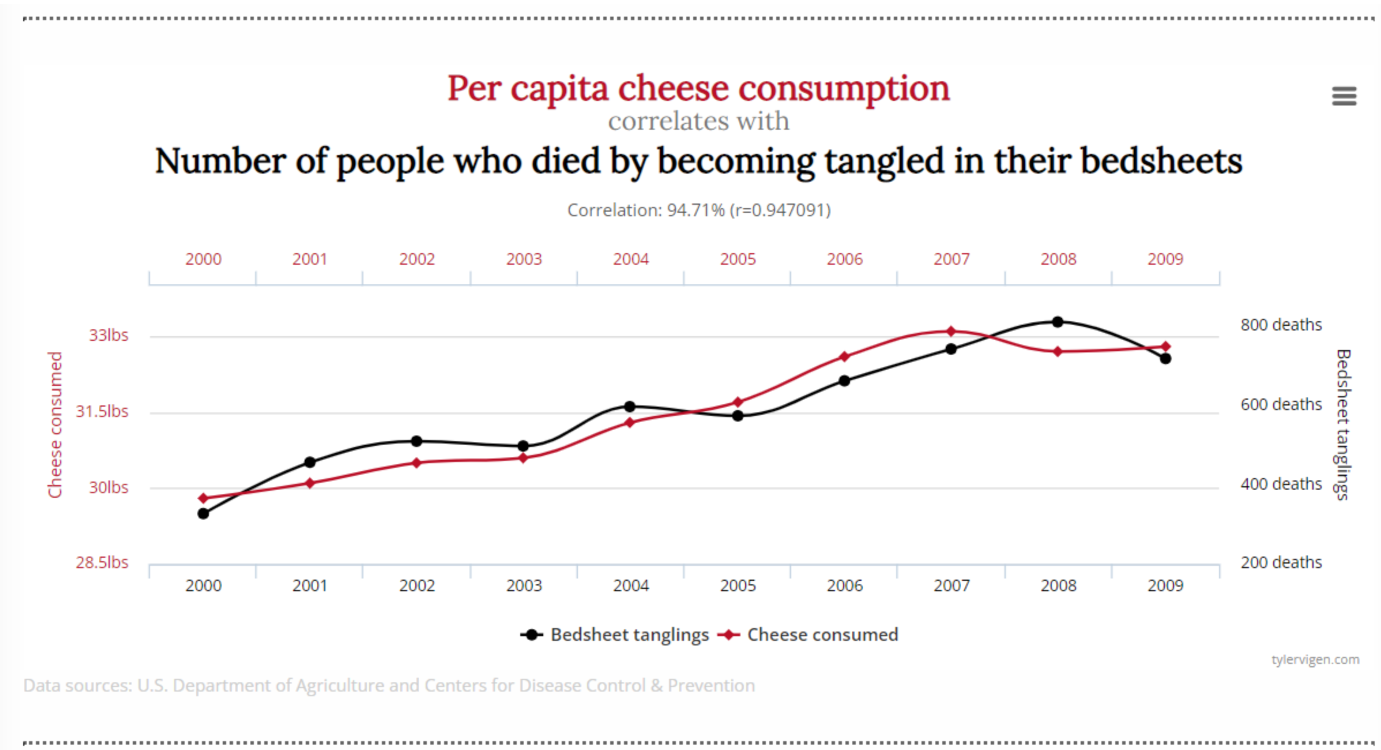
Exercise 7

Homework exercise

To be solved at home before the exercise session.

1. a. Go to the website which lists pairs of variables that have no causal relationship but still exhibit a large correlation. Pick one of the datasets and figure out how the data is presented, i.e., how are the plots constructed from the (x_i, y_i)-data (the plots are *not* scatter plots of the two variables in question), how are individual pairs (x_i, y_i) represented in the plots and what are the lines going through the points?



How the plots are constructed: There are two datasets: (xi, yi)1 represents (cheese consumption per capita, year) and (xi, y1)2 represents (Death by bedsheets, year). Then these two plots are merged together, creating a sense of correlation. The lines going through the points is the change of yi (death counts and chess consumption) with respect to the years

b. Let x, y, ε be random variables such that,

$$y = x + \varepsilon,$$

where $\mathrm{Var}(x) = 1$, $\mathrm{Var}(\varepsilon) = \sigma^2 > 0$ and x and ε are independent (interpretation: x and y have a perfect linear relationship but the observed value of y is contaminated with the noise/measurement error ε having variance σ^2). Compute the Pearson correlation ρ between x and y and investigate how it behaves when σ^2 is increased. Interpret this behavior.

We have: $var(x) = E[x^2] - E[x]^2 = 1 \Rightarrow \sigma_x = \sqrt{var(x)} = 1$

$var(\varepsilon) = E[\varepsilon^2] - E[\varepsilon]^2 = \sigma^2 \Rightarrow \sigma_\varepsilon = \sigma$

The Pearson correlation coefficient: $\rho(x,y) = \dfrac{cov(xy)}{\sigma_x \sigma_y} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

- We have: $var(y) = var(x+\varepsilon) = var(x) + var(\varepsilon) + 2\,cov(x,\varepsilon)$
$= 1 + \sigma^2$ ($cov(x,\varepsilon) = 0$ because independent)

$\Rightarrow \sigma_y = \sqrt{var(y)} = \sqrt{1+\sigma^2}$

- For $cov(x,y)$: $cov(x,y) = cov(x, x+\varepsilon)$

$\Rightarrow cov(x,y) = E[x(x+\varepsilon)] - E[x]E[x+\varepsilon]$
$= E[x^2] + E[x\varepsilon] - E[x]^2 - E[x]E[\varepsilon]$
$= E[x^2] - E[x]^2 + E[x]E[\varepsilon] - E[x]E[\varepsilon]$
$= var(x) = 1$ (Since x & ε are independent)

$\Rightarrow \rho(x,y) = \dfrac{cov(x,y)}{\sigma_x \sigma_y} = \dfrac{1}{1(\sqrt{1+\sigma^2})} = \dfrac{1}{\sqrt{1+\sigma^2}}$

$\Rightarrow$ If σ^2 is increased, $\rho(x,y)$ will decrease. This behavior means that the more varied the noise error ε, the less y is correlated with x.

Indeed: $\lim_{\sigma^2 \to \infty} \dfrac{1}{\sqrt{1+\sigma^2}} = 0 \Rightarrow$ if σ^2 is very big, x and y are not correlated at all