Roman Fries
Guillaume Joyet

Project Proposal

Analysis of the NHEFS Dataset using a differentially private Database

Group members: Roman Fries, Guillaume Joyet

Dataset: *NHEFS complete* (Complete-Data National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study). This dataset is available in the python package *causaldata*) and contains entries of 1556 individuals over 67 demographic, medical and lifestyle variables.

Goal of Analysis: We want to analyse predictors of a set of health factors, as well as consequences of lifestyle choices. Here are some example questions that we could include in our analysis:

- Are factors like education, marital status, income, etc. correlated with the probability of an individual being a smoker?
- Is being a smoker correlated with the risk of developing illnesses as asthma, cancer, etc.?
- Is quitting smoking correlated with a subsequent weight loss or gain?
- etc.

Privacy mechanism: We will run the analysis as if we were data analysts accessing the data only over a curator who returns differentially private queries. Said curator will use a combination of the Laplace and exponential mechanisms.

Libraries: apart from the typical python data science libraries (*NumPy*, *SciPy*, *Pandas*, etc.), we will only need the *causaldata* library.