SISCER Module 2: Causal Inference with Observational Data: Common Designs and Statistical Methods

Ting Ye & Richard Guo

Department of Biostatistics, UW

Day 1, Lecture 1: Causal inference and randomized controlled trials

July 8, 2024

Overview

Association

Causality and association

Causality and association

- Causality is central to human knowledge.

Causality and association

- Causality is central to human knowledge.
- The major part of classic statistics is about association (e.g., Pearson correlation, regression coefficient) rather than causation.
  - Association/correlation describes the statistical relationship in the data, indicating difference in one variable is associated with difference in another.

Causality and association

- Causality is central to human knowledge.
- The major part of classic statistics is about association (e.g., Pearson correlation, regression coefficient) rather than causation.
  - Association/correlation describes the statistical relationship in the data, indicating difference in one variable is associated with difference in another.
  - Association / correlation does not imply causation.

Causality and association

- Causality is central to human knowledge.
- The major part of classic statistics is about association (e.g., Pearson correlation, regression coefficient) rather than causation.
  - Association/correlation describes the statistical relationship in the data, indicating difference in one variable is associated with difference in another.
  - Association / correlation does not imply causation.
  - May be good for prediction but not enough for causation.

Causality and association

- Causality is central to human knowledge.
- The major part of classic statistics is about association (e.g., Pearson correlation, regression coefficient) rather than causation.
  - Association/correlation describes the statistical relationship in the data, indicating difference in one variable is associated with difference in another.
  - Association / correlation does not imply causation.
  - May be good for prediction but not enough for causation.

Causality and association

- Causality is central to human knowledge.
- The major part of classic statistics is about association (e.g., Pearson correlation, regression coefficient) rather than causation.
  - Association/correlation describes the statistical relationship in the data, indicating difference in one variable is associated with difference in another.
  - Association / correlation does not imply causation.
  - May be good for prediction but not enough for causation.
- Causation requires mechanistic understanding, indicating **intervention** in one variable leads to change in another.

Simpson's paradox

Simpson's paradox

- Consider a kidney stone example from Charig et al. (1986)

	Success	Failure
Open surgical procedure	273	77
Small puncture procedure	289	61

Simpson's paradox

- Consider a kidney stone example from Charig et al. (1986)

	Success	Failure
Open surgical procedure	273	77
Small puncture procedure	289	61

- The estimated risk difference is

$$\widehat{\text{RD}} = \underbrace{\frac{273}{273+77}}_{\text{open}} - \underbrace{\frac{289}{289+61}}_{\text{small}} = 78\% - 83\% = -5\% < 0.$$

Simpson's paradox

- Consider a kidney stone example from Charig et al. (1986)

	Success	Failure
Open surgical procedure	273	77
Small puncture procedure	289	61

- The estimated risk difference is

$$\widehat{\text{RD}} = \underbrace{\frac{273}{273+77}}_{\text{open}} - \underbrace{\frac{289}{289+61}}_{\text{small}} = 78\% - 83\% = -5\% < 0.$$

- Success rate is higher among the small puncture group (association)

Simpson's paradox

- Consider a kidney stone example from Charig et al. (1986)

	Success	Failure
Open surgical procedure	273	77
Small puncture procedure	289	61

- The estimated risk difference is

$$\widehat{\text{RD}} = \underbrace{\frac{273}{273+77}}_{\text{open}} - \underbrace{\frac{289}{289+61}}_{\text{small}} = 78\% - 83\% = -5\% < 0.$$

- Success rate is higher among the small puncture group (association)
- But is small puncture procedure better? (causation)

Simpson's paradox

Simpson's paradox

- Patients were **not** randomized into the two procedures

Simpson's paradox

- Patients were **not** randomized into the two procedures
- Patients receiving open surgical tend to have large stones, whereas patients receiving small puncture tend to have small stones.

Patients with small stones	Success	Failure
Open surgical procedure	81	6
Small puncture procedure	234	36

Patients with large stones	Success	Failure
Open surgical procedure	192	71
Small puncture procedure	55	25

Simpson's paradox

- Patients were **not** randomized into the two procedures
- Patients receiving open surgical tend to have large stones, whereas patients receiving small puncture tend to have small stones.

Patients with small stones	Success	Failure
Open surgical procedure	81	6
Small puncture procedure	234	36

Patients with large stones	Success	Failure
Open surgical procedure	192	71
Small puncture procedure	55	25

  - Simpson's paradox:

$$\widehat{\mathrm{RD}}_S = \frac{81}{8+6} - \frac{234}{234+46} = 6\% > 0, \quad \widehat{\mathrm{RD}}_L = \frac{192}{192+71} - \frac{55}{55+25} = 4\% > 0, \quad \text{but } \widehat{\mathrm{RD}} = -5\% < 0.$$

Simpson's paradox

- Patients were **not** randomized into the two procedures
- Patients receiving open surgical tend to have large stones, whereas patients receiving small puncture tend to have small stones.

Patients with small stones	Success	Failure
Open surgical procedure	81	6
Small puncture procedure	234	36

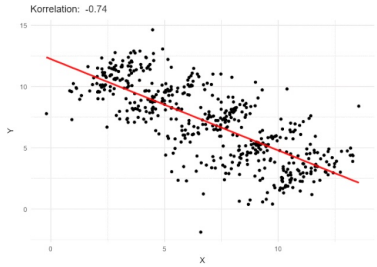
Patients with large stones	Success	Failure
Open surgical procedure	192	71
Small puncture procedure	55	25

  - Simpson's paradox:

$$\widehat{\mathrm{RD}}_S = \frac{81}{8+6} - \frac{234}{234+46} = 6\% > 0, \quad \widehat{\mathrm{RD}}_L = \frac{192}{192+71} - \frac{55}{55+25} = 4\% > 0, \quad \text{but } \widehat{\mathrm{RD}} = -5\% < 0.$$

  - Confounding: stone size affects both treatment assignment and success rate.

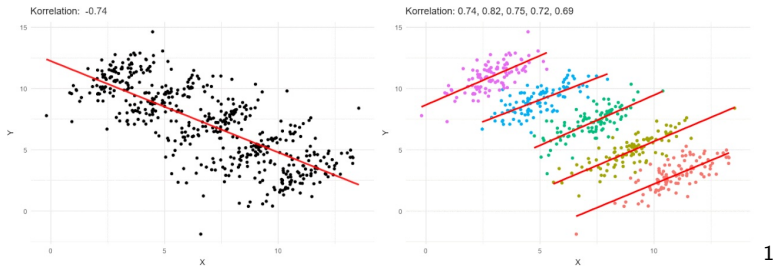
Simpson's paradox



1

[1] By Pace svwiki - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=62007681

Simpson's paradox



[1]

[1] By Pace svwiki - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=62007681

Causality and potential outcomes

Potential outcomes

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.
- Consider two levels of treatment / exposure / intervention (e.g., 1 for treatment; 0 for control).
  - Example: 1 for open surgical procedure; 0 for small puncture procedure

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.
- Consider two levels of treatment / exposure / intervention (e.g., 1 for treatment; 0 for control).
  - Example: 1 for open surgical procedure; 0 for small puncture procedure
- Unit i has two potential outcomes (i.e., 'counterfactuals') $Y_i(1)$ and $Y_i(0)$

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.
- Consider two levels of treatment / exposure / intervention (e.g., 1 for treatment; 0 for control).
  - Example: 1 for open surgical procedure; 0 for small puncture procedure
- Unit i has two potential outcomes (i.e., 'counterfactuals') $Y_i(1)$ and $Y_i(0)$
  - $Y_i(1)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the treatment

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.
- Consider two levels of treatment / exposure / intervention (e.g., 1 for treatment; 0 for control).
  - Example: 1 for open surgical procedure; 0 for small puncture procedure
- Unit i has two potential outcomes (i.e., 'counterfactuals') $Y_i(1)$ and $Y_i(0)$
  - $Y_i(1)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the treatment
  - $Y_i(0)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the control

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.
- Consider two levels of treatment / exposure / intervention (e.g., 1 for treatment; 0 for control).
  - Example: 1 for open surgical procedure; 0 for small puncture procedure
- Unit i has two potential outcomes (i.e., 'counterfactuals') $Y_i(1)$ and $Y_i(0)$
  - $Y_i(1)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the treatment
  - $Y_i(0)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the control
  - Unit i's individual treatment effect: $Y_i(1) - Y_i(0)$

Potential outcomes

- Potential outcomes framework was developed by Neyman (1923) and Rubin (1974); see Holland (1986) for more background and history.
- Consider two levels of treatment / exposure / intervention (e.g., 1 for treatment; 0 for control).
  - Example: 1 for open surgical procedure; 0 for small puncture procedure
- Unit i has two potential outcomes (i.e., 'counterfactuals') $Y_i(1)$ and $Y_i(0)$
  - $Y_i(1)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the treatment
  - $Y_i(0)$ is the outcome (e.g., 1: success; 0: failure) that unit i would experience if takes the control
  - Unit i's individual treatment effect: $Y_i(1) - Y_i(0)$
- A clear definition of causal effect and actionable information:
  ☞ What can we conclude if we knew $Y_i(1) - Y_i(0) = 1$, 0 or -1?

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments
   - Can be violated in infectious diseases or network experiments.

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments
   - Can be violated in infectious diseases or network experiments.
   - Example: if my friends receive flu shots, my chance of getting the flu decreases even if I do not receive the flu shot.

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments
   - Can be violated in infectious diseases or network experiments.
   - Example: if my friends receive flu shots, my chance of getting the flu decreases even if I do not receive the flu shot.
   - Causal inference with interference is an active research area (Hudgens & Halloran, 2008).

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments
   - Can be violated in infectious diseases or network experiments.
   - Example: if my friends receive flu shots, my chance of getting the flu decreases even if I do not receive the flu shot.
   - Causal inference with interference is an active research area (Hudgens & Halloran, 2008).
2. **No multiple versions of treatment**: Treatment levels are well-defined and each level has only one 'version' of the treatment.

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments
   - Can be violated in infectious diseases or network experiments.
   - Example: if my friends receive flu shots, my chance of getting the flu decreases even if I do not receive the flu shot.
   - Causal inference with interference is an active research area (Hudgens & Halloran, 2008).
2. **No multiple versions of treatment**: Treatment levels are well-defined and each level has only one 'version' of the treatment.
   - Can be violated, e.g., $Y_i(1)$ (1: open surgical) depends on the surgeon who performs the operation.

$Y_i(0)$, $Y_i(1)$: Implicit assumptions

1. **No interference**: Unit i's potential outcomes do not depend on other units' treatments
   - Can be violated in infectious diseases or network experiments.
   - Example: if my friends receive flu shots, my chance of getting the flu decreases even if I do not receive the flu shot.
   - Causal inference with interference is an active research area (Hudgens & Halloran, 2008).
2. **No multiple versions of treatment**: Treatment levels are well-defined and each level has only one 'version' of the treatment.
   - Can be violated, e.g., $Y_i(1)$ (1: open surgical) depends on the surgeon who performs the operation.

▶ These two assumptions together is called **SUTVA** (Stable Unit Treatment Value Assumption) (Rubin, 1980).

Fundamental problem of causal inference

- Under SUTVA, potential outcomes of all units in the study is called the 'Science Table' (Rubin, 2005).

i	$Y_i(0)$	$Y_i(1)$
1	$Y_1(0)$	$Y_1(1)$
2	$Y_2(0)$	$Y_2(1)$
$\vdots$	$\vdots$	$\vdots$

Fundamental problem of causal inference

- Under SUTVA, potential outcomes of all units in the study is called the 'Science Table' (Rubin, 2005).

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	$Y_1(0)$	$Y_1(1)$	0	
2	$Y_2(0)$	$Y_2(1)$	1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- A_i: treatment received by unit i; Y_i: unit i's outcome.

Fundamental problem of causal inference

- Under SUTVA, potential outcomes of all units in the study is called the 'Science Table' (Rubin, 2005).

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	$Y_1(0)$	$Y_1(1)$	0	$Y_1(0)$
2	$Y_2(0)$	$Y_2(1)$	1	$Y_2(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- A_i: treatment received by unit i; Y_i: unit i's outcome.

Consistency: $Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i) Y_i(0).$

Fundamental problem of causal inference

- Under SUTVA, potential outcomes of all units in the study is called the 'Science Table' (Rubin, 2005).

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	$Y_1(0)$	$Y_1(1)$	0	$Y_1(0)$
2	$Y_2(0)$	$Y_2(1)$	1	$Y_2(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- A_i: treatment received by unit i; Y_i: unit i's outcome.

Consistency: $Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i) Y_i(0).$

- Fundamental problem of causal inference:

$Y_i(0)$ and $Y_i(1)$ **cannot** be both observed for a unit i.

Sample average treatment effect (finite population)

☞ Although we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any **individual** i, we can statistically infer the effect for a **population** on average.

Sample average treatment effect (finite population)

☞ Although we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any **individual** i, we can statistically infer the effect for a **population** on average.

Suppose the study consists of units $i = 1, \ldots, n$.

- The **SATE** (Sample Average Treatment Effect) is

$$\boxed{\text{SATE} := \frac{1}{n}\sum_i Y_i(1) - \frac{1}{n}\sum_i Y_i(0) = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0)).}$$

Sample average treatment effect (finite population)

☞ Although we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any **individual** i, we can statistically infer the effect for a **population** on average.

Suppose the study consists of units $i = 1, \ldots, n$.

- The **SATE** (Sample Average Treatment Effect) is

$$\boxed{\text{SATE} := \frac{1}{n}\sum_i Y_i(1) - \frac{1}{n}\sum_i Y_i(0) = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0)).}$$

  - $\frac{1}{n}\sum_i Y_i(1)$: mean outcome had all units received treatment.

Sample average treatment effect (finite population)

☞ Although we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any **individual** i, we can statistically infer the effect for a **population** on average.

Suppose the study consists of units $i = 1, \ldots, n$.

- The **SATE** (Sample Average Treatment Effect) is

$$\boxed{\text{SATE} := \frac{1}{n}\sum_i Y_i(1) - \frac{1}{n}\sum_i Y_i(0) = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0)).}$$

- $\frac{1}{n}\sum_i Y_i(1)$: mean outcome had all units received treatment.
- $\frac{1}{n}\sum_i Y_i(0)$: mean outcome had all units received control.

▶ **intervention**

Sample average treatment effect (finite population)

☞ Although we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any **individual** i, we can statistically infer the effect for a **population** on average.

Suppose the study consists of units $i = 1, \ldots, n$.

- The **SATE** (Sample Average Treatment Effect) is

$$\boxed{\text{SATE} := \frac{1}{n}\sum_i Y_i(1) - \frac{1}{n}\sum_i Y_i(0) = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0)).}$$

- $\frac{1}{n}\sum_i Y_i(1)$: mean outcome had all units received treatment.
- $\frac{1}{n}\sum_i Y_i(0)$: mean outcome had all units received control.
- $\frac{1}{n}\sum_i (Y_i(1) - Y_i(0))$: the average of individual treatment effects

▶ **intervention**

Sample average treatment effect (finite population)

☞ Although we cannot observe both $Y_i(1)$ and $Y_i(0)$ for any **individual** i, we can statistically infer the effect for a **population** on average.

Suppose the study consists of units $i = 1, \ldots, n$.

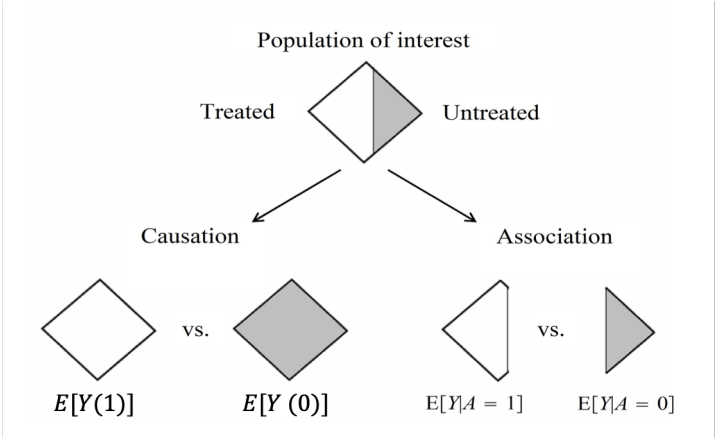
▶ The **SATE** (Sample Average Treatment Effect) is

$$\text{SATE} := \frac{1}{n}\sum_i Y_i(1) - \frac{1}{n}\sum_i Y_i(0) = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0)).$$

☞ This is a **fixed** quantity pertaining to the n units in the study.

- $\frac{1}{n}\sum_i Y_i(1)$: mean outcome had all units received treatment.
- $\frac{1}{n}\sum_i Y_i(0)$: mean outcome had all units received control.
- $\frac{1}{n}\sum_i (Y_i(1) - Y_i(0))$: the average of individual treatment effects

▶ **intervention**



[2]

[2] Adapted from Fig 1.1 in Hernan and Robins.

RCT: Randomization inference

Randomized controlled trials (RCTs): gold standard of causal inference

Observed data

- There are n units in the experiment
- Binary treatment: $A_i = 1$ is the treatment (e.g., open surgical) and $A_i = 0$ the control (e.g., small puncture)
- After treatment assignment, we measure an outcome variable Y_i

Potential outcomes

- $Y_i(0), Y_i(1)$, potential outcomes for i-th unit under control and treatment
- $Y_i = Y_i(A_i)$ (consistency)

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	?	8	1	8
2	-6	?	0	-6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Estimate SATE from RCT

Estimate SATE from RCT

- **Complete Randomization:** Fix treatment group size n_1. Choose n_1 units uniformly at random to receive treatment; the rest $n_0 = n - n_1$ units receive control.

Estimate SATE from RCT

- **Complete Randomization:** Fix treatment group size n_1. Choose n_1 units uniformly at random to receive treatment; the rest $n_0 = n - n_1$ units receive control.

☞ Then, using consistency, (What is $\mathbb{E}[\cdot]$ taken over?)

$$\mathbb{E}\underbrace{[\bar{Y}_1]}_{\text{trt group mean}} = \mathbb{E}\left[\frac{1}{n_1}\sum_i A_i Y_i\right] = \mathbb{E}\left[\frac{1}{n_1}\sum_i A_i Y_i(1)\right] = \frac{1}{n_1}\sum_i \mathbb{E}[A_i]Y_i(1) = \frac{1}{n}\sum_i Y_i(1)$$

Estimate SATE from RCT

- **Complete Randomization:** Fix treatment group size n_1. Choose n_1 units uniformly at random to receive treatment; the rest $n_0 = n - n_1$ units receive control.

☞ Then, using consistency, (What is $\mathbb{E}[\cdot]$ taken over?)

$$\mathbb{E}\underbrace{[\bar{Y}_1]}_{\text{trt group mean}} = \mathbb{E}\left[\frac{1}{n_1}\sum_i A_i Y_i\right] = \mathbb{E}\left[\frac{1}{n_1}\sum_i A_i Y_i(1)\right] = \frac{1}{n_1}\sum_i \mathbb{E}[A_i]Y_i(1) = \frac{1}{n}\sum_i Y_i(1)$$

$$\mathbb{E}\underbrace{[\bar{Y}_0]}_{\text{ctrl group mean}} = \mathbb{E}\left[\frac{1}{n_0}\sum_i (1-A_i) Y_i\right] = \mathbb{E}\left[\frac{1}{n_0}\sum_i (1-A_i) Y_i(0)\right] = \frac{1}{n}\sum_i Y_i(0).$$

Estimate SATE from RCT

- **Complete Randomization:** Fix treatment group size n_1. Choose n_1 units uniformly at random to receive treatment; the rest $n_0 = n - n_1$ units receive control.
  - Also holds for **simple randomization**, $A_i \sim$ coin flip.

☞ Then, using consistency, (What is $\mathbb{E}[\cdot]$ taken over?)

$$\mathbb{E}\underbrace{[\bar{Y}_1]}_{\text{trt group mean}} = \mathbb{E}\left[\frac{1}{n_1}\sum_i A_i Y_i\right] = \mathbb{E}\left[\frac{1}{n_1}\sum_i A_i Y_i(1)\right] = \frac{1}{n_1}\sum_i \mathbb{E}[A_i]Y_i(1) = \frac{1}{n}\sum_i Y_i(1)$$

$$\mathbb{E}\underbrace{[\bar{Y}_0]}_{\text{ctrl group mean}} = \mathbb{E}\left[\frac{1}{n_0}\sum_i (1-A_i) Y_i\right] = \mathbb{E}\left[\frac{1}{n_0}\sum_i (1-A_i) Y_i(0)\right] = \frac{1}{n}\sum_i Y_i(0).$$

☞ The **difference-in-means**

$$\boxed{\widehat{\tau} := \bar{Y}_1 - \bar{Y}_0.}$$

is **unbiased** for $\mathrm{SATE} = \frac{1}{n}\sum_i (Y_i(1) - Y_i(0))$.

- association = causation

An RCT of fish oil diet (Knapp & FitzGerald, 1989)

An RCT of fish oil diet (Knapp & FitzGerald, 1989)

- Researchers used 7 red and 7 black playing cards to randomly assign 14 volunteer males with high blood pressure to one of two diets for four weeks: a fish oil diet ($A = 1$) and a standard oil diet ($A = 0$).

An RCT of fish oil diet (Knapp & FitzGerald, 1989)

- Researchers used 7 red and 7 black playing cards to randomly assign 14 volunteer males with high blood pressure to one of two diets for four weeks: a fish oil diet ($A = 1$) and a standard oil diet ($A = 0$).
- The reductions in diastolic blood pressure (DBP) after four weeks among the 14 men are shown below.

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	?	8	1	8
2	?	12	1	12
3	?	10	1	10
4	?	14	1	14
5	?	2	1	2
6	?	0	1	0
7	?	0	1	0

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	-6	?	0	-6
2	0	?	0	0
3	1	?	0	1
4	2	?	0	2
5	-3	?	0	-3
6	-4	?	0	-4
7	2	?	0	2

Result from RCT

Result from RCT

- For those assigned to fish oil diet, average DBP reduction after four weeks is

$$\bar{Y}_1 := \frac{1}{7} \sum_i A_i Y_i = 6.6$$

Result from RCT

- For those assigned to fish oil diet, average DBP reduction after four weeks is

$$\bar{Y}_1 := \frac{1}{7}\sum_i A_i Y_i = 6.6$$

- For those assigned to standard oil diet, average DBP reduction after four weeks is

$$\bar{Y}_0 := \frac{1}{7}\sum_i (1 - A_i) Y_i = -1.1$$

Result from RCT

- For those assigned to fish oil diet, average DBP reduction after four weeks is

$$\bar{Y}_1 := \frac{1}{7}\sum_i A_i Y_i = 6.6$$

- For those assigned to standard oil diet, average DBP reduction after four weeks is

$$\bar{Y}_0 := \frac{1}{7}\sum_i (1 - A_i) Y_i = -1.1$$

- Our SATE estimate is $\bar{Y}_1 - \bar{Y}_0 = 7.7$ (95% CI: [2.7, 12.7]), indicating fish oil diet for four weeks on average lead to an additional 7.7 reduction in DBP compared to standard oil diet.

$\widehat{\tau}$ as an estimator of SATE

- Potential outcomes are fixed; **randomness** solely comes from **treatment assignment**.
- SATE: $\tau := \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$
- Notation: $\bar{Y}(1) = \frac{1}{n}\sum_{i=1}^{n} Y_i(1)$, $\bar{Y}(0) = \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$, $S^2(1) = \frac{1}{n-1}\sum_{i=1}^{n}\{Y_i(1) - \bar{Y}(1)\}^2$ and $S^2(0) = \frac{1}{n-1}\sum_{i=1}^{n}\{Y_i(0) - \bar{Y}(0)\}^2$, all four quantities are unobserved.

Theorem

(a) $\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0$ is unbiased for SATE $= \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$

(b) $\widehat{\tau}$ has variance $V = Var(\widehat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}$, where $S^2(\tau) = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \tau)^2$ measures the treatment effect heterogeneity.

(c) $\widehat{V} = S_1^2/n_1 + S_0^2/n_0$ is a conservative variance estimator in the sense that $E(\widehat{V}) - Var(\widehat{\tau}) = S^2(\tau)/n \geq 0$, with equality holding if and only if $\tau_i = \tau$ for all units, where $S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n} A_i(Y_i - \bar{Y}_1)^2$ and $S_0^2 = \frac{1}{n_0-1}\sum_{i=1}^{n}(1 - A_i)(Y_i - \bar{Y}_0)^2$.

Randomization inference: Lady tasting tea

Randomization inference: Lady tasting tea

- Fisher described the following famous experiment of *lady tasting tea* in his 1935 book *The Design of Experiments*.

Randomization inference: Lady tasting tea

- Fisher described the following famous experiment of *lady tasting tea* in his 1935 book *The Design of Experiments*.
- A lady claimed that she could tell the difference between the two ways of making milk tea: one with milk added first, and the other with tea added first.

Randomization inference: Lady tasting tea

- Fisher described the following famous experiment of *lady tasting tea* in his 1935 book *The Design of Experiments*.
- A lady claimed that she could tell the difference between the two ways of making milk tea: one with milk added first, and the other with tea added first.
- As a statistician, Fisher designed an experiment to test whether the lady could actually tell the difference between the two ways of making milk tea.

Randomization inference: Lady tasting tea

- Fisher described the following famous experiment of *lady tasting tea* in his 1935 book *The Design of Experiments*.
- A lady claimed that she could tell the difference between the two ways of making milk tea: one with milk added first, and the other with tea added first.
- As a statistician, Fisher designed an experiment to test whether the lady could actually tell the difference between the two ways of making milk tea.
- He made 8 cups of tea, 4 with milk added first and the other 4 with tea added first. Then he presented these 8 cups of tea in a random order to the lady and asked the lady to pick up the 4 with milk added first.

Quiz 1: Lady tasting tea

- Fisher described the following famous experiment of *lady tasting tea* in his 1935 book *The Design of Experiments*.
- A lady claimed that she could tell the difference between the two ways of making milk tea: one with milk added first, and the other with tea added first.
- As a statistician, Fisher designed an experiment to test whether the lady could actually tell the difference between the two ways of making milk tea.
- He made 8 cups of tea, 4 with milk added first and the other 4 with tea added first. Then he presented these 8 cups of tea in a random order to the lady and asked the lady to pick up the 4 with milk added first.
- Exercise: What are the units, treatment, and outcome in this experiment?

Lady tasting tea

Lady tasting tea

- Units: cup $i = 1, \ldots, 8$
- Treatment A_i for the i-th cup (0: milk added first; 1: tea added first)
- Outcome Y_i for the i-th cup (0: lady determines milk added first; 1: lady determines tea added first)
- Potential outcomes $Y_i(0), Y_i(1)$ for the i-th cup (What does $Y_3(0) = 1$ represent?)

Lady tasting tea

- Units: cup $i = 1, \ldots, 8$
- Treatment A_i for the i-th cup (0: milk added first; 1: tea added first)
- Outcome Y_i for the i-th cup (0: lady determines milk added first; 1: lady determines tea added first)
- Potential outcomes $Y_i(0), Y_i(1)$ for the i-th cup (What does $Y_3(0) = 1$ represent?)

▶ For $a = 0, 1$ and $y = 0, 1$, let N_{ay} be the number of cups with $A_i = a$ and $Y_i = y$.

		Outcome Y		
		milk first (lady)	tea first (lady)	Total
Treatment A	milk first	N_{00}	N_{01}	4
	tea first	N_{10}	N_{11}	4
	Total	4	4	8

Lady tasting tea

- Units: cup $i = 1, \ldots, 8$
- Treatment A_i for the i-th cup (0: milk added first; 1: tea added first)
- Outcome Y_i for the i-th cup (0: lady determines milk added first; 1: lady determines tea added first)
- Potential outcomes $Y_i(0), Y_i(1)$ for the i-th cup (What does $Y_3(0) = 1$ represent?)

▶ For $a = 0, 1$ and $y = 0, 1$, let N_{ay} be the number of cups with $A_i = a$ and $Y_i = y$.

		Outcome Y		
		milk first (lady)	tea first (lady)	Total
Treatment A	milk first	N_{00}	N_{01}	4
	tea first	N_{10}	N_{11}	4
	Total	4	4	8

(Why are the margins fixed?)

Lady tasting tea

- For $a = 0, 1$ and $y = 0, 1$, let N_{ay} be the number of cups with $A_i = a$ and $Y_i = y$.

		Outcome Y		
		milk first (lady)	tea first (lady)	Total
Treatment A	milk first	N_{00}	$4 - N_{00}$	4
	tea first	$4 - N_{00}$	N_{00}	4
	Total	4	4	8

Lady tasting tea

- For $a = 0, 1$ and $y = 0, 1$, let N_{ay} be the number of cups with $A_i = a$ and $Y_i = y$.

		Outcome Y		
		milk first (lady)	tea first (lady)	Total
Treatment A	milk first	N_{00}	$4 - N_{00}$	4
	tea first	$4 - N_{00}$	N_{00}	4
	Total	4	4	8

- Fisher's sharp null H_{0F} posits that the lady cannot tell the difference of any cup:

$$\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$$

For every cup, whether or not the milk was added first to the cup, the lady would say the same thing: either $Y_i(0) = Y_i(1) = 0$ (milk added first) or $Y_i(0) = Y_i(1) = 1$ (tea added first).

Lady tasting tea

- Fisher's sharp null H_{0F}: $\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$

Lady tasting tea

- Fisher's sharp null H_{0F}: $\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$
- In the real example, an extreme value of $N_{00} = 4$ was recorded.

Lady tasting tea

- Fisher's sharp null H_{0F}: $\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$
- In the real example, an extreme value of $N_{00} = 4$ was recorded.

'Science Table' under H_{0F}

cup i	1	2	3	4	5	6	7	8
milk-first $Y_i(0)$	0	0	0	0	1	1	1	1
tea-first $Y_i(1)$	0	0	0	0	1	1	1	1

Lady tasting tea

- Fisher's sharp null H_{0F}: $\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$
- In the real example, an extreme value of $N_{00} = 4$ was recorded.

'Science Table' under H_{0F}

cup i	1	2	3	4	5	6	7	8
milk-first $Y_i(0)$	0	0	0	0	1	1	1	1
tea-first $Y_i(1)$	0	0	0	0	1	1	1	1

Lady tasting tea

- Fisher's sharp null H_{0F}: $\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$
- In the real example, an extreme value of $N_{00} = 4$ was recorded.

'Science Table' under H_{0F}

cup i	1	2	3	4	5	6	7	8
milk-first $Y_i(0)$	0	0	0	0	1	1	1	1
tea-first $Y_i(1)$	0	0	0	0	1	1	1	1

Lady tasting tea

- Fisher's sharp null H_{0F}: $\boxed{Y_i(0) = Y_i(1), \quad i = 1, \ldots, 8.}$
- In the real example, an extreme value of $N_{00} = 4$ was recorded.

'Science Table' under H_{0F}

cup i	1	2	3	4	5	6	7	8
milk-first $Y_i(0)$	0	0	0	0	1	1	1	1
tea-first $Y_i(1)$	0	0	0	0	1	1	1	1

☞ Under H_{0F}, that is to say among 4 cups of milk-first and 4 cups of tea-first, randomly draw 4 cups and all of them are milk-first. There is only $\binom{4}{4} = 1$ such occasion out of $\binom{8}{4} = 70$ possibilities, so the p-value is $1/70 = 0.014 < 0.05$.

RCT: Super-population inference

Average treatment effect (super-population)

However, SATE is a quantity for the n units in the study; it may **not** generalize someone outside the study.

Average treatment effect (super-population)

However, SATE is a quantity for the n units in the study; it may **not** generalize someone outside the study.

- **Super-population**: Imagine the units in the study are drawn randomly from a larger, super-population. That is, ☞ Science table is now random.

$$\boxed{(Y_i(0), Y_i(1))_{i=1}^n \text{ are iid copies of } (Y(1), Y(0)).}$$

Average treatment effect (super-population)

However, SATE is a quantity for the n units in the study; it may **not** generalize someone outside the study.

▶ **Super-population**: Imagine the units in the study are drawn randomly from a larger, super-population. That is, ☞ Science table is now random.

$$\boxed{(Y_i(0), Y_i(1))_{i=1}^n \text{ are iid copies of } (Y(1), Y(0)).}$$

It makes sense to infer the **ATE** (Average Treatment Effect) with respect to this population

$$\boxed{\mathrm{ATE} := \mathbb{E}\, Y(1) - \mathbb{E}\, Y(0) = \mathbb{E}\, Y_i(1) - \mathbb{E}\, Y_i(0).}$$

Average treatment effect (super-population)

However, SATE is a quantity for the n units in the study; it may **not** generalize someone outside the study.

- **Super-population**: Imagine the units in the study are drawn randomly from a larger, super-population. That is, ☞ Science table is now random.

$$\boxed{(Y_i(0), Y_i(1))_{i=1}^n \text{ are iid copies of } (Y(1), Y(0)).}$$

It makes sense to infer the **ATE** (Average Treatment Effect) with respect to this population

$$\boxed{\mathrm{ATE} := \mathbb{E}\, Y(1) - \mathbb{E}\, Y(0) = \mathbb{E}\, Y_i(1) - \mathbb{E}\, Y_i(0).}$$

- **Identification in RCT**: $\mathbb{E}[Y_i \mid A_i = 1] = \mathbb{E}[Y_i(1) \mid A_i = 1] = \mathbb{E}[Y_i(1)]$ by $A_i \perp\!\!\!\perp Y_i(0), Y_i(1)$.

$$\boxed{\mathrm{ATE} = \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0].}$$

☞ association $=$ causation.

ATE estimation

- There are many estimators for

$$\boxed{\mathrm{ATE} = \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0].}$$

ATE estimation

- There are many estimators for

$$\boxed{\mathrm{ATE} = \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0].}$$

- The simplest is still the **difference-in-means**

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0,$$

which is also the ANOVA with respect to a binary variable A.

ATE estimation

- There are many estimators for

$$\boxed{\mathrm{ATE} = \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0].}$$

- The simplest is still the **difference-in-means**

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0,$$

  which is also the ANOVA with respect to a binary variable A.
- Can we estimate ATE more accurately (i.e., efficiently) and in a robust way?

Yes, covariate adjustment!

In RCT, **baseline covariates** (i.e., those that are measured **before randomization** of treatment) can be used as precision variables (i.e., variables that are related to the outcome but not the treatment) to gain efficiency.

We can adjust for them to improve efficiency in a robust way.

Yes, covariate adjustment!

In RCT, **baseline covariates** (i.e., those that are measured **before randomization** of treatment) can be used as precision variables (i.e., variables that are related to the outcome but not the treatment) to gain efficiency.

We can adjust for them to improve efficiency in a robust way.

There are three approaches:

1. Regression adjustment
2. Post-stratification
3. Inverse probability weighting (IPW)

Regression adjustment: ANCOVA

[3] Cassel et al., 1976; Lin, 2013; Tsiatis et al., 2008; Yang and Tsiatis, 2001 among others

Regression adjustment: ANCOVA

- Historically, Fisher (1925) proposed to adjust for baseline covariates X_i using the analysis of covariance (ANCOVA) to improve efficiency. This remains a standard strategy in many fields.

[3] Cassel et al., 1976; Lin, 2013; Tsiatis et al., 2008; Yang and Tsiatis, 2001 among others

Regression adjustment: ANCOVA

- Historically, Fisher (1925) proposed to adjust for baseline covariates X_i using the analysis of covariance (ANCOVA) to improve efficiency. This remains a standard strategy in many fields.
- He suggested running a linear regression of Y_i on $(1, A_i, X_i)$. If we trust the linear model, the coefficient of A_i is an estimator the treatment effect.

[3] Cassel et al., 1976; Lin, 2013; Tsiatis et al., 2008; Yang and Tsiatis, 2001 among others

Regression adjustment: ANCOVA

- Historically, Fisher (1925) proposed to adjust for baseline covariates X_i using the analysis of covariance (ANCOVA) to improve efficiency. This remains a standard strategy in many fields.
- He suggested running a linear regression of Y_i on $(1, A_i, X_i)$. If we trust the linear model, the coefficient of A_i is an estimator the treatment effect.
- Later, it was discovered by many statisticians[3] that **in RCT**, the ANCOVA estimator correctly estimates ATE even linear model is wrong!
  - Regression adjustment does not change the estimand
  - For example, if Y_i and X_i do not have a linear relationship, or even if Y_i is binary, the ANCOVA procedure is still correct in RCT.
  - Without randomization, this magic disappears.

[3] Cassel et al., 1976; Lin, 2013; Tsiatis et al., 2008; Yang and Tsiatis, 2001 among others

Why does it work?

☞ If we fit OLS

$$\boxed{Y_i = \mu + \theta A_i + \beta X_i + \varepsilon_i,}$$

the coefficient of A_i is

$$\widehat{\theta} = \underbrace{\bar{Y}_1 - \bar{Y}_0}_{\text{difference in means}} - \widehat{\beta}\underbrace{(\bar{X}_1 - \bar{X}_0)}_{A \perp\!\!\!\perp X},$$

so, morally speaking, the second term is **asymptotically mean zero**.

Example: Iron intake on academic achievement

- Chong et al. (2016) conducted an RCT to evaluate the impact of iron intake on academic achievement.

Example: Iron intake on academic achievement

- Chong et al. (2016) conducted an RCT to evaluate the impact of iron intake on academic achievement.
- They recruited students of age 11 to 19 in a rural area of Cajamarca, Peru, where many adolescents suffer from iron deficiency.
- $n = 219$ students were assigned to one of the following three promotional videos:
  - Video 1: A soccer player is encouraging iron supplements to maximize energy;
  - Video 2: A physician is encouraging iron supplements for overall health;
  - Video 3: A dentist encouraging oral hygiene without mentioning iron at all (placebo)

Example: Iron intake on academic achievement

- Chong et al. (2016) conducted an RCT to evaluate the impact of iron intake on academic achievement.
- They recruited students of age 11 to 19 in a rural area of Cajamarca, Peru, where many adolescents suffer from iron deficiency.
- $n = 219$ students were assigned to one of the following three promotional videos:
  - Video 1: A soccer player is encouraging iron supplements to maximize energy;
  - Video 2: A physician is encouraging iron supplements for overall health;
  - Video 3: A dentist encouraging oral hygiene without mentioning iron at all (placebo)
- Goal: whether showing different promotional videos to the students can improve their academic performance through increased iron intake.

Example: Iron intake on academic achievement

- Chong et al. (2016) conducted an RCT to evaluate the impact of iron intake on academic achievement.
- They recruited students of age 11 to 19 in a rural area of Cajamarca, Peru, where many adolescents suffer from iron deficiency.
- $n = 219$ students were assigned to one of the following three promotional videos:
  - Video 1: A soccer player is encouraging iron supplements to maximize energy;
  - Video 2: A physician is encouraging iron supplements for overall health;
  - Video 3: A dentist encouraging oral hygiene without mentioning iron at all (placebo)
- Goal: whether showing different promotional videos to the students can improve their academic performance through increased iron intake.
- Primary outcome: academic achievement, measured by a standardized average of the student's grades in math, foreign language, social sciences, science, and communications in a semester.

Example: Iron intake on academic achievement

- Chong et al. (2016) conducted an RCT to evaluate the impact of iron intake on academic achievement.
- They recruited students of age 11 to 19 in a rural area of Cajamarca, Peru, where many adolescents suffer from iron deficiency.
- $n = 219$ students were assigned to one of the following three promotional videos:
  - Video 1: A soccer player is encouraging iron supplements to maximize energy;
  - Video 2: A physician is encouraging iron supplements for overall health;
  - Video 3: A dentist encouraging oral hygiene without mentioning iron at all (placebo)
- Goal: whether showing different promotional videos to the students can improve their academic performance through increased iron intake.
- Primary outcome: academic achievement, measured by a standardized average of the student's grades in math, foreign language, social sciences, science, and communications in a semester.
- Baseline covariate: baseline anemia status (1/0), grade (5 levels)

Example: Iron intake on academic achievement

Table: Estimate, standard error (SE), and p-value (reproduced from Ye et al., 2023)

		Physician versus placebo			Soccer star versus placebo		
Method	X	Estimate	SE	p-value	Estimate	SE	p-value
ANOVA		0.386	0.211	0.067	-0.068	0.205	0.739
ANCOVA	Grade, Anemia status	0.437	0.199	0.028	-0.085	0.201	0.672

- Point estimates are similar, suggesting that they both estimate ATE
- Including baseline grade and anemia status in the working model are useful to reduce the standard error
- Compared to the control group, the promotional video by the soccer player does not seem to have a positive effect on the academic achievement. In contrast, the video of the physician promoting iron supplements appears to have a positive effect.

Regression adjustment: ANHECOVA

- With reasonably large sample size, it is even better to include treatment-by-covariate interactions (with centered covariates).
- Specifically, we fit a linear model

$$\boxed{Y_i = \mu + \theta A_i + \beta(X_i - \bar{X}) + \delta A_i(X_i - \bar{X}) + \varepsilon_i,}$$

  where $\bar{X}$ is the average of X_i for all participants. We call this Analysis of Heterogeneous covariance (ANHECOVA) (Ye et al., 2023).
- The coefficient of A_i is

$$\bar{Y}_1 - \bar{Y}_0 - (\widehat{\beta} + \widehat{\delta})(\bar{X}_1 - \bar{X}) + \widehat{\beta}(\bar{X}_0 - \bar{X})$$

- The ANHECOVA estimator is **always no less efficient** than ANOVA even when the linear model is wrong.
- The R package `RobinCar` can calculate robust variance estimators.

Example: Iron intake on academic achievement

Table: Estimate, standard error (SE), and p-value (reproduced from Ye et al., 2023)

		Physician versus placebo			Soccer star versus placebo		
Method	X	Estimate	SE	p-value	Estimate	SE	p-value
ANOVA		0.386	0.211	0.067	-0.068	0.205	0.739
ANCOVA	Grade, Anemia status	0.437	0.199	0.028	-0.085	0.201	0.672
ANHECOVA	Grade, Anemia status	0.481	0.193	0.013	-0.046	0.195	0.815

- Point estimates are similar, suggesting that they all estimate the ATE
- Including baseline grade and anemia status in the working model are useful to reduce the standard error
- ANHECOVA has smaller SE than ANCOVA and ANOVA

Regression adjustment: beyond linear regression

Regression adjustment: beyond linear regression

- In general, we can use other non-linear models to estimate ATE, using a method called g-computation (i.e., standardization)

Regression adjustment: beyond linear regression

- In general, we can use other non-linear models to estimate ATE, using a method called g-computation (i.e., standardization)
  - e.g., logistic regression for binary outcomes, Poisson regression for count outcomes

Regression adjustment: beyond linear regression

- In general, we can use other non-linear models to estimate ATE, using a method called g-computation (i.e., standardization)
  - e.g., logistic regression for binary outcomes, Poisson regression for count outcomes
- Take logistic regression as an example, we may fit a logistic model

$$P(Y = 1 \mid A, X) = \frac{e^{\mu+\theta A+\beta X}}{1 + e^{\mu+\theta A+\beta X}},$$

then we can estimate ATE using

$$\widehat{\mathrm{ATE}}_{\mathrm{g}} = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}}{1 + e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}} - \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\widehat{\mu}+\widehat{\beta}X_i}}{1 + e^{\widehat{\mu}+\widehat{\beta}X_i}}.$$

Regression adjustment: beyond linear regression

- In general, we can use other non-linear models to estimate ATE, using a method called g-computation (i.e., standardization)
  - e.g., logistic regression for binary outcomes, Poisson regression for count outcomes
- Take logistic regression as an example, we may fit a logistic model

$$P(Y = 1 \mid A, X) = \frac{e^{\mu+\theta A+\beta X}}{1 + e^{\mu+\theta A+\beta X}},$$

then we can estimate ATE using

$$\widehat{\mathrm{ATE}}_{\mathrm{g}} = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}}{1 + e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}} - \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\widehat{\mu}+\widehat{\beta}X_i}}{1 + e^{\widehat{\mu}+\widehat{\beta}X_i}}.$$

- G-computation can robustly estimate ATE with linear/logistic/Poisson regression (Freedman, 2008; Guo & Basse, 2021).

☞ For other models (e.g., negative binomial), it may be biased when model is wrong but this can be fixed.

Why does it work?

The difference-in-means

$$\widehat{\mathrm{ATE}} = \bar{Y}_1 - \bar{Y}_0$$

is always **unbiased**. By the property of logistic regression, it holds that

$$\widehat{\mathrm{ATE}} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1} \sum_{i:A_i=1} \frac{e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}}{1+e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}} - \frac{1}{n_0} \sum_{i:A_i=0} \frac{e^{\widehat{\mu}+\widehat{\beta}X_i}}{1+e^{\widehat{\mu}+\widehat{\beta}X_i}}.$$

So the difference between the two is

$$\widehat{\mathrm{ATE}}_{\mathrm{g}} - \widehat{\mathrm{ATE}} = \left(\frac{1}{n} \sum_i \frac{e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}}{1+e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}} - \frac{1}{n_1} \sum_{i:A_i=1} \frac{e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}}{1+e^{\widehat{\mu}+\widehat{\theta}+\widehat{\beta}X_i}} \right) - \left(\frac{1}{n} \sum_i \frac{e^{\widehat{\mu}+\widehat{\beta}X_i}}{1+e^{\widehat{\mu}+\widehat{\beta}X_i}} - \frac{1}{n_0} \sum_{i:A_i=0} \frac{e^{\widehat{\mu}+\widehat{\beta}X_i}}{1+e^{\widehat{\mu}+\widehat{\beta}X_i}} \right),$$

which is **asymptotically mean zero** because $A \perp\!\!\!\perp X$.

▶ More on this on Day 2, Lecture 3.

Post-stratification

- **Discrete baseline covariates** can be adjusted by *post-stratification* (Fuller, 2011; Miratrix et al., 2013)
- Suppose we can create K strata, we calculate outcome mean difference separately within each stratum, then do an weighted average:

$$\frac{1}{n}\sum_{k=1}^{K}\frac{n_k}{n}(\bar{Y}_{1k}-\bar{Y}_{0k})$$

  where n_k is the size of stratum k, and $\bar{Y}_{1k}, \bar{Y}_{0k}$ are the mean outcomes under treatment and control, respectively, in stratum k.
- In the iron intake example, we can create $5 \times 2 = 10$ strata based on baseline grade (5 levels) and anemic indicator (2 levels), and use the post-stratification method.

Inverse probability weighting (IPW)

- Propensity score were introduced in Rosenbaum and Rubin (1983) as a tool to estimate the causal effect in observational studies.
- Propensity score is the probability of receiving treatment conditional on covariates $P(A = 1 \mid X)$.
- Although in RCT, the true propensity score is known by design, e.g., $P(A = 1 \mid X) = 0.5$. Nevertheless, we can still fit a model (e.g. logistic regression) of A on X.
- Intuition: in RCT, there may be chance imbalance in covariates, which is modeled using propensity score. Weighting by the propensity score leads to increased balance of these covariate and thus less variability and increased precision of the ATE estimate (Williamson et al., 2014).

Relationships among regression adjustment, post-stratification, and IPW in RCT

- Regression adjustment is a general-purpose and robust approach to improve efficiency in estimating ATE. Efficiency gains increase with model quality, but all models are valid.
- Post-stratification is the same as adjusting for the strata indicators using ANHECOVA regression adjustment.
- IPW (fit logistic regression for $A \sim X$) is asymptotically equivalent to adjusting for X using ANHECOVA regression adjustment (Shen et al., 2014). However, IPW may be unstable in small sample size if weights are close to 0 or 1.

In practice

- Choice of covariates: choosing key **baseline covariates** that can **predict** the outcome. For valid statistical inference, covariates should be pre-specified.

DO NOT adjust for post-randomization covariates!

- The previous discussion focuses on **simple randomization** (coin flipping). All results we discussed still hold for a slightly different scheme called **complete randomization** (e.g., randomly assign 50 out of 100 to treatment).
- There are more complicated assignment mechanisms to balance covariates between the treatment and controls, e.g., covariate-adaptive randomization (Ye et al., 2023). Covariate adjustment methods have been extended to theses assignment mechanisms.

Practical 1

References I

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, *63*(3), 615–620. https://doi.org/10.1093/biomet/63.3.615

Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy.. *Br Med J (Clin Res Ed)*, *292*(6524), 879–882.

Chong, A., Cohen, I., Field, E., Nakasone, E., & Torero, M. (2016). Iron deficiency and schooling attainment in peru. *American Economic Journal: Applied Economics*, *8*(4), 222–55. https://doi.org/10.1257/app.20140494

References II

- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh by Oliver; Boyd, 1st edition.
- Freedman, D. A. (2008). Randomization does not justify logistic regression. *Statistical Science*, *23*(2), 237–249. https://doi.org/10.1214/08-STS262
- Fuller, W. A. (2011). *Sampling statistics*. John Wiley & Sons.
- Guo, K., & Basse, G. (2021). The generalized oaxaca-blinder estimator. *Journal of the American Statistical Association*, *116*, 1–13. https://doi.org/10.1080/01621459.2021.1941053 doi: 10.1080/01621459.2021.1941053.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

References III

- Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.
- Knapp, H. R., & FitzGerald, G. A. (1989). The antihypertensive effects of fish oil. *New England Journal of Medicine*, *320*(16), 1037–1043.
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics*, *7*(1), 295–318. https://doi.org/10.1214/12-AOAS583
- Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(2), 369–396.

References IV

- Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.. *Statistical Science*, *5*(4), 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *6*(5), 688–701.
- Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *Journal of the American statistical association*, *75*(371), 591–593.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

References V

- Shen, C., Li, X., & Li, L. (2014). Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in medicine*, *33*(4), 555–568.
- Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, *27*(23), 4658–4677. https://doi.org/https://doi.org/10.1002/sim.3113 https://doi.org/10.1002/sim.3113.
- Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine*, *33*(5), 721–737.

References VI

Yang, L., & Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, *55*(4), 314–321. https://doi.org/10.1198/000313001753272466 doi: 10.1198/000313001753272466.

Ye, T., Shao, J., Yi, Y., & Zhao, Q. (2023). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, *117*(544), 2370–2382.