Lecture 12: Causality & Experiments, Part 2
Modeling Social Data, Spring 2017
Columbia University

April 21, 2017

Notes from cms2286

1 Random Experiments

1.1 Confounds

Reviewing the example from last lecture, if we want to assess the effect of going to hospitals on health, we need to account for confounds. The cause and effect of the outcome may be affected (confounded) by a common cause, and this impacts the result. As seen in Figure 1, a person's health today has an effect on both on whether they go to hospital and their health tomorrow,

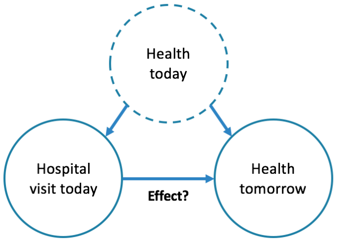


Figure 1: *Source: Lecture 11*

1.2 Random Assignment

Random Assignment is the process of assigning human participants into different groups using randomization.

Using random assignment breaks the link between confounds and the results. By assigning different treatments randomly, we can account for the confounds.

For example, if the people who would go to the hospital were chosen by a coin flip, we would be able to isolate the effect of going to the hospital without the confounds, as shown in Figure 2.

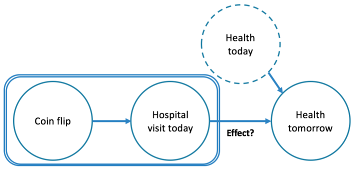


Figure 2: *Source: Lecture 11*

1.2.1 Problems with Random Assignment

While it is considered a "gold standard" of causal inference, it may be misleading. There are 4 main issues with random assignment: small sample sizes, researcher degrees of freedom, publication bias, and p-hacking.

Small sample sizes make the sample more subject to statistical variation, leading to the potential of observing several flukes. α, the significance level, and Power $(1-\beta)$, the chance of detecting a real effect if one exists, also have an effect on the experiment. To choose these values for an experiment, it is often useful to run a **pilot study**, a smaller, preliminary version of the experiment.

Researcher degrees of Freedom may affect the results, as the flexibility in analysis is greater than the flexibility in the data. A researcher may have decided to run a different test on different data, for example, and they may also decide when to stop collecting samples such that the data fits a desired result.

Publication bias

p-hacking

1.2.2 Other Limitations

- Randomization often is possible or ethical
- Experiments are expensive (in terms of both time and money)
- It may be difficult to create feasible parallel worlds
- People may be non-compliers: they may not follow the rules of the test

1.2.3 Measuring Compliance Rates

To measure compliance rates, measure the fraction of people that accept treatment in the treatment group, and the fraction that gets treated anyway in the control group.

1.3 Counterfactuals

Counterfactuals are events or outcomes that did not happen, "what if?" questions. However, to isolate causal effect, we have to change one and only one thing and compare outcomes, so we are not able to see the counterfactual.

2 Natural Experiments

Sometimes, experiments happen naturally, without needing to be formally designed. There are four main kinds of Natural Experiment: as-if random, instrumental variables, discontinuities, and difference in differences.

2.1 As-if Random

Nature randomly assigns conditions, and people comply An example of this is the Cholera outbreak in London in 1854, in which people were exposed to different water sources.

2.2 Instrumental Variables

An instrument independently shifts the distribution of the treatment For example, a lottery determines the military draft. It functions like a random assignment, and helps obtain results.

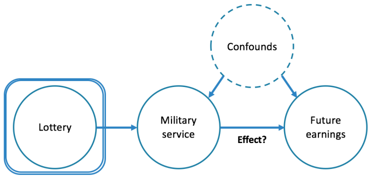


Figure 3: *Source: Lecture 11*

2.3 Regression Discontinuities

Things change around an arbitrarily chosen threshold Yelp reviews, for example, round the stars arbitrarily, resulting in the discontinuity seen in Figure 4.

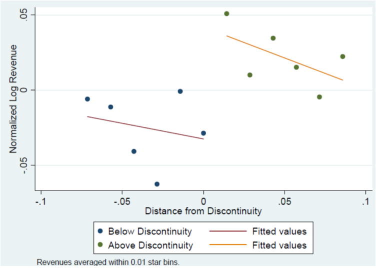


Figure 4: Yelp star ratings

2.4 Difference in Differences

Compare differences after a sudden change with trends in a control group If minimum wage changes in one state, for example, the effect may be visible in the difference in the trend line, as shown in Firgure 5.

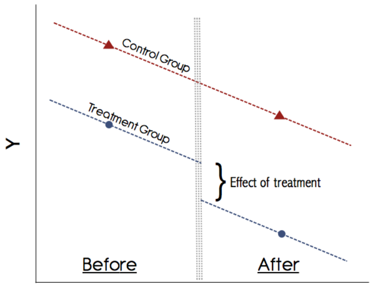


Figure 5: Difference in Treatments can be seen in Trendlines

Notes from jv2488

1 Introduction

Last class was an introduction to randomized and natural experiments. Today, we'll dig deeper into causality and experiments.

2 Randomization

2.1 Confounding variables

(Review from last week) How do hospital visits today impact health tomorrow? Health today impacts BOTH hospital visits today and health tomorrow so we cannot estimate effect of one thing changing (health tomorrow) when two things change together (hospital visit today and health today).

2.2 Randomization breaks the influence of confounds

(Review from last week) What we actually want are cloned worlds of "reality" and "counterfactual" where reality is what happened if someone went to the hospital and counterfactual is what would have happened if they did not go to the hospital (or vice versa).

- Random assignment mimics this because groups only differ in which hospital assignment they receive
- Coin flip that determines hospital visit today is independent of health today
- Random assignment breaks the link between confounds and the thing you care about

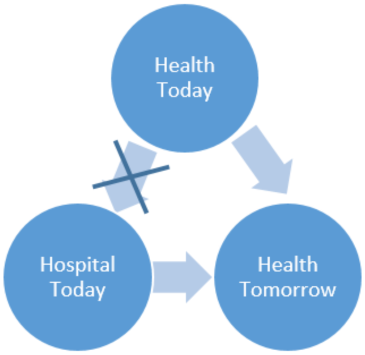
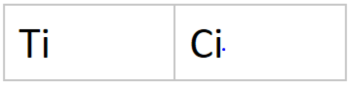


Figure 6: Breaking the confound allows one to study how hospital visits today impact health tomorrow.

2.3 Randomization as tickets



Ti	Ci

Figure 7: Think of a person represented as a ticket. Ti and Ci are what would happen to that person under treatment and control respectfully. (Idea from Dunning's adaptation from Friedman)

Randomization process is like drawing a random sample of tickets

- In treatment group, just get to observe treatment outcome aka not what would have happened in control
- In control group, just get to observe control outcome aka not what would have happened in treatment
- Average of random sample is unbiased estimate, so the average treatment outcome (estimate) is $Avg(Ti)$ and average control outcome (estimate) is $Avg(Ci)$. Therefore, the average treatment effect (estimate) is

$$ATE = Avg(Ti) - Avg(Ci)$$

3 Problems with Randomized Experiments

3.1 Reproducibility crisis

This is ongoing right now in social science (psychology) where many previously published studies not reproducible/wrong.

3.2 Small sample sizes

Small sample sizes are more likely to yield flukes in results due to statistical variation between samples.

Example: Imagine some area of science with 1000 scientists. Let's say 30 percent of things they are studying have real effects (aka there is a difference between treatment and control) but we do not know that it is 30 percent - no way to know.

$$N = samplesize$$

$$Alpha = significancelevel = P(significanceintest|norealeffect)$$

(Not p-value, but the level you decide p-value needs to be under in order for you to reject null hypothesis; alternatively, the false positive rate)

$$Power = 1 - beta = P(significance|realeffect)$$

(The the chance of detecting a real effect; kind of like sensitivity)
Result: Given Alpha = 0.05, Power = 0.35, and N = 1000, there are 105 real positives and 35 false positives

$$FalseDiscoveryRate = 0.25$$

$$Whatwewant : P(realeffect|significance)$$

Takeaway: Power matters! Even if small percent of flukes, if small power and small number of real effects, the FDR will still be high.

3.3 Degrees of freedom of researcher, publication bias, p-hacking

While these can be accidental, they can also be purposeful and devious with people gaming the system.

Example: Let's say you run an experiment. It's expensive to set up, hard to get people into lab, etc. The data does not match the initial hypothesis. What now? Maybe you try slicing data in a particular way, decide to run a different test, etc. This influences the results because had you gotten a different set of data, you could have not only gotten a different test outcome, but also, you may have run a different test.
Takeaway: This can induce as much variability as small sample size

3.4 Caveats/limitations

- May not be feasible/ethical
- Costly - time/money
- Hard to create parallel worlds
- Non compliance

Example: Facebook trying to test video chat. This is hard because it involves pairs of people communicating with each other. How does one find a group to expose randomly and another to not expose if everyone is connected? That's why last week, New Zealand was mentioned as a good "isolated" test area similar to US. (Difficulty in creating parallel worlds)
Takeaway: Independence in what you do to one unit and what you do with another unit is often broken

4 Recent work with Randomized Experiments

4.1 UPenn music study

Study 1 - listening to music changes how old you feel e.g. listening to children's song makes people report feeling older. Study 2 - listening to music changes how old you are e.g. listening to when I'm sixty four made people report being younger.

- Strange points of experiment: Study 1 and Study 2 had different sample sizes (both of which were small), different control songs, and slightly different survey methods.

What actually happened

- They ran a bunch of different versions of the experiment and measured a bunch of different outcomes, a couple of which turned out to be significant. (As they introduced more and more flexibility, by dumb luck, one turned out to be significant.)
- They did not share everything about their methods: they added observations and stopped when got results they wanted; they tested 3 songs in the experiment and only told you about 2 for each case
- A clever way of showing how significance can be misleading in a paper

What should be done

- Problem is exactly like overfitting to your experimental data set in machine learning; exploratory data analysis is fine, but cannot decide the test you do based on that data
- Solution is in real life holdout set where need to decide test first (like via a pilot experiment) and THEN do test on new data, once pilot errors ironed out
- Hofman @ Microsoft: do experiment, then do it again with new set of people without changing parameters to ensure robustness

4.2 Experimental evidence of massive-scale emotional contagion through social networks

Facebook's 7000 person experiment where they had a list of sad/happy words from research done in the 50s, promoted some amount of posts with these words on people's feeds, and measured how much they used the words after. The result was that seeing a friend express an emotion gives you a higher prevalence of expressing that emotion without social interaction. This study was very controversial and generated backlash.
Ethics

- Google, Facebook, etc. run experiments all the time; for example, a button on gmail was tested in 40 shades of blue to see what is optimal for click through. People don't seem to be upset about those because "not manipulated"
- With this experiment, bias influenced by machine learning algorithm could be a lot larger than the actual manipulation and there was a big exaggeration of the effect.
- Current assumption that what Facebook is showing you is true is false; algorithm picks/chooses

Institutional problems

- Informed consent: Technically in Facebook's agreement for signing up, but not in a way that people are aware. No one was notified after the experiment.
- Publishing data: Plausible? Raw data is on hundreds of servers. For other experiments, Facebook has released final tables that could reproduce plots in their papers which is but still, do we know it's true?

5 R Simulations

Source: Yakir 11.2.3, Intro to Statistical Thinking without Calculus with R

What is probability of a coin landing on heads (p)? Forget statistics, just run the experiment a huge number of times (N) to see what happens. We can see that sampling distribution gets narrower around p and taller as N increases (aka there is less variation around the true average).

Confidence intervals If pick a 0.95 confidence interval, this means that 0.95 of the time, our true result will be in the interval. We don't know whether the truth is in the interval we have.

Hypothesis testing *Single coin*: how does our coin compare to the null distribution of a fair coin? Since it's very unlikely to get our result with a fair coin, reject null hypothesis that it is fair. *Two coins*: guess p1 = .12 and p2 = .08. Are they different? Since it's not unlikely enough that they are different to get our result, do not reject null hypothesis that they are the same.

Power.prop.test (also see power.t.test) What sample size do we need if we want a certain power and alpha? Gives you sample size of 700; we used 500 people, so may have missed a real effect because N too small. This would be expensive to do with a simulation.

6 Natural Experiments

6.1 As if random

You did not set it up, but it is really like a randomized trial

Example: Cholera from last week

6.2 Discontinuities

Consequences of an arbitrary discontinuity: can have real consequences, or can be statistically significant but practically insignificant

Example: Star ratings get arbitrarily rounded.
How different are 3.48 and 3.51? Estimates, uncertainty on them anyway so not real a real difference. Yet, some get rounded to 3 and some to 4. Does this have monetary impact?

Example: National test creates arbitrary inequality in access top schools.
There is a threshold; if someone scores above it, they have access to certain schools/opportunity/publicity. However, people right above and below are identical. What is the effect of being above/below threshold?

6.3 Differences in differences

Similar to discontinuities but accounts for trends over time (at least in our example below)

Example: Minimum wage changes in one state (NJ changes, PA does not)
Argue PA and NJ are the same - in employers, etc. Check trends in unemployment in PA and NJ before and after wage change. Both states have decreasing unemployment, but in NJ after wage change, there is a huge sudden drop.

6.4 Instrumental variables

Touched upon this last week. There is a variable that we control that influences the assignment to the treatment but does not determine the treatment

Example: Military draft
Lottery: Influences who serves in military but does not determine it; for example, there are people who want to go anyway, people can't go for medical reasons or professional reasons, etc. Question is either "what is the effect of the policy on future earnings" OR "what is the effect of military service on future earnings"

Strategy

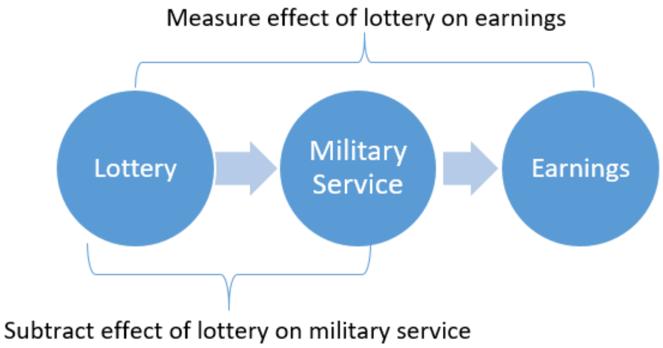
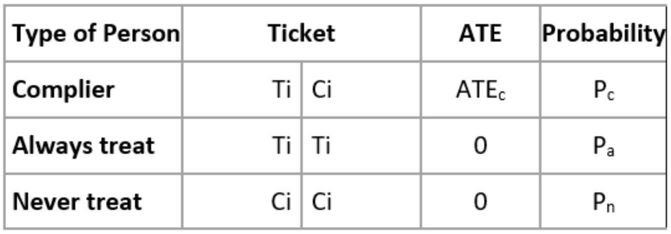

Measure effect of lottery on earnings
Lottery
Military Service
Earnings
Subtract effect of lottery on military service

Figure 8: Measuring the effect of lottery on earnings and subtracting the effect of lottery on military service is easy because the information is there and gives us the effect of military service on earnings, which is what we want.

Average treatment effect: effect of assignment to treatment on earnings (not of treatment itself)



Type of Person	Ticket		ATE	Probability
Complier	Ti	Ci	ATE_c	P_c
Always treat	Ti	Ti	0	P_a
Never treat	Ci	Ci	0	P_n

Figure 9: Types of people (excluding defiers).

$$ATEtotal = Pc(ATEc) + Pa(ATEa) + Pn(ATEn) = Pc(ATEc)$$

If we know Pc, we can find$ATEtotal = Pc(ATEc)$

Set of people you assigned to control AND who serve: Always treats

Set of people you assigned to treat AND who don't serve: Never treats

Fraction of people who accept treatment in the treatment group $= Pc + Pa$

Fraction of people who accept treatment in the control group $= Pa$

$$(Pc + Pa) - Pa = Pc$$

Conceptual notes on instrumental variables

- Make sure none of the confounds influence the instrument. This is hard to prove, so people tend to use a certain set of accepted instruments like weather.
- Make sure the instrument does not impacts outcome directly; e.g. if someone felt unlucky by being drafted and stopped trying, the lottery impacted their earnings
- One could use an instrumental variable in a randomized experiment with issues like non-compliance to "patch it up"

7 Caveats with Natural Experiments

- Hard to find
- Untestable assumptions
- Set of people you treat may not be the ones you care about

Potential solution: use additional data/algorithms to find natural experiments

8 Recent work with Natural Experiments

8.1 Exercise contagion in a global social network

Cannot realistically find a social network, make some people exercise, and see if friends do too. Therefore, use weather as an instrumental variable. If it rains in NY, SF residents see less running from friends in NY. Does this impact running in SF?

Conditions for this to work

- Rain in SF and in NY is independent; otherwise rain in NY could impact rain in SF which would then influence running in SF.
- Rain in a city does indeed impact running in that city.
- This design isolates peer influence; e.g. it excludes things like weekend effect and weather in SF.

8.2 Amazon: How much traffic does a recommendation cause?

Someone was looking for a winter hat and saw a recommendation for gloves. Did this cause them to buy gloves, or would they have bought them anyway?

- **Confound**: Demand for winter items could be confounding variable.
- **Randomization**: Cannot do an AB test to get rid of this because do not work at Amazon.
- **Solution**: Can look at an external shock that causes a sudden influx of users (Canonical example is Oprah features a book, which causes a huge influx in people buying it)

What we care about: Once they show up, do they click through to recommended item?

- **Instrumental Variable**: Sudden influx of users acts as an instrument
- **Potential side biases**: Even if there is click-through, it is because the people looking are already interested in similar works/authors
- **Solution**: As long as other traffic to the recommended product is constant, then can look at added people buying that product and attribute it to click-throughs due to the shock

Result: 75 percent of activity on recommended products would have happened anyway

9 Conclusion

While observational data is great for predictive models of a static world, it is difficult to figure out causal effects from it. Randomized experiments are like customized datasets great for figuring out causal mechanisms, but can easily go wrong in practice. Looking at additional data/algorithms as a way to discover natural experiments is an exciting and developing practice that will be important going forward.

Notes from meh2243

1 Introduction/Recap

Last week the class was introduced to causality and experiments by Andrew and Amit, Microsoft Researchers who were the guest lecturers. Today, the lecture focuses more in-depth on these concepts. First we discussed random experiments, highlighting why they are great in theory, why they can be troublesome in practice, and what you can do to not fall prey to these problems. Next we discussed different designs of natural experiments, as well as some caveats. Finally, we explored some recent work Professor Hofman has been undertaking with Amit regarding discovering natural experiments.

2 Random Experiments

2.1 Review example from last lecture

If we want to assess the effect of hospitalization on health, we must account for confounds: the cause and effect may be confounded by a common cause, and be changing together as a result. Figure 1 displays this idea: our health today has an effect on our hospitalization today, as well as our health tomorrow.

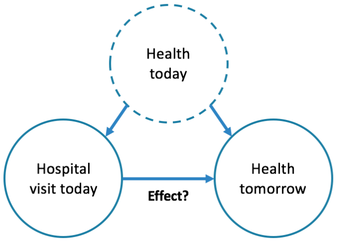


Figure 10: Effect and cause may be confounded by a common cause *Source: lecture 11 slides*

As Box explains, "to find out what happens when you change something, it is necessary to change it."

2.2 Random assignment

Through **random assignment** we are able to break this link between the confound and the second event (the thing we care about). Through intervention, say a coin flip as in figure 2, we are able to determent treatment independent of any confounds.

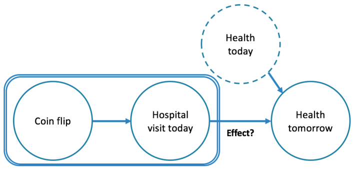


Figure 11: Through intervention, we can break the link between the confounded thing and what we care about *Source: lecture 11 slides*

Under this approach, we are not able to see what didn't happen, a concept that is referred to as **counterfactuals**. To isolate the causal effect, we have to change one and only one thing, and then compare the outcomes. All else must be equal; while this is possible with experiments where we can ensure identical formats and procedures are followed, we cannot do this with people.

Random Assignment is the process of assigning one of two worlds to an instance and observing what happens. The groups in each of these two worlds are only different in what treatment they receive. Neyman's Model does a great job in understanding why randomization works, which will later help us in understanding instrumental variables in natural experiments. In Figure 3, each person i is represented by two tickets: T_i representing what happens under treatment, and C_i representing what happens under control. When we draw a random sample, for those in the treatment group, we only observe the T_i outcome, and vice versa for the control group. Random assignment is usually presented as conditional expectations, but Neyman's model is a simpler and intuitive explanation.

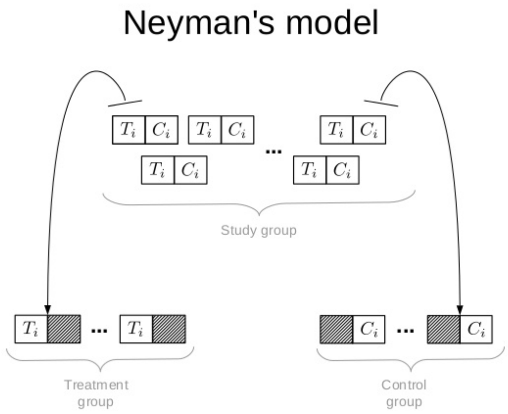


Figure 12: Neyman's Model *Source: https://www.slideshare.net/ChaToX/natural-experiments*

Here are some useful calculations. Note, since the average of a random sample is an unbiased estimator, $\overline{T}$ and $\overline{C}$ are unbiased.

- Average treatment outcome: $\overline{T} = \frac{1}{N_T}\sum_i T_i$
- Average control outcome: $\overline{C} = \frac{1}{N_C}\sum_i C_i$
- Average treatement effect: $ATE = \overline{T} - \overline{C}$

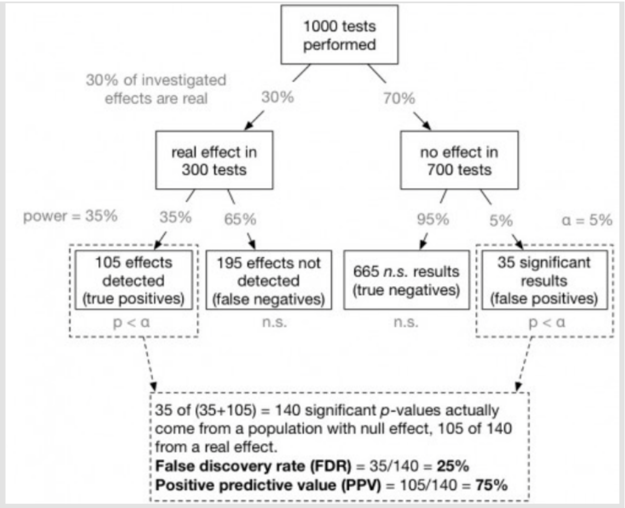
2.3 Problems with Random Assignment

While random assignment is the gold standard for casual inference, it can be misleading under different circumstances. We are currently amidst a reproducibility crisis whereby published papers are wrong. While this is positive in that we are learning lots, this is disheartening and there is lost faith in the scientific process.

2.3.1 Small sample sizes

When sample sizes are small, the sample is more subject to statistical variation. This sampling variability can lead to the observation of "flukes". Figure 4 observes an instance where 30% of investigated effects are real, and the other details are as follows:

- Sample size (N) is 1000
- The significance level (α), the false positive rate, is 5%
- Power ($1 - \beta$), the change of detecting a real effect if one exists, is 35%


1000 tests performed
30% of investigated effects are real
30%
70%
real effect in 300 tests
no effect in 700 tests
power = 35%
35%
65%
95%
5%
α = 5%
105 effects detected (true positives)
195 effects not detected (false negatives)
665 *n.s.* results (true negatives)
35 significant results (false positives)
p < α
n.s.
n.s.
p < α

35 of (35+105) = 140 significant *p*-values actually come from a population with null effect, 105 of 140 from a real effect.
False discovery rate (FDR) = 35/140 = **25%**
Positive predictive value (PPV) = 105/140 = **75%**

Figure 13: False Positives *Source: http://bit.ly/2ouDzUC*

Keep in mind...

- We are trying to calculate P(effect | significance)
- $\alpha = P$(significance | no effect)
- Power $= P$(significance | effect)

We can use R to simulate experiments and observe the effects of sample size, power, and significance levels. Check out the code from this week's lecture for more details, but some of the main ideas discussed were:

- How does variation change as we change the number of experiments (i.e. the sample size)? As sample size increases, sampling variation decreases. However, doubling the sample size does not result in a halving of the width of the distribution. The standard error as a function of N decreases according to $\frac{1}{\sqrt{N}}$
- After computing the upper and lower confidence intervals, it is useful to heck how often the true proportion is contained in the confidence interval.
- With hypothesis testing, we simulate the null distribution, and then compare this to the real experiment that we do just once. We then set a threshold, α, which is the minimum size the p-value can be to reject the null.
- In comparing two proportions, we can run experiments for both and then look at the distribution of the outcomes.
- Use a **power proportion test** to compute the N you need to obtain a certain power and significance level. This will tell us what to set N to in order to detect something $x\%$ of the time (here x is the power level you set). Note, that we see diminishing returns at a point: as we keep increasing N, we don't get much more power. In psychological studies, power is typically in the 30% range.

A practical tip is to run a **pilot study** of your experiment. This is a small version of your experiment, which you can use to get (i) an estimate of the proportions, and (ii) a sense of what you did wrong before you execute experiment at scale. Then, you can use the power proportion test to see what size N you need. Make sure you are well-powered, so that you have a lower FDR. A caveat here is that in some industries, sometimes more thing have a real effect so you don't need power to be as high in order to achieve a low FDR.

2.3.2 Researcher degrees of freedom (plus publication bias and p-hacking)

Researcher degrees of freedom implies that decisions about what statistical tests to compute are made based on the data. Thus, there is flexibility in the analysis, beyond the flexibility in the sampling data. In this case, if there was different data, the researcher may have decided to run a different test, and we see this variability. This is an important point that is often missed, as we do not think about what we would have done in a different scenario with different data.

Under this topic, we discussed the UPenn paper titled, *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*. The objective of this paper was to show that by increasing the flexibility of the researcher, the more and more flukes you observe. The authors measured lots of different outcomes of their study, and sliced it in different ways to see the likelihood of obtaining a false positive. This experimental flexibility can be likened to overfitting a regression.

Andrew Gelman published a paper on this topic, titled *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. It is important to maintain a bit of skepticism when you see a random experiment.

2.3.3 Limitations of random assignment

Random assignment has some caveats/limitations:

- Randomization often isn't feasible/ethical
- Experiments are costly (time, money)
- Hard to create convincingly parallel worlds (don't forget the independence assumption)
- Inevitably people deviate from their random assignments: non-compliers

Ron Kohavi's work, *Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO* includes suggestions for controlling experiments on the web, including running A/A tests, pilot testing and considering day of week effects.

Next we discussed the paper *Experimental evidence of massive-scale emotional contagion through social networks*, which outlines details of a Facebook experiment. In this experiment, Facebook manipulated users feed and would either mute or promote words related to various emotions. Thus, though this manipulated feed, users would see slightly happier or sadder content than the unaltered posts, and then Facebook measured how often users used these emotion-related words. There was much backlash with this paper as users were uneasy about the manipulation of their feed, and without explicit consent. While the Facebook terms do legally cover the researchers, there still exist ethical and fairness concerns.

3 Natural Experiments

Sometimes we get lucky and nature or policy effectively runs experiments for us. The types of natural experiments we discussed were: as-if random, instrumental variables, discontinuities, and difference in differences.

3.0.4 As-if random

Idea: Nature randomly assigns conditions (and people also comply).

An example of this is the Cholera outbreak in London in 1854, where people were randomly exposed to different water sources, some of which were contaminated.

3.0.5 Instrumental variables

Idea: An instrument independently shifts the distribution of a treatment

For example, if you want to assess the effect of military service on future earnings, intervening with a lottery allows you to measure this effect. The instrumental variable must systematically shift the distribution. It is important to note that those assigned to treatment is not the same as the treatment people receive (this will be discussed later).

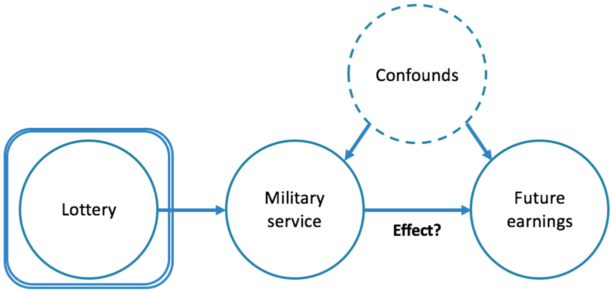


Figure 14: Instrumental variables: a lottery influences military service *Source: lecture 11 slides*

3.0.6 Regression discontinuities

Idea: Things change around an arbitrarily chosen threshold.

The example we discussed in class is related to Figure 5, where Yelp star ratings were arbitrarily rounded and the result is this jump in revenue.

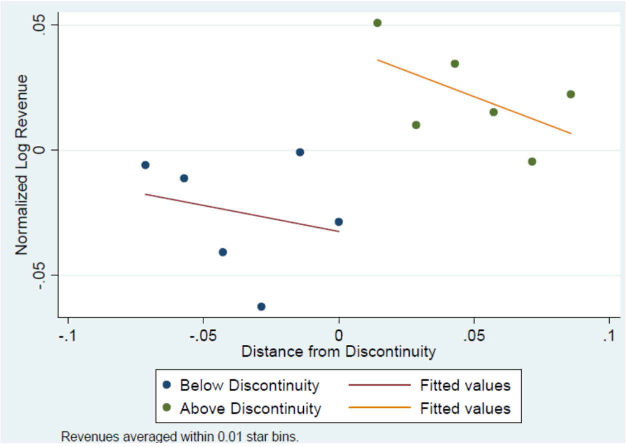


Figure 15: Regression discontinuities with Yelp ratings *Source: http://hbs.me/28NLHND*

3.0.7 Difference in differences

Idea: Compare differences after a sudden change with trends in a control group.

If wages change in just one state, for example, Figure 6 depicts this shifted trend line where we can see the effect of the treatment. It is important to note that this change must be sudden, and any anticipation will hurt the validity of the natural experiment results. This idea is similar to regression discontinuity, but differs in that with difference in differences there was already a trend before.

3.1 Instrumental Variables: In-depth

3.1.1 Calculating average treatment effect

As briefly noted above, assignment to treatment does not equal receipt of treatment. It is important to piece out noncompliance, and the ticket model described in the random assignment section can be used to do this. Consider three types of people: compliers (c), always-treats (a), and never-treats (n). We ignore defiers in this

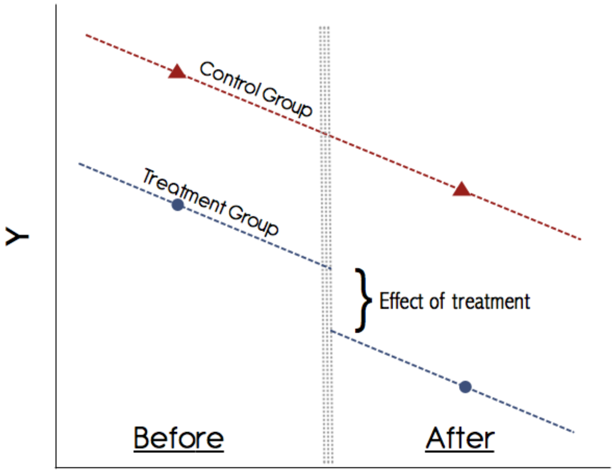


Figure 16: Difference in differences trend *Source: http://bit.ly/2p3YI9o*

world, and assume that people do not do the opposite of what they are told. Like before, compliers have both tickets: T_i and C_i. Always-treats on the other hand have two T_i tickets, and Never-treats have two C_i tickets.

In assessing the average treatment effect (ATE), we observe that for always-treats and never-treats, it is zero. Now, to calculate the ATE for compliers, we just use simple algebra:

$$ATE_{total} = P_c ATE_c + P_a ATE_a + P_n ATE_n$$

$$ATE_{total} = P_c ATE_c + 0 + 0$$

$$ATE_c = \frac{ATE_{total}}{P_c}$$

Now, in order to calculate the compliance rate, P_c we look at the fraction of those that accept treatment in a group $(P_c + P_a)$ less those that accept treatment in the control group (P_a). As a reminder, in the treatment group, those with T_i tickets $= P_c + P_a$, and in the control group, those with T_i tickets $= P_a$.

3.1.2 Example: Exercise contagion in a global social network

As an example of a natural experiment with an instrumental variable, we discussed Christos Nicolaides and Sinan Aral's paper *Exercise contagion in a global social network*. The idea behind this experiment was to see if social networks have an effect on exercise, using weather patterns as an instrument. For example, the weather in city A instrumentally shifts the probability of running in city A (i.e. if it is raining, less likely to run). The researchers then measured the effect of their friends in city B running. In order for this to be valid, it is important to argue that the weather in city B is not a confound for the weather in city A. Overall, this experiment changes on if your peers run or not, using the weather as an instrument.

3.1.3 Caveats

In practice, it is hard to find instruments. The instrument cannot have a direct influence on the "effect" action, like future earnings in the military example or running in city B in the exercise contagion example. Moreover, instrumental variables rely on many untestable assumptions, and you must be certain that no confounds influence the instrument. Lastly, with natural instrumental variables, the treated population may not be the one you are interested in investigating.

4 Discovering Natural Experiments

We can use additional data and algorithms to discover natural experiments. For example, online retailers are likely interested in learning how much traffic a recommender causes: was a recommender clickthrough causal, or a convenience click.

To set up this experiment as we have before, we are interested in seeing if a direct view for a focal product has an effect on a referred-click through. Both are confounded by demand for the products, but an external shock in traffic to the focal product can act as an instrument variable. By considering this sudden influx of users as an instrument, we are able to measure the marginal effect. Therefore, we are automatically discovering a natural experiment: first consider a big spike in traffic, then look for a small spike in recommender traffic (this represents an influx of traffic looking at the recommended item) and see if this spike also exists in direct traffic.

We are able to use data to find these shocks, without knowing what caused it. In the end, the casual clickthrough is marginal: to calculate the causal effect we look at the change in recommender clicks over the size of the shock

5 Concluding Remarks

Large scale observational data is useful for building predictive models, but without random variable it is hard to find causal effects. We've seen that random experiments are like custom data sets that we can use to answer specific questions. Further, data and algorithms have help us discover and analyze these examples in the wild.

Notes from mw3148

1 Problems of Experiments

Random assignment is the “gold standard” for causal inference, but it has some limitations:
1. Small sample size
2. Researcher’s degree of freedom
3. Publication bias: only those who reach statistically significant could be published.
4. P hacking

Factors influence power:
N = sample size
Alpha = significance level
Effect size (how strong is the effect? i.e. Cohen’s D)
P(significance — no effect) = "False Alarm"
Power = 1- Beta = chance of detecting a real effect if one exists.
P(significance — effect) = "Hit"

How to explain 95 percent confidence level:
if we replicate the “experiments” infinite time, 95 percent results would contain the parameter (true value).

2 Caveat and Limitations

1.randomization often is not feasible or ethical
2.experiments are costly in terms of time and money
3.it’s difficult to create convincing parallel worlds
4.inevitably people deviate from their random assignment

3 Natural Experiments

Sometimes we get lucky and nature effectively runs experiments for us e.g.:
1. As-if random: people are randomly exposed to water sources
2. Instrumental variables: a lottery influences military services:
IV influences treatment, but no association with the errors.
3. Regression-discontinuities:
idea: things change around an arbitrary chosen threshold.
4. Difference in differences:
idea: compare difference after a sudden change with trends in a control group

T C : “Compliers”
T T : “Always treats”
C C : “Never treats”

ATE = pcATEc + paATEa + pnATEn

(Fraction accept treatment in treatment group) – (Fraction accept treatment in control group) = (Pc + Pa) – Pa = Pc

4 Natural Experiments' limitations

1. Good natural experiments are hard to find
2. They rely on naturally-occurring event rather than controlled manipulation
3. Difficult to control for possible alternative explanations
4. Limited sample size

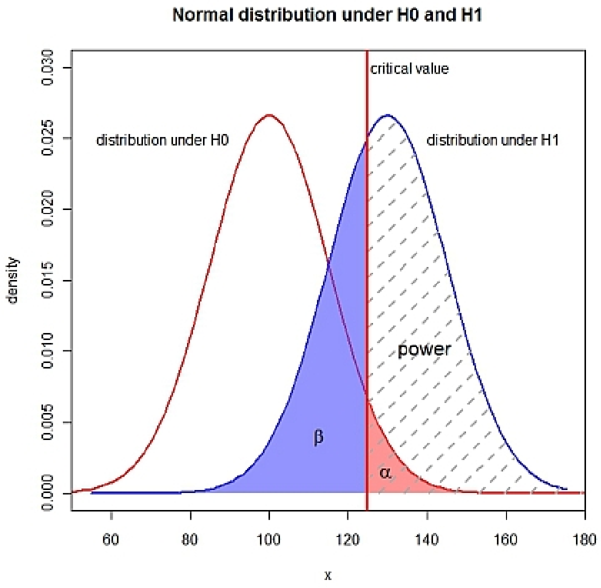


Figure 17: This figure's source is as below:
http://jamescaldwell.info/optimization/the-maths-behind-statistically-significant-sample-sizes/.

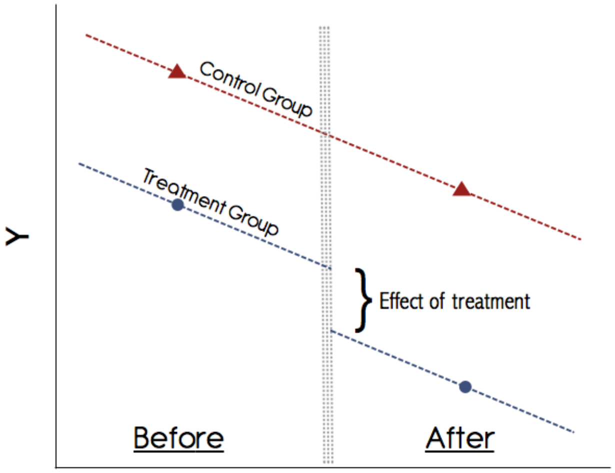


Figure 18: Example plot for Difference in Differences.source: https://i.stack.imgur.com/J7P3p.png

Notes from ww2440

- Random Assignment

$$\begin{array}{c}\hline T_i\ |C_i \\ \hline\end{array}$$

↗ ↖

what happens under treatment

what happens under control

Theoretically, in reality, you can never measure them at the same time.

Treatment group: $\begin{array}{c}\hline T_i\ |||| \\ \hline\end{array}$

Average treatment outcome:

$$\hat{\bar{\mathrm{T}}} = \frac{1}{N_T}\sum_i T_i$$

Control group: $\begin{array}{c}\hline |||||C_i \\ \hline\end{array}$

Average control outcome:

$$\hat{\bar{\mathrm{C}}} = \frac{1}{N_C}\sum_i C_i$$

Average of random sample is an unbiased estimate, and the average treatment effect is:

$$A\hat{T}E = \hat{\bar{\mathrm{T}}} - \hat{\bar{\mathrm{C}}}$$

- Problems

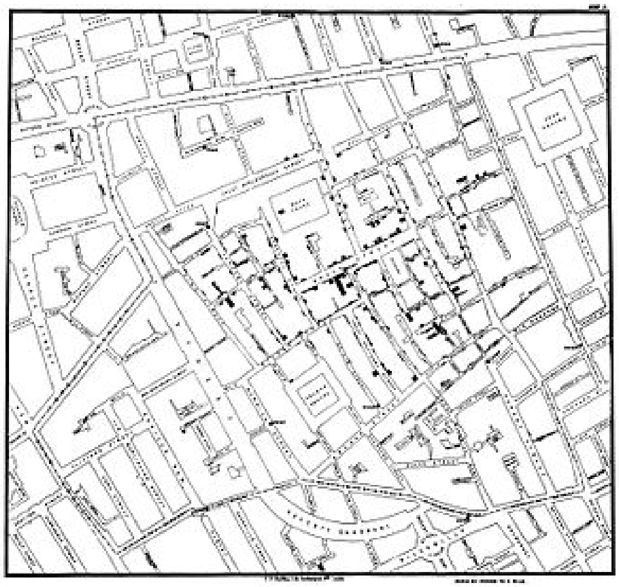


Figure 19: Example plot for natural experiment.
source: http://thelancet.com/journals/lancet/article/PIIS0140-6736(13)60830-2/fulltext?rss

  - Small sample size: large sample size can reduce SE and lower the chance of the estimate being way too off.
  - Researcher degrees of freedom: tend to utilize various method to "mine" the data for a nice result. The hypotheses and analysis method should be set before touching the data.
  - Publication bias: the reproducibility is questionable, especially in a field where the power is relatively low.
  - P-hacking

- Power Analysis
N = sample size
α = significance level = P(significance |no effect)
power = 1 - β = P(significance |effect) $\Longleftarrow$ chance of detecting a real effect if one exists

  - How to get hypothesized p-value? Run a pilot study!

- Limitation

  - Sometimes isn't feasible/ethical
  - Costly in terms of time and money
  - Difficult to create convincing parallel world
  - People inevitably deviate from assignment

- Non-Compliance

$T_i \; \| C_i$	Compliers	ATE_c
$T_i \; \| T_i$	Always treats	$ATE_a = 0$

C_i $|C_i$ Never treats $ATE_n = 0$

$Overall\,ATE = p_c ATE_c + p_a ATE_a + p_n ATE_n = p_c ATE_c$
Therefore $ATE_c = \frac{Overall\,ATE}{p_c}$

In the assigned-to-treatment group:

T_i |||| Compliers or Always-treats

|||||C_i Never-treats $\Longleftarrow$ tell people to serve but some don't

In the assigned-to-control group:

|||||C_i Compliers or Never-treats

T_i |||| Always-treats $\Longleftarrow$ tell people **not** to serve but some do serve

Fraction accept treatment in treatment group: $p_c + p_a$
Fraction accept treatment in control group: p_a
Therefore we can get p_c by deducting the second one from the first one: $p_c = (p_c + p_a) - p_a$

- Instrumental Variable
  The effect will break when:
  - Confound variable influences instrumental variable
  - Instrumental variable influences DV

  Another example of instrumental variable:
  $\boxed{Weather\,in\,city\,A} \Longrightarrow \boxed{Running\,in\,city\,A} \Longrightarrow \boxed{Running\,in\,city\,B}$

  The instrumental variable (weather in city A) only changes the probability of IV (Running in city A) so that we can figure out if IV **causes** DV.

Notes from yh2825

1 Random Assignments

1.1 Definition

Random Assignment is an experimental technique for assigning participants to different groups in an experiment using randomization.

1.2 Example: Neyman's Model

Random assignment breaks the link between the confounds and the results. The following example discusses the details of random assignment using Neyman's Model, and provides ground for understanding instrumental variables (section 2.2).

In the experiment, each person is represented by a ticket, $T_i|C_i$. T_i represents the outcome under treatment, and C_i represents the outcome under control. By randomly picking participants into treatment/control group, we want to observe the average treatment effect, ATE, by calculating the outcome differences. In each group, we only observe the corresponding outcome.

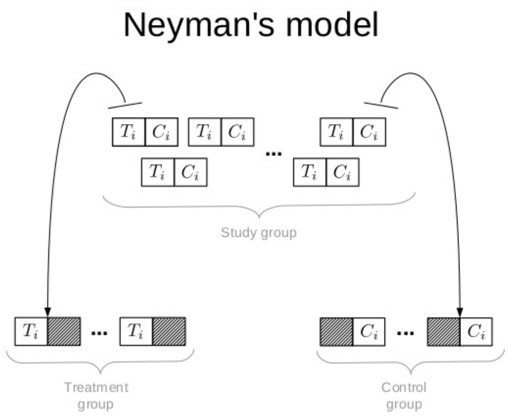


Figure 20: Neyman's Model

Since the sample is randomly selected, the average of the sample is an unbiased estimator of the true average.

- Average treatment outcome: $\hat{\overline{T}} = \frac{1}{N}\sum T_i$
- Average control outcome: $\hat{\overline{C}} = \frac{1}{N}\sum C_i$
- Average treatment effect: $\hat{\overline{ATE}} = \hat{\overline{T}} - \hat{\overline{C}}$

1.3 Problems of Random Assignment

Random Assignment does have several problems:

- **Small sample sizes**: When the sample size is small, the experiment is subject to statistical variation. Denote α to be the significance level, and power, $(1-\beta)$, to be the chance of detecting a real effect if the effect exists. In the following figure, α is 5%, and $(1-\beta)$ is 35%. Assume that 30% of the investigated

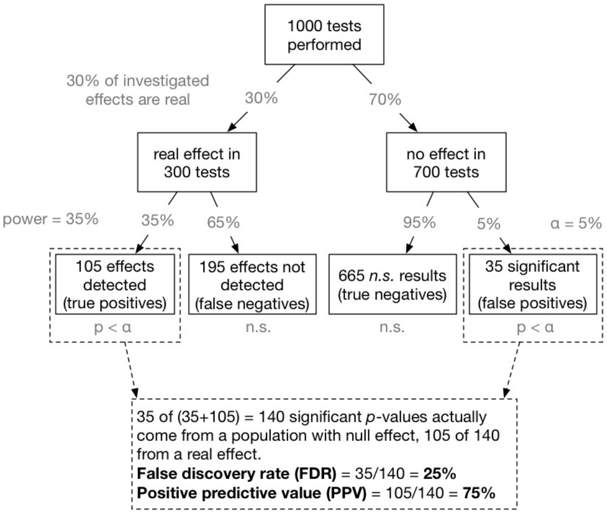

1000 tests performed
30% of investigated effects are real
30%
70%
real effect in 300 tests
no effect in 700 tests
power = 35%
35%
65%
95%
5%
α = 5%
105 effects detected (true positives)
195 effects not detected (false negatives)
665 *n.s.* results (true negatives)
35 significant results (false positives)
p < α
n.s.
n.s.
p < α
35 of (35+105) = 140 significant *p*-values actually come from a population with null effect, 105 of 140 from a real effect.
False discovery rate (FDR) = 35/140 = **25%**
Positive predictive value (PPV) = 105/140 = **75%**

Figure 21: Small Sample

effects are real, the false discovery rate is 25%.
Small sample sizes is commonly seen in many experiments today in social sciences. A good counter measure to it is to run a pilot study along with a power proportion test to discover the sample size needed. Usually, the lower the rate of real investigated effects is, the higher the sample size should be.

- **Researchers' degree of freedom, Publication Bias, p-hacking**: The variation of researchers' choice to run different tests or choose different sets of data is sometimes even larger than the variation in the data or experiment.

2 Natural Experiments

2.1 Definition

Natural Experiment is an empirical study in which individuals (or clusters of individuals) exposed to the experimental and control conditions are determined by nature or by other factors outside the control of the investigators, but the process governing the exposures arguably resembles random assignment.

2.2 Instrumental Variables

An instrumental variable in natural experiments is independent of the confounds and changes the distribution in the results. In the case of studying the correlation of military service and future earnings, the lottery of assigning military service is the instrumental variable. However, the assignment is different from the actual, and we need to consider the case of non-compliance.

Figure 3 shows the case of non-compliance in the previous treatment example. Consider the following three types of people: (a) Compliers, who complies to the given assignment (b) Always-treats, who always receive treatment regardless of the assignment (c) Never-treats, who always do not receive treatment.

Denote ATE_c, ATE_a, ATE_n to be the average treatment effect of the groups (a),(b) and (c), p_c, p_a, p_n to be the proportion of the three groups in the sample. Since group (b) and (c) will show no different in effect regardless of their assignments, the ATE for these groups are 0. Therefore, we have: $\overline{\hat{ATE}} = \overline{\hat{ATE_c}}p_c$. Note that the number we care about is ATE_c.

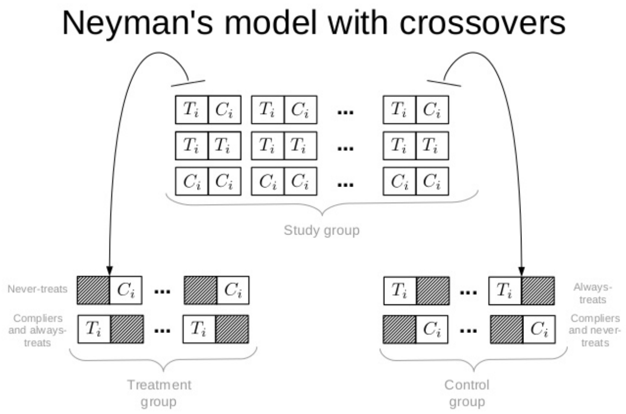


Figure 22: Non-compliance

To estimate p_c, we can observe the rate of accepting assignment in treatment and control group and get the proportion by subtraction.

2.3 Regression Discontinuities

Sometimes in experiments, an arbitrary threshold would have an effect in the regression output.

In class we are given an example of Yelp rating. The rounded value of rating presented to users caused a jump in regression result around the rounding threshold.