Lectures on causal inference and experimental methods

Macartan Humphreys

Section 1

Roadmap

Road Map

Day 1: Intro

- 1.1 Course outline, tools,
- 1.2 Introduction to Declare design

Day 2: Causality

- 2.1 Fundamental problems and basic solutions
- 2.2 General inquiries and causal identification

Day 3: Estimation and Inference

- 3.1 Frequentist
- 3.2 Bayesian

Day 4:

- 4.1 Experimental Design
- 4.2 Design evaluation

Subsection 1

Getting started

Plan

- General aims and structure
- Expectations
- Pointers for exercises
- Quick `declaredesign` intro

Aims

- Deep understanding of key ideas in causal inference
- Transportable tools for understanding how to evaluate and improve design
- Applied skills for design and analysis
- Exposure to open science practices

Syllabus

https://macartan.github.io/ci/syllabus.pdf

The topics

Day 1: Intro

- 1.1 Course outline, tools,
- 1.2 Introduction to Declare design

Day 2: Causality

- 2.1 Fundamental problems and basic solutions
- 2.2 General inquiries and causal identification

Day 3: Estimation and Inference

- 3.1 Frequentist
- 3.2 Bayesian

Day 4:

- 4.1 Experimental Design
- 4.2 Design evaluation

Expectations

- 5 tasks
- (Required) Work in four "exercise teams": 1 team per session $\times 4$
- (Optional) Prepare a research design or short paper, perhaps building on existing work. Typically this contains:
  - a problem statement
  - a description of a method to address the problem
  - analytic or simulation based results describing properties of the solution
  - a discussion of implications for practice. A passing paper will illustrate subtle features of a method; a good paper will identify unknow properties of a method; en excellent paper will develop a new method.
- Plus general reading and participation.

Exercise team job

Teams should prepare 15 - 20 minute presentations on set puzzles. Typically the task is to:

- Take a puzzle, theorem, claim
- Declare and diagnose a design that shows the claim operating (e.g. some estimator produces unbiased estimates under some condition)
- Modify the design to show behavior when conditions are violated
- Share a report with the class. Best in self-contained documents for easy third party viewing. e.g. `.html` via `.qmd` or `.Rmd`

See example in `git`.

Good coding rules

- https://bookdown.org/content/d1e53ac9-28ce-472f-bc2c-f499f18264a3/code.html
- https://www.r-bloggers.com/2018/09/r-code-best-practices/

Good coding rules

- Metadata first
- Call packages at the beginning: use `pacman`
- Put options at the top
- Call all data files once, at the top. Best to call directly from a public archive, when possible.
- Use functions and define them at the top: comment them; useful sometimes to illustrate what they do
- Replicate first, re-analyze second. Use sections.
- Have subsections named after specific tables, figures or analyses

Aim

Nothing local, everything relative: so please do not include hardcoded paths to your computer

- First best: if someone has access to your `.Rmd`/`.qmd` file they can hit render or compile and the whole thing reproduces first time.
- But: often you need ancillary files for data and code. That's OK but aims should still be that with a self contained folder someone can open a `master.Rmd` file, hit compile and get everything. I usually have an `input` and an `output` subfolder.

Collaborative coding / writing

- Do not get in the business of passing attachments around
- Share self contained folders; folders contain a small set of live documents plus an archive. Old versions of documents are in archive. Only one version of the most recent document is in a main folder.
- Data is self contained folder (`in`) and is never edited directly
- Update to github frequently

Section 2

DeclareDesign

Subsection 1

Roadmap

Roadmap

1. The MIDA framework and the declaration-diagnosis-redesign cycle
2. `DeclareDesign`: key resources
3. The Declare-Diagnose-Redesign life cycle
4. Using designs
5. Hands-on declaration and diagnosis
6. An illustration of power
7. A deeper dive into declaration functionality
8. Topics and Exercises
9. Solutions

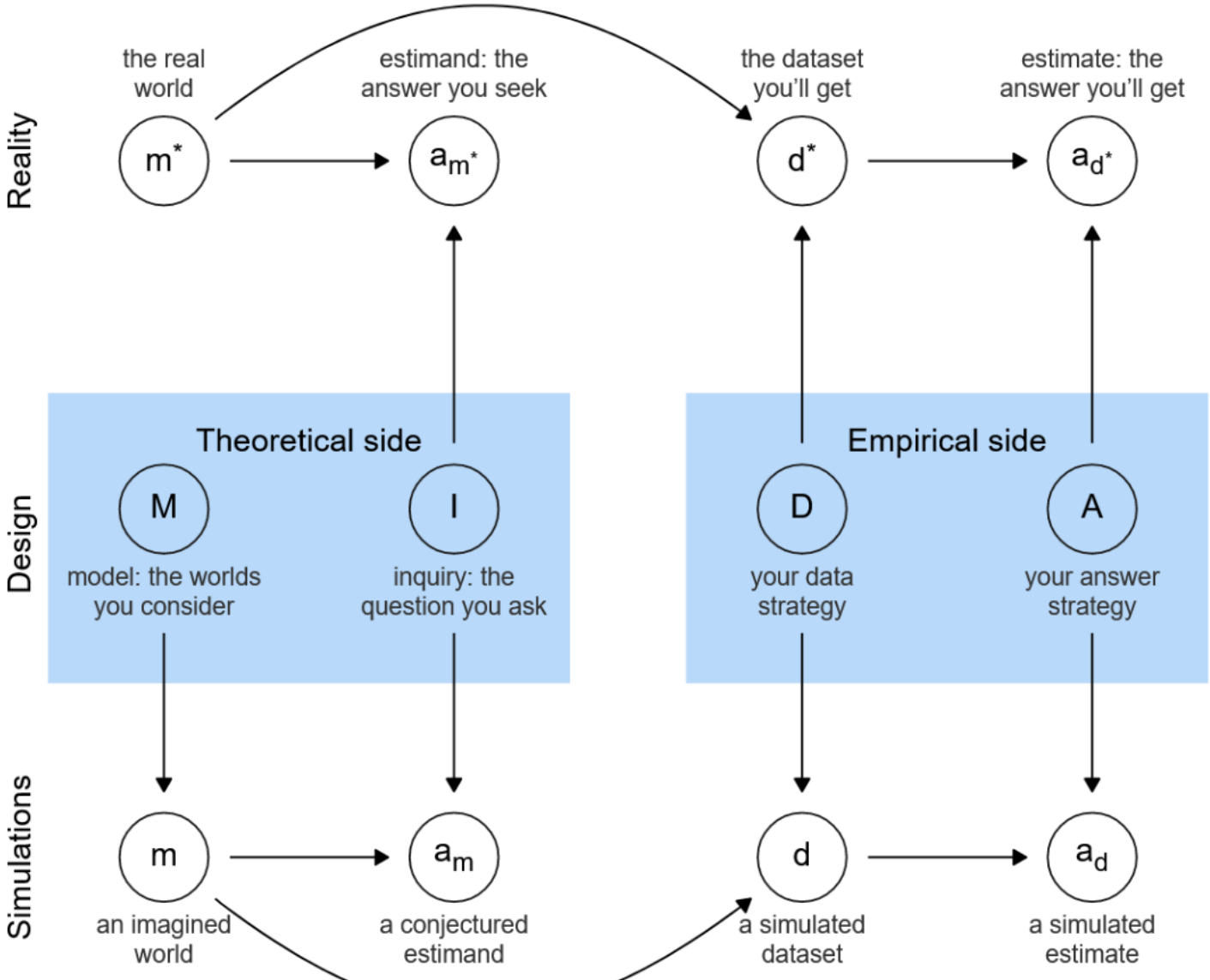
Subsection 2

The MIDA Framework

Four elements of any research design

- `Model`: set of models of what causes what and how
- `Inquiry`: a question stated in terms of the model
- `Data strategy`: the set of procedures we use to gather information from the world (sampling, assignment, measurement)
- `Answer strategy`: how we summarize the data produced by the data strategy

Four elements of any research design



Declaration

Design declaration is telling the computer (and readers) what `M`, `I`, `D`, and `A` are.

Diagnosis

Design diagnosis is figuring out how the design will perform under imagined conditions.

Estimating "diagnosands" like power, bias, rmse, error rates, ethical harm, amount learned.

Diagnosis take account of model uncertainty: it seems to identify models for which the design works well and models for which it does not

Redesign

Redesign is the fine-tuning of features of the data and answer strategies to understand how changing them affects the diagnosands

- Different sample sizes
- Different randomization procedures
- Different estimation strategies
- Implementation: effort into compliance versus more effort into sample size

Very often you have to simulate!

- Doing all this is often too hard to work out from rules of thumb or power calculators
- Specialized formulas exist for some diagnosands, but not all

Subsection 3

DeclareDesign: Overview of key functions and resources

Key commands for making a design

- `declare_model()`
- `declare_inquiry()`
- `declare_sampling()`
- `declare_assignment()`
- `declare_measurement()`
- `declare_estimator()`

and there are more `declare_` functions!

Key commands for using a design

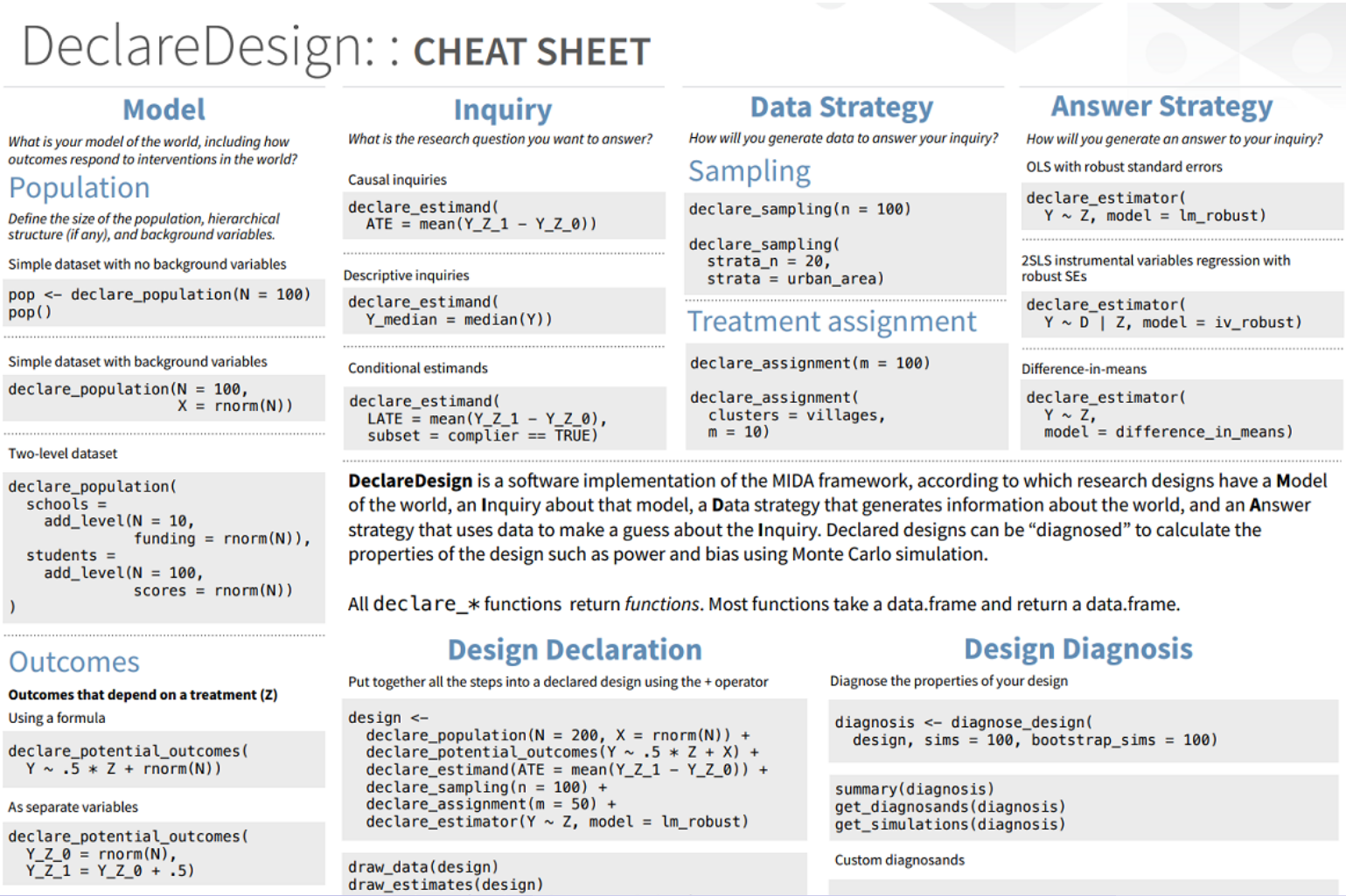
- `draw_data(design)`
- `draw_estimands(design)`
- `draw_estimates(design)`
- `get_estimates(design, data)`
- `run_design(design)`, `simulate_design(design)`
- `diagnose_design(design)`
- `redesign(design, N = 200)`
- `compare_designs()`, `compare_diagnoses()`

Pipeable commands

```
design |>
  redesign(N = c(200, 400)) |>
  diagnose_designs() |>
  tidy() |>
  ggplot(...)
```

Cheat sheet

https://raw.githubusercontent.com/rstudio/cheatsheets/master/declaredesign.pdf



DeclareDesign: : CHEAT SHEET

Model

What is your model of the world, including how outcomes respond to interventions in the world?

Population

Define the size of the population, hierarchical structure (if any), and background variables.

Simple dataset with no background variables

```
pop <- declare_population(N = 100)
pop()
```

Simple dataset with background variables

```
declare_population(N = 100,
                   X = rnorm(N))
```

Two-level dataset

```
declare_population(
  schools =
    add_level(N = 10,
              funding = rnorm(N)),
  students =
    add_level(N = 100,
              scores = rnorm(N))
)
```

Outcomes

Outcomes that depend on a treatment (Z)

Using a formula

```
declare_potential_outcomes(
  Y ~ .5 * Z + rnorm(N))
```

As separate variables

```
declare_potential_outcomes(
  Y_Z_0 = rnorm(N),
  Y_Z_1 = Y_Z_0 + .5)
```

Inquiry

What is the research question you want to answer?

Causal inquiries

```
declare_estimand(
  ATE = mean(Y_Z_1 - Y_Z_0))
```

Descriptive inquiries

```
declare_estimand(
  Y_median = median(Y))
```

Conditional estimands

```
declare_estimand(
  LATE = mean(Y_Z_1 - Y_Z_0),
  subset = complier == TRUE)
```

Data Strategy

How will you generate data to answer your inquiry?

Sampling

```
declare_sampling(n = 100)

declare_sampling(
  strata_n = 20,
  strata = urban_area)
```

Treatment assignment

```
declare_assignment(m = 100)

declare_assignment(
  clusters = villages,
  m = 10)
```

Answer Strategy

How will you generate an answer to your inquiry?

OLS with robust standard errors

```
declare_estimator(
  Y ~ Z, model = lm_robust)
```

2SLS instrumental variables regression with robust SEs

```
declare_estimator(
  Y ~ D | Z, model = iv_robust)
```

Difference-in-means

```
declare_estimator(
  Y ~ Z,
  model = difference_in_means)
```

DeclareDesign is a software implementation of the MIDA framework, according to which research designs have a **M**odel of the world, an **I**nquiry about that model, a **D**ata strategy that generates information about the world, and an **A**nswer strategy that uses data to make a guess about the **I**nquiry. Declared designs can be "diagnosed" to calculate the properties of the design such as power and bias using Monte Carlo simulation.

All `declare_*` functions return *functions*. Most functions take a data.frame and return a data.frame.

Design Declaration

Put together all the steps into a declared design using the + operator

```
design <-
  declare_population(N = 200, X = rnorm(N)) +
  declare_potential_outcomes(Y ~ .5 * Z + X) +
  declare_estimand(ATE = mean(Y_Z_1 - Y_Z_0)) +
  declare_sampling(n = 100) +
  declare_assignment(m = 50) +
  declare_estimator(Y ~ Z, model = lm_robust)
```

```
draw_data(design)
draw_estimates(design)
```

Design Diagnosis

Diagnose the properties of your design

```
diagnosis <- diagnose_design(
  design, sims = 100, bootstrap_sims = 100)
```

```
summary(diagnosis)
get_diagnosands(diagnosis)
get_simulations(diagnosis)
```

Custom diagnosands

Other resources

- The website: https://declaredesign.org/
- The book: https://book.declaredesign.org
- The console: `?DeclareDesign`

Subsection 4

Design declaration-diagnosis-redesign workflow: Design

The simplest possible (diagnosable) design?

```
mean <- 1
simplest_design <-
  declare_model(N = 100, Y = rnorm(N, mean)) +
  declare_inquiry(Q = mean) +
  declare_estimator(Y ~ 1)
```

- We draw 100 units from a standard normal distribution, we define our inquiry as the population expectation (0), we estimate the average using a regression with a constant term.

The simplest possible design?

```
simplest_design <-
  declare_model(N = 100, Y = rnorm(N)) +
  declare_inquiry(Q = 0) +
  declare_estimator(Y ~ 1)
```

- This design has three steps, with steps connected by a +
- The design itself is just a list of steps and has class design

```
str(simplest_design)
```

```
List of 3
 $ model    :design_step:    declare_model(N = 100, Y = rnorm(
 $ Q        :design_step:    declare_inquiry(Q = 0)
 $ estimator:design_step:    declare_estimator(Y ~ 1)
 - attr(*, "call")= language construct_design(steps = steps)
 - attr(*, "class")= chr [1:2] "design" "dd"
```

The simplest possible design? It's a pipe

Each step is a function (or rather: a function that generates functions) and each function presupposes what is created by previous functions.

- The ordering of steps is quite important
- Most steps take the `main` data frame in and push the `main` dataframe out; this data frame normally builds up as you move along the pipe.
- `declare_estimator` steps take the `main` data frame in and send out an `estimator_df` dataframe; `declare_inquiry` steps take the `main` data frame in and send out an `estimand_df` dataframe.

The simplest possible design? It's a pipe

- You can run these functions one at a time if you like.
- For instance the third step presupposes the data from the first step:

```
df <- simplest_design[[1]]()
A  <- simplest_design[[3]](df)

A |> kable(digits = 2)
```

estimator	term	estimate	std.error	statistic	p.value	conf.low
estimator	(Intercept)	-0.1	0.09	-1.2	0.23	-0.27

```
Estimand  <- simplest_design[[2]](df)

Estimand |> kable(digits = 2)
```

inquiry	estimand
Q	0

The simplest possible design? Run it once

You can also just run through the whole design once by typing the name of the design:

```
simplest_design
```

```
Research design declaration summary

Step 1 (model): declare_model(N = 100, Y = rnorm(N)) ---------

Step 2 (inquiry): declare_inquiry(Q = 0) ---------------------

Step 3 (estimator): declare_estimator(Y ~ 1) -----------------

Run of the design:

 inquiry estimand estimator          term estimate std.error sta
```

The simplest possible design? Run it again

Or by asking for a run of the design

```
one_run <- simplest_design |> run_design()
one_run |> kable(digits = 2)
```

inquiry	estimand	estimator	term	estimate	std.error	statistic
Q	0	estimator	(Intercept)	0	0.09	-0.03

A single run creates data, calculates estimands (the answer to inquiries) and calculates estimates plus ancillary statistics.

The simplest possible design?: Simulation

Or by asking for a run of the design

```
some_runs <- simplest_design |> simulate_design(sims = 1000)
some_runs |> kable(digits = 2)
```

design	sim_ID	inquiry	estimand	estimator	term	e
simplest_design	1	Q	0	estimator	(Intercept)	
simplest_design	2	Q	0	estimator	(Intercept)	
simplest_design	3	Q	0	estimator	(Intercept)	
simplest_design	4	Q	0	estimator	(Intercept)	
simplest_design	5	Q	0	estimator	(Intercept)	
simplest_design	6	Q	0	estimator	(Intercept)	
simplest_design	7	Q	0	estimator	(Intercept)	
simplest_design	8	Q	0	estimator	(Intercept)	
simplest_design	9	Q	0	estimator	(Intercept)	
simplest_design	10	Q	0	estimator	(Intercept)	
simplest_design	11	Q	0	estimator	(Intercept)	

The simplest possible design?: Diagnosis

Once you have simulated many times you can "diagnose".

This is the next topic

Subsection 5

Design declaration-diagnosis-redesign workflow: Diagnosis

Diagnosis by hand

Once you have simulated many times you can "diagnose".

For instance we can ask about bias: the average difference between the estimand and the estimate:

```
some_runs |> mutate(error = estimate - estimand) |>
  summarize(mean_estimate = mean(estimate),
            mean_estimand = mean(estimand),
            bias = mean(error)) |>
  kable(digits= 2)
```

mean_estimate	mean_estimand	bias
0	0	0

The simplest possible design?

`diagnose_design()` does this in one step for a set of common "diagnosands":

```
diagnosis <-
  simplest_design |>
  diagnose_design()
```

Design	N Sims	Mean Estimand	Mean Estimate	Bias	SD
simplest_design	500	0.00	-0.00	-0.00	0.1
		(0.00)	(0.00)	(0.00)	(0.

What is the diagnosis object?

The diagnosis object is also a list; of class `diagnosis`

```
names(diagnosis)
```

```
[1] "simulations_df"      "diagnosands_df"      "diagnosand_
[4] "group_by_set"        "parameters_df"       "bootstrap_
[7] "bootstrap_sims"      "duration"
```

```
class(diagnosis)
```

```
[1] "diagnosis"
```

What is the diagnosis object?

```
diagnosis$simulations_df |>
  head() |> kable(digits = 2)
```

design	sim_ID	inquiry	estimand	estimator	term	e
simplest_design	1	Q	0	estimator	(Intercept)	
simplest_design	2	Q	0	estimator	(Intercept)	
simplest_design	3	Q	0	estimator	(Intercept)	
simplest_design	4	Q	0	estimator	(Intercept)	
simplest_design	5	Q	0	estimator	(Intercept)	
simplest_design	6	Q	0	estimator	(Intercept)	

What is the diagnosis object?

```
diagnosis$diagnosands_df |>
  head() |> kable(digits = 2)
```

design	inquiry	estimator	outcome	term	mean_estir
simplest_design	Q	estimator	Y	(Intercept)	

What is the diagnosis object?

```
diagnosis$bootstrap_replicates |>
  head() |> kable(digits = 2)
```

design	bootstrap_id	inquiry	estimator	outcome	term
simplest_design	1	Q	estimator	Y	(Intercep
simplest_design	2	Q	estimator	Y	(Intercep
simplest_design	3	Q	estimator	Y	(Intercep
simplest_design	4	Q	estimator	Y	(Intercep
simplest_design	5	Q	estimator	Y	(Intercep
simplest_design	6	Q	estimator	Y	(Intercep

Diagnosis: Bootstraps

- The bootstraps dataframe is produced by resampling from the simulations dataframe and producing a diagnosis dataframe from each resampling.
- This lets us generate estimates of uncertainty around our diagnosands.
- It can be controlled thus:

```
diagnose_design(
  ...,
  bootstrap_sims = 100
)
```

After Diagnosis

It's reshapeable: as a tidy dataframe, ready for graphing

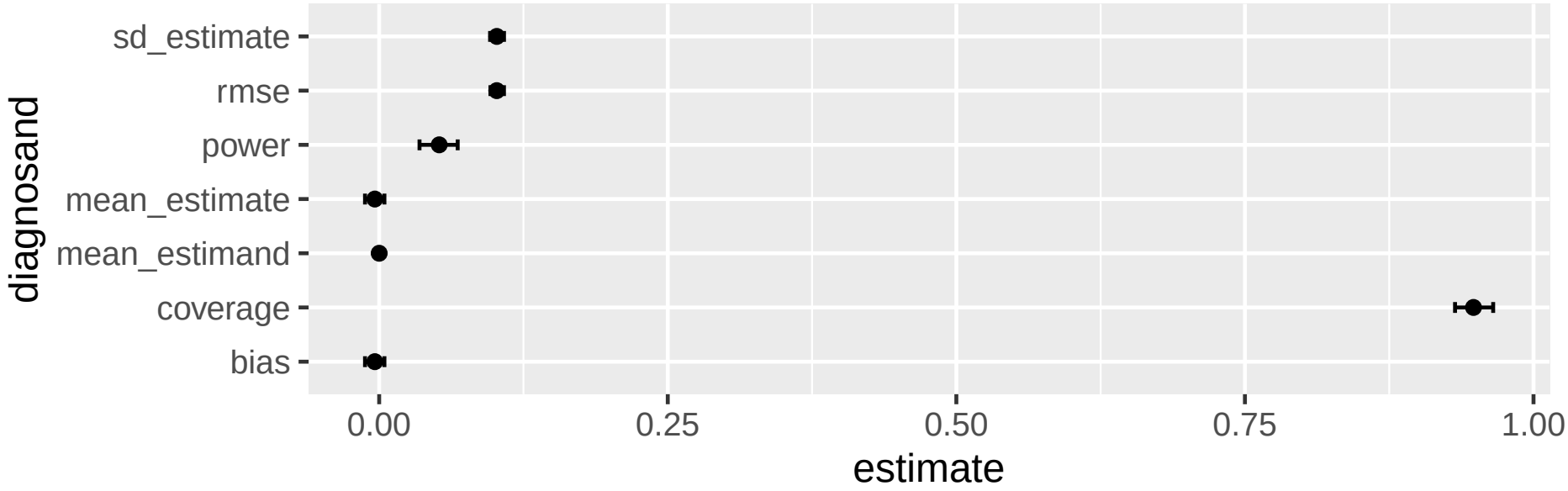
```
diagnosis |>
  tidy() |> kable(digits = 2)
```

design	inquiry	estimator	outcome	term	diagnosand
simplest_design	Q	estimator	Y	(Intercept)	mean_estir
simplest_design	Q	estimator	Y	(Intercept)	mean_estir
simplest_design	Q	estimator	Y	(Intercept)	bias
simplest_design	Q	estimator	Y	(Intercept)	sd_estimat
simplest_design	Q	estimator	Y	(Intercept)	rmse
simplest_design	Q	estimator	Y	(Intercept)	power
simplest_design	Q	estimator	Y	(Intercept)	coverage

After Diagnosis

It's reshapeable: as a tidy dataframe, ready for graphing

```
diagnosis |>
  tidy() |>
  ggplot(aes(estimate, diagnosand)) + geom_point() +
  geom_errorbarh(aes(xmax = conf.high, xmin = conf.low, height
```



After Diagnosis: Tables

Or turn into a formatted table:

```
diagnosis |>
  reshape_diagnosis() |> kable()
```

Design	Inquiry	Estimator	Outcome	Term	N Sims	
simplest_design	Q	estimator	Y	(Intercept)	500	

Advanced Diagnosis: Variations

```
DeclareDesign:::default_diagnosands
```

```
    mean_estimand <- mean(estimand)
    mean_estimate <- mean(estimate)
    bias <- mean(estimate - estimand)
    sd_estimate <- sd(estimate)
    rmse <- sqrt(mean((estimate - estimand)^2))
    power <- mean(p.value <= alpha)
    coverage <- mean(estimand <= conf.high & estimand >= conf
```

Advanced Diagnosis: Other diagnosands

```
    mean_se = mean(std.error)
    type_s_rate = mean((sign(estimate) != sign(estimand))[p.va
    exaggeration_ratio = mean((estimate/estimand)[p.value <= a
    var_estimate = pop.var(estimate)
    mean_var_hat = mean(std.error^2)
    prop_pos_sig = estimate > 0 & p.value <= alpha
    mean_ci_length = mean(conf.high - conf.low)
```

Advanced Diagnosis: Custom diagnosands

```
my_diagnosands <-
  declare_diagnosands(median_bias = median(estimate - estiman

diagnose_design(simplest_design, diagnosands = my_diagnosands
  reshape_diagnosis() |> kable()
```

Design	Inquiry	Estimator	Outcome	Term	N Sims	
simplest_design	Q	estimator	Y	(Intercept)	10	

Advanced Diagnosis: Adding diagnosands to a design

```
simplest_design <-
  set_diagnosands(simplest_design, my_diagnosands)

simplest_design |> diagnose_design(sims = 10)|>
  reshape_diagnosis() |> kable()
```

Design	Inquiry	Estimator	Outcome	Term	N Sims
simplest_design	Q	estimator	Y	(Intercept)	10

Advanced Diagnosis: Diagnosing multiple designs

You can diagnose multiple designs or a list of designs

```
list(dum = simplest_design, dee = simplest_design) |>
  diagnose_design(sims = 5) |>
  reshape_diagnosis() |>
  kable()
```

Design	Inquiry	Estimator	Outcome	Term	N Sims	Median B
dum	Q	estimator	Y	(Intercept)	5	-0.08
						(0.08)
dee	Q	estimator	Y	(Intercept)	5	-0.08
						(0.08)

Advanced Diagnosis: Diagnosing in groups

You can partition the simulations data frame into groups before calculating diagnosands.

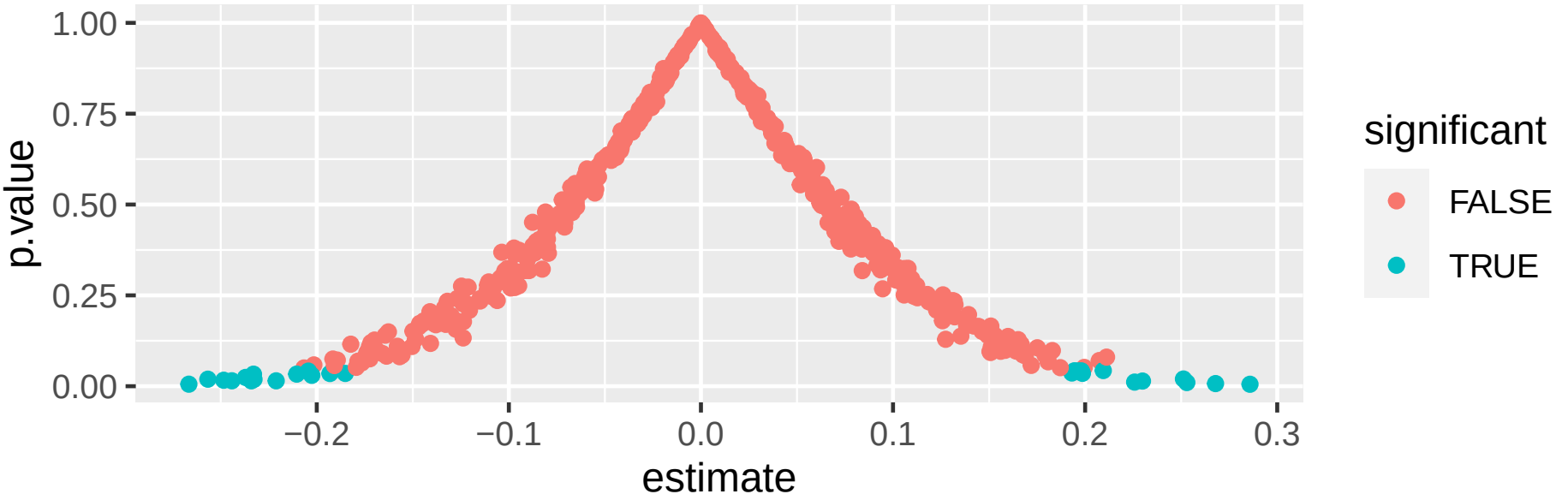
```
grouped_diagnosis <-

  simplest_design |>
  diagnose_design(
    make_groups = vars(significant = p.value <= 0.05),
    sims = 500
  )
```

Design	Significant	N Sims	Mean Estimand	Mean Estimate	Bias
design_1	FALSE	474	0.00	-0.00	-0.00
			(0.00)	(0.00)	(0.00
design_1	TRUE	26	0.00	-0.02	-0.02
			(0.00)	(0.04)	(0.04

Note especially the mean estimate, the power, the coverage, the RMSE

Significance filter

```
grouped_diagnosis$simulations_df |>
  ggplot(aes(estimate, p.value, color = significant)) + geom_p
```



Advanced Diagnosis: Multistage simulation

- Usually a design simulation simulates "from the top": going from the beginning to the end of the design in each run and repeating
- But sometimes you might want to follow a tree like structure and simulate different steps a different number of times

Consider for instance this sampling design:

```
sampling_design <-

  declare_model(N = 500, Y = rnorm(N, sd = 10)) +
  declare_inquiry(Q = mean(Y)) +
  declare_sampling(S = complete_rs(N = N, n = 100)) +
  declare_estimator(Y ~ 1)
```

Advanced Diagnosis: Multistage simulation

Compare these two diagnoses:

```
diagnosis_1 <- diagnose_design(sampling_design, sims = c(5000,
diagnosis_2 <- diagnose_design(sampling_design, sims = c(1, 50
```

diagnosis	N Sims	Mean Estimand	Mean Estimate	Bias	SD Est
diagnosis_1	5000	0.00	0.00	-0.00	1.01
diagnosis_1		(0.01)	(0.01)	(0.01)	(0.01)
diagnosis_2	5000	0.22	0.22	-0.00	0.91
diagnosis_2		(0.00)	(0.00)	(0.00)	(0.00)

Spotting design problems with diagnosis

Diagnosis alerts to problems in a design. Consider the following simple alternative design.

```
simplest_design_2 <-

  declare_model(N = 100, Y = rnorm(N)) +
  declare_inquiry(Q = mean(Y)) +
  declare_estimator(Y ~ 1)
```

Here we define the inquiry as the average Y, but otherwise things stay the same.

What do we think of this design?

Spotting design problems with diagnosis

Here is the diagnosis

Design	N Sims	Mean Estimand	Mean Estimate	Bias
simplest_design_2	500	-0.00	-0.00	0.00
		(0.00)	(0.00)	(0.00)

- Why is power 5%? is that OK?
- Why is coverage so high? is that OK?
- Why is the RMSE 0 but the mean standard error > 0? is that OK?
  - Is it because the RMSE is too low?
  - Or the standard error is too large?

It depends on the inquiry

- If we are really interested in the sample average then our standard error is too low.
- If we are really interested in the population average then our inquiry is badly defined.

Subsection 6

Design declaration-diagnosis-redesign workflow: Redesign

Redesign

Redesign is the process of taking a design and modifying it in some way.

There are a few ways to do this:

1. Just make a new design using modified code
2. Take a design and alter some steps using `replace_step`, `insert_step` or `delete_step`
3. Modify a design *parameter* using `redesign`

we will focus on the third approach

Redesign

- A design parameter is a modifiable quantity of a design.
- These quantities are objects that were in your global environment when you made your design, get referred to explicitly in your design, and got scooped up when the design was formed.
- In our simplest design above we had a fixed `N`, but we could make `N` a modifiable quantity like this:

```
N <- 100

simplest_design_N <-

  declare_model(N = N, Y = rnorm(N)) +
  declare_inquiry(Q = 0) +
  declare_estimator(Y ~ 1)
```

Redesign

```
N <- 100

simplest_design_N <-

  declare_model(N = N, Y = rnorm(N)) +
  declare_inquiry(Q = 0) +
  declare_estimator(Y ~ 1)
```

Note that N is defined in memory; and it gets called in one of the steps. It has now become a parameter of the design and it can be modified using redesign.

Simple Redesign

Here is a version of the design with N = 200:

```
design_200 <- simplest_design_N |> redesign(N = 200)

design_200 |> draw_data() |> nrow()
```

```
[1] 200
```

Redesigning to a list

Here is a list of three different designs with different *N*s.

```
design_Ns <- simplest_design_N |> redesign(N = c(200, 400, 800

design_Ns |> lapply(draw_data) |> lapply(nrow)
```

```
$design_1
[1] 200

$design_2
[1] 400

$design_3
[1] 800
```

Redesigning to a list

The good thing here is that it is now easy to diagnose over multiple designs and compare diagnoses. The parameter names then end up in the `diagnosis_df`

Consider this:

```
N <- 100
m <- 0

design <-
  declare_model(N = N, Y = rnorm(N, m)) +
  declare_inquiry(Q = m) +
  declare_estimator(Y ~ 1)
```

Then:

```
designs <-  redesign(design, N = c(100, 200, 300), m = c(0, .1
```

Redesigning to a list

Output:

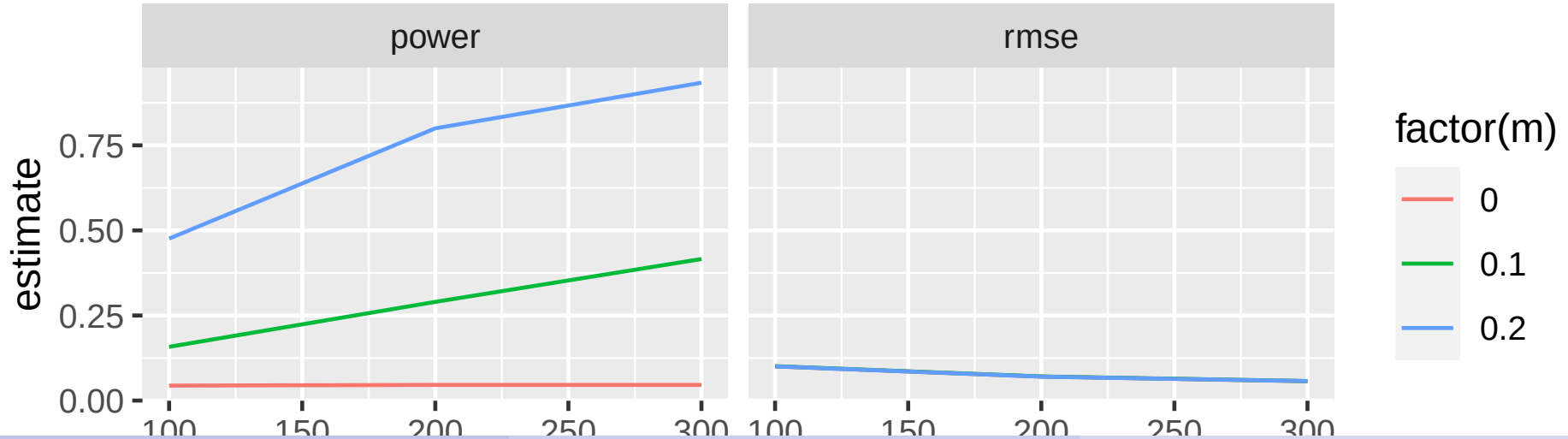
```
designs |> diagnose_design() |> tidy()
```

N	m	diagnosand	estimate	std.error	conf.low	conf.high
100	0.0	mean_estimand	0.0	0	0.00	0.00
100	0.0	mean_estimate	0.0	0	-0.01	0.01
200	0.0	mean_estimand	0.0	0	0.00	0.00
200	0.0	mean_estimate	0.0	0	-0.01	0.00
200	0.1	mean_estimand	0.1	0	0.10	0.10
200	0.1	mean_estimate	0.1	0	0.09	0.10

Redesigning to a list

Graphing after redesign is especially easy:

```
designs |> diagnose_design() |>
  tidy() |>
  filter(diagnosand %in% c("power", "rmse")) |>
  ggplot(aes(N, estimate, color = factor(m))) +
  geom_line() +
  facet_wrap(~diagnosand)
```



Redesign with vector arguments

When redesigning with arguments that are vectors, use `list()` in redesign, with each list item representing a design you wish to create

```
prob_each <- c(.1, .5, .4)

design_multi  <-
  declare_model(N = 10) +
  declare_assignment(Z = complete_ra(N = N, prob_each = prob_e

## returns two designs

designs <- design_multi |> redesign(prob_each = list(c(.2, .5,

designs |> lapply(draw_data)
```

Redesign warnings

A parameter has to be called correctly. And you get no warning if you misname.

```
simplest_design_N  |> redesign(n = 200) |> draw_data() |> nrow
```

```
[1] 100
```

why not 200?

Redesign warnings

A parameter has to be called explicitly

```
N <- 100

my_N <- function(n = N) n

simplest_design_N2 <-

  declare_model(N = my_N(), Y = rnorm(N)) +
  declare_inquiry(Q = 0) +
  declare_estimator(Y ~ 1)

simplest_design_N2 |> redesign(N = 200) |> draw_data() |> nrow
```

```
[1] 100
```

why not 200?

Redesign warnings

A parameter has to be called explicitly

```
N <- 100

my_N <- function(n = N) n

simplest_design_N2 <-

  declare_model(N = my_N(N), Y = rnorm(N)) +
  declare_inquiry(Q = 0) +
  declare_estimator(Y ~ 1)

simplest_design_N2 |> redesign(N = 200) |> draw_data() |> nrow
```

```
[1] 200
```

OK

Redesign with a function

Here is an example of redesigning where the "parameter" is a function

```
new_N <- function(n, factor = 1.31) n*factor

simplest_design_N2 |> redesign(my_N = new_N) |> draw_data() |>
```

```
[1] 131
```

Subsection 7

Using a design

Using a design

What can you do with a design once you have it?

We will start with a very simple experimental design (more on the components of this later)

```
b <-1
N <- 100
design <-
  declare_model(N = N, U = rnorm(N), potential_outcomes(Y ~ b
  declare_assignment(Z = simple_ra(N), Y = reveal_outcomes(Y
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_estimator(Y ~ Z, inquiry = "ate", .method = lm_robus
```

Make data from the design

```
data <- draw_data(design)

data |> head () |> kable()
```

ID	U	Y_Z_0	Y_Z_1	Z	Y
001	0.8939241	0.8939241	1.8939241	1	1.8939241
002	1.3350334	1.3350334	2.3350334	1	2.3350334
003	0.8329075	0.8329075	1.8329075	1	1.8329075
004	-0.2886946	-0.2886946	0.7113054	0	-0.2886946
005	-0.3062044	-0.3062044	0.6937956	1	0.6937956
006	0.6443779	0.6443779	1.6443779	1	1.6443779

Make data from the design

Play with the data:

```
lm_robust(Y ~ Z, data = data) |>
  tidy() |>
  kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.21	0.14	-1.50	0.14	-0.48	0.07
Z	1.27	0.19	6.69	0.00	0.89	1.65

Draw estimands

```
draw_estimands(design) |>
  tidy() |>
  kable(digits = 2)
```

```
Error in UseMethod("tidy"): no applicable method for 'tidy' ap
```

Draw estimates

```
draw_estimates(design) |>
  tidy() |>
  kable(digits = 2)
```

```
Error in UseMethod("tidy"): no applicable method for 'tidy' ap
```

Get estimates

Using your actual data:

```
get_estimates(design, data) |>
  tidy() |>
  kable(digits = 2)
```

```
Error in UseMethod("tidy"): no applicable method for 'tidy' ap
```

Simulate design

```
simulate_design(design, sims = 3) |>
  kable(digits = 2)
```

design	sim_ID	inquiry	estimand	estimator	term	estimate	std.er
design	1	ate	1	estimator	Z	1.50	0
design	2	ate	1	estimator	Z	1.27	0
design	3	ate	1	estimator	Z	0.87	0

Diagnose design

```
design |>
  diagnose_design(sims = 100)
```

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
1.00	0.00	0.19	0.19	1.00	0.95
(0.02)	(0.02)	(0.01)	(0.01)	(0.00)	(0.02)

Redesign

```
new_design <-

  design |> redesign(b = 0)
```

- Modify any arguments that are explicitly called on by design steps.
- Or add, remove, or replace steps

Compare designs

```
redesign(design, N = 50) %>%

  compare_diagnoses(design)
```

diagnosand	mean_1	mean_2	mean_difference	conf.low	conf.h
mean_estimand	0.50	0.50	0.00	0.00	0
mean_estimate	0.48	0.50	0.02	-0.01	0
bias	-0.02	0.00	0.02	-0.01	0
sd_estimate	0.28	0.20	-0.08	-0.10	-0
rmse	0.28	0.20	-0.08	-0.10	-0
power	0.38	0.71	0.32	0.26	0
coverage	0.97	0.96	-0.01	-0.04	0

Illustration of power calculation

Recall?: The power of a design is the *probability* that you will reject a null hypothesis

```
N <- 100
b <- .5

design <-
  declare_model(N = N,
    U = rnorm(N),
    potential_outcomes(Y ~ b * Z + U)) +
  declare_assignment(Z = simple_ra(N),
                     Y = reveal_outcomes(Y ~ Z)) +
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_estimator(Y ~ Z, inquiry = "ate", .method = lm_robus
```

"Run" the design once

```
run_design(design)
```

Table 1: Summary of a single 'run' of the design

inquiry	estimand	estimator	term	estimate	std.error	statistic	p.v
ate	0.5	estimator	Z	0.57	0.2	2.88	

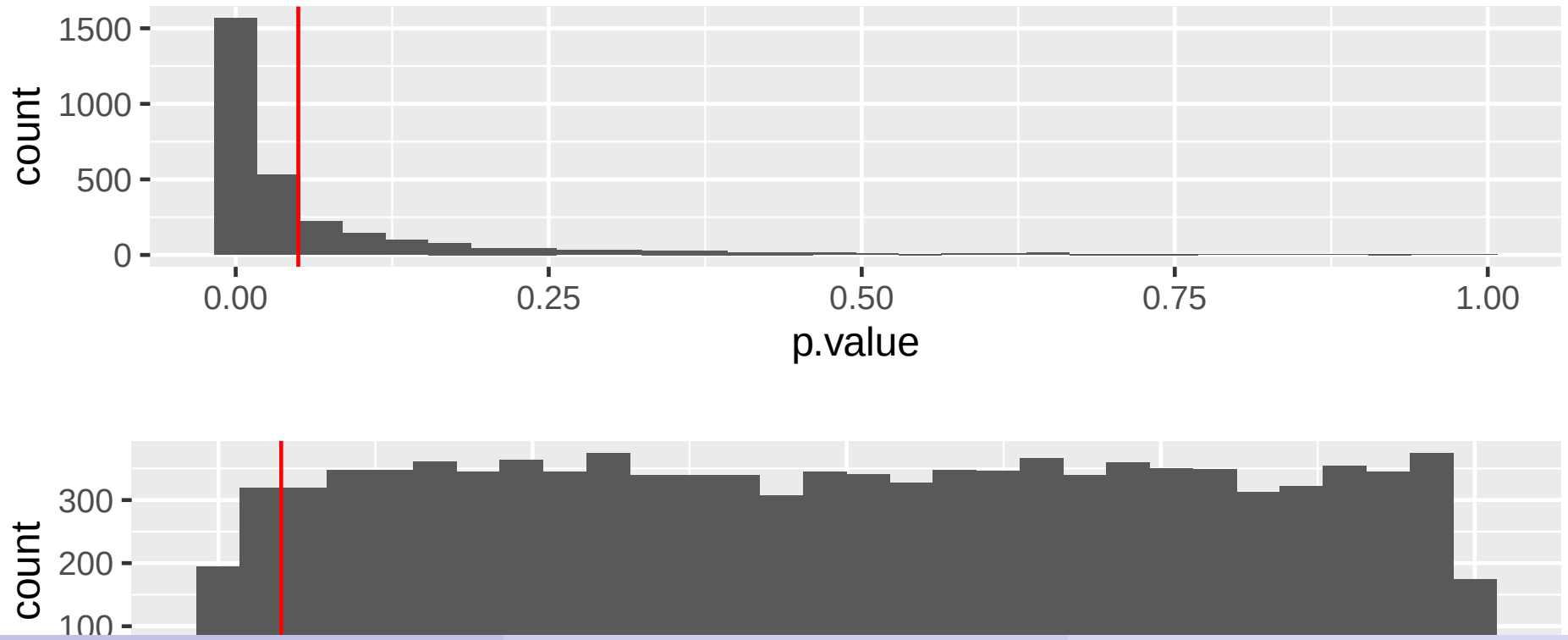
Run it many times

```
sims_1 <- simulate_design(design)

sims_1 |> select(sim_ID, estimate, p.value)
```

sim_ID	estimate	p.value
1	0.81	0.00
2	0.40	0.04
3	0.88	0.00
4	0.72	0.00
5	0.38	0.05
6	0.44	0.02

Power is mass of the sampling distribution of decisions under the model

```
sims_1 |>
  ggplot(aes(p.value)) +
  geom_histogram() +
  geom_vline(xintercept = .05, color = "red")
```



Design diagnosis does it all (over multiple designs)

```
diagnose_design(design)
```

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0.50	0.00	0.20	0.20	0.70	0.95
(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Design diagnosis does it all

```
design |>
  redesign(b = c(0, 0.25, 0.5, 1)) |>
  diagnose_design()
```

b	Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0	-0.00	-0.00	0.20	0.20	0.05	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.25	0.25	-0.00	0.20	0.20	0.23	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.5	0.50	0.00	0.20	0.20	0.70	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
1	1.00	0.00	0.20	0.20	1.00	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Subsection 8

Declaration a deeper dive (Reference)

Declaration a deeper dive (Reference)

We start with a simple experimental design and then show ways to extend.

- Variations to M and I are supported by the `fabricatr` package (and others)
- Variations to D are supported by the `randomizr` package (and others)
- Variations to A are supported by the `estimatr` package (and others)

Steps: A simple experimental design

```
N <- 100
b <- .5

design <-
  declare_model(N = N, U = rnorm(N),
                potential_outcomes(Y ~ b * Z + U)) +
  declare_assignment(Z = simple_ra(N), Y = reveal_outcomes(Y
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_estimator(Y ~ Z, inquiry = "ate", .method = lm_robus
```

A few new elements here:

- `declare_model` can be used much like `mutate` with multiple columns created in sequence
- the `potential_outcomes` function is a special function that creates potential outcome columns
- when you assign a treatment that affects an outcome you can use

Steps: A simple experimental design

```
N <- 100
b <- .5

design <-
  declare_model(N = N, U = rnorm(N),
                potential_outcomes(Y ~ b * Z + U)) +
  declare_assignment(Z = simple_ra(N), Y = reveal_outcomes(Y
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_estimator(Y ~ Z, inquiry = "ate", .method = lm_robus
```

A few new elements here:

- when you declare an estimator you should normally associate an inquiry with the estimator and provide the method to be used; `lm_robust` is default
- you should generally label estimators as you may have many

Steps: Order matters

e.g. If you sample before defining the inquiry you get a different inquiry to if you sample after you define the inquiry

```
design_1 <-
  declare_model(N = 1000, X = rep(0:1, N/2), Y = X + rnorm(N))
  declare_sampling(S= strata_rs(strata = X, strata_prob = c(.2
  declare_inquiry(m = mean(Y))

design_1 |> draw_estimands()
```

```
  inquiry  estimand
1       m 0.7907839
```

Steps: Order matters

e.g. If you sample before defining the inquiry you get a different inquiry to if you sample after you define the inquiry

```
design_2 <-
  declare_model(N = 1000, X = rep(0:1, N/2), Y = X + rnorm(N))
  declare_inquiry(m = mean(Y)) +
  declare_sampling(S= strata_rs(strata = X, strata_prob = c(.2

design_2 |> draw_estimands()
```

```
  inquiry  estimand
1       m 0.5467558
```

Key extensions to model declaration

You can generate hierarchical data like this:

```
M <-
  declare_model(
    households = add_level(
      N = 100,
      N_members = sample(c(1, 2, 3, 4), N,
                         prob = c(0.2, 0.3, 0.25, 0.25), repla
    ),
    individuals = add_level(
      N = N_members,
      age = sample(18:90, N, replace = TRUE)
    )
  )
```

Key extensions to model declaration

You can generate hierarchical data like this: ::: {.cell}

```
M() |> head() |> kable(digits = 2)
```

households	N_members	individuals	age
001	1	001	57
002	4	002	34
002	4	003	40
002	4	004	57
002	4	005	31
003	3	006	41

:::

Key extensions to model declaration

You can generate panel data like this:

```
M <-
  declare_model(
    countries = add_level(
      N = 196,
      country_shock = rnorm(N)
    ),
    years = add_level(
      N = 100,
      time_trend = 1:N,
      year_shock = runif(N, 1, 10),
      nest = FALSE
    ),
    observation = cross_levels(
      by = join_using(countries, years),
      observation_shock = rnorm(N),
```

Key extensions to model declaration

You can generate panel data like this: ::: {.cell}

```
M() |> head() |> kable(digits = 2)
```

countries	country_shock	years	time_trend	year_shock	observation
001	-1.01	001	1	7.24	00001
002	1.59	001	1	7.24	00002
003	0.18	001	1	7.24	00003
004	-2.07	001	1	7.24	00004
005	0.22	001	1	7.24	00005
006	-0.37	001	1	7.24	00006

:::

You can pull in preexisting data

```
M <-
  declare_model(
    data = baseline_data,
    attitudes = sample(1:5, N, replace = TRUE)
  )
```

A simple experimental design

You can repeat steps and play with the order, always conscious of the direction of the pipe

```
design <-
  declare_model(N = N, X = rep(0:1, N/2)) +
  declare_model(U = rnorm(N), potential_outcomes(Y ~ b * Z * X
  declare_assignment(Z = block_ra(blocks = X), Y = reveal_outc
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_inquiry(cate = mean(Y_Z_1[X==0] - Y_Z_0[X==0])) +
  declare_estimator(Y ~ Z, inquiry = "ate", label = "ols") +
  declare_estimator(Y ~ Z*X, inquiry = "cate", label = "fe")
```

You can generate multiple columns together

```
M2 <-
  declare_model(
    draw_multivariate(c(X1, X2) ~ MASS::mvrnorm(
      n = 1000,
      mu = c(0, 0),
      Sigma = matrix(c(1, 0.3, 0.3, 1), nrow = 2)
    )))
```

You can generate multiple columns together

```
M2() |> head() |> kable()
```

X1	X2
-1.3706858	-0.9653579
1.3264821	-0.0392575
-0.5585624	1.3270622
0.2907566	-1.2893750
0.4111638	-0.5837608
-1.2206595	-1.1646952

Cluster structures with cluster correlations

```
M <-
  declare_model(households = add_level(N = 1000),
                individuals = add_level(
                  N = 4,
                  X = draw_normal_icc(
                    mean = 0,
                    clusters = households,
                    ICC = 0.65
                  )
                ))
```

Cluster structures with cluster correlations

```
model <- lm_robust(X ~ households, data = M())
model$adj.r.squared
```

```
[1] 0.6709427
```

Assignment schemes

The `randomizr` package has a set of functions for different types of block and cluster assignments.

- Simple random assignment: "Coin flip" or Bernoulli random assignment. All units have the same probability of assignment: `simple_ra(N = 100, prob = 0.25)`
- Complete random assignment: Exactly m of N units are assigned to treatment, and all units have the same probability of assignment m/N `complete_ra(N = 100, m = 40)`

Assignment schemes

- Block random assignment: Complete random assignment within pre-defined blocks. Units within the same block have the same probability of assignment m_b / N_b `block_ra(blocks = regions)`
- Cluster random assignment: Whole groups of units are assigned to the same treatment condition. `cluster_ra(clusters = households)`
  * Block-and-cluster assignment: Cluster random assignment within blocks of clusters `block_and_cluster_ra(blocks = regions, clusters = villages)`

Assignment schemes

You can combine these in various ways. For examples with saturation random assignment first clusters are assigned to a saturation level, then units within clusters are assigned to treatment conditions according to the saturation level:

```
saturation = cluster_ra(clusters = villages, conditions = c(0,
block_ra(blocks = villages, prob_unit = saturation)
```

Inquiries

Many causal inquiries are simple summaries of potential outcomes:

Inquiry	Units	Code
Average treatment effect in a finite population (PATE)	Units in the population	`mean(Y_D_1 - Y_D_0)`
Conditional average treatment effect (CATE) for X = 1	Units for whom X = 1	`mean(Y_D_1[X == 1] - Y_D_0[X == 1])`
Complier average causal effect (CACE)	Complier units	`mean(Y_D_1[D_Z_1 > D_Z_0] - Y_D_0[D_Z_1 > D_Z_0])`
Causal interactions of D_1 and D_2	Units in the population	`mean((Y_D1_1_D2_1 - Y_D1_0_D2_1) - (Y_D1_1_D2_0 - Y_D1_0_D2_0))`

Inquiries

Often though we need to define inquiries as a function of continuous variables. For this generating a potential outcomes function can make life easier. This helps for:

- Continuous quantities
- Spillover quantities
- Complex counterfactuals

Inquiries: Complex counterfactuals

Here is an example of using functions to define complex counterfactuals:

```
f_M <- function(X, UM) 1*(UM < X)
f_Y <- function(X, M, UY) X + M - .4*X*M + UY

design <-
  declare_model(N = 100,
                X = simple_rs(N),
                UM = runif(N),
                UY = rnorm(N),
                M = f_M(X, UM),
                Y = f_Y(X, M, UY)) +
  declare_inquiry(Q1 = mean(f_Y(1, f_M(0, UM), UY) - f_Y(0, f_

design |> draw_estimands() |> kable()
```

inquiry	estimand

Inquiries: Complex counterfactuals

Here is an example of using functions to define effects of continuous treatments.

```
f_Y <- function(X, UY) X - .25*X^2 + UY

design <-
  declare_model(N = 100,
                X  = rnorm(N),
                UY = rnorm(N),
                Y = f_Y(X, UY)) +
  declare_inquiry(
    Q1 = mean(f_Y(X+1, UY) - f_Y(X, UY)),
    Q2 = mean(f_Y(1, UY) - f_Y(0, UY)),
    Q3 = (lm_robust(Y ~ X)|> tidy())[2,2]
    )

design |> draw_estimands() |> kable()
```

Answers: terms

By default `declare_estimates()` assumes you are interested in the *first term after the constant* from the output of an estimation procedure.

But you can say what you are interested in directly using `term` and you can also associate different terms with different quantities of interest using `inquiry`.

```
design <-
  declare_model(N = 100,
                X1 = rnorm(N),
                X2 = rnorm(N),
                X3 = rnorm(N),
                Y = X1 - X2 + X3 + rnorm(N)) +
  declare_inquiries(ate_2 = -1, ate_3 = 1) +
  declare_estimator(Y ~ X1 + X2 + X3, term = c("X2", "X3"), in

design  |> run_design()  |> kable(digits = 2)
```

Answers: terms

Sometimes it can be confusing what the names of a term is but you can figure this by running the estimation strategy directly. Here's an example where the names of a term might be confusing.

```
lm_robust(Y ~ A*B,
          data = data.frame(A = rep(c("a",  "b"), 3),
                            B = rep(c("p", "q"), each = 3),
                            Y = rnorm(6))) |>
  coef() |> kable()
```

	x
(Intercept)	0.984547
Ab	-1.172676
Bq	-1.976603
Ab:Bq	2.115862

The names are they appear in the output here is the name of the term that `declare_estimator` will look for

Answers: other packages

DeclareDesign works natively with estimatr but you you can use whatever packages you like. You do have to make sure though that estimatr gets as input a nice tidy dataframe of estimates, and that might require some tidying.

```
design <-
  declare_model(N = 1000, U = runif(N),
                potential_outcomes(Y ~ as.numeric(U < .5 + Z/3
  declare_assignment(Z = simple_ra(N), Y = reveal_outcomes(Y
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_estimator(Y ~ Z, inquiry = "ate",
                    .method = glm,
                    family = binomial(link = "probit"))
```

Note that we passed additional arguments to glm; that's easy.

It's not a good design though. Just look at the diagnosis:

Answers: other packages

```
diagnose_design(design)
```

```
if(run)
  diagnose_design(design) |> write_rds("saved/probit.rds")

read_rds("saved/probit.rds") |> reshape_diagnosis() |> kable()
```

Design	Inquiry	Estimator	Term	N Sims	Mean Estimand	Mean Es
design	ate	estimator	Z	500	0.33	0.97
					(0.00)	(0.00)

Why is it so terrible?

Answers: other packages

Because the probit estimate does not target the ATE directly; you need to do more work to get there.

You essentially have to write a function to get the estimates, calculate the quantity of interest and other stats, and turn these into a nice dataframe.

Luckily you can use the `margins` package with `tidy` to create a `.summary` function which you can pass to `declare_estimator` to do all this for you

```
tidy_margins <- function(x) broom::tidy(margins::margins(x, da

design <- design +
  declare_estimator(Y ~ Z, inquiry = "ate",
                    .method = glm,
                    family = binomial(link = "probit"),
                    .summary = tidy_margins,
                    label = "margins")
```

Answers: other packages

```
if(run)
  diagnose_design(design) |> write_rds("saved/probit_2.rds")

read_rds("saved/probit_2.rds") |> reshape_diagnosis() |> kable
```

Design	Inquiry	Estimator	Term	N Sims	Mean Estimand	Mean Es
design	ate	estimator	Z	500	0.33	0.97
					(0.00)	(0.00)
design	ate	margins	Z	500	0.33	0.31
					(0.00)	(0.00)

Much better

Section 3

Causality. What's a cause?

Subsection 1

Potential outcomes and the counterfactual approach

Potential outcomes and the counterfactual approach

Causation as difference making

Motivation

The *intervention* based motivation for understanding causal effects:

- We want to know if a particular intervention (like aid) caused a particular outcome (like reduced corruption).
- We need to know:
  1. What happened?
  2. What would the outcome have been if there were no intervention?
- The problem:
  1. ... this is hard
  2. ... this is impossible

The problem in 2 is that you need to know what would have happened if things were different. You need information on a **counterfactual**.

Potential Outcomes

- For each unit, we assume that there are two **post-treatment** outcomes: $Y_i(1)$ and $Y_i(0)$.
- For example, $Y(1)$ is the outcome that *would* obtain *if* the unit received the treatment.
- The **causal effect** of Treatment (relative to Control) is: $\tau_i = Y_i(1) - Y_i(0)$
- Note:
  - The causal effect is defined at the *individual level*.
  - There is no "data generating process" or functional form.
  - The causal effect is defined relative to something else, so a counterfactual must be conceivable (did Germany cause the second world war?).
  - Are there any substantive assumptions made here so far?

Potential Outcomes

Idea: A causal claim is (in part) a claim about something that did not happen. This makes it metaphysical.

Potential Outcomes

Now that we have a concept of causal effects available, let's answer two **questions**:

- TRANSITIVITY: If for a given unit A causes B and B causes C, does that mean that A causes C?

Potential Outcomes

Now that we have a concept of causal effects available, let's answer two **questions**:

- TRANSITIVITY: If for a given unit A causes B and B causes C, does that mean that A causes C?
- A boulder is flying down a mountain. You duck. This saves your life.
- So the boulder caused the ducking and the ducking caused you to survive.
- So: *did the boulder cause you to survive?*

Potential Outcomes

CONNECTEDNESS Say A causes B — does that mean that there is a spatiotemporally continuous sequence of causal intermediates?

Potential Outcomes

CONNECTEDNESS Say A causes B — does that mean that there is a spatiotemporally continuous sequence of causal intermediates?

- Person A is planning some action Y; Person B sets out to stop them; person X intervenes and prevents person B from stopping person A. In this case Person A may complete their action, producing Y, without any knowledge that B and X even exist; in particular B and X need not be anywhere close to the action. So: *did X cause Y*?

Causal claims: Contribution or attribution?

The counterfactual model is all about contribution, not attribution, except in a very conditional sense.

- Focus is on non-rival contributions
- Not: what caused Y but what is the effect of X?
- At most it provides a conditional account

Causal claims: Contribution or attribution?

Consider an outcome Y that might depend on two causes X_1 and X_2:

$$Y(0,0) = 0$$
$$Y(1,0) = 0$$
$$Y(0,1) = 0$$
$$Y(1,1) = 1$$

What caused Y? Which cause was most important?

Causal claims: Contribution or attribution?

The counterfactual model is about attribution in a very conditional sense.

- Focus is on non-rival contributions
- Not: what caused Y but what is the effect of X?
- At most it provides a conditional account
- This is problem for research programs that define "explanation" in terms of figuring out the things that cause Y
- Real difficulties conceptualizing what it means to say one cause is more important than another cause. What does that mean?

Causal claims: Contribution or attribution?

Erdogan's increasing authoritarianism was the most important reason for the attempted coup

- More important than Turkey's history of coups?
- What does that mean?

Causal claims: No causation without manipulation

- Some seemingly causal claims not admissible.
- To get the definition off the ground, manipulation must be imaginable (whether practical or not)
- This renders thinking about effects of race and gender difficult
- What does it mean to say that Aunt Pat voted for Brexit because she is old?

Causal claims: No causation without manipulation

- Some seemingly causal claims not admissible.
- To get the definition off the ground, manipulation must be imaginable (whether practical or not)
- This renders thinking about effects of race and gender difficult
- **Compare**: What does it mean to say that Southern counties voted for Brexit because they have many old people?

Causal claims: Causal claims are everywhere

- Jack exploited Jill
- It's Jill's fault that bucket fell
- Jack is the most obstructionist member of Congress
- Melania Trump stole from Michelle Obama's speech
- Activists need causal claims

Causal claims: What is actually seen?

- We have talked about what's potential, now what do we *observe*?
- Say Z_i indicates whether the unit i is assigned to treatment $(Z_i = 1)$ or not $(Z_i = 0)$. It describes the treatment process. Then what we observe is:

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

This is sometimes called a "switching equation"

In `DeclareDesign` Y is realised from potential outcomes and assignment in this way using `reveal_outcomes`

Causal claims: What is actually seen?

- Say Z is a random variable, then this is a sort of data generating process. BUT the key thing to note is
  - Y_i is random but the randomness comes from Z_i — the potential outcomes, $Y_i(1)$, $Y_i(0)$ are fixed
  - Compare this to a regression approach in which Y is random but the X's are fixed. eg:

$$Y \sim N(\beta X, \sigma^2) \text{ or } Y = \alpha + \beta X + \epsilon, \epsilon \sim N(0, \sigma^2)$$

Causal claims: The estimand and the rub

- The causal effect of Treatment (relative to Control) is:

$$\tau_i = Y_i(1) - Y_i(0)$$

- This is what we want to estimate.
- BUT: We never can observe both $Y_i(1)$ and $Y_i(0)$!
- This is the **fundamental problem** (@holland1986statistics)

Causal claims: The rub and the solution

- Now for some magic. We really want to estimate:

$$\tau_i = Y_i(1) - Y_i(0)$$

- BUT: We never can observe both $Y_i(1)$ and $Y_i(0)$
- Say we lower our sights and try to estimate an *average* treatment effect:

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

- Now make use of the fact that

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

- In words: *The average of differences is equal to the difference of averages*; here, the average treatment effect is equal to the difference in average outcomes in treatment and control units.
- The magic is that *while we can't hope to measure the differences; we*

Causal claims: The rub and the solution

- So we want to estimate $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$.
- We know that we can estimate averages of a quantity by taking the average value from a random sample of units
- To do this here we need to select a random sample of the $Y(1)$ values and a random sample of the $Y(0)$ values, in other words, we **randomly assign** subjects to treatment and control conditions.
- When we do that we can in fact estimate:

$$\mathbb{E}_N[Y_i(1)|Z_i = 1) - \mathbb{E}_N(Y_i(0)|Z_i = 0]$$

which in expectation equals:

$$\mathbb{E}[Y_i(1)|Z_i = 1 \text{ or } Z_i = 0] - \mathbb{E}[Y_i(0)|Z_i = 1 \text{ or } Z_i = 0]$$

- This highlights a deep connection between **random assignment** and **random sampling**: when we do random assignment *we are in fact randomly sampling from different possible worlds*.

Causal claims: The rub and the solution

This provides a **positive argument** for causal inference from randomization, rather than simply saying with randomization "everything else is controlled for"

Let's discuss:

- *Does the fact that an estimate is unbiased mean that it is right?*
- *Can a randomization "fail"?*
- *Where are the covariates?*

Idea: random assignment is random sampling from potential worlds: to understand anything you find, you need to know the sampling weights

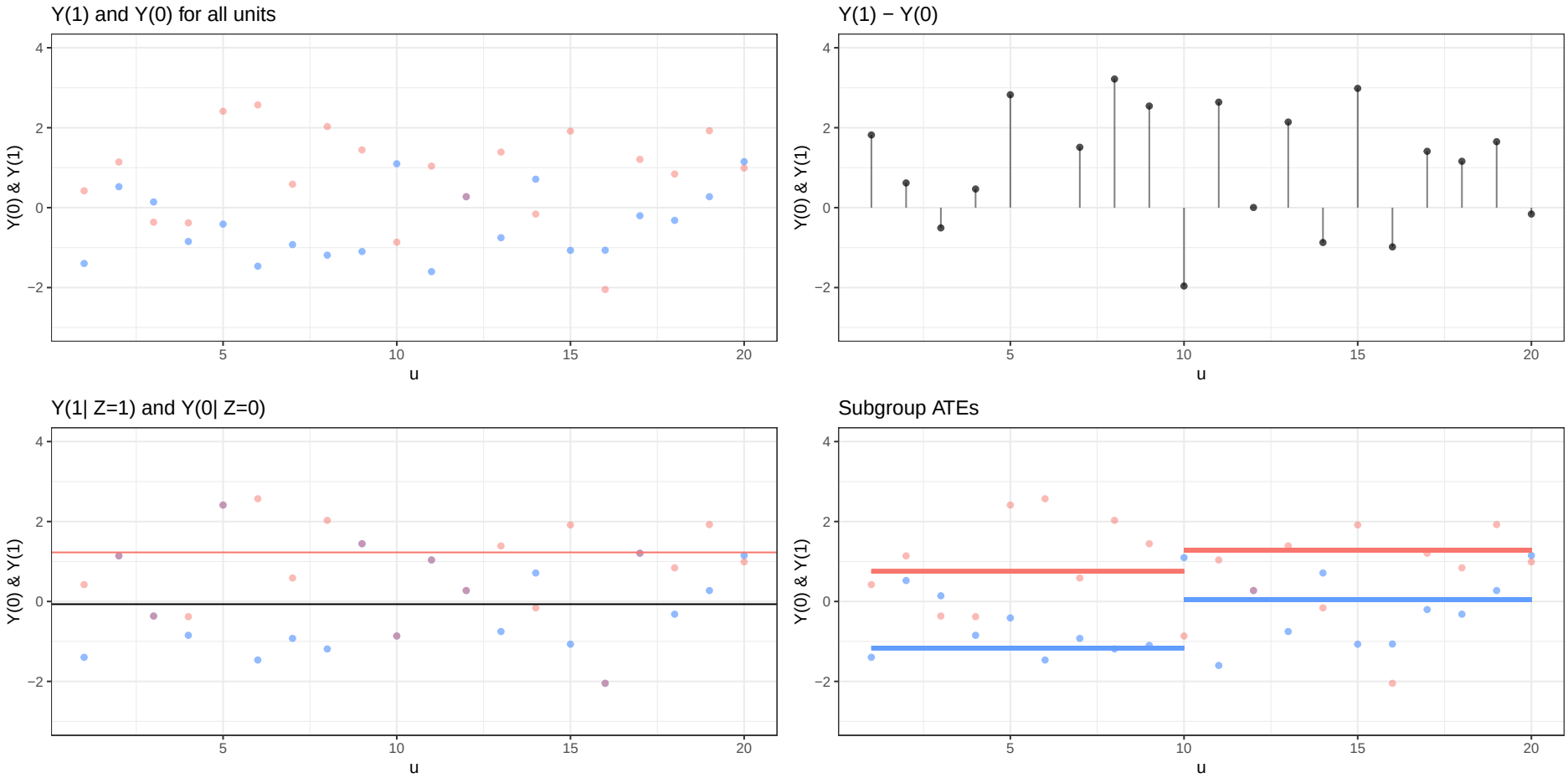
Reflection

Idea: We now have a *positive* argument for claiming unbiased estimation of the average treatment effect following random assignment

But is the average treatment effect a quantity of *social scientific* interest?

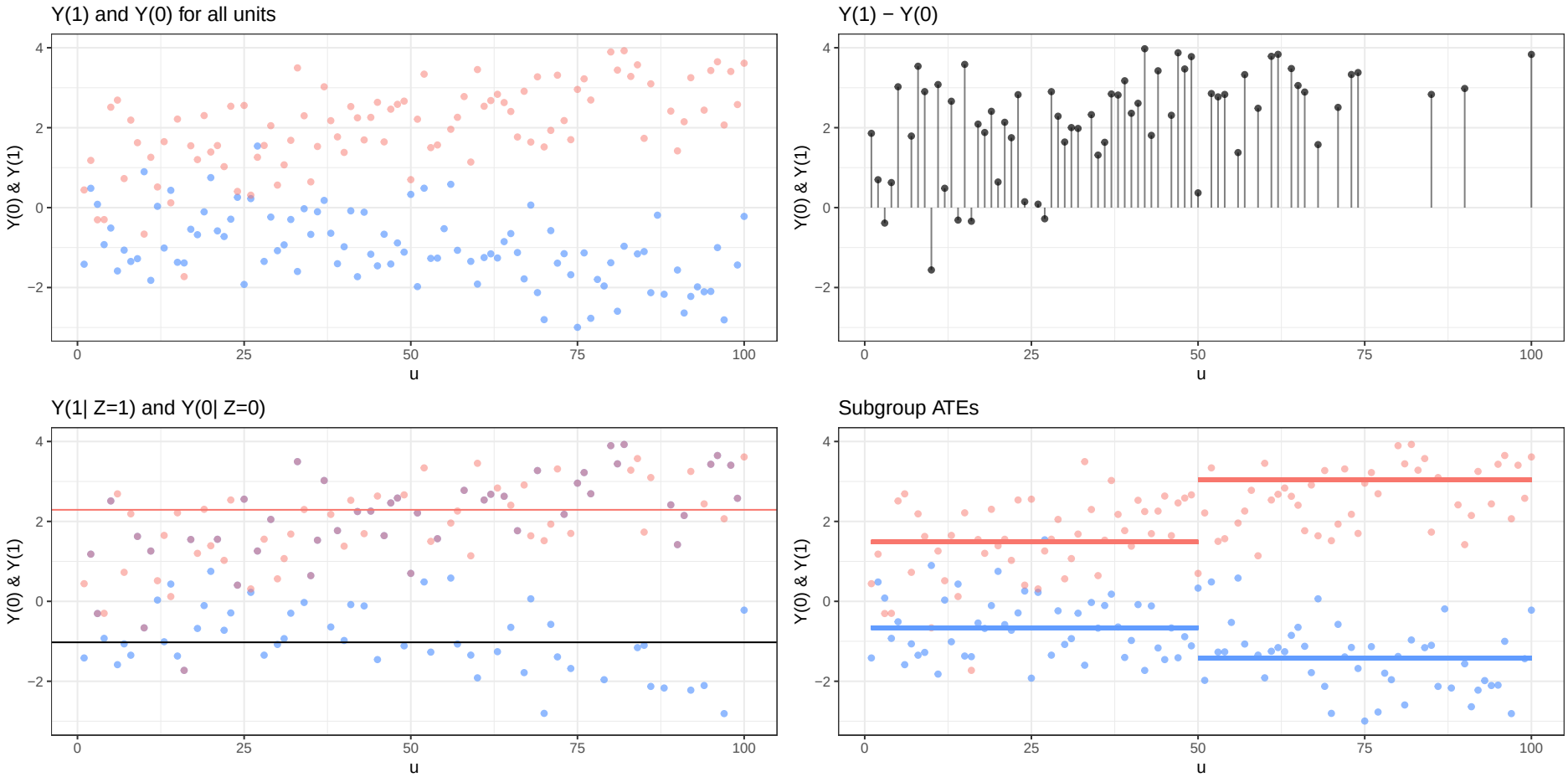
Potential outcomes: why randomization works

The average of the differences $\approx$ difference of averages



Potential outcomes: heterogeneous effects

The average of the differences $\approx$ difference of averages



Potential outcomes: heterogeneous effects

Question: $\approx$ or $=$?

Exercise your potential outcomes 1

Consider the following potential outcomes table:

Unit	Y(0)	Y(1)	τ_i
1	4	3	
2	2	3	
3	1	3	
4	1	3	
5	2	3	

Questions for us: What are the unit level treatment effects? What is the average treatment effect?

Exercise your potential outcomes 2

Consider the following potential outcomes table:

In treatment?	Y(0)	Y(1)
Yes		2
No	3	
No	1	
Yes		3
Yes		3
No	2	

Questions for us: Fill in the blanks.

- Assuming a constant treatment effect of $+1$
- Assuming a constant treatment effect of -1
- Assuming an *average* treatment effect of 0

What is the actual treatment effect?

Subsection 2

Endogeneous subgroups

Endogeneous Subgroups

Experiments often give rise to endogenous subgroups. The potential outcomes framework can make it clear why this can cause problems.

Heterogeneous Effects with Endogeneous Categories

- Problems arise in analyses of subgroups when the categories themselves are affected by treatment
- Example from our work:
  - You want to know if an intervention affects reporting on violence against women
  - You measure the share of all subjects that experienced violence that file reports
  - The problem is that which subjects experienced violence is itself a function of treatment

Heterogeneous Effects with Endogeneous Categories

It is possible that in truth no one's reporting behavior has changed, what has changed is the propensity of people with different propensities to report to experience violence:

```
\begin{table} \scriptsize
    \centering
    \begin{tabular}{rcc|cc|cc}

        & \multicolumn{ 2}{c}{Violence(Treatment)} & \multicol

        &       V(0) &       V(1) &     R(0,1) &     R(1,1) &

        Type 1 (reporter) &          1 &          1 &

        Type 2 (non reporter) &          1 &          0 &
    \end{tabular}
\end{table}
```

Heterogeneous Effects with Endogeneous Categories

```
|                            | **Violence(Treatment)**                 | **F
```

|———————–|—————————–|———————————|———————
————|————————————|————————————|———————
—–| | | V(0) | V(1) | R(0,1) | R(1,1) | R(0,0) | R(1,0) |
|———————–|—————————–|———————————|———————
————|————————————|————————————|———————
—–| | Type 1 (reporter) | 1 | 1 | 1 | 1 | 0 | 0 | | Type 2 (non reporter)| 1 |
0 | 0 | 0 | 0 | 0 |

Expected reporting given violence in control = Pr(Type 1)

Expected reporting given violence in treatment = 100%

Question: What is the actual effect of treatment on the propensity to report violence?

Heterogeneous Effects with Endogeneous Categories

It is possible that in truth no one's reporting behavior has changed, what has changed is the propensity of people with different propensities to report to experience violence:

	Reporters		Non reporters		
	Experience Violence		Experience Violence		
	No	Yes	No	Yes	% Report
Control	25	25	25	25	$\frac{25}{25+25} = 50\%$
Treatment	25	25	50	0	$\frac{25}{25+0} = 100\%$

Heterogeneous Effects with Endogeneous Categories

This problem can arise as easily in seemingly simple field experiments. Example:

- In one study we provided constituents with information about performance of politicians
- we told politicians in advance so that they could take action
- we wanted to see whether voters punished poorly performing politicians
- what's the problem?

Heterogeneous Effects with Endogeneous Categories

Question for us:

Setting:

```
* Quotas for women are randomly placed in a set of constit
* In year 2 these quotas are then lifted.
```

Questions Which problems face an endogenous subgroup issue?:

1. You want to estimate the likelihood that a woman will stand for reelection in treatment versus control areas in year 2.
2. You want to estimate how much incumbents are more likely to be reelected in treatment versus control areas in year 2.
3. You want to estimate how much treatment areas have more relected incumbents in elections in year 2 compared to control.

Heterogeneous Effects with Endogeneous Categories

In such cases you can:

- Examine the joint distribution of multiple outcomes
- Condition on pretreatment features only
- Engage in mediation analysis

Missing data can create an endogeneous subgroup problem

- It is well known that missing data can undo the magic of random assignment.
- One seemingly promising approach is to match into pairs *ex ante* and drop pairs together *ex post*.
- Say potential outcomes looked like this (four units divided into two pairs):

Pair	I	I	II	II	
Unit	1	2	3	4	Average
Y(0)	0	0	0	0	
Y(1)	-3	1	1	1	
τ	-3	1	1	1	

Missing data

- Say though that cases are likely to drop out of the sample if things go badly (eg they get a negative score or die)
- Then you might see no attrition in cases in which people that are likely to drop out if treated do not get treated.
- You might assume you have no problem (after all, no attrition).
- No missing data when the normal cases happens to be selected

Pair	I	I	II	II	
Unit	1	2	3	4	Average
Y(0)	0		0		0
Y(1)		1		1	1
$\widehat{\tau}$					1

Missing data

- But in cases in which you have attrition, dropping the pair doesn't necessarily help.
- The problem is potential missingness still depends on potential outcomes
- The kicker is that the method can produce bias even if (*in fact*) there is no attrition!

Missing data when the vulnerable cases happens to be selected

Pair	I	I	II	II	
Unit	1	2	3	4	Average
Y(0)		[0]	0		0
Y(1)	[-3]			1	1
$\widehat{\tau}$					1

Missing data

Note: The right way to think about this is that bias is a property of the strategy over possible realizations of data and not normally a property of the estimator conditional on the data.

Multistage games

Multistage games can also present an endogenous group problem since collections of late stage players facing a given choice have been created by early stage players.

Multistage games

Question: Does **visibility** alter the extent to which subjects follow norms to punish antisocial behavior (and reward prosocial behavior)? Consider a trust game in which we are interested in how information on receivers affects their actions

Table 8: Return rates given investments under different conditions

		% invested	Average % returned	
		(average)	...when 10% invested	...when 50% invested
Treatment	Masked information on respondents	30% (avg)	20%	40%
	Full information on respondents	30% (avg)	0%	60%

What do we think? Does visibility make people react more to investments?

Multistage games

Imagine you could see all the potential outcomes, and they looked like this:

Table 9: Potential outcomes with (and without) identity protection

		Responder's return decision (given type)						Avg.
		Nice 1	Nice 2	Nice 3	Mean 4	Mean 4	Mean 6	
Offerer behavior	Invest 10%:	60%	60%	60%	0%	0%	0%	30%
	Invest 50%:	60%	60%	60%	0%	0%	0%	30%

Conclusion: Both the offer and the information condition are **completely irrelevant** for all subjects.

Multistage games

Unfortunately you only see a sample of the potential outcomes, and that looks like this:

Table 10: Outcomes when respondent **is visible**

		Responder's return decision (given type)						Avg.
		Nice 1	Nice 2	Nice 3	Mean 4	Mean 4	Mean 6	
Offerer behavior	Invest 10%:				0%	0%	0%	0%
	Invest 50%:	60%	60%	60%				60%

False Conclusion: When not protected, responders condition behavior *strongly* on offers (because offerers can select on type accurately)

Multistage games

Unfortunately you only see a sample of the potential outcomes, and that looks like this:

Table 11: Outcomes when respondent **is not visible**

		Responder's return decision (given type)						Avg.
		Nice 1	Nice 2	Nice 3	Mean 4	Mean 4	Mean 6	
Offerer behavior	Invest 10%:			60%		0%	0%	20%
	Invest 50%:	60%	60%		0%			40%

False Conclusion: When protected, responders condition behavior less strongly on offers (because offerers can select on type less accurately)

Multistage games

What to do?

Solutions?

1. Analysis *could* focus on the effect of treatment on respondent behavior, directly.
   - This would get the correct answer but to a different question [Does information affect the share of contributions returned by subjects on average? No]
2. **Strategy method** can sometimes help address the problem, **but** that is also (a) changing the question and (b) putting demands on respondent imagination and honesty
3. First mover action could be **directly manipulated**, but unless deception is used that is also changing the question
4. First movers could be **selected** because they act in predictable ways (bordering on deception?)

Idea: Proceed with extreme caution when estimating effects beyond the

Subsection 3

DAGs

Subsection 4

Key insight

Key insight

The most powerful results from the study of DAGs are procedures for figuring out when conditioning aids or hinders causal identification.

- You can read off a **confounding** variable from a DAG.
  - You have to condition on such a variable for causal identification.
- You can read off **"colliders"** from a DAG
  - Sometimes you have *avoid* conditioning on these
- Sometimes a variable might be both, so
  - you have to condition on it
  - you have to avoid conditioning on it
  - Ouch.

Subsection 5

Key resource

Key resource

- Pearl's book *Causality* is the key reference. @pearl2009causality (Though see also older work such as @pearl1985graphoids)
- There is a lot of excellent material on Pearl's page http://bayes.cs.ucla.edu/WHY/
- See also excellent material on Felix Elwert's page http://www.ssc.wisc.edu/~felwert/causality/?page_id=66

Subsection 6

Challenge for us

Challenge for us

- Say you don't like graphs. Fine.
- Consider this causal structure:
  - $Z = f_1(U_1, U_2)$
  - $X = f_2(U_2)$
  - $Y = f_3(X, U_1)$

Say Z is temporally prior to X; it is correlated with Y (because of U_1) and with X (because of U_2).

Question: Would it be useful to "control" for Z when trying to estimate the effect of X on Y?

Subsection 7

Challenge for us

Challenge for us

- Say you don't like graphs. Fine.
- Consider this causal structure:
  - $Z = f_1(U_1, U_2)$
  - $X = f_2(U_2)$
  - $Y = f_3(X, U_1)$

Question: Would it be useful to "control" for Z when trying to estimate the effect of X on Y?

Answer: Hopefully by the end of today you should see that that the answer is obviously (or at least, plausibly) "no."

Subsection 8

Conditional independence

Conditional independence

Variable sets A and B are conditionally independent, given C if for all a, b, c:

$$\Pr(A = a|C = c) = \Pr(A = a|B = b, C = c)$$

Informally; given C, knowing B tells you nothing more about A.

Subsection 9

Causal graphs basics 1

Causal graphs basics 1

- Consider a situation with variables $X_1, X_2, ... X_n$
- The probability of outcome x can always be written in the form $P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1)(X_3 = x_3|X_1 = x_1, X_2 = x_2)....$
- This can be done with any ordering of variables.
- However the representation can be greatly simplified if you can make use of a set of "parentage" relationships
- Given an ordering of variables, the **Markovian parents** of variable X_j are the minimal set of variables such that when you condition on these, X_j is independent of all other prior variables in the ordering
- In this case we can write: $P(x) = \prod_j (x_j|pa_j)$
- No graphs yet

Subsection 10

Causal graphs basics 2

Causal graphs basics 2

- We want to use causal graphs to represent these relations of conditional independence.
- Informally, an arrow, $A \rightarrow B$ means that A is a cause of B: that is, under some conditions, a change in A produces a change in B.
  - Arrows carry no information about the type of effect; e.g. sign, size, or whether different causes are complements or substitutes
- We say that arrows point from *parents* to *children*, and by extension from *ancestors* to *descendants*.
- These are parents *on the graph*; but we will connect them to Markovian parents in a probability distribution P.

Subsection 11

Causal graphs basics 2

Causal graphs basics 2

- A DAG is just a graph in which some or all nodes are connected by *directed* edges (arrows) and there are no cyclical paths along these directed edges.
- Consider a DAG, G, and consider the ancestry relations implied by G: the distribution P is *Markov relative to the graph* G if every variable is independent of its nondescendants (in G) conditional on its parents (in G).
  - This is the **Markov condition**: conditional on its parents, a variable is independent of its non-descendants.
- OK now we have a link from probability distributions to graphs. But we have not talked about causality.

Subsection 12

Causal graphs basics 3

Causal graphs basics 3

We want the graphs to be able to represent the effects of interventions.

Pearl uses do notation to capture this idea.

$$\Pr(X_1, X_2, ... | do(X_j = x_j))$$

or

$$\Pr(X_1, X_2, ... | \hat{x}_j)$$

denotes the distribution of X when a particular node (or set of nodes) is intervened upon and forced to a particular level, x_j.

Subsection 13

Causal graphs basics 3

Causal graphs basics 3

Note, in general:

$$\Pr(X_1, X_2, ... | do(X_j = x'_j)) \neq \Pr(X_1, X_2, ... | X_j = x'_j)$$

as an example we might imagine a situation where for men binary X always causes $Y = 1$ and for women $Y = 1$ regardless of X. We imagine that $X = 1$ for men only.

In that case $\Pr(Y = 1|X = 1) = 1$ but $\Pr(Y = 1|do(X = 1)) = .5$

Subsection 14

Causal graphs basics 3

Causal graphs basics 3

- Let P_z denote the resulting distribution on all variables that arises when vector Z is "set" (forced, controlled...) to the value z. That is when we have `do(Z=z)`.
- Let P_* denote the set of all such distributions that can result from any set of interventions on variables.
- A DAG, G, is a **causal Bayesian network compatible with** P_* iff, for all interventions z:
  1. P_z is Markov relative to G
  2. $P_z(x_i) = 1$ for all x_i consistent with z
  3. $P_z(x_j|pa_j) = P(x_j|pa_j)$ for all $x_j \notin Z$ when pa_j is consistent with z

Subsection 15

Causal graphs basics 3

Causal graphs basics 3

- That all means that the probability distribution resulting from setting some set X_i to $\hat{x'}_i$ (i.e. `do(X=x')`) is:

$$P_{\hat{x}_i} = P(x_1, x_2, ... x_n | \hat{x}_i) = \prod_{-i} P(x_j | pa_j) \mathbb{1}(x_i = x'_i)$$

This means that there is only probability mass on vectors in which $x_i = x'_i$ (reflecting the success of control) and all other variables are determined by their parents, given the values that have been set for x_i.

Subsection 16

Conditional Independence and d-separation

Conditional Independence and d-separation

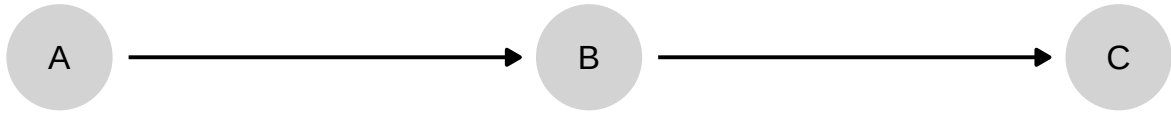
- We now have a well defined sense in which the arrows on a graph represent a causal structure and capture the conditional independence relations implied by the causal structure.
- Of course any graph might represent many different probability distributions P
- We can now start reading off from a graph when there is or is not conditional independence between sets of variables

Subsection 17

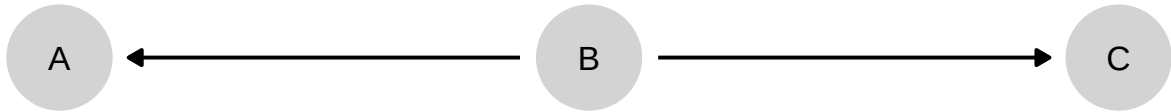
Conditional independence on paths

Conditional independence on paths

(1) A path of arrows pointing in the same direction



(2) A forked path



(3) An inverted fork (collision)

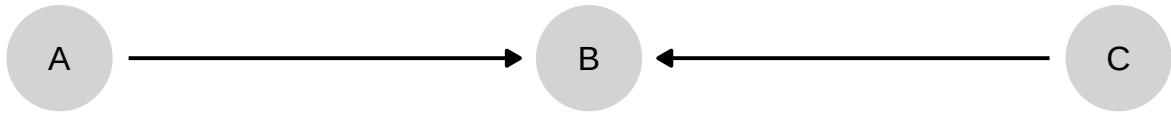


Figure 2: Three elemental relations of conditional independence.

Subsection 18

Conditional independence

Conditional independence

A and B are *conditionally independent*, given C if on *every* path between A and B:

- there is some chain ($\bullet \rightarrow \bullet \rightarrow \bullet$ or $\bullet \leftarrow \bullet \leftarrow \bullet$) or fork ($\bullet \leftarrow \bullet \rightarrow \bullet$) with the central element in C,

or

- there is an inverted fork ($\bullet \rightarrow \bullet \leftarrow \bullet$) with the central element (and its descendants) *not* in C

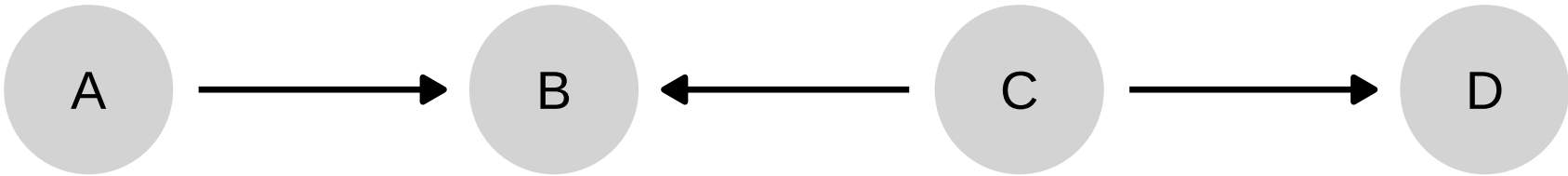
Notes:

- In this case we say that A and B are d-separated by C.
- A, B, and C can all be sets
- Note that a path can involve arrows pointing any direction
  $\bullet \rightarrow \bullet \rightarrow \bullet \leftarrow \bullet \rightarrow \bullet$

Subsection 19

Test yourself

Test yourself



Are A and D unconditionally independent:

- if you do not condition on anything?
- if you condition on B?
- if you condition on C?

Subsection 20

Back to this example

Back to this example

* $Z = f_1(U_1, U_2)$
* $X = f_2(U_2)$
* $Y = f_3(X, U_1)$

1. Let's graph this
2. Now: say we removed the arrow from X to Y
   - Would you expect to see a correlation between X and Y if you did not control for Z
   - Would you expect to see a correlation between X and Y if you did control for Z

Subsection 21

From graphs to Causal Models

From graphs to Causal Models

A "**causal model**" is:

1.1: An ordered list of n endogenous nodes, $\mathcal{V} = (V^1, V^2, \ldots, V^n)$, with a specification of a range for each of them

1.2: A list of n exogenous nodes, $\Theta = (\theta^1, \theta^2, \ldots, \theta^n)$

2: A list of n functions $\mathcal{F} = (f^1, f^2, \ldots, f^n)$, one for each element of $\mathcal{V}$ such that each f^i takes as arguments θ^i as well as elements of $\mathcal{V}$ that are *prior* to V^i in the ordering

and

3: A probability distribution over Θ

Subsection 22

From graphs to Causal Models

From graphs to Causal Models

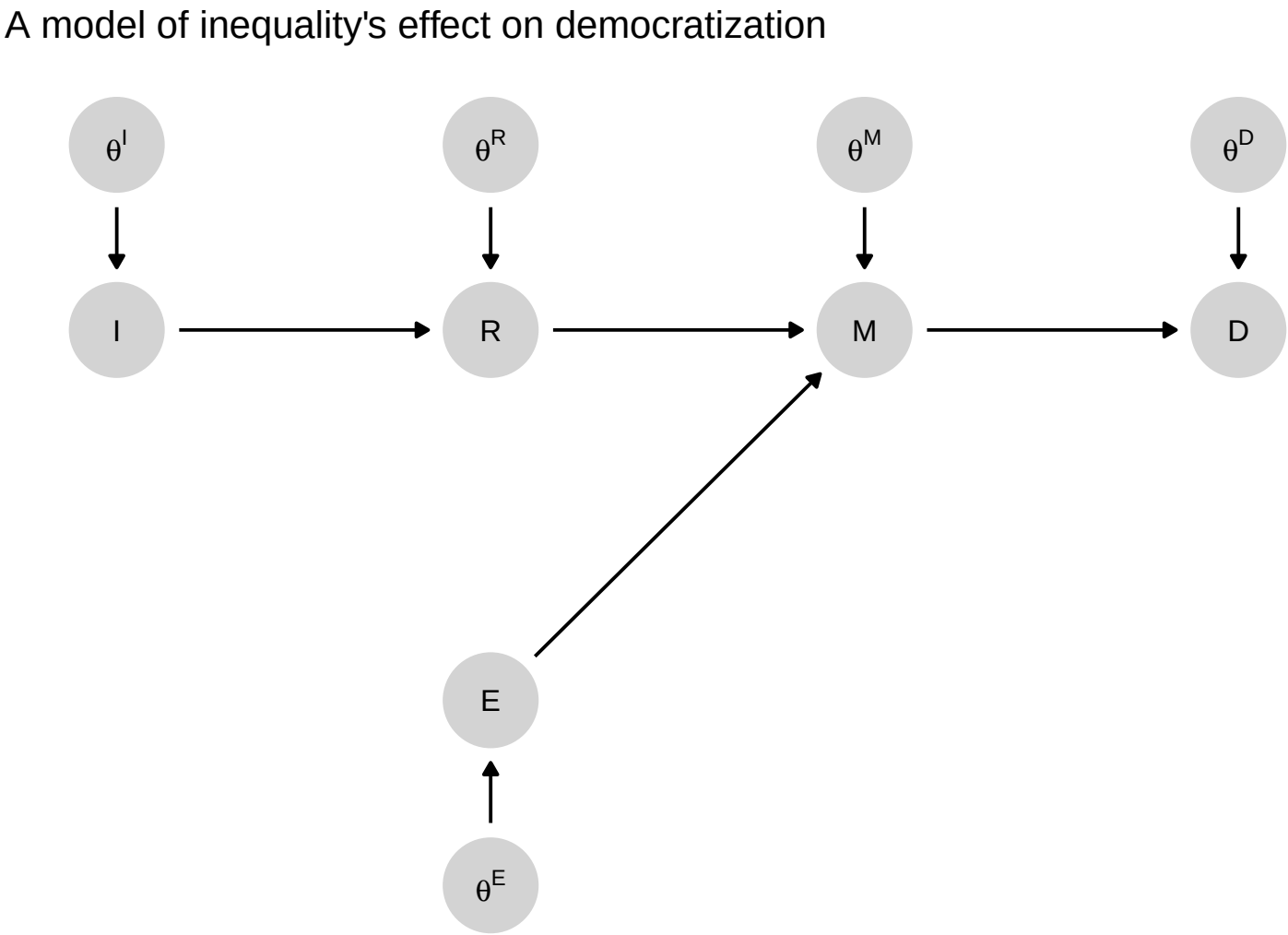


Figure 3: A simple causal model in which high inequality (I) affects democratization (D) via redistributive demands (R) and mass mobilization (M)

Effects on a DAG

Learning about effects *given* a model means learning about F and *also* the distribution of shocks (Θ).

For discrete data this can be reduced to a question about learning about the distribution of Θ only.

Subsection 23

Recap: Key features of graphs

Recap: Key features of graphs

- Directed
- Acyclic
- The missing arcs are the really important ones
- Implicitly there are shocks going into every node
- These graphs represent Nonparametric structural equation models NPSEMs
- But you cannot read off the size or direction of effects from a DAG

Recap: Ten things you need to know about causal inference

1. A causal claim is a statement about what didn't happen.
2. There is a fundamental problem of causal inference.
3. You can estimate average causal effects even if you cannot observe any individual causal effects.
4. If you know that A causes B and that B causes C, this does not mean that you know that A causes C.
5. The counterfactual model is primarily about contribution, and about attribution in a limited sense.
6. X can cause Y even if there is no "causal path" connecting X and Y.
7. Correlation is not causation.
8. X can cause Y even if X is not a necessary condition or a sufficient condition for Y.
9. Estimating average causal effects does not require that treatment and control groups are identical.
10. There is no causation without manipulation.

http://egap.org/resources/guides/causality/

Section 4

Inquiries

Subsection 1

Estimands and inquiries

Estimands and inquiries

- Your inquiry is your question and the estimand is the true (generally unknown) answer to the inquiry
- The estimand is the thing you want to estimate
- If you are estimating something you should be able to say what your estimand is
- You are responsible for your estimand. Your estimator will not tell you what your estimand is
- Just because you can calculate something does not mean that you have an estimand
- You can test a hypothesis without having an estimand

Read: II ch 4, DD, ch 7

Estimands: ATE, ATT, ATC, S-, P-, C-, ITT, LATE

Say that units are randomly assigned to treatment in different strata (maybe just one); with fixed, though possibly different, shares assigned in each stratum. Then the key estimands and estimators are:

Estimand	Estimator
$\tau_{ATE} \equiv \mathbb{E}[\tau_i]$	$\hat{\tau}_{ATE} = \sum_x \frac{w_x}{\sum_j w_j}\hat{\tau}_x$
$\tau_{ATT} \equiv \mathbb{E}[\tau_i \mid Z_i = 1]$	$\hat{\tau}_{ATT} = \sum_x \frac{p_x w_x}{\sum_j p_j w_j}\hat{\tau}_x$
$\tau_{ATC} \equiv \mathbb{E}[\tau_i \mid Z_i = 0]$	$\hat{\tau}_{ATC} = \sum_x \frac{(1-p_x) w_x}{\sum_j (1-p_j) w_j}\hat{\tau}_x$

where x indexes strata, p_x is the share of units in each stratum that is treated, and w_x is the size of a stratum.

Here:

- ATE is Average Treatment Effect (all units)
- ATT is Average Treatment Effect on the Treated

Estimands: ATE, ATT, ATC, S-, P-, C-

In addition, each of these can be targets of interest:

- for the **population**, in which case we refer to PATE, PATT, PATC and $\widehat{PATE}, \widehat{PATT}, \widehat{PATC}$
- for a **sample**, in which case we refer to SATE, SATT, SATC, and $\widehat{SATE}, \widehat{SATT}, \widehat{SATC}$

And for different subgroups,

- given some value on a covariate, in which case we refer to CATE (conditional average treatment effect)

Broader classes of estimands: LATE/CATE

The CATEs are **conditional** average treatment effects, for example the effect for men or for women. These are straightfoward.

However we might also imagine conditioning on unobservable or counterfactual features.

- The LATE (or CACE: complier average causal effect) asks about the effect of a treatment (X) on an outcome (Y) *for people that are responsive to an encouragement* (Z)

$$LATE = \frac{1}{|C|}\sum_{j \in C}(Y_j(X=1) - Y_j(X=0))$$

$$C := \{j : X_j(Z=1) > X_j(Z=0)\}$$

We will return to these in the study of instrumental variables.

Quantile estimands

Other ways to condition on potential outcomes:

- A *quantile* treatment effect: You might be interested in the difference between the median $Y(1)$ and the median $Y(0)$ (@imbens2015causal 20.3.1)
- or even be interested in the median $Y(1) - Y(0)$. Similarly for other quantiles.

Model estimands

Many inquiries are averages of individual effects, even if the groups are not known, but they do not have to be:

- The RDD estimand is a statement about what effects *would be* at a threshold; it can be defined under a model even if no actual individuals are at the threshold. We imagine average potential outcomes as a function of treatment Z and running variable X, $f(z, x)$ and define:

$$\tau_{RDD} := f(1, x^*) - f(0, x^*)$$

Distribution estimands

Many inquiries are averages of individual effects, even if the groups are not known,

But they do not have to be:

- Inquiries might relate to distributional quantities such as:
  - The effect of treatment on the variance in outcomes: $var(Y(1)) - var(Y(0))$
  - The variance of treatment effects: $var(Y(1) - Y(0))$
  - Other inequality measures (e.g. Ginis; (@imbens2015causal 20.3.2))

You might even be interested in $\min(Y_i(1) - Y_i(0))$.

Spillover estimands

There are lots of interesting "spillover" estimands.

Imagine there are three individuals and each person's outcomes depends on the assignments of all others. For instance $Y_1(Z_1, Z_2, Z_3$, or more generally, $Y_i(Z_i, Z_{i+1(\text{mod } 3)}, Z_{i+2(\text{mod } 3)})$.

Then three estimands might be:

- $\frac{1}{3}\left(\sum_i Y_i(1,0,0) - Y_i(0,0,0)\right)$
- $\frac{1}{3}\left(\sum_i Y_i(1,1,1) - Y_i(0,0,0)\right)$
- $\frac{1}{3}\left(\sum_i Y_i(0,1,1) - Y_i(0,0,0)\right)$

Interpret these. What others might be of interest?

Differnces in CATEs and interaction estimands

A difference in CATEs is a well defined estimand that might involve interventions on one node only:

- $\mathbb{E}_{\{W=1\}}[Y(X=1) - Y(X=0)] - \mathbb{E}_{\{W=0\}}[Y(X=1) - Y(X=0)]$

It captures differences in effects.

An *interaction* is an effect on an effect:

- $\mathbb{E}[Y(X=1, W=1) - Y(X=0, W=1)] - \mathbb{E}[Y(X=1, W=0) - Y(X=0, W=0)]$

Note in the latter the expectation is taken over the whole population.

Mediation estimands and complex counterfactuals

Say X can affect Y directly, or indirectly through M. then we can write potential outcomes as:

- $Y(X = x, M = m)$
- $M(X = x)$

We can then imagine inquiries of the form:

- $Y(X = 1, M = M(X = 1)) - Y(X = 0, M = M(X = 0))$
- $Y(X = 1, M = 1) - Y(X = 0, M = 1)$
- $Y(X = 1, M = M(X = 1)) - Y(X = 1, M = M(X = 0))$

Interpret these. What others might be of interest?

Mediation estimands and complex counterfactuals

Again we might imagine that these are defined with respect to some group:

- $A = \{i | Y_i(1, M(X = 1)) > Y_i(0, M(X = 0))\}$
- $\frac{1}{|A|} \sum_{i \in A} (Y(1, 1) > Y(0, 1))$

here, among those for whom X has a positive effect on Y, for what share would there be a positive effect if M were fixed at 1.

Causes of effects and effects of causes

In qualitative research a particularly common inquiry is "did $X = 1$ cause $Y = 1$?

This is often given as a probability, the "probability of causation" (though at the case level we might better think of this probabiliy as an estimate rather than an estimand):

$$\Pr(Y_i(0) = 0 | Y_i(1) = 1, X = 1)$$

Causes of effects and effects of causes

Intuition: What's the probability $X = 1$ caused $Y = 1$ in an $X = 1, Y = 1$ case drawn from a large population with the following experimental distribution:

	Y=0	Y=1	All
X=0	1	0	1
X=1	0.25	0.75	1

Causes of effects and effects of causes

Intuition: What's the probability $X = 1$ caused $Y = 1$ in an $X = 1, Y = 1$ case drawn from a large population with the following experimental distribution:

	Y=0	Y=1	All
X=0	0.75	0.25	1
X=1	0.25	0.75	1

Actual causation

Other inquiries focus on distinguishing between causes.

For the Billy Suzy problem [@hall2004two], @halpern2016actual focuses on "actual causation" as a way to distinguish between Suzy and Billy:

> *Imagine Suzy and Billy, simultaneously throwing stones at a bottle. Both are excellent shots and hit whatever they aim at. Suzy's stone hits first, knocks over the bottle, and the bottle breaks. However, Billy's stone would have hit had Suzy's not hit, and again the bottle would have broken. Did Suzy's throw cause the bottle to break? Did Billy's?*

Actual causation

Actual Causation:

1. $X = x$ and $Y = y$ both happened;
2. there is some set of variables, $\mathcal{W}$, such that if they were fixed at the levels that they *actually took* on in the case, and if X were to be changed, then Y would change (where $\mathcal{W}$ can also be an empty set);
3. no strict subset of X satisfies 1 and 2 (there is no redundant part of the condition, $X = x$).

Actual causation

- Suzy: Condition 2 is met if Suzy's throw made a difference, counterfactually speaking—with the important caveat that, in determining this, we are permitted to condition on Billy' stone not hitting the bottle.
- Billy: Condition 2 is not met.

An inquiry: for what share in a population is a possible cause an actual cause?

Subsection 2

Pearl's ladder

Subsection 3

Inquiries as statements about principal strata

Subsection 4

Identification

Identification

What it is. When you have it. What it's worth.

Identification

Informally a quantity is "identified" if it can be "recovered" once you have enough data.

Say for example average wage is x in some very large population. If I gather lots and lots of data on the wages of individuals and take the average then then my estimate will ultimately let be figure out x. If x is 1 then by estimate will end up centered on \$1. If it is \$2 it will end up centered on \$2.

Essentially: Each underlying value produces a unique data distribution. When you see that distribution you recover the parameter.

Identification (Example without identification)

Informally a quantity is "identified" if it can be "recovered" once you have enough data.

- Say for example average wage is x^m for men and x^w for women (in some very large population).
- If I gather lots and lots of data on the wages of (male and female) couples, e.g. $x_i^c = x_i^m + x_i^w$ then, although this will be informative, it will never be sufficient to recover x^m for men and x^w.
- I can recover x^c, but there are too many combinations of possible values of x^m and x^w consistent with the observed data.

Identification : Goal

Our goal in causal inference is to estimate quantities such as:

$$\Pr(Y|\hat{x})$$

where $\hat{x}$ is interpreted as X set to x by "external" control. Equivalently: $do(X = x)$ or sometimes $X \leftarrow x$.

If this quantity is **identifiable** then we can recover it with infinite data.

If it is not identifiable, then, even in the best case, we are not guaranteed to get the right answer.

Are there general rules for determining whether this quantitiy can be identified? Yes.

Identification : Goal

Note first, identifying

$$\Pr(Y|x)$$

is easy.

But we are not interested in identifying the distribution of Y given observed values of x, but rather, the distribution of Y if X is *set* to x.

Subsection 5

Levels and effects

Levels and effects

If we can identify the controlled disteribution we can calculate other causal quantities of interest.

For example for a binary X, Y the causal effect of X on the probability that $Y = 1$ is:

$$\Pr(Y = 1|\hat{x} = 1) - \Pr(Y = 1|\hat{x} = 0)$$

Again, **this is not the same as**:

$$\Pr(Y = 1|x = 1) - \Pr(Y = 1|x = 0)$$

It's the difference between seeing and doing.

When to condition? What to condition on?

The key idea is that you want to find a set of variables such that when you condition on these you get what you would get if you used a do operation.

Intuition:

- You could imagine creating a "mutilated" graph by removing all the arrows leading *out* of X
- Then select a set of variables, Z, such that X and Y are d-separated by Z on the the mutilated graph
- When you condition on these you are making sure that any covariation between X and Y is covariation that is due to the effects of X

Illustration

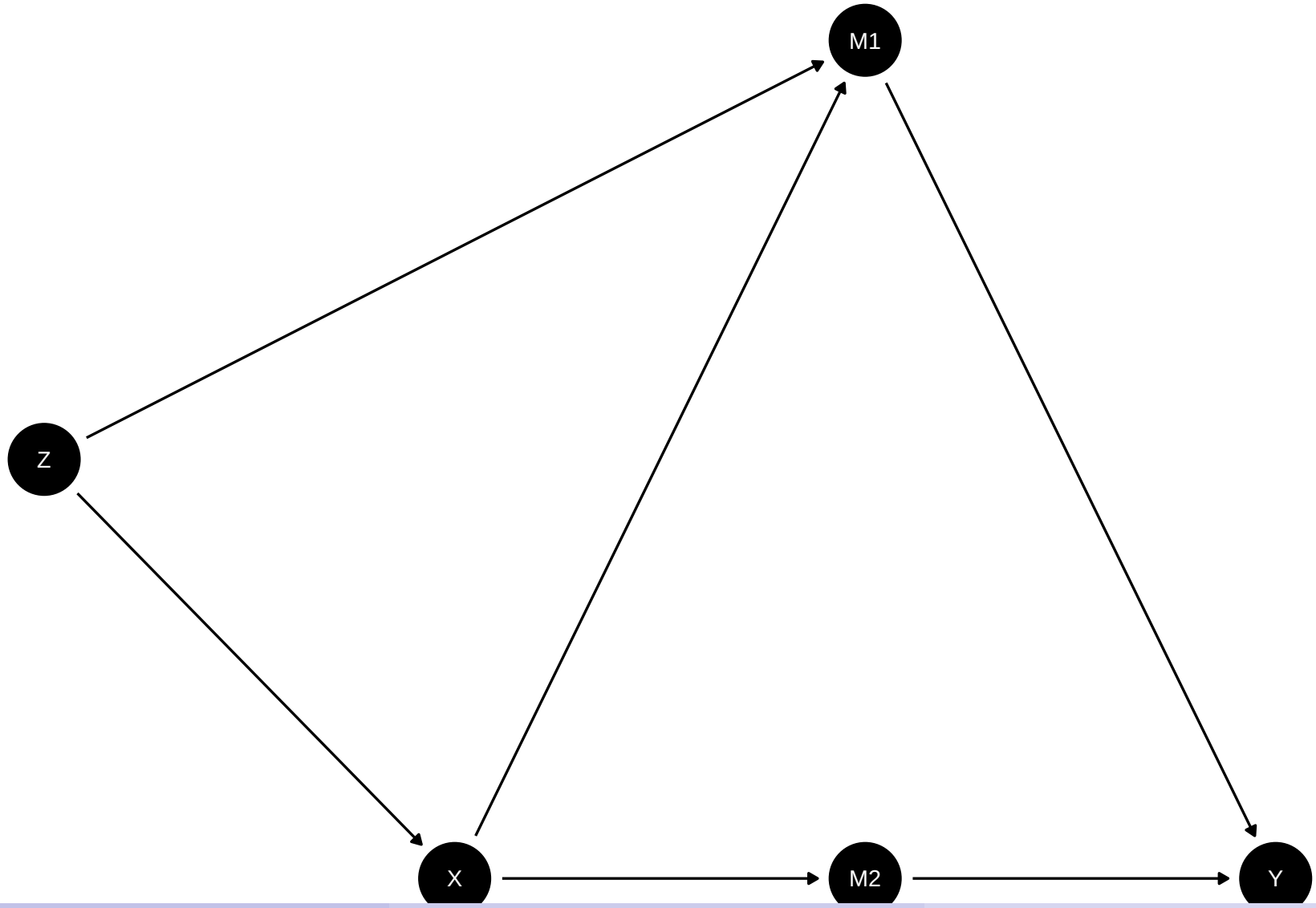


Illustration: Remove paths out

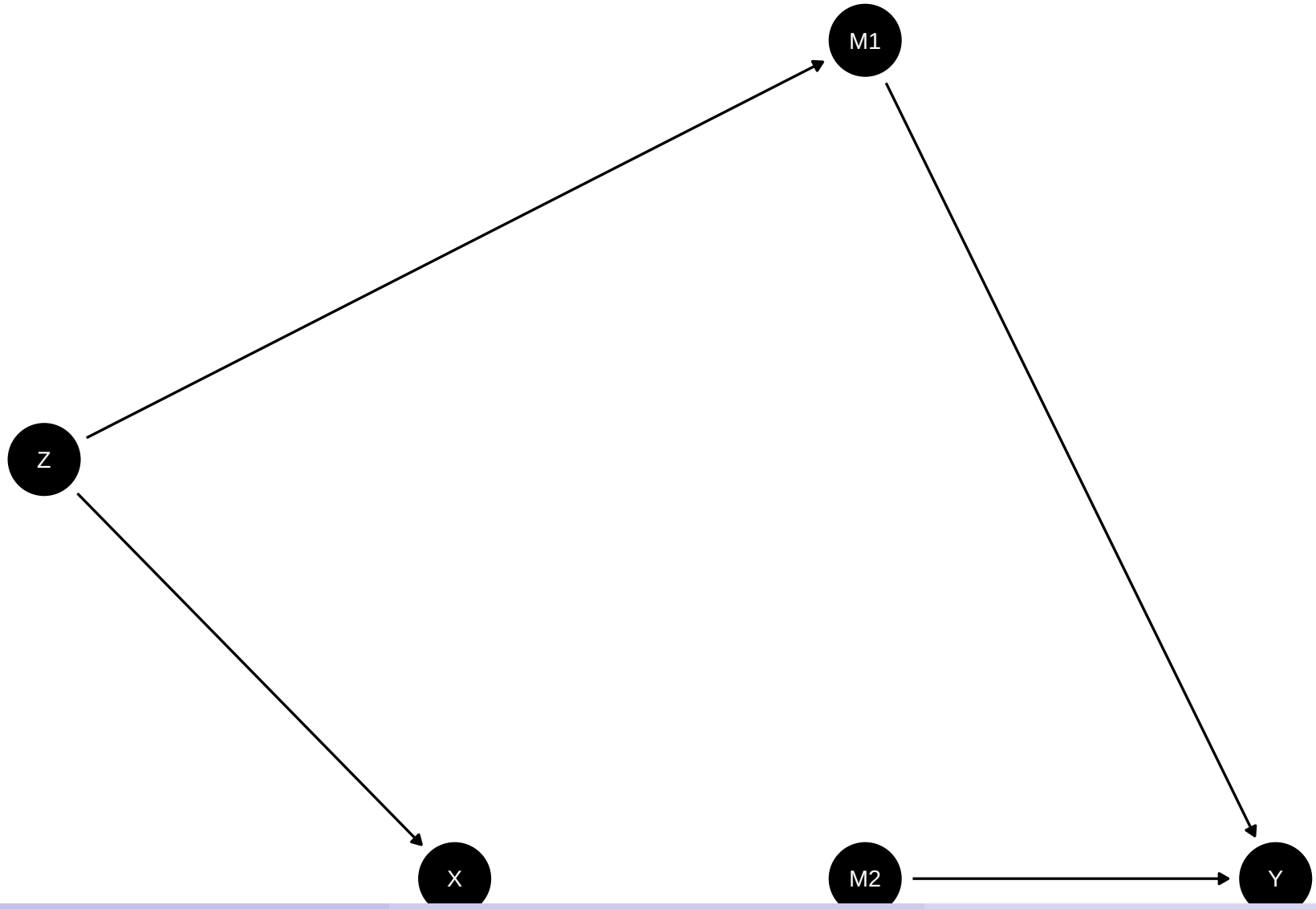


Illustration: Block backdoor path

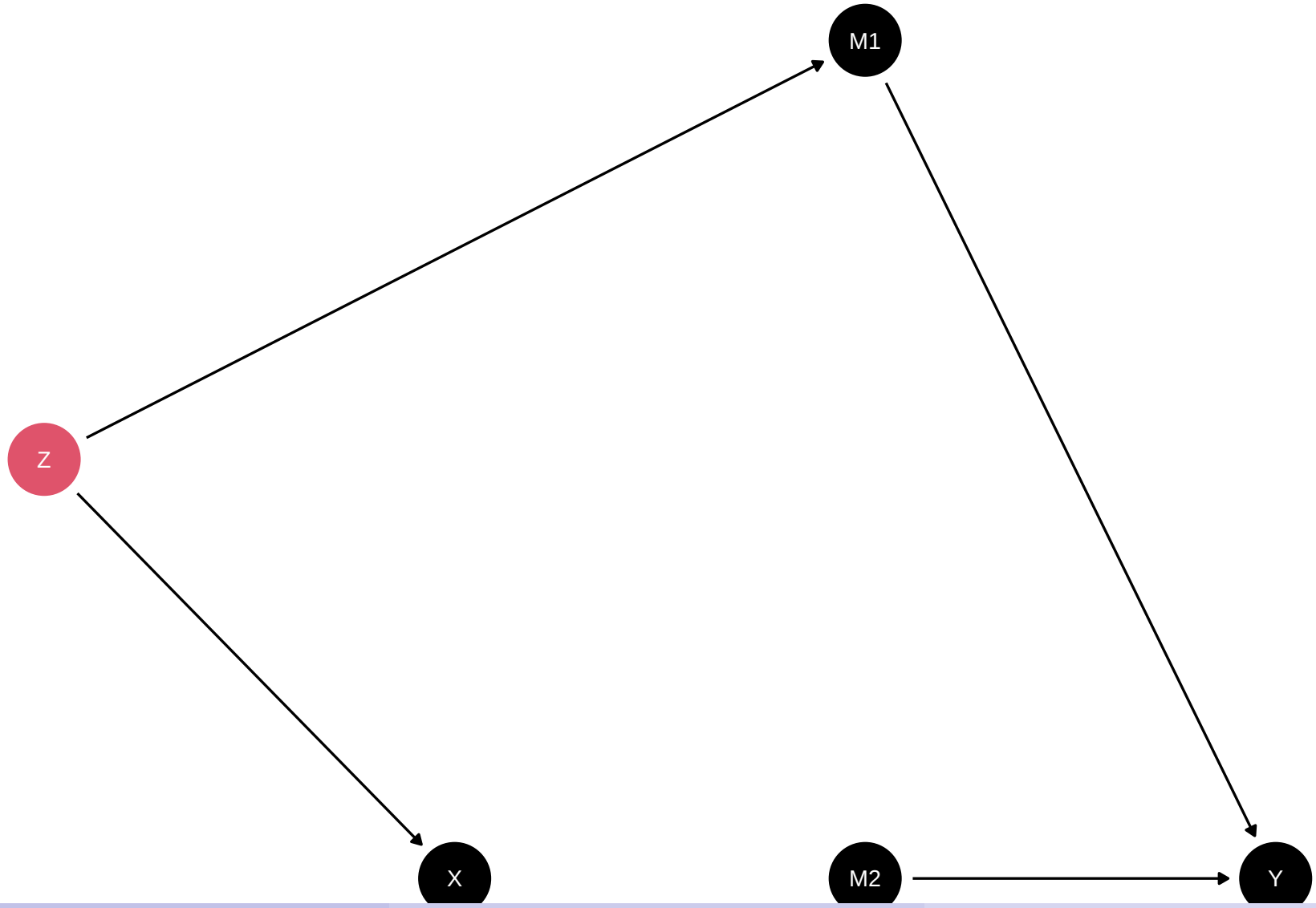
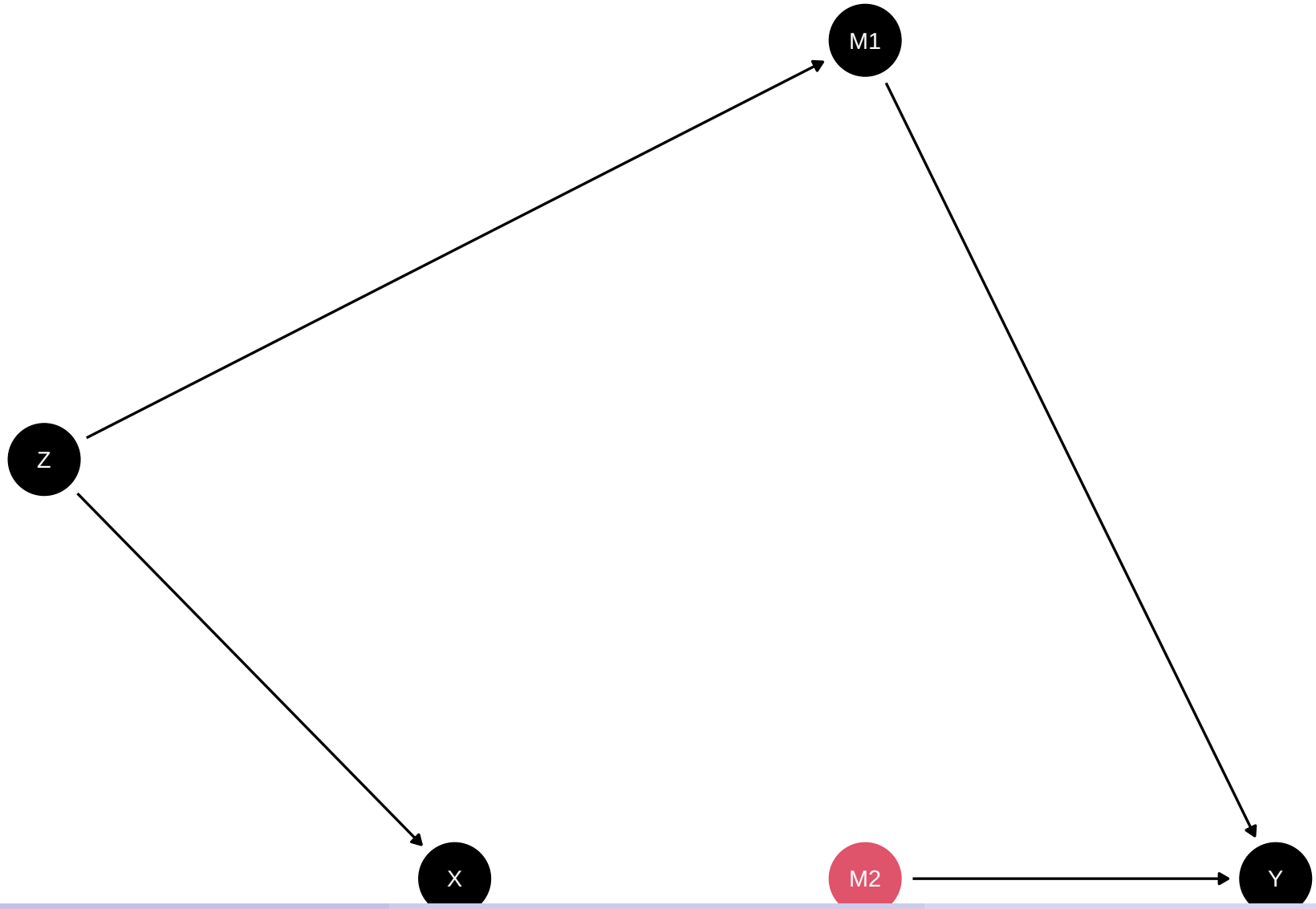


Illustration: Why not like this?



Identification

- Three results ("Graphical Identification Criteria")
  - Backdoor criterion
  - Adjustment criterion
  - Frontdoor criterion
- There are more

Backdoor Criterion: (Pearl 1995)

The **backdoor criterion** is satisfied by Z (relative to X, Y) if:

1. No node in Z is a descendant of X
2. Z blocks every **backdoor** path from X to Y (i.e. every path that contains an arrow into X)

In that case you can identify the effect of X on Y by conditioning on Z:

$$P(Y = y|\hat{x}) = \sum_z P(Y = y|X = x, Z = z)P(z)$$

(This is eqn 3.19 in Pearl (2000))

Backdoor Criterion: (Pearl 1995)

$$P(Y = y|\hat{x}) = \sum_z P(Y = y|X = x, Z = z)P(z)$$

- Note notion of a linear control of anything like that; idea really is like blocking: think lots of discrete data and no missing patterns
- Note this is a formula for a (possibly counterfactual) *level*; a counterfactual difference would be given in the obvious way by:

$$P(Y = y|\hat{x}) - P(Y = y|\hat{x}') = \sum_z P(Y = y|X = x, Z = z)P(z) - \sum_z P(Y =$$

Backdoor Proof

Following Pearl (2009), Chapter 11. Let T denote the set of parents of X: $T := pa(X)$, with (possibly vector valued) realizations t.

If the backdoor criterion is satisfied, we have:

1. Y is independent of T, given X and observed data, Z (since Z blocks backdoor paths)
2. X is independent of Z given T. (Since Z includes only nondescendents)

- From the DAG we have:

$$p(y|\hat{x}) = \sum_{t \in T} p(t)p(y|\hat{x}, t)$$

Backdoor Proof

- But we do not observe T, rather we observe Z. OK, but we can write:

$$p(y|\hat{x}) = \sum_{t \in T} p(t) \sum_{z} p(y|\hat{x}, t, z) p(z|\hat{x}, pa(X))$$

- Then using the two conditions above:
  1. replace $p(y|\hat{x}, pa(X), z)$ with $p(y, \hat{x}, z)$
  2. replace $p(z|\hat{x}, pa(X))$ with $p(z|\hat{x})$

This gives:

$$p(y|\hat{x}) = \sum_{pa(X)} p(pa(X)) \sum_{z} p(y|\hat{x}, z) p(z|pa(X))$$

Now Clean up:

$$p(y|\hat{x}) = \sum_{pa(X)} p(pa(X)) \sum_{z} p(y|\hat{x}, z) p(z|pa(X))$$

$$\leftrightarrow$$

$$p(y|\hat{x}) = \sum_{z} p(y|\hat{x}, z) \sum_{pa(X)} p(pa(X)) p(z|pa(X)) = \sum_{z} p(y|\hat{x}) p(z)$$

Adjustment criterion

See @shpitser2012validity

The adjustment criterion is satisfied by Z (relative to X, Y) if:

1. no element of Z is a descendant (in the mutilated graph[1]) of any variable $W \notin X$ which lies on a proper causal path from X to Y[2]
2. Z blocks all **noncausal paths** from X to Y

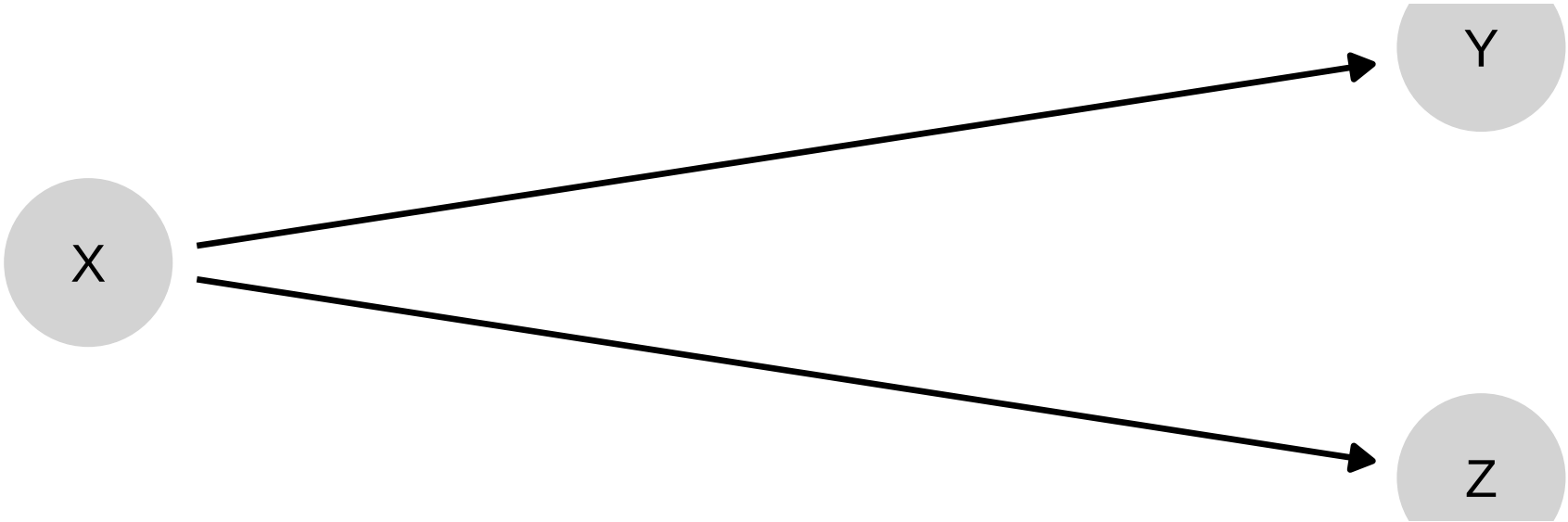
[1] remove arrows pointing into X

[2] A *proper* causal pathway nodes in X to nodes in Y only intersects X at the endpoint

These are different. Simple illustration.

Here Z satisfies the adjustment criterion but not the backdoor criterion:

Controlling for Z is OK



Z is descendant of X but it does not a descendant of a node on a path from X to Y. No harm adjusting for Z here, but not necessary either.

Frontdoor criterion

In code: Dagitty

There is a package for this

```
library(dagitty)
```

Then define a dag using dagitty syntax:

```
g <- dagitty("dag{X -> M -> Y ; Z -> X ; Z -> R -> Y}")
```

There is then a simple command to check whether two sets are d-separated by a third set:

```
dseparated(g, "X", "Y", "M")
```

```
[1] FALSE
```

```
dseparated(g, "X", "Y", c("Z","M"))
```

```
[1] TRUE
```

Dagitty: Find adjustment sets

And a simple command to identify the adjustments needed to identify the effect of one variable on another:

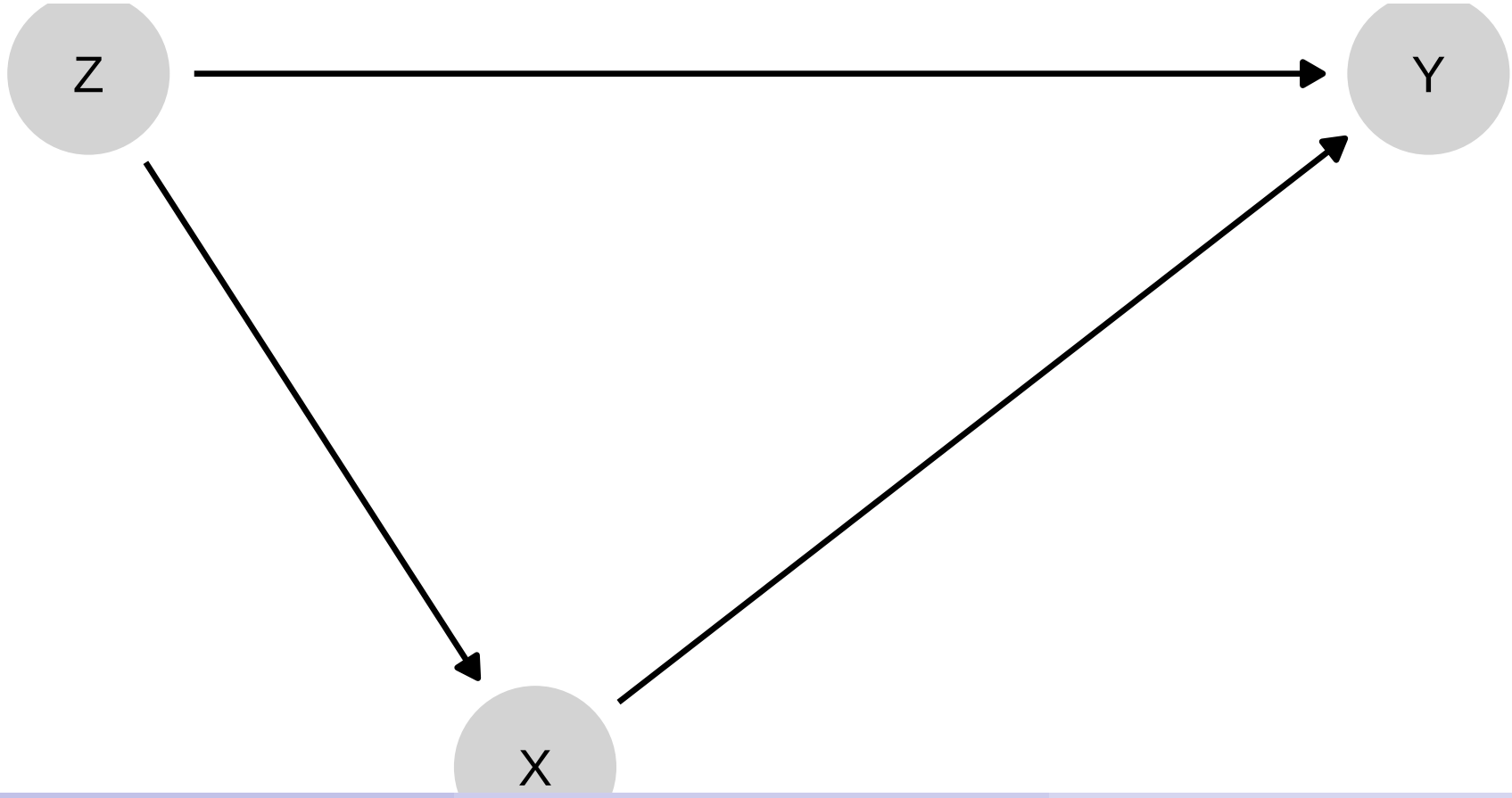
```
adjustmentSets(g, exposure = "X", outcome = "Y")
```

```
{ R }
{ Z }
```

Important Examples : Confounding

Example where Z is correlated with X and Y and is a confounder

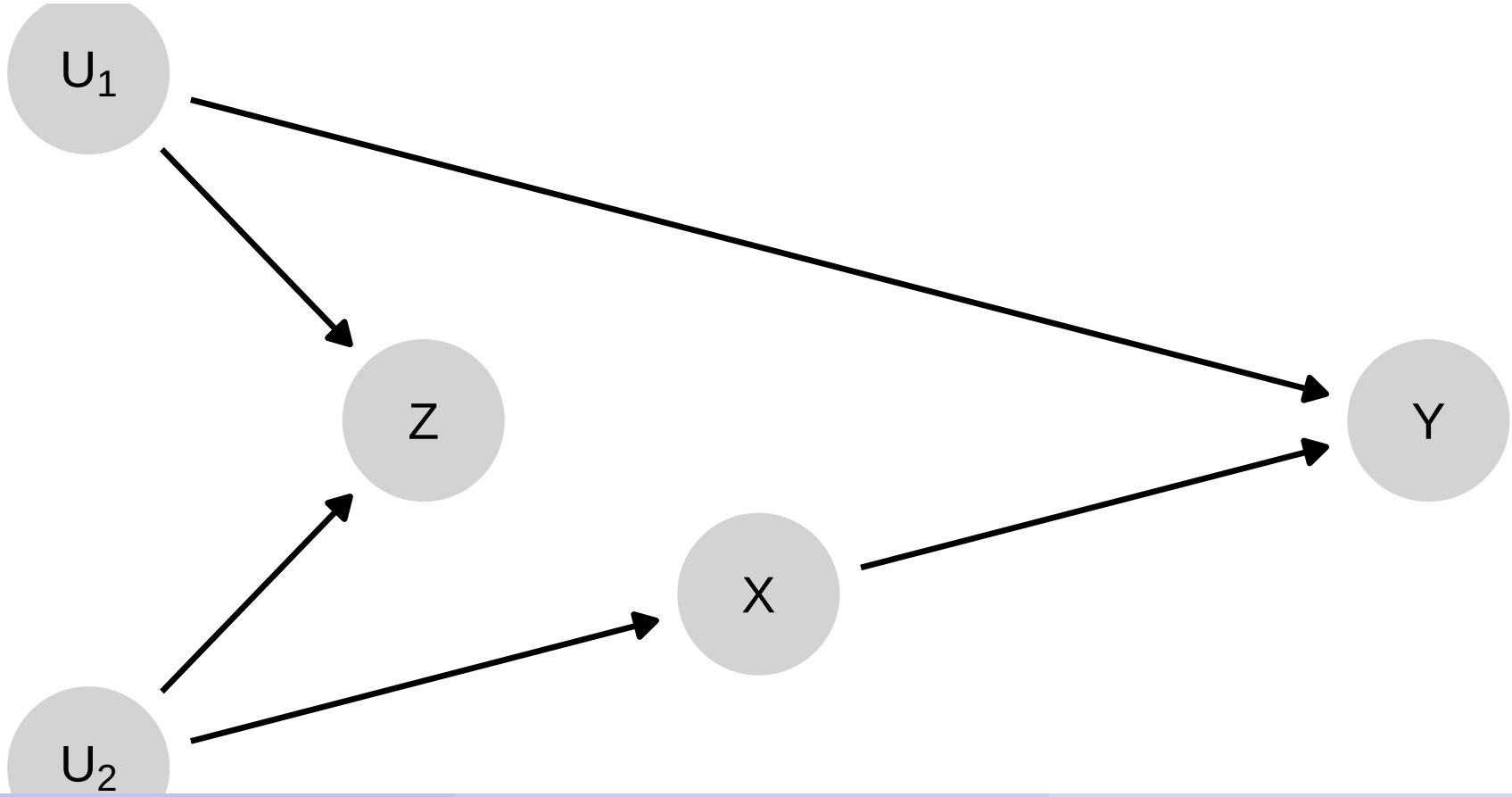
Controlling for Z can remove bias



Confounding

Example where Z is correlated with X and Y but it is *not* a confounder

Unbiased without controlling for Z



Important Examples : Collider

But controlling can also cause problems. In fact conditioning on a temporally pre-treatment variable could cause problems. Who'd have thunk? Here is an example from Pearl (2005):

Controlling for Z can induce bias

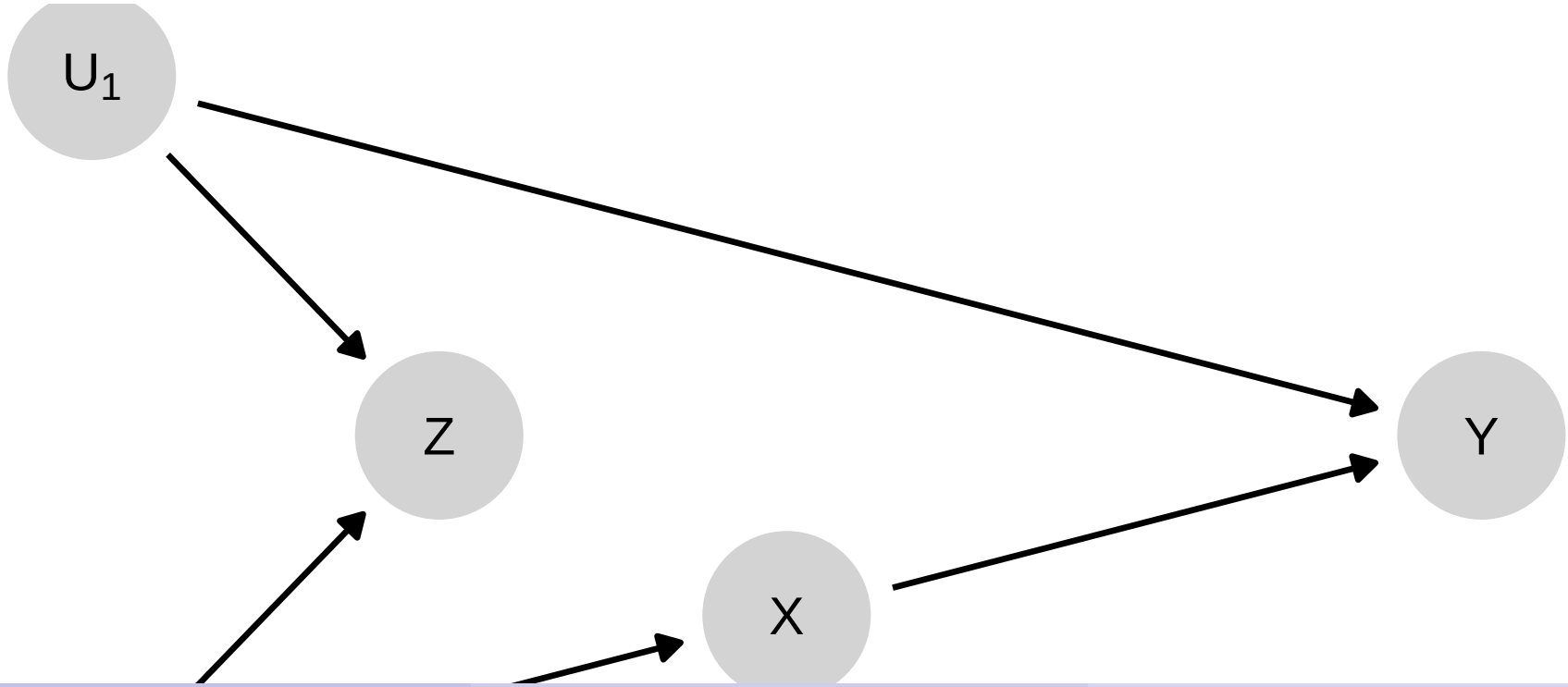


Illustration of identification failure from conditioning on a collider

```
U1 <- rnorm(10000);  U2 <- rnorm(10000)
Z  <- U1+U2
X  <- U2 + rnorm(10000)/2
Y  <- U1*2 + X

lm_robust(Y ~ X) |> tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.02	0.02	-0.85	0.39	-0.06	0.02
X	0.98	0.02	54.27	0.00	0.94	1.02

```
lm_robust(Y ~ X + Z) |> tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.01	0.01	-0.65	0.51	-0.02	0.01
X	-0.33	0.01	-35.01	0.00	-0.35	-0.31

Let's look at that in dagitty

```
g <- dagitty("dag{U1 -> Z  ; U1 -> y ; U2 -> Z ; U2 -> x  -> y
adjustmentSets(g, exposure = "x", outcome = "y")
```

```
 {}
```

```
isAdjustmentSet(g, "Z", exposure = "x", outcome = "y")
```

```
[1] FALSE
```

```
isAdjustmentSet(g, NULL, exposure = "x", outcome = "y")
```

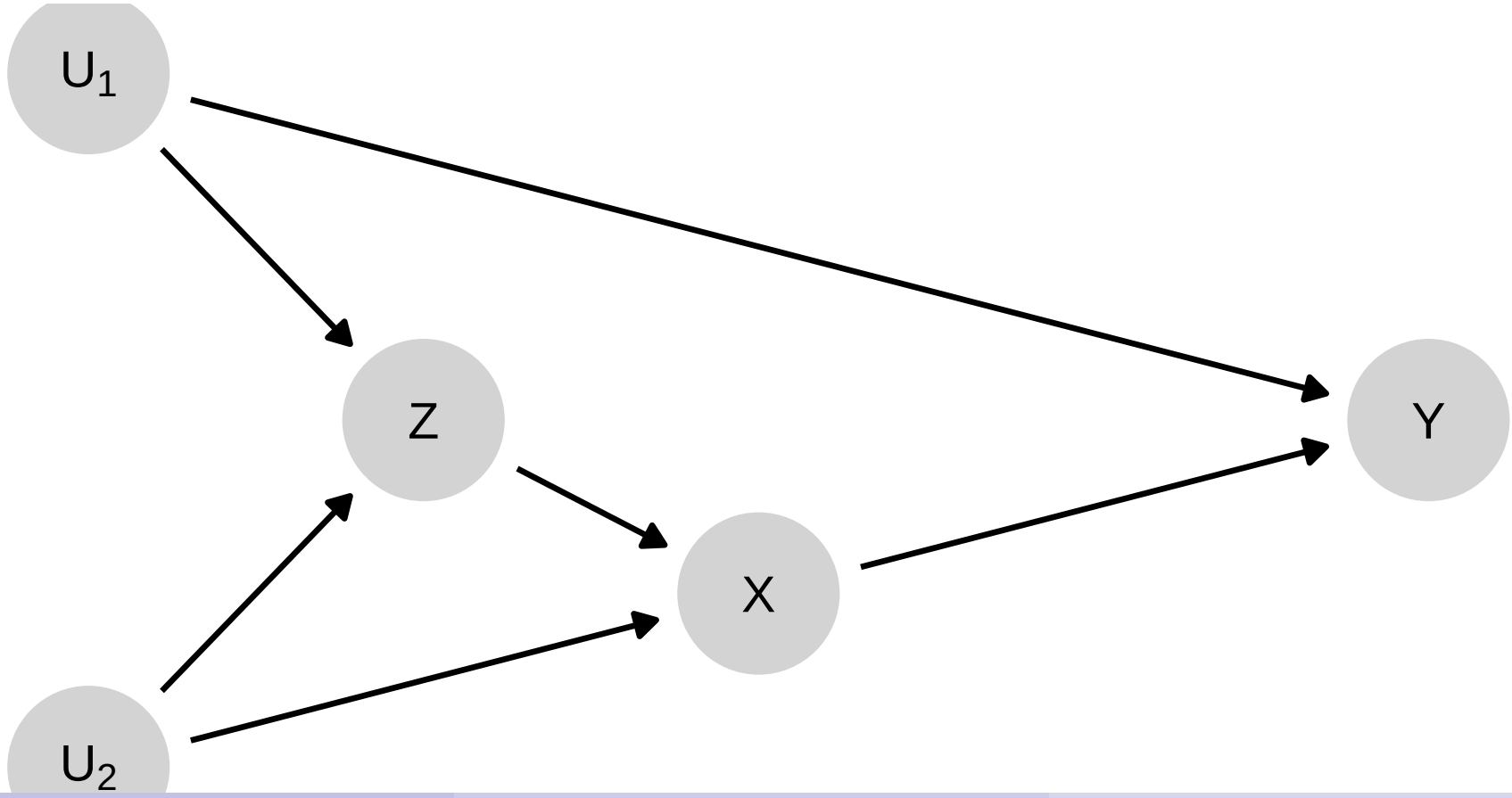
```
[1] TRUE
```

Which means, no need to condition on anything.

Collider & Confounder

A bind: from Pearl 1995.

Z is a confound but controlling for it can induce bias



Let's look at that in dagitty

```
g <- dagitty("dag{U1 -> Z  ; U1 -> y ;
              U2 -> Z ; U2 -> x  -> y;
              Z -> x}")
adjustmentSets(g, exposure = "x", outcome = "y")
```

```
{ U1 }
{ U2, Z }
```

which means you have to adjust on an unobservable. Here we double check that including or not including "Z" is enough:

```
isAdjustmentSet(g, "Z", exposure = "x", outcome = "y")
```

```
[1] FALSE
```

```
isAdjustmentSet(g, NULL, exposure = "x", outcome = "y")
```

```
[1] FALSE
```

Section 5

Frequentist Analysis

Frequentist Analysis

▸ Top

Subsection 1

Basic Analysis

Basic Analysis

- Simple estimates from experimental data
- Weighting, blocking
- Doubly robust estimation
- Design based variance estimates
- Design based p values
- Reporting

ATE: DIM

Unbiased estimates of the (sample) average treatment effect can be estimated (**whether or not there imbalance on covariates**) using:

$$\widehat{ATE} = \frac{1}{n_T}\sum_{T} Y_i - \frac{1}{n_C}\sum_{C} Y_i,$$

ATE: DIM in practice

```
df <- fabricatr::fabricate(N = 100, Z = rep(0:1, N/2), Y = rnc

# by hand
df |>
  summarize(Y1 = mean(Y[Z==1]),
            Y0 = mean(Y[Z==0]),
            diff = Y1 - Y0) |> kable(digits = 2)
```

Y1	Y0	diff
1.07	-0.28	1.35

```
# with estimatr
estimatr::difference_in_means(Y ~ Z, data = df) |>
  tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	
Z	1.35	0.17	7.94	0	1.01	1.68	97.9

ATE: DIM in practice

We can also do this with regression:

```
estimatr::lm_robust(Y ~ Z, data = df) |>
  tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.28	0.12	-2.33	0.02	-0.51	-0.04
Z	1.35	0.17	7.94	0.00	1.01	1.68

See @freedman2008regression on why regression is fine here

ATE: Blocks

Say now different strata or blocks $\mathcal{S}$ had different *assignment probabilities*. Then you could estimate:

$$\widehat{ATE} = \sum_{S \in \mathcal{S}} \frac{n_S}{n} \left(\frac{1}{n_{S1}} \sum_{S \cap T} y_i - \frac{1}{n_{S0}} \sum_{S \cap C} y_i \right) \tag{1}$$

Note: you cannot just ignore the blocks because assignment is no longer independent of potential outcomes: you might be sampling units with different potential outcomes with different probabilities.

However, the formula above works fine because selecting is random *conditional* on blocks.

Subsection 2

ATE: Blocks in practice

ATE: Blocks in practice

Data with heterogeneous assignments:

```
df <- fabricatr::fabricate(
  N = 500, X = rep(0:1, N/2),
  prob = .2 + .3*X,
  Z = rbinom(N, 1, prob),
  ip = 1/(Z*prob + (1-Z)*(1-prob)), # discuss
  Y = rnorm(N) + Z*X)
```

True effect is 0.5, but:

```
estimatr::difference_in_means(Y ~ Z, data = df) |>
  tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	
Z	0.9	0.1	9.32	0	0.71	1.09	377

Subsection 3

ATE: Blocks in practice

ATE: Blocks in practice

Averaging over effects in blocks ::: {.cell}

```
# by hand
estimates <-
  df |>
  group_by(X) |>
  summarize(Y1 = mean(Y[Z==1]),
            Y0 = mean(Y[Z==0]),
            diff = Y1 - Y0,
            W = n())

estimates$diff |> weighted.mean(estimates$W)
```

```
[1] 0.7236939
```

```
# with estimatr
estimatr::difference_in_means(Y ~ Z, blocks = X, data = df) |>
  tidy() |> kable(digits = 2)
```

ATE with IPW

This also corresponds to the difference in the weighted average of treatment outcomes (with weights given by the inverse of the probability that each unit is assigned to treatment) and control outcomes (with weights given by the inverse of the probability that each unit is assigned to control).

- The average difference in means estimator is the same as what you would get if you weighted inversely by shares of units in different conditions inside blocks.

ATE with IPW in practice

```
# by hand
df |>
  summarize(Y1 = weighted.mean(Y[Z==1], ip[Z==1]),
            Y0 = weighted.mean(Y[Z==0],  ip[Z==0]), # note !
            diff = Y1 - Y0)|>
  kable(digits = 2)
```

Y1	Y0	diff
0.59	-0.15	0.74

```
# with estimatr
estimatr::difference_in_means(Y ~ Z, weights = ip, data = df)
  tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df
Z	0.74	0.11	6.65	0	0.52	0.96	498

ATE with IPW

- But **inverse propensity weighting** is a more general principle, which can be used even if you do not have blocks.
- The intuition for it comes straight from **sampling weights** — you weight up in order to recover an unbiased estiamte of the potential outcomes for all units, whether or not they are assigned to treatment.
- With sampling weights however you can include units even if their weight was 1. *Why can you not include these units when doing inverse propensity weighting?*

Illustration: Estimating treatment effects with terrible treatment assignments: Fixer

Say you made a mess and used a randomization that was correlated with some variable, X. For example:

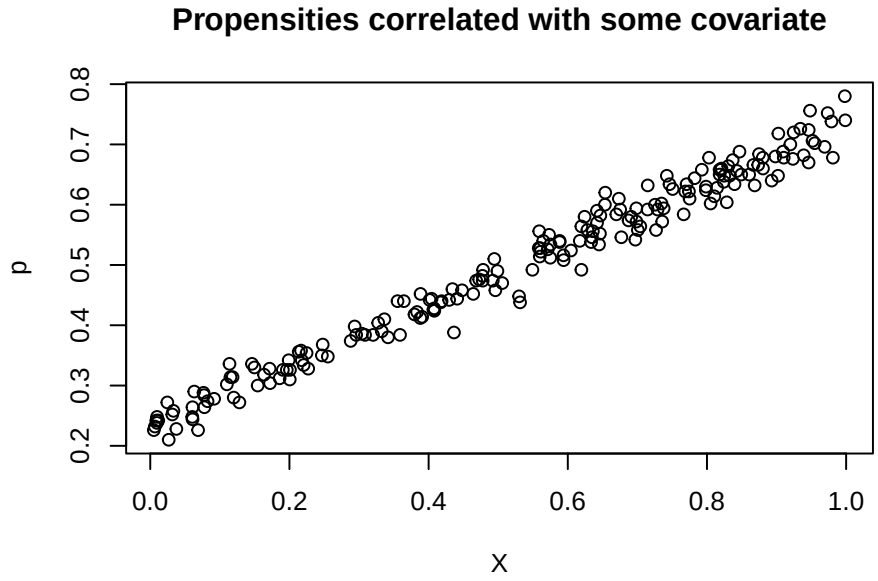
- The randomization is done in a way that introduces a correlation between Treatment Assignment and Potential Outcomes
- Then possibly, even though there is no true causal effect, we naively estimate a large one — enormous bias
- However since we know the assignment procedure we can **fully** correct for the bias
- In the next example, we do this using "**inverse propensity score weighting**." This is exactly analogous to standard survey weighting — since we selected different units for treatment with different probabilities, we weight them differently to recover the average outcome among treated units (same for control).

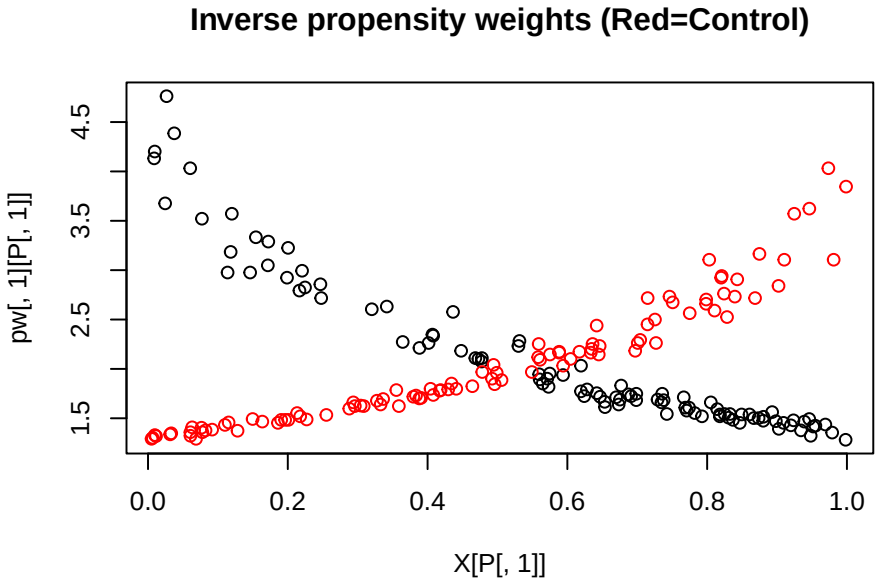
Basic randomization: Fixer

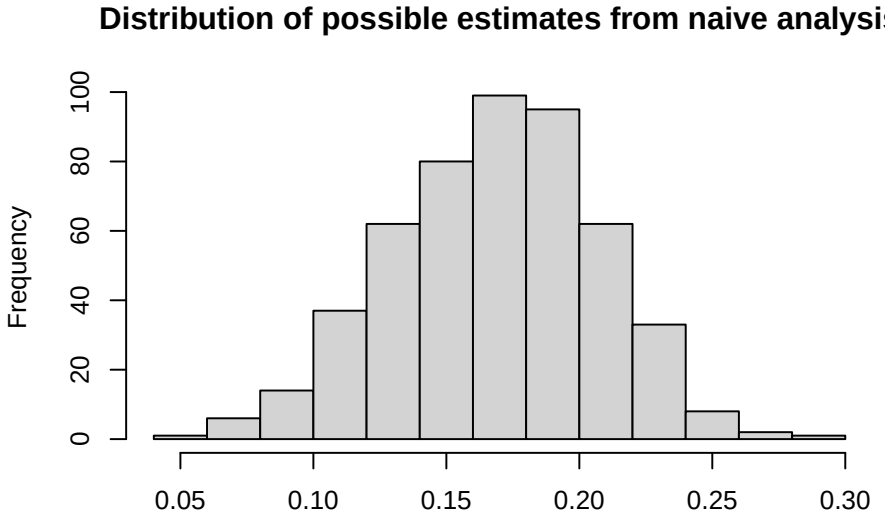
Code to generate bad assignment but proper propensity weights:

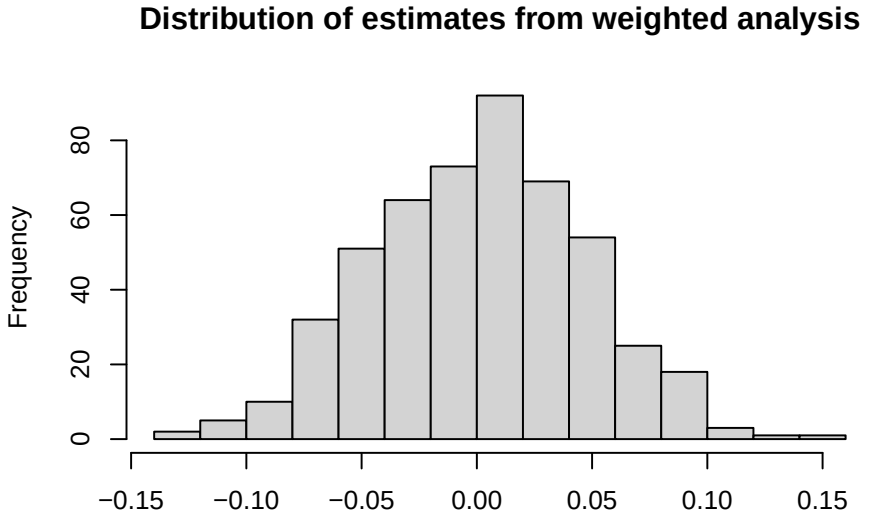
```
# design <-
#   declare_model(N = 200,
#                 X = runif(N),
#                 Y0 = X,
#                 Y1 = X,
#                 Y = X)

n  <- 200; reps <- 500; X  <- runif(n)      # Create a covariat
Y  <- Y1 <- Y0 <- X                         # Say X completely
Z  <- function(i) rank(X+2*runif(n))>(n/2) # Bad randomizatior
P  <- sapply(1:reps, Z)                     # Lots of possible
p  <- apply(P, 1, mean)                     # Recreate propensi
pw <- (!P)*(1/(1-p)); pw[P]=(P*(1/p))[P]    # Create inv prop w

naive   <- sapply(1:ncol(P),function(i) {
```

Basic randomization: Fixer









IPW with one unit!

This example is surprising but it helps you see the logic of why inverse weighting gets unbiased estimates (and why that might not guarantee a reasonable answer)

Imagine there is one unit with potential outcomes $Y(1) = 2, Y(0) = 1$. So the unit level treatment effect is 1.

You toss a coin.

- If you assign to treatment you estimate: $\hat{\tau} = \frac{2}{0.5} = 4$
- If you assign to control you estimate: $\hat{\tau} = -\frac{1}{0.5} = -2$

SO your expected estimate is:

$$0.5 \times 4 - 0.5 \times (-2) = 1$$

Great on average but always lousy

Subsection 4

Example

Covariate Adjustment

Consider for example this data.

- You randomly pair offerers and receivers in a dictator game (in which offerers decide how much of $1 to give to receivers).
- Your population comes from two groups (80% Baganda and 20% Banyankole) *so in randomly assigning partners you are randomly determining whether a partner is a coethnic or not*.
- **You find that in non-coethnic pairings 35% is offered, in coethnic pairings 48% is offered.**

Should you believe it?

Covariate Adjustment

- Population: randomly matched Baganda (80% of pop) and Banyankole (20% of pop)
- You find: in non-coethnic pairings 35% is offered, in coethnic pairings 48% is offered.
- But a closer look at the data reveals ...

		To: Baganda	To: Banyankole
Offers by	Baganda	64%	16%
	Banyankole	16%	4%

Table 15: Number of Games

		To: Baganda	To: Banyankole
Offers by	Baganda	50	50
	Banyankole	20	20

Table 16: Average Offers

So that's a problem

Covariate Adjustment

Control?

- With such data you might be tempted to 'control' for the covariate (here: ethnic group), using regression.
- But, perhaps surprisingly, it turns out that regression with covariates does not estimate average treatment effects.
- It does estimate an average of treatment effects, but specifically a minimum variance estimator, not necessarily an estimator of your estimand.

Covariate Adjustment

Compare:

- $\hat{\tau}_{ATE} = \sum_x \frac{w_x}{\sum_j w_j} \hat{\tau}_x$
- $\hat{\tau}_{OLS} = \sum_x \frac{w_x p_x (1-p_x)}{\sum_j w_j p_j (1-p_j)} \hat{\tau}_x$

Instead you can use formula above for $\hat{\tau}_{ATE}$ to estimate ATE

alternatively...

Covariate adjustment via saturated regression

Alternatively you can use a regression that includes both the treatment and the treatment *interacted* with the covariates.

In practice this is best done by *demeaning* the covariates; doing this lets you read off the average effect from the main term. Key resource: @lin2012agnostic

Covariate adjustment via saturated regression

Returning to prior example:

```
df <- fabricatr::fabricate(
  N = 500,
  X = rep(0:1, N/2),
  Z = rbinom(N, 1, .2 + .3*X),
  Y = rnorm(N) + Z*X)

lm_robust(Y ~ Z*X_c, data = df |> mutate(X_c = X - mean(X))) |>
  tidy() |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.10	0.06	-1.70	0.09	-0.22	0.02
Z	0.59	0.10	5.83	0.00	0.39	0.78
X_c	-0.18	0.12	-1.48	0.14	-0.41	0.06
Z:X_c	0.86	0.20	4.27	0.00	0.46	1.26

```
lm_lin(Y ~ Z, ~ X, data = df) |>
```

Demeaning and saturating

Demeaning interactions

- Say you have a factorial design with treatments X1 and X2 (or obervational data with two covariates)
- You analyse with a model that has main terms and interaction terms
- Interpreting coefficients can be confusing, but sometimes demeaning can help. What does demeaning do?

Let's:

- Declare a factorial design in which Y is generated according to

```
f_Y <- function(X1, X2, u) .1 + .2*X1 + .3*X2 + u*X1*X2
```

where u is distributed $U[0, 1]$.

- Specify estimands carefully
- Run analyses in which we do and do not demean the treatments; compare and explain results

Demeaning interactions

```
f_Y <- function(X1, X2, u) .1 + .2*X1 + .3*X2 + u*X1*X2

design <-
  declare_model(N = 1000,
                u = runif(N),
                X1 = complete_ra(N),
                X2 = block_ra(blocks = X1),
                X1_demeaned = X1 - mean(X1),
                X2_demeaned = X2 - mean(X2),
                Y = f_Y(X1, X2, u)) +
  declare_inquiry(
    base = mean(f_Y(0, 0, u)),
    average = mean(f_Y(0, 0, u) + f_Y(0, 1, u)  + f_Y(1, 0, u)
    CATE_X1_given_0 = mean(f_Y(1, 0, u) - f_Y(0, 0, u)),
    CATE_X2_given_0 = mean(f_Y(0, 1, u) - f_Y(0, 0, u)),
    ATE_X1 = mean(f_Y(1, X2, u) - f_Y(0, X2, u)),
```

Demeaning interactions: Solution

```
f_Y <- function(X1, X2, u) .1 + .2*X1 + .3*X2 + u*X1*X2
```

Inquiry	Estimator	Term	Mean Es
ATE_X1	demeaned	X1_demeaned	0.45
			(0.00)
ATE_X2	demeaned	X2_demeaned	0.55
			(0.00)
average	demeaned	(Intercept)	0.48
			(0.00)
base	natural	(Intercept)	0.10
			(0.00)
CATE_X1_given_0	natural	X1	0.20
			(0.00)
CATE_X2_given_0	natural	X2	0.30
			(0.00)
I_X1_X2	demeaned	X1_demeaned:X2_demeaned	0.50

Summary

If you have different groups with different assignment propensities you can do any or all of these:

1. Blocked differences in means
2. Inverse propensity weighting
3. Saturated regression (Lin)

We will compare the performances of these different approaches later.

You cannot (reliably):

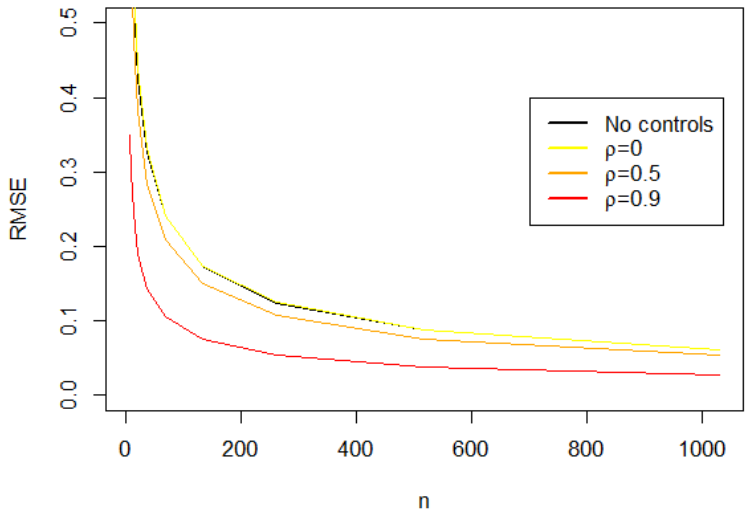
1. Ignore the groups
2. Include them in a regression (without interactions)

Covariate Adjustment

- Even though randomization ensures no bias, you may sometimes **want** to "**control**" for covariates in order to improve efficiency (see the discussion of blocking above).
- Or you may **have** to take account of the fact that the assignment to treatment is correlated with a covariate.

Conditional Bias and Precision Gains from Controls

Controls can do reduce noise and improve precision. This is an argument for using variables that are correlated with the output (not with the treatment).



Conditional Bias and Precision Gains from Controls

Introducing controls can create complications

As argued by Freedman (summary from @lin2012agnostic), we can get: "worsened asymptotic precision, invalid measures of precision, and small-sample bias"[3]

These adverse effects are essentially removed with an interacted model

See discussions in @imbens2015causal (7.6, 7.7) and especialy Theorem 7.2 for the aymptotic variance of the estimator

[3] though note that the precision concern does not hold when treatment and control groups are equally sized

Conditional Bias and Precision Gains from Controls

Note also including controls when treatment is correlated with covariates can induce "conditional bias." Doing this will change your estimates so be sure not to fish!

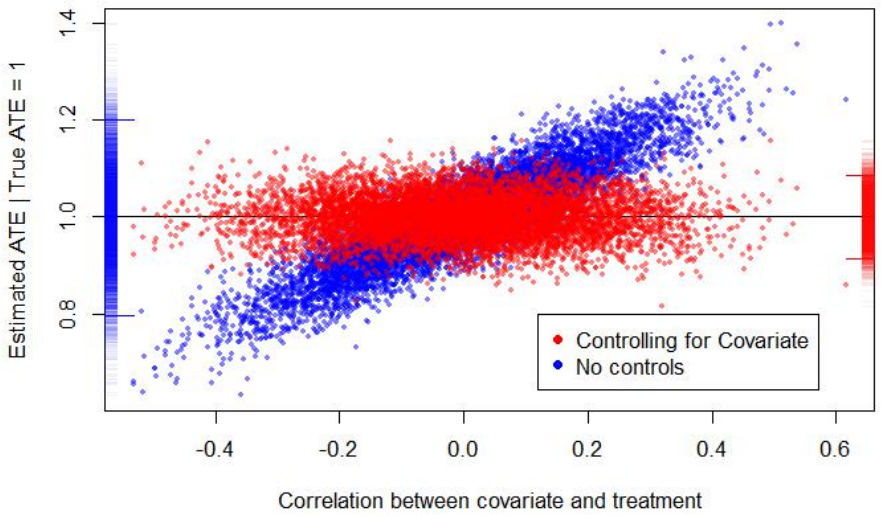


Figure 4: Advantages of controlling for vars that are correlated with outcomes

There are more or less sophisticated ways of doing this....

Subsection 5

Doubly robust estimation

Doubly robust estimation

Doubly robust estimation combines:

1. A model for how the covariates predict the potential outcomes
2. A model for how the covariates predict assignment propensities

Using both together to estimate potential outcomes using propensity weighting lets you do well even if either model is wrong.

Each part can be done using nonparameteric methods resulting in an overall semi-parametric procedure.

- $\pi(Z) = \Pr(Z = 1|X)$: Estimate $\hat{\pi}$
- $Y_z = \mathbb{E}[Y|Z = z, X]$: Estimate $\hat{Y}_z$
- Estimate of causal effect:
  $\frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{Z_i}{\hat{\pi}_i}(Y_i - \hat{Y}_{i1}\right) - \left(\frac{1-Z_i}{1-\hat{\pi}_i}(Y_i - \hat{Y}_{i0}\right) + \left(\hat{Y}_{i1} - \hat{Y}_{i0}\right)\right)$

Doubly robust estimation

- Estimate of causal effect:
$\frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{Z_i}{\hat{\pi}_i}(Y_i-\hat{Y}_{i1}\right)-\left(\frac{1-Z_i}{1-\hat{\pi}_i}(Y_i-\hat{Y}_{i0}\right)+\left(\hat{Y}_{i1}-\hat{Y}_{i0}\right)\right)$
- Note that if $\hat{Y}_{iz}$ are correct then the first parts drop out and we we get the right answer.
- So if you can impute the potential outcomes, you are good (though hardly surprising)

Doubly robust estimation

- More subtly say the $\widehat{pi}$s are correct, but your imputations are wrong; then we again have an unbiased estimator.

To see this imagine with probability π we assign unit 1 to treatment and 2 to control (otherwise 1 to control and 2 to treatment).

Then our *expected* estimate is:

$$\frac{1}{2}\pi\left(\left(\frac{1}{\pi}(Y_{11}-\hat{Y}_{11}\right)-\left(\frac{1}{\pi}(Y_{20}-\hat{Y}_{20}\right)\right)+(1-$$
$$\pi)\left(\left(\frac{1}{1-\pi}(Y_{21}-\hat{Y}_{21}\right)-\left(\frac{1}{1-\pi}(Y_{10}-\hat{Y}_{10}\right)\right)+\left(\hat{Y}_{11}-\hat{Y}_{20}\right)+\left(\hat{Y}_{21}-\hat{Y}_{10}\right)$$

$$\frac{1}{2}\left(Y_{11}-Y_{10}+Y_{21}-Y_{20}+\pi\left(\left(\frac{1}{\pi}(-\hat{Y}_{11}\right)-\left(\frac{1}{\pi}(-\hat{Y}_{20}\right)\right)+(1-\pi)\left(\left(\frac{1}{1-\pi}(-\right.\right.\right.$$
$$\left.\left.\left.\hat{Y}_{11}-\hat{Y}_{20}\right)+\left(\hat{Y}_{21}-\hat{Y}_{10}\right)\right.\right.$$

$$\frac{1}{2}\left(Y_{11}-Y_{10}+Y_{21}-Y_{20}\right)$$

@robins1994estimation

Subsection 6

Doubly robust estimation illustration

Data with confounding

Consider this data:

```
# df with rue treatment effect of 1
# (0.5 if race = 0; 1.5 if race = 1)

df <- fabricatr::fabricate(
  N = 5000,
  class = sample(1:3, N, replace = TRUE),
  race = rbinom(N, 1, .5),
  Z = rbinom(N, 1, .2 + .3*race),
  Y = .5*Z + race*Z + class + rnorm(N),
  qsmk = factor(Z),
  class = factor(class),
  race = factor(race)
)
```

Simple approaches

Naive regression produces biased estimates, even with controls. Lin regression gets the right result however.

```
# Naive
lm_robust(Y ~ Z, data = df)$coefficients[["Z"]]
```

```
[1] 1.257443
```

```
# OLS with controls
lm_robust(Y ~ Z + class + race, data = df)$coefficients[["Z"]]
```

```
[1] 1.121328
```

```
# Lin
lm_lin(Y ~ Z,  ~ class + race, data = df)$coefficients[["Z"]]
```

```
[1] 1.002136
```

Doubly robust estimation

`drtmle` is an R package that uses doubly robust estimation to compute "marginal means of an outcome under fixed levels of a treatment."

```
library(SuperLearner)
library(drtmle)
drtmle_fit <- drtmle(
  W = df |> select(race, class),
  A = df$Z,
  Y = df$Y,
  SL_Q = c("SL.glm", "SL.mean", "SL.glm.interaction"),
  SL_g = c("SL.glm", "SL.mean", "SL.glm.interaction"),
  SL_Qr = "SL.glm",
  SL_gr = "SL.glm",
  maxIter = 1
)
```

Doubly robust estimation

```
# "Marginal means"
drtmle_fit$drtmle$est
```

```
[1] 1.983348 2.985222
```

```
# Effects
ci(drtmle_fit, contrast = c(-1,1))
```

```
$drtmle
                    est   cil   ciu
E[Y(1)]-E[Y(0)] 1.002 0.937 1.067
```

```
wald_test(drtmle_fit, contrast = c(-1,1))
```

```
$drtmle
                         zstat pval
H0:E[Y(1)]-E[Y(0)]=0 30.301    0
```

Resource: https://muse.jhu.edu/article/883477

Subsection 7

Assessing performance

Assessing performance

Challenge: Use `DeclareDesign` to compare performance of `drtmle` and `lm_lin`

Subsection 8

Randomization Inference

Calculate a p value in your head

- Illustrating p values via "randomization inference"
- Say you randomized assignment to treatment and your data looked like this.

Unit	1	2	3	4	5	6	7	8	9	10
Treatment	0	0	0	0	0	0	0	1	0	0
Health score	4	2	3	1	2	3	4	8	7	6

Then:

- Does the treatment improve your health?
- What's the p value for the null that treatment had no effect on anybody?

Calculate a p value in your head

- Illustrating p values via "randomization inference"
- Say you randomized assignment to treatment and your data looked like this.

Unit	1	2	3	4	5	6	7	8	9	10
Treatment	0	0	0	0	0	0	0	0	1	0
Health score	4	2	3	1	2	3	4	8	7	6

Then:

- Does the treatment improve your health?
- What's the p value for the null that treatment had no effect on anybody?

Randomization Inference: Some code

- In principle it is very easy.
- These few lines generate data, produce the regression estimate and then an RI estimate of p:

```
# data
df <- fabricate(N = 200, X = rep(c(FALSE,TRUE), N/2), Y=   .1*X

# test stat
t <- function(df) with(df, mean(Y[X])- mean(Y[!X]))

# test stat distribution
ts <- replicate(1000, df |> mutate(X = sample(X)) |> t())

# test
mean(ts >= t(df))    # One sided p value
```

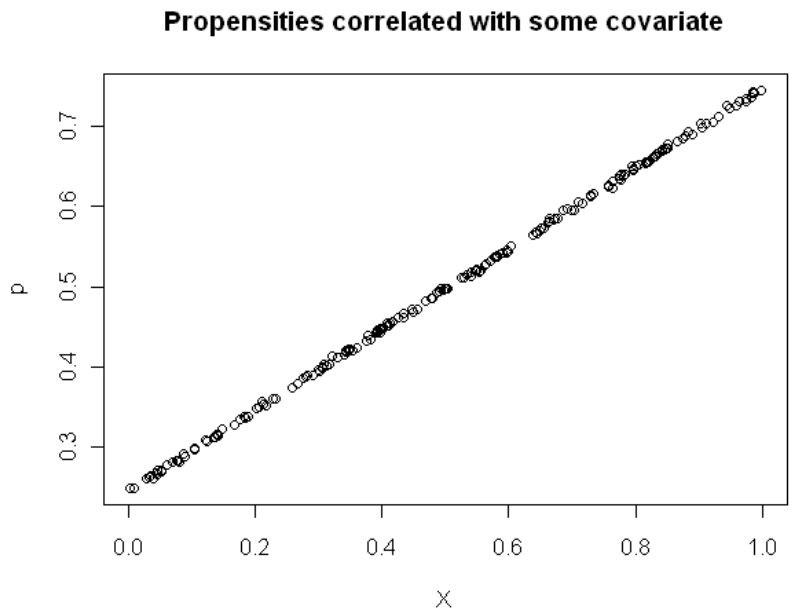
```
[1] 0.349
```

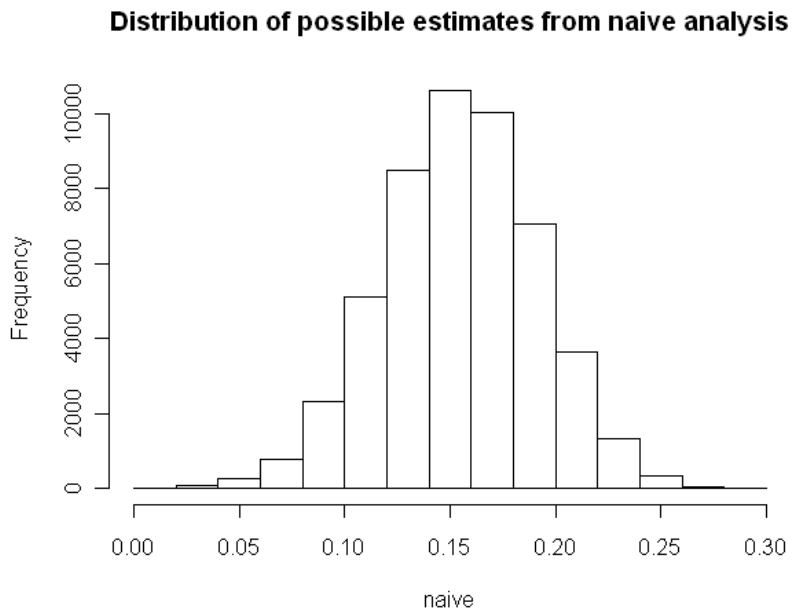
Randomization Inference

In practice it is a good idea to create a P matrix when you do your randomization (although note: if the null is about one treatment, then you are interested only in the randomization of that treatment, not the joint randomization of all)

Randomization Inference

- Say you had a silly randomization procedure and forgot to take account of it in your estimates.





- You estimate .15. *Does the treatment improve your health?*
- $p =$?

Randomization Inference

- Randomization procedures are sometimes funky in lab experiments
- Using randomization inference would force a focus on the true assignment of individuals to treatments
- Fake (but believable) example follows

Randomization Inference

Table 19: Optimal assignment to treatment given constraints due to facilities

		Capacity	T1	T2	T3
Session	Thursday	40	10	30	0
	Friday	40	10	0	30
	Saturday	10	10	0	0
		90	30	30	30

Table 20: Constraints due to subjects

Subject Type	N	Available
A	30	Thurs, Fri
B	30	Thurs, Sat
C	30	Fri, Sat

Randomization Inference

If you think hard about assignment you might come up with an allocation like this.

Table 21: Assignment of people to days

			Allocation		
Subject Type	N	Available	Thurs	Fri	Sat
A	30	Thurs, Fri	15	15	
B	30	Thurs, Sat	25		5
C	30	Fri, Sat		25	5

That allocation balances as much as possible. Given the allocation you might randomly assign individuals to different days as well as randomly assigning them to treatments within days. If you then figure out assignment propensities, this is what you would get:

			Assignment Probabilities		
Subject Type	N	Available	T1	T2	T3
A	30	Thurs, Fri	0.25	0.375	0.375
B	30	Thurs, Sat	0.375	0.625	0
C	30	Fri, Sat	0.375		0.625

Randomization Inference

Even under the assumption that the day of measurement does not matter, these assignment probabilities have big implications for analysis.

			Assignment Probabilities		
Subject Type	N	Available	T1	T2	T3
A	30	Thurs, Fri	0.25	0.375	0.375
B	30	Thurs, Sat	0.375	0.625	0
C	30	Fri, Sat	0.375		0.625

- Only the type A subjects could have received any of the three treatments.
- There are no two treatments for which it is possible to compare outcomes for subpopulations B and C
- A comparison of $T1$ versus $T2$ can only be made for population $A \cup B$
- However subpopulation A is assigned to A (versus B) with probability 4/5; while population B is assigned with probability 3/8
- **Implications for design**: need to uncluster treatment delivery
- **Implications for analysis**: need to take account of propensities

Idea: Wacky assignments happen but if you know the propensities you can do the analysis.

Randomization Inference

- Randomization inference can get quite a bit more complicated when you want to test a null other than the sharp null of no effect.
- Say you wanted to test the null that the effect is 2 for all units. How do you do it?
- Say you wanted to test the null that an *interaction effect* is zero. How do you do it?
- In both cases by filling in a potential outcomes schedule given the hypothesis in question and then generating a test statistic

Observed		Under null that effect is 0		Under null that effect is 2	
Y(0)	Y(1)	Y(0)	Y(1)	Y(0)	Y(1)
1	?	1	1	1	3
2	?	2	2	2	4
?	4	4	4	2	4
?	3	3	3	1	3

Subsection 9

Design Based Estimation of Variance

Var(ATE)

- Recall that the treatment effect is gotten by taking a sample of outcomes under treatment and comparing them to a sample of outcomes under control
- Say that there is no "error"
- Why would this procedure produce uncertainty?

Var(ATE)

- Why would this procedure produce uncertainty?
- The uncertainty comes from being uncertain about the average outcome under control from observations of the control units, and from being uncertain about the average outcome under treatment from obervation of the treated units
- In other words, it comes from the variance in the treatment outcomes and variance in the control outcomes (and not, for example, from variance in the treatment effect)

Var(ATE)

You can also estimate variance straight from the data. From Freedman Prop 1 (using combinatorics!) we have:

$$V(\widehat{ATE}) = \frac{1}{n-1}\left[\frac{n_C}{n_T}V(Y(1)) + \frac{n_T}{n_C}V(Y(0))\right] + 2C\left(Y(1), Y(0)\right)$$

Usefully rewritten as:

$$V(\widehat{ATE}) = \frac{n}{n-1}\left[\frac{V(Y(1))}{n_T} + \frac{V(Y(0))}{n_C}\right] - \frac{1}{n-1}\left[V(Y(1)) + V(Y(0)) - 2C\left(Y(1), Y(0)\right)\right.$$

...where V denotes variance and C covariance

Var(ATE)

Note:

- We can use the sample estimates $s^2(\{Y_i\}_{i \in C})$ and $s^2(\{Y_i\}_{i \in T})$ for the first part.
- But $C(Y(1), Y(0))$ cannot be estimated from data.
- The **Neyman estimator** ignores the second part (and so is conservative).
- Tip: for STATA users, use `, robust` (see @samii2012equivalencies)

ATE and Var(ATE)}

For the case with blocking, the conservative estimator is:

$$V(\widehat{ATE}) = \sum_{S \in \mathcal{S}} \left(\frac{n_S}{n} \right)^2 \left(\frac{s^2_{S1}}{n_{S1}} + \frac{s^2_{S0}}{n_{S0}} \right)$$

Illustration of Neyman Conservative Estimator

An illustration of *how* conservative the conservative estimator of variance really is (numbers in plot are correlations between $Y(1)$ and $Y(0)$.

We confirm that:

1. the estimator is conservative
2. the estimator is more conservative for negative correlations between $Y(0)$ and $Y(1)$ — eg if those cases that do particularly badly in control are the ones that do particularly well in treatment %, and
3. with τ and $V(Y(0))$ fixed. high positive correlations are associated with highest variance.

Illustration of Neyman Conservative Estimator

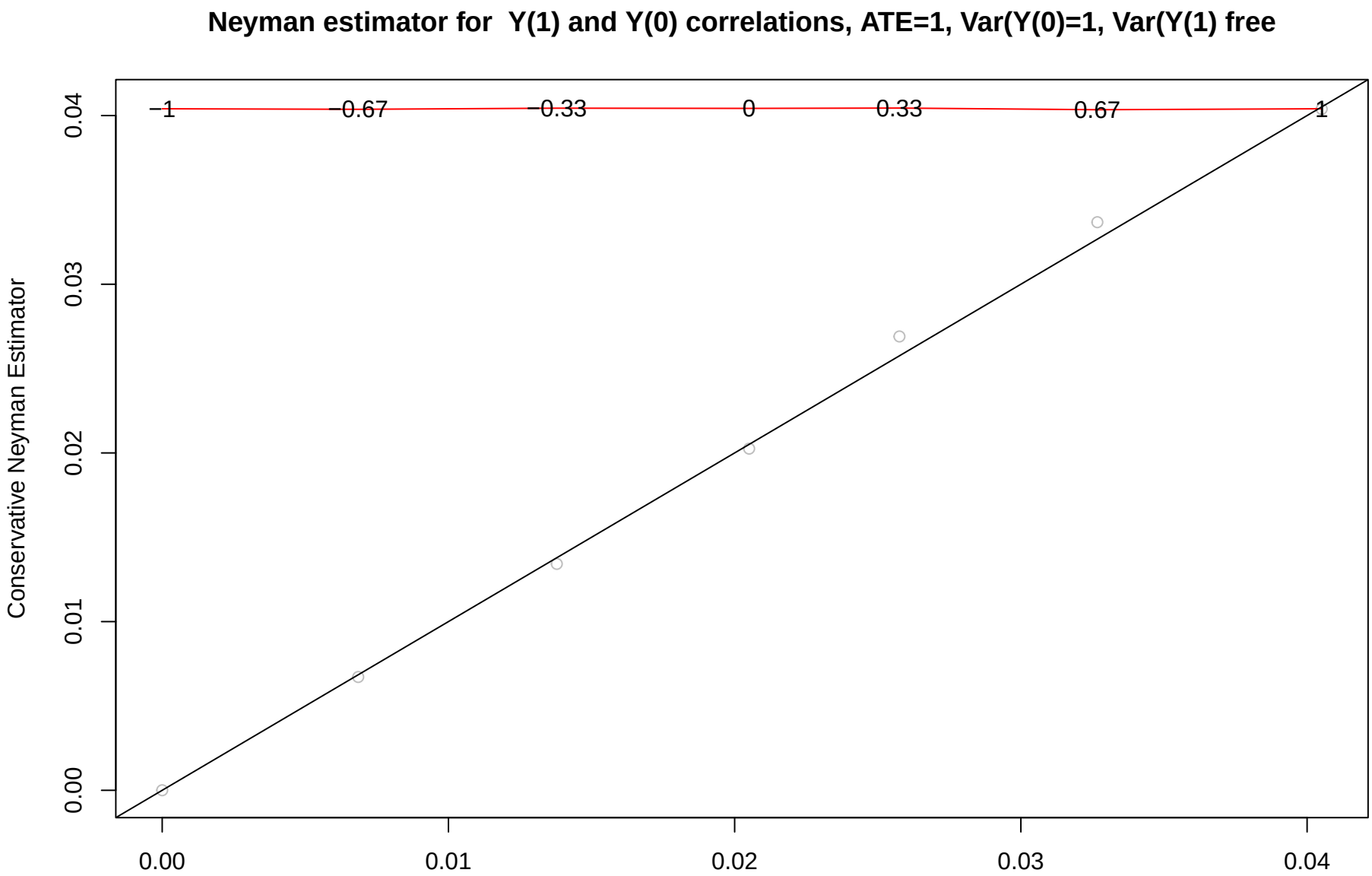


Illustration of Neyman Conservative Estimator

τ	ρ	$\sigma^2_{Y(1)}$	Δ	σ^2_τ	$\hat{\sigma}^2_\tau$	$\hat{\sigma}^2_{\tau(\text{Neyman})}$
1.00	-1.00	1.00	-0.04	0.00	-0.00	0.04
1.00	-0.67	1.00	-0.03	0.01	0.01	0.04
1.00	-0.33	1.00	-0.03	0.01	0.01	0.04
1.00	0.00	1.00	-0.02	0.02	0.02	0.04
1.00	0.33	1.00	-0.01	0.03	0.03	0.04
1.00	0.67	1.00	-0.01	0.03	0.03	0.04
1.00	1.00	1.00	0.00	0.04	0.04	0.04

Tighter Bounds On Variance Estimate

The conservative variance comes from the fact that you do not know the covariance between $Y(1)$ and $Y(0)$.

- But as Aronow, Green, and Lee (2014) point out, you *do* know something.
- Intuitively, if you know that the variance of $Y(1)$ is 0, then the covariance also has to be zero.
- This basic insight opens a way of calculating bounds on the variance of the sample average treatment effect.

Tighter Bounds On Variance Estimate

Example:

- Take a million-observation dataset, with treatment randomly assigned
- Assume $Y(0) = 0$ for everyone and $Y(1)$ distributed normally with mean 0 and standard deviation of 1000.
- Note here the covariance of $Y(1)$ and $Y(0)$ is 0.
- Note the true variance of the estimated sample average treatment effect should be (approx) $\frac{Var(Y(1))}{\sqrt{1000000}} + \frac{Var(Y(0))}{\sqrt{1000000}} = 1$.
- But using the Neyman estimator (or OLS!) we estimate (approx) $\frac{Var(Y(1))}{\sqrt{1000000/2}} + \frac{Var(Y(0))}{\sqrt{1000000/2}} = \sqrt{2}$.
- But we can recover the truth knowing the covariance between $Y(1)$ and $Y(0)$ is 0.

Tighter Bounds On Variance Estimate: Code

```
sharp_var <- function(yt,yc,N=length(c(yt,yc)),upper=TRUE){
 m <- length(yt);  n <- m + length(yc)
 V <- function(x,N) {
       (N-1)/(N*(length(x)-1)) * sum((x - mean(x))^2)}
 yt <- sort(yt)
 if(upper) {yc <- sort(yc)
       } else {yc <- sort(yc,decreasing=TRUE)}
 p_i <- unique(sort(c(seq(0,n-m,1)/(n-m),seq(0,m,1)/m)))-
        .Machine$double.eps^.5
 p_i[1] <- .Machine$double.eps^.5
 yti <- yt[ceiling(p_i*m)]; yci <- yc[ceiling(p_i*(n-m))]
 p_i_minus <- c(NA,p_i[1: (length(p_i)-1)])
 return(((N-m)/m * V(yt,N) + (N-(n-m))/(n-m)*V(yc,N) +
       2*sum(((p_i-p_i_minus)*yti*yci)[2:length(p_i)]) -
       2*mean(yt)*mean(yc))/(N-1))}
```

Illustration

```
n   <- 1000000
Y   <- c(rep(0,n/2), 1000*rnorm(n/2))
X   <- c(rep(0,n/2), rep(1, n/2))
ols <- round(coef(summary(lm(Y~X)))[2,],3)
kable(t(as.matrix(ols)))
```

```
Error in eval(substitute(expr), data, enclos = parent.frame()
```

```
sharp <- round(c(sharp_var(Y[X==1], Y[X==0], upper = FALSE),
                 sharp_var(Y[X==1], Y[X==0], upper = TRUE)),3)
sharp
```

```
[1] 1 1
```

Subsection 10

Principle: Keep the reporting close to the design

Design based analysis

- Report the analysis that is implied by the design.

		T2 N	T2 Y	T2 All	Diff
T1	N	$\overline{y}_{00}$ (sd)	$\overline{y}_{01}$ (sd)	$\overline{y}_{0x}$ (sd)	$d_2\|T1=0$ (sd)
	Y	$\overline{y}_{10}$ (sd)	$\overline{y}_{10}$ (sd)	$\overline{y}_{1x}$ (sd)	$d_2\|T1=1$ (sd)
	All	$\overline{y}_{x0}$ (sd)	$\overline{y}_{x1}$ (sd)	y (sd)	d_2 (sd)
	Diff	$d_1\|T2=0$ (sd)	$d_1\|T2=1$ (sd)	d_1 (sd)	$d_1 d_2$ (sd)

This is instantly recognizable from the design and returns all the benefits of the factorial design including all main effects, conditional causal effects, interactions and summary outcomes. It is much clearer and more informative than a regression table.

Section 6

Bayesian approaches

Subsection 1

Bayes Basics

Bayes Rule

- Bayesian methods are just sets of procedures to figure out how to update beliefs in light of new information.
- We begin with a prior belief about the probability that a hypothesis is true.
- New data then allow us to form a posterior belief about the probability of the hypothesis.

Bayesian inference takes into account:

- the consistency of the evidence with a hypothesis
- the uniqueness of the evidence to that hypothesis
- background knowledge about the problem.

Illustration 1

I draw a card from a deck and ask *What are the chances it is a Jack of Spades?*

- Just 1 in 52.

Now I tell you that the card is indeed a spade. What would you guess?

- 1 in 13

What if told you it was a heart?

- No chance it is the Jack of Spades

What if I said it was a face card and a spade.

- 1 in 3.

Illustration 1

These answers are applications of Bayes' rule.

In each case the answer is derived by assessing what is possible, given the new information, and then assessing how likely the outcome of interest among the states that are possible. In all the cases you calculate:

$$\text{Prob Jack of Spades} \mid \text{Info} = \frac{\text{Is Jack of Spades Consistent w/ Info?}}{\text{How many cards are consistent w/ Info?}}$$

Illustration 2 **Interpreting Your Test Results**

You take a test to see whether you suffer from a disease that affects 1 in 100 people. The test is good in the following sense:

- if you have the disease, then with a 99% probability it will say you have the disease
- if you do not have it, then with a 99% probability, it will say that you do not have it

The test result says that you have the disease. What are the chances you have it?

Illustration 2 **Interpreting Your Test Results**

- It is *not* 99%. 99% is the probability of the result given the disease, but we want the probability of the disease given the result.
- The right answer is 50%, which you can think of as the share of people that have the disease among all those that test positive. For example
- e.g. if there were 10,000 people, then 100 would have the disease and 99 of these would test positive. But 9,900 would not have the disease and 99 of these would test positive. So the people with the disease that test positive are half of the total number testing positive.

Illustration 2. An illustration

```
p=.9; ## power of test and prior prob healthy
s=2000  ## population size
col5 = "red"
col0 = "black"

par(mfrow=c(1,2))
  plot(c(0,1), c(0,1), axes=F, type="n", ann=F)
  title(main ="Healthy Circles")
  points(.175*rnorm(p*p*s)+.5, .175*rnorm(p*p*s) + .5, col =
  points(.1*rnorm((1-p)*p*s)+.5, .1*rnorm((1-p)*p*s) + .5, c
  box()

  plot(c(0,1), c(0,1), axes=F, type="n", ann=F)
  title(main ="Sick squares")
  points(.1*rnorm(p*(1-p)*s)+ .5, .1*rnorm(p*(1-p)*s) + .5,
  points(.175*rnorm((1-p)*(1-p)*s)+.5, .175*rnorm((1-p)*(1-p
```

Illustration 2. More formally.

As an equation this might be written:

$$\text{Prob You have the Disease} \mid \text{Pos} = \frac{\text{How many have the disease and test pos}}{\text{How many people test pos?}}$$

Two Child Problem

Consider last an old puzzle found described @gardner1961second.

- Mr Smith has two children, A and B.
- At least one of them is a boy.
- What are the chances they are both boys?

To be explicit about the puzzle, we will assume that the information that one child is a boy is given as a truthful answer to the question "*is at least one of the children a boy?*"

Assuming also that there is a 50% probability that a given child is a boy.

Two Child Problem

As an equation:

$$\text{Prob both boys} \mid \text{Not both girls} = \frac{\text{Prob both boys}}{\text{Prob not both girls}} = \frac{\text{1 in 4}}{\text{3 in 4}}$$

Bayes Rule

Formally, all of these equations are applications of Bayes' rule which is a simple and powerful formula for deriving updated beliefs from new data.

The formula is given as:

$$\Pr(H|\mathcal{D}) = \frac{\Pr(\mathcal{D}|H)\Pr(H)}{\Pr(\mathcal{D})} \tag{2}$$

$$= \frac{\Pr(\mathcal{D}|H)\Pr(H)}{\sum_{H'}\Pr(\mathcal{D}|H')\Pr(H'))} \tag{3}$$

Bayes Rule

Formally, all of these equations are applications of Bayes' rule which is a simple and powerful formula for deriving updated beliefs from new data.

For continuous distributions and parameter vector θ:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_{\theta'} p(\mathcal{D}|\theta')p(\theta')d\theta}$$

Useful Distributions: Beta and Dirichlet Distributions

- Bayes rule requires the ability to express a prior distribution but it does not require that the prior have any particular properties other than being probability distributions.
- Sometimes however it can be useful to make use of "off the shelf" distributions.

Consider **the share of people in a population that voted**. This is a quantity between 0 and 1.

- Two people might may both believe that the turnout was around 50% but differ in how certain they are about this claim.
- One might claim to have no information and to believe any turnout rate between 0 and 100% is equally likely; another might be completely confident that the number if 50%.

Here the parameter of interest is a *share*. The **Beta** and **Dirichlet** distributions are particularly useful for representing beliefs on shares.

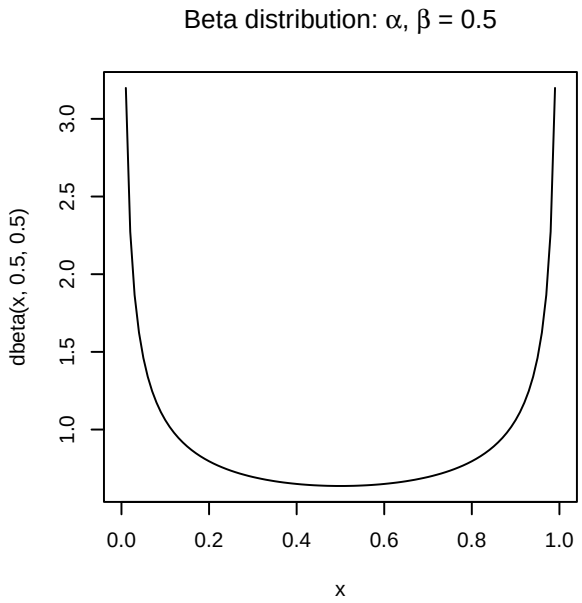
Beta

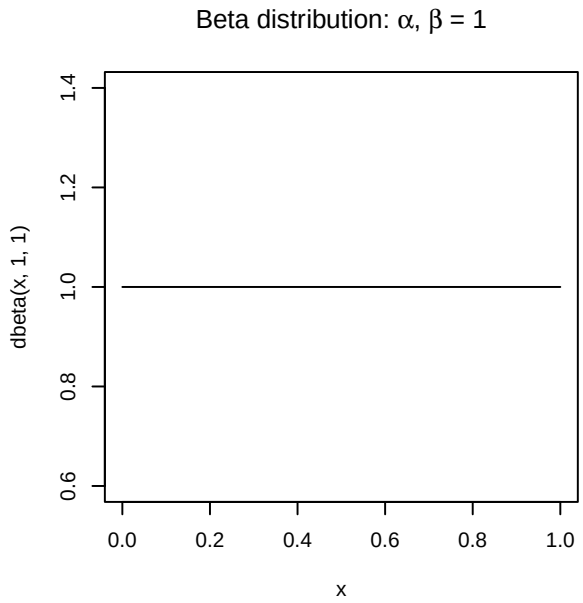
- The Beta distribution is a distribution over the $[0, 1]$ that is governed by two parameters, α and β.
- In the case in which both α and β are 1, the distribution is uniform – all values are seen as equally likely.
- As α rises large outcomes are seen as more likely
- As β rises, lower outcomes are seen as more likely.
- If both rise proportionately the expected outcome does not change but the distribution becomes tighter.

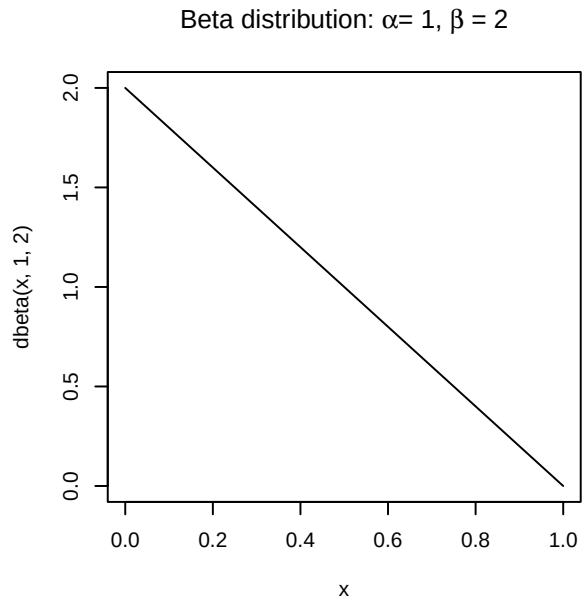
An attractive feature is that if one has a prior Beta(α, β) over the probability of some event, and then one observes a positive case, the Bayesian posterior distribution is also a Beta with with parameters $\alpha + 1, \beta$. Thus if people start with uniform priors and build up knowledge on seeing outcomes, their posterior beliefs should be Beta.

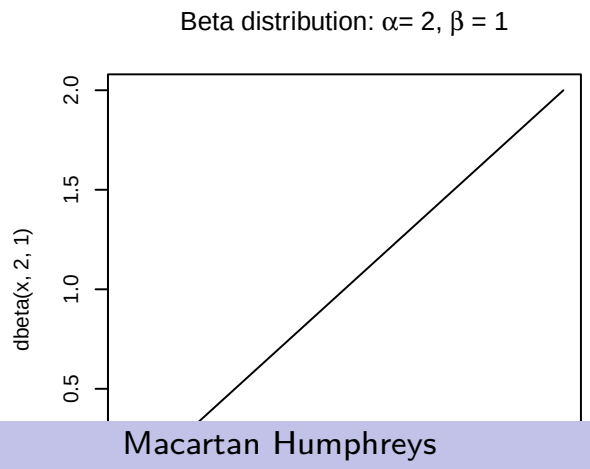
Beta

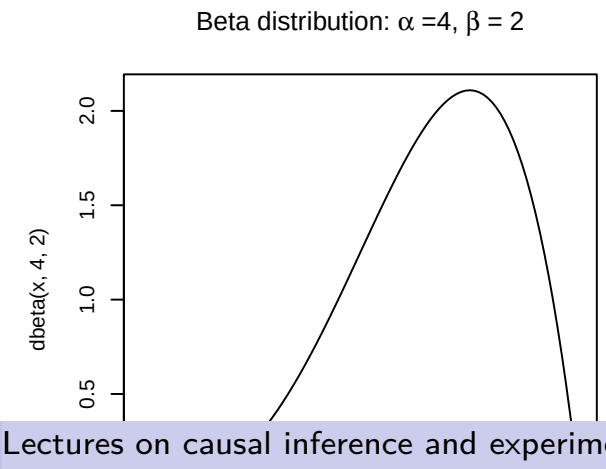
Here is a set of such distributions.

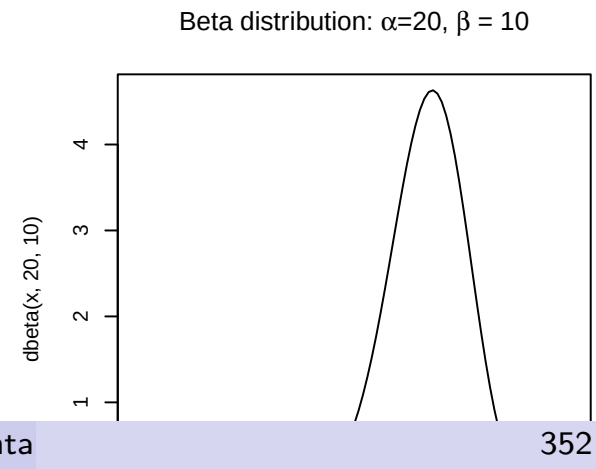












Dirichlet distributions.

The Dirichlet distributions are generalizations of the Beta to the situation in which there are beliefs not just over a proportion, or a probability, but over collections of probabilities.

- If four outcomes are possible and each is likely to occur with probability p_k, $k = 1, 2, 3, 4$ then beliefs are distributions over a three dimensional unit simplex.
- The distribution has as many parameters as there are outcomes and these are traditionally recorded in a vector, α.
- As with the Beta distribution, an uninformative prior (Jeffrey's prior) has α parameters of $(.5, .5, .5, \dots)$ and a uniform ("flat") distribution has $\alpha = (1, 1, 1, , \dots)$.
- The Dirichlet updates in a simple way. If you have a Dirichlet prior with parameter $\alpha = (\alpha_1, \alpha_2, \dots)$ and you observe outcome 1, for example, then then posterior distribution is also Dirichlet with

Subsection 2

Stan

Plan

In this short lecture we:

- fire up stan
- implement a simple linear model and talk through the main model blocks
- implement a simple hierarchical model
- describe a behavioral game and set up a model to recover some parameters of interest, given the game

Getting set up

The good news: There is lots of help online. Start with: https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started

We will jump straight into things and work through a session.

1. Install the stan package and fire up. Useful to set options so that multiple cores are being used:

```
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
```

One variable model: Simple example

2. Now lets consider the simplest one var linear model.
   - We will need model code
   - And data

A simple model: Code

To implement a stan model you should write the code in a text editor and save it as a text file. You can also write it directly in your script. You can then bring the file into R or call the file directly.

- There are many examples of stan models here:
  - https://github.com/stan-dev/example-models/tree/master/ARM/Ch.4
  - https://github.com/stan-dev/example-models/wiki

A simple model: Code

I saved a simple model called `one_var.stan` locally. Here it is:

```
readLines("assets/one_var.stan", warn = FALSE) |>
  cat(sep = "\n")
```

```
data {
  int<lower=0> N;
  vector[N] Y;
  vector[N] X;
}
parameters {
  real a;
  real b;
  real<lower=0> sigma;
}
model {
  Y ~ normal(a + b * X, sigma);
```

A simple model: Code

The key features here are (read from bottom up!):

- Y is assumed to be normally distributed with mean `a + bX` and standard deviation `sigma`.
- There are then three parameters: `a`, `b`, `sigma`.
- There are no priors placed on these but sigma is constrained to be positive. Without priors improper flat priors are assumed.
- Stan expects a data set that contains three things: a scalar, `N` and `X1,Y‘` data

Simple model: Data

We feed data to the model in the form of a list. The idea of a list is that the data can include all sorts of objects, not just a single dataset.

```
X = rnorm(20)

some_data <- list(
 N = 20,
 X = X,
 Y = X + rnorm(20)
 )
```

Simple model: Now let's Run It

```
M <- stan(file = "assets/one_var.stan",
          data = some_data)
```

When you run the model you get a lot of useful output on the estimation and the posterior distribution. Here though are the key results: ::: {.cell} ::: {.cell-output-display}

	mean	sd	Rhat
a	-0.179	0.214	1
b	0.738	0.183	1
sigma	0.950	0.175	1

::: :::

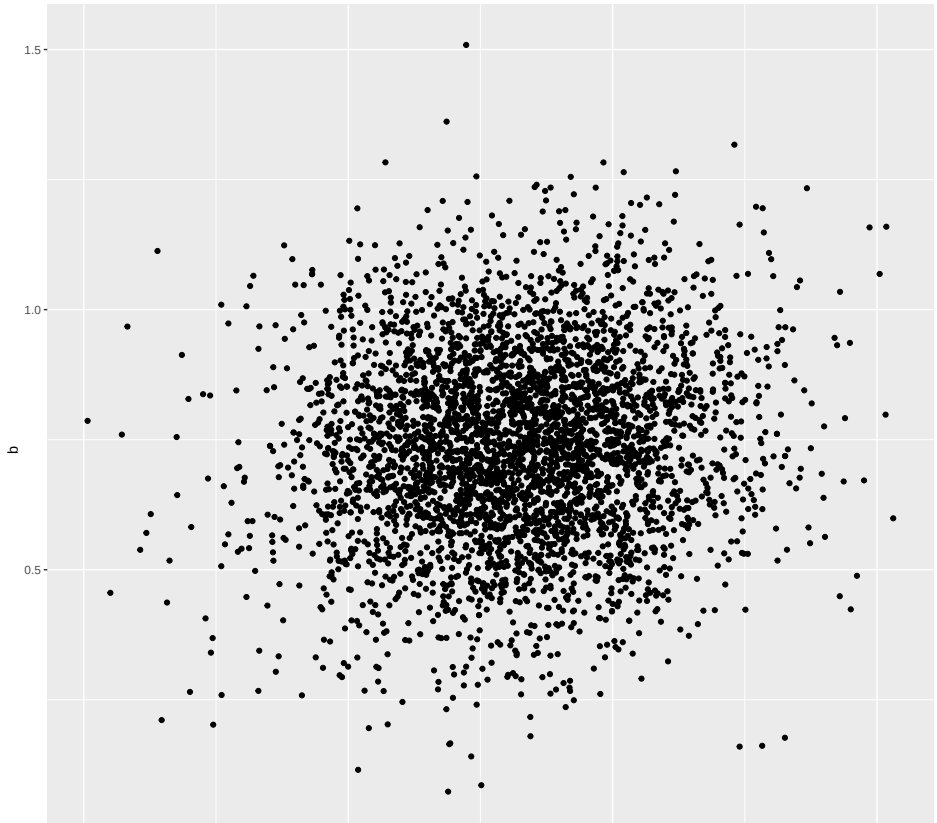
These look good.

The Rhat at the end tells you about convergence. You want this very close to 1.

A simple model: Now lets use it

The model output contains the full posterior distribution.

```
my_posterior <- M |> extract() |> data.frame()

my_posterior |> ggplot(aes(a,b)) + geom_point()
```



A simple model: Now lets use it

With the full posterior you can can look at marginal posterior distributions over arbitrary transformations of parameters.

```
summary((my_posterior$a + my_posterior$b)/my_posterior$a) |> t
  kable(digits = 2)
```

```
Error in UseMethod("tidy"): no applicable method for 'tidy' ap
```

Building up

Let's go back to the code.

There we had three key blocks: `data`, `parameters`, and `model`

More generally the blocks you can specify are:

- `data` (define the vars that will be coming in from the data list)
- `transformed data` (can be used for preprocessing)
- `parameters` (required: defines the parameters to be estimated)
- `transformed parameters` (transformations of parameters useful for computational reasons and sometimes for clarity)
- `model` (give priors and likelihood)
- `generated quantities` (can be used for post processing)

Parameters block

The parameters block declared the set of parameters that we wanted to estimate. In the simple model these were `a`, `b`, and `sigma`. Note in the declaration we also:

- said what kind of parameters they (vectors, matrices, simplices etc)
- gave possible constraints

Parameters block

Instead of defining: ::: {.cell}

```
real a;
real b;
```

:::

We could have defined ::: {.cell}

```
vector[2] coefs;
```

:::

and then referenced `coef[1]` and `coef[2]` in the model block.

Parameters block

Or we could also have imposed the constraint that the slope coefficient is positive by defining: ::: {.cell}

```
real a;
real<lower = 0> b;
```

:::

Model Block

In the model block we give the likelihood

But we can also give the priors (if we want to). If priors are not provided, flat (possibly improper) priors are assumed

In our case for example we could have provided something like

```
model {
  b ~ normal(-10, 1);
  Y ~ normal(a + b * X, sigma);
}
```

This suggests that we start off believing b is centered on -10. That will surely matter for our conclusions. Lets try it:

Version 2

This time I will write the model right in the editor:

```
new_model <- '
data {
  int<lower=0> N;
  vector[N] Y;
  vector[N] X;
}
parameters {
  real a;
  real b;
  real<lower=0> sigma;
}
model {
  b ~ normal(-10,1);
  Y ~ normal(a + b * X, sigma);
}
```

Estimation 2

```
M2 <- stan(model_code = new_model, data = some_data)
```

	mean	sd	Rhat
a	-1.338	2.444	1.003
b	-7.875	1.172	1.003
sigma	10.988	2.499	1.001

Note that we get a much lower estimate for `b` with the same data.

A multilevel model

Now imagine a setting in which there are 10 villages, each with 10 respondents. Half in each village are assigned to treatment $X = 1$, and half to control $X = 0$.

Say that there is possible a village specific average outcome: $Y_v = a_v + b_v X$ where a_v and b_v are each draw from some distribution with a mean and variance of interest. The individual outcomes are draws from a village level distribution centered on the village specific average outcome.

This all implies a multilevel structure.

A ml model

Here is a model for this

```
ml_model <- '
data {
  vector[100] Y;
  int<lower=0,upper=1> X[100];
  int village[100];
}
parameters {
  vector<lower=0>[3] sigma;
  vector[10] a;
  vector[10] b;
  real mu_a;
  real mu_b;
}
transformed parameters {
  vector[100] Y_vx;
  for (i in 1:100) Y_vx[i] = a[village[i]] + b[village[i]] * X[i];
}
model {
  a ~ normal(mu_a, sigma[1]);
  b ~ normal(mu_b, sigma[2]);
  Y ~ normal(Y_vx, sigma[3]);
}
'
```

Here is a slightly more general version:

https://github.com/stan-dev/example-models/blob/master/ARM/Ch.17/17.1_radon_vary_inter_slope.stan

Multilevel model: Data

Lets create some multilevel data. Looking at this, can you tell what is the typical village level effect? How much heterogeneity is there?

```
village   <- rep(1:10, each = 10)
village_b <- 1 + rnorm(10)
X         <- rep(0:1, 50)
Y         <- village_b[village]*X + rnorm(100)

ml_data <- list(
  village = village,
  X = X,
  Y = Y)
```

Multilevel Results

```
M_ml <- stan(model_code = ml_model, data = ml_data)
```

	mean	sd	Rhat
mu_a	-0.29	0.25	1.00
mu_b	1.89	0.26	1.00
sigma[1]	0.61	0.25	1.00
sigma[2]	0.47	0.28	1.01
sigma[3]	0.98	0.08	1.00

A game and a structural model

Say that a set of people in a population are playing sequential prisoner's dilemmas.

In such games selfish behavior might suggest defections by everyone everywhere. But of course people often cooperate. Why might this be?

- One possible reason is that some people are irrational, in the sense that they simply choose to cooperate, ignoring the payoffs.
- Another possibility is that rational people think that others are irrational, in the sense that they think that others will reciprocate when they observe cooperative action

Model

We will capture some of this intuition with a behavioral type model in which

- each player has a "rationality" propensity of r_i – this is the probability with which they choose to do the rational thing, rather than the generous thing
- $r_i \sim U[0, \theta]$ for $\theta > .5$.
- A player with rationality propensity of r_i believes $r_j \sim [0, r_i]$. So everyone assumes that they are the most rational people in the room...
- The game is such that: * second mover: a second mover with rationality propensity r_i will cooperate with probability $1 - r_i$ if the first mover cooperated; otherwise they defect * first mover: a first mover with r_i will cooperate nonstrategically with probability $(1 - r_i)$; however with probability r_i they will also cooperate *strategically* if they think that the second mover has $r_j < .25$.

Expectations from model

In all, this means that a player with propensity $r_i > .5$ will cooperate with probability $1 - r_i$; a player with propensity $r_i < .5$ will cooperate with probability 1.

Interestingly the not-very-rational people sometimes cooperate strategically but the really rational people never cooperate strategically because they think it won't work.

Event Probabilities

What then are the probabilities of each of the possible outcomes?

- There will be cooperation by *both* players with probability $(\int_0^{.5} p(r_i)dr_i + \int_{.5}^1 p(r_i)(1-r_i)dr_i) \int_0^1 p(r_i)(1-r_i)dr_i$
- There will be cooperation by player 1 only with probability $(\int_0^{.5} p(r_i)dr_i + \int_{.5}^1 p(r_i)(1-r_i)dr_i)(\int_0^1 p(r_i)(r_i)dr_i)$
- There will be cooperation by neither with probability: $1 - \int_0^{.5} p(r_i)dr_i - \int_{.5}^1 p(r_i)(1-r_i)dr_i$

where p is the density function on r_i given θ

Event probabilities

Given the assumption on p

- There will be cooperation by *both* players with probability $(1 + .25/\theta - .5\theta)(1 - .5\theta)$
- There will be cooperation by player 1 only with probability $(1 + .25/\theta - .5\theta)(.5\theta)$
- There will be cooperation by neither player with probability $(.5\theta - .25/\theta)$

Data

- We have data on the actions of the first movers and the second movers and are interested in the distribution of the p_is.
- Lets collapse that data into a simple list of the number of each type of game outcome:
- And say we start off with a uniform prior of θ.
- What should we conclude about θ?

Model

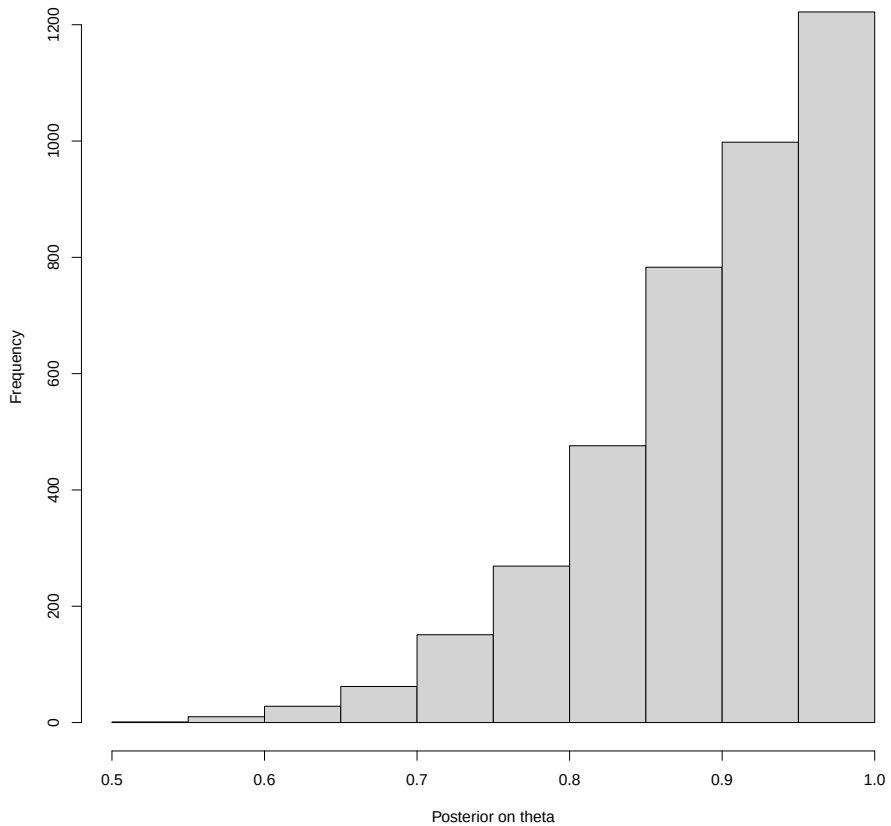
Here's a model:

```
game_model <- '
data {
  int<lower=0> play[3];
}
parameters {
  real<lower=.5, upper=1> theta;
}
transformed parameters {
simplex[3] w;
 w[1] = (1+.25*theta - .5*theta)*(1-.5*theta);
 w[2] = (1+.25*theta - .5*theta)*(.5*theta);
 w[3] = (-.25*theta  + .5*theta);
}
model {
  play ~ multinomial(w);
}
'
```

Note we define event weights as transformed parameters on a simplex. We also constrain θ to be $> .5$. Obviously we are relying *a lot* on our model.

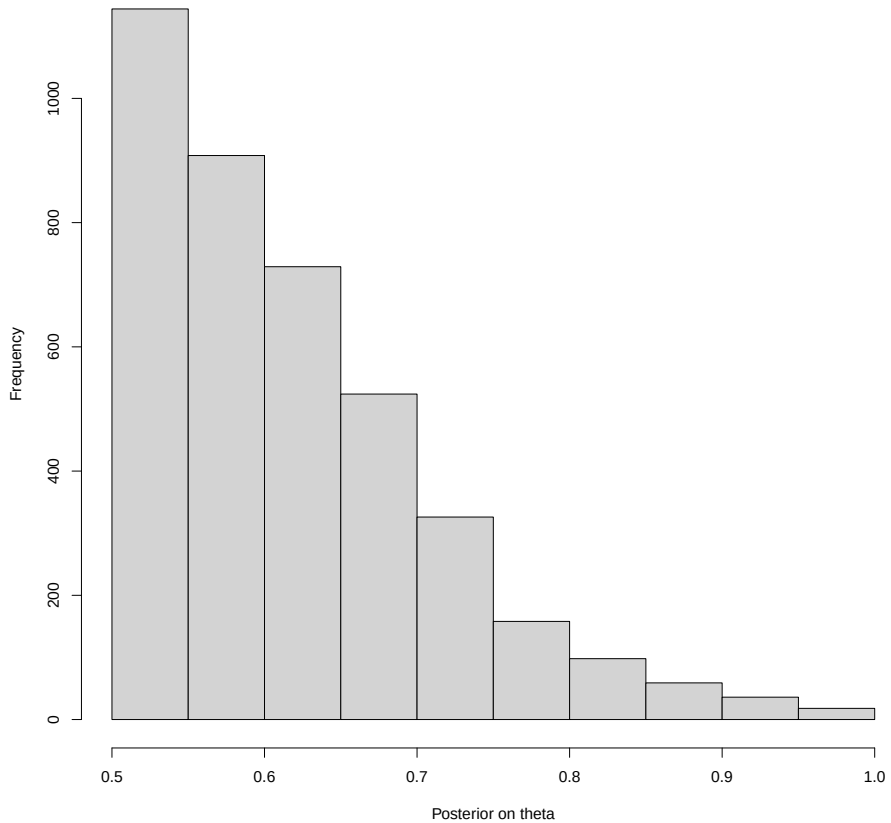
Plot posterior on θ

```
M3 <- stan(model_code = game_model,
           data = list(play = c(10,10,10)))
```



Plot posterior on θ

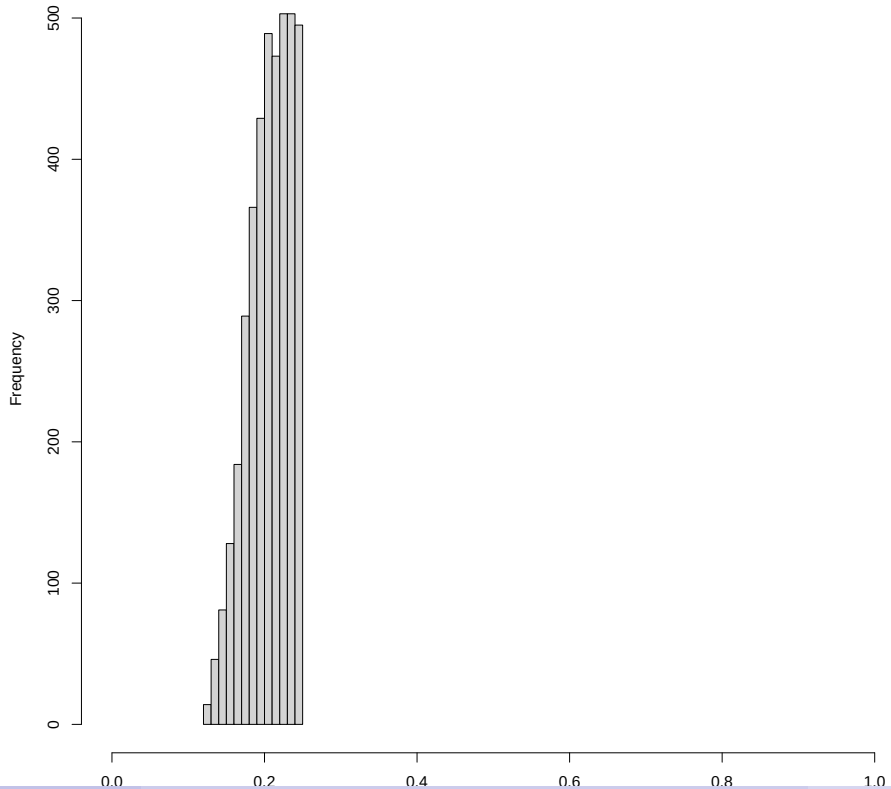
```
M4 <- stan(model_code = game_model,
           data = list(play = c(20,6,4)))
```



Posterior on a quantity of interest

What is the probability of observing *strategic* first round cooperation?

A player with rationality r_i will cooperate strategically with probability r_i if $r_i < .5$ and 0 otherwise. Thus we are interested in $\int_0^{.5} r_i/\theta dr_i = .125/\theta$



Section 7

Design

Subsection 1

Topics

Topics

- Sampling schemes
- Randomization schemes

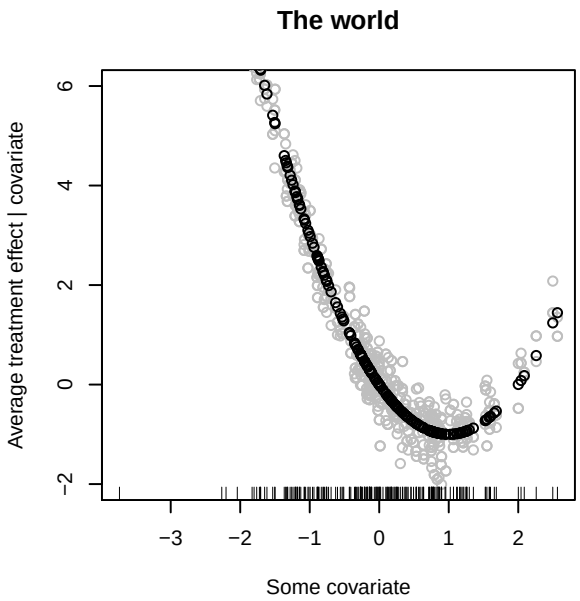
Experiments

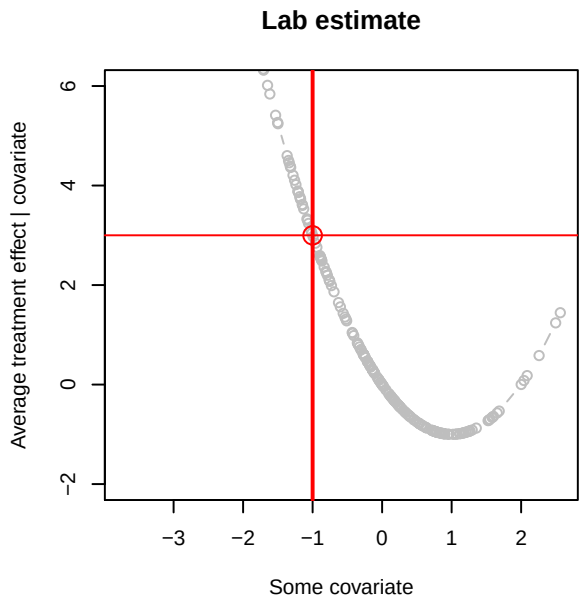
- Experiments are investigations in which an intervention, in all its essential elements, is under the control of the investigator. (Cox & Reid)
- Two major types of control:

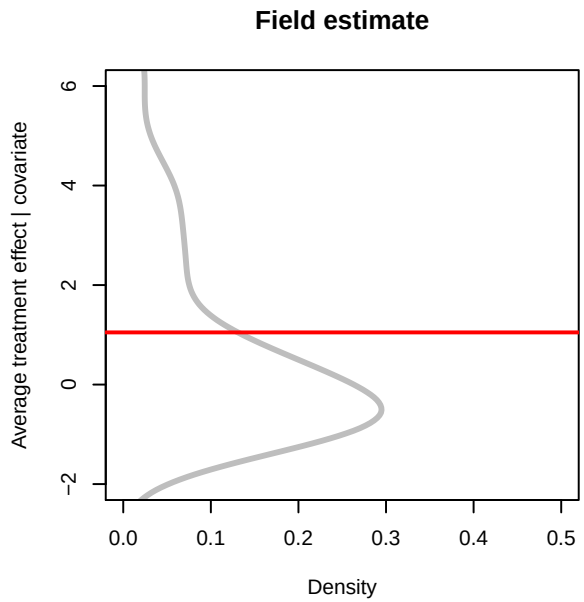
```
1. control over assignment to treatment -- this is a
2. control over the treatment itself -- this is at t
```

- Main focus today is on 1 and on the question: *how does control over assignment to treatment allow you to make reasonable statements about causal effects?*

Experiments







Basic randomization

▸ Top

- Basic randomization is very simple. For example, say you want to assign 5 of 10 units to treatment. Here is simple code:

```
1:10 %in% sample(1:10, 5)
```

```
 [1] FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TR
```

...should be replicable

In general you might want to set things up so that your randomization is **replicable**. You can do this by setting a **seed**:

```
set.seed(20111112)
1:10 %in% sample(1:10, 5)
```

```
 [1] FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FAI
```

```
set.seed(20111112)
1:10 %in% sample(1:10, 5)
```

```
 [1] FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FAI
```

Basic randomization

Even better is to set it up so that it can reproduce **lots of possible draws** so that you can check the propensities for each unit.

```
set.seed(20111112)
P <- sapply(1:1000, function(i) 1:10 %in% sample(1:10, 5))
apply(P, 1, mean)
```

```
 [1] 0.495 0.493 0.512 0.522 0.505 0.522 0.479 0.510 0.444 0.5
```

Here the P matrix gives 1000 possible ways of allocating 5 of 10 units to treatment. We can then confirm that the average propensity is 0.5.

- A huge advantage of this approach is that if you make a mess of the random assignment; **you can still generate the P matrix and use that for all analyses**!

Do it in advance

- Unless you need them to keep subjects at ease, leave your spinners and your dice and your cards behind.
- Especially when you have multiple or complex randomizations you are generally much better doing it with a computer in advance

	IDENTIF	VARIATIONS														
Order	ID	V1 SURVEY MP	V1 NAME OF SURVEY MP	V1 PARTY	V2 INTIMIDATIO	V3 FRAUD	V4 ELEMT	V5 CARD	V5 SCORE TYPE	V5 SCORE	V5 ENDORSE	V5 WHICH ENDORSE	V6 SMS	V6 PRICE	V6 SEE	V7 FIND	V7 MP LCV
1	573411	Constituency	Musumba Isaac Isanga	NRM	A	D	Yes	Yes	CONSTIT	E	No		No		.	Yes	LCV
2	573412	Constituency	Musumba Isaac Isanga	NRM	B	D	No	Yes	PLENY	E	Yes	G	Yes	50 Sh	Yes	No	
3	573422	Constituency	Musumba Isaac Isanga	NRM	B	C	Yes	Yes	PLENY	E	No		Yes	Full Price	Yes	No	
4	573421	Constituency	Musumba Isaac Isanga	NRM	A	C	No	Yes	CONSTIT	E	Yes	C	No		.	No	
1	717221	Women	Alitwala Kadaga Rebecca	NRM	B	D	No	Yes	CONSTIT		No		No		.	Yes	WOM
2	717211	Women	Alitwala Kadaga Rebecca	NRM	A	D	Yes	Yes	PLENY		Yes	E	Yes	Free	No	No	
3	717212	Constituency	Balikowa Henry	NRM	B	C	Yes	No	.		.		No		.	No	
4	717222	Constituency	Balikowa Henry	NRM	A	C	No	No	.		.		Yes	Free	No	No	
1	717421	Women	Alitwala Kadaga Rebecca	NRM	A	C	Yes	Yes	CONSTIT		No		Yes	Full Price	Yes	Yes	WOM
2	717412	Women	Alitwala Kadaga Rebecca	NRM	B	C	No	Yes	PLENY		Yes	A	No		.	No	
3	717411	Constituency	Balikowa Henry	NRM	B	D	Yes	Yes	CONSTIT	C	Yes	D	No		.	No	

Figure 6: A survey dictionary with results from a complex randomization presented in a simple way for enumerators

Did the randomization "work"?

* People often wonder: did randomization work? * Common practice is to implement a set of t-tests to see if there is balance * This makes no sense.

* If you doubt whether it was **implemented** properly do an F test * If you worry about **variance** specify controls in advance as a function of relation with outcomes (more on this later) * If you worry about **conditional bias** then look at substantive differences between groups, not t–tests

Subsection 2

Cluster Randomization

Cluster Randomization

- Simply place units into groups (clusters) and then randomly assign the groups to treatment and control.
- All units in a given group get the same treatment

Note: clusters are part of your design, not part of the world.

Cluster Randomization

- Often used if intervention has to function at the cluster level *or* if outcome defined at the cluster level.
- **Disadvantage:*** } loss of statistical power
- However: perfectly possible to assign *some* treatments at cluster level and then *other* treatments at the individual level
- **Principle:** (unless you are worried about spillovers) generally make clusters as small as possible
- **Principle:** Surprisingly, variability in cluster size makes analysis harder. (See analysis section)
- **Be clear** about whether you believe effects are operating at the cluster level or at the individual level. This matters for power calculations.
- **Be clear** about whether spillover effects operate only within clusters or also across them. If within only you might be able to interpret treatment as the effect of being in a treated cluster...

Cluster Randomization: Block by cluster size

Surprisingly, if clusters are of different sizes the difference in means estimator is *not* unbiased, even if all units are assigned to treatment with the same probability. **Here's the intuition.** Say there are two clusters each with homogeneous

treatment effects:

Cluster	Size	Y0	Y1
1	1000000	0	1
2	1	0	0

Then: * What is the true average treatment effect? * What do you expect to estimate from cluster random assignment?

The solution is to block by cluster size. For more see: http://gking.harvard.edu/files/cluster.pdf

Subsection 3

Blocked assignments and other restricted randomizations

Blocking

There are more or less **efficient** ways to randomize.

- Randomization helps ensure good balance on all covariates (observed and unobserved) *in expectation*.
- But balance may not be so great *in realization*
- Blocking can help ensure balance ex post on observables

Consider a case with four units and two strata. There are 6 possible assignments of 2 units to treatment:

ID	X	Y(0)	Y(1)	R1	R2	R3	R4	R5	R6
1	1	0	1	1	1	1	0	0	0
2	1	0	1	1	0	0	1	1	0
3	2	1	2	0	1	0	1	0	1
4	2	1	2	0	0	1	0	1	1
$\widehat{\tau}$:				0	1	1	1	1	2

Even with a constant treatment effect and everything uniform within blocks, there is variance in the estimation of $\widehat{\tau}$. This can be eliminated by excluding R1 and R6.

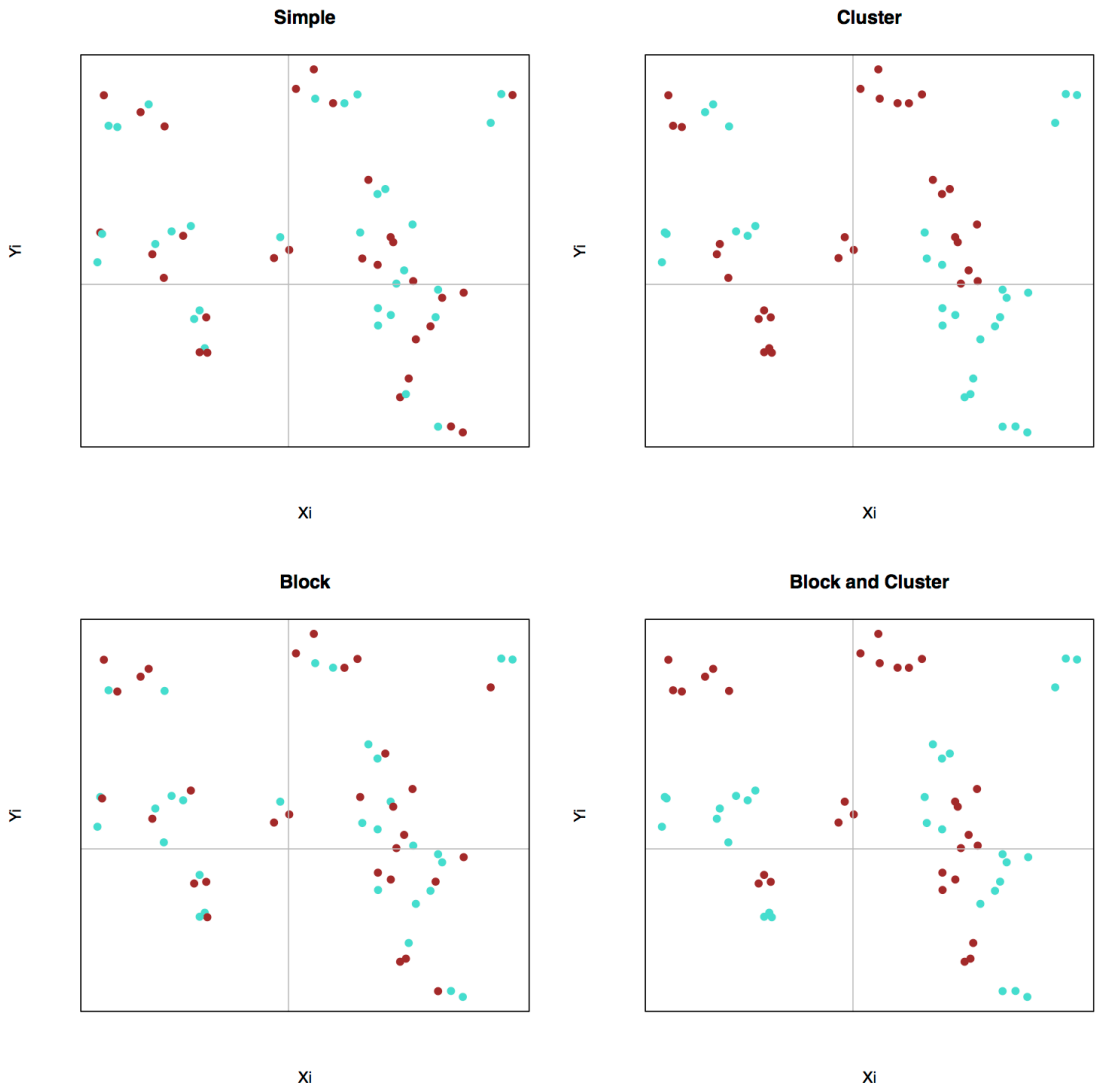
Blocking

Simple blocking in R (5 pairs):

```
sapply(1:5, function(i) rank(runif(2))==1)
```

1	2	3	4	5
TRUE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	TRUE

Of blocks and clusters



Blocking

- Blocking is a case of **restricted randomization**. Although each unit is sampled with equal probability, the *profiles* of possible assignments are not.
- You have to take account of this when doing analysis
- There are many other approaches.
  - **Matched Pairs** are a particularly fine approach to blocking
  - You could also randomize and then **replace the randomization** if you do not like the balance. This sounds tricky (and it is) but it is OK as long as you understand the true lottery process you are employing and incorporate that into analysis
  - It is even possible to block on **covariates for which you don't have data** ex ante, by using methods in which you allocate treatment over time as a function of features of your sample (also tricky)

Other types of restricted randomization

- Really you can set whatever criterion you want for your set of treated units to have (eg no treated unit beside another treated unit; at least 5 from the north, 10 from the south, guaranteed balance by some continuous variable etc)
- You just have to be sure that you understand the random process that was used and that you can use it in the analysis stage
- But here be dragons
  - The more complex your design, the more complex your analysis.
  - General injunction (Senn 2004 "as ye randomize so shall ye analyze' ')
  - In general you should make sure that a given randomization procedure coupled with a given estimation procedure will produce an unbiased estimate. `DeclareDesign` can help with this.

Subsection 4

Factorial Designs

Factorial Designs

- Often when you set up an experiment you want to look at more than one treatment.
- Should you do this or not? How should you use your power?

Factorial Designs

- Often when you set up an experiment you want to look at more than one treatment.
- Should you do this or not? How should you use your power?

	$T2=0$	$T2=1$
$T1=0$	50%	0%
$T1=1$	50%	0%

	$T2=0$	$T2=1$
$T1=0$	25%	25%
$T1=1$	25%	25%

	$T2=0$	$T2=1$
$T1=0$	33.3%	33.3%
$T1=1$	33.3%	0%

Factorial Designs

- Surprisingly adding multiple treatments does not eat into your power (unless you are decomposing a complex treatment – then it can. Why?)
- Especially when you use a fully crossed design like the middle one above.
- Fisher: "No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken."
- However – adding multiple treatments *does* alter the **interpretation** of your treatment effects. If T2 is an unusual treatment for example, then half the T1 effect is measured for unusual situations.

Factorial Designs: In practice

- In practice if you have a lot of treatments it can be hard to do full factorial designs – there may be too many combinations.
- In such cases people use **fractional factorial designs**, like the one below (5 treatments but only 8 units!)

Variation	T1	T2	T3	T4	T5
1	0	0	0	1	1
2	0	0	1	0	0
3	0	1	0	0	1
4	0	1	1	1	0
5	1	0	0	1	0
6	1	0	1	0	1
7	1	1	0	0	0
8	1	1	1	1	1

- Then randomly assign units to rows. Note columns might also be blocking covariates.
- In R, look at `{library(survey); hadamard(7)}`

Factorial Designs: In practice

- But be careful: you have to be comfortable with possibly not having any simple counterfactual unit for any unit (invoke sparsity-of-effects principle).

Unit	T1	T2	T3	T4	T5
1	0	0	0	1	1
2	0	0	1	0	0
3	0	1	0	0	1
4	0	1	1	1	0
5	1	0	0	1	0
6	1	0	1	0	1
7	1	1	0	0	0
8	1	1	1	1	1

- In R, look at `{library(survey); hadamard(7)}`

Subsection 5

External Validity: Can randomization strategies help?

Principle: Address **external validity** at the design stage

Anything to be done on randomization to address external validity concerns?

- **Note 1**: There is little or nothing about field experiments that makes the external validity problem greater for these than for other "sample based'' research
- **Note 2**: Studies that use up the available universe (cross national studies) actually have a distinct external validity problem
- Two ways to think about external validity issues:
  1. Are things likely to operate in other units like they operate in these units? (even with the same intervention)
  2. Are the processes in operation in this treatment likely to operate in other treatments? (even in this population)

Principle: Address **external validity** at the design stage

- Two ways to think about external validity issues:
  1. Are things likely to operate in other units like they operate in these units? (even with the same intervention) 2.Are the processes in operation in this treatment likely to operate in other treatments? (even in this population)
- Two approaches for 1.
  - Try to sample cases and estimate *population average treatment effects*
  - Exploit internal variation: block on features that make the case unusal and assess importance of these (eg is unit poor? assess how effects differ in poor and wealthy components)
- 2 is harder and requires a sharp identification of context free primitives, if there are such things.

Subsection 6

Assignments with ‘DeclareDesign“

A design: Multilevel data

A design with hierarchical data and different assignment schemes.

```
design <-
  declare_model(
    school = add_level(N = 16,
                       u_school = rnorm(N, mean = 0)),
    classroom = add_level(N = 4,
                      u_classroom = rnorm(N, mean = 0)),
    student =  add_level(N = 20,
                         u_student = rnorm(N, mean = 0))
    ) +
  declare_model(
    potential_outcomes(Y ~ .1*Z + u_classroom + u_student + u_
    ) +
  declare_assignment(Z = simple_ra(N)) +
  declare_measurement(Y = reveal_outcomes(Y ~ Z))  +
  declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0)) +
```

Sample data

Here are the first couple of rows and columns of the resulting data frame.

```
my_data <- draw_data(design)
kable(head(my_data), digits = 2)
```

school	u_school	classroom	u_classroom	student	u_student	Y_Z
01	-0.77	01	-0.06	0001	0.36	-0
01	-0.77	01	-0.06	0002	0.16	-0
01	-0.77	01	-0.06	0003	1.04	0
01	-0.77	01	-0.06	0004	1.54	0
01	-0.77	01	-0.06	0005	-0.99	-1
01	-0.77	01	-0.06	0006	-0.70	-1

Sample data

Here is the distribution between treatment and control:

```
kable(t(as.matrix(table(my_data$Z))),
      col.names = c("control", "treatment"))
```

```
Error in eval(substitute(expr), data, enclos = parent.frame()
```

Complete Random Assignment using the built in function

```
assignment_complete <-   declare_assignment(Z = complete_ra(N)

design_complete <-
  replace_step(design, "assignment", assignment_complete)
```

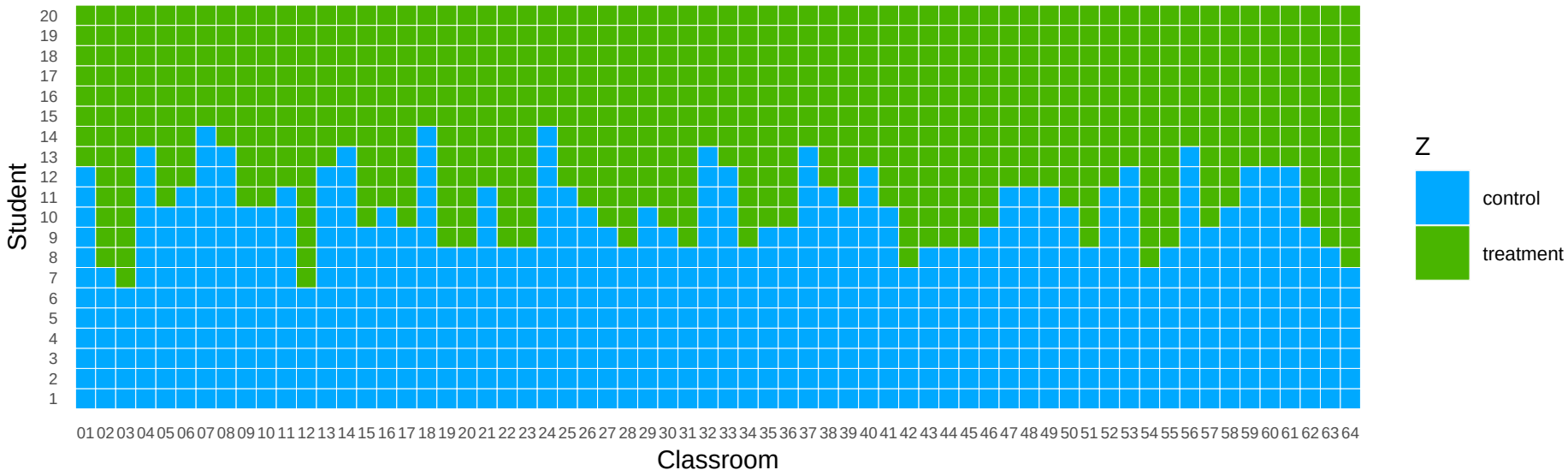
Data from complete assignment

We can draw a new set of data and look at the number of subjects in the treatment and control groups.

```
set.seed(1:5)
data_complete <- draw_data(design_complete)

kable(t(as.matrix(table(data_complete$Z))))
```

```
Error in eval(substitute(expr), data, enclos = parent.frame()
```

Plotted


Student
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
Classroom
Z
control
treatment

Block Random Assignment

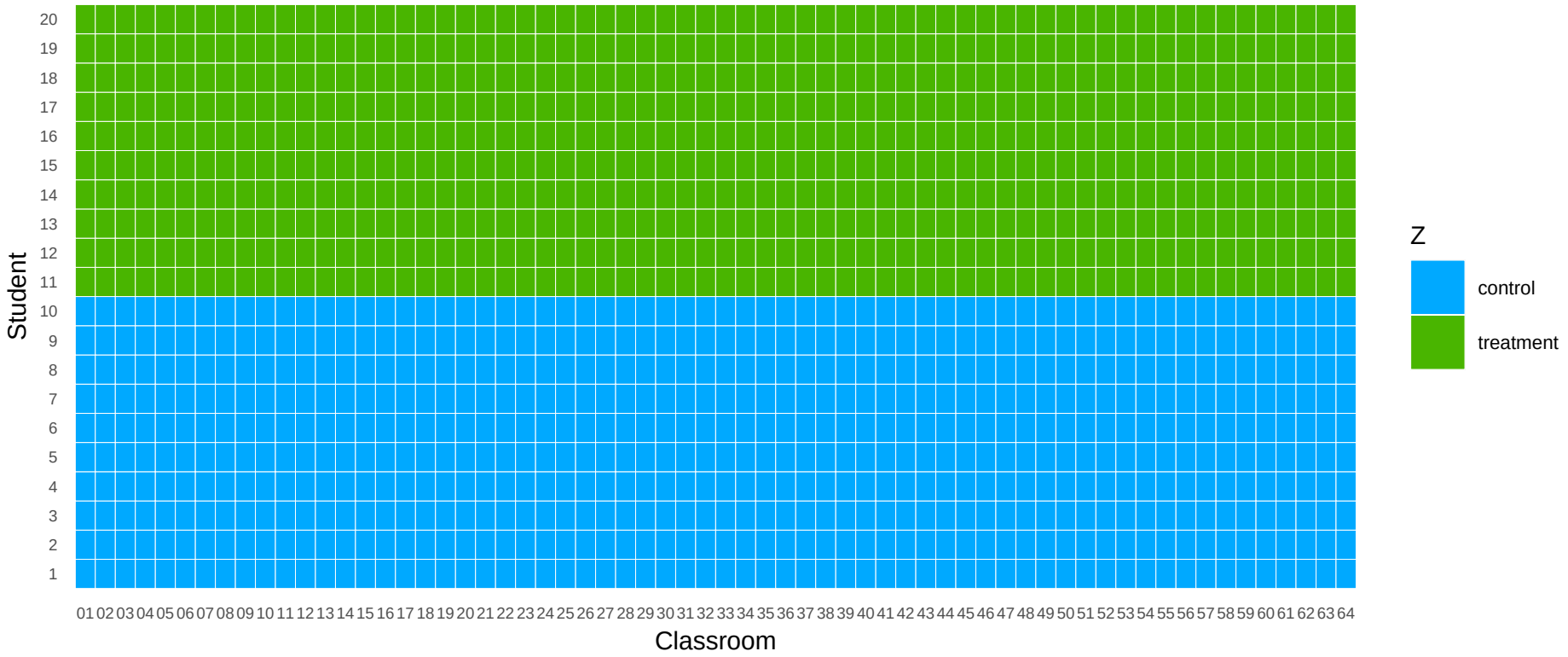
- The treatment and control group will **in expectation** contain the same share of students in different classrooms.
- But as we saw this does necessarily hold in **realization**
- We make this more obvious by sorting the students by treatment status with schools

Blocked design

```
assignment_blocked <-
  declare_assignment(Z = block_ra(blocks = classroom))

estimator_blocked <-
  declare_estimator(Y ~ Z, blocks = classroom,
                    .method = difference_in_means)

design_blocked <-
  design |>
  replace_step("assignment", assignment_blocked) |>
  replace_step("estimator", estimator_blocked)
```

Illustration of blocked assignment

- Note that subjects are sorted here after the assignment to make it easier to see that in this case blocking ensures that exactly 5 students within each classroom are assigned to treatment.

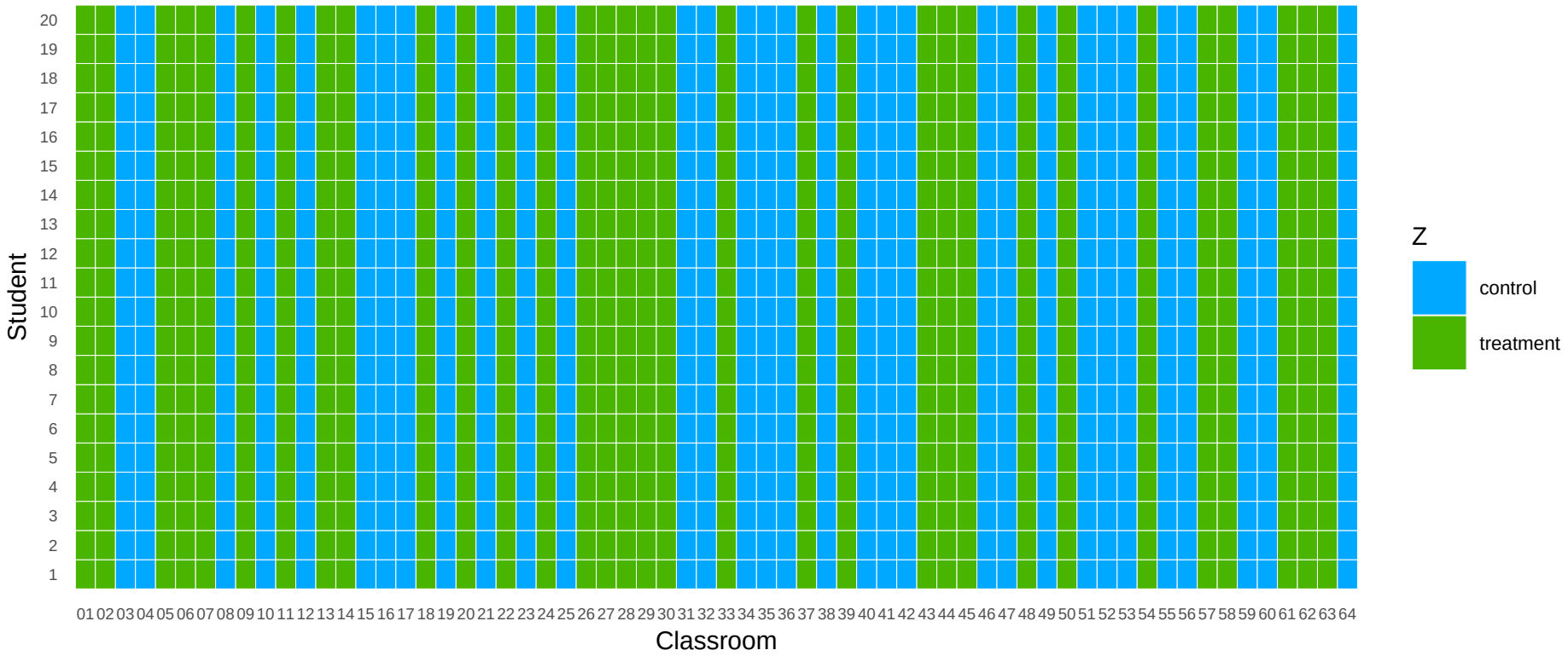


Clustering

But what if all students in a given class have to be assigned the same treatment?

```
assignment_clustered <-
  declare_assignment(Z = cluster_ra(clusters = classroom))
estimator_clustered <-
  declare_estimator(Y ~ Z, clusters = classroom,
                    .method = difference_in_means)


design_clustered <-
  design |>
  replace_step("assignment", assignment_clustered) |>
  replace_step("estimator", estimator_clustered)
```

Illustration of clustered assignment



Clustered and Blocked

```
assignment_clustered_blocked <-
  declare_assignment(Z = block_and_cluster_ra(blocks = school,
                                              clusters = class
estimator_clustered_blocked <-
  declare_estimator(Y ~ Z, blocks = school, clusters = classro
                    .method = difference_in_means)


design_clustered_blocked <-
  design |>
  replace_step("assignment", assignment_clustered_blocked) |>
  replace_step("estimator", estimator_clustered_blocked)
```

Illustration of clustered and blocked assignment

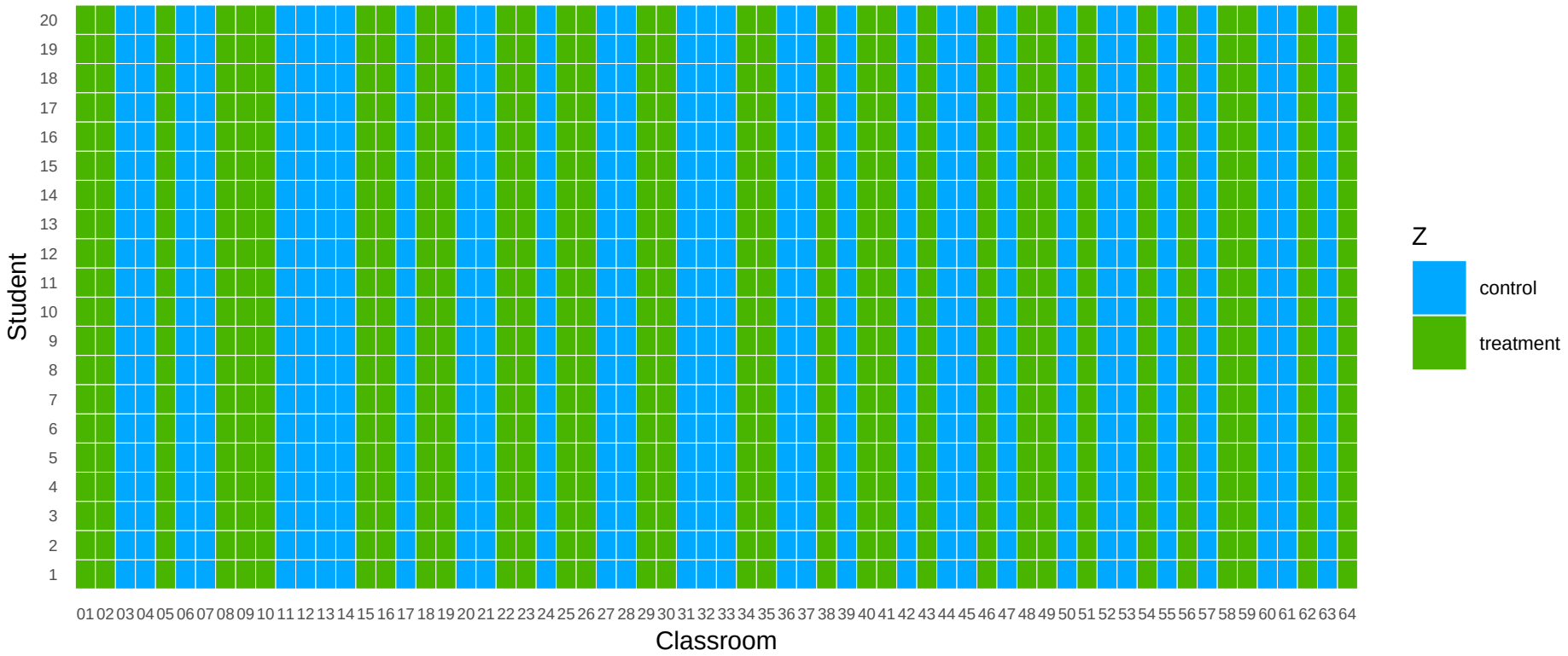


Illustration of efficiency gains from blocking

```
designs <-
  list(
    simple = design,
    complete = design_complete,
    blocked = design_blocked,
    clustered = design_clustered,
    clustered_blocked = design_clustered_blocked)
```

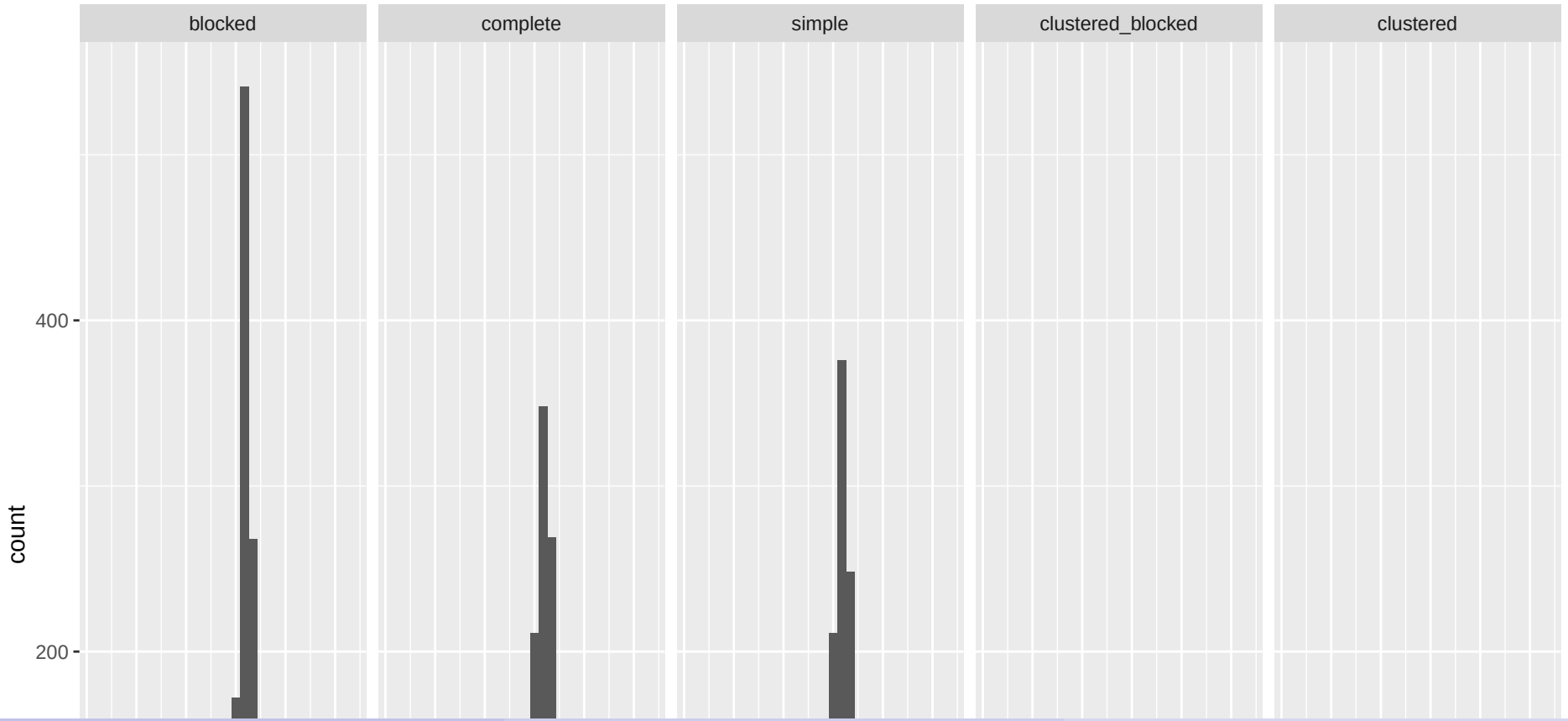
```
diagnoses <- diagnose_design(designs)
```

Illustration of efficiency gains from blocking

Design	Power	Coverage
simple	0.16	0.95
	(0.01)	(0.01)
complete	0.20	0.96
	(0.01)	(0.01)
blocked	0.42	0.95
	(0.01)	(0.01)
clustered	0.06	0.96
	(0.01)	(0.01)
clustered_blocked	0.08	0.96
	(0.01)	(0.01)

Sampling distributions

```
diagnoses$simulations_df |>
  mutate(design = factor(design, c("blocked", "complete", "sim
  ggplot(aes(estimate)) +
  geom_histogram() + facet_grid(~design)
```



Nasty integer issues

Section 8

Design diagnosis

Subsection 1

Outline

Outline

1. Tests review
2. p values and significance
3. Power
4. Sources of power
5. Advanced applications

Subsection 2

Tests

Review

In the classical approach to testing a hypothesis we ask:

How likely are we to see data like this if indeed the hypothesis is true?

- If the answer is "not very likely" then we treat the hypothesis as suspect.
- If the answer is *not* "not very likely" then the hypothesis is maintained (some say "accepted" but this is tricky as you may want to "maintain" multiple incompatible hypotheses)

How unlikely is "not very likely"?

Weighing Evidence

When we test a hypothesis we decide first on what sort of evidence we need to see in order to decide that the hypothesis is not reliable.

- **Othello** has a hypothesis that Desdemona is innocent.
- **Iago** confronts him with evidence:
  - See how she looks at him: would she look a him like that if she were innocent?
  - ... would she defend him like that if she were innocent?
  - ... would he have her handkerchief if she were innocent?
  - Othello, the chances of all of these things arising if she were innocent is surely less than 5%

Hypotheses are often rejected, sometimes maintained, but rarely accepted

- Note that Othello is focused on the probability of the events if she were innocent but not the probability of the events if Iago were trying to trick him.
- He is not assessing his belief in whether she is faithful, but rather how likely the data would be if she were faithful.

So:

- He assesses: $\Pr(\text{Data}|\text{Hypothesis is TRUE})$
- While a Bayesian would assess: $\Pr(\text{Hypothesis is TRUE}|\text{Data})$

Recap: Calculate a p value in your head

- Illustrating p values via "randomization inference"
- Say you randomized assignment to treatment and your data looked like this.

Unit	1	2	3	4	5	6	7	8	9	10
Treatment	0	0	0	0	0	0	0	1	0	0
Health score	4	2	3	1	2	3	4	8	7	6

Then:

- Does the treatment improve your health?
- What's the p value for the null that treatment had no effect on anybody?

Calculate a p value in your head

- Illustrating p values via "randomization inference"
- Say you randomized assignment to treatment and your data looked like this.

Unit	1	2	3	4	5	6	7	8	9	10
Treatment	0	0	0	0	0	0	0	0	1	0
Health score	4	2	3	1	2	3	4	8	7	6

Then:

- Does the treatment improve your health?
- What's the p value for the null that treatment had no effect on anybody?

Subsection 3

Power

Subsection 4

What power is

What power is

Power is just the probability of ~~getting a significant result~~ rejecting a hypothesis.

Simple enough but it presupposes:

- A well defined hypothesis
- An actual stipulation of the world under which you evaluate the probability
- A procedure for producing results and determining of they are significant / rejecting a hypothesis

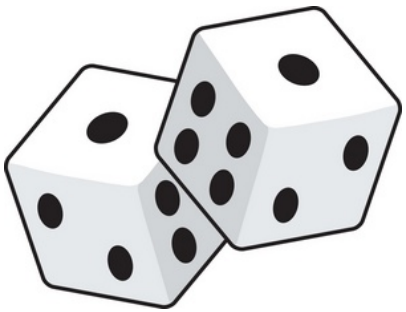
By hand

I want to test the hypothesis that a six never comes up on this dice.

Here's my **test**:

- I will roll the dice **once**.
- If a six comes up I will reject the hypothesis.

What is the power of this test?



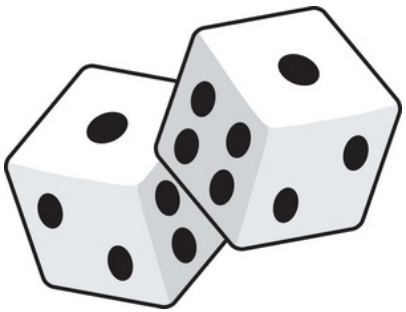
By hand

I want to test the hypothesis that a six never comes up on this dice.

Here's my **test**:

- I will roll the dice **twice**.
- If a six comes up **either time** I will reject the hypothesis.

What is the power of *this* test?



Two probabilities

Power sometimes seems more complicated because hypothesis rejection involves a calculated probability and so you need the probability of a probability.

I want to test the hypothesis that this dice is *fair*.

Here's my **test**:

- I will roll the dice **1000** times and if I see fewer than x 6s or more than y 6s I will reject the hypothesis.

Now:

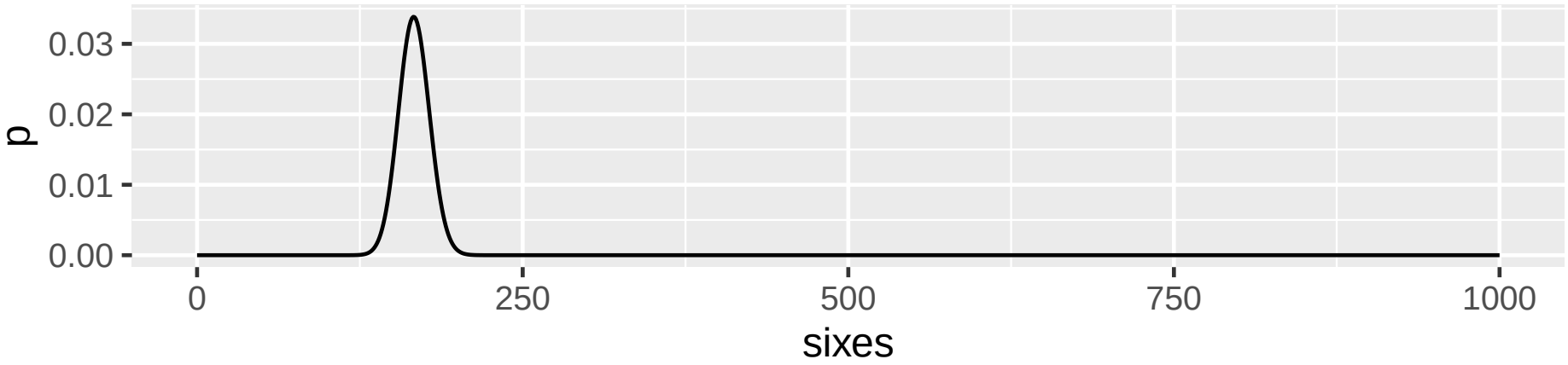
- What should x and y be?
- What is the power of this test?

Step 1: When do you reject?

For this we need to figure a rule for rejection. This is based on identifying events that should be unlikely under the hypothesis.

Here is how many 6's I would expect if the dice is fair:

```
fabricate(N = 1001, sixes = 0:1000, p = dbinom(sixes, 1000, 1/
  ggplot(aes(sixes, p)) + geom_line()
```



Step 1: When do you reject?

I can figure out from this that 143 or fewer is really very few and 190 or more is really very many:

```
c(lower = pbinom(143, 1000, 1/6), upper = 1 - pbinom(189, 100
```

```
     lower      upper
0.02302647 0.02785689
```

Step 2: What is the power?

- Now we need to stipulate some belief about how the world really works—this is not the null hypothesis that we plan to reject, but something that we actually take to be true.
- For instance: we think that *in fact* sixes appear 20% of the time.

Now what's the probability of seeing at least 190 sixes?

```
1 - pbinom(189, 1000, .2)
```

```
[1] 0.796066
```

So given I think 6s appear 20% of the time, I think it likely I'll see at least 190 sixes and reject the hypothesis of a fair dice.

Rule of thumb

- 80% or 90% is a common rule of thumb for "sufficient" power
- but really, how much power you need depends on the purpose

Think about

- Are there other tests I could have implemented?
- Are there other ways to improve this test?

Last subtleties

- Is a significant result from an underpowered study less credible? (only if there is a significance filter)
- What significance level should you choose for power? (Obviously the stricter the level the lower the power, so use what you will use when you actually implement tests)
- Do you really have to know the effect size to do power analysis? (No, but you should know at least what effects sizes you would want to be sure about picking up if they were present)
- Power is just one of many possible diagnosands
- What's power for Bayesians?

Power via design diagnosis

```
N <- 100
b <- .5

design <-
  declare_model(N = N,
    U = rnorm(N),
    potential_outcomes(Y ~ b * Z + U)) +
  declare_assignment(Z = simple_ra(N),
                     Y = reveal_outcomes(Y ~ Z)) +
  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_estimator(Y ~ Z, inquiry = "ate", .method = lm_robus
```

"Run" the design once

```
run_design(design)
```

Table 25: Summary of a single 'run' of the design

inquiry	estimand	estimator	term	estimate	std.error	statistic	p.v
ate	0.5	estimator	Z	0.42	0.17	2.45	

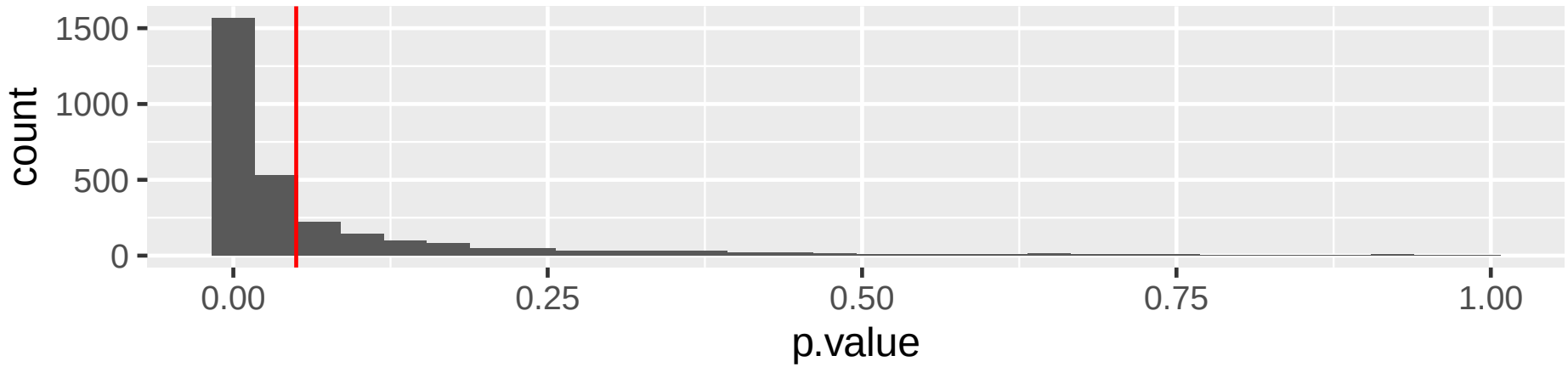
Run it many times

```
sims_1 <- simulate_design(design)

sims_1 |> select(sim_ID, estimate, p.value)
```

sim_ID	estimate	p.value
1	0.81	0.00
2	0.40	0.04
3	0.88	0.00
4	0.72	0.00
5	0.38	0.05
6	0.44	0.02

Power is mass of the sampling distribution of decisions under the model

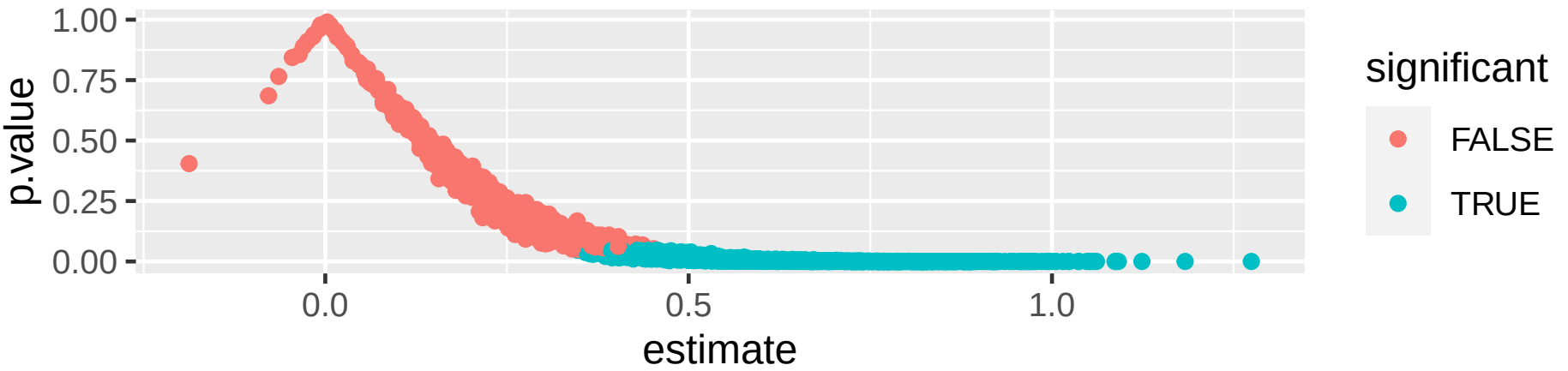
```
sims_1 |>
  ggplot(aes(p.value)) +
  geom_histogram() +
  geom_vline(xintercept = .05, color = "red")
```



Power is mass of the sampling distribution of decisions under the model

Obviously related to the estimates you might get

```
sims_1 |>
  mutate(significant = p.value <= .05) |>
  ggplot(aes(estimate, p.value, color = significant)) +
  geom_point()
```



Check coverage is correct

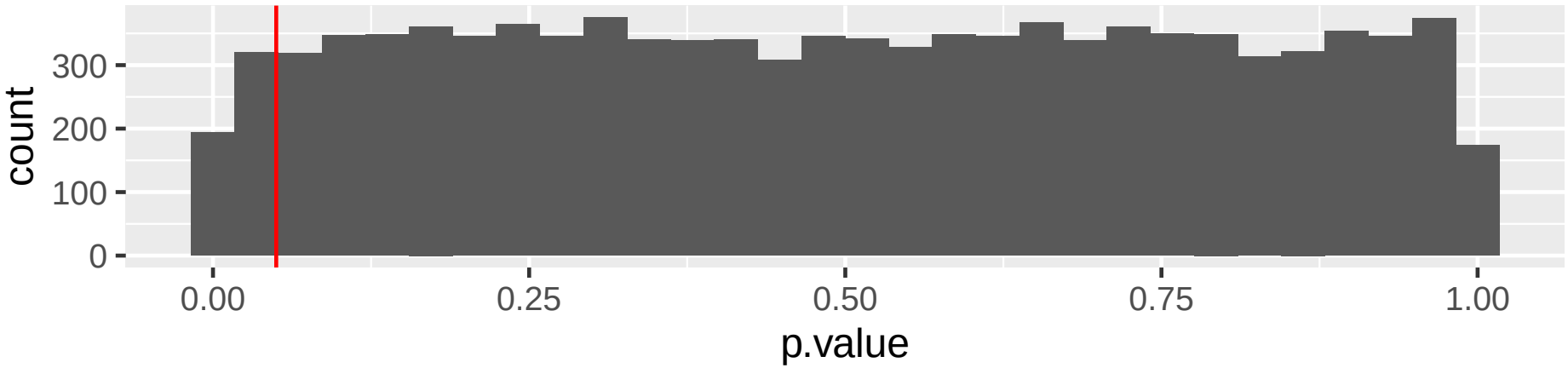
```
sims_1 |>
  mutate(within = (b > sims_1$conf.low) & (b < sims_1$conf.hig
  pull(within) |> mean()
```

```
[1] 0.9573333
```

Check validity of p value

A valid p-value satisfies $\Pr(p \leq x) \leq x$ for every $x \in [0, 1]$ (under the null)

```
sims_2 <-

  redesign(design, b = 0) |>

  simulate_design()
```



Design diagnosis does it all (over multiple designs)

```
diagnose_design(design)
```

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0.50	0.00	0.20	0.20	0.70	0.95
(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Design diagnosis does it all

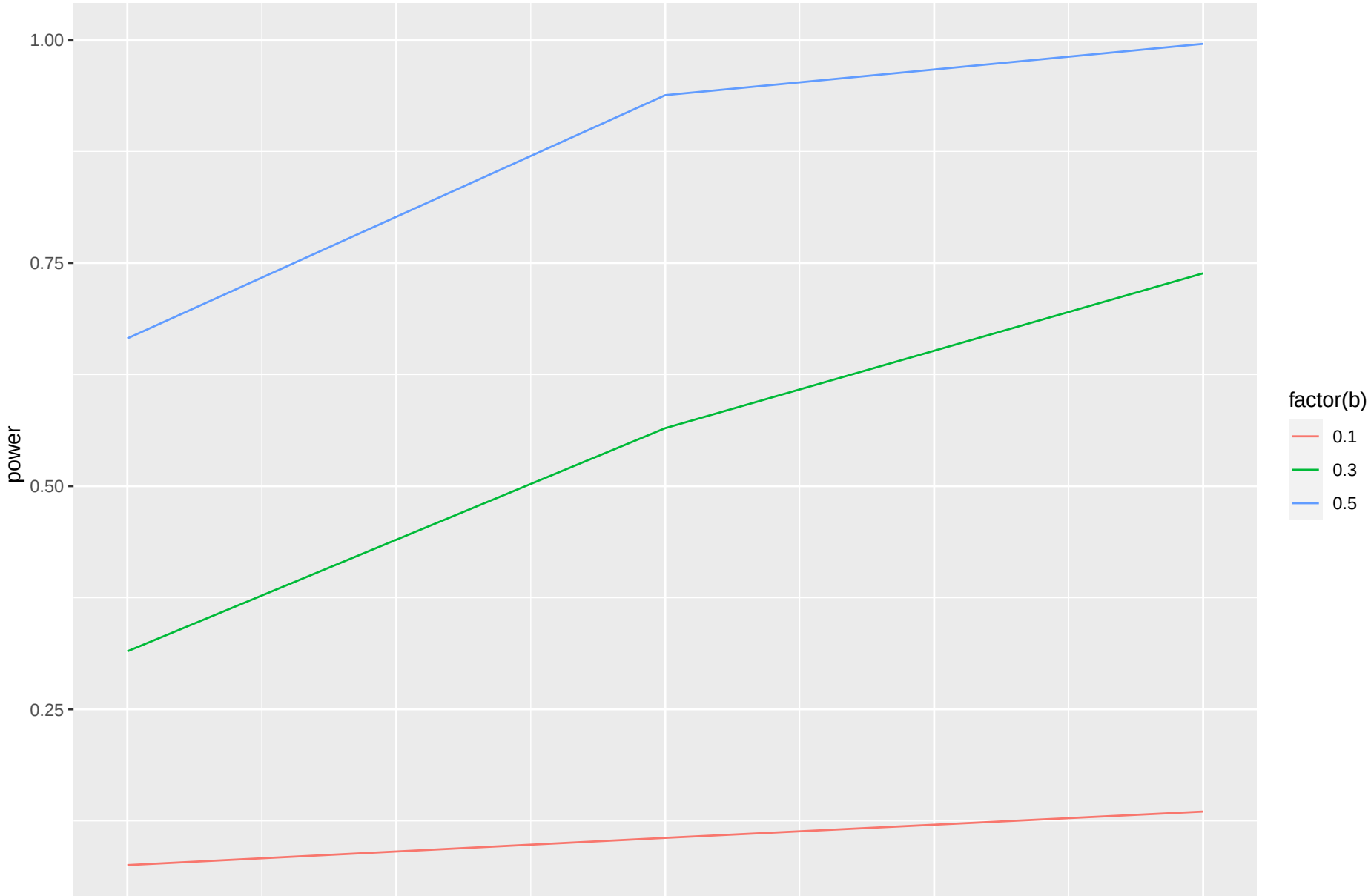
```
design |>
  redesign(b = c(0, 0.25, 0.5, 1)) |>
  diagnose_design()
```

b	Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0	-0.00	-0.00	0.20	0.20	0.05	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.25	0.25	-0.00	0.20	0.20	0.23	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.5	0.50	0.00	0.20	0.20	0.70	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
1	1.00	0.00	0.20	0.20	1.00	0.95
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Diagnose over multiple moving parts (and ggplot)

```
design |>
  ## Redesign
  redesign(b = c(0.1, 0.3, 0.5), N = 100, 200, 300) |>
  ## Diagnosis
  diagnose_design() |>
  ## Prep
  tidy() |>
  filter(diagnosand == "power") |>
  ## Plot
  ggplot(aes(N, estimate, color = factor(b))) +
  geom_line()
```

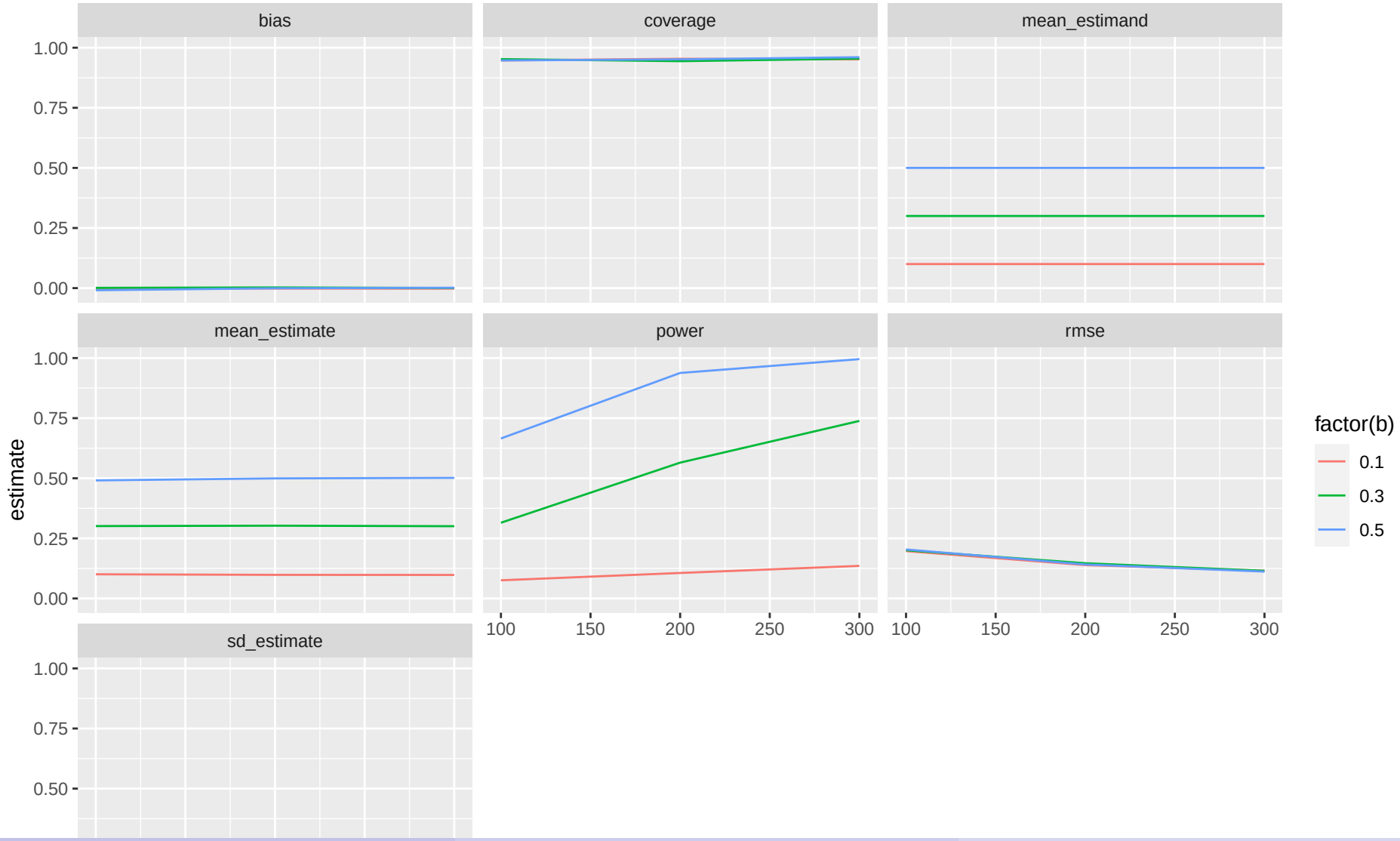
Diagnose over multiple moving parts (and ggplot)



Diagnose over multiple moving parts and multiple diagnosands (and ggplot)

```
design |>

  ## Redesign
  redesign(b = c(0.1, 0.3, 0.5), N = 100, 200, 300) |>

  ## Diagnosis
  diagnose_design() |>

  ## Prep
  tidy() |>

  ## Plot
  ggplot(aes(N, estimate, color = factor(b))) +
  geom_line()+
  facet_wrap(~diagnosand)
```

Diagnose over multiple moving parts and multiple diagnosands (and ggplot)



Subsection 5

Beyond basics

Power tips

coming up:

- power everywhere
- power with bias
- power with the wrong standard errors
- power with uncertainty over effect sizes
- power and multiple comparisons

Power depends on all parts of MIDA

We often focus on sample sizes

But

Power also depends on

- the model – obviously signal to noise
- the assignments and specifics of sampling strategies
- estimation procedures

Power when estimates are biased

```
bad_design <-

  declare_model(N = 100,
    U = rnorm(N),
    potential_outcomes(Y ~ 0 * X + U, conditions = list(X = 0
    X = ifelse(U > 0, 1, 0)) +

  declare_measurement(Y = reveal_outcomes(Y ~ X)) +

  declare_inquiry(ate = mean(Y_X_1 - Y_X_0)) +

  declare_estimator(Y ~ X, inquiry = "ate", .method = lm_robus
```

Power when estimates are biased

You can see from the null design that power is great but bias is terrible and coverage is way off.

```
diagnose_design(bad_design)
```

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
1.59	1.59	0.12	1.60	1.00	0.00
(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)

Power without unbiasedness corrupts, absolutely

Power with a more subtly biased experimental design

```
another_bad_design <-

  declare_model(
    N = 100,
    female = rep(0:1, N/2),
    U = rnorm(N),
    potential_outcomes(Y ~ female * Z + U)) +

  declare_assignment(
    Z = block_ra(blocks = female, block_prob = c(.1, .5)),
    Y = reveal_outcomes(Y ~ Z)) +

  declare_inquiry(ate = mean(Y_Z_1 - Y_Z_0)) +

  declare_estimator(Y ~ Z + female, inquiry = "ate",
                    .method = lm_robust)
```

Power with a more subtly biased experimental design

You can see from the null design that power is great but bias is terrible and coverage is way off.

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0.76	0.26	0.24	0.35	0.84	0.85
(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)

Power with the wrong standard errors

```
clustered_design <-
  declare_model(
    cluster = add_level(N = 10, cluster_shock = rnorm(N)),
    individual = add_level(
        N = 100,
        Y_Z_0 = rnorm(N) + cluster_shock,
        Y_Z_1 = rnorm(N) + cluster_shock)) +
  declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0)) +
  declare_assignment(Z = cluster_ra(clusters = cluster)) +
  declare_measurement(Y = reveal_outcomes(Y ~ Z)) +
  declare_estimator(Y ~ Z, inquiry = "ATE")
```

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
-0.00	-0.00	0.64	0.64	0.79	0.20
(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

What alerts you to a problem?

Let's fix that one

```
clustered_design_2  <-
  clustered_design |> replace_step(5,
  declare_estimator(Y ~ Z, clusters = cluster))
```

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0.00	-0.00	0.66	0.65	0.06	0.94
(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)

Power when you are not sure about effect sizes (always!)

- you can do power analysis for multiple stipulations
- or you can design with a distribution of effect sizes

```
design_uncertain <-
  declare_model(N = 1000, b = 1+rnorm(1), Y_Z_1 = rnorm(N), Y_
  declare_assignment(Z = complete_ra(N = N, num_arms = 3, conc
  declare_measurement(Y = reveal_outcomes(Y ~ Z)) +
  declare_inquiry(ate = mean(b)) +
  declare_estimator(Y ~ factor(Z), term = TRUE)

draw_estimands(design_uncertain)
```

```
  inquiry estimand
1     ate 1.523312
```

```
draw_estimands(design_uncertain)
```

```
  inquiry  estimand
```

Multiple comparisons correction (complex code)

Say I run two tests and want to correct for multiple comparisons.

Two approaches. First, by hand:

```
b = .2

design_mc <-
  declare_model(N = 1000, Y_Z_1 = rnorm(N), Y_Z_2 = rnorm(N) +
  declare_assignment(Z = complete_ra(N = N, num_arms = 3, conc
  declare_measurement(Y = reveal_outcomes(Y ~ Z)) +
  declare_inquiry(ate = b) +
  declare_estimator(Y ~ factor(Z), term = TRUE)
```

Multiple comparisons correction (complex code)

```
design_mc |>
  simulate_designs(sims = 1000) |>
  filter(term != "(Intercept)") |>
  group_by(sim_ID) |>
  mutate(p_bonferroni = p.adjust(p = p.value, method = "bonfer
         p_holm = p.adjust(p = p.value, method = "holm"),
         p_fdr = p.adjust(p = p.value, method = "fdr")) |>
  ungroup() |>
  summarize(
    "Power using naive p-values" = mean(p.value <= 0.05),
    "Power using Bonferroni correction" = mean(p_bonferroni <=
    "Power using Holm correction" = mean(p_holm <= 0.05),
    "Power using FDR correction" = mean(p_fdr <= 0.05)
    )
```

Power using naive p-values	Power using Bonferroni correction	Power usi
0.7374	0.6318	

Multiple comparisons correction (approach 2)

The alternative approach (generally better!) is to design with a custom estimator that includes your corrections.

```
my_estimator <- function(data)
  lm_robust(Y ~ factor(Z), data = data) |>
  tidy() |>
  filter(term != "(Intercept)") |>
  mutate(p.naive = p.value,
         p.value = p.adjust(p = p.naive, method = "bonferroni"



design_mc_2 <- design_mc |>
  replace_step(5, declare_estimator(handler = label_estimator(

run_design(design_mc_2) |>
  select(term, estimate, p.value, p.naive) |> kable()
```

Multiple comparisons correction (Null model case)

Lets try same thing for a null model (using `redesign(design_mc_2, b = 0)`)

```
design_mc_3 <-
  design_mc_2 |>
  redesign(b = 0)

run_design(design_mc_3) |> select(estimate, p.value, p.naive)
```

estimate	p.value	p.naive
-0.0363484	1	0.6297813
-0.0020170	1	0.9787214

Multiple comparisons correction (Null model case)

...and power:

Mean Estimate	Bias	SD Estimate	RMSE	Power	Coverage
0.00	0.00	0.08	0.08	0.02	0.95
(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
-0.00	-0.00	0.08	0.08	0.02	0.96
(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)

bothered?

You might try

- Power for an interaction (in a factorial design)
- Power for a binary variable (versus a continuous variable?)
- Power from adding covariates in the analysis stage
- Power gains from blocked randomization
- Power losses from clustering at different levels
- Controlling the ICC directly? (see book cluster designs section)

Big takeaways

- Power is affected not just by sample size, variability and effect size but also by you data and analysis strategies.
- Try to estimate power under multiple scenarios
- Try to use the same code for calculating power as you will use in your ultimate analysis
- Basically the same procedure can be used for any design. If you can declare a design and have a test, you can calculate power
- Your power might be right but misleading. For confidence:
  - Don't just check power, check bias and coverage also
  - Check power especially *under the null*
- Don't let a focus on power distract you from more *substantive* diagnosands

Section 9

Topics

Topics

Top

Subsection 1

Noncompliance and the LATE estimand

LATE—Local Average Treatment Effects

Sometimes you give a medicine but only a non random sample of people actually try to use it. Can you still estimate the medicine's effect?

	$X=0$	$X=1$
$T=0$	$\overline{y}_{00}$ (n_{00})	$\overline{y}_{01}$ (n_{01})
$T=1$	$\overline{y}_{10}$ (n_{10})	$\overline{y}_{11}$ (n_{11})

Say that people are one of 3 types:

1. n_a "always takers'' have $X=1$ no matter what and have average outcome $\overline{y}_a$
2. n_n never takers have $X=0$ no matter what with outcome $\overline{y}_n$
3. n_c compliers have $X=T$ and average outcomes $\overline{y}_c^1$ if treated and $\overline{y}_c^0$ if not.

LATE—Local Average Treatment Effects

Sometimes you give a medicine but only a non random sample of people actually try to use it. Can you still estimate the medicine's effect?

	$X=0$	$X=1$
$T=0$	$\overline{y}_{00}$ (n_{00})	$\overline{y}_{01}$ (n_{01})
$T=1$	$\overline{y}_{10}$ (n_{10})	$\overline{y}_{11}$ (n_{11})

We can figure something about types:

	$X=0$	$X=1$
$T=0$	$\frac{\frac{1}{2}n_c}{\frac{1}{2}n_c+\frac{1}{2}n_n}\overline{y}_c^0 + \frac{\frac{1}{2}n_n}{\frac{1}{2}n_c+\frac{1}{2}n_n}\overline{y}_n$	$\overline{y}_a$
$T=1$	$\overline{y}_n$	$\frac{\frac{1}{2}n_c}{\frac{1}{2}n_c+\frac{1}{2}n_a}\overline{y}_c^1 + \frac{\frac{1}{2}n_a}{\frac{1}{2}n_c+\frac{1}{2}n_a}\overline{y}_a$

LATE—Local Average Treatment Effects

You give a medicine to 50% but only a non random sample of people actually try to use it. Can you still estimate the medicine's effect?

	$X=0$	$X=1$
$T=0$	$\frac{n_c}{n_c+n_n}\overline{y}_c^0 + \frac{n_n}{n_c+n_n}\overline{y}_n$	$\overline{y}_a$
(n)	$(\frac{1}{2}(n_c+n_n))$	$(\frac{1}{2}n_a)$
$T=1$	$\overline{y}_n$	$\frac{n_c}{n_c+n_a}\overline{y}_c^1 + \frac{n_a}{n_c+n_a}\overline{y}_a$
(n)	$(\frac{1}{2}n_n)$	$(\frac{1}{2}(n_a+n_c))$

Average in $T=0$ group: $\frac{n_c\overline{y}_c^0+(n_n\overline{y}_n+n_a\overline{y}_a)}{n_a+n_c+n_n}$

Average in $T=1$ group: $\frac{n_c\overline{y}_c^1+(n_n\overline{y}_n+n_a\overline{y}_a)}{n_a+n_c+n_n}$

Difference: $ITT = (\overline{y}_c^1 - \overline{y}_c^0)\frac{n_c}{n}$

So: $LATE = ITT \times \frac{n}{n_c}$

The good and the bad of LATE

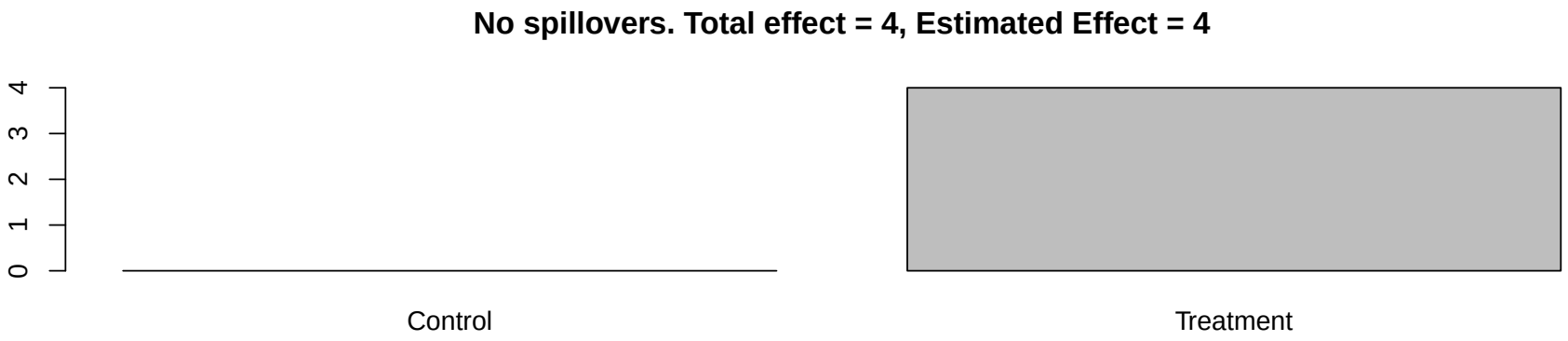
- You get a well-defined estimate even when there is non-random take-up
- May sometimes be used to assess mediation or knock-on effects
- But:
  - You need assumptions (monotonicity and the exclusion restriction – *where were these used above*?)
  - Your estimate is only for a subpopulation
  - The subpopulation is not chosen by you and is unknown
  - Different encouragements may yield different estimates since they may encourage different subgroups

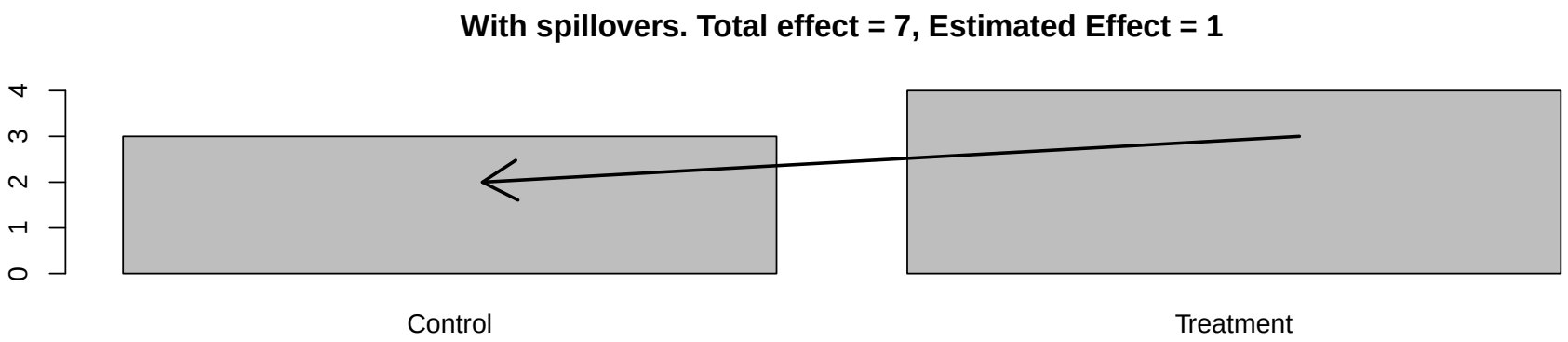
Subsection 2

Spillovers

SUTVA violations (Spillovers)

Spillovers can result in the estimation of weaker effects when effects are actually stronger.





The key problem is that $Y(1)$ and $Y(0)$ are not sufficient to describe potential outcomes

SUTVA violations

More completely specified potential outcomes (and estimands)

		0		1		2		3		4	
Unit	Location	$D_\emptyset$	$y(D_\emptyset)$	D_1	$y(D_1)$	D_2	$y(D_2)$	D_3	$y(D_3)$	D_4	$y(D_4)$
A	1	0	0	1	3	0	1	0	0	0	0
B	2	0	0	0	3	1	3	0	3	0	0
C	3	0	0	0	0	0	3	1	3	0	3
D	4	0	0	0	0	0	0	0	1	1	3
$\bar{y}_{\text{treated}}$			-		3		3		3		3
$\bar{y}_{\text{untreated}}$			0		1		4/3		4/3		1
$\bar{y}_{\text{neighbors}}$			-		3		2		2		3
$\bar{y}_{\text{pure control}}$			0		0		0		0		0
ATT (direct effect)			-		3		3		3		3
ATT (indirect effect)			-		3		2		2		3

Table 26: Potential outcomes for four units for different treatment profiles, D_1-D_4. D_i represents an allocation to treatment and $y_j(D_i)$ is the potential outcome for (row) unit j given (column) allocation i.

SUTVA violations

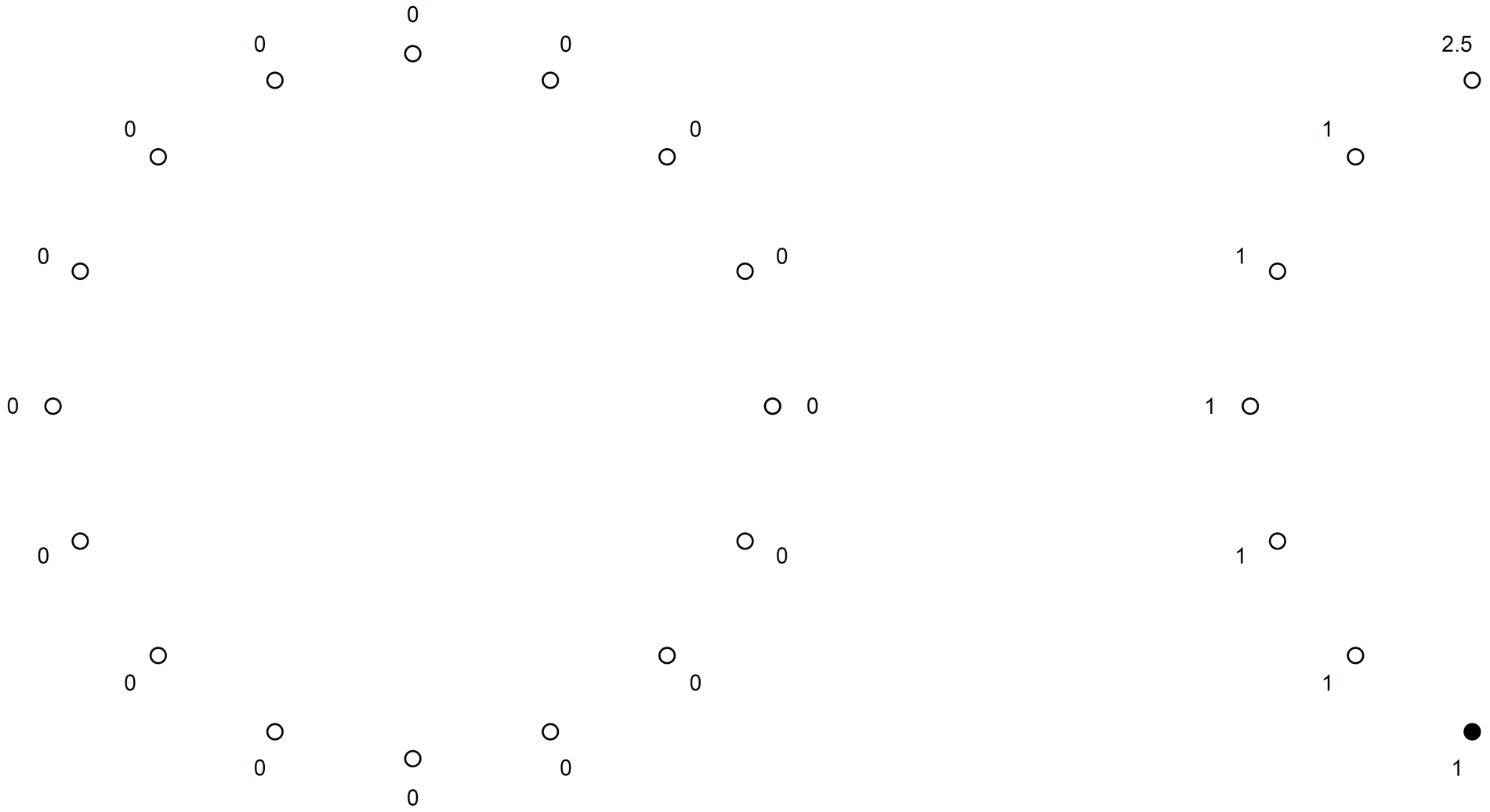
		0		1		2		3		4	
Unit	Location	$D_\emptyset$	$y(D_\emptyset)$	D_1	$y(D_1)$	D_2	$y(D_2)$	D_3	$y(D_3)$	D_4	$y(D_4)$
A	1	0	0	1	3	0	1	0	0	0	0
B	2	0	0	0	3	1	3	0	3	0	0
C	3	0	0	0	0	0	3	1	3	0	3
D	4	0	0	0	0	0	0	0	1	1	3

Table 27: Potential outcomes for four units for different treatment profiles, D_1-D_4. D_i represents an allocation to treatment and $y_j(D_i)$ is the potential outcome for (row) unit j given (column) allocation i.

- The key is to think through the structure of spillovers.
- Here immediate neighbors are exposed
- In this case we can **define a direct treatment** (being exposed) and **an indirect treatment** (having a neighbor exposed) and we can work out *the **propensity** for each unit of receiving each type of treatment*
- These may be non uniform (here central types are more likely to have teated neighbors); but we can still use the randomization to assess effects

SUTVA violations

Even still, to estimate effects you need some SUTVA like assumption.

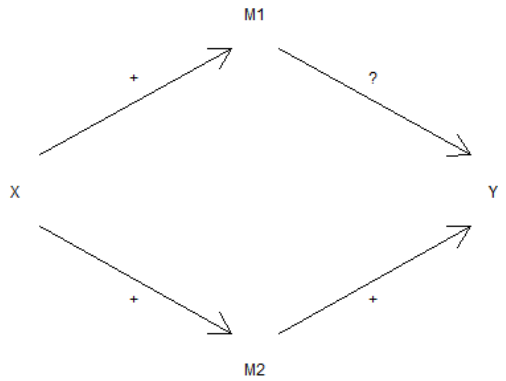


Subsection 3

Mediation

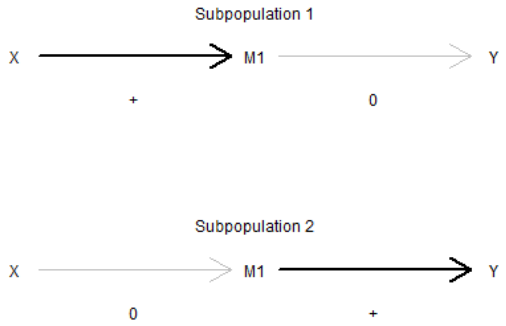
The problem of unidentified mediators

- Consider a causal system like the below.
- The effect of X on M1 and M2 can be measured in the usual way.
- But unfortunately, if there are multiple mediators, the effect of M1 (or M2) on Y is not identified.
- The 'exclusion restriction' is obviously violated when there are multiple mediators (unless you can account for them all).



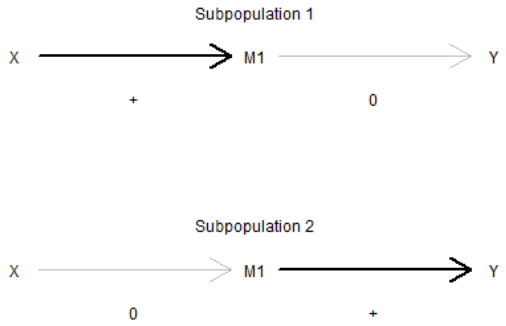
The problem of unidentified mediators

* An obvious approach is to first examine the (average) effect of X on M1 and then use another manipulation to examine the (average) effect of M1 on Y. * But **both of these average effects may be positive (for example) even if there is no effect of X on Y through M1**.
\end{itemize}



The problem of unidentified mediators

* An obvious approach is to first examine the (average) effect of X on M1 and then use another manipulation to examine the (average) effect of M1 on Y. * Similarly **both of these average effects may be zero even if X affects on Y through M1 for every unit!**.
\end{itemize}



The problem of unidentified mediators

* Another somewhat obvious approach is see how the effect of X on Y in a regression is reduced when you control for M. If the effect of X on Y passes through M then surely there should be no effect of X on Y after you control for M. * But this common strategy is also not guaranteed to produce reliable results
* See Imai on better ways to think about this problem and designs to address it
\end{itemize}

The problem of unidentified mediators: Quantities

- In the potential outcomes framework we can describe a **mediation effect** as (see Imai et al):

$$\delta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \text{ for } t = 0, 1$$

- The **direct effect** is:

$$\psi_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \text{ for } t = 0, 1$$

- This is a **decomposition**, since:

$$Y_i(1, M_i(1)) - Y_1(0, M_i(0)) = \frac{1}{2}(\delta_i(1) + \delta_i(0) + \psi_i(1) + \psi_i(0))$$

- If (and a big if), there are no interaction effects—ie $\delta_i(1) = \delta_i(0), \psi_i(1) = \psi_i(0)$, then

$$Y_i(1, M_i(1)) - Y_1(0, M_i(0)) = \delta_i + \psi_i$$

- The bad news is that although a single experiment might identify the total effect, it can not identify these elements of the direct effect.

The problem of unidentified mediators: Solutions?

- Check **formal requirement** for identification under single experiment design ("sequential ignorability"—that, conditional on actual treatment, it is as if the value of the mediation variable is randomly assigned relative to potential outcomes). But this is strong (and in fact unverifiable) and if it does not hold, bounds on effects always include zero (Imai et al)
- You can use **interactions** with covariates **if you are willing to make assumptions on no heterogeneity of direct treatment effects** over covariates. eg you think that money makes people get to work faster because they can buy better cars; you look at the marginal effect of more money on time to work for people with and without cars and find it higher for the latter. This might imply mediation through transport but only if there is no direct effect heterogeneity (eg people with cars are less motivated by money).

The problem of unidentified mediators: Solutions?

- Weaker assumptions justify **parallel design**
  - Group A: T is randomly assigned, M left free.
  - Group B: divided into four groups $T \times M$ (requires two more assumptions (1) that the **manipulation** of the mediator only affects outcomes through the mediator (2) **no interaction**, for each unit, $Y(1, m) - Y(0, m) = Y(1, m') - Y(0, m')$.)

Idea 5: Understanding mechanisms is harder than you think. Figure out what assumptions fly.

Subsection 4

Differences in differences

Differences in differences

New challenges, new developments

Subsection 5

Regression discontintuity

Regression discontintuity

Errors and diagnostics

Section 10

Experimentation: Processes and Workflows

Experimentation: Processes and Workflows

- Scope for experimentation
- Ethics of experiments
- Open science workflows

Subsection 1

When to experiment

Prospects

- Whenever someone is uncertain about *something* they are doing (all the time)
- Whenever someone hits scarcity constraints
- When people have incentives to demonstrate that they are doing the right thing (careful...)

Prospects

- **Advice**: If you can, **start from theory** and find an intervention, rather than the other way around.
- **Advice**: If you can, go for *structure* rather than *gimmicks*
- **Advice**: In attempts to parse, beware of generating unnatural interventions (how should a voter think of a politician that describes his policy towards Korea in detail but does not mention the economy? Is not mentioning the economy sending an unintended message?)

Prospects & Potential

- Randomization of where police are stationed (India)
- Randomization of how government tax collectors get paid (do they get a share?) (Pakistan)
- Randomization of the voting rules for determining how decisions get made (Afghanistan)
- Random assignment of populations to peacekeepers (Liberia)
- Random assignment of ex-combatants out of their networks (Indonesia)
- Randomization of students to ethnically homogeneous or ethnically diverse schools (anywhere?)

Subsection 2

Ethics

Constraint: Is it ethical to manipulate subjects for research purposes?

- There is no foundationless answer to this question. So let's take some foundations from the Belmont report and seek to ensure:
  1. Respect for persons
  2. Beneficence
  3. Justice
- Unfortunately, operationalizing these requires further ethical theories. Let's assume that (1) is operationalized by informed consent (a very liberal idea). We are a bit at sea for (2) and (3) (the Belmont report suggests something like a utilitarian solution).
- The major focus on (1) by IRBs might follow from the view that if subjects consent, then they endorse the ethical calculations made for 2 and 3 — *they* think that it is good and fair.
- This is a little tricky, though, since the study may not be good or fair

Is it ethical to manipulate subjects for research purposes?

- The problem is that many (many) field experiments have nothing like informed consent.
- For example, whether the government builds a school in your village, whether an ad appears on your favorite radio show, and so on.
- Consider three cases:
  1. You work with a nonprofit to post (true?) posters about the crimes of politicians on billboards to see effects on **voters**
  2. You hire confederates to offer bribes to **police officers** to see if they are more likely to bend the law for coethnics
  3. The British government asks you to work on figuring out how the use of water cannons helps stop **rioters** rioting

Is it ethical to manipulate subjects for research purposes?

- Consider three cases:
  - You work with a nonprofit to post (true?) posters about the crimes of politicians on billboards to see effects on **voters**
  - You hire confederates to offer bribes to **police officers** to see if they are more likely to bend the law for coethnics
  - The British government asks you to work on figuring out how the use of water cannons helps stop **rioters** rioting
- In all cases, there is **no consent** given by subjects.
- In 2 and 3, the treatment is **possibly harmful** for subjects, and the results might also be harmful. But even in case 1, there could be major unintended harmful consequences.
- In cases 1 and 3, however, the "intervention" is within the sphere of **normal activities** for the implementer.

Constraint: Is it ethical to manipulate subjects for research purposes?

- Sometimes it is possible to use this point of difference to make a "spheres of ethics" argument for "embedded experimentation."
- **Spheres of Ethics Argument**: Experimental research that involves manipulations that are not normally appropriate for researchers may nevertheless be ethical if:
  - Researchers and implementers agree on a **division of responsibility** where implementers take on responsibility for actions
  - Implementers have **legitimacy** to make these decisions within the sphere of the intervention
  - Implementers are indeed **materially independent** of researchers (no swapping hats)
- Difficulty with this argument:
  - **Question begging**: How to determine the legitimacy of the implementer? (Can we rule out Nazi doctors?)

Subsection 3

Transparency & Experimentation

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Analytic replication.** This should be a no brainer. Set everything up so that replication is easy. Use rmarkdown, or knitr or sweave. Or produce your replication code as a package.

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Data.** How soon should you make your data available? **My view**: as soon as possibe. Along with working papers and before publication. Before it affects policy in any case. Own the ideas not the data.
    - Hard core: no citation without (analytic) replication. Perhaps. Non-replicable results should not be influencing policy.
- **Where should you make your data available?** Dataverse is focal for political science. Not personal website (mea culpa)
- **What data should you make available?** Disagreement is over how raw your data should be. **My view**: as raw as you can but at least post cleaning and pre-manipulation.

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Should you register?**: Hard to find reasons against. But case strongest in testing phase rather than exploratory phase.
- **Registration**: When should you register? **My view:** Before treatment assignment. (Not just before analysis, mea culpa)
- **Registration**: Should you deviate from an preanalysis plan if you change your mind about optimal estimation strategies. **My view:** Yes, but make the case and describe both sets of results.

Contentious Issues

Experimental researchers are deeply engaged in the movement towards more transparency social science research.

Contentious issues (mostly):

- **Registration**: When should you register? **My view:** Before treatment assignment. (Not just before analysis, mea culpa)
- **Registration**: Should you deviate from an preanalysis plan if you change your mind about optimal estimation strategies. **My view:** Yes, but make the case and describe both sets of results.

Subsection 4

Pre-registration rationales and structures

Two distinct rationales for registration

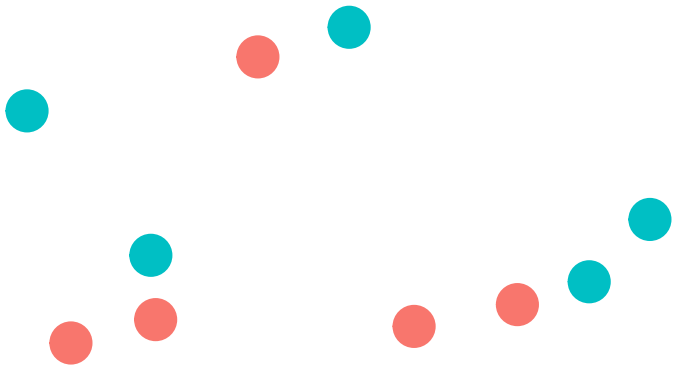
- File drawer bias (Publication bias)
- Analysis bias (Fishing)

File drawer bias

– Say in truth X affects Y in 50% of cases.

– Researchers conduct multiple excellent studies. But they only write up the 50% that produce "positive" results.

– Even if each individual study is indisputably correct, the account in the research record – that X affects Y in 100% of cases – will be wrong.

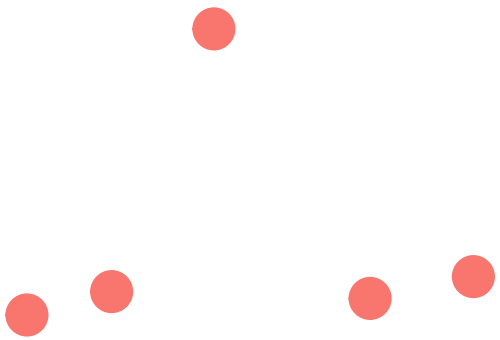


File drawer bias

– Say in truth X affects Y in 50% of cases.

– Researchers conduct multiple excellent studies. But they only write up the 50% that produce "positive" results.

– Even if each individual study is indisputably correct, the account in the research record – that X affects Y in 100% of cases – will be wrong.



File drawer bias

Exacerbated by:

– Publication bias – the positive results get published

– Citation bias – the positive results get read and cited

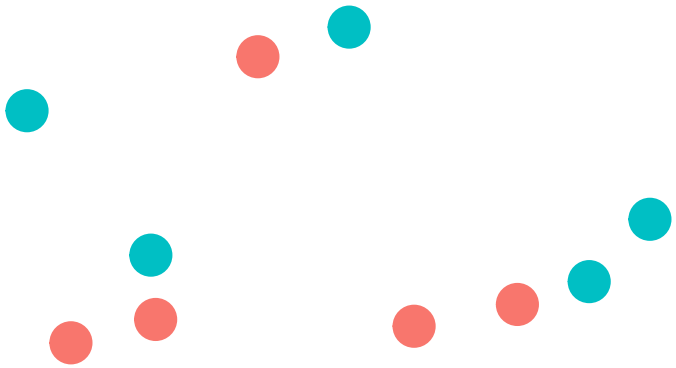
– Chatter bias – the positive results gets blogged, tweeted and TEDed.

Analysis bias (Fishing)

– Say in truth X affects Y in 50% of cases.

– But say that researchers enjoy discretion to select measures for X or Y, or enjoy discretion to select statistical models after seeing X and Y in each case.

– Then, with enough discretion, 100% of analyses may report positive effects, even if all studies get published.

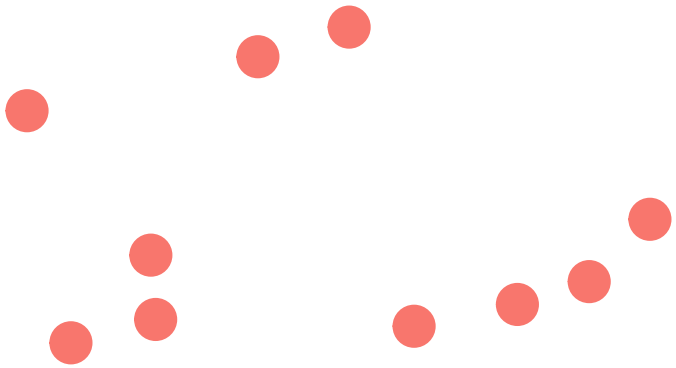


Analysis bias (Fishing)

– Say in truth X affects Y in 50% of cases.

– But say that researchers enjoy discretion to select measures for X or Y, or enjoy discretion to select statistical models after seeing X and Y in each case.

– Then, with enough discretion, 100% of analyses may report positive effects, even if all studies get published.



Analysis bias (Fishing)

– Try the exact fishy test An Exact Fishy Test (https://macartan.shinyapps.io/fish/)

– What's the problem with this test?

Evidence-Proofing: Illustration

- When your conclusions do not really depend on the data
- Eg – some evidence will always support your proposition – some interpretation of evidence will always support your proposition
- Knowing the mapping from data to inference in advance gives a handle on the false positive rate.

Evidence Proofing: Bayesian Illustration

- Say choice of two pieces of evidence to bring to bear, $K1$ or $K2$

Table 28: Likelihoods

(a)

(b) If TRUE

(c)

(d) If FALSE

	K_1 = No	K_1 = Yes	All	K_1 = No	K_1 = Yes
K_2 = No	0.9	0.05	0.95	0	
K_2 = Yes	0.05	0	0.05	0.05	
All	0.95	0.05	1	0.05	

- Posterior | K1 = Posterior | K2 = 95%
- Probability positive claim | H is false; evidence randomly selected (p) = 5%
- Probability positive claim | H is false; evidence is fished (p) = 10%

Evidence Proofing: Bayesian Illustration

- Say choice of two pieces of evidence to bring to bear, $K1$ or $K2$

Table 29: Likelihoods

(a)

(b) If TRUE

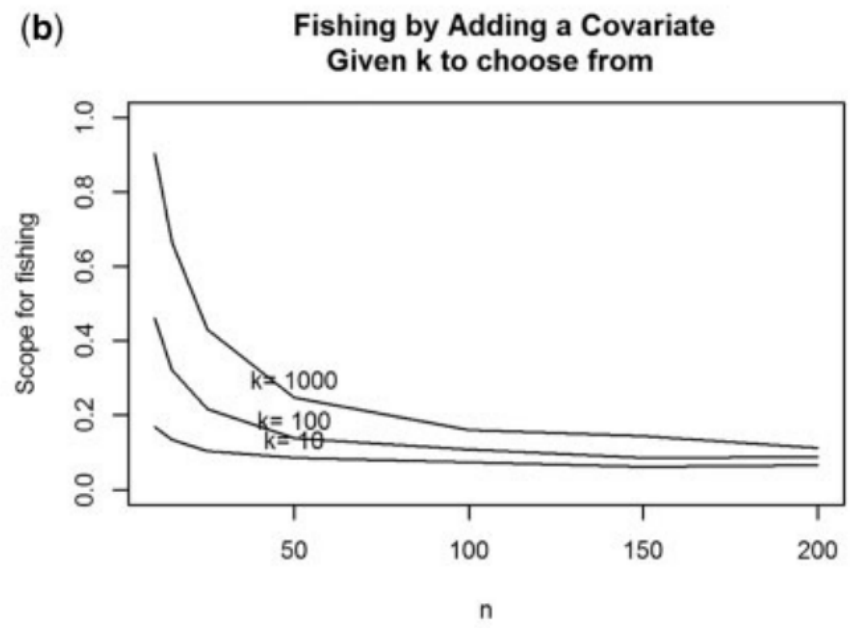
(c)

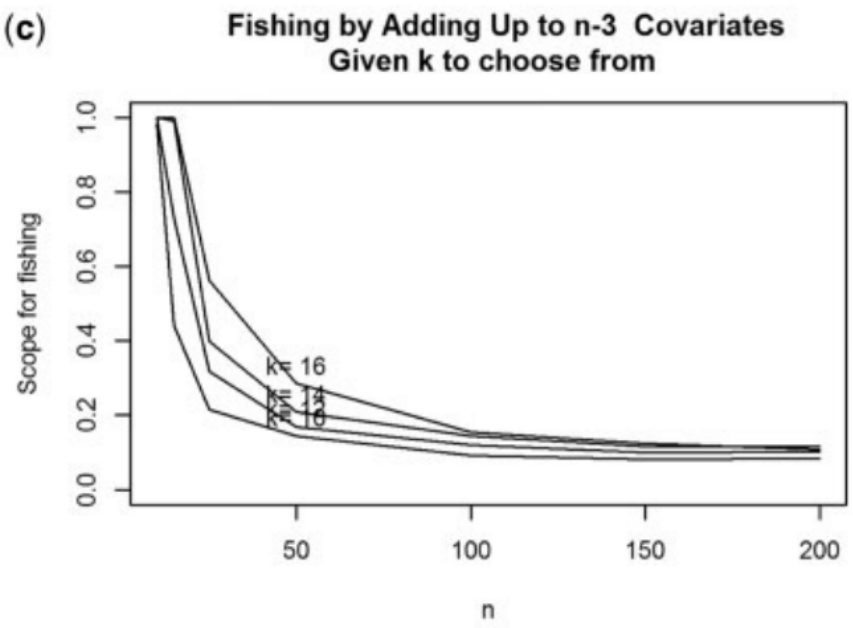
(d) If FALSE

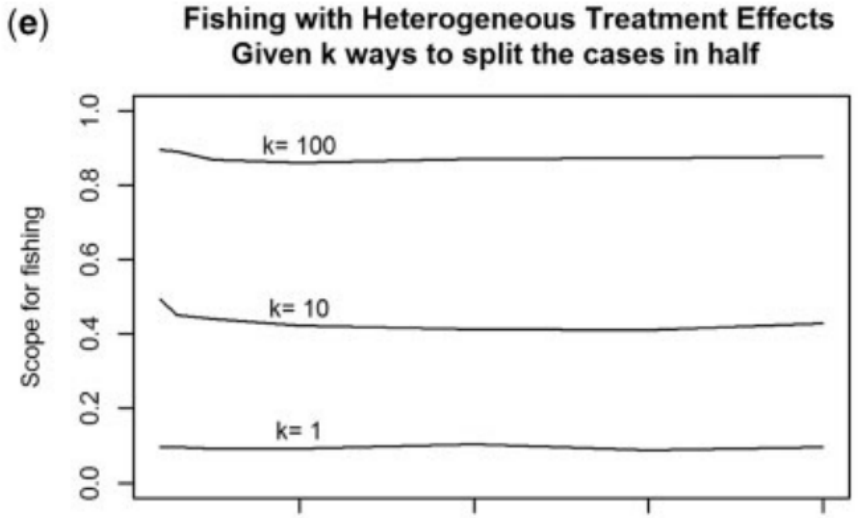
	K_1 = No	K_1 = Yes	All	K_1 = No	K_1 = Yes
K_2 = No	0.9	0.05	0.95	0	
K_2 = Yes	0.05	0	0.05	0.05	
All	0.95	0.05	1	0.05	

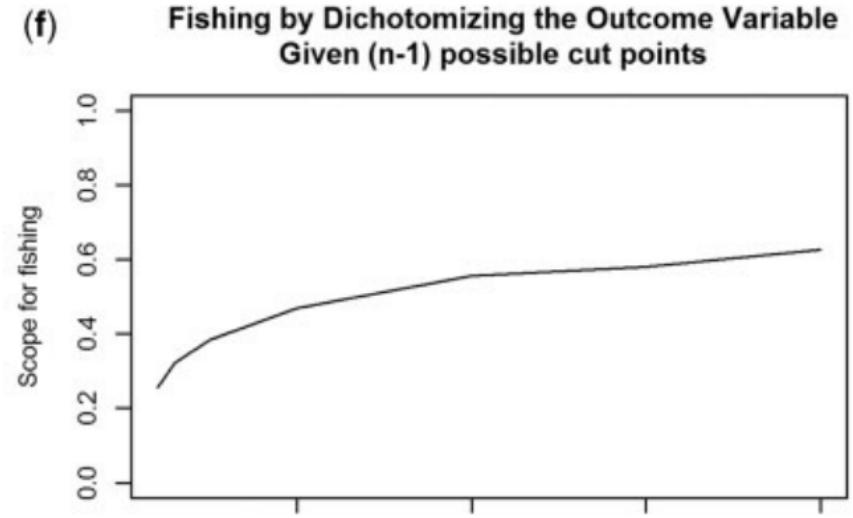
- What's the truly correct inference if you KNOW that researcher is a fisher?
- Depends: say you thought K1 and K2 were sought in order. Then if K2 evidence is presented this means K1 not found. So posterior |K2

The scope for fishing

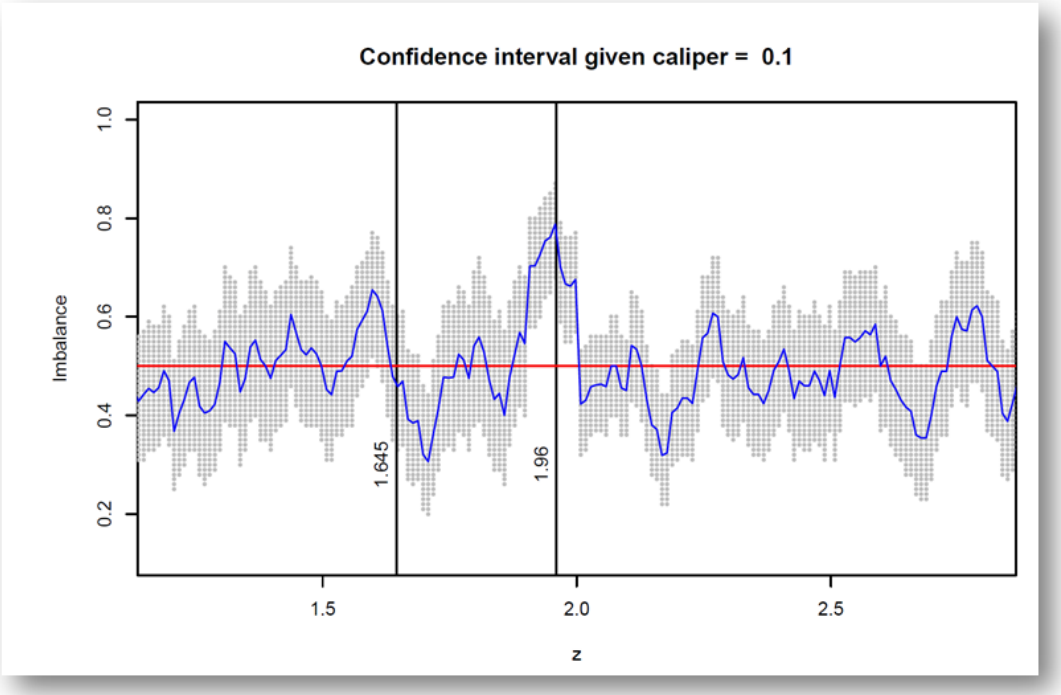








Evidence from political science



Source: Gerber and Malhotra

More evidence from TESS

- Malhotra tracked 221 TESS studies.
- 20% of the null studies were published. 65% not even written up (file drawer or anticipation of publication bias)
- 60% of studies with strong results were published.

Implications are:

- population of results not representative
- (subtler) individual published studies are also more likely to be overestimates

The problem

- Summary: we do not know when we can or cannot trust claims made by researchers.
- [Not a tradition specific claim]

Registration as a possible solution

Simple idea:

– It's about communication: – just say what you are planning on doing before you do it – if you don't have a plan, say that – If you do things differently from what you were planning to do, say that

Bells and whistles

– To be really useful a registry would have to have some credibility, some searchability, and some consistency in fields.

Registration as a possible solution

Elements:

- Make it a facility
- Non-mandatory
- Non-binding
- But comprehensive
- Report whether registered or not
- Report changes in plans

What's the Right Scope:

For discussion: but claims of "tests" seem like a good start

Bells and Whistles: Certificaton?

Center for Open Science Badges

- A public date-time stamped registration is in an institutional registration system (e.g., ClinicalTrials.gov, Open Science Framework)
- Registration pre-dates realization of the outcomes
- Registered design and analysis plan corresponds directly to reported design and analysis
- Full disclosure of results following the registered plan

Notations PR (peer review certified), DE (data exist), and TC (transparent changes)

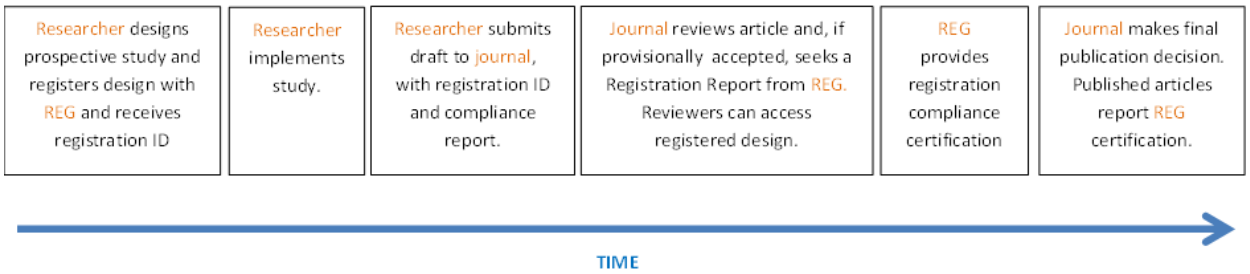
- https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/

Bells and Whistles: Certificaton?

Center for Open Science Badges

- Requires validating body (eg journal or registry)
- Validation on process not on quality
- Emphasis on transparency

Possible Cycle



Possible Models

- Journal-led model: let the registry follow

– Hard form – Medical sciences for RCTs – Soft form – Medical sciences for observational studies

- Professional Association led model

– AEA for RCTs – APSA?

- Funder led model (more mandatory) – RIDIE, NSF?
- Bottom up model? – Eg established by APSA sections; CQRM, PolMeth, Experiments? No formal journal recognition.

But even the simple idea is not everywhere welcome. There are many worries and some myths.

Subsection 5

Worries and Myths around registration

Myth: Concerns about fishing presuppose researcher dishonesty

- Fishing can happen in very subtle ways, and may seem natural and justifiable.
- Example:

– I am interested in whether more democratic institutions result in better educational outcomes. – I examine the relationship between institutions and literacy and between institutions and school attendance. – The attendance measure is significant and the literacy one is not. Puzzled, I look more carefully at the literacy measure and see various outliers and indications of measurement error. As I think more I realize too that literacy is a slow moving variable and may not be the best measure anyhow. I move forward and start to analyze the attendance measure only, perhaps conducting new tests, albeit with the same data.

Structural challenge

Our journal review process is largely organized around advising researchers how to adjust analysis in light of findings in the data.

Myth: Fishing is technique specific

- Frequentists can do it
- Bayesians can do it too.
- Qualitative researchers can also do it.
- You can even do it with descriptive statistics

Myth: Fishing is estimand specific

- You can do it when estimating causal effects
- You can do it when studying mechanisms
- You can do it when estimating counts

Myth: Registration only makes sense for experimental studies, not for observational studies

- The key distinction is between prospective and retrospective studies.
- Not between experimental and observational studies.
- A reason (from the medical literature) why registration is especially important for experiments: because you owe it to subjects
- A reason why registration is less important for experiments: because it is more likely that the intended analysis is implied by the design in an experimental study. Researcher degrees of freedom may be greatest for observational qualitative analyses.

Worry: Registration will create administrative burdens for researchers, reviewers, and journals

- Registration will produce some burden but does not require the creation of content that is not needed anyway

– It does shift preparation of analyses forward – And it also can increase the burden of developing analyses plans even for projects that don't work. But that is in part, the point.

- Upside is that ultimate analyses may be much easier.

Worry: Registration will force people to implement analyses that they know are wrong

- Most arguments for registration in social science advocate for non-binding registration, where deviations from designs are possible, though they should be described.
- Even if it does not prevent them, a merit of registration is that it makes deviations visible.

Myth: Replication (or other transparency practices) obviates the need for registration

- There are lots of good things to do, including replication.
- Many of these do not substitute for each other. (How to interpret a fished replication of a fished analysis?)
- And they may likely act as complements
- Registration can clarify details of design and analysis and ensure early preparation of material. Indeed material needed for replication may be available even before data collection

Worry: Registration will put researchers at risk of scooping

- But existing registries allow people to protect registered designs for some period
- Registration may let researchers lay claim to a design

Worry: Registration will kill creativity

- This is an empirical question. However, under a nonmandatory system researchers could:
- Register a plan for structured exploratory analysis
- Decide that exploration is at a sufficiently early stage that no substantive registration is possible and proceed without registration.

Implications:

- In neither case would the creation of a registration facility prevent exploration.
- What it might do is make it less credible for someone to claim that they have tested a proposition when in fact the proposition was developed using the data used to test it.
- Registration communicates when researchers are angage in exploration or not. We love exploration and should be proud of it.

The challenge of historical data

- Does registering analyses of historical data make sense?
- The problem is not just that researchers might have already seen the testing data; but that they have seen data that is correlated with it.

Historical data: Illustration

- Consider historical proposition H. – Say we start with a prior of .5 that H is true. – Say that if H is true then we observe K1 with probability 0.8 but if it is false we observe K1 with probability 0.2 ("double decisiveness") – Similarly if H is true then we observe K2 with probability 0.8 but if it is false we observe K2 with probability 0.2 ("double decisiveness" again)
- Say we observe K1 (some collection of facts)
- We then update our belief in H...

Historical data: Illustration

- Our updated belief is:

$$\Pr(H|K1) = \Pr(K1|H)\Pr(H)/\Pr(K1) = \frac{.8 * .5}{.8 * .5 + .2 * .5} = 80\%$$

- We are now 80% confident in proposition H.
- We decide to look for evidence K2. And we find it!
- Our posterior is now:

$$\Pr(H|K2) = \Pr(K2|H)\Pr(H)/\Pr(K2) = .8*.8/(.8*.8+.2*.2) = 94\%$$

- Or is it?

Historical data: Illustration

- Our updated belief is:

$$\Pr(H|K1) = \Pr(K1|H)\Pr(H)/\Pr(K1) = .8*.5/(.8*.5+.2*.5) = 80\%$$

- We are now 80% confident in proposition H.
- We decide to look for evidence K2. And we find it!
- Our posterior is now:

$$Pr(H|K2) = Pr(K2|H)Pr(H)/Pr(K2) = .8*.8/(.8*.8+.2*.2) = 94$$

- Or is it?

Historical data: Illustration

- What if there are correlated probabilities?
- Then

$$\Pr(H|K1\&K2) = .76 \times .5/(.76 \times .5 + .16 \times .5) = 83\%$$

Historical data: Illustration

- In a sense the fishing has already happened.
- How so?
- Say the proposition is FALSE but K1 is still observed
- A decision is then made to seek "new data" K2
- Now K2 will be observed with 80% probability even though H is false

Historical data: Illustration

- Naïve inference (using a prior of 80% due to K1): 94% if K2; 50% if not K2
- Inference if K1 used to decide on search for K2 but prior is "reset" to .5 80% if K2; 20% if not K2
- Sophisticated inference: 83% if K2; 50% if not K2
- This sophisticated inference is unchanged if you take explicit account of the fact that searching for K2 was conditional on K1; either way it is still $\Pr(H|K1, K2)$.
- *It requires assessing the probability of knowing what you know now and finding out what you will find, if the proposition is true or false.*

Historical data: Illustration

- Naïve inference (using a prior of 80% due to K1): 94% if K2; 50% if not K2
- Inference if K1 used to decide on search for K2 but prior is "reset" to .5 80% if K2; 20% if not K2
- Sophisticated inference: 83% if K2; 50% if not K2
- This sophisticated inference is unchanged if you take explicit account of the fact that searching for K2 was conditional on K1; either way it is still Pr(H | K1, K2).
- *Can such beliefs be elicited?* Perhaps.

Will it make a difference?

- Striking paucity of evidence.

How to?

- Let's look at an example
- Design declaration idea

Subsection 6

Reconciliation

Reconciliation

Incentives and strategies

Reconciliation

Table 22.1: Illustration of an inquiry reconciliation table.

Inquiry	In the preanalysis plan	In the paper	In the appendix
Gender effect	X	X	
Age effect			X

Table 22.2: Illustration of an answer strategy reconciliation table.

Inquiry	Following *A* from the PAP	Following *A* from the paper	Notes
Gender effect	estimate = 0.6, s.e = 0.31	estimate = 0.6, s.e = 0.25	Difference due to change in control variables [provide cross references to tables and code]

Reconciliation

Subsection 7

Replication files