



In-class exercises

GRA 4142 Data Management and Python Programming

Jan Kudlicka (jan.kudlicka@bi.no)

Department of Data Science and Analytics

September 17, 2021

Use NumPy to solve the following set of equations:

$$x + y + z = 25$$

$$5x + 3y + 2z = 0$$

$$y - z = 6$$

1. Prepare a data set which we will use in the second step in a linear regression analysis.
 - ▶ Generate a column vector $\mathbf{x} = \{x_i\}$ consisting of 8 samples x_i drawn from a uniform distribution between 0 and 10.
 - ▶ Generate a column vector $\mathbf{y} = \{y_i\}$, where $y_i = 20 + 2x_i + \epsilon_i$, where ϵ_i denotes noise drawn from a Gaussian distribution with zero mean and standard deviation 1.5.

Plot the points (x_i, y_i) in a figure together with the line $y = 20 + 2x$.

2. Use the normal equation

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

to estimate the coefficient vector of the least-squares line. In the equation, $\mathbf{X}$ denotes a matrix that consists of two columns: the elements in the first column are all ones, the second column is the vector $\mathbf{x}$.

Add the least-squares line $y = \beta_0 + \beta_1 x$ to the figure (use a different color).

Given a group of n people, what is the probability that at least two people in the group will celebrate the birthday on the same day? It might sound surprising that even in a small group with $n = 23$ people, the probability is over 50%. With $n = 70$ people, the probability is close to 99.9%.

Write a Python function that accepts n and estimates this probability for a group of n people using the Monte Carlo method. Specifically:

- ▶ Sample the birthdays of n people uniformly at random and check if there are at least two people who have birthday the same day.
- ▶ Repeat this experiment 100 000 times and estimate the probability as the number of how many times you found at least two people having birthday the same day divided by 100 000.