w241: Experiments and Causality

Applications of Experimental Design

David Reiley, David Broockman, D. Alex Hughes
UC Berkeley, School of Information
Updated: 2021-07-09

Introduction:

Applications of Experimental Design

Week 13 - Applications of Experimental Design

DATASCI W241: Experiments and Causality

Quiz 1

Week 13 - Applications of Experimental Design

DATASCI W241: Experiments and Causality

Oversampling a Subpopulation

Introduction

Example: MN Dept of Revenue tax-compliance experiment

- Anticipated heterogeneous treatment effects
  - People with high income have more incentive to hide income
  - People with more opportunity to hide income (self-employed) more likely to under-report
- Without oversampling, treatment group of 1,500 taxpayers contained *only 7* high-income, high-opportunity earners
  - This is the population we are most interested in
  - Oversampling increased this to 80 people

When to Oversample

When we have a particularly interesting subpopulation

- Some examples:
  - High-opportunity taxpayers
  - Girls in computer-programming course
  - Minority legislators in study about response to minority constituents

When important subpopulation represents a small fraction of overall population

When important subpopulation has relatively high variance of Y

- Standard deviation is many times the mean
- Large samples can overcome huge variance

Adaptive Sampling

Adaptive Sampling: Adjust experiment based on responses

- For example:
  - If an A/B test reveals one sub-population has high variance of purchases, then
  - Oversample this high-variance group

Summary

If a group is either **very small** or has **high variance**, oversampling is prudent

Reading Assignment

Next we will discuss how we measure outcomes.

Read Section 12.3 on catalog mailing experiment of *Simester et al. (2009)*

Quiz 2

How Broadly Should We Define the Outcome?

Simester, *et al.* (2009) Experiment

Catalog mailing experiment showing how long-run effects differ from short-run effects

- Sending extra 5 catalogs in 8-month experiment produced positive return on investment
- Measured effect smaller with more broadly-defined outcome
  - Majority of effect from accelerating purchases forward in time
  - Negative treatment effect on purchases for 8 months following experiment
  - Negative treatment effect for website purchases for same retailer

Comparison 1:

Lewis and Reiley Advertising Experiment

Outcome defined more broadly

- Online ads had positive treatment effect on online sales, as well as brick-and-mortar stores
- Study included examination of purchases for 10 weeks after conclusion of 2-week ad campaign
  - No evidence of acceleration of purchases

Lesson: pay attention to outcomes of actual concern (in this case total sales)

Comparison 2:

Employee Reciprocity Study: Labor Economics

Lab experiments with two people in job market scenario

- Employer, employee
- Employer: decides how much to pay employee
- Employee: decides how hard to work for employer
  - Effort level: 0 - 10
  - Cash payout to employee subtracts effort from what employer paid

Game theory predictions:

- Employee has received wage; no incentive to engage in costly effort
- Employer has no incentive to pay more than minimum wage

Comparison 2:

Employee Reciprocity Study: Labor Economics

Actual results:

- Most employers set wages well above minimum
- Employees who receive higher wage offers tend to work harder for employer than employees who receive low wage offers
  - We know this is treatment effect of offer received
  - Evidence of reciprocity effect between workers and wage-setting employees

Comparison 3:

Gneezy and List (2006) field experiment

- Wanted to see what would happen when people are being paid actual money for actual work, rather than in a laboratory setting
- Students hired to solicit charitable donations
- Half received higher-than-advertised wage
- In first hour, higher-paid students brought in more donations (positive reciprocity treatment effect)
- A few hours (or days) later, treatment effect diminished

Lesson: We may need to redefine the outcome to be more broad

- Redefining outcome more broadly here lead to different conclusion

Conclusion

Long-run effects are difficult to measure in laboratory experiments

Treatment effects may wear off after longer period of time

Reading Assignment

Read *Field Experiments* textbook Section 12.4

- Bertrand & Mullainathan Audit Study
- **Audit Study:** Field experiment where research doesn't necessarily follow through with a completed transaction

Quiz 3

Audit Studies

Recap:

Bertrand and Mullhainathan job discrimination study

- Drew from identical resumes with white-sounding names vs. black-sounding names
- Showed clear evidence of discrimination against black-sounding candidate names
- Proves causal effect of race

This is an audit study because we didn't perform the call-back interviews based on responses from employers

Bertrand and Mullainathan Audit Study:

Advantages

- Drawing from variety of names and variety of high/low quality resumes $\longrightarrow$ more generalizable treatment effect
- Sending four resumes (black-sounding, white-sounding, high- and low- quality) per firm to identify some within-employer effects
  - Which candidates were called back?
  - What did distribution of discrimination look like?
- Is double-blind
  - If we had done face-to-face interviews it would have required actors to go to interviews, could no longer be double blind

Aside: Car Dealer Audit Study

Scenario: black and white consumers visited car dealers

- Followed script to counter with lower price than price offered by dealer
- Result: Black consumers ended up with a higher negotiated price
- Problems:
  - Actors behaved differently from one another
  - Actors were small in number
  - Treatment effect could erroneously be ascribed if one white actor had exceptional negotiating skills

Bertrand and Mullainathan Audit Study:

Advantages

- Drawing from variety of names and variety of high/low quality resumes $\longrightarrow$ more generalizable treatment effect
- Sending four resumes (black-sounding, white-sounding, high- and low- quality) per firm to identify some within-employer effects
  - Which candidates were called back?
  - What did distribution of discrimination look like?
- Is double-blind
  - If we had done face-to-face interviews it would have required actors to go to interviews, could no longer be double blind
- Used wide variety of resumes (inexpensive way to expand the study)
- Used factorial design
  - 2x2 design (two races, two quality levels)
  - More discrimination for high-quality or low-quality resumes?

Next: Schneider (2012)Audit Study

Does expectation of repeat business cause an auto mechanic to behave differently?

Example

Schneider (2012) Audit Study

Schneider (2012) Audit Study

Overview

- Investigated effect of expected repeat business on behavior of auto mechanics
- Hypothesis: a firm is likely to work harder when expecting repeat business or referrals
  - Doing high-quality work for you is important because of the expectation of generation of new business
- Schneider broke his own car in ways that should be repaired for safety during a road trip, then visited mechanics for inspection and estimated cost of repairs
  - Loose battery cable, low coolant, missing tail light, etc.

Schneider (2012) Audit Study

Details

- Went to 40 repair shops and assigned each to one of two treatments:
  - Low reputation: "I'm moving from New Haven to Chicago in two weeks."
  - High reputation: "I'm taking a vacation to Montreal in two weeks."
- Destinations are similar distances from origination city
- Only difference in treatment is the expectation that he will vs. will not be living in New Haven in the future
- Requested estimate for repairs
- Never intended to purchase repairs, which is what makes this an audit study

Schneider (2012) Audit Study

Results

- Charged more than twice as much for estimate in low-reputation treatment vs. high-reputation treatment
- No significant difference in diagnosing problems between treatments

Example:

Bandiera *et al.* Personnel Economics

Example: Bandiera *et al.* Experiment

Overview

Personnel Economics: Study of employee behavior

- British farm employed Eastern Europeans to pick fruit in Summer 2005
- Initial employee wages tied to *relative* performance
  - Farm wanted to implement productivity incentives
  - Farm manager reluctant to set piece rate (pay per unit picked) because picking conditions vary considerably
  - Payments relative to quantities picked by others
- Researchers wanted to try *absolute* performance payments
  - Fixed amount per piece picked
  - Implementation half-way through Summer
  - Deliberate intervention designed to measure causal effects
  - Before/After design constraint related to concern regarding spill-over effects from employees talking to one another

Example: Bandiera *et al.* Experiment

Results

- Productivity increased substantially from first to second half of summer
- Productivity per worker increased under *absolute* vs *relative* payment
- Variance of output per worker much higher

Proposed explanation:

- High-ability workers were holding back under *relative* payment
  - Lived and socialized almost exclusively with co-workers
  - Had fear of co-workers opinions
  - Worried that out-picking others lowered compensation for all

Evidence that relative performance scheme can encourage people to work less

Example: Bandiera *et al.* Experiment

Evaluation

Study was not randomized

Study included three comparisons:

- Late part of Summer vs. early part: Sharp increase in productivity
- That year vs. previous year (same time frame): *Absolute* payment outperformed *relative*, had higher variance
- Pickers of strawberries vs. raspberries
  - Raspberries on tall bushes with briers limited visibility
  - Above results held for pickers of strawberries but not raspberries
  - Raspberries functioned as control

Example:

Response to Incentives

Charness and Gneezy (2009)

Overview

Are people more likely to excercise if they are paid to go to the gym?

- Charness and Gneezy (2009) worked with UCSD student gymnasium
- Introduced three conditions:
  - Control
  - Low: $25 for one visit in a week
  - High: $100 for 8 additional visits in next 4 weeks
- Collected data on gym attendance for 7 weeks post-experiment
- Collected data on gym attendance pre-experiment

Charness and Gneezy (2009)

Results

- People in high-incentive condition developed exercise habit that extended beyond experiment
- Heterogeneous treatment effect
  - Highest effect in participants who did not previously exercise at gym at all
  - Low effect in participants who were already going to gym

Charness and Gneezy (2009)

Limitations

- Examination of carry-over effect to subsequent semesters was not possible due to inaccessibility of data
- Attempt to replicate similar results for the long-term didn't work
  - Places limit on generalizability regarding habit formation

Additional Incentive Examples

"Can we pay people to..."

- Lose weight?
  - From behavioral economics perspective, what constitutes sufficient incentive?
- Quit smoking?
- Take blood pressure medication consistently?
  - From behavioral economics perspective, what constitutes sufficient incentive?
- Get better grades?
  - Fryer's attempt to pay students to get better scores was ineffective

Incentives, Unintended Consequences

Gneezy and Rustichini (2000) experiment with Israeli daycare centers

- Imposed fine for late pick-ups
  - Number of late pick-ups increased
- Added 10 additional centers to experiment
  - Six centers had fines imposed for late pick-ups
  - Four centers functioned as controls
- Number of late pick-ups *increased*

Behavioral economics explanation

- Prior to fine, parents were loath to impose on daycare workers
  - Fine was interpreted as a price
  - Fine was insufficient to deter unwanted behavior

Summary: Experimental Design

Strive for originality and creativity

Think more broadly

Consider ideal experiment to address a research question

Draw from other experiments

- Audit studies
- Factorial design
- Double blindness
- Variety in treatment to create more generalizability
- Subpopulation oversampling

Case Study:

Study on LGBT Canvassing by David Broockman

Case Study:

Results of LaCour and Green (2014) on Gay-Marriage Canvassing

Broockman Smells a Rat

Detective Work:

How Can We Tell When Data Is Fake?

This Time For Real:

Transgender Canvassing in Miami

Final Perspectives:

LGBT Canvassing and Research Integrity

Further Reading:

Broockman's LGBT Canvassing Research

Further Reading:

Broockman's LGBT Canvassing Research

If you are interested in reading more about this topic, we recommend two articles:

Jesse Singal, *New York Magazine*, The Case of the Amazing Gay-Marriate Data: How a Graduate Student Reluctantly Uncovered a Hughe Scientific Fraud

John Bohannon, *Science*, For Real This Time: Talking to people about gay and transgender issues can change their prejudices