Class 3: Home Alarm LTV, Testing, and Beyond

Professor Song Yao
Olin Business School

Customer Analytics

Conducting the A/B test is conceptually straightforward

HOW THE A/B TEST WORKS

The target group is <u>targeted</u> with the initiative

The control group is <u>not targeted</u> with the initiative

<u>Compare outcomes</u> between target and control group

<u>Difference</u> is the change in outcome caused by the initiative

But experiments/randomization can be hard to implement in practice

DIFFICULTIES

Can be technically difficult to implement

Can be expensive

- Measurement = comparison = withholding best ideas from some group for some period of time
- If one is sure that idea is good, experimentation is "wasteful"
- Can be slow

What to do when true experiments are not possible:

1. Matching
2. Difference-in-Differences
3. Regression Discontinuity (RD)

Do we really need probabilistically equivalent groups?

EXAMPLE: For-profit education industry (E.g., Coursera, Udacity)

Student attrition is a major problem

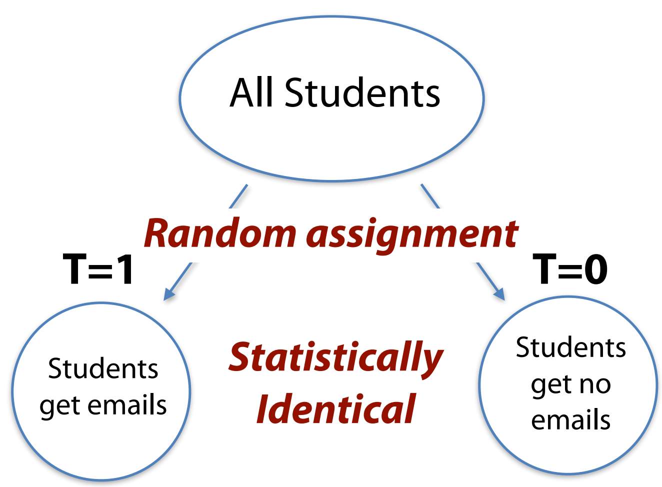
What keeps students on track?

One key relationship:

- Students who sign up weekly emails (e.g., updates about progress, assignments due, upcoming exams, etc.) are significantly more likely to graduate.
- But is this causal?

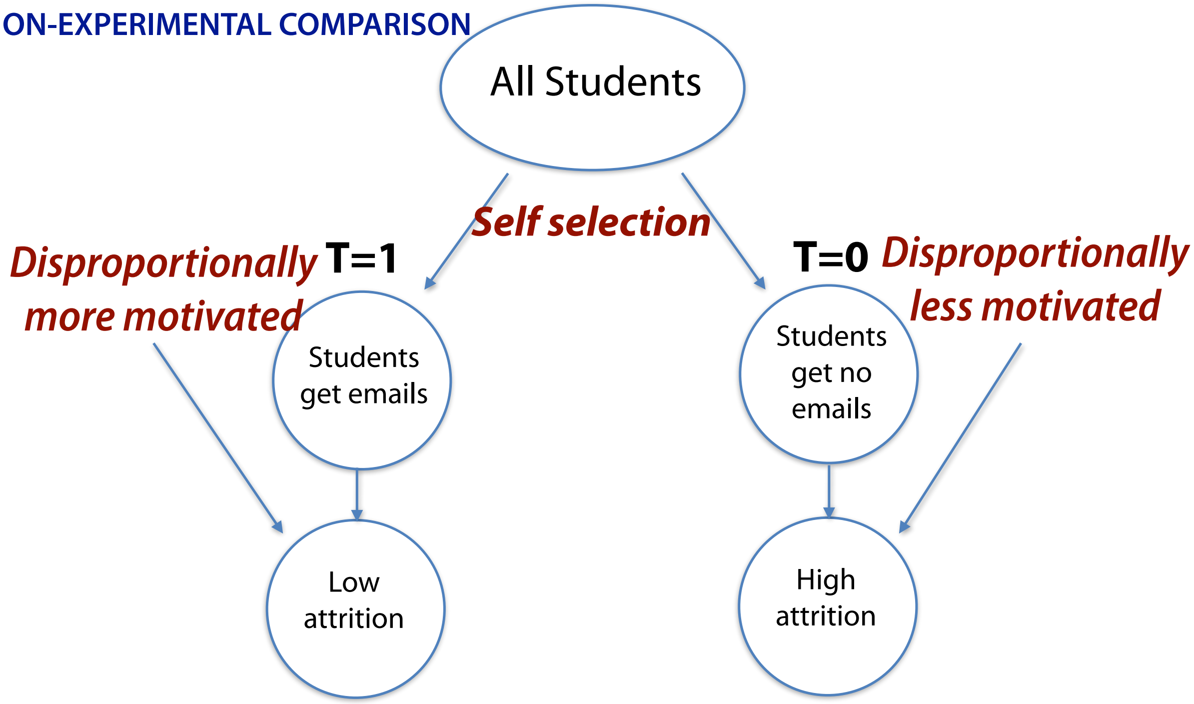
To find out whether the email has a causal effect, a randomized experiment would be ideal

EXPERIMENTAL SETUP



If students who get emails have lower attrition, do you feel comfortable interpreting this as a causal effect?

NON-EXPERIMENTAL COMPARISON



Email Data

ID	Email?	Complete?
2618643	0	0
3199888	0	1
2844779	0	0
3118111	0	0
2775543	0	0
3325986	0	0
3382231	0	0
3680449	0	1
1833421	0	0
2409691	0	1
1800623	1	1
1946297	1	1
2955486	1	0
3338292	1	0
3838594	1	1
2051366	1	0
2702835	1	1
2913122	1	1
3136996	1	1
2093118	1	1

When we can't run an experiment, matching is trying to create *"functionally equivalent"* groups

Email Data

ID	Email?	Complete?
2618643	0	0
3199888	0	1
2844779	0	0
3118111	0	0
2775543	0	0
3325986	0	0
3382231	0	0
3680449	0	1
1833421	0	0
2409691	0	1
1800623	1	1
1946297	1	1
2955486	1	0
3338292	1	0
3838594	1	1
2051366	1	0
2702835	1	1
2913122	1	1
3136996	1	1
2093118	1	1

No Email = 40% completion

Email = 70% completion

Email Data

ID	Email?	Complete?	Sex	GPA	Age
2618643	0	0	Female	3	23
3199888	0	1	Male	3.7	34
2844779	0	0	Male	2.7	42
3118111	0	0	Male	2.5	44
2775543	0	0	Female	2.9	35
3325986	0	0	Male	4	26
3382231	0	0	Female	3.9	29
3680449	0	1	Female	3.6	30
1833421	0	0	Male	3.1	34
2409691	0	1	Female	2.6	28
1800623	1	1	Male	3.1	41
1946297	1	1	Male	3.8	32
2955486	1	0	Male	2.8	42
3338292	1	0	Female	3.9	28
3838594	1	1	Male	3.6	25
2051366	1	0	Female	2.8	33
2702835	1	1	Male	2.3	46
2913122	1	1	Male	3.8	24
3136996	1	1	Male	3.9	30
2093118	1	1	Female	2.8	29

Email Data

ID	Email?	Complete?	Sex	GPA	Age
2618643	0	0	Female	3	23
3199888	0	1	Male	3.7	34
2844779	0	0	Male	2.7	42
3118111	0	0	Male	2.5	44
2775543	0	0	Female	2.9	35
3325986	0	0	Male	4	26
3382231	0	0	Female	3.9	29
3680449	0	1	Female	3.6	30
1833421	0	0	Male	3.1	34
2409691	0	1	Female	2.6	28
1800623	1	1	Male	3.1	41
1946297	1	1	Male	3.8	32
2955486	1	0	Male	2.8	42
3338292	1	0	Female	3.9	28
3838594	1	1	Male	3.6	25
2051366	1	0	Female	2.8	33
2702835	1	1	Male	2.3	46
2913122	1	1	Male	3.8	24
3136996	1	1	Male	3.9	30
2093118	1	1	Female	2.8	29

Email Data

ID	Email?	Complete?	Sex	GPA	Age
2618643	0	0	Female	3	23
3199888	0	1	Male	3.7	34
2844779	0	0	Male	2.7	42
3118111	0	0	Male	2.5	44
2775543	0	0	Female	2.9	35
3325986	0	0	Male	4	26
3382231	0	0	Female	3.9	29
3680449	0	1	Female	3.6	30
1833421	0	0	Male	3.1	34
2409691	0	1	Female	2.6	28
1800623	1	1	Male	3.1	41
1946297	1	1	Male	3.8	32
2955486	1	0	Male	2.8	42
3338292	1	0	Female	3.9	28
3838594	1	1	Male	3.6	25
2051366	1	0	Female	2.8	33
2702835	1	1	Male	2.3	46
2913122	1	1	Male	3.8	24
3136996	1	1	Male	3.9	30
2093118	1	1	Female	2.8	29

Email Data

ID	Email?	Complete?	Sex	GPA	Age
3199888	0	1	Male	3.7	34
2844779	0	0	Male	2.7	42
3118111	0	0	Male	2.5	44
2775543	0	0	Female	2.9	35
3382231	0	0	Female	3.9	29
2409691	0	1	Female	2.6	28
1946297	1	1	Male	3.8	32
2955486	1	0	Male	2.8	42
3338292	1	0	Female	3.9	28
2051366	1	0	Female	2.8	33
2702835	1	1	Male	2.3	46
2093118	1	1	Female	2.8	29

No Email = 33% completion

Email = 50% completion

What to do when true experiments are not possible:

1. Matching
2. Difference-in-Differences
3. Regression Discontinuity (RD)

We can back off from random assignments while still being able to show whether or not an initiative works

WAYS TO CREATE TARGET AND CONTROL GROUPS

Assign people to groups ...

1. completely randomly
2. **by geography (Miami vs. Chicago)**
3. **by time (Winter vs. Summer)**
4. **by time <u>and</u> geography**

Assign people to groups based on time <u>and</u> geography

HOW IT WORKS

- Pick 2 geographic regions with *<u>similar time trends</u>* (by ZIP, region, etc.)
- Pick 2 time periods (by week, month, etc.)

	Region 1	Region 2
Period 1	Outcome in Region 1 before test	Outcome in Region 2 before test
Period 2	Outcome in Region 1 during test	Outcome in Region 2 during test
	Target Group (A)	
	Control Group (B)	

Assign people to groups based on time and geography

Digital coupons example

	Region 1	Region 2
Period 1	Average spending: 110	Average spending: 90
Period 2	Average spending: 120	Average spending: 140
	Target Group (A)	
	Control Group (B)	

Assign people to groups based on time and geography

Digital coupons example

	Region 1	Region 2
Period 1	Average spending: 110	Average spending: 90
Period 2	Average spending: 120	Average spending: 140
	Target Group (A)	
	Control Group (B)	

Difference between ? 120 - 110 = 10

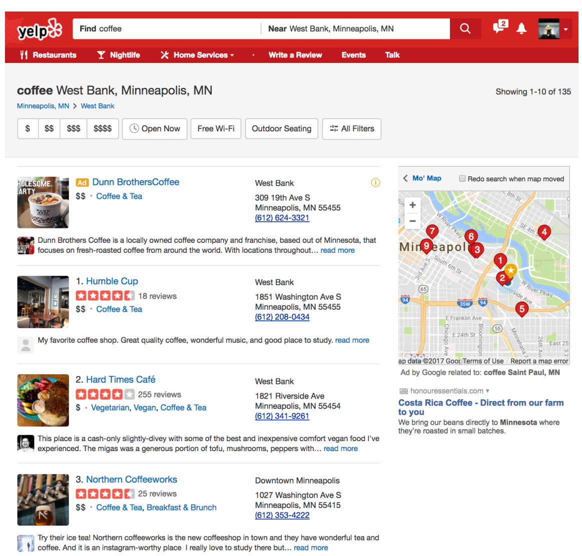
What would happen to Region 2 if there was no promotion? 90 + 10 = 100

What is the effect of the promotion? 140 - 100 = **40**

What to do when true experiments are not possible:

1. Matching
2. Difference-in-Differences
3. Regression Discontinuity (RD)

Does online reviews rating affect sales?



An amazing new chef

- Better online review ratings
- More customers

Yelp (and many other websites) rounds ratings!

Key assumption:

- Restaurants with 4.25 rating and with 4.24 rating are "identical" in all other dimensions pertaining to quality.
  - (E.g., similarly amazing chefs!)

The difference in their sales can be attributed to the ratings of 4.3 vs. 4.2

Diff-in-Diff Example

Does social media increase demand?

Social Media Marketing

- 83% of Fortune 500 companies have an active presence
- Spending expects to soon exceed 20% of marketing budgets

85% of surveyed marketing executives were unsure about the effectiveness

Testing the effectiveness of tweeting is not easy!

* Source: Seiler, Yao, and Wang (2017): "Does Online Word-of-Mouth Increase Demand? (and How?) Evidence from a Natural Experiment", *Marketing Science*, 36(6), pp. 838–861.

We tested the effectiveness of tweets using a unique setting of Tweeting in China

* Source: Seiler, Yao, and Wang (2017): "Does Online Word-of-Mouth Increase Demand? (and How?) Evidence from a Natural Experiment"

Testing the effectiveness of tweeting is not easy!

CORE CHALLENGE

How do you randomly construct treatment vs. control groups that have different levels of online WOM?

The Chinese government did us a big favor!

A political scandal shut down Sina Weibo for 3 days in 2012

Instead of one pair of treatment and control groups, we have multiple pairs at our disposal

DIFF-IN-DIFF NATURAL EXPERIMENT

- Hong Kong
- 24 Mainland Cities (or Shenzhen, Guangdong Province)
  - *Regular time:* Ratings (Viewerships) of HK and Mainland China gives geographic difference
  - *During the block:* Ratings (Viewerships) of HK gives seasonal effect.

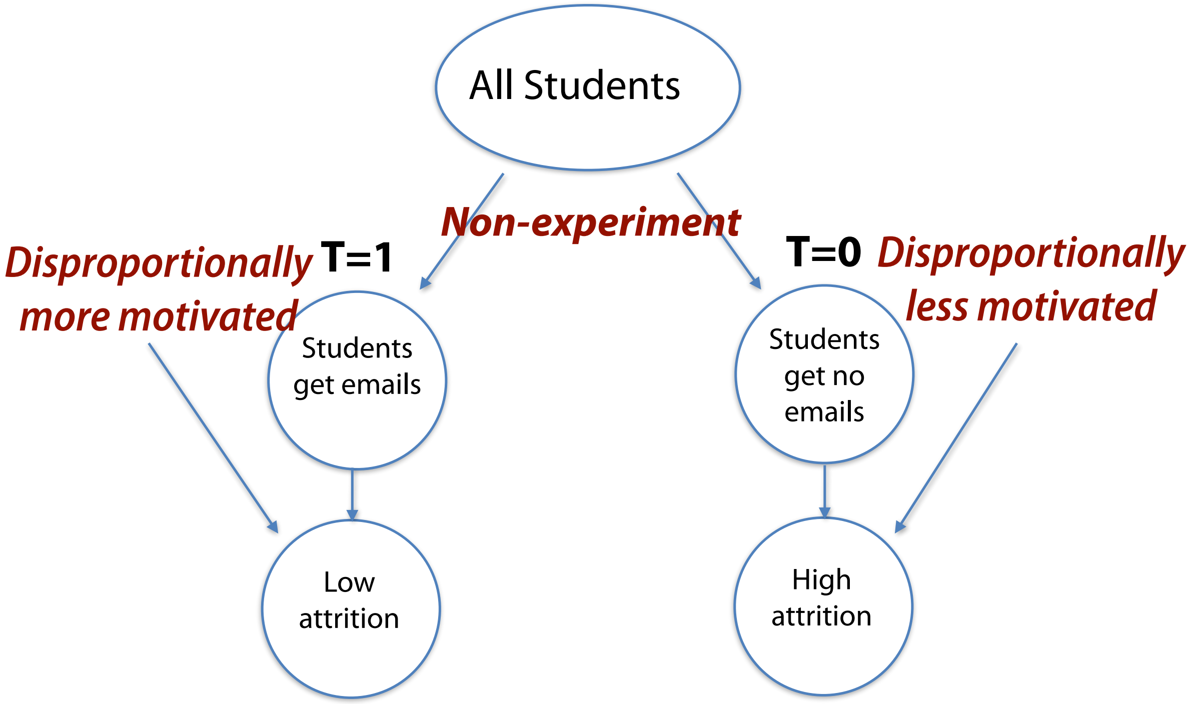
Instead of one pair of treatment and control groups, we have multiple pairs at our disposal

DIFF-IN-DIFF NATURAL EXPERIMENT

- Hong Kong: The censorship has NO EFFECT
- 24 Mainland Cities: Rating (Viewerships) dropped 3.1%

Correlation is not Causation!!

Correlation is not Causation!! Think about the Students Email case



Beyond the intuition—Let's start with some math

Common language of the causal inference clan

- T_i or $D_i = \begin{cases} 1 & \text{if unit/subject i received the treatment} \\ 0 & \text{otherwise} \end{cases}$
- Y_i is the observed outcome of unit i.
- Y_{i0} or $Y_i(0)$ is the *potential outcome* if unit i is not treated
- Y_{i1} or $Y_i(1)$ is the *potential outcome* if **the same unit i** is treated
  - *We can only observe either Y_{i1} or Y_{i0}, but not both*
- The treatment effect of i is $Y_{i1} - Y_{i0}$

The treatment effect of i is $Y_{i1} - Y_{i0}$, BUT…

Never observe i with *AND* without treatment at the same time!

- Average Treatment Effect (ATE) = $E(Y_1) - E(Y_0)$
- Average Treatment Effect on the Treated (ATT or ATET) = $E(Y_1 | T = 1) - E(Y_0 | T = 1)$

Still, it is tricky to estimate ATE and ATT. See the toy example below

Cust.	Treated	Y1 (spending if treated)	Y0 (spending if not treated)	Treatment Effect
1	1	700	600	100
2	1	500	450	50
3	0	550	500	50
4	0	450	400	50

NOTE: Red are what we can observe in data. Grey are not observed

- ATE=(100+50+50+50)/4=62.5
- ATT=(100+50)/2=75
- What we can observe is
  - $E(Y | T = 1) - E(Y | T = 0)$ =(700+500)/2-(500+400)/2=150

Why the substantial differences among the estimates?

BIAS! But where does the bias come from?

Observed and can be calculated

- $E(Y_1 | T = 1) - E(Y_0 | T = 0)$

Let's add and subtract $E(Y_0 | T = 1)$, the "counterfactual" of the treated if they were not treated

$$\begin{aligned} & E(Y_1 | T = 1) - E(Y_0 | T = 0) \\ = & E(Y_1 | T = 1) - E(Y_0 | T = 0) + E(Y_0 | T = 1) - E(Y_0 | T = 1) \\ = & E(Y_1 | T = 1) - E(Y_0 | T = 1) \Longleftarrow \text{ATT} \\ & + E(Y_0 | T = 1) - E(Y_0 | T = 0) \Longleftarrow \text{Bias} \end{aligned}$$

- Source of the bias? Treated and control groups already differ in their potential outcomes even without the treatment!

How can we address the bias?

$$E(Y_1 | T=1) - E(Y_0 | T=0) = E(Y_1 | T=1) - E(Y_0 | T=1) \quad \text{ATT}$$
$$+E(Y_0 | T=1) - E(Y_0 | T=0) \quad \text{Bias}$$

The bias would dissappear if $E(Y_0 | T=1) = E(Y_0 | T=0)$

- But what does $E(Y_0 | T=1) = E(Y_0 | T=0)$ mean?
- The treatment ($T = 1$ or 0) has no relationship with potential outcome Y_0
  - E.g., If Pentathlon sends more emails (T) to customers already having higher spending (Y), the estimated effect using observed data is biased.

The "Unconfoundedness" Requirement/Assumption

For causal inference using observational data, one of the most important requirements: **Unconfoundedness**

- $E(Y_0 | T=1) = E(Y_0 | T=0)$: The treatment has no relationship with potential outcome Y_0
- $E(Y_1 | T=1) = E(Y_1 | T=0)$: The treatment has no relationship with potential outcome Y_1
- Together, these two are called **Unconfoundedness**
  - $Y_1, Y_0 \perp T \Rightarrow \begin{cases} E(Y_0 | T=1) = E(Y_0 | T=0) \\ E(Y_1 | T=1) = E(Y_1 | T=0) \end{cases}$
- Further, if the condition $Y_1, Y_0 \perp T$ is true,

$$E(Y_1 | T=1) - E(Y_0 | T=1) \quad \text{ATT}$$
$$= E(Y_1) - E(Y_0) \quad \text{ATE}$$

A/B Tests vs. Non-experiment Settings

- A/B tests satisfy the condition $Y_1, Y_0 \perp T$ because of random assignment.
- For non-experiment settings, we must rely on various modeling approaches
  - E.g., Matching—trying to reduce the (observed) attribute differences between treated and control groups, subsequently, reducing the differences in potential outcomes between the treated and control groups.

ID	Email?	Complete	Sex	GPA	Age
3199888	0	1	Male	3.7	34
2844779	0	0	Male	2.7	42
3118111	0	0	Male	2.5	44
2775543	0	0	Female	2.9	35
3382231	0	0	Female	3.9	29
2409691	0	1	Female	2.6	28
1946297	1	1	Male	3.8	32
2955486	1	0	Male	2.8	42
3338292	1	0	Female	3.9	28
2051366	1	0	Female	2.8	33
2702835	1	1	Male	2.3	46
2093118	1	1	Female	2.8	29

No Email = 33% completion

Email = 50% completion