Poznámky ke střední hodnotě

A study of whether cancer pamphlet information is written at an appropriate level to be read and understood by cancer patients.

The data consist of a sample of 63 patients whose reading level was determined and a sample of 30 pamphlets whose readability level was assessed on the same scale.

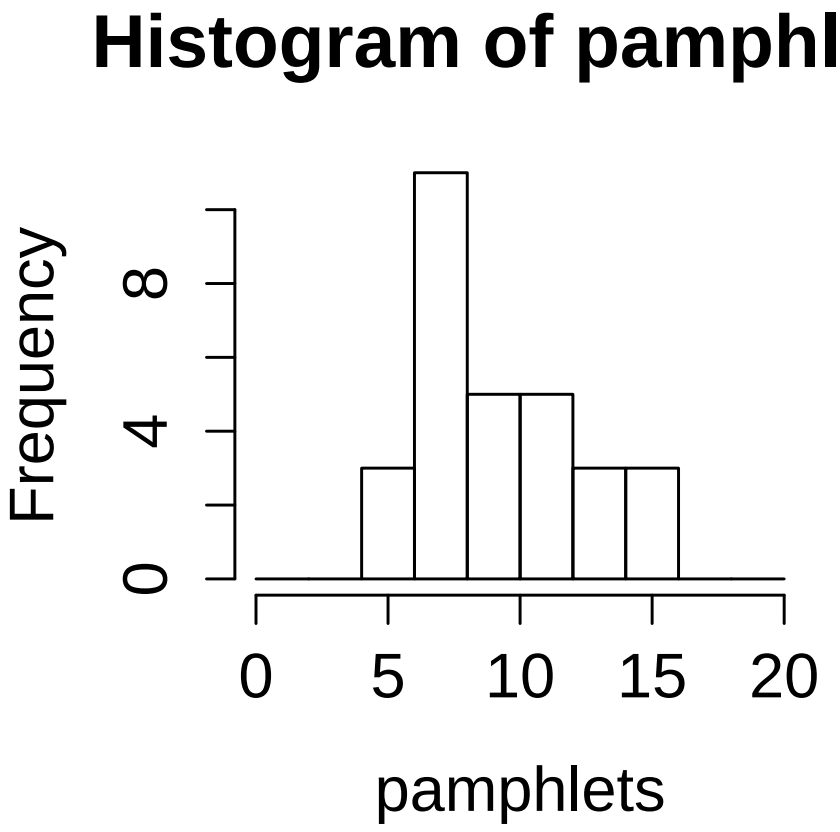
průměr (letáky) = 9.8; medián (letáky) = 9

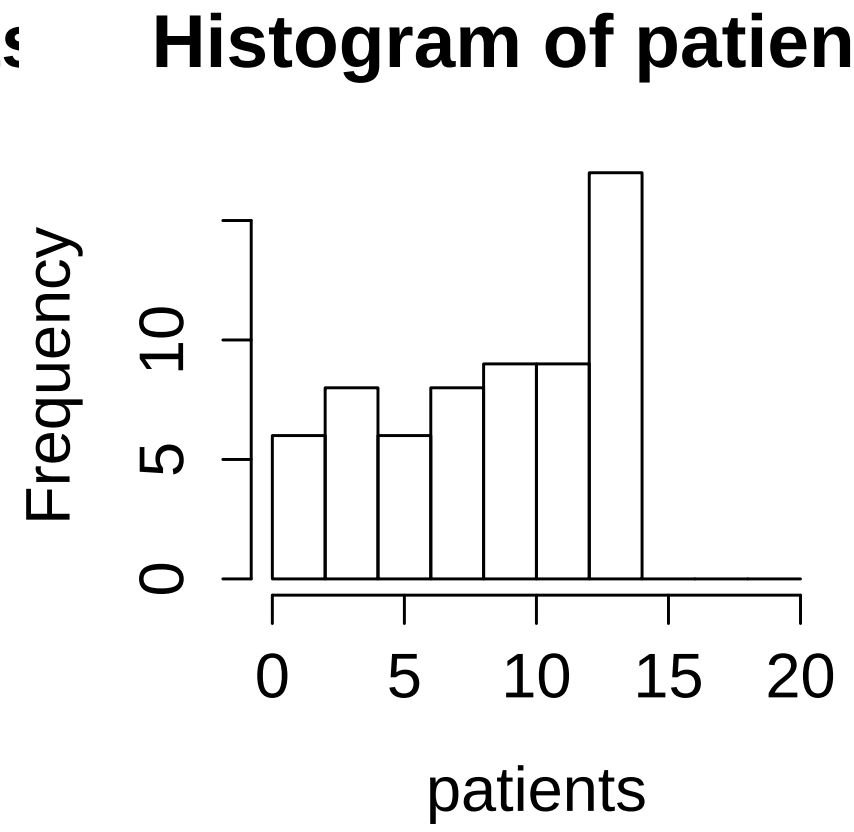
průměr (pacienti) = 8.6; medián (pacienti) = 9

Srovnej: *průměrný plat.*

K diskusi: Odpovídá složitost letáčků potřebám a dovednostem pacientů?

Poznámky ke střední hodnotě





Poznámky ke střední hodnotě – rozptyl

Míra polohy: $\mathbb{E}X$

Míra variability: $\mathbb{E}\,(X - \mathbb{E}X)^2$

rozptyl (letáky) = 8.5

rozptyl (pacienti) = 15.1

Síla závislosti mezi náhodnými veličinami

Sílu *lineárního vztahu* mezi dvojicí náhodných veličin X, Y měříme pomocí *korelačního koeficientu*:

$$\rho = \frac{\mathbb{E}\,(X - \mathbb{E}X)\,(Y - \mathbb{E}Y)}{\sqrt{\mathbb{E}\,(X - \mathbb{E}X)^2}\,\sqrt{\mathbb{E}\,(Y - \mathbb{E}Y)^2}}$$

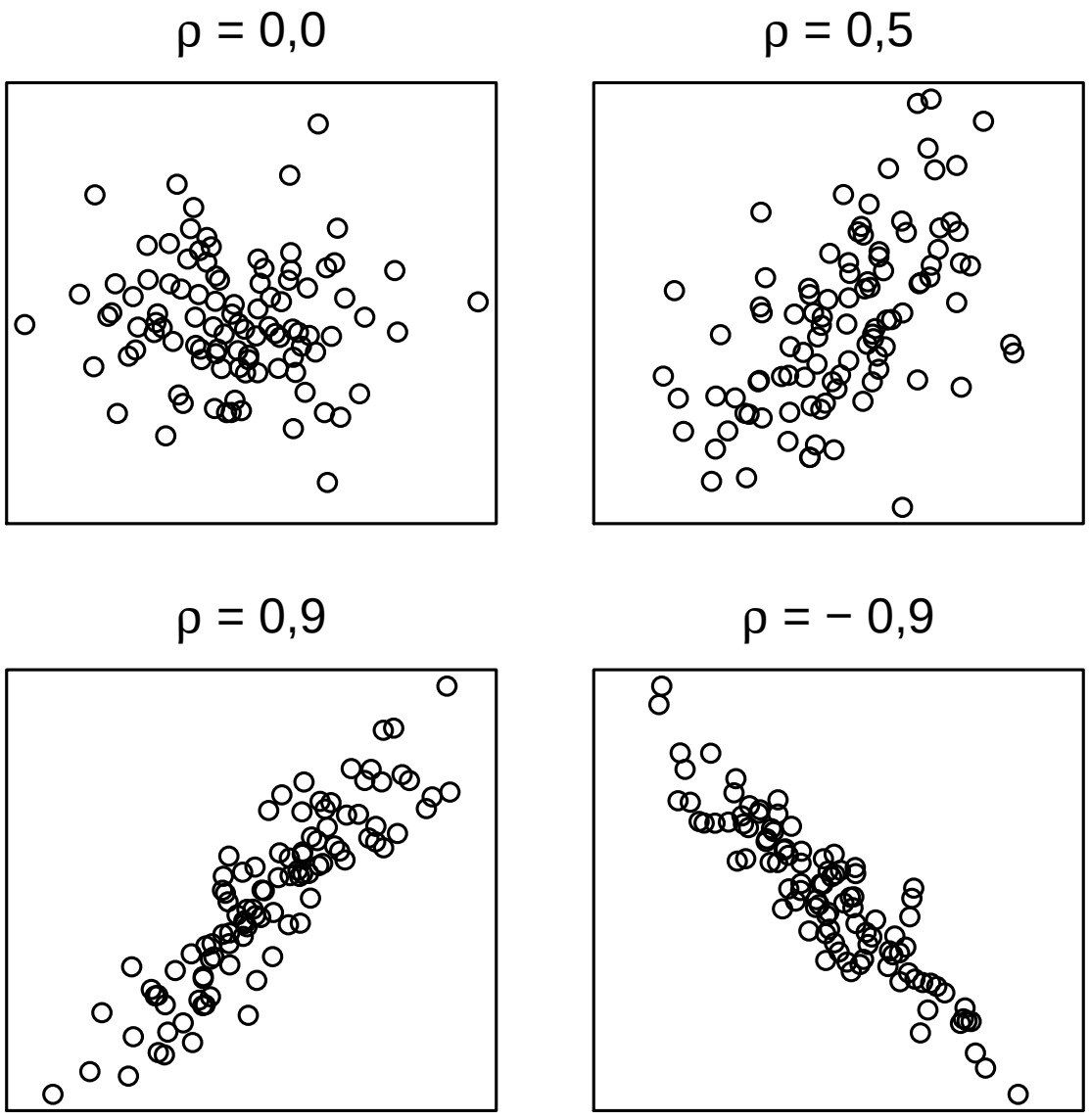
$\rho \in [-1; 1]$

Perfektní lineární závislost typu $Y = a + bX$, $b > 0 \Rightarrow \rho = 1$,

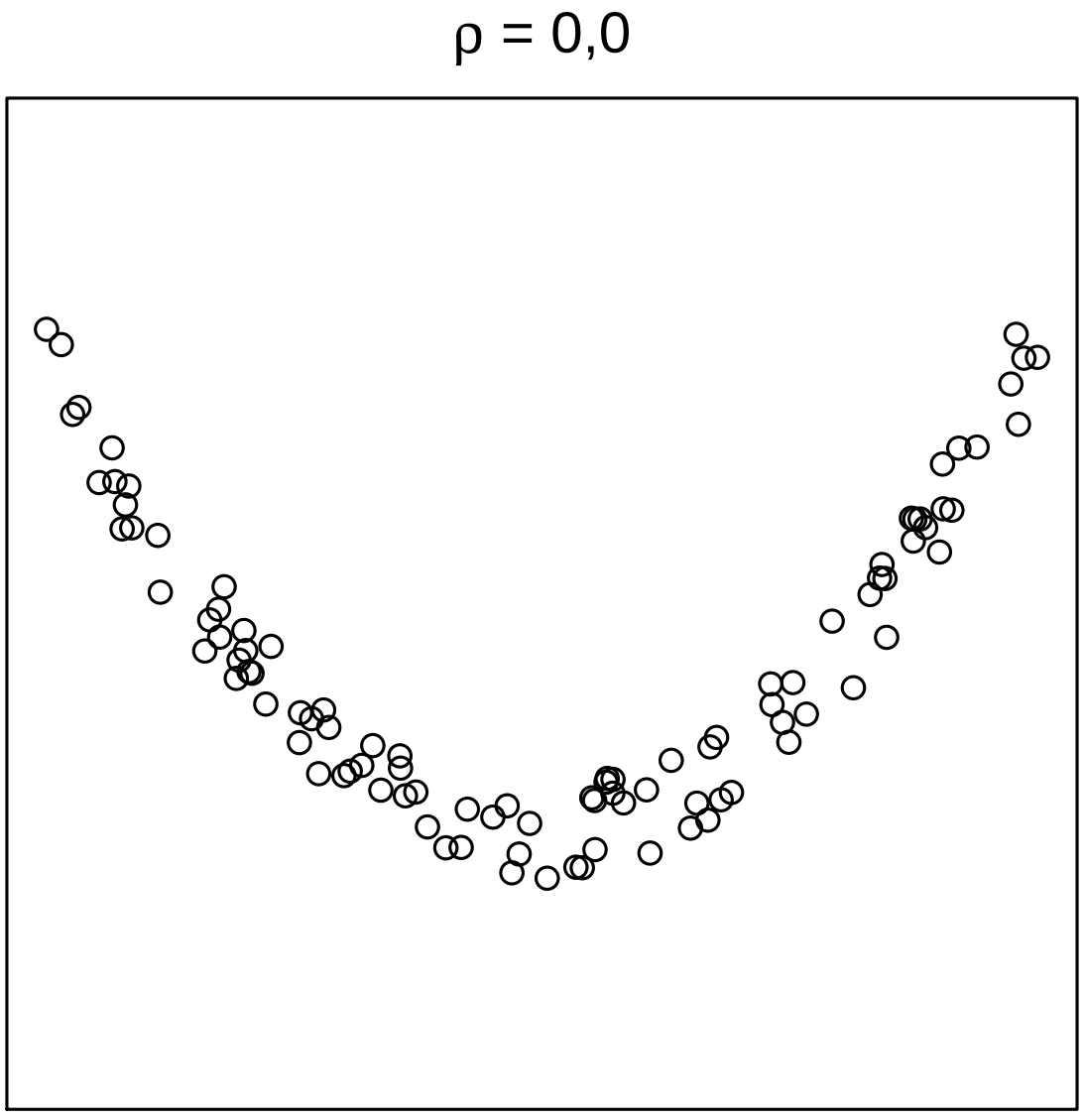
perfektní lineární závislost typu $Y = a + bX$, $b < 0 \Rightarrow \rho = -1$,

nezávislost náhodných veličin X, $Y \Rightarrow \rho = 0$.

Ukázka 1



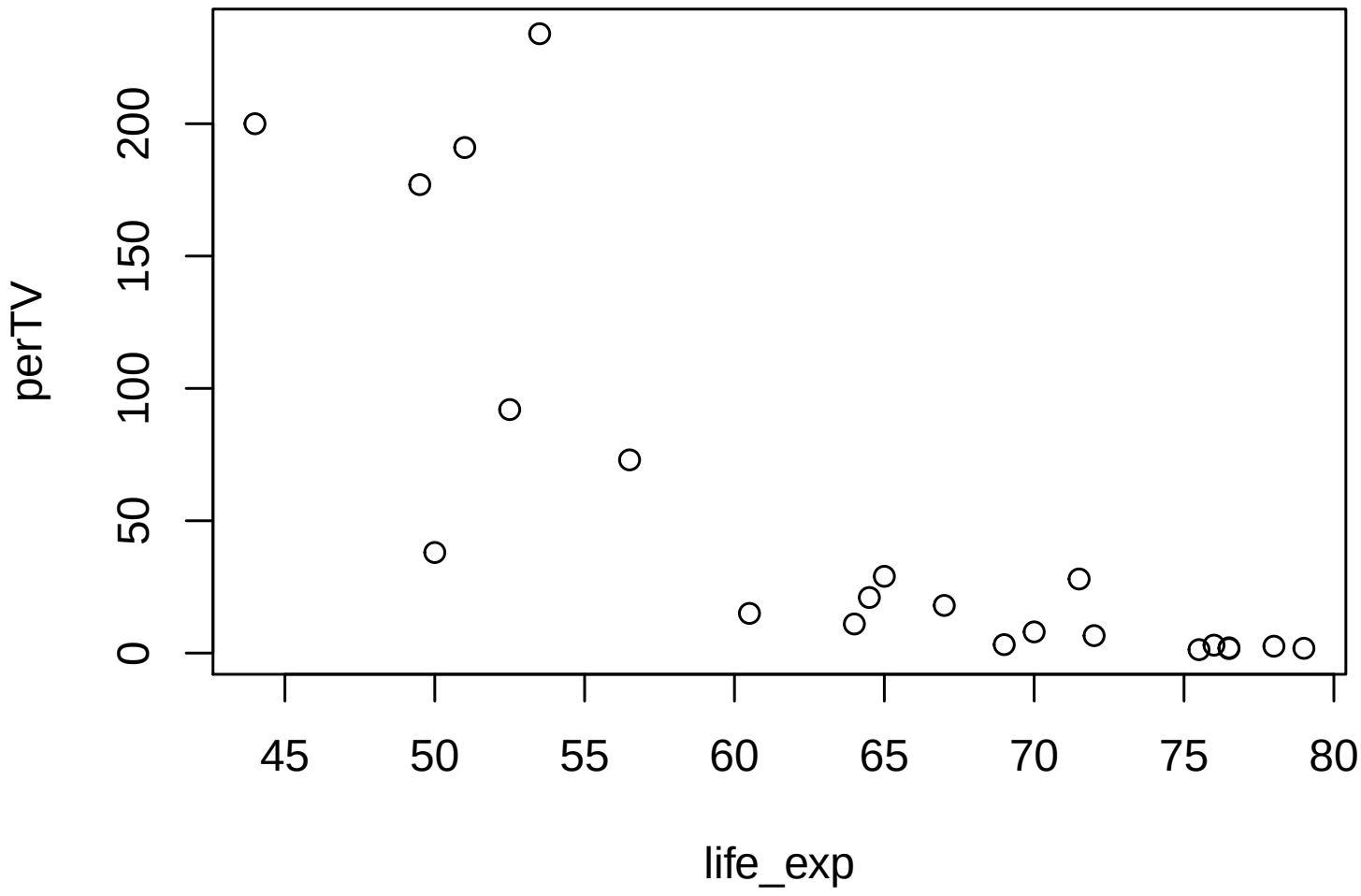
Ukázka 2



Snadné příklady

- Za větší balení čokolády zaplatím více peněz.
- Vysocí lidé váží více.
- Vyšší rychlost větru znamená vyšší výkon větrné elektrárny.

- Starší muži mají méně vlasů.
- Čím vyšší sněhová pokrývka, tím méně aut na silnicích.
- Čím jedu pomaleji, tím déle mi cesta potrvá.



Očekávaná délka života vs. počet obyvatel na jednu televizi ($\widehat{\rho} \doteq -0{,}80$; p-hodnota $7 \cdot 10^{-6}$, viz příští seminář).

Hlasovací otázka 11

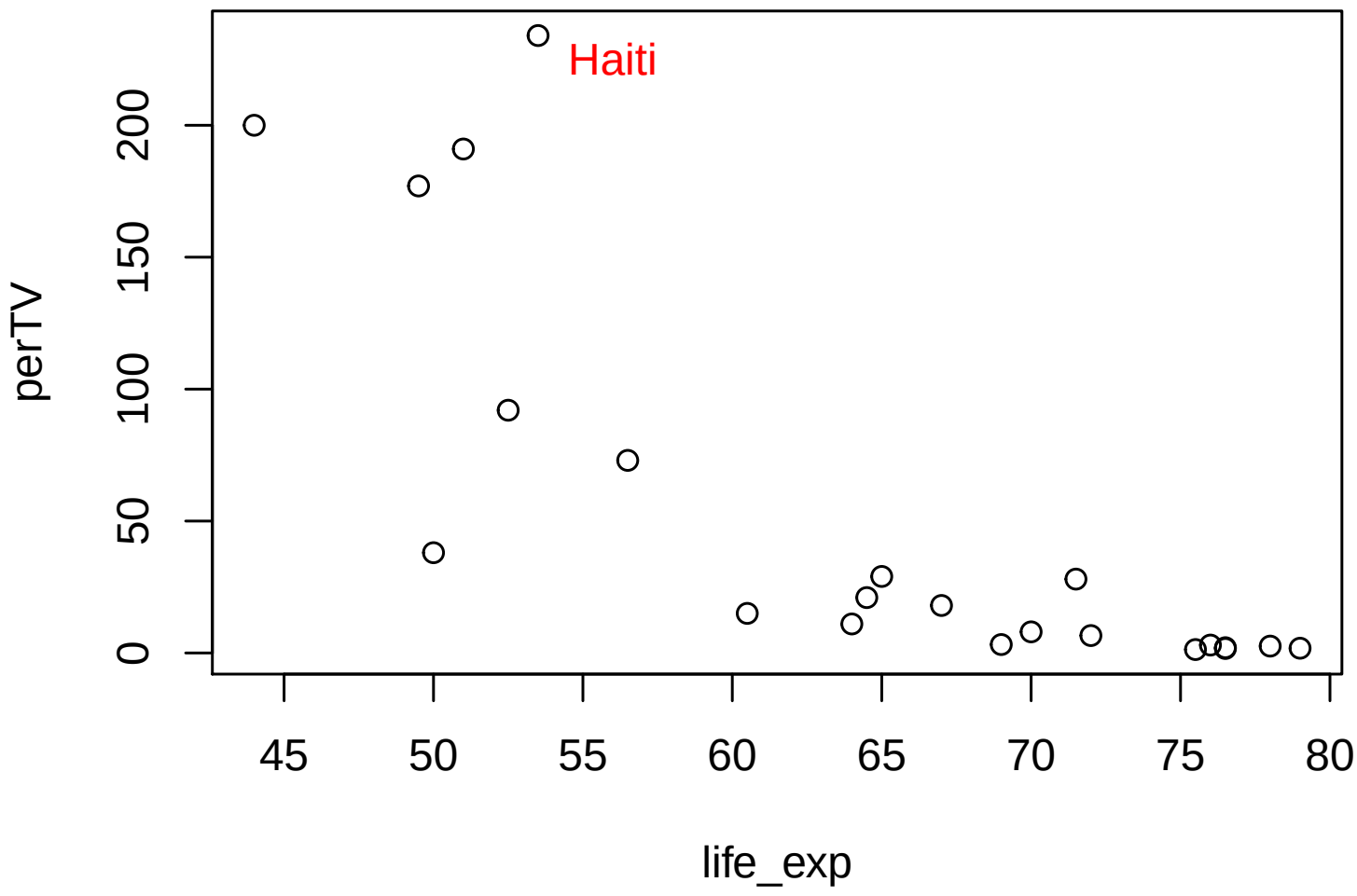
Jak spolu souvisí očekávaná délka života v dané zemi a počet obyvatel na jednu televizi?

A) Vysoký počet obyvatel na jednu televizi snižuje očekávanou délku života,

B) vysoká očekávaná délka života snižuje počet obyvatel na jednu televizi,

C) veličiny spolu souvisí jinak,

D) veličiny spolu nesouvisí.



Obrázek: Očekávaná délka života vs. počet obyvatel na jednu televizi.

Nejdůležitější slide dnešního semináře

Ne každá (statisticky) významná korelace je (věcně) důležitá!

Změny hodnot jedné veličiny nemusí způsobovat/vysvětlovat změny hodnot druhé veličiny.

Varovné příklady I.

Počet titulů Ph.D. udělených v matematických oborech ve Spojených státech:

- kladně koreluje s množstvím uranu uskladněným v jaderných elektrárnách tamtéž ($\widehat{\rho} > 0{,}95$),[1]
- záporně koreluje s roční spotřebou plnotučného mléka na jednoho obyvatele USA ($\widehat{\rho} < -0{,}94$).[2]

Počet filmů (za rok), ve kterých se objevil Nicolas Cage:

- kladně koreluje s počtem lidí (opět za rok), kteří utonuli po pádu do bazénu ($\widehat{\rho} > 0{,}66$),[3]
- negativně koreluje s počtem lidí, kteří utonuli po pádu z rybářské lodi ($\widehat{\rho} < -0{,}54$).[4]

---

[1] `http://tylervigen.com/view_correlation?id=1100`
[2] `http://tylervigen.com/view_correlation?id=1103`
[3] `http://tylervigen.com/view_correlation?id=359`
[4] `http://tylervigen.com/view_correlation?id=10037`

Varovné příklady II.

- Čím více hasičů dorazí k požáru, tím větší je celková škoda.
- Počet utonutí v jednotlivých měsících silně koreluje s celkovým prodejem zmrzlinářských výrobků.
- U žáků prvního stupně ZŠ je kladná korelace mezi úrovní čtenářských dovedností a velikostí bot.
- Velikost dlaně je záporně korelovaná s očekávanou délkou života dané osoby.

Příklady – srovnání

Varovné příklady I. a II. mají rozdílnou povahu:

I. falešné korelace (náhoda, vybíráme z obrovského množství dat),

II. vliv jiné veličiny.

Snadné příklady odpovídaly kauzální vazbě mezi veličinami.

Snadné příklady (připomenutí)

- Za větší balení čokolády zaplatím více peněz.
- Vysocí lidé váží více.
- Vyšší rychlost větru znamená vyšší výkon větrné elektrárny.

- Starší muži mají méně vlasů.
- Čím vyšší sněhová pokrývka, tím méně aut na silnicích.
- Čím jedu pomaleji, tím déle mi cesta potrvá.

Shrnutí

Pozorovaná korelace může být způsobena:

- přímým kauzálním vztahem,
- vlivem jiné veličiny (zprostředkovaný vztah),
- náhodou (nepřítomnost kauzálního vztahu).

Hodnota korelačního koeficientu nedokáže prokázat (ne)přítomnost kauzálního vztahu.

Pro odhalování kauzálních vztahů bychom potřebovali provádět kontrolované experimenty, v nichž je možné cíleně měnit hodnotu jedné veličiny a sledovat změny hodnot veličiny druhé, při zachování všech ostatních vlivů beze změny.