

(08) Náhodné vektory II.

Plán prednášok o náhodných vektoroch:

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

1. Kovariancia: definícia

Príklad: Našiel sa odtlačok topánky v pôde, chceme na základe toho odhadnúť, ako vysoký bol človek, ktorý tade prešiel (napríklad kvôli lepšiemu popisu páchatel'a nejakého trestného činu). Najskôr by nás ale zaujímalo, nakoľko dobre sa to dá, aká silná je závislosť medzi týmito dvoma veličinami.

Dáta:

https://sebastiansauer.github.io/pradadata/reference/wo_men.html
<https://osf.io/66fvm>

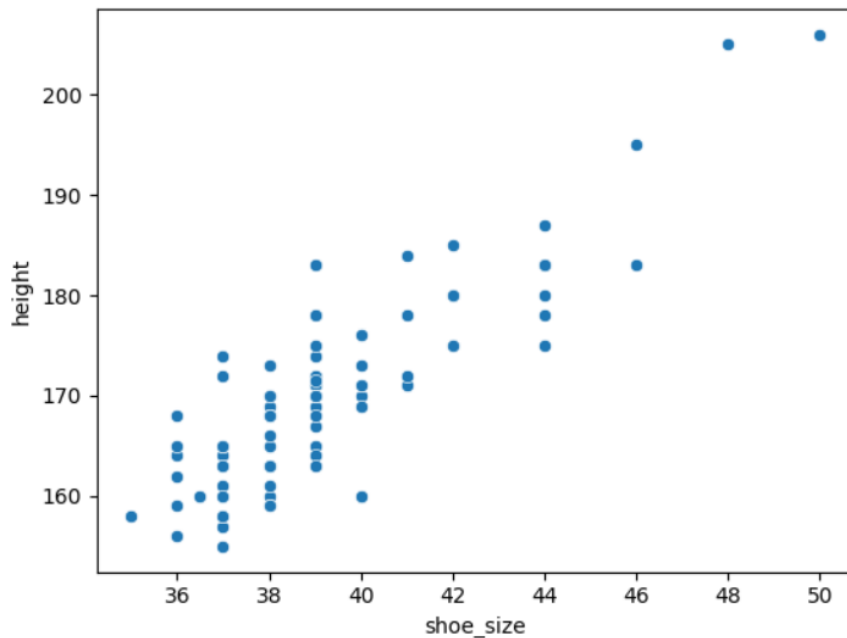
	time	sex	height	shoe_size
0	04.10.2016 17:58:51	woman	160.0	40.0
1	04.10.2016 17:58:59	woman	171.0	39.0
2	04.10.2016 18:00:15	woman	174.0	39.0
3	04.10.2016 18:01:17	woman	176.0	40.0
4	04.10.2016 18:01:22	man	195.0	46.0

Po úprave dát (jedna zjavne nesprávne zadaná hodnota a niekoľkokrát výška uvedená v metroch namiesto centimetrov) si dáta vykreslíme:

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.scatterplot(x='shoe_size', y='height', data=df)
plt.show()
```

(Prečo **seaborn**, uvidíme o chvíľu + bude v tohtotýždňovom spoločnom bonuse na Datacampe.)



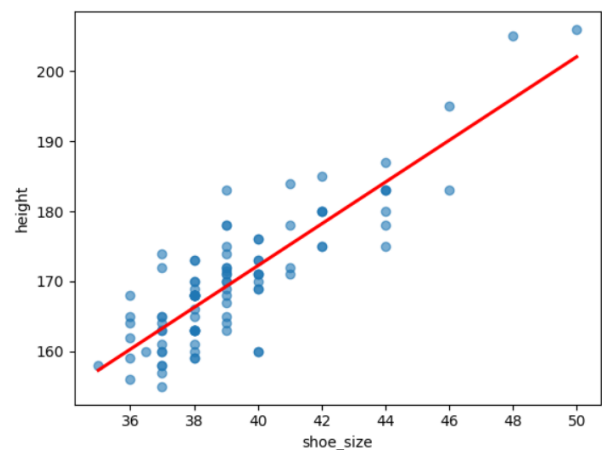
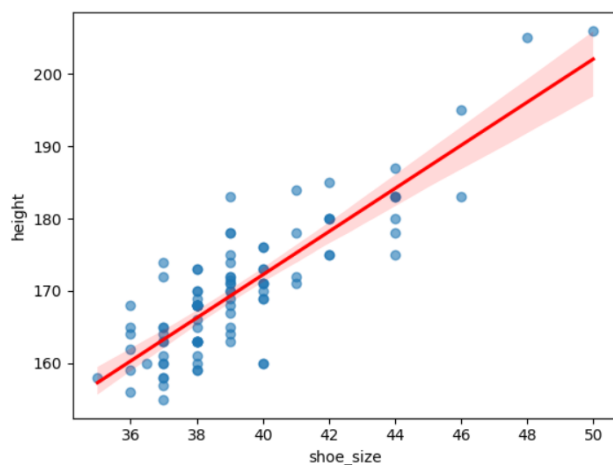
Pozorovania:

(1) Je celkom prirodzené, že závislosť je rastúca, typická situácia je taká, že väčšej topánke zodpovedá väčšia výška.

(2) Zdá sa, že táto závislosť by sa dala dobre aproximovať priamkou. Tomu sa budeme podrobnejšie venovať neskôr, zatiaľ len ukážka funkcie **regplot**:

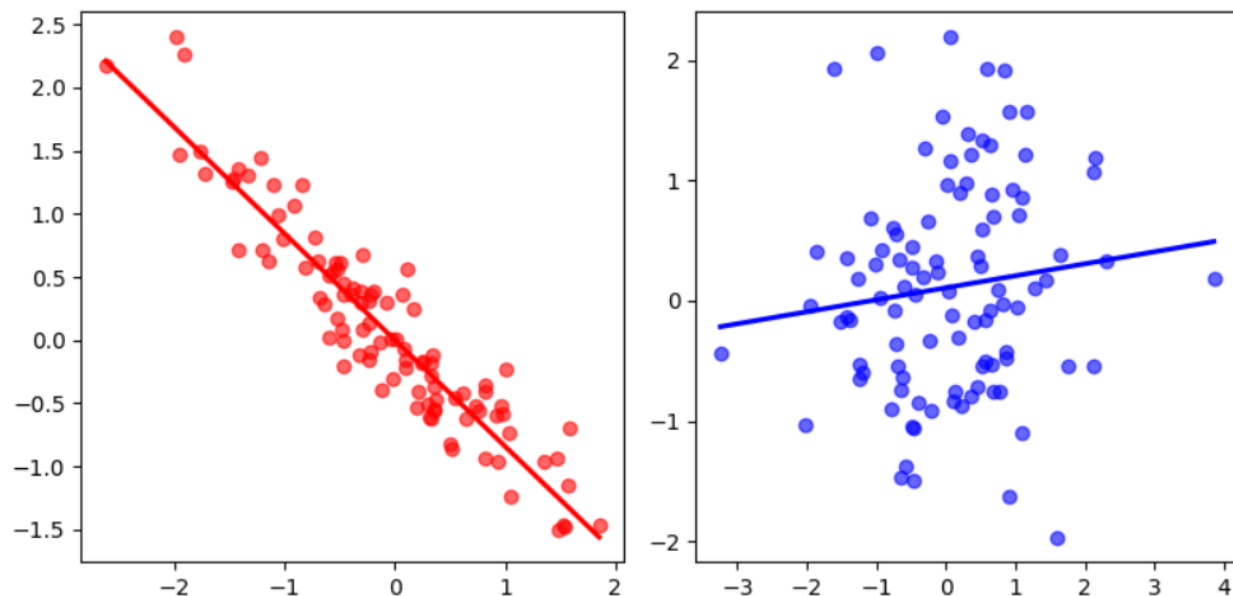
```
sns.regplot(x="shoe_size", y="height", data=df,
            scatter_kws={"alpha": 0.6},
            line_kws={"color": "red"})
plt.show()
```

Respektíve bez tej vyznačenej plochy (na neskorších prednáškach si ju vysvetlíme) pomocou parametra `ci=None`

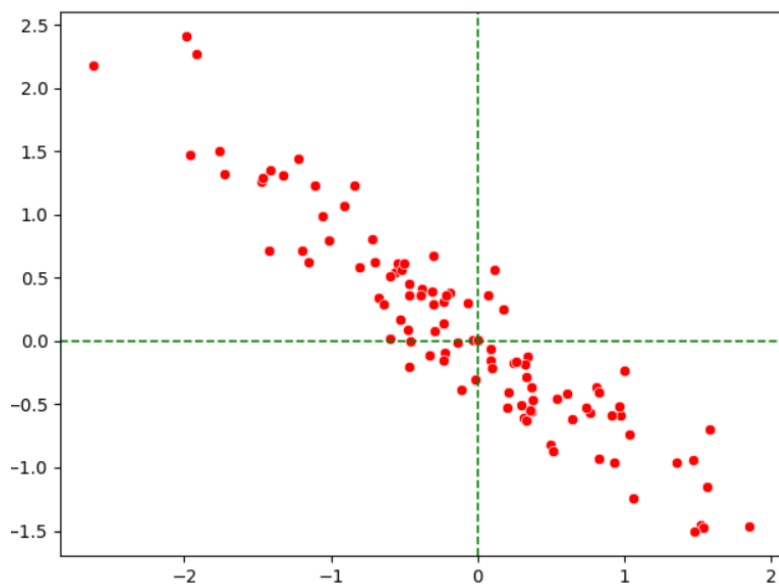


Chceli by sme nejako číselne popísať túto vlastnosť dát, keďže nie vždy to takto musí byť. Závislosť môže byť klesajúca, alebo tiež nemusí byť viditeľná žiadna, resp. veľmi slabá (**regplot** nejakú priamku nakreslí, ale vidíme, že dáta nijako nevystihuje).

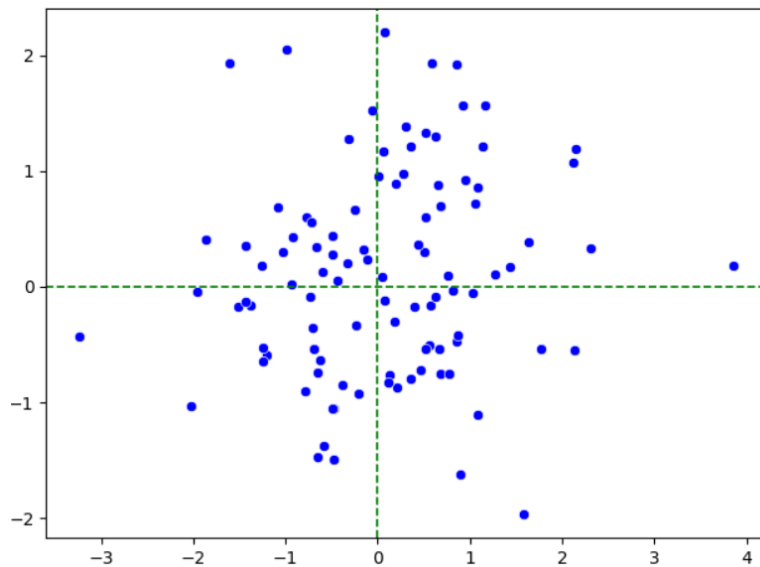
Ilustračné obrázky:



Tieto dáta majú nulový priemer, zakreslíme si do obrázku x-ovú a y-ovú os:

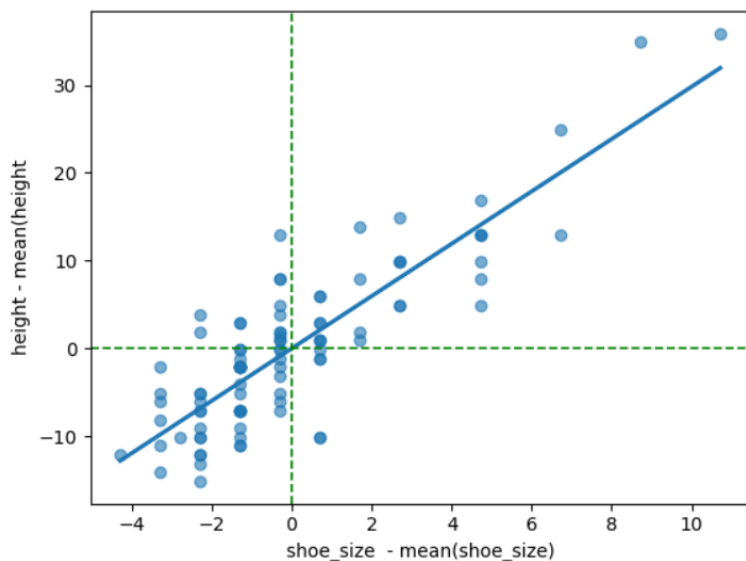


Vidíme, že väčšina hodnôt je v pravom dolnom a ľavom hornom kvadrante. Premenné x , y teda majú väčšinou opačné znamienka a teda ich súčin je väčšinou záporný. Priemer takýchto súčinov bude záporný.



Hodnoty sú vo všetkých kvadrantoch. Súčin x, y má rôzne znamienka, priemer takýchto súčinov bude asi blízky nule.

Aby sme mohli spraviť podobnú analýzu pre dáta o veľkosti topánok a výške, posunieme ich tak, aby sme dostali dáta s nulovým priemerom. Tým sa nezmení vzťah medzi nimi. Pracujeme teda s dátami **shoe_size - mean(shoe_size), height - mean(height)**.



Väčšina hodnôt je potom v pravom hornom a ľavom dolnom kvadrante. Premenné teda majú väčšinou rovnaké znamienka a teda ich súčin je väčšinou záporný. Priemer takýchto súčinov bude kladný.

Ak máme náhodné premenné, nie dáta, namiesto priemeru sa berie stredná hodnota a definuje sa **kovariancia náhodných premenných X, Y** ako:

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

2. Kovariancia: vlastnosti

1. Kovariancia náhodnej premennej so sebou je jej disperzia:
$$\text{Cov}(X, X) = E[(X - E(X))(X - E(X))] = E[(X - E(X))^2] = D(X)$$
2. Užitočný vzťah pre kovarianciu:
$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] = \\ &= E[XY - X \cdot E(Y) - E(X) \cdot Y + E(X) \cdot E(Y)] = \\ &= E[XY] - \underbrace{E[X \cdot E(Y)]}_{\text{konst.}} - \underbrace{E[E(X) \cdot Y]}_{\text{konst.}} + \underbrace{E[E(X) \cdot E(Y)]}_{\text{konst.}} = \\ &= E[XY] - E(Y) \cdot E(X) - E(X) \cdot E(Y) + E(X) \cdot E(Y) = \\ &= E(XY) - E(X) \cdot E(Y)\end{aligned}$$
3. Pre nezávislé náhodné premenné sme (bez dôkazu) uvádzali vlastnosť $E(XY) = E(X)E(Y)$. Preto z bodu 2 vyplýva dôležitá vlastnosť: **kovariancia nezávislých náhodných premenných je nulová.**
4. Naopak to neplatí: **náhodné premenné s nulovou kovarianciou nemusia byť nezávislé.**

Príklad 1. Nech X je náhodná premenná s rovnomerným rozdelením na intervale $(-1, 1)$ a Y je jej druhá mocnina. Potom X, Y nie sú nezávislé. Ukážeme však, že majú nulovú kovarianciu.

$$X \sim R(-1, 1) \Rightarrow f(x) = \frac{1}{2} \text{ na intervale } (-1, 1), \text{ inde } 0$$

$$Y = X^2$$

$$XY = X \cdot X^2 = X^3$$

$$E(X) = 0 \quad (\text{stred intervalu } (-1, 1))$$

$$E(Y) = E(X^2) = \int_{-1}^1 x^2 \cdot \frac{1}{2} dx = \frac{1}{2} \left[\frac{x^3}{3} \right]_{-1}^1 = \frac{1}{6} (1^3 - (-1)^3) = \frac{2}{6} = \frac{1}{3}$$

$$E(XY) = E(X^3) = \int_{-1}^1 x^3 \cdot \frac{1}{2} dx = \frac{1}{2} \left[\frac{x^4}{4} \right]_{-1}^1 = \frac{1}{8} (1^4 - (-1)^4) = 0$$

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y) = 0 - 0 \cdot \frac{1}{3} = 0$$

Vlastnosti - pokračovanie:

5. Symetria : $\boxed{\text{cov}(X, Y) = \text{cov}(Y, X)}$ - vyplýva priamo z definície

6.

Ak $a, b \in \mathbb{R}$ sú konštanty :

$$\begin{aligned}\boxed{\text{cov}(aX, bY)} &= E(aX \cdot bY) - E(aX) \cdot E(bY) = \\ &= ab E(XY) - aE(X) \cdot bE(Y) = \\ &= ab [E(XY) - E(X)E(Y)] = \boxed{ab \cdot \text{cov}(X, Y)}\end{aligned}$$

7.

$$\begin{aligned}\boxed{\text{cov}(X+Y, U+V)} &= E((X+Y)(U+V)) - E(X+Y) \cdot E(U+V) = \\ &= E(XU + XV + YU + YV) - [E(X) + E(Y)] \cdot [E(U) + E(V)] = \\ &= E(XU) + E(XV) + E(YU) + E(YV) - \\ &\quad - [E(X) \cdot E(U) + E(X) \cdot E(V) + E(Y) \cdot E(U) + E(Y) \cdot E(V)] = \\ &= \boxed{\text{cov}(X, U) + \text{cov}(X, V) + \text{cov}(Y, U) + \text{cov}(Y, V)}\end{aligned}$$

Príklad 3.

$$\text{cov}(2X+Y, 3Z+W-V) = \text{cov}(2X, 3Z) + \text{cov}(2X, W) + \text{cov}(2X, -V) + \text{cov}(Y, 3Z) + \text{cov}(Y, W) + \text{cov}(Y, -V) =$$

roznásobuje sa
každý člen s každým

$$= 2 \cdot 3 \cdot \text{cov}(X, Z) + 2 \cdot \text{cov}(X, W) + 2 \cdot (-1) \text{cov}(X, -V) + 3 \text{cov}(Y, Z) + \text{cov}(Y, W) + (-1) \cdot \text{cov}(Y, V)$$

Vlastnosti - pokračovanie:

8. Ak $a, b \in \mathbb{R}$ sú konštanty :

$$\begin{aligned}\boxed{\text{cov}(X+a, Y+b)} &= E\left[\underbrace{(X+a - E(X+a))}_{X+a-E(X)-E(a)} \cdot \underbrace{(Y+b - E(Y+b))}_{Y-E(Y)}\right] = \\ &= E((X-E(X)) \cdot (Y-E(Y))) = \boxed{\text{cov}(X, Y)}\end{aligned}$$

Príklad 4.

$$\begin{aligned} \text{cov}(2X+3, -3Y-2) &= \text{cov}(2X, -3Y) = 2 \cdot (-3) \text{cov}(X, Y) = \\ &= -6 \text{cov}(X, Y) \end{aligned}$$

Disperzia súčtu náhodných premenných

- Už sme používali vlastnosť disperzie, podľa ktorej sa disperzia súčtu nezávislých náhodných premenných rovná súčtu disperzií.
- Takýto výsledok platí nielen pre nezávislé, ale aj pre nekorelované náhodné premenné.
- Vo všeobecnosti platí:

$$\begin{aligned} D(X+Y) &= \text{cov}(X+Y, X+Y) = \underbrace{\text{cov}(X, X)}_{=D(X)} + \underbrace{\text{cov}(X, Y)}_{=\text{cov}(X, Y)} + \underbrace{\text{cov}(Y, X)}_{=\text{cov}(X, Y)} + \underbrace{\text{cov}(Y, Y)}_{=D(Y)} = \\ &= D(X) + D(Y) + 2\text{cov}(X, Y) \end{aligned}$$

Príklad 5.

Let X and Y be the number of hours that a randomly selected person watches movies and sporting events, respectively, during a three-month period. The following information is known about X and Y :

$$E(X) = 50, \quad E(Y) = 20, \quad \text{Var}(X) = 50, \quad \text{Var}(Y) = 30, \quad \text{Cov}(X, Y) = 10.$$

One hundred people are randomly selected and observed for these three months. Let T be the total number of hours that these one hundred people watch movies or sporting events during this three-month period. Approximate the value of $P(T < 7100)$.

- (A) 0.62 (B) 0.84 (C) 0.87 (D) 0.92 (E) 0.97

Riešenie:

$T = T_1 + \dots + T_{100}$, kde $T_i = X_i + Y_i$ sú nezávislé
Z centrálnej limitnej vety vyplýva, že T má približne
normálne rozdelenie.

$$\begin{aligned} E(T) &= E(T_1) + \dots + E(T_{100}) = 100 \cdot E(T_i) = 100 \cdot (E(X_i) + E(Y_i)) = \\ &= 100(50 + 20) = 7000 \end{aligned}$$

$$D(T) \underset{\substack{\uparrow \\ \text{nezávislosť } T_i}}{=} D(T_1) + \dots + D(T_{100}) = 100 \cdot D(T_i) = 100 \cdot D(X_i + Y_i) =$$

$$= 100 [D(X_i) + D(Y_i) + 2\text{cov}(X_i, Y_i)] = 100 [50 + 30 + 2 \cdot 10] = 100 \cdot 100$$

$$P(T < 7100) = P\left(\frac{T - 7000}{\sqrt{100 \cdot 2}} < \frac{7100 - 7000}{\sqrt{100 \cdot 2}}\right) = P\left(N(0,1) < \frac{100}{100}\right) = \Phi(1)$$

kde Φ je distribučná f. $N(0,1)$

3. Kovariančná matica náhodného vektora

Ak $\mathbf{X}$ je n -rozmerný náhodný vektor, tak jeho **kovariančná matica** je definovaná ako

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Cov}(X_n, X_n) \end{pmatrix}$$

Všimnime si, že **na diagonále sú disperzie** zložiek náhodného vektora.

Odhadovanie z dát:

<https://numpy.org/devdocs/reference/generated/numpy.cov.html>

Poznámka:

Rovnako ako pri odhadovaní disperzie sa rozlišuje prípad, že ide o náhodný výber (menovateľ $N-1$) a že ide o disperziu spočítanú z celej populácie (menovateľ N)

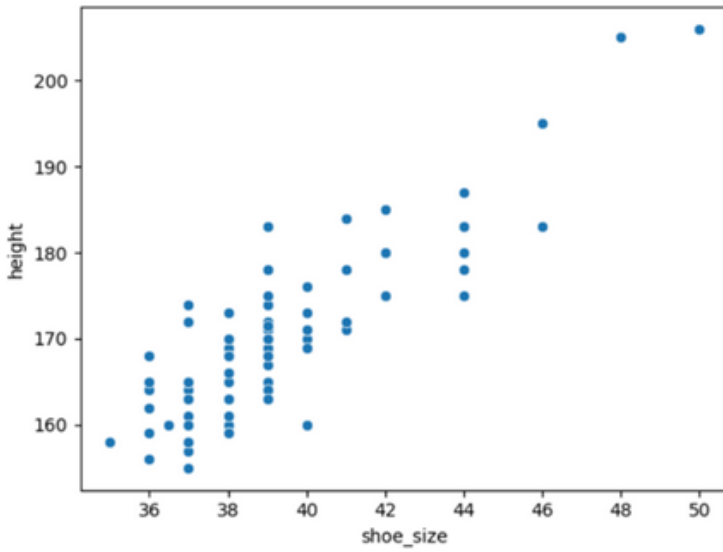
bias : *bool, optional*

Default normalization (False) is by $(N - 1)$, where N is the number of observations given (unbiased estimate). If *bias* is True, then normalization is by N .

Príklad výpočtu:

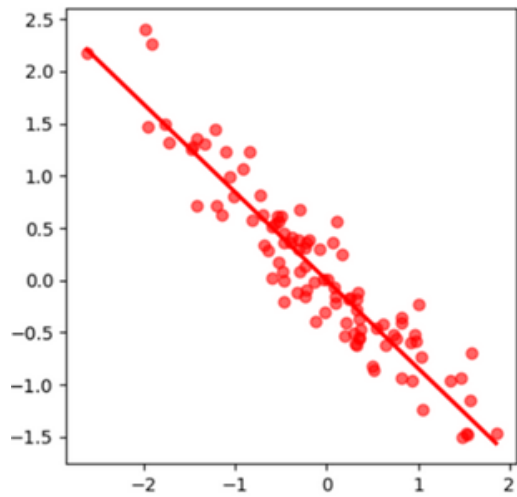
```
import numpy as np

cov_red = np.cov(X_red, Y_red)
cov_blue = np.cov(X_blue, Y_blue)
cov_df = np.cov(df["shoe_size"], df["height"])
```



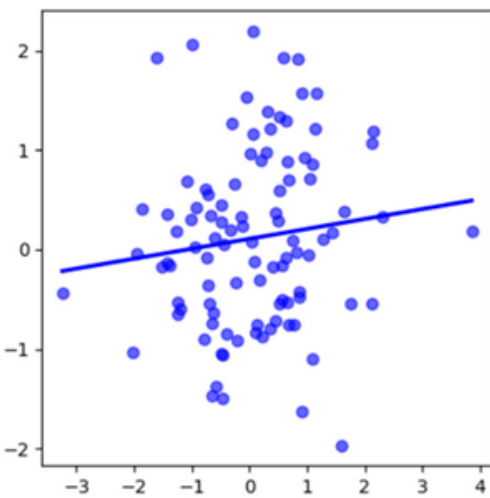
```
print(cov_df)
```

```
[[ 7.21474954 21.52017625]
 [21.52017625 87.53200371]]
```



```
print(cov_red)
```

```
[[ 0.82476989 -0.69526217]
 [-0.69526217  0.66642035]]
```



```
print(cov_blue)
```

```
[[1.17566937 0.1173186 ]
 [0.1173186  0.79334234]]
```

Alebo:

```
import pandas as pd

kovariancna_matica = df[["shoe_size", "height"]].cov()
print(kovariancna_matica)
```

```
      shoe_size  height
shoe_size  7.214750  21.520176
height     21.520176  87.532004
```

4. Korelácia a korelačná matica

Nevýhodou kovariancie je, že závisí od jednotiek. Ak v našom úvodnom príklade ponecháme premennú vyjadrujúcu veľkosť topánok bez zmeny a výšku človeka vyjadríme namiesto centimetrov v metroch, kovariancia sa zmenší stokrát. Pritom závislosť dát zostáva rovnaká, na pôvodnom grafe by stačilo doplniť desatinné čiarky na y-ovú os. Preto sa definuje **korelácia**, pri ktorej tento problém nevzniká. **Korelácia náhodných premenných** (tiež sa nazýva **korelačný koeficient**) sa definuje ako kovariancia normalizovaných náhodných premenných (teda odpočítame strednú hodnotu a vydělíme štandardnou odchýlkou):

$$\begin{aligned} \text{cov}\left(\frac{X - E(X)}{\sqrt{D(X)}}, \frac{Y - E(Y)}{\sqrt{D(Y)}}\right) &= \frac{1}{\sqrt{D(X)}} \cdot \frac{1}{\sqrt{D(Y)}} \text{cov}\left(\overset{\text{konst.}}{X - E(X)}, \overset{\text{konst.}}{Y - E(Y)}\right) = \\ &= \frac{\text{cov}(X, Y)}{\sqrt{D(X) \cdot D(Y)}} \end{aligned}$$

Teda **korelácia** bude:

$$\rho_{X,Y} = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{D(X) D(Y)}}$$

Výpočet:

```
import numpy as np

korelacna_matica_1 = np.corrcoef(df["shoe_size"], df["height"])
print(korelacna_matica_1)
```

```
[[1.          0.8563509]
 [0.8563509  1.          ]]
```

Alebo:

```
import pandas as pd

korelacna_matica_2 = df[["shoe_size", "height"]].corr()
print(korelacna_matica_2)
```

	shoe_size	height
shoe_size	1.000000	0.856351
height	0.856351	1.000000

Ak nás zaujíma len jedna korelácia:

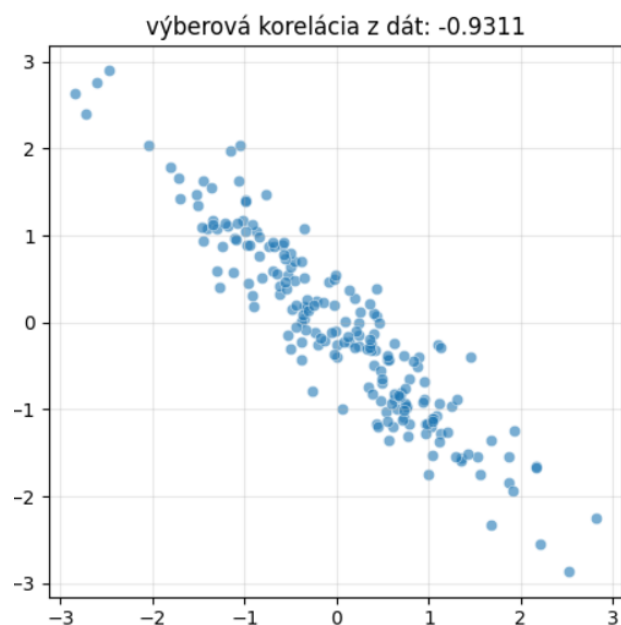
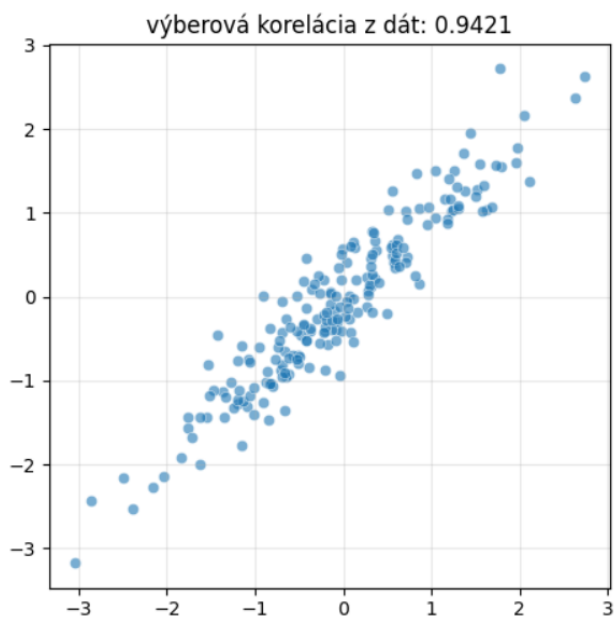
```
import pandas as pd

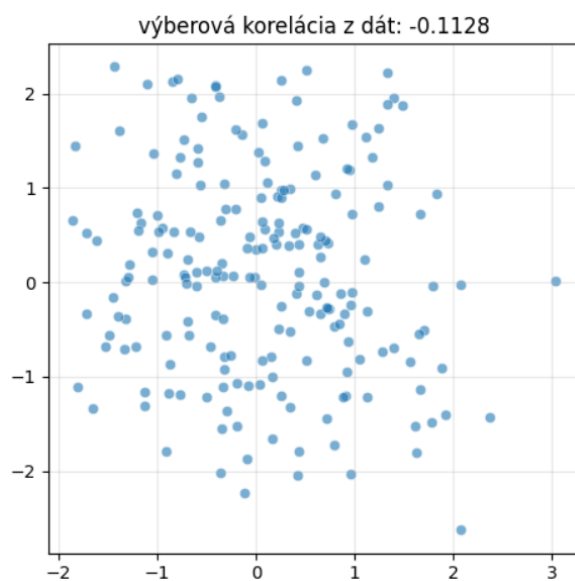
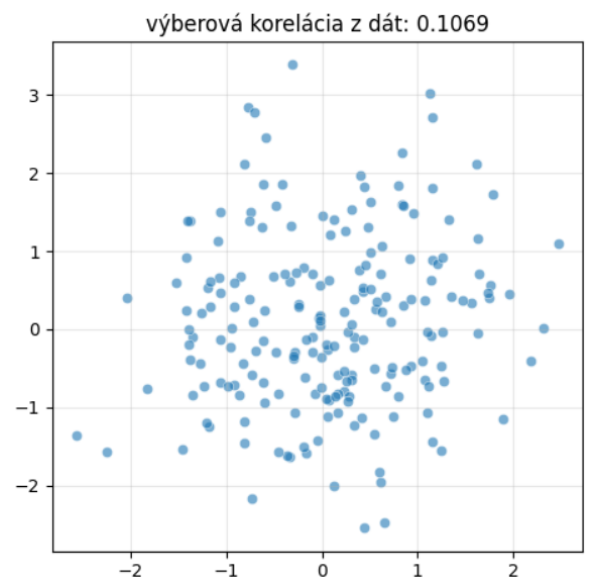
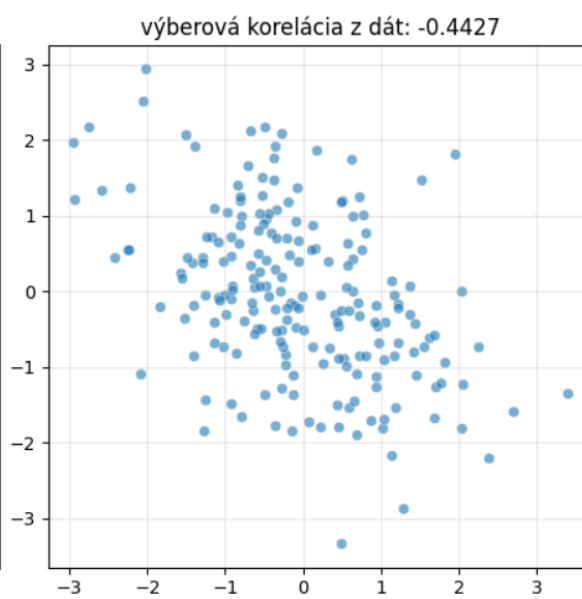
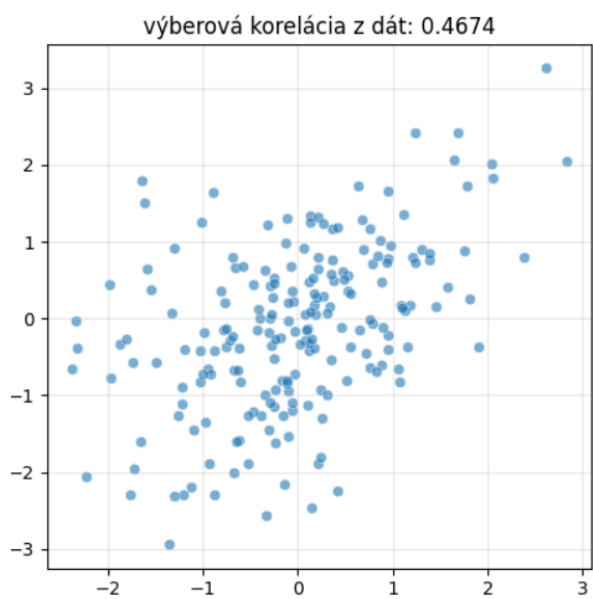
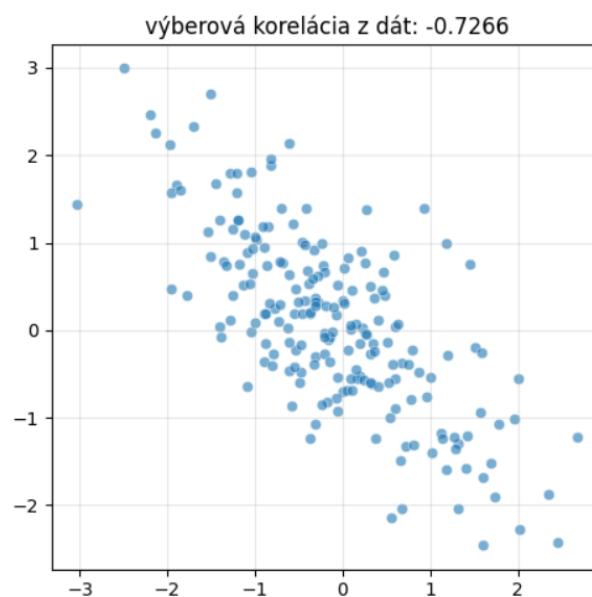
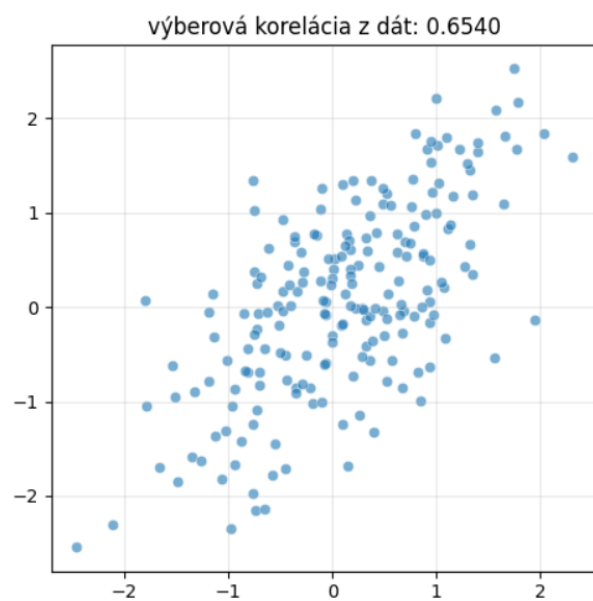
korelacia = df["shoe_size"].corr(df["height"])
print(f"{korelacia:.5f}")
```

0.85635

Vlastnosti korelácie:

- Korelácia nadobúda **hodnoty medzi -1 a +1** (vrátane).
- Je mierou **lineárnej závislosti**.
 - **Znamienko určuje rastúcosť (+), resp. klesajúcosť (-)**
 - **Absolútna hodnota určuje silu závislosti**





Takže keď nám **seaborn** nakreslí priamku do dát, korelácia hovorí o tom, ako dobre táto priamka vystihuje dáta.

Vlastnosti korelácie - pokračovanie:

Z vlastností kovariancie vyplýva:

- Ak sú náhodné premenné nezávislé, tak majú nulovú koreláciu.
- Z nulovej korelácie ešte nevyplýva, že sú náhodné premenné nezávislé.

Platí však (zdôvodniť to budeme vedieť, keď budeme hovoriť o náhodných vektoroch s normálnym rozdelením) :

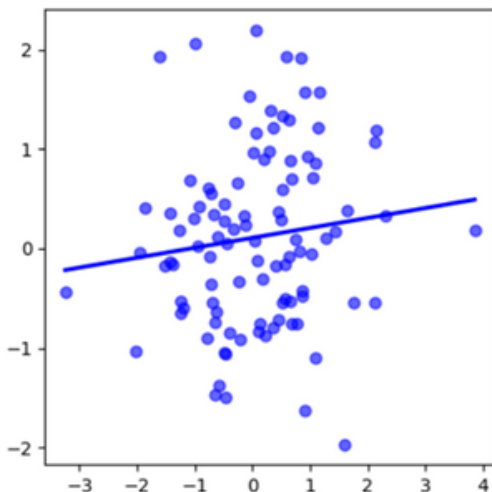
- Ak majú náhodné premenné normálne rozdelenie, tak z nulovej korelácie vyplýva, že sú náhodné premenné nezávislé.

5. Testovanie, či je korelácia nulová (normálne rozdelené dáta)

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html>

```
korelacia_blue = np.corrcoef(X_blue, Y_blue)[0, 1] # prvok korelacionej matice
print(f"{korelacia_blue:.5f}")
```

0.12148



Názov funkcie znamená "Pearson r", čo je Pearsonov korelačný koeficient – tá korelácia, s ktorou sme doteraz pracovali.

Nulová hypotéza je, že korelácia sa rovná nule.

Testovacia štatistika hodnota odhadnutej korelácie dát, p-hodnota sa počíta z pravdepodobnostného rozdelenia odhadnutej korelácie, ak je skutočná hodnota nulová.

Vidíme, že v tomto prípade sa nulová hypotéza nezamieta (na štandardnej hladine významnosti 0,05).

```
from scipy.stats import pearsonr
```

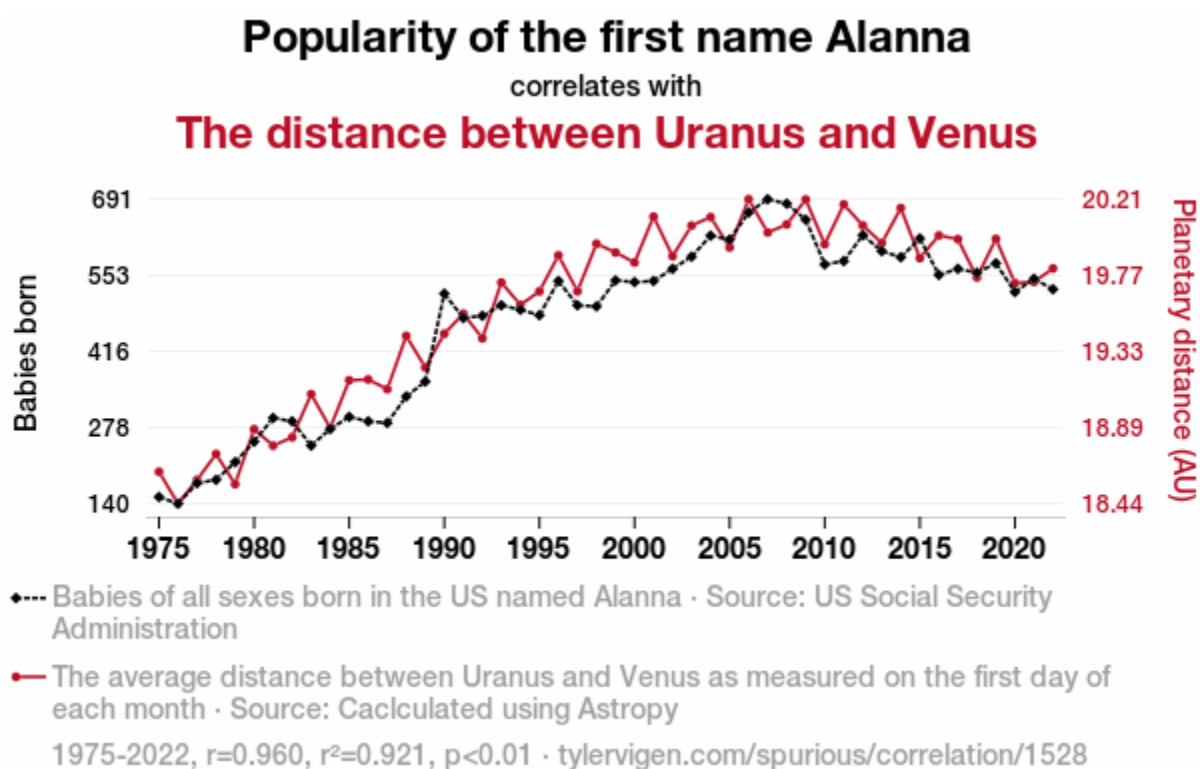
```
corr, p_value = pearsonr(X_blue, Y_blue)
print(f"Pearson r = {corr:.4f}, p = {p_value:.4f}")
```

Pearson r = 0.1215, p = 0.2286

Korelácia meria mieru **lineárnej závislosti**. Teda:

- Nemeria hocijakú závislosť (videli sme príklad, keď rovnomerné rozdelenie a jeho druhá mocnina mali nulovú koreláciu).
- Nemeria ani hocijakú monotónnu závislosť (na to existujú modifikácie korelačného koeficientu, to je aj dôvod, prečo pri testovaní hypotézy v predchádzajúcom príklade je *Pearsonov* korelačný koeficient – na odlíšenie od korelačného koeficientu, ktorý meria silu ľubovoľnej *monotónnej* závislosti).

6. Korelácia neznamená kauzalitu



Viac príkladov na <https://www.tylervigen.com/spurious-correlations>