Applied Micro-Econometrics,Fall 2023

Lecture 1: Causal Inference in Social Science

Zhaopeng Qu
Nanjing University Business School
9/27/2023

Review the Last Lecture

The Previous Lecture

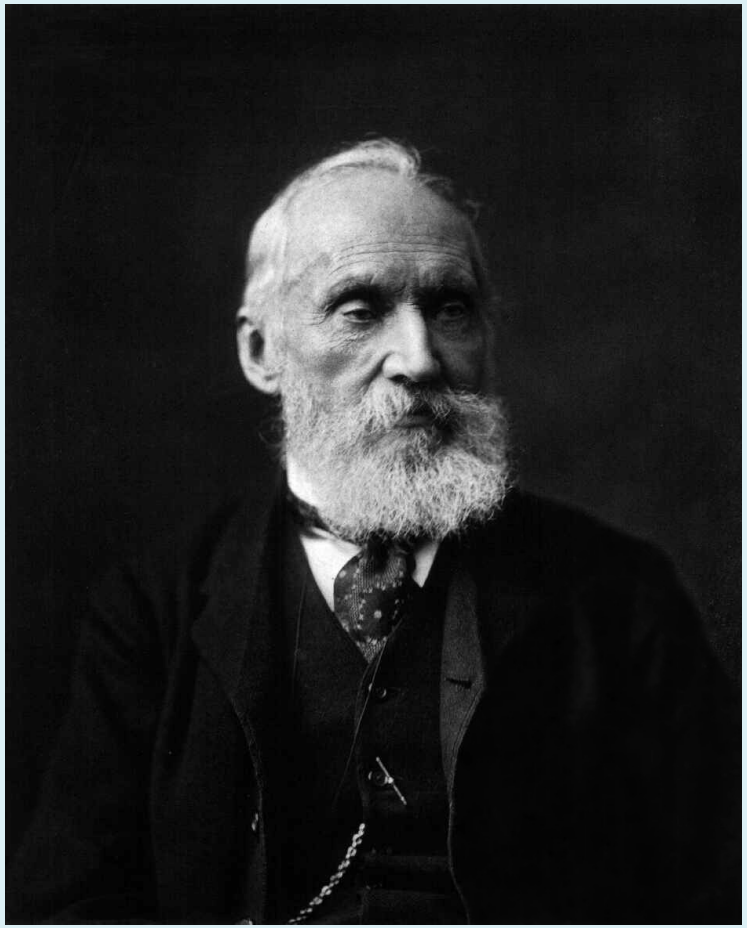
- A Scientific Framework of Making Rational Choice
  - Econometrical Analysis plays a key role
- What is Econometrics
  - Econometrics and Big Data
- Logistics to the Course

The Previous Lecture

- Evaluation(you care about most)
  - Class Participation (10%)
  - Midterm: A Proposal and Presentation (50%)
  - Final Exam: Preliminary Results (30%)
  - Final Draft(10%)
- The Procedure
  - The First Part: My lecture
  - The Second Part: Presentation for Papers(?)
  - The Third Part: Your Own Research(two presentations)

Causal Inference in Social Science

The Purposes of Empirical Studies



- **Lord Kelvin(1824-1907)**

- British mathematical physicist and engineer

> **"The objective of science is the discovery of the relations"**.

- In most cases,we often want to explore the relationship between two variables in one study.
  - education and wage
- Then, in simplicity, there are two relationships between two variables.
  - Correlation(相关)V.S. Causality (因果)

A Classical Example: Hemline Index(裙边指数)

- **George Taylor**, an economist in the United States, made up the phrase it in the 1920s. The phrase is derived from the idea that hemlines on skirts are shorter or longer depending on the economy.

- **Therefore what is about now?** Short shirt is resorting?

Causality and Big Data

- Some Big Data researchers think causality is not important any more in our times.

Viktor Mayer-Schönberger is the OII's Professor of Internet Governance and Regulation. His research focuses on the role of information in a networked economy.

> **"Look at correlations. Look at the 'what' rather than the 'why', because that is often good enough." by Viktor Mayer-Schonberger(2013)**

Causality and Econometrics

- Most empirical economists think that correlation only tell us the **superficial**, even **false** relationship while causal inference can provide solid evidence of the real relationship.

Joshhua Angrist(MIT)

Jörn-Steffen Pischke(LSE)

> **"the most interesting and challenging research in social science is about cause and effect"— Angrist and Pischke(2009)**

Causality and Econometrics

- Machine learning is a set of data-driven algorithms that use data to predict or classify some variable Y as a function of other variables X.
- Machine learning is mostly about prediction.
  - Having a good prediction does work sometimes but does NOT mean understanding causality.
- The biggest difference between machine learning and econometrics(or causal inference).
- Although two fields have developed in parallel for a while, a view to incorporating advantages of both methodologies is emerging.
  - eg. Causal Machine Learning

The Central Question of Causality

The Central Question of Causality(I)

- A simple example: **Do hospitals make people healthier?**
  - (Q: Dependent variable and Independent variable?)
- A naive solution:
  - Comparing the health status of those *who have been to the hospital* to the health of those *who have not*.
- Two key questions are documented by the questionnaires from The National Health Interview Survey(NHIS)

1. > "During the past 12 months, was the respondent a patient in a hospital overnight?"
2. > "Would you say your health in general is excellent, very good, good ,fair and poor" and scale it from the number "1" to "5" respectively.

The Central Question of Causality(II)

Group	Sample Size	Mean Health Status	S.D
Hospital	7774	**2.79**	0.014
No Hospital	90049	**2.07**	0.003

- In favor of the non-hospitalized, WHY?
- Hospitals not only cure but also hurt people.
    1. hospitals are full of other sick people who might infect us.
    2. dangerous machines and chemicals that might hurt us.
- More important: **People with poorer health tend to visit hospitals.**
- This simple case exhibits that it is not easy to answer a causal question in reality, so let us **formalize a model** to show where the problem is.

The Central Question of Causality(III)

- A right way to answer the question is by constructing **a counterfactual world**

> What if ..., then

- For any respondent, we want to compare health outcomes between two states
  - Health status if he/she see the doctor.
  - Health status if he/she **had not** see the doctor.
- **Treatment** D_i is a dummy that indicate whether individual i receive treatment or not

$$D_i = \begin{cases} 1 & \text{if individual i received the treatment} \\ 0 & \text{otherwise} \end{cases}$$

- Examples
  - Go to college or not
  - Have health insurance or not
  - Join a training program or not

Formalization: Potential Outcomes

- A potential outcome is the outcome that would be realized if the individual received a specific value of the treatment.
- For each individual, two potential outcomes, Y_{1i} and Y_{0i}, one for each value of the treatment
  - Y_{1i}: Potential outcome for an individual i **with treatment**.
  - Y_{0i}: Potential outcome for an individual i **without treatment**.

$$\text{Potential Outcomes} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Stable Unit Treatment Value Assumption (SUTVA)

- Then, the observed outcomes are realized as

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i)$$

- Implies that potential outcomes for an individual i are unaffected by the treatment status of other individual j
  - Individual i's potential outcomes are only affected by his/her own treatment.
- Rules out possible treatment effect from other individuals (spillover effect/externality)
  - Contagion
  - Displacement

Formalization: Causal Effects

- To know the difference between Y_{1i} and Y_{0i}, which can be said to be the causal effect of seeing a doctor on health for individual i. (Do you agree with it?)
- Definition: **Causal effect** is a comparison of counterfactuals under different treatment conditions on the same set of units. It also call *Individual Treatment Effect(ICE)*

$$\delta_i = Y_{1i} - Y_{0i}$$

- Further, knowing individual effect is not our final goal. As a social scientist, we would like more to know the average effect as a social pattern.
- Therefore it makes us focus on the average health status for a group of people.
  - How can we get the **average health benefits** from seeing a doctor?

Math Reveiw: Conditional Expectation Function

- **Expectation**: We usually use $E[Y_i]$ (the expectation of a variable Y_i) to denote population average of Y_i
  - Suppose we have a population with N individuals

$$E[Y_i] = \frac{1}{N}\Sigma_{i=1}^{N} Y_i$$

- **Conditional Expectation**: the expected value of a random variable given certain conditions or information.
  - The average health status for those who see a doctor:

$$E[Y_i|D_i = 1]$$

  - The average health status for those who did not see a doctor:

$$E[Y_i|D_i = 0]$$

Average Causal Effects

- Average Treatment Effect(ATE) is the average of ICEs **over the population**.

$$\alpha_{ATE} = E[\delta_i] = E[Y_{1i} - Y_{0i}]$$

- Average Treatment Effect on the Treated(ATT) is the average of ICEs **over the treated population**.

$$\alpha_{ATT} = E[\delta_i | D_i = 1] = E[Y_{1i} - Y_{0i} | D_i = 1]$$

- **Difficulty**: we can never directly observe causal effects (ICE, ATE or ATT)
  - Because we can never observe both potential outcomes Y_{0i}, Y_{1i} for any individual.
- Our aims is to compare **potential outcomes**, but we only have **observed outcomes**.
- By this view, causal inference refers to a series of methods that are used to restore or construct counterfactuals in order to address the missing data problem.

Observed Association and Selection Bias

- By using **observed data**, we can only establish **association(correlation)**, which is the observed difference in average outcome between those getting treatment and those not getting treatment.

$$\begin{aligned}\alpha_{corr} &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\end{aligned}$$

- The first term on the right side is actually ATT

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$$

- The second term on the right side is called as **Selection Bias(SB)**

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

  - which implies the potential outcomes of treatment and control groups are different even if both groups did not receive the same treatment.

- Observed association is neither necessary nor sufficient for causality for **SB**

$$\alpha_{corr} \neq ATT$$

Causal Inference and Identification Strategy

- Causal inference is the process of estimating a comparison of counterfactuals under different treatment conditions on the same set of units.
- The main goal of identification strategy is to eliminate the selection bias and construct a more proper counterfactual using the observable data.
- **Question**:
  - **how to eliminate the selection bias?**

Experimental Design as a Benchmark

How to Solves the Selection Problem

- **Answer**: Random assignment of treatment D_i can eliminates selection bias.
- Mathematically ,it makes D_i **independent** of potential outcomes, thus

$$D_i \perp (Y_{0i}, Y_{1i})$$

- **Independence**: Two random variables are said to be **independent** if knowing the outcome of one provides no useful information about the outcome of the other.

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$
$$\text{or } P(Y = y|X = x) = P(Y = y)$$

- Knowing outcome of $D_i(0, 1)$ does not help us understand what potential outcomes Y_{0i}, Y_{1i} will be.
  - In other words, the potential outcomes Y_{0i}, Y_{1i} are not correlated with $D_i(0, 1)$(actually stronger)

Random Assignment Solves the Selection Problem

- **Math**: Expectation for Independent Random Variables

$$E[Y \mid X = x] = \sum_{y \in R_y} xP_{Y|X}(y \mid x) = \sum_{x \in R_y} yP_Y(y) = E[Y]$$

- For D is independent to Y,

$$E[Y_{0i}|D_i = 1] = E(Y_{0i}) = E[Y_{0i}|D_i = 0]$$

- Thus the Selection Bias equals to ZERO.Then Observed Association equals to ATT and ATE

- Because

$$\begin{aligned} E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] \end{aligned}$$

- No matter what assumptions we make about the distribution of Y,we can always estimate it with the difference in means.

Warp up

- Think of causal effects is about comparing counterfactuals or potential outcomes. However, we can never observe both counterfactuals — fundamental problem of causal inference.
- To construct the counterfactuals, we could use two broad categories of empirical strategies.

  1.Random Controlled Trials/Experiments:

  - It can eliminates selection bias which is the most important bias arises in empirical research.
  - If we could observe the counterfactual directly, then just simply difference.
  - The data collected by RCTs is called experimental data, which is selection-bias free.

  2.Nonexperimental Methods

  - The data collected is **ex-post** data or naturally-occuring data which can not be selection-bias free.

Randomized Controlled Trials(RCTs)

- A randomized controlled trial (RCT) is a form of investigation in which units of observation (e.g. individuals, households, schools, states) are randomly assigned to **treatment** and **control** groups.
- RCT has two features that can help us hold other things equal and then eliminates selection bias

  1.Random assign treatment:

  - Randomly assign treatment (such as a coin flip) ensures that every observation has the same probability of being assigned to the treatment group.
  - Therefore, the probability of receiving treatment is unrelated to any other confounding factors.

  2.Sufficient large sample

  - Large sample size can ensure that the group differences in individual characteristics wash out.
- RCTs are considered the **gold standard** for establishing a causal link between an intervention and change.

RCTs in History: The first one in record

James Lind(1716-1794)

a Scottish physician in the Royal Navy.

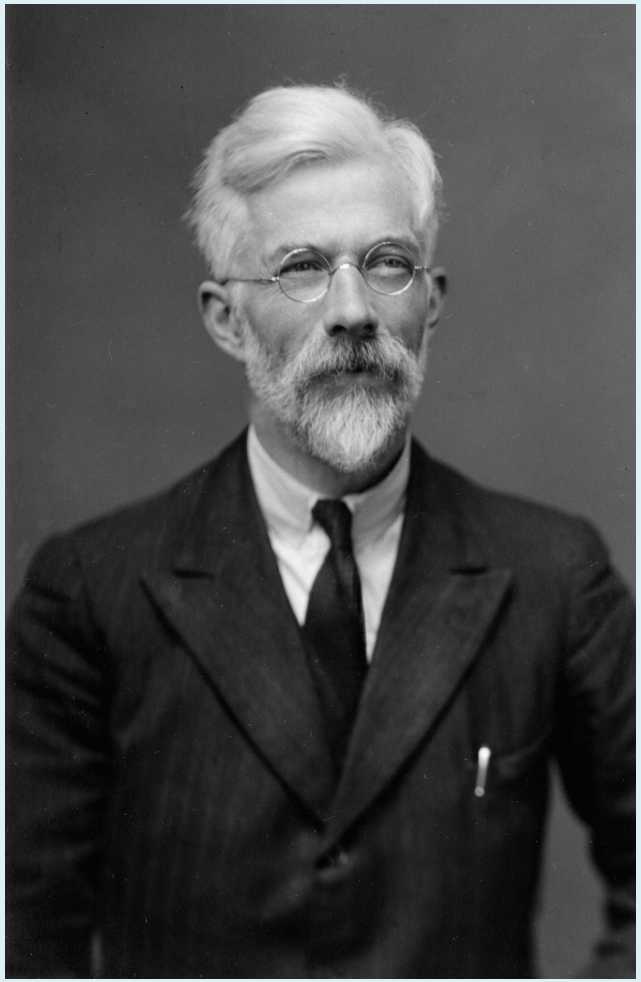
- The first recorded RCT was done in 1747.
- Scurvy(败血症) is a terrible disease caused by Vitamin C deficiency.
- Lind took 12 sailors with scurvy and split them into six groups of two.
- Groups were assigned:

1. 1 qt cider(苹果酒)
2. 25 drops of vitriol(硫酸）
3. 6 spoonfuls of vinegar,
4. 1/2 pt of sea water,
5. garlic,mustard and barley water（大麦汤）
6. 2 oranges and 1 lemon

RCTs in History: The first one in record

- Only Group 6 (citrus fruit) showed substantial improvement.

RCTs in History: Modern era



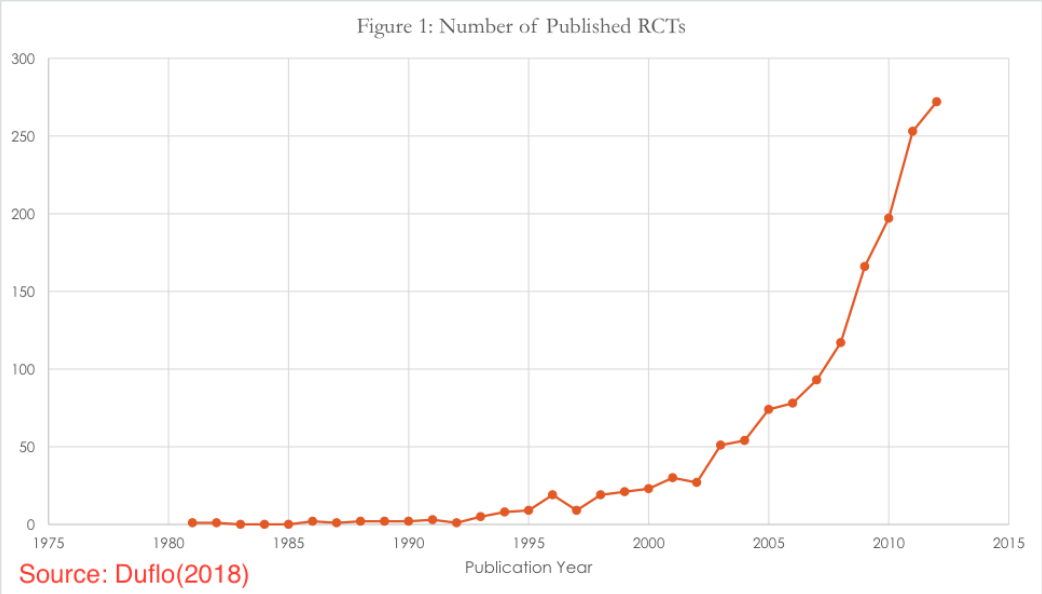
Ronald A. Fisher(1890-1962)

British statistician and geneticist who pioneered the application of statistical procedures to the design of scientific experiments.

> "a genius who almost single-handedly created the foundations for modern statistical science."

- Rothamsted Experimental Station is one of the oldest agricultural research institutions in the world, having been founded in 1843.

RCTs in Economics



Published Papers

Noble Prize 2019

RCTs in Public Policies

- According to Boruch et al(1978), 245 randomized field experiments had been conducted in U.S for social policies evaluations up to 1978.†
- The huge effort has been prompted by the 1% part of every social budget devoted to evaluation.
- Some of them were ambitious and very costly, and affected different kind of policies.
  - the Perry Preschool Program in 1961
  - The Rand Health Insurance Experiment from 1974-1982.

† Boruch RF, Mcsweeny AJ, Soderstrom EJ. Randomized field experiments for program planning, development, and evaluation: an illustrative bibliography. Eval Q. 1978 Nov;2(4):655-95.

RCTs in Public Policies

Education: the Perry Preschool Program

- 123 children born between 1958 and 1962 in Michigan
- Half of them (drawn at random) entered the Perry school program at 3 or 4 years old.
- Education by skilled professionals in nurseries and kindergarten.
- Program duration circle 30 weeks
- follow-up survey (age : 14, 15, 19, 27 and 40 years old)

Health Care: The Rand Health Insurance Experiment

- 5809 people randomly assigned in 1974 to different insurance programs with 0%, 25%, 50% and 75% sharing.
- They were followed until 1982.
- Main results : paying a portion of health cost make people give up some “superfluous” cares, with little harm on their health.
- But some heterogeneity : not true for poor people.

RCTs in Public Policies

Scott Rozelle(Standford)

- "One egg a day" program in rural China by REAP at Stanford.
  - One egg a day
- "Free-lunch" program in primary schools at Western China.
  - Free Lunch
- bilibili: 中国农村儿童发展怎样影响未来中国

RCT in Business

- An interesting question: What is the optimal color for taxis?

Taxi in NYC

Taxi in NJ

- Ho, Chong and Xia(2017), Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue,PNAS

RCT in Business

- Another Critical Question for business: Is Working at Home is better than Working at Office?

James Liang

- Bloom, Liang, Roberts and Ying,(2015), “Does Working from Home Work? Evidence from a Chinese Experiment”, The Quarterly Journal of Economics
- Bloom, Han and Liang(2022),"How Hybrid Working From Home Works Out",NBER working papers w30292

Types of RCTs

- Lab Experiments
  - eg: students evolves a experiment in a classroom.
  - eg: computer game for gamble in Lab
- Field Experiments
  - eg: the role of women in household's decision or fake resumes in job application
- Quasi-Experiment or Natural Experiments: some unexpected institutional change or natural shock
  - eg: Germany Reunion, Great Famine in China and U.S Bombing in Vietnam.

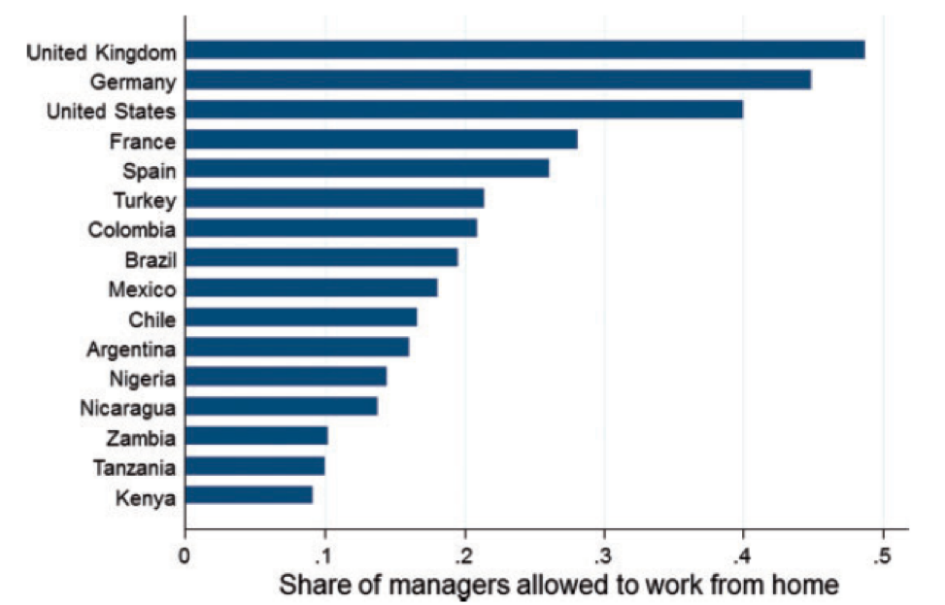
An Example of Randomized Controlled Trials

Working from Home(WFH) v.s Working from Office

- “Does Working from Home Work? Evidence from a Chinese Experiment”,by Nicholas A. Bloom, James Liang, John Roberts, Zhichun Jenny Ying The Quarterly Journal of Economics,February 2015, Vol. 130, Issue 1, Pages 165-218.
- Basic Question: WFH=SFH?
    - SFH(Shirking from Home)?

WFH v.s WFO

- Working from home is a modern management practice which appears to be stochastically spreading in the US and Europe.
  - 20 million people in US report working from home at least once per week



- Little evidence on the effect of workplace flexibility
  - productivity(shirking)
  - employee satisfaction

Ctrip Experiment

- Ctrip, China's largest travel-agent, with 16,000 employees, $6bn NASDAQ in 2015.
- James Liang,Co-founder of Ctrip,was an Econ PhD at Stanford and decided to run a experiment to test WFH at his own company.
- The experiment runs on airfare & hotel departments in Shanghai.
- Main Work: Employees take calls and make bookings.

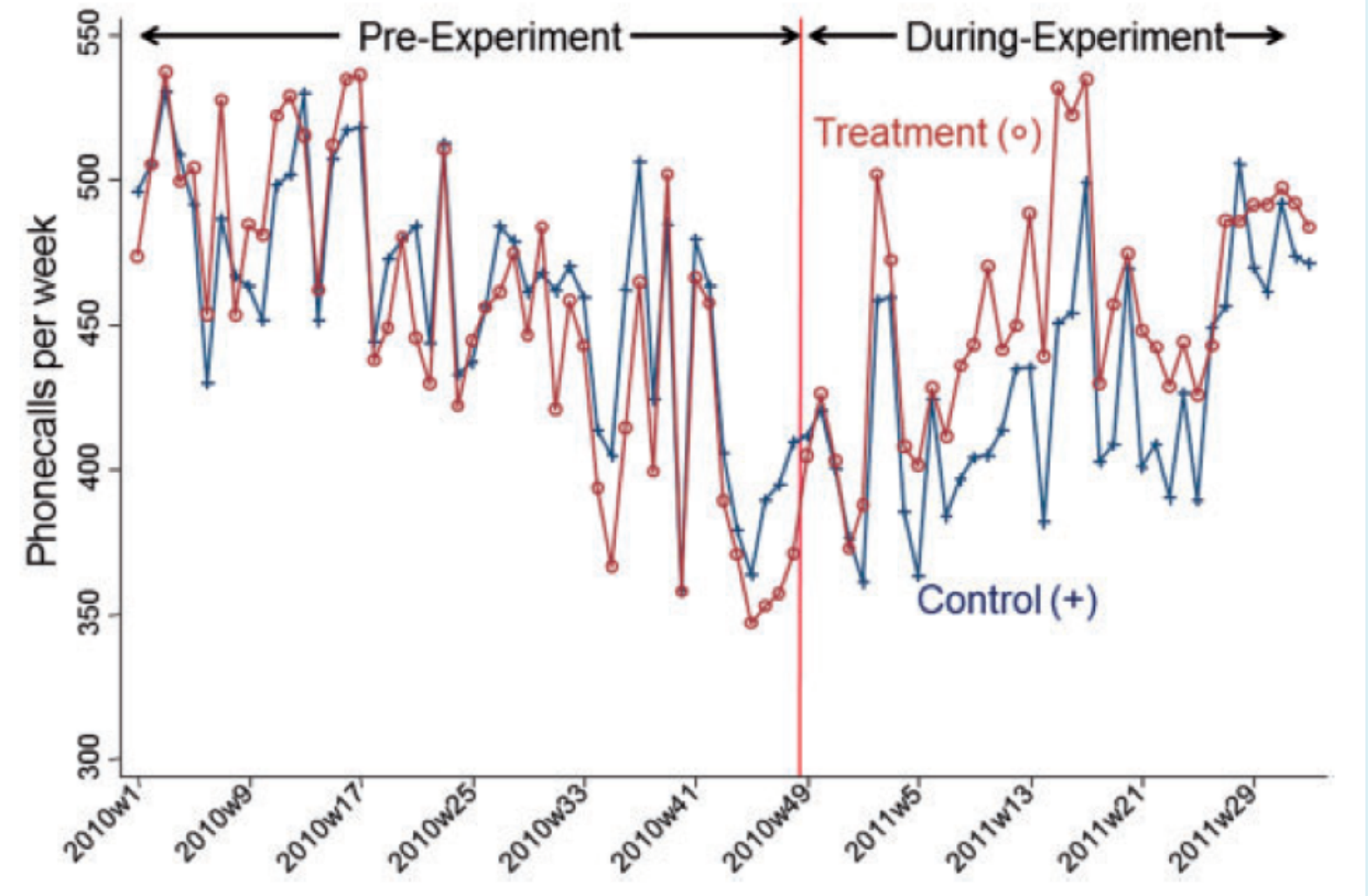
Citrip SH

The Experimental Design

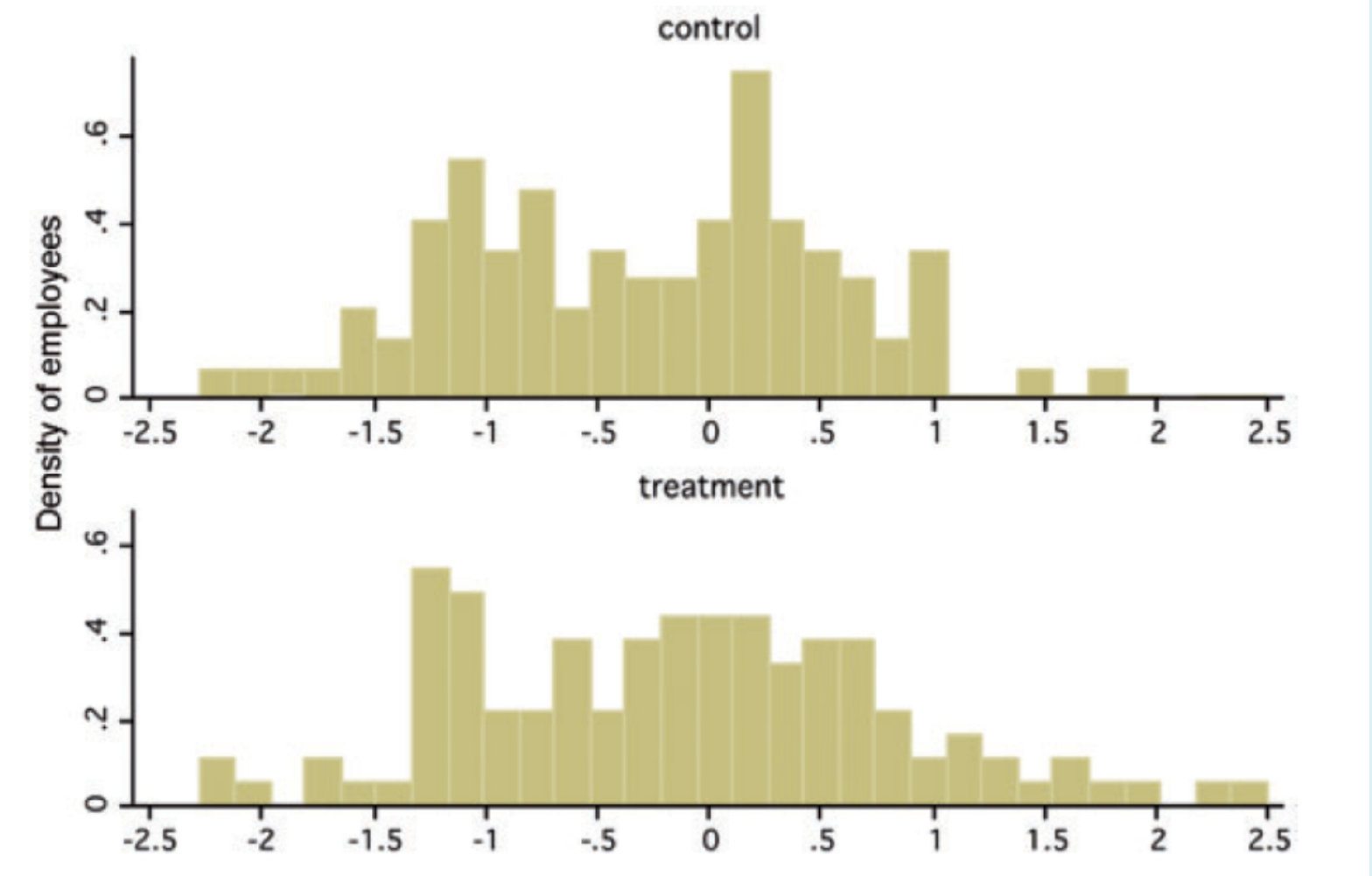
- Treatment: work 4 shifts (days) a week at home and to work the 5th shift in the office.
- Control: work in the office on all 5 days.
- Timeline
  - early Nov.2010, employees were informed of the WFH program(994 employees).
  - 503 (51%) volunteered for the experiment,249 (50%) of the employees are eligible.
  - The treatment and control groups were then determined from this group of 249 employees through a **public lottery**.

Citrip SH

Results:the number of receiving calls



Results:Working hours



Results

Variables	(1) Minutes on the phone	(2) Minutes on the phone/days worked	(3) Days worked	(4) Minutes on the phone	(5) Minutes on the phone/days worked	(6) Days worked
$Experiment_t$*$Treatment_i$	0.088***	0.063***	0.025**	0.069**	0.049*	0.021
	(0.027)	(0.024)	(0.012)	(0.030)	(0.027)	(0.013)
$Experiment_t$*$Treatment_i$* [total commute > 120 min]$_i$				0.069*	0.055*	0.014
				(0.036)	(0.031)	(0.017)
Number of employees	134	134	134	134	134	134
Number of weeks	85	85	85	85	85	85
Observations	9,426	9,426	9,426	9,426	9,426	9,426

Notes. The regressions are run at the individual by week level, with a full set of individual and week fixed effects. *Experiment*treatment* is the interaction of the period of the experimentation (December 6, 2010, until August 14, 2011) by an individual having an even birthdate (2nd, 4th, 6th, etc. day of the month). The pre-experiment period refers to January 1, 2010, until November 28, 2010. During the experiment period refers to December 6, 2010, to August 14, 2011. In columns (4)–(6), *Experiment* × *Treatment* is further interacted with a dummy variable indicating whether an employee's total daily commute (to and from work) is longer than 120 minutes (21.3% of employees have a commute longer than 120 minutes). Standard errors are clustered at the individual level. Once employees quit they are dropped from the data. *** denotes 1% significance, ** 5% significance, and * 10% significance. Minutes on the phone are recorded from the call logs.

Results

Variables	(1) Minutes on the phone	(2) Minutes on the phone/days worked	(3) Days worked	(4) Minutes on the phone	(5) Minutes on the phone/days worked	(6) Days worked
$Experiment_t$*$Treatment_i$	0.088***	0.063***	0.025**	0.069**	0.049*	0.021
	(0.027)	(0.024)	(0.012)	(0.030)	(0.027)	(0.013)
$Experiment_t$*$Treatment_i$* [total commute > 120 min]$_i$				0.069*	0.055*	0.014
				(0.036)	(0.031)	(0.017)
Number of employees	134	134	134	134	134	134
Number of weeks	85	85	85	85	85	85
Observations	9,426	9,426	9,426	9,426	9,426	9,426

Notes. The regressions are run at the individual by week level, with a full set of individual and week fixed effects. *Experiment*treatment* is the interaction of the period of the experimentation (December 6, 2010, until August 14, 2011) by an individual having an even birthdate (2nd, 4th, 6th, etc. day of the month). The pre-experiment period refers to January 1, 2010, until November 28, 2010. During the experiment period refers to December 6, 2010, to August 14, 2011. In columns (4)–(6), *Experiment* × *Treatment* is further interacted with a dummy variable indicating whether an employee's total daily commute (to and from work) is longer than 120 minutes (21.3% of employees have a commute longer than 120 minutes). Standard errors are clustered at the individual level. Once employees quit they are dropped from the data. *** denotes 1% significance, ** 5% significance, and * 10% significance. Minutes on the phone are recorded from the call logs.

Results

Variables	(1) Minutes on the phone	(2) Minutes on the phone/days worked	(3) Days worked	(4) Minutes on the phone	(5) Minutes on the phone/days worked	(6) Days worked
$Experiment_t$*$Treatment_i$	0.088***	0.063***	0.025**	0.069**	0.049*	0.021
	(0.027)	(0.024)	(0.012)	(0.030)	(0.027)	(0.013)
$Experiment_t$*$Treatment_i$*[total commute > 120 min]$_i$				0.069*	0.055*	0.014
				(0.036)	(0.031)	(0.017)
Number of employees	134	134	134	134	134	134
Number of weeks	85	85	85	85	85	85
Observations	9,426	9,426	9,426	9,426	9,426	9,426

Notes. The regressions are run at the individual by week level, with a full set of individual and week fixed effects. *Experiment*treatment* is the interaction of the period of the experimentation (December 6, 2010, until August 14, 2011) by an individual having an even birthdate (2nd, 4th, 6th, etc. day of the month). The pre-experiment period refers to January 1, 2010, until November 28, 2010. During the experiment period refers to December 6, 2010, to August 14, 2011. In columns (4)–(6), *Experiment* × *Treatment* is further interacted with a dummy variable indicating whether an employee's total daily commute (to and from work) is longer than 120 minutes (21.3% of employees have a commute longer than 120 minutes). Standard errors are clustered at the individual level. Once employees quit they are dropped from the data. *** denotes 1% significance, ** 5% significance, and * 10% significance. Minutes on the phone are recorded from the call logs.

Results

Variables	(1) Minutes on the phone	(2) Minutes on the phone/days worked	(3) Days worked	(4) Minutes on the phone	(5) Minutes on the phone/days worked	(6) Days worked
$Experiment_t$*$Treatment_i$	0.088***	0.063***	0.025**	0.069**	0.049*	0.021
	(0.027)	(0.024)	(0.012)	(0.030)	(0.027)	(0.013)
$Experiment_t$*$Treatment_i$*[total commute > 120 min]$_i$				0.069*	0.055*	0.014
				(0.036)	(0.031)	(0.017)
Number of employees	134	134	134	134	134	134
Number of weeks	85	85	85	85	85	85
Observations	9,426	9,426	9,426	9,426	9,426	9,426

Notes. The regressions are run at the individual by week level, with a full set of individual and week fixed effects. *Experiment*treatment* is the interaction of the period of the experimentation (December 6, 2010, until August 14, 2011) by an individual having an even birthdate (2nd, 4th, 6th, etc. day of the month). The pre-experiment period refers to January 1, 2010, until November 28, 2010. During the experiment period refers to December 6, 2010, to August 14, 2011. In columns (4)–(6), *Experiment* × *Treatment* is further interacted with a dummy variable indicating whether an employee's total daily commute (to and from work) is longer than 120 minutes (21.3% of employees have a commute longer than 120 minutes). Standard errors are clustered at the individual level. Once employees quit they are dropped from the data. *** denotes 1% significance, ** 5% significance, and * 10% significance. Minutes on the phone are recorded from the call logs.

Conclusion: Very positive

- They found a highly significant 13% increase in employee performance from WFH,
  - of which about 9% was from employees working more minutes of their shift period (fewer breaks and sick days)
  - and about 4% from higher performance per minute.
- Home workers also reported substantially higher work satisfaction and psychological attitude scores, and their job attrition rates fell by over 50%.

Limitations of RCTs

RCTs are not easy in practice!

- High Costs, Long Duration
- Small sample: Student Effect
- Hawthorne effect(霍桑效应）： The subjects are in an experiment can change their behavior.
- Attrition（样本流失）： It refers to subjects dropping out of the study after being randomly assigned to the treatment or control group.
- Failure to randomize or failure to follow treatment protocol: People don't always do what they are told.
    - eg. Wearing glasses program in Western Rural China.

Limitations of RCTs

RCTs are far from perfect!

- Limited Generalizability
- RCTs allow us to gain knowledge about causal effects but without knowing the mechanism.
- Potential Ethical Problems:

> “Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomized controlled trials."

- Some classical examples
  - Milgram Experiment
  - Stanford Prison Experiment
  - Monkey Experiment

Nonexperimental methods

- We can generate the data of our interest by controlling experiments just as physical scientists or biologists do. However, it is quite obvious that we face more difficult and controversial situations than those in any other sciences.
- The various approaches using naturally-occurring data provide alternative methods of constructing the proper counterfactual
  - **Econometrics**
- Congratulation! We work and study in a field that is **tougher and more challenging than others**, which includes a vast amount of scientific knowledge..
- When conducting empirical research, regardless of the methods we use, we should consider **randomized experimental methods as our benchmark**.

Program Evaluation Econometrics

- Since non-experimental data suffer from selection bias inherently, or in terms of "endogeneity," building a reasonable counterfactual world using naturally occurring data to find proper control groups is the core of econometric methods.
- Here you Furious Seven Weapons in Applied Econometrics(七种盖世武器）

1. Regression(回归)
2. Matching and Propensity Score（匹配）
3. Decomposition（分解）
4. Instrumental Variable（工具变量）
5. Regression Discontinuity（断点回归）
6. Differences in Differences（双差分）
7. Synthetic Control（合成控制）

A Simple Guide of Causal Inference

A Simple Guide of Causal Inference

General Questions

1. Why is your question important/interesting?
2. Why is the current literature lacking or nonexistant?
3. How do you propose to advance the literature?

Angrist and Pischke's FAQs†

1. What is the **causal relationship of interest**?
2. How would an **ideal experiment** capture this causal effect of interest?
3. What is your **identification strategy**?
4. What is your **mode of inference**?

† See *MHE,* chapter 1.

A Simple Guide of Causal Inference

FAQ1: What is the causal relationship of interest?

What are the variables that you are most concerned about in your project?

- What are the independent and dependent variables?
- What are the possible control variables?
- Is possible to access some data which includes information about these variables.

FAQ2: What is the ideal experiment for this setting?

Describing the *ideal experiment* helps us formulate

- the exact causal question(s)
- the dimensions we want to manipulate
- the factors we need to hold constant
- These *ideal experiments* are generally hypothetical, but if you can't describe the ideal, it will probably be hard to come up with data and plausible research designs in real life.
- Angrist and Pischke call questions without ideal experiments *fundamentally unanswerable questions* (FUQs).

FAQ2: What is the ideal experiment for this setting?

Examples of potentially answerable questions...

- **The effect of education on wages:** Randomize scholarships or incentives to remain in school.
- **Institution and development:** Arbitrarily assign institutional types to countries as they receive independence.
- **Environmental cleanups:** Ask EPA to randomly clean toxic sites.

Examples of challenging questions to answer (potentially unanswerable?)...

- How does gender affect eventual career paths?
- What role does race play in one's wages?

FAQ2: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

> Do children perform better by starting school at an older age?

Central problem: Mechanical links between ages and time in school.

$$\text{(Start Age)} = \text{(Current Age)} - \text{(Time in School)}$$

No experiment can separate these effects (for school-age children).

FAQ3: What's your identification strategy?

This question describes how you plan to recover/observe *as good as random* assignment of your variable of interest (approximating your ideal experiment) **in real life**.

Examples

- Compulsory school-attendance laws *interacted with* quarter of birth
- Vietnam War draft
- Thresholds for the Clean Air Act violations
- Notches in income-tax policies
- Judge assignments
- Randomly assigned characteristics on résumés

FAQ4: What is your mode of inference?

Historically, inference—standard errors, confidence intervals, hypothesis tests, *etc.*—has received much less attention than point estimates. It's becoming more important (more than an afterthought).

- Which **population** does your sample represent?
- How much **noise** (error) exists in your estimator (and estimates)?
- How much **variation** do you actually have in your variable of interest?

Without careful inference, we don't know the difference between

- 21% ± 2.3%
- 21% ± 20.3%


Let's Start Our Journey