Introduction to Econometrics

Lecture 1: Introduction to Causal Inference and Randomized Controlled Trials

Zhaopeng Qu

Business School,Nanjing University

March 2,2023

Outlines

1. Review the Last Lecture
2. Causal Inference in Social Science
3. Experimental Design as a Benchmark
4. Two Cases
5. Limitations of RCTs and Econometrics

Review the Last Lecture

The Last Lecture

- A Scientific Framework of Rational Acknowledge
  - Why we need empirical analysis.
- Two Revolutions in Social Science
  - Econometrical Analysis plays a key role
- What is Econometrics?
  - Two Missions
    1. Causal Inference
    2. Scientific Prediction

The Last Lecture

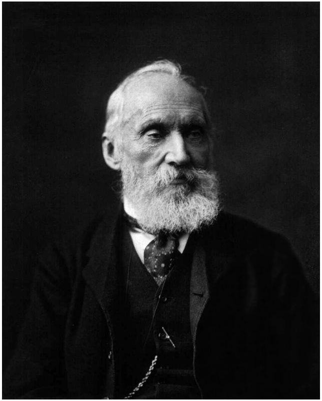
- Logistics to the Course
  - Evaluation(probably you care most)
    - Class Participation (10%)
    - Homework(30%)
    - Team Project: A research proposal(20%)
    - Final Exam: (40%)

The Last Lecture

- The Structure of Economic Data
    - Cross-sectional data
    - Time series data
    - Pooled cross-sectional data
    - Panel data

Causal Inference in Social Science

The Purposes of Science



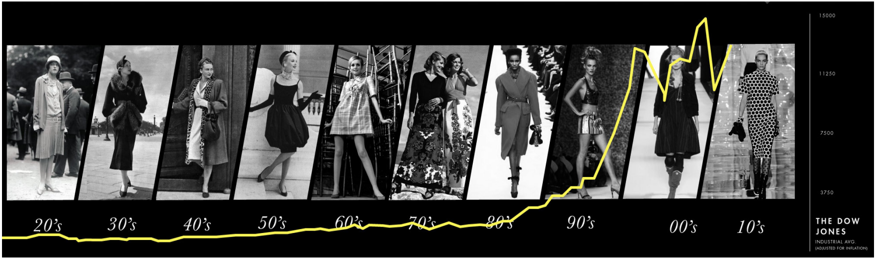
- **Lord Kelvin(1824-1907)**:
- "*The objective of science is the discovery of the relations*"

- In most cases, we often want to explore the relationship between **two** variables in one study.
  - eg. education and wage
- Then, in simplicity, there are **two** simple relationships between two variables.
  - Correlation(相关)
  - 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.
  - Before 1930s, fashion women favored middle skirts most.
  - In 1929, long skirts became popular. While the *Dow Jones Industrial Index(DJII)* plunged from about 400 to 200 and to 40 two years later.
  - In 1960s, DJII rushed to 1000. At the same time, short skirts showed up.
  - In 1970s, DJII fell to 590 and women began to wear long skirts again.
  - In 1990s, mini skirt debuted, DJII rushed to 10000.
  - In 2000s, bikini became a nice choice for girls, DJII was high up to 13000.
  - So what is about now? Long skirt is resorting?

Hemline Index:1920s-2010s



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.*" —**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.

- Today, empirical economists care more about the causal relationship of their interests than ever before.
    - "*the most interesting and challenging research in social science is about cause and effect*" —**Joshua Angrist** and **Jörn-Steffen Pischke**(2009)

Causality v.s. Forecasting

- Traditionally **Machine learning** can be seen as a set of data-driven algorithms that use data to predict or classify some variable Y as a function of other variables X.
    - There are many machine learning algorithm. The best methods vary with the particular data application
- Machine learning is mostly about **prediction**.
    - Having a good prediction does work sometimes but does **NOT** mean understanding causality.

Causality v.s. Forecasting

- Even though forecasting need not involve causal relationships, economic theory suggests patterns and relationships that might be useful for forecasting.
    - Econometric analysis(times series) allows us to quantify historical relationships suggested by economic theory, to check whether those relationships have been stable over time, to make quantitative forecasts about the future, and to assess the accuracy of those forecasts.
- 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 recently.
    - Casual Machine Learning

The Central Question of Causality(I)

- A simple example: Do hospitals make people healthier? (Q: Dependent variable and Independent variable?)
- A naive solution: compare 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（问卷）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)

Hospital v.s. No Hospital

Group	Sample Size	Mean Health Status	Std.Dev
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 having worse health tends to visit hospitals.
- This simple case exhibits that it is not easy to answer an causal question, so let us formalize an model to show where the problem is.

The Central Question of Causality(III)

- So A right way to answer a causal questions is construct a counterfactual world, thus “What If ....then” , Such as
- An example: *How much wage premium you can get from college attendance(*上大学使工资增加多少？)
  - For any worker, we want to compare
    - Wage if he have a college degree (上了大学后的工资)
    - Wage if he had not a college degree （假设没上大学，工作的工资）
  - Then make a difference. This is the right answer to our question.
- Difficulty in Identification: only one state can be observed

Formalization: Rubin Causal Model

- Treatment : D_i is a **dummy** that indicate whether individual i receive treatment or not

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

- Examples:
    - Go to college or not
    - Have health insurance or not
    - Join a training program or not
    - Make an online-advertisement or not
    - ....

Formalization: Treatment

- Treatment : D_i can be a **multiple valued**(continues) variable

$$D_i = s$$

- Examples:
    - Schooling years
    - Number of Children
    - Number of advertisements
    - Money Supply
- For simplicity, we assume treatment variable D_i is just a **dummy**.

Formalization: Potential Outcomes

- A potential outcome is the outcome that would be realized if the individual received a specific value of the treatment.
    - Annual earnings if attending to college
    - Annual earnings if not attending to college
- For each individual, we has 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 with treatment.

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

Stable Unit Treatment Value Assumption (SUTVA)

- 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 j 's potential outcomes are only affected by his/her own treatment.
- Rules out possible treatment effect from other individuals (**spillover effect/externality**)
  - Contagion
  - Displacement

Causal effect for an Individual

- To know the difference between Y_{1i} and Y_{0i}, which can be said to be the **causal effect** of going to college for individual i. (Do you agree with it?)

Definition

Causal inference is the process of estimating a comparison of counterfactuals under different treatment conditions on the same set of units. It also call Individual Treatment Effect(ICE)

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

Formalization: Estimate ICE

- Due to unobserved counterfactual outcome, we need to make strong assumptions to estimate ICE.
  - Rule out that the ICE differs across individuals (*heterogeneity effect*)
- 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**.
- So it make us focus on the average wage for a group of people.
  - How can we get the average wage premium for college attendance?

Conditional Expectation

- **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 average wage for those who attend college: $E[Y_i|D_i = 1]$
  - The average wage for those who did not attend college: $E[Y_i|D_i = 0]$

Average Causal Effects

Average Treatment Effect (ATE)

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

- It is average of ICEs over **the population**.

Average treatment effect on the treated(ATT)

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

- Average of ICEs over the **treated population**

Fundamental Problem of Causal Inference

- 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.
- We need to compare **potential outcomes**, but we only have **observed outcomes**
- So by this view, causal inference is a **missing data** problem.

Fundamental Problem of Causal Inference

- Imagine a population with 4 people

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1} - Y_{0i}$
Tom	3	?	3	1	?
Jerry	2	?	2	1	?
Scarlett	?	1	1	0	?
Nicole	?	1	1	0	?

- What is Individual causal effect (ICE) of attending college for Tom? for Nicole?

Individual Causal Effect

- Suppose we can observe counterfactual outcomes

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1}-Y_{0i}$
Tom	3	2	3	1	1
Jerry	2	1	2	1	1
Scarlett	1	1	1	0	0
Nicole	1	1	1	0	0

- The ICE for Tom

$$\delta_{Tom} = 3 - 2 = 11$$

- The ICE for Nicole

$$\delta_{Nicole} = 1 - 1 = 0$$

Average Treatment Effect(ATE)

- Missing data problem also arises when we estimate ATE

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1} - Y_{0i}$
Tom	3	?	3	1	?
Jerry	2	?	2	1	?
Scarlett	?	1	1	0	?
Nicole	?	1	1	0	?
$E[Y_{1i}]$	?				
$E[Y_{0i}]$		?			
$E[Y_{1i} - Y_{0i}]$					?

- What is the effect of attending college on average wage of population(ATE)

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

Average Treatment Effect(ATE)

- Missing data problem also arises when we estimate ATE

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1} - Y_{0i}$
Tom	3	2	3	1	1
Jerry	2	1	2	1	1
Scarlett	1	1	1	0	0
Nicole	1	1	1	0	0
$E[Y_{1i}]$	$\frac{3+2+1+1}{4} = 1.75$				
$E[Y_{0i}]$		$\frac{2+1+1+1}{4} = 1.25$			
$E[Y_{1i} - Y_{0i}]$					0.5

- What is the effect of attending college on average wage of *the population*(ATE)

$$\alpha_{ATE} = E[\delta_i] = E[Y_{1i} - Y_{0i}] = \frac{1+1+0+0}{4} = 0.5$$

Average Treatment Effect on the Treated(ATT)

- Missing data problem arises when we estimate ATT

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1} - Y_{0i}$
Tom	3	?	3	1	?
Jerry	2	?	2	1	?
Scarlett	?	1	1	0	?
Nicole	?	1	1	0	?
$E[Y_{1i}\|D_i = 1]$	?				
$E[Y_{0i}\|D_i = 1]$		?			
$E[Y_{1i} - Y_{0i}\|D_i = 1]$					?

- What is the effect of attending college on average wage for *those who attend college*(ATT)

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

Average Treatment Effect on the Treated(ATT)

- Missing data problem also arises when we estimate ATE

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1}-Y_{0i}$
Tom	3	2	3	1	1
Jerry	2	1	2	1	1
Scarlett	1	1	1	0	0
Nicole	1	1	1	0	0
$E[Y_{1i}\|D_i=1]$	$\frac{3+2}{2}=2.5$				
$E[Y_{0i}\|D_i=1]$		$\frac{2+1}{2}=1.5$			
$E[Y_{1i}-Y_{0i}\|D_i=1]$					1

- The effect of attending college on average wage for *those who attend college*(ATT)

$$\alpha_{ATE}=E[Y_{1i}-Y_{0i}|D_i=1]=\frac{1+1}{2}=1$$

Observed Association and Selection Bias

- Causality is defined by **potential outcomes**, not by **realized (observed) outcomes**.
- In fact, we can not observe all potential outcomes .Therefore, we can not estimate the above causal effects without further assumptions.
- 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.

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

College vs Non-College Wage Differentials:

- Comparing the average wage in labor market who went to college and did not go.

College vs Non-College Wage Differentials:

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

- Question 1: Which one defines the causal effect of college attendance?

Formalization: Rubin Causal Model

- **Selection Bias(SB)** implies the potential outcomes of treatment and control groups are different even if both groups receive the same treatment

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

- Question 2: Selection Bias is positive or negative in the case?
- This means two groups could be quite different in other dimensions: other things are not equal.
- Observed association is *neither necessary nor sufficient for causality.*

Observed Association:College vs Non-College

- Missing data problem also arises when we estimate ATE

i	Y_{i1}	Y_{0i}	Y_i	D_i	$Y_{i1}-Y_{0i}$
Tom	3	?	3	1	?
Jerry	2	?	2	1	?
Scarlett	?	1	1	0	?
Nicole	?	1	1	0	?
$E[Y_{1i}\|D_i=1]$	$\frac{3+2}{2}=2.5$				
$E[Y_{0i}\|D_i=0]$		$\frac{1+1}{2}=1$			
$E[Y_{1i}\|D_i=1]-E[Y_{0i}\|D_i=0]$					1.5

- The Observed Association of attending college on average wage

$$\alpha_{corr}=2.5-1=1.5$$

Observed Association and Selection Bias

- But we are interested in causal effect, here is ATT

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

- So the selection bias

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] = 0.5$$

- The **Selection Bias** is positive: *Those who attend college could be more intelligent so they can earn more even if they did not attend college.*
- Many Many Other examples
    - the effect of job training program on worker's earnings
    - the effect of class size on students performance
    - ....

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.
- **Identification strategy** tells us what we can learn about a causal effect from the observable data.
- The main goal of identification strategy is **to eliminate the selection bias** and **construct a more proper counterfactual** using the observable data.
- When you use a certain method, you have to make some certain **assumptions**.
- "Your identification is good or bad" = "Your paper is good or bad" = if the assumptions that allow you to claim you've estimated a causal effect are valid?

Experimental Design as a Benchmark

Randomized Controlled Trial

- 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
    - 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.
    - Sufficient large sample
        - Large sample size can ensure that the group differences in individual characteristics wash out

How to Solves the Selection Problem

- Random assignment of treatment D_i can eliminates selection bias. It means that the treated group is a random sample from the population.
- Being a random sample, we know that those included in the sample are **the same, on average,** as those not included in the sample on any measure.
- Mathematically ,it makes D_i **independent** of potential outcomes, thus

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

- **Independence**: Two variables are said to be independent if knowing the outcome of one provides no useful information about the outcome of the other.
    - Knowing outcome of $D_i(0, 1)$ does not help us understand what potential outcomes of (Y_{0i}, Y_{1i}) will be

Random Assignment Solves the Selection Problem

- So we have

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

- Thus the **Selection Bias** equals to **ZERO**.
- Then **ATT** equals **Observed Association** because the

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

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

Our Benchmark: Randomized Experiments

- Think of causal effects in terms of 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.
  - **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 there is no evaluation problem, just **simply difference**.

Randomized Controlled Trials(RCTs)(随机可控试验)

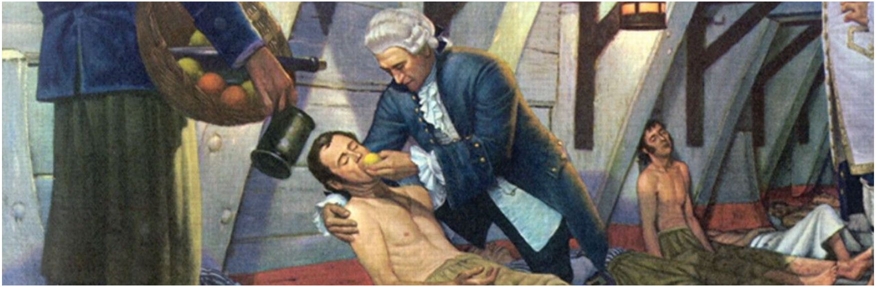
- In essence, an RCT is an **experiment** carried out on **two or more groups** where participants are randomly assigned to receive an intervention or not.
    - Participants are randomly assigned to either an **treatment group** who are given the intervention, or a **control group** who are not..
- In RCTs, each group is tested at the end of the trial and the results from the groups are compared to see if the intervention has made a difference and achieved its desired outcome. If the randomized groups are large enough, you can be confident that differences observed are due to the intervention and not some other factors.
- RCTs are considered the *gold standard* for establishing a causal link between an intervention and change.

RCT in History

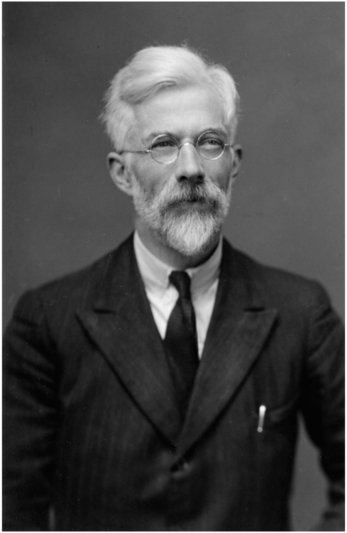
- First recorded RCT was done in 1747 by **James Lind**,a Scottish physician in the Royal Navy.
- Scurvy(败血症) is a terrible disease caused by Vitamin C deficiency.
- Lind took 12 sailors with scurvy and split them into six groups of two.
- Groups were assigned: (1) 1 qt cider(苹果酒) (2) 25 drops of vitriol(硫酸) (3) 6 spoonfuls of vinegar, (4) 1/2 pt of sea water, (5) garlic, mustard(芥末)and barley water (大麦汤)

RCT in History

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

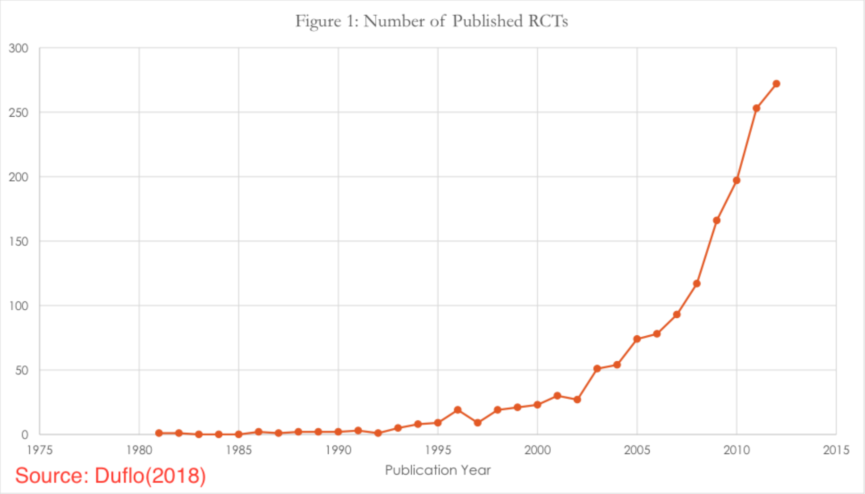


RCT in History



- **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

RCTs in Economics



Source: Duflo(2018)

Randomized Experimental Methods: Noble Prize 2019

EKONOMIPRISET 2019
THE PRIZE IN ECONOMIC SCIENCES 2019

KUNGL.
VETENSKAPS-
AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Abhijit Banerjee

Esther Duflo

Michael Kremer

"för deras experimentella ansats för att mildra global fattigdom"

"for their experimental approach to alleviating global poverty"

RCTs in Social Policies

- According to Baruch (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.

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 China

- “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

RCT in Business

- An interesting question: What is the optimal color for taxis?
  - Ho, Chong and Xia(2017), Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue, PNAS.

RCT in Business

RCT in Business

- Another Critical Question for business: Is Working at Home is better than Working at Office?
  - Bloom, Liang, Roberts and Ying,(2015), "Does Working from Home Work? Evidence from a Chinese Experiment" , The Quarterly Journal of Economics

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(德国统一), Great Famine in China(1959-1961 年大饥荒) and U.S Bombing in Vietnam(美国轰炸越南).

Two Cases

A Case: the California School

- Draw schools ($n = 420$) randomly from all school in California
- Variables:
  - 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
  - Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

Summary Table: Descriptive Statistics

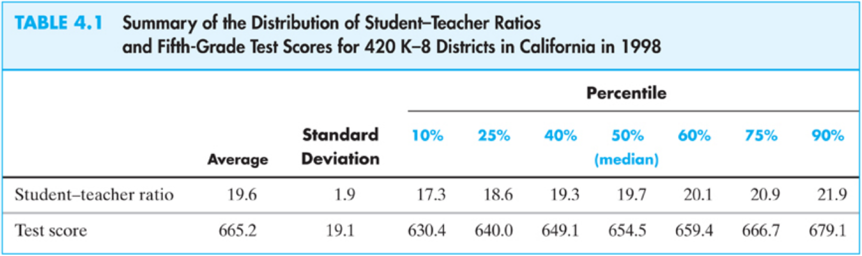
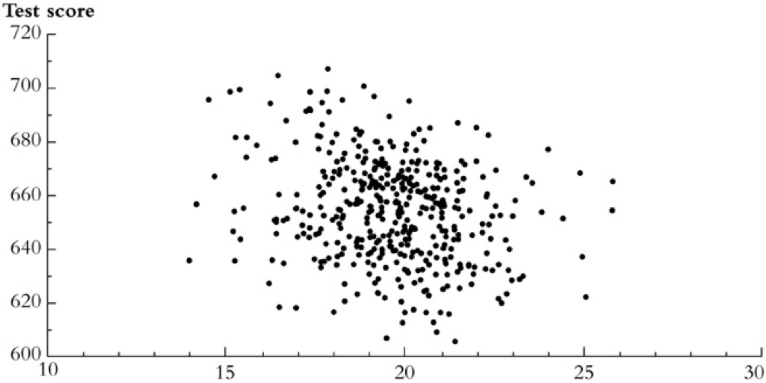


TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

- Does this table tell us anything about the relationship between test scores and the STR?

Scatterplot: test score v. student-teacher ratio



- What does this figure show? and it may suggest...?

The California Test Score

- We need to get some **numerical** evidence on whether districts with low STRs have higher test scores.
- But how?
  1. Compare average test scores in districts with low STRs to those with high STRs (“estimation”)
  2. Test the “null” hypothesis that the mean test scores in the two types of districts are the same, against the “alternative” hypothesis that they differ (“hypothesis testing”)
  3. Estimate an interval for the difference in the mean test scores, high v. low STR districts (“confidence interval”)

The California Test Score

- Compare districts with "small" and "large" class sizes:

Small v.s. Large

Class Size	Average score($\overline{Y}$)	Standard deviation	N
Small($STR < 20$)	657.4	19.4	238
Large($STR \geqslant 20$)	650.0	17.9	182

-
  1. Estimation of Δ= difference between group means
  2. Test the hypothesis that $\Delta = 0$
  3. Construct a confidence interval for Δ

Comparing Means from Different Populations

- In an RCT, we would like to estimate the average causal effects over the population

$$ATE = ATT = E\{Y_i(1) - Y_i(0)\}$$

- We only have random samples and random assignment to treatment, then what we can estimate instead

$$\textit{difference in mean} = \overline{Y}_{treated} - \overline{Y}_{control}$$

- Under randomization, *difference-in-means* is a good estimate for the ATE.

Hypothesis Tests for the Difference Between Two Means

- To illustrate a test for the difference between two means, let μ_s be the mean scores in the population of small classes and let μ_l be the population mean scores for the large classes.
- Then the **null hypothesis** and **the two-sided alternative hypothesis** are

$$\begin{aligned} H_0 : \mu_s &= \mu_l \\ H_1 : \mu_s &\neq \mu_l \end{aligned}$$

- Consider the null hypothesis that mean scores for these two populations differ by a certain amount, say d_0. The null hypothesis that large classes and small classes have the same mean scores corresponds to $H_0 : d_0 = \mu_s - \mu_l = 0$

The Difference Between Two Means

- Suppose we have samples of n_s classes and n_l classes drawn at random from the population of CA. Let the sample average scores be $\overline{Y}_s$ for the small and $\overline{Y}_l$ for the large. Then an estimator of $\mu_s - \mu_l$ is $\overline{Y}_s - \overline{Y}_l$.
- Let us discuss the distribution of $\overline{Y}_s - \overline{Y}_l$.
    - Recall $\overline{Y}_s$ is approximately distributed $N(\mu_s, \frac{\sigma_s^2}{n_s})$ and $\overline{Y}_l$ is approximately distributed $N(\mu_l, \frac{\sigma_l^2}{n_l})$ according to the C.L.T.
    - Then $\overline{Y}_s - \overline{Y}_l$ is distributed as

$$\sim N(\mu_s - \mu_l, \frac{\sigma_s^2}{n_s} + \frac{\sigma_l^2}{n_l})$$

The Difference Between Two Means

- If $\sigma_s^2$and σ_l^2 are known, then the this approximate normal distribution can be used to compute p-values for the test of the null hypothesis. In practice, however, these population variances are typically unknown so they must be estimated using the variance of the sample mean.
- Thus the *standard error* of $\overline{Y}_s - \overline{Y}_l$ is

$$SE(\overline{Y}_s - \overline{Y}_l) = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

The Difference Between Two Means

- The t-statistic for testing the null hypothesis is constructed analogously to the t-statistic for testing a hypothesis about a single population mean, thus *a simplest **t-statistic*** for comparing two means is

$$t_{act} = \frac{\overline{Y}_s - \overline{Y}_l - d_0}{SE(\overline{Y}_s - \overline{Y}_l)}$$

- If both $n_s$and n_l are large, then this t-statistic has a standard normal distribution when the null hypothesis is true,thus $\overline{Y}_m - \overline{Y}_w = 0$.

Confidence Intervals for the Difference Between Two Means

- the 95% two-sided confidence interval for d consists of those values of d within ± 1.96 standard errors of $\overline{Y}_m - \overline{Y}_w$, thus $d = \mu_m - \mu_w$ is

$$(\overline{Y}_m - \overline{Y}_w) \pm 1.96 SE(\overline{Y}_m - \overline{Y}_w)$$

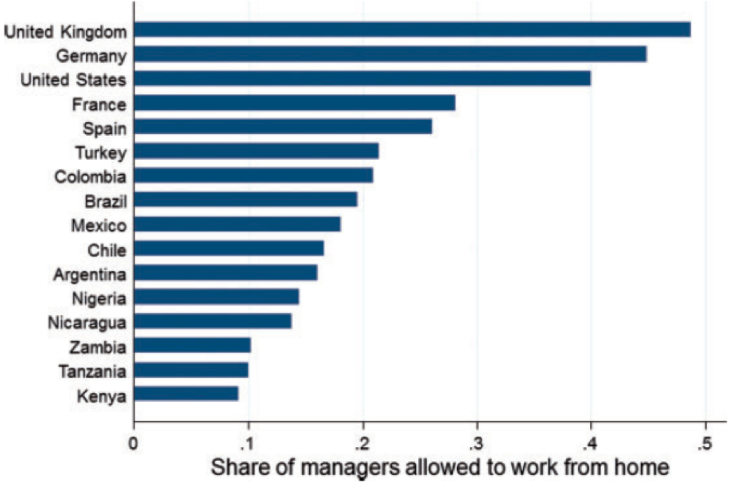
Hypothesis Test of the Difference Between Two Means

- Reject the null hypothesis if
  - $\mid t^{act} \mid = \mid \frac{\overline{Y}_m - \overline{Y}_w - d_0}{SE(\overline{Y}_m - \overline{Y}_w)} \mid > \textit{critical value}$
  - or if $p - value < significance\ level$

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)?

Working from Home(WFH) is a trend internationally



Motivations

- 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
  - employee satisfaction
  - shirking

Ctrip Experiment

- Ctrip, China's largest travel-agent, with16,000 employees, $6bn NASDAQ.
- Co-founder of Ctrip, **James Liang**, was an Econ PhD at Stanford and decided to run a experiment to test WFH.

Ctrip Experiment: A call center in Shanghai

- The experiment runs on airfare & hotel departments in Shanghai.
- Main Work: Employees take calls and make bookings.

Headquarters in Shanghai

Main Lobby

The Experimental Design

- **Treatment**: work 4 shifts (days) a week at home and to work the 5th shift in the office on a fixed day.
- **Control**: work in the office on all 5 days.

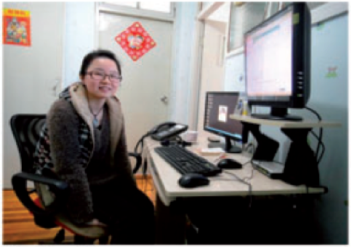
The Experimental Design: Timeline

- In early November 2010, employees in the airfare and hotel booking departments were informed of the WFH program.
- Of the 994 employees in the airfare and hotel booking departments, 503 (51%) volunteered for the experiment.
- Among the volunteers, 249 (50%) of the employees met the eligibility requirements and were recruited into the experiment.
- The treatment and control groups were then determined from this group of 249 employees through a public lottery.

The Experimental Design

Treatment groups were determined by a lottery

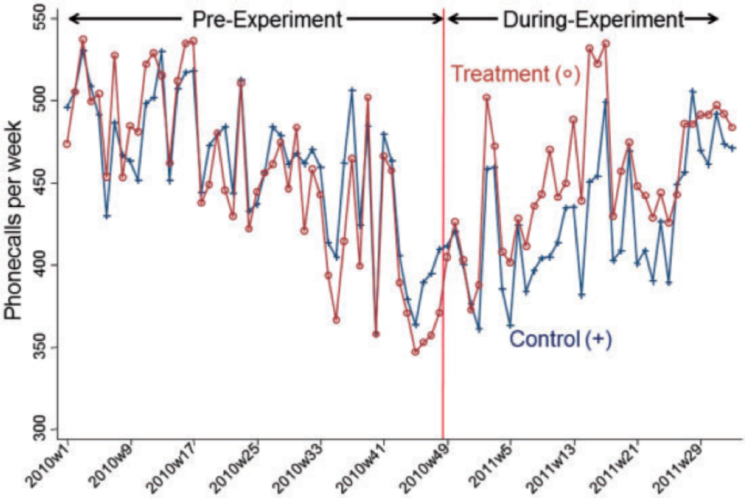
Working at home



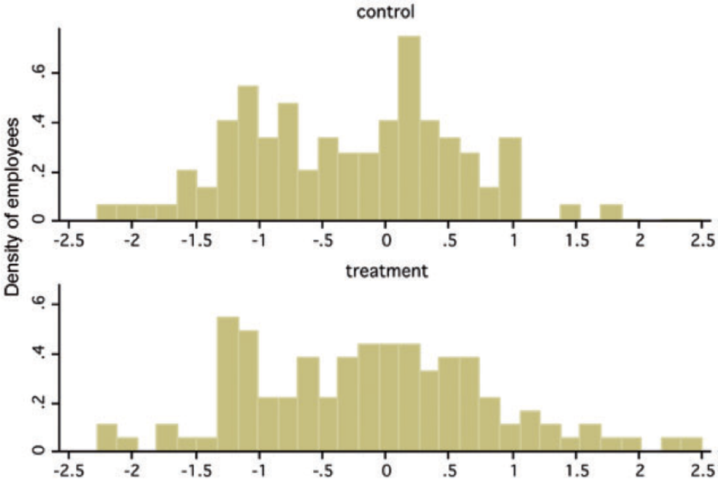
Working at home

Working at home

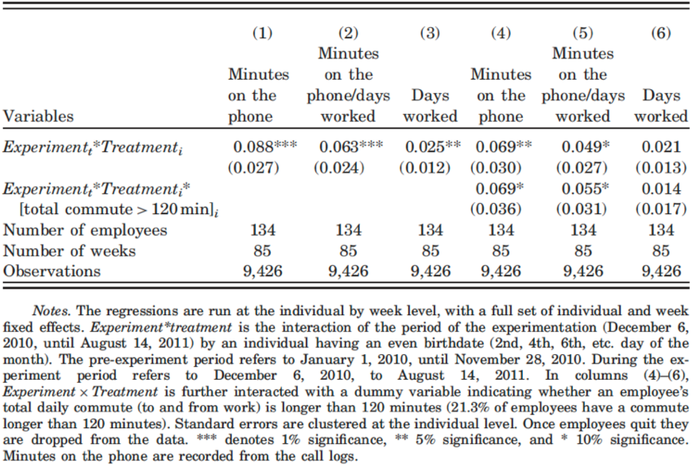
Results: the number of receiving calls



Results: Working hours



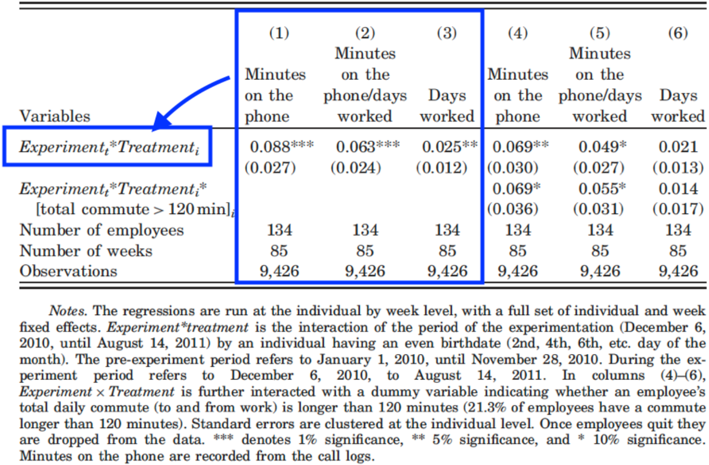
Results:Many Outcomes



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$_i$**				0.069*	0.055*	0.014
[total commute > 120 min]$_i$				(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.

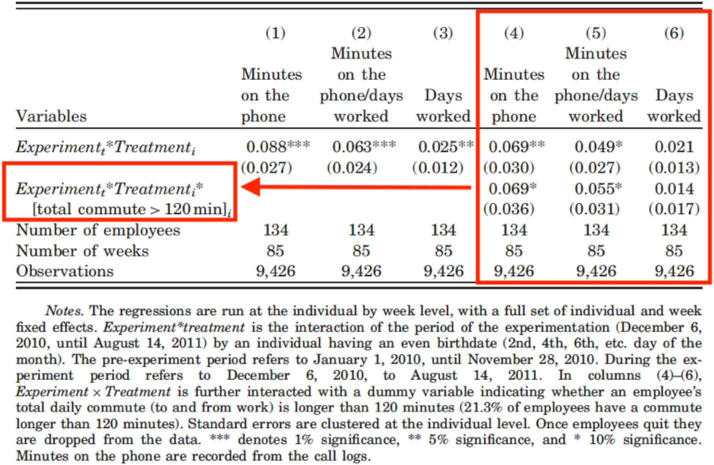
Results:Many Outcomes



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$_i$**				0.069*	0.055*	0.014
[total commute $>$ 120 min]$_i$				(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.

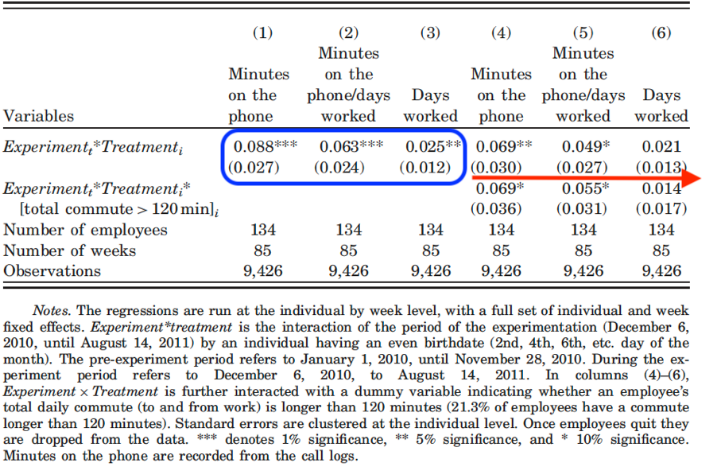
Results:Many Outcomes



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)
$Experiment_t{*}Treatment_i{*}$				0.069*	0.055*	0.014
$[\text{total commute} > 120\,\text{min}]_i$				(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:Many Outcomes



	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Minutes on the phone	Minutes on the phone/days worked	Days worked	Minutes on the phone	Minutes on the phone/days worked	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)
$Experiment_t$*$Treatment_i$*				0.069*	0.055*	0.014
$[\text{total commute} > 120\,\text{min}]_i$				(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 and Econometrics

RCT are far from perfect!

- High Costs, Long Duration
- Potential Ethical Problems: "Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomized controlled trials."
    - Milgram Experiment
    - Stanford Prison Experiment
    - Monkey Experiment
- Limited Generalizability
- RCTs allow us to gain knowledge about causal effects but without knowing the mechanism.

Potential Problems in Practice

- 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.
  - Wearing glasses program in Western Rural China.

Program Evaluation Econometrics

- We can generate the data of our interest by controlling experiments just as physical scientists or biologists do.
- But too obviously, we face more difficult and controversy situation than those in any other sciences.
- **Question**: How to do empirical research scientifically when we can not do experiments?
  - It means that we always have selection bias in our data, or in term of endogeneity.

Program Evaluation Econometrics

- We can generate the data of our interest by controlling experiments just as physical scientists or biologists do.
  - **Answer:** Build a reasonable counterfactual world by naturally occurring data to find or construct a proper control group is the core of econometrical methods.
  - The various approaches using naturally-occurring data provide alternative methods constitutes
    - Econometrics
  - Though we can not always do experiments, we should take the randomized experimental methods as our benchmark when we do empirical research whatever the methods we apply.

Program Evaluation Econometrics

- Question: How to do empirical research scientifically when we can not do experiments? It means that we always have selection bias in our data, or in term of “endogeneity”.
  - Answer: Build a reasonable counterfactual world by naturally occurring data to find a proper control group is the core of econometrical methods.
  - Here you Furious Seven Weapons in Applied Econometrics(七种盖世武器)
    1. Random Controlled Trials (RCT)（随机实验）
    2. Regression(回归)
    3. Matching and Propensity Score（匹配）
    4. Instrumental Variable（工具变量）
    5. Regression Discontinuity（断点回归）
    6. Differences in Differences（双差分）
    7. Synthetic Control （合成控制法）

Program Evaluation Econometrics

- These Furious Seven are the most basic and popular methods in applied econometrics and so powerful that
  - even if you just master one, you may finish your empirical paper and get a good score.
  - if you master several ones, you could have opportunity to publish your paper.
  - If you master all of them, you might to teach applied econometrics class just as what I am doing now.
- We will introduce essentials of these methods in the class as many as possible. Let's start our journey together.

An Amazing But Tough Journey