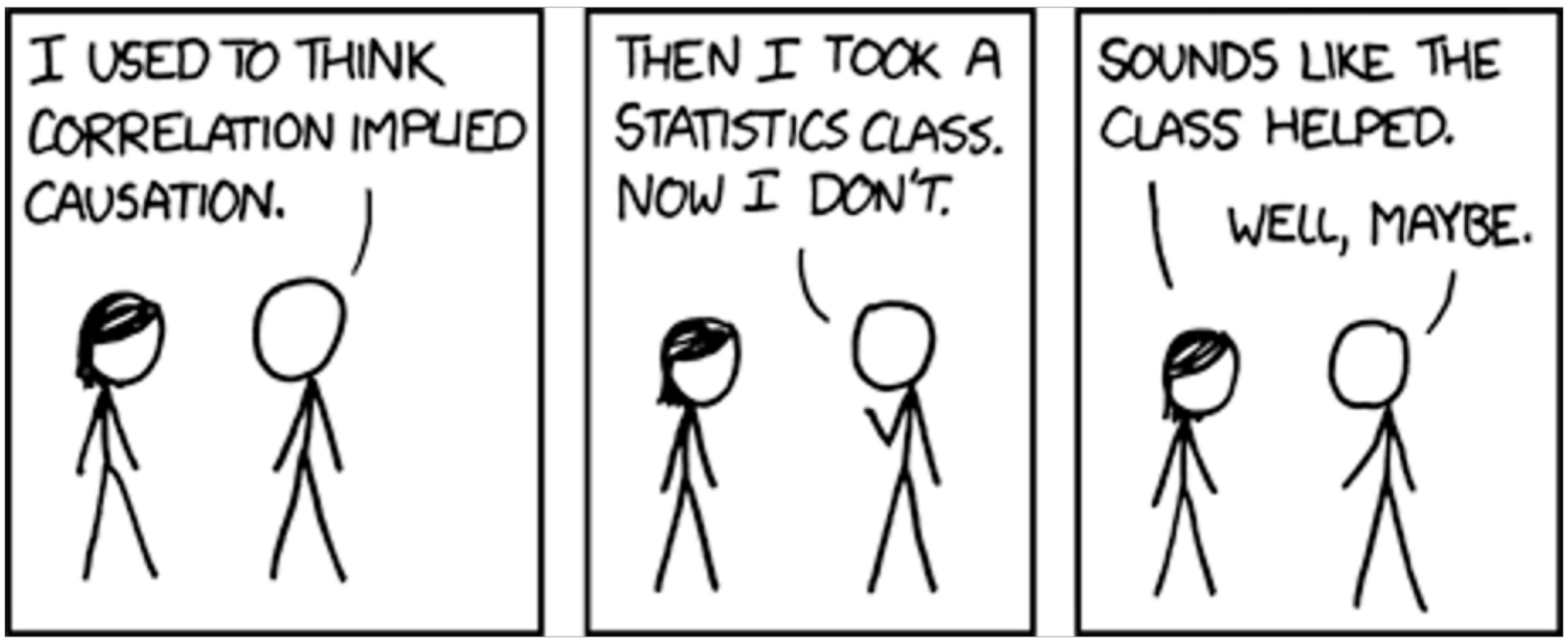
6.3720/6.3722
INTRODUCTION TO STAT DATA ANALYSIS

Lecture 25

Causal inference



Today's agenda

- Setting up the causal question
  - Why is the causal question hard to answer?
  - Common assumptions: When can we estimate causal effect?

Today's agenda

- Setting up the causal question
  - Why is the causal question hard to answer?
  - Common assumptions: When can we estimate causal effect?

- Experimentation: A/B testing
  - Why is randomization the gold standard?
  - How can we use hypothesis testing to assess whether the average treatment effect is significant?

Today's agenda

- Setting up the causal question
  - Why is the causal question hard to answer?
  - Common assumptions: When can we estimate causal effect?

- Experimentation: A/B testing
  - Why is randomization the gold standard?
  - How can we use hypothesis testing to assess whether the average treatment effect is significant?

- Observational studies
  - Covariate adjustment: How can we use regression to estimate the conditional average treatment effect?
  - Covariate matching: How can we use nearest neighbor matching to estimate causal effect?

Why is the causal question hard to answer?

- What is the effect of treatment T on outcome Y?

- Example: If a person in a car accident is taken to the hospital in a helicopter instead of an ambulance, will that improve their chance of survival?

Why is the causal question hard to answer?

- What is the effect of treatment T on outcome Y?

- Example: If a person in a car accident is taken to the hospital in a helicopter instead of an ambulance, will that improve their chance of survival?

	Died	Survived	% Died
Ambulance	260	840	24%
Helicopter	64	136	32%

Why is the causal question hard to answer?

- What is the effect of treatment T on outcome Y?

- Example: If a person in a car accident is taken to the hospital in a helicopter instead of an ambulance, will that improve their chance of survival?

	Died	Survived	% Died
Ambulance	260	840	24%
Helicopter	64	136	32%

- Difference is statistically significant: p-value of .018 from two-sided t-test and .014 from GLRT
- Helicopter transports are more likely to die! Should we recommend taking an ambulance?

Simpson's paradox

- Let's separate car accident victims by severity:

Severe	Died	Survived	% Died
Ambulance	60	40	60%
Helicopter	48	52	48%

Simpson's paradox

- Let's separate car accident victims by severity:

Severe	Died	Survived	% Died
Ambulance	60	40	60%
Helicopter	48	52	48%

Light	Died	Survived	% Died
Ambulance	200	800	20%
Helicopter	16	84	16%

Simpson's paradox

- Let's separate car accident victims by severity:

Severe	Died	Survived	% Died
Ambulance	60	40	60%
Helicopter	48	52	48%

Light	Died	Survived	% Died
Ambulance	200	800	20%
Helicopter	16	84	16%

- In both categories, helicopter transports have better outcomes!
- That's because severity of accident affects both the mode of transportation and the outcome
- We must account for severity of accident to avoid Simpson's paradox when examining causal effect!

Causal question: Revised

- For a person with features X,
  what is the effect of treatment T on outcome Y?



- Example: If a person in a car accident is taken to the hospital in a helicopter instead of an ambulance, will that improve their chance of survival?
- Features: Consciousness, head injury, age, wearing seat belt, car speed

Revised causal question: Facebook friend recommendations

- For a person with features X,
  what is the effect of treatment T on outcome Y?

- Example: If Facebook gives a user more friend recommendations, will the user spend more time on Facebook?

Revised causal question: Facebook friend recommendations

- For a person with features X,
  what is the effect of treatment T on outcome Y?

- Example: If Facebook gives a user more friend recommendations, will the user spend more time on Facebook?
- Observed features: Current number of minutes on Facebook per day, current number of Facebook friends, age

Revised causal question: Facebook friend recommendations

- For a person with features X,
  what is the effect of treatment T on outcome Y?

- Example: If Facebook gives a user more friend recommendations, will the user spend more time on Facebook?
- Observed features: Current number of minutes on Facebook per day, current number of Facebook friends, age
- Challenge #1: Some features that affect the number of recommendations and time spent on Facebook aren't observed by Facebook! (hopefully)
- Examples: Sociableness, number of friends in real life, number of minutes per day chatting with friends in real life

Today's agenda

- Setting up the causal question
  - Why is the causal question hard to answer?
  - Common assumptions: When can we estimate causal effect?

- Experimentation: A/B testing
  - Why is randomization the gold standard?
  - How can we use hypothesis testing to assess whether the average treatment effect is significant?

- Observational studies
  - Covariate adjustment: How can we use regression to estimate the conditional average treatment effect?
  - Covariate matching: How can we use nearest neighbor matching to estimate causal effect?

Conditional average treatment effect

$X = x$

$T = 1$

$Y_1(x) = 1$

$T = 0$

$Y_0(x) = 0$

- Conditional average treatment effect: For a person with features $X = x$, what is the expected difference in outcomes if they were to receive $T = 1$ instead of $T = 0$?

$$CATE(x) = \mathbb{E}\left[Y_1(x) \mid X = x\right] - \mathbb{E}\left[Y_0(x) \mid X = x\right]$$

- Why can't we compute this effect?

Potential outcomes

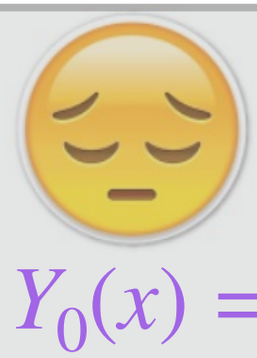
$X = x$

$T = 1$

$Y_1(x) = 1$

Factual outcome is observed

$T = 0$



$Y_0(x) = 0$

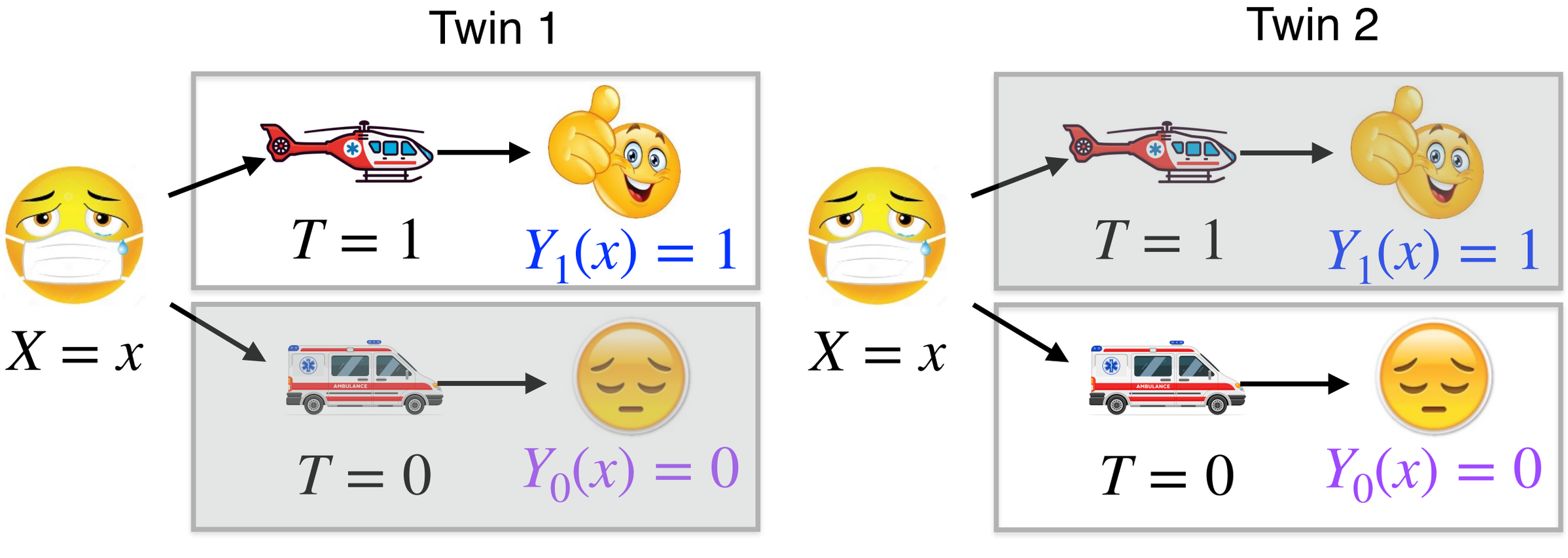
Challenge # 2: We don't observe this counterfactual outcome!

- Both are called potential outcomes. How can we estimate the counterfactual?

$$CATE(x) = \mathbb{E}\left[Y_1(x) \mid X = x\right] - \mathbb{E}\left[Y_0(x) \mid X = x\right]$$

- Fun fact: Proposed by the same Neyman who created Neyman-Pearson lemma and confidence intervals!

Hypothetical twin

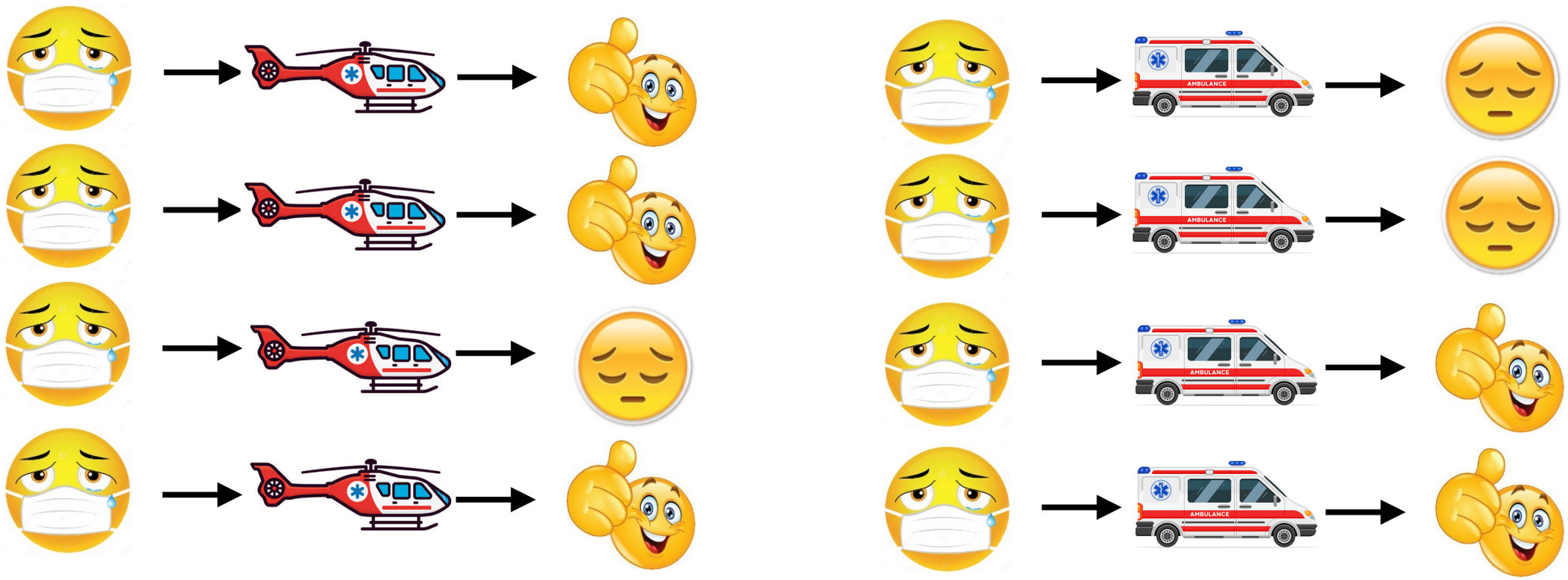


- Estimate CATE using factual outcomes from both twins

$$CATE(x) = \mathbb{E}\left[Y_1(x) \mid X = x\right] - \mathbb{E}\left[Y_0(x) \mid X = x\right]$$

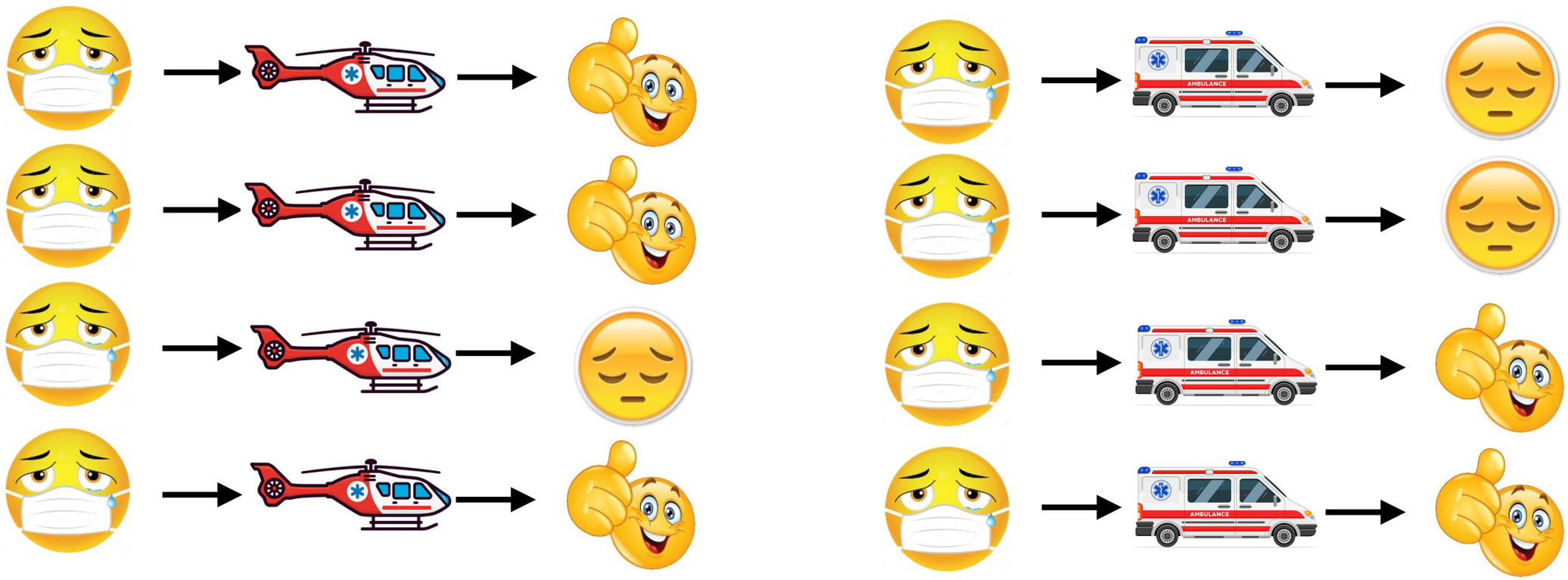
- What if you're a unique person with no twin?

Estimating causal effects from data



- Suppose $\mathbb{E}\left[Y_t(x) \mid X = x\right] = f(x, t)$. Let's estimate f from multiple people.
- Challenge #3: What is the correct specification for f ?

Estimating causal effects from data



- Suppose $\mathbb{E}\left[Y_t(x) \mid X = x\right] = f(x, t)$. Let's estimate f from multiple people.
- Challenge #3: What is the correct specification for f ?
- In reality, we are estimating $\mathbb{E}\left[Y \mid X = x, T = t\right]$ from the observations
- Challenge #4: When can we assume the expected potential outcome is the expected factual outcome?

Summarizing the challenges for causal inference

- Challenge #1: Some features that affect both the treatment and the outcome aren’t observed

Summarizing the challenges for causal inference

- Challenge #1: Some features that affect both the treatment and the outcome aren’t observed
- Challenge #2: Counterfactual outcome is not observed

Summarizing the challenges for causal inference

- Challenge #1: Some features that affect both the treatment and the outcome aren’t observed
- Challenge #2: Counterfactual outcome is not observed
- Challenge #3: True causal model is unknown

Summarizing the challenges for causal inference

- Challenge #1: Some features that affect both the treatment and the outcome aren't observed
- Challenge #2: Counterfactual outcome is not observed
- Challenge #3: True causal model is unknown
- Challenge #4: Assuming $\mathbb{E}\left[Y_t(x) \mid X = x\right] = \mathbb{E}\left[Y \mid X = x, T = t\right]$ might not hold if the treated and control populations have different potential outcomes

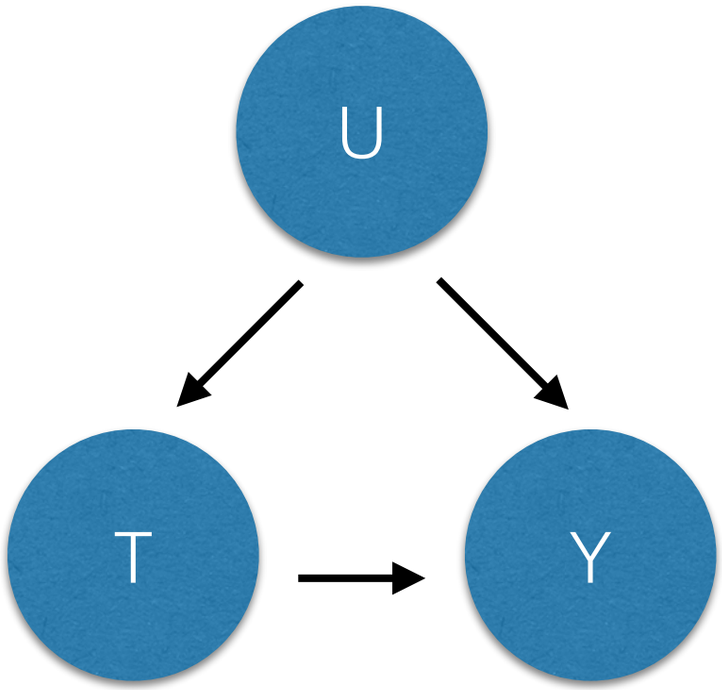
Summarizing the challenges for causal inference

- Challenge #1: Some features that affect both the treatment and the outcome aren't observed
- Challenge #2: Counterfactual outcome is not observed
- Challenge #3: True causal model is unknown
- Challenge #4: Assuming $\mathbb{E}\left[Y_t(x) \mid X = x\right] = \mathbb{E}\left[Y \mid X = x, T = t\right]$ might not hold if the treated and control populations have different potential outcomes

- Address these challenges by designing a study where we can:
  - Assume we observe all relevant features
  - Assume the treated and control populations are similar

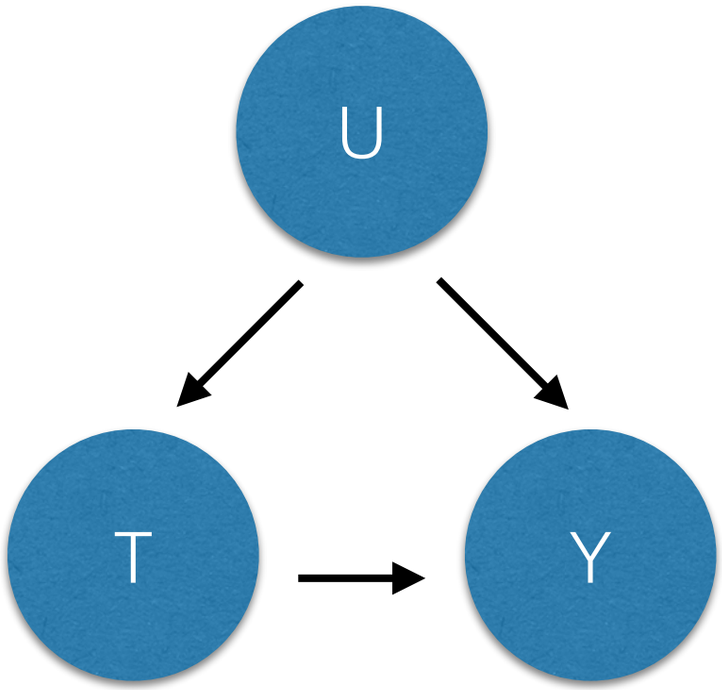
Assumption #1: No unobserved confounding

- Confounder: Factor that affects both treatment T and outcome Y
- Example from prior lecture: Myopia in parents is a confounder that affects night light usage and myopia in children
- Confounder U is often drawn in a causal graph as a parent of both treatment T and outcome Y



Assumption #1: No unobserved confounding

- Confounder: Factor that affects both treatment T and outcome Y
- Example from prior lecture: Myopia in parents is a confounder that affects night light usage and myopia in children
- Confounder U is often drawn in a causal graph as a parent of both treatment T and outcome Y



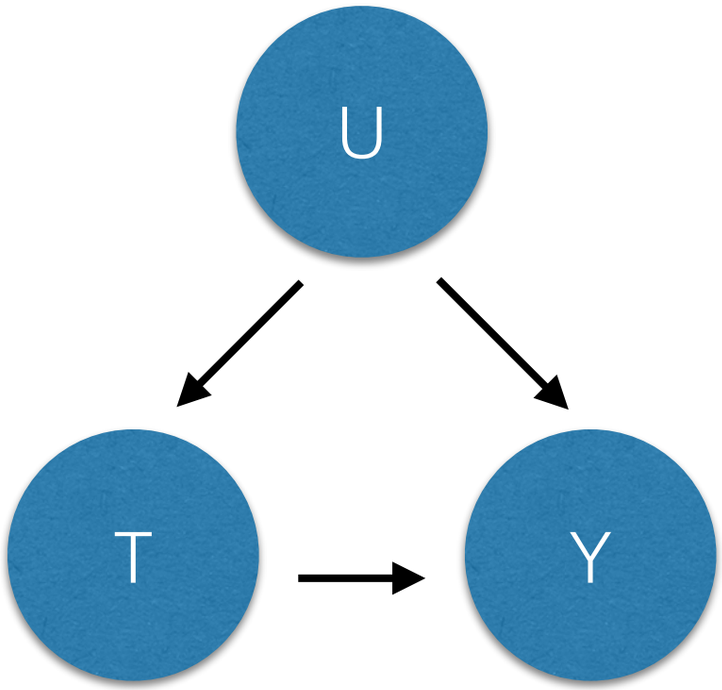
- When there are no confounders:

$$Y_0, Y_1 \perp\!\!\!\perp T \mid X$$

- Known as strong ignorability

Assumption #1: No unobserved confounding

- Confounder: Factor that affects both treatment T and outcome Y
- Example from prior lecture: Myopia in parents is a confounder that affects night light usage and myopia in children
- Confounder U is often drawn in a causal graph as a parent of both treatment T and outcome Y



- When there are no confounders:

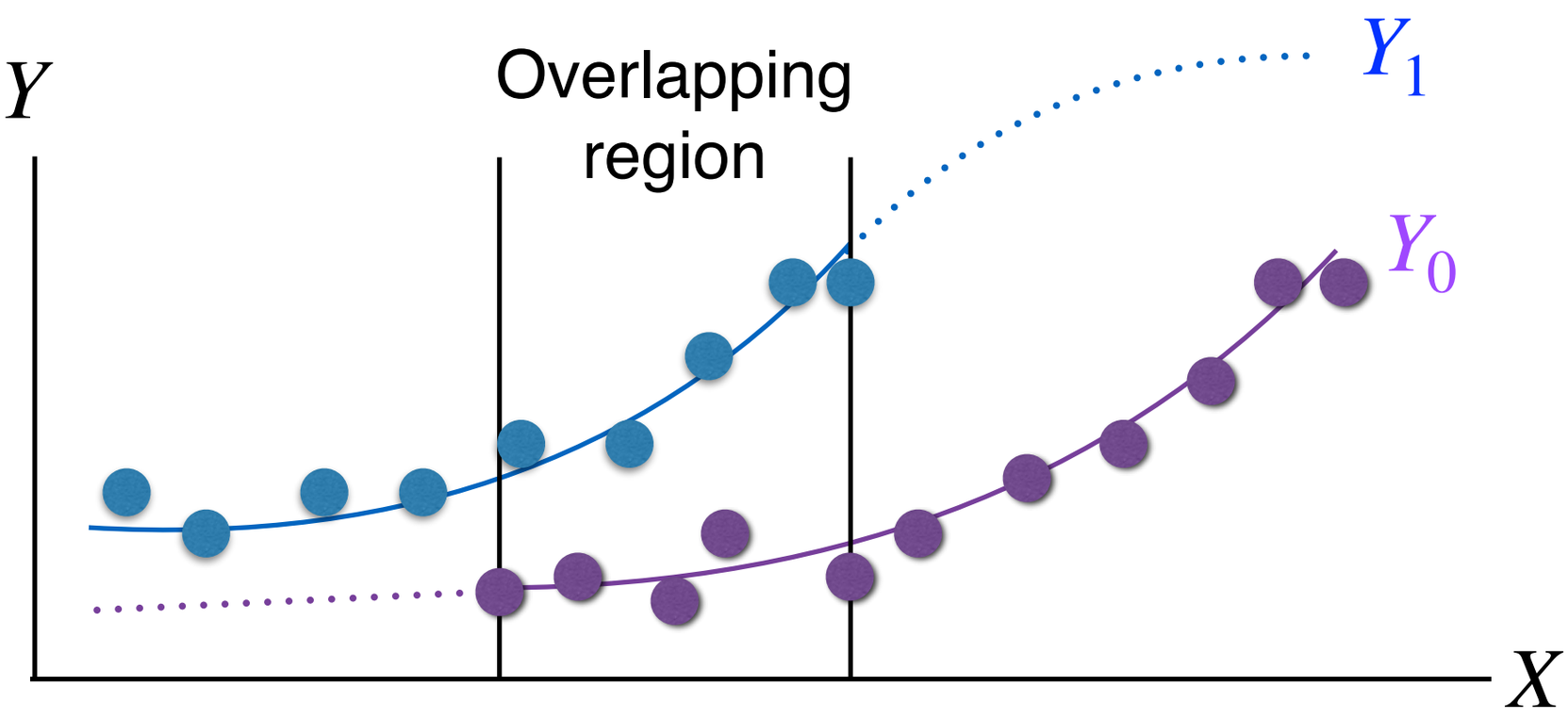
$$Y_0, Y_1 \perp\!\!\!\perp T \mid X$$

- Known as strong ignorability

- Solution:
  - All confounders must be included in feature set X so we have a chance of specifying the correct causal model
  - This will address challenges #1 and #3

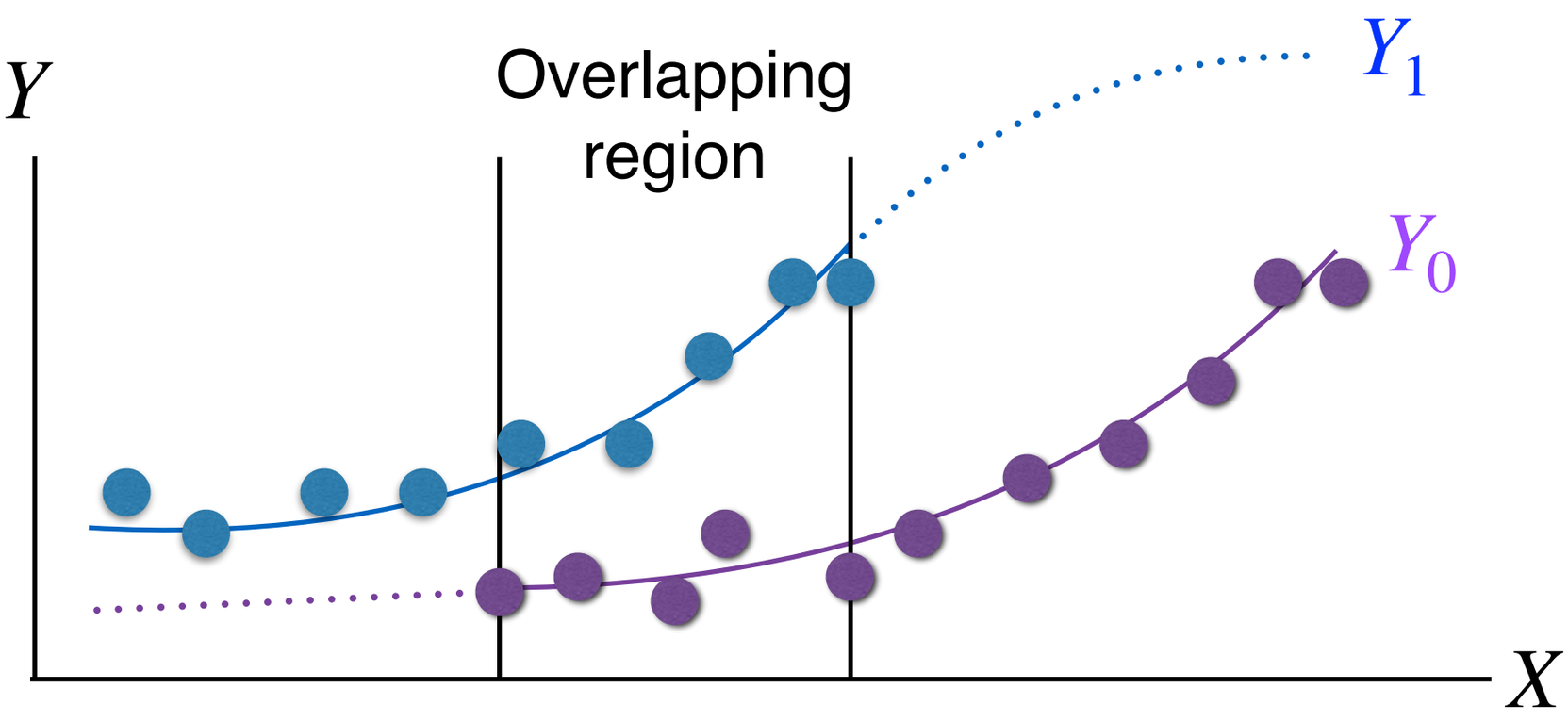
Assumption #2: Overlap between treated and control groups

- Because of challenge #3 (true causal model unknown), can't extrapolate effect estimate to where no people are given that treatment



Assumption #2: Overlap between treated and control groups

- Because of challenge #3 (true causal model unknown), can't extrapolate effect estimate to where no people are given that treatment



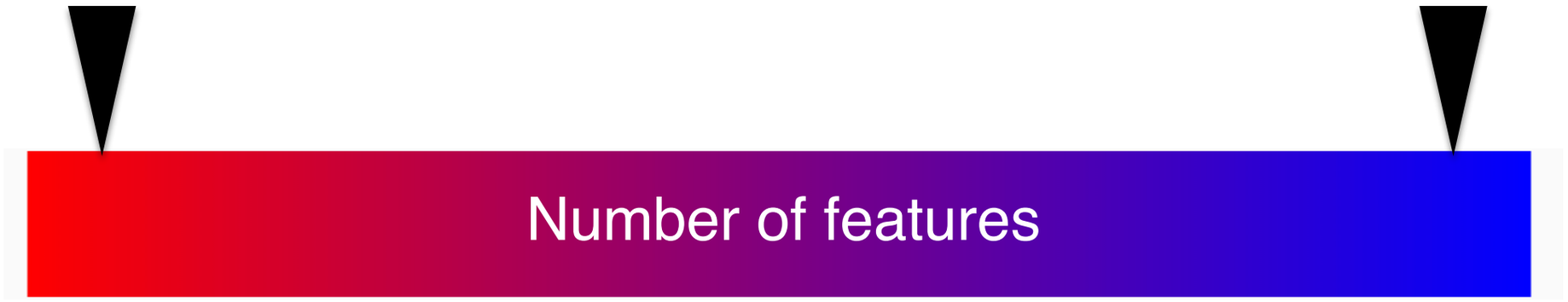
- Solution:
  - Define similarity based on selected features X
  - Report treatment effects only for overlapping population
  - Addresses challenges #2 and #4: Factual outcomes would be counterfactual outcomes for similar people!
  - Also known as positivity or common support

Trade-off between assumptions #1 and #2

- Assumption #1 suggests add as many potential confounders as possible to feature set
- Assumption #2 is easier to satisfy with fewer features: With too many features defining each category, no two patients would be similar!

Few features
- Pro: High overlap
- Con: Missing confounders

Many features
- Con: Little overlap
- Pro: All confounders measured



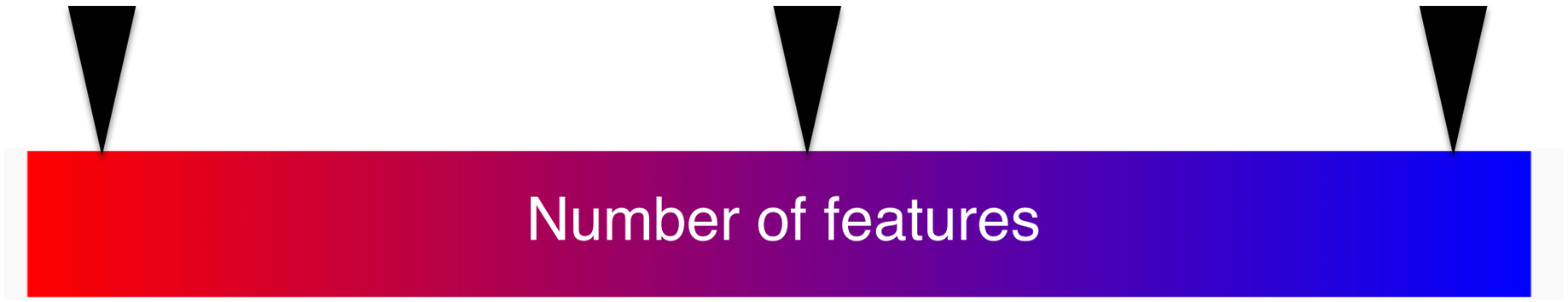
Trade-off between assumptions #1 and #2

- Assumption #1 suggests add as many potential confounders as possible to feature set
- Assumption #2 is easier to satisfy with fewer features: With too many features defining each category, no two patients would be similar!
- Solution:
  - Use domain knowledge to select which features to include

Few features
- Pro: High overlap
- Con: Missing confounders

Select some features to balance the two

Many features
- Con: Little overlap
- Pro: All confounders measured



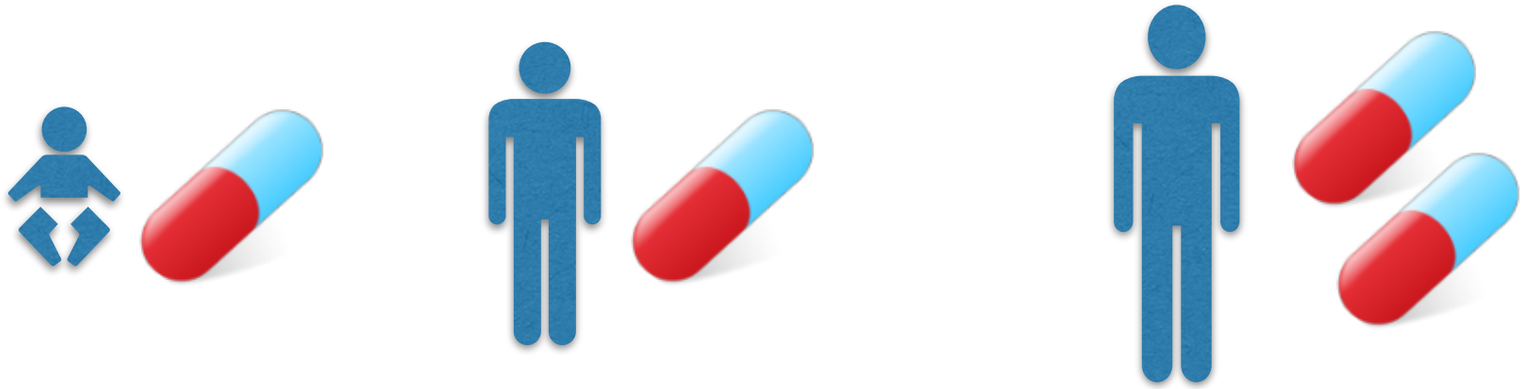
Assumption #3: Stable unit treatment value part 1

- Up until now, we've assumed all treated patients get the same treatment and all control patients get the same treatment
- That's not so obvious!

Assumption #3: Stable unit treatment value part 1

- Up until now, we've assumed all treated patients get the same treatment and all control patients get the same treatment
- That's not so obvious!

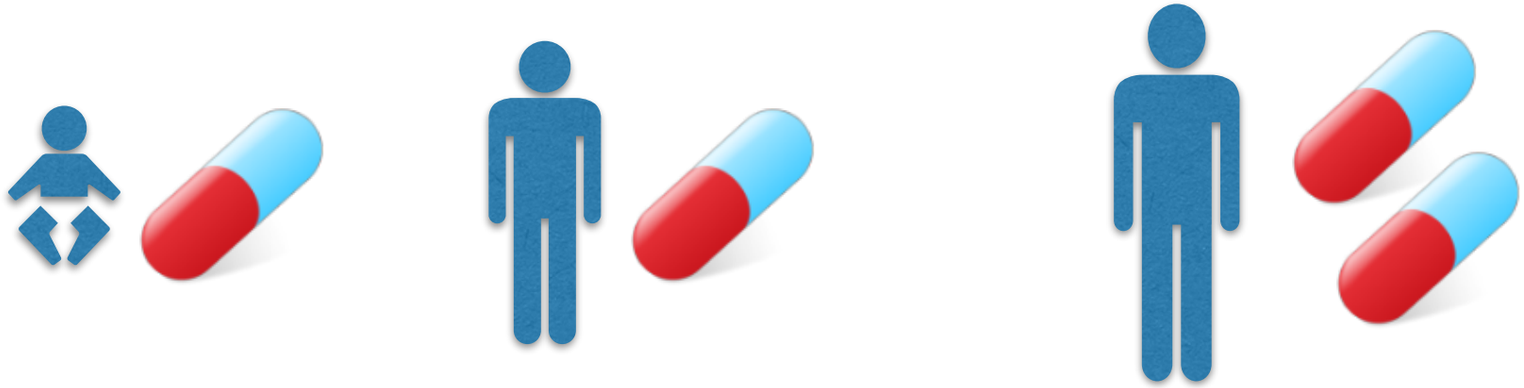
- Are these dosages equivalent?



Assumption #3: Stable unit treatment value part 1

- Up until now, we've assumed all treated patients get the same treatment and all control patients get the same treatment
- That's not so obvious!

- Are these dosages equivalent?

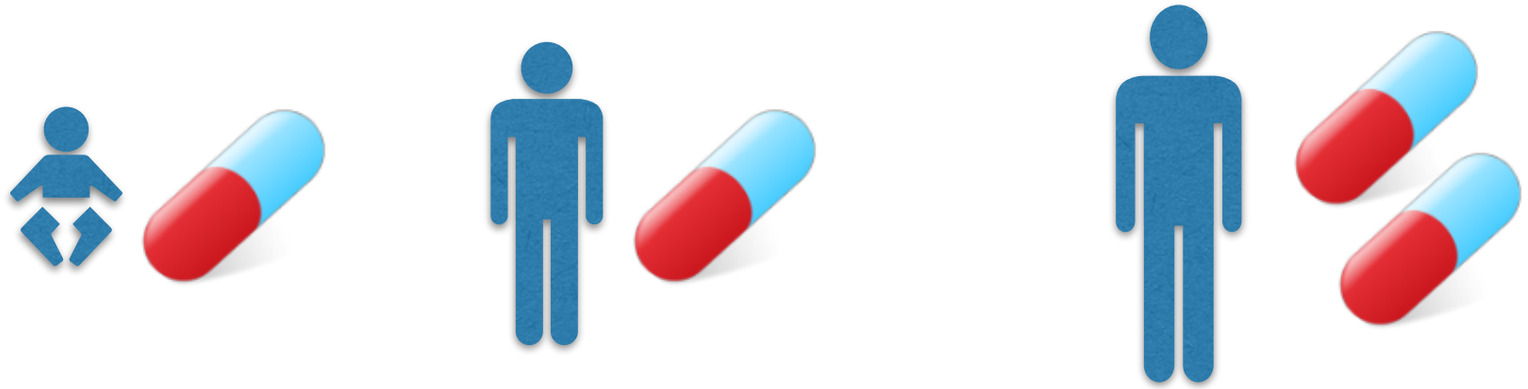


- Are these friend recommendations equivalent?

Assumption #3: Stable unit treatment value part 1

- Up until now, we've assumed all treated patients get the same treatment and all control patients get the same treatment
- That's not so obvious!

- Are these dosages equivalent?

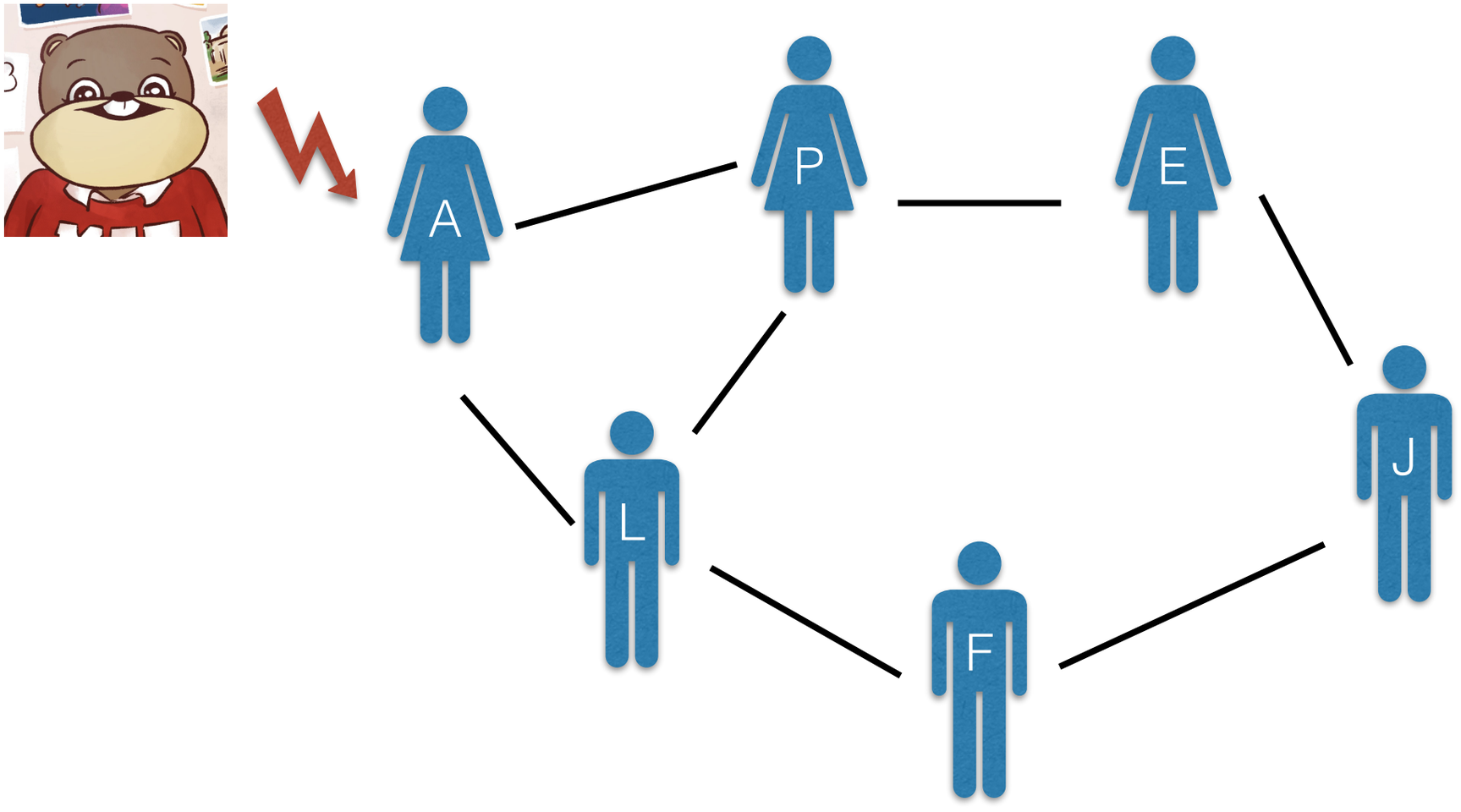


- Are these friend recommendations equivalent?

- Causal consistency: $T = t$ is the same treatment for all patients. If $T = t$ is given, the corresponding potential outcome $\mathbb{E}\left[Y \mid X = x, T = t\right] = Y_t(x)$ is observed

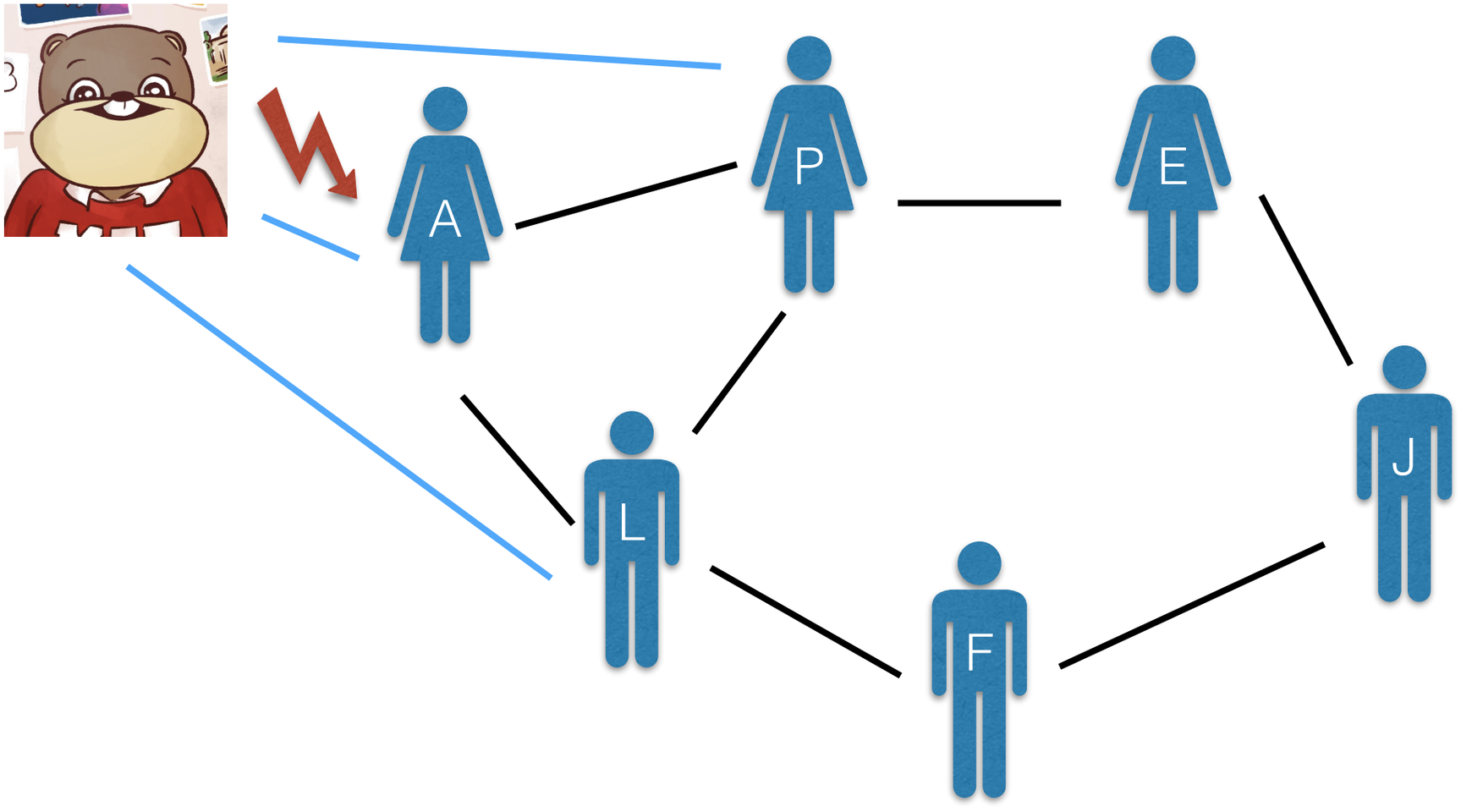
Assumption #3: Stable unit treatment value part 2

- No interference: One person's treatment does not affect another person's outcome
- Violations in Facebook friend recommendation example:
  - User A is recommended to friend user B. User A's treatment affects user B's outcome.
  - User A is recommended a really cool friend. Then User A introduces her new friend to her other friends.



Assumption #3: Stable unit treatment value part 2

- No interference: One person's treatment does not affect another person's outcome
- Violations in Facebook friend recommendation example:
  - User A is recommended to friend user B. User A's treatment affects user B's outcome.
  - User A is recommended a really cool friend. Then User A introduces her new friend to her other friends.



Summarizing the causal assumptions

- Assumption #1: All confounders are included in the feature set.

Summarizing the causal assumptions

- Assumption #1: All confounders are included in the feature set.

- Assumption #2: We are only inferring causal effects where the treated and control cohorts overlap.

Summarizing the causal assumptions

- Assumption #1: All confounders are included in the feature set.
- Assumption #2: We are only inferring causal effects where the treated and control cohorts overlap.
- Assumption #3a: Treatments are consistent for everyone.

Summarizing the causal assumptions

- Assumption #1: All confounders are included in the feature set.
- Assumption #2: We are only inferring causal effects where the treated and control cohorts overlap.
- Assumption #3a: Treatments are consistent for everyone.
- Assumption #3b: One person's treatment does not interfere another person's outcome.

Summarizing the causal assumptions

- Assumption #1: All confounders are included in the feature set.
- Assumption #2: We are only inferring causal effects where the treated and control cohorts overlap.
- Assumption #3a: Treatments are consistent for everyone.
- Assumption #3b: One person's treatment does not interfere another person's outcome.

When designing a study to estimate causal effect, we need to satisfy these assumptions.

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Treatment

- $T = 1$: pass 6.3720
- $T = 0$: never enrolled or attended
- Satisfies causal consistency
- For no interference, let's assume 6.3720 students aren't teaching their friends the course material

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Treatment

- $T = 1$: pass 6.3720
- $T = 0$: never enrolled or attended
- Satisfies causal consistency
- For no interference, let's assume 6.3720 students aren't teaching their friends the course material

Outcome

- Total compensation in first year after obtaining bachelor's degree

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Treatment

- $T = 1$: pass 6.3720
- $T = 0$: never enrolled or attended
- Satisfies causal consistency
- For no interference, let's assume 6.3720 students aren't teaching their friends the course material

Outcome

- Total compensation in first year after obtaining bachelor's degree

- What features would you include to satisfy no unobserved confounding?

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Treatment

- $T = 1$: pass 6.3720
- $T = 0$: never enrolled or attended
- Satisfies causal consistency
- For no interference, let's assume 6.3720 students aren't teaching their friends the course material

Outcome

- Total compensation in first year after obtaining bachelor's degree

- What features would you include to satisfy no unobserved confounding?

Features

- Major
- Prior coursework
- Prior internships

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Treatment

- $T = 1$: pass 6.3720
- $T = 0$: never enrolled or attended
- Satisfies causal consistency
- For no interference, let's assume 6.3720 students aren't teaching their friends the course material

Outcome

- Total compensation in first year after obtaining bachelor's degree

- What features would you include to satisfy no unobserved confounding?

Features

- Major
- Prior coursework
- Prior internships

- How would you define your cohort to ensure overlap?

Example of causal set-up: Effect of taking 6.3720 on your salary

- Causal question: What is the effect of taking 6.3720 on a student's first post-graduation salary?

Treatment

- $T = 1$: pass 6.3720
- $T = 0$: never enrolled or attended
- Satisfies causal consistency
- For no interference, let's assume 6.3720 students aren't teaching their friends the course material

Outcome

- Total compensation in first year after obtaining bachelor's degree

- What features would you include to satisfy no unobserved confounding?

Features

- Major
- Prior coursework
- Prior internships

- How would you define your cohort to ensure overlap?

Cohort

- Restrict to course 6 students so students who take 6.3720 and don't take 6.3720 are more

Goal of causal inference

- Under these assumptions, estimate the conditional average treatment effect for a person with particular features X:

$$CATE(x) = \mathbb{E}\left[Y_1(x) \mid X = x\right] - \mathbb{E}\left[Y_0(x) \mid X = x\right]$$

Goal of causal inference

- Under these assumptions, estimate the conditional average treatment effect for a person with particular features X:

$$CATE(x) = \mathbb{E}\left[Y_1(x) \mid X = x\right] - \mathbb{E}\left[Y_0(x) \mid X = x\right]$$

- We might also be interested in the average treatment effect across a cohort with feature distribution $\mathbb{P}$:

$$ATE = \mathbb{E}_{X \sim \mathbb{P}}\left[\mathbb{E}\left[Y_1(X) \mid X\right] - \mathbb{E}\left[Y_0(X) \mid X\right]\right]$$

Today's agenda

- Setting up the causal question
  - Why is the causal question hard to answer?
  - Common assumptions: When can we estimate causal effect?

- Experimentation: A/B testing
  - Why is randomization the gold standard?
  - How can we use hypothesis testing to assess whether the average treatment effect is significant?

- Observational studies
  - Covariate adjustment: How can we use regression to estimate the conditional average treatment effect?
  - Covariate matching: How can we use nearest neighbor matching to estimate causal effect?

Goal of experimentation

- Causal question: What is the average effect of a treatment in a cohort? Is this effect statistically significant?

Goal of experimentation

- Causal question: What is the average effect of a treatment in a cohort? Is this effect statistically significant?
- How an experiment works:
  - We determine which treatment each person receives
  - Then we measure the difference between the average outcomes in the treated and control groups

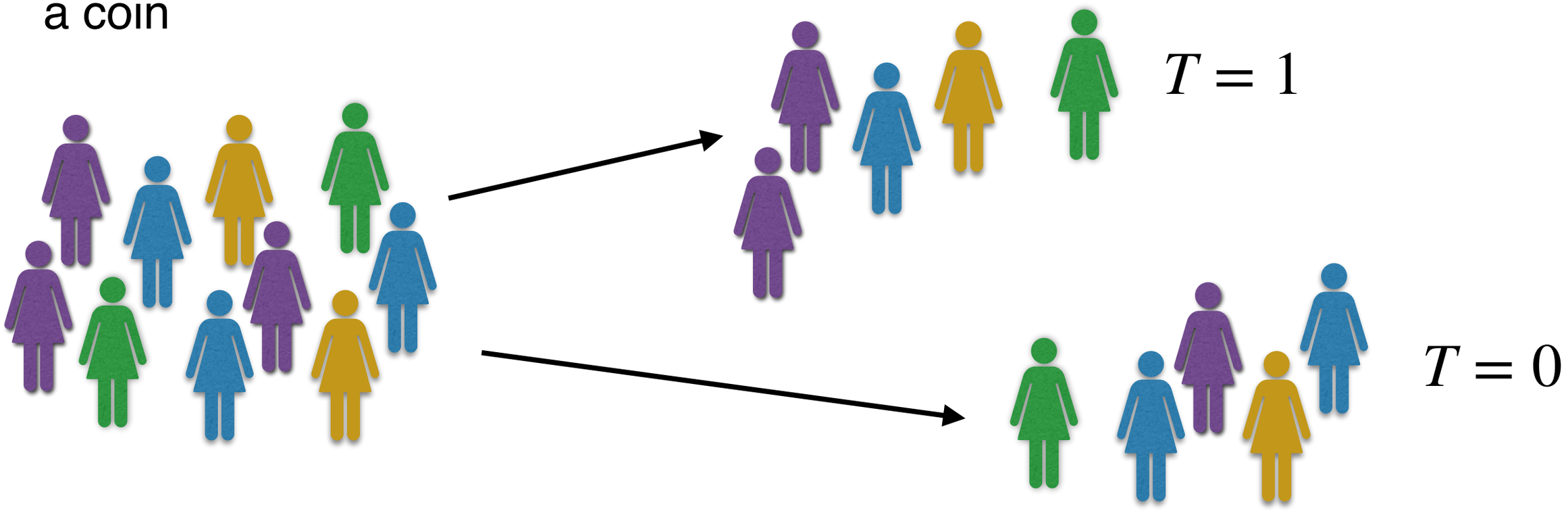
Goal of experimentation

- Causal question: What is the average effect of a treatment in a cohort? Is this effect statistically significant?
- How an experiment works:
  - We determine which treatment each person receives
  - Then we measure the difference between the average outcomes in the treated and control groups

- Example: Clinical trials
  - $T = 1$: Randomly assign some patients to receive new treatment
  - $T = 0$: Randomly assign some patients to receive placebo / standard of care
  - Measure differences in outcomes between the two groups
  - Significant difference required for approval of new treatment

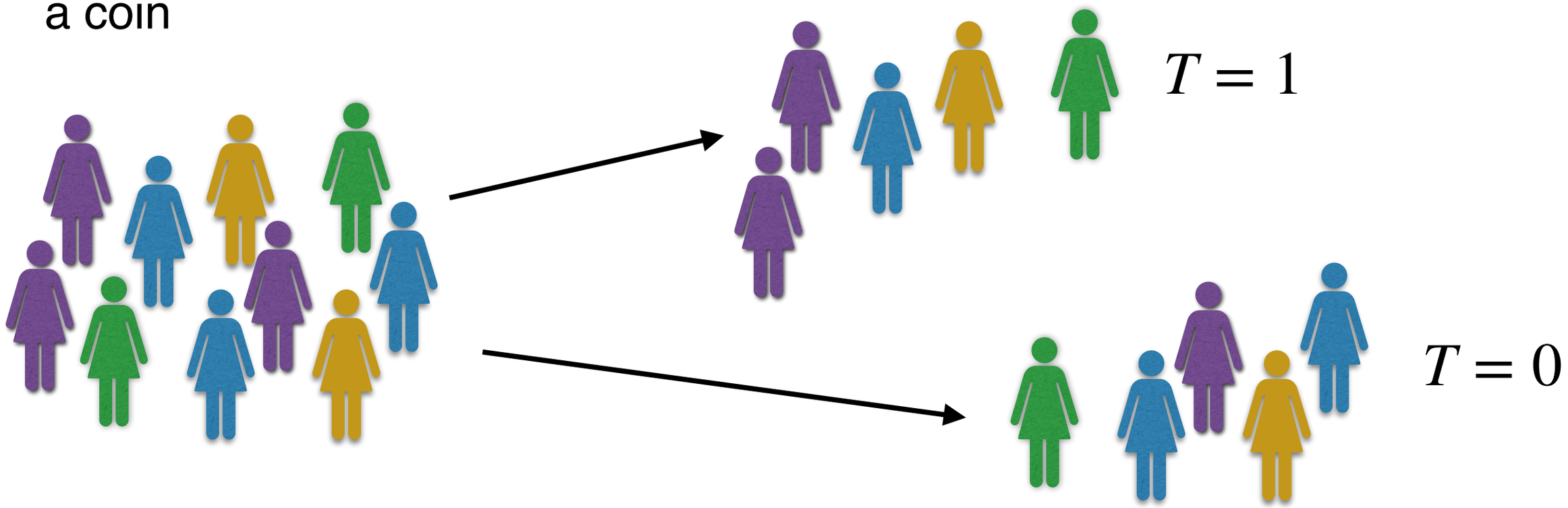
Randomization satisfies causal assumptions

- Randomization: Every person is assigned $T = 1$ or $T = 0$ by flipping a coin



Randomization satisfies causal assumptions

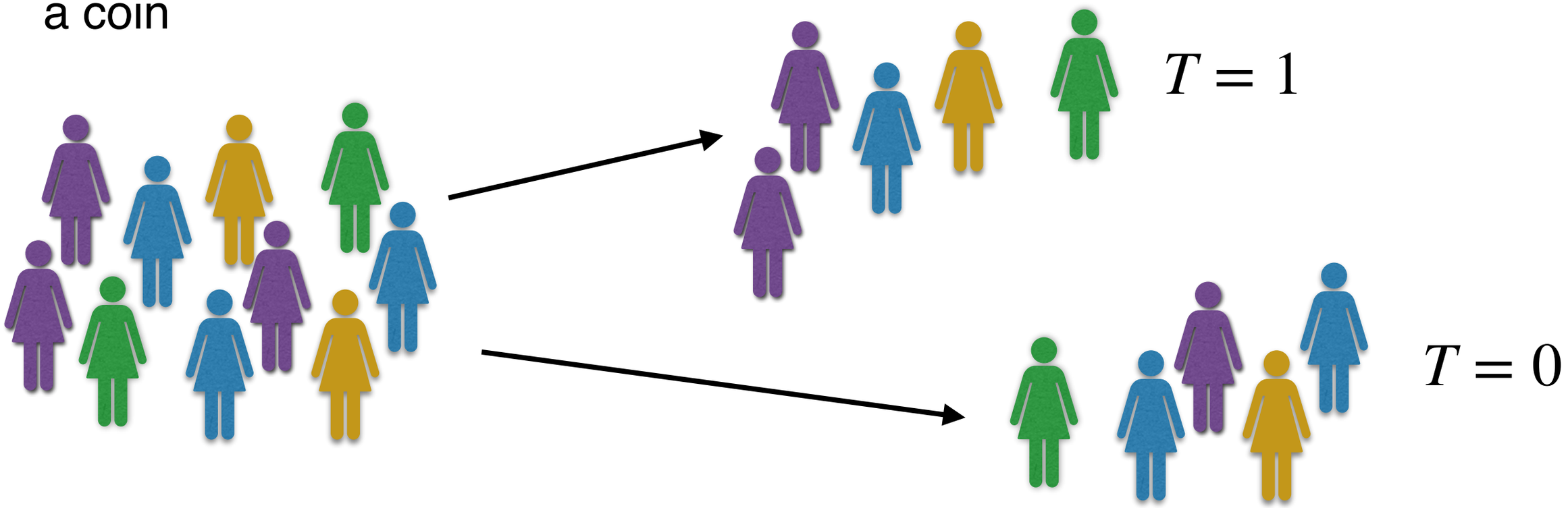
- Randomization: Every person is assigned $T = 1$ or $T = 0$ by flipping a coin



- Satisfies causal assumptions:
  - No unobserved confounding (assumption #1): Patient features have no impact on treatment!
  - Overlap between treated and control (assumption #2): Same feature distributions expected in both groups

Randomization satisfies causal assumptions

- Randomization: Every person is assigned $T = 1$ or $T = 0$ by flipping a coin



- Satisfies causal assumptions:
  - No unobserved confounding (assumption #1): Patient features have no impact on treatment!
  - Overlap between treated and control (assumption #2): Same feature distributions expected in both groups

- Features are only used to define the cohort we are computing the average treatment effect for

Example of experiment: Effect of Netflix recommendations

- Causal question: What is the effect of recommending continuing a show versus starting a new show on time spent on Netflix?

- How would you define the treatments?
  - Criteria: Treatments are *consistent* across all users

Example of experiment: Effect of Netflix recommendations

- Causal question: What is the effect of recommending continuing a show versus starting a new show on time spent on Netflix?

- How would you define the treatments?
  - Criteria: Treatments are *consistent* across all users
  - $T = 0$ Recommend continuing: Show 5 most recent shows at the top of the home screen every time a user opens Netflix for 1 week
  - $T = 1$ Recommend new show: Show top 5 unseen recommendations at the top of the home screen every time a user opens Netflix for 1 week

Example of experiment: Effect of Netflix recommendations

- Causal question: What is the effect of recommending continuing a show versus starting a new show on time spent on Netflix?

- How would you define the treatments?
  - Criteria: Treatments are *consistent* across all users
  - $T = 0$ Recommend continuing: Show 5 most recent shows at the top of the home screen every time a user opens Netflix for 1 week
  - $T = 1$ Recommend new show: Show top 5 unseen recommendations at the top of the home screen every time a user opens Netflix for 1 week

- How would you measure the outcome?
  - Criteria: Start measuring outcome *after* treatment starts

Example of experiment: Effect of Netflix recommendations

- Causal question: What is the effect of recommending continuing a show versus starting a new show on time spent on Netflix?

- How would you define the treatments?
  - Criteria: Treatments are *consistent* across all users
  - $T = 0$ Recommend continuing: Show 5 most recent shows at the top of the home screen every time a user opens Netflix for 1 week
  - $T = 1$ Recommend new show: Show top 5 unseen recommendations at the top of the home screen every time a user opens Netflix for 1 week

- How would you measure the outcome?
  - Criteria: Start measuring outcome *after* treatment starts
  - Y: Number of minutes watched that week

Example of experiment: Effect of Netflix recommendations

- Causal question: What is the effect of recommending continuing a show versus starting a new show on time spent on Netflix?

- How would you define the treatments?
  - Criteria: Treatments are *consistent* across all users
  - $T = 0$ Recommend continuing: Show 5 most recent shows at the top of the home screen every time a user opens Netflix for 1 week
  - $T = 1$ Recommend new show: Show top 5 unseen recommendations at the top of the home screen every time a user opens Netflix for 1 week

- How would you measure the outcome?
  - Criteria: Start measuring outcome *after* treatment starts
  - Y: Number of minutes watched that week

- What restrictions do we want to place on the cohort?
  - Criteria: 1) Treatments must be possible. 2) Who do we expect to see an effect for?

Example of experiment: Effect of Netflix recommendations

- Causal question: What is the effect of recommending continuing a show versus starting a new show on time spent on Netflix?

- How would you define the treatments?
  - Criteria: Treatments are *consistent* across all users
  - $T = 0$ Recommend continuing: Show 5 most recent shows at the top of the home screen every time a user opens Netflix for 1 week
  - $T = 1$ Recommend new show: Show top 5 unseen recommendations at the top of the home screen every time a user opens Netflix for 1 week

- How would you measure the outcome?
  - Criteria: Start measuring outcome *after* treatment starts
  - Y: Number of minutes watched that week

- What restrictions do we want to place on the cohort?
  - Criteria: 1) Treatments must be possible. 2) Who do we expect to see an effect for?
  - Only include viewers who have at least 5 ongoing shows and opened Netflix at least once that week

A/B test: Are the experiment results statistically significant?

- Suppose these were the results of the Netflix experiment

Treatment group	# people	Mean minutes watched over week (Y)	Standard deviation (minutes)
Continuing recs (T = 0)	1000	240	30
New recs (T = 1)	1000	300	60

- How do we know if this difference is significant?

A/B test: Are the experiment results statistically significant?

- Suppose these were the results of the Netflix experiment

Treatment group	# people	Mean minutes watched over week (Y)	Standard deviation (minutes)
Continuing recs (T = 0)	1000	240	30
New recs (T = 1)	1000	300	60

- How do we know if this difference is significant?
  - Use a two-sample t-test:
    - H: No difference in outcomes between the two groups
    - K: Treated group has better average outcome

$$t = \frac{300 - 240}{\sqrt{\frac{30^2}{1000} + \frac{60^2}{1000}}} \approx 28.3,\ p < .00001$$

- Called an A/B test: Default treatment is A, and new treatment is B.

A/B test example: Herceptin clinical trial

- Does adding Herceptin to chemotherapy result in more tumor reduction for breast cancer patients?
- Actual clinical trial results:

Treatment group	Reduction (Y = 1)	No reduction (Y = 0)
Herceptin + chemo (T = 1)	45	190
Chemo (T = 0)	29	205

- Based on these results, should Herceptin be approved?

A/B test example: Herceptin clinical trial

- Does adding Herceptin to chemotherapy result in more tumor reduction for breast cancer patients?
- Actual clinical trial results:

Treatment group	Reduction (Y = 1)	No reduction (Y = 0)
Herceptin + chemo (T = 1)	45	190
Chemo (T = 0)	29	205

$$\hat{\mu}_H \approx .1915$$

$$\hat{\mu}_C \approx .1239$$

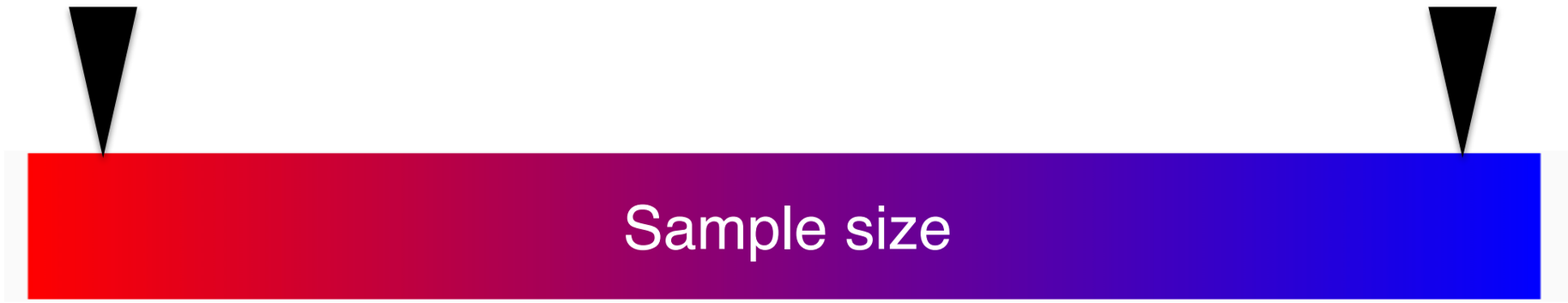
- Based on these results, should Herceptin be approved? Yes!

$$t = \frac{\hat{\mu}_H - \hat{\mu}_C}{\sqrt{\frac{\hat{\sigma}_H}{n_H} + \frac{\hat{\sigma}_C}{n_C}}} \approx 2.01, \quad p \approx .02$$

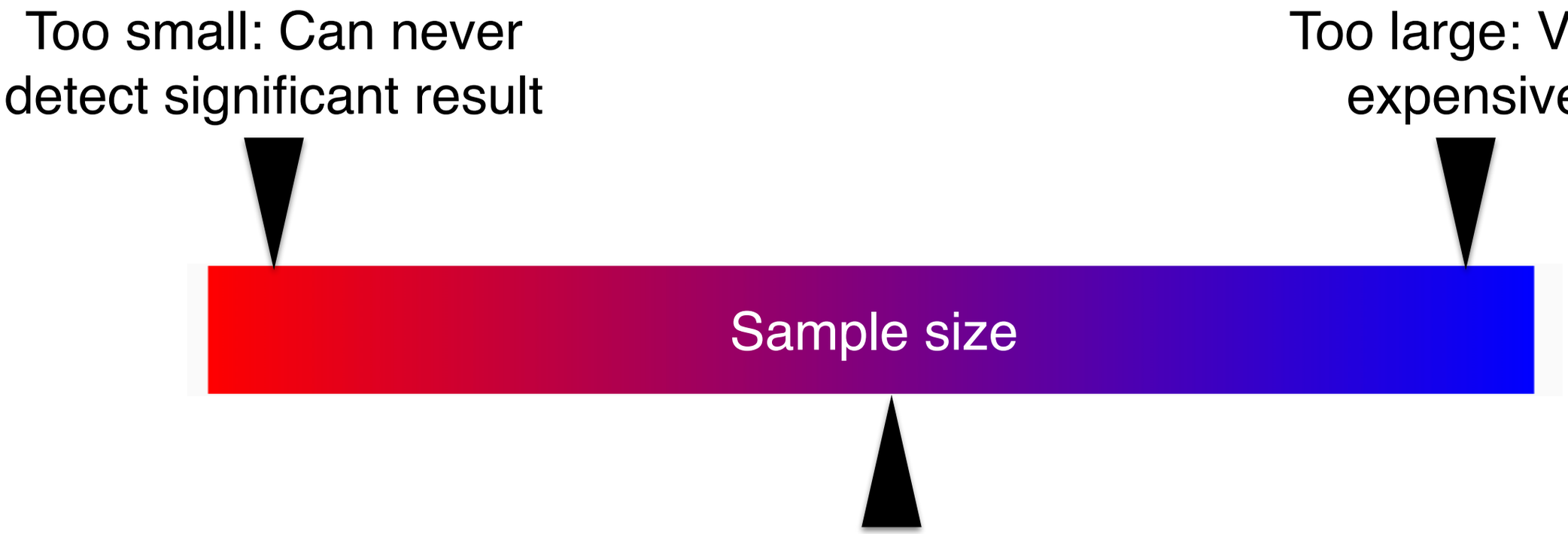
Determining the experiment sample size

Too small: Can never detect significant result

Too large: Very expensive

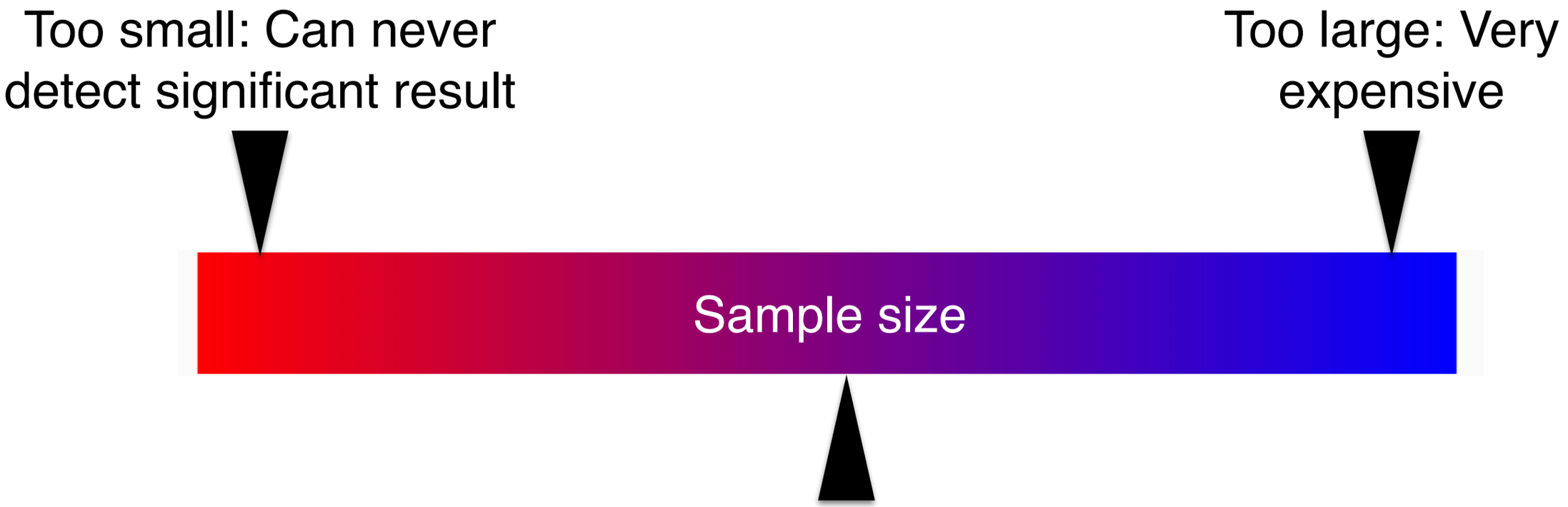


Determining the experiment sample size



- How do we set the sample size before running the experiment?
  - Based on desired significance threshold and power for the test
  - Need an estimated effect size and standard deviation based on prior knowledge
  - See notes for derivation and example

Determining the experiment sample size



- How do we set the sample size before running the experiment?
  - Based on desired significance threshold and power for the test
  - Need an estimated effect size and standard deviation based on prior knowledge
  - See notes for derivation and example
- If calculated sample size is too large, consider secondary outcomes with larger differences

Caveats about experiments

- Non-compliance to treatments:
  - $T = 1 \rightarrow T = 0$: Treatment causes intolerable side effects or patient forgets to take treatment
  - $T = 0 \rightarrow T = 1$: Patient gets treatment from another source
  - Even if actual treatment is recorded, no longer randomized
  - Conduct an observational study instead: Intention to treat can be an instrument for actual treatment

Caveats about experiments

- Non-compliance to treatments:
  - $T = 1 \rightarrow T = 0$: Treatment causes intolerable side effects or patient forgets to take treatment
  - $T = 0 \rightarrow T = 1$: Patient gets treatment from another source
  - Even if actual treatment is recorded, no longer randomized
  - Conduct an observational study instead: Intention to treat can be an instrument for actual treatment

- Non-random withdrawals:
  - More patients on new treatment experience significant side effects and are lost to follow up
  - Observed outcomes are biased
  - Experiment may need to be rerun with a better design

Today's agenda

- Setting up the causal question
  - Why is the causal question hard to answer?
  - Common assumptions: When can we estimate causal effect?

- Experimentation: A/B testing
  - Why is randomization the gold standard?
  - How can we use hypothesis testing to assess whether the average treatment effect is significant?

- Observational studies
  - Covariate adjustment: How can we use regression to estimate the conditional average treatment effect?
  - Covariate matching: How can we use nearest neighbor matching to estimate causal effect?

Motivation for observational studies

- If randomization satisfies the causal assumptions, why not do it all the time?
  - Experiments are expensive to run
  - Some interventions are unethical or unsafe
  - Companies may be wary of testing new ideas that might lose business

Motivation for observational studies

- If randomization satisfies the causal assumptions, why not do it all the time?
  - Experiments are expensive to run
  - Some interventions are unethical or unsafe
  - Companies may be wary of testing new ideas that might lose business
- Example of unethical experiment for evaluating effect of taking 6.3720 on 1st post-graduation salary:
  - $T = 1$: Randomly require some students take 6.3720
  - $T = 0$: Randomly prohibit some students from taking 6.3720
  - Measure differences in salaries between the two groups
  - Not feasible to prohibit students from taking the class!

Motivation for observational studies

- If randomization satisfies the causal assumptions, why not do it all the time?
  - Experiments are expensive to run
  - Some interventions are unethical or unsafe
  - Companies may be wary of testing new ideas that might lose business
- Example of unethical experiment for evaluating effect of taking 6.3720 on 1st post-graduation salary:
  - $T = 1$: Randomly require some students take 6.3720
  - $T = 0$: Randomly prohibit some students from taking 6.3720
  - Measure differences in salaries between the two groups
  - Not feasible to prohibit students from taking the class!
- A large amount of observational data is available from everyday activities. It's a great existing resource!

Using regression to estimate conditional average treatment effect

- Assume we have selected features, defined treatments, and restricted the cohort to satisfy all the causal assumptions
- Choose a regression model for $Y_T = f(X, T)$
  - Examples: Linear regression, causal forest (random forest with a different splitting criterion), neural network
- Fit the regression model

Using regression to estimate conditional average treatment effect

- Assume we have selected features, defined treatments, and restricted the cohort to satisfy all the causal assumptions
- Choose a regression model for $Y_T = f(X, T)$
  - Examples: Linear regression, causal forest (random forest with a different splitting criterion), neural network
- Fit the regression model

- Conditional average treatment effect:

$$CATE(x) = \hat{f}(x,1) - \hat{f}(x,0)$$

Using regression to estimate conditional average treatment effect

- Assume we have selected features, defined treatments, and restricted the cohort to satisfy all the causal assumptions
- Choose a regression model for $Y_T = f(X, T)$
  - Examples: Linear regression, causal forest (random forest with a different splitting criterion), neural network
- Fit the regression model

- Conditional average treatment effect:

$$CATE(x) = \hat{f}(x,1) - \hat{f}(x,0)$$

- Average treatment effect across samples $i = 1,\ldots,n$:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}\left(x^{(i)},1\right) - \hat{f}\left(x^{(i)},0\right)\right)$$

Using regression to estimate conditional average treatment effect

- Assume we have selected features, defined treatments, and restricted the cohort to satisfy all the causal assumptions
- Choose a regression model for $Y_T = f(X, T)$
  - Examples: Linear regression, causal forest (random forest with a different splitting criterion), neural network
- Fit the regression model

- Conditional average treatment effect:

$$CATE(x) = \hat{f}(x,1) - \hat{f}(x,0)$$

- Average treatment effect across samples $i = 1, \ldots, n$:

$$ATE = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}\left(x^{(i)},1\right) - \hat{f}\left(x^{(i)},0\right) \right)$$

- This method is known as covariate adjustment

Analyzing the linear case

- In the special case of a linear regression, ATE is coefficient γ on T
- In the regression lecture, the coefficient was interpreted as increasing TV advertisement spending by 1 unit is *associated* with a γ increase in sales. Cautioned against claiming causality!

Analyzing the linear case

- In the special case of a linear regression, ATE is coefficient γ on T
- In the regression lecture, the coefficient was interpreted as increasing TV advertisement spending by 1 unit is *associated* with a γ increase in sales. Cautioned against claiming causality!

- Why can we claim the coefficient is the causal effect under the 3 assumptions?
- Start with how the best prediction for quadratic loss is the conditional mean:

$$\begin{aligned}\gamma &= \hat{f}(x,1) - \hat{f}(x,0) \\ &= \underbrace{\mathbb{E}\left[Y|X=x,T=1\right] - \mathbb{E}\left[Y|X=x,T=0\right]}_{\text{Observed difference in outcomes}}\end{aligned}$$

Analyzing the linear case

- In the special case of a linear regression, ATE is coefficient γ on T
- In the regression lecture, the coefficient was interpreted as increasing TV advertisement spending by 1 unit is *associated* with a γ increase in sales. Cautioned against claiming causality!

- Why can we claim the coefficient is the causal effect under the 3 assumptions?
- Start with how the best prediction for quadratic loss is the conditional mean:

$$\begin{aligned}\gamma &= \hat{f}(x,1) - \hat{f}(x,0) \\ &= \underbrace{\mathbb{E}\left[Y|X=x,T=1\right] - \mathbb{E}\left[Y|X=x,T=0\right]}_{\text{Observed difference in outcomes}}\end{aligned}$$

- Need overlap to estimate these two quantities for all values of x

Analyzing the linear case

- In the special case of a linear regression, ATE is coefficient γ on T
- Why can we claim the coefficient is the causal effect under the 3 assumptions?

- Apply causal consistency assumption: Average observed outcome is expected potential outcome

$$\gamma = \underbrace{\mathbb{E}\left[Y|X=x,T=1\right] - \mathbb{E}\left[Y|X=x,T=0\right]}_{\text{Observed difference in outcomes}}$$

$$= \underbrace{\mathbb{E}\left[Y_1|X=x,T=1\right] - \mathbb{E}\left[Y_0|X=x,T=0\right]}_{\text{Difference in potential outcomes within treatment groups}}$$

Analyzing the linear case

- In the special case of a linear regression, ATE is coefficient γ on T
- Why can we claim the coefficient is the causal effect under the 3 assumptions?
- No unobserved confounding states $Y_0, Y_1 \perp\!\!\!\perp T \,|\, X$
- This implies $\mathbb{E}\left[Y_t \,|\, X = x, T\right] = \mathbb{E}\left[Y_t \,|\, X = x\right]$

$$\gamma = \overbrace{\mathbb{E}\left[Y_1 \,|\, X = x, \cancel{T = 1}\right] - \mathbb{E}\left[Y_0 \,|\, X = x, \cancel{T = 0}\right]}^{\text{Difference in potential outcomes within treatment groups}}$$

$$= \underbrace{\mathbb{E}\left[Y_1 \,|\, X = x\right] - \mathbb{E}\left[Y_0 \,|\, X = x\right]}_{\text{Causal effect}}$$

Analyzing the linear case

- What happens if there are unobserved confounders?
- Let's add and subtract the same term $\mathbb{E}\left[Y_0 \,|\, X = x, T = 1\right]$

$$\gamma = \underbrace{\mathbb{E}\left[Y_1 \,|\, X = x, T = 1\right] - \mathbb{E}\left[Y_0 \,|\, X = x, T = 0\right]}_{\text{Difference in potential outcomes within treatment groups}}$$

$$= \underbrace{\mathbb{E}\left[Y_1 \,|\, X = x, T = 1\right] - \mathbb{E}\left[Y_0 \,|\, X = x, T = 1\right]}_{\text{Conditional average treatment effect on the treated}}$$

$$+ \underbrace{\mathbb{E}\left[Y_0 \,|\, X = x, T = 1\right] - \mathbb{E}\left[Y_0 \,|\, X = x, T = 0\right]}_{\text{Selection bias}}$$

Take-away from analysis of linear case

- Coefficient γ on T is the average treatment effect only if all 3 assumptions hold and $\hat{f}(X, T)$ is a good estimate of $\mathbb{E}\left[Y \mid X, T\right]$

$$\underbrace{\gamma = \mathbb{E}\left[Y_1 \mid X = x\right] - \mathbb{E}\left[Y_0 \mid X = x\right]}_{\text{Causal effect}}$$

- Select the features, treatments, outcomes, and cohort carefully to satisfy the assumptions

Estimating average treatment effect without a parametric function

- Instead of using a parametric function f to compute ATE:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}\left(x^{(i)},1\right) - \hat{f}\left(x^{(i)},0\right)\right)$$

- Can we take the difference between the average observed outcomes?

$$ATE = \frac{1}{n_1}\sum_{i:T^{(i)}=1} Y^{(i)} - \frac{1}{n_0}\sum_{i:T^{(i)}=0} Y^{(i)}$$

Estimating average treatment effect without a parametric function

- Instead of using a parametric function f to compute ATE:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}\left(x^{(i)},1\right) - \hat{f}\left(x^{(i)},0\right)\right)$$

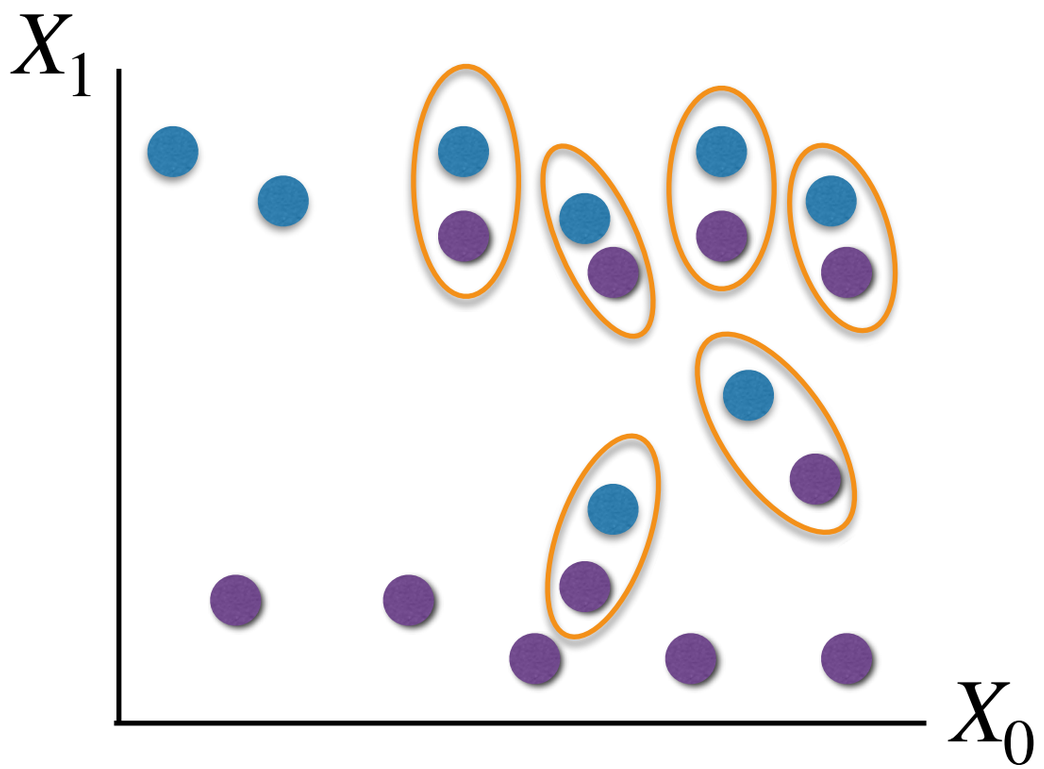
- Can we take the difference between the average observed outcomes?

$$ATE = \frac{1}{n_1}\sum_{i:T^{(i)}=1} Y^{(i)} - \frac{1}{n_0}\sum_{i:T^{(i)}=0} Y^{(i)}$$

- This only works if the treated and control cohorts are “similar” across all potential confounders
- How can we restrict the cohorts to be more similar?
  - Find hypothetical twins in the dataset

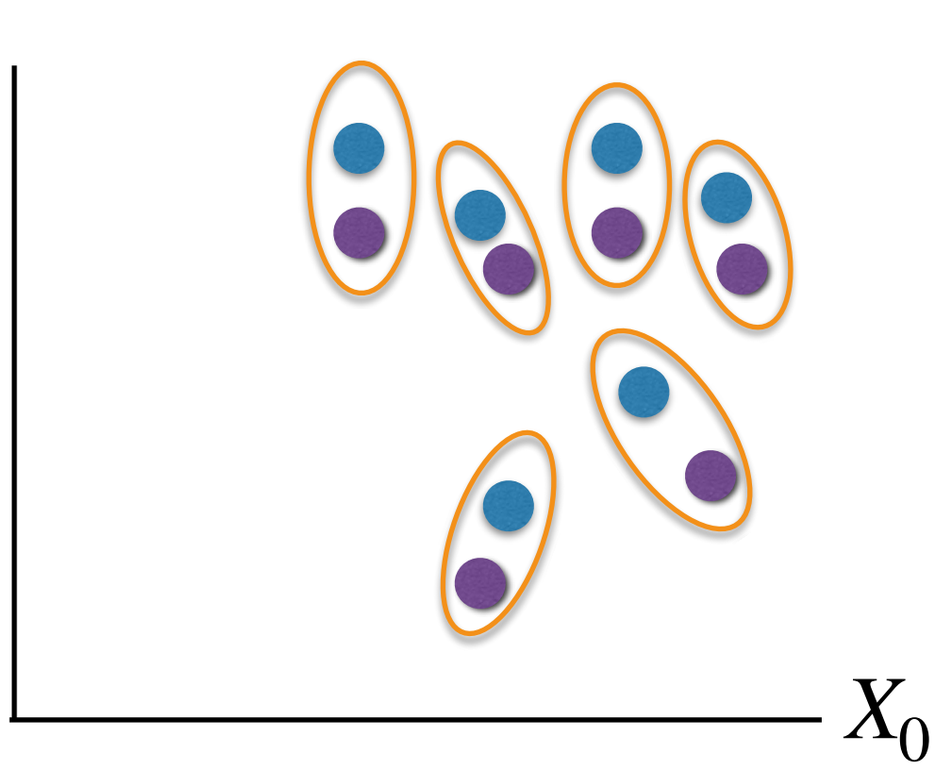
Finding nearest neighbors to estimate average treatment effect

- Find the nearest control sample to each treated sample.
  - If they are close enough, add the pair to the cohort.
  - Once a control sample has been added, it cannot be the nearest neighbor for another treated sample.



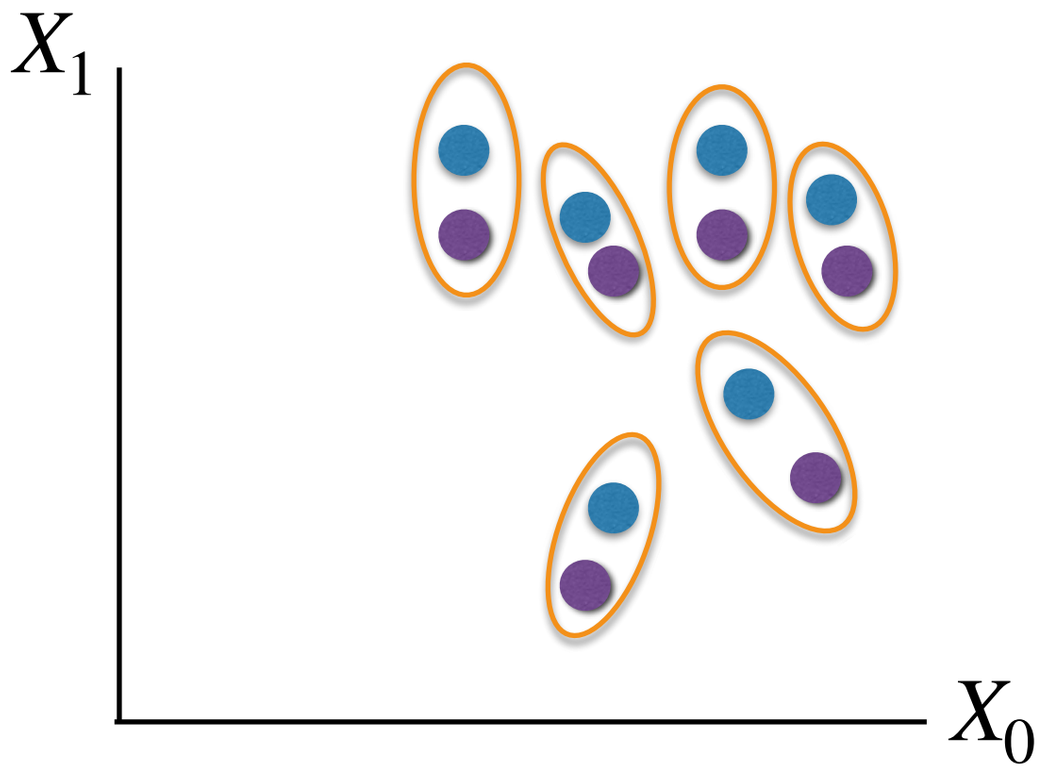
Finding nearest neighbors to estimate average treatment effect

- Discard treated samples without close neighbors or unmatched control samples



Finding nearest neighbors to estimate average treatment effect

- Discard treated samples without close neighbors or unmatched control samples

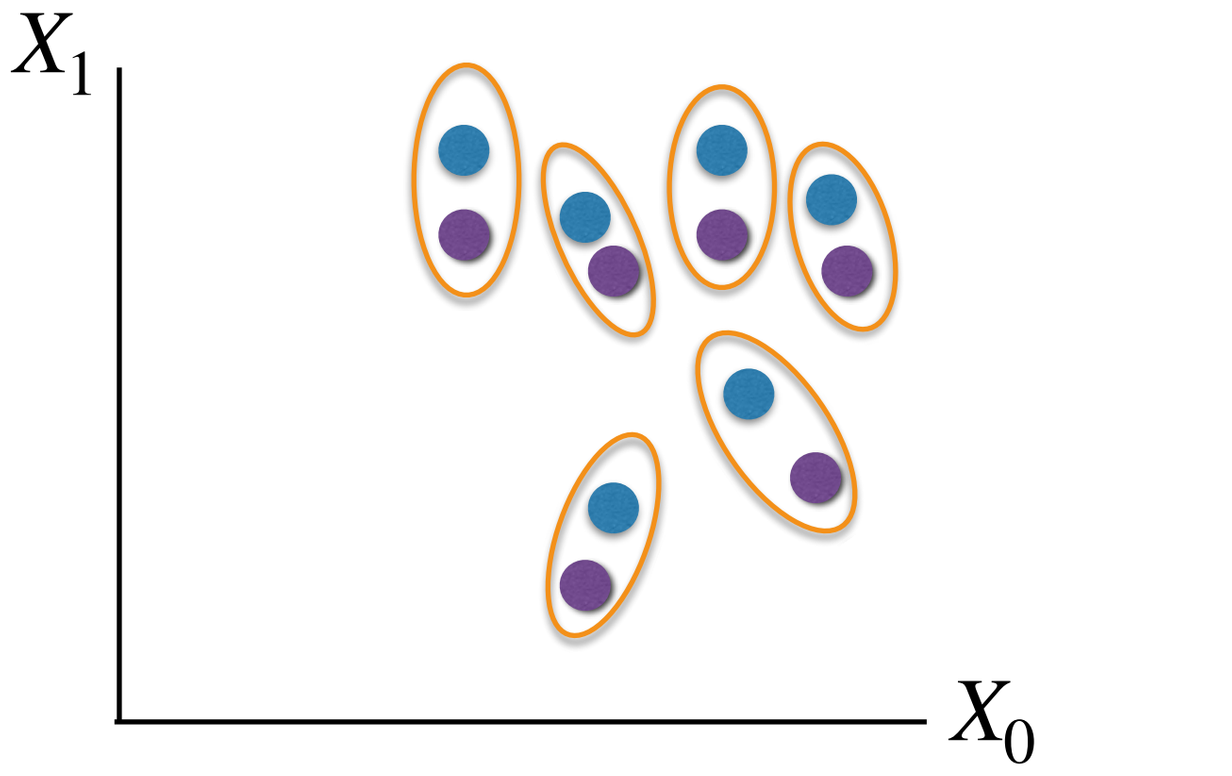


- Estimate average treatment effect as difference between average factual outcomes:

$$ATE = \frac{1}{n_1} \sum_{i:T^{(i)}=1} Y^{(i)} - \frac{1}{n_0} \sum_{i:T^{(i)}=0} Y^{(i)}$$

Finding nearest neighbors to estimate average treatment effect

- Discard treated samples without close neighbors or unmatched control samples



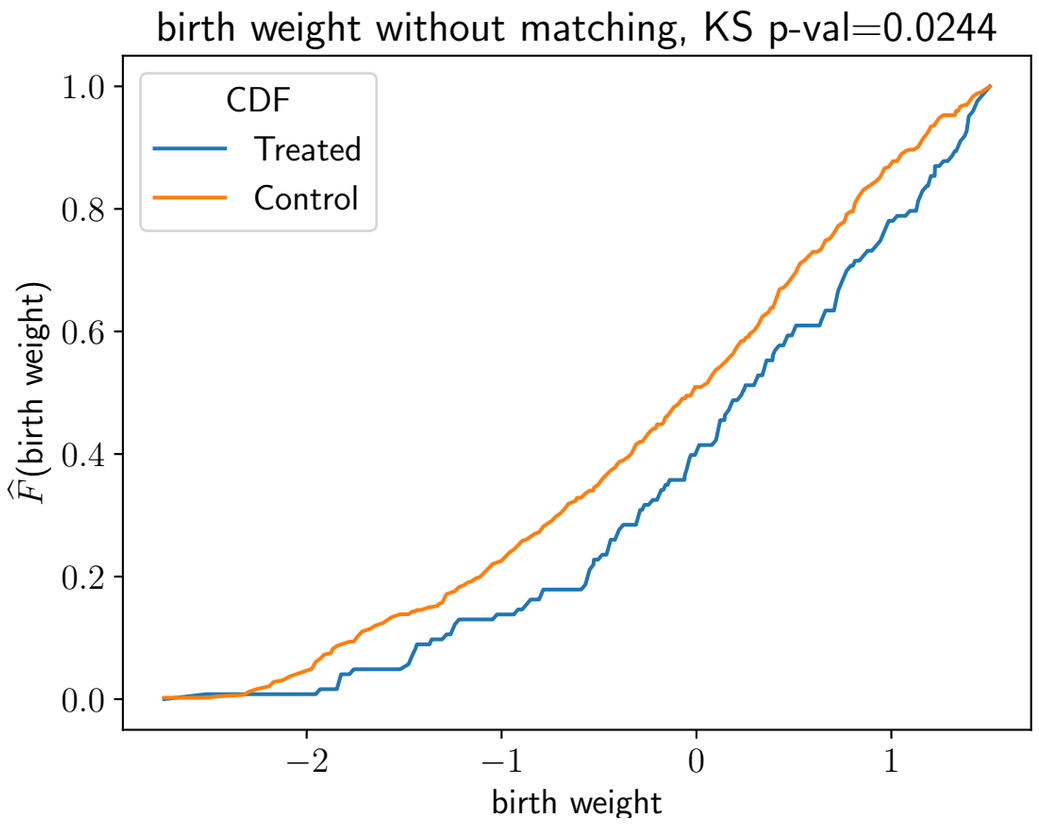
- Estimate average treatment effect as difference between average factual outcomes:

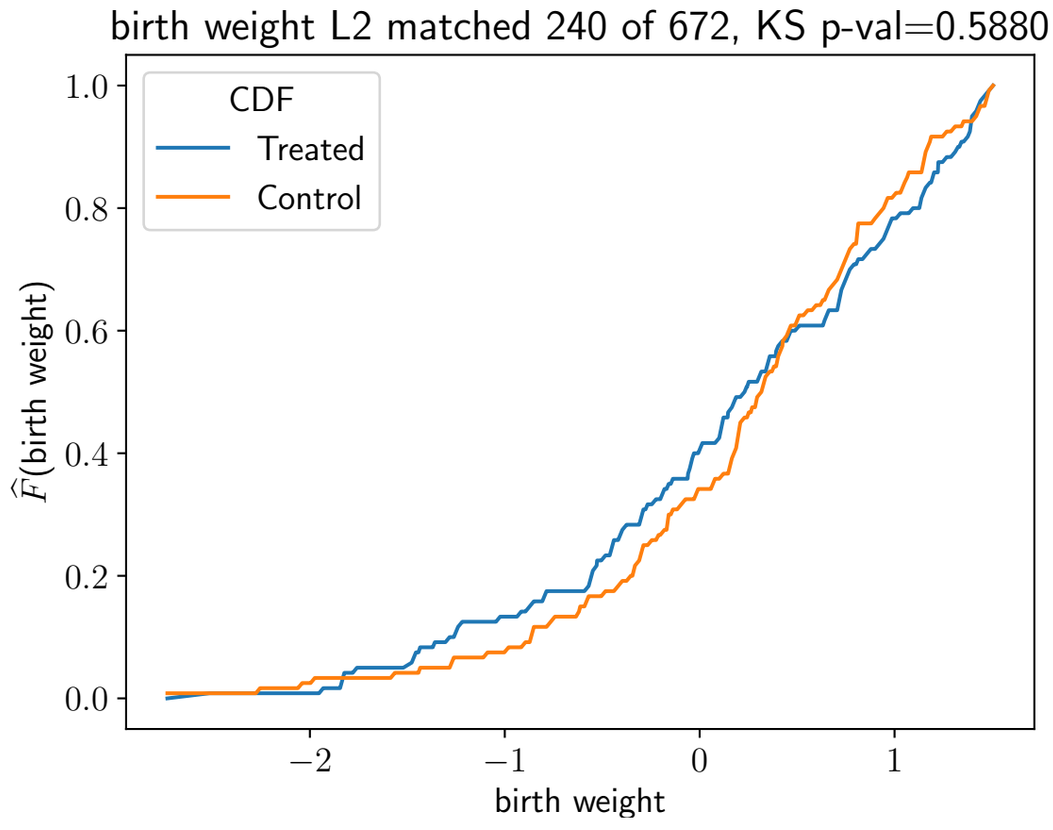
$$ATE = \frac{1}{n_1} \sum_{i:T^{(i)}=1} Y^{(i)} - \frac{1}{n_0} \sum_{i:T^{(i)}=0} Y^{(i)}$$

- This method is known as covariate matching
- Caution: Estimated average treatment effect only applies to restricted cohort!

Check covariate matching improves overlap

- Run a KS test:
  - H: Treated and control cohorts have same feature distribution
  - KS rejects null without matching and does not reject null after matching





Example of covariate matching: Netflix recommendations

- What is the effect of recommending new shows vs continuations?
- Apply covariate matching to genre and # hours in prior week

Viewer	Genre	Prior # hrs	Rec	Post # hrs
1	Action	8	New	14
2	Action	10	New	10
3	Comedy	5	New	5
4	Comedy	7	New	9
5	Horror	2	New	1
6	Action	6	Continue	8
7	Action	10	Continue	8
8	Comedy	9	Continue	8
9	Comedy	12	Continue	10
10	Horror	10	Continue	2

Example of covariate matching: Netflix recommendations

- What is the effect of recommending new shows vs continuations?
- Apply covariate matching to genre and # hours in prior week

Viewer	Genre	Prior # hrs	Rec	Post # hrs
1	Action	8	New	14
2	Action	10	New	10
3	Comedy	5	New	5
4	Comedy	7	New	9
5	Horror	2	New	1
6	Action	6	Continue	8
7	Action	10	Continue	8
8	Comedy	9	Continue	8
9	Comedy	12	Continue	10
10	Horror	10	Continue	2

Example of covariate matching: Netflix recommendations

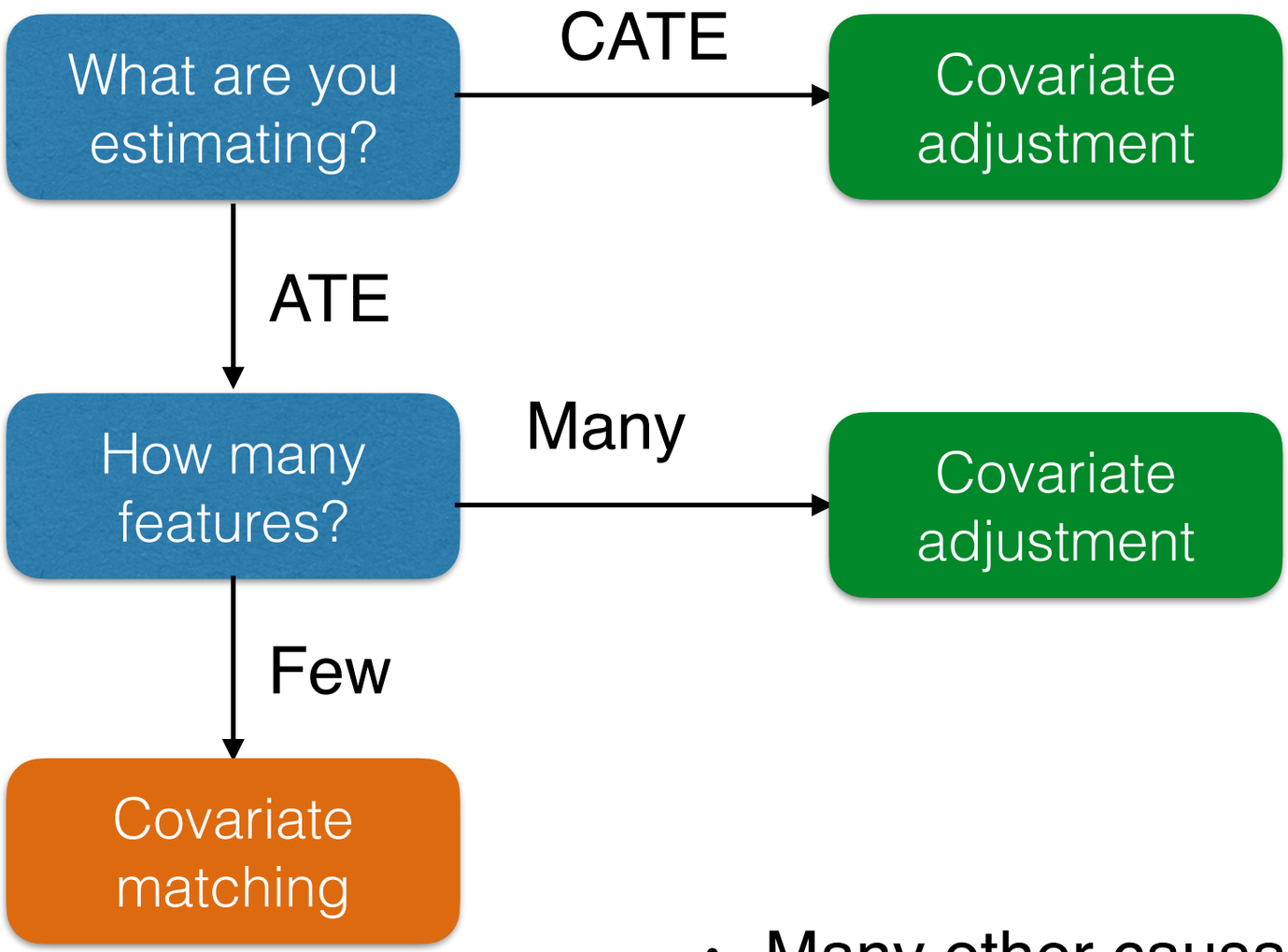
- What is the effect of recommending new shows vs continuations?
- Apply covariate matching to genre and # hours in prior week

Viewer	Genre	Prior # hrs	Rec	Post # hrs
1	Action	8	New	14
2	Action	10	New	10
4	Comedy	7	New	9
6	Action	6	Continue	8
7	Action	10	Continue	8
8	Comedy	9	Continue	8

$$ATE = \frac{1}{3}(14 + 10 + 9) - \frac{1}{3}(8 + 8 + 8) = 3$$

- Recommending new shows leads to 3 additional hours of viewing
- Estimate only applies to action and comedy viewers

Selecting an observational study approach



- Many other causal inference approaches exist depending on what kind of data you have

Key take-aways for causal inference

- For a person with features X,
  what is the effect of treatment T on outcome Y?

Key take-aways for causal inference

- For a person with features X,
  what is the effect of treatment T on outcome Y?



- Design the features, treatment, outcome, and cohort to satisfy assumptions:
  - No unobserved confounding
  - Overlap between treated and control cohorts
  - Consistent treatments with no interference

Key take-aways for causal inference

- For a person with features X,
  what is the effect of treatment T on outcome Y?



- Design the features, treatment, outcome, and cohort to satisfy assumptions:
  - No unobserved confounding
  - Overlap between treated and control cohorts
  - Consistent treatments with no interference
- If you can run an experiment, randomization is essential!
  - Analyze whether results are significant with a two-sample t-test

Key take-aways for causal inference

- For a person with features X,
  what is the effect of treatment T on outcome Y?



- Design the features, treatment, outcome, and cohort to satisfy assumptions:
  - No unobserved confounding
  - Overlap between treated and control cohorts
  - Consistent treatments with no interference
- If you can run an experiment, randomization is essential!
  - Analyze whether results are significant with a two-sample t-test
- If you have observational data, we learned two methods:
  - Covariate adjustment: Fit a regression to estimate $Y_t = f(X, t)$
  - Covariate matching: Pair neighboring treated and control samples