

(09) Náhodné vektory III.

Plán prednášok o náhodných vektoroch:

- všeobecné vlastnosti
- korelácia zložiek náhodného vektora
- **diskrétné náhodné vektory**
- spojité náhodné vektory

1. Diskrétné náhodné vektory

S diskretnými náhodnými vektormi sme už prakticky pracovali, keď sme mali viac náhodných premenných. Napríklad: hádžeme dvakrát pravidelnou kockou a definujeme M ako maximum z čísel, ktoré padli a S ako súčet. Potom vieme počítať napríklad $P(S < 10, M = 5)$ tým, že si vypíšeme všetky možnosti, ako tie podmienky môžu nastať. V skutočnosti tak už pracujeme s pravdepodobnosťami náhodného vektora (S, M) .

Príklad 1.

- $P(S < 10, M = 5) = 8/36$
- Pravdepodobnostné rozdelenie náhodného vektora (S, M) :

a	b					
	1	2	3	4	5	6
2	1/36	0	0	0	0	0
3	0	2/36	0	0	0	0
4	0	1/36	2/36	0	0	0
5	0	0	2/36	2/36	0	0
6	0	0	1/36	2/36	2/36	0
7	0	0	0	2/36	2/36	2/36
8	0	0	0	1/36	2/36	2/36
9	0	0	0	0	2/36	2/36
10	0	0	0	0	1/36	2/36
11	0	0	0	0	0	2/36
12	0	0	0	0	0	1/36

Marginálne rozdelenie náhodného vektora je pravdepodobnostné rozdelenie jeho zložiek. V prípade diskretného náhodného vektora, ktorého hodnoty máme dané v tabuľke, sčítame pravdepodobnosti v jednotlivých riadkoch, resp. stĺpcoch

Príklad 1 - pokračovanie.

a	b						$p_S(a)$
	1	2	3	4	5	6	
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
$p_M(b)$	1/36	3/36	5/36	7/36	9/36	11/36	

Opakovanie – nezávislosť diskretných náhodných premenných:

Matematické vyjadrenie nezávislosti diskretných náhodných premenných X, Y :

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i) \mathbb{P}(Y = y_j) \quad \forall x_i, y_j$$

Príklad 2. (Zdroj: *A Modern Introduction to Probability and Statistics*)

9.2 $\boxplus$ The joint probability distribution of two discrete random variables X and Y is partly given in the following table.

b	a			$P(Y = b)$
	0	1	2	
-1	...	...	...	1/2
1	...	1/2	...	1/2
$P(X = a)$	1/6	2/3	1/6	1

- Complete the table.
- Are X and Y dependent or independent?

Riešenie:

b	a			P(Y = b)
	0	1	2	
-1	1/6	1/6	1/6	1/2
1	0	1/2	0	1/2
P(X = a)	1/6	2/3	1/6	1

V prvom kroku sme doplnili modré čísla, v druhom kroku červené.

X, Y nie sú nezávislé, napríklad $P(X=0, Y=1)$ je nula, ale $P(X=0)P(Y=1)$ nula nie je.

Príklad 2 – pokračovanie.

Vypočítajte kovarianciu náhodných premenných X a Y.

Riešenie:

$$\begin{array}{c|c|c|c|c}
 k & -2 & -1 & 0 & 1 \\
 \hline
 P(XY=k) & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{2}
 \end{array}$$

$$E(XY) = -2 \cdot \frac{1}{6} + (-1) \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{2} = -\frac{2}{6} - \frac{1}{6} + \frac{1}{2} = 0$$

$$E(X) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{3} + 2 \cdot \frac{1}{6} = 1$$

$$E(Y) = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$$

$$\text{cov}(X, Y) = E(XY) - E(X) \cdot E(Y) = 0 - 1 \cdot 0 = \underline{\underline{0}}$$

Príklad 3 (Zdroj: A Modern Introduction to Probability and Statistics)

9.7 To investigate the relation between hair color and eye color, the hair color and eye color of 5383 persons was recorded. The data are given in the following table:

Eye color	Hair color		
	Fair/red	Medium	Dark/black
Light	1168	825	305
Dark	573	1312	1200

Source: B. Everitt and G. Dunn. *Applied multivariate data analysis*. Second edition Hodder Arnold, 2001; Table 4.12. Reproduced by permission of Hodder & Stoughton.

Eye color is encoded by the values 1 (Light) and 2 (Dark), and hair color by 1 (Fair/red), 2 (Medium), and 3 (Dark/black). By dividing the numbers in the table by 5383, the table is turned into a joint probability distribution for random variables X (hair color) taking values 1 to 3 and Y (eye color) taking values 1 and 2.

- Determine the joint and marginal probability distributions of X and Y .
- Find out whether X and Y are dependent or independent.

Riešenie:

$Y \backslash X$	1	2	3	
1	$1168/5383$	$825/5383$	$305/5383$	$2298/5383$
2	$573/5383$	$1312/5383$	$1200/5383$	$3085/5383$
	$1741/5383$	$2137/5383$	$1505/5383$	

$$P(X=1, Y=1) = \frac{1168}{5383} \approx 0,21698$$

$$P(X=1) \cdot P(Y=1) = \frac{1741}{5383} \cdot \frac{2298}{5383} \approx 0,13807$$

Vidíme, že sú závislé.

2. Testovanie nezávislosti

Je prirodzené predpokladať, že hodnoty v tabuľkách, ako je tá v príklade 3, nepredstavujú všetky hodnoty, ale výsledky náhodného výberu. Vyberieme náhodnú skupinu ľudí, u ktorých zaznamenáme farbu vlasov a farbu očí. Preto ani v prípade skutočnej nezávislosti nedostaneme po zapísaní do tabuľky nezávislé náhodné premenné - môžu tam byť malé odchýlky.

Štatistickými metódami vieme testovať, či sú odchýlky významné alebo ich môžeme pripísať náhodnosti. Ako to spraviť v tomto konkrétnom prípade – použijeme chí-kvadrát test dobrej zhody, o ktorom sme už hovorili v súvislosti s testovaním pravdepodobnostných rozdelení. V tomto prípade:

- Nulová hypotéza: nezávislosť
- Pozorované hodnoty: máme
- Očakávané hodnoty: vyplníme analogickú tabuľku pri predpoklade nezávislosti
- Štandardným spôsobom sa spočíta testovacia štatistika.
- Počet stupňov voľnosti: $(m-1)(n-1)$, kde m je počet riadkov tabuľky a n je počet stĺpcov

- Tak ako v prípade všeobecného chí-kvadrát testu potrebujeme očakávané početnosti aspoň 5.

V Pythone:

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

```
import numpy as np
from scipy.stats import chi2_contingency

tabulka = np.array([[1168, 825, 305], [573, 1312, 1200]])
test_nezavislosti = chi2_contingency(tabulka)
test_nezavislosti

Chi2ContingencyResult(statistic=np.float64(747.4876816409442),
pvalue=np.float64(4.842975659802979e-163), dof=2,
expected_freq=array([[ 743.23202675,  912.28422813,  642.48374512],
[ 997.76797325, 1224.71577187,  862.51625488]]))
```

Ako vznikli tie očakávané počty:

Máme dané marginálne rozdelenia spočítané z dát a predpoklad nezávislosti. Potom napríklad:

$$P(X=1, Y=1) \stackrel{\text{nezávislosť}}{=} \underbrace{P(X=1)}_{\text{dané}} \cdot \underbrace{P(Y=1)}_{\text{dané}} = \frac{1741}{5383} \cdot \frac{2298}{5383}$$

$$5383 \text{ pokusov} \Rightarrow \text{očakávaný počet} = 5383 \cdot P(X=1, Y=1) = \frac{1741 \cdot 2298}{5383} \doteq 743,232$$

Analogicky ostatné.

Štatistika a p-hodnota:

```
print(f"štatistika: {test_nezavislosti.statistic:.5f}")
print(f"p-hodnota: {test_nezavislosti.pvalue:.5f}")
```

```
štatistika: 747.48768
p-hodnota: 0.00000
```

Hypotézu o nezávislosti teda zamietame.

Poznámka: V prípade 2x2 tabuľky zodpovedá tomuto postupu nastavenie parametra **correction=False**. Defaultne sa v tomto prípade robí určitá korekcia, čo zodpovedá **correction=True**

3. Spojité náhodné vektory

Ide o analógiu spojitej náhodnej premennej:

- Plocha pod hustotou náhodnej premennej na určitej množine sa rovná pravdepodobnosti, že náhodná premenná nadobudne hodnotu z tej množiny. Dá sa vypočítať pomocou integrálu.
- Objem pod hustotou dvojrozmerného náhodného vektora na určitej množine sa rovná pravdepodobnosti, že náhodný vektor nadobudne hodnotu z tej množiny. Dá sa vypočítať pomocou tzv. dvojrozmerného integrálu.
- Pravdepodobnosť, že n -rozmerný náhodný vektor nadobudne hodnotu z určitej množiny je hodnota n -rozmerného integrálu z hustoty na danej množine.

Ilustračný obrázok:

Zdroj: *A Modern Introduction to Probability and Statistics*

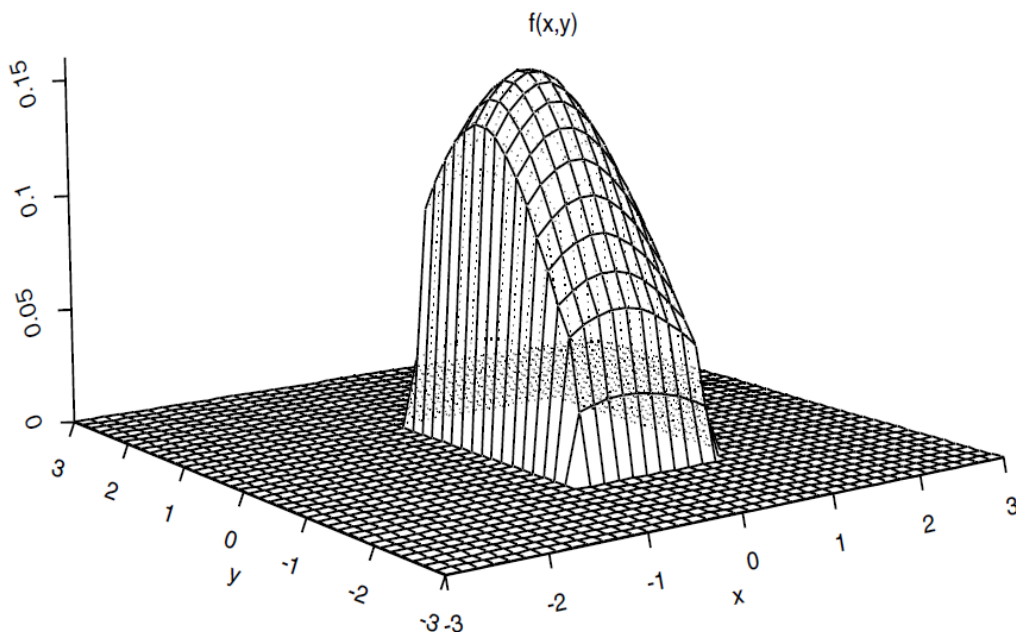


Fig. 9.1. Volume under a joint probability density function f on the rectangle $[-0.5, 1] \times [-1.5, 1]$.

Terminológia:

- Hustota náhodného vektora sa nazýva **zružená hustota**.
- Hustota jeho zložky sa nazýva **marginálna hustota**

V prípade diskrétného náhodného vektora sme pravdepodobnostné rozdelenie jeho zložiek získali sčítaním hodnôt v tabuľke. Spojité náhodné premenné a vektory nadobúdajú nespočítateľne veľa hodnôt, analógiou sčítovania je v tomto prípade integrovanie.

Príklad 4 (Zdroj: *A Modern Introduction to Probability and Statistics*)

density function

$$f(x, y) = \frac{2}{75} (2x^2y + xy^2) \quad \text{for } 0 \leq x \leq 3 \text{ and } 1 \leq y \leq 2,$$

and $f(x, y) = 0$ otherwise; see Figure 9.3.

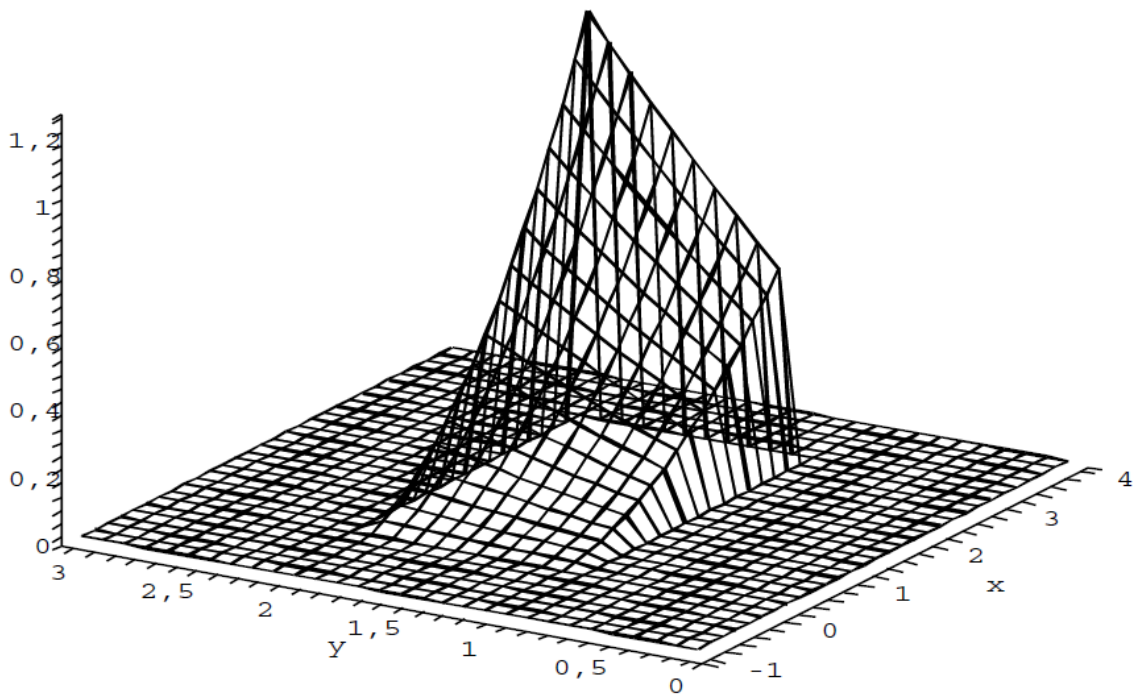


Fig. 9.3. The probability density function $f(x, y) = \frac{2}{75} (2x^2y + xy^2)$.

Marginálna hustota X je:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \frac{2}{75} \int_1^2 (2x^2y + xy^2) dy = \frac{2}{225} (9x^2 + 7x)$$

pre x z intervalu $[0, 3]$, inde je nulová.

(Ak integrujeme podľa y , na premennú x sa pozeráme ako na konštantu.)

Ako dostať z distribučnej funkcie spojitého náhodného vektora jeho hustotu:

- V prípade náhodnej premennej: Hustotu $f(x)$ získame zderivovaním distribučnej funkcie $F(x)$.

- V prípade dvojrozmerného náhodného vektora: Hustotu $f(x,y)$ získame tak, že distribučnú funkciu $F(x,y)$ zderivujeme podľa x a potom podľa y (alebo v opačnom poradí).
- V prípade n -rozmerného náhodného vektora získame hustotu tak, že distribučnú funkciu postupne zderivujeme podľa jednotlivých premenných.

Dôsledok:

- Vieme, že nezávislosť zložiek náhodného vektora znamená, že jeho distribučná funkcia je súčin distribučných funkcií jeho zložiek.
- Z výpočtu hustoty vyplýva, že **zložky spojitého náhodného vektora sú nezávislé práve vtedy, keď je združená hustota súčinom marginálnych hustôt.**

Príklad 5

Majme distribučnú funkciu spojitého náhodného vektora, ktorá sa pre kladné x, y rovná

$$F(x, y) = 1 - e^{-2x} - e^{-y} + e^{-(2x+y)}$$

a inde je nulová.

- Nájdeme:
 - hustotu náhodného vektora
 - marginálne hustoty jeho zložiek
- Zistíme, či sú zložky náhodného vektora nezávislé.

Riešenie:

$$F(x, y) = 1 - e^{-2x} - e^{-y} + e^{-(2x+y)}$$

$$f(x, y) = \left(\underbrace{-e^{-2x} \cdot (-2) + e^{-(2x+y)} \cdot (-2)}_{\text{derivácia } F \text{ podľa } x} \right)'_y = -2e^{-(2x+y)}(-1) \quad \text{pre } x, y > 0$$

este ju musíme
zderivovať podľa y

pre $x > 0$:

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^{\infty} 2e^{-(2x+y)} dy = 2e^{-2x} \underbrace{\int_0^{\infty} e^{-y} dy}_{=1} = 2e^{-2x}$$

pre $y < 0$
je $f=0$

pre $y > 0$:

$$f_y(y) = \int_{-\infty}^{\infty} f(x,y) dx = \int_0^{\infty} 2e^{-(2x+y)} dx = 2e^{-y} \underbrace{\int_0^{\infty} e^{-2x} dx}_{= \frac{1}{2}} = e^{-y}$$

Platí $f(x,y) = f_x(x) \cdot f_y(y) \Rightarrow x, y$ sú nezávislé

4. Združená a marginálna distribučná funkcia

Majme distribučnú funkciu náhodného vektora, chceme z nej získať distribučnú funkciu jeho zložiek.

V dvojrozmernom prípade:

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$$

$$F_X(x) = \mathbb{P}(X \leq x) = \lim_{y \rightarrow \infty} \mathbb{P}(X \leq x, Y \leq y) = \lim_{y \rightarrow \infty} F_{X,Y}(x,y)$$

Teda ako keby sme povedali, že "Y má byť menšie alebo rovné nekonečnu" v zmysle "môže byť hocikaké".

Rovnakou úvahou dostaneme, že v n-rozmernom prípade idú do nekonečna všetky premenné okrem tej, ktorej distribučnú funkciu chceme získať.

Príklad 6

Majme distribučnú funkciu spojitého náhodného vektora z predchádzajúceho príkladu, ktorá sa pre kladné x, y rovná

$$F(x,y) = 1 - e^{-2x} - e^{-y} + e^{-(2x+y)}$$

a inde je nulová. Nájdeme marginálne distribučné funkcie.

Riešenie:

$$F_x(x,y) = 1 - e^{-2x} - e^{-y} + e^{-(2x+y)}$$
$$F_y(y) = \lim_{x \rightarrow \infty} F(x,y) = \lim_{x \rightarrow \infty} (1 - \underbrace{e^{-2x}}_{\rightarrow 0} - e^{-y} + \underbrace{e^{-2x} \cdot e^{-y}}_{\rightarrow 0}) = 1 - e^{-y} \text{ pre } y > 0$$

$$F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x,y) = \lim_{y \rightarrow \infty} (1 - e^{-2x} - e^{-y} + e^{-2x} e^{-y}) = 1 - e^{-2x} \text{ pre } x > 0$$

$$F_X(x) \cdot F_Y(y) = (1 - e^{-2x})(1 - e^{-y}) = 1 - e^{-2x} - e^{-y} + e^{-2x-y} = F_{X,Y}(x,y)$$

$\Rightarrow X, Y$ sú nezávislé

5. Dôležité rozdelenia: multinomické (diskrétné)

Pripomeňme si **binomické rozdelenie** (predpona **bi** = dva):

- Robíme **N** nezávislých pokusov.
- Každý pokus skončí úspechom alebo neúspechom.
- Pravdepodobnosť úspechu je v každom pokuse rovná **p**.
- Náhodná premenná vyjadruje počet úspechov v uskutočnených pokusoch.

Napríklad: Hádzame mincou a za úspech označíme padnutie hlavy.

Teraz definujeme zovšeobecnenie – **multinomické rozdelenie** (predpona **multi** = veľa):

- Robíme **N** nezávislých pokusov.
- Každý pokus skončí jednou z **k** možností
- Pravdepodobnosť *i*-tej možnosti je v každom pokuse rovná **p_i**.
- Náhodný vektor má **k** zložiek, *i*-ta zložka vyjadruje počet pokusov, v ktorých nastala *i*-ta možnosť.

Napríklad: Hádzame kockou a zapisujeme počet jednotiek, dvojek, ... atď.

Pravdepodobnosti (ak je súčet **x_i** rovný **n**, teda počtu zrealizovaných pokusov):

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

Z kombinatorickej interpretácie týchto pravdepodobností vyplýva, že marginálne rozdelenia jednotlivých zložiek sú binomické (počet pokusov **N**, pravdepodobnosť úspechu **p_i**).

6. Dôležité rozdelenia: viacrozmerné normálne (spojité)

Náhodný vektor, ktorého zložky sú nezávislé náhodné premenné s normálnym rozdelením má hustotu (pripomeňme si, že hustota je vďaka nezávislosti súčinom marginálnych hustôt):

$$f(x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left\{ -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right\} = \frac{1}{(2\pi)^{n/2} \prod_{i=1}^n \sigma_i} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right\}$$

Špeciálne v 2D:

$$f(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y} \exp \left\{ -\frac{(x - \mu_X)^2}{2\sigma_X^2} - \frac{(y - \mu_Y)^2}{2\sigma_Y^2} \right\}$$

Zovšeobecnenie:

$$\frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \exp \left(-\frac{1}{2(1 - \rho^2)} \left\{ \left(\frac{x - \mu_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x - \mu_X}{\sigma_X} \right) \left(\frac{y - \mu_Y}{\sigma_Y} \right) + \left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 \right\} \right)$$

- Marginálne rozdelenia zložiek X a Y sú normálne, stredná hodnota a štandardná odchýlka majú príslušné indexy.
- Korelácia ρ musí byť z intervalu $(-1, 1)$, teda nemôže byť rovná $+1$ alebo -1 . Inak povedané, jedna zložka vektora nemôže byť lineárnou funkciou druhej.

Vyjadrenie kovariancie (potrebujeme do kovariančnej matice) pomocou korelácie a štandardných odchýliek (tie majú lepšiu interpretáciu):

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \Rightarrow \text{cov}(X, Y) = \rho_{XY} \sigma_X \sigma_Y$$

Simulovanie v Pythone:

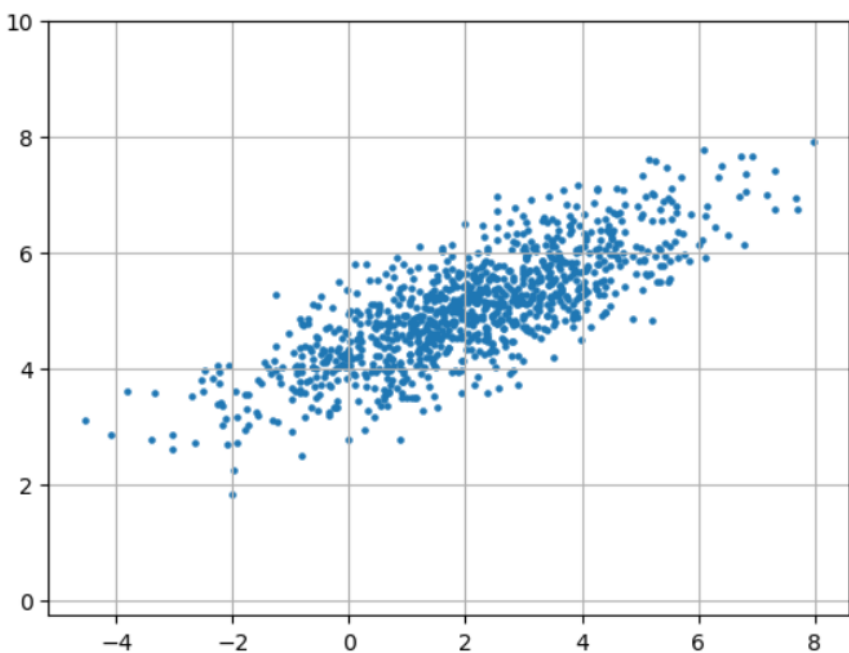
```
import numpy as np
import matplotlib.pyplot as plt

mu_X, mu_Y = 2, 5
sigma_X, sigma_Y = 2, 1
rho = 0.8

mu = [mu_X, mu_Y]
Sigma = [[sigma_X**2, rho*sigma_X*sigma_Y],
          [rho*sigma_X*sigma_Y, sigma_Y**2]]

data = np.random.multivariate_normal(mu, Sigma, size=1000)

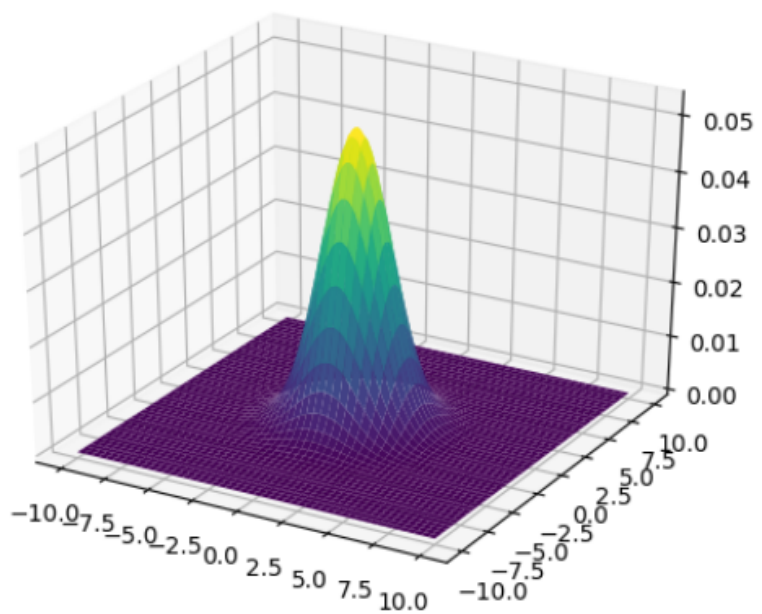
plt.scatter(*data.T, s=5)
plt.grid(True)
plt.axis('equal')
plt.show()
```



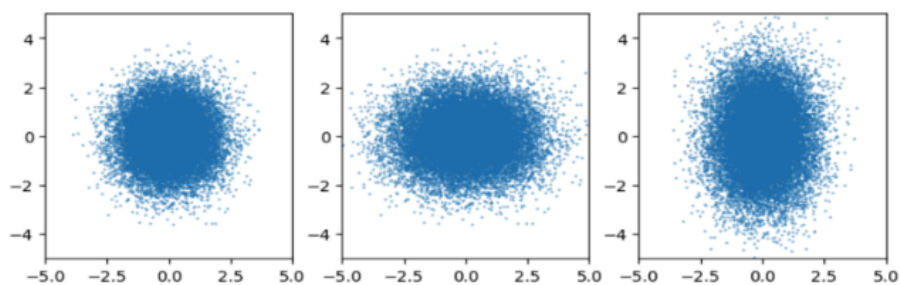
Zdroj nasledujúcich obrázkov:

https://web.stanford.edu/class/archive/cs/cs109/cs109.1218/files/student_drive/5.9.pdf

Hustota:



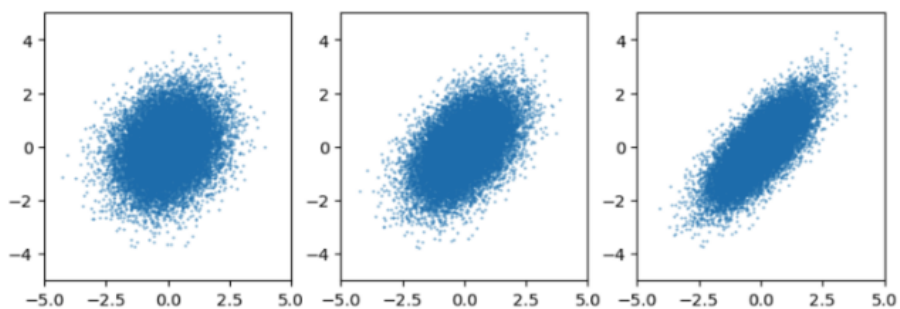
Generované body (středná hodnota je vždy (0, 0)):



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

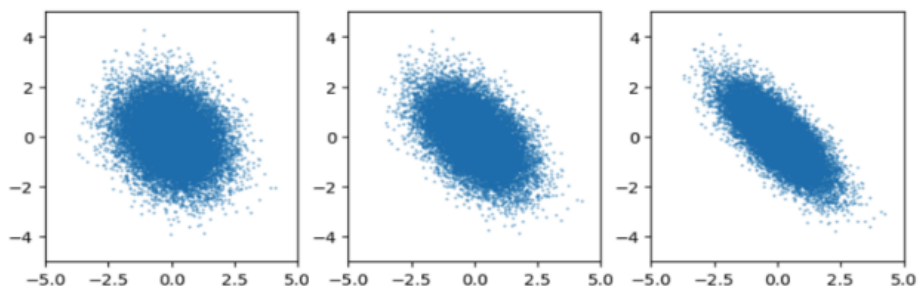
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

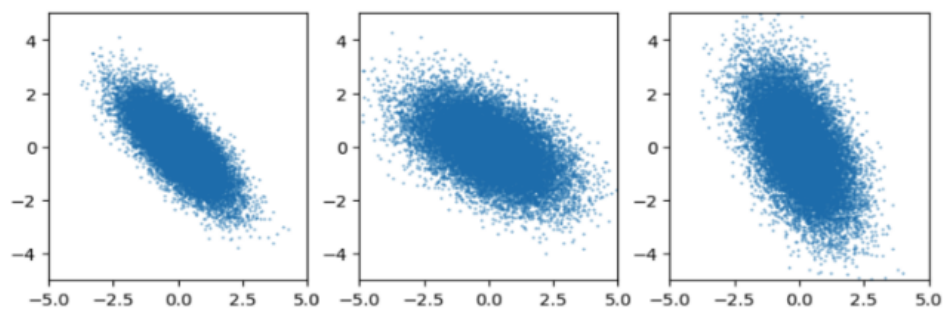
$$\Sigma = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.25 \\ -0.25 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2 & -0.75 \\ -0.75 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 2 \end{bmatrix}$$

Zovšeobecnenie: n-rozmerné normálne rozdelenie

- Marginálne rozdelenia sú normálne.
- Je daný vektor stredných hodnôt $\boldsymbol{\mu}$ kovariančná matica $\boldsymbol{\Sigma}$.
- Podmienka: žiadna zložka náhodného vektora sa nedá napísať ako lineárna kombinácia ostatných.

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$