MS-C1620 Statistical Inference

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?

   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.

Class exercise

To be solved at the exercise session.

1. The file `data_dependency.txt` contains seven bivariate data sets (the columns `xi` and `yi`, where $i = 1, 2, \ldots, 7$, always form a pair).
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot for each pair of variables.
   c. Calculate the Pearson and Spearman correlations of the pairs and compare them to the scatter plots.
   d. The underlying distributions of the samples 5-7 are the same up to the variance of `yi` (the variance is highest in sample 7). What happens to the correlation coefficients as the variance increases and why?

2. The file `data_tobacco.txt` contains data on cigarette consumption and lung cancer incidences from 11 different countries. The variable `consumption` describes the yearly consumption of cigarettes per capita in 1930 and the variable `incidence` tells the lung cancer incidence rates per 100 000 people in 1950. We use correlation to study the connection between these two.
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot of `consumption` and `incidence` which also shows the country names.
   c. Using the scatter plot, make an educated guess on the signs and magnitudes of the Pearson and Spearman correlations of the two variables.
   d. Calculate the Pearson and Spearman correlations.
   e. Use permutation test to test whether the two correlations differ significantly from zero, using the significance level 5%.
   f. Drop USA from the data, redo the previous analysis and compare the results to those obtained with the full data. What happened?

3. **(Optional)** Use also the tests given on slides 6.16 and 6.20 to test the null hypothesis $H_0 : \rho = 0$ for Pearson correlation in problem 2e. How do the results compare to the permutation test?

4. **(Optional)** Simulate the distribution of the sample Pearson correlation $\hat{\rho}$ under normality by generating multiple datasets of size `n` from a bivariate normal distribution of your choice. Then transform the sample Pearson correlations as $\hat{\rho} \mapsto \operatorname{arctanh}(\hat{\rho})$ and inspect the distribution of the transformation. Does it look normal? (it should for large n, as per slide 6.13)