IEOR E4650 Business Analytics

Session 17: Prescriptive Analytics: Difference-in-Difference Methodology and Real-time A/B Testing

Spring 2018

Prof. Adam Elmachtoub

DiD Session Outline

- Evaluating the Buy Online Pickup in Store (BOPS) program at a *Home and Kitchen*
  - Analyzing the impact
  - Prescription (keep or drop)

  Source: "Integration of Online and Offline Channels in Retail: The Impact of Sharing Reliable Inventory Availability Information," 2014. Gallino, S., Moreno, A., *Management Science*.

- Difference-in-Differences (DiD) Method

- Measuring impact of search engine marketing (SEM) at eBay

Prescriptive Analytics: Testing

$$\text{decision/treatment} \Longrightarrow \text{outcome}$$

- How can we quantify the impact of a treatment?
- Looking at the outcome alone ignores *confounding factors*
- Testing seeks to isolate the treatment's *causal effect*
- Common techniques:
  - A-B Testing (randomized trials)
  - Difference-in-Differences (DiD) Method

Decision Time

Home and Kitchen should:

(1) Drop the BOPS initiative

(2) Move ahead and deploy BOPS to Canada

Home and Kitchen Sales Results

	Time period	**Online sales ($)** Average Sales per week per DMA	**B&M sales ($)** Average Sales per week per store
Before BOPS	Apr 11-Oct 11	14,738	67,646
After BOPS	Oct 11-Apr 12	12,734	60,101
	Change	−2,004	−7,545

What can we conclude about

BOPS from these results?

What Might We Be Missing?

Many other factors could be affecting sales between the before and after periods

- Seasonality (holidays, back-to-school, summer moving season)
- Macro-economic factors (growth, shocks)
- Systemic company-wide factors (product selection, marketing)
- Systemic competitive factors (entrance of new competitor)

How can we isolate the effect of BOPS from all these other confounding factors?

Isolating the Impact of BOPS on Online Sales

General idea of the difference in differences (DiD) approach

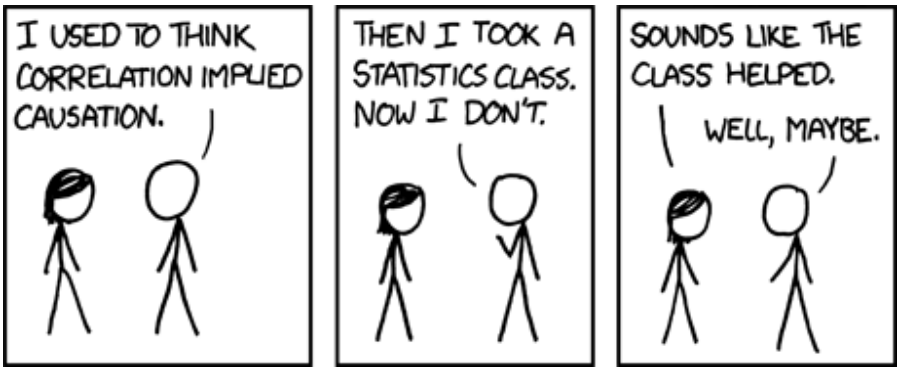
- Identify a *control group*
  - Similar to the test group and subject to the same common factors
  - Not exposed to the treatment
- Compare the change in outcome in the control group to the change in outcome in the test group

Example of DiD

- A recent study at Columbia College reports that freshmen who participate in club sports gain an average of 3.6 pounds during their first year of college
- Does participating in sports cause weight gain?

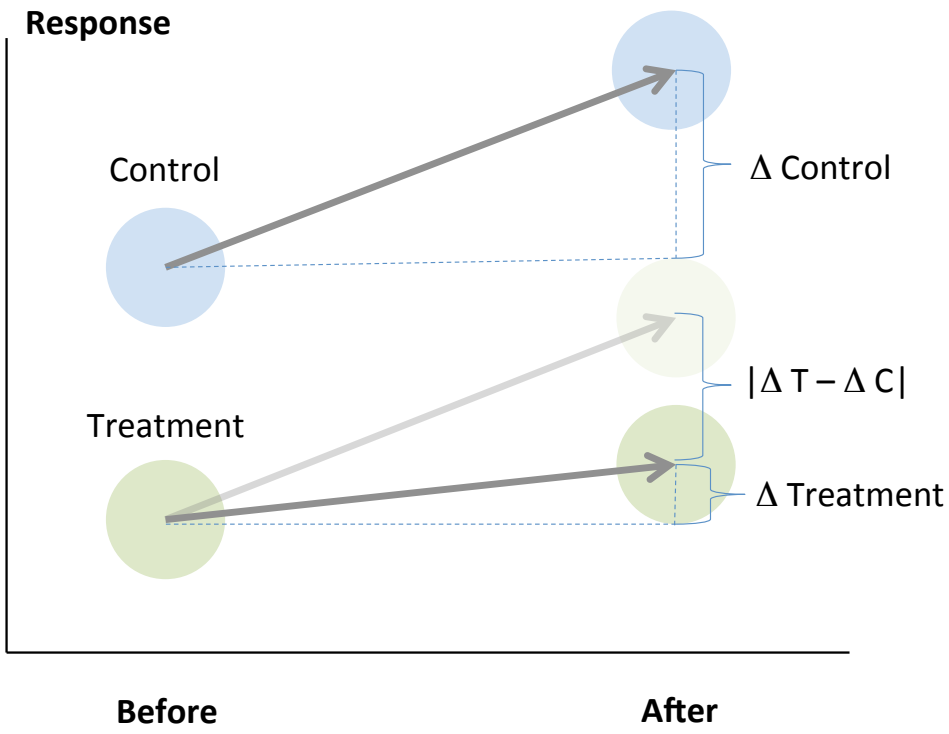
Student group	Average starting weight	Average ending weight	Difference
Club sports	144.3	147.9	3.6
No club sports	149.2	156.3	7.1
		DiD	−3.5

Correlation is not Causation



1. A recent article in the New England Journal of Medicine reports a striking correlation between chocolate consumption and number of Nobel prize recipients across countries. What are some possible explanations for this pattern?
2. A recent Credit Suisse study found that companies with women on their board of directors outperform companies with all male boards.
3. Scientific studies report that married men live longer than their single counterpart.

The Concept Behind DiD



What About Using a Randomized Trial?

A-B Testing: Membership in treatment and control groups selected randomly

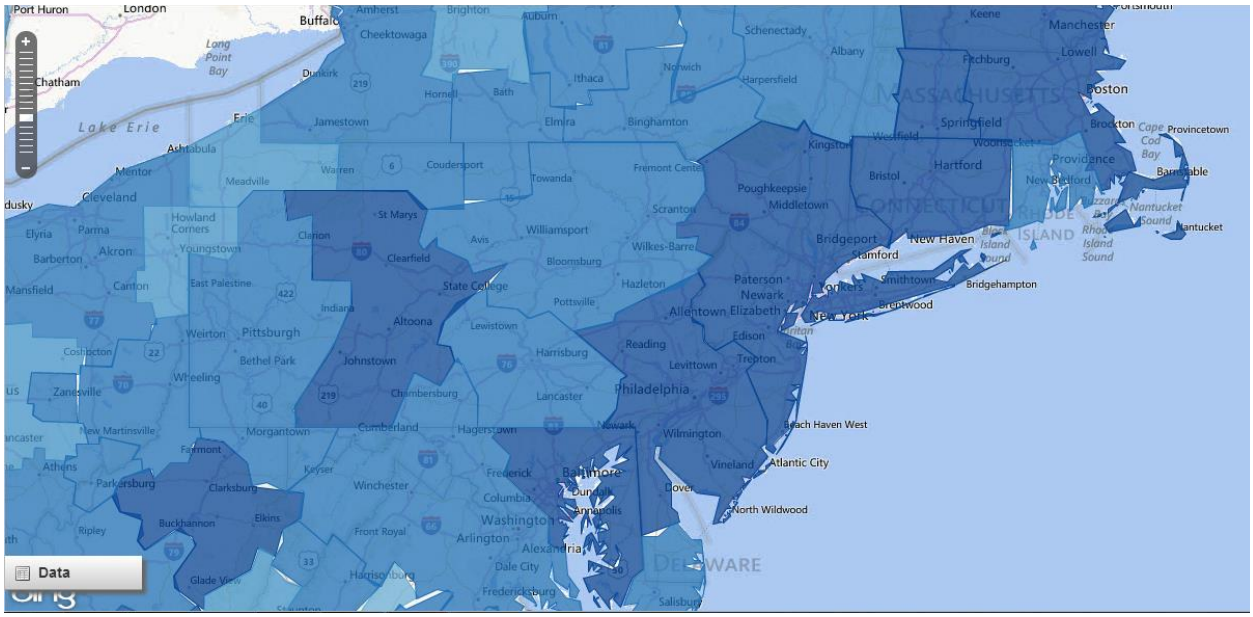
- Pros
  - Little chance of systematic differences between treatment & control
  - Allows us to isolate the true effect
- Cons
  - Can be difficult to operationalize
  - Smaller sample size
  - Hawthorne effect: groups may modify behavior knowing they are being observed
  - Opportunity cost: can we afford to wait for the results of randomized test?
- What are the alternatives?

  Finding the best control group possible *ex post*

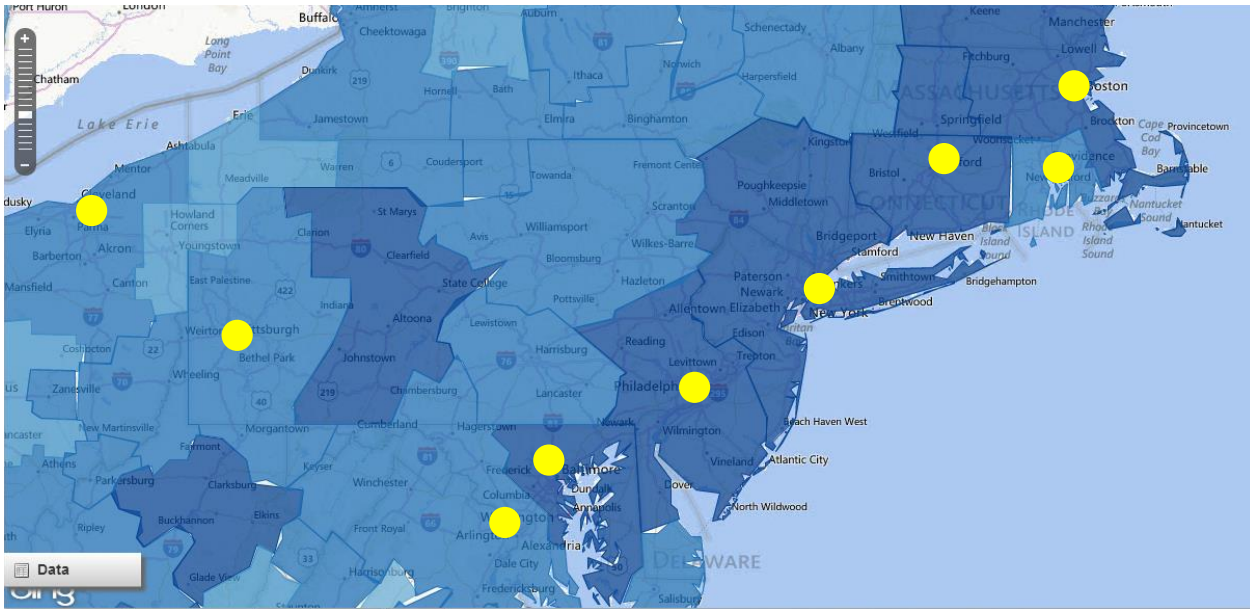
How can we apply DiD to analyze on-line sales?

What are the treatment and control groups?

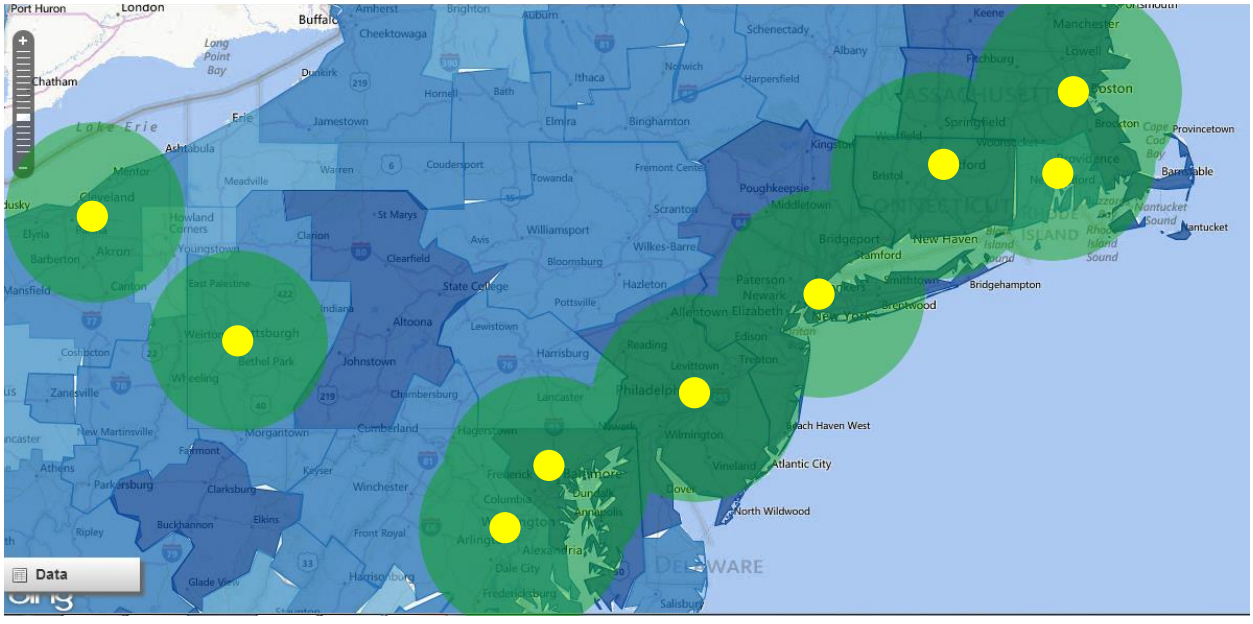
Which DMAs are Affected by BOPS?



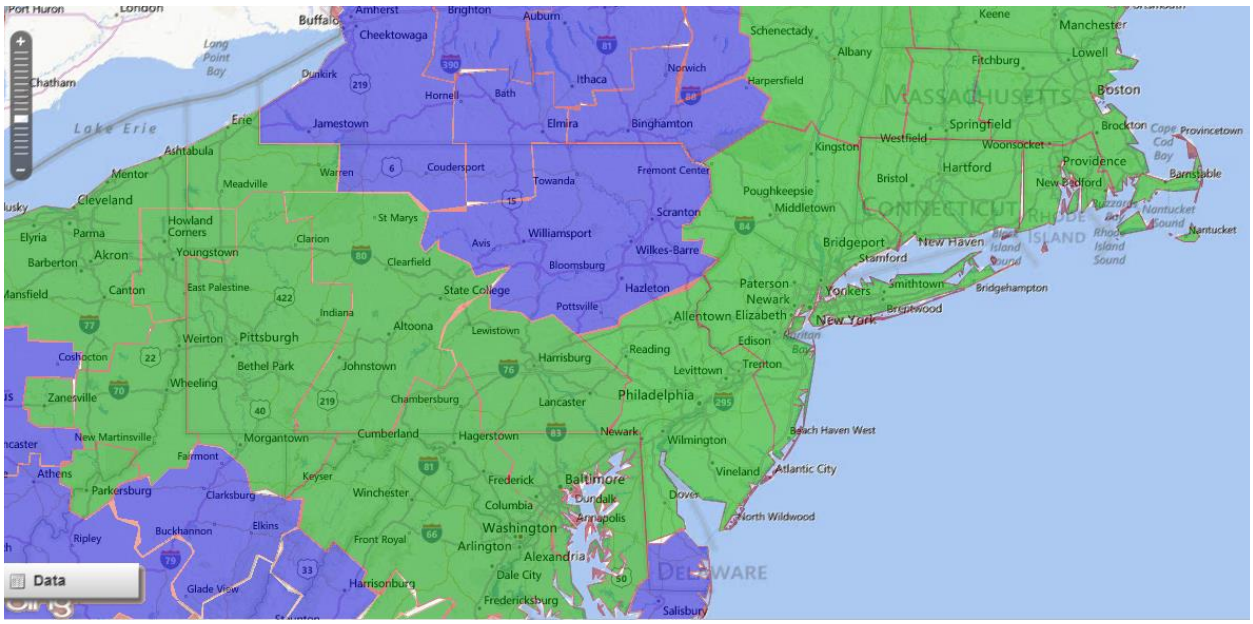
Store Locations



50 Mile Radius From Stores



"Close" and "Far" DMAs



Close DMAs (treatment)

Far DMAs (control)

DiD for Online Sales Analysis

Treatment: DMA has a store close to it (within 50 miles)

Control: DMA does not have a store close to it (further than 50 miles)

	Online Sales ($M)		**Difference**	
	Before	After	$ M	%
Far DMAs (no BOPS)	44.4	37.5	−6.9	−15.4%
Close DMAs (BOPS)	36.1	29.3	−6.8	−18.7%
Total	80.5	66.8		
			DiD	−3.3%

What about the effect on B&M stores?

How Can We Apply DiD to B&M Stores?

What are the treatment and control groups?

Impact on B&M Store Sales

	Sales ($M)		Difference	
	Before	After	$ M	%
CAN (no BOPS)	30.7	25.8	−4.8	−15.8%
US (BOPS)	122.7	110.5	−12.3	−10.0%
Total	153.4	136.3		
			DiD	5.8%

Aggregate Impact of BOPS on Sales

Estimated impact of BOPS on *Home and Kitchen* sales

- Online sales affected by BOPS: \$36.1M $\times$ −3.3% = \$ −1.2M
- B&M sales affected by BOPS: \$123M $\times$ 5.8% = \$7.1M

Estimated aggregate impact on sales

- \$7.1M −\$1.2M = \$5.9M
- 2.9% increase in company-wide revenues

Decision

Home and Kitchen should:

(1) Drop the BOPS initiative

(2) Move ahead and deploy BOPS to Canada

Other considerations?

DiD: A General Approach to Measurement of Effects

Two inputs

- A treatment was implemented at a point in time and we want to evaluate its impact
- Some units received treatment, some did not (treatment and control groups)

Examples

- A firm changed suppliers and suddenly sees fewer quality problems for the end product
- A firm launched an advertising campaign and sales declined

Another possible approach is A-B testing. DiD can be a practical alternative.

DiD in Linear Regression Analysis

Obtaining DiD Results Through Linear Regression

The DiD method computed the impact of the BOPS treatment according to:

$$(\%\text{TotalSalesChange})_{\text{TREATED}} - (\%\text{TotalSalesChange})_{\text{CONTROL}}$$

This aggregates changes across all units (stores or DMAs) in the two groups. It makes the assumption that the two groups are similar, aside from the treatment.

An alternative is to explain the change on a unit-by-unit basis, using linear regression.

This alternative gives us flexibility to correct for confounding differences between the two groups.

Obtaining DiD Results Through Linear Regression

Regression equation

$$\%\text{SalesChange}_i = a + b \times \text{TREATED}_i + \text{error}$$

- i: index of unit (DMAs or stores)
- TREATED_i: dummy variable indicating whether unit i (DMA or store) received treatment (1), or not (i.e., the DMA was close to a store for online sales, or the store was located in the US for B&M sales)
- b measures the impact of the treatment

Online Sales: Data and Regression Results

	A	B	C	D	E	F	G	H	I	J	K
1	**id (DMA)**	**Before**	**After**	**Sales Change**	**Affected**						
2	1	650,039	531,296	-18.27%	1		**SUMMARY OUTPUT**				
3	2	1,818,505	1,976,250	8.67%	0						
4	3	517,513	346,929	-32.96%	1		*Regression Statistics*				
5	4	84,948	74,002	-12.88%	1		**R Square**	1.7%			
6	5	892,664	549,045	-38.49%	0		**Observations**	210			
7	6	316,062	237,479	-24.86%	0						
8	7	1,915,247	1,585,285	-17.23%	1			*Coefficients*	*Standard Error*	*t Stat*	*P-value*
9	8	782,996	580,016	-25.92%	1		**Intercept**	-0.170	0.010	-17.4	0.00
10	9	274,908	242,939	-11.63%	1		**Affected**	-0.027	0.014	-1.9	0.06
11	10	2,297,303	1,750,102	-23.82%	1						
12	11	861,079	779,517	-9.47%	0						
13	12	78,058	63,798	-18.27%	0						
14	13	185,225	161,969	-12.56%	1						
15	14	10,402	11,049	6.22%	0						

-3.3% lies in the 95% confidence interval

Note: the results of the regression are slightly different than the aggregate analysis because DMAs have different sizes

B&M Sales: Data and Regression Results

	A	B	C	D	E	F	G	H	I	J	K
1	**id (stores)**	**Before**	**After**	**Sales Change**	**Affected (US)**						
2	1	3,426,216	3,067,961	-10.46%	0		**SUMMARY OUTPUT**				
3	3	1,286,236	1,138,918	-11.45%	1						
4	5	2,724,176	2,518,139	-7.56%	1		*Regression Statistics*				
5	7	2,220,210	1,772,500	-20.17%	1		**R Square**	14.1%			
6	9	2,647,521	2,617,902	-1.12%	1		**Observations**	84			
7	11	1,725,954	1,570,215	-9.02%	1						
8	13	1,466,022	1,280,101	-12.68%	1			*Coefficients*	*Standard Error*	*t Stat*	*P-value*
9	15	2,547,007	2,305,288	-9.49%	1		**Intercept**	-0.159	0.014	-11.4	0.00
10	17	2,610,319	2,210,091	-15.33%	1		**Affected (US)**	0.057	0.016	3.7	0.00
11	19	1,292,153	1,150,925	-10.93%	0						
12	21	837,044	835,597	-0.17%	1						
13	23	1,947,191	1,540,438	-20.89%	0						
14	25	816,350	682,600	-16.38%	0						

5.8% lies in the 95% confidence interval

Note: the results of the regression are slightly different than the aggregate analysis because stores have different sizes

DiD Using Linear Regression: Adding Explanatory Variables

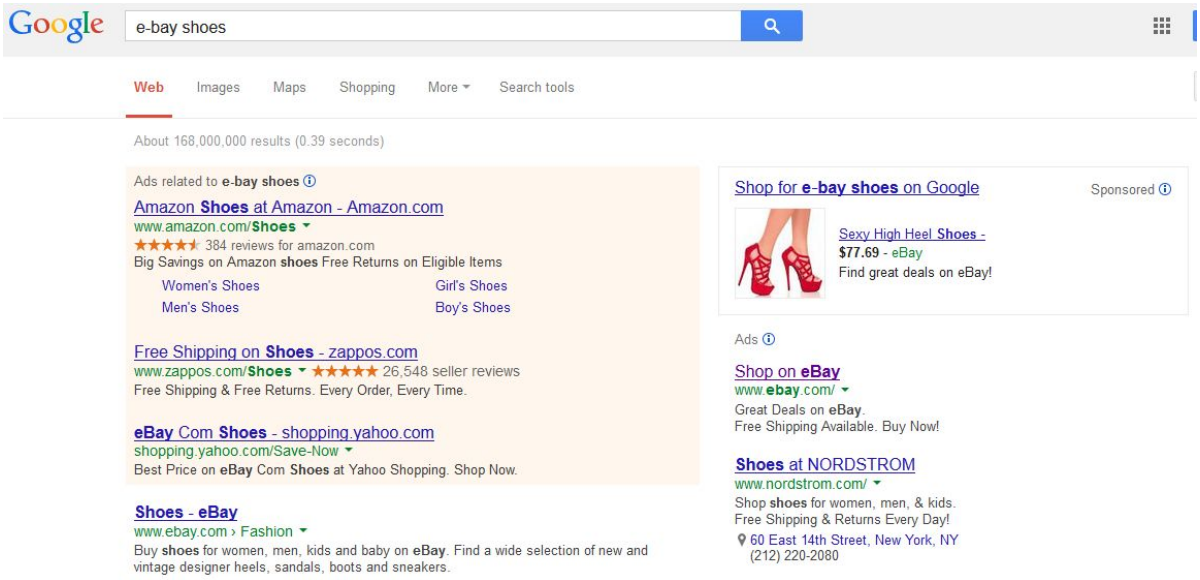
Suppose, in some stores, Home and Kitchen competes with Bed, Bath, & Beyond. This can be accounted for in the regression method:

$$\begin{aligned}\%\text{SalesChange}_i \quad = \quad & a \; + \; b \times \text{TREATED}_i + c \times \text{BBB}_i \\ & + \text{other explanatory variables} \; + \; \text{error}\end{aligned}$$

- i: index of store
- BBB_i: dummy variable indicating whether Bed, Bath, & Beyond is a competitor to store i (1), or not (0)
- b measures the impact of the treatment
- other explanatory variables can similarly be added (e.g., store specific variables, seasonality, etc.) to refine the impact estimate and correct for confounding variables

DiD: Other Applications

Search Engine Marketing (SEM) at eBay



Two types of keywords:

- Branded (all queries that include the word eBay)
- Non-branded (other queries)

Measuring the Impact of Advertising at eBay

Context: In 2010, eBay spent $4 million per month on "search engine marketing" (SEM) (also known as sponsored search advertising)

Cost of SEM: pay a fee each time a customer clicks on the ad

Question: How to compute the ROI of SEM?

Measuring the Impact of Advertising: Google's Advice

Calculating your ROI for sales

Determining your AdWords ROI is straightforward if your business goal is web-based sales. You'll need 3 numbers:

- Your revenue made via AdWords advertising
- Costs related to the products you sold
- Your AdWords costs (available in the Campaigns tab of your AdWords account)

Calculate your net profit by subtracting your costs from your AdWords revenue for a given time period. Then divide your net profit by your AdWords costs to get your AdWords ROI for that time period. Here's an example:

($1300	-	$1000)	/	$1000	=	0.3
Your revenue (measured by conversions)		Your overall costs		Your AdWords costs		Your ratio of profit to advertising cost is 30% -- this is your AdWords ROI.

https://support.google.com/adwords/answer/1722066

Comments?

Measuring the Impact of Advertising at eBay

How to measure the impact of advertising?

Approach: testing

(Focus on non-branded keywords for this example)

Measuring the Impact of SEM at eBay

Experiment

- Construct treatment/control groups through DMAs
  Easy to restrict ads by geographic area
- **Treated group:**
  Out of 210 DMAs, selected randomly 65 DMAs where SEM would be turned off on Google for two months
- **Control group:**
  Create a control group of DMAs that match the previous 65 DMAs (similar traffic seasonality)

Measuring the Impact of SEM at eBay

Estimate the impact of SEM on Sales by using Differences in Differences:

(Difference in Sales in treated group) − (Difference in Sales in control group)

Result:

SEM increases sales by about 0.44% (and this is not statistically significant)

ROI estimate: −63% (short term estimate)

Source: "*Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment*," *Econometrica*, 2015. Blake, T., Nosko, C and Tadelis, S.
http://conference.nber.org/confer/2013/EoDs13/Tadelis.pdf

Takeaways

- Live A/B testing provides a valuable way to make decisions with data, but
  - Can be expensive
  - Not always possible (but sometimes it is, next class!)
- Naive before-after comparisons to evaluate a change can be misleading
- Difference in differences (DiD) provides a way to isolate the impact of a change and obtain more accurate estimates of impact
- DiD requires
  - Treatment group
  - Control group, that has similar trends at the treatment group

Motivation

- You are considering adding a new feature to your website or app and have a few different ideas (font, size, location, etc.)

How do you pick the best layout?

- You are going to buy advertising slots from Google, and have a few different ads you can show (different text and/or promotions)

Which ad is the best? Does it depend on the user?

Problem Setup

- Start with no data about the effectiveness of each alternative (ad, layout)
- We would like to know the success probability of each action
- Want to maximize the total success rate (clicks, conversions) over some time horizon (day, week, or month)
- Every time a decision must be made for a user, we need to pick one of our alternatives
- Want to always pick the best alternative, but we also need to experiment to learn what is the best!

Other Examples

- Email campaigns
  - You are going to start an email campaign with different possible promotions or messages to drive traffic to your website.
  - Which promotions should we email our customers? Personalize based on user?
- Clinical Trials
  - You are running a trial for a treatment and have different dosing or treatment possibilities.
  - What's the best dose or treatment? Personalize based on patient?
- Online Education
  - You have several different ways of creating online teaching videos (different lecturer, content, pace)
  - Which one works best? Personalize based on student?

Real world software

- `https://support.google.com/analytics/answer/2844870?hl=en`
- `https://mailchimp.com/features/ab-testing/`
- `https://www.optimizely.com/ab-testing/`
- `http://www.adobe.com/marketing-cloud/testing-targeting/ab-testing.html`

The Multi-armed Bandit Framework I

- There are K different actions (ad, layout, promotion, etc.) decision-maker can take
- In each time step t, the decision-maker observes a user, chooses an action, and then observes a response (click/no click, success/fail)
- Decision-maker's objective is to maximize total number of successes over T periods .

The Multi-armed Bandit Framework II

- The probability of success for each action is $a = 1, \ldots, K$ is p_a, which is unknown!
- If decision-maker uses action a_t on the t'th user, they observe $Y_t \sim Ber(p_{a_t})$. ($Y_t = 1$ w.p. p_{a_t}, $Y_t = 0$ w.p. $1 - p_{a_t}$)
- We never observe p_{a_t}, just the binary indicator of success!

Goal: Maximize Expected Total Successes

$$\mathsf{E}[\sum_{t=1}^{T} Y_t] = \sum_{t=1}^{T} p_{a_t}$$

Equivalent Goal: Minimize Cumulative Regret

$$T \max_{a=1,\ldots,K} p_a - \sum_{t=1}^{T} p_{a_t}$$

Exploration vs. Exploiation (Learning vs. Earning)

- What if the expected payouts for each machine were known?
  - The decision-maker always chooses the action with the highest probability of success and earns $T \max_{a=1,\ldots,K} p_a$
- We assume the decision-maker has *no prior knowledge* about each action's success rate.
  - However, the decision-maker can collect data and predict success rates of each action.

Convenient Notation

- Let $n_{t,a}$ be the number of times action a chosen up to time t, i.e.,

$$\hat{n}_{t,a} = \sum_{s=1}^{t-1} \mathbb{I}(a_s = a)$$

- Let $\hat{p}_{t,a}$ be the estimated probability of success of action a at time t, i.e.

$$\hat{p}_{t,a} = \frac{\sum_{s=1}^{t-1} Y_s \mathbb{I}(a_s = a)}{n_{t,a}}$$

MAB Algorithms

1. Pure Exploration (bad)
   - Try a random action in each period.
2. Pure Exploitation (bad)
   - Choose the action according to $\text{argmax}_a \hat{p}_{t,a}$
3. Explore then Exploit (meh)
   - Do Pure Exploration for $T_0 < T$ users, then do Pure Exploitation for remaining $T - T_0$ users
4. Epsilon-Greedy (good)
   - In period t, do Pure Exploration w.p. $1/t$, and Pure Exploitation w.p. $1 - 1/t$.
5. Upper Confidence Bound (UCB) (great)
   - Choose the action according to $\text{argmax}_a \hat{p}_{t,a} + \text{“}std.err.(\hat{p}_{t,a})\text{”}$

Pure Exploration Policy

- Pure Exploration: The decision-maker chooses the actions sequentially.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = ?$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = ?$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = ?$

Payouts

Pure Exploration Policy

- Pure Exploration: The decision-maker chooses the actions sequentially.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = ?$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = ?$

Payouts

0		

Pure Exploration Policy

- Pure Exploration: The decision-maker chooses the actions sequentially.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = ?$

Payouts

0	0	

Pure Exploration Policy

- Pure Exploration: The decision-maker chooses the actions sequentially.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = 1$

Payouts

0	0	1

Pure Exploration Policy

- Pure Exploration: The decision-maker chooses the actions sequentially.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = 1$

Payouts

0	0	1
0		

Pure Exploitation Policy

- Pure Exploitation: The decision-maker always chooses the action with the highest estimated probability of success.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = ?$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = ?$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = ?$

Payouts

Pure Exploitation Policy

- Pure Exploitation: The decision-maker always chooses the action with the highest estimated probability of success.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = ?$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = ?$

Payouts

0		

Pure Exploitation Policy

- Pure Exploitation: The decision-maker always chooses the action with the highest estimated probability of success.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = ?$

Payouts

0	0	

Pure Exploitation Policy

- Pure Exploitation: The decision-maker always chooses the action with the highest estimated probability of success.

$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$

$\hat{p}_{t,1} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$

$\hat{p}_{t,2} = 0$

$Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,3} = 1$

Payouts

0	0	1

Pure Exploitation Policy

- Pure Exploitation: The decision-maker always chooses the action with the highest estimated probability of success.



$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$ $Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$ $Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,1} = 0$ $\hat{p}_{t,2} = 0$ $\hat{p}_{t,3} = 1/2$

Payouts

0	0	1
		0

Pure Exploitation Policy

- Pure Exploitation: The decision-maker always chooses the action with the highest estimated probability of success.



$Y_t \sim \text{Ber}\left(\frac{1}{16}\right)$ $Y_t \sim \text{Ber}\left(\frac{1}{2}\right)$ $Y_t \sim \text{Ber}\left(\frac{1}{4}\right)$

$\hat{p}_{t,1} = 0$ $\hat{p}_{t,2} = 0$ $\hat{p}_{t,3} = 1/3$

Payouts

0	0	1
		0
		0

Exploration-Exploitation Trade-off

- Policies which purely exploit can converge to a sub-optimal action.
- However, too much exploring can waste a lot of opportunities for success.
- Exploration-Exploitation trade-off: in each time step, the decision-maker faces a trade-off between:
  1. maximizing estimated expected payout (exploiting), and
  2. maximizing information gain (exploring)

Explore-Then-Exploit Policy

- Input parameter: $T_0 \in 1, ..., T$
- Explore-Then-Exploit Algorithm: Choose the actions sequentially for the first T_0 time steps (explore). Then, choose the action which maximizes expected reward (exploit) until time T.

Epsilon-Greedy Policy

- Input parameter: $\epsilon \in [0, 1]$
- Epsilon-Greedy Algorithm: In each time step t, choose a random action with probability ϵ (explore). Otherwise, choose the action which maximizes expected reward (exploit), i.e., $a_t = \text{argmax}_a \hat{p}_{t,a}$.
- Epsilon is typically either:
  1. a constant (e.g., $\epsilon = 0.1$), or
  2. a decreasing function of t (e.g., $\epsilon_t = 1/t$)

Upper Confidence Bounding

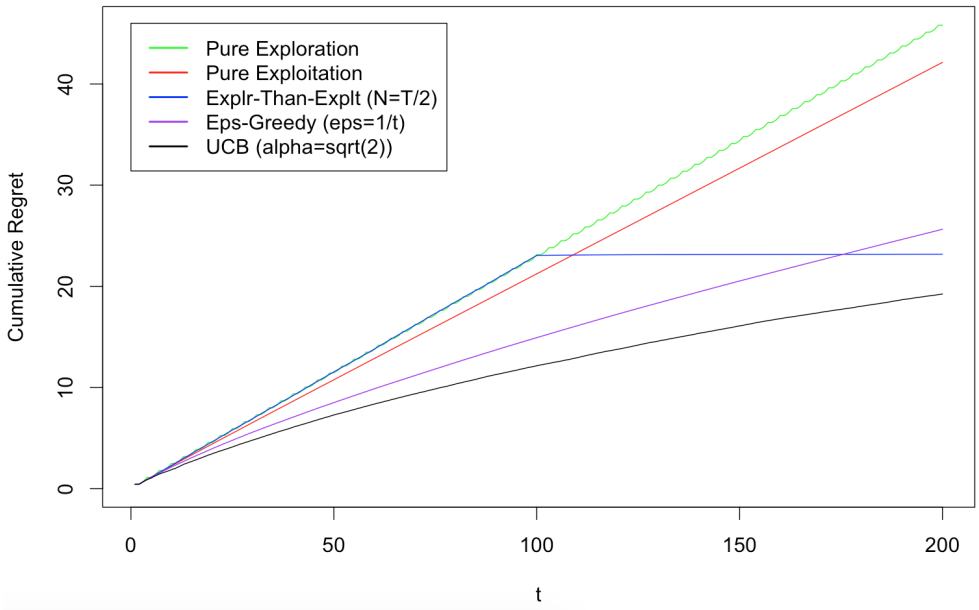
- Input parameter: $\alpha \in [0, \infty)$ (commonly set to $\sqrt{2}$)
- Upper Confidence Bounding (UCB): In each time step t, choose the action a_t which maximizes the UCB score:

$$UCB_{t,a} = \hat{p}_{t,a} + \alpha \sqrt{\frac{\log(t)}{n_{t,a}}}$$

  - $\hat{p}_{t,a}$: estimated probability of success for action a
  - $n_{t,a}$: number of times action a has been selected through t periods
- The second term represents our *uncertainty* about $\hat{\mu}_{t,a}$.
- Thus, α directly controls the trade-off between maximizing expected reward and maximizing information gain.

Experimental Results

- All algorithms were run on our earlier slot machine example
  - $K = 3, T = 200, p_1 = \frac{1}{16}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$
  - $Y_t \sim \text{Bernoulli}(p_{a_t})$



Wrap Up

- Online A/B testing is very useful to quickly learn which actions are optimal without too much experimentation
- The key idea is to tradeoff learning about the quality of each action with earning as much as possible, and we use the Multi-Arm Bandit framework to do this
- Regular MAB should be applied when users cannot be distinguished (or are not allowed to be), e.g., website layout
- Simple algorithms tend to work very well, and naturally incorporate exploration-exploitation tradeoff directly
- Next Time: Contextual MAB problem which can applied we can choose action based on the user's features, e.g., email campaign or web advertising