

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.

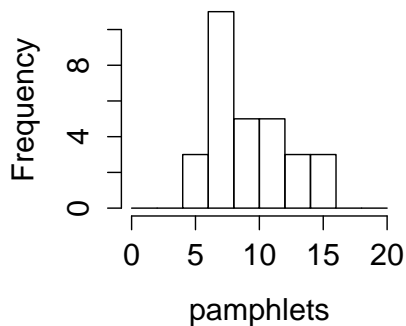
průměr (letáky) = 9.8; medián (letáky) = 9

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

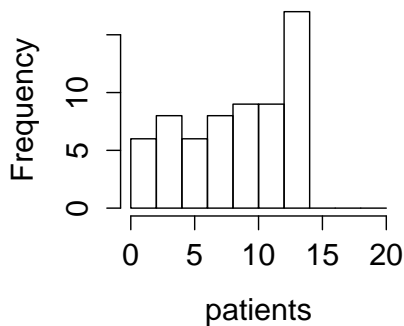
Srovnej: *průměrný plat.*

K diskusi: Odpovídá složitost letáčků potřebám a dovednostem pacientů?

Histogram of pamphlets



Histogram of patients



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*:

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

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

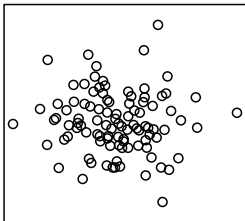
Perfektní lineární závislost typu $Y = a + bX$, $b > 0 \Rightarrow \rho = 1$,

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

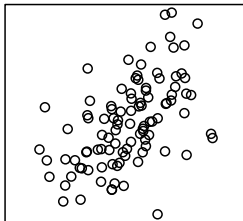
nezávislost náhodných veličin $X, Y \Rightarrow \rho = 0$.

Ukázka 1

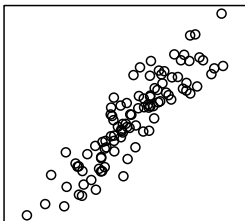
$\rho = 0,0$



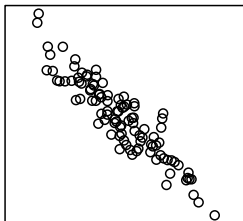
$\rho = 0,5$



$\rho = 0,9$

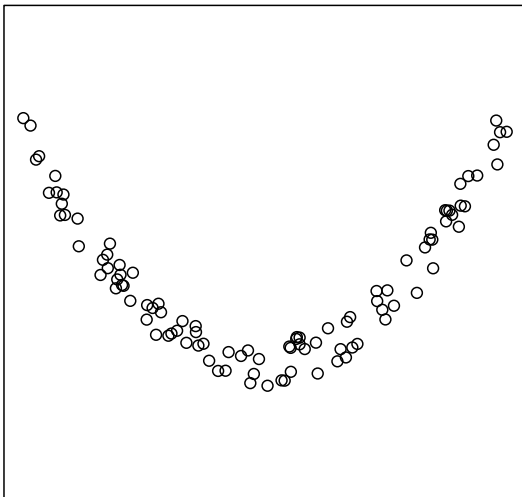


$\rho = -0,9$



Ukázka 2

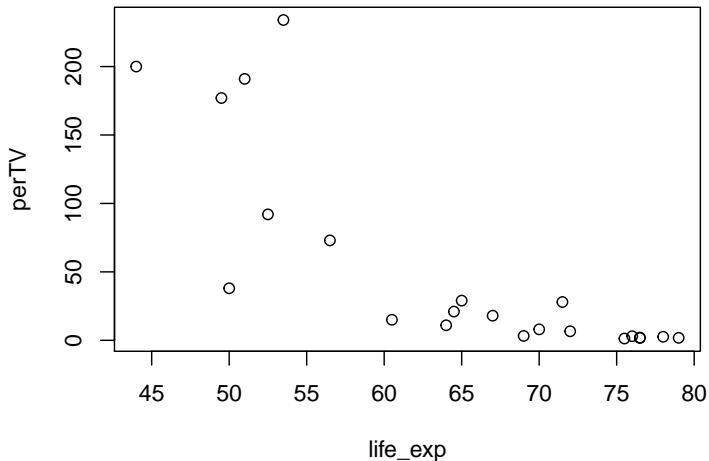
$\rho = 0,0$



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.

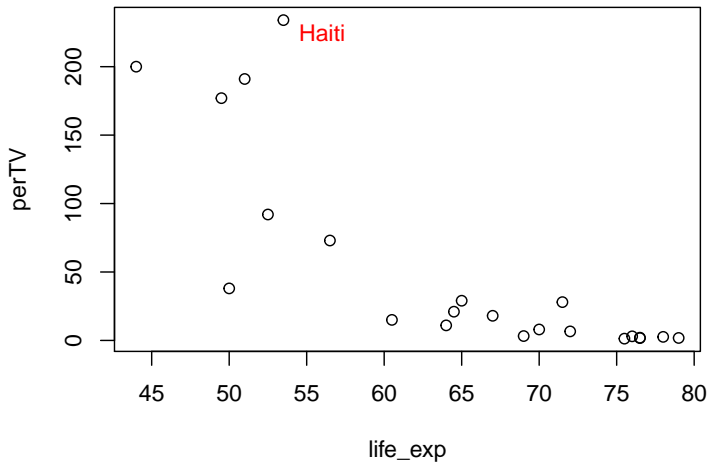


Očekávaná délka života vs. počet obyvatel na jednu televizi
($\hat{\rho} \doteq -0,80$; p-hodnota $7 \cdot 10^{-6}$, viz příští seminář).

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.

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

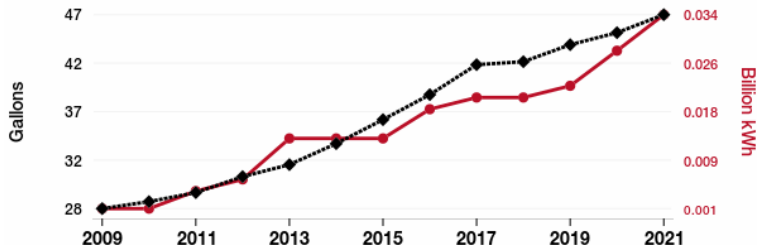
Změny hodnot jedné veličiny nemusí způsobovat/vysvětlovat změny hodnot druhé veličiny.

Varovné příklady I.

US Bottled Water Consumption per Person

correlates with

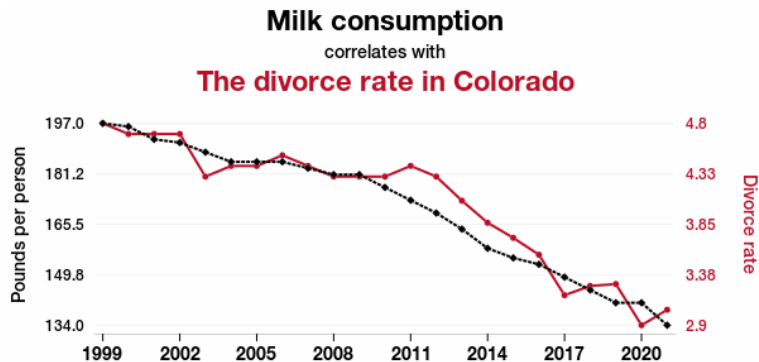
Solar power generated in Sudan



◆ US Bottled Water Consumption per Person · Source: Statista

● Total solar power generated in Sudan in billion kWh · Source: Energy Information Administration

2009-2021, $r=0.967$, $r^2=0.935$, $p<0.01$ · tylervigen.com/spurious/correlation/1898



◆ Per capita consumption of Fluid beverage milk in the US · Source: USDA

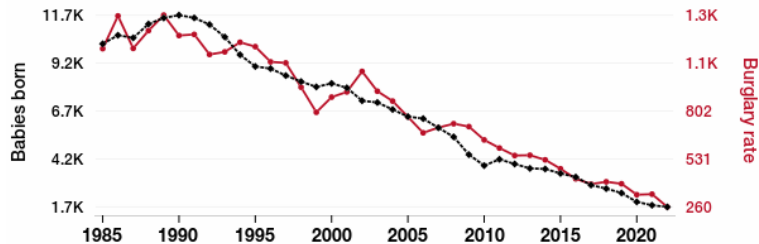
■ The divorce rate in Colorado · Source: CDC National Vital Statistics

1999-2021, $r=0.965$, $r^2=0.932$, $p<0.01$ · tylervigen.com/spurious/correlation/1038

Popularity of the first name Katherine

correlates with

Burglaries in Hawaii



●--- Babies of all sexes born in the US named Katherine · Source: US Social Security Administration

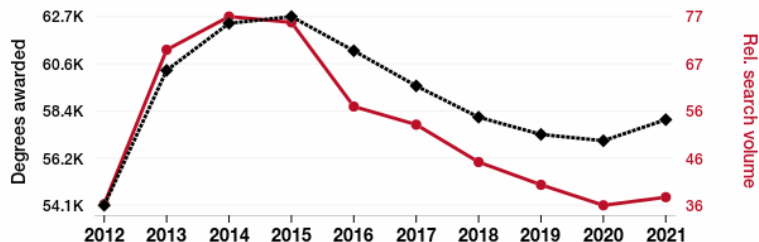
— The burglary rate per 100,000 residents in Hawaii · Source: FBI Criminal Justice Information Services

1985-2022, $r=0.975$, $r^2=0.951$, $p<0.01$ · tylervigen.com/spurious/correlation/3951

Bachelor's degrees awarded in law enforcement

correlates with

Google searches for 'sleepwalking'



◆ Bachelors degrees conferred by postsecondary institutions, in field of study: Homeland security, law enforcement, and firefighting · Source: National Center for Education Statistics

● Relative volume of Google searches for 'sleepwalking' (Worldwide, without quotes) · Source: Google Trends

2012-2021, $r=0.903$, $r^2=0.815$, $p<0.01$ · tylervigen.com/spurious/correlation/1532

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.

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.

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.