w241: Experiments and Causality

Unit 1

David Reiley, David Broockman, D. Alex Hughes
UC Berkeley, School of Information
Updated: 2021-08-13

Importance of Experimentation

- Large scale datasets are nice... but the majority of this data is not useful for assessing causal questions: *Does this cause that?*
- More data is not a solution for causal questions:
  - More of the wrong type of data is not better than less of the wrong type of data.
  - **Example**: randomized controlled trials ("RCTs")
  - **Treatment Group**: A group of subjects provided a new drug to determine efficacy -- but compared to whom?
  - **Control Group**: A group of subjects provided *no* new drug -- compared to the treatment group.
- Results from a well-constructed randomized experiment -- even if small -- contains information that even the largest, most dynamic observational data cannot.

Hormone Replacement Therapy (HRT)

Reading Assignment

Please read this article about *Hormone Replacement Therapy*

Think about the following questions as you read

- What is the difference in *data generation* strategy between the two studies?
    - **Nurses Health Study** *key finding*: HRT reduces heart attacks among postmenopausal women
    - **Women's Health Initiative** *key finding* HRT increases heart disease (as well as strokes and breast cancer).

The two results are not compatible

- Which of these results do be *believe* and why?
- HRT either increases, decreases, or doesn't change risk of heart attacks. How then can the Nurses' Health Study and Womens' Health Initiative disagree?

WHI results are more credible. Why?

Nurses' Health Study

- Epidemiological, (i.e. observational)
- Large sample of women
- Statistically significant correlations
- "*Women who (choose to) take HRT are less likely to have heart attacks.*"

Women's Health Initiative

- Designed experiments
- Smaller sample size (vis-a-vis Nurses' Health Study)
- "*Women who (are randomly assigned to) take HRT are more likely to have heart attacks.*"

Unmeasured confound

- Nurses' Health Study: Those who *choose* to take HRT have different underlying health conditions and motivations than those who *choose not* to take HRT.
- Womens' Health Initiative: Those who *are assigned* to undergo HRT are statistically indistinguishable form those who *are assigned* **to not** undergo HRT.

The oat bran example

Unmeasured confounders

- Potential invisible or unmeasurable factors that affect results
- These factors can be avoided by conducting the experiment with identical populations

Topic overview for the week

- Observation vs. intervention
- What experiments can tell us
- Kinds of experiments in the "natural" and "social" sciences.
- *Example*: "Magic on the Internet" auction experiments
- *Reading Assignment*: Read *Field Experiments* (Gerber and Green, 2012) Chapter 1.

Reading: Key Points

- Causal questions are crucial in a varietey of areas
  - Business
  - Public policy
  - Individual decisionmaking
- Decisions should be concerned with counterfactuals
  - We observe the state of the world if we have done `X`. What would we observed of we had done `not X`? Or, if we had done `Y`?
- This causal inference is difficult because we cannot observe *both* of these states of the world.
- Arguments based on intuition and anecdotes provide no solid ground for disputation and resolution. They usually end in stalemates.
  - Should the government extend unemployment benefits if there is as pandemic that causes a recession? Perhaps this could be settled with an experiment.
  - Should our organization continue to purchase display ads?

Reading: Key Points, cont'd

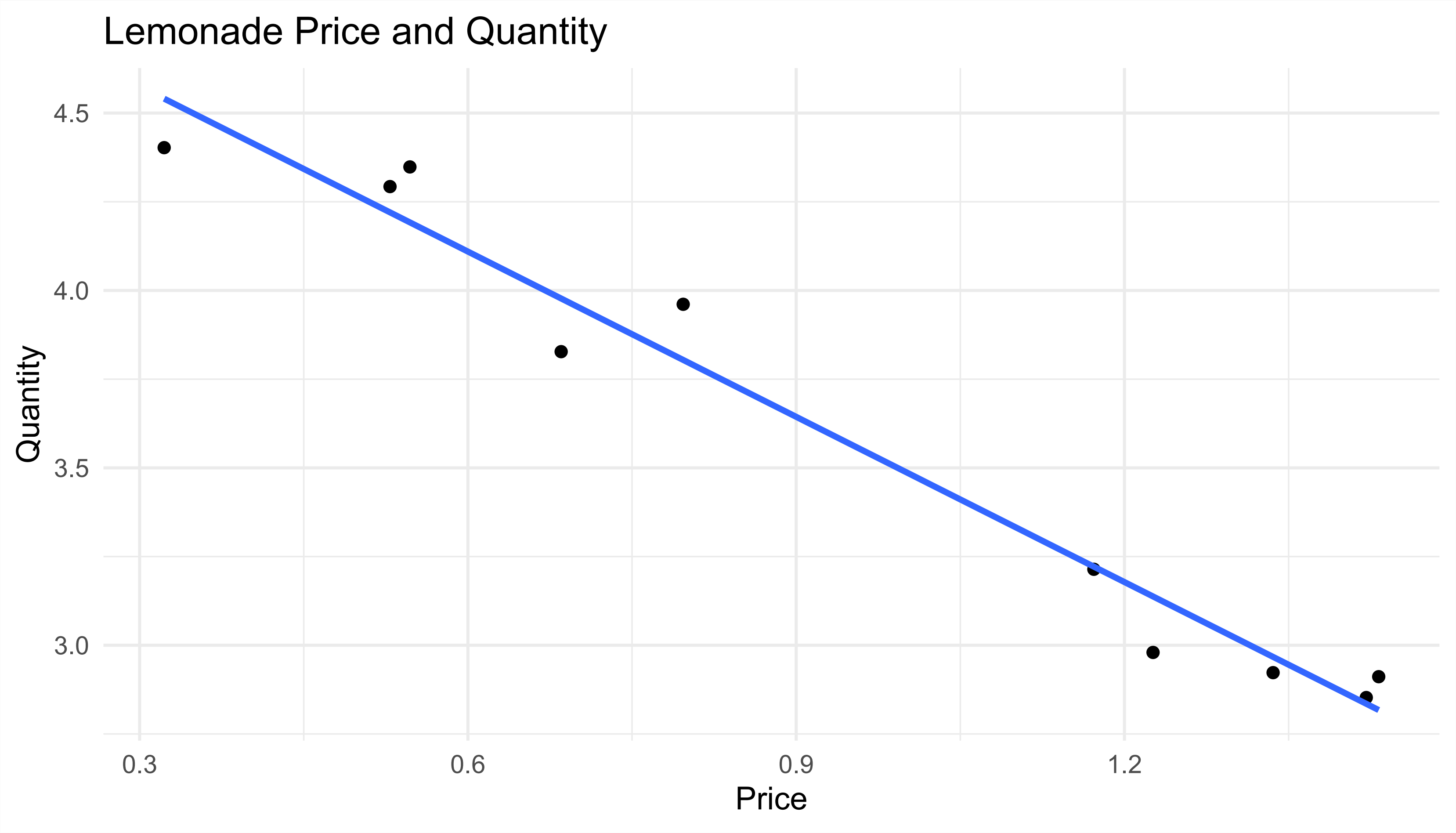
- Causal questions are settled with experiments in a way that avoid stalemates.
- Causal questions are harder to get correct in social and data sciences than in physical science, because of underlying heterogeneity
  - All electrons are the same.
  - Not all humans are the same.
  - In fact, perhaps all humans are distinct.
- One should be skeptical of causal inferences based on observational (i.e. non-experimental) data because of the possibility of unobserved heterogeneity.

Examples of causal questions

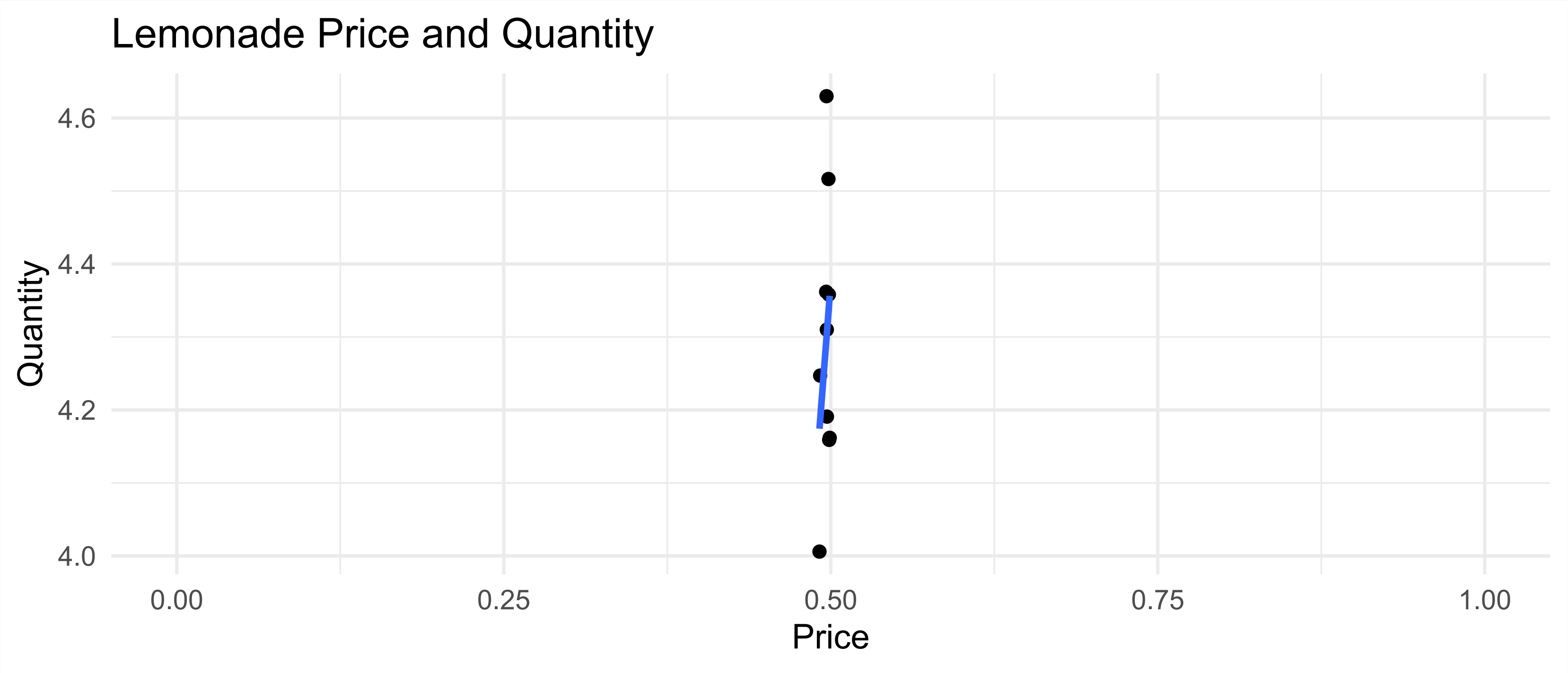
- Does boiling drinking water prevent people from contracting cholera?
- Do mandatory seat belt laws reduce traffic fatalities?
- Does TV advertising increase purchases? If so, by how much?
- Does Chanel No. 5 have a downward-sloping demand curve?
- Does having children give people a more satisfying life?

Experiment: An intervention that creates variation in order to teach us causal questions.

Lemonade Stand Exmple, Part 1



Lemonade Stand Example, Part 2



- We cannot learn anything about the demand curve for lemonade if the price remains the same!
- To learn about the demand curve, it is necessary for there to be *some* variation in prices.
- When we conduct an experiment, we **deliberately create variation, but in a way that is not related to other features.**

Intervention and Randomization

- Experiments do *not* have to involve randomization
- Intervention is the key element of an experiment.
- Randomization can sometimes be difficult to implement
  - Randomly assign price to every customer?
  - Might take too much time
  - Might irritate customers
- But we learn more -- and have better guarantees -- with deliberate variation than without it.

Why is randomization useful?

Consider the lemonade stand example

- Charging a different price every day
- Charging more on a very hod day, and more lemonade sold
    - Temperature is a confounding factor: like price, termperature also affects lemonade sales
    - Might incorrectly assume the price increase (rather than the temperature) caused the increase in sales.
- One wouldn't learn the true effects of a price change if any factor that also influences sales is correlated with price change.

Why is randomization useful? (cont'd)

- We can control for confounds by repeating the experiment many times
- The treatment, *price*, should be independent of every other feature that might influence lemonade sales: temperature, sun, holidays

Randomization produces the guarantee of independence.


Lemonade Price and Quantity
Price Randomly Assigned
3.0
2.5
2.0
1.5
Quantity
0.4
0.8
1.2
Price

Pitfalls of naturally occurring variation

Possible pitfalls in making causal inferences from observational data

- Nurses' Health Study on HRT:
  - *Possible story*: Women who receive hormones generally tended to care more about their health, and/or follow doctors' recommendations.
  - Lower incedence of heart attacks possibly caused by *these* factors, ratehr than HRT.
- Lemonade stand data
  - The higher the price, the more lemonade sold
  - Turns out, the kid is a ruthless capitalist! They charged more when the temperature was higher.
  - Temperatuer is a confounding factor in this case.

What is a natural experiment?

- Green and Gerber are not fond of using the term, *natural experiment*.
  - A *natural experiment* is naturally occurring data -- i.e. the researcher did not themselves design the data -- that a researcher reasons has the same properties as a true, designed and controlled experiment.

An analogy: Herschel's Gardern

- William Herschel: astronomer who discovered Uranus
- Experiments on stars are not possible -- but one can still learn something by studying various stars at different stages of development
- **Herschel's Garden**: the idea of viewing the night sky as a garden that features the same types of *plants* at different stages of development
  - Perhaps this idea can be applied into the social data sciences
  - Natural experiments in the social data sciences are likely to be less clear, due to un-observable heterogeneity

Observational studies

Best

- Data from *naturally occurring experiments* (i.e. situations where variation was produced by something like random assignment)
    - Example: charter school lottery deciding applicant acceptance
    - Data about applicants possibly analyzed to study causal effects of charte school education
    - Same characteristics in both groups, those who attened the charger school and those who did not
    - Example: Vietnam draft lottery that causd some people to get more college education
    - Group that got more college education is otherwise identical to group that did not

Observational studies, cont'd

Less ideal

- Decent reasons to believe that those who received an intervention are otherwise identical to those who did not
    - Example: Want to know the effect of plowing snow on business in a college town. Can't randomly assign plowing, but the plows clear different streets on alternating days. The days of the week might not be otherwise identical in terms of foot traffic.
    - Example: Does raising the minimum wage *actually* cause a decrease in employment? We cannot randomly set the minimum wage, but Seattle raised its minimum wage to $15, while Tacoma did not.

Observational studies, cont'd

Non-credible

- No reason to believe that those who received an intervention are otherwise identical to those who did not
    - Example: Charging more for lemonade on hot days than cool days; women receiving HRT care more about their health.

Magic on the internet

Magic on the internet

- Using field experiments to test equivalence between auction formats
- Observational data holds no guarantees that auction format is independent of card value.

Early Online Magic Auctions

- People trading *Magic* cards via Usenet newsgroup
- New transaction mechanism, before eBay
  - Currency: Cash
  - Question: How should one value a *Magic* card?
- Beginning of *Magic* auctions, with bids updated via email
- Cards sold -- auction closed -- if there were no price increasese in three days.

Why conduct this experiment?

- Auction *Magic* cards in a controlled experiment to learn about *auction theory*.
  - Note that the context is magic cards, but the larger question is broader than this particular good; rather it is about auctions as a concept.
- Why is auction theory important?
- Why study revenue equivalence between auction formats?
- Which format generates higher selling prices? Ascending-bid format, or sealed-bit format?
- There isn't actually a prediction about which should generate higher revenues, in expectation.

Lab researchc on auctions

- Students bid on a fictitious good -- a good that they don't particularly want, and are not particularly invested in
- So experimenters, assign different values to different subjects
  - Each subject knows their own value, but they do not know the values that others assign to the good
  - Goal is to maximize the difference between the price you pay, and the value that you ascribe to the good.
  - Importantly, there is *no* value in this system that is beyond the assigned value. Does this match reality?
- With this setup, researchers can examine ascending-bid and sealed-bid auctions
- Find interesting violations of theory
- These lab violations raised the question: are these inconsistencies with theoretical predictions an artifact of the labratory, or are these more general when people are acting within a *real* auction market?

Vickrey: Auction Formats

1. English auctions: ascending bids
2. Sealed-bid auctions: one-time highest bids
3. Dutch auctions: decending price clock
4. Vickrey's "second-price" auction: second-highest bid determines winner's price (like eBay proxy bidding)

Revenue Equivalence: Two Kinds

- Strategic equivalence: strong prediction
  - Dutch and first-price auctions are strategically equivalent because they have the same amount of information
  - English and second-price auctions are strategically equivalent under *private value* (like those assigned in the lab)
  - Thus, there is a dominant-strategy mechanism for trutheful revelation of valuations.
  - Regardless of others' strategies, one's own optimal stretegy is the same bid: bid my maximum willingness to pay
- General revenue equivalence: weaker prediction
  - Expected revenue of all four formats should be the ame if people are risk-neutral

Lab research: Violation to theory

Cox et al. (1982, 1983)

- First price auctions raise more revenue than Dutch uactions

Kagel et al (1987, 1993)

- With private values, subjects overbid in second-price auctions
- So, yielded higher revenues than English auctions

Why?

- Could this be risk aversion?

Field Experiments

Similar to observational studies

- Auctions for real good, by real people, who are really accustomed to bidding for these goods

Similar to lab experiments

- Deliberate, designed interventions into a *real* system, without actors being aware of the interventions.
- Two different auction formats with the same good and the same bidders.

Similar to real world

- Cannot control individuals valuations of the good
- Cannot control risk, time, or other preferences of the subjects

Background on Magic: The Gathering

- First sold in July 1993, with a first printing of 10 million cards
- To date, more than 20 billion cards printed
- Estimated 1995 wholesale revenues: greater than $100 million USD
- Cards are sold in a random assortment, which generates a large aftermarket
- Creates a real world market laboratory

The experiment

- Four pairs of auctions designed for within-card comparisons
- Auctioned sold the same card twice (e.g. in both a Dutch and first-price auctions
- Sealed-bid auction: one week to submit bids
- English auction:
  - Bid any time
  - Daily update on each card's high bid
  - Like the Usenet marketplace, three days without a new bid closes the auction

The experiment, cont'd

- Dutch auction: start at a high price, announce a decrease of 5% each day via e-mail; same bid increments as other auction formats
- To control for order effects -- perhaps there is only a market for one card, and so once it sells, the market has little demand for another -- each experiment was run twice
  - FD: First-price followed by Dutch
  - DF: Dutch followed by first-price
  - FD and DF sets were run at the same time

Result 1: Card-level data

- Violates Dutch/first-price revenue equivalence
- FD and FD experiments 173 matched pairs of cards
  - Dutch format: 122 yielded higher revenue
  - First-price format: 34 yielded higher revenue
- On average, cards yielded $0.32 more in Dutch auction (24% of total card value)
- Sign-rank test: price-per-card differences are highly significant
- No qualitative difference between FD and DF results; that is, order of treatment doesn't matter
- Opposite of the violation that had been observed in the lab.

Result 2: Bid-level data

- Treatment unit: what is an observation?
- Here: an individual bid, vs. a card the receives multiple bids
- Bid-level data weakly supports (note the loose language here) violation of strategic equivalence: Dutch and first price auctions should have generated the same bids; but some evidence this might not be true.
- Compare bids by the same bidder in two matched auctions
- Data censoring -- we don't see most people's revealed prices because there is only one bid that is made, which closes the auction
- Of 38 observations with bids observed in both Dutch and first-price auctions:
  - 30 favored the Dutch; mean difference $2.52
  - 4 favored first-price, mean difference -$0.50

Result 3: English auction

- Seems to produce slightly more revenue than second-price auction
- Card level data: 164 pairs observed
  - 1.8% more revenue than second-price auction (not statistically significant difference)
  - Cannot reject null hypothesis of revenue equivalence
- Bid-level data: 112 matched bid pairs
  - 75 cases had higher bids in English auction
  - 29 cases had higher bids in second-price
  - On average, English bids were 3% higher
- Estimates opposite to lab results, but with limited statistical power

Result 4: First/Dutch vs. English/Second

- See how auction revenues deviate from reference price ("cloister price") for each good
- Pool together Dutch/first-price data and English/second price data: total of 370 observations
- On average DF auctions raise 12% more revenue than ES auctions
- Lower difference for higher-priced cards; cards above $13 show no differences
- Results are consistent with prior lab research
- Could this be risk aversion? If so, then why is the effect smaller for higher-priced cards?

Conclusions

Revenue ranking

1. Dutch
2. First-price
3. English
4. Second-price

Field data vs. laboratory data

- Opposite violations of the FD and ES strategic violations from the lab results
- Same FD > ES effects as lab results

Questions

- What cause results to be different from those in the lab?
  - Real goods mechanism?
  - Cash payout mechanism?
  - Simultaneous vs. sequential mechanism?
  - Clock speed mechanism?
- New lab experiments (Katok and Kwasnica, 2003): Slower clock speeds lead to higher revenue in Dutch auctions -- people are impatient?

Lessons to Remember

Lessons to Remember

Importance of Interventions: Holding secondary factors constant while trying alternate methods

The Spirit of Experimentation

- Try new, uncertain things!
- Be willing to live with short-term loss; this can generate longer-term gains!

Questions

- What is *an* experiment?
- Was *this* an experiment?
- Would *Green and Gerber* be skeptical? If so, why?
- Do *you* believe the results of what you see?
- Are the results from *this* experiment application to *other* environments?

Intervention vs. Observation

Example: US Forest Service

- The US Forest Service auctioned off timber-logging rights
  - For some auctions, they used sealed-bid auctions
  - For other auctions, they used English ascending-bid auctions

Analysis

- One format generated higher revenue than the other
- But, this result could be confounded by timer-quality variations that were not measured in the original data
  - Unobserved secondary variables often affect outcomes
  - All (observed or unobserved) secondary variables must be equal for experimental comparisons to produce unbiased estimates

What to Remember

1. Causal Questions cannot be answered with intervention to induce variation in *input features*
2. If "interventions" are naturally occurring (i.e. observational), then systematic differences may exist between those units that *do* and *do not* receive the intervention
   - Random assignment solves this problem to produce interventions that are not correlated with any other feature
3. People make choices about what experience they receive. So, a *healthy dose* of skepticism is in order when causal claims are made without a well-designed experiment.