

Introduction to the Design and Analysis of Randomized Experiments

Class 1: Experiments, Potential Outcomes, and Treatment Effects

Jake Bowers

July 28, 2024

Overview and Review

Introductions

Experiments in context: Research questions, theories, and research designs

Why experiment?

Experiments and the Counterfactual Approach to Causal Inference

Random assignment vs Random sampling

Key assumptions for randomized experiments

Estimation, Estimators, Bias, Consistency, Given Randomization

Summary and Overview

Appendix

```
## here() starts at /Users/jwbowers/repos/CLASSES/short_course_experi

## Loading required package: knitr

## -- Attaching core tidyverse packages -----
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr    1.5.1
## v ggplot2     3.5.1      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.1
## v purrr       1.0.2
## -- Conflicts -----
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to f
## Loading required package: randomizr
##
## Loading required package: fabricatr
##
## Loading required package: estimatr
##
##
## Attaching package: 'DeclareDesign'
##
```

Overview and Review

Overview of the class I

At the end be able to answer questions like:

- **Why choose a randomized experiment?**
- **What is a counterfactual quantity and why might we care about this?**
- How can we randomize using R?
- **Why test? Why estimate? How test and estimate? What is randomization based inference and why should we care?** How can we use R to test hypotheses about counterfactual quantities? How can we use R to estimate average causal effects (differences of averages of counterfactual quantities)?
- How might we know when it is worth randomizing and when it is not? (Statistical power) **What is statistical power? Why should we care?**

Overview of the class I

- How should we make choices about certain common designs:
 - ▶ Why might one randomize within strata? What does this imply for analysis and interpretation?
 - ▶ Why might one randomize an intervention to groups or clusters of units? What does this imply for analysis and interpretation?
 - ▶ How might we use randomization to learn about the effects of interventions even when we cannot control who receives a full dose of the treatment? (LATE/CACE)

Overall: This class is for you. Please ask questions!

Today

1. Introductions and our projects (An experiment you are planning, an experiment you'd like to do, an experiment you are doing, an experiment you have done.)
2. Why experiment?
3. Causal inference in randomized experiments and the idea of only partially observed **potential outcomes**. The idea that we can use what we observe to learn about what we cannot observe.
4. Statistical inference for causal effects in randomized experiments via the Fisher and Neyman approaches (Rosenbaum, 2010, Chap 2), (Gerber and Green, 2012, Chap 1-3): Estimation, Estimators, Tests, Testing.

Note: You can download (and contribute to) course materials at
https://github.com/bowers-illinois-edu/short_course_experiments

Hay recursos en español aquí:

https://egap.github.io/theory_and_practice_of_field_experiments_spanish/

Overly Ambitious Plan for the day

- Introduction to Jake
- Introduction to the idea of the course: roadmap
- Introductions to you and a project that you might work on during the week. (See perhaps <https://egap.github.io/learningdays-resources/Exercises/design-form.html> for help if you want to develop a project this week.)
- Jake introduces concepts: experiments, potential outcomes, and addresses questions from the class.
- Coffee Break
- Questions about the lecture
- Exercise 1: How big of an experiment should we have to feel comfortable about covariate balance?
- Break
- Lecture
- Open Discussion and/or work on projects with small group/individual consultations

Introductions

Some Experiments

My main experience is with policy experiments.

- <https://oes.gsa.gov>
- <https://thepolicylab.brown.edu>
- <https://egap.org>
- <https://thelabprojects.dc.gov>

Class introductions

- Name
- Where are you currently studying/working?
- Possible/Existing randomized experiment?

Experiments in context: Research questions, theories, and research designs

What makes a research question good?

- The answer to a good research question should produce knowledge that people will care about, that will change beliefs, that will inspire action.
- Addressing the question should (help) solve a problem, (help) make a decision, or clarify/challenge our understanding of/explanations about the world.
- That is, a good question arises in the context of a theory and in the context of values.
- But an interesting question is not enough to make the change in the world that we'd like to see

We also need a good research design

- A good research design is a practical plan for research that makes the best use of available resources and produces a credible answer.
- The quality of a research design can be assessed by how well it produces results that can be used to guide policy and improve science:
 - ▶ A great research design produces results that clearly point in certain directions that we care about.
 - ▶ A poor research design produces results that leave us in the dark — results with confusing interpretation, ambiguity in interpretation.
- The point of most social and behavioral science experimental research design is to learn about theory not about the world per se. (This point is often not appreciated in policy experiments. The confusion arises often in discussions of “external validity” or “generalizability”.)

The importance of theory I

All research design involves theory, whether implicit or explicit.

- **Why do the research?** We have implicit theories and values which guide the questions we ask. Our questions are value laden: For example, social scientists studied marijuana use in the 1950s as a form of “deviance”, the questions focused on “why are people making such bad decisions?” or “how can policy makers prevent marijuana use?” (see Becker, [1998](#)).
- **Why do the research?** We might want to change how scientists explain the world and/or change the policy decisions in (a) one place and time and/or (b) in other places and times.
- Research focused on learning the causal effect of an intervention, X , on an outcome, Y , requires a model of the world: *how* might intervention X might have an effect on some outcome Y , and *why*, and *how large* might be the effect? It helps us think about how a different intervention or targeting different recipients might lead to different results.

The importance of theory II

- Designing research will often clarify where we are less certain about our theories. Our theories will point to problems with our design. And questions arising from the process of design may indicate a need for more work on explanation and mechanism.

Designing or selecting your treatment

- Your treatment and control need to clearly connect to your research question.
- The treatment you're interested in might be a bundle of multiple components. If your research question is about one specific component, then the control should be different from the treatment in just that component. Everything else should be the same.

An example

A campaign where someone visits a home to talk with a family for 15 minutes to share health information.

- If you're interested in the effect of the specific information, then your control should still have all the other components (home visit with 15 minutes duration, similar visitor, etc.) but have different information. This design will not teach you about the effect of visits, just about the effect of information.
- If your question focuses on the effect of visits, then you need a control group without a visit. But this design will not do a good job answering specific questions about information (visits and information are bundled together).

Interpretation

- Sometimes it's not possible to separate out a specific component of your treatment.
- For example, your partner community health organization that visits homes may not be interested in visiting homes and sharing non-health information. Then your control might be no visit.
- You must be careful to interpret your effects as the effect of the information delivered in this particular way.
- You will not be able to conclude that you have estimated the effect of only the information.
 - ▶ This might be fine for certain policy purposes: maybe the policy question is about the visits as an implicit bundle of treatments.
 - ▶ But it is difficult to interpret the results of this design as telling us something clear about information alone.

The Research Process: Questions, theory, and credibility

- Research starts with our values and theories about how the world works.
- It continues by articulating questions that can be clearly addressed by observation (in this course, using randomized experimentation).
- Good questions have consequential answers: changing scientific explanations, changing policy decisions.
- Good designs tick all the boxes and give readers reason to believe the results.
- Not all randomized experiments are good designs. And not all good designs are randomized experiments.

Why experiment?

Policy experiments

A method for:

- Putting beliefs about what works to the test
- Enabling policy decisions based on data

An example experiment (Wantchekon et al)

- Question: are programmatic policies or clientelistic policies more effective at mobilizing voters?
- Why we care: Programmatic policies can be more equitable, pro-poor
- Idea: politicians believe clientelistic policies are more effective – and maybe they would change campaigns if they knew they were wrong
- Research partner: four main parties in Benin
- Intervention: programmatic policy or clientelistic policy promoted by party at election rallies in a district
- Experimental design: campaign randomly assigned to districts
- Outcome: vote share for the party
- Results: on average voters prefer clientelistic campaigns, but women more likely to prefer programmatic policies

An example experiment (Wantchekon et al)

Random assignment → **highly credible evidence** and **easy to interpret evidence**
that the change to messages at rallies led to a change in vote share

If politicians believe results → decide to change campaigns

Long history of the method

- Late 1700s, early 1800s: early double-blind experiments with comparison group (not randomized)
- 1920s: first randomized experiments in agriculture, education, and political science
- 1965: first clinical randomized trial (Streptomycin for TB)
- 2019: Nobel prize for popularizing randomized experiments in economics

Widespread use

- Clinical trials required by regulators for vaccines, medicines, and treatments
- Political campaigns test mobilization and persuasion strategies
- Tech companies test website features to find most lucrative ones
- Governments test policies using RCTs

Opposing existing beliefs I (EGAP Accountability Metaketa, Dunning et al)

- Belief: providing information about incumbent politicians will change votes
- Intervention: information about politicians' corruption, job attendance
- Experimental design: randomly assign districts to receive information or not
- Outcome: vote share for incumbent (administrative data)
- Result: no effect

Opposing existing beliefs II

- Belief: local democratic institutions improve delivery of development aid
- Partner: International Rescue Committee, CARE
- Intervention: two years of democratic institutions and development program
- Experimental design: randomly assign village clusters to treatment or not
- Outcomes: corruption, government practices chosen after intervention
- Result: no change

Opposing existing beliefs III (EGAP Community Policing Metaketa)

- Belief: engagement between police & citizens improves trust, lowers crime
- Partner: Police agencies in six countries
- Intervention: “community policing” (townhalls, beat patrols, etc.)
- Experimental design: randomly assign police beats to receive/not
- Outcomes: citizen trust in and cooperation with police, crime
- Result: no change in any main outcome

Supporting existing theories beliefs

- Belief: conditional cash transfers change behavior, improve welfare
- Partner: Mexico public health ministry
- Intervention: cash to mothers conditioned on children attending school, going to health clinics
- Experimental design: first phase of rollout randomly assigned to communities
- Outcomes: poverty, school attendance, health
- Result: improvements in every outcome

Incomplete list of places with CCTs following this Progresa study:

Argentina, Bangladesh, Brazil, Cambodia, Chile, Colombia, Egypt, Guatemala, Honduras, Indonesia, Jamaica, Mexico, Nicaragua, Panama, Peru, Philippines, Turkey, US

Limitations to experiments

Some questions don't need an experiment

- Does smoking cause poor health? (This took many years of observational studies including laboratory and epidemiology studies)
- Do parachutes help when you jump out of a plane? (Back to theory ...)

Limitations to experiments

Some questions shouldn't have an experiment

- Does changing the interest rate affect inflation?
- Were the funds from the American Rescue Plan distributed equitably?
- What is the poverty rate?
- Why didn't information affect accountability? Why didn't community policing change trust?

Important role for descriptive research, theory generation, qualitative investigation

Limitations to experiments

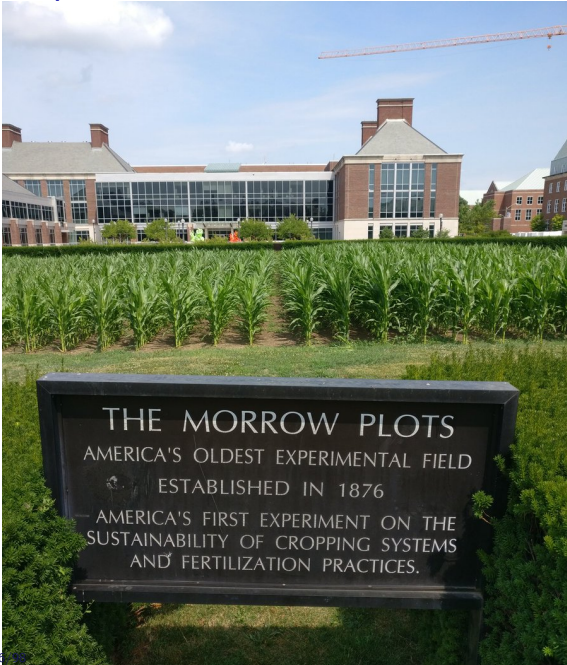
- Ethics — is this sort of manipulation ethical? Sometimes not.
- Must be done in real time, ahead of the intervention roll-out. (Can't run a new randomized experiment on data that has already been collected.)
- Reduced flexibility for a partner organization (problem for any prospective evaluation).
- Maybe Cost if it implies more administration and/or new measurement.

Why experiments

- Put beliefs about effects of policies to the test.
- Offer clear evidence for and/or against existing theories.

Experiments and the Counterfactual Approach to Causal Inference

Experiments and Counterfactual Causal Inference



Did the fertilizer cause more corn to grow?

Each plot either receives the fertilizer or status quo.

What does it mean for fertilizer to **cause** more corn to grow? Perhaps someone suggests a causal mechanism (fertilizer \rightarrow ... \rightarrow corn)? Perhaps someone suggests another mechanism that does not involve fertilizer (?sun light? ?music?)

How to provide evidence about the causal mechanism? One implication of the theory: if the plot of land had not received fertilizer, then less corn would have grown.

“X causes Y” is a claim about what didn't happen

- In the counterfactual approach: “If X had not occurred, then Y would not have occurred.”
- Experiments help us learn about counterfactual and manipulation-based claims about causation.
- It's not wrong to *conceptualize* “cause” in another way (for example, whenever I see fertilized fields, I expect more corn.). But the counterfactual framework has made discussion of the design and analysis of experiments easier (Henry E Brady, 2008a).
- For example, we can have mechanistic explanation linking fertilizer to corn growth (fertilizer does something to the corn cells to trigger increased growth) but also a covering law theory (without nutrients, corn cannot grow). The counterfactual idea helps us relate specific implications to specific observations.

How to interpret “X causes Y” in this approach

1. “X causes Y” need not imply that other variables W and V do not cause Y: X is a part of the story, not the whole story. (The whole story is not necessary in order to learn about whether X causes Y). Music could be a part of the cause. Or sunlight.
2. “X causes Y” requires a **context**: matches cause flame but require oxygen; small classrooms improve test scores but require experienced teachers and funding (Cartwright and Hardie, [2012](#)); corn requires soil and oxygen and reasonable temperatures.
3. “X causes Y” can mean “With X, the probability of Y is higher than would be without X.” or “Without X there is no Y.” Either is compatible with the counterfactual idea.
4. It is not necessary to know the mechanism to establish that X causes Y. The mechanism can be complex, and it can involve probability: X causes Y sometimes because of A and sometimes because of B. (In fact, in science we often have “X causes Y” and then decades of research work to narrow down the causal mechanism.)

Exercise: Colds and Honey

- Your friend says drinking honey water reduces the duration of colds.
- If we take a counterfactual approach, what does this statement implicitly claim about the counterfactual?
- If we saw that people who drank honey water also tended to have shorter colds, would that confirm the causal claim? What other counterfactuals might be possible and why?

Potential outcomes notation for counterfactual causal effects

- Given two possible treatments, for each unit we formalize the simple version of the counterfactual idea by writing that there are two **post-treatment** outcomes: $Y_i(1)$ and $Y_i(0)$.
- $Y_i(1)$ is the outcome that **would** obtain *if* the unit received the treatment ($T_i = 1$).
- $Y_i(0)$ is the outcome that **would** obtain *if* the unit received the treatment ($T_i = 0$).

A unit can be a plot of land at a moment in time, a village, a person, etc.

Definition of causal effect

- An additive **causal effect** of treatment (relative to control) is:
$$\tau_i = Y_i(1) - Y_i(0)$$
- We could also write $\tau_i = Y_i(1)/Y_i(0)$ — how much did person i earn after the training program as a ratio?
- Or other functions relating the potential outcomes.

Key features of this definition of causal effect

1. You have to define two conditions to define a causal effect.
 - ▶ Say $T = 1$ means a community meeting to discuss public health. Is $T = 0$ no meeting at all? Is $T = 0$ a community meeting on a different subject? Is $T = 0$ a flyer on public health?
 - ▶ The phrase “causal effect of T on Y ” doesn’t make sense in counterfactual terms without knowing what it means to not have T (for $T = 0$).
2. Each individual unit i has its own causal effect τ_i .
3. But we can’t measure the individual-level causal effect, because we can’t observe both $Y_i(1)$ and $Y_i(0)$ at the same time. This is known as the **fundamental problem of causal inference**. What we observe is Y_i :

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

Imagine we know both $Y_i(1)$ and $Y_i(0)$ (this is never true!)

i	$Y_i(1)$	$Y_i(0)$	τ_i
Andrei	1	1	0
Bamidele	1	0	1
Claire	0	0	0
Deepal	0	1	-1

- We have the (additive) treatment effect for each individual.
- Note that individual-level treatment effects differ.
- But we only have at most one potential outcome for each individual, which means we don't directly see these treatment effects. We can only **infer** them.

Recap: Notation and Concepts for Counterfactual Causal Inference

- *Treatment* or *Intervention* $T_i = 1$ for treatment and $T_i = 0$ for control for units i . (We mostly assume that all units **could have** $T_i = 1$ or $T_i = 0$. That it is not impossible for any unit to have either value.) (Q: What is a unit? Examples of interventions?)
- Each unit has a pair of *potential outcomes* $(y_{i,T_i=1}, y_{i,T_i=0})$ (also written $(y_{i,1}, y_{i,0})$) (given the Stable Unit Treatment Value Assumption (SUTVA)).
 - ▶ Without the SUTVA assumption, and with 4 units, with two having $T_i = 1$, unit $i = 1$ would have the following potential outcomes:
 $(y_{i,1100}, y_{i,1010}, y_{i,1001}, y_{i,0101}, y_{i,0011})$
- *Causal Effect under SUTVA* when $y_{i,1} \neq y_{i,0}$, $\tau_i = f(y_{i,1}, y_{i,0})$ ex. $\tau_i = y_{i,1} - y_{i,0}$.
- *Fundamental Problem of (Counterfactual) Causality* We only see one potential outcome $Y_i = T_i * y_{i,1} + (1 - T_i)y_{i,0}$ (Examples: Argentina having elected Mieli versus not? Chile under a new constitution versus not?)
- *Covariates*, $\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix}$ is a matrix containing background information about the units that might predict $(y_{i,1}, y_{i,0})$ or Z (but that don't predict Z if Z is randomized as in an experiment).

The observation and unobserved causal comparisons

We can learn about unobserved but theorized causal mechanisms by observing the world (Henry E. Brady, [2008b](#)):

- **Persistent association** “We always/mostly see $Y = 1$ when $X = 1$ and $Y = 0$ when $X = 0$.”
- **Counterfactual Difference** “If X had not been this value, then Y would not have been that value.”
- **Difference after manipulation** “When we change X from one value to another value (and maybe call it T for “Treatment”), then Y changes from one value to another value.” (establishes causal priority of X over Y , implied that Y would not have changed.).
- **Difference after operation of a mechanism** “Once upon a time A changed X , and then one day X changed B , and because of that B changed C , and finally C changed Y .”

All approaches are useful. This week we are focusing on the counterfactual approach.

Randomization in Action: Honey and Colds

Your friend explains a causal mechanism that eating raw honey reduces the duration of colds. What kinds of **alternative** explanations might we come up with for this result?

Imagine these were the underlying potential outcomes with two covariates x_1 and x_2 representing two of those explanations and that $x_1 \rightarrow y_0$ and that $x_2 \nrightarrow y_0$. (Say, x_2 is “reading the newspaper” and x_1 is “going often to the doctor”)

id	x_1	x_2	y_0	y_1	τ
1	1	3	6.25	5.25	-1
2	1	8	10.25	9.25	-1
3	2	8	10.00	8.00	-2
4	3	2	12.25	9.25	-3
5	1	6	12.25	11.25	-1
6	0	6	12.00	12.00	0
7	0	7	9.00	9.00	0
8	0	1	5.00	5.00	0
9	0	8	10.00	10.00	0
10	2	7	10.00	8.00	-2

The true, unobserved, average (additive) causal effect is: -1.

Let us run a randomized experiment and see how we do:

An RCT:

Let's imagine that we randomized honey to 5 of the people in the friendship group. How would we know whether `complete_ra` worked as it should? (First we asked ChatGPT what `complete_ra` does!)

```
library(randomizr)
set.seed(12345)
dat$Z <- complete_ra(N = 10, m = 5)
dat$Y <- with(dat, Z * y1 + (1 - Z) * y0)
```

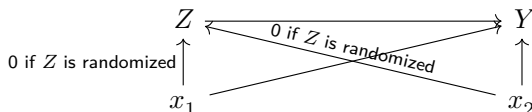

An RCT:

We see the potential outcomes (y_0 and y_1) because this is a simulation. In general, we only see Y and Z and the covariates x_1 and x_2 .

id	x_1	x_2	y_0	y_1	τ	Z	Y
1	1	3	6.25	5.25	-1	0	6.25
2	1	8	10.25	9.25	-1	1	9.25
3	2	8	10.00	8.00	-2	0	10.00
4	3	2	12.25	9.25	-3	0	12.25
5	1	6	12.25	11.25	-1	1	11.25
6	0	6	12.00	12.00	0	1	12.00
7	0	7	9.00	9.00	0	1	9.00
8	0	1	5.00	5.00	0	1	5.00
9	0	8	10.00	10.00	0	0	10.00
10	2	7	10.00	8.00	-2	0	10.00

Assessing randomization I

We expect that the distributions of x_1 and x_2 would be (nearly) the same between the treated and control groups. But how nearly the same? We write “0” below, but in fact, randomization does not make those relationships exactly 0.



Here, just looking at means:

```
dat %>%  
  group_by(Z) %>%  
  reframe(mnx1 = mean(x1), mnx2 = mean(x2))
```

```
# A tibble: 2 x 3  
      Z  mnx1  mnx2  
  <int> <dbl> <dbl>  
1     0   1.6   5.6  
2     1   0.4   5.6
```

Assessing randomization II

```
library(RItools)
bal1 <- balanceTest(Z ~ x1 + x2, data = dat)
bal1$results[, 1:4, ]
```

	stat			
vars	Control	Treatment	std.diff	adj.diff
x1	1.6	0.4	-1.342	-1.2
x2	5.6	5.6	0.000	0.0

Is this what we should expect from a well-conducted experiment? Have we messed up the code?

What **should** we expect from an experiment?

...in regards covariate balance? Lets simulate to learn: (let's explain this code)

```
new_exp <- function(Z) {  
  newZ <- sample(Z)  
  return(newZ)  
}  
  
diff_means <- function(x, Z) {  
  mean(x[Z == 1]) - mean(x[Z == 0])  
}  
  
all_cov_bal <- replicate(10000, diff_means(  
  x = dat$x1,  
  Z = new_exp(dat$Z)  
))  
  
summary(all_cov_bal)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.6000	-0.4000	0.0000	0.0025	0.4000	1.6000

```
obs_cov_bal <- diff_means(dat$x1, dat$Z)  
obs_cov_bal
```

```
[1] -1.2
```

What **should** we expect from an experiment?

So:

1. experiments do not guarantee exact equality of covariates and
2. we can **know** (or closely approximate) what kind of covariate differences a given experimental design would generate.

```
obs_cov_bal
```

```
[1] -1.2
```

```
quantile(all_cov_bal, seq(0, 1, .1))
```

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-1.6	-0.8	-0.4	-0.4	0.0	0.0	0.0	0.4	0.4	0.8	1.6

```
sd(all_cov_bal)
```

```
[1] 0.6687
```

```
mean(all_cov_bal <= obs_cov_bal)
```

```
[1] 0.066
```

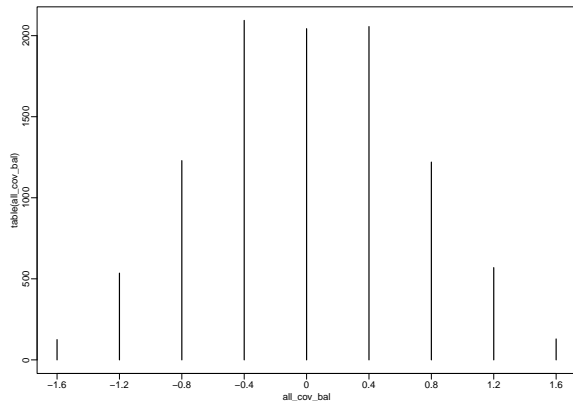
```
mean(all_cov_bal >= obs_cov_bal)
```

```
[1] 0.9875
```

What **should** we expect from an experiment?

1. experiments do not guarantee exact equality of covariates and
2. we can **know** (or closely approximate) what kind of covariate differences a given experimental design would generate.

These are the ways that the difference in mean x_1 can occur in our randomized design.



Maybe an exercise I

This part might be best done interactively using R on your own computers.

If -1.2 is small in statistical terms ($p = .07$), but seems large in substantive terms, what could we do to reduce this? Would a larger experiment help? Let's try using `DeclareDesign` for this — I'd like to see how large of an experiment we'd need to reduce the covariate difference in `x1`.

Maybe an exercise II

This is a function to create a design

```
create_design <- function(N) {  
  set.seed(12345)  
  design <- declare_model(  
    N = N,  
    id = 1:N,  
    x1 = rpois(N, lambda = 0.5) * 2,  
    x2 = sample(1:9, size = N, replace = TRUE),  
    u0 = rpois(N, lambda = 8),  
    u1 = rpois(N, lambda = 8),  
    y0 = x1 + 0.25 * sd(u0) * x1^2 + u0,  
    tau = x1 + u1,  
    y1 = y0 - tau  
  ) +  
  declare_assignment(Z = complete_ra(N = N, m = N / 2)) +  
  declare_inquiry(  
    diff_mean_x1 = mean(x1[Z == 1] - x1[Z == 0]),  
    diff_mean_x2 = mean(x2[Z == 1] - x2[Z == 0])  
  ) +  
  declare_measurement(Y = ifelse(Z == 1, y1, y0)) +  
  declare_estimator(x1 ~ Z, .method = lm_robust, inquiry = "diff_me
```


Using DeclareDesign: make fake data

```
## A simulation is one random assignment of treatment
```

```
sim_dat_N10 <- draw_data(des_N_10)
```

```
## Simulated data (just the first 6 lines)
```

```
head(sim_dat_N10)
```

	ID	id	x1	x2	u0	u1	y0	tau	y1	Z	Y
1	01	1	2	8	7	7	11.33	9	2.331	1	2.331
2	02	2	2	2	9	11	13.33	13	0.331	0	13.331
3	03	3	2	6	8	12	12.33	14	-1.669	1	-1.669
4	04	4	2	6	6	9	10.33	11	-0.669	0	10.331
5	05	5	0	7	8	5	8.00	5	3.000	1	3.000
6	06	6	0	1	10	10	10.00	10	0.000	0	10.000

Using DeclareDesign: repeat the experiment I

```
sim_N10 <- simulate_design(des_N_10, sims = 100)
head(sim_N10)
```

	design	sim_ID	inquiry	estimand	estimator	term	estimate	st
1	des_N_10	1	diff_mean_x1	-1.2	calc_dm_x1	Z	-1.200e+00	
8.485e-01	0.4208	-4.4612	2.061	8	x1			
2	des_N_10	2	diff_mean_x1	0.8	calc_dm_x1	Z	8.000e-	
01	0.8000	1.000e+00	0.3466	-1.0448	2.645	8	x1	
3	des_N_10	3	diff_mean_x1	0.4	calc_dm_x1	Z	4.000e-	
01	1.0954	3.651e-01	0.7245	-2.1261	2.926	8	x1	
4	des_N_10	4	diff_mean_x1	1.2	calc_dm_x1	Z	1.200e+00	
0.7567	3.157	8	x1					
5	des_N_10	5	diff_mean_x1	0.0	calc_dm_x1	Z	4.213e-	
16	0.6928	6.081e-16	1.0000	-1.5976	1.598	8	x1	
6	des_N_10	6	diff_mean_x1	-0.8	calc_dm_x1	Z	-8.000e-	
01	1.0198	-7.845e-01	0.4554	-3.1517	1.552	8	x1	

Using DeclareDesign: repeat the experiment II

```
sim_N10 %>%  
  group_by(inquiry, estimator) %>%  
  reframe(  
    pctlile = c(0, .1, .5, .9, 1),  
    quantiles = round(quantile(estimate, pctlile), 5)  
  ) %>%  
  print(n = 100)
```

A tibble: 10 x 4

	inquiry <chr>	estimator <chr>	pctlile <dbl>	quantiles <dbl>
1	diff_mean_x1	calc_dm_x1	0	-2
2	diff_mean_x1	calc_dm_x1	0.1	-1.2
3	diff_mean_x1	calc_dm_x1	0.5	0
4	diff_mean_x1	calc_dm_x1	0.9	0.8
5	diff_mean_x1	calc_dm_x1	1	2
6	diff_mean_x2	calc_dm_x2	0	-4.6
7	diff_mean_x2	calc_dm_x2	0.1	-2.22
8	diff_mean_x2	calc_dm_x2	0.5	-0.1

Using DeclareDesign: repeat the experiment III

```
9 diff_mean_x2 calc_dm_x2      0.9      2.02
10 diff_mean_x2 calc_dm_x2      1      3.8
```

```
sim_N10 %>%
  filter(inquiry == "diff_mean_x1") %>%
  summarize(p_x1 = 2 * min(mean(estimate >= 1.2), mean(estimate <= 1.2)))
```

```
  p_x1
1 0.12
```

Now try for some larger Ns (here is where DeclareDesign makes our lives much easier):

```
the_designs <- redesign(des_N_10, N = c(10, 16, 20))
sims <- simulate_designs(the_designs, sims = 1000)
```

```
sims %>%
  filter(inquiry == "diff_mean_x1") %>%
  group_by(N) %>%
  summarize(p_x1 = 2 * min(mean(estimate >= 1.2), mean(estimate <= 1.2)))
```

Using DeclareDesign: repeat the experiment IV

```
# A tibble: 3 x 2
```

```
      N  p_x1  
<int> <dbl>
```

```
1     10  0.21
```

```
2     16  0.1
```

```
3     20  0.09
```

```
cov_bal_res <- sims %>%
```

```
  group_by(inquiry, N) %>%
```

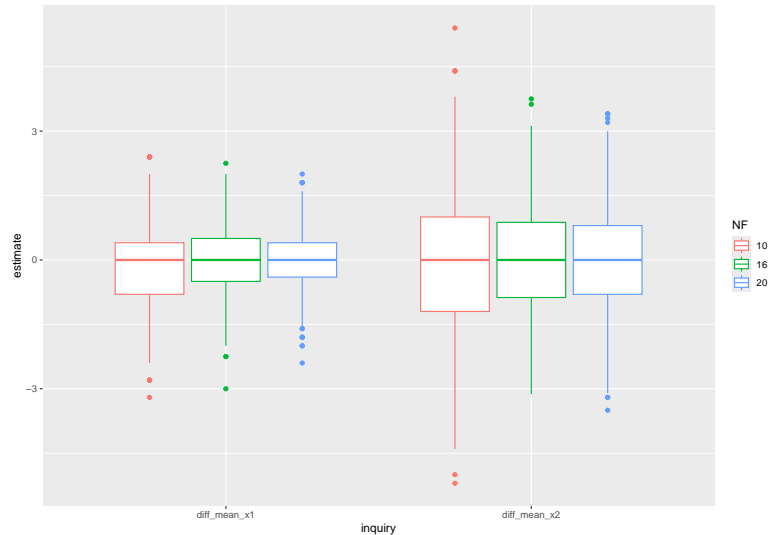
```
  reframe(  
    pctl = c(0, .1, .9, 1),  
    quantiles = round(quantile(estimate, pctl), 5)  
  )
```

```
sims$NF <- factor(sims$N)
```

```
g <- ggplot(sims, aes(x = inquiry, groups = NF, color = NF, y = estimate))  
  geom_boxplot()
```

```
g
```

Using DeclareDesign: repeat the experiment V



Summary:

1. Randomization breaks the systematic relationship between treatment and background covariates.
 - Notice that the median of the differences in means of x_1 and x_2 is zero for all N above in the boxplot.
 - Notice also that this is true even if a covariate is a strong driver of the potential outcomes.
2. Randomization allows for covariate differences in any given experiment. But as the sample size increases, the maximum size of the differences goes down.
3. Randomization allows us to **know** the sizes of the possible covariate differences (this is because we can **repeat the experimental design** and **use simulation**). It also allows us to reason about whether a given observed difference is really weird or not so weird. (For example, 1.2 is not that strange to see if $N = 10$, but becomes stranger to see as N increases.)
4. Randomization makes those who were randomly selected to not receive the intervention to be good stand-ins for the counterfactuals for those who were randomly selected to receive the treatment, and vice versa.

Random assignment vs Random sampling

Randomization of treatment assignment

- Randomization means that every observation has a known probability of assignment to experimental conditions *between* 0 and 1.
 - ▶ No unit in the experimental sample is assigned to treatment with certainty (probability = 1) or to control with certainty (probability = 0).
- Units can vary in their probability of treatment assignment.
 - ▶ For example, the probability might vary by group: women might have a 25% probability of being assigned to treatment while men have a different probability.
 - ▶ It can even vary across individuals, though that would complicate your analysis.

Random assignment of treatment vs. random sampling from a population

- Randomization (of treatment): assigning subjects with known probability to experimental conditions.
 - ▶ This random assignment of treatment can be combined with any kind of sample (random sample, convenience sample, etc.).
 - ▶ But the size and other characteristics of your sample will affect your power and your ability to extrapolate from your finding to other populations.
- Random sampling (from population): selecting subjects into your sample from a population with known probability.

Random sampling

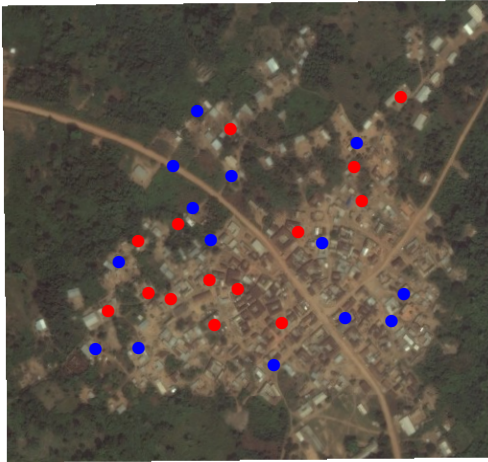


Potential outcomes

Each household i has $Y_i(1)$ and $Y_i(0)$.



Random assignment to red (1) or blue (0) condition



Key assumptions for randomized experiments

Three key assumptions

To make causal claims with an experiment (or to judge whether we believe a study's claims), we need three core assumptions:

- Random assignment of subjects to treatment, which implies that receiving the treatment is statistically independent of subjects' potential outcomes.
- Stable unit treatment value assumption (SUTVA).
- Excludability, which means that a subject's potential outcomes respond only to the defined treatment, not other extraneous factors that may be correlated with treatment.

Key assumption: SUTVA, part 1

1. No interference – A subject's potential outcome reflects only whether that subject receives the treatment himself/herself. It is not affected by how treatments happen to be allocated to other subjects.
 - ▶ A classic violation is the case of vaccines and their spillover effects.
 - ▶ Say I am in the control condition (no vaccine). If whether I get sick ($Y_i(0)$) depends on other people's treatment status (whether they take the vaccine), it's like I have two different $Y_i(0)$ (one for when I'm near a vaccinated person and one for when I am not)!
 - ▶ SUTVA (= stable unit treatment value assumption)

Key assumption: SUTVA, part 2

2. No hidden variations or types of the treatment

- ▶ Say treatment is taking a vaccine, but there are two kinds of vaccines and they have different ingredients.
- ▶ An example of a violation is when whether I get sick when I take the vaccine: $Y_i(1)$ depends on which vaccine I got. We would have two different $Y_i(1)$ — one for vaccine 1 and another for vaccine 2!

If I want to write $\tau_i = Y_i(T_i = 1) - Y_i(T_i = 0)$ as the causal effect of T on Y for person i . Recall: without the SUTVA assumption, and with 4 units, with two having $T_i = 1$, unit $i = 1$ would have the following potential outcomes:

$(y_{i=1,1100}, y_{i=1,1010}, y_{i=1,1001}, y_{i=1,0101}, y_{i=1,0011})$

Key assumption: Excludability

- Treatment assignment has no effect on outcomes except through its effect on whether treatment was received.
 - ▶ Important to also maintain symmetry between treatment and control groups (e.g., through blinding, having the same data collection procedures for all study subjects, etc.), so that treatment assignment only affects the treatment received, not other things like interactions with researchers that you don't want to define as part of the treatment.
 - ▶ Example: people in the treatment group (who know they got a vaccine) rush out to go dancing and then we see more sickness in that group.
 - ▶ Example: village not receiving payments (the control group) get targeted by an NGO for increased funds and support.

Estimation, Estimators, Bias, Consistency, Given Randomization

Randomization and an unbiased estimator

- Say we want the ATE, $\bar{\tau}_i = \overline{Y_i(1) - Y_i(0)}$.
- We will make use of the fact that the average of differences equals the difference of averages to write it down:

$$\text{ATE} = \overline{Y_i(1) - Y_i(0)} = \overline{Y_i(1)} - \overline{Y_i(0)}$$

Randomization and an unbiased estimator

- Let's *randomly assign* some of our units to the treatment condition. For these treated units, we measure the outcome $Y_i|T_i = 1$, which is the same as $Y_i(1)$ for these units.
- Because these units were randomly assigned to treatment, these observed $Y_i = Y_i(1)$ for the treated units represent the $Y_i(1)$ for all our units.
- In expectation (or on average across repeated experiments (written $E_R[\cdot]$)):

$$E_R[\overline{Y_i}|T_i = 1] = \overline{Y_i(1)}.$$

- $\overline{Y}|T_i = 1$ is an unbiased estimator of the population mean of potential outcomes under treatment.
- The same logic applies for units randomly assigned to control:

$$E_R[\overline{Y_i}|T_i = 0] = \overline{Y_i(0)}.$$

Randomization and an unbiased estimator

- The same logic applies when we measure Y_i for the control units ($Y_i|T_i = 0$). So $\overline{Y_i|T_i = 0}$, which we can calculate, gives us an unbiased estimate of $\overline{Y_i(0)}$.
- So we can write down an estimator for the ATE:

$$\hat{\tau}_i = (\overline{Y_i(1)}|T_i = 1) - (\overline{Y_i(0)}|T_i = 0)$$

- In expectation, or on average across repeated experiments, $\hat{\tau}_i$ equals the ATE:

$$E_R[Y_i|T_i = 1] - E_R[Y_i|T_i = 0] = \overline{Y_i(1)} - \overline{Y_i(0)}.$$

- So we can just take the difference of these unbiased estimators of $\overline{Y_i(1)}$ and $\overline{Y_i(0)}$ to get an unbiased estimate of the ATE.

Estimating the ATE in an RCT:

Here are two proposals for estimating the ATE. How would we know whether either or both of them work well (trick question)? (What do we want estimators to do for us? Recall “unbiased”. What does this mean?)

```
est1 <- function(Z, Y) {  
  mean(Y[Z == 1]) - mean(Y[Z == 0])  
}
```

```
est2 <- function(Z, Y) {  
  coef(lm(Y ~ Z))["Z"]  
}
```

```
with(dat, est1(Z = Z, Y = Y))
```

```
[1] -0.4
```

```
with(dat, est2(Z = Z, Y = Y))
```

```
[1] -0.4
```

How does randomization help us trust our estimators?

This is a simulation assessing **estimation bias** (and hinting at **consistency**)

```
## The truth:
```

```
with(dat, mean(y1 - y0))
```

```
[1] -1
```

```
new_exp <- function(Z) {  
  ## This next shuffles Z  
  newZ <- sample(Z)  
  return(newZ)  
}
```

```
new_est <- function(newZ, y0, y1, the_est) {  
  newY <- newZ * y1 + (1 - newZ) * y0  
  result <- the_est(Z = newZ, Y = newY)  
}
```

```
set.seed(1235)  
dist_est1 <- with(dat, replicate(100, new_est(newZ = new_exp(Z), y0 = y0,  
mean(dist_est1)
```

```
[1] -0.806
```


How does randomization help us trust our estimators?

Note: (1) Different simulations give slightly different results and (2) more simulations differ from each other less.

```
set.seed(1235)
dist_est1a <- with(dat, replicate(100, new_est(newZ = new_exp(Z), y0
mean(dist_est1a)
```

```
[1] -0.806
```

```
dist_est1b <- with(dat, replicate(100, new_est(newZ = new_exp(Z), y0
mean(dist_est1b)
```

```
[1] -1.019
```

```
dist_est2a <- with(dat, replicate(10000, new_est(newZ = new_exp(Z), y
mean(dist_est2a)
```

```
[1] -0.9883
```

```
dist_est2b <- with(dat, replicate(10000, new_est(newZ = new_exp(Z), y
mean(dist_est2b)
```

```
[1] -0.989
```

How does randomization help us trust our estimators?

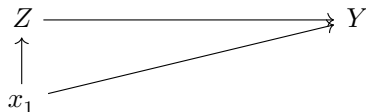
What did randomization provide here?

1. Grounds for repetition (i.e. we **knew** how to repeat the generation of Z),
2. No need to mention x_1 (we could check to see if we should worry about x_1).
3. an unbiased estimator.

What about if we didn't know exactly how Z was assigned?

|

Imagine that x_1 causes Z (here, Z is randomized but x_1 changes Z before revealing y_0 or y_1). This might be called a “broken experiment” because Z is no longer independent of the potential outcomes because it is changed by a covariate that itself is a driver of potential outcomes.



```
new_biased_exp <- function(Z, x1) {  
  newZ1 <- sample(Z)  
  # newZ <- newZ1*rbinom(10,size=1,prob=(x1+1)/4)  
  newZ <- pnorm(x1 + newZ1) > pnorm(median(x1 + newZ1))  
  return(as.numeric(newZ))  
}  
  
trueATE <- with(dat, mean(y1 - y0))  
with(dat, est1(new_biased_exp(Z, x1), Y))
```

What about if we didn't know exactly how Z was assigned?
||

```
[1] 0.9
```

```
set.seed(1235)
dist_est_biased <- with(dat, replicate(10000, new_est(newZ = new_bias
summary(trueATE - dist_est_biased)
```

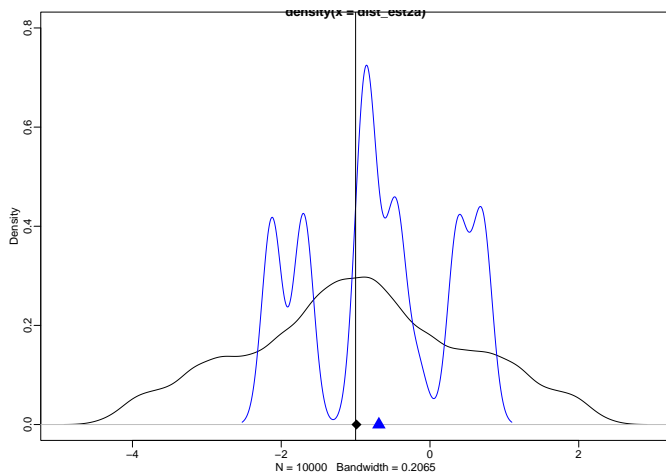
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.700	-1.375	-0.167	-0.313	0.700	1.125

```
## And recall our previous distribution of our estimator across random
## This next is unbiased (mean \approx 0)
summary(trueATE - dist_est2a)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.300	-1.000	-0.100	-0.012	1.000	3.300

What about if we didn't know exactly how Z was assigned?

Imagine that x_1 causes Z (here, Z is randomized but x_1 changes Z before revealing y_0 or y_1):



What about if we didn't know how Z was assigned at all?

How would we assess bias if we didn't know that x_1 caused Z ?

- We cannot simply shuffle Z . Because we don't know how Z arose.
- Do we know how x_1 was generated? If so, we could re-generate x_1 and *hope* that our x_1 to Z function is right
- We could repeatedly re-generate Y itself if we **knew** how it was created.
- We could resample the entire dataset if we **knew** how it was sampled.

So: **known randomization allows us to assess bias.**

Summary and Overview

Benefits of Randomized Designs

Randomization makes $y_1, y_0, \mathbf{X} \perp Z$. How to make use of this fact in a randomized experiment?

1. Interpretable comparisons (lack of omitted variable bias, confounding, selection bias)
 - ▶ Can I interpret differences in outcome as caused by Z and not some covariate like x_1 ? Is it easy to confuse the effect of Z with the effects of a covariate like x_1 ?
 - ▶ How does randomization do this? How does randomization eliminate **alternative explanations**? (Recall that it does not exactly balance all of the covariates in every randomization.)
2. Reliable statistical inferences (estimators and tests)
 - ▶ The idea of **design-based** versus **model-based** statistical inference (next few slides). We've hinted at this already.

Design Based Approach: Estimate Averages

1. Notice that the observed Y_i are a sample from the (small, finite) population of $(y_{i,1}, y_{i,0})$.
2. Decide to focus on the average, $\bar{\tau}$, because sample averages, $\hat{\bar{\tau}}$ are unbiased and consistent estimators of population averages under random sampling (where no covariate determines the sample inclusion probability or assignment to T_i).
3. Estimate $\bar{\tau}$ with the observed difference in means.



I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	$y_{i,1}$	$y_{i,0}$
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i,1}$	$\bar{y}_{i,0}$

$$\begin{aligned}\widehat{ATE} &= \bar{Y}_i | Z_i = 1 - \bar{Y}_i | Z_i = 0 \\ &= \frac{22+14}{2} - \frac{16+7}{2} = 6.5\end{aligned}$$

Design Based Approach: Estimate Averages



I don't know the truth, but I can provide a good guess of the average causal effect.

i	z_i	y_i	y_{i1}	y_{i0}
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\overline{y_{i1}}$	$\overline{y_{i0}}$

$$\widehat{ATE} = \overline{y}_{i|z_i=1} - \overline{y}_{i|z_i=0}$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

Lingering Questions?

Questions arising?

Project and R time.

Appendix

Design Based Approach: Test Hypotheses

1. Make a guess about τ_i .
2. Then measure surprise or consistency of data with this guess given the design.
(Given all of the ways this experiment could have occurred, how many look more extreme than what we observe? Does our observation look typical or rare?)

	C	fully observed		part observed	
		Z_i	Y_i	$y_{i,z_i=1}$	$y_{i,z_i=0}$
units	A	0	16	?	16
	B	1	22	22	? 22
	C	0	7	?	7
	D	1	14	14	? 14
mean diff			6.5		


we see $\bar{Y}|Z=1 = \frac{22+14}{2} = 18$
 $\bar{Y}|Z=0 = \frac{16+7}{2} = 11.5$

6.5 compared to what?

Possible Z's

0	1	1	1	0	0
0	1	0	0	1	1
1	0	1	0	1	0
1	0	0	1	0	1
<hr/>					
-8.5	8.5	-6.5	.5	-.5	6.5

Possible mean diff if $y_{i,1} = y_{i,0}$.



Design Based Approach: Test Hypotheses

	i	fully observed		part observed	
		z_i	y_i	$y_{i,z_i=1}$	$y_{i,z_i=0}$
units	A	0	16	?	16
	B	1	22	22	? 22
	C	0	7	?	7
	D	1	14	14	? 14

mean diff
6.5

We see $\bar{y}|z=1 = \frac{22+14}{2} = 18$
 $\bar{y}|z=0 = \frac{16+7}{2} = 11.5$

6.5 compared to what?

Possible z 's

0	1	1	1	0	0
0	1	0	0	1	1
1	0	1	0	1	0
1	0	0	1	0	1
<hr/>					
-8.5	8.5	-6.5	.5	-.5	6.5

Possible mean diff if
 $y_{i,1} = y_{i,0}$.



Design Based Approach: Test Hypotheses

	i	fully observed		part observed	
		z_i	y_i	$y_{i,z_i=1}$	$y_{i,z_i=0}$
units	A	0	16	?	16
	B	1	22	22	? 22
	C	0	7	?	7
	D	1	14	14	? 14

mean
diff

6.5

We see $\bar{y}|z=1 = \frac{22+14}{2} = 18$
 $\bar{y}|z=0 = \frac{16+7}{2} = 11.5$

6.5 compared to what?

$p(\text{mean diff})$
 $\frac{1}{6}$



Possible z 's

0	1	1	1	0	0
0	1	0	0	1	1
1	0	1	0	1	0
1	0	0	1	0	1
-8.5	8.5	-6.5	.5	-.5	6.5

Possible mean diff if
 $y_{i,1} = y_{i,0}$.

$p = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$







References

References

EGAP Policy Brief 40: Development Assistance and Collective Action Capacity

EGAP Policy Brief 58: Does Bottom-Up Accountability Work?

- 
- Becker, Howard S. (1998).
- Tricks of the Trade : How to Think about Your Research While You're Doing It*
- (Chicago Guides to Writing, Editing, and Publishing). University Of Chicago Press. ISBN: 0226041247.

 Brady, Henry E (2008a). "Causation and explanation in social science". In: *The Oxford Handbook of Political Science*. URL: <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199286546.001.0001/oxfordhb-9780199286546-e-10>. — (2008b). "Causation and explanation in social science". In: ed. by Janet M Box-Steffensmeier, Henry E Brady, and David Collier. Oxford University Press, pp. 217–270. Cartwright, Nancy and Jeremy Hardie (2012). *Evidence-based policy: a practical guide to doing it better*. Oxford University Press. Gerber, Alan S and Donald P Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton. Rosenbaum, Paul R (2010). *Design of Observational Studies*. New York, NY: Springer.