Lecture 11: Causality & Experiments, Part 1
Modeling Social Data, Spring 2017
Columbia University

April 14, 2017

Notes from bk2628

1 Prediction vs Causation

1.1 Prediction

Making a forecast without changing anything. For example: seeing your neighbor with an umbrella might predict rain

1.2 Causation

Make a change in the current scenario, i.e., the current state of the world and anticipate what will happen.

1.2.1 Reverse causal inference

Finding the cause of something that has already happened. For example: what caused my kid to get sick? Reverse causal inference is generally quite hard.

1.2.2 Forward causal inference

Forward causal inference is more "what happens if we do a certain thing?".

1.2.3 Example : Hospitalization on health

What's the effect of going to the hospital today on your health tomorrow? If we try to estimate the above model based on just observational data we may be missing out an unobserved common cause like the health of the person on the day of the hospital visit.

Observational estimates

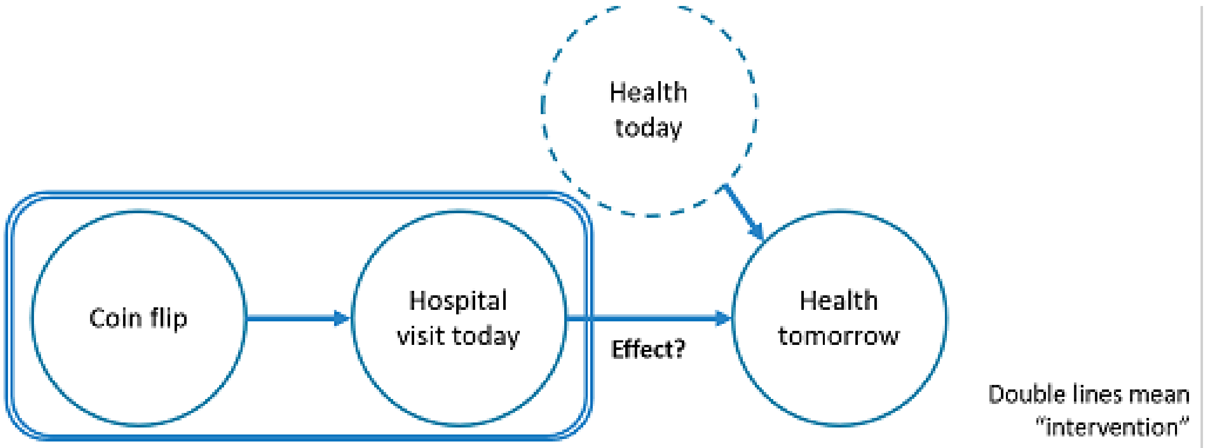
Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home
Observed difference in health tomorrow = (Sick and went to hospital) – (Healthy and stayed home)
Observed difference in health tomorrow = [(Sick and went to hospital) – (Sick if stayed home)] + [(Sick if stayed home) - (Healthy and stayed home)]
Causal effect = (Sick and went to hospital) – (Sick if stayed home)
Selection bias = (Sick if stayed home) - (Healthy and stayed home)
Observed difference in health tomorrow = Casual effect - Selection bias
There is a selection bias in this example because the people visiting the hospital are not random as today's health of the person affects the decision of visiting the hospital today.

2 Predictive Systems

Aim : predict the future activity for a user.

2.1 Search engine example

Say you wanted to buy a toy from amazon and so, you search for it on Google. Google would show you an ad for Amazon.
But the search results also have Amazon. This raises the following question:
The counterfactual question is "would have i still landed on Amazon.com even without the ad?"
Maybe or maybe not.
There can be hidden causes for the observed results, and ignoring such causes can lead to completely different conclusions.



Double lines mean "intervention"

In the above example, people may have anyway used Amazon.com to order the toy even without the ad. But our observations may suggest that the activity was because of the ad.

2.2 Simpson's Paradox

Selection bias can be so large that observational and causal estimates give opposite effects.
(example: going to hospitals makes you less healthy)
Other examples : Comparing an old algorithm with a new one.

Old Algorithm	New Algorithm
5%(50/1000)	5.4%(54/1000)

After dividing the group into low and high activity users
Low activity users

Old Algorithm	New Algorithm
2.5%(10/400)	2%(4/200)

High activity users

Old Algorithm	New Algorithm
6.6%(40/600)	6.2%(50/800)

New algorithm has a better success rate overall but when we divide the data, old algorithm seems to have a higher success rate.
Conclusion: we can add as many variables as we want to the model and different algorithms would perform different under different variables.

Going back to the hospital example:
How can we check if going to the hospital makes you less healthy?
One way is to "clone" the world which means just change the one thing that we want to measure while keeping everything else the same.
Guy goes to the hospital Vs Guy doesn't go to the hospital everything else remains the same.
This works well in theory but how do we actually clone the world?
Take the data set(sick or healthy) and use a coin flip to decide who goes into which version of the world. Since the decision is based on a coin flip, the expected result would be to get similar samples in both the worlds.
Basic identity of causal inference:
observed difference = causal effect - selection bias
Selection bias is zero since there in no difference, on average, between the two worlds.
observed difference = causal effect

Random sampling makes sure that health today has no effect on the health tomorrow. The only thing we are checking is "does going to the hospital today affects you health tomorrow?".

2.2.1 Experiments : Caveats/Limitations

- Randomizations is not always feasible or ethical.
- Experiments cost a lot of money and time.
- Anyone can flip a coin, but convincing parallel worlds are hard to simulate.
- It's inevitable that some people would deviate from their random assignments.

2.2.2 Experiment : Experimental drug trial

Two goals for this experiment:
1) Internal Validity: Could anything other than the treatment have produced results?
One cause could be that doctors gave the medicine to some special patients, thus breaking the randomization.
2) External Validity: Do the results of the experiment hold in a setting that we care about?
Would the medicine be as successful in the real world as it was in the clinical trial?
One example is birth control pills which may not be as effective in the real world as they are during the test on account of people forgetting to take them on same days.

3 How we conduct behavioral experiments?

3.1 Lab Experiment

3.1.1 Better internal validity

- Greatest procedural control: we can define the setting the experiment takes place.
- Can carefully curate situations

3.1.2 Less external validity

- Artificial context, simple tasks :
  The user may have behaved differently in a "natural" environment.
- Demand effects
- Homogeneous (WEIRD) subject pool
  WEIRD stands for Western, educated, and from industrialized, rich, and democratic.
- Time/scale limitations

3.2 Field Experiment

3.2.1 Better generalization

- Experiment findings apply to at least one real world setting

3.2.2 But:

- Less Control, more potential confounds
- Demand of experiment conflicts with goals of real organizations
- More effort to conduct and manage
- More room for error

4 Examples of Causal Inference

4.1 Natural Experiments:

Sometimes nature runs experiment for us, e.g.:

4.1.1 As-if random:

People are randomly exposed to water sources (Snow, 1854)

4.1.2 Instrumental variables:

An instrument independently shifts the distribution of a treatment. Example: A lottery influences military service (Angrist, 1990)

What can we try to infer? 1) Effect of lottery in making people join the military. 2) What would be somebody's future earnings if your name was called in the lottery?

Natural experiments: Caveats

- Good natural experiments are hard to find.
- They may have many (untested) assumptions.
- The treated population may not be the one of interest.

Notes from ch3216

1 Prediction Vs. Causation

Prediction: predict future activity for a user from their past data/activity
Causation: agency or efficacy that connects one process (the cause) with another process or state (the effect). But what exactly does causation mean? For example:

- What makes an email spam?
- How does the stock market works?
- Why is Apple, Harry Potter so successful?

But there are so many parts, reasons, aspect, possibilities that make one successful. It is almost impossible to figure it out in a scientific sense.

2 Why is it so hard?

At any point you have an event, imagine there are so many other things that could have happened. Just by looking at the data, it is not possible to find the root cause. Another more practical way to think about causality: What we can do now and make sense of it? How does one decision have impact on the future? E.g. Should I take this route? How does education actually have impact on your career development?

Simple Example: What's the effect on your health by going to a hospital?
Let start by looking at people who come to hospital and get treated, and see what's their health looks like tomorrow. Underlying question: *does hospital cause the change of people's health tomorrow?*
Think about: what kinds of people would go to the hospital? People who are less healthy.
Take into account: Health of the people today really determines whether or not they go the hospital.
What can we do? Assume sick people go to hospital, and healthy people do not go to the hospital.
If you just look at the data: going to the hospital, decrease your chance of being healthy.
In reality: people who go to the hospital are usually sick, and healthy people don't usually go to the hospital.

$$\text{Observed difference} = (\text{Sick and went to hospital} - \text{healthy and stayed home}) + (\text{sick if stayed home} - \text{healthy and stayed home})$$

Key take away: there is, in any kind of data, an inherent bias

$$\text{Observed difference} = \text{causal} - \text{bias}$$

3 Why do we care?

If you do not know the effect, you cannot make recommendation/decision. We really want to know: If you pick a person, and assume he take this action, what will be the outcome without him actually doing it. It is different compared to regression, because in regression get some data, have some outcome you already measured, you just say let me see how many patterns I can find.

Another example: Xbox user activity
In Regression only want to find out: *how much activity they will produce in the future?*
In Causality we care about: *What do we do to increase user activity?*
But this can be really complicated:

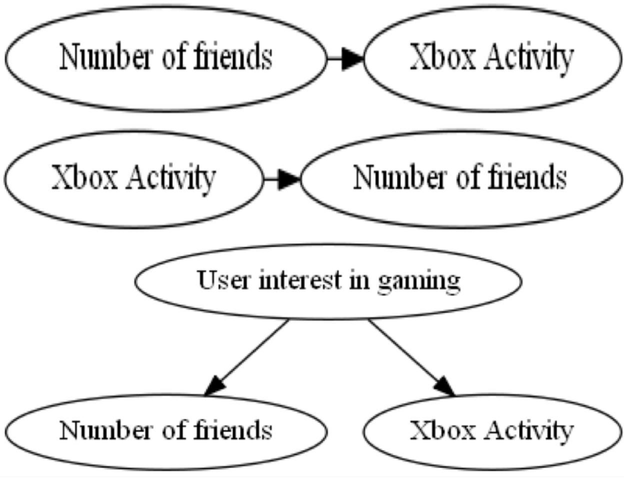
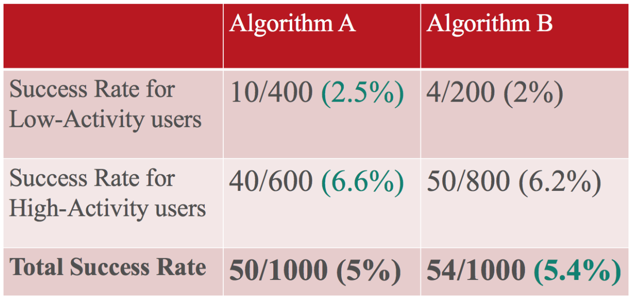


Figure 1: There can be many causes leading to different outcomes

4 Simpson's Paradox (Law?)

When you have any kind of data, if you don't condition on the right variables, not only you will get the wrong answer, you will get completely flipped answer.
Simplest example: google breast screening test went wrong.
One Other Example: recommendation engine — Two algorithms which one is better?
Key Question: how do you know, when you change one algorithm, it works better?
Let's look at the data. Random 1000 session for each algorithm. Measure Click Through Rate (CTR).



	Algorithm A	Algorithm B
Success Rate for Low-Activity users	10/400 (2.5%)	4/200 (2%)
Success Rate for High-Activity users	40/600 (6.6%)	50/800 (6.2%)
Total Success Rate	**50/1000 (5%)**	**54/1000 (5.4%)**

Figure 2: Is Algorithm A better?

Answer: As usual, may be, may be not.
A different control variable gives the same data a very different outcome

- Different kinds of user (low/high activity user)
- Time of day
- Sample size

Stratification — this thing we changes will have different effects on different groups, so we will break everything into different groups, and within each group we will do experiment, and look effects within each stratum. Picking which group we are going to do beforehand. You want to stratify a lot of things, Fixing different variables, getting exponential number of cases.

5 How do we solve Simpson's Paradox?

Let's go back to the hospital example:
What we wanna do: take the world, make a clone of it (parallel universe) and one case bring this guy to the hospital, and one case we don't. Everything else is the same, everything is identical except the change. We know nothing could affect the treatment outcome, other than the one thing that we changed. (Idea of counter-factual, but we can't observe it)
In reality: can't make copies of the universe. Instead, for each person we are interested, we will flip a coin, and tell them whether they go to the hospital or not going to the hospital. Treatment vs. control. Most important thing here, is that we are flipping the coin. (even if someone's healthy, we are going to send him to the hospital; some one's sick don't send them to hospital)
Key point: randomly send people to hospital or make them stay home. Reason: The reason they go to hospital is nothing but the result of coin flip, this produces a controlled situation. (similar to the idea of parallel universe)
Key Recipe: TRUE RANDOMNIZATION
When we are flipping the coin (creating world A, world B), we are taking away the relationship any reason that people might have chosen to be in the treatment, with all the other condition they had before.
MAKING SELECTION BIAS ZERO.

$$\text{CAUSAL EFFECT} = \text{OBSERVED DIFFERENCE}$$

A—B TESTING: (sometimes called split testing) a randomized experiment with two variants, A and B, which are the control and variation in the controlled experiment. It is much easier to measure causal effect. The bigger size the experiment you run, greater statistical power, more able to identify cause. However, in reality it is very hard to create an equal world 1 and world 2.
Many challenges and difficulties:

- **Experimenter effect:** if people know they are in an experiment, they will change the outcome.
- Difficult to create convincing world: when we actually flip a coin, people in two groups are the same.

6 Internal and External Validity Trade-off

Internal validity: experiment is RIGHT, anything but the coin flips we did change the outcome
e.g. ethical dilemma in the hospital
External Validity: validity to generalize, take this experiment, apply it to practice, does the result carry over? e.g. if we decide to sell this drug on market, will it be as effective as it were in the experiment? Maybe clinical trials, we rigorously monitor patients, but in practice does not adhere.
Internal: could anything other than the treatment have produced this outcome?
External: do the results of the experiment hold in settings we care about?
How social science experiments are done?

1. Lab Experiment: held at university, bring people in, make them do experiment, measure the result.
   (a) High degree of procedural control
   (b) Optimized for causal inference
   (c) Artificial environment
   (d) Simple tasks, demand effects
   (e) Homogeneous WEIRD* subject pools (Western, educated, industrialized, rich, democratic countries)
   (f) Time/scale limitations

2. Field Experiment:go to real world organization, having people flipping coins, doing experiments
   (a) High degree of procedural control
   (b) Optimized for causal inference
   (c) Fewer constraints on location
   (d) Longer periods of time
   (e) More samples of data
   (f) More effort

   You are choosing between complexity (more precise control) vs. realism (more generalizable)

7 Natural Experiments

What can you do if you don't have experiments? Not much, you can try a few strategy Even today we don't know if we can not do experiments. What is something you can exploit? Understand problem, understand data, understand what you can exploit.

1) find experiment in nature, as-if random

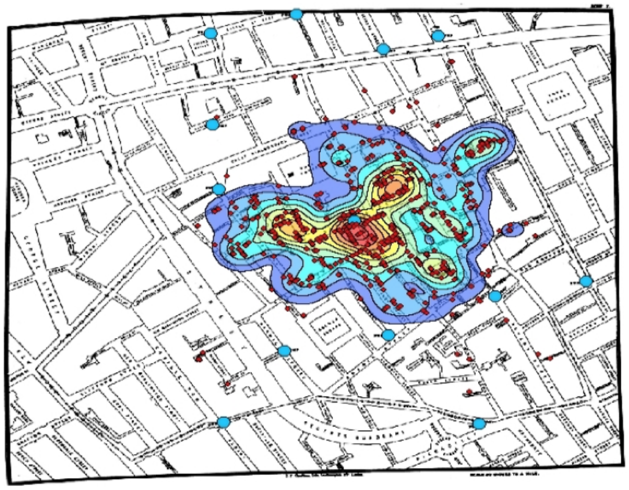


Figure 3: 1854: London was having a devastating cholera outbreak

Example: Jon Snow and his Maps. But if you think about it: that just shows people who live near the pond got sick, maybe people who live by pond got bad habits, bad neighborhood, air by the pond is bad.
2nd thing Jon Snow did: he showed there is as-if random. There are two companies, one getting water downstream (contaminated), one getting water upstream(uncontaminated), by just looking at some person who went to first/second company) Have sufficient evidence to convince yourself and others that this difference could have not happened without the corresponding changes (TREATMENT AND CONTROL)
Why was Snow's study so convincing?

- Choice of water company cannot cause cholera.
- Choice of water company was not related to people's neighborhood or its air quality.

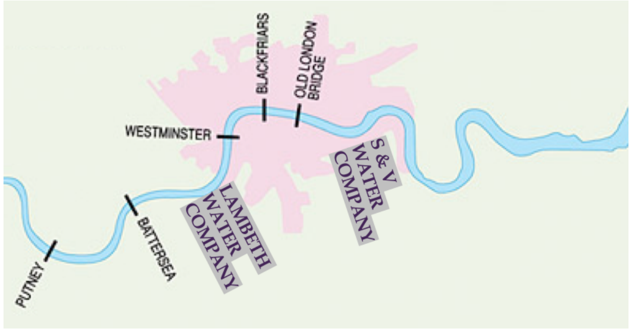


Figure 4: Two major water companies for London: one upstream and one downstream.

2) Instrumental variables

You don't change water (root cause) itself, you do something else that is instrumentally changing where people are having their water from. Change the distribution of the cause.
e.g. lottery system how military service affect future earnings When veterans come back, will they be a good earner? People can make all kinds of decisions, once we know the lottery is random, if that's true. "An important question in economics research is what determines earnings. Angrist (1990) evaluated the effects of military service on lifetime earnings. Using statistical methods developed in econometrics, Angrist capitalized on the approximate random assignment of the Vietnam War draft lottery, and used it as an instrumental variable associated with eligibility (or non-eligibility) for military service. Because many factors might predict whether someone serves in the military, the draft lottery frames a natural experiment whereby those drafted into the military can be compared against those not drafted because the two groups should not differ substantially prior to military service. Angrist found that the earnings of veterans were, on average, about 15 percent less than the earnings of non-veterans." (From WikiPedia)

3) Regression discontinuities:

Regression discontinuity design (RDD)is a quasi-experimental pretest-post test design that elicits the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is unfeasible

Notes from cx2187

Causality and Experiments

Prediction: Make a forecast, while leaving the world as it is (e.g. seeing someone with an umbrella might indicate that rain is likely) Causation: Anticipate what will happen when you make a change in the world (e.g. giving someone an umbrella doesn't cause it to rain)

We want to make **reverse causal inferences** by finding out what causes an observable change, such as "why did the stock market drop", or "why is this email spam". However, there are often many factors that go into an outcome. We can also look at **forward causal inferences** where we look how changing a variable can impact future outcomes, such as "how does education impact future earnings", or "what is the effect of advertising on sales".

Hospitals -¿ Better health? There may be confounding variables, e.g. people who visit hospitals tends to have worse health to begin with, but this is unobservable. In such a case, our observational estimate – e.g. the number of sick people who went to the hospital minus the number of healthy people who stay home – might seem to indicate that hospitals lead to worse health, which is not true. There are also alternative observational estimates we can measure, e.g. difference between those who opted into treatment, and those who didn't. This is often attributable to selection bias.

Observed difference = causal effect - selection bias, although the difficulty is in determining the extent of selection bias.

Example: How do predictive systems work? We attempt to predict future activity for a user based on their profile and past activity. We posit that future activity = f(number of friends, past logins), but need to turn this into **actionable insight**, i.e. how to increase user activity.

We should always ask the counterfactual, i.e. what would happen if an action were not performed. E.g. websites see a high click rate when they place ads on search engines, but they may achieve the same click rate even when they don't post sponsored content.

Simpson's Paradox: Given a large selection bias, observational and causual estimates might even lead to opposite results, e.g. that going to hospitals makes you less healthy

With Simpson's Paradox, it is possible that overall a new model may perform better on overall data, but worse in all subsets of a data. E.g. a new algorithm might successfully predict 54/1000 outcomes, and 50/1000 outcomes. However, if we looks at low activity users and high activity users, it's possible for the new algorithm to underperform both if our sample sizes for each subset are different.

Simpson's Paradox in Reddit: Average comment length falls over time, but if we plot this based on when a user joins Reddit, their comment length increases instead. There might be several reasons for this, e.g. more people were joining Reddit later, and tended to comment less even though their rate of comments would still increase over time. This brought down the overall average comment length.

The easiest way may be simply to change a variable and observe an outcome, e.g. an experiment or A/B testing. This helps us isolate a causual effect, and establish a counterfactual. Because we cannot create a copy of the real world, we instead use random assignment to create treatment and control groups, based on a randomized coin flip for each individual regardless of their original condition.

Random Assignment

If people know they are being measured, they might behave differently from their normal condition, which will influence the test result.

Two goals for experiments

Internal validity (care about confound)

Ex: Did doctors give the experimental drug to some especially sick patients hoping that it would save them?

External validity (care about results)

Ex: Would this medication be just as effective outside of a clinical trial when usage is less rigorously monitored?

Notes from jyl2164

1 Simpson's Paradox

Selection bias can be so large that observational and causal estimates give opposite effects (e.g., going to hospitals makes you less healthy)

1.1 Comparing old vs. new algorithm

Two algorithms, A (production) and B (new) are running on the system. From the system logs, data is collected for 1000 sessions for each algorithm. Measure CTR. Which algorithm is better?

Old Algorithm (A)	New Algorithm (B)
50/1000(5%)	54/1000(5.4%)

Frequent users of the Store tend to be different from new users. So lets look at CTR separately. The Simpson's paradox:

CTR	Old Algorithm (A)	New Algorithm (B)
CTR for Low-Activity users	10/400(2.5%)	4/200(2%)
CTR for High-Activity users	40/600(5%)	50/800(6.2%)
Total CTR	50/1000(5%)	54/1000(5.4%)

Is Algorithm A better? Answer (as usual): may be, may be not. Algorithm A could have been shown at different times than B. You could have also targeted it to people with high activity. There are also much more people in the new algorithm. Essentially, you can play around and find different algorithms that work well. There are many other hidden casual variations, like income level and stratification (we think that the thing we are changing will have different effects on different groups).

1.2 Example: Simpson's paradox in Reddit

If you look at the average number of words on Reddit and condition on different years you will notice that it is decreasing over time. That is, the average comment length is decreasing over time, which could be a worrying sign if you are Reddit. This would mean that Reddit would need to find ways to make people write more. However, if you condition on when a person joined Reddit you will find that the number of comments increases over time. For some reason people who join later seem to comment less and the early adopters of Reddit are the most active. Now Reddit's problem changes. They don't want people to write more, what they want is to attract more of those "good" people.

2 Counterfactuals

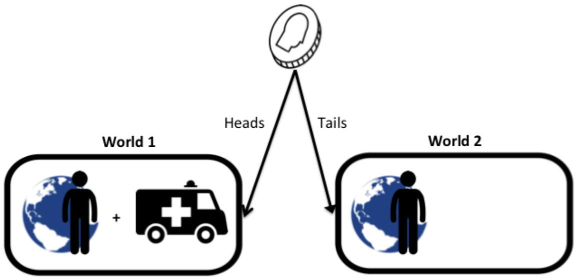
To isolate the causal effect, we have to change one and only one thing (hospitals visits), and compare outcomes. That is, in an ideal world we would take the world and make a clone world. In one case we would bring the guy to the hospital and in the other case we wouldn't. The only thing that differs between the two cases is whether or not we brought the guy to the hospital, everything else is the same. This is the core idea of what we try to do in an experiment. However, in practice it hard to do an experiment and say that world 1 and world 2 are equal except for the thing that we changed.

3 Random assignment

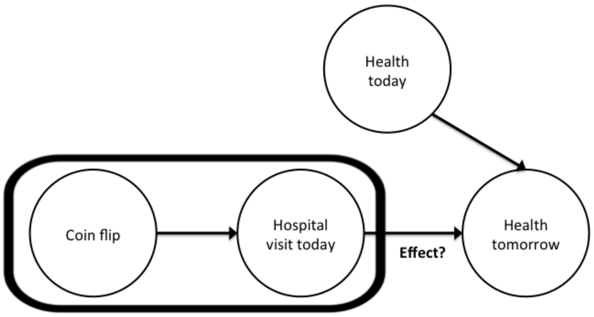
We can use randomization to create two groups that differ only in which treatment they receive, restoring symmetry. For each person, we will flip a coin and randomly assign them to going to the hospital or not going to the hospital. There are going to be healthy and sick people in each group. On average the people who get

Reality
+
(what happened)
Vs.
Counterfactual
(what would have happened)

sent and not sent to the hospital is the same, i.e. those groups should be identical because nothing else affected whether or not they went to the hospital. We can then measure the effect of people getting treated at hospitals or not.



Random assignment determines the treatment independent of any confounds.



4 Basic identity of causal inference

The observed difference is now the causal effect: Observed difference = Causal effect - Selection bias = Causal effect

Selection bias is zero, since there's no difference, on average between those who were hospitalized and those who weren't.

5 Experiments

5.1 Caveats/limitations

Random assignment is the "gold standard" for causal inference, but it has some limitations:

- Randomization often isn't feasible and/or ethical (often happens in medical settings)

- Experiments are costly in terms of time and money
- It's difficult to create convincing parallel worlds
- Inevitably people deviate from their random assignments

5.2 Two goals for experiments

1. Internal validity: Could anything other than the treatment (i.e. a confound) have produced this outcome?
    - Did doctors give the experimental drug to some especially sick patients (breaking randomization) hoping that it would save them?
2. External validity (Generalization): Do the results of the experiment hold in settings we care about?
    - Would this medication be just as effective outside of a clinical trial, when usage is less rigorously monitored?

5.3 How we conduct behavioral experiments and write academic papers across the social sciences

1. Lab Experiment - what are held at colleges
    - Better internal validity (correctness): Greatest procedural control, can carefully curate situations
    - But less external validity (generalization): Artificial context, simple tasks, demand effects, homogeneous (WEIRD - Western Educated Industrialized Rich Democratic) subject pools, time/ scale limitation - You can't keep people there and you can only bring in a certain number of people
2. Field Experiment - run an experiment in the real world
    - Better generalization: Experiment findings apply to at least one real-world setting
    - But: Less control, more potential confounds, demand of experiment conflict with goals of real organizations, more effort to conduct and manage, more room for error

These kinds of issues also come up in systems as well. Even in the systems that we build ourselves it is not that simple. Internal validity is a problem at companies like Microsoft and Google. If your system has an A/B test you should wait at least 6 months because there are so many things that happen before your code reaches the user. Generalization also comes in. Suppose you run your code and it works well in New York, will it also work well in another city, in another country?

5.4 Natural experiments

Sometimes we get luck and nature effectively runs experiments for us e.g:

- As-if random: People are randomly exposed to water sources
- Instrumental variables: A lottery influences military service
    - Idea: An instrument independently shifts the distribution of a treatment
- Discontinuities: Star ratings get arbitrarily rounded
- Difference in differences: Minimum wage changes in just one state

5.5 Behavioral science labs are very limiting

Pros:

- High degree of procedural control, Optimized for causal inference

Cons (limitations):

- Artificial environment, Simple tasks, demand effects, Homogeneous (WEIRD) subject pools, Time/scale limitations, Expensive, difficult to set up

5.6 Experiments are underpowered

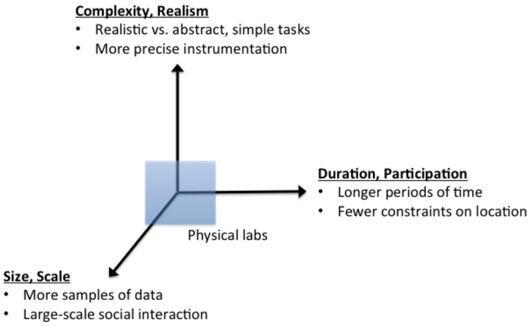
Two-thirds of psychology studies don't replicate! This shows how hard it is to do experiments properly. And if you do do it properly you should be able to run the experiment again and get the same results.

5.7 Most social science experiments aren't social

The vast majority of experiments are done with individuals as it is harder to do experiments on groups of people. Therefore, experiments are focused on individual behavior: logistically it's just easier. We actually don't know the causal effects of policies in many large-scale, collective behavior settings. Some examples include: economic inequality, systems of governance, and the black box of macroeconomic policy.

5.8 So, about them experiments

They are costly run but are limited in the types of questions they can answer. Large-scale observational data is useful for building predictive models of a static world. Randomized experiments are like custom-made datasets to answer a specific question. Moreover, published experimental research is probably wrong and they are far from answering many big important questions. So how do we fix this? By expanding the experiment design space.


Complexity, Realism
• Realistic vs. abstract, simple tasks
• More precise instrumentation
Duration, Participation
• Longer periods of time
• Fewer constraints on location
Physical labs
Size, Scale
• More samples of data
• Large-scale social interaction

Notes from rgn2112

1 Prediction vs. Causation

So far, we've been trying to understand data to make a forecast. Causation, however, is to try and anticipate what will happen. For example, people out on the street holding umbrellas might predict rain, but it does not cause rain.

1.1 Types of Causal Inference

Not all types of causal inference are created equal. For example, asking "What makes Apple so successful?" brings with it different challenges than the question "What effects does hospitalization have on health?"

Reverse Causal Inference: Looking at the Causes of Effects. Generally quite hard to do, nearly impossible. There are tendencies to have tons of effects, making it difficult to isolate out causes. The Apple question is an example of Reverse Causal Inference.

Forward Causal Inference: Looking at the Effects of Causes. While it's still difficult to do properly, it's less contentious and can lead to actual conclusions. Most science relies on Forward Causal Inference to connect causes to their effects because properly conducted experiments can isolate out the causal effects in data. The hospitalization question is an example of this.

1.2 The Problem with Forward Causal Inference

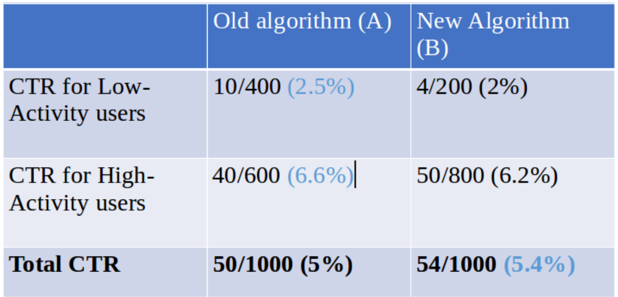
Great! Now that we have our method, we should be able to grab any old dataset and do some Forward Caussal Inference magic and we're good to go! Well, not exactly. If we use an arbitrarily gathered dataset, that dataset can inherently contain biases that distort out the causal effects we hope to study. These are known as *confounds*. For example, returning to the hospitalizaion example, health of an individual determines whether or not they end up going to the hospital. In the extreme case, where all sick people go to the hospital and all healthy people stay home, we would find that hopitalization reduces your hance of survival! The problem here is that we can't observe the effect of sick people not going to the hopsital.
The above example gets into the issue of **selection bias**. In any kind of data, there's a selection bias present, whereby the group of people has been pre-selected, causing problems with trying to isolate causal relationships. In all situations, the observed difference between two groups can be described by the following formula:

$$\Delta_{obs} = causal\ effect - selection\ bias \tag{1}$$

Selection bias is usually negative.

1.3 Simpson's Paradox

Selection bias can be so large that observational and causal estimates give opposite effects. In other words, you can get completely flipped answers by not conditioning on the right variables. For example, when comparing two algorithms, you can find that one algorithm apparently outperforms the other, but, in reality, the opposite is true: Now what? Simpson's Paradox seems to suggest that, if we're not careful, we're going to be sent down endless rabbit holes, never knowing when exactly we've reached the conclusion of our analysis. Fear not! Science has an answer!



	Old algorithm (A)	New Algorithm (B)
CTR for Low-Activity users	10/400 (2.5%)	4/200 (2%)
CTR for High-Activity users	40/600 (6.6%)	50/800 (6.2%)
Total CTR	**50/1000 (5%)**	**54/1000 (5.4%)**

Figure 5: Simpson's Paradox in action

2 Experiments

2.1 The Light at the End of the Tunnel

Counterfactuals: To isolate the causal effect in the data, we have to change one things and one thing only. We try to clone the current universe and just change one thing in the new universe. This allows us to compare reality to the counterfactual. Where reality is what happened, whereas the counterfactual is what would've happened.

To accomplish this feat, we introduce a sense of randomness into the world. We randomly assign individuals to World A and World B where the only difference between the two is the one change we've made. In this matter, we've reframed the previous formula:

$$\Delta_{obs} = causal\ effect \tag{2}$$

Selection bias has, for the most part, been removed due to the introduction of randomization. With this technique, then, we have constructed a means of measuring causal effects.

2.2 The Trade-Off: Internal Validity vs. External Validity

Although we managed to generate a means of measuring causation in datasets, the fact of the matter is that everything isn't perfect. Randomization is easy to write about but difficult to accomplish in practice, running into ethical or feasibility concerns. Experiments often are costly and take time to set up. Finally, it is often *painstakingly difficult* to create convincing parallel worlds. Due to this issues, scientists make trade-offs when constructing experiments.

Internal Validity: Ensuring that the observed changes are the result of the change of the world and that change early. The question becomes: Could anything other than the treatment (aka the change) produce this outcome? Lab Experiments are great at controlling for Internal Validity. Experimenters have great control over introducing changes in the population. Trying to measure internval validity outside of a lab environment is difficult, however, since the real world introduces many different variables that could influence the result.

External Validity: Ensuring that the observed changes are not just observable in the experiment but hold in other settings as well. For example, would the tested medication be effective outside of a clinical trial? Field experiments that take place outside of a lab environment are great for external validity. They generalize better and apply to real-world situations. Lab experiments, on the other hand, run into problems with generalzation, since tested populations are homogeneous (see W.E.I.R.D.) and the environment is artifical.

2.3 The Promise of the Internet

Research has historically been bounded by the previous trade-off, but the introduction of the Internet has the potential to rework this system entirely. Properly conduted Internet experiments have hte potential to weave together the benefits of the lab and the field, allowing researchers to turn the historical trade-off into more of a spectrum. The hope is that, one day, researchers will develop a system that allow the mass-testing of various behavioral experiments online.

3 Natural Experiments

3.1 When Reseachers get Lucky

Sometimes, nature does the experimenter's job for them. Some systems have introduced features of experiments that effectively isolate causal effects. Although this is a rare occurence, it is useful to keep an eye out for and allows for otherwise difficult questions to be answered.

As-if Random: Sometimes people are randomly distributed already without the intervention of the experimenter. The best example of this is John Snow's study of the cholera outbreak in London. Because he found that people were exposed to water sources as-if randomly, he was able to demonstrate that there was a causal link between contaminated water and sickness. The goal for this type of study is to find the features in nature that are random in themselves.

Instrumntal Variables: Sometimes human systems randomly assign human beings to different fates. In otherwords, an instrument independently shifts the distribution of the treatment. An example of this is the draft lottery, whereby people were randomly assigned to military service or not. With this situation, it is possible to ask what the effects of military service are on an individual's life. This situation doesn't necesarily have complete randomness, but the system has *some randomness* that allows you, as a researcher, to infer out the causal relationship.

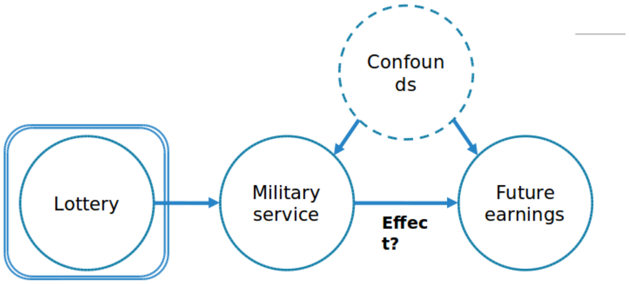


Figure 6: Example of Instrumental Variable in Military Lottery

Notes from sl4095

1 Intro to Causal Inference

Prediction: Make a forecast, leaving the world as it is. For example, seeing my neighbor with an umbrella might predict rain.
Causation: Anticipate what will happen when you make a change in the world. However, for the same example in prediction, seeing my neighbor bringing an umbrella doesn't cause the rain.
It's very tempting to ask "what caused Y", but this is "reverse causal inference", and is generally quite hard. Alternatively, we can ask "what happens if we do X?" For example,

- How does education impact future earnings?
- What is the effect of advertising on sales?
- How does hospitalization affect health?

This is "forward causal inference". It's still hard, but less contentious.

2 Hospitalization on Health

Suppose now we have data of people who visit hospital today, and people's health tomorrow, we want to know the effect of going to hospital today over the health of that person tomorrow. Estimating the effect of going to hospital using these observational data is wrong, because the effect and cause might be *confounded* by a common cause, and be changing together as a result. We have people who are healthy today and therefore don't go to hospital, and people who are not healthy today but still don't go to the hospital. We also know that people who go to the hospital typically are the people who are sick. If we only get to observe the data of hospital visit today and health tomorrow changing together, we can't estimate the effect of hospitalization changing alone.
Let's say all sick people in our dataset went to the hospital today, and healthy people stayed home. The observed difference in health tomorrow is:

$$\Delta_{obs} = (\text{Sick and went to hospital} - \text{Sick if stayed home}) + (\text{Sick if stayed home} - \text{Healthy and stayed home}) \tag{3}$$

The first part is the causal effect, while the second part is the selection bias.

3 From data to prediction

This is a fundamental problem across science and industry. For science, we want to know what is the effect of doing X, while in industry we would like to know when we should do X.
How does the predictive system work?
First, we see data about user profiles and past activities. For example, for any user, we might see their age, gender, past activity and their social network. People have higher activity tend to have more friends, and people who have lower activity tend to have fewer friends. We would use these correlations to make a predictive model, for example, Future Activity $= f($number of friends, logins in past month$)$. Number of friends can predict activity with high accuracy. Now, we want to have "actionable insights", that is, we want to know how do we increase activity of users. There are multiple explanations, and it's hard for us to know what causes what. In order to increase activity, would it make sense to launch a campaign to increase friends?
Another example: Search engines uses ad targeting to show relevant ads. We have prediction model based on user's search query. Search ads have the highest click-through rate in online ads. However, even without search ads, it's also highly possible that we will reach the same website advertised. Therefore without reasoning about causality, we may overestimate the effectiveness of ads.

4 Simpson's paradox

Selection bias can be so large that observational and causal estimates give opposite effects. Here's an excellent article from Reddit.
https://www.reddit.com/r/Entrepreneur/comments/60ob8w/is_your_data_lying_to_you_the_simpsons_paradox/

5 Experiments!

To isolate the causal effect, we have to change on and only one thing, and compare outcomes. For example, in hospital case, we change hospital visits and compare the health tomorrow. In order to avoid selection bias, we randomly assign people who go to the hospital and people who don't. Then the observed difference is just the causal effect.
Even though random assignment is the "gold standard" for causal inference, it has some limitations:

- Randomization often isn't feasible and/or ethical
- Experiments are costly in terms of time and money
- It's difficult to create convincing parallel worlds.
- Inevitably people deviate from their random assignments

Natural experiments: Sometimes we get lucky and nature effectively runs experiments for us.

- As-if random: People are randomly exposed to water sources
- Instrumental variables: A lottery influences military service
- Discontinuities: Star ratings get arbitrarily rounded
- Difference in differences: minimum wage changes in just one state

These natural experiments are great, but they are hard to find, and there are many untestable assumptions.