# ATE in randomized controlled trials and observational studies

This Lab illustrates notions and concepts presented during the first class, with a focus on randomized controlled trials. We will work with both synthetic and real data.

- Synthetic data will allow you to discover and understand how a theory can be checked as data are simulated with a known generative process.
- You will also apply method on real data for which the data generation process in not under our control. Specifically, we will use the data from a randomized controlled trial, the Tennessee Student/Teacher Achievement Ratio (STAR) study. This RCT is a pioneering randomized study from the domain of education (Angrist et al. 2008), started in 1985, and designed to estimate the effects of smaller classes in primary school on learning. This experiment showed a strong payoff to smaller classes. Using methods detailed in class, you will compute the average effect of this public policy on the academic results with two different estimators.

## Recover the properties of $\widehat{\tau}_{DM}$ and $\widehat{\tau}_{OLS}$ with simulated data

You are provided with a script that produces for simulations of randomized controlled trials. You can find two functions, one following a linear generative model and one following a non-linear generative model. The data are generated following the two generative models,

$$Y = 3X_1 + 2X_2 - 2X_3 - 0.8X_4 + T(2X_1 + 5X_3 + 3X_4) + \varepsilon,$$

and

$$Y = 3X_1 + 2X_2^2 - 2X_3 - 0.8X_4 + 10T + \varepsilon,$$

where in both cases the random variables $(X_1, X_2, X_3, X_4, T, \varepsilon)$ are jointly independent. The covariates $(X_1, X_2, X_3, X_4)$ are both Gaussian with unit mean and unit variance and $T$ is a Bernoulli random variable with parameter $p$. The $\varepsilon$ is a Gaussian centered noise whose variance $\sigma^2$ has to be fixed.

A Python is provided you and you can use the two functions `linear_simulation` or `non_linear_simulation` to generate either the linear model or the non linear one.

**Question 1**
Compute the simple difference in mean (DM) estimator with confidence intervals for each of the simulations. Which estimated ATEs do you obtain?
Tips: The asymptotic distribution of the DM estimator is

$$\sqrt{n}(\widehat{\tau}_{DM} - \tau) \to \mathcal{N}(0, V_{DM})$$

with

$$V_{DM} = \frac{Var(Y_i(0))}{\mathbb{P}[T_i = 0]} + \frac{Var(Y_i(1))}{\mathbb{P}[T_i = 1]}$$

A asymptotic confidence interval can be given by:

$$\mathbb{P}\left[\tau \in \left(\tau - \phi(1-\alpha/2) \cdot \sqrt{\widehat{V}_{DM}/n}; \tau + \phi(1-\alpha/2) \cdot \sqrt{\widehat{V}_{DM}/n}\right)\right]$$

You can use a plug-in estimator for the variance estimating $\mathbb{P}(T = t)$ as $n_t/n$.
Remark: we advise you to wrap the difference-in-means implementation in a function for the rest of the lab.

**Question 2**

Which is the value of the ATE for the two simulations? Is it in accordance with what do you obtain?

**Question 3**

(3.1) Implement the OLS estimator $\widehat{\tau}_{OLS}$.
(3.2) Compute the 95% confidence intervals of $\tau$. You can use that

$$\sqrt{n}(\widehat{\tau}_{OLS} - \tau) \to \mathcal{N}(0, V_{OLS})$$

and deduce the following expression of the asymptotic variance

$$V_{OLS} = V_{DM} - (\beta_1 + \beta_0)^T var(X)(\beta_1 + \beta_0)$$

where the notations are those of the lecture.

**Question 4**

(4.1) Recover the properties of both estimators performing a simulation study. Do you recover the theoritical results? Why?
Tip: repeat 100 times the process, that is, estimate the ATE on for each simulation using the functions `linear_simulation()` and `non_linear_simulation()`, and store the results.
(4.2) Propose a vizualisation with boxplots.

## STAR data set: effect of class size reduction on children performances

We propose to consider the Tennessee Student/Teacher Achievement Ratio (STAR) trial. This RCT is a pioneering randomized study from the domain of education (Angrist et al. 2008), started in 1985, and designed to estimate the effects of smaller classes in primary school, on the results. Note that the questions are key public policy questions, even in France as you can see in this report from the DEPP (Direction de l'évaluation, de la prospective, et de la performance). The following code loads the STAR data with preprocess similar as in Kallus et al. 2018. The main outcome is computed as the average of three grades and denoted $Y$. The treatment, here class size, is denoted $T$. The preprocessing steps used in their paper are given below. The covariates are:

- `gender`
- `age` (in month)
- `g1freelunch` being the number of lunchs provided to the child per day, that can be considered as a proxy for a socio economic status
- `g1surban` the localisation of the school (inner city or rural)
- `ethnicity`

All these covariates are supposed to have an impact on the outcome, that is predict the children success.

One can keep only complete observations. But in practice, several methods exist to handle missing values, could it be through an imputation procedure, or using prediction methods that can deal with missing values (such as forest). You can explore the pandas library to understand more

**Question 1**

(1.1) Explore the data. Which type of data do you observe? In Python you might want to use the following pandas methods: `head()` to peek into data as well as the `info()`, `describe()` to get summary statistics.

(1.2) Since we are working with a RCT, we are concerned with the appropriate covariates balance between the two groups of children. Propose a visualization to highlight possible imbalances. Tips: for example, you can overlay the distribution of age between both groups or repartition of children receiving or not lunch to control the balance. In Python you can use the `seaborn` library.

**Question 2**

Draw a categorical plot, to represent the mean of each value per treatment group.

**Question 3**

To mimic the habits of econometrics and medical papers, first propose a so-called Table 1 of the different covariates present in the data set. In Python you can use the `tableone` library (see `https://pypi.org/project/tableone/` for details).

**Question 4**

(4.1) Estimate the ATE with the two estimators and give confidence intervals. What do you oberve?

(4.2) For the DM estimator, you can compute the intervals based on the asymptotic variance vs those based on the bootstrap.

Note that in next classes we will see how to study heterogeneity in the effect, which is a key information for any policy maker or when it comes to personalized medicine.