Causal Inference: Randomized Experiments

Dr. Ria Ivandić

24th January 2022

Last week

- You learned why causal inference with observational data is hard
- You learned the differences between correlation and causation (and the role of confounders)
- You learned what a counterfactual is and became more comfortable with Potential Outcomes notation
- You were introduced to the concept of ATE
- You learned how observed mean differences across treated and control groups can contain selection bias

This week

- Recap of potential outcomes framework and randomization
- How can randomization solve the selection bias issue
- Important challenges to RCTs
    1. Power
    2. Design
    3. Non-Compliance
- Exercise: RQ: Does information increase turnout?

Let's begin with a recap!

Real world selection bias: health insurance

- Health insurance in the US used to be voluntary. Obama's "Affordable Care Act" required citizens to buy insurance. Trump rescinded the mandate in 2019. Is mandatory health insurance a good policy? Similarly, is the NHS in the UK a good policy?
- **Question: What is the effect of health insurance on expenditures and health outcomes?**
- The US spends a larger fraction of GDP on health care than most other developed countries, but Americans are not necessarily healthier. Health insurance supposedly has health benefits. Yet, heath care is expensive and requiring it may be considered to violate freedom.

What is the Effect of Health Insurance on Health?

We want to compare:

- The health of someone with insurance.
- The health **the same person** without insurance.

We begin by comparing health outcomes between those with and without insurance.
Should we do this? Are there any potential issues?

What is the Effect of Health Insurance on Health?

	Some Health Insurance	No Health Insurance	Difference
	A. Health		
Health Index	4.01	3.70	0.31 (0.03)
	B. Characteristics		
Non-White	0.16	0.17	-0.01 (0.01)
Age	43.98	41.26	2.71 (0.29)
Education	14.31	11.56	2.74 (0.10)
Employed	0.92	0.85	0.07 (0.01)
Family Income	106,467	45,656	60,810 (1,355)

The observed difference in the outcome for the treatment and control group are:

$$E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=0] = \underbrace{E[Y_{1i}-Y_{0i}|D_i=1]}_{ATT} + \underbrace{E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]}_{SelectionBias}$$

What is the Effect of Health Insurance on Health?

What can cause a difference in health outcomes for individuals with and without health insurance?

What is the Effect of Health Insurance on Health?

What can cause a difference in health outcomes for individuals with and without health insurance?

1. Causation: having health insurance directly leads to better health.
2. Reverse causality: the less (or more) healthy are more likely to buy insurance.
3. Confounders: e.g., the more educated tend to buy insurance more often and they know how to live healthier.
4. Any other ideas?

What is the Effect of Health Insurance on Health? Selection bias

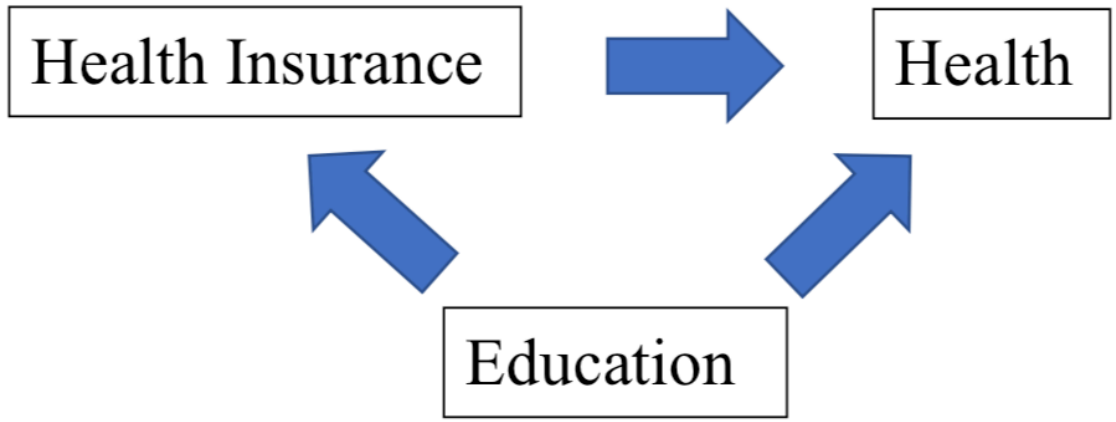
The observed difference in the outcome for the treatment and control group are:

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{SelectionBias}$$

Selection bias is the difference in the potential untreated outcomes between those who were treated and who were not treated. Can we approximate the selection bias here? Is it positive or negative?

$$\underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{SelectionBias}$$

What is the Effect of Health Insurance on Health? Selection bias



Selection bias is when treatment is assigned in a manner that also affects the outcome. In our example confounders, e.g., education levels, may affect health. Education may also affect the choice to attain insurance. Thus, **potential outcomes differ for individuals with and without insurance.**

From observed mean differences to ATT

We would like to turn this:

$$\underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{SelectionBias}$$

Into this:

$$\underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{0}_{SelectionBias}$$

This means that:

$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$$

This equality would make $ATT = ATE$. What assumptions do we need to claim this equality and how do we get there?

From observed mean differences to the causal effect

The causal effect for any given subject is not directly observable. However, experiments provide unbiased estimates of the average treatment effect (ATE) among all subject when certain conditions are met. These **three assumptions** (Gerber & Green, 2012) are:

- **Random Assignment:** implies that assignment to treatment is statistically independent of the subjects' potential outcomes (observable or unobservable characteristics)

From observed mean differences to the causal effect

- **Excludability:** potential outcomes solely respond to the treatment, not the random assignment to the treatment or any by-products of random assignment. This can be violated when i) different procedures are used to measure outcomes in the treatment and control groups ii) other third-party interventions differentially affect the treatment and control groups. Examples.
- **Non-interference or Stable Unit Treatment Value Assumption (SUTVA)** - the value of the potential outcome for unit i depends only upon whether that unit did or did not receive the treatment. This can be violated when i) subjects are aware of the treatments other subjects receive ii) treatments can be transmitted from treated to untreated subjects iii) resources used to treat one set of subjects diminish resources that would be available to other subjects. Examples

Randomization

Randomization and Selection Bias

With randomly assigned D_i there is no Selection Bias! Units i are similar on all (un)observed traits and only differ in terms of D_i.

Because both conditional expectations $E[Y_{1i}|D = 1]$ and $E[Y_{0i}|D = 0]$ come from the same underlying population, we can claim that when D_i is randomly assigned, the units are interchangeable.

Is randomization always possible? Motivation for causal inference.

Randomization

Why Randomization?

Random Assignment addresses the counterfactual problem by creating two subsamples that are identical prior to the intervention. When treatments are randomly assigned, $D_i = 1$ is a random sample of all units i and thus the potential outcomes for the treated units are identical to the potential outcomes of all units. The same for units in $D_i = 0$.

In expectation,
$E[Y_{1i}|D_i = 1] = E[Y_{1i}]$ but also $E[Y_{0i}|D_i = 0] = E[Y_{0i}]$

Now because the treatment does not affect the potential outcomes (recall the independence assumption), we are able to say that:

$$ATE = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Randomization and Causal Estimates

If D_i was assigned at random; this would mean that:

1. $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$
2. $E[ATT] = ATE$.
3. $DIGM = ATT$

With randomly assigned treatments there is No Selection Bias!

So revisiting the difference in group means (DIGM) :

$$\underbrace{E[Y_{i1} - Y_{i0}|D_i = 1]}_{ATT} + \underbrace{E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]}_{SelectionBias}$$

Is equal to:

$$\underbrace{E[Y_{i1} - Y_{i0}|D_i = 1]}_{ATT} + \underbrace{0}_{SelectionBias}$$

Randomization and the Experimental Ideal

You might be wondering...

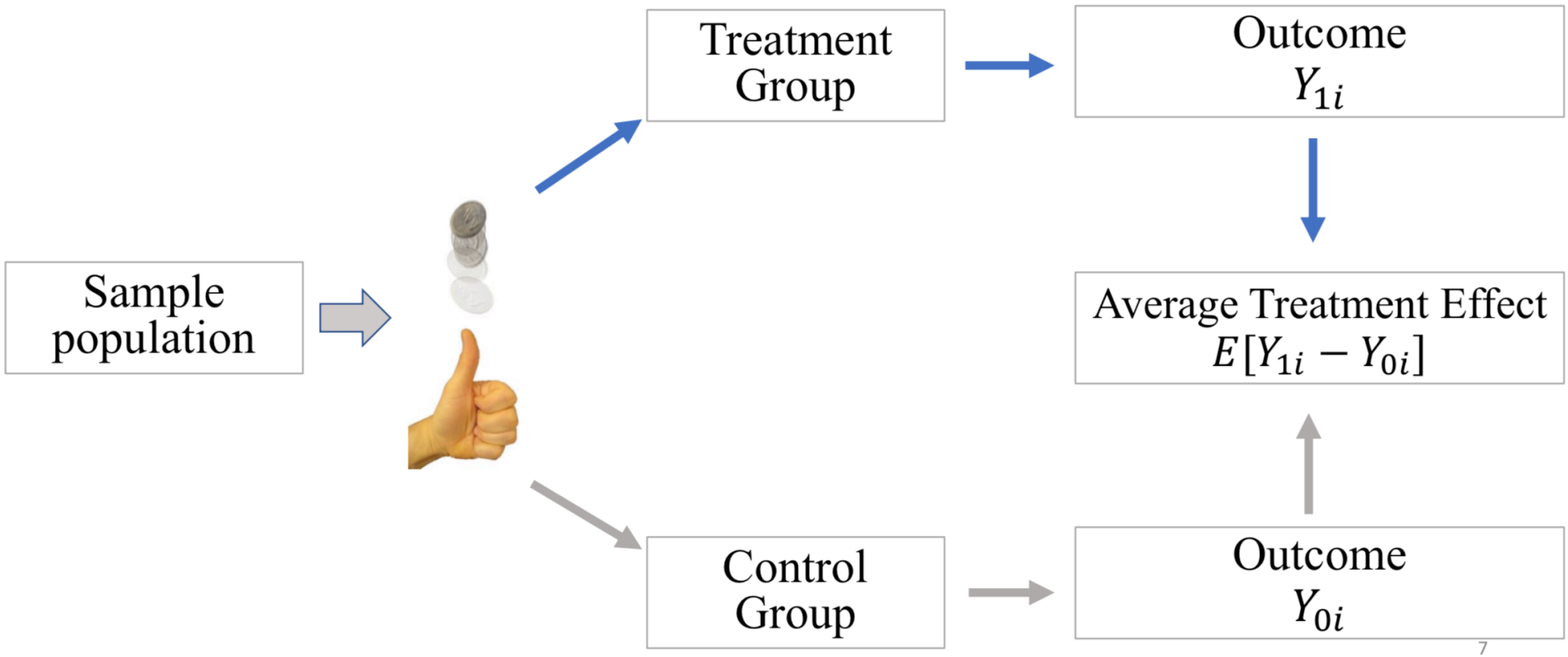
- Will I have to run an experiment to find causal estimates?
- Can I assign my units to treatment and control?
- Am I wasting my time learning about Causal Inference?

Causal Inference is not only about RCTs! Causal Inference is about the experimental ideal!

The aim is in most cases the same; we try to find interchangeable units (counterfactuals) that only differ in terms of their treatment status.

Always think about how to **approximate the experimental setting** (random assignment to treatment) by using observational data.

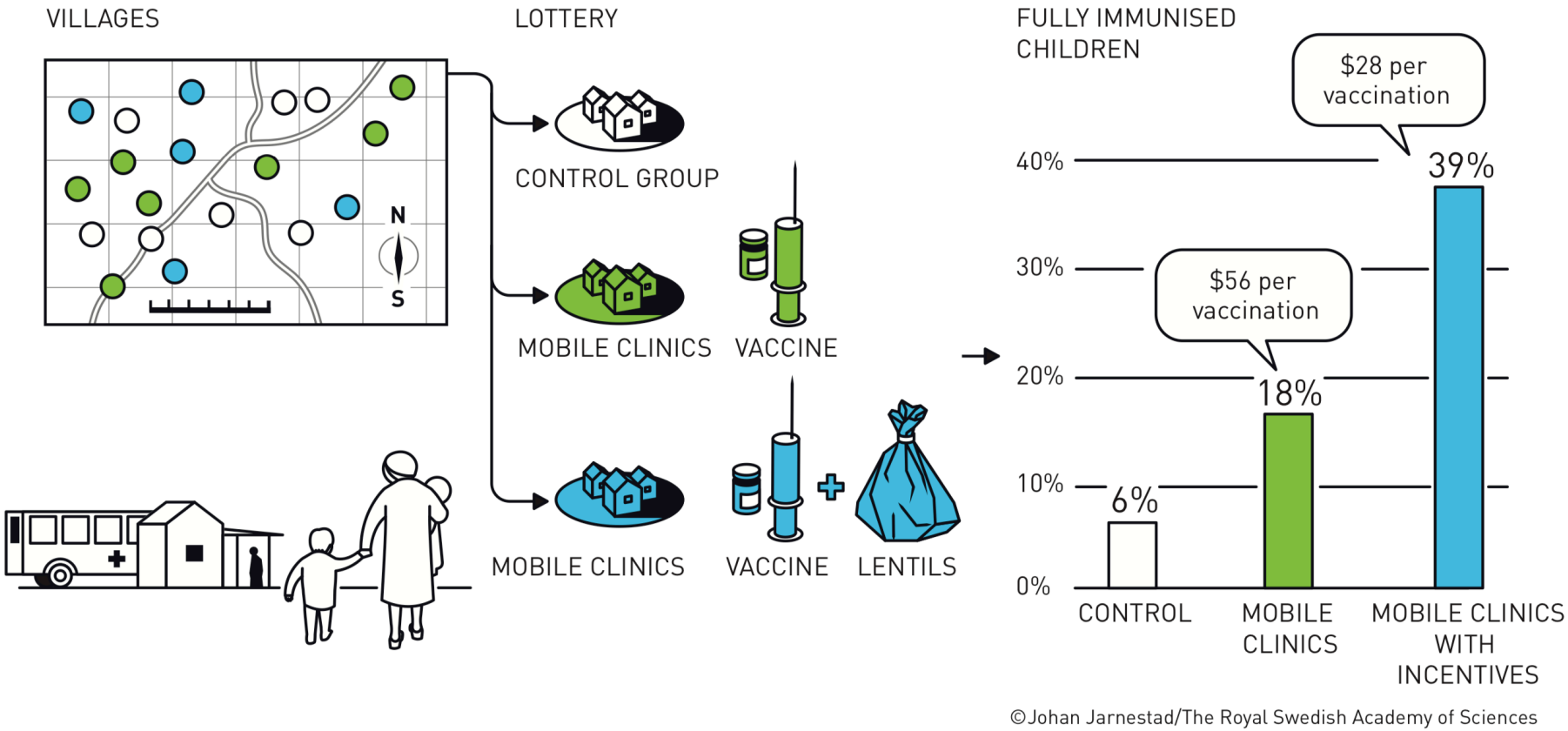
The Mechanics of Experiments



The Power of Experiments

- The Prize in Economic Sciences 2019 was awarded to Abhijit Banerjee, Esther Duflo and Michael Kremer "for their experimental approach to alleviating global poverty"
- They have shown that these smaller, more precise, questions are often best answered via carefully designed experiments among the people who are most affected.
- As a direct result of one of their studies, more than five million Indian children have benefitted from effective programmes of remedial tutoring in schools.
- Another example is the heavy subsidies for preventive healthcare that have been introduced in many countries.

The Power of Experiments



Roadmap

1. Translating your research question to a treatment and measurable outcome
2. Randomization of the treatment
3. Power of the experiment
4. Validity of the design
5. Correct treatment and mechanisms
6. Non-Compliance (in field experiments)

Randomizing the Treatment

The results we have seen showing no selection bias hinge on correctly executed **random assignment**:

- **Simple random assignment:** a procedure, such as a dice roll or coin toss, that gives each subject an identical probability of being assigned to treatment. Disadvantages when small N?
- **Complete random assignment**: m of N units are assigned to the treatment group with equal probability (one practical way to do it is permuting the order of all N subjects and label the first m subjects as the treatment group)
- **Stratified randomization or block randomization:** Suppose we observe some covariate x_j, and we know that the outcome y_i varies with x_j Then any difference in the covariate between the treatment groups will lead to a difference in the average outcome, unrelated to the actual treatment effect. This issue can be prevented by **balancing treatment assignment on these covariates**. It means partitioning the sample into groups (blocks) of different x_j, and then carrying out permutation randomization within each block.
- In practical terms, randomization is best done using statistical software like R (Randomizr)

Correct Randomization?

- Important: randomization ensures that all pre-treatment covariates (observable, unobservable) are balanced in expectation (across randomizations).
- Selection bias can't be tested (why not?), but **balance tests are often used to check for differences in observable pre-treatment covariates between the treatment and control groups**:
  - covariate-by-covariate comparison of means (e.g. via t-tests)
  - multivariate regression of treatment status (DV) on all covariates
- Balance tests useful to detect botched randomization, but can never say anything about balance in unobserved covariates, so "potentially misleading" (Imai, King, and Stuart 2008)
- What covariates would you check if you are running an experiment where you randomize free university tuition on graduating university?

Can Experiments Uncover Social Relationships?

The challenges of experimentation

1. Power
   - Noise (Y_i)
   - Specificity (D_i)
   - Sample Size
2. Validity
   - Internal
   - External
3. Mechanisms (?)

Design of experiments: considerations

1. Designing an experiment one often picks n (sample size). How to choose?
   - **Reality**: If you have £X and the cost per subject is c, $n = X/c$
   - **Ideal Power analysis: choose n such that in 80% of randomizations you would reject the null hypothesis at 95% confidence level, given assumptions about size of ATE and variance of outcomes in treatment and control group. (See DeclareDesign package for R.)**
2. Types of treatments:
   - Baseline case: $k \geq 2$ distinct alternatives, one per unit/subject (between-subjects design)
   - In some survey and lab experiments, multiple treatments assigned in sequence (within-subjects comparisons)

Power in Randomized Experiments

- Running experiments is expensive, so we do not want our work to be ex-ante futile (an effect that could never be statistically significant)
- Power Analysis involves guess work and strong assumptions
  - **The main idea is the following; by making an educated guess about the size of the effect (ATE) and the variation of the estimates, we are able to calculate a sample size for each treatment group that would give us power to correctly reject the null around -say- 8 out of 10 times.**
  - In other words, we seek to calculate the necessary sample size that would give us confidence that we have enough power to retrieve the ATE we hypothesized
- *Note:* When calculating a mean of a population from a sample, extreme observations will have large effects on our estimate if the N is small, and a negligible effect if the N is large.

Power in Randomized Experiments

The ingredients

1. Power (measured in Z): Most researchers (and research councils) opt for a power of 0.8 (i.e. reject the null correctly 8 out of 10 times)
2. α level (measured in Z): Our typical level of statistical significance for the ATE (we reject the null if $p - value \leq 0.05$)
3. standardized treatment d: This is our expected ATE standardized with the pooled standard deviations.

$$d = \frac{ATE}{\frac{\sqrt{\sigma^2_{D_{i1}} + \sigma^2_{D_{i0}}}}{2}}$$

Formula for two sample t-test Power analysis

$$N = 2 \times \frac{(Z_\alpha + Z_{power})^2}{d^2}$$

Power: An Example

Let's consider a small effect (btw, small=0.2, medium=0.5 large=0.8) with a power of 0.8 and 95% significance level

```
library(pwr)
pwr.t.test(n = , d =0.2 ,
sig.level = 0.05 ,
power = 0.8, type ="two.sample")

Two-sample t test power calculation

              n = 393.4057
              d = 0.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Power: An Example

```
z1=qnorm(0.975)
z2=qnorm(0.80)
d=0.2

samplesize= 2*(((z1+z2)^2)/d^2)

> samplesize
[1] 392.444
```

Power in Randomized Experiments

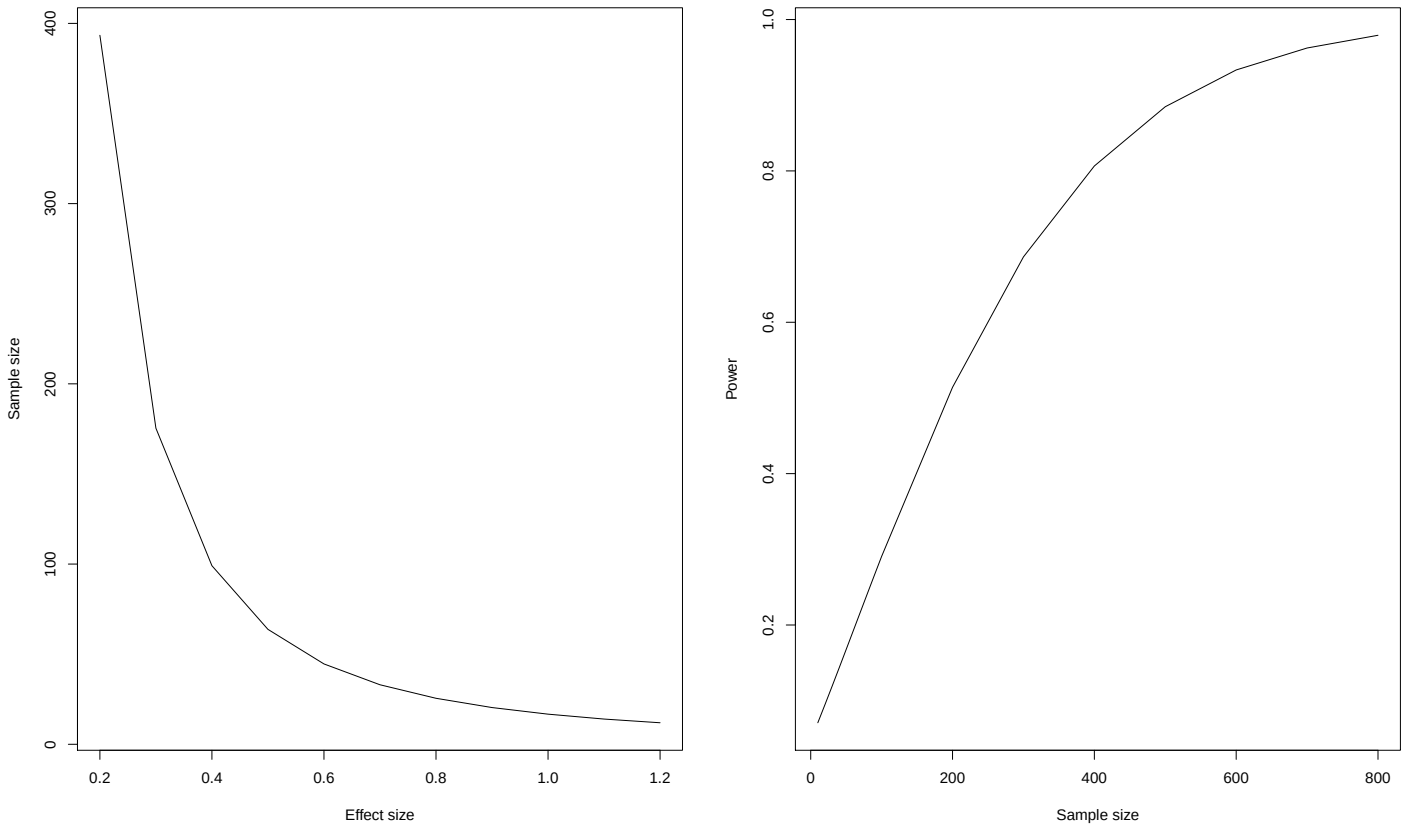


Figure: Power Analysis Simulation L: N and d, R: Power and N

Key Challenges in Power Analysis

Some thoughts...

- Note that the formula to be used depends on the type of outcome you are measuring (i.e. it is different for proportions)
- As it was mentioned it is not easy to know the ATE or the variances. We can use information from past experiments or with small N pilots
- Although it is still possible to assume a medium effect, the whole exercise requires too many assumptions.
- Yet, it is also good practice; it makes you think about the treatment you are using and the measurement of your outcome.

Improving Experiments and Power

- The more precise the measurement of the outcome and the stronger the treatment, the stronger the power of the experiment

Model Specification in Randomized Control Trials (RCTs)

This is the clear advantage of randomized control trial!

If the RCTs has a simple (one treatment and one control group, complete randomization) then the difference in group means estimator will be unbiased (can be interpreted causally as there is no selection bias).

This is the equivalent of running a regression:

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

More complicated experimental research designs will be a variation of the regression above.

Roadmap

1. Translating your research question to a treatment and measurable outcome ✓
2. Randomization of the treatment ✓
3. Power of the experiment ✓
4. Validity of the design
5. Correct treatment and mechanisms
6. Non-Compliance (in field experiments)

Validity

Two types

1. Internal
2. External

Internal Validity

An experiment is internally valid when a certain factor (say X) has a causal effect on Y can be identified by the experimental process. It is high when estimate of ATE is expected to be close to the true ATE.

External Validity

An experiment is externally valid when the causal effect is generalizable to different contexts or/and the wider population

Are all randomized experiments the same?

Typology

- Lab experiments: lab setting, contrived interventions and outcomes (but can be incentivized with money)
- Survey Experiments: survey setting, "weak" interventions (information, priming; forced choice among different alternatives), survey outcomes.
- Field Experiments: natural setting, real interventions, real outcomes.

All types of experiments have something to offer. At the same time, they all have serious constraints.

Are all randomized experiments the same?

Typology

- Lab experiments high Internal/low External Validity
- Survey Experiments good Internal/good External Validity
- Field Experiments good Internal/high External Validity

All types of experiments have something to offer. At the same time, they all have serious constraints.

Good Research Design Matters!

- A survey experiment for example can be really bad in terms of both internal and external validity
- A lab experiment can be realistic in some contexts
- Field experiments are expensive and hard to measure the outcomes (in most cases).
- **The bottom line is that all these assessments about validity are conditional on the quality (and the context) of the experiment.**

Are we manipulating the hypothesized cause?

With experiments, we use a treatment that supposedly captures the causal effect of one variable on the outcome

1. What if the treatment can capture more than one variables/causes?
2. How can we know if we captured the intended variable/cause?

Compound treatment effects are very common. E.g. researchers are interested in the micro foundations of democratic peace theory. They treat subjects with information about **democracies or autocracies** and they unintentionally prime units to think about the **wealth** of a given nation (democracies are on average wealthier than autocracies) or how much **press freedom** there is. Was this the intention of the experimenter? What are implications for internal and external validity?

This is a key question to ask yourself when designing an experiment.

A check-list for population-based experiments

1. Measure pretreatment covariates. Make sure you don't measure variables that resemble the treatment Why?
2. Spend time thinking about measurement of the outcome. Single items might contain measurement error.
3. Spend equal amounts of time thinking about the treatment
   - 3.1 Are we measuring the underlying concept?
   - 3.2 Is it easy for respondents to comprehend the treatment?
   - 3.3 Are we deceiving them?
   - 3.4 Is there a chance we are measuring more than one concepts?
4. Crucial: Add manipulation checks to test what is being captured by the treatment

A Challenge in Field Experimentation

- In field experiments we have little control over our units/subjects
- Often times, we assign units to the treatment group, but they are not eventually treated

We call this Non-Compliance

- Often times, our inability to treat units can correlate with the outcome or the treatment
- This can have serious consequences for the estimation of the treatment effects

Typical Example: Canvassing and Turnout!

Naive Solutions to Non-Compliance

Naive Solutions

1. Discard those not treated although assigned to the treatment group: compliance is often self-selected
2. Ignore random assignment: randomization is broken

- *In other types of experiments we are certain that our units were exposed to the treatment (treatment status=treatment assignment, thus mean differences [i.e. the ATE] are sufficient)*

Solutions to Non-Compliance: Intention to treat

- One less naive solution is to preserve randomization and examine the effect of treatment assignment (not status) on the outcome.
- It is, in any case, an interesting quantity of interest (think of an employment program)
- We will revisit this in more detail later in the term (including assumptions and regression based techniques to retrieve uncertainty around the CACE estimates (2SLS))

Intent-to-Treat

$$ITT = E[Y_{1i}] - E[Y_{0i}]$$

Example: Field Experiment looking at the Effects of Canvassing on Turnout

- Research by Gerber and Green (2000): The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experimen
- During the weeks leading up to the 1998 general elections, they assigned 7,090 single person households (registered voters) in New Haven to treatment and control groups
- Treatment group: visited by canvassers who stressed the importance of voting
- Control group: no contacts from the campaign Does this mean no contacts overall?
- Outcome: public records on which registered voters turned out to vote
- Did everyone open the door?

Calculating Effects in Field Experiments

Complier Average Causal Effect

$$CACE = \frac{ITT}{E[D_{1i}]}$$

i.e. the ATE among compliers

Example from GG(2012), pp 150-1:

	T	C
Turnout by those contacted	54.43 (395)	
Turnout among those not contacted	36.48 (1,050)	37.54 (5,645)
Overall Turnout	41.38 (1,445)	37.54 (5,645)

Treatment Status (the CACE denominator):

$$D_{1i} = \frac{395}{1445} - \frac{0}{5,645} = 0.273$$

i.e. 27,3% actually treated from those assigned to treatment

Calculating Effects in Field Experiments

	T	C
Turnout by those contacted	54.43(395)	
Turnout among those not contacted	36.48(1,050)	37.54 (5,645)
Overall Turnout	41.38 (1,445)	37.54 (5,645)

Treatment Status (the CACE denominator):

$$D_{i1} = \frac{395}{1445} - \frac{0}{5,645} = 0.273$$

The ITT (i.e. the CACE numerator) can be calculated as

$$41.38 - 37.54 = 3.84\%$$

and the effect among compliers as

$$CACE = \frac{3.84}{0.273} = 14.1\%$$

Interesting papers using experiments

- Olken, 2010: Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia: This article presents an experiment in which 49 Indonesian villages were randomly assigned to choose development projects through either representative-based meetings or direct election-based plebiscites
- Kalla amd Broockman, 2020: Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments: In these experiments, 230 canvassers conversed with 6,869 voters across 7 US locations. In Experiment 1, face-to-face conversations deploying arguments alone had no effects on voters' exclusionary immigration policy or prejudicial attitudes, but otherwise identical conversations also including the non-judgmental exchange of narratives durably reduced exclusionary attitudes for at least four months (d = 0.08).