Jsou náhodné veličiny X, Y nezávislé? Stanovte střední hodnotu náhodných veličin X, Y a dále určete $\mathrm{var}(X,Y)$ a $\mathrm{corr}(X,Y)$.