# CAUSAL INFERENCE:
## Using statistics to explain cause and effect

## COURSE DESCRIPTION

The basic idea of causal inference is to think about a question (theory, hypothesis) by keeping in mind the following three things: (1) potential outcomes, (2) selection bias, and (3) random assignment into treatment and control groups.

For example, what is the effect of going to college on earnings later in life? The fundamental problem of causal inference is that you can only observe one **potential outcome** for one set of actions. You can either go to college or not, you can't do both.

Can we observe two students from the same generation where one goes to college and the other doesn't? No, this is **selection bias**. We cannot make comparisons between different people who go or don't go to college because there could be a whole number of differences between them that would make their outcomes (in this case earnings) different regardless of being at college (students with better high school grades, greater working capacity and willingness to work, parental wealth and family background, etc.).

How do we eliminate selection bias when comparing different choices of different people? We **randomize**! How? And why? This is what you will learn in this course.

This course will teach you the value of using statistical methods to make and evaluate causal claims. After this course you will be able to: distinguish between correlation and causality (if A happens before B, it must be that A *caused* B?!); do your own experiments (you will be encouraged to do so); figure out whether an economic policy really works (e.g., does mandatory health insurance improve health outcomes?); and tell when a research paper in the social sciences is actually good science in that it *proves* its hypothesis and when it is just a good guess.

In one sentence, **this course will change the way you think about cause and effect.**

## PREREQUISITES

Students are strongly recommended to be familiar with at least undergraduate level statistics. Usually this course follows introductory lectures in statistics. It is recommended for students of economics, political science, and business, but it is for anyone who wants to do empirical research in the social sciences and anyone interested in public policy.

## COURSE LECTURER

**Vuk Vuković** has a **DPhil in political economy from the University of Oxford**. His research is focused on uncovering and reducing local government corruption and showing how the collusion between politicians and corporate executives (an interaction defined as elite networks) affects the distribution of top incomes. He taught methods courses (causal inference, advanced statistics, econometrics), as well as basic economics courses (econ 101, international econ, public finance, political economy). He holds an MSc from the London School of Economics, was a visiting scholar at Cambridge in 2013, and did summer schools at Harvard and Berkeley in 2010 and 2009. He is happily married and a father of three.

He is also the CEO and co-founder of [Oraclum Intelligence Systems Ltd](), a UK-based data company that uses the power of social networks, big data, and machine learning to predict election outcomes, market movements, product demand, and consumer behavior.

Reach him at [https://www.vukvukovic.org/](https://www.vukvukovic.org/) or via email at [vuk@oraclum.co.uk](mailto:vuk@oraclum.co.uk)

## COURSEWORK AND ASSESSMENT

The course has one final exam worth 60% of the grade. The exam will have four essay questions in which students will be given a research design from different social science papers (economics and political science mostly) and they will be asked to replicate the methodology of the paper (datasets will be provided) and offer their own methodological assessment.

The class will meet twice a week, for lectures and for seminars. Seminars will require usage of a computer room, but students are encouraged to bring their laptops. On the seminars we will use R-studio and Stata to replicate some of the papers we discuss in class. Students will read the papers in advance. Seminar attendance provides a great way for students to "learn while doing". In other words seminars will enable the students to fully understand how to perform each new methodology we will cover in class.

In addition to the take-home exam each student is expected to submit a proposal (research design) for their own experiment (survey or field) or a research idea they would like to test empirically. A lot of emphasis will be placed on how to do this correctly and how to draft up a correct research design. This part will carry 40% of the final grade.

## COURSE TOPICS

1. Randomized and natural experiments
2. Regression and covariate adjustment with matching
3. Instrumental variables
4. Regression discontinuity design
5. Differences-in-differences design and panel data
6. Evaluating the plausibility of as-if random
7. Heterogenous treatment effects
8. How to design an experiment?

# COURSE OUTLINE AND READING LIST

## WEEK 1 & 2

**Randomized and natural experiments** – the first part introduces the concepts and shows the logic of why randomization is good to solve the selection bias issue. We will use potential outcome notation (Rubin-Neyman model); explain average treatment effects, selection bias, randomization, use many examples of true randomized experiments. Then we turn to natural experiments and talk about how to design optimal experiments.

Recommended:

- Dunning (2012) (chapters 1 and 2)
- Angrist & Pischke (2015): Mastering Metrics (chapter 1)
- Angrist & Pischke (2009): Mostly Harmless Econometrics (chapters 1 and 2)

Papers we will use or replicate in class:

- Finkelstein et al (2012) "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics*
- Chattopadhyay and Duflo (2004) "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India". *Econometrica*
- Dehejia and Wahba (1999) "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*

## WEEK 3 & 4

**Regression and matching** – why use regressions? What is and why to use OLS (we will not focus on asymptotics but rather on applications); Covariate adjustment with matching (a model-based approach to generate equal and thus comparable treatment and control groups)

Recommended:

- Angrist & Pischke (2015): Mastering Metrics (chapter 2)
- Angrist & Pischke (2009): Mostly Harmless Econometrics (chapter 3)

Papers we will use or replicate in class:

- Eggers and Hainmueller (2009) "MPs for Sale?", *American Political Science Review*.
- Kocher, Pepinsky, & Kalyvas (2011) "Aerial Bombings and Counterinsurgency in the Vietnam War." *American Journal of Political Science*

## WEEK 5

**Instrumental variables (IV)** – imposing (as-if) randomization by finding some variable that is allocated/distributed randomly and predicts the behavior of your independent variable

Recommended:

- Angrist & Pischke (2015): Mastering Metrics (chapter 3)
- Dunning (2012) (chapters 4 and 5)
- Angrist & Pischke (2009): Mostly Harmless Econometrics (chapter 4)

Papers we will use or replicate in class:

- Angrist (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery." *American Economic Review*
- Acemoglu, Johnson, & Robinson (2001) "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*
- Nunn (2008) "The Long-Term Effects of Africa's Slave Trade" *Quarterly Journal of Economics*
- Card (1995) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling"

## WEEK 6

**Regression discontinuity design (RDD)** – imposing (as-if) randomization by taking advantage of a clear exogenous rule that separates observations into treatment and control groups

Recommended:

- Angrist & Pischke (2015): Mastering Metrics (chapter 4)
- Dunning (2012) (chapters 3 and 5)
- Angrist & Pischke (2009): Mostly Harmless Econometrics (chapter 6)

Papers we will use or replicate in class:

- Lee (2008) "Randomized Experiments from Non-random Selection in US House Elections" *Journal of Econometrics*
- Brollo, Nannicini, Perotti, and Tabellini (2009) "The Political Resource Curse" *American Economic Review*
- Vuković (2019) "Corruption and Re-election: How Much Can Politicians Steal Before Getting Punished." *Journal of Comparative Economics*

**WEEK 7**

**Differences-in-differences (DID) and panel data** – exploiting the element of time by looking at how changes in behavior occur before and after certain exogenous shocks

Recommended:

- Angrist & Pischke (2015): Mastering Metrics (chapter 5)
- Angrist & Pischke (2009): Mostly Harmless Econometrics (chapter 5)

Papers we will use or replicate in class:

- Snow, J. [1855] (1965) *On the Mode of Communication of Cholera.* London
- Card & Krueger (1994) "Minimum Wages and Employment" *American Economic Review*
- Bechtel & Hainmueller (2011) "How Lasting is Voter Gratitude? An Analysis of the Short and Long Term Electoral Returns to Beneficial Policy" *American Journal of Political Science*
- Bertrand, Duflo, and Mullainathan (2004) "How much should we trust DID estimates?" *Quarterly Journal of Economics*

**WEEK 8 & 9**

**Evaluating the plausibility of as-if random** – if you're doing as-if random estimation (i.e. not a randomized experiment) you will need to convince the reader that your design is plausible: introducing placebo tests, sorting and density tests, and various other causal inference diagnostics. Explaining SUTVA. Introducing **heterogenous treatment effects** – sometimes the treatment effect varies across groups

Recommended:

- Dunning (2012) (chapter 5, and chapters 8-10)
- Gerber & Green (2012) (chapters 8, 9)

Papers we will use or replicate in class:

- Returning to some of the papers we discussed to see how they justify their design
- For HTE: Lupu (2012) "Party Brands and Partisanship: Theory with Evidence from a Survey Experiment in Argentina." *American Journal of Political Science*

**WEEK 10**

**How to design an experiment and write a great empirical paper** – in the first session of the final week you will be given a blueprint of how to write a good empirical paper that will satisfy journal reviewers; or just a blueprint to help you prove a causal impact of a particular action.

In the seminar session the **students will present the research designs of their experiments**. The instructor will provide feedback which they could use to improve it before submitting it as part of their final exam.

Recommended:

- Gerber & Green (2012) (chapters 12-13)
- Dunning (2012) (chapter 11)

## READING LIST

Most of the course material will be available online (datasets for replication, papers that we read in class, instructor's notes, etc.) but the course will roughly follow a few books, all of which are highly recommended to the students.

Readings:

- Angrist, Joshua, and Pischke, Jorn-Steffen (2015) *Mastering Metrics. The Path from Cause to Effect*. Princeton University Press
- Angrist, Joshua, and Pischke, Jorn-Steffen (2009) *Mostly Harmless Econometrics*. Princeton University Press
- Dunning, Thad (2012) *Natural Experiments in the Social Sciences*. Cambridge University Press
- Gerber, Alan and Green, Donald (2012) *Field Experiments*. WW Norton