Day 05: Power Analysis and DeclareDesign

Erin Rossiter

February 8, 2022

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Today's plan

1. Tidy up
   - James, will you:
     - » remind me to take a break at 4:30-4:45
     - » jot notes on typos & email after class
   - Questions about HW4?
   - HW5 (the last one!) will give you an opportunity to start designing your experiment
     - » I'll distribute after seeing HW4
   - A note on pre-reg readings for next week
2. Power analysis lecture and lab
3. DeclareDesign paradigm and lab

Recap

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

ATE and $\hat{ATE}$

- The average treatment effect (ATE) is often an interesting, important, useful estimand.
- Difference-in-means is an **unbiased** estimator of the ATE assuming:
  - **Randomization of treatment**
    - $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
    - $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$
    - We can estimate left-hand terms using our observed data!
- What about **uncertainty** with our $\hat{ATE}$?

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Sampling distribution of $\hat{ATE}$

Distribution of all $\hat{ATE}$'s across all hypothetical random assignments

- There's some true standard error: $SE(\hat{ATE})$
  - (Standard deviation of the sampling distribution)
  - Our design choices influence this!
  - Beneficial to reduce standard error in design phase.
- We must estimate the standard error: $\hat{SE}(\hat{ATE})$
  - Be careful of known bias in estimators of $SE(\hat{ATE})$
    - i.e., pooling when it was a blocked design
- Now that we've quantified uncertainty about $\hat{ATE}$, we can test our hypotheses

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Null hypothesis testing

1. We state a null hypothesis, usually $H_0 : ATE = 0$
2. We choose a test statistic
3. Determine the distribution of the test statistic **under the null**
4. Calculate the probability (p-value) of observing our test statistic **under the null**

What are the major differences between these two hypothesis testing procedures?

- t-test
- Randomization inference

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $\hat{ATE}$
- ✓ Quantifying uncertainty; reducing $SE(\hat{ATE})$ in our design
- ✓ Hypothesis testing; can we reject the null?
- → **Power**
  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...
- → **Power analysis**
  - Process of generating a guess of how prepared we are to reject the null given our design

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $A\hat{T}E$
- ✓ Quantifying uncertainty; reducing $SE(A\hat{T}E)$ in our design
- ✓ Hypothesis testing; can we reject the null?
- → **Power**
  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...
- → **Power analysis**
  - Process of generating a guess of how prepared we are to reject the null given our design

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $A\hat{T}E$
- ✓ Quantifying uncertainty; reducing $SE(A\hat{T}E)$ in our design
- ✓ Hypothesis testing; can we reject the null?

→ **Power**

  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...

→ **Power analysis**

  - Process of generating a guess of how prepared we are to reject the null given our design

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $A\hat{T}E$
- ✓ Quantifying uncertainty; reducing $SE(A\hat{T}E)$ in our design
- ✓ Hypothesis testing; can we reject the null?
- → **Power**
  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...
- → **Power analysis**
  - Process of generating a guess of how prepared we are to reject the null given our design

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $A\hat{T}E$
- ✓ Quantifying uncertainty; reducing $SE(A\hat{T}E)$ in our design
- ✓ Hypothesis testing; can we reject the null?
- → **Power**
  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...
- → **Power analysis**
  - Process of generating a guess of how prepared we are to reject the null given our design

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $A\hat{T}E$
- ✓ Quantifying uncertainty; reducing $SE(A\hat{T}E)$ in our design
- ✓ Hypothesis testing; can we reject the null?
- → **Power**
  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...
- → **Power analysis**
  - Process of generating a guess of how prepared we are to reject the null given our design

Roadmap

- ✓ ATE as our estimand; DIM as unbiased estimator of $A\hat{T}E$
- ✓ Quantifying uncertainty; reducing $SE(A\hat{T}E)$ in our design
- ✓ Hypothesis testing; can we reject the null?
- → **Power**
  - A measure of how often would we reject the null if we could re-run our experiment and hypothesis test over and over...
- → **Power analysis**
  - Process of generating a guess of how prepared we are to reject the null given our design

Power

Power and hypothesis testing

- Hypothesis testing in polisci often assumes conventional $p < 0.05$ is "statistically significant"
  - Note: nothing to do with statistical theory!
  - We know p value is a measure of how surprising our results are given the null, so are $p = .049$ and $p = .051$ really that different?
  - Just convention
  - But, a convention so deeply entrenched we should be prepared to speak to our results in this framework
  - How? **Statistical power**

Power and hypothesis testing

- Hypothesis testing in polisci often assumes conventional $p < 0.05$ is "statistically significant"
  - Note: nothing to do with statistical theory!
  - We know p value is a measure of how surprising our results are given the null, so are $p = .049$ and $p = .051$ really that different?
  - Just convention
  - But, a convention so deeply entrenched we should be prepared to speak to our results in this framework
  - How? **Statistical power**

Power and hypothesis testing

- Hypothesis testing in polisci often assumes conventional $p < 0.05$ is "statistically significant"
  - Note: nothing to do with statistical theory!
  - We know p value is a measure of how surprising our results are given the null, so are $p = .049$ and $p = .051$ really that different?
  - Just convention
  - But, a convention so deeply entrenched we should be prepared to speak to our results in this framework
  - How? **Statistical power**

Power and hypothesis testing

- Hypothesis testing in polisci often assumes conventional $p < 0.05$ is "statistically significant"
  - Note: nothing to do with statistical theory!
  - We know p value is a measure of how surprising our results are given the null, so are $p = .049$ and $p = .051$ really that different?
  - Just convention
  - But, a convention so deeply entrenched we should be prepared to speak to our results in this framework
  - How? **Statistical power**

Power and hypothesis testing

- Hypothesis testing in polisci often assumes conventional $p < 0.05$ is "statistically significant"
  - Note: nothing to do with statistical theory!
  - We know p value is a measure of how surprising our results are given the null, so are $p = .049$ and $p = .051$ really that different?
  - Just convention
  - But, a convention so deeply entrenched we should be prepared to speak to our results in this framework
  - How? **Statistical power**

Power and hypothesis testing

- Hypothesis testing in polisci often assumes conventional $p < 0.05$ is "statistically significant"
  - » Note: nothing to do with statistical theory!
  - » We know p value is a measure of how surprising our results are given the null, so are $p = .049$ and $p = .051$ really that different?
  - » Just convention
  - » But, a convention so deeply entrenched we should be prepared to speak to our results in this framework
  - » How? **Statistical power**

Big picture

Power is the ability to distinguish signal from noise

Compare two experiments where the true effect of each treatment is a 1pp increase in average outcome

1. Some treatment's effect on rare disease
2. Some treatment's effect on income

Do you think we will find the signal in the noise? Why or why not?

(From Alexander Coppock's EGAP 10 Things to Know)

Big picture

Power is the ability to distinguish signal from noise

Compare two experiments where the true effect of each treatment is a 1pp increase in average outcome

1. Some treatment's effect on rare disease
2. Some treatment's effect on income

Do you think we will find the signal in the noise? Why or why not?

(From Alexander Coppock's EGAP 10 Things to Know)

Big picture

Power is the ability to distinguish signal from noise

Compare two experiments where the true effect of each treatment is a 1pp increase in average outcome

1. Some treatment's effect on rare disease
2. Some treatment's effect on income

Do you think we will find the signal in the noise? Why or why not?

(From Alexander Coppock's EGAP 10 Things to Know)

Big picture

Power is the ability to distinguish signal from noise

Compare two experiments where the true effect of each treatment is a 1pp increase in average outcome

1. Some treatment's effect on rare disease
2. Some treatment's effect on income

Do you think we will find the signal in the noise? Why or why not?

(From Alexander Coppock's EGAP 10 Things to Know)

Big picture

Power is the ability to distinguish signal from noise

Compare two experiments where the true effect of each treatment is a 1pp increase in average outcome

1. Some treatment's effect on rare disease
2. Some treatment's effect on income

Do you think we will find the signal in the noise? Why or why not?

(From Alexander Coppock's EGAP 10 Things to Know)

Big picture

Power is the ability to distinguish signal from noise

Compare two experiments where the true effect of each treatment is a 1pp increase in average outcome

1. Some treatment's effect on rare disease
2. Some treatment's effect on income

Do you think we will find the signal in the noise? Why or why not?

(From Alexander Coppock's EGAP 10 Things to Know)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Big picture

The danger of **false negatives** is a major concern

- Type 2 error
  - β = probability of "retaining" (or fail to reject) false null
- True causal effect goes undetected
- All the work to randomize treatment, collect data, analyze results... and we're unable detect results that are there!
- **Address ability to detect effects in the design phase by considering statistical power**
  - Power $= 1 - \beta$
  - Probability of rejecting a false null (what we hope to do!)

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

Power analysis goal

The question we're trying to answer:

- Supposing there truly is a treatment effect, imagine you could re-run your experiment a huge number of times, how often would you get a statistically significant result?

Power analysis is done *before* experiment, so answering this question requires informed guesswork (i.e., assumptions), including:

- treatment effect size
- N
- $\alpha = .05$

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

How to generate your assumptions

We haven't run the experiment yet! How do we guess these things (i.e., treatment effect size)?

- Previous studies
- Your own expertise!
- Information about the sample
- etc.

The good news–it is easy to find out how much your conclusions depend on your assumptions

- Vary your assumptions and see how power varies
  - i.e., $ATE = 1, ATE = 2, ATE = 3...$
- Are my conclusions about power very sensitive to this assumption? How much?

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta$ = statistical power
- μ_c = mean (normally distributed) outcome in control group
- μ_t = mean (normally distributed) outcome in treatment group
- σ = standard deviation of the outcome of each group
- α = desired level of statistical significance (usually .05)
- $\Phi(\cdot)$ = normal CDF
- $\Phi^{-1}(\cdot)$ = inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta$ = statistical power
- μ_c = mean (normally distributed) outcome in control group
- μ_t = mean (normally distributed) outcome in treatment group
- σ = standard deviation of the outcome of each group
- α = desired level of statistical significance (usually .05)
- $\Phi(\cdot)$ = normal CDF
- $\Phi^{-1}(\cdot)$ = inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Common formula to calculate power

$$\text{Power} = 1 - \beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

- $1 - \beta =$ statistical power
- $\mu_c =$ mean (normally distributed) outcome in control group
- $\mu_t =$ mean (normally distributed) outcome in treatment group
- $\sigma =$ standard deviation of the outcome of each group
- $\alpha =$ desired level of statistical significance (usually .05)
- $\Phi(\cdot) =$ normal CDF
- $\Phi^{-1}(\cdot) =$ inverse of normal CDF
- Assumes all subjects have equal prob of being in treatment or control

See pg 93 of GG

Example

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}

power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)

## [1] 0.7981752
```

Example

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

```
power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

```
## [1] 0.7981752
```

Example

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

```
power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

```
## [1] 0.7981752
```

Varying N

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as N increases?

```
power_func(N = c(500, 1000, 1500),
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

Varying N

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as N increases?

```
power_func(N = c(500, 1000, 1500),
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

Varying N

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as N increases?

```
power_func(N = c(500, 1000, 1500),
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

Varying N

```
power_func(N = c(500, 1000, 1500),
           mu_t = 65,
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

```
## [1] 0.7981752 0.9768629 0.9980196
```

Varying treatment effect size

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as $|\mu_t - \mu_c|$ increases?

```
power_func(N = 500,
           mu_t = c(65,66,67),
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

Varying treatment effect size

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as $|\mu_t - \mu_c|$ increases?

```
power_func(N = 500,
           mu_t = c(65,66,67),
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

Varying treatment effect size

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as $|\mu_t - \mu_c|$ increases?

```
power_func(N = 500,
           mu_t = c(65,66,67),
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

Varying treatment effect size

```
power_func(N = 500,
           mu_t = c(65,66,67),
           mu_c = 60,
           sigma = 20,
           alpha = .05)
```

```
## [1] 0.7981752 0.9183620 0.9745994
```

Varying standard deviation in outcomes

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as σ increases?

```
power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = c(20,22,24),
           alpha = .05)
```

Varying standard deviation in outcomes

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as σ increases?

```
power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = c(20,22,24),
           alpha = .05)
```

Varying standard deviation in outcomes

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

```
power_func <- function(N, mu_t, mu_c, sigma, alpha){
  term1 <- (abs(mu_t - mu_c)*sqrt(N))/(2*sigma)
  term2 <- 1-(alpha/2)
  pnorm(term1 - qnorm(term2))
}
```

What happens as σ increases?

```
power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = c(20,22,24),
           alpha = .05)
```

Varying standard deviation in outcomes

```
power_func(N = 500,
           mu_t = 65,
           mu_c = 60,
           sigma = c(20,22,24),
           alpha = .05)
```

```
## [1] 0.7981752 0.7193873 0.6440381
```

Power analysis in practice

Maybe this formula & the assumptions it makes will work for some of your work

When will it work well? When might it not?

Power analysis in practice

Maybe this formula & the assumptions it makes will work for some of your work

When will it work well? When might it not?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

Power analysis in practice

Canned formulas are tough to tweak

- Three conditions instead of two
- Blocked or clustered random assignment
- Outcomes not expected to be normally distributed
- Expected attrition
- ...

Answer? Simulation

- More flexible
- More intuitive (?)
- Simulate M runs of the experiment where you...
  - state potential outcomes
  - allocate treatment
  - test hypothesis
  - **a p value is generated**
- Power = proportion of M simulations where $p < \alpha$
- Sound familiar?

DeclareDesign

```
N <- 500
population <- declare_population(N = 500)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))

estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)

design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
diagnosis <- diagnose_design(design, sims = 1000)

diagnosis$diagnosands_df$power

## [1] 0.807
```

DeclareDesign

```
N <- 500
population <- declare_population(N = 500)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))
```

```
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
```

```
design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
diagnosis <- diagnose_design(design, sims = 1000)

diagnosis$diagnosands_df$power

## [1] 0.807
```

DeclareDesign

```
N <- 500
population <- declare_population(N = 500)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))
```

```
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
```

```
design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
diagnosis <- diagnose_design(design, sims = 1000)
```

```
diagnosis$diagnosands_df$power
```

```
## [1] 0.807
```

DeclareDesign

```
N <- 500
population <- declare_population(N = 500)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))
```

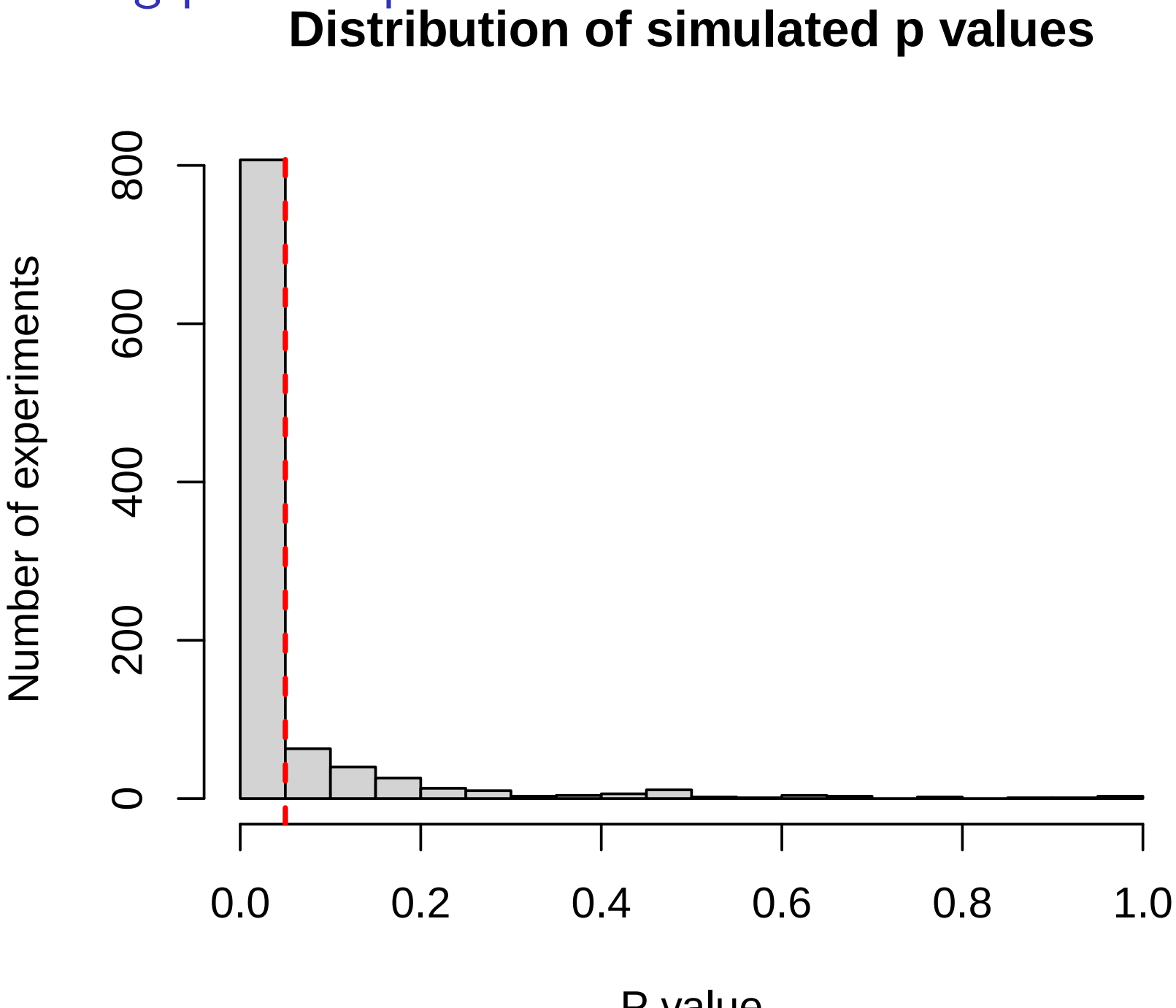
```
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
```

```
design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
diagnosis <- diagnose_design(design, sims = 1000)
```

```
diagnosis$diagnosands_df$power
```

```
## [1] 0.807
```

Visualizing possible p values



Varying design features is easy

```
N <- 500 ## Let's vary N!
population <- declare_population(N = N)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
```

```
design_Ns <- redesign(design, N = seq(400,600,50)) # Varying N

diagnosis_Ns <- diagnose_design(design_Ns, sims = 500)
diagnosis_Ns$diagnosands_df$power

## [1] 0.704 0.770 0.804 0.840 0.842
```

Varying design features is easy

```
N <- 500 ## Let's vary N!
population <- declare_population(N = N)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
```

```
design_Ns <- redesign(design, N = seq(400,600,50)) # Varying N
```

```
diagnosis_Ns <- diagnose_design(design_Ns, sims = 500)
diagnosis_Ns$diagnosands_df$power
```

```
## [1] 0.704 0.770 0.804 0.840 0.842
```

Varying design features is easy

```
N <- 500 ## Let's vary N!
population <- declare_population(N = N)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
design <- (population + po +
             estimand + assignment + reveal_Y + estimator)
```

```
design_Ns <- redesign(design, N = seq(400,600,50)) # Varying N
```

```
diagnosis_Ns <- diagnose_design(design_Ns, sims = 500)
diagnosis_Ns$diagnosands_df$power
```

```
## [1] 0.704 0.770 0.804 0.840 0.842
```

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**
   - Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
   - Ex: Carey's Rooney Center talk, video vs. script
2. **Background noise**
   - Selecting and measuring outcomes with lower variability
   - Ex: feeling thermometers
   - (Of course, background noise is inherent in social science)
3. **Experimental design**
   - How many treatment groups?
   - How is randomization conducted?
   - Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**
   - Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
   - Ex: Carey's Rooney Center talk, video vs. script
2. **Background noise**
   - Selecting and measuring outcomes with lower variability
   - Ex: feeling thermometers
   - (Of course, background noise is inherent in social science)
3. **Experimental design**
   - How many treatment groups?
   - How is randomization conducted?
   - Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about *N*

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

Three main ingredients of statistical power

Power is not all about N

1. **Strength of treatment**

- Ex: if your treatment were giving every subject $100K, we would be able to see differences in behavior between the treatment and control groups!
- Ex: Carey's Rooney Center talk, video vs. script

2. **Background noise**

- Selecting and measuring outcomes with lower variability
- Ex: feeling thermometers
- (Of course, background noise is inherent in social science)

3. **Experimental design**

- How many treatment groups?
- How is randomization conducted?
- Will other factors be controlled for in analysis?

In practice

- You'll often see a standard of .80 power
  - Another convention not rooted in statistical theory
  - If you can reach .85 or .95, why not?
  - If the best you can do is .75, why not?

In practice

- You'll often see a standard of .80 power
  - Another convention not rooted in statistical theory
  - If you can reach .85 or .95, why not?
  - If the best you can do is .75, why not?

In practice

- You'll often see a standard of .80 power
  - Another convention not rooted in statistical theory
  - If you can reach .85 or .95, why not?
  - If the best you can do is .75, why not?

In practice

- You'll often see a standard of .80 power
  - Another convention not rooted in statistical theory
  - If you can reach .85 or .95, why not?
  - If the best you can do is .75, why not?

In practice

- You'll often see a standard of .80 power
  - Another convention not rooted in statistical theory
  - If you can reach .85 or .95, why not?
  - If the best you can do is .75, why not?

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

   – Give yourself the best shot possible!
   – Tweak *in the design phase*
      » Ex: interaction's power is coming back at .20?
      » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

   – Seems like a time cost, but simulation forces you to think through *all* steps of design
   – Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

   – Give yourself the best shot possible!
   – Tweak *in the design phase*
      » Ex: interaction's power is coming back at .20?
      » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

   – Seems like a time cost, but simulation forces you to think through *all* steps of design
   – Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

- Give yourself the best shot possible!
- Tweak *in the design phase*
  - » Ex: interaction's power is coming back at .20?
  - » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

- Seems like a time cost, but simulation forces you to think through *all* steps of design
- Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

- Give yourself the best shot possible!
- Tweak *in the design phase*
  - » Ex: interaction's power is coming back at .20?
  - » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

- Seems like a time cost, but simulation forces you to think through *all* steps of design
- Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

- Give yourself the best shot possible!
- Tweak *in the design phase*
  - » Ex: interaction's power is coming back at .20?
  - » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

- Seems like a time cost, but simulation forces you to think through *all* steps of design
- Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

- Give yourself the best shot possible!
- Tweak *in the design phase*
  - » Ex: interaction's power is coming back at .20?
  - » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

- Seems like a time cost, but simulation forces you to think through *all* steps of design
- Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

- Give yourself the best shot possible!
- Tweak *in the design phase*
  - » Ex: interaction's power is coming back at .20?
  - » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

- Seems like a time cost, but simulation forces you to think through *all* steps of design
- Analysis after fielding is then straightforward

Summing up

1. The hard truth: why conduct an experiment you have no realistic chance in answering?

- Give yourself the best shot possible!
- Tweak *in the design phase*
  - » Ex: interaction's power is coming back at .20?
  - » Increase N, collapse conditions, change outcome measure, etc. but don't field as is!

2. Pre-registration (next week) is increasingly expected. Power analysis is a part of that.

- Seems like a time cost, but simulation forces you to think through *all* steps of design
- Analysis after fielding is then straightforward

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

   – You'll be expected to answer if you are powered to reject the null with that effect size and N

   – I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

   – Ex: experiment with 100K people and 0.1pp increase in outcome
      » Statistical significance trivially easy to achieve
      » Substantive significance needs to be questioned

   – Ex: first of its kind experiment fails to reach $p < .05$
      » Best guess at true *ATE*
      » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Summing up

3. Trying to write up a null is *tough*

- You'll be expected to answer if you are powered to reject the null with that effect size and N
- I'm going through this now! Powered to detect effect size of 4 not 3. . . . $p = .15$. . .

4. Power isn't everything

- Ex: experiment with 100K people and 0.1pp increase in outcome
  - » Statistical significance trivially easy to achieve
  - » Substantive significance needs to be questioned
- Ex: first of its kind experiment fails to reach $p < .05$
  - » Best guess at true *ATE*
  - » Further investigation warranted

Lab 1

DeclareDesign

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Recap

- We've be relying on DeclareDesign to build intuition
  - A lot has been handwavy!
- Time today devoted to understanding what I call the DeclareDesign "paradigm"
- Resources
  - Book
    - Note the "Research Design Library" section
  - Website
    - Templates, dicussion board, mostly package-oriented
  - APSR

Sec 1 of Research Design: Declaration, Diagnosis, Redesign

The core idea is the MIDA framework, in which a research design is characterized by four elements: a model, an inquiry, a data strategy, and an answer strategy... The design encodes your beliefs about the world, it describes your questions, and it lays out how you go about answering those questions, both in terms of what data you collect and how you analyze it...

We think of designs as objects that can be interrogated. *Each of the four design elements can be "declared" in computer code and – if done right – the information provided is enough to "diagnose" the quality of the design through computer simulation. Researchers can then select the best design for their purposes by "redesigning" over alternative, feasible designs.*

→ paradigm to **construct** and **critique** your research design

Sec 1 of Research Design: Declaration, Diagnosis, Redesign

The core idea is the MIDA framework, in which a research design is characterized by four elements: a model, an inquiry, a data strategy, and an answer strategy... The design encodes your beliefs about the world, it describes your questions, and it lays out how you go about answering those questions, both in terms of what data you collect and how you analyze it...

We think of designs as objects that can be interrogated. *Each of the four design elements can be "declared" in computer code and – if done right – the information provided is enough to "diagnose" the quality of the design through computer simulation. Researchers can then select the best design for their purposes by "redesigning" over alternative, feasible designs.*

→ paradigm to **construct** and **critique** your research design

Sec 1 of Research Design: Declaration, Diagnosis, Redesign

The core idea is the MIDA framework, in which a research design is characterized by four elements: a model, an inquiry, a data strategy, and an answer strategy... The design encodes your beliefs about the world, it describes your questions, and it lays out how you go about answering those questions, both in terms of what data you collect and how you analyze it...

We think of designs as objects that can be interrogated. *Each of the four design elements can be "declared" in computer code and – if done right – the information provided is enough to "diagnose" the quality of the design through computer simulation. Researchers can then select the best design for their purposes by "redesigning" over alternative, feasible designs.*

→ paradigm to **construct** and **critique** your research design

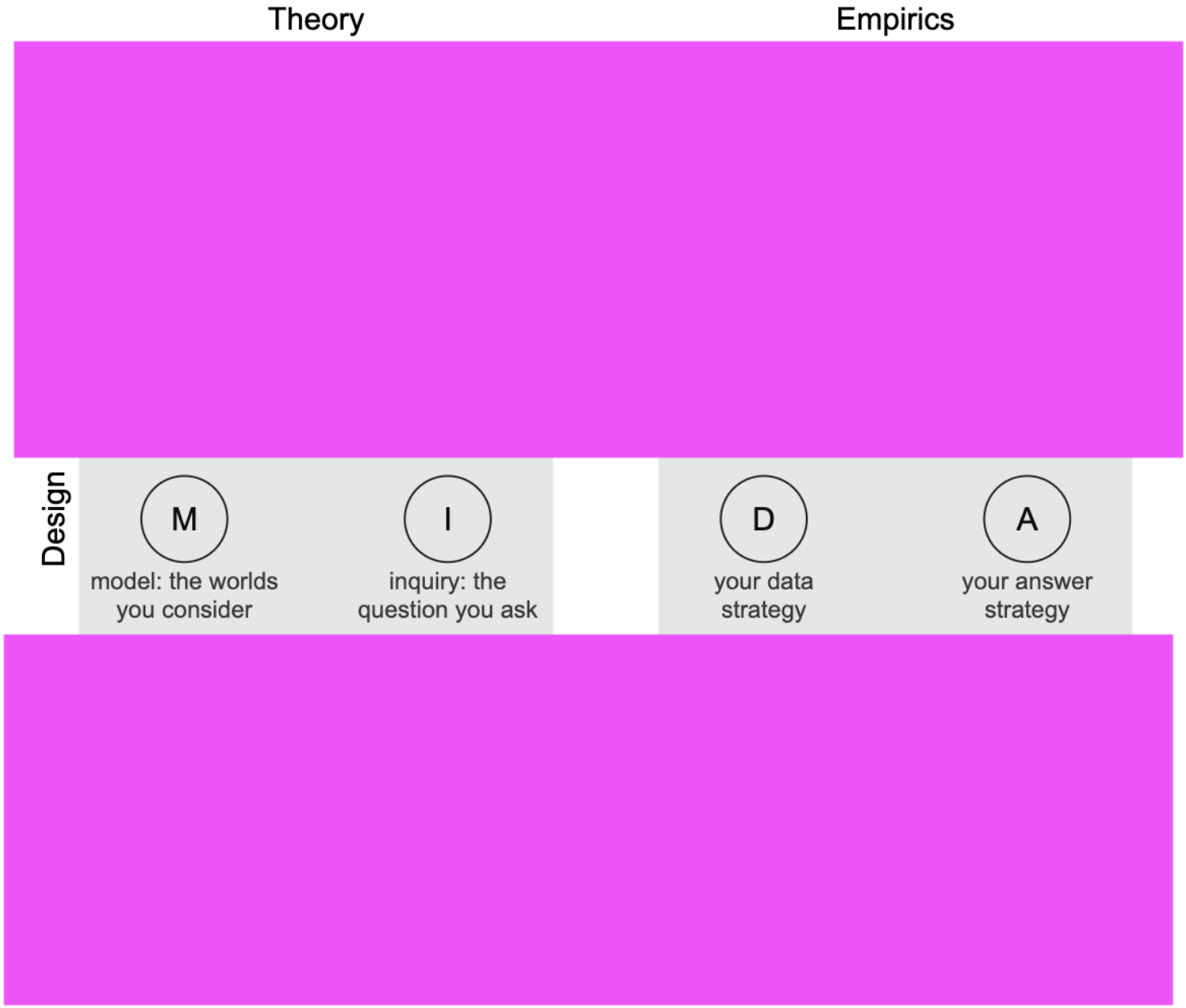
Sec 1 of Research Design: Declaration, Diagnosis, Redesign

The core idea is the MIDA framework, in which a research design is characterized by four elements: a model, an inquiry, a data strategy, and an answer strategy... The design encodes your beliefs about the world, it describes your questions, and it lays out how you go about answering those questions, both in terms of what data you collect and how you analyze it...

We think of designs as objects that can be interrogated. *Each of the four design elements can be "declared" in computer code and – if done right – the information provided is enough to "diagnose" the quality of the design through computer simulation. Researchers can then select the best design for their purposes by "redesigning" over alternative, feasible designs.*

$\rightarrow$ paradigm to **construct** and **critique** your research design

MIDA


Theory
Empirics
Design
M
model: the worlds you consider
I
inquiry: the question you ask
D
your data strategy
A
your answer strategy

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

M Models of how the world works

- Not necessarily your beliefs, just *if* it worked this way
- We're trying to understand if research design is good despite many ways the world could work
- Theoretical work
  - » Reflect on existing theories and studies, maybe pilot data, etc.
- Steps in the DeclareDesign process? Examples?

MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

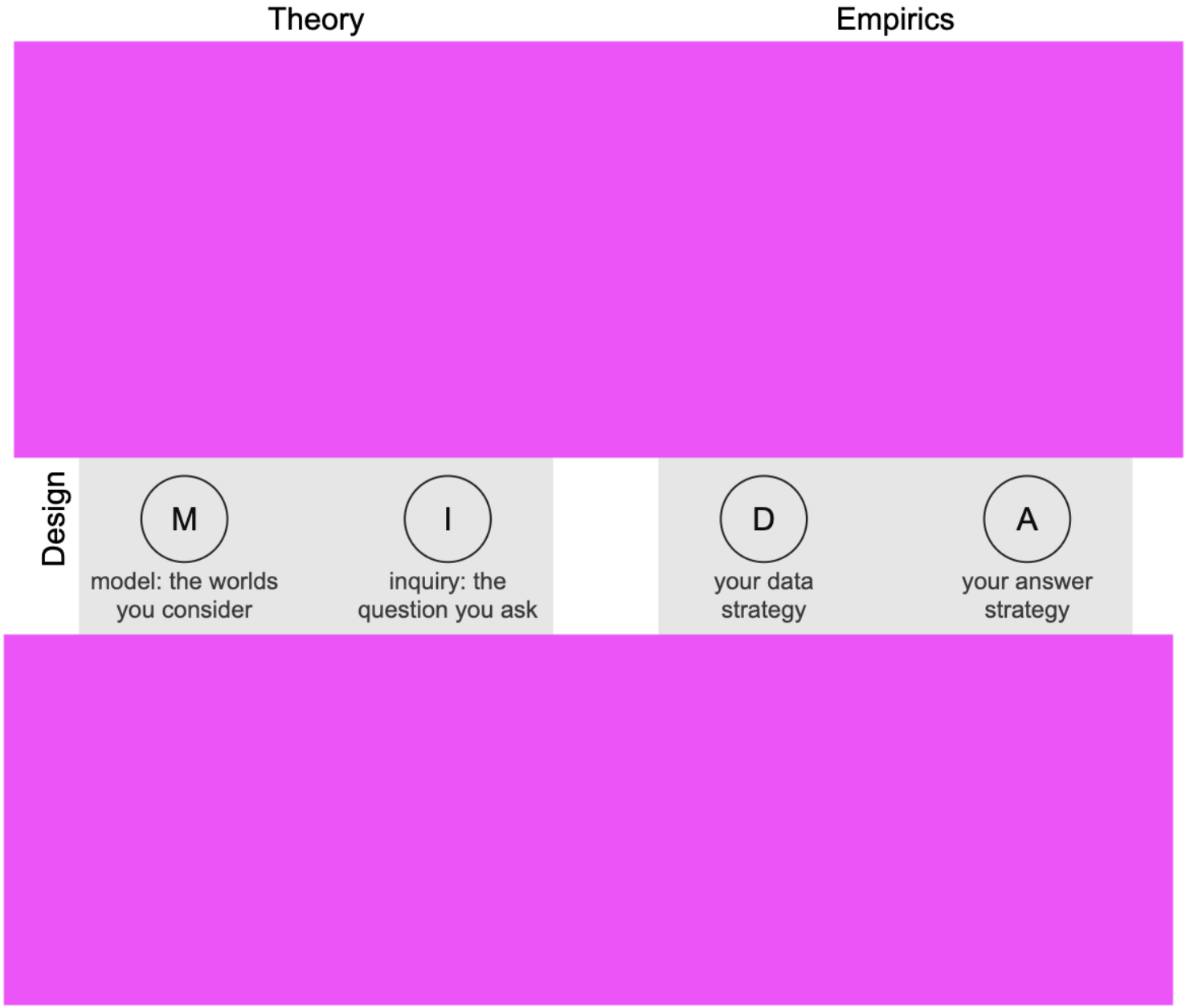
MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

MIDA, the theoretical half

- Inquiry we want to ask
  - Inquiry = question, function of the model of the world
  - Does gaining political office make divorce more likely?
  - So far, our inquiry has been about average treatment effects
  - Estimand (ATE) is the answer to inquiry
    - More complex inquiries to come, like ATT, heterogeneous effects, etc.
  - Steps in the DeclareDesign process?

MIDA


Theory
Empirics
Design
M
model: the worlds you consider
I
inquiry: the question you ask
D
your data strategy
A
your answer strategy

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

MIDA, the empirical half

D Data strategy

- Procedures to gather information about the world
- How units are sampled, conditions assigned, and variables measured
- Steps in the DeclareDesign process?

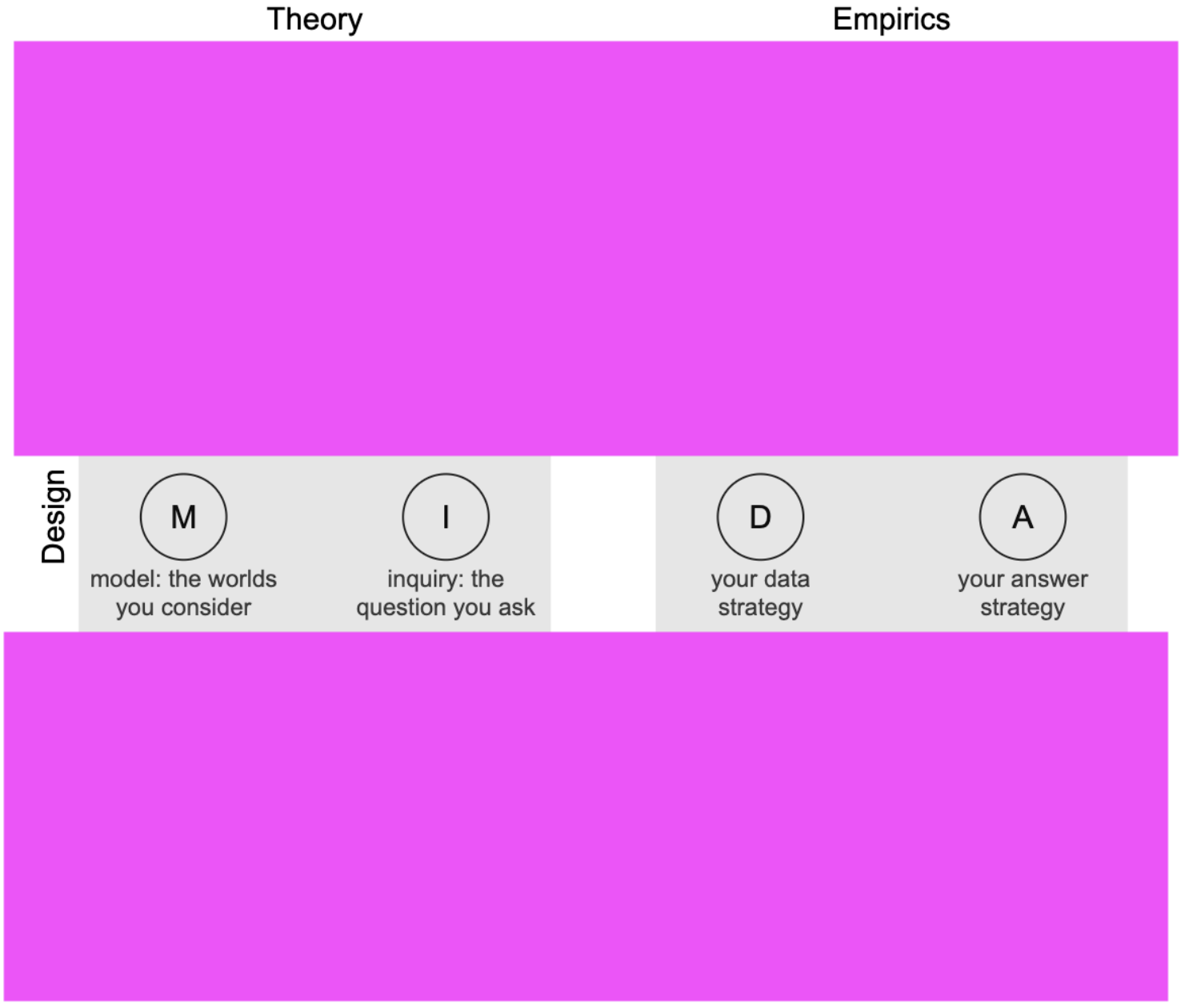
A Answer strategy

- Procedures for summarizing the data
- Cleaning, estimation, interpretation, etc.
- Steps in the DeclareDesign process?

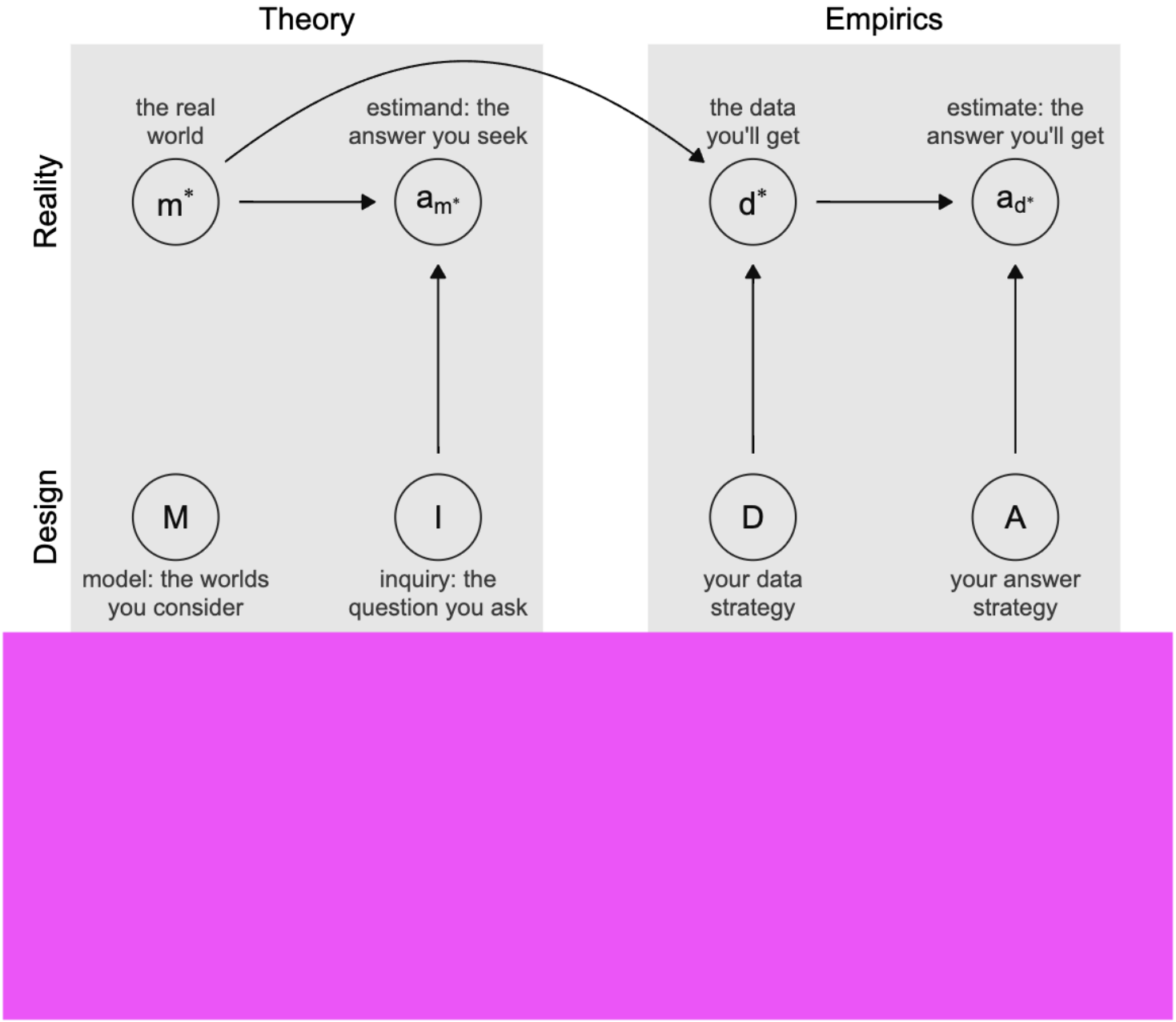
MIDA

```
# M
N <- 500
population <- declare_population(N = N)
po <- declare_potential_outcomes(Y_Z_0 = rnorm(N, 60, 20),
                                 Y_Z_1 = rnorm(N, 65, 20))

# I
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))

# D
assignment <- declare_assignment(Z = complete_ra(N, m = N/2))
reveal_Y <- declare_reveal()

# A
estimator <- declare_estimator(Y ~ Z,
                               model = estimatr::difference_in_means)
```

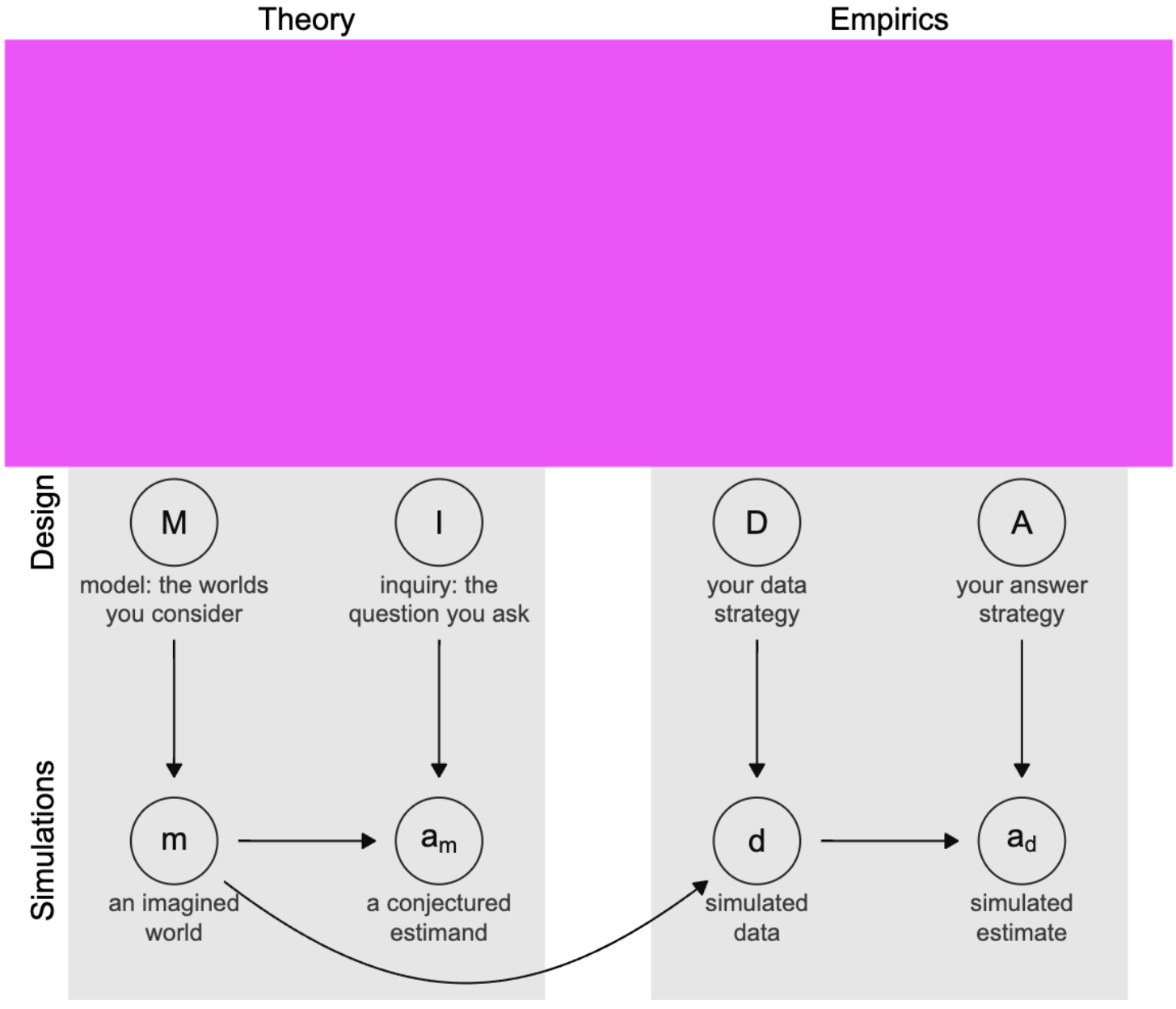
MIDA


Theory
Empirics
Design
M
model: the worlds you consider
I
inquiry: the question you ask
D
your data strategy
A
your answer strategy

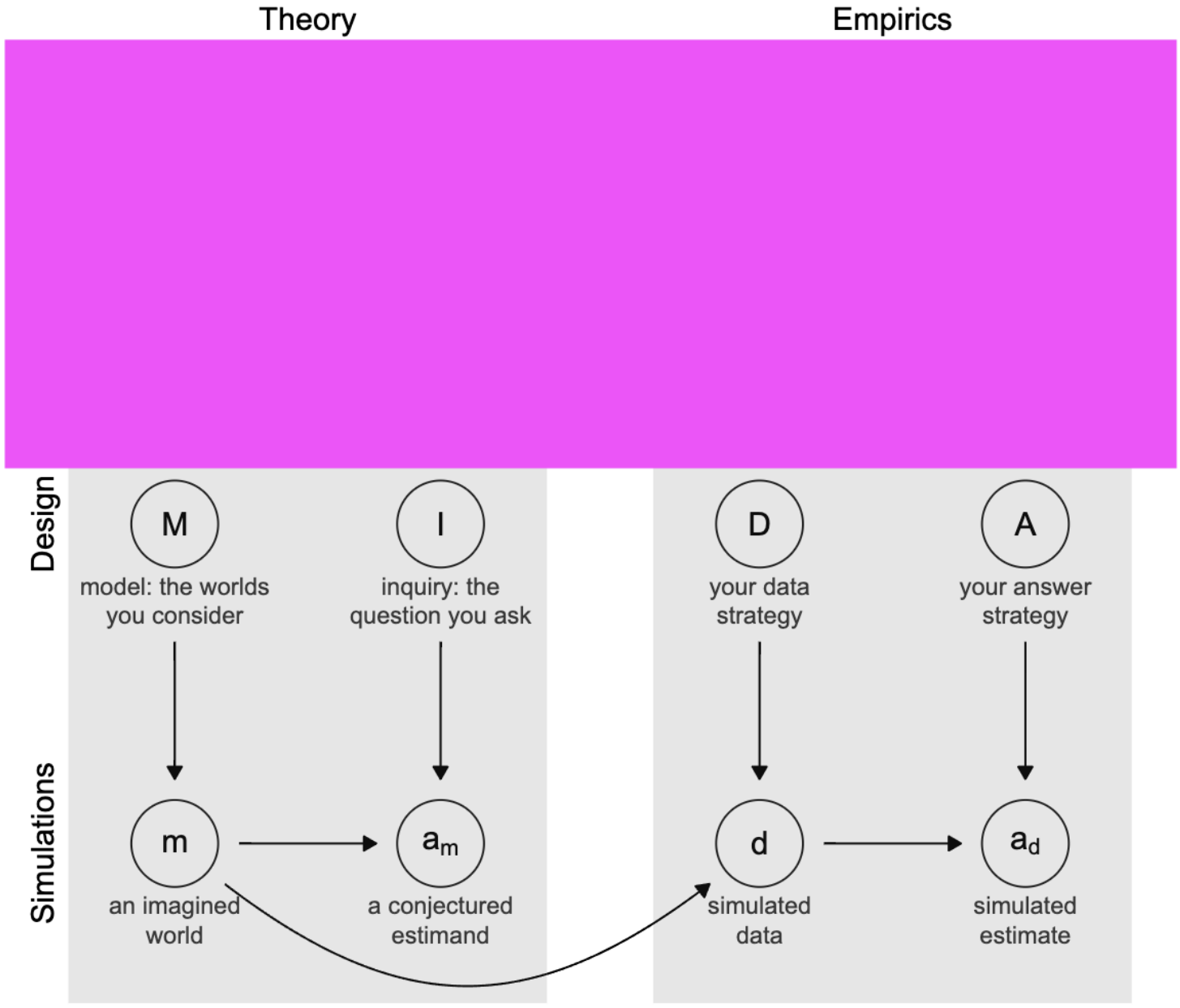
MIDA – How it maps on to the usual research process



MIDA – How it guides us via simulation



MIDA


Theory
Empirics
Design
Simulations
M
model: the worlds you consider
I
inquiry: the question you ask
D
your data strategy
A
your answer strategy
m
an imagined world
a_m
a conjectured estimand
d
simulated data
a_d
simulated estimate

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - We are pros at this
  - MIDA as functions
- **Diagnose**
  - Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - Does increasing N help?
  - What if the treatment is stronger?
  - Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - We are pros at this
  - MIDA as functions
- **Diagnose**
  - Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - Does increasing N help?
  - What if the treatment is stronger?
  - Should I simplify those two conditions into one?

In practice

- **Declare**
  - » We are pros at this
  - » MIDA as functions
- **Diagnose**
  - » Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - » Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - » Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - » Does increasing N help?
  - » What if the treatment is stronger?
  - » Should I simplify those two conditions into one?

In practice

- **Declare**
  - We are pros at this
  - MIDA as functions
- **Diagnose**
  - Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - Does increasing N help?
  - What if the treatment is stronger?
  - Should I simplify those two conditions into one?

In practice

- **Declare**
  - We are pros at this
  - MIDA as functions
- **Diagnose**
  - Diagnostic statistic
    - Summary statistic from one simulation
    - Ex: e = Estimand$_i$ - Estimate$_i$
  - Diagnosand
    - Summary of distribution of diagnostic statistic
    - Ex: Bias = Mean($\mathbf{e}$)
  - Simulation error
    - Diagnosands are approximated, so accompanied by $S\hat{E}()$
- **Redesign**
  - Does increasing N help?
  - What if the treatment is stronger?
  - Should I simplify those two conditions into one?

Key take-aways

- Design early
  - » *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - » What happens if there's attrition?
  - » What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - » Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - » Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - » *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - » Practice in this class

Key take-aways

- Design early
  - » *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - » What happens if there's attrition?
  - » What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - » Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - » Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - » *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - » Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - » *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - » What happens if there's attrition?
  - » What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - » Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - » Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - » *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - » Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - » *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - » What happens if there's attrition?
  - » What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - » Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - » Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - » *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - » Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Key take-aways

- Design early
  - *The deeper reason than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better.*
- Entertain many models
  - What happens if there's attrition?
  - What happens if these two variables are not as highly correlated as I thought?
- Have a clear (answerable) inquiry
  - Ex: Is there a treatment effect on average?
    - DIM is answer strategy for ATE estimand
  - Ex: Did New Zealand do well against Covid-19 because Prime Minister Ardern was a woman?
    - Difficult to answer empirically...
- Design to share
  - *Helps communicates your work, justify your decisions, and contribute to the scientific enterprise*
  - Practice in this class

Lab