w241: Experiments and Causality

Problems and Diagnostics

David Reiley, David Broockman, D. Alex Hughes
UC Berkeley, School of Information
Updated: 2021-06-30

Problems and Diagnostics Introduction

Common Implementation Problems

Review from previous weeks

Compliance Problems

- Incomplete control over treatment delivery
  - Solutions: Intent to treat or placebo design
- Accidental delivery to control group

New Implementation Problems

- Randomization system doesn't work as intended
- Treatment not delivered or not received
  - Eg. letters not read or emails not opened
- Unintended effects of treatment

Hawthorne Effect

The Effect of change or novelty

Social experiment conducted at the Hawthorne factory

- Effects on worker productivity
  - Brightness of lights
  - Temperature in factory
- Every variable **seemed** to affect productivity
  - Effects were only **Temporary**

Websites

- Changes to a site cause users to act differently
- User activity often normalizes after adjusting to change
  - Eg. Change in purchase flow might cause **temporary** conversion increase

Demand Effects

Subjects who are aware of being studied often give different answers

- Subjects may give answers they think the researcher wants to hear
- Treatment can alert subjects to what researchers are looking for
- Example: Wearing a Barack Obama sticker may change how people respond to the question "Who do you plan on voting for?"
- Example: study to determine price for new product
  - Subjects might lie if they feel they can affect the price of the new product
- Researchers strive to conceal the connection between survey and treatment

Blind Trials

Single-blind Trial: Subject doesn't know

Double-blind trial: Experimenter doesn't know

Triple-blind trial: Analyst doesn't know group assignments

- Eliminates fishing expedition problem
- Biased analysis may cause study to be halted prematurely

Unexpected Implementation Results

Abstract concepts may result in missed real-world contexts

- **Example:** Restorative justice meetings
  - Attempt at reconciliation between criminals and victims
  - Criminals didn't show up, further damaging victims
  - Pilot study could have prevented problem
- **Example:** Field workers refuse to randomize or think they are helping
  - Canvassers who knock on every door unintentionally ruin control group
- **Example:** TiVo's effect on people remembering ads
  - People wouldn't accept the free TiVo device, treatment group was smaller than expected
  - Started giving TiVo's to people in control group as well

Detecting Errors

Detecting Errors: Overview

Pilot studies: Discover unanticipated problems

Placebo Tests: No effect desired

Manipulation checks: Was treatment successfully delivered?

Pilot Studies

Always run a pilot study, no excuses!

- Unanticipated reactions
  - **Example:** Fake emails to state legislators
  - Intended to study responses based on geographic origin of messages
  - Spillover between legislators
- Flawed power calculations
  - Likelihood of effects of different sizes
  - Baseline
  - Response rates
  - Variance of y
- Training Staff for correct implementation
- Determining if systems really work
- New ideas, potential for improvement

Yoga: Detecting Errors

Review: Detecting Errors

Placebo Tests

Placebo Tests

If experiment/experimentation system works, there should be **no difference** in a variable

- Difference in variable suggests a problem with the sytsem
- **A/A test:** Tests a treatment against itself to detect any difference in outcome
- **Traditional placebo test:** Checks other outcomes that treatment shouldn't affect

A/A Placebo Test

Example: Website optimizer code

In-house system randomly directed to page A or B when page A loaded

- Control: page A loaded
- Treatment: User redirected to page B
- Measured outcome: purchases on page
- Issues:
  - Load time was longer for treatment group
  - Treatment code didn't run on all web browsers. Some users directed to treatment page, but when it didn't load their purchase assigned to control group
  - Unrealized manipulations, Eg. delay loading web page
  - *Page A consistently bested B despite being identical*
  - Redirect delay and issues made the treatment perform better, not the content

Placebo Test

- Great tool for questioning observational studies
  - To prove bias, look for differences that shouldn't exist if assumption is right
- **Example:** Krueger (1993)
- Research question: Does using computers cause people to earn more?
- Workers who use computers earn 15-20% more
- Ideal experiment: Computers increase productivity, warranting higher wages
- DiNardo and Pischke saw potential bias
  - Found workers who use pencils, pens, and phones also earn more
  - Workers who use screwdrivers earn less
  - Wages depend on *types of jobs that use these tools*, not the tools themselves

Another Placebo Test Example

Gerber and Green Placebo Treatment: blood drive non-compliance

- Can't compare compliers to everyone in control group
- Compliers tend to be systematically different
- Compare entire treatment group to entire control group OR treatment compliers to control compliers

Social Influence Example

Do people influence their friends?

- Many studies ask for a person's weight and weight of friend
- Social network models find that being fatter causes friends to get fatter (causal claim)
  - Likely **not** causal relationship
- What can be controlled for?
- Repeat same procedure where there cannot be a causal affect
  - Same model finds being tall causes friends to get taller, even though this is impossible
  - Social network models struggle to recover causal effects

Placebo tests can find differences where, if your procedure works correctly, there should be no difference

Manipulation Checks and Covariate Balance

Manipulation Checks

- In placebo tests we wanted to find no difference. If we did, there might be an issue with our experiment
- Manipulation checks are the opposite
- If experiment worked, there **should** be differences in this variable
- **Example:** Broockman and Bulter (2014)
  - Treatment: Send letters from legislators
  - Control group: received no letters
  - Did constituents receive letters? Did they learn their legislator's positions?
  - Asked people if they recalled recently receiving a letter from legislator
  - People in treatment group more likely to report having received letter, understand legilator's position
  - Can't use this information in estimation

Manipulation Checks (Cont'd)

- **Example:** Sending postcards to voters causes 10% increase in understanding candidate's position
  - Can't use information in same way as one-sided non-compliance
  - Sanity check, making sure treatment had expected effect
- Particularly important for proving null effect
  - Otherwise it may be believed that your treatment simply wasn't received
- Returning to Broockman and Butler example
  - If no effect, skeptics may believe letters not opened
  - Information not assigned directly to people, it is assigned to their mailbox.

Covariate Balance Checks

- Check for problems in implementation of experiments
- Experiments guarantee balance in observable and unobservable characteristics
- Check balance on observable covariates to ensure random assignment was done correctly
- Example of balance not holding: users in treatment group are more active internet users
  - Might give impression of not conducting clean random assignment
- Especially important when randomization scheme is complex
  - Blocking, clustering, different probabilities
  - Complex systems between researchers and subjects
- See regression unit for more on this

Example of Failing Covariate Balance

- **Example:** Lewis and Reiley's first dataset
- Strange patterns: negative treatment effect
- Large imbalance on pretreatment sales
  - Control group bought more in preperiod than treatment group
  - Reason: vendor truncated data by number of sales
  - Treatment group sales data lost
- This is an example of a problem that would never be expected, but there are all kinds of small things like this that could go wrong
- Really important to check covariable balance as a sanity check

Summary

1. Conduct a pilot study
2. Manipulation check to measure delivery of treatment
3. Placebo test
4. Check for covariate balance

Advocating Experimentation

Advocating Experimentation

1. Increase perceived benefits
2. Decrease perceived costs

Increase Perceived Benefits

Stimulate curiosity

- Intellectual interest
- Get people excited

Vivid examples of current data leading to bad decisions

- Conduct placebo test and show practices failing
- Present potential conclusions experiment could produce and ways this could chance practices
- Tell story of why causal inference from observational data might be wrong
  - Eg. Playing outside improves children's eyesight
  - How did people receive treatment?
  - What's different about those people? What are the reasons?

Increase Perceived Benefits (cont'd)

"Investment in information for future decisions"

- Eg. Firm worries that holding back advertising for sake of experimenting will cost them money
- Short-term costs pay off in future
- Down payment

Build rapport

- Personal connections lead to willingness to run experiments

Do small studies as proofs of concept

- Example: Broockman and Butler (2014)
  - One congressman signing on for small study led to results that garnered interest for larger study
- Helps secure greater cooperation

Decrease Perceived Costs

Administration, unfairness, giving up potential gains from treatment (eg. ads)

Delay for some units

- Randomize order of mailings
- Campaign contributions
  - Spread out donations to allow experimentation

Limited resources

- Example: Charity gives 5,000 bed nets to people in Africa
  - Witholding bed nets seems morally questionable
  - Expanding population creates control group without witholding bed nets

Experimentation as investment in information

The Long View

Building Knowledge Over Time

- Try pilot studies
- System should produce covariate balance
- Parameters should look as expected
  - Eg. 80/20 treatment/control split
  - Example: Internet explorer skewed data
- Understand parameters for power analysis
- When creating a system, studies should be useful

Pooling Results

Pooling Results

Pooling Results

Pooling Results

Pooling Results

Pooling Results

Pooling Results

Pooling Results

- Precision-weighted average:

$1/SE^2$

- Double the standard error gets $1/4$ the weight
- 1/4 sample size means $1/4$ the information
- If variance is the same, weight by number of subjects
- Standard error used to represent adjustment of overall view of treatment
- Meta-analysis summarizes data

Yoga: Combining or Separating

Review: Combining or Separating