University of Iowa
Observational Data Boot Camp
Week 3: Causal Inference

Duncan Ermini Leaf, Ph.D.

University of Southern California
Leonard D. Schaeffer Center for Health Policy & Economics

October 17, 2023

Today's topics

- The core causal inference problem with observational data
- Some popular solutions

Potential outcomes setup

For each person in a study we have:

- Y: the outcome of interest
- A: treatment assignment
  - $A = 1$ "treatment" condition
  - $A = 0$ "control" condition
- Without looking at any data, there are two **potential outcomes**:
  - Y^1 : the person's outcome if they receive the treatment
  - Y^0 : the person's outcome if they do not receive the treatment
- If $A = a$ ($a = 0$ or $a = 1$) is the actual treatment received, we observe the outcome $Y = Y^a$.
- The unobserved Y^{1-a} is called a **counterfactual** outcome.
- Causal inference studies the **treatment effect**: $Y^1 - Y^0$

Two examples

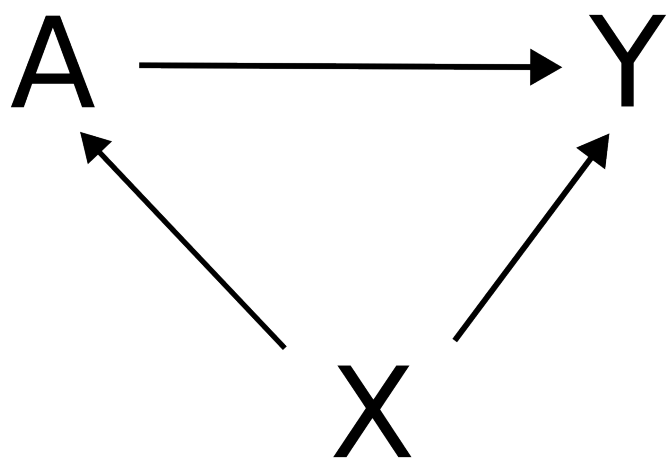
Example 1: A drug trial

- Y : patient's age at death
- Y^1: age at death if this patient is given experimental drug
- Y^0: age at death if this patient is given standard (control) drug
- $Y^1 - Y^0$: effect of experimental drug on this patient's longevity

Example 2: The benefit of education

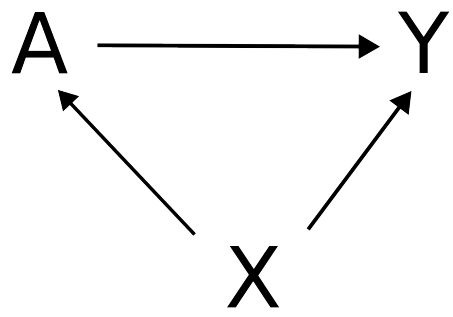
- Y : person's age at death
- Y^1: age at death if this person goes to college
- Y^0: age at death if this person only graduates high school
- $Y^1 - Y^0$: effect of college education on this person's longevity

Pre-treatment confounding



In addition to outcome Y and treatment A, we have one or more "pre-treatment" variables, X. Treatment assignment A depends on X. Also, for a given treatment assignment $A = a$, the outcome Y depends on X.

Pre-treatment confounding examples

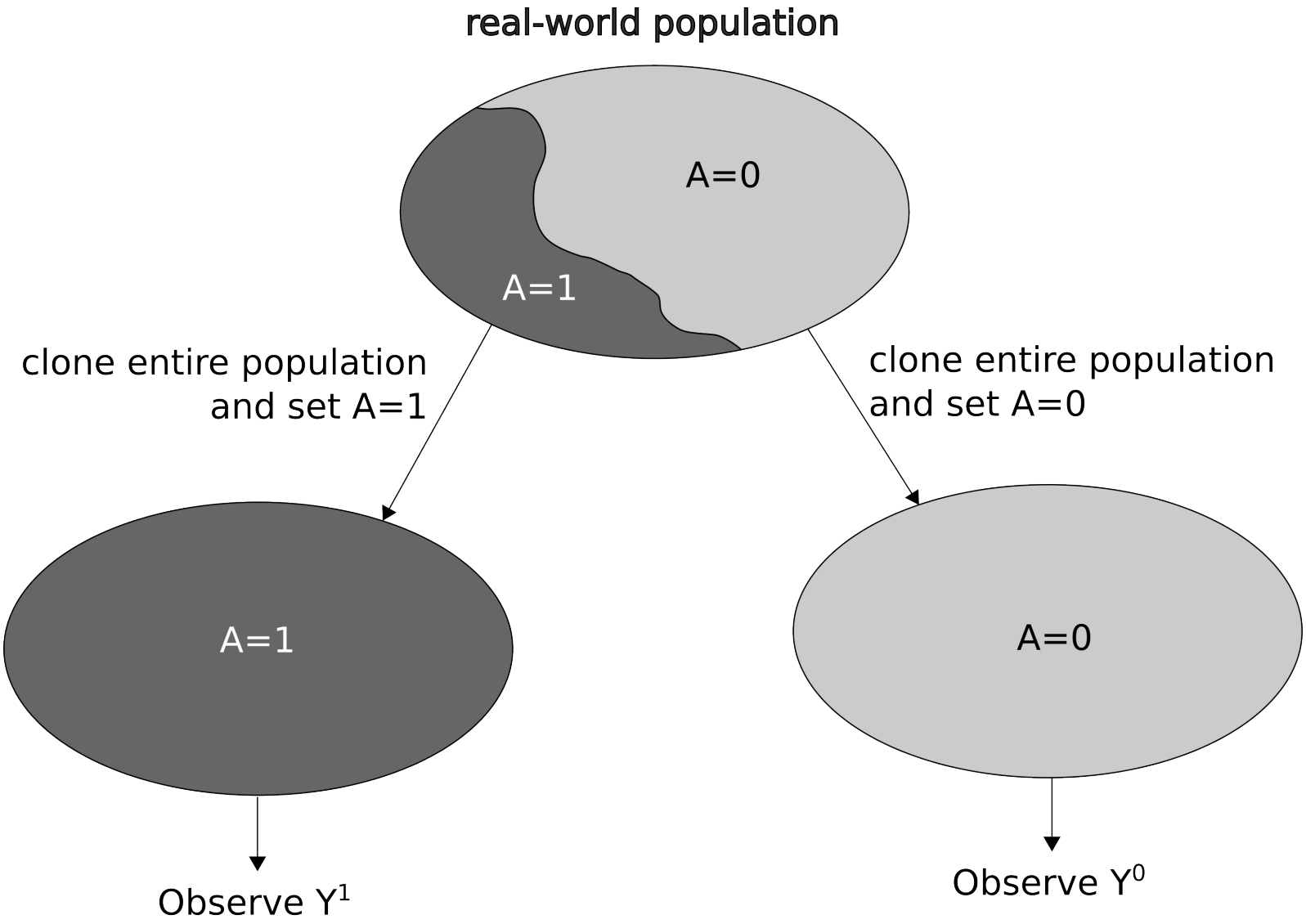


Example 1: Mother's characteristics (X) influence both the decision to smoke during pregnancy (A) and the baby's birth weight (Y).

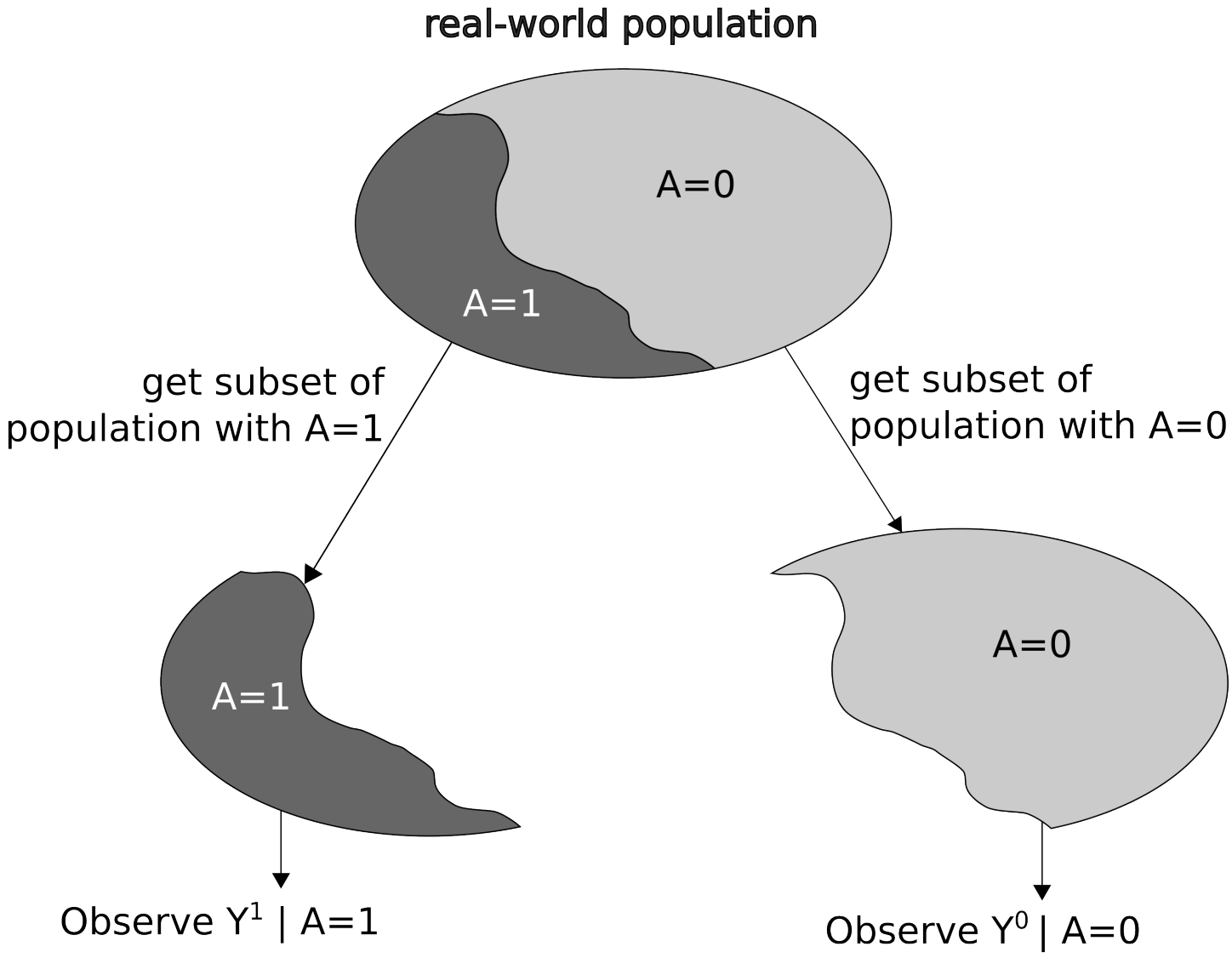
Example 2: Pre-existing health conditions (X) influence both the intensity of treatment (A) for disease such as cancer and longevity after treatment (Y).

Example 3: Adverse childhood experiences (X) influence both going to college (A) and later-life health conditions that affect longevity (Y).

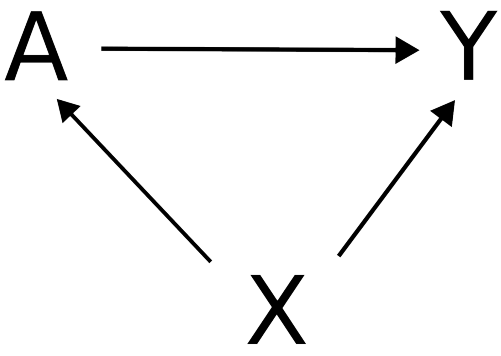
Target populations for causal inference



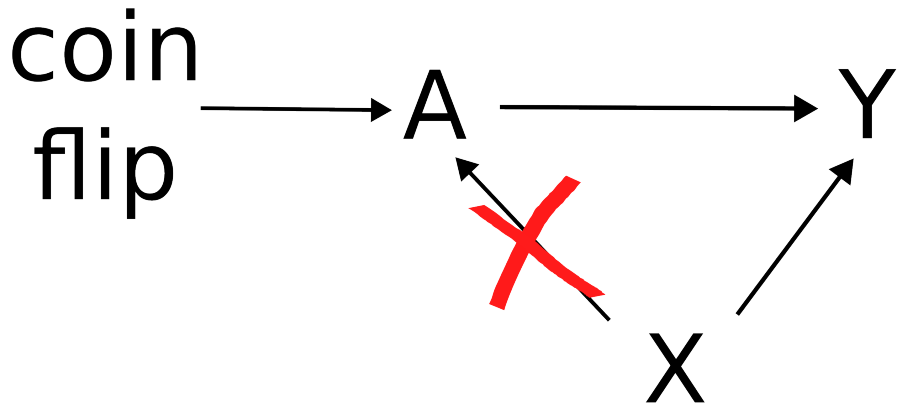
Conditioning on A doesn't work



Experimental randomization works (in theory)

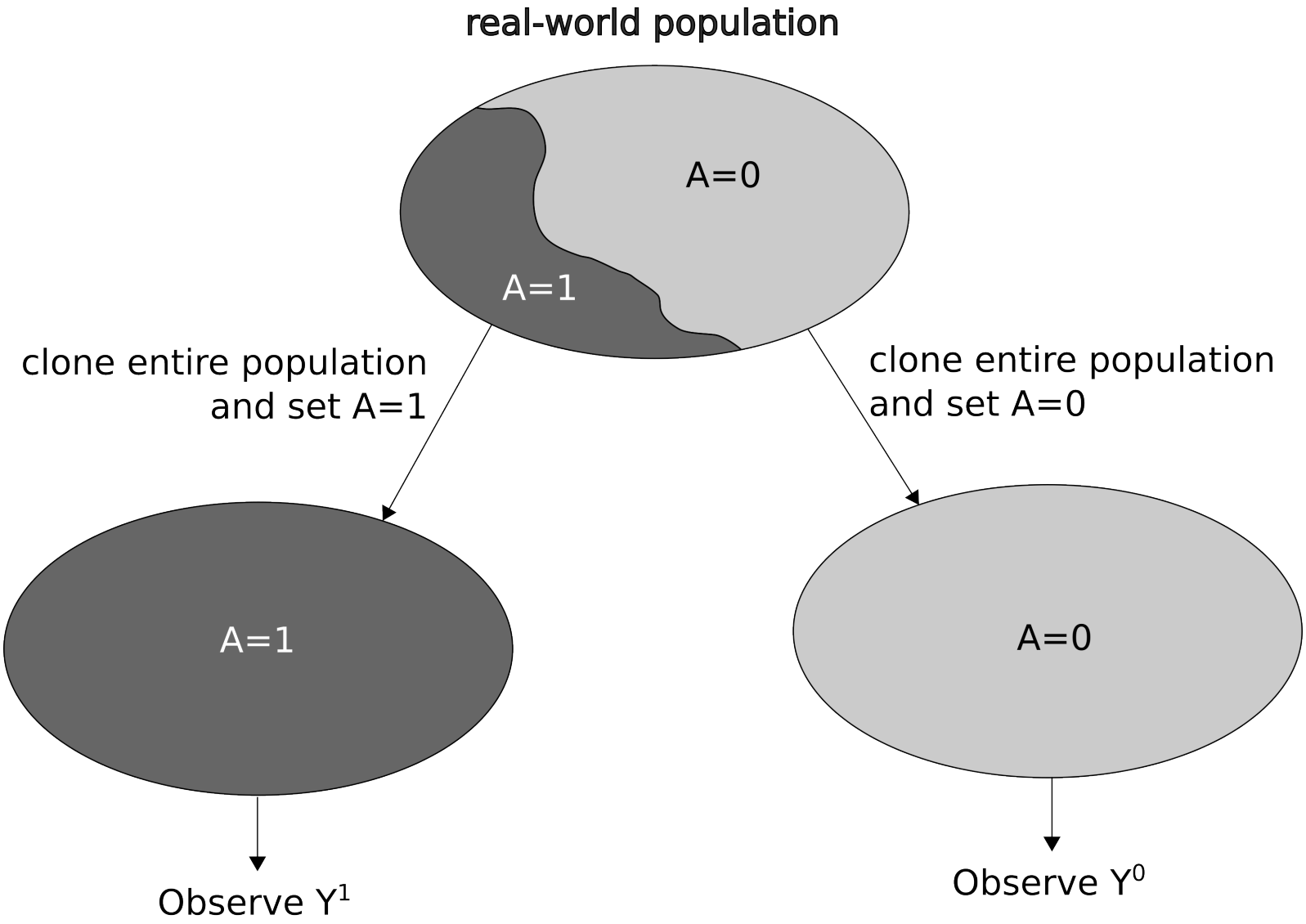


Observed
in nature

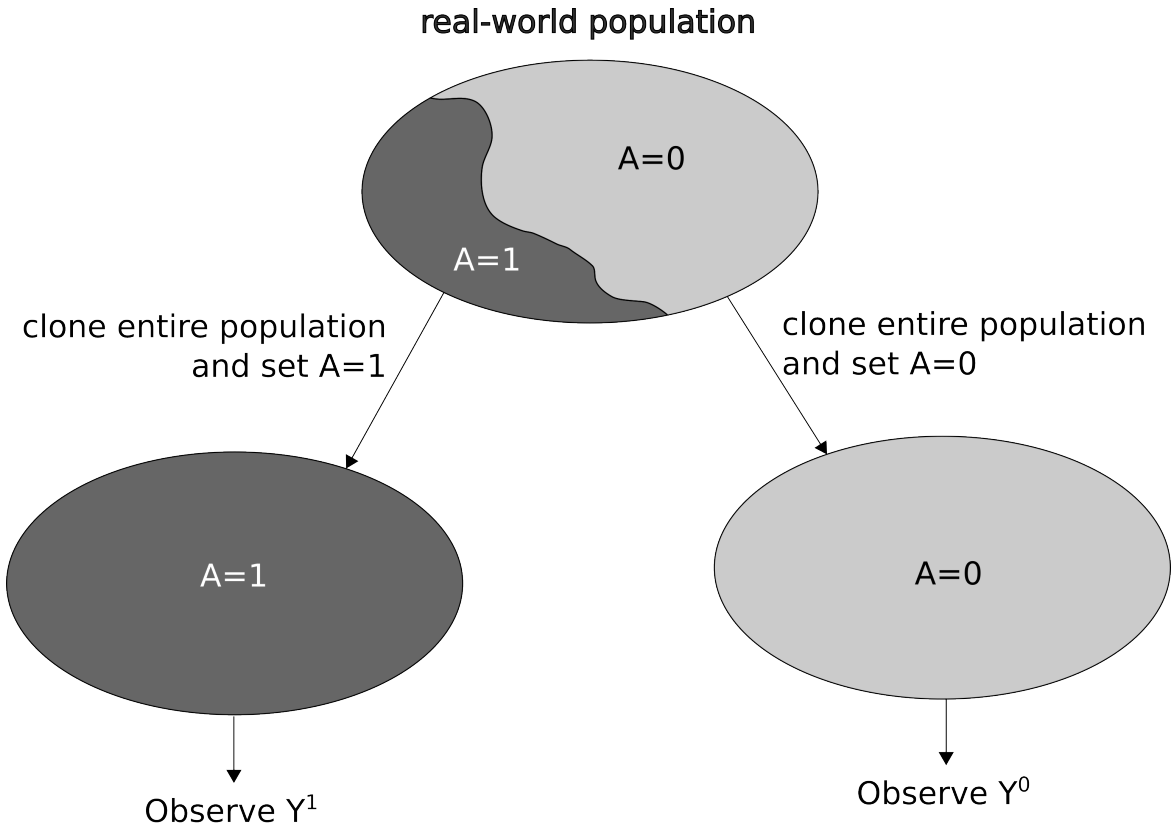


Randomized
experiment

Randomized experiment samples target populations directly



Causal inference with observational data



Problem: Compare these two populations using one observational data set.

Solution strategy depends on whether we observe X or not.

Assumptions when X is observed

In order to study the case when X is observed, we will use several assumptions:

- **Stable Unit Treatment Value Assumption** (**SUTVA**): Each unit's potential outcomes are unaffected by other units' treatment assignments.
- **Ignorability**: For each value of $X = x$, the potential outcomes Y^0 and Y^1 are independent of the observed treatment assignment, A.

$$Y^0 \perp\!\!\!\perp A \mid X = x \quad \text{and} \quad Y^1 \perp\!\!\!\perp A \mid X = x \quad \text{for all } x$$

- **Positivity**: For each value of $X = x$, there is a chance to receive either of the treatment assignments.

$$\Pr(A = 0 \mid X = x) > 0 \quad \text{and} \quad \Pr(A = 1 \mid X = x) > 0 \quad \text{for all } x$$

From assumptions to a causal estimate

From these assumptions we get:

$$\mathrm{E}[Y^1 \mid X = x] = \mathrm{E}[Y \mid X = x, A = 1]$$

Marginalizing over X gives:

$$\sum_x \mathrm{E}[Y \mid X = x, A = 1]\Pr(X = x) = \sum_x \mathrm{E}[Y^1 \mid X = x]\Pr(X = x)$$
$$= \mathrm{E}[Y^1]$$

We can estimate $\mathrm{E}[Y \mid X = x, A = 1]$ and $\Pr(X = x)$ in our data, so we can estimate $\mathrm{E}[Y^1]$.

From assumptions to a causal estimate

Apply the same reasoning to $A = 0$ cases:

$$\sum_x \mathrm{E}[Y \mid X = x, A = 0]\Pr(X = x) = \mathrm{E}[Y^0]$$

Put all the pieces together to estimate the average treatment effect in the population:

$$\mathrm{E}[Y^1 - Y^0] = \mathrm{E}[Y^1] - \mathrm{E}[Y^0]$$

Approach #1 when X is observed: model the outcome

We could try a linear model:

$$\mathrm{E}[Y^a \mid X = x] = \mathrm{E}[Y \mid X = x, A = a] = \beta_0 + \beta_1 x + \beta_2 a$$

This gives β_2 as the average treatment effect in the population.

Problems:

- Trying to solve two problems in one model: design and analysis.
- Are we confident in our model specification?
- If X is high dimensional (β_1 is a vector), we will use many degrees of freedom estimating β_1 when we really only care about β_2.

Approach #2 when X is observed: reweight the sample

The reweighting approach uses the **propensity score**: the probability of assignment to the treatment condition for a given value of X

$$\Pr(A = 1 \mid X = x)$$

You can estimate this from observed data using logistic regression.

Inverse Probability of Treatment Weighting (IPTW): Each unit i in the **treatment condition** $(A = 1)$ with $X = x$ is weighted by the inverse probability of being treated:

$$w_i = \frac{1}{\Pr(A = 1 \mid X = x)}$$

What does IPTW do?

With these weights, the sample with $A = 1$ and $X = x$ is the same size as the entire unweighted sample with $X = x$.

Example: Suppose there are 100 units in the sample with $X = x$, 25 of which are in the treatment condition ($A = 1$). This gives:

$$\Pr(A = 1 \mid X = x) = 25/100 = 1/4$$

and

$$w_i = \frac{1}{\Pr(A = 1 \mid X = x)} = 4$$

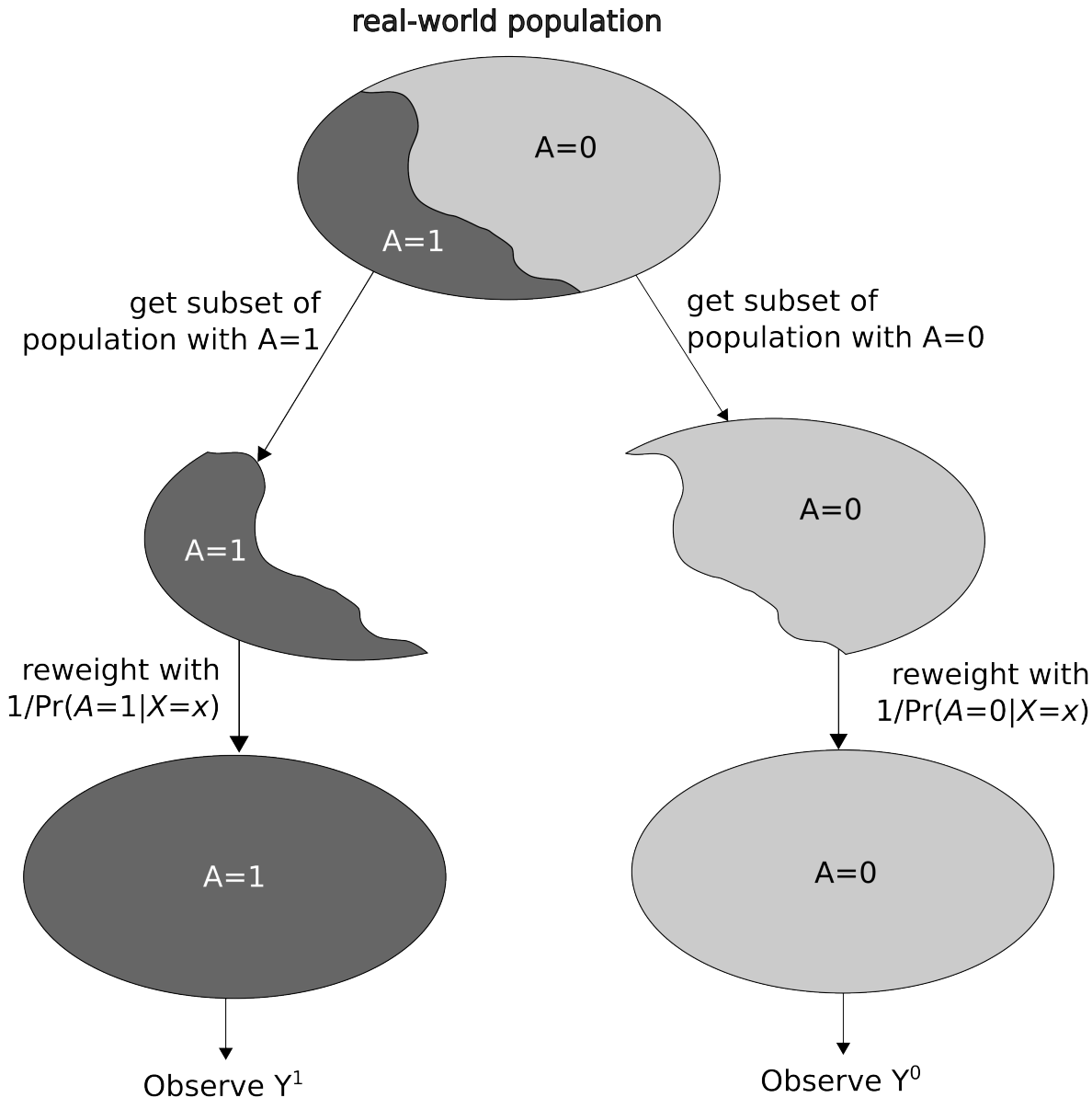
Using a weight of 4, the 25 units with $X = x$ and $A = 1$ now represent all 100 units with $X = x$.

IPTW for the control condition ($A = 0$)

Each unit i in the **control condition** ($A = 0$) with $X = x$ is weighted by the inverse probability of being in the control condition:

$$w_i = \frac{1}{\Pr(A = 0 \mid X = x)}$$

What does IPTW do?



Balance

The weighted distribution of X should look approximately the same in the treatment and control conditions. You can check this by:

- comparing summary statistics (standardized mean difference)
- comparing graphical summaries
- t-tests
- Kolmogorov-Smirnov tests

If the distribution of X is not balanced between treatment and control, revisit your propensity score model specification.

The IPTW estimate of treatment effect

Let Y_i be the observed outcome for unit i, let A_i be their treatment assignment, and let x_i be their observed pre-treatment variable values.

$\mathrm{E}[Y^1]$ and $\mathrm{E}[Y^0]$ are estimated as:

$$\widehat{\mathrm{E}[Y^1]} = \frac{\sum_i A_i w_i Y_i}{\sum_i A_i w_i}$$

$$\widehat{\mathrm{E}[Y^0]} = \frac{\sum_i (1 - A_i) w_i Y_i}{\sum_i (1 - A_i) w_i}$$

The estimate of the average treatment is $\widehat{\mathrm{E}[Y^1]} - \widehat{\mathrm{E}[Y^0]}$.

Extreme weights can be a problem

Very large values of w_i make for noisy estimates of $\widehat{\mathrm{E}[Y^1]}$ or $\widehat{\mathrm{E}[Y^0]}$.

Recall the positivity assumption:

$$\Pr(A = 0 \mid X = x) > 0 \quad \text{and} \quad \Pr(A = 1 \mid X = x) > 0 \quad \text{for all } x$$

A very large w_i means either $\Pr(A = 1 \mid X = x_i) \approx 0$ or $\Pr(A = 0 \mid X = x_i) \approx 0$. This unit had almost no chance of experiencing their treatment assignment.

Remedies:

- Check your propensity score model.
- Drop these units because there is essentially nothing to compare in your observed data for $X = x_i$.
- Truncate the weights.

Combining IPTW with survey weights

Survey data sets like NHANES already have weights based on:

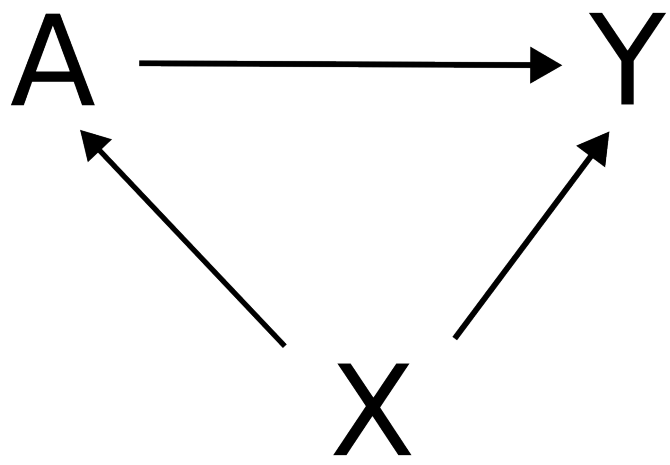
- the probability of being selected for the sample
- the probability of participating in the survey
- post-stratification to match census population estimates

Ridgeway et al. (2018) advise to do both of these:

1. use the survey weights when estimating the propensity score model
2. use the IPT weight multiplied by the survey weight as the final weight for analysis

Ridgeway, G., Kovalchik, S.A., & Griffin, B.A. (2018) Propensity score analysis with survey weighted data. *Journal of Causal Inference*, *3*(2), 237–249.

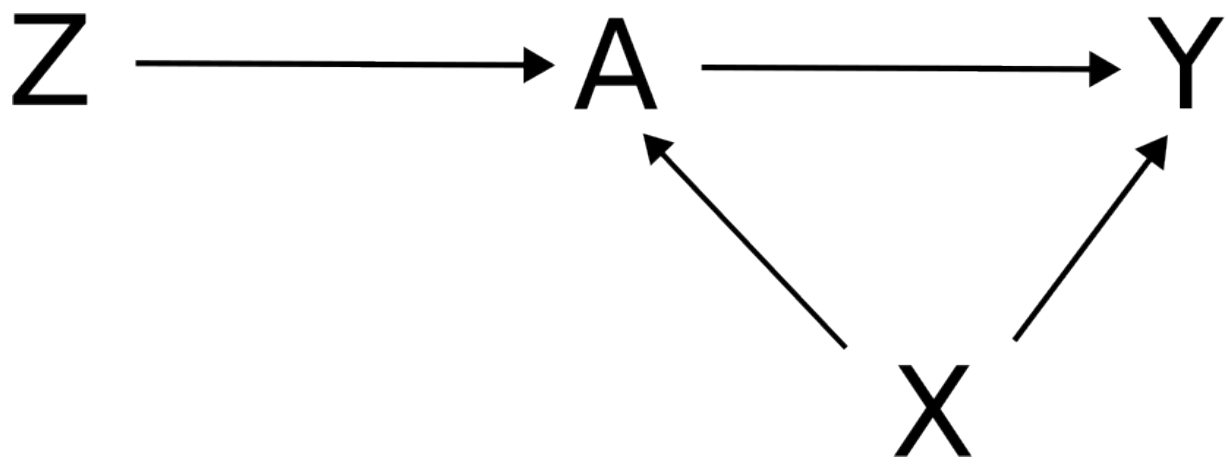
What if X isn't observed?



Popular Methods:

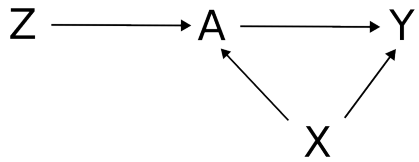
- Regression discontinuity design
- **Instrumental variable**
- Difference-in-difference

Instrumental variable setup



- We don't observe confouding variables X
- We do observe an **instrumental variable** Z
  - Z is independent of confounding variables X
  - Z "encourages" units to be in the treatment condition: $\Pr(A = 1 \mid Z = z)$ increases with z.
  - For a given treatment assignment ($A = 0$ or $A = 1$), the outcome Y is independent of Z.

Instrumental variable examples



Example 1: Mendelian randomization of genes (Z) could put someone at higher risk for a disease diagnosis (A) via a biological process. Genes (Z) are independent of their unobserved health behaviors (X) that also put them at risk for disease (A). Both disease diagnosis (A) and health behaviors (X) influence longevity (Y). The genes associated with the disease (Z) only affect longevity (Y) through disease diagnosis (A).

Example 2: Whether or not a mother smokes cigarettes during pregnancy (A) and the baby's birth weight (Y) could both depend on many unobserved characteristics of the mother (X). The amount of cigarette taxes (Z) influences whether or not a mother smokes (A), but is independent of the mother's characteristics (X) and the baby's birth weight (Y).

Regression in the IV setting

Suppose there is a linear relationship between the outcome Y and the assigned treatment A:

$$Y = \beta_0 + \beta_1 A + \epsilon$$

The unobserved confounder (X) makes the distribution of ϵ dependent on A. This breaks an assumption of **ordinary least squares** regression.

We can use **two-stage least squares** (**2SLS**) regression instead.

Two-stage least squares (2SLS) estimation

Stage 1: First, estimate the effect of the instrumental variable (Z) on observed treatment (A)

$$A = \alpha_0 + \alpha_1 Z + \eta$$

using ordinary least squares. η is the error term. Yes, this is a linear model for a binary outcome.

Next, predict the treatment assignment for each unit i using their observed value of the instrumental variable, z_i:

$$\hat{a}_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i$$

Two-stage least squares (2SLS) estimation

Stage 2: Estimate the effect of the predicted treatment assignment ($\hat{a}_i$) on the outcome (Y)

$$y_i = \beta_0 + \beta_1 \hat{a}_i + \epsilon_i$$

using ordinary least squares. This works because the independent variable ($\hat{a}_i$) is a function of Z which is independent of ϵ.

The causal estimate of the average effect of A on Y is $\hat{\beta}_1$.

2SLS estimation with survey data in R

Suppose you've renamed the variables in your NHANES data set to **z**, **a**, and **y**, corresponding to the notation in these slides. Use *svyivreg* in the *survey* package.

```
library(survey)
nhanes.svyd = svydesign(data=nhanes, id=~SDMVPSU,
    strata=~SDMVSTRA, weights=~wtmec2yr, nest=TRUE)
svyivreg(y~a | z, nhanes.svyd)
```

For further study

- Coursera online courses
  - Entry level: *A Crash Course in Causality: Inferring Causal Effects from Observational Data*
    `https://www.coursera.org/learn/crash-course-in-causality`
  - More technical: *Causal Inference 1* and *2*
    `https://www.coursera.org/learn/causal-inference`
    `https://www.coursera.org/learn/causal-inference-2`
- Textbook: *Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction* by Imbens and Rubin
- Judea Pearl's graph-based approach
  - Entry level: *The Book of Why: the New Science of Cause and Effect* by Pearl and MacKenzie
  - More advanced: *Causality: Models, Reasoning, and Inference* by Pearl
  - See also: *Causal Inference in Statistics: a Primer* by Pearl, Glymour, and Jewell

Homework

1. Your last homework looked at associations observed in nature that might not be interpretable as causal effects. Look at the subpopulation with diabetes (diagnosed and undiagnosed). Take diagnosis (yes/no) as the outcome of interest (Y). Consider each variable that predicts diagnosis as a treatment assignment (A). If the variable is continuous, you can make it into a binary low/high indicator for this exercise. Is it feasible to intervene on this variable in the real world? What are some "pre-treatment" variables (X) in NHANES that could influence both the treatment assignment (A) and diagnosis (Y) within a given treatment assignment? Use IPTW (including survey weights) to get a causal estimate of the effect of the predictor (A) on diagnosis (Y).
2. Document what you did in this boot camp's homeworks. What data did you use and how did you use it? What methods? What were the results? This documentation is invaluable if you have to return to a project later or hand it off to someone else. As you prepare this document, think about what to present and how to present to a non-technical, decision-making audience.